DISSERTATIONS IN
HEALTH
SCIENCES

Tuomo Kalliokoski

Accelerating Three-Dimensional Virtual Screening

New Software and Approaches for Computer-Aided Drug Discovery

Publications of the University of Eastern Finland
Dissertations in Health Sciences

UNIVERSITY OF
EASTERN FINLAND

TUOMO KALLIOKOSKI

Accelerating Three-Dimensional Virtual Screening

New Software and Approaches for Computer-Aided Drug Discovery

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland for public examination in Auditorium MET, Mediteknia building, University of Eastern Finland on Saturday 23rd of October 2010, at 12 noon.

Publications of the University of Eastern Finland
Dissertations in Health Sciences
22

School of Pharmacy
Faculty of Health Sciences
University of Eastern Finland
Kuopio
2010

Kopijyvä Oy
Kuopio 210

Distribution:

University of Eastern Finland
Kuopio Campus Library/Sales of Publications
P.O. Box 1627, FI-70211 Kuopio, FINLAND
http://www.uef.fi/kirjasto

ISBN: 978-952-61-0181-1
ISBN: 978-952-61-0182-8 (PDF)
ISSN: 1798-5706
ISSN: 1798-5714 (PDF)
ISSNL: 1798-5706

Authors address:

School of Pharmacy
Faculty of Health Sciences
University of Eastern Finland
P.O.Box 1627,
FI-70211 Kuopio, Finland
tkalliok@gmail.com

Supervisors:

Professor Antti Poso, PhD
School of Pharmacy
Faculty of Health Sciences
University of Eastern Finland
Kuopio, Finland

Toni Rönkkö, PhD
Kuopio, Finland

Reviewers:

Professor Gerhard Wolber, PhD
Computer-Aided Drug Design
Institute of Pharmacy
Freie Universität Berlin
Berlin, Germany

Richard Cramer, PhD
Tripos, A Certara Company
St. Louis, Missouri, USA

Opponent:

Professor Anders Karlén, PhD
Department of Medicinal Chemistry
Faculty of Pharmacy
Uppsala University
Uppsala, Sweden

Kalliokoski, Tuomo. Accelerating Three-Dimensional Virtual Screening: New Software and Approaches for Computer-Aided Drug Discovery. Publications of the University of Eastern Finland. Dissertations in Health Sciences, 22. 2010. 174 p.

ABSTRACT

Computers are routinely used in the drug discovery process. Virtual screening is defined to be the selection of compounds by evaluating their desirability in a computational model. Usually the predicted property is the bioactivity of a compound in an in vitro assay. Based on the classic Fischer lock and key-model, virtual screening is either ligand- or structure-based. In three-dimensional virtual screening, models of ligands and/or target proteins are used. In the ligand-based approach, the similarity of known ligands is used in the search for novel structures, whereas in structure-based virtual screening, compounds are docked into a protein model of the drug target. A consideration of all three-dimensions increases the computational expense of virtual screening considerably. The predictions need to be fast, as the commonly used data sets consist of many thousands, even millions of compounds. As virtual screening is a relatively new field of science, there is a need for novel methods and for the improvement of existing virtual screening protocols. In this thesis, a novel ligand-based virtual screening method called FieldChopper was developed. FieldChopper can be used when multiple, similarly binding active compounds are known. This novel method outperformed techniques based on single ligand similarities in a comparative study. In addition, the effects of tautomerism and protonation in structure-based virtual screening were studied with large data sets. It does appear that current methods are not yet accurate enough for separating between different tautomers and protonation sets and therefore the use of multiple forms of molecules in structure-based virtual screening is simply a waste of resources.

The effect of conformational analysis approaches on ligand-based virtual screening using shape-based overlay techniques was investigated. It was shown that with GPU computing and single conformation databases that even large databases can be screened on a regular desktop computer.

National Library of Medicine Classification: QU 34, QV 26.5, QV 744
Medical Subject Headings: Drug Discovery; Models, Molecular; Molecular Structure; Molecular Conformation; Ligands; Computer-Aided Design; Software; Software Design

Kalliokoski, Tuomo. Kolmiuloitteisen virtuaaliseulonnan nopeuttaminen: uusia ohjelmia ja lähestymistapoja tietokoneavusteiseen lääkeainesuunnitteluun. Itä-Suomen yliopiston julkaisuja. Terveystieteiden tiedekunnan väitöskirjat, 22. 2010. 174 p.

TIIVISTELMÄ

Virtuaaliseulonnalla tarkoitetaan yhdisteiden pisteyttämistä halutun ominaisuuden suhteen tietokoneen avulla. Yleensä ennustetaan yhdisteen biologista aktiivisuutta in vitro -kokeessa perustuen joko vertaamalla samankaltaisuutta tunnettuihin aktiivisiin yhdisteisiin (ns. ligandi-pohjainen virtuaaliseulonta) tai kohdeproteiiniin rakenteeseen telakoimalla (ns. rakenne-pohjainen virtuaaliseulonta).

Kolmi-ulotteisessa virtuaaliseulonnassa sekä pieniä molekyylejä että proteiineja käsitellään joustavina kolmiulotteisina rakenteina. Tämä lisää seulonnan laskennallista vaativuutta huomattavasti. Koska käsiteltäviä molekyylejä on yleensä tuhansia ja aikataulut lääkekehitysprojekteissa tiukkoja, on virtuaaliseulontamenetelmien oltava nopeita.

Tässä väitöskirjatyössä kehitettiin uusi ligandi-pohjainen nopea virtuaaliseulontamenetelmä FieldChopper, jota voidaan käyttää, kun tunnetaan useita samaan sitoutumistaskuun vaikuttavia yhdisteitä. Alustavien tulosten mukaan FieldChopper voi olla hyödyllinen molekyyliseula.

Uusia lähestymistapoja kehitettiin sekä rakenne- että ligandi-pohjaiseen virtuaaliselontaan. Ligandien tautomerian ja erilaisten protonaatiomuotojen vaikutusta molekyylitelakointiin on arvioitu aikaisemmin suureksi. Tässä tutkimuksessa havaittiin, että nykyisillä telakointiohjelmilla erot ovat luultua pienempiä ja seulontaprosessia voidaan yksinkertaistaa lisänopeuden saamiseksi. Viimeisessä osatyössä selvitettiin konformaatioanalyysin vaikutusta muotoon perustuvassa, ligandi-pohjaisessa virtuaaliseulonnassa. Yleisimmin käytetty menetelmä, jossa hakumolekyyliä käsitellään jäykkänä rakenteena ja tietokantamolekyylejä joustavina, ei tulosten mukaan ole välttämättä optimaalinen ratkaisu.

Yleinen suomalainen asiasanasto: lääkkeet; lääkeaineet; molekyylit; rakenne; kolmiulotteisuus; tietokoneavusteinen suunnittelu; ohjelmistokehitys

Quantity has a quality all of its own.
A remark usually attributed to Joseph Stalin (1878-1953)

Acknowledgements

The research was carried out in University of Eastern Finland during 2004-2010. I wish to thank my main supervisor Prof. Antti Poso for providing me solid funding and complete academic freedom. I also thank my second supervisor Dr. Toni Rönkkö for his altruistic attitude towards my work.

I am grateful to Prof. Anders Karlén for accepting the invitation to serve as my opponent. I wish to thank my esteemed reviewers Dr. Richard Cramer and Prof. Gerhard Wolber for their kind comments. Dr. Ewen McDonald is acknowledged for proof-reading the English in this dissertation.

While this thesis is mostly result of a solitary effort, I laud my co-authors Heikki Salo and Dr. Maija Lahtela-Kakkonen for smooth collaboration on the docking paper. I also wish to thank Dr. Pekka Tiikkainen for being good company during the notorious conference trips around Europe, Africa and Asia.

I wish to thank Dr. Sanni Matero for her opinion on some scientific issues and Heikki Käsnänen for digging up articles for me. CSC is acknowledged for computing resources.

This research was funded mainly by the Finnish Funding Agency for Technology (TEKES). Some parts are based on the research that I conducted while working for BCK/BCF. I wish to thank the Faculty of Health Sciences for the grant to finish my dissertation.

"Cheers!" to my friends Tero M, Jussi T and Antti S.
Finally, I want to thank my family: mother, father, Laura (thanks for the tip on phdcomics.com), Antti and my dear wife Riikka.

Tuomo Kalliokoski
Kuopio, September 2010

List of original publications

This doctoral thesis is based on the following original publications:

I Kalliokoski T, Rönkkö T, Poso A: FieldChopper, a New Tool for Automatic Model Generation and Virtual Screening Based on Molecular Fields. Journal of Chemical Information and Modeling **2008**, 48, 1131-1137.

II Kalliokoski T#, Salo HS#, Lahtela-Kakkonen M, Poso A: The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. Journal of Chemical Information and Modeling **2009**, 49, 2742-2748.

III Kalliokoski T, Rönkkö T, Poso A: Increasing the Throughput of Shape-Based Virtual Screening with GPU Processing and Single Conformation Databases. Molecular Informatics **2010**, 29, 293-296

#Equal contribution.

All the publications were adapted with the permission of copyright owners.

Contents

Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
ACE	Angiotensin-Converting Enzyme
ACHE	Acetylcholinesterase
ADA	Adenosine reductase
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
ALR2	Aldose Reductase
AMPC	AmpC beta-lactamase
AR	Androgen Receptor
AUC	Area Under Curve
BEDROC	Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic
CAS	Chemical Abstracts Service
CB1	Cannabinoid Receptor 1
CCDC	Cambridge Crystallographic Data Centre
CDK	Chemical Development Kit
CDK2	Cyclic-Dependent Kinase 2
CFF	Color Force Field
CMC	Comprehensive Medicinal Chemistry
CoMASA	Comparative Molecular Active Site Analysis
CoMFA	Comparative Molecular Field Analysis
COMSIA	Comparative Molecular Similarity Indices Analysis
COMT	Catecholamine-O-MethylTransferase
CORAL	Conformational Analysis, ROCS Alignment
COX1	Cyclooxygenase-1
COX2	Cyclooxygenase-2
CPU	Central Processing Unit
DHFR	Dihydrofolate Reductase
DNA	Deoxyribonucleic Acid
DUD	Database of Useful Decoys
EF	Enrichment Factor
EGFR	Epidermal Growth Factor Receptor
EMEA	European Medicines Agency

ER	Estrogen Receptor
FDA	Food and Drug Agency
FF	Force Field
FGFR1	Fibroblast Growth Factor Receptor Kinase
FXA	Factor Xa
GA	Genetic Algorithm
GART	Glycinamide Ribonucleotide Transformylase
GPB	Glycogen Phosphorylase Beta
GPCR	G-Protein Coupled Receptor
GPGPU	General Purpose Computing on Graphics Processing Units
GPL	GNU General Public License
GPU	Graphics Processing Unit
GR	Glucocorticoid Receptor
GRIND	GRID-Independent Descriptors
HERG	Human Ether-a-go-go
HIV	Human Immunodeficiency Virus
HIVPR	Human Immunodeficiency Virus Protease
HIVRT	Human Immunodeficiency Virus Reverse Transcriptase
HMGR	Hydromethylglutaryl-CoA Reductase
HPC	High-Performance Computing
HSP90	Human Heat Shock Protein 90
HTS	High Throughput Screening
INHA	Enoyl ACP Reductase
IUPAC	International Union of Pure and Applied Chemistry
LBVS	Ligand-Based Virtual Screening
LGA	Lamarckian Genetic Algorithm
MC	Monte Carlo
MD	Molecular Dynamics
MDDR	MDL Drug Data Report
MM	Molecular Mechanics
MMFF	Merck Molecular Force Field
MOE	Molecular Operating Environment
MR	Mineralocorticoid Receptor
MUV	Maximum Unbiased Validation

MW	Molecular Weight
NA	Neuraminidase
NAADP	Nicotinic Acid Adenine Dinucleotide Phosphate
NASA	National Aeronautics and Space Administration
NMR	Nuclear Magnetic Resonance
NPR	Normalized PMI Ratio
P38	P38 Mitogen Activated Protein
PARP	Poly (ADP-ribose) Polymerase
PBSA	Poisson-Boltzmann Solvent Accessible Surface Area
PDB	Protein DataBank
PDE5	Phosphodiesterase 5
PDGFRB	Platelet Derived Growth Factor Receptor Kinase
PLS	Partial Least Squares
PMI	Principal Moment of Inertia
PNP	Purine Nucleoside Phosphorylase
PPARG	Peroxisome Proliferator Activate Receptor Gamma
PR	Progesterone Receptor
QSAR	Quantitative Structure Activity Relationships
RAM	Random Access Memory
RECAP	Retrosynthetic Combinatorial Analysis Procedure
RIE	Rapid Initial Enrichment
RMSD	Root-Mean Square Deviation
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
ROCS	Rapid Overlay of Chemical Structures
RXRA	Retinoic X Receptor
SAHH	S-Adenosyl-Homocysteine Hydrolase
SBVS	Structure-Based Virtual Screening
SRC	Tyrosine Kinase SRC
SVD	Singular Value Decomposition
TK	Thymidine Kinase
USR	Ultrafast Shape Recognition
VEGFR2	Vascular Endothelial Growth Factor Receptor
VRAM	Video Random Access Memory
ZINC	ZINC Is Not Commercial

1 Introduction

Drug discovery and development is a long and expensive process, taking on average 12-15 years and costing 0.8-1.7 billion US dollars (DiMasi et al. 2003; Adams and Brantner 2006; Paul et al. 2010). The process is illustrated with a simplified flowchart on Figure 1.1. Initially, there must be a disease or symptom for which there is need for a novel drug. Since the drug development is so expensive, financial aspects must be also considered.

The first step in the actual development process is the drug target identification and validation. Novel drug targets are often identified via basic research by analyzing various molecular pathways. After a potential drug target has been identified, a cell-based assay needs to be developed in order to measure the biological activity of chemicals for the target.

Large chemical libraries have been created with combinatorial chemistry techniques. Natural sources like plants and bacteria provide also useful sources for drug molecules. The chemical libraries are evaluated for the drug target in a process called High-Throughput Screening (HTS), which is conducted by robots. An alternative to this rather laborious and expensive HTS method is to use computers for the prediction of biological activity (virtual HTS).

When a biologically active compound is identified from the initial screening, it is then tested in more sophisticated assays and thus selected as a lead molecule for the drug development process. The lead molecule is modified into a drug candidate by improving its pharmacokinetic and pharmacodynamic properties by synthesizing numerous analogues of the main compound. After animal testing, the molecule is tested on human volunteers. The most expensive parts in the drug development are the clinical experiments that demonstrate the

efficacy of the new drug molecule. Finally, drug must pass through a rigorous regulatory procedure before it can reach the market.

In addition to being an extremely expensive and long process, development of a new drug molecule is also risky as about nine out of ten candidate molecules fail to complete the course before they are accepted as drugs (Shah and Federoff 2009). The extra money spent in research and development has not increased the number of new chemical entities entering the market (Tralau-Stewart et al. 2009). Therefore, novel and preferably cheap methods are urgently needed by the pharmaceutical industry in order to boost its productivity (Paul et al. 2010). Computer-based methods are one such strategy. As the selection of a reasonable lead structure is a critical step for the successful development of a drug, the lead identification step has received considerable attention recently (Köppen 2009; Paul et al. 2010).

International Union of Pure and Applied Chemistry (IUPAC) have defined virtual screening (also called *in silico* screening) as the *"selection of compounds by evaluating their desirability in a computational model"* (Maclean et al. 1999). In this thesis, it is assumed that the number of compounds screened will be large, from thousands to millions of molecules (virtual high-throughput screening). The focus of this study has been in the development of novel rapid virtual screening software and the acceleration of current methods.

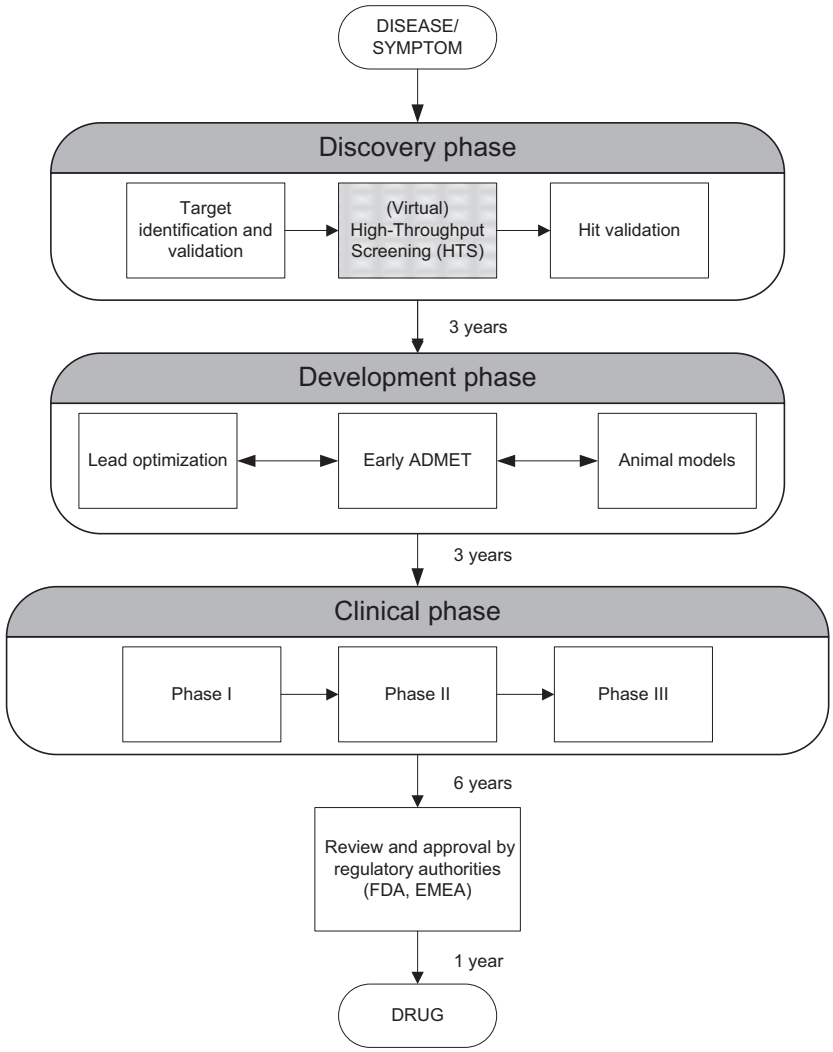


Figure 1.1: The drug development process (O'Driscoll 2004)

2 3D-Virtual Screening

Most drug molecules act via interactions with the various target proteins that exist in the organism, e.g. receptors, ion-channels, enzymes and transport proteins. In modern drug discovery, these targets are typically identified by the genetic analysis of the molecular pathways involved in a disease state (Zhu and Cuozzo 2009).

A compound that binds to a protein is called a ligand (Nelson and Cox 2005). It binds to the active site of the protein, which is complementary to the ligand in its steric and electrostatic properties. The specific nature of the binding can be compared to a key (ligand) and lock (protein), the model first proposed in the end of 19th century (Fischer 1894). The model is illustrated in Figure 2.1. It shows protein P, which has a triangular-shaped active site. The ligand A is a triangle, so it fits the active site and thus is able to bind to the protein. However, inactive compound B is a circle, so it will not fit.

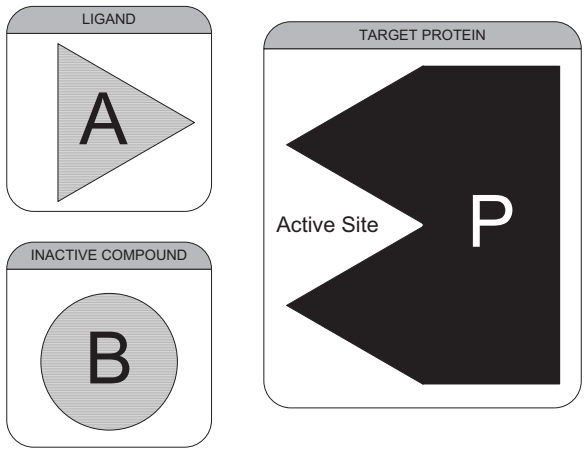


Figure 2.1: The Fischer's key and lock model. Ligand A fits the active site of target protein P, whereas inactive compound B does not.

Fischer's key and lock model is however overly simplistic, as it implies that ligands and proteins are rigid objects. The binding of a protein and ligand often induces a change in the binding site, which is called "induced fit" (Koshland 1958; Koshland 1996). Both the ligand and target protein adapt their conformation for better complementary (Jorgensen 1991; Nelson and Cox 2005). The induced fit theory is illustrated in Figure 2.2. At first, ligand C does not match the active site of protein P, but after undergoing induced fit, it is able to bind. There a better metaphor for the protein-ligand process is a hand (ligand) and glove (protein) instead of the rigid objects like a key and lock (Rao 2005).

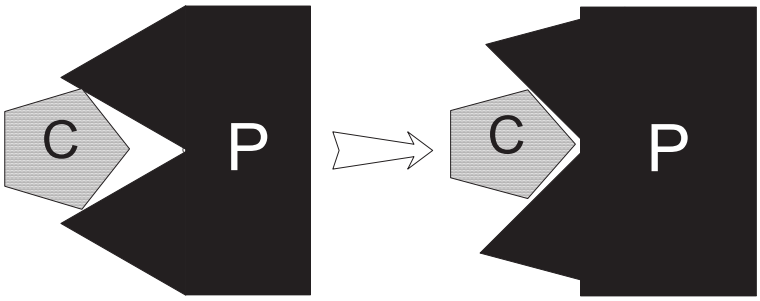


Figure 2.2: The Koshland induced fit theory. The ligand C induces a change in protein P's conformation, which makes the binding site more complementary to the ligand.

The induced fit theory does not explain all observed phenomena relating to the protein-ligand binding and recently a theory called "conformational selection" has emerged (Bosshard 2001; Boehr et al. 2009). It is illustrated on Figure 2.3 as a thermodynamic circle. Protein P can exist in two conformations in solution (P1 and P2). The binding conformation P2 pre-exists in solution before the ligand D is added. The kinetic constants K1 and K2 define, in addition to thermodynamic factors, if the binding of ligand D is via induced fit or conformational selection.

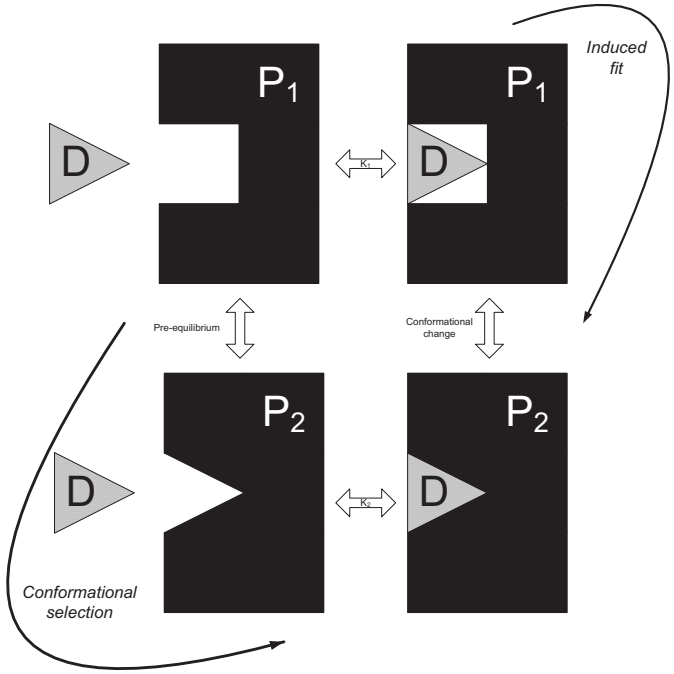


Figure 2.3: Conformational selection theory (adapted from Boehr 2009). The process can viewed as a simplified thermodynamic cycle.

In virtual screening, often the simplest Fischer theory of a rigid protein is considered due to computational demands (McInnes 2007). However, it has been recommended to be cautious and keeping the complexity of the biomolecular recognition process in mind in order not to over-interpret results from VS studies (Bissantz et al. 2010).

The overall view of the virtual screening process is shown in Figure 2.4. Virtual screening can be divided into two major strategies: ligand-based virtual screening (LBVS) and structure-based or target-based virtual screening (SBVS) (Rester 2008). Both approaches can be applied simultaneously provided that enough information is available. As with any modeling procedure, experimental data is required before predictions can be made. In LBVS, the information about other similarly bioactive compounds ("keys") is used, whereas in SBVS 3D-models of the target proteins ("locks") are utilized. The 3D-models of target proteins are either derived from X-ray

crystallography and Nuclear Magnetic Resonance (NMR) experiments or homology modeling, where the existing experimental data is used to build comparative models of proteins from their amino acid sequence. The chemical libraries that are screened are usually created using combinatorial chemistry techniques or they are built from natural products, such as chemicals extracted from plants. The result of a virtual screen is a hit list that is a prioritized list of compounds suitable for biological testing (*in vitro* evaluation). It is hoped that the top of the hit list contains more bioactive compounds than could be obtained from a random selection.

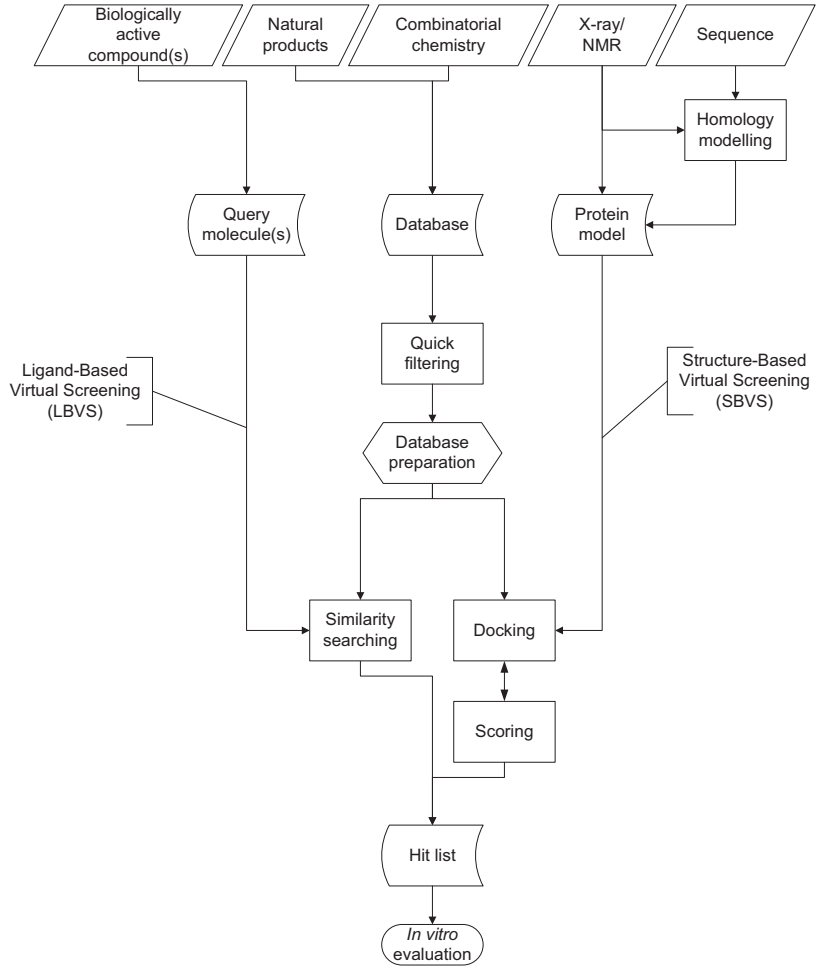


Figure 2.4: Overview of virtual screening

Virtual screening has many attractive qualities. The number of compounds that can be screened is much larger than with biological screening (Figure 2.5). In an academic setting, usually only virtual screening is possible due to the high costs related to HTS. The size of medicinal chemistry space is almost infinite, estimated to be 10^{60} molecules (Nicholls 2008; Köppen 2009). This is truly a staggering number: for comparison NASA Glenn Research Center has estimated that the total mass of all the stars in the observable universe is $3 * 10^{52}$ kg (NASA 2009). In a typical academic virtual screening study, one to ten million compounds will be screened for their biological potential and approximately 100-1000 molecules are tested *in vitro*. The large number of compounds to be screened means that virtual screening methods need to be fast in order to be truly useful for drug development.

As the price of high-performance computing has plummeted due to advances in both hardware and software, virtual screening costs only a small fraction of HTS. One can also predict bioactivity for molecules that can be readily made, but do not yet exist (virtual libraries). This strategy is often applied in the lead optimization phase.

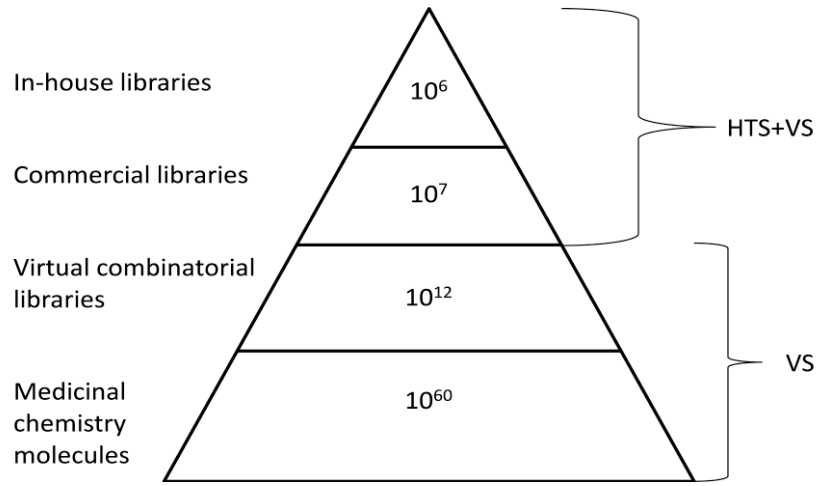


Figure 2.5: The numbers of molecules available from different sources (Köppen 2009).

Next, an overview for LBVS- and SBVS-methods is given with short introductions to high-performance computing (HPC) and database preparation. It should be noted that as there are thousands of different methods proposed in the literature (Todeschini and Consonni 2009); only some of the most used publicly available techniques are discussed. As the experimental part of this thesis is based on the acceleration of 3D-based virtual screening methods, the 2D-methods are only briefly mentioned. The different methods are viewed from practitioner's view and not discussed in algorithmic detail. This literature review hopes to consider most of the readily available 3D-virtual screening tools available on May 2010.

2.1 HIGH-PERFORMANCE COMPUTING IN VIRTUAL SCREENING

Since a large amount of data is processed in virtual screening, High-Performance Computing (HPC) is required for most real-life applications. HPC is based on massively parallel computing using supercomputers and computer clusters. Most algorithms used in virtual screening are trivial to parallelize by splitting the data into smaller pieces.

In the past, HPC required specialized and expensive hardware. Due to the availability of cheap multicore processors and free operating systems like Linux, this is no longer the case. Even a single person can build and maintain an HPC system with a small budget. The 56-CPU cluster located in University of Eastern Finland is an example of such a computer. It was built by the author without previous knowledge about how to set up such a system and it was ready for production in less than a month.

Recently, the power of graphical processing units (GPUs) has become readily available for scientific computing via general-purpose computing on graphics processing units (GPGPU). Originally developed for 3D graphics, the modern GPUs can perform parallel general-purpose calculations extremely fast

compared to regular CPUs. In addition to having high-performance, GPU hardware is also cheap and readily available as it is used for computer gaming. The downside of GPUs is that they are difficult to program and achieving high performance requires a thorough understanding of hardware details. However, the hard work pays off. The higher throughput combined with the cheap price of GPU-hardware allows the screening of extremely large databases with regular desktop computers instead of supercomputers (Giupponi et al. 2008).

2.2 LIGAND-BASED VIRTUAL SCREENING (LBVS)

Ligand-based virtual screening is based on “the similarity principle” that states that similar molecules tend to have similar biological properties (Eckert and Bajorath 2007). Molecular similarity is a subjective concept like beauty and molecules can be “similar” in many different ways (Maggiora and Shanmugasundaram 2004; Sheridan and Kearsley 2002). Although the term “ligand-based virtual screening” has only recently appeared in the literature, the idea is not new (Bohm et al. 2004). For instance, bioisosteric modifications are small modifications to molecules that are based on rules like "hydrogen can be changed to fluorine without losing the biological activity" (Patani and LaVoie 1996).

The aim of LBVS is usually scaffold hopping. LBVS methods can be also helpful in drug repurposing, where new targets and diseases are sought for existing drug molecules (Ashburn and Thor 2004). Scaffold hopping can be defined as the identification of isofunctional molecular structures with significantly different molecular backbones (Schneider et al. 1999). Although "scaffold hopping" is the most commonly used term (Fitzgerald et al. 2007), "leapfrogging" (Stanton et al. 1999), "scaffold searching" (Hert et al. 2006) and "lead hopping" (Cramer et al. 2004) have also been used to describe this strategy.

Some examples of different motivations and successes of scaffold hopping are presented in Table 2.1. Since peptides make

very poor drug molecules for various reasons (e.g. flexibility, proteolytic stability), it is desirable to replace the peptidic scaffold of a bioactive molecule (Bohm et al. 2004). Several successful cases have been published where peptides have been substituted by other structures (Ripka and Rich 1998).

Poor absorption, distribution, metabolism, excretion and toxicity (ADMET) properties may also be the reasons for scaffold hopping (Rush et al. 2005). If a lipophilic scaffold can be changed to a more polar one, this will increase the solubility of the compound, which is often a major problem in contemporary drug discovery programs (Lipinski 2000; Paul et al. 2010).

Scaffold hopping has also been used for intellectual property issues. When a "breakthrough-drug" is introduced onto the market by a pharmaceutical company, its competitors try to develop molecules with similar biological but a dissimilar chemical structure ("me-too" drugs).

Table 2.1: Some examples of different motivations and successes of scaffold hopping

Target	**Issues with original ligand**	**Reference**
Histamine H3	Chemical/metabolic instability, hERG-channel inhibition	Lau et al. 2006
Activator protein-1 (AP-1)	Peptide	Tsuchida et al. 2006
HIV TAR RNA	Pharmacokinetics (multiple charges, size)	Renner et al. 2005
Cholecystokinin-2 (CCK2)	High levels of biliary elimination	Low et al. 2005
HIV-1 reverse transcriptase	Metabolic instability	O'Meara et al. 2007
Sphingosine 1-phosphate-3 Receptor (S1P3)	Potency and selectivity	Koide et al. 2007
Glycogen Synthase Kinase-3 (GSK-3)	Not suitable for further optimization	Naerum et al. 2002
5-lipoxygenase (5-LO)	Lack of selectivity	Franke et al. 2007
Tau protein aggregation	Toxicity, cell penetration	Larbig et al. 2007
Histamine H4	Very short half-life	Smits et al. 2008
Glutamate racemase (Murl)	Restricted antibacterial spectrum	Breault et al. 2008
Trypanothione Reductase	Potency and selectivity	Perez-Pineiro et al. 2009
Kinases	Undesirable thiourea linker	Tasler et al. 2009

Even though popular, scaffold hopping is an ill-defined term and highly subjective concept (Brown and Jacoby 2006; Bohm et al. 2004). There are various definitions for a scaffold (Roberts et al. 2000; Xu 2002; Jenkins et al. 2004; Krier et al. 2006; Barker et al. 2006; Wilkens et al. 2005). One of the first definitions of scaffold was made in a patent by Markush (Markush 1924; Brown and Jacoby 2006). It defined a set of dye chemicals*:"…dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline"*. Markush structures are used by drug companies to protect chemical series around a promising molecule, even though not all of the structures are even possible to synthesize, let alone having any biological

effect whatsoever. Therefore, Markush structures are more of a legal tool than a scientific concept (Brown and Jacoby 2006).
The most commonly used scaffold concept is based on the work of Bemis and Murcko, where they analyzed the properties of known drugs using the Comprehensive Medicinal Chemistry (CMC) database (Bemis and Murcko 1996). These scaffolds are sometimes referred to as "Murcko's scaffolds" or "molecular frameworks" (Krier et al. 2006; Lipkus et al. 2008). The classification is based on a hierarchical description of molecules, illustrated in Figures 2.6 and 2.7. A molecule consists of a scaffold that has side chains, whereas a scaffold consists of a ring system and linkers. Murcko's scaffolds have the obvious pitfall that only cyclic scaffolds that were included in the CMC datasets can be detected. Recently, Lipkus and co-workers analyzed the scaffolds found in the CAS registry using a similar approach to Bemis and Murcko. They found out that half of the 24 million organic compounds in CAS could be described by only 143 scaffolds. Other general classifications are the maximum common substructures (McGregor and Willett 1981), maximum rigid fragments (Su et al. 2001) and RECAP fragments (Lewell et al. 1998). The problem of scaffold definition has not yet been satisfactorily solved and it will be discussed also in Chapter 3.3.

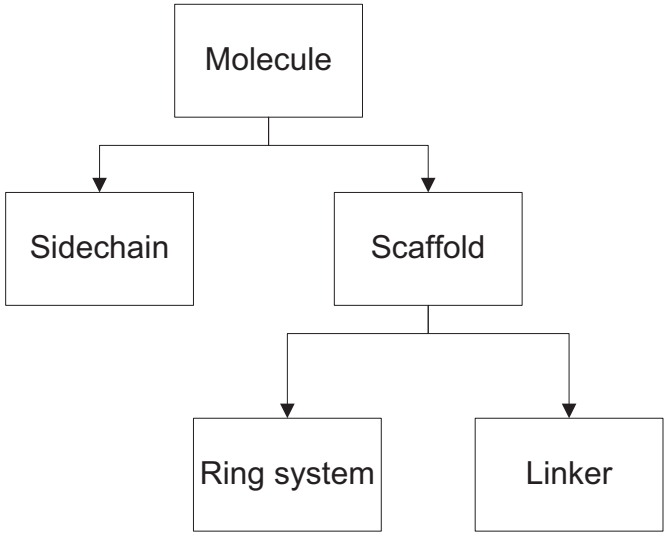


Figure 2.6: Hierarchical description of molecules (adapted and modified from Bemis and Murcko 1996).

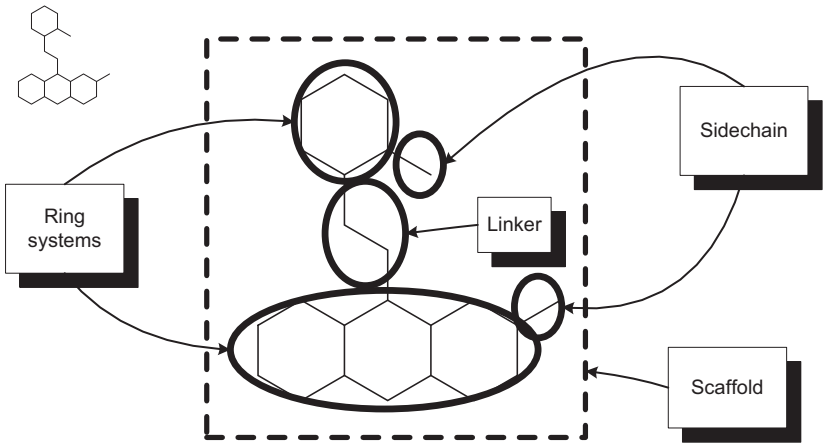


Figure 2.7: Detecting scaffolds using concepts of ring systems, linkers and side chains (adapted and modified from Bemis and Murcko 1996).

One example scaffold hopping is shown in Figure 2.8, where there are the two similarly bioactive compounds that have completely different scaffolds. Hypothesis for their similar activity is based on matching three-dimensional shape of the molecules.

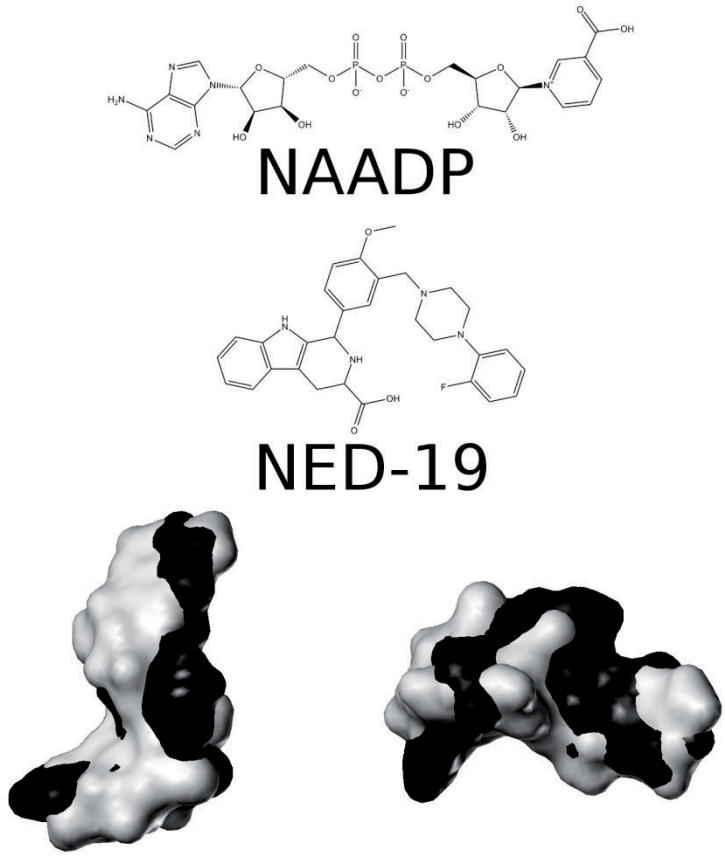


Figure 2.8: Example of scaffold hopping. NAADP and NED-19 have similar bioactivity even though their scaffolds are completely different. Both molecules are similar in their 3D surfaces (black and white shapes) (Connolly 1983). Analysis is based on the findings of Naylor and co-workers (Naylor et al. 2009).

There are many similarity methods which have been developed for LBVS. Some of the commonly used approaches are presented in Table 2.2. For a comprehensive listing, the interested reader is referred to the recent book of Todeschini and Consonni (Todeschini and Consonni 2009).

Table 2.2: Some of the commonly used LBVS approaches (Eckert and Bajorath 2007; Poptodorov et al. 2006; Rester 2008)

Method	Example(s)	Approach
0D/1D descriptors	Atom counts	Generated from molecular graph
2D fingerprints	MACCS	Quantitative comparison of bit strings
3D descriptors	UNITY3D, NPR, USR, ESshape3D, GRIND	Generated using intramolecular distances
Pharmacophores	Catalyst	Common features of active molecules are detected
3D similarity based on pair-wise alignment	ROCS/EON, BRUTUS, ShaEP, FlexS	Comparison of superimposed molecules

2.2.1 0D-2D descriptors

The simplest ways of describing molecules are the one- and two-dimensional descriptors like the number of carbon atoms or molecular indexes based on graph theory (Hall and Kier 1991). These kinds of descriptors are easy to calculate with modeling tools like MOE (Chemical Computing Group). Despite their simplicity, they have been shown to be surprisingly effective in virtual screening. For example, in the study of Bender and Glen, a large data set of over 100000 compounds containing 11 activity classes was screened using the number of atoms per chemical element as a molecular descriptor (Bender and Glen 2005). Enrichment factors over random selection of around four were achieved and also diverse chemical scaffolds were detected in the active group.

The commonly used two-dimensional fingerprints are binary strings that encode the presence or absence of sub-structural fragments (Willett 2006). A set of chemical features is defined and then a bit is set to either zero (0) or one (1), depending on whether the substructure exists in the molecule or not. A fingerprint is a long bit string, which can also be expressed as an

integer. An example of a two-dimensional fingerprint is shown in Figure 2.9, which illustrates the MACCS-fingerprint for citalopram.

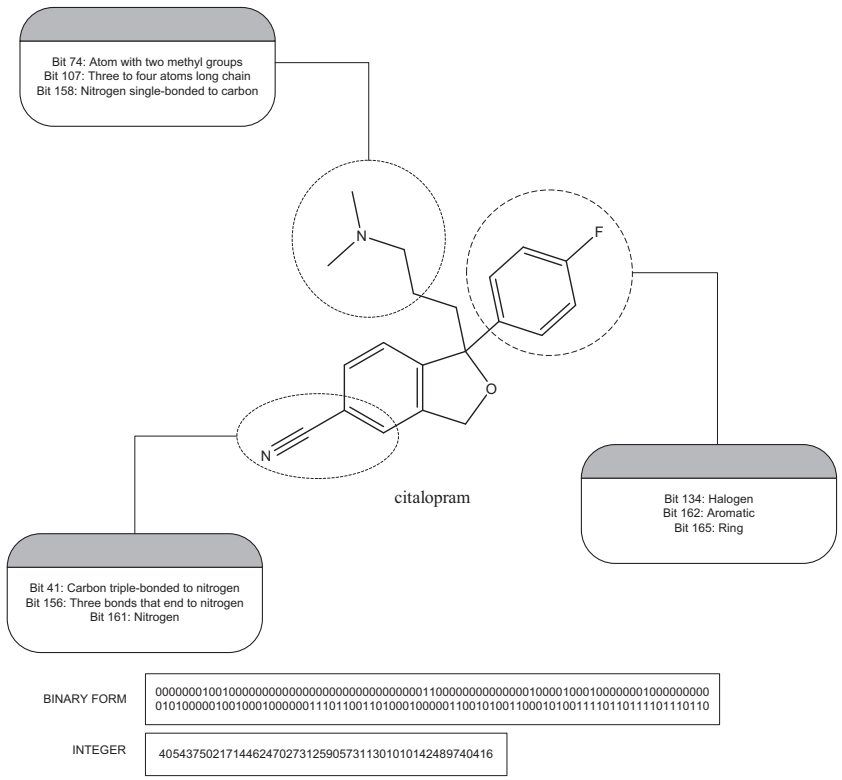


Figure 2.9: Example of 2D fingerprint: MACCS structural keys for citalopram. For clarity, only some of the defined bits are shown. Fingerprint generated with OpenBabel 2.2.3 (Guha et al. 2006).

There are many 2D fingerprint methods available but it is scientifically difficult to accept any 2D fingerprint as a golden standard (Eckert and Bajorath 2007). The most commonly used fingerprints are UNITY from Tripos Inc (for example, Schuffenhauer et al. 2000; Raymond and Willett 2002; Holliday et al. 2003), MACCS/MDL Keys from MDL (for example, Koehler et al. 1999; Wild and Blankley 2000; Durant et al. 2002) and Daylight from Daylight Chemical Information Systems (for example, Kogej et al. 2006; Capelli et al. 2006; Stiefl and Zaliani 2006).

Usually a single fingerprint is compared with a database in order to retrieve similar compounds. However, it is also possible to form fusion fingerprints based on multiple fingerprints from several query molecules (Willett 2006).

There are numerous similarity coefficients for measuring the similarity between two 2D fingerprints (Todeschini and Consonni 2009). For example, Holliday and co-workers have compared 22 different coefficients with UNITY fingerprints (Holliday et al. 2003). The most widely used similarity coefficient was formulated by Tanimoto in 1957 (Willett et al. 1998):

$$T_c = \frac{c}{a + b - c}$$

where a is the number of bits set to one in the first fingerprint, b is the number of bits set to one in the second fingerprint and c is the number of bits set at identical positions in both fingerprints. The Tanimoto coefficient is between 0 (completely different) and 1 (completely similar).

There has been much debate on the approriate value for the Tanimoto coefficient in similarity searching with some workers attempting to use a fixed threshold (0.85) for all screens (Patterson et al. 1996; Matter 1997). However, this has been proven to be an inefficient approach (Martin et al. 2002). It seems that compound class specific effects strongly affect fingerprint calculations and proper thresholds have to be set on a case-by-case basis (Godden et al. 2005).

Even though 2D fingerprints have proved to be useful tools in drug discovery projects, they suffer from several drawbacks (Raymond and Willett 2002). For example, a single atom change in a ring structure may change the fingerprint from being nearly similar to almost completely different. Moreover, as is shown in Figure 2.8, two compounds that have very different topologies can nonetheless adopt a similar orientation and thus could have similar biological effects. Since this thesis is about 3D-virtual screening, the reader interested in 2D methods is referred to a comprehensive review on the subject (Willett 2006).

2.2.2 3D descriptors

3D fingerprints (also known as pharmacophore keys) encode 3D relationships in a molecule as a bit string (Matter 1997; Good et al. 2004a; Leach 2001). An example of such an algorithm is the UNITY 3D fingerprints (Tripos 2009). The basic idea is presented in Figure 2.10, where there are two different conformations of disulfiram. The combinations of features are enumerated with the distances between them. In a 3D-fingerprint, each bit encodes a distance between specific groups. For example, bit 0 could be ”donor-donor with distance 2-2.5” and bit 1 ”donor-donor with distance 2.5-3” etc. The number of features used in combinations varies from two up to nine (Martin and Hoeffel 2000). However, the size of a fingerprint increases rapidly with the number of features used.

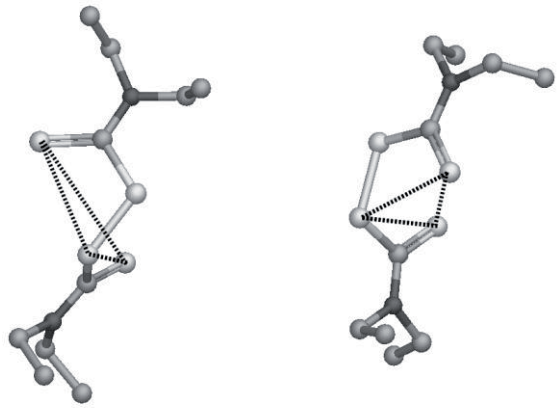


Figure 2.10: Two conformations of disulfiram. The three-point pharmacophoric feature is different in the two conformations. Conformations generated with OPLS_2005 force field implemented in MacroModel (Schrödinger Inc). Image created with Maestro (Schrödinger Inc).

The basic problem with 3D fingerprints (and with other 3D methods as well) is of course conformational sampling, since the number of possible conformations *N* increases very rapidly with

the number of rotating bonds n (so called combinatorial explosion) (Kitchen et al. 2004; Boström and Grant 2008):

$$N = \left(\frac{360}{m}\right)^n$$

where m is the size of the rotational angle in degrees. For example, a molecule with six rotatable bonds has 2985984 possible conformations with a rotational angle increment of 30 degrees. It is therefore not possible to use all possible conformations in similarity calculations for most molecules. The problem of conformational analysis will be discussed more in Chapter 2.5.3.

Shape-based descriptors encode the shape of the molecule into numbers. The shape complementarness of the ligand to the active site is a prerequisite for the drug action, so several approaches for describing this important feature have been developed (Putta and Beroza 2007). If compared to the 3D fingerprints, which describe molecules as sets of atoms, the shape-based descriptors consider molecules as volumes and surfaces (Nicholls et al. 2010). The normalized ratio of principal moments of inertia (PMI) is an example of a shape-based descriptor (Sauer and Schwarz 2003). PMIs are easily calculated with molecular modeling packages like SYBYL and they have been widely used to assess molecular shape, geometry and conformation. Three principal components are calculated and assigned by ascending order to I_1, I_2 and I_3. These are normalized by dividing the lower values I_1 and I_2 by I_3. The normalization eliminates the dependency on the size of the molecules. These normalized PMI ratios (NPRs) fulfill the following relation due to the intrinsic characteristic of the inertia tensor:

$$\frac{I_2}{I_3} \geq \max\left(\frac{I_1}{I_3}, 1 - \frac{I_1}{I_3}\right)$$

Therefore, the resulting plot against each other is an isosceles triangle onto which all molecules can be placed. The three

corners correspond to archetype shapes of spheres, disks and rods (Figure 2.11). Compounds are mapped to different parts of the triangle according to their shape.

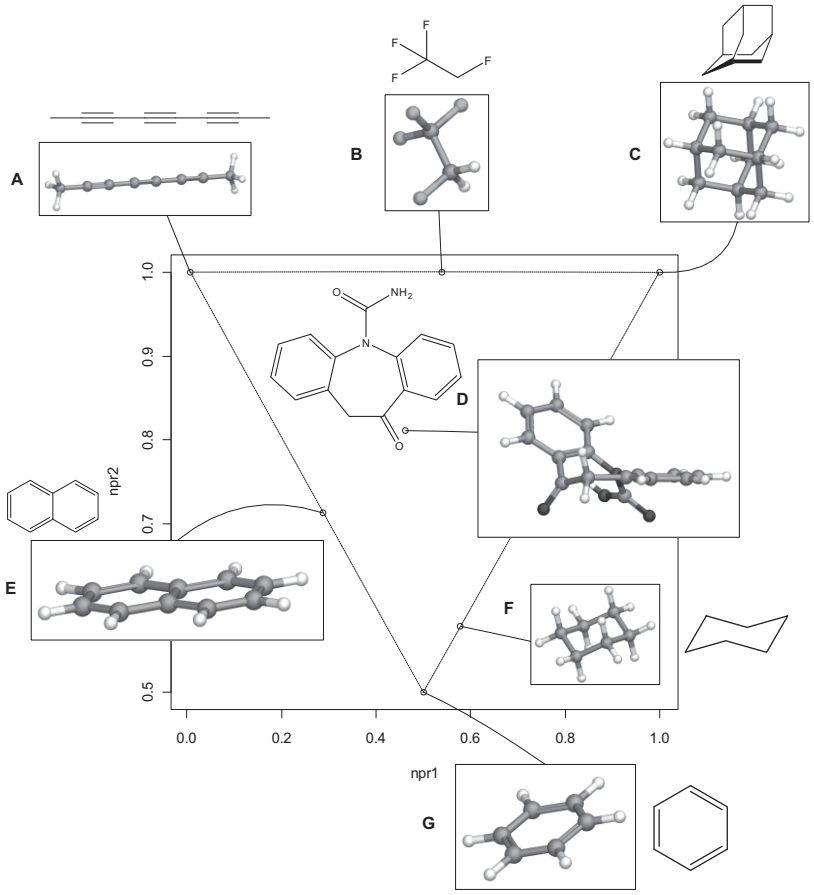


Figure 2.11: Some examples of normalized PMI ratios and corresponding structures (adapted from Sauer and Schwarz 2003). Conformations were calculated with CORINA (Molecular Networks GmBH) and NPRs were calculated with MOE (Chemical Computing Group). Molecules were visualized with Maestro (Schrödinger Inc)

However, this methodology has obvious serious flaws. For example, completely different molecules like methane and fullerene would be classified as similar because they adopt a similar sphere-like conformation.

The shape-based descriptors that are more relevant to the virtual screening are ESShape3D (implemented in MOE by Chemical Computing Group) and Ultrafast Shape Recognition (USR). ESshape3D is formed by first measuring the Euclidean distance between all pairs of the heavy atoms in the molecule (Henry A, personal communication, Jan 12th, 2010; Ballester et al. 2009). Then, the eigenvalues are calculated from this distance matrix. The signed square roots of these eigenvalues are smoothened with a Gaussian function and stored in a histogram with 122 bins containing values between -30 and 30. The similarity between two ESshape3D descriptors is calculated from the distance between the values for each of the histogram bins. For example, if we assume that there are two molecules with distances *A* and *B* (three bins instead of the 122 used for clarity):

$$A = [10,20,30]$$
$$B = [30,20,10]$$

The difference between *A* and *B* would be [-20, 0, 20]. The distance *D* is the square root of the sum of the squared differences (√800). The similarity *S* is calculated from the distance *D* (*S*=0.714):

$$S = \frac{2}{2 + \frac{D}{1000}}$$

USR is based on atomic intramolecular distances from four molecular locations that are used to form a 12 element vector (Ballester and Richards 2007a; Ballester and Richards 2007b). It is one order of magnitude faster to calculate than the ESShape3D descriptor (Ballester et al. 2009). The similarity between two descriptors *A* and *B* is calculated from:

$$S = \frac{1}{1 + \frac{1}{12}\sum_{i=1}^{12} |A_i - B_i|}$$

USR is implemented in Chemical Development Kit (CDK) by Guha (DistanceMoment class). A command line user interface was written by the author to conduct a virtual screening and to calculate descriptors with USR (available via http://www.uku.fi/~tkalliok/usr).

Since both steric and electrostatic properties are important in protein-ligand complementary, the accuracy of shape-based descriptors for virtual screening is limited (Nicholls et al. 2010). Examples of a descriptor that encodes both shape and electrostatic properties are Grid-Independent descriptors (GRIND) (Pastor et al. 2000). The descriptors are derived from a collection of GRID molecular interaction fields computed using different chemical probes based on the work of Goodford (Goodford 1985). These fields are then discretized by finding "the hot spots" of interactions. The relative position of "hot spots" is then encoded into descriptors called correlograms. Principal component analysis of the correlograms is then used for the similarity calculations. The algorithm for the calculation of GRIND descriptors has evolved over the years (Fontaine et al. 2004; Durán et al. 2008; Durán et al. 2009). The most recent version of the method is implemented in Pentacle (available from Molecular Discovery Ltd).

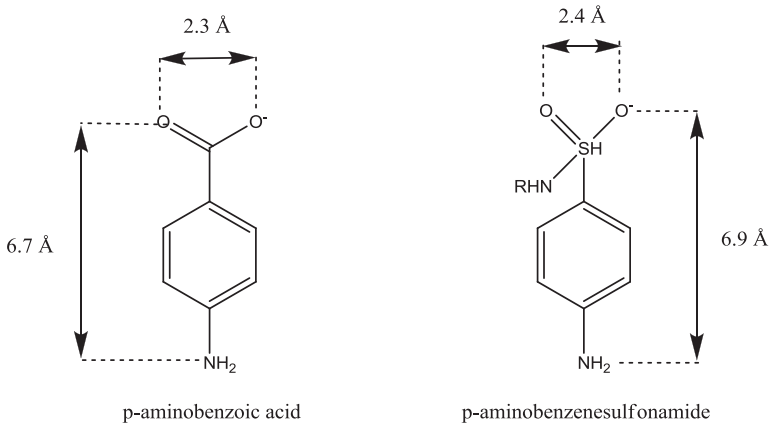
2.2.3 Pharmacophores

The term pharmacophore was introduced by Paul Ehrlich in 1909 (Ehrlich 1909; Triballeau et al. 2006). The modern IUPAC definition dates from 1998: *"A pharmacophore 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."* (Wermuth et al. 1998)

A pharmacophore is an abstract concept that describes the interaction capability of either one or a group of compounds toward a drug target instead of a real molecule or real association of functional groups (Wermuth 2006). Pharmacophores can be also constructed from protein models (Wolber and Langer 2005). The main advantage of

pharmacophore methods is that it is possible to find very diverse compounds. The early pharmacophores were constructed manually in the 1940's with the knowledge of the bond lengths and the van der Waals radii of atoms (Figure 2.12). Such simple constraints could be used as a crude filtering criterion for large set of compounds to weed out clearly unsuitable molecules.

Sulfonamides and p-aminobenzoic acid (Woods and Fildes 1940):



Estrogen pharmacophore (Dodds and Lawson 1938; Schueler 1946):

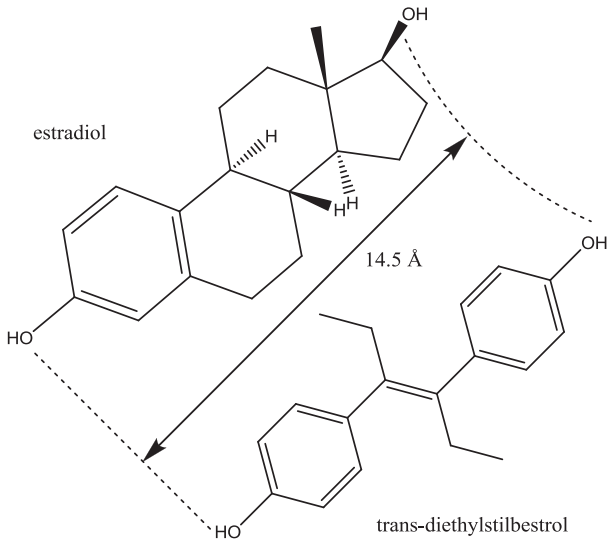


Figure 2.12: Two early pharmacophores with example molecules (adapted from Wermuth 2006)

The pioneers in the modern computational pharmacophore identification are Marshall and co-workers who developed the so-called Active Analog Approach in the 1970's (Marshall et al. 1979). The core algorithm is illustrated in Figure 2.13. The number of conformations for a flexible molecule is reduced by

the geometry of a rigid reference molecule. Pharmacophores are then derived from these alignments. This approach forms the basis of many existing automated pharmacophore generation methods (van Drie 2004; Poptodorov et al. 2006).

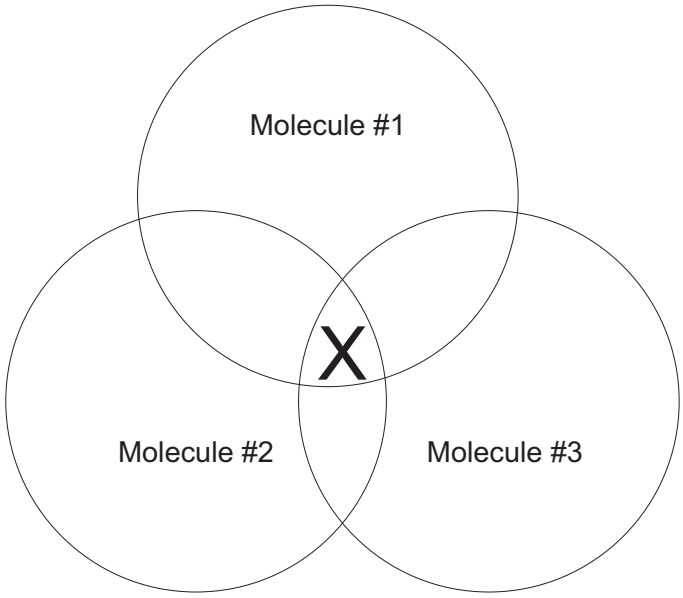


Figure 2.13: The Active Analogue approach by Marshall and co-workers (adapted from van Drie 2004). The circles represent the available conformational space. The intersection X is the area from which the common pharmacophores will be found.

The work flow for general pharmacophore modeling is presented in Figure 2.14. Several compounds that have similar biological activities are needed to form a hypothesis. Some methods also allow incorporation of activity data. An important assumption is that all compounds in the pharmacophore have a similar binding mode and thus they can be superimposed. After compounds are superimposed, common features of the molecules can be detected. A pharmacophore can almost always be generated, but it must be validated by using an external data set before use. After a reasonable pharmacophore is formed, the virtual screening step itself is fast.

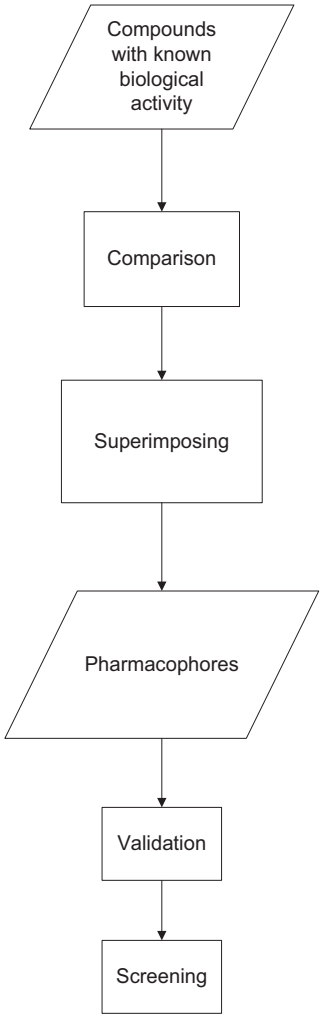


Figure 2.14: General pharmacophore modeling workflow (adapted and modified from Poptodorov et al. 2006)

To some extent, pharmacophores have been neglected and the development of new methods has been extremely slow (Langer and Hoffmann 2006). This might be due to the strong emphasis on SBVS in recent years. Since SBVS methods have not been as successful as was originally anticipated, there has been increasing interest in using the pharmacophore approach (Kolb et al. 2009; Leach et al. 2010).

Geometry- and feature-based pharmacophore methods usually consider compounds as sets of connected features like

hydrophobic and H-bond acceptors/donors (Poptodorov et al. 2006). These features are important for selective binding of drug molecules as they describe hydrogen bonding, electrostatics and hydrophobic interactions. As a practical example of these kinds of chemical function definitions, Greene et al. (1994) proposed a set of features based on atom types (Figure 2.15). A similar set is used in most modern pharmacophore programs. This set, originally implemented in Catalyst software, is not completely satisfactory, as for example it will describe incorrectly both oxygen atoms in esters as "hydrogen bond acceptors".

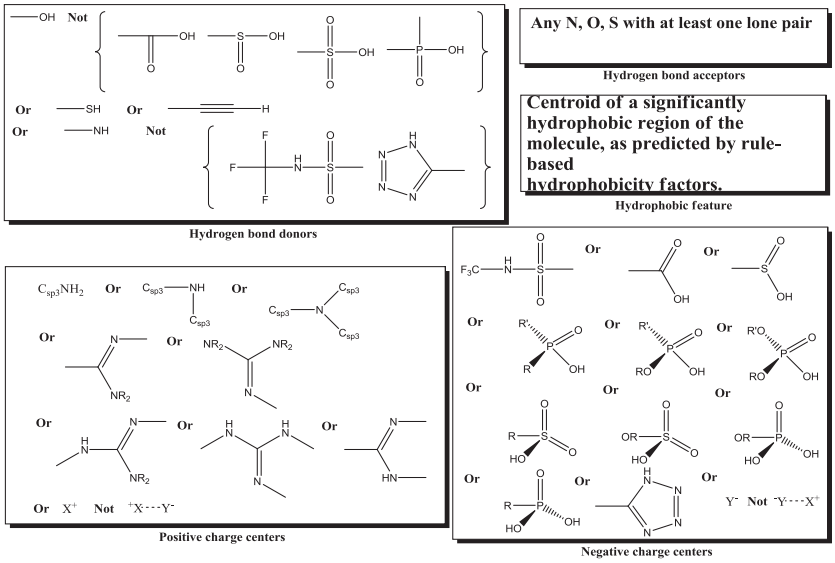


Figure 2.15: Pharmacophore features as proposed by Greene et al. (1994).

The most widely used geometry- and feature-based pharmacophore elucidation method is Catalyst from Accelrys, which is currently a part of the Discovery Studio package (some recent applications of the methodology are presented in Table 2.3). Catalyst is an integrated set of algorithms for conformation generation (ConFirm), molecular superimposition (HipHop), pharmacophore generation (HypoGen) and database searching (Info). HipHop and HypoGen provide two approaches for automatic pharmacophore generation. HipHop identifies pharmacophores by aligning the chemical features of active

molecules (Barnum et al. 1996). Each conformation of each molecule is used as a reference for alignment and every configuration is scored. HypoGen is designed to correlate structure and activity for automatic pharmacophore generation (Poptodorov et al. 2006), working in three steps. In the first step, common features are detected between the two most active compounds. In the second step, those features that are common between active and inactive compounds are removed from the pharmacophore. The last step is an optimization phase where simulated annealing is used to improve the predictive power of the pharmacophore. An exclusion volume can be added to HypoGen pharmacophores to filter out too large molecules from the search. For a more detailed description of Catalyst and other feature-based pharmacophore generation methods, the reader is referred to recent review articles (Poptodorov et al. 2006; Leach et al. 2010).

Table 2.3: Some recent examples of Catalyst-based pharmacophores for virtual screening

Target	**Reference**
Human adenosine kinase	Bhutoria and Ghoshal 2010
Bovine viral diarrhea virus	Tonelli et al. 2010
11β-hydroxysteroid dehydrogenase 1	Rollinger et al. 2010; Schuster et al. 2006
Phosphodiesterase type-5	Chen 2010
Peroxisome proliferator-activated receptor gamma	Fakhrudin et al 2010; Markt et al. 2008; Markt et al. 2007
5-Lipoxygenase	Aparoy et al. 2010
Human tyrosyl-DNA phosphodiesterase	Weidlich et al. 2010
Plasmodium falciparum dihydrofolate reductase	Adane et al. 2009
Nuclear factor-kappa B	Tsai et al. 2009
ZAP-70	Sanam et al. 2009
Transforming growth factor-β Type I Receptor (ALK5)	Ren et al. 2009
Caspase-3	Laksmi et al. 2009
Various cancer cell lines	Chiang et al. 2009
Aromatase	Neves et al. 2009
Monoamine Oxidase B	Boppana et al. 2009
Spleen tyrosine kinase	Xie et al. 2009
Aurora B kinase	Wang et al. 2009
Cannabinoid receptor 2	Markt et al. 2009
Raf-1 kinase	Li et al. 2009
Glycogen Synthase Kinase 3beta	Vadivelan et al. 2009

3D-Quantitative Structure Activity Relationships (QSAR) methods can be considered as field-based automatic pharmacophore generation methods (Poptodorov et al. 2006). The most frequently used 3D-QSAR method is Comparative Molecular Field Analysis (CoMFA) devised by Cramer and co-workers (Cramer et al. 1988). Other widely used 3D-QSAR methods are CoMSIA (Klebe et al. 1994) and GRID/GOLPE (Cruciani and Watson 1994).

Even though there are hundreds of CoMFA studies published (PubMed lists over 900 citations with keyword "CoMFA"), most of these studies are mostly retrospective analyses and have very

little predictive value that could be used in prospective virtual screening of new biologically active molecules (Doweyko 2004). Also, the superimposing step is a major limitation for virtual screening applications, as the compounds to be screened need to have a common scaffold to permit automatic alignment (Hillebrecht and Klebe 2008). It could be therefore concluded that CoMFA is more a tool for lead optimization rather than a virtual screening method for large databases. There is also Topomer-CoMFA available, which is easier to use than the traditional CoMFA (Cramer 2003).

CoMFA has however inspired various other field-based virtual screening methods, including FieldChopper described in this thesis, and it has been used in conjunction with other methods to find novel compounds (for an example, see Zhang et al. 2007), so it serves as an example of a field-based virtual screening method. An outline of the method is presented in Figure 2.16. The molecular field is presented as a lattice. Compounds are superimposed and their activity values, steric and electrostatic potentials are recorded in the QSAR table. From this table, an equation is derived with Partial Least Squares (PLS) data analysis method (Wold et al. 1984). This equation can then be used in the prediction of activity for compounds outside the model. Although the basic idea is rather straightforward, the correct use of the method is difficult, as the results are critically dependent on conformation and superimposition of the compounds. Furthermore, the chemical parameters used to generate fields and the statistical evaluation methods have a large influence on the models.

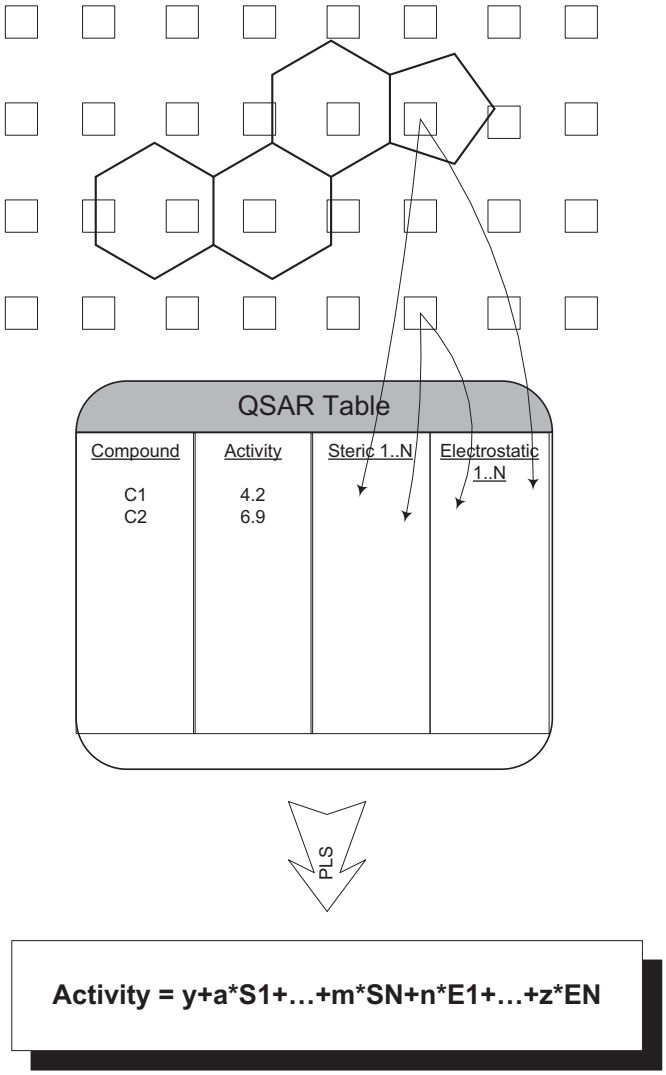


Figure 2.16: Comparative Molecular Field Analysis (CoMFA) (adapted from Cramer et al. 1988).

2.2.4 3D similarity based on pair-wise alignment

In pharmacophore methods, a set of compounds is compared in order to find common features, which are then matched to a set of compounds in a database. One can also try to match the whole query molecule to database molecules by aligning them in a pair-wise manner. It is easier to find a reasonable alignment for a pair of molecules than for diverse set of molecules. Some of

the recent virtual screening successes using pair-wise alignment have been listed on Table 2.4.

Table 2.4: Some recent examples of alignment-based virtual screening

Target	**Reference**
ZipA-FtsZ protein-protein interaction	Rush et al. 2005
Biological role of NAADP	Naylor et al. 2009
CB1 receptor	Boström et al. 2007
Metabotropic glutamate receptor 2	Tresadern et al. 2010
Nipah virus envelope protein	Niedermeier et al. 2009
Melanin Concentrating Hormone	Oyarzabal et al. 2009
Neurotensin NTS1 receptor	Fan et al. 2008
Androgen receptor	Trump et al. 2007
γ-secretase	Gundersen et al. 2005

The problem of molecular alignment is a complex issue due to the degrees of freedom involved and is comprehensively discussed in a recent doctoral dissertation (Rönkkö 2009). From a practical point of view, there are several high-throughput molecular alignment methods publicly available (Table 2.5).

Table 2.5: High-throughput small molecule alignment-based similarity methods suitable for virtual screening. License abbreviations: O=Open Source, F=Free, FA=Free for Academic use and C=Commercial.

Program	**Reference**	**Lic.**	**Website**
ROCS	Grant et al. 1995	FA	www.eyesopen.com
EON	Nicholls et al. 2004	FA	www.eyesopen.com
PAPER	Haque and Pande 2009	O	simtk.org/home/paper
BRUTUS	Rönkkö et al. 2006	C	www.visipoint.fi
ShaEP	Vainio et al. 2009	F	users.abo.fi/mivainio/shaep
FlexS	Lemmen et al. 1998b	C	www.biosolveit.de

The most widely used molecular alignment method for virtual screening is Rapid Overlay of Chemical Structures (ROCS) from OpenEye Scientific Software (Grant et al. 1995; Kirchmair et al. 2009). In this method, molecules are superimposed with a smooth Gaussian function representing the molecular volume. ROCS optimizes this function by rigidly translating and rotating

the molecule with respect to the query molecule. In the original version, optimization started with four initial orientations, but the current version has some undisclosed improvements for generating the starting positions (Nicholls et al. 2004).
The similarity *S* between two molecules A and B is calculated from the volumes of the molecules (ShapeTanimoto score):

$$S = \frac{O_{AB}}{O_{AA} + O_{BB} - O_{AB}}$$

where O_{AA} is the volume of molecule A, O_{BB} is the volume of molecule B and O_{AB} is the overlapping volume between these molecules.
In addition to shape, ROCS considers the electrostatic properties of molecules via "Color Force Field" (CFF). The CFF is based on 1D-atom rules that define chemically important areas and has rules about how such centers should interact. The hydrogen-bonding rules are derived from the crystal survey of the Cambridge Structural Database (Mills and Dean 1996). Usually, the Color Tanimoto value is used in combination with ShapeTanimoto (TanimotoCombo).
EON from OpenEye Scientific Software is a more sophisticated electrostatic similarity method (Nicholls et al. 2004). It creates electrostatic fields around a pair of aligned molecules and calculates the similarity between the two fields. ShapeTanimoto is often combined with this electrostatic score.
Recently an open-source, GPU-accelerated version of ROCS was developed called PAPER (Haque and Pande 2009). In addition to having the advantage of being free software, PAPER is over one order of magnitude faster on a single desktop PC than the commercial ROCS package due to the high performance of the GPU computing. It does not however have the CFF implemented.
BRUTUS is an automated computer program for rigid-body molecular superimposition which considers molecular fields (Rönkkö et al. 2006; Rönkkö 2009). It is based on rotating and translating molecular fields instead of the molecules. This

removes the need to re-calculate the fields during the optimization process. In addition, BRUTUS uses a simple interpolation algorithm for estimating the energy between grid points and this allows the use of coarse energy fields for alignment. These factors make BRUTUS fast enough to screen large databases. The similarity between the molecular volume and electrostatic field is computed separately with the Hodgkin index (Hodgkin and Richards 1987). The volumic similarity S_V and electrostatic similarity S_e thus computed are combined to the total similarity S:

$$S = wS_v + (1 - w)S_e$$

where w is a weighting factor (0.5 is used by default).

ShaEP is an alignment algorithm based on shape and electrostatic potential (Vainio et al. 2009). First, initial alignments are produced by a matching algorithm on graphs that represent the electrostatic potential of the molecule. Then, the alignments are optimized by maximization of the volumic overlap using Gaussian functions. It uses a similar total similarity score as BRUTUS.

Incremental construction is implemented in FlexS (Lemmen et al. 1998b). Initially, molecules are partitioned into fragments and an anchor fragment for the incremental construction procedure is either selected by the program or manually by the user. The anchor fragment is then placed on the reference ligand and the remaining fragments are added iteratively. For virtual screening, FlexS uses the RIGFIT algorithm to place the base-fragment onto the reference molecule (Lemmen et al. 1998a). FlexS also uses the Hodgkin index in order to compute the similarity between two molecules.

2.3 STRUCTURE-BASED VIRTUAL SCREENING (SBVS)

Structure-Based Virtual Screening (SBVS) is usually based on molecular docking (Kitchen et al. 2004; Moitessieri et al. 2008).

In molecular docking, a small molecule is fitted into the protein model's active site. As an example, oseltamvir was docked into N1 neuraminidase with Glide (Figure 2.17). The binding mode predicted by docking calculation is remarkably close to the one observed in the crystal structure. Several successful virtual screening studies have been reported for various proteins (Table 2.6).

Table 2.6: Some recent examples of successful SBVS studies

Protein	**Reference**
β2-adrenergic receptor	Kolb et al. 2009
A2A-adenosine receptor	Carlsson et al. 2010
Histamine H4-receptor	Kiss et al. 2008
JAK2 kinase	Kiss et al. 2009
JAK3 kinase	Kim et al. 2010
Ubiquitin C-terminal hydrolase L1	Mitsui et al. 2010
FGFR1	Ravindranathan et al. 2010
Death-associated protein kinase 1	Okamoto et al. 2009
B2 subunit of V-ATPase	Ostrov et al. 2009
Falcipain-2	Li et al. 2009
Hepatitis C virus nonstructural protein 3	Chen et al. 2009
Thermolysin	Khan et al. 2009
D-alanine:D-alanine ligase	Kovac et al. 2008
Mycobacterium tuberculosis APSR	Cosconati et al. 2008
SecA ATPase	Li et al. 2008
SRC	Lee et al. 2009
ATP-dependent Mur ligases MurD and MurF	Turk et al. 2009
Sarco/endoplasmic reticulum calcium ATPase	Deye et al. 2009
SARS-3CL(pro)	Mukherjee et al. 2008
E. coli enoyl-ACP-reductase	Yao et al. 2010
HSP90	Hong et al. 2009
Cdc25B phosphatase	Park et al. 2009
Beta-secretase	Xu et al. 2009
Extracellular signal-regulated kinase 2	Park et al. 2008
CK1 delta	Cozza et al. 2008
Phosphatase of regenerating liver-3	Park et al. 2008
S. pneumoniae VicR/K	Li et al. 2009
PPARG	Salam et al. 2008
V. harveyi LuxP	Li et al. 2008
VHR Phosphatase	Park et al. 2008
ErmC Methyltransferase	Feder et al. 2008
Protein tyrosine phosphatase 1B	Park et al. 2009
Insulin-regulated aminopeptidase	Albiston et al. 2008
Human PEBP4	Qiu et al. 2010
H. pylori UPPS	Kuo et al. 2008
Mammalian proteasome 20S	Basse et al. 2010
Aurora kinase A	Coumar et al. 2009

Aldo-keto reductase 1C1	Brozic et al. 2009
Chemokine receptor CCR4	Bayry et al. 2008

The aim of docking is to predict the structure of the complex [P+L] = [PL] under equilibrium conditions in water and to estimate the Gibbs energy of binding ΔG. ΔG can be described by the equation $\Delta G=\Delta H-T\Delta S$ (Whitesides and Krishnamurthy 2005). Enthalpic factors (ΔH) include steric and electrostatic complementary, hydrogen-bonding, protein strain and also ligand strain, if the ligand is flexible. Desolvation, rotational and translational entropy are important factors in entropy (ΔS).

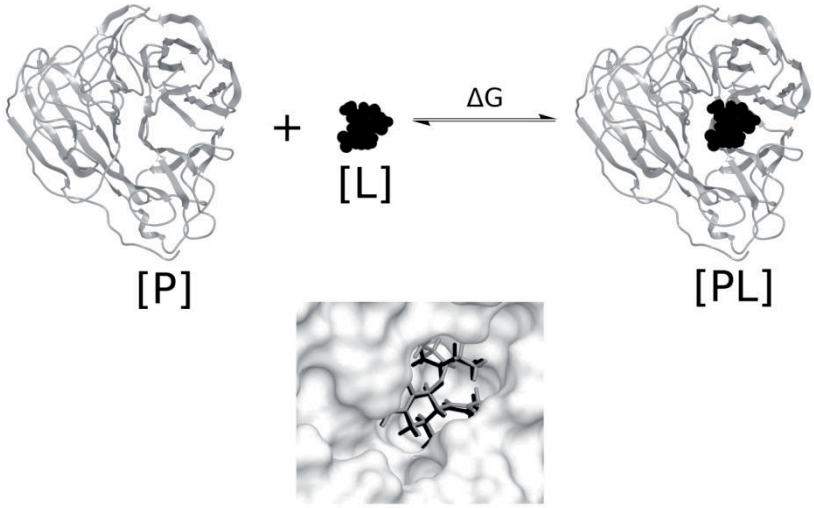


Figure 2.17: The concept of molecular docking. Ligand (L) is docked to the protein (P) to form a protein-ligand complex (PL). PDB-Complex 2HU4 (Russell et al. 2006) formed by oseltamivir and N1 neuraminidase is closely reproduced by a docking program. The docking programs best scored solution is shown in black and that experimentally observed in gray. Images created with GLIDE and Maestro (Schrödinger Inc).

There are two major components in a docking program: a search algorithm that produces relevant binding modes (poses), and a scoring function, which should be able to predict the affinity of the docked compound to the protein i.e. estimate ΔG. The

searching problem has been basically solved, but the scoring problem persists.

Due to the number of atoms involved in the protein-ligand interaction, the problem is extremely complex. A typical approximation in order to speed up the calculations is to use a rigid protein and torsionally flexible ligand instead of a fully flexible protein and ligand. Even with these simplifications, molecular docking is still a time consuming process compared to the ligand-based virtual screening methods.

There are over 60 docking programs and more than 30 scoring functions described in the literature (Moitessier et al. 2008; Viji et al. 2009). However, only a fraction of the proposed methods are readily available for virtual screening studies. The currently available software is listed on Table 2.7 (the references on the table are to the latest versions of the programs). Most of the docking software is commercial, so licensing might represent a rate-limiting step in a virtual screening study even though supercomputing capability is available. Commonly used docking methods include AutoDock, DOCK, LigandFit, FlexX, FRED, GLIDE and GOLD.

Table 2.7: Currently available docking programs (adapted and modified from Moitessier et al. 2008). The most commonly used programs are shown in italics (McInnes 2007). License abbreviations: O=Open Source, F=Free, FA=Free for Academic use and C=Commercial.

Program	**Reference**	**Search algorithm**	**Lic.**	**Website**
AutoDock	Morris et al. 2009	LGA	O	autodock.scripps.edu
AutoDock Vina	Trott and Olson 2010	Gradient optimization based	FA	vina.scripps.edu
CDOCKER of Discovery-Studio	Wu et al. 2003	MD/simulated-annealing	C	www.accelrys.com
CHARMM (Galgor)	Vieth et al. 1998	GA/MC	C	www.charmm.org
DOCK	Lang et al. 2009	Sphere matching	FA	dock.compbio.ucsf.edu
DockIt	-	Sphere matching	C	www.metaphorics.com
eHiTs	Zsoldos et al. 2006	Rigid docking of fragments	C	www.simbiosys.com
DAIM-SEED-FFLD	Kolb and Caflisch 2006; Majeux et al. 2001; Budin et al. 2001	Docking of fragments	F	www.biochem-caflisch.uzh.ch
FITTED	Corbeil et al. 2007	GA	FA	www.fitted.ca
LibDock of Discovery-Studio	Diller et al. 2001	Pregenerated ligand conformations with gradient-based optimization	C	www.accelrys.com
LigandFit of Discovery-Studio	Venkatachalam et al. 2003	Shape-based method with MC	C	www.accelrys.com
FlexX	Rarey et al. 1996	Incremental construction	C	www.biosolveit.de
FlipDock	Zhao and Sanner 2007	GA	FA	flipdock.scripps.edu
FRED	McGann et al. 2003	Gaussian docking function	FA	www.eyesopen.com

FTDock	Gabb et al. 1997	Fourier correlation algorithm	O	bmm.cancerresearchuk.org
GEMDOCK	Yang and Chen 2004	GA	F	gemdock.life.nctu.edu.tw
GlamDock	Tietze and Apostolakis 2007	MC	FA	www.chil2.de
GLIDE	Friesner et al. 2004	MC	C	www.schrodinger.com
GOLD	Verdonk et al. 2003	GA	C	www.ccdc.cam.ac.uk
HADDOCK	de Vries et al. 2007	Mainly for protein-protein docking	FA	www.nmr.chem.uu.nl
MolDock	Thomsen and Christensen 2006	Heuristic search	C	www.molegro.com
PatchDOCK	Schneidman-Duhovny et al. 2005	Shape complementary	FA	bioinfo3d.cs.tau.ac.il
PLANTS	Korb et al. 2009	Ant colony optimisation	FA	www.tcd.uni-konstanz.de
ICM	Abagyan et al. 1994	MC	C	www.molsoft.com
rDock	Morley and Afshar 2004	MC	F	www.ysbl.york.ac.uk/rDock
ROSETTA-LIGAND	Meiler and Baker 2006	MC	FA	www.rosettacommonds.org
SLIDE	Schnecke and Kuhn 2000	Incremental construction	FA	www.bch.msu.edu/~kuhn
SODOCK	Chen et al 2007	Particle swarm optimization for AutoDock 3.05	F	iclab.life.nctu.edu.tw/sodock
Surflex-Dock	Jain 2007	Incremental construction	FA	www.biopharmics.com
MOE-Dock	-	MC	C	www.chemcomp.com

In the following section, the different types of searching algorithms and scoring functions are described briefly. Computationally intensive simulation-based methods such as free energy perturbation or Molecular Mechanics/Poisson-Boltzmann-Surface-Area (MM/PBSA) are not discussed here, as they cannot be used in high-throughput virtual screening (at least not yet).

2.3.1 Searching algorithms

Different approaches for the docking pose generation have been applied. The methods can be roughly divided into three main types: rigid-body, incremental construction and stochastic algorithms.

Rigid-body docking

Rigid-body docking algorithms use either single or multi-conformation databases to account for ligand flexibility (Moitessier et al. 2008). The molecules are fitted into the binding sites of proteins by shape complementary or by interaction matching algorithms. These are the fastest structure-based virtual screening methods, but their accuracy may be limited due to the fact that ligand conformation is not refined at the binding site. They are also highly dependent on the method used to create the conformations (see Chapter 2.5). It has been suggested that these kinds of methods should be used only for initial screening of large libraries.

An example of rigid-body docking software is FRED from OpenEye Scientific Software (McGann et al. 2003). It uses pre-generated multi-conformation database as its input. First, all possible poses of the ligand around the active site are enumerated for each of the conformations. These poses are then filtered, based on the volume of the active site. The remaining poses are then scored with a scoring function. FRED is one the fastest docking program currently available, as it requires just a few seconds per ligand. Its performance was also comparable to

the more computationally intensive GLIDE in an 11 target study (McGaughey et al. 2007).

Incremental construction docking

There are also docking programs based on incremental construction algorithms. These programs build up the ligand in the active site. First, the ligand is fragmented and one fragment is selected as the anchor fragment. The anchor fragment is then rigidly docked into the active site and the other fragments are connected with the knowledge of preferred conformations.
FlexX is an example of a program that is based on incremental construction (Rarey et al. 1996). It uses a pose-clustering technique similar to those used in pattern recognition.

Stochastic docking

Both multi-conformation and incremental construction docking algorithms are deterministic. There are also stochastic docking algorithms available that have a random element in them. Therefore, they do not usually produce exactly the same results in every run. The two most widely used stochastic approaches are Monte Carlo methods and genetic algorithms.
Monte Carlo methods are based on repeated random sampling. The ligand to be docked is randomly rotated and translated one parameter at the time. The modified conformation is then evaluated by a scoring function. If the new conformation has a lower energy than the previous one, it is kept. The process is repeated until a satisfactory pose has been generated. A typical example of Monte Carlo docking method is ICM (Abagyan et al. 1994). GLIDE has also a Monte Carlo element, as final poses from hierarchical filtering are generated by the Monte Carlo method (Friesner et al. 2004).
Genetic algorithms are based on Darwin's theory of evolution (Moitessier et al. 2008). A docking pose is stored in a data structure called a "chromosome", which is made up of numbers called "genes" that store each translational angle, rotation and

translation of the ligand. Chromosomes then evolve through a process of reproduction and are altered by genetic operators like mutation and crossover. The next generation is then selected by the survival of the fittest, where the two lowest energy chromosomes are kept. Lamarckian Genetic Algorithm (LGA) is a modification of the genetic algorithm that is used in AutoDock (Morris et al. 2009). LGA is hybrid method which contains an adaptive global optimizer with a local search. The local search method uses a random search optimization, which is allowed to change the chromosome of the global optimizer. The use of LGA instead of the regular genetic algorithm increases the performance of AutoDock (Morris et al. 1998).

2.3.2 Scoring functions

The scoring functions can be roughly divided into force field-, empirical and knowledge-based (Kitchen et al. 2004; Moitessier et al. 2008). Scoring functions can be also hybrids of molecular mechanics and empirical terms (Table 2.8).

Table 2.8: Currently available scoring functions (adapted and modified from Moitessier et al. 2008).

Scoring function	**Reference**	**Type**	**Software or website**
ChemScore	Eldrige et al. 1997	Empirical	GOLD, FRED, CScore
ShapeGauss	McGann et al. 2003	Empirical	FRED
ChemGauss3	-	Empirical	FRED
eHiTs	Zsoldos et al. 2006	Empirical	eHiTs
GlideScore	Friesner et al. 2004	Empirical	Glide
FlexX	Rarey et al. 1996	Empirical	FlexX
Hammerhead	Pham and Jain 2006	Empirical	Surflex-Dock, DiscoveryStudio
LigScore	Krammer et al. 2005	Empirical	DiscoveryStudio
PLP	Verkivker et al. 2000	Empirical	DiscoveryStudio, FRED, DockIt
RankScore	Moitessier et al. 2006	Empirical/FF	FITTED
ScreenScore	Stahl and Rarey 2001	Empirical/consensus	FRED
SLIDE SCORE	Schnecke and Kuhn 2000	Empirical	SLIDE
X-Score	Wang et al. 2003	Empirical/consensus	sw16.im.med.umich.edu/software/xtool
AutoDock4 SF	Huey et al. 2007	FF/Empirical	AutoDock, SODOCK
DockScore	Meng et al. 1992	FF	DOCK, Cscore
Zou GB/SA Score	Liu et al. 2004	GB/SA	DOCK
GoldScore	Jones et al. 1997	FF	GOLD, Cscore
HADDOCK	van Dijk et al. 2006	FF	HADDOCK
ICM	Abagyan et al. 1994	FF	ICM

DrugScoreCSD	Velec et al. 2005	Knowledge based	pc1664.pharmazie.uni-marburg.de/drugscore
DrugScorePDB	Gohlke et al. 2000	Knowledge based	pc1664.pharmazie.uni-marburg.de/drugscore
M-Score	Yang et al. 2006	Knowledge based	sw16.im.med.umich.edu/lab/members/chaoyie
PMF	Muegge 2006	Knowledge based	DiscoveryStudio, DockIt, Cscore
Zapbind	Grant et al. 2001	Empirical/PBSA	FRED, DOCK
Astex Scoring Potential	Mooij and Verdonk 2005	Knowledge based	GOLD
Cscore	Clark et al. 2002	Consensus	SYBYL
LUDI	Böhm et al. 1998	Empirical	DiscoveryStudio
ASE	-	Gaussian	MOE
London dG	-	Empirical/FF	MOE

Force field-based scoring functions

Molecular mechanics force fields are used in scoring functions to calculate the protein-ligand interaction energy and the internal ligand energy. The two factors contributing to the energy are van der Waals and electrostatic terms. van der Waals energy is most often described by a Lennard-Jones potential (also known as the 12-6 potential):

$$E_{vdw}(r) = \sum_{j=1}^{N_A} \sum_{i=1}^{N_B} 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

where N_A and N_B are the number of atoms in molecules A and B, r is the distance between atoms i and j, σ is the collision

diameter between atoms i and j, and ε is the well depth of the potential.

Different modifications of Lennard-Jones potential have been formulated. For example, the 12-10 potential is used in AutoDock to model hydrogen bonding (Morris et al. 2009).

The electrostatic potential energy is calculated from a Coulombic equation:

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

where N_x are the number of atoms in molecule x, ε_0 the electric constant and q_y is the charge of each atom y.

These kinds of descriptions suffer from obvious serious limitations such as modeling protein-ligand binding in water, as they were originally formulated to model gas-phase interactions and do not take solvation or entropy into account. Furthermore, arbitrary cut-off values are required for modeling of non-bonded interactions, which complicates the estimation of long-distance interactions.

Given these limitations, additional terms besides van der Waals and Coulombic energy have been added to the scoring functions. The AutoDock scoring function includes a desolvation potential Esol based on the general approach by Wesson and Eisenberg (Wesson and Eisenberg 1992; Huey et al. 2007). It has an atomic solvation parameter S_i and volume V_i of the atoms surrounding given atom i:

$$S_i = A_i + Q * |q_i|$$

$$E_{sol} = \sum_{i,j} \left(S_i V_j + S_j V_i\right) e^{\left(\frac{-r_{ij}^2}{2\sigma^2}\right)}$$

where σ is a distance weighting factor, A_i and Q are atomic solvation parameters calibrated using six atom types.

Empirical scoring functions

Scoring functions can also take advantage of existing experimental data (Kitchen et al. 2004; Moitessier et al. 2008). Empirical scoring functions are derived with regression analysis from determined binding energies and/or crystallography data. The concept was originally implemented in *de novo* design program LUDI (Böhm 1992) and since then, several empirical scoring functions have been proposed (Wang et al. 2002). Empirical scoring functions are very simple to calculate, but obviously their performance is highly dependent on the data set from which they are derived.

Knowledge-based scoring functions

Knowledge-based scoring functions are also very quick to calculate (Kitchen et al. 2004). They are designed to reproduce experimentally observer structures instead of devising predictions of affinity like empirical scoring functions. As the name implies, knowledge-based scoring functions use data about protein-ligand interactions. Pre-defined atom-pair interactions are used to evaluate the docking pose. Similar to empirical scoring functions, knowledge-based scoring functions are limited by the availability of experimental structures.
Given the limitations of the current scoring functions (see Chapter 2.6), there have been many attempts made to combine several scoring functions to improve the accuracy (consensus scoring) (Clark et al. 2002). The debate still is on-going if consensus scoring is actually useful (Brooijmans and Humblet 2010), as there are both positive (Charifson et al. 1999; Krovat and Langer 2004) and negative findings (Verdonk et al. 2004; Stahl and Rarey 2001).

2.4 DATABASE PREPARATION

Whenever one chooses ligand- or structure-based approach for virtual screening, a molecular database of available compounds has to be prepared. While at first glance this step seems trivial, there are potential pitfalls for the unwary modeler.

One must make sure that the molecular database to be screened is up-to-date and at least most of the compounds are readily available for purchase. It makes little sense to find an interesting molecule that one cannot acquire, as the chemical synthesis is usually extremely laborious and rarely justified simply on the basis of initial virtual screening results.

A comprehensive and free source for virtual screening databases is ZINC, available at zinc.docking.org (Irwin and Shoichet 2005). It contains pre-processed databases from the most prominent chemical vendors. The current version 10 of ZINC has over 13 million compounds that are readily available.

2.4.1 Prefiltering

Before embarking on computationally intensive 3D virtual screening, prefiltering of the database is useful. Simple properties like molecular weight or number of rotatable bonds can be used to remove compounds that are not wanted. The most famous of these simple filters is “the rule of five” (Lipinski et al. 1997), which is used to a evaluate compound’s so-called druglikeness. It states that an orally active drug should possess no more than one violation of the following criteria:

- less than six hydrogen bond donors
- less than 11 hydrogen bond acceptors
- molecular weight below 500 Da
- CLogP less than five

It may not make sense to use “the rule of five” as a filter in the early lead discovery, as it has been shown that actual drug molecules are different from lead molecules (Teague et al. 1999; Ohno et al. 2010) and chemical probes (Oprea et al. 2007). A good lead needs to be less complex than the actual drug

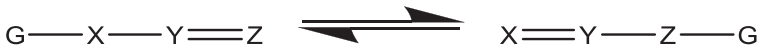
molecule, so that there is room for optimization. Oprea and co-workers have defined lead-like properties (Oprea et al. 2001):

- molecular weight below 450 Da
- ClogP between -3.5 and 4.5
- less than five ring structures
- less than 11 nonterminal single bonds
- less than six hydrogen bond donors
- less than nine hydrogen bond acceptors

One should also remove promiscuous compounds (also known as "frequent hitters"), molecules that show up as false positives in HTS, independent of the target due to reasons not related to the protein-ligand interaction (Baell et al. 2010). Reactive and undesirable groups should be also removed in the database preparation phase (for example, see Lagorce et al. 2008).

2.4.2 Tautomerism, protonation states and stereoisomerism

Tautomerism also has an effect on molecular databases (Pospisil et al. 2003; Knox et al. 2005). Tautomerism is isomerism of the form:



where the isomers are readily interconvertible (IUPAC 2010). This kind of isomer is called a tautomer. The atoms connecting groups X,Y,Z are usually carbon, hydrogen, oxygen or sulphur. Group G becomes an electrofuge or nucleofuge during isomerization. The most commonly known tautomeric phenomenon is the proton migration (prototropy), where the hydrogen atom moves between different sites on the same molecule. This is not to be confused with ionization or protonation where the hydrogen atom leaves or comes from *another* molecule. These different protonation states are sometimes called protomers in the virtual screening literature (not to be confused with the official definition of a protomer, which is a structural unit of an oligomeric protein). Other examples of tautomerism are keto-enol and ring-chain tautomerism.

An equilibrium exists between the different tautomeric forms. In the aqueous medium, several factors such as temperature and pH affect the tautomeric equilibrium. It is therefore non-trivial to predict the relative stabilities of different tautomers. Several programs are available for rapid tautomer and protonation state enumeration, such as QuacPac (OpenEye Scientific Software) and LigPrep (Schrödinger Inc.). QuacPac simply enumerates all reasonable tautomeric and protonation states in an aqueous solvent. LigPrep's tautomer tool is also based on pre-defined tautomeric groups and their assumed probabilities. EpiK (Schrödinger Inc.) and MoKa Suite (Molecular Discovery Ltd) can also predict the most likely tautomeric and protonation state instead of simply enumerating all forms. EpiK is based on Taft and Hammett equation parameterized by values from the literature and proprietary data. In a comparative study on currently available pKa-prediction programs, it was postulated that the training set used in EpiK is too small for diverse set of predictions (Manchester et al. 2010). The MoKa Suite is based on recursive enumeration of tautomers and an empirical tautomeric stability prediction method (Milletti et al. 2009). First, tautomers are generated by knowledge and aromaticity rules. Then, the stability of different tautomers is predicted by using empirical data. The predictions are adjusted with pKa-values predicted by MoKa (Milletti et al. 2007). MoKa also generates the relevant protonation states over a given pH range.

In addition to tautomerism and protonation states, there is stereochemistry to be considered. Normally the molecules that are purchasable from chemical vendors are racemic mixtures. Therefore, all stereoisomers need to be considered in docking as all forms will be present also in the bioassay (Brooks et al. 2008).

2.4.3 Conformational analysis

As previously stated, most small molecules are flexible and have several three-dimensional conformations. Conformation generators perform conformational analysis efficiently for virtual screening purposes. Efficiency in this context means that they produce a small total number of biologically relevant

conformations per molecule in a reasonable time for a large number of compounds (Sadowski and Gasteiger 1993; Watts et al. 2010). There are two possible outputs from a generator. Either it produces a single, low-energy conformation or an ensemble of diverse conformations. The selection of the virtual screening method then determines which kind of database one should use in the screening process.

Several alternatives for rapid conformation analysis are publicly available (Table 2.9). Conformation generators are generally based on either numerical methods such as distance geometry or more commonly, empirical data (Sadowski and Gasteiger 1993).

Table 2.9: Publicly available conformation generators that are suitable for virtual screening studies. S=Single, E=Ensemble, L=License, F=Free, FA=Free for Academic use and C=Commercial.

Program	Ref.	Algorithm	S	E	L	Website
Catalyst (Discovery Studio)	-	Systematic Search/MM/Random Search	y	y	C	www.accelrys.com
OMEGA	Hawkins et al. 2010	Fragment-based/Torsion library/MM	y	y	FA	www.eyesopen.com
MOE/Conf	-	MM/Fragment-based/Systematic Search	y	y	C	www.chemcomp.com
Balloon	Vainio and Johnson 2007	Distance Geometry/GA/MM	y	y	F	users.abo.fi/mivainio/balloon
CORINA	Gasteiger et al. 1990	Fragment-based	y	n	C	www.molecular-networks.com
ROTATE	Schwab 2003; Renner et al. 2003	Torsion library/MM	y	y	C	www.molecular-networks.com
ConfGen	Watts et al. 2010	MM/Empirical heuristics	y	y	C	www.schrodinger.com
CONCORD	-	Empirical/MM	y	n	C	www.tripos.com
CONFORT	-	MM	n	y	C	www.tripos.com
DG-AMMOS	Lagorce et al. 2010	Distance Geometry/MM	y	n	F	www.mti.univ-paris-diderot.fr/fr/downloads.html
Multiconf-DOCK	Sauton et al. 2008	Incremental construction, Systematic Search	n	y	F	www.mti.univ-paris-diderot.fr/fr/downloads.html

Balloon is an example of a modern conformation generator based on numerical methods (Vainio and Johnson 2007). It creates the initial conformation using distance geometry and additional conformations are generated with a genetic algorithm designed to preserve the diversity of conformations. In the postprocessing step, the conformations are relaxed using a MMFF94-like force field. Since Balloon does not utilize of any empirical data, it is rather slow compared to other conformation generators.

OMEGA from OpenEye Scientific Software is a hybrid method combining empirical information about fragment conformations and calculations of molecular mechanics (Hawkins et al. 2010). First, the initial conformation is constructed with a fragment library. Then, all rotatable torsions are sampled using a knowledge-based list of reasonable angles. Finally, the set of conformations is sampled with geometric and molecular mechanics criteria. It is extremely fast, on average generating a conformation for a molecule in 0.2 seconds (Lagorce et al. 2009).

ConfGen (Schrödinger Inc) is derived from the conformational analysis part of the docking program Glide (Watts et al. 2010). It uses a combination of molecular mechanics calculations and a set of empirical heuristic rules to generate diverse conformations. ConfGen has four different levels for conformational analysis: very fast, fast, intermediate and comprehensive. By default, the fast mode is used, which is designed for virtual screening purposes. The fast mode produced 13 conformations per molecule in approximately one second on average with a modern Intel Core2 2.4GHz system.

Some comparison studies on conformation generators have been published (for example, see Kirchmair et al. 2006 and Chen et al. 2008). However it is difficult to state with certainty which method is the best. There also seem to be some data issues in most studies published so far (Hawkins et al. 2010). Even though such studies are easy to conduct after a suitable benchmark set has been created, most conformation generators are commercial and licensing issues complicate those kinds of studies.

2.5 THE LIMITATIONS OF VIRTUAL SCREENING

Even though virtual screening has been successful in drug discovery projects, there are some fundamental limitations in both LBVS and SBVS that are good to keep in mind when designing new experiments. The issues relating to the validation and benchmarking will be discussed in the next chapter, as they do not directly link to the virtual screening methods themselves.

2.5.1 Limitations of LBVS

The first limitation of LBVS is the classical chicken and an egg problem: at least one biologically relevant molecule must be identified before database can be screened. This is a major limitation as there are many potential targets for which there are known ligands available.

It is unreasonable to expect something completely different from a methodology that is based on searching for similar molecules. The issue is illustrated on Figure 2.18, which shows two inhibitors for the catecholamine-O-methyltransferase (COMT) enzyme. They have both low 2D- and 3D-similarities even though they have similar biological activities. Total similarity based on a single molecule is therefore a relatively limited technique. This problem is alleviated by the fact that often several active molecules are known.

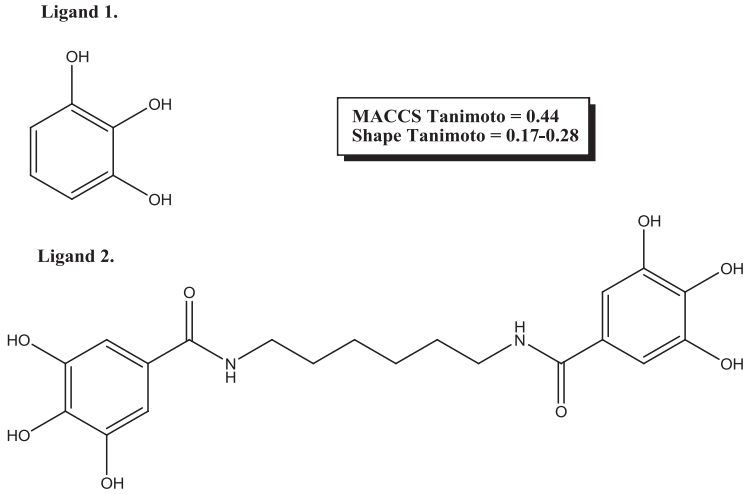


Figure 2.18: Global similarity metrics miss sub-structural similarity. The two ligands of the COMT-enzyme have both low 2D- and 3D-similarities.

There is also a clear paradox in the whole fundamental idea of finding novel bioactive molecules from LBVS, since there is the similarity principle that states that structurally related compounds display similar biological activities (Eckert and Bajorath 2007). This of course means that the more different compounds that there are, the less likely they are going to have similar activity (Bohm et al. 2004). Even though there are various ways to measure the similarity between two molecules, there is always a tradeoff between scaffold hopping and the probability of finding an active compound (Figure 2.19). It depends on the project if one wishes to find rather similar compounds with a high probability of being active or simply a large number of diverse compounds (Triballeau et al. 2006).

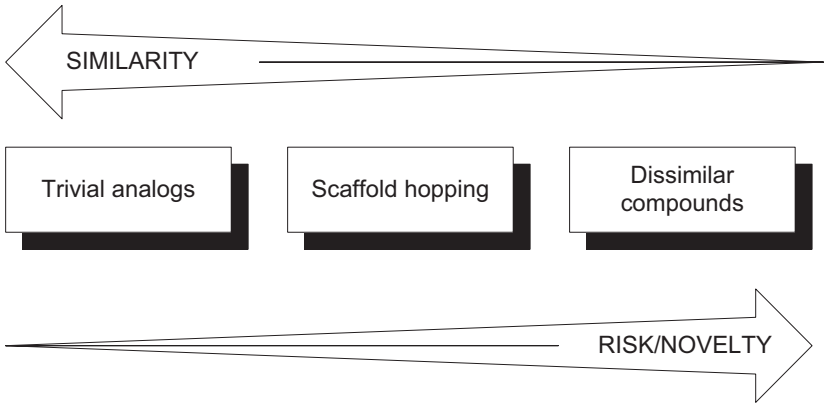


Figure 2.19: The tradeoff between chemical similarity and the probability of finding an active compound.

LBVS methods that require molecular alignment of multiple compounds, such as pharmacophores, assume that all of the active molecules bind in a similar conformation. Aligning several active conformations simultaneously is far from trivial, as the crystallized structures of protein-ligand complexes have well demonstrated. Two commonly used inhibitors of phosphodiesterase 5(PDE5), sildenafil and tadalafil, both have the same binding pocket, but the alignment is not obvious from the molecular structures alone (Figure 2.20).

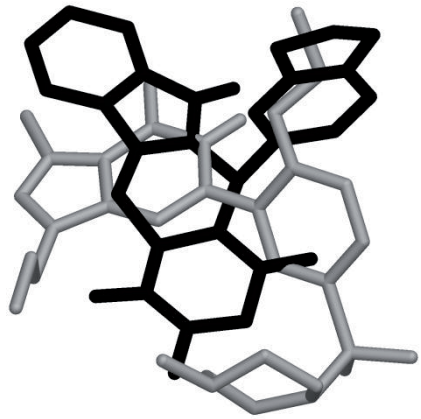


Figure 2.20: The binding conformations of two inhibitors of PDE5-enzyme. Sildenafil is in gray (PDB 2H42) and Tadalafil in black (PDB 1XOZ) (Wang et al. 2006; Card et al. 2004).

2.5.2 Limitations of SBVS

X-ray crystallography is a rather difficult and laborsome science and therefore, it is not surprising that the crystal structures of most drug targets are not available. The structures of only a few G-protein Coupled Receptors (GPCRs) have been successfully solved, even though this class accounts for approximately 30% of targets of all marketed drugs (Sela et al. 2010). Homology modeling-based structures have been used instead, but it is still unclear if such models are truly suitable for virtual screening. In a recent GPCR modeling and docking contest, most of the 29 homology models submitted were not accurate enough to permit virtual screening (Michino et al. 2009).

A protein model based on X-ray crystallography is an interpretation of experimental data (Davis et al. 2008). Two crystallographers may reach different conclusions from the same diffraction data. For example, a functional group of the bound ligand might be confused with a water molecule. This subjective nature of X-ray crystallography is often ignored when utilizing structures from Protein Databank.

In addition to the issues related to X-ray crystallography, there are major problems with current docking methods. The assumption that there is a rigid protein over-simplifies the modeling of protein-ligand interaction. The inductive effects are rarely considered and therefore the binding pocket may be of the wrong shape. A greater problem is that a macromolecular complex is not a single structure, but an ensemble of structures (Bissantz et al. 2010). Changes in conformations of both ligand and protein during the binding have a significant impact on the binding energy.

Scoring functions assume that binding free energy can be formulated by additive terms from various protein-ligand interactions. In reality, different molecular interactions are nonadditive and should be designated with different amounts of Gibbs energy in different contexts (Dill 1997).

Another serious deficiency in docking is that it does not take enthalpy-entropy compensation properly into account (Whitesides and Krishnamurthy 2005). An increase in entropy can compensate for a loss in enthalphy (Krishnamurthy et al. 2006; Ladbury et al. 2010). A good example of this phenomenon is the study of Christof and co-workers on a pair of thrombin inhibitors (Christof et al. 2007). The cyclopentyl group of the first compound was switched to cyclohexyl group in the second molecule. Both compounds had identical binding affinity even though X-ray crystallography indicated that the cyclopentyl group was located inside the binding pocket, whereas the cyclohexyl group was not. This similar binding affinity with a different binding mode was caused by enthalphy-entropy compensation as revealed by isothermal titration calorimentry. It is highly doubtful that this phenomenon would have been detected from molecular docking studies.

One can indeed wonder how docking can work at all, given all of these problems (Whitesides and Krishnamurthy 2005; Kolb and Irwin 2009). There are successful structure-based virtual screening studies where novel biologically active compounds have been identified, but rarely has the docking pose been experimentally validated by comparing it to the crystallized

structure (Kolb and Irwin 2009; Bissantz et al. 2010). It is therefore possible that at least some of the reported findings are either based on crude features like molecular shape or just sheer luck. Indeed, for more sophisticated tasks like lead optimization, molecular docking does not seem to be a reliable enough technique (Warren et al. 2006; Tirado-Rivers and Jorgensen 2006; Leach et al. 2006; Enyedy and Egan 2008).

3 Validation and evaluation of VS methods

"From combinatorial chemistry to genomics, new concepts or technologies that claim to help accelerate drug development have arguably been too rapidly embraced without true validation."
(Quote from the Editorial of Nature Reviews Drug Discovery 6, 3, 2007)

There have been proposals about hundreds of different virtual screening methods. It is rather difficult to say which methods are truly useful in finding novel bioactive compounds. There are two approaches for validation. In retrospective validation, data from the literature is used to evaluate the performance of a method, whereas in a prospective validation, the method is validated by the discovery of novel bioactive compounds.
In most cases, the methods have been validated by retrospective virtual screening and no prospective results are provided. The risk of retrospective studies is that the method may work artificially well with certain data sets and that the results gained are not generally applicable. However, prospective studies are not conclusive either as active molecules can be found simply by luck. After all, history is filled with examples of serendipitous drug discovery (Ban 2006).
When evaluating a virtual screening method, there are two points to consider (Sheridan and Kearsley 2002). First, how good the methods are at selecting active molecules from a database i.e. what is the *quantity* of hits? Secondly, how novel are the chemical structures of the molecules that are predicted to be active i.e. what is the *quality* of hits? This is not trivial matter because there is no standardized test set or even a metric available to measure the performance of a new method (Geppert et al. 2010; Triballeau et al. 2005; Edgar et al. 2000) and the

validation of new methods is often rather limited (Good et al. 2004b; Kolb and Irwin 2009). To paraphrase Lord Kelvin (1824-1907): one cannot improve current virtual screening methods if one cannot measure the performance.

The output of a virtual screening method is a hit list, which contains the database molecules ordered according to their likeness to be active. Ideally, active and inactive molecules are separated by the score produced in the virtual screening method with some threshold *T* (Figure 3.1). A true positive is a molecule that was correctly predicted as being active, while a false positive is an inactive molecule that was predicted to be active. A true negative is a compound that was predicted correctly to be inactive and a false negative is an active compound that was predicted to inactive. Generally speaking, virtual screening methods tend to produce a high number of false positives.

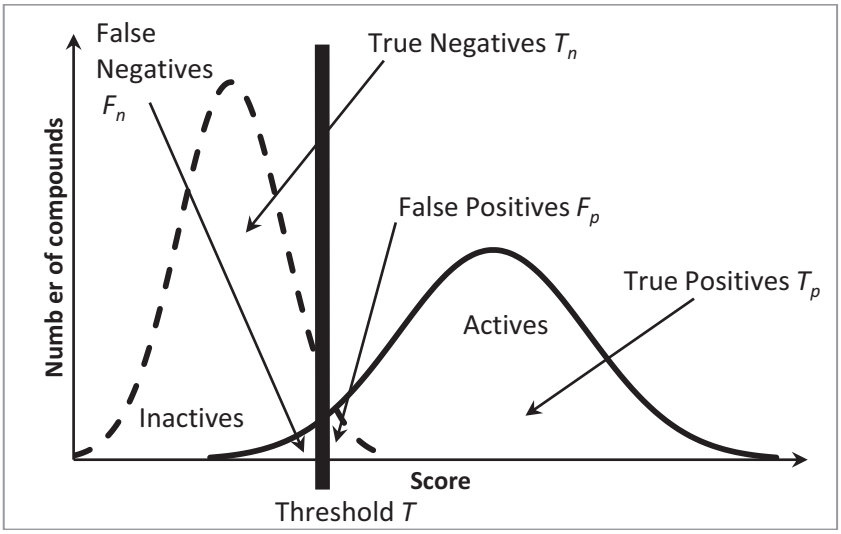


Figure 3.1: Idealized example of a virtual screening hit list (adapted from Triballeau et al. 2006)

3.1 PUBLICLY AVAILABLE DATA SETS FOR VS EVALUATION

Usually virtual screening methods are validated by retrospective screening of a data set, where a number of known

actives (ligands) are mixed with supposedly inactive compounds (decoys). This is very problematic, since the results depend strongly on the data set composition (van Drie 2004; Eckert and Bajorath 2007). Some targets are easier than others. Decoy molecules can be trivially different from ligands. For example, if decoy molecules are much larger than ligands, it is trivial to separate the two groups based on a simple descriptor like molecular weight. An imbalanced data set can be compared to a police identity parade with one black male suspect in a row otherwise filled with white females (Nicholls 2008). In addition, the different data set composition makes the reliable comparison of methods between different studies impossible.

In order to tackle these problems, Huang and co-workers created a publicly available data set called the "Database of Useful Decoys" (DUD) (Huang et al. 2006). It contains a large and diverse test set of forty targets belonging to various protein families such as nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, etc. Each target has a set of ligand and decoy molecules. 2950 ligands in total were gathered from the literature. The decoy molecules were selected from Lipinski-compliant subset of the ZINC database using 2D-fingerprints. These molecules are assumed to be inactive due to their dissimilarity to the ligand set. From this set, 36 decoy molecules per ligand that had similar calculated physical properties were selected. Molecular weight, hydrogen bond acceptors, hydrogen bond donors, CLogP and the number of rotatable bonds were considered. DUD seems to provide a more stringent test than the previously commonly used MDL Drug Data Report (MDDR) and it has been utilized in numerous recent studies (for example: Faver et al. 2010; Cross et al. 2009; Venkatraman et al. 2009; Englebienne et al. 2009; Clark et al. 2009; Cosconati et al. 2009; von Korff et al. 2009; Hartmann et al. 2009; Pham and Jain 2008).

Some important aspects of DUD have been raised after the publication of the data set. One of the authors of DUD explicitly stated afterwards that DUD is *only* for benchmarking molecular docking and nothing else (Irwin 2008). Indeed, as the DUD

decoys have been selected to be 2D-dissimilar from active compounds, DUD is clearly unsuitable for benchmarking 2D-methods. It also produces overoptimistic results for 3D-methods, since there are many targets where all active compounds are trivial analogs of a central structure. There is also an imbalance in formal charges between active and inactive molecules: 42% of active and 15% of inactive molecules is charged. This creates anomalies in enrichment studies.

There is a filtered version of the DUD available called DUD LIB VS that is intended for benchmarking ligand-based virtual screening methods (Good and Oprea 2008; Jahn et al. 2009). A lead-like filter and a clustering algorithm were applied to remove trivial analogs and molecules that would not have passed a normal database preparation step. The imbalance between formal charges is however still present on the data set and the ligand chemical diversity of some targets is still rather modest. Since the original publication, DUD has been also clusterized for scaffold hopping analyses by Andrew Good. The scaffolds are detected with reduced graph assemblies using the method of Barker and co-workers (Barker et al. 2003).

As there is clearly room for improvement in DUD, alternative benchmarking data sets have also been proposed. The Maximum Unbiased Validation (MUV) data set is based on PubChem bioactivity data for both ligands and decoys (Rohrer and Baumann 2009). In MUV, decoy molecules have been selected to resemble ligands on the basis of simple descriptors. These descriptors are vectors containing various atom counts combined with hydrogen-bond acceptors/donors, logP, the number of chiral centers and the number of ring systems. It has been proven extremely challenging for current ligand-based virtual screening methods (Tiikkainen et al. 2009).

There is the recently published ChEMBL (available at http://www.ebi.ac.uk/chembldb), a database containing approximately 500 000 bioactive compounds, which should provide a good starting point for the building of future virtual screening benchmarking data sets. It contains information about bioactive compounds, their targets and screening data extracted

from various high-impact journals. All data is manually evaluated and is therefore of higher quality than the PubChem data used in MUV (Bender 2010).

3.2 MEASURING THE QUANTITY: EVALUATING THE HIT RATE

When measuring the quantity of active molecules from a virtual screening method, the recommended metric is the area under the curve for Receiver Operating Characteristic plot (ROC AUC) (Jain and Nicholls 2008). ROC analysis was developed during World War II for radar applications and since then it has been applied in many fields of science. It is a visual as well as numerical method for evaluation of different virtual screening methods (Triballeau et al. 2005; Sonego et al. 2008).

ROC analysis can be applied to any binary classification problem. In virtual screening, the compounds in the hit list must be assigned as being either active (1) or inactive (0). Many benchmarking sets already have this classification, as they are divided into ligands and decoys. A confusion matrix is generated for each threshold in the hitlist, from which sensitivity *Se* and specificity *Sp* are calculated (Figure 3.2). Finally, sensitivity is plotted as a function of 1-specificity to form the ROC curve. The integral of this curve (area under curve, AUC) is a single numerical measure of ranking performance (Sonego et al. 2008). Random ranking produces a diagonal curve with AUC of 0.5, while a perfect AUC is 1.0. There is no absolute AUC threshold for "good performance", but a virtual screening method should at least produce AUC higher than the random ranking (0.5).

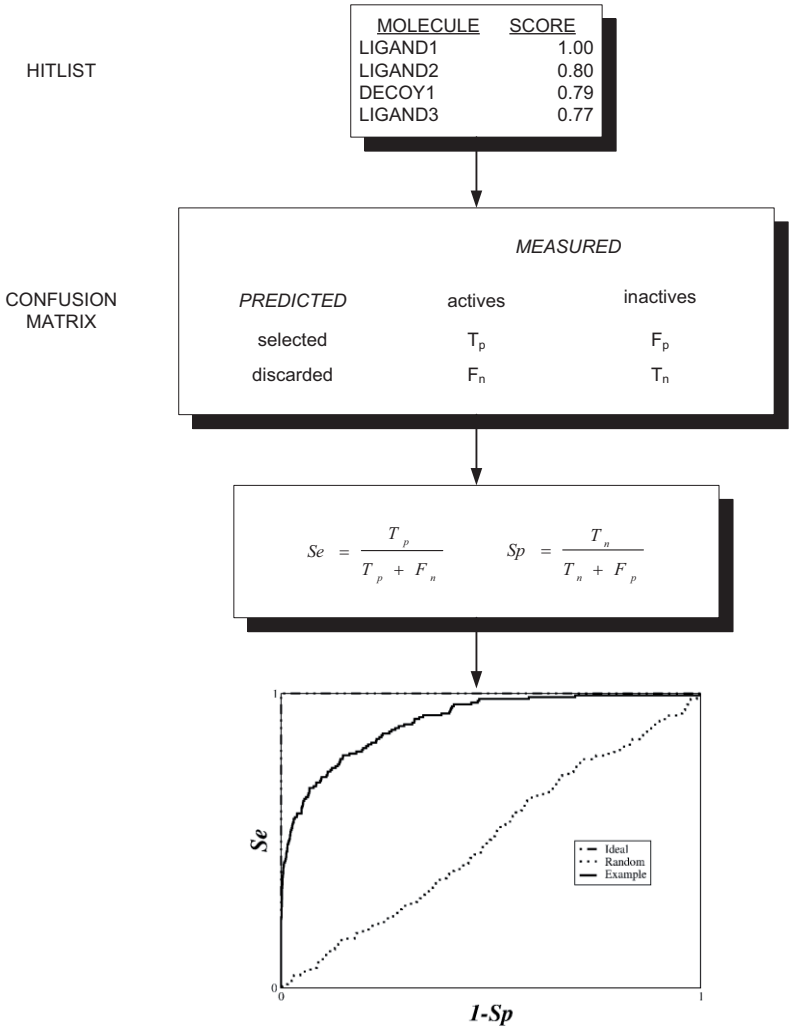


Figure 3.2: Generation of ROC curves for virtual screening (Triballeu et al. 2005).

The application of ROC AUC to the evaluation of virtual screening methods has been criticized, because ROC AUC does not take the "early enrichment problem" into account (Truchon and Bayly 2007; Kirchmair et al. 2008). As only the top of the hitlist can be normally tested for biological activity, early enrichment is an important issue. The problem is illustrated on Figure 3.3, where there are two idealized curves with the same ROC AUC. Both curves have the same AUC of 0.5 even though the hitlists are clearly different. The dashed line is simply random and the solid line is from a hitlist where 50% of actives

are retrieved at the top and 50% at the bottom. It is however debatable if such extreme biphasic behavior could really be observed in real life virtual screening scenarios (Nicholls 2008).

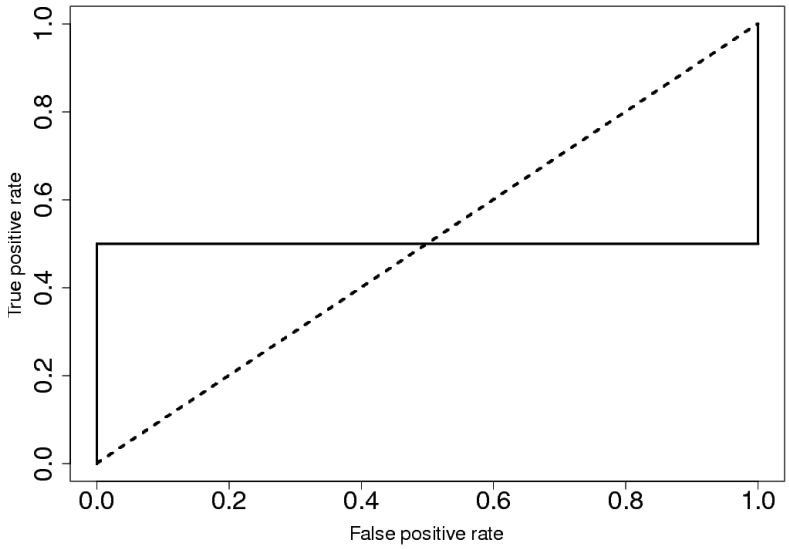


Figure 3.3: "The early enrichment" problem. Both solid and dashed curves have the same ROC AUC.

The most common metric to measure the early enrichment is the Enrichment Factor (EF) (Jacobsson et al. 2003; Hecker et al. 2002; Diller et al. 2003; Triballeau et al. 2006):

$$EF = \frac{\frac{T_p}{N_{subset}}}{\frac{L_{database}}{N_{database}}}$$

where T_p is the number of true positives, N_{subset} the number of molecules in a given cutoff, $L_{database}$ the total number of actives in the data set and $N_{database}$ the total number of molecules in the data set.

There are two problems associated with EF i.e. it relies on an arbitrary cutoff (the subset size) and it is dependent on the ratio

of the active molecules in the database, which makes comparisons between different studies impossible (Kirchmair et al. 2008; Nicholls 2008). The cutoff issue can be examined by calculating several EFs at different cutoffs, but this complicates the interpretation of the results.

To reduce the effect of the arbitrary cut-off value, Robust Initial Enrichment (RIE) was developed by Sheridan and co-workers (2001):

$$S = \sum_{i}^{N_{actives}} e^{-\frac{R_i}{N_{actives}}}$$

$$RIE = \frac{S}{\langle S \rangle}$$

where $N_{actives}$ is the number of active molecules in the hit list, R_i is the rank of the active compound I and <S> is the mean S calculated from 1000 trials where the ranks of active compounds are randomized.

RIE has its own limitations, e.g. it is difficult to reliably compare two RIE values, and therefore a new metric called the Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC) was developed (Truchon and Bayly 2007). It is a weighted ROC AUC which means that the top of the hit list is weighted more than the rest. The mathematics behind BEDROC are complicated, but the authors provide C++ and Python codes for calculating BEDROC values from simple hit lists. However, it is not clear if the RIE or BEDROC provides any extra value over ROC AUC when evaluating virtual screening methods (Nicholls 2008; Geppert et al. 2010).

The metrics used for measuring the retrieval effectiveness of information retrieval systems can be readily applied in virtual screening validation (Table 3.1) (Edgar et al. 2000; Triballeau et al. 2006). Similar to EF, these metrics suffer from the problem of an arbitrary cut-off for subset selection.

Table 3.1: Different metrics for the assessment of virtual screening performance (Edgar et al. 2000; Triballaeu et al. 2006; Bradley et al. 2000, 2003; Matthews 1975; Diller et al. 2001; Guha and Jurs 2005; Weston et al. 2003; Güner and Henry 2000). The variables in the formulas: the number of selected molecules N_{subset}, the number of screened molecules $N_{database}$, the number of active molecules in the database $L_{database}$ and the number of active molecules in the hit list L_{subset}.

Metric	**Formula**
Recall/sensitivity (R, Se)	$\frac{L_{subset}}{L_{database}}$
Precision/hitrate (P)	$\frac{L_{subset}}{N_{subset}}$
Fallout (F)	$\frac{N_{subset} - L_{subset}}{N_{database} - L_{database}}$
Generality (G)	$\frac{L_{database}}{N_{database}}$
Vickery	$\frac{1}{\left(\frac{2}{P}\right) + \left(\frac{2}{R} - 3\right)}$ $\frac{1}{\left(\frac{1}{P}\right) + \left(\frac{1}{R} - 1\right)}$
van Rijsbergern (α is the relative importance of the precision)	$\frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$
Cosine coefficient	$\sqrt{PR}$
Discrimination ratio	$\frac{Se}{Sp}$
Information content	$T_p \log\frac{T_p}{F_p} + F_n \log\frac{F_n}{T_n}$
Matthews correlation coefficient	$\frac{T_p T_n - F_n F_p}{\sqrt{(T_n + F_n)(T_n + F_p)(T_p + F_n)(T_p + F_p)}}$
GH score	$\left(\frac{3}{4}P + \frac{1}{4}R\right)Sp$
Ford's M (ω is an weighting factor), "balanced labeling performance" when ω=0.5	$\omega Se + (1-\omega)Sp$
Statistical significance	$\sum_{k=T_p}^{L_{database}} \frac{\binom{L_{database}}{k}\binom{N_{database} - L_{database}}{N_{subset} - k}}{\binom{N_{database}}{N_{subset}}}$
Analysis of efficiency (U is the n of mols. with unkown activity)	$\frac{1}{2}(Se + Sp)(1 - \frac{U_{subset}}{U_{database}})$

3.3 MEASURING THE QUALITY: EVALUATING THE CHEMICAL DIVERSITY AND SCAFFOLD HOPPING

When evaluating if scaffold hopping has occurred or not, a well-defined criterion for scaffolds should be available (Jenkins et al. 2004; Brown and Jacoby 2006; Schneider et al. 2006; Mackey and Melville 2009). The scaffold criterion should also be independent of the algorithm used in the virtual screening. There are two aspects to this problem: first, the definition of a scaffold and secondly, the quantification of scaffold retrieval (Figure 3.4). Most publications seem to define scaffolds based on molecular frameworks – the concept introduced by Bemis and Murcko (Chapter 2.3). Original Bemis and Murcko molecular frameworks are simply graphs without any atom information, but often different heterocyclic structures are considered as different scaffolds (Lipkus et al. 2008). This makes sense from both the synthetic chemistry and chemical information point of view. Connecting two piperidine rings is different than connecting two cyclohexane rings. Both structures also have clearly different electrostatic properties and thus probably different biological properties. However, the heterocyclic scaffold definition introduces some new issues. For example, should piperidine and piperazine be considered as different scaffolds?

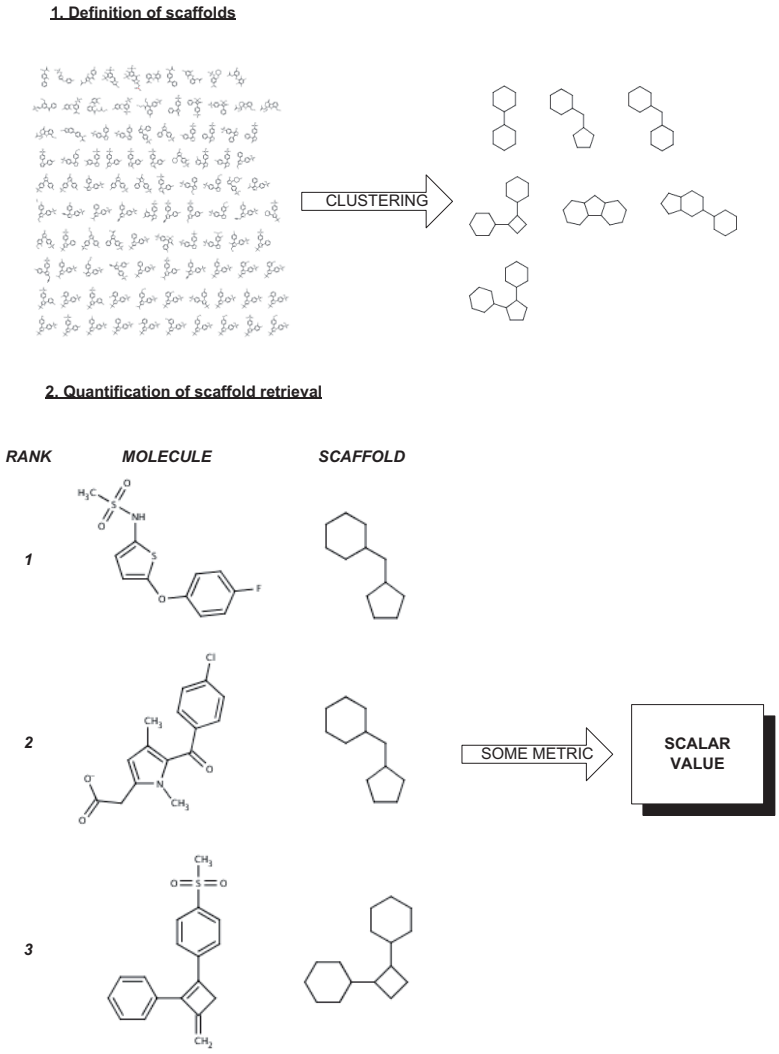


Figure 3.4: The two steps required in objective scaffold hopping quantification.

Most of the current objective scaffold definition methods are based on 2D-properties and do not take 3D-similarity into account. Scaffolds are still often subjectively defined and not numerically measured (Bender et al. 2004; Jenkins et al. 2004; Zhang and Muegge 2006; Williams 2006; Good et al. 2004b). Some studies have calculated 2D-fingerprints and Tanimoto coefficients in order to show scaffold hopping (for example, Saeh et al. 2005), but there is the problem of choosing the optimal threshold value i.e. when the two scaffolds are different.

Clustering of 2D-properties has been used to define scaffolds in several recent studies (Stiefl et al. 2006; Vidal et al. 2006; Krier et al. 2006). Selector from Tripos (part of the SYBYL molecular modeling package) is one of the first classification programs for undertaking library diversity analysis (Martin 2001). Selector can use several diversity metrics such as UNITY 2D fingerprints, 3DFlex fingerprints, atom pair distances, CLOGP, QSAR derived parameters, substituent constants and CoMFA columns. There are also three different classification algorithms implemented in Selector (hierarchical clustering, Jarvis-Patrick clustering and reciprocal nearest neighbor clustering). Other similar tools are available from ChemAxon (Borosy et al. 2001; Vargyas et al. 2006) and Simulations Plus (Krier et al. 2006). The problem with all automatic clustering programs is that two different similarity metrics define different scaffolds and the scaffolds may appear arbitrary (Medina-Franco et al. 2009). The issue of scaffold definition is clearly a complex subject which needs to be clarified.

There is still the issue of quantification of the scaffold retrieval even if one uses pre-defined scaffold definitions. Such metric would be a very useful tool in the validation and evaluation of virtual screening methodologies (Good et al. 2004b). The intuitive "the fraction of retrieved scaffolds in the top of hitlist" has been used in most studies assessing scaffold hopping in virtual screening, but it also has some issues as recently pointed out (Mackey and Melville 2009). For example, the scaffolds that are present in a large number of compounds are more likely to be found by random than those scaffolds that are only in few molecules in the data set. Overall, there is still not any generally accepted metric available for the objective evaluation of scaffold hopping, but recently some metrics have been proposed that may perhaps represent at least a partial solution to the problem (Krier et al. 2006; Clark and Webster-Clark 2008; Mackey and Melville 2009; Medina-Franco et al. 2009).

Krier and co-workers proposed a new metric called PC50C (Krier et al. 2006). It is computed from the percentage of scaffold classes accounting for 50% of the classified compounds. Medina-

Franco and co-workers further developed this line of thought with scaffold retrieval curves and a metric called Scaled Shannon Entropy (Medina-Franco et al. 2009). The way that these methodologies could be applied to the virtual screening would be to analyze the diversity of the active molecules in the top of the hitlist selected. However, this strategy has not yet been investigated.

Clark and Webster-Clark suggested the addition of weights to ROC curves accounting for the measurement of scaffold hopping (Clark and Webster-Clark 2008). Their idea was recently refined in a study by researchers at Cresset Biomolecular Discovery (Mackey and Melville 2009). In this study, scaffold hopping weighted ROC AUC, EF and other weightings for enrichment metrics were developed. It was shown that the difference between the DUD COX2 active molecules (relatively few scaffolds) and FXA active molecules (more diverse set) could be detected. These two extreme cases were easy to identify, but for other targets the quantification was not so clear. More studies will be required to confirm the usefulness of this interesting technique. As Mackey and Melville provide free software to calculate their metrics, conducting such a study might be rather straightforward.

Given the controversial current status of scaffold hopping quantification (Geppert et al. 2010), pre-defined 2D-scaffold definitions with a simple metric "the fraction of retrieved scaffolds in the top of the hitlist" are used in assessing of early scaffold hopping in the experimental part of this thesis:

$$F_X = \frac{S_{found}}{S_{all}}$$

where S_{found} is the number of the scaffolds found out of all possible scaffolds S_{all} in the top X of the hit list. This metric is a "First Found" technique since each scaffold is counted only once (Mackey and Melville 2009).

4 Aims of the study

The aim of study was to improve current 3D-virtual screening protocols by reducing the number of required calculations, while maintaining roughly the same accuracy as measured by enrichment and chemical diversity. The more specific aims of the study were:

1. to develop a fast molecular field-based method for ligand-based virtual screening that could be used when multiple similarly binding ligands are known in order to overcome the limitations of total similarity (FieldChopper)
2. to validate and benchmark FieldChopper with a diverse data set by retrospective virtual screening and apply the technique for ADMET-predictions
3. to assess the impact of ligand-based tautomer and protonation state prediction on molecular docking in order to speed up the structure-based virtual screening
4. to increase the throughput of shape-based virtual screening with GPU-computing and reduced conformational analysis

5 Development and validation of FieldChopper

FieldChopper is a novel method for LBVS that can be used when multiple, similarly binding active ligands are known. Here, the algorithm is described and its performance is evaluated. For a more detailed description of the implementation, the reader is referred to the licentiate thesis (Kalliokoski 2008). The chapter is adapted with permissions from: Kalliokoski T, Rönkkö T, Poso A: FieldChopper, A New Tool for Automatic Model Generation and Virtual Screening Based on Molecular Fields. Journal of Chemical Information and Modeling 48: 1131-1137, 2008.

5.1 INTRODUCTION

Molecular fields describe the properties of a compound by the potential around the molecule. They have been applied for virtual screening. The seminal work in this area is CoMFA, which is the mostly commonly used 3D-QSAR method (see Chapter 2.3.3). Putta and co-workers (2002) developed a method in which molecules are represented in a binary shape-feature descriptor space as bit-strings and the molecule's relative activity is used to identify the subset of the bit-string that is most relevant to that activity. This subset is then used as a model for virtual screening. The method was evaluated using two retrospective virtual screenings of thrombin inhibitors (Srinivasan et al. 2002). The model was constructed using 38

active molecules and 2418 inactive molecules. In the first virtual screening, this model was used to filter the MDDR database. MDDR was pre-filtered down to a set of 35462 molecules, which contained 540 known thrombin inhibitors. The shape-feature based method selected 507 compounds, from which 181 were known active molecules. The average enrichment was 2.7 times greater compared to 2D-fingerprints. In a second screening, the same model was used to screen a small, in-house synthetic library of 634 compounds, which contained 64 known active molecules. The shape-feature based method selected 109 compounds, from which 15 were active. The enrichment ratio was 1.4 and considering that a random selection should on average result in enrichment of 1.0, one can conclude that these results are exploratory at the best.

Jain (2004) has developed Surflex-Sim, a method for ligand-based structural hypotheses for use in virtual screening. These hypotheses are built by aligning a set of active compounds by using a molecular fragmentation and incremental construction algorithm. The algorithm is rather computationally expensive, since it takes several hours on a modern desktop computer. Only two to three compounds are used to build a model. Virtual screening is done by superimposing the new compound against each active compound used in the model. The superimposition that has maximum mean similarity against all active compounds in the model is returned after being given a score from 0 to 1. Surflex-Sim was evaluated using a diverse test set of 22 targets (Cleves and Jain 2006). A total of 979 active drug molecules were mixed with 850 inactive molecules from Available Chemicals Directory. Models were built using one to three active compounds. The ability of the models to identify cognate drugs against a background of screening molecules showed excellent enrichment in 20 out of 22 cases.

Comparative Molecular Active Site Analysis (CoMASA) uses a set of active compounds, which are used to generate a 3D map (Kotani and Higashiura 2004). This map can then be used for building of queries for virtual screening. CoMASA has not been

developed any further, but the method is conceptually similar to FieldChopper described in this chapter.

An overview of FieldChopper is shown in Figure 5.1. First, a template molecule for superimposition is selected which is used in both model generation and scoring algorithms. A model is built by superimposing a set of bioactive molecules onto the template molecule and running the model generation algorithm. Then, a 3D multi-conformation database is screened by superimposing the compounds onto the same template and scoring them against the model. Finally, the scores are saved in a hit-list that can be used in the selection of compounds for in vitro testing. For practical purposes, the superimposing method has to be fast enough to handle thousands of molecules within a reasonable time. For example, the previously described ROCS and BRUTUS are these kinds of methods.

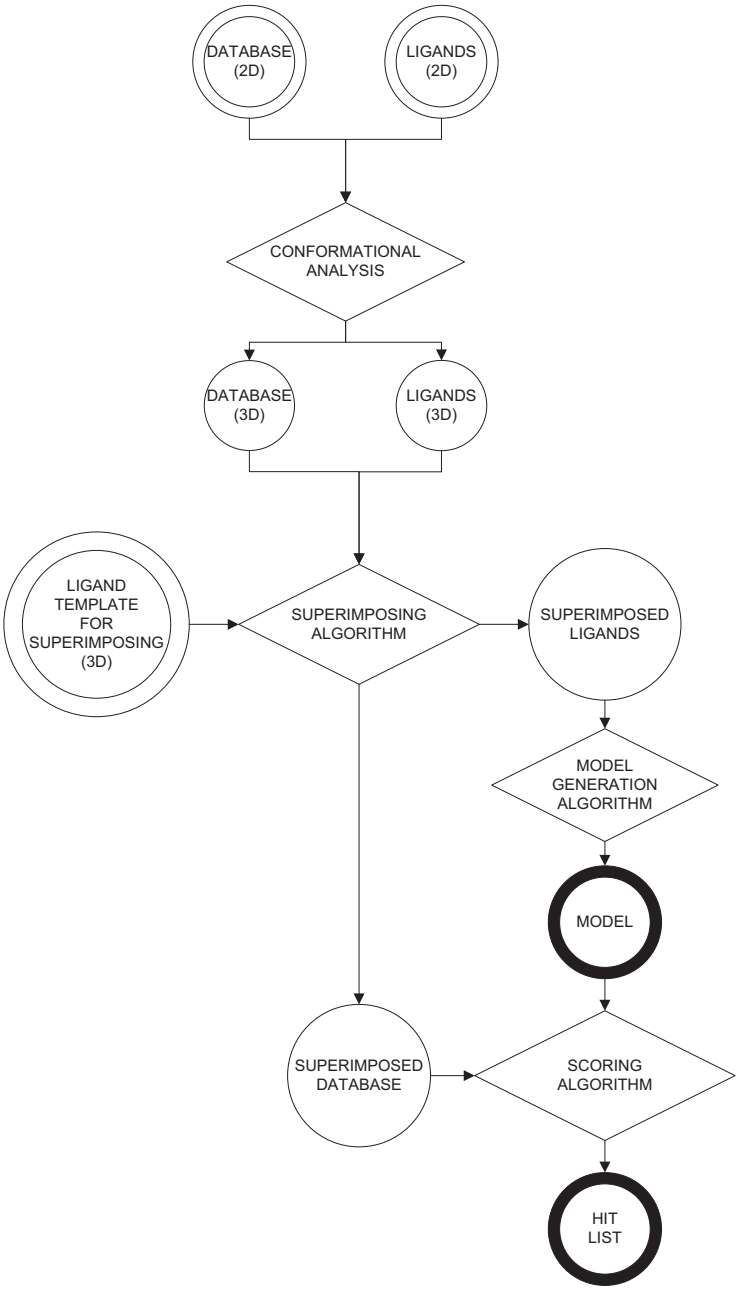


Figure 5.1: The overview of the FieldChopper method.

5.2 PREPARATION OF THE DATA SET

The method was developed and validated by a data set prepared from the DUD. After the publication of this study, several issues have been raised concerning DUD (see Chapter 3.1). With the benefit of hindsight, it is likely that the results obtained here are over-optimistic and lower enrichments would be observed in reality. However, the main conclusions of the original publication are still valid.

5.2.1 Selection of targets

Twelve targets from DUD's 40 targets were selected for this study based on the number of ligands (Table 5.1), having at least 50 active ligands. All other protein classes are included except for the metalloenzymes, which were excluded because there were not enough ligands in DUD for any of the metalloenzymes.

Table 5.1: Selected Targets for Retrospective Virtual Screening and Number of Ligands and Decoys. The ligands used in the models (15 per target) are not included.

Target	Ligands	Decoys
ACHE	90	3714
AR	59	2628
COX2	333	12464
DHFR	186	7145
EGFR	429	14894
$ER_{agonist}$	52	2355
FGFR1	103	4205
FXA	127	5095
GR	63	2797
INHA	70	3035
P38	241	8387
SRC	140	5793

5.2.2 Decoy Sets

As suggested in the original DUD paper, two decoy sets were created. The "own decoys" set contains the DUD decoys selected for the target, and the "combined decoys" set consists of all the decoys used in the selected 12 targets combined together. This "combined decoys" consists of about 60% of the DUD decoys, so direct comparisons cannot be made with the "amalgamated test set" in the original DUD paper. The "combined decoys" set is used to simulate a virtual screening scenario, where there would be a larger number of heterogeneous compounds available.

5.2.3 Conformation Generation and Calculation of Partial Charges

Since all the methods used in this study consider molecules as rigid structures, conformations had to be pregenerated using OMEGA (see Chapter 2.5.3) with the MMFF94s force field. The number of conformations was limited to ten for the virtual screening sets. Partial charges were assigned using the MMFF94 method implemented in MolCharge from OpenEye Scientific Software.

5.2.4 Molecule Superimpositioning

To study the effect of different alignment algorithms on FieldChopper's virtual screening accuracy, the alignments were produced with BRUTUS and ROCS (see Chapter 2.3.4). The three highest scoring molecular alignments for each conformation were included into the screening database for FieldChopper. As previously discussed, the main difference between these two alignment methods is the way that molecular energy fields are represented. ROCS employs a set of analytic Gaussian functions, while BRUTUS is a grid-based method. ComboScore with default ImplicitMillsDean force field was used to score the superimposed structures with ROCS and the default total score was used in scoring of BRUTUS alignments.

Template molecules for alignment were taken in their bioactive conformation from the Protein Databank using the same crystal

structures as in the original DUD paper. Atom and bond types of template molecules were corrected with SYBYL version 7.1. Hydrogen atoms and partial charges with MMFF94 were added using the same software.

The template molecule selected for alignment might not have been ideal in some of the cases studied here. For example, the template molecule for FGFR1 seems to be on average smaller than a molecule in the ligand set (Table 5.2). Most likely the prediction accuracy could have been improved by considering other ligands as template molecule for alignment. However, then the results would have been difficult to compare due to the arbitrary selection of the alignment template molecule. In general it seems that the template molecules being used are roughly of the same size or slightly larger than the ligand molecule on average.

Table 5.2: Molecular weight (MW), surface area (SA) and number of heavy atoms (HA) for the alignment templates and ligand sets

	Template molecule			**Ligand molecule average**		
Target	**MW**	**SA**	**HA**	**MW**	**SA**	**HA**
ACHE	393	684	28	344	613	25
AR	284	494	21	327	509	23
COX2	458	590	26	383	584	26
DHFR	469	723	33	339	563	24
EGFR	399	712	29	357	561	24
$ER_{agonist}$	333	527	24	286	286	21
FGFR1	345	586	25	430	682	30
FXA	470	683	31	452	711	33
GR	392	546	28	384	585	27
INHA	407	653	30	373	611	27
P38	549	875	39	369	606	27
SRC	506	770	36	411	649	28

.

5.2.5 Model building

The active molecules that are selected have a large impact on the model and its performance. A FieldChopper model should have a diverse selection of active compounds with different chemical

scaffolds. They should also be superimposable into each other. Subjective selection of compounds by a human investigator would introduce significant bias into the study. To ensure objective selection of compounds for the FieldChopper models, the following semi-automatic protocol was applied. Since the activities for DUD ligands were not readily available, compounds were only considered to be either active or inactive. To maximize the chemical diversity of the model compounds, GRIND-descriptors for the ligands were calculated using three probes (DRY, O carbonyl, and N amide). Then, two-component principal component analysis was performed with ALMOND. From these analyses, 15 compounds were selected for each target using the Kennard-Stone uniform subset selection algorithm as implemented by Daszykowski and co-workers with GNU Octave (Daszykowski et al. 2002). The compounds used in the FieldChopper models were removed from all virtual screening data sets.
BRUTUS was used to superimpose the compounds used in the model. Three possible solutions were generated for each molecule. From these superimpositions, the best one for the model was selected by visual inspection.

5.3 ALGORITHMS

The most important FieldChopper algorithms are those constructed for model generation and scoring. FieldChopper uses the electrostatic potentials and van der Waals volumes to describe molecules. One common way to represent these fields is to use a rectilinear 3D-lattice that is equally spaced and to calculate the interaction between the molecule and a probe atom in each grid point. This approach was utilized in the previously described CoMFA. FieldChopper uses grid-spacing of 1 Å by default and a sp3 carbon with a charge of +1.0 as a probe atom. The size of the grid is determined according to the molecules being examined in the model.
The van der Waals volume is approximated as

$$E_{vdW} = \frac{v^6}{r^6}$$

where v is the van der Waals radius for the atom; and r is the distance between the grid point and the atom.
The electrostatic potential can be calculated with the following equation based on Coulomb's law (Dill and Bromberg 2002):

$$E_{ele} = \frac{kQ}{r^2}$$

where k is a conversion factor; Q is the partial charge on the atom; and r is the distance between the grid point and the atom.

5.3.1 Model Generation Algorithm

In standard pharmacophore methods that result in a model, all of the compounds used to form the hypothesis on the possible binding mode should have similar, high biological activity (Poptodorov et al. 2006), although it has been suggested that it might also be useful to use low activity compounds (Dixon et al. 2006). On the other hand, 3D-QSAR methods (like CoMFA) require a diverse activity range for ligands contributing to the model (Höltje et al. 2008). FieldChopper considers molecules to be either active or inactive. The user must decide what an"active" compound is and thus what is an"inactive" molecule. For example, some kind of potency, e.g. an IC50 value, could be used to distinguish between active and inactive compounds.
A reference molecule to act as the template molecule for superimposition is needed. This molecule should be in the bioactive conformation and be large enough so that the whole set of active compounds can be superimposed onto it. The same molecule is used to superimpose molecules during the database screening phase.
The model generation algorithm detects similar grid points between active compounds. Each grid point is analyzed one at a time. The values from active molecules are classified into three bins (Table 5.3) for each point. This results in the creation of a

three-class histogram from which the peaks are detected. If the class frequency is over one-third of the sum of classes, then it is classified as a peak. The amount of the peaks can vary from zero to two peaks within one histogram.

Table 5.3: Classification limits for the van der Waals and Electrostatic Potentials

Interaction	**Bin 1**	**Bin 2**	**Bin 3**
van der Waals	inside (>=1)	near (>=0.001, <=1.0)	outside (<0.001)
electrostatic	negative (<-0.1)	near zero (>=-0.1, <0.1)	positive (>=0.1)

The peaks are used in the scoring algorithm. Since most of the electrostatic grid is empty, important grid points for the activity are detected using the van der Waals histograms. The grid points having a peak in their "Near Molecule" bin are taken into the electrostatic scoring, and all other points are excluded. For the van der Waals volumes, all grid points are used in the scoring, since a van der Waals volume describes the overall shape of the binding site. Classification limits for the van der Waals histograms are selected so that compounds larger than those used in the model are punished in the scoring algorithm.
In order to obtain an overview of the FieldChopper models, an analysis of peak distributions was performed. The numbers of different peaks are presented in Tables 5.4 and 5.5. The models displayed a very similar distribution of the van der Waals peaks. The reason for this phenomenon is that the template molecules used for alignment are roughly of the same size. The differences in "Outside" peaks in the van der Waals scoring are attributable to larger grid boxes for certain targets. No histograms with "Inside&Outside" or "None" peak cases were found in van der Waals material. With respect to the electrostatics, the nuclear hormone targets (AR, $ER_{agonist}$, and GR) had distributions different from the other targets. This reflects their partial charges which are close to zero in most cases. Positive electrostatics seem to be dominant, which is due to the

positively charged nitrogen atoms in the original DUD data. This revealed the bias in the DUD data set, which was later also identified by others (Irwin 2008).

Table 5.4: van der Waals Peak Distributions in Models (1Å Resolution). I=Inside, N=Near, O=Outside

Target	Points	I	I&N	N	N&O	O
ACHE	24025	272	272	4307	1191	17983
AR	16675	240	200	3582	850	11803
COX2	15341	243	238	3631	854	10375
DHFR	15525	210	270	3586	1286	10173
EGFR	18975	259	248	4035	1196	13237
$Er_{agonist}$	14283	235	191	3468	791	9598
FGFR1	22599	256	301	4199	1535	16308
FXA	27869	292	313	4442	1374	21448
GR	16875	292	247	4029	969	11338
INHA	19251	287	239	4084	1217	13424
P38	22707	213	354	3825	1941	16374
SRC	25839	220	316	3920	1656	19727

.

Table 5.5: Electrostatics Peak Distributions in Models (1Å Resolution). Only those peaks are shown that are included in the scoring process (points near the surface).

Target	Points	-	- / 0	- / +	0	0 / +	+	N
ACHE	24025	0	2	31	4	211	4791	1
AR	16675	180	628	38	2123	857	211	21
COX2	15341	262	321	287	1025	1313	920	42
DHFR	15525	59	188	406	276	2175	1111	149
EGFR	18975	20	28	231	27	2132	2238	9
$Er_{agonist}$	14283	95	333	79	2513	729	145	19
FGFR1	22599	12	29	840	30	1149	3024	18
FXA	27869	2	2	105	16	501	4604	3
GR	16875	61	531	90	2833	700	265	20
INHA	19251	128	135	272	106	3024	1096	8
P38	22707	1	21	153	17	2633	2156	9
SRC	25839	2	32	175	55	2349	2229	16

Since there were positively charged compounds in the model, active and inactive compounds were mostly differentiated by fitting into "positive" peak. This is an anomaly caused by the careless preparation of the DUD set which biases all comparisons made with electrostatic methods and DUD.

5.3.2 Scoring Algorithm

The scoring algorithm requires a previously generated model and a superimposed 3D-molecule as input data. The van der Waals volume and the electrostatic potential are generated, and each grid point is scored. First, the grid point is classified into one of three classes described in the model generation algorithm. Then, this class is compared with the peaks in the model and scored using the scoring matrices (Tables 5.6 and 5.7). The score for a field is simply the sum of values from the scoring matrices. The total score is defined as

$$S = W_v P_p + W_E P_e$$

where S is the total score; W_V is the weight for the van der Waals score; P_V is the van der Waals score; W_E is the weight for the electrostatic score; and P_E is the electrostatic score.

Table 5.6: van der Waals Scoring Matrix.

Model	**Molecule**		
Peak(s)	**Inside**	**Near**	**Outside**
Inside	1	-5	-5
Inside / Near	1	1	-5
Inside & Outside	1	-5	1
Near	-10	1	-5
Near / Outside	-10	1	1
Outside	-10	-5	1
None	1	1	1

Table 5.7: Electrostatic Scoring Matrix.

Model	**Molecule**		
Peak(s)	**-**	**0**	**+**
-	2	0	-2
- / 0	1	1	0
- / +	1	0	1
0	0	1	0
0 / +	0	1	1
+	-2	0	2
None	0	0	0

Since there are fewer points in the electrostatic score than in the van der Waals score, the latter score needs to be scaled down. In this study, arbitrary values of 0.2 for W_V and 1.0 for W_E were selected. It should be noted that these values are probably not optimal, and the weights should be modified according to the nature of the target.

The effect of grid spacing was studied on one target from each protein family that had an ROC AUC value approximately equal to 0.8 or higher at 1 Å resolution (Table 5.8). It seems that the spacing of 1 Å or 2 Å is optimal for most cases. Surprisingly, grid spacing of 5 Å still yields high ROC AUC for COX2 and FXA. This illustrates the artificial nature of the DUD data set.

Also, such crude spacing leads to several ties in the hit-list and thus complicates the selection of the top compounds.

Table 5.8: Effect of Grid Spacing on Selected Models: ROC AUCs for "Own Decoys" Sets.

	Grid Spacing			
Target	**0.5Å**	**1.0Å**	**2.0Å**	**5.0Å**
AR	0.802	0.803	0.807	0.640
COX2	0.901	0.896	0.897	0.868
DHFR	0.830	0.830	0.835	0.733
EGFR	0.805	0.798	0.814	0.691
FXA	0.914	0.915	0.915	0.907
mean	0.851	0.849	0.853	0.768
median	0.830	0.830	0.835	0.733

The classification limits can also be adjusted. The default classification limits were used in retrospective screening, since they produced the highest average ROC AUC (Table 5.9). However, the differences in accuracy between different classification limits are small. It is possible that different limits should be used for different kinds of targets, but that would require undertaking a completely new study.

Table 5.9: Different Classification Limits: ROC AUCs for "Own Decoys" Set.

Target	**0.5*limits**	**Normal**	**2*limits**
AR	0.809	0.803	0.762
COX2	0.879	0.896	0.903
DHFR	0.797	0.830	0.833
EGFR	0.738	0.798	0.810
FXA	0.911	0.915	0.919
mean	0.827	0.849	0.845
median	0.809	0.830	0.833

The orientation of molecules in the model is a critical step. This is illustrated in Figure 5.2, which shows two different FieldChopper models for AR. Both had the same crystal structure as a starting point. In one of the models, the coordinates of crystal structure were transferred to another

position. The two models have different performances, which is probably due to the differences in molecular alignments. This is a major problem with FieldChopper. However, a similar problem exists with other grid-based methods like CoMFA (Doweyko 2004).

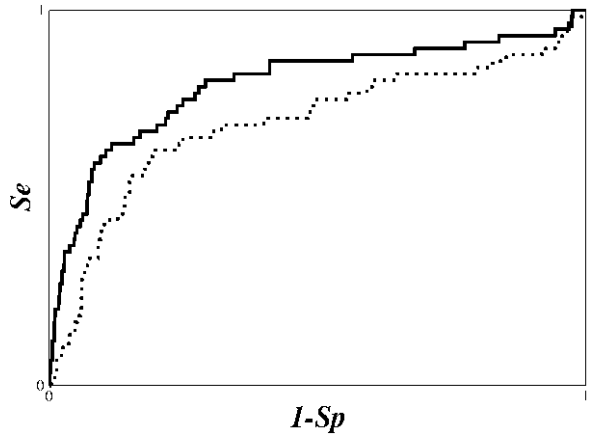


Figure 5.2: The effect of orientation of the molecules in the model. ROC curves for two different FieldChopper models (AR, own decoys set). The original model (solid line) outperforms the new model (dotted line).

5.4 RETROSPECTIVE VIRTUAL SCREENING

Since FieldChopper requires information on several active compounds instead of a single ligand, it should outperform similarity metrics which rely only on a single active conformation. EON (see Chapter 2.3.4) was selected for an example of this kind of ligand based-virtual screening method. EON has been successfully used in virtual screening to identify novel bioactive ligands (Muchmore et al. 2006; Naylor et al. 2009) and its throughput is similar to FieldChopper.

In this study, compounds were ranked with EON using the default ET_Combo score. The molecules used for alignment from Protein Databank were also used in the EON similarity calculations.

The highest ranked alignment was used as a score for the compounds. ROC curves and ROC AUCs were calculated with

the ROCR package. The Wilcoxon signed ranks test was used to assess the statistical significance of the results ($\alpha = 0.05$) (Demsar 2006). In order to evaluate early enrichment in this study, enrichment factors at the top 1% of the hit-list (EF1) were calculated. Since the top 1% of the "own decoys" set is smaller than the number of active compounds, only the enrichment factors for the "combined decoys" set are presented.

Perhaps more importantly than just producing simple enrichment is that virtual screening should reveal unique chemical structures for lead discovery i.e. be able to do scaffold hopping. However, as the ligands in the data set were not clustered, this kind of analysis was not feasible. The classification of Good cannot be applied here, as it is intended for a filtered subset of the original DUD.

FieldChopper ROC AUC averages were quite similar with both BRUTUS and ROCS superimpositions, even though there is a statistically significant difference in the combined decoys set (Table 5.10). The difference in FieldChopper's accuracy between the two algorithms could be explained by the fact that BRUTUS was used to produce the alignments for the models. It seems that FieldChopper could be used with both methods. This is not surprising, since both methods have been shown to produce reasonable molecular alignments.

Table 5.10: FieldChopper ROC AUCs with alignments from BRUTUS and ROCS. Wilcoxon signed pairs test: combined decoys set P = 0.01 < 0.05, own decoys set P = 0.5186 > 0.05.

	FieldChopper(combined)		**FieldChopper (own)**	
Target	**BRUTUS**	**ROCS**	**BRUTUS**	**ROCS**
ACHE	0.874	0.863	0.516	0.567
AR	0.930	0.898	0.803	0.810
COX2	0.909	0.867	0.896	0.891
DHFR	0.860	0.939	0.830	0.957
EGFR	0.822	0.732	0.798	0.777
$ER_{agonist}$	0.933	0.915	0.775	0.767
FGFR1	0.620	0.472	0.585	0.515
FXA	0.928	0.868	0.915	0.906
GR	0.912	0.627	0.814	0.594
INHA	0.804	0.767	0.832	0.826
P38	0.777	0.740	0.735	0.801
SRC	0.640	0.513	0.702	0.644
mean	0.834	0.767	0.767	0.755
median	0.867	0.815	0.801	0.789

ROC curves for EON and FC sets are shown in Figures 5.3 and 5.4. The early enrichment measured by enrichment factors at 1% (EF1) of a ranked database for FieldChopper and EON are shown in Tables 5.11 and 5.12. Both methods displayed a similar overall performance when one examines average and median. FieldChopper outperformed EON on nuclear hormone targets (AR, $ER_{agonist}$, and GR), whereas EON exhibited higher enrichment on ACHE and INHA. There was high enrichment on COX2, which was also reported in the previous study on EON by Nicholls and co-workers (Nicholls et al. 2004). The huge difference in performance on FXA is caused by the bias in DUD. As the ligands are charged differently than the decoys, FieldChopper can easily distinguish the ligands and decoys due to the charge alone.

Table 5.11: ROC AUCs for FieldChopper and EON. FieldChopper outperforms EON in both data sets (Wilcoxon signed rank test: combined decoys P = 0.001465 < 0.05, own decoys P = 0.02100 < 0.05).

	combined		**own**	
Target	**FieldChopper**	**EON**	**FieldChopper**	**EON**
ACHE	0.516	0.778	0.874	0.910
AR	0.803	0.676	0.930	0.677
COX2	0.896	0.884	0.909	0.878
DHFR	0.830	0.769	0.860	0.767
EGFR	0.798	0.702	0.822	0.713
$ER_{agonist}$	0.775	0.619	0.933	0.739
FGFR1	0.585	0.456	0.620	0.298
FXA	0.915	0.399	0.928	0.249
GR	0.814	0.725	0.912	0.593
INHA	0.832	0.758	0.804	0.715
P38	0.735	0.596	0.777	0.572
SRC	0.702	0.292	0.640	0.338
mean	0.767	0.638	0.834	0.621
median	0.801	0.689	0.867	0.695

Table 5.12: : Enrichment Factors at 1% of Ranked Database. Both methods have similar overall performance. FieldChopper outperforms EON on nuclear hormone targets (AR, $ER_{agonist}$, and GR), while EON has a higher enrichment factor on ACHE and INHA. The maximum enrichment factor is 100.

	combined	
Target	**FieldChopper**	**EON**
ACHE	39.79	64.47
AR	66.19	64.47
COX2	54.41	45.69
DHFR	1.08	3.76
EGFR	9.10	10.73
$ER_{agonist}$	61.62	28.88
FGFR1	0.00	0.00
FXA	4.73	0.79
GR	30.20	6.36
INHA	11.43	28.57
P38	0.42	3.32
SRC	0.72	4.29
mean	23.14	18.10
median	10.26	8.55

.

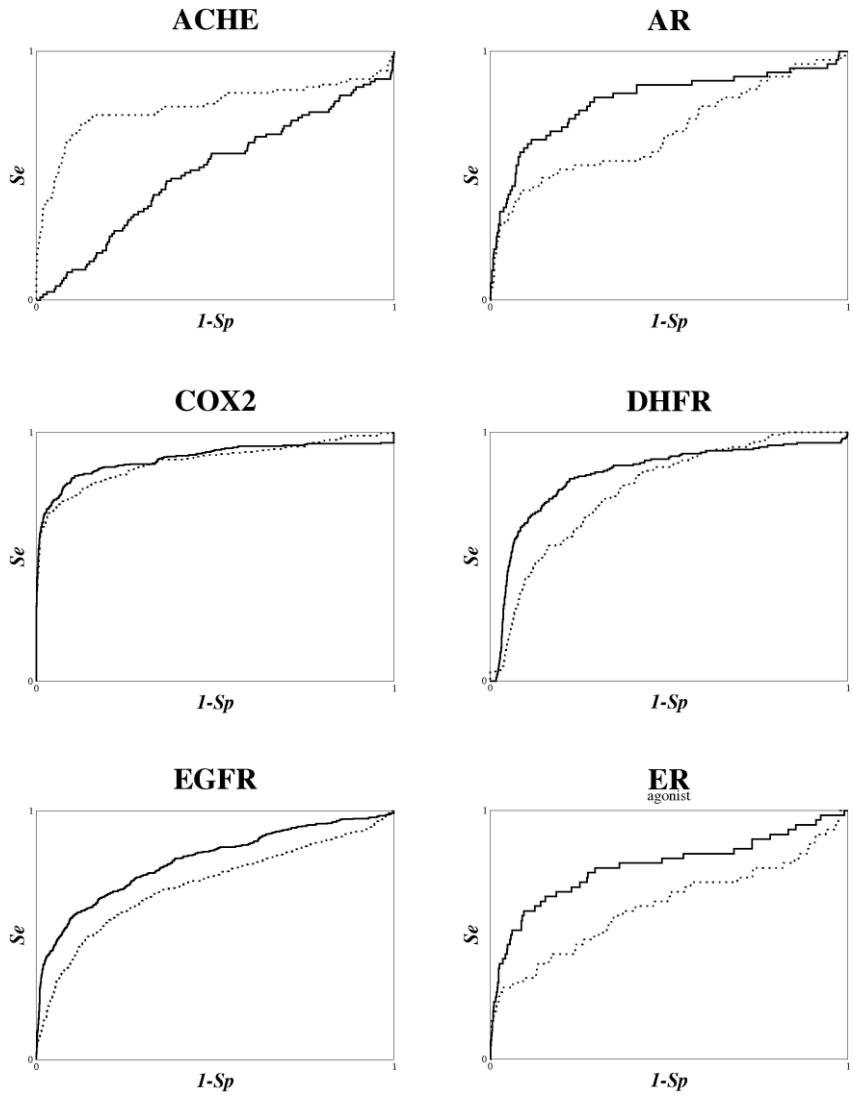


Figure 5.3: ROC curves for the own set. FieldChopper (solid line) outperforms EON (dotted line) in most cases. (1 of 2)

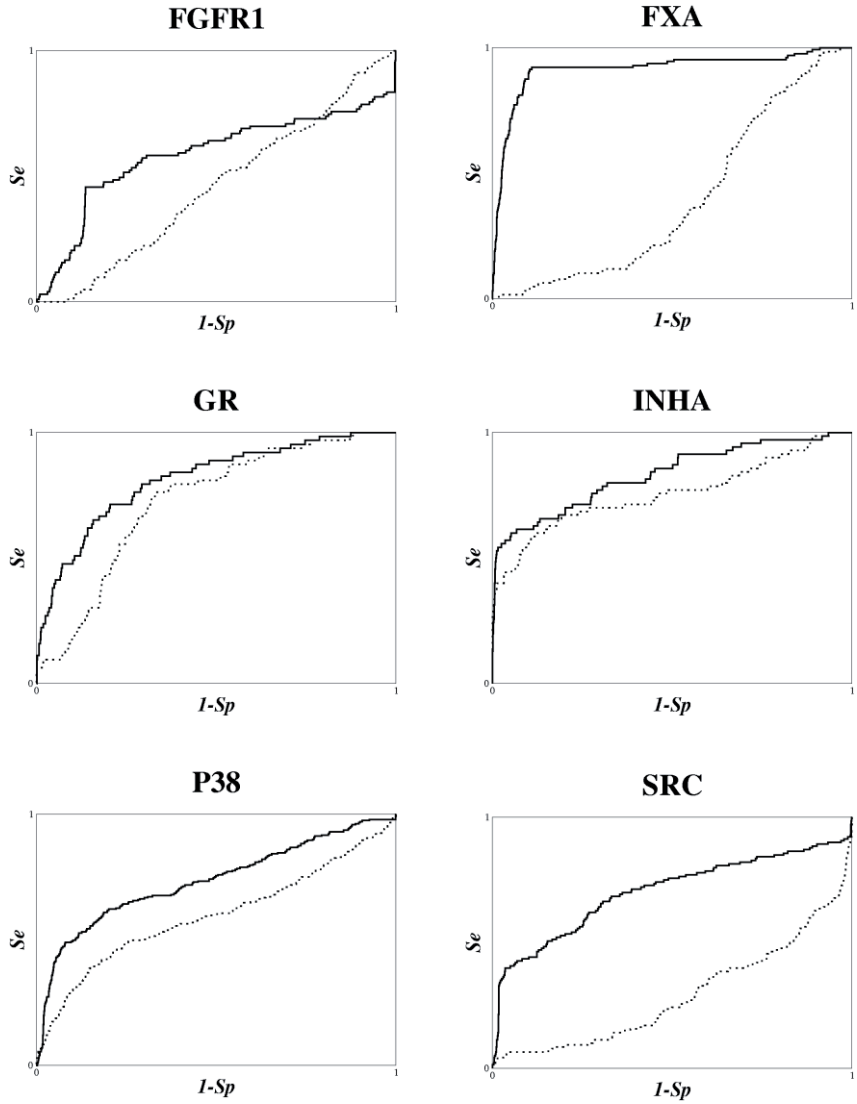


Figure 5.3: ROC curves for the own set. FieldChopper (solid line) outperforms EON (dotted line) in most cases. (2 of 2)

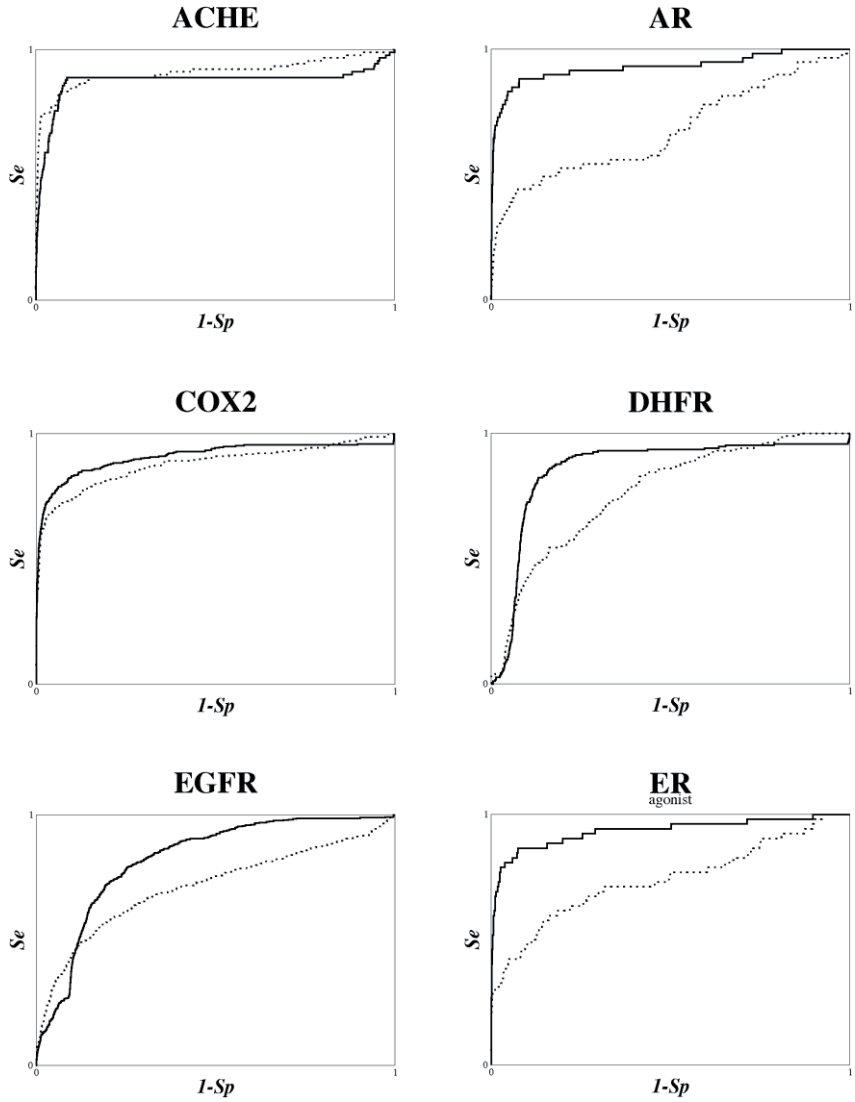


Figure 5.4: ROC curves for the combined set. FieldChopper (solid line) outperforms EON (dotted line) on most cases. (1 of 2)

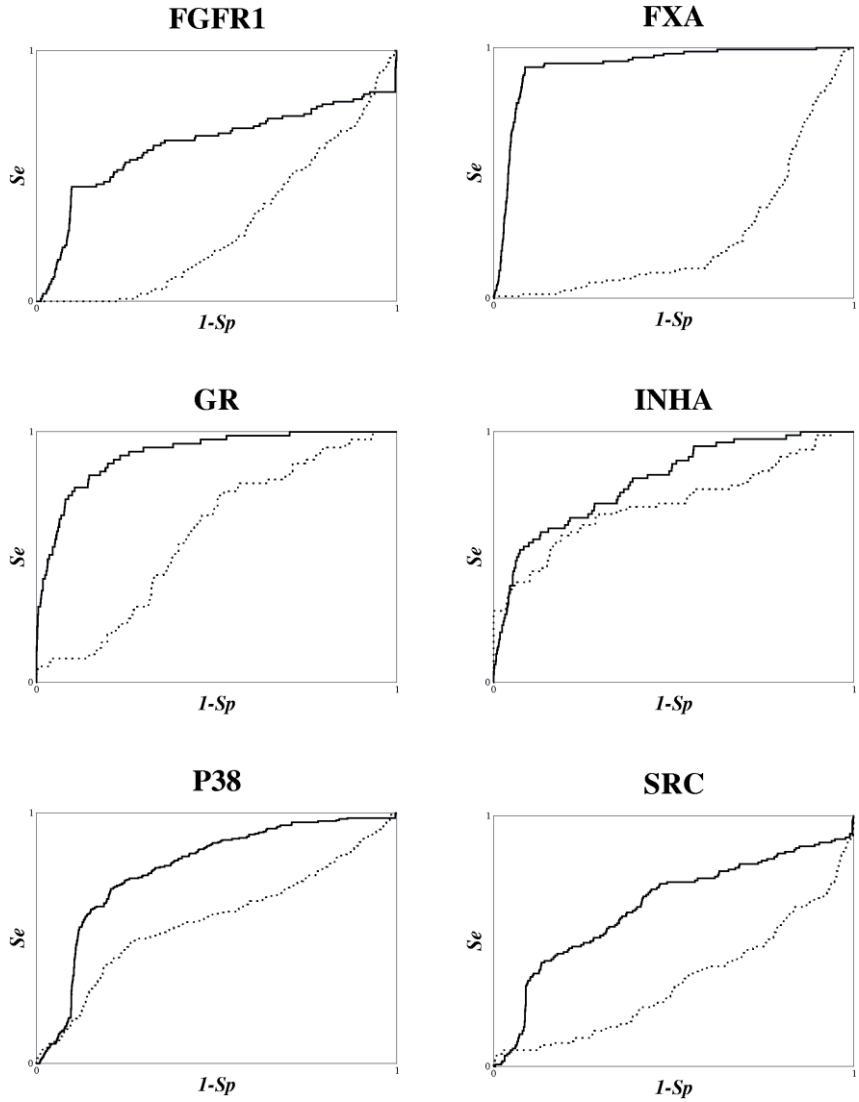


Figure 5.4: ROC curves for the combined set. FieldChopper (solid line) outperforms EON (dotted line) in most cases. (2 of 2)

6 The effect of tautomerism and protonation on SBVS

As tautomerism and ionization may significantly change the interaction possibilities between a ligand and a target protein, these phenomena could have an effect on structure-based virtual screening. However, there is very little information published on the effect of tautomeric and protonation state enumeration on the enrichment of active molecules in structure-based virtual screening. In this next chapter, the impact of this database preparation step is examined with retrospective virtual screening. The chapter is adapted with permissions from: Kalliokoski T, Salo HS, Lahtela-Kakkonen M, Poso A: The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. Journal of Chemical Information and Modeling 49: 2742-2748, 2009.

6.1 INTRODUCTION

From the computational point of view, the different tautomers of the same compound are all different molecules. Tautomeric and protonation-state enumeration ensures that the state with optimal interaction possibilities is always included in the screening process, as the predicted state may not always be the optimal binder. However, the enumeration of all possible forms of a compound increases the computing time considerably with larger data sets as a significant proportion of molecules in chemical vendor databases are tautomeric. Milletti and co-

workers analyzed four chemical databases containing 683,862 molecules for tautomerism from which 29% were tautomeric (Milletti et al. 2009).

In addition to tautomerism, protonation and ionization affect many drug molecules since many are either weak acids or bases. Predicting the correct protonation state (protomer) for a molecule is also not straightforward, as it requires estimation of pKa-values of acidic and basic functional groups in the molecule.

An alternative to enumeration is to predict the most likely tautomeric form a molecule and to discard the other forms (see Chapter 2.5). The effect of the prediction of tautomers and protomers in the virtual screening context has been surprisingly little studied. Several articles emphasize the importance of tautomers (Pospisil et al. 2003; Kirchmair et al. 2008; Martin 2009), but there are only a few studies where the effect of tautomerism on SBVS has been explicitly studied. Knox and co-workers studied the impact of various aspects of database preprocessing to SBVS using ER-alpha antagonists as the target (Knox et al. 2005). They concluded that the enumeration of tautomers increased the number of false positive compounds. That study, as well as the other studies on tautomerism and SBVS, suffers from the small data set: only one target was used to study the phenomena. The performance of a docking program is however highly dependent on the target protein (Cross et al. 2009) and thus general conclusions can only be made from a diverse set of proteins. Polgár and co-workers studied the effect of ligand protonation with beta-secretase BACE1 using several docking programs (Polgár et al. 2007). Their conclusion was that consideration of all possible protomers does not necessarily increase screening efficiency and may be just a waste of computing resources. During the preparation of this thesis, ten Brink and Exner published a study on the influence of protonation, tautomeric, and stereoisomeric states on SBVS (ten Brink and Exner 2009). Their data set was also rather limited as well; only 15 active ligands seeded with 735 properties matched decoys from ZINC. It was suggested

that enumeration of different forms was detrimental to the screening performance due to the creation of unreasonable protonation states.
This present study is the first to investigate the issue of different tautomeric and protonation states in SBVS using a diverse, publicly available data set.

6.2 PREPARATION OF THE DATA SET

6.2.1 Target Selection and Protein Structure Preparation

From the 40 targets in the original DUD data set, a subset of 28 proteins was initially selected for this study. Metalloproteins as a group were excluded from this study, since for these targets, the binding site's microenvironment for a ligand is clearly different from that of a solution or vacuum due to the presence of the metal ion (Pospisil et al. 2003; ten Brink and Exner 2009). No ligand-based tautomerism and protonation prediction method can make sensible predictions without taking into account the interaction of the protein with these targets. The current version of MoKa does not provide a solution for this problem, and thus, metalloproteins are clearly beyond the scope of this investigation. Other structures that were excluded displayed problems with the crystal structures, such as broken/missing ligands (CDK2, VEGFR2, HIVPR), covalent ligand binding (thrombin), or missing experimental details (DHFR, TK). There was also one homology model in the DUD data set (PDGFRB), which was removed from the data set.
Protein structures were downloaded from the DUD Web site (http://dud.docking.org, accessed June 1, 2009). The structures were used as is, except for the addition of hydrogen atoms with the Protonate3D method implemented in MOE (Labute 2009). The temperature and pH parameters for Protonate3D were taken from the PDB file (values from the crystallization process). Protonate3D is a method for predicting hydrogen geometry, ionization, and tautomeric states of macromolecular structures. It uses a unary quadratic optimization algorithm to optimize the

Gibbs energy of the system and to find the optimal configuration of all possible tautomeric and ionization states. The energy model used in the optimization includes van der Waals, Coulomb, solvation, rotamer, tautomer, and titration effects.

6.2.2 Ligand and Decoy Molecule Preparation

The molecular databases used in this study were built using the DUD. As DUD ligands and decoys suffer from an imbalance between charged molecules (42% ligands are charged, as compared to 15% of decoys), only molecules that had multiple forms in the MoKa suite were selected for this study (Irwin 2008). It was also expected that the effect of tautomer and protomer enumeration would be more clearly visible in this way. The disadvantage of this procedure was that it made it impossible to compare the results directly to other SBVS studies conducted on the DUD data set.

Ligands and decoys were downloaded in single enantiomer SMILES format from the ZINC database (http://zinc.docking.org, accessed June 10, 2009) (Irwin and Schoichet 2005). Two sets of molecules were generated using the MoKa suite version 1.10 (Molecular Discovery Ltd). The enumerated set contained all of the tautomeric and protonation states, whereas the predicted set included only a single form for each compound. The pH values for the MoKa predictions were taken from the PDB files. The initial three-dimensional (3D) conformations for the docking were calculated using CORINA version 3.20 (Molecular Networks GmBH).

6.3 THE DOCKING PROTOCOL

The AutoDock version 4.0 was used for docking (Huey et al. 2004; Huey et al. 2007). The program is widely used in docking studies and can also be utilized for virtual screening, provided that supercomputing resources are available (Park et al. 2006;

Jacq et al. 2008; Trott and Olson 2010). AutoDock is described in Chapter 2.4.

AutoDock is computationally exceedingly demanding for virtual screening of larger databases and, thus, requires extensive supercomputing resources. However, it is the only GNU General Public License (GPL) licensed docking program currently available. The license allows its use on both academic and commercial projects without limitations or fees.

Proteins and small molecules were prepared for docking with AutoDockTools version 1.5.4 (Sanner 1999). The docking grid was centered on the cocrystallized ligand, and default values were used for docking.

In order to validate the docking procedure, the cocrystallized ligand was redocked in a MoKa predicted form, and the RMSD between the docked and crystallized pose was calculated. The validation dockings were performed twice, since AutoDock uses a genetic algorithm which is prone to problems with sampling (ten Brink and Exner 2009). Targets that were not correctly docked were removed from the virtual screening phase. The limit for RMSD was set to 2 Å (Warren et al. 2006; Watts et al. 2010) and no significant difference between the two runs was allowed.

The results from the crystal structure dockings are presented in Table 6.1. The RMSDs of protein-ligand complexes for ACHE, AMPC, EGFR, FXA, HSP90, PARP, PPARG and trypsin were over 2 Å. GPB could not be reliably docked, as there was almost a 2 Å difference between the two runs. These targets were removed from the virtual screening phase. A total of 19 targets were selected for the SBVS phase.

Table 6.1: Root-Mean-Square Deviations (RMSD, Å) and Highest Ranking Energies (kcal/mol) from the docking validation step. The removal criteria values are in italics.

Target	**RMSD (Å)**	**energy (kcal/ mol)**	**ΔRMSD (Å)**	**Δenergy (kcal/mol)**
ACHE	*4.44*	-10.32	1.84	0.41
ALR2	0.54	-7.52	0.68	0.20
AMPC	*2.09*	-7.41	0.93	0.14
AR	0.47	-10.60	0.04	0.00
COX1	0.70	-8.31	0.17	0.01
COX2	1.39	-10.24	0.13	0.02
EGFR	*3.56*	-6.52	1.88	0.37
$ER_{agonist}$	0.70	-10.85	0.12	0.04
$ER_{antagonist}$	1.25	-13.10	0.28	0.14
FGFR1	0.97	-7.05	0.04	0.07
FXA	*2.13*	-10.11	0.33	0.91
GART	1.54	-11.57	0.03	0.45
GPB	0.62	-7.01	*1.99*	0.10
GR	0.78	-11.12	0.06	0.02
HIVRT	0.44	-9.00	0.33	0.02
HMGR	1.44	-8.34	0.08	0.42
HSP90	*4.60*	-7.12	1.81	0.03
INHA	0.47	-11.98	0.09	0.11
MR	0.58	-12.09	0.02	0.11
NA	1.89	-13.00	0.07	0.12
P38	0.95	-13.84	0.09	0.17
PARP	*2.04*	-8.20	0.01	0.04
PNP	0.42	-10.87	0.02	0.01
PPARG	*2.75*	-10.34	0.26	1.19
PR	0.87	-13.36	0.13	0.04
RXRA	0.75	-14.04	0.02	0.02
SAHH	0.65	-8.13	0.06	0.06
Trypsin	*2.52*	-6.88	1.84	0.09

.

ten Brink and Exner have previously discussed the effect of different protonated, tautomeric, and stereoisomeric forms on the pose prediction, using the high-quality CCDC/ ASTEX data set, thus the focus of this study was to evaluate enrichment on SBVS and therefore, the redocking results were not analyzed.

The issues relating to the SBVS have been discussed on Chapter 2.6.

6.4 RETROSPECTIVE VIRTUAL SCREENING

The docking time per molecule was limited in order to keep the computational time feasible. The maximum time allowed was double the time used for the cocrystallized ligand. If the molecule was not docked within the time limit, then it was removed from the assessment. Each molecule was docked 10 times to the protein, and the highest ranked pose (the one with the lowest energy) was used in the final hit list. For the predicted data set, only the predicted form was used. The dockings were calculated using a 2176 CPU Linux cluster. The numbers of molecules docked are shown in Table 6.2.

Table 6.2: Data Sets Used in Virtual Screening Experiments. N_{lig} is the number of ligand compounds, N_{dec} is the number of decoy compounds, D_{lig} is the number of different ligand forms docked, D_{dec} is the number of different decoy forms docked, F_{enum} is the fraction of ligands in the numerated set, and F_{pred} is the fraction of ligands in the predicted set

Target	pH	T (K)	N_{lig}	N_{dec}	D_{lig}	D_{dec}	F_{en} %	F_{pr} %
ALR2	6.2	273	19	677	127	4168	3.1	2.8
AR	7.9	93	63	2234	548	14465	3.8	2.8
COX1	6.7	180	6	280	22	1180	1.9	2.1
COX2	8.0	113	189	9174	610	48140	1.3	2.1
$ER_{agonist}$	8.8	103	62	1695	278	11004	2.5	3.7
$ER_{antagonist}$	7.0	100	16	896	69	3746	1.8	1.8
FGFR1	6.5	110	107	3313	2383	31659	7.5	3.2
GART	7.2	94	12	195	792	13775	5.8	6.2
GR	8.0	100	67	2241	302	9834	3.1	3.0
HIVRT	5.0	100	35	1083	197	6685	3.0	3.2
HMGR	7.5	123	31	994	130	8337	1.6	3.1
INHA	6.8	120	70	2450	279	14897	1.9	2.9
MR	7.5	100	13	467	283	4366	6.5	2.8
NA	7.8	100	48	1308	251	14066	1.8	3.7
P38	7.4	100	230	6801	2271	45532	5.0	3.4
PNP	8.0	140	22	639	834	18221	4.6	3.4
PR	6.5	100	27	835	99	3570	2.8	3.2
RXRA	7.0	93	18	463	57	2480	2.3	3.9
SAHH	5.6	100	31	817	521	21343	2.4	3.8

.

The SBVS results were evaluated using two commonly used metrics: Enrichment factor (EF) and ROC AUC. ROC AUCs were calculated using ROCR (Sing et al. 2006). As there are relatively few decoys per ligand in the data set, the cutoff of 5% was selected as the enrichment factor. This means that the maximum enrichment factor is 20.

The enrichment metrics are shown in Table 6.3, and the ROC curves are shown in Figure 6.1. There is no major difference between the enumerated and the predicted sets in terms of the enrichment metrics. This finding is in line with the previous SBVS studies on the effect of tautomerism and protonation that

used single targets (Pólgar et al. 2007; ten Brink and Exner 2009). However, there is a vast difference in computing time per compound. With a rather slow docking method, like AutoDock, the extra time spent in considering the enumerated tautomers and protomers becomes quickly very significant. The enumeration can also be an issue with large databases incorporating millions of compounds. Therefore, the use of a single, reasonable form of the molecule for structure-based virtual screening of molecular databases is recommended.

*Table 6.3: ROC AUCs, EFs at 5% (EF5) and Mean Times Per Compound Used in minutes (t) for Docking Shown for the Enumerated and the Predicted Set. There is no statistical difference (as measured by Wilcoxon signed rank test) between ROC AUCs between the two sets (Demsar 2006). *=For GART, 255 runs was also used instead of the standard 10, illustrates the sampling problem.*

	Enumerated set			**Predicted set**		
Target	**ROC AUC**	**EF_5**	**t**	**ROC AUC**	**EF_5**	**t**
ALR2	0.516	1.05	50	0.538	3.14	8
AR	0.712	4.44	46	0.719	4.12	8
COX1	0.390	3.41	36	0.334	0	8
COX2	0.851	8.36	52	0.877	9.95	10
$ER_{agonist}$	0.868	10.95	50	0.883	10.63	8
$ER_{antagonist}$	0.862	16.11	59	0.841	14.87	14
FGFR1	0.321	1.12	132	0.381	1.12	14
GART*	0.881	6.90	793	0.639/ 0.859	5.18/ 10.35	17/ 690
GR	0.598	4.19	38	0.647	5.09	10
HIVRT	0.363	2.28	55	0.395	1.71	9
HMGR	0.886	5.84	106	0.829	5.19	13
INHA	0.389	4.57	65	0.423	4.57	11
MR	0.845	3.08	87	0.765	0	9
NA	0.838	7.48	107	0.843	8.31	10
P38	0.585	3.21	73	0.554	1.65	11
PNP	0.528	4.55	248	0.516	4.55	9
PR	0.630	7.43	36	0.634	5.20	9
RXRA	0.969	11.13	65	0.967	11.13	12
SAHH	0.418	0	216	0.548	0.65	9
mean	0.655	5.58	122	0.649	5.11	11
median	0.630	4.55	65	0.639	4.57	10

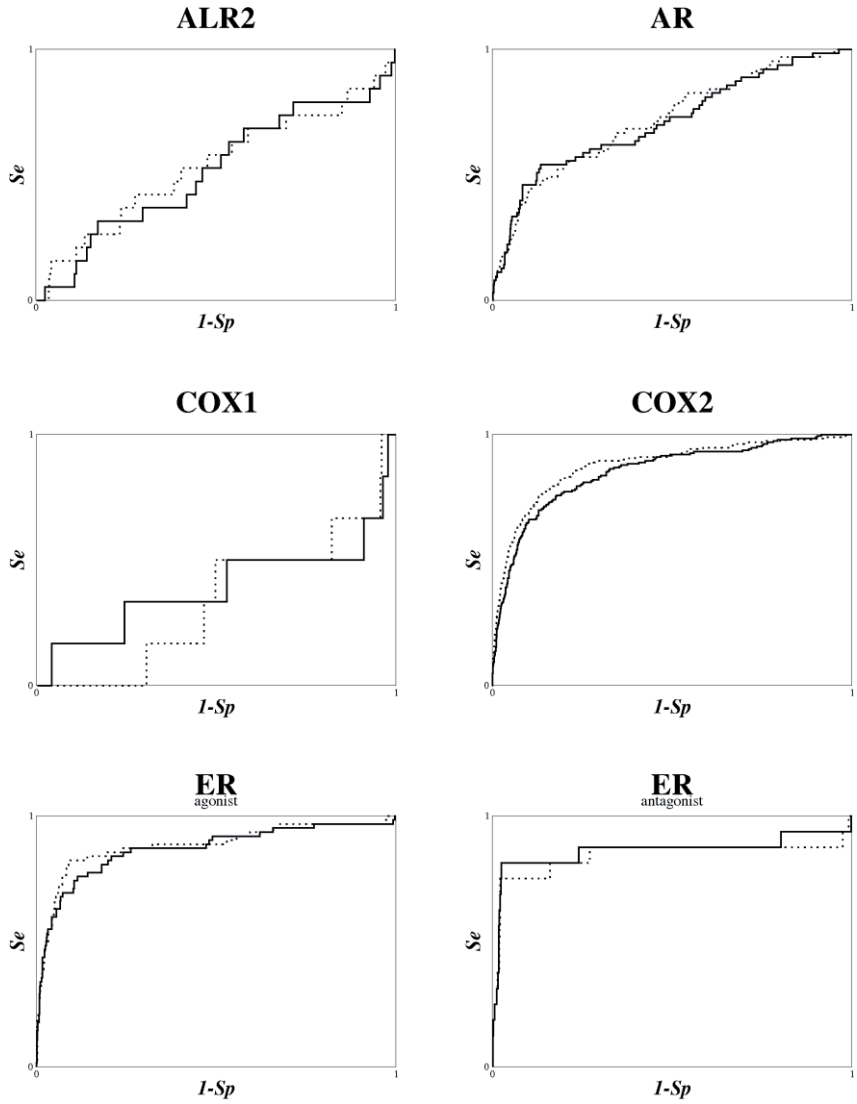


Figure 6.1: ROC curves for the enumerated (solid line) and predicted (dotted line) sets. For GART, the predicted set performance increased significantly (dashed line) with increased sampling. (1 of 4)

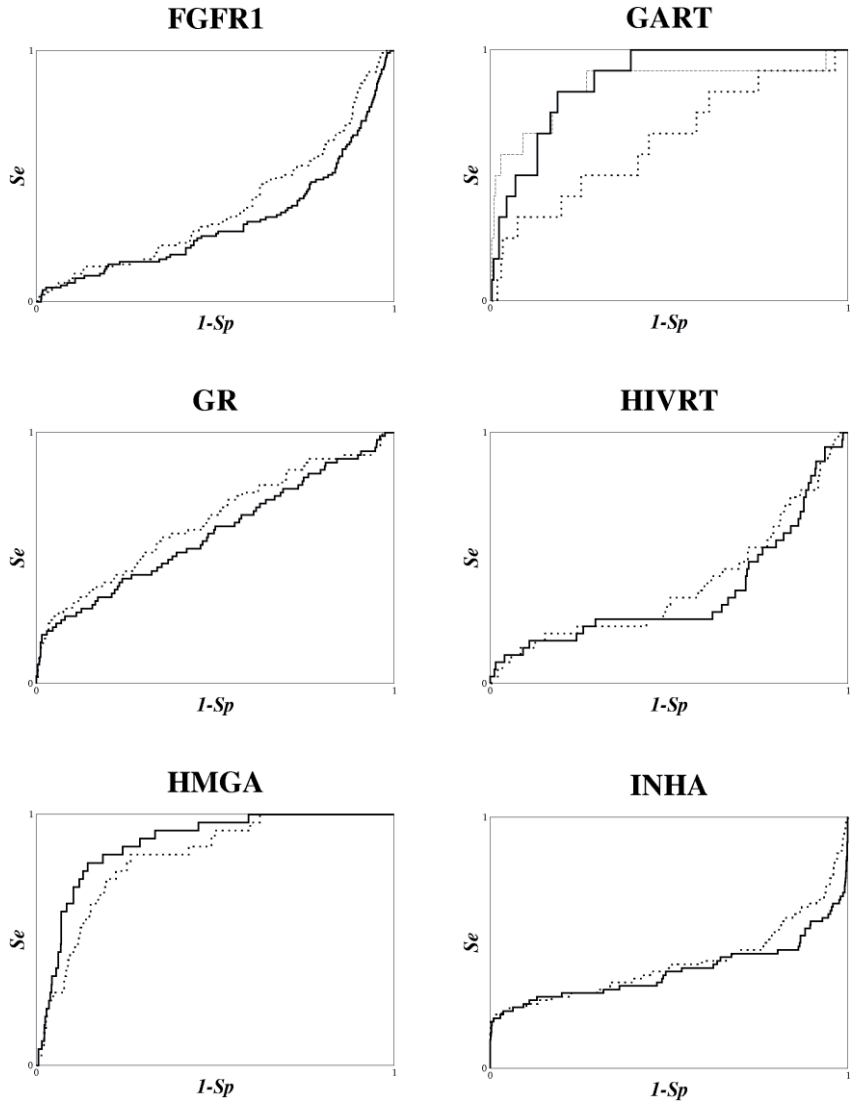


Figure 6.1: ROC curves for the enumerated (solid line) and predicted (dotted line) sets. For GART, the predicted set performance increased significantly (dashed line) with increased sampling. (2 of 4)

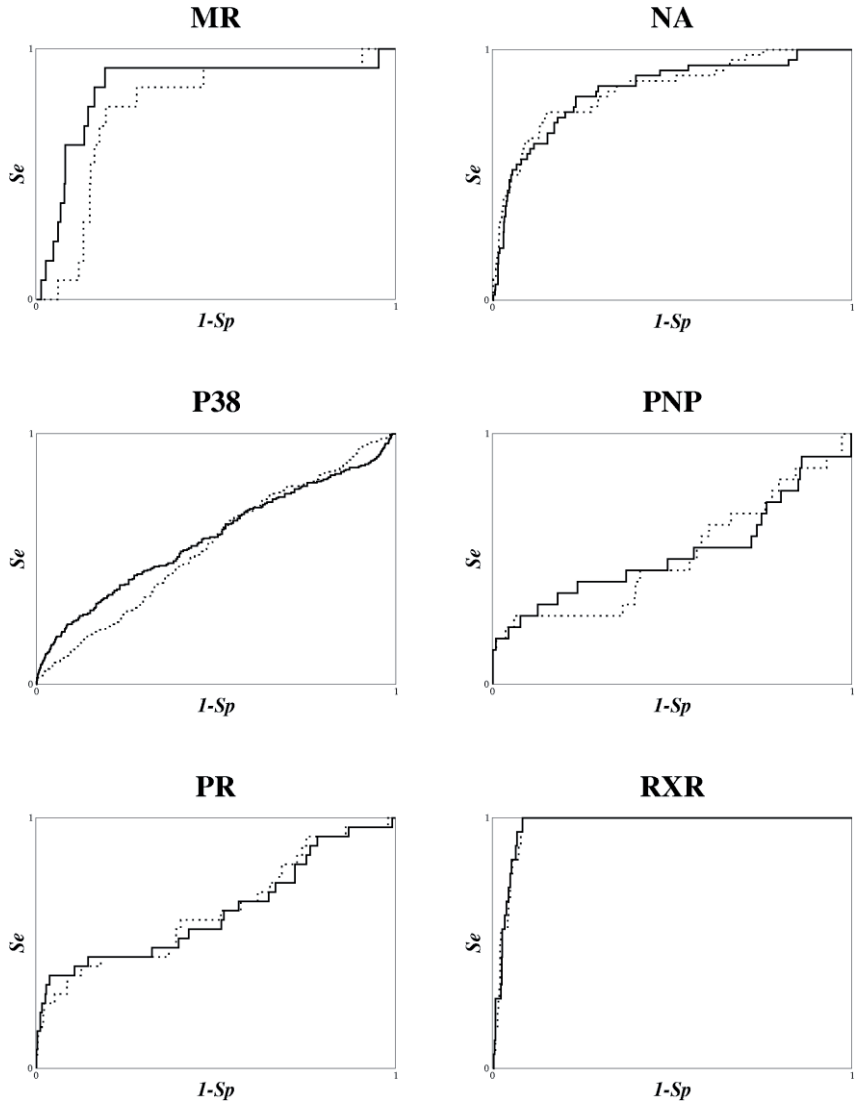


Figure 6.1: ROC curves for the enumerated (solid line) and predicted (dotted line) sets. For GART, the predicted set performance increased significantly (dashed line) with increased sampling. (3 of 4)

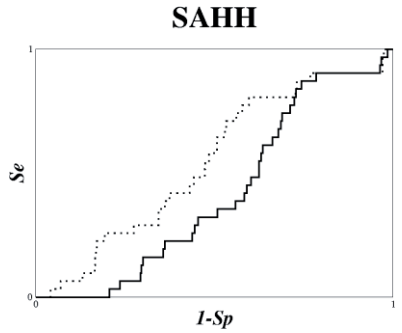


Figure 6.1: ROC curves for the enumerated (solid line) and predicted (dotted line) sets. For GART, the predicted set performance increased significantly (dashed line) with increased sampling. (4 of 4)

There is a striking difference between the enumerated and the predicted set on GART. The potential reasons for this are inadequate sampling and exceptionally large numbers of different tautomers and protomers per compound on this target. In order to determine if this is really the case, the predicted set was redocked with an increased number of docking runs. The number of LGA runs was increased from 10 to 255. The ROC AUC of the predicted set rose from 0.639 to 0.859, which is much closer to the ROC AUC of the enumerated set (0.881). The enumerated set was not redocked with the new settings because this would have been computationally extremely demanding, requiring approximately 18 CPU years.

Even though the data set used here is not the DUD itself, some rough comparisons can be made to other benchmarking studies which have utilized DUD. Cross and co-workers compared several commonly used docking methods with the whole 40 protein data set from the DUD (Cross et al. 2009). The mean ROC AUC values for the whole 40 protein data set varied from 0.55 to 0.77, depending on the docking method. As the mean ROC AUC for the data sets in this study is approximately 0.65, AutoDock's performance seems to be typical for a docking program. There are 9 cases out of 19 where clear enrichment can be seen (ROC AUC > 0.70). ER, NA and RXRA have been shown previously to represent easy targets for docking programs, whereas the kinase targets FGFR1 and P38 are very challenging

for the current docking programs (Huang et al. 2006; Cross et al. 2009).

It has been suggested that the number of false positives may increase on the enumerated set due to the strongly charged and unlikely forms of decoy molecules that receive high scores (ten Brink and Exner 2009). The mean energies for ligands and decoys were calculated to verify this assumption (Table 6.4). It can be seen that the energy difference is usually larger for the decoys between the enumerated and the predicted sets than with that of ligands. The effect is more clearly visible on those targets where there is good enrichment ($ER_{agonist}$, $ER_{antagonist}$, COX2, GART, and RXRA). However, this change in the energy differences is so small that it does not translate into any major differences between the ROC AUC values of the two sets.

*Table 6.4: Average Energies from Docking Results. *=the predicted data set energies are from the 255 LGA run.*

	Enumerated set		**Predicted set**			
Target	**E_{lig}**	**E_{dec}**	**E_{lig}**	**E_{dec}**	**ΔE_{lig}**	**ΔE_{dec}**
ALR2	-7.35	-7.53	-6.99	-7.07	0.36	0.46
AR	-9.44	-8.40	-9.08	-7.83	0.36	0.57
COX1	-7.05	-7.42	-6.66	-7.18	0.39	0.24
COX2	-9.79	-8.54	-9.60	-8.07	0.19	0.47
$ER_{agonist}$	-8.63	-7.50	-8.35	-7.07	0.28	0.43
$ER_{antagonist}$	-11.58	-9.32	-11.26	-8.83	0.32	0.49
FGFR1	-7.20	-7.59	-6.47	-6.77	0.73	0.82
GART*	-11.33	-8.96	-11.54	-8.61	0.24	0.35
GR	-9.34	-8.79	-9.07	-8.34	0.27	0.45
HIVRT	-8.68	-8.96	-8.09	-8.39	0.59	0.57
HMGR	-7.96	-6.63	-7.19	-6.01	0.77	0.62
INHA	-8.67	-8.86	-8.43	-8.46	0.24	0.45
MR	-10.37	-8.91	-9.42	-8.46	0.95	0.40
NA	-8.43	-6.78	-7.87	-6.04	0.56	0.74
P38	-9.78	-9.52	-9.26	-9.05	0.52	0.47
PNP	-7.56	-7.49	-6.83	-6.70	0.73	0.79
PR	-8.95	-8.34	-8.65	-8.03	0.30	0.31
RXRA	-12.20	-8.64	-11.98	-8.23	0.22	0.41
SAHH	-7.19	-7.57	-6.26	-6.03	0.93	1.54
mean	-9.03	-8.20	-8.38	-7.54	0.47	0.56
median	-8.68	-8.40	-8.35	-7.83	0.36	0.47

As revealed in the study of Warren and co-workers (Warren et al. 2006), the docking programs may be capable of reproducing crystal structures and identifying active molecules from a pool of inactive molecules, but they are not able to rank properly closely related molecules. Tautomers and protomers of a molecule can be considered as closely related molecules from the docking program's point of view. The accuracy of the scoring functions might not be sufficient to separate different tautomers and protomers in virtual screening programs. Todorov and co-workers studied the dependence of docking results on the tautomeric and protonation states of the ligand on three protein-ligand complexes (Todorov et al. 2006). The

differences in the protonation pattern occurred at positions where they only had a limited impact on the binding energy, and also the flexible bonding groups permitted a greater number of hydrogen bonds to be formed than were found in the crystal structures. It was concluded that ligand binding is rather insensitive to changes in the tautomeric and protonation states. This could also explain the small difference between the enumerated and the predicted set observed in this study.

7 GPUs and single conformation databases in LBVS

Recently, a GPU-version called PAPER of widely-used ROCS algorithm was published (Paque and Hande 2010). In order to study the applicability of the PAPER algorithm to ligand-based virtual screening, a command line interface for the PAPER algorithm intended to facilitate virtual screening was developed and the effect of conformation analysis of both query and database molecules was investigated. The chapter is based on the following publication: Kalliokoski T, Rönkkö T, Poso A: Increasing the throughput of shape-based virtual screening with GPU processing and single conformation databases. Molecular Informatics 29: 293-296, 2010.

7.1 INTRODUCTION

Shape-based virtual screening is based on the ranking of molecules according to their shape similarity (see Chapter 2.3.2). As small molecules are typically flexible and can adopt several conformations, conformational analysis is required before the screening process. Usually a single conformation that is assumed to be the "bioactive" conformation i.e. the conformation observed in the protein-ligand crystal complex is used for the query, whereas a conformational ensemble is created for the database molecules. There are several arguments why this might not be the optimal approach for shape-based virtual screening. The use of an *in silico* conformation instead of

one from an X-ray crystallography model has been shown to produce equal results (Hawkins et al. 2007). However, as ligands and proteins are usually rather flexible structures, they often have several possible conformations and there is no simple way to say which of the conformations are bioactive from all the possibilities (Borodina et al. 2007; Watts et al. 2010). It is therefore reasonable to propose that multiple conformations for the query molecule could improve the results as the risk of overlooking the bioactive conformation would be decreased. On the other hand, virtual screening is simply similarity searching. The task is to find molecules that are similar to the query, not to predict how they might bind to the target protein. The use of single conformation databases might be also therefore feasible.

Given that shape-based virtual screening is widely used in industry and academia, the effect of different conformational analysis approaches on accuracy has been surprisingly little studied. Tawa and co-workers proposed technique called CORAL (Conformational analysis, Rocs Alignment) (Tawa et al. 2009). It is based on the assumption that the ligands share the same conformational space, which contains the bioactive conformations. This is illustrated in Figure 7.1, where the five compounds have different individual conformations, but they share the bioactive conformation space.

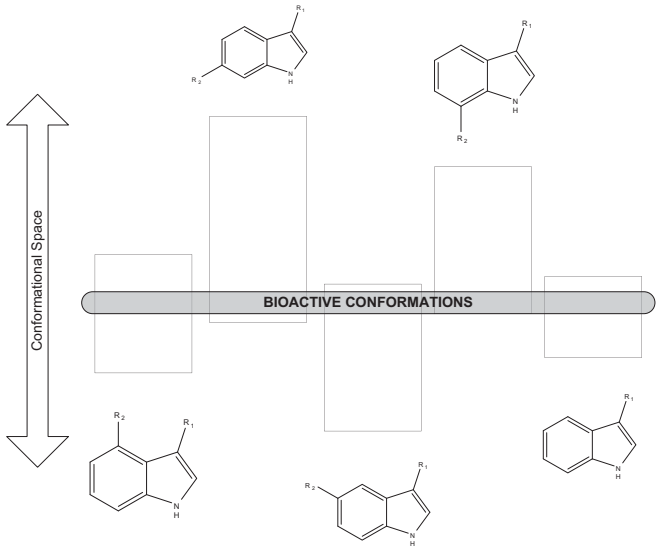


Figure 7.1: The idea of CORAL (adapted from Tawa et al. 2009).

CORAL requires multiple known ligands. First, a conformational expansion of each ligand is performed. It is assumed that all ligands share the same binding mode. Then, every conformation of every ligand is superimposed with each other to form matrix *O* of the similarity scores:

$$O = \begin{bmatrix} 1 & O_{12} & \dots & O_{1j} \\ O_{21} & 1 & \dots & O_{2j} \\ \dots & \dots & \dots & \dots \\ O_{i1} & O_{i2} & \dots & 1 \end{bmatrix}$$

where the O_{ij} is the similarity score (see Chapter 2.3.4) with between conformations i and j. Vector A is calculated from matrix *O* by averaging every row:

$$A = \begin{bmatrix} \frac{1}{N}\sum_{j=1}^{N} O_{1j} \\ \frac{1}{N}\sum_{j=1}^{N} O_{2j} \\ \cdots \\ \frac{1}{N}\sum_{j=1}^{N} O_{ij} \end{bmatrix}$$

where N is the total number of conformations. CORAL query with index i_{CORAL} is identified from vector A:

$$i_{CORAL} = \max [A]$$

The authors of CORAL compared the virtual screening performance of ROCS with minimum energy conformation against CORAL conformation using DUD as the data set. It was concluded that using CORAL conformation instead of minimum energy conformation could be beneficial, as, on average, ROC AUC increased from 0.835 to 0.842. The obvious disadvantage of CORAL is that it requires multiple known ligands that are assumed to bind in exactly same binding mode. When using each ligand separately as a single query, this assumption is not required and as since shape based virtual screening is quite fast, it is not computationally prohibitive strategy either. As an example, the 15 actives used in FieldChopper models were used as queries for EON and superior enrichments were observed compared to the single query results (Table 7.1).

Table 7.1: ROC AUCs for EON with 15 query molecules per target using the datasets from the FieldChopper study.

Target	ROC AUC
ACHE	0.827
AR	0.912
COX2	0.968
DHFR	0.977
EGFR	0.922
$ER_{agonist}$	0.905
FGFR1	0.972
FXA	0.930
GR	0.976
INHA	0.942
P38	0.981
SRC	0.906
mean	0.935
median	0.936

Also, if one assumes that several ligands can be superimposed meaningfully, then one could also use pharmacophore methods and overcome the limitations of total similarity scoring.

Kirchmair and co-workers studied the effect of using conformational ensembles as queries using ROCS and the DUD data set (Kirchmair et al. 2009). In this approach, all conformations of the query molecule are scored against the database molecule and the highest scored pair is retained. The 40 targets in DUD were screened using the PDB co-crystallized ligand as the query molecule. One, three, five and ten conformations were used for each query molecule. It was concluded that the use of multiple query conformations did not increase the virtual screening accuracy of ROCS significantly, as the average ROC AUCs for the screens with different number of query conformations increased from 0.72 to 0.74. However, this study has some issues. Firstly, the use of fixed numbers of query conformations between one and ten for all query molecules is problematic, as the number of *reasonable* and *diverse* conformations per molecule is clearly different. This can be seen from the number of conformations generated by ConfGen using

the FAST-setting (see Chapter 2.5.3 for details). As illustrated in Figure 7.2, for half of the queries, ten conformations per molecule might not be enough, while 25% of query molecules have less than ten reasonable conformations. This leads to situation where in most cases, the number of conformations for the query molecule might not optimal. One should use *all* of the query's conformations and not use arbitrary cutoffs. Secondly, the query molecule from the PDB crystal structure in some cases produces extremely poor ROC AUC, which is related to the data set composition: the query is simply too different from the ligands. In such cases, it is unreasonable to expect the use of multiple query conformations to overcome this fundamental limitation of LBVS (see Chapter 2.6). It would be better to use all of the ligands as a query instead using just a single molecule when analyzing the effect of the number of query's conformations. Obviously, addressing these two issues would increase the computing time considerably and supercomputing facilities would be required.

There is also the question of whether or not the use of multiple conformations for the database molecules increases the accuracy of virtual screening. The use of single conformation databases would significantly reduce the computing time and permit the screening of larger databases compared to the multi-conformation databases. This strategy has not been studied on shape-based virtual screening, but Renner and co-workers have compared single and multi-conformation databases with CATS3D descriptors (Renner et al. 2006). In their study, the use of multi-conformation database only slightly improved the virtual screening performance and therefore the extra effort of conformational analysis was not justified.

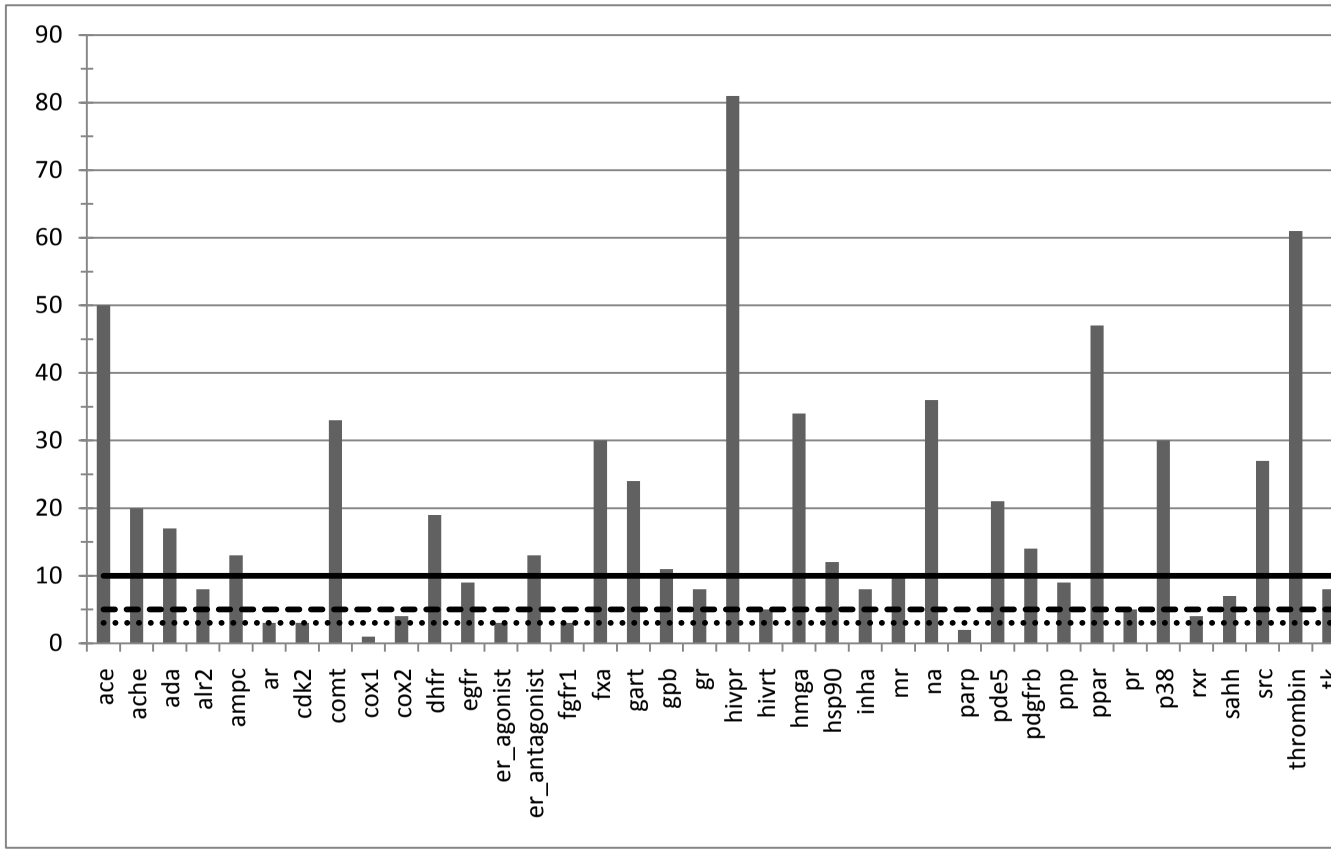


Figure 7.2: The number of diverse conformations for the queries used in the Kirchmair study. The horizontal lines are the selected cutoffs f number of conformations used in the screening. Conformations were generated with ConfGen using the 'FAST' preset (Schrödinger Inc).

7.2 DEVELOPMENT OF COMMAND-LINE INTERFACE FOR PAPER

Even though the source code for PAPER is freely available, the available version is not directly suitable for virtual screening, as it only outputs a 4x4 transformation matrix and does not handle multiple molecules. Therefore, a user interface is required before the algorithm can be used for virtual screening.

The PAPER GPU kernel was wrapped into a command-line interface named WeedyControl for PAPER (WCPAPER) without modifying the algorithm itself at all and by using as much as possible of the OpenBabel library. The simplified flowchart of the program is shown at Figure 7.3. First, the query molecule is loaded into RAM. Since the memory capacity varies between different GPU hardware, there is an adjustable parameter GPU_MOLS, which controls the number of molecules kept in VRAM at one time. After the template and database molecules have been read into VRAM, the molecules are aligned with PAPER. The overlap volumes and transformation matrices are copied from VRAM to RAM for similarity scoring and optional alignment output. ShapeTanimoto similarity S between molecules A and B is calculated from (Haque and Pande 2010):

$$S = \frac{O_{AB}}{(O_{AA} - O_{BB} + O_{AB})}$$

where O_{xy} is the overlap volume between molecules x and y, calculated by PAPER using a spherical Gaussian function. The 4x4 transformation matrix M generated by PAPER contains both the translation and rotation (Haque and Pande 2010):

$$M = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rotation matrix *R* and translation vector *T* are formed:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{r_{11}} & \boldsymbol{r_{12}} & \boldsymbol{r_{13}} \\ \boldsymbol{r_{21}} & \boldsymbol{r_{22}} & \boldsymbol{r_{23}} \\ \boldsymbol{r_{31}} & \boldsymbol{r_{32}} & \boldsymbol{r_{33}} \end{bmatrix}$$

$$T = [t_x \quad t_y \quad t_z]$$

Matrix *R* and vector *T* are applied to every atom of the molecule using OpenBabel by first applying Rotate()-function and then the Translate()-function in OBMol-class. Finally, the ShapeTanimoto scores and aligned molecules are outputted.

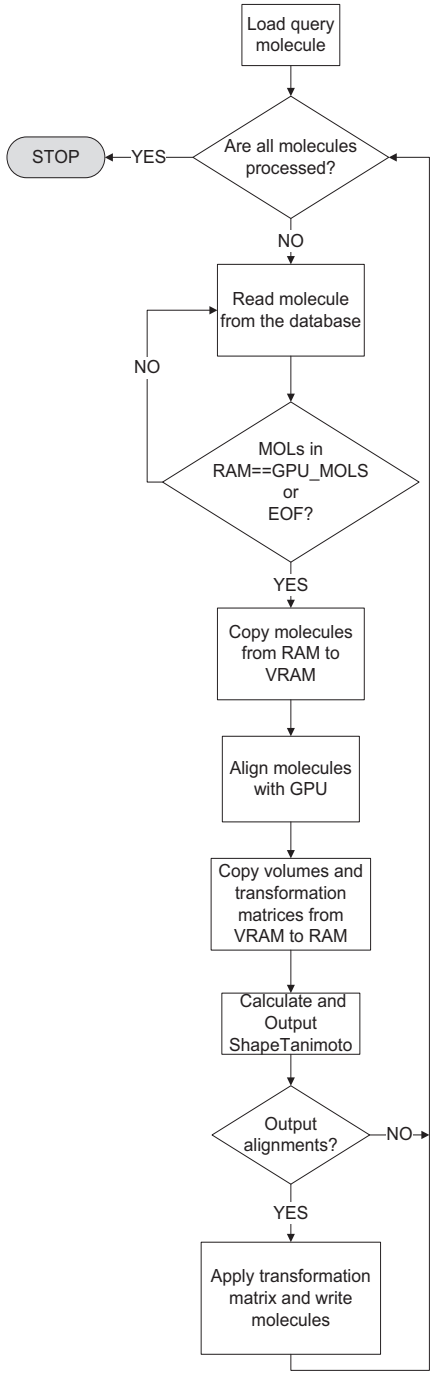


Figure 7.3: Simplified flowchart of WCPAPER.

PAPER algorithm assumes that molecules have been previously oriented by the singular value decomposition of the point cloud made up of the atom centers. This SVD-based preprocessing step is handled externally in a Python script, as it is only required once for each of the database molecules. The script supplied with the public version of PAPER uses the commercial OEChem library. The code was modified to use OpenBabel instead.

As the number of starting positions is directly linked to the execution time of optimization algorithm, a low number of diverse positions would be preferable. There are nine different modes implemented in the PAPER code with an increasing number of starting positions *n*, although only four of them are described in the publication (Table 7.2). The cyclic translations in modes seven and eight are determined by a procedure that first decomposes the molecule into cyclic and acyclic components. Then, the centroid of each cyclic component is used as a translational starting point. Mode 1 is used for the alignment process in WCPAPER, as was recommended for virtual screening applications by Haque and Pande.

Table 7.2: Different initialization modes implemented in PAPER. The mode 1 is used in WCPAPER (the original mode proposed by Grant and co-workers).

Mode	n	Description
0	1	Inertial overlay
1	4	Mode 0 + 180° degree rotations around each axis
2	12	Mode 1 + 90° degree rotations around each axis
3	68	Mode 1 + moving the center of molecule of each molecule onto a corner of the bounding box of the other
4	204	Mode 2 + moving the center of molecule of each molecule onto a corner of the bounding box of the other
5	30	30 random orientations
6	12	12 random orientations
7	4*RS	180° degree rotations around each axis for each cyclic translation
8	12*RS	90° and 180° degree rotations around each axis for each cyclic translation

In order to find the optimum for GPU_MOLS, COX2 data set was screened on two different computers with four different values (100, 1000, 5000 and 10000) using a typical query molecule with 22 heavy atoms (Table 7.3). The value of 1000 seems optimal, as one must take into account the fact that some databases may have larger molecules that consume more memory than those in the COX2 data set. The value of 100 is recommended for graphics adapters with little memory, such as those found in laptop computers.

Table 7.3: Running times in seconds for WCPAPER on two different systems and four GPU_MOLS values. There was not enough memory in 8800GT for 10000 molecules. OS=Operating System, GA=Graphics Adapter

OS	CPU	GA	Cores	VRAM	100	1k	5k	10k
Linux (CentOS 5.4, 64bit)	2.67 Ghz Intel Core i5	Nvidia GeForce 295GTX	2 x 240	2 x 896 MB	273	158	158	159
Mac OS X (10.6)	2.8 Ghz Intel Xeon	Nvidia GeForce 8800GT	112	512 MB	298	235	230	-

CPU/GPU architecture, system libraries and compilers can influence virtual screening accuracy: sometimes even incorrect results are produced (Feher and Williams 2009). Since GPU computing has been only recently introduced and both hardware and software are changing very rapidly, it is likely that GPU applications will be especially vulnerable for such anomalies. The impact of the computing platform is visible on WCPAPER. The two different computers used in performance testing produced slightly different ShapeTanimoto values (Figure 7.4). Even though these differences are quite subtle, they clearly have an effect on the virtual screening performance (Figure 7.5). As the source code is exactly same in both cases, this difference originates from hardware, system software or compilers. It is therefore important to use the same platform in comparative studies.

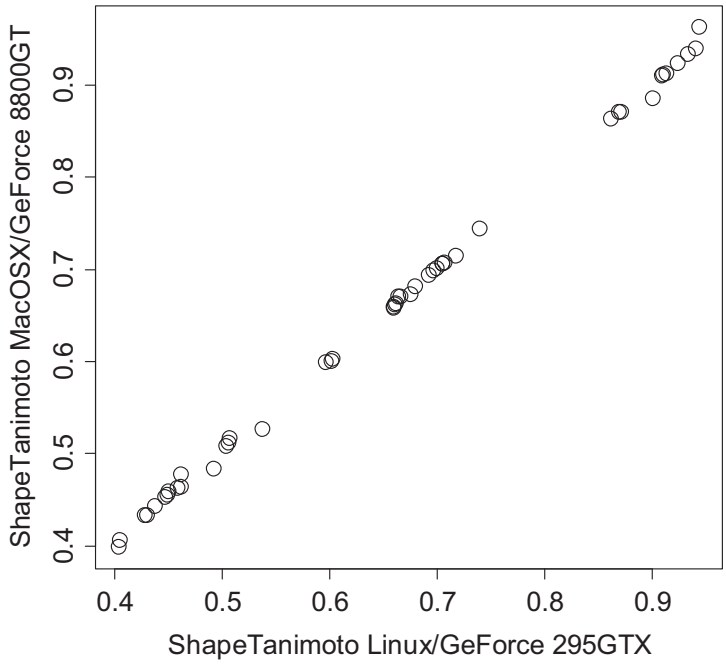


Figure 7.4: The ShapeTanimoto values from two different computer platforms. Pearson correlation coefficient is 0.999481.

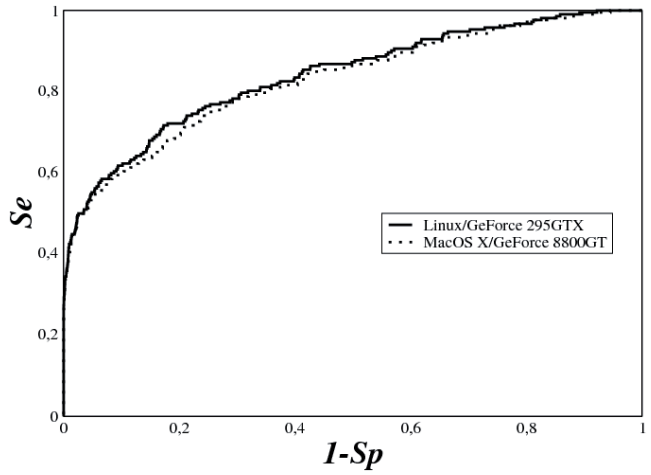


Figure 7.5: ROC curves produced by the two computers.

7.3 PREPARATION OF THE DATA SET

The data set used in this study was built from DUD LIB VS (see Chapter 3.1). The ligand and decoy molecules with just a single conformation were removed in order to amplify the effect of different conformational approaches. The number of conformations for the molecules is shown in Table 7.4.

Table 7.4: The number of molecules and conformations for each of the targets used in the study.

Target	S	L	LC	C/L	D	DC	C/D
ACE	19	45	615	14	1722	25703	15
ACHE	16	96	1458	15	3651	46685	13
ADA	8	23	283	12	809	9483	12
ALR2	13	21	123	6	847	6229	7
AMPC	6	21	203	10	695	5136	7
AR	6	37	163	4	2346	13590	6
CDK2	31	46	431	9	1758	23900	14
COMT	2	11	222	20	395	4122	10
COX1	10	21	66	3	754	4189	6
COX2	39	190	851	4	11375	105895	9
DHFR	14	184	1383	8	7014	108543	15
EGFR	40	348	2496	7	14297	164230	11
$ER_{agonist}$	10	60	275	5	2150	14874	7
$Er_{antagonist}$	8	18	317	18	1016	17401	17
FGFR1	12	64	389	6	3186	61218	19
FXA	19	64	1370	21	1883	44928	24
GART	5	8	240	30	118	3641	31
GPB	9	51	919	18	1824	27287	15
GR	8	23	109	5	2233	14731	7
HIVPR	3	4	73	18	9	253	28
HIVRT	12	27	215	8	1388	13317	10
HMGA	4	25	557	22	1192	22955	19
HSP90	4	23	243	11	849	11475	14
INHA	23	57	550	10	2436	24308	10
MR	2	8	36	4	496	3525	7
NA	7	49	402	8	1580	26406	17
P38	19	114	862	8	5883	52815	9
PARP	6	16	67	4	915	4550	5
PDE5	21	25	163	7	1550	23591	15
PDGFRB	21	102	627	6	5209	61716	12
PNP	3	23	149	6	812	9050	11
PPARG	6	6	97	16	38	716	19
PR	3	19	60	3	772	4054	5
RXRA	3	18	85	5	545	6077	11
SAHH	2	33	390	12	1124	14393	13
SRC	20	86	638	7	5206	86405	17

Thrombin	13	23	361	16	1059	27985	26
TK	7	22	236	11	757	7373	10
Trypsin	6	8	117	15	663	17795	27
VEGFR2	25	40	408	10	2466	34805	14

Four different conformational analysis approaches were investigated: single conformation query with single conformation database (SINGLE_SINGLE, SS), single conformation query with multi-conformation database (SINGLE_MULTI, SM), multi-conformation query with single conformation database (MULTI_SINGLE, MS) and multi-conformation query with multi-conformation database (MULTI_MULTI, MM). From these, SINGLE_MULTI is the most commonly used methodology. Single conformations were generated with MacroModel version 9.7 using OPLS_2003 force-field and multiple conformations were calculated with ConfGen version 2.1 using the 'FAST' preset and an energy cut-off of 25kcal/mol was applied.

7.4 RETROSPECTIVE VIRTUAL SCREENING

Every ligand was used as a query one at a time for the screening. The query was removed from the ligands and the highest scored similarity value was used for each of the molecules in the database.

ROC AUC values were used to measure the accuracy. As the ROC AUC measures overall performance and does not take into account the early enrichment or the chemical diversity of the hit molecules, the fractions of the possible scaffolds retrieved were also calculated (see Chapters 3.2 and 3.3).

As the targets in DUD all have different molecules, the results were also analyzed by calculating median values for each of the targets. The median was used instead of average, because it was expected that there are some queries in every target that perform either exceptionally well or poorly compared to others.

Box plots of ROC AUCs of queries (Figure 7.6) and of target medians (Figure 7.7) show that the differences in screening accuracy between the conformational analysis approaches are negligible (the hinges in the figures are versions of the first and third quartiles). The values of different target ROC AUCs are shown in Table 7.5.

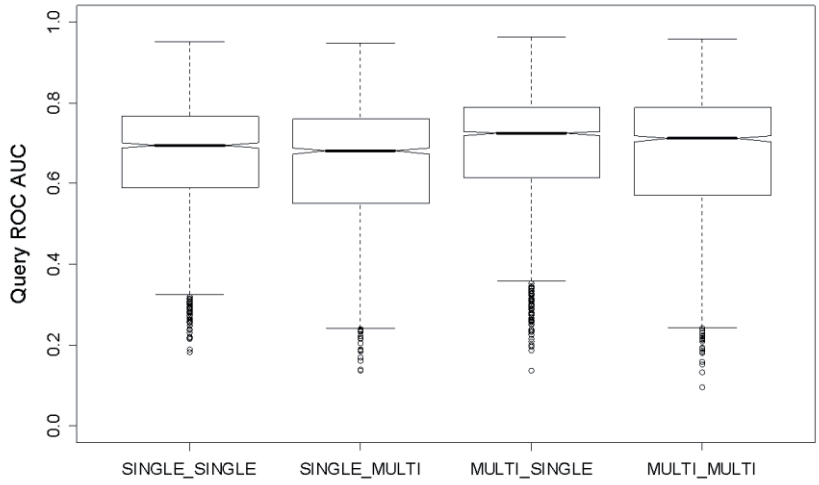


Figure 7.6: Box plot of query ROC AUCs.

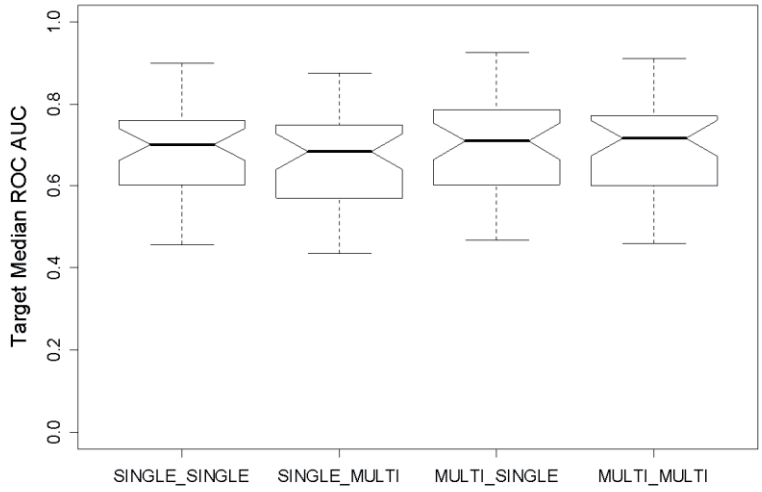


Figure 7.7: Box plot of target median ROC AUC.

Table 7.5: Medians (Med) and averages (Avg) of target ROC AUCs.

	SS		SM		MS		MM	
Target	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**
ACE	0.52	0.47	0.67	0.48	0.47	0.45	0.46	0.45
ACHE	0.69	0.64	0.75	0.68	0.71	0.64	0.73	0.66
ADA	0.59	0.59	0.56	0.57	0.61	0.59	0.62	0.61
ALR2	0.55	0.47	0.44	0.42	0.49	0.45	0.47	0.40
AMPC	0.76	0.64	0.80	0.75	0.59	0.59	0.80	0.75
AR	0.85	0.82	0.83	0.79	0.85	0.83	0.84	0.80
CDK2	0.63	0.58	0.61	0.56	0.60	0.58	0.58	0.55
COMT	0.60	0.51	0.58	0.50	0.60	0.49	0.59	0.49
COX1	0.56	0.55	0.46	0.46	0.55	0.56	0.47	0.47
COX2	0.90	0.86	0.88	0.81	0.93	0.88	0.91	0.84
DHFR	0.66	0.65	0.68	0.67	0.74	0.73	0.80	0.77
EGFR	0.70	0.67	0.67	0.62	0.73	0.69	0.70	0.65
$ER_{agonist}$	0.79	0.77	0.73	0.68	0.79	0.77	0.74	0.68
$ER_{antagonist}$	0.79	0.79	0.74	0.73	0.78	0.79	0.74	0.73
FGFR1	0.76	0.68	0.62	0.57	0.76	0.69	0.63	0.58
FXA	0.72	0.68	0.77	0.69	0.79	0.70	0.79	0.70
GART	0.65	0.67	0.69	0.68	0.66	0.66	0.75	0.72
GPB	0.76	0.66	0.75	0.66	0.75	0.65	0.75	0.64
GR	0.75	0.72	0.70	0.64	0.75	0.70	0.68	0.63
HIVPR	0.59	0.56	0.56	0.57	0.85	0.75	0.72	0.64
HIVRT	0.46	0.49	0.54	0.52	0.49	0.49	0.54	0.51
HMGA	0.61	0.60	0.63	0.62	0.67	0.64	0.74	0.70
HSP90	0.74	0.73	0.72	0.71	0.76	0.76	0.69	0.72
INHA	0.64	0.65	0.69	0.66	0.67	0.66	0.68	0.66
MR	0.88	0.88	0.85	0.83	0.88	0.86	0.85	0.82
NA	0.75	0.69	0.71	0.64	0.78	0.71	0.73	0.67
P38	0.66	0.62	0.69	0.62	0.71	0.66	0.72	0.65
PARP	0.54	0.53	0.55	0.55	0.54	0.53	0.57	0.57
PDE5	0.83	0.79	0.82	0.74	0.88	0.83	0.85	0.78
PDGFRB	0.72	0.70	0.63	0.61	0.71	0.68	0.62	0.60
PNP	0.71	0.66	0.63	0.64	0.76	0.75	0.74	0.75
PPARG	0.70	0.68	0.81	0.77	0.62	0.63	0.71	0.73
PR	0.84	0.81	0.81	0.76	0.84	0.82	0.82	0.77
RXRA	0.79	0.75	0.72	0.69	0.81	0.80	0.82	0.78
SAHH	0.83	0.77	0.86	0.84	0.86	0.85	0.90	0.88
SRC	0.64	0.66	0.54	0.55	0.69	0.68	0.57	0.58

Thrombin	0.72	0.67	0.62	0.62	0.70	0.64	0.66	0.61
TK	0.57	0.53	0.64	0.63	0.60	0.58	0.63	0.62
Trypsin	0.66	0.63	0.50	0.53	0.60	0.62	0.56	0.57
VEGFR2	0.56	0.57	0.54	0.49	0.61	0.57	0.51	0.49
average	0.69	0.66	0.52	0.64	0.70	0.67	0.69	0.66
median	0.70	0.66	0.68	0.64	0.71	0.67	0.72	0.66

As previously stated, the number of retrieved actives in the hit list is not as important in shape-based virtual screening as the chemical diversity of the top ranked compounds (Geppert et al. 2010). All approaches yield the same chemical diversity of top hits (Figures 7.8 and 7.9, Table 7.6).

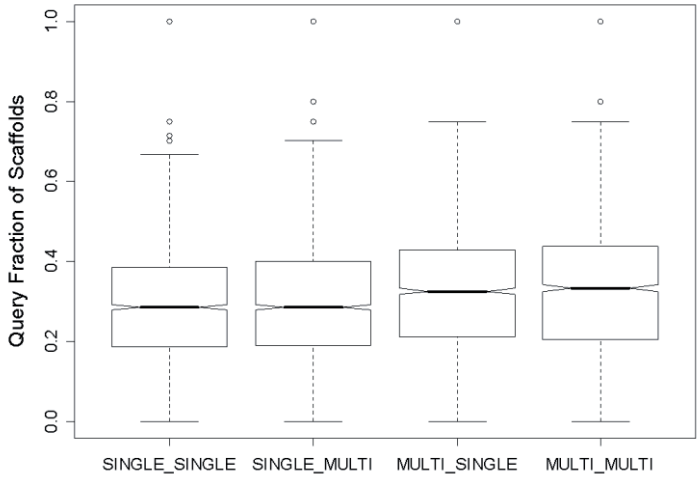


Figure 7.8: Box plot of fraction of retrieved scaffolds.

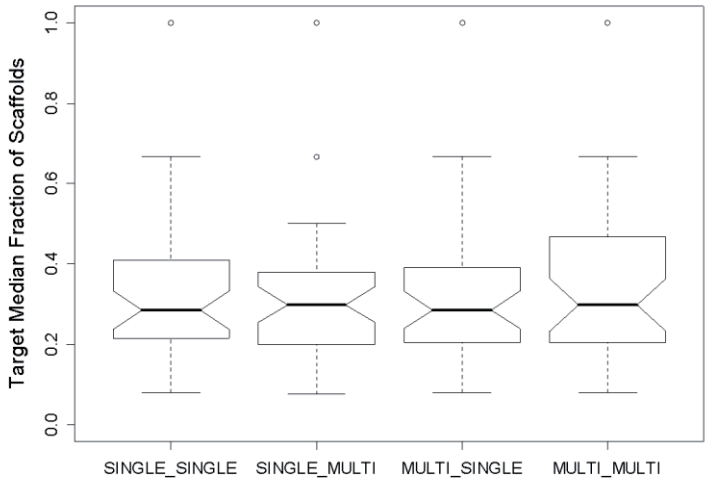


Figure 7.9: Box plot of target median fraction of retrieved scaffolds.

Table 7.6: Medians (Med) and averages (Avg) of retrieved scaffolds.

	SS		SM		MS		MM	
Target	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**
ACE	0.16	0.17	0.16	0.16	0.16	0.16	0.16	0.15
ACHE	0.25	0.25	0.31	0.29	0.31	0.29	0.31	0.29
ADA	0.38	0.30	0.38	0.31	0.25	0.32	0.38	0.37
ALR2	0.15	0.14	0.08	0.13	0.15	0.14	0.15	0.15
AMPC	0.33	0.29	0.33	0.32	0.33	0.32	0.33	0.33
AR	0.17	0.16	0.17	0.17	0.17	0.18	0.17	0.18
CDK2	0.21	0.20	0.19	0.19	0.19	0.21	0.19	0.20
COMT	0.50	0.41	0.50	0.46	0.50	0.41	0.50	0.46
COX1	0.20	0.21	0.10	0.17	0.20	0.21	0.20	0.17
COX2	0.36	0.35	0.39	0.36	0.41	0.38	0.44	0.39
DHFR	0.29	0.31	0.29	0.27	0.36	0.34	0.36	0.34
EGFR	0.33	0.31	0.35	0.34	0.38	0.36	0.40	0.37
$ER_{agonist}$	0.50	0.46	0.50	0.47	0.50	0.46	0.50	0.47
$ER_{antagonist}$	0.25	0.29	0.38	0.31	0.25	0.31	0.38	0.36
FGFR1	0.33	0.30	0.25	0.28	0.33	0.34	0.25	0.29
FXA	0.18	0.19	0.21	0.19	0.21	0.20	0.21	0.20
GART	0.40	0.30	0.20	0.18	0.20	0.15	0.20	0.15
GPB	0.22	0.26	0.33	0.28	0.22	0.25	0.33	0.29
GR	0.25	0.21	0.25	0.22	0.25	0.20	0.25	0.24
HIVPR	0.50	0.50	0.33	0.42	0.67	0.67	0.67	0.67
HIVRT	0.17	0.13	0.08	0.12	0.08	0.12	0.08	0.14
HMGA	0.50	0.45	0.50	0.45	0.50	0.54	0.50	0.57
HSP90	0.50	0.46	0.50	0.52	0.50	0.51	0.50	0.51
INHA	0.17	0.20	0.22	0.22	0.22	0.22	0.22	0.23
MR	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
NA	0.29	0.27	0.29	0.28	0.29	0.31	0.29	0.32
P38	0.37	0.31	0.32	0.31	0.37	0.34	0.37	0.34
PARP	0.25	0.23	0.33	0.25	0.25	0.23	0.25	0.25
PDE5	0.24	0.25	0.19	0.20	0.29	0.28	0.19	0.23
PDGFRB	0.29	0.30	0.26	0.29	0.29	0.31	0.29	0.31
PNP	0.67	0.70	0.67	0.58	0.67	0.75	0.67	0.75
PPARG	0.42	0.36	0.50	0.44	0.33	0.33	0.58	0.53
PR	0.33	0.32	0.33	0.32	0.33	0.33	0.33	0.32
RXRA	0.67	0.61	0.67	0.57	0.67	0.57	0.67	0.61
SAHH	1.00	0.76	1.00	0.77	1.00	0.83	1.00	0.85
SRC	0.25	0.28	0.20	0.23	0.25	0.29	0.25	0.26

Thrombin	0.31	0.26	0.23	0.26	0.15	0.21	0.23	0.28
TK	0.14	0.21	0.29	0.27	0.29	0.28	0.29	0.29
Trypsin	0.25	0.23	0.17	0.21	0.17	0.19	0.17	0.19
VEGFR2	0.08	0.10	0.08	0.09	0.08	0.12	0.08	0.11
average	0.33	0.31	0.33	0.31	0.33	0.33	0.35	0.34
median	0.29	0.31	0.30	0.28	0.29	0.31	0.30	0.30

In order to investigate the differences between the different targets, average ROC curves (Nicholls 2008) were plotted. In ten cases, there is no difference or it is extremely small. The data set is too small for six targets (GART, HIVPR, MR, PPARG, RXRA and TRYPSIN). The remaining 24 ROC curves are shown in Figure 7.10. In DHFR, HMGA, PNP, SAHH data sets, the MULTI_MULTI approach clearly outperforms others. There are also some cases where the use of single conformation databases yields better results ($ER_{agonist}$, FGFR1, PDGFRB and SRC). Overall, there is no clear pattern between target type and the approach with highest ROC AUC.

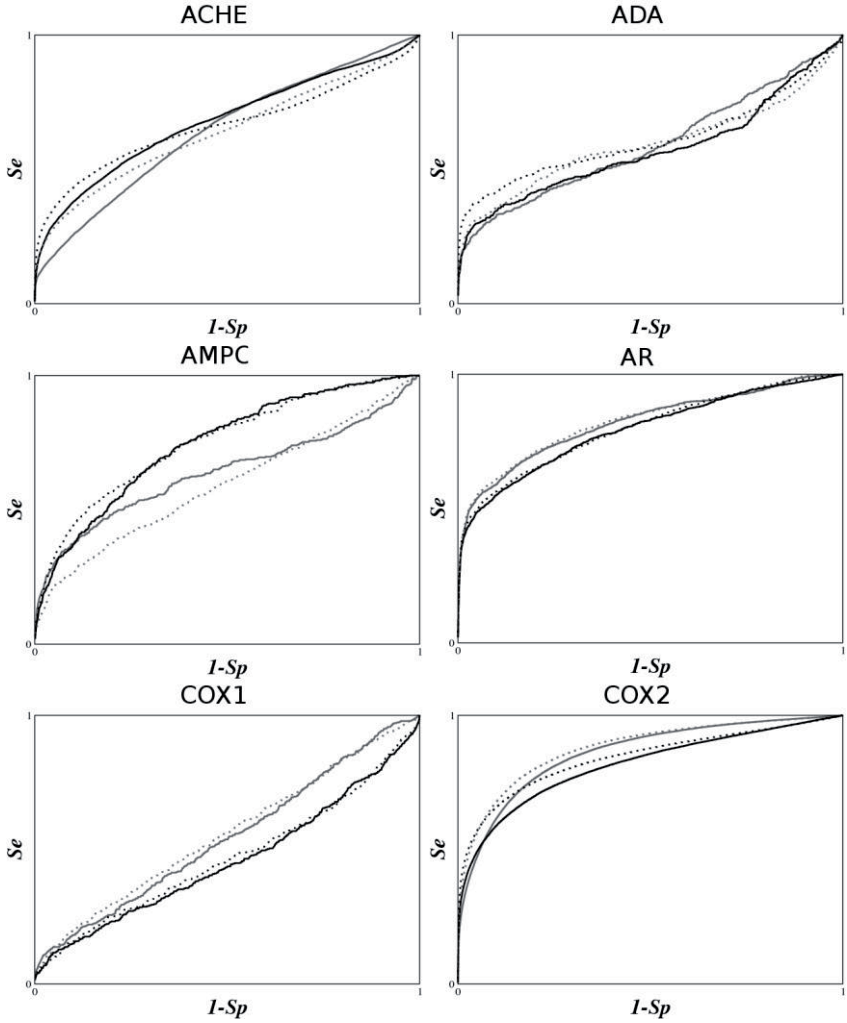


Figure 7.10: Average ROC curves for SINGLE_SINGLE (solid gray line), SINGLE_MULTI (solid black line), MULTI_SINGLE (dotted gray line) and MULTI_MULTI (dotted black line). (1 of 4)

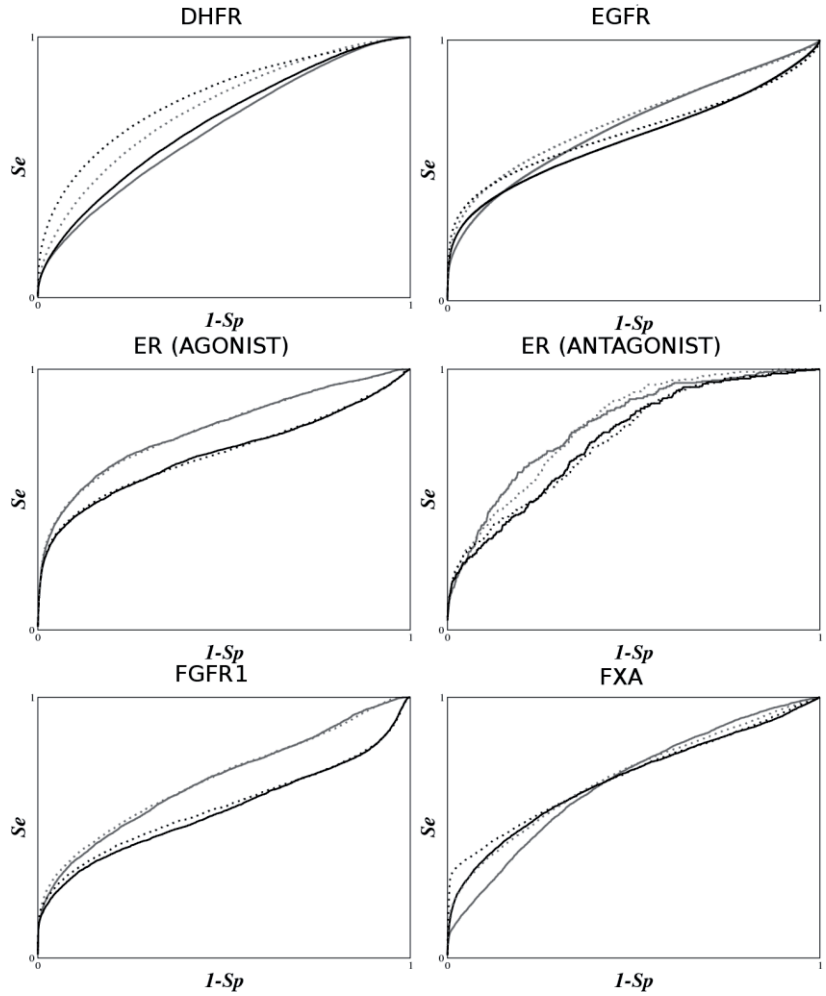


Figure 7.10: Average ROC curves for SINGLE_SINGLE (solid gray line), SINGLE_MULTI (solid black line), MULTI_SINGLE (dotted gray line) and MULTI_MULTI (dotted black line). (2 of 4)

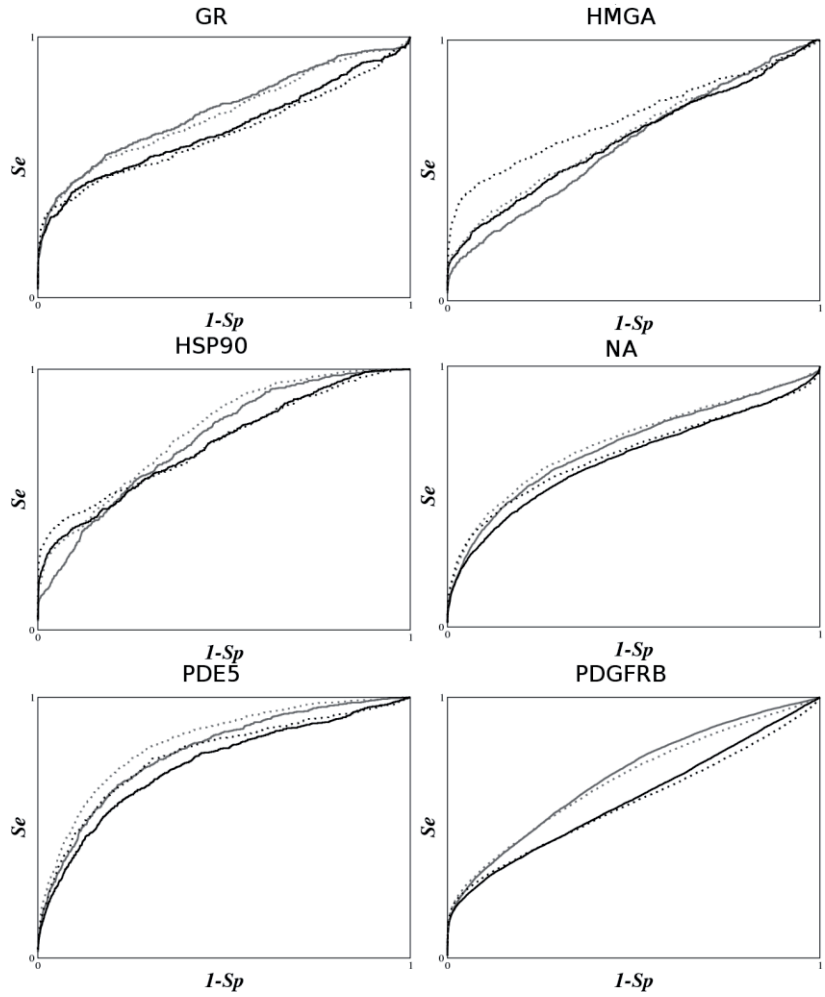


Figure 7.10: Average ROC curves for SINGLE_SINGLE (solid gray line), SINGLE_MULTI (solid black line), MULTI_SINGLE (dotted gray line) and MULTI_MULTI (dotted black line). (3 of 4)

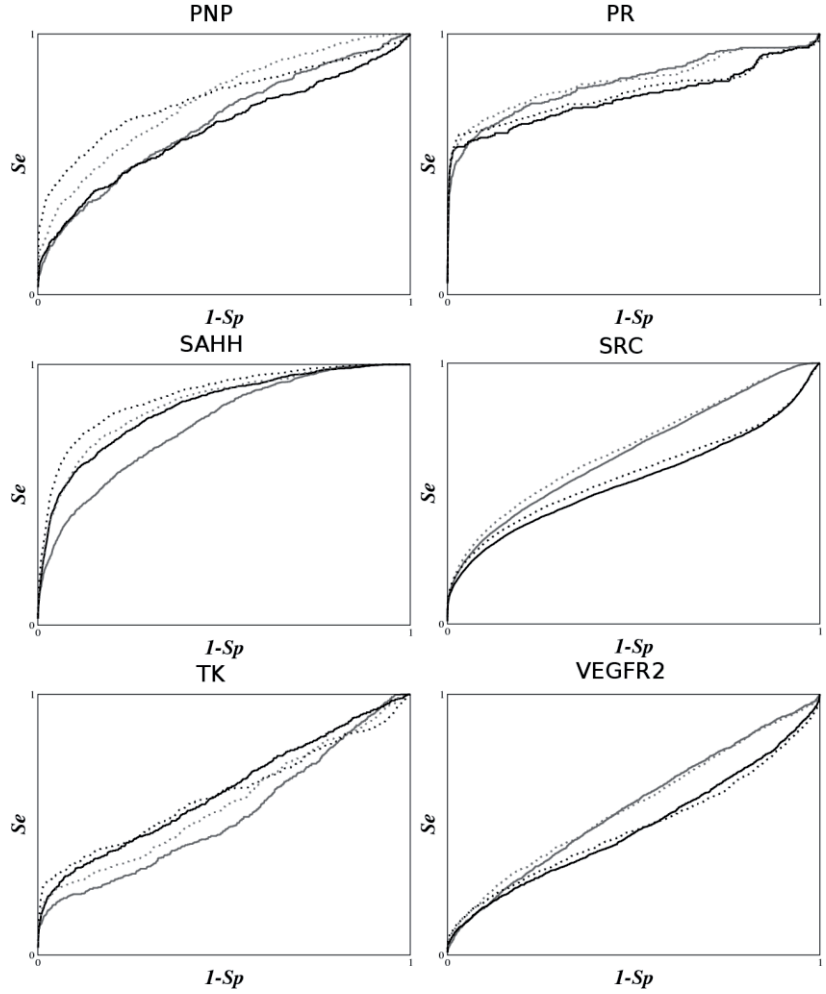


Figure 7.10: Average ROC curves for SINGLE_SINGLE (solid gray line), SINGLE_MULTI (solid black line), MULTI_SINGLE (dotted gray line) and MULTI_MULTI (dotted black line). (4 of 4)

To determine whether the arbitrarily selected cutoff for the top molecules (two times the number of ligands per target) had any effect on the results, average curves of fraction of retrieved scaffolds in different cutoffs were plotted. The maximum curve in this plot is the case, where every unique scaffold is retrieved from the top of the hitlist. A random curve is generated from a shuffled hitlist. It can be concluded that the selection of the cutoff did not have any effect on the conclusions, as the graphs of different approaches are similar for most targets. Graphs for two targets (COX2 and EGFR) with large numbers of ligands and scaffolds are shown in Figure 7.11 as an example.

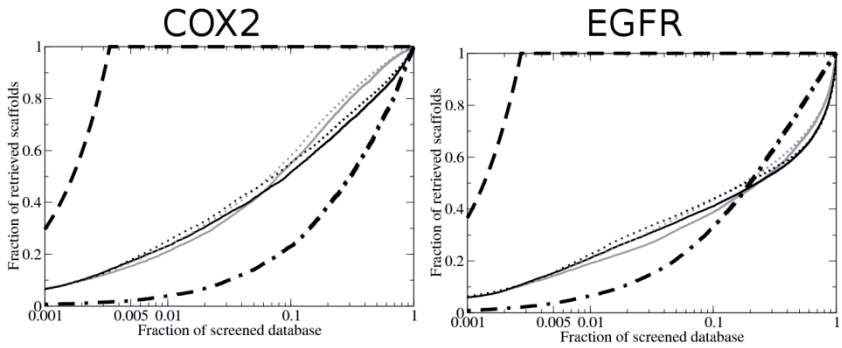


Figure 7.11: Fraction of retrieved scaffolds represented as a function of fraction of screened database for COX2 and EGFR. SINGLE_SINGLE (solid gray), SINGLE_MULTI (solid black), MULTI_SINGLE (dotted gray) and MULTI_MULTI (dotted black) produce similar results. Maximum is drawn with dashed line and random with dashed dotted line.

The small variation between the approaches might be related to the issue of the number of starting positions. Different conformations of the same molecule most likely have a similar effect on the results as increasing the number of starting positions, as there were minor differences found between various initialization modes in the PAPER article by Haque and Pande.

Conformation generation revealed an imbalance in DUD LIB VS, as the ligands had fewer conformations (8.9) than the decoys (12.4) on average and it is possible that this has skewed the results. However, no correlation was found between the

difference in conformations per molecule and the target median ROC AUC of any of the conformational analysis approaches (Figure 7.12).

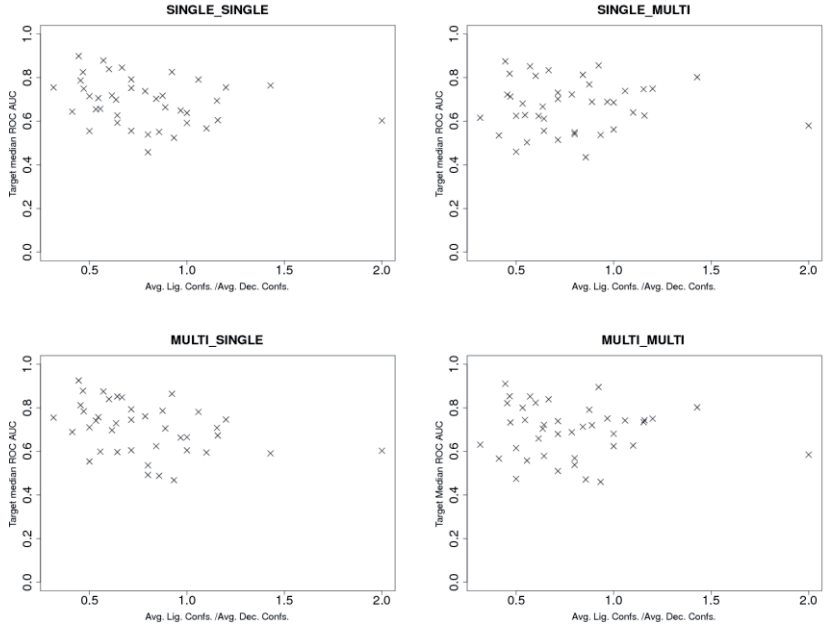


Figure 7.12: Target median ROC AUC vs. the fraction between average number of ligand conformations and the average number of decoy conformations.

Various query molecules for different targets produce extremely poor ROC AUC values, which are independent of the conformational analysis strategy. These ROC AUCs are often related to small query molecules. As previously discussed, it makes little sense to align a small query molecule against much larger molecules (see Chapter 2.6). Some kind of quick pre-filtering step should be applied to the database before the actual shape-based virtual screening in order to eliminate these kinds of pairs, so that the computationally more intensive molecular alignment process could be omitted for these cases.

Even though the ROC AUC values and fraction of retrieved scaffolds are rather similar, the ShapeTanimoto scores of different approaches are clearly different with different conformational analysis approaches (Table 7.7). The more

computation effort that is used, the higher are the ShapeTanimoto scores. However, the difference between ligand and decoy sets stays approximately the same, which explains the similar virtual screening accuracy (Table 7.8). Similar observations have been made when comparing simple shape descriptors like USR and ROCS (Nicholls et al. 2010). It is possible that the ligands and decoys are separated to some other factors that are not very sensitive to the shape of the molecules. Therefore enrichment metrics seem to be a poor measure of the quality of the alignment algorithm.

Table 7.7: Average ShapeTanimoto similarity for ligand (L) and decoy (D) sets.

	SS		SM		MS		MM	
Target	**L**	**D**	**L**	**D**	**L**	**D**	**L**	**D**
ACE	0.49	0.48	0.57	0.56	0.56	0.56	0.62	0.62
ACHE	0.56	0.49	0.64	0.56	0.64	0.58	0.70	0.63
ADA	0.66	0.61	0.70	0.67	0.70	0.66	0.77	0.72
ALR2	0.57	0.58	0.62	0.63	0.61	0.62	0.66	0.67
AMPC	0.60	0.54	0.69	0.61	0.69	0.66	0.77	0.71
AR	0.75	0.63	0.77	0.68	0.77	0.65	0.79	0.69
CDK2	0.56	0.53	0.62	0.61	0.62	0.59	0.67	0.66
COMT	0.53	0.55	0.57	0.59	0.57	0.60	0.61	0.63
COX1	0.63	0.61	0.65	0.65	0.65	0.63	0.67	0.67
COX2	0.64	0.50	0.69	0.57	0.69	0.53	0.73	0.60
DHFR	0.59	0.54	0.69	0.64	0.68	0.62	0.76	0.69
EGFR	0.63	0.57	0.68	0.64	0.68	0.61	0.72	0.68
$ER_{agonist}$	0.73	0.64	0.74	0.69	0.74	0.66	0.75	0.70
$ER_{antagonist}$	0.55	0.42	0.61	0.54	0.61	0.52	0.66	0.59
FGFR1	0.55	0.48	0.60	0.57	0.59	0.52	0.64	0.60
FXA	0.47	0.41	0.58	0.52	0.58	0.51	0.69	0.60
GART	0.52	0.46	0.63	0.58	0.63	0.58	0.76	0.69
GPB	0.66	0.59	0.71	0.65	0.71	0.65	0.75	0.69
GR	0.60	0.53	0.63	0.58	0.63	0.56	0.66	0.61
HIVPR	0.41	0.38	0.50	0.48	0.50	0.46	0.61	0.56
HIVRT	0.51	0.52	0.59	0.59	0.59	0.59	0.66	0.65
HMGA	0.50	0.45	0.57	0.53	0.57	0.53	0.66	0.59
HSP90	0.57	0.49	0.66	0.58	0.66	0.56	0.73	0.64
INHA	0.58	0.52	0.65	0.60	0.65	0.60	0.71	0.66
MR	0.78	0.59	0.78	0.63	0.78	0.61	0.80	0.65
NA	0.58	0.51	0.62	0.58	0.62	0.55	0.67	0.61
P38	0.58	0.52	0.64	0.58	0.64	0.57	0.69	0.62
PARP	0.66	0.63	0.70	0.67	0.69	0.67	0.72	0.70
PDE5	0.54	0.45	0.60	0.53	0.59	0.50	0.64	0.57
PDGFRB	0.63	0.56	0.67	0.63	0.66	0.60	0.69	0.66
PNP	0.64	0.58	0.71	0.66	0.71	0.62	0.77	0.70
PPARG	0.47	0.43	0.56	0.51	0.55	0.52	0.64	0.58
PR	0.68	0.56	0.71	0.62	0.72	0.60	0.76	0.65
RXRA	0.63	0.51	0.66	0.58	0.67	0.54	0.72	0.61
SAHH	0.73	0.63	0.80	0.70	0.60	0.54	0.64	0.61
SRC	0.55	0.49	0.60	0.58	0.60	0.54	0.64	0.61

Thrombin	0.52	0.44	0.57	0.53	0.57	0.52	0.64	0.60
TK	0.68	0.65	0.75	0.71	0.75	0.71	0.80	0.76
Trypsin	0.53	0.44	0.56	0.54	0.56	0.52	0.65	0.61
VEGFR2	0.53	0.51	0.60	0.60	0.60	0.57	0.65	0.65
mean	0.59	0.53	0.65	0.60	0.65	0.59	0.70	0.64
median	0.58	0.52	0.64	0.59	0.65	0.58	0.69	0.64

Table 7.8: Average difference in ShapeTanimoto between ligand and decoy sets.

Target	SS	SM	MS	MM
ACE	0.008	0.003	0.001	-0.004
ACHE	0.068	0.077	0.060	0.066
ADA	0.050	0.032	0.034	0.050
ALR2	-0.005	-0.019	-0.009	-0.016
AMPC	0.062	0.073	0.024	0.060
AR	0.118	0.095	0.119	0.097
CDK2	0.032	0.019	0.027	0.018
COMT	-0.020	-0.023	-0.027	-0.021
COX1	0.018	-0.008	0.017	-0.005
COX2	0.142	0.118	0.156	0.129
DHFR	0.050	0.050	0.063	0.071
EGFR	0.062	0.042	0.067	0.048
$ER_{agonist}$	0.086	0.048	0.082	0.049
$ER_{antagonist}$	0.103	0.071	0.090	0.069
FGFR1	0.075	0.032	0.074	0.038
FXA	0.060	0.061	0.073	0.091
GART	0.057	0.053	0.046	0.075
GPB	0.071	0.056	0.059	0.053
GR	0.071	0.051	0.068	0.052
HIVPR	0.029	0.016	0.037	0.051
HIVRT	-0.003	0.000	0.003	0.007
HMGA	0.044	0.041	0.048	0.066
HSP90	0.083	0.077	0.095	0.089
INHA	0.052	0.052	0.049	0.050
MR	0.187	0.151	0.172	0.151
NA	0.071	0.046	0.072	0.057
P38	0.061	0.055	0.070	0.065
PARP	0.025	0.025	0.022	0.025
PDE5	0.087	0.066	0.097	0.074
PDGFRB	0.074	0.038	0.067	0.036
PNP	0.059	0.044	0.082	0.078
PPARG	0.044	0.059	0.037	0.060
PR	0.116	0.093	0.120	0.107
RXRA	0.119	0.085	0.122	0.116
SAHH	0.093	0.096	0.104	0.107
SRC	0.058	0.020	0.058	0.027
Thrombin	0.081	0.039	0.051	0.045

TK	0.031	0.042	0.033	0.044
Trypsin	0.083	0.020	0.040	0.040
VEGFR2	0.019	-0.001	0.021	0.002
mean	0.063	0.047	0.061	0.055
median	0.062	0.047	0.060	0.053

A low ShapeTanimoto value does not necessarily mean an unreasonable alignment. This is illustrated in Figure 7.13, where there are two COX2 inhibitors superimposed with different conformational analysis approaches. Even though the SINGLE_SINGLE alignment looks reasonable enough, it has a low ShapeTanimoto score of 0.683 because the benzene rings are in different orientations on both molecules.

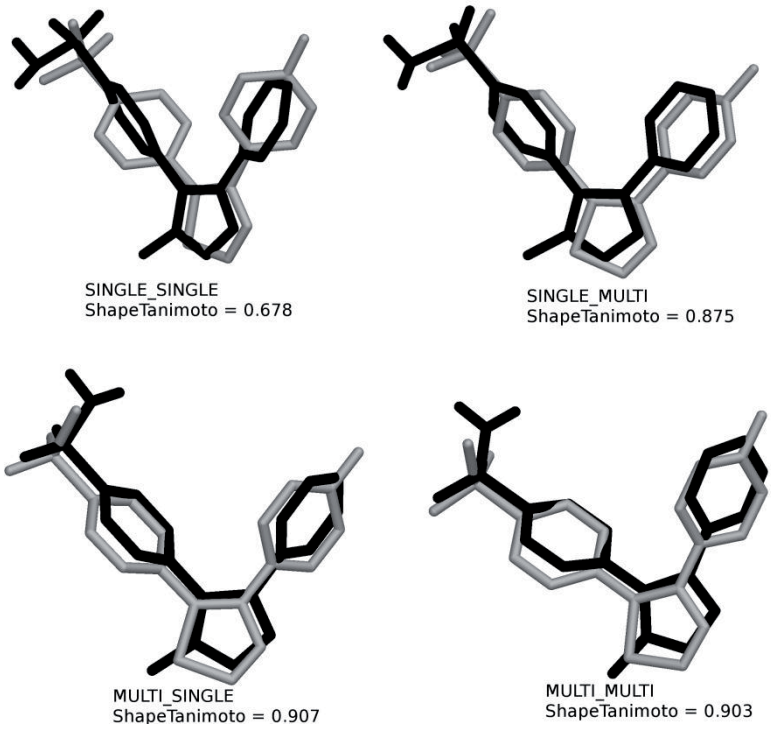


Figure 7.13: Valdecoxib (black) superimposed to ZINC00006596 (gray).

Although PAPER algorithm is extremely fast (0.1-0.3 ms per alignment), the large variations in the numbers of alignments create significant differences in the required computation time

between the different conformational analysis approaches (Table 7.9). By using a multi-conformation database with a single conformation query, one must align ten times more molecules than with single conformation database. The use of a multi-conformation query with multi-conformation database increases the number of alignments by almost two orders of magnitude compared to the simplest approach. The generation of multi-conformation databases also adds to the computational expense of SINGLE_MULTI and MULTI_MULTI approaches.

Table 7.9: The number of alignments in this study for each conformational analysis approach.

Approach	Alignments	Increase Factor
SINGLE_SINGLE	13253166	1.0
SINGLE_MULTI	144375674	10.89
MULTI_SINGLE	97157157	7.33
MULTI_MULTI	1114373810	84.08

8 Conclusions

Ligand-based virtual screening based on alignment and simple models derived from molecular fields might be feasible. A novel virtual screening method called FieldChopper was developed. It is based on the discretization of the electrostatic and van der Waals field into three classes. The results from retrospective virtual screening experiments suggest that FieldChopper is competitive with more complex descriptors and could be used as a molecular sieve when multiple ligands are known. However, it is obvious that additional work would be required to make the software more relevant to drug discovery projects. A major obstacle to the further development of FieldChopper is the lack of high quality data sets that fulfill the requirement of similar ligand binding mode. This effectively prevented the study of using FieldChopper for the rapid prediction of ADMET-properties (notably metabolism), which was one planned application area of the original project.

The use of several query ligands in alignment-based virtual screening improves results considerably. In the FieldChopper evaluation, it was discovered that by using several query molecules with EON, clearly superior results compared to single alignments with FieldChopper could be achieved. However, this strategy increased the computation time by approximately 1500% i.e. it requires considerably more computing resources as the number of active compounds and the size of the database increase. After the publication of this study, Kirchmair and co-workers reported that this observation applies to all targets in DUD (Kirchmair et al. 2009).

Tautomerism prediction is not an issue in current structure-based virtual screening. It was shown that more accurate treatment of tautomerism did not have a dramatic effect on a

current structure-based virtual screening method. The culprit for poor performance must be sought elsewhere. Given the limited accuracy of current scoring functions, the use of multiple tautomeric and protonation states of the ligands is simply a waste of time and resources.

The use of single conformation databases may be feasible in shape-based virtual screening. The use of single conformation databases for the PAPER method yields comparable results to more elaborate multi-conformational virtual screening strategies, as measured by ROC AUC and the fraction of retrieved scaffolds. By using single conformation databases, one can significantly decrease the need for computing resources, especially when working with large databases containing very flexible molecules. This is however only an initial observation and needs to be investigated in more detail. During the preparation of this thesis, a perspective article by Nicholls and co-workers was published (Nicholls et al. 2010). They showed that even though shape-descriptor USR and ROCS ShapeTanimoto had approximately the same ROC AUC values for DUD, the correlation between the two similarity scores was poor. It was suggested that there is some other feature in the data set in addition to the shape that differentiates ligands from decoys. Whatever the reality may be, new virtual screening benchmarks are urgently needed to study such peculiar observations.

Enrichment metrics can be misleading in virtual screening method development. Although the aim of virtual screening is always to find active ligands from a large pool of inactive molecules, enrichment metrics are problematic in method development. The quality of alignments was significantly poorer when using single conformational databases with PAPER, but this was not evident from simply calculating the enrichments. In addition, the problem of analog bias and scaffold definitions should be investigated in more detail.

GPU-computing has a great potential for both ligand- and structure-based methods. As a side product from this study, a publicly available command line interface was developed for PAPER, which makes it possible for anyone to align large sets of molecules on regular desktop computers. At the time of writing of this thesis, GPU software for virtual screening was virtually unavailable. It is very likely that these kinds of programs will become commonly available in the near future.

Structure-based virtual screening has serious limitations. The same issues remain in molecular docking from year to year, which can be seen by comparing review articles from the last eight years (Lyne 2002; Kitchen et al. 2004; Köppen 2009; Kolb and Irwin 2009). There is still the fundamental question if docking is actually useful or are the results obtained from prospective screens more or less due to chance, as the hits from the screens are rarely validated by experimental procedures (Leach et al. 2006; Nicholls 2008; Kolb and Irwin 2009). Perhaps the huge increase in parallel computing in recent years may alleviate this issue by allowing more sophisticated methods to be used than the current scoring functions. However, as the understanding of protein-ligand interactions is still rather limited (Whitesides and Krishnamurthy 2005), there is still much basic research to be done to overcome the current limitations. Given that docking is still a rather computationally intensive task, it might be wise to first to use ligand-based techniques if possible.

3D-methods should be used evaluated more. There is still an on-going discussion about whether the computationally more demanding 3D methods can actually confer any extra value compared to simple 2D-methods like fingerprints (Eckert and Bajorath 2007; Zhang and Muegge 2006; Brown and Martin 1996, 1997; Makara 2001; Sciabola et al. 2007; McGregor and Muskal 1999, 2000; Jenkins et al. 2004; Good et al. 2004b; Nettles et al. 2006; Tiikkainen et al. 2009). It is clear that additional

investigations are needed to establish the putative benefits of 3D-virtual screening.

Finally, currently the development and evaluation of virtual screening methods is challenging due to the lack of standards. In order to improve current methods, it is imperative that such guidelines are quickly established by the scientific community. The author would like to end this thesis by quoting Anthony Nicholls of OpenEye Scientific Software: *"Whether the modeling community has the will to enact such measures may well determine whether future generations of scientists look back and see a field that became essential to drug discovery or one that became a mere footnote"* (Nicholls 2008).

9 References

Abagyan RA, Totrov MM, Kuznetsov DA: ICM: A New Method For Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J Comput Chem* **15**:488-506, 1994.

Adams CP, Brantner W: Estimating the Cost of New Drug Development: Is It Really $802 Million? *Health Aff* **25**:420-428, 2006.

Adane L, Bharatam PV, Sharma V: A common feature-based 3D-pharmacophore model generation and virtual screening: identification of potential PfDHFR inhibitors. *J Enzyme Inhib Med Chem* doi: 10.3109/14756360903393817, 2009.

Albiston AL, Morton CJ, Ng HL, Pham V, Yeatman HR, Ye S, Fernando RN, De Bundel D, Ascher DB, Mandelsohn FA, Parker MW, Chai SY: Identification and characterization of a new cognitive enhancer based on inhibition of insulin-regulated aminopeptidase. *FASEB J* **22**:4209-4217, 2008.

Aparoy P, Kumar Reddy K, Kalangi SK, Chandramohan Reddy T, Reddanna P: Pharmacophore modeling and virtual screening for designing potential 5-Lipoxygenase inhibitors. *Bioorg Med Chem Lett* **20**:1013-1018, 2010.

Ashburn TT, Thor KB: Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* **3**:673-683, 2004.

Baell JB, Holloway GA: New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem* **53**:2719-2740, 2010.

Ballester PJ, Richards WG: Ultrafast shape recognition for similarity search in molecular databases. *Proc R Soc A* **463**:1307–1321, 2007a.

Ballester PJ, Richards WG: Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* **28**:1711–1723, 2007b.

Ballester PJ, Finn PW, Richards WG: Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *J Mol Graph Model* **27**:836-845, 2009.

Ban TA: The role of serendipity in drug discovery. *Dialogues Clin Neurosc* **8**:335-344, 2006.

Barker EJ, Gardiner EJ, Gillett VJ, Kitts P, Morris J: Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J Chem Inf Comput Sci* **43**:346-356, 2003.

Barker EJ, Buttar D, Cosgrove DA, Gardiner EJ, Kitts P, Willett P, Gillet VJ: Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J Chem Inf Model* **46**:503–511, 2006.

Barnum D, Greene J, Smellie A, Sprague P: Identification of Common Functional Configurations among Molecules. *J Chem Inf Comput Sci* **36**:563–571, 1996.

Basse N, Montes M, Maréchai X, Qin L, Bouvier-Durand M, Genin E, Vidal J, Villoutreix BO, Reboud-Ravaux M: Novel organic proteoasome inhibitors identified by virtual and in vitro screening. *J Med Chem* **53**:509-513, 2010.

Bayry J, Tchilian EZ, Davies MN, Forbes EK, Draper SJ, Kaveri SV, Hill AV, Kazatchikine MD, Beverley PC, Flower DR, Tough DF: In silico identified CCR4 antagonists target regulatory T cells and exert

adjuvant activity in vaccination. *Proc Natl Acad Sci USA* **105**:10221-10226, 2008.

Bemis GW, Murcko MA: The properties of known drugs. 1. Molecular Frameworks. *J Med Chem* **39**:2887–2893, 1996.

Bender A, Mussa HY, Gill GS, Glen RC: Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT3D). *J Med Chem* **47**:6569–6583, 2004.

Bender A, Glen RC: A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J Chem Inf Model* **45**:1369-1375, 2005.

Bender A: Compound bioactivities go public. *Nat Chem Biol* **6**:309, 2010.

Bhutoria S, Ghoshal N: Deciphering ligand dependent degree of binding site closure and its implication in inhibitor design: A modeling study on human adenosine kinase. *J Mol Graph Model* **28**:577-591, 2010.

Bissantz C, Kuhn B, Stahl M: A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem* doi: 10.1021/jm100112j, 2010.

Boehr DD, Nussinov R, Wright PE: The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **5**:789-796, 2009.

Bohm HJ, Flohr A, Stahl M: Scaffold hopping. *Drug Discovery Today: Technologies* **1**:217–224, 2004.

Boppana K, Dubey PK, Jagarlapudi SA, Vadivelan S, Rambabu G: Knowledge based identification of MAO-B selective inhibitors using pharmacophore and structure based virtual screening models. *Eur J Med Chem* **44**:3584-3590, 2009.

Borodina YV, Bolton E, Fontaine F, Bryant SH: Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space. *J Chem Inf Model* **47**:1428-1437, 2007.

Borosy A, Csizmadia F, Volford A: Structure Based Clustering of NCI's Anti-HIV Library. *EuroCombi-1, First Symposium of the European Society of Combinatorial Science*, July 1-5, 2001.

Bosshard HR: Molecular Recognition by Induced Fit: How Fit is the Concept? *News Physiol Sci* **16**:171-173, 2001.

Boström J, Berggren K, Elebring T, Greasley PJ, Wilstermann M: Scaffold hopping, synthesis and structure–activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: A novel series of CB1 receptor antagonists. *Bioorg Med Chem* **15**:4077-4084, 2007.

Boström J, Grant A: Exploiting Ligand Conformations in Drug Design. In: *Molecular Drug Properties*, Ed. Mannhold R, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2008.

Bradley EK, Beroza P, Penzotti JE, Grootenhuis PDJ, Spellmeyer DC, Miller JL: A Rapid Computational Method for Lead Evolution: Description and Application to α1-Adrenergic Antagonists. *J Med Chem* **43**:2770-2774, 2000.

Bradley EK, Miller JL, Saiah E, Grootenhuis PDJ: Informative Library Design as an Efficient Strategy to Identify and Optimize Leads: Application to Cyclin-Dependent Kinase 2 Antagonists. *J Med Chem* **46**:4360-4364, 2003.

Breault GA, Comita-Prevoir J, Eyermann CJ, Geng B, Petrichko R, Doig P, Gorseth E, Noonan B: Exploring 8-benzyl pteridine-6,7-diones as inhibitors of glutamate racemase (MurI) in gram-positive bacteria. *Bioorg Med Chem Lett* **18**:6100-6103, 2008.

Brooijmans N, Humblet C: Chemical space sampling by different scoring functions and crystal structure. *J Comput Aided Mol Des* doi: 10.1007/s10822-010-9356-2, 2010.

Brooks WH, Daniel KG, Sung SS, Guida WC: Computation validation of the importance of absolute stereochemistry in virtual screening. *J Chem Inf Model* **48**:639-645, 2008.

Brown N, Jacoby E: On scaffolds and hopping in medicinal chemistry. *Mini Rev Med Chem* **6**:1217–1229, 2006.

Brown RD, Martin YC: Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem Inf Comput Sci* **36**:572–584, 1996.

Brown RD, Martin YC: The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J Chem Inf Comput Sci* **37**:1–9, 1997.

Brozic P, Turk S, Lanisnik TR, Gobec S: Discovery of new inhibitors of aldo-keto reductase 1C1 by structure-based virtual screening. *Mol Cell Endocrinol* **301**:245-250, 2009.

Budin N, Majeux N, Caflisch A: Fragment-Based flexible ligand docking by evolutionary optimization. *Biol Chem* **382**:1365-1372, 2001.

Böhm HJ: LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**:593-606, 1992.

Böhm HJ: Prediction of binding constants of protein ligands: a fast method for the priorization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* **12**:309-323, 1998.

Capelli AM, Feriani A, Tedesco G, Pozzan A: Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands. *J Chem Inf Model* **46**:659–664, 2006.

Card GL, England BP, Suzuki Y, Fong D, Powell B, Lee B, Luu C, Tabrizizad M, Gillette S, Ibrahim PN, Artis DR, Bollag G, Milburn MV, Kim SH, Schlessinger J, Zhang KYJ: Catalytic domain of human phosphodiesterase 5A in complex with tadalafil. *Structure* **12**:2233-2247, 2004.

Carlsson Y, Yoo L, Gao ZG, Irwin JJ, Shoichet BK, Jacobson KA: Structure-based discovery of A2A adenosine receptor ligands. *J Med Chem* **53**:3748-3755, 2010.

Charifson PS, Corkery JJ, Murcko MA, Walters WP: Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**:5100-5109, 1999.

Chen CS, Chiou CT, Chen GS, Chen SC, Hu CY, Chi WK, Chu YD, Hwang LH, Chen PJ, Chen DS, Liaw SH, Chern JW: Structure-based discovery of triphenylmethane derivatives as inhibitors of hepatitis C virus helicase. *J Med Chem* **52**:2716-2723, 2009.

Chen CY: Virtual screening and drug design for PDE-5 receptor from traditional Chinese medicine database. *J Biomol Struct Dyn* **27**:627-640, 2010.

Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY: SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J Comput Chem* **28**:612-623, 2007.

Chen IJ, Foloppe N: Conformational sampling of druglike molecules with MOE and Catalyst: Implications for pharmacophore modeling and virtual screening. *J Chem Inf Model* **48**:1773-1780, 2008.

Chiang YK, Kuo CC, Wu YS, Chen CT, Coumar MS, Wu JS, Hsieh HP, Chang CY, Jseng HY, Wu MH, Leou JS, Song JS, Chang JY, Lyu PC, Chao YS, Wu SY: Generation of ligand-based pharmacophore model and virtual screening for identification of novel tubulin inhibitors with potent anticancer activity. *J Med Chem* **52**:4221-4233, 2009.

Christof G, Smolinski M, Steuber H, Sotriffer CA, Heine A, Hangauer DG, Klebe G: Thermodynamic Inhibition Profile of a Cyclopentyl and a

Cyclohexyl Derivative towards Thrombin: The Same but for Different Reasons. *Angew Chem Int Ed* **46**:8511-8514, 2007.

Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB: Consensus scoring for ligand/protein interactions. *J Mol Graph Model* **20**:281-295, 2002.

Clark RD, Webster-Clark DJ: Managing bias in ROC curves. *J Comput Aided Mol Des* **22**:141-146, 2008.

Clark RD, Shepphird JK, Holliday J: The effect of structural redundancy in validation sets on virtual screening performance. *J Chemometr* **23**:471-478, 2009.

Cleves AE, Jain AN: Robust Ligand-Based Modeling of the Biological Targets of Known Drugs. *J Med Chem* **49**:2921–2938, 2006.

Connolly ML: Analytical molecular surface calculation. *J Appl Crystallogr* **16**:548-558, 1983.

Corbeil CR, Englebienne P, Moitessier: Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of FITTED 1.0. *J Chem Inf Model* **47**:435-449, 2007.

Cosconati S, Hong JA, Novellino E, Carroll KS, Goodsell DS, Olson AJ: Structure-based virtual screening and biological evaluation of Mycobacterium tuberculosis adenosine 5'-phosphosulfate reductase inhibitors. *J Med Chem* **51**:6627-6630, 2008.

Cosconati S, Marinelli L, La Motta C, Sartini S, Da Settimo F, Olson AJ, Novellino E: Pursuing Aldose Reductase Inhibitors through in Site Cross-Docking and Similarity-Based Virtual Screening. *J Med Chem* **52**:5578-5591, 2009.

Coumar MS, Leou JS, Shukla P, Wu JS, Dixit AK, Lin WH, Chang CY, Lien TW, Tan UK, Chen CH, Hsu JT, Chao YS, Wu SY, Hsieh HP: Structrure-based drug design of novel Aurora kinase A inhibitors: structural basis for potency and specificity. *J Med Chem* **52**:1050-1062, 2009.

Cozza G, Gianoncelli A, Montopoli M, Caparrotta L, Venerando A, Meggio F, Pinna LA, Zagotto G, Moro S: Identification of novel protein kinase CK1 delta (CK1delta) inhibitors through structure-based virtual screening. *Bioorg Med Chem Lett* **18**:5672-5675, 2008.

Cramer RD, Patterson DE, Bunce JD: Comparative Field Analysis (CoMFA). 1 Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* **110**:5959–5967, 1988.

Cramer RD: Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J Med Chem* **46**:374-388, 2003.

Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RD: "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J Med Chem* **47**:6777–6791, 2004.

Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C: Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model* **49**:1455-1474, 2009.

Cruciani G, Watson KA: Comparative Molecular Field Analysis Using GRID Force-Field and GOLPE Variable Selection Methods in a Study of Inhibitors of Glycogen Phosphorylase B. *J Med Chem* **37**:2589–2601, 1994.

Daszykowski M, Walczak B, Massart DL: Representative subset selection. *Anal Chim Acta* **468**:91-103, 2002.

Davis AM, St-Gallay SA, Kleywegt GJ: Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discovery Today* **13**:831-841, 2008.

Demsar J: Statistical comparisons on classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**:1-30, 2006.

Deye J, Elam C, Lape M, Ratliff R, Evans K, Paula S: Structure-based virtual screening for novel inhibitors of the sarco/emdoplasmic reticulum

calcium ATPase and their experimental evaluation. *Bioorg Med Chem* **17**:1353-1360, 2009.

de Vries SJ, van Dijk ADJ, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJ: HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on CAPRI targets. *Proteins Struct Funct Bioinf* **69**:726-733, 2007.

Dill KA: Additivity principles in biochemistry. *J Biol Chem* **272**:701-704, 1997.

Dill KA, Bromberg S: *Molecular driving forces: Statistical thermodynamics in chemistry and biology*, Garland Science, New York, USA, 2003.

Diller DJ, Kenneth J, Merz M: High throughput docking for library design and library prioritization. *Proteins: Struct Funct Genet* **43**:113-124, 2001.

Diller DJ, Li R: Kinases, Homology Models, and High Throughput Docking. *J Med Chem* **46**:4638-4647, 2003.

DiMasi JA, Hansen RW, Grabowski HG: The price of innovation: new estimates of drug development costs. *J Health Econ* **22**:151-185, 2003.

Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA: PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comp Aided Mol Des* **20**:647-671, 2006.

Dodds EC, Lawson W: Molecular structure in relation to oestrogenic activity. Compounds without phenanthrene nucleus. *Proc R Soc Lond Ser B* **125**:122-132, 1940.

Doweyko AM: 3D-QSAR illusions. *J Comp Aided Mol Des* **18**:587–596, 2004.

Durant JL, Leland BA, Henry DR, Nourse JG: Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **42**:1273–1280, 2002.

Durán A, Martinez GC, Pastor M: Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in Molecular Interaction Fields. *J Chem Inf Model* **48**:1813-1823, 2008.

Durán A, Zamora I, Pastor M: Suitability of GRIND-Based Principal Properties for the Description of Molecular Similarity and Ligand-Based Virtual Screening. *J Chem Inf Model* **49**:2129-2138, 2009.

Eckert H, Bajorath J: Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **12**:225–233, 2007.

Edgar SJ, Holliday JD, Willett P: Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures. *J Mol Graph Model* **18**:343-357, 2000.

Ehrlich P: Über den jetzigen Stand der Chemotherapie. *Ber Dtsch Chem Ges* **42**:17–47, 1901.

Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP: Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aid Mol Des* **11**:425-445, 1997.

Englebienne P, Moitessier N: Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins. *J Chem Inf Model* **49**:2564-2571, 2009.

Enyedy IJ, Egan WJ: Can we use docking and scoring for hit-to-lead optimization? *J Comput Aided Mol Des* **22**:161-168, 2002.

Fakhrudin N, Ladurner A, Atanasov AG , Heiss EH, Baumgartner L, Markt P, Schuster D, Ellmerer EP, Wolber G, Rollinger JM, Stuppner H, Dirsch VM: Computer-aided discovery, validation and mechanistic characterization of novel neolignan activators of PPARγ. *Mol Pharm* doi: 10.1124/mol.109.062141, 2010.

Fan Y, Lai MH, Sullivan K, Popiolek M, Andree TH, Dollings P, Pausch MH: The identification of neurotensin NTS1 receptor partial agonists through a ligand-based virtual screening approach. *Bioorg Med Chem Lett* **18**:5789-5791, 2008.

Faver J, Merz KM: Utility of the Hard/Soft Acid-Base Principle via the Fukui Function in Biological Systems. *J Chem Theory Comput* doi: 10.1021/ct9005085, 2010.

Feder M, Purta M, Koscinski L, Cubrilo S, Vlahovicek GM, Bujnicki JM: Virtual screening and experimental verification to identify potential inhibitors of the ErmC Methyltransferase responsible for bacterial resistance against macrolide antibiotics. *ChemMedChem* **3**:316-322, 2008.

Feher M, Williams CI: Effect of Input Differences on the Results of Docking Calculations. *J Chem Inf Model* **49**:1704-1714, 2009.

Fischer E: Einfluss der Konfiguration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges* **27**:2985-2993, 1894.

Fitzgerald SH, Sabat M, Geysen HM: Survey of the diversity space coverage of reported combinatorial libraries. *J Comb Chem* **9**:724-734, 2007.

Fontaine F, Pastor M, Sanz F: Incorporating molecular shape into the alignment-free Grid-Independent Descriptors. *J Med Chem* **47**:2805-2815, 2004.

Franke L, Schwarz O, Muller-Kuhrt L, Hoernig C, Fischer L, George S, Tanrikulu Y, Schneider P, Werz O, Steinhilber D, Schneider G: Identification of Natural-Product-Derived Inhibitors of 5-Lipoxygenase Activity by Ligand-Based Virtual Screening. *J Med Chem* **50**:2640–2646, 2007.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS: Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**:1739-1749, 2004.

Gabb HA, Jackson RM, Sternberg JME: Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272**:106-120, 1997.

Gasteiger J, Rudolph C, Sadowski J: Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput Methodol* **3**:537-547, 1990.

Geppert H, Vogt M, Bajorath J: Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J Chem Inf Model* doi: 10.1021/ci900419k, 2010.

Giupponi G, Harvey MJ, De Fabritiis G: The impact of accelerator processors for high-throughput molecular modeling and simulation. *Drug Discovery Today* **13**:1052-1058, 2008.

Godden JW, Stahura FL, Bajorath J: Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J Chem Inf Model* **45**:1812–1819, 2005.

Gohlke, Hendlich M, Klebe G: Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295**:337-356, 2000.

Good AC, Cho SJ, Mason JS: Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. *J Comput Aided Mol Des* **18**:523–527, 2004a.

Good AC, Hermsmeier MA, Hindle SA: Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J Comp Aided Mol Des* **18**:529–536, 2004b.

Good AC, Oprea TI: Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: a Help or Hindrance in Tool Selection? *J Comput Aided Mol Des* **22**:169-178, 2008.

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

Grant JA, Gallardo MA, Pickup BT: A fast method of molecular shape comparison: a simple application of a Gaussian description molecular shape. *J Comp Chem* **17**:1653–1666, 1995.

Grant JA, Pickup BT, Nicholls A: A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J Comput Chem* **22**:608-640, 2001.

Greene J, Kahn S, Savoj H, Sprague P, Teig S: Chemical Function Queries for 3D Database Search. *J Chem Inf Comput Sci* **34**:1297–1308, 1994.

Guha R, Jurs PC: Determining the Validity of a QSAR Model − A Classification Approach. *J Chem Inf Model* **45**:65-73, 2005.

Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner JK, Willighagen: The Blue Obelisk – Interoperability in Chemical Informatics. *J Chem Inf Model* **46**:991-998, 2006.

Gundersen E, Fan K, Haas K, Huryn D, Jacobsen JS, Kreft A, Martone R, Mayer S, Sonnenberg-Reines J, Sun SC, Zhou H: Molecular modeling-based design, synthesis, and activity of substituted piperidines as γ-secretase inhibitors. *Bioorg Med Chem Lett* **15**:1891-1894, 2005.

Güner OF, Henry DR: Metric for Analyzing Hit Lists and Pharmacophores. In: *Pharmacophore Perception, Development, and Use in Drug Design*, pp. 193-210. Ed. Güner OF, International University Line, La Jolla, U.S.A, 2000.

Hall LH, Kier LB: The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in Computational Chemistry*, **2**:367–415, 1991.

Haque IS, Pande VS: PAPER – Accelerating Parallel Evaluations of ROCS. *J Comput Chem* **31**:117-132, 2010.

Hartmann C, Antes I, Lengauer T: Docking and scoring with alternative side-chain conformations. *Proteins: Struct Funct Bioinf* **74**:712-726, 2009.

Hawkins PCD, Skillman AG, Nicholls A: Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J Med Chem* **50**:74-82, 2007.

Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MA: Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* doi: 10.1021/ci100031x, 2010.

Hecker EA, Duraiswami C, Andrea TA, Diller DJ: Use of Catalyst Pharmacophore Models for Screening of Large Combinatorial Libraries. *J Chem Inf Comput Sci* **42**:1204-1211, 2002.

Hert J, Willett P, Wilton DJ: New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model* **46**:462-470, 2006.

Hillebrecht A, Klebe G: Use of 3D QSAR Models for Database Screening: A Feasibility Study. *J Chem Inf Model* **48**:384-396, 2008.

Hodgkin EE, Richards WG: Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int J Quantum Chem* **32**:105-110, 1987.

Holliday JD, Salim N, Whittle M, Willett P: Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J Chem Inf Comput Sci* **43**:819–828, 2003.

Hong TJ, Park H, Kim YJ, Jeong JH, Hahn JS: Identification of new Hsp90 inhibitors by structure-based virtual screening. *Bioorg Med Chem Lett* **19**:4839-4842, 2009.

Huang N, Schoichet BK, Irwin JJ: Benchmarking Sets for Molecular Docking. *J Med Chem* **49**:6789–6801, 2006.

Huey R, Goodsell DS, Morris GM, Olson AJ: Grid-based hydrogen potentials with improved directionality. *Lett Drug Des Discovery* **1**:178-183, 2004.

Huey R, Morris GM, Olson AJ, Goodsell DS: A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**:1145-1152, 2007.

Höltje HD, Sippl W, Rognan D, Folkers G: *Molecular modeling: basic principles and applications*, 3rd edition, Wiley-VCH Verlag GmbH & Co, Weinheim, Germany, 2008.

Irwin JJ, Schoichet BK: ZINC – Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model* **45**:177-182, 2005.

Irwin JJ: Community benchmarks for virtual screening. *J Comput Aided Mol Des* **22**:193-199, 2008.

IUPAC: *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book"). doi: doi:10.1351/goldbook, 2010.

Jacobsson M, Lidén P, Stjernschantz E, Boström H, Norinder U: Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J Med Chem* **46**:5781-5789, 2003.

Jacq N, Salzemann J, Jacq F, Legré E, Montagnat J, Maaβ A, Reichstadt M, Schwichtenberg H, Sridhar M, Kasam V, Zimmermann M, Hoffmann M, Breton V: Grid-enabled Virtual Screening Against Malaria. *J Grid Comput* **1**:29-43, 2008.

Jahn A, Hinselmann G, Fechner N, Zell A: Optimal assignment methods for ligand-based virtual screening. *J Cheminf* **1**:14, 2009.

Jain AN: Ligand-Based Structural Hypotheses for Virtual Screening. *J Med Chem* **47**:947–961, 2004.

Jain AN: Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des* **21**:281-306, 2007.

Jain AN, Nicholls A: Recommendations for evaluation of computational methods. *J Comput Aided Mol Des* **22**:133-139, 2008.

Jenkins JL, Glick M, Davies JW: A 3D Similarity Method for Scaffold Hopping from known drugs or natural ligands to new chemotypes. *J Med Chem* **47**:6144–6159, 2004.

Jones G, Willett P, Glen RC, Leach AR, Taylor: Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**:727-748, 1997.

Jorgensen WL: Rusting of the lock and key model for protein-ligand binding. *Science* **254**:954-955, 1991.

Kalliokoski T: *Ligand-based virtual screening using molecular fields*. Licentiate Thesis in Pharmacy, University of Kuopio, Finland, 2008.

Kalliokoski T, Rönkkö T, Poso A: FieldChopper, a New Tool for Automatic Model Generation and Virtual Screening Based on Molecular Fields. *J Chem Inf Model* **48**:1131-1137, 2008.

Kalliokoski T, Salo HS, Lahtela-Kakkonen M, Poso A: The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J Chem Inf Model* **49**:2742-2748, 2009.

Kalliokoski T, Rönkkö T, Poso A: Increasing the throughput of shape-based virtual screening with GPU processing and single conformation databases. *Mol Inf* **29**:293-296, 2010.

Khan MT, Fuskevåg OM, Sylte I: Discovery of potent thermolysin inhibitors using structure based virtual screening and binding assays. *J Med Chem* **52**:48-61, 2009.

Kim BH, Jee JG, Yin CH, Sandoval C, Jayabose S, Kitamura D, Bach EA, Baeg GH: NSC114792, a novel small molecule identified through structure-based computational database screening, selectively inhibits JAK3. *Mol Cancer* **9**:36, 2010.

Kirchmair J, Wolber G, Laggner C, Langer T: Comparative Performance Assessment of the Conformational Model Generators Omega and

Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J Chem Inf Model* **46**:1848-1861, 2006.

Kirchmair J, Markt P, Distinto S, Wolber G, Langer T: Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes? *J Comput Aided Mol Des* **22**:213-228, 2008.

Kirchmair J, Distinto S, Markt P, Schuster D, Spitzer GM, Liedl KR, Wolber G: How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J Chem Inf Model* **49**:678-692, 2009.

Kiss R, Kiss B, Könczöl Á, Szalai F, Jelinek I, Lásló V, Noszál B, Falus A, Keseru GM: Discovery of novel human histamine H4 receptor ligands by large-scale structure-based virtual screening. *J Med Chem* **51**:3145-3153, 2008.

Kiss R, Polgár T, Kirabo A, Sayyah J, Figueroa NC, List AF, Sokol L, Zuckerman KS, Gali M, Bisht KS, Sayeski PP, Keseru GM: Identification of a novel inhibitor of JAK2 tyrosine kinase by structure-based virtual screening. *Bioorg Med Chem Lett* **19**:3598-3601, 2009.

Kitchen DB, Decornez H, Furr JR, Bajorath J: Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**:935–949, 2004.

Klebe G, Abraham U, Mietzner T: Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J Med Chem* **37**:4130–4146, 1994.

Knox AJS, Meegan MJ, Carta G, Lloyd DG: Considerations in Compound Database Preparation – "Hidden" Impact on Virtual Screening Results. *J Chem Inf Model* **45**:1908-1919, 2005.

Koehler RT, Dixon SL, Villar HO: LASSOO: A Generalized Directed Diversity Approach to the Design and Enrichment of Chemical Libraries. *J Med Chem* **42**:4695–4704, 1999.

Kogej T, Engkvist O, Blomberg N, Muresan S: Multifingerprint based similarity searches for targeted class compound selection. *J Chem Inf Model* **46**:1201–1213, 2006.

Koide Y, Uemoto K, Hasegawa T, Sada T, Murakami A, Takasugi H, Sakurai A, Mochizuki N, Takahashi A, Nishida A: Pharmacophore-Based Design of Sphingosine 1-phosphate-3 Receptor Antagonists That Include a 3,4-Dialkoxybenzophenone Scaffold. *J Med Chem* **50**:442–454, 2007.

Kolb P, Caflish A: Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J Med Chem* **14**:7384-7392, 2006.

Kolb P, Ferreira RS, Irwin JJ, Shoichet BK: Docking and chemoinformatic screens for new ligands and targets. *Curr Opin Biotechnol* **20**:429-436, 2009.

Kolb P, Irwin JJ: Docking screens: right for the right reasons? *Curr Top Med Chem* **9**:755-770, 2009.

Kolb P, Rosenbaum DM, Irwin JJ, Fung JJ, Kobilka BK, Shoichet BK: Structure-based discovery of beta2-adrenergic receptor ligands. *Proc Natl Acad Sci USA* **106**:6843-6848, 2009.

Korb O, Stützle T, Exner TE: Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J Chem Inf Model* 49:84-96, 2009.

Koshland DE, Jr.: Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U. S. A.* **44**:98-104, 1958.

Koshland DE, Jr.: How to Get Paid for Having Fun. *Annu Rev Biochem* **65**:1-13, 1996.

Kotani T, Higashiura K: Comparative Molecular Active Site Analysis (CoMASA). 1. An Approach to Rapid Evaluation of 3D QSAR. *J Med Chem* **47**:2732-2742, 2004.

Kovac A, Konc J, Vehar B, Bostock JM, Chopra I, Janezic, Gobec S: Discovery of new inhibitors of D-alanine:D-alanine ligase by structure-based virtual screening. *J Med Chem* **51**:7442-7448, 2008.

Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M: LigScore: a novel scoring function for predicting binding affinities. *J Mol Graph Model* **23**:395-407, 2005.

Krier M, Bret G, Rognan D: Assessing the Scaffold Diversity of Screening Libraries. *J Chem Inf Model* **46**:512–524, 2006.

Krishnamurthy VM, Bohall BR, Semetey V, Whitesides GM: The Paradoxical Thermodynamic Basis for the Interaction of Ethylene Glycol, Glycine, and Sarcosine Chains with Bovine Carbonic Anhydrase II: An Unexpected Manifestation of Enthalpy/Entropy Compensation. *J Am Chem Soc* **128**:5802-5812, 2006.

Krovat EM, Langer T: Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J Chem Inf Comput Sci* **44**:1123-1129, 2004.

Kuo CJ, Guo RT, Lu IL, Liu HG, Wu SY, Ko TP, Wang AH, Liang PH: Structure-based inhibitors exhibit differential activities against Helicobacter pylori and Escherichia coli undecaprenyl pyrophosphate synthases. *J Biomed Biotechnol* **2008**:841312, 2008.

Köppen H: Virtual screening – What does it give us? *Curr Opin Drug Discovery Dev* **12**:397-407, 2009.

Ladbury JE, Klebe G, Freire E: Adding calorimetric data to decision making in lead discovery: a hot tip. *Nat Rev Drug Discov* **9**:23-27, 2010.

Labute P: Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **75**:187-205, 2009.

Lagorce D, Sperandio O, Galons H, Miteva MA, Villoutreix BO: FAF-Drugs2: Free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* **9**:396, 2008.

Lagorce D, Pencheva T, Villoutreix BO, Miteva MA: DG-AMMOS: A New tool to generate 3D conformation of small molecules using Distance Geometry and Automated Molecular Mechanics Optimization for in silico screening. *BMC Chemical Biology* **9**:6, 2009.

Lakshmi PJ, Kumar BV, Nayana RS, Mohan MS, Bolligarla R, Das SK, Bhanu MU, Kondapi AK, Ravikumar M: Design, synthesis, and discovery of novel non-peptide inhibitor of Caspase-3 using ligand based and structure based virtual screening approach. *Bioorg Med Chem* **17**:6040-6047, 2009.

Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID: DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **15**:1219-1230, 2009.

Langer T, Hoffmann RD: A personal foreword. In: *Pharmacophores and pharmacophore searches*, p. 15. Eds. Langer T, Hoffman RD, Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, Germany, 2006.

Larbig G, Pickhardt M, Lloyd DG, Schmidt B, Mandelkow E: Screening for Inhibitors of Tau Protein Aggregation into Alzheimer Paired Helical Filaments: A Ligand Based Approach Results in Successful Scaffold Hopping. *Curr Alzheimer Res* **4**:315-323, 2007.

Lau JF, Jeppesen CB, Rimvall K, Hohlweg R: Ureas with histamine H3-antagonist receptor activity, a new scaffold discovered by lead-hopping from cinnamic acid amides. *Bioorg Med Chem Lett* **16**:5303–5308, 2006.

Leach AR: *Molecular modelling: Principles and Applications*. Prentice Hall, 2nd edition, 2001.

Leach AR, Shoichet BK, Peishoff CE: Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J Med Chem* **49**:5851-5855, 2006.

Leach AR, Gillett VJ, Lewis RA, Taylor R: Three-dimensional Pharmacophore Methods in Drug Discovery. *J Med Chem* **53**:539-558, 2010.

Lee K, Kim J, Jeong KW, Lee KW, Lee Y, Song JY, Kim MS, Lee GS, Kim Y: Structure-based virtual screening of Src kinase inhibitors. *Bioorg Med Chem* **17**:3152-3161, 2009.

Lemmen C, Hiller C, Lengauer T: RigFit: A new approach to superimposing ligand molecules. *J Comput Aided Mol Des* **12**:491-502, 1998a.

Lemmen C, Lengauer T, Klebe G: FLEXS: A Method for Fast Flexible Ligand Superposition. *J Med Chem* **41**:4502–4520, 1998b.

Lewell XQ, Judd DB, Watson SP, Hann MM: RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J Chem Inf Comput Sci* **38**:511–522, 1998.

Li H, Huang J, Chen L, Liu X, Chen T, Zhu J, Lu W, Shen X, Li J, Hilgenfeld R, Jiang H: Identification of novel falcipain-2 inhibitors as potential antimalarial agents through structure-based virtual screening. *J Med Chem* **52**:4936-4940, 2009.

Li HF, Lu T, Zhu T, Jiang YJ, Rao SS, Hu LY, Xin BT, Chen YD: Virtual screening for Raf-1 kinase inhibitors based on pharmacophore model of substituted ureas. *Eur J Med Chem* **44**:1240-1249, 2009.

Li M, Huang YJ, Tai PC, Wang B: Discovery of the first SecA inhibitors using structure-based virtual screening. *Biochem Biophys Res Commun* **368**:839-845, 2008.

Li M, Ni N, Chou HT, Lu CD, Tai PC, Wang B: Structure-based discovery and experimental verification of novel Al-2 quorum sensing inhibitors against Vibrio harveyi. *ChemMedChem* **3**:1242-1249, 2008.

Li N, Wang F, Niu S, Cao J, Wu K, Li Y, Yin N, Zhang X, Zhu W, Yin Y: Discovery of novel inhibitors of Streptococcus pneumoniae based on the virtual screening with the homology-modeled structure of histidine kinase (VicK). *BMC Microbiol* **9**:129, 2009.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv* **23**:3-25, 1997.

Lipinski CA: Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods*, **44**:235–249, 2000.

Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt WF, Schenck RJ, Trippe AJ: Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS registry. *J Org Chem* **73**:4443-4451, 2008.

Liu HY, Kuntz ID, Zou X: Pairwise GB/SA Scoring Function for Structure-based Drug Design. *J Phys Chem B* **108**:5453-5462, 2004.

Low CMR, Buck IM, Cooke T, Cushnir JR, Kalindjian SB, Kotecha A, Pether MJ, Shankley NP, Vinter JG, Wright L: Scaffold Hopping with Molecular Field Points: Identification of a Cholecystokinin-2 (CCK2) Receptor Pharmacophore and Its Use in the Design of a Prototypical Series of Pyrrole- and Imidazole-Based CCK2 Antagonists. *J Med Chem* **48**:6790–6802, 2005.

Lyne PD: Structure-based virtual screening: an overview. *Drug Discovery Today* **7**:1047-1055, 2002.

Mackey MD, Melville JL: Better than Random? The Chemotype Enrichment Problem. *J Chem Inf Model* **49**:1154-1162, 2009.

Maclean D, Baldwin JJ, Ivanov VT, Kato Y, Shaw A, Schneider P, Gordon M: Glossary of Terms Used in Combinatorial Chemistry. *Pure Appl Chem* **71**:2349-2365, 1999.

Maggiora GM, Shanmugasundaram V: Molecular Similarity Principles. In: *Methods in Molecular Biology vol. 275: Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, pp. 1-49. Ed. Bajorath J, Humana Press, New Jersey, U.S.A., 2004.

Majeux N, Scarsi M, Caflisch A: Efficient electrostatic solvation model for protein-fragment docking. *Proteins: Struct Funct Genet* **42**:256-268, 2001.

Makara GM: Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J Med Chem* **44**:3563–3571, 2001.

Manchester J, Walkup G, Rivin O, You Z: Evaluation of pKa Estimation Methods on 211 Druglike compounds. *J Chem Inf Model* doi: 10.1021/ci100019p, 2010.

Markt P, Petersen RK, Flindt EN, Kristiansen K, Kirchmair J, Spitzer G, Distinto S, Schuster D, Wolber G, Laggner C, Langer T: Discovery of novel PPAR ligands by a virtual screening approach based on pharmacophore modeling, 3D shape, and electrostatic similarity screening. *J Med Chem* **51**:6303-6317, 2008.

Markt P, Feldmann C, Rollinger JM, Raduner S, Schuster D, Kirchmair J, Distinto S, Spitzer GM, Wolber G, Laggner C, Altmann KH, Langer T, Gertsch J: Discovery of novel CB2 receptor ligands by a pharmacophore-based virtual screening workflow. *J Med Chem* **52**:369-378, 2009.

Markt P, Schuster D, Kirchmair J, Laggner C, Langer T: Pharmacophore modeling and parallel screening for PPAR ligands. *J Comput Aided Mol Des* **21**:575-590, 2007.

Markush EA: Pyrazolone Dye and Process of Making the Same. *U.S. Patent* 1506316, 1924.

Marshall GR, Barry CD, Bosshard HE, Dammkoehler R, Dunn DA: The conformational Parameter in Drug Design: The Active Analog Approach. In: *Computer-Assisted Drug Design*, pp.205-226. Eds. Olson E, Christoffersen RE, American Chemical Society, Columbus, U.S.A., 1979.

Martin EJ, Hoeffel TH: Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. *J Mol Graph Model* **18**:383–403, 2000.

Martin YC: Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J Comb Chem* **3**:231–250, 2001.

Martin YC, Kofron JL, Traphagen LM: Do structurally similar molecules have similar biological activity? *J Med Chem* **45**:4350–4358, 2002.

Martin YC: Let's not forget tautomers. *J Comput Aided Mol Des* **23**:693-704, 2009.

Matter H: Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J Med Chem* **40**:1219–1229, 1997.

Matthews BW: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**:442-451, 1975.

McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK: Gaussian docking fuctions. *Biopolymers* **68**:76-90, 2003.

McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD: Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J Chem Inf Model* **47**:1504-1519, 2007.

McGregor JJ, Willett P: Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J Chem Inf Comput Sci* **21**:137–140, 1981.

McGregor MJ, Muskal SM: Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J Chem Inf Comput Sci* **39**:569–574, 1999.

McGregor MJ, Muskal SM: Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J Chem Inf Comput Sci* **40**:117–125, 2000.

McInnes C: Virtual screening strategies in drug discovery. *Curr Opin Chem Biol* **11**:494-502, 2007.

Medina-Franco JL, Martínez-Mayorga KM, Bender A, Scior A: Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb Sci* **28**:1551-1560, 2009.

Meiler J, Baker D: ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins: Struct Funct Bioinf* **65**:538-548, 2006.

Meng EC, Shoichet BK, Kuntz ID: Automated docking with grid-based energy evaluation. *J Comp Chem* **13**:505-524, 1992.

Michino M, Abola E, Brooks CL, Dixon JS, Moult J, Stevens RC: Community-wide assessment of GPCR structure modeling and ligand docking: GPCR Dock 2008. *Nat Rev Drug Discov* **8**:455-463, 2009.

Milletti F, Storchi L, Sforna G, Cruciani G: New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J Chem Inf Model* **47**:2172-2181, 2007.

Milletti F, Storchi L, Sforna G, Cross S, Cruciani G: Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J Chem Inf Model* **49**:68-75, 2009.

Mills JEJ, Dean PM: Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J Comp Aided Mol Des* **10**:607-622, 1996.

Mitsui T, Hirayama K, Aoki S, Nishikawa K, Uchida K, Matsumoto T, Kabuta T, Wada K: Identification of a novel chemical potentiator and inhibitors of UCH-L1 by in silico drug screening. *Neurochem Int* **56**:679-686, 2010.

Moitessier N, Therrien E, Hanessian S: A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic β-secretase (BASE 1) inhibitors. *J Med Chem* **49**:5885-5894, 2006.

Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR: Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol* **153**:S7-S26, 2008.

Mooij WTM, Verdonk ML: General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct Funct Bioinf* **61**:272-287, 2005.

Morley SD, Afshar M: Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock. *J Comp Aided Mol Des* **18**:189-208, 2004.

Morris GM, Goodsell DS, Halliday RD, Huey R, Hart WE, Belew RK, Olson AJ: Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J Comput Chem* **19**:1639-1662, 1998.

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**:2785-2791, 2009.

Muchmore S, Souers AJ, Akritopoulou-Zanze I: The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chem Biol Drug Des* **67**:174-176, 2006.

Muegge I: PMF scoring revisited. *J Med Chem* **49**:5895-5902, 2006.

Mukherjee P, Desai P, Ross L, White EL, Avery MA: Structure-based virtual screening against SARS-3CL(pro) to identify novel non-peptidic hits. *Bioorg Med Chem* **16**:4138-4149, 2008.

Naerum L, Norskov-Lauritsen L, Olesen PH: Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg Med Chem Lett* **12**:1525–1528, 2002.

NASA: On the expansion of the Universe. NASA Glenn Research Centre, U.S.A., 2009. http://www.grc.nasa.gov/WWW/K-12/Numbers/Math/documents/ON_the_EXPANSION_of_the_UNIVERSE.pdf.

Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R, Mizote A, Rosen D, Thomas JM, Izumi M, Ganesan A, Galione A, Churchill GC: Identification of a chemical probe for NAADP by virtual screening. *Nat Chem Biol* **5**:220-226, 2009.

Nelson DL, Cox MM. *Lehninger Principles of Biochemistry*, 4th edition. pp. 157-158. W.H. Freeman and Company, New York, U.S.A., 2005.

Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M: Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J Med Chem* **49**:6802–6810, 2006.

Neves MA, Dinis TC, Colombo G, Sá e Melo ML: An efficient steroid pharmacophore-based strategy to identify new aromatase inhibitors. *Eur J Med Chem* **44**:4121-4127, 2009.

Nicholls A, MacCuish NE, MacCuish JD: Variable selection and model validation of 2D and 3D molecular descriptors. *J Comput Aided Mol Des* **18**:451-474, 2004.

Nicholls A: What Do We Know and When Do We Know It? *J Comput Aided Mol Des* **22**:239-255, 2008.

Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B: Molecular Shape and Medicinal Chemistry: A Perspective. *J Med Chem* doi: 10.1021/jm900818s, 2010.

Niedermeider S, Singethan K, Sebastian G, Matz M, Kossner M, Diederich S, Maisner A, Schmitz J, Hiltensperger G, Baumann K, Holzgrabe U, Schneider-Schaulies J: A small-molecule inhibitor of Nipah virus envelope protein-mediated membrane fusion. *J Med Chem* **52**:4257-4265, 2009.

O'Driscoll C: A Virtual Space Odyssey. In: *The 4th Horizon Symposium*, U.S.A., 2004. www.nature.com/horizon/chemicalspace/background/pdf/odyssey.pdf.

Ohno K, Nagahara Y, Tsunoyama K, Orita M: Are There Differences between Launched Drugs, Clinical Candidates, and Commercially Available Compounds? *J Chem Inf Model* doi: 10.1021/ci100023s, 2010.

Okamoto M, Takayama K, Shimizu T, Ishida K, Takahashi O, Furuya T: Identification of death-associated protein kinases inhibitors using structure-based virtual screening. *J Med Chem* **52**:7323-7327, 2009.

O'Meara JA, Jakalian A, LaPlante S, Bonneau PR, Coulombe R, Faucher AM, Guse I, Landry S, Racine J, Simoneau B, Thavonekham B, Yoakima C: Scaffold hopping in the rational design of novel HIV-1 non-nucleoside reverse transcriptase inhibitors. *Bioorg Med Chem Lett* **17**:3362–3366, 2007.

Oprea TI, Davis AM, Teague SJ, Leeson PD: Is There a Difference between Leads and Drugs? A Historical Perspective. *J Chem Inf Comput Sci* **41**:1308-1315, 2001.

Oprea TI, Allu TK, Fara DC, Rad RF, Ostopovici L, Bologa CG: Lead-like, drug-like or "Pub-like": how different are they? *J Comput Aided Mol Des* **21**:113-119, 2007.

Ostrov DA, Magis TA, Wronski TJ, Chan EK, Toro EJ, Donatelli RE, Sajek K, Haroun IN, Nagib MI, Piedrahita A, Harris A, Holliday LS: Identification of Enoxacin as an Inhibitor of Osteoclast Formation

and Bone Resorption by Structure-Based Virtual Screening. *J Med Chem* **52**:5144-5151, 2009.

Oyarzabal J, Howe T, Alcazar J, Andrés JI, Alvarez RM, Dautzenberg F, Iturrino L, Martínez S, Van der Linden I: Novel approach for Chemotype Hopping Based on Annotated Databases of Chemically Feasible Fragments and a prospective case study: new melanin concentrating hormone antagonists. *J Med Chem* **52**:2076-2089, 2009.

Park H, Lee J, Lee S: Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* **65**:459-554, 2006.

Park H, Bahn YJ, Jeong DG, Woo EJ, Kwon JS, Ryu SE: Identification of novel inhibitors of extracellular signal-regulated kinase 2 based on the structure-based virtual screening. *Bioorg Med Chem Lett* **18**:5372-5376, 2008.

Park H, Bhattarai BR, Ham SW, Cho H: Structure-based virtual screening approach to identify novel classes of PTP1B inhibitors. *Eur J Med Chem* **44**:3280-3284, 2009.

Park H, Jung SK, Jeong DG, Ryu SE, Kim SJ: Discovery of novel PRL-3 inhibitors based on the structure-based virtual screening. *Bioorg Med Chem Lett* **18**:2250-2255, 2008.

Park H, Jung SK, Jeong DG, Ryu SE, Kim SJ: Discovery of VHR phosphatase inhibitors with micromolar activity based on structure-based virtual screening. *ChemMedChem* **3**:877-880, 2008.

Park H, Li M, Choi J, Cho H, Ham SW: Structure-based virtual screening approach to identify novel classes of Cdc25B phosphatase inhibitors. *Bioorg Med Chem Lett* **19**:4372-4375, 2009.

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

Patani GA, LaVoie EJ: Bioisosterism: A Rational Approach in Drug Design. *Chem Rev* **96**:3147–3176, 1996.

Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J Med Chem* **39**:3049–3059, 1996.

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**:203-214, 2010.

Pham TA, Jain AN: Parameter estimation for scoring protein-ligand interactions using negative training data. *J Med Chem* **49**:5856-5868, 2006.

Pham TA, Jain AN: Customizing scoring functions for docking. *J Comput Aided Mol Des* **22**:269-286, 2008.

Perez-Pineiro R, Burgos A, Jones DC, Andrew LC, Rodriguez H, Suarez M, Fairlamb AH, Wishart DS: Development of a novel virtual screening cascade protocol to identify potential trypanothione reductase inhibitors. *J Med Chem* **52**:1670-1680, 2009.

Poptodorov K, Luu T, Hoffman RD: Pharmacophore Model Generation Software Tools. In: *Pharmacophores and Pharmacophore Searches*, pp. 17-47. Eds. Langer T, Hoffmann RD, Wiley-VCH Verlag, Weinheim, Germany, 2006.

Pospisil P, Ballmer P, Scapozza L, Folkers G: Tautomerism in Computer-Aided Drug Design. *J Recept Signal Transduction Res* **23**:361-371, 2003.

Pólgar T, Magyar C, Simon I, Keserü GM: Impact of Ligand Protonation on Virtual Screening against β-Secretase (BACE1). *J Chem Inf Model* **47**:2366-2373, 2007.

Putta S, Lemmen C, Beroza P, Greene J: A novel shape-feature based approach to virtual library screening. *J Chem Inf Comput Sci* **42**:1230-1240, 2002.

Putta S, Beroza P: Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery. *Cutt Top Med Chem* **7**:1514-1524, 2007.

Qiu J, Xiao J, Han C, Li N, Shen X, Jiang H, Cao X: Potentiation of tumor necrosis factor-alpha-induced tumor cell apoptosis by a small molecule inhibitor for anti-apoptotic protein hPEBP4. *J Biol Chem* **285**:12241-12247, 2010.

Rao DG: *Introduction to Biochemical Engineering*. pp. 84-86. Tata McGraw-Hill Publishing Company Limited, New Delhi, India, 2005.

Rarey M, Kramer B, Lengauer T, Klebe G: A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**:470-489, 1996.

Ravindranathan KP, Mandiyan V, Ekkati AR, Bae JH, Schlessinger J, Jorgensen WL: Discovery of novel fibroblast growth factor receptor 1 kinase inhibitors by structure-based virtual screening. *J Med Chem* **53**:1662-1672, 2010.

Raymond JW, Willett P: Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J Comput Aided Mol Des* **16**:59-71, 2002.

Ren JX, Li LL, Zou J, Yang L, Yang JL, Yang SY: Pharmacophore modeling and virtual screening for the discovery of new transforming growth factor-beta type I receptor (ALK5) inhibitors. *Eur J Med Chem* **44**:4259-4265, 2009.

Renner S, Ludwig V, Boden O, Scheffer U, Gobel M, Schneider G: New Inhibitors of the Tat-TAR RNA Interaction Found with a "Fuzzy" Pharmacophore Mode. *ChemBioChem* **6**:1119–1125, 2005.

Renner S, Schwab CH, Gasteiger J, Schneider G: Impact of Conformational Flexibility on Three-Dimensional Similarity Searching Using Correlation Vectors. *J Chem Inf Model* **46**:2324-2332, 2006.

Rester U: From virtuality to reality – Virtual screening in lead discovery and lead optimization: A medicinal chemistry perspective. *Curr Opin Drug Discovery Dev* **11**:559-568, 2008.

Ripka AS, Rich DH: Peptidomimetic design. *Curr Opin Chem Biol* **2**:441–452, 1998.

Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE: LeadScope: Software for Exploring Large Sets of Screening Data. *J Chem Inf Comput Sci* **40**:1302–1314, 2000.

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

Rollinger JM, Kratschmar DV, Schuster D, Pfisterer PH, Gumy C, Aubry EM, Brandstötter S, Stuppner H, Wolber G, Odermatt A: 11β-Hydroxysteroid dehydrogenase 1 inhibiting constituents from Eriobotrya japonica revealed by bioactivity-guided isolation and computational approaches. *Bioorg Med Chem* doi:10.1016/j.bmc.2010.01.010, 2010.

Rush TS, Grant JA, Mosyak L, Nicholls A: A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **48**:1489–1495, 2005.

Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP, Blackburn GM, Hay AJ, Gamblin SJ, Skehel JJ: The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**:45-49, 2006.

Rönkkö T, Tervo AJ, Parkkinen J, Poso A: BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II.

Description and characterization. *J Comput Aided Mol Des* **20**:227–236, 2006.

Rönkkö T: *Brutus – A Molecular Energy Field Superposition Algorithm for Virtual Screening*. Doctoral Dissertation, University of Kuopio, Finland 2009.

Sadowski J, Gasteiger J: From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem Rev* **93**:2567-2581, 1993.

Saeh JC, Lyne PD, Takasaki BK, Cosgrove DA: Lead hopping using SVM and 3D pharmacophore fingerprints. *J Chem Inf Model* **45**:1122–1133, 2005.

Salam NK, Huang TH, Kota BP, Kim MS, Li Y, Hibbs DE: Novel PPAR-gamma agonists identified from a natural product library: a virtual screening, induced-fit docking and biological assay study. *Chem Biol Drug Des* **71**:57-70, 2008.

Sanam R, Vadivelan S, Tajne S, Narasu L, Rambabu G, Jagarlapudi SA: Discovery of potential ZAP-70 kinase inhibitors: pharmacophore design, database screening and docking studies. *Eur J Med Chem* **44**:4793-4800, 2009.

Sanner MF: Python: a programming language for software integration and development. *J Mol Graph Model* **17**:57-61, 1999.

Sauer WHB, Schwarz MK: Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J Chem Inf Comput Sci* **43**:987–1003, 2003.

Sauton N, Lagorce D, Villoutreix BO, Miteva MA: MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics* **9**:184, 2008.

Schuffenhauer A, Gillet VJ, Willett P: Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J Chem Inf Comput Sci* **40**:295–307, 2000.

Sciabola S, Morao I, de Groot MJ: Pharmacophoric Fingerprint Method (TOPP) for 3D-QSAR Modeling: Application to CYP2D6 Metabolic Stability. *J Chem Inf Model* **47**:76–84, 2007.

Schnecke V, Kuhn LA: Virtual screening with solvation and ligand-induced complementary. *Persp Drug Discov Des* **20**:171-190, 2000.

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

Schneider G, Schneider P, Renner S: Scaffold-Hopping: How Far Can You Jump? *QSAR Comb Sci* **12**:1162-1171, 2006.

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ: PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* **33**:W363-W367, 2005.

Schueler FW: Sex hormone action and chemical constitution. *Science* **103**:221-223, 1946.

Schuster D, Maurer EM, Laggner C, Nashev LG, Wilckens T, Langer T, Odermatt A: The Discovery of New 11β-Hydroxysteroid Dehydrogenase Type 1 Inhibitors by Common Feature Pharmacophore Modeling and Virtual Screening. *J Med Chem* **49**:3454-3466, 2006.

Schwab CH: Conformational Analysis and Searching. In: *Handbook of Chemoinformatics*, pp. 262-301, Ed. Gasteiger J, Wiley-VCH, New York, U.S.A, 2003.

Sela I, Golan G, Strajbl M, Rivenzon-Segal D, Bar-Haim S, Bloch I, Inbal B, Shitrit A, Ben-Zeev E, Fichman M, Markus Y, Marantz Y, Senderowitz H, Kalid O: G Protein Coupled Receptors - In Silico Drug Discovery and Design. *Curr Top Med Chem* **10**:638-656, 2010.

Shah S, Federoff, HJ: Drug discovery dilemma and Cura Quartet collaboration. *Drug Discovery Today* **14**:1006-1010, 2009.

Sheridan RP, Singh SB, Fluder EM, Kearsley SK: Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J Chem Inf Comput Sci* **41**:1395-1406, 2001.

Sheridan RP, Kearsley SK: Why do we need so many chemical similarity search methods? *Drug Discovery Today* **7**:903-911, 2002.

Sing T, Sander O, Beerenwinkel N, Lengauer T: ROCR: visualizing classifier performance in R. *Bioinformatics* **21**:3940-3941, 2006.

Smits RA, de Esch IJ, Zuiderveld OP, Broeker J, Sansuk K, Guaita E, Coruzzi G, Adami M, Haaksma E, Leurs R: Discovery of quinazolines as histamine H4 receptor inverse agonists using a scaffold hopping approach. *J Med Chem* **51**:7855-7865, 2008.

Sonego P, Kocsor A, Pongor S: ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief Bioinform* **9**:198-209, 2008.

Srinivasan J, Castellino A, Bradley EK, Eksterowicz JE, Grootenhuis PD, Putta S, Stanton RV: Evaluation of a novel shape-based computational filter for lead evolution: application to thrombin inhibitors. *J Med Chem* **45**:2494-2500, 2002.

Stahl M, Rarey M: Detailed Analysis of Scoring Functions for Virtual Screening. *J Med Chem* **44**:1035-1042, 2001.

Stanton DT, Morris TW, Roychoudhury S, Parker CN: Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J Chem Inf Comput Sci* **39**:21–27, 1999.

Stiefl N, Watson IA, Baumann K, Zaliani A: ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J Chem Inf Model* **46**:208–220, 2006.

Stiefl N, Zaliani A: A knowledge-based weighting approach to ligand-based virtual screening. *J Chem Inf Model* **46**:587–596, 2006.

Su AI, Lorber DM, Weston GS, Baase WA, Matthews BW, Shoichet BK: Docking molecules by families to increase the diversity of hits in database screens: Computational strategy and experimental evaluation. *Proteins* **42**:279–293, 2001.

Tasler S, Müller O, Wieber T, Herz T, Krauss R, Totzke F, Kubbutat MH, Schächtele C: N-substituted 2'-(aminoaryl)benzothiazoles as kinase inhibitors: hit identification and scaffold hopping. *Bioorg Med Chem Lett* **19**:1349-1356, 2009.

Tawa GJ, Baber JC, Humblet C: Computation of 3D queries for ROCS based virtual screens. *J Comput Aided Mol Des* **23**:853-868, 2009.

Teague SJ, Davis AM, Leeson PD, Oprea T: The Design of Leadlike Combinatorial Libraries. *Angew Chem Int Ed* **38**:3743-3748, 1999.

ten Brink T, Exner TE: Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J Chem Inf Model* **49**:1535-1546, 2009.

Thomsen R, Christensen MH: MolDock: A New Technique for High-Accuracy Molecular Docking. *J Med Chem* **49**:3315-3321, 2006.

Tietze S, Apostolakis J: Glamdock: Development and validation of a new docking tool on several thousand protein-ligand complexes. *J Chem Inf Model* **47**:1657-1672, 2007.

Tiikkainen P, Markt P, Wolber G, Kirchmair J, Distinto S, Poso A, Kallioniemi O: Critical Comparison of Virtual Screening Methods against the MUV Data Set. *J Chem Inf Model* **49**:2168-2178, 2009.

Tirado-Rivers J, Jorgensen WL: Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein−Ligand Binding. *J Med Chem* **49**:5880-5884, 2006.

Todeschini R, Consonni V: *Molecular Descriptors for Chemoinformatics*. Wiley-VCH Verlag, Weinheim, Germany, 2009.

Todorov NP, Monthoux PH, Alberts IL: The Influence of Variations of Ligand Protonation and Tautomerism on Protein-Ligand Recognition and Binding Energy Landscape. *J Chem Inf Model* **46**:1134-1142, 2006.

Tonelli M, Boido V, La Colla P, Loddo R, Posocco P, Paneni MS, Fermeglia M, Pricl S: Pharmacophore modeling, resistant mutant isolation, docking, and MM-PBSA analysis: combined experimental/computer-assisted approaches to identify new inhibitors of the Bovine Viral Diarrhea Virus (BVDV). *Bioorg Med Chem* In press, 2010.

Tralau-Stewart CJ, Wyatt CA, Kleyn DE, Ayad A: Drug discovery: new models for industry–academic partnerships. *Drug Discovery Today* **14**:95-101, 2009.

Tresadern H, Cid JM, Macdonald GJ, Vega JA, de Lucas AI, García A, Matesanz E, Linares ML, Oehlrich D, Lavreysen H, Biesmans I, Trabanco AA: Scaffold hopping from pyridines to imizao[1,2-a]pyridines. New positive allosteric modulators of metabotropic glutamate 2 receptor. *Bioorg Med Chem Lett* **20**:175-179, 2010.

Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO: Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J Med Chem* **48**:2534-2547, 2005.

Triballeau N, Bertrand HO, Acher F: Are You Sure You Have a Good Model? In: *Pharmacophores and Pharmacophore Searches*, pp. 325-364. Eds. Langer T, Hoffmann RD, Wiley-VCH Verlag, Weinheim, Germany, 2006.

Tripos: UNITY Manual (SYBYL-X). 2009. www.tripos.com.

Trott O, Olson AJ: AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comp Chem* **31**:455-461, 2010.

Truchon JF, Bayly CI: Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J Chem Inf Model* **47**:488-508, 2007.

Trump RP, Blanc JBE, Stewart EL, Brown PJ, Caivano M, Gray DW, Hoekstra WJ, Willson TM, Han B, Turnbull P: Design and synthesis of an array of selective androgen receptor modulators. *J Comb Chem* **9**:107-114, 2007.

Tsai KC, Teng LW, Shao YM, Chen YC, Lee YC, Li M, Hsiao NW: The first pharmacophore model for potent NF-kappaB inhibitors. *Bioorg Med Chem Lett* **19**:5665-5669, 2009.

Tsuchida K, Chaki H, Takakura T, Kotsubo H, Tanaka T, Aikawa Y, Shiozawa S, Hirono S: Discovery of Nonpeptidic Small-Molecule AP-1 Inhibitors: Lead Hopping Based on a Three-Dimensional Pharmacophore Model. *J Med Chem* **49**:80–91, 2006.

Turk S, Kovac A, Boniface A, Bostock JM, Chopra I, Bianot D, Gobec S: Discovery of new inhibitors of the bacterial peptidoglycan biosynthesis enzymes MurD and MurF by structure-based virtual screening. *Bioorg Med Chem* **17**:1884-1889, 2009.

Vadivelan S, Sinha BN, Tajne S, Jagarlapudi SA: Fragment and knowledge-based design of selective GSK-3beta inhibitors using virtual screening models. *Eur J Med Chem* **44**:2361-2371, 2009.

Vainio MJ, Johnson MS: Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J Chem Inf Model* **47**:2462-2474, 2007.

Vainio MJ, Puranen JS, Johnson MS: ShaEP: molecular overlay based on shape and electrostatic potential. *J Chem Inf Model* **49**:492-502, 2009.

van Dijk M, van Dijk AD, Hsu V, Boelens R, Bonvin AM: Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res* **34**:3317-3325, 2006.

van Drie JH: Pharmacophore discovery: A critical review. In: *Computational Medicinal Chemistry for Drug Discovery*, pp. 437–455. Eds. Bultinck P, De Winter H, Langenaeker W, Tollenare JP, Marcel Dekker, New York & Basel, U.S.A., 2004.

Vargyas M, Papp J, Csizmadia F, Csepregi S, Allardyce A, Vadasz P: Maximum Common Substructure Based Hierarchical Clustering. *American Chemical Society National Meeting*, 11-14 September, 2006.

Velec HFG, Gohlke H, Klebe G: DrugScoreCSD-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem* **48**:6296-6303, 2005.

Venkatachalam CM, Jiang X, Oldfield T, Waldman M: Ligandfit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites. *J Mol Graphics Model* **4**:289-307, 2003.

Venkatraman V, Chakravarthy PR, Kihara D: Application of 3D Zernike descriptors to shape-based ligand similarity searching. *J Cheminf* **1**:19, 2009.

Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD: Improved protein-ligand docking using GOLD. *Proteins: Struct Funct Bioinf* **52**:609-623, 2003.

Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P: Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **44**:793-806, 2004.

Verkivker GM, Bouzida D, Gehlaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW: Deciphering common failures in molecular docking of ligand-protein complexes. *J Comput Aid Mol Des* **14**:731-751, 2000.

Vidal D, Thormann M, Pons M: A novel search engine for virtual screening of very large databases. *J Chem Inf Model* **46**:836–846, 2006.

Vieth M, Hirst JD, Dominy BN, Daigler H, Brooks III CL: Assessing search strategies for flexible docking. *J Comput Chem* **19**:1623-1631, 1999.

Viji SN, Prasad PA, Gautham N: Protein−Ligand Docking Using Mutually Orthogonal Latin Squares (MOLSDOCK). *J Chem Inf Model* **49**:2687-2694, 2009.

von Korff M, Freyss J, Sander T: Comparison of ligand- and structure-based virtual screening on the DUD data set. *J Chem Inf Model* **49**:209-231, 2009.

Wang H, Liu Y, Huai Q, Cai J, Zaraghi R, Francis SH, Corbin JD, Robinson H, Xin Z, Lin G, Ke H: Multiple conformation of phosphodiesterase-5: implications for enzyme function and drug development. *J Biol Chem* **281**:21469-21479, 2006.

Wang HY, Li LL, Cao ZX, Luo SD, Wei YQ, Yang SY: A specific pharmacophore model of Aurora B kinase inhibitors and virtual screening studies based on it. *Chem Biol Drug Des* **7**:115-126, 2009.

Wang R, Lai L, Wang S: Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **16**:11-26, 2002.

Wang R, Lu Y, Wang S: Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J Med Chem* **46**:2287-2303, 2003.

Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS: A critical assessment of docking programs and scoring functions. *J Med Chem* **49**:5912-5931, 2006.

Watts KS, Dalal P, Murphy RB, Sherman W, Friesner RA, Shelley JC: ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformations. *J Chem Inf Model* doi: 10.1021/ci100015j, 2010.

Weidlich IE, Dexheimer T, Marchand C, Antony S, Pommier Y, Nicklaus MC: Inhibitors of human tyrosyl-DNA phospodiesterase (hTdp1) developed by virtual screening using ligand-based pharmacophores. *Bioorg Med Chem* **18**:182-189, 2010.

Wermuth CG: Pharmacophores: Historical perspective and viewpoint from a medicinal chemist. In: *Pharmacophores and Pharmacophore Searches*, pp. 3-13. Eds. Langer T, Hoffmann RD, Wiley-VCH Verlag, Weinheim, Germany, 2006.

Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA: Glossary of terms used in medicinal chemistry. *Pure & Appl Chem* **70**:1129–1143, 1998.

Wesson L, Eisenberg D: Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci* **1**:227-235, 1992.

Weston J, Perez-Cruz F, Bousquet O, Chapelle O, Elisseef A, Scholkopf B: Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics* **19**:764-771, 2003.

Whitesides GM, Krishnamurthy VM: Designing ligands to bind proteins. *Q Rev Biophys* **38**:385-395, 2005.

Wild DJ, Blankley CJ: Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J Chem Inf Comput Sci* **40**:155–162, 2000.

Wilkens SJ, Janes J, Su AI: HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J Med Chem* **48**:3182–3193, 2005.

Willett P, Barnard JM, Downs GM: Chemical similarity searching. *J Chem Inf Comput Sci* **38**:983–996, 1998.

Willett P: Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **11**:1046-1053, 2006.

Williams C: Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol Divers* **10**:311–332, 2006.

Wolber G, Langer T: LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model* **45**:160-169, 2005.

Wold S, Ruhe A, Wold H, Dunn WJ: The colinearity problem in linear regression: the partial least squares (PLS) approach to generalised inverses. *SIAM J Sci Stat Comput* **5**:735–743, 1984.

Woods DD, Fildes P: The anti-sulphanilamide activity (in vitro) of p-aminobenzoic acid and related compounds. *Chem Ind* **59**:133-134, 1940.

Wu G, Robertson DH, Brooks CL, Vieth M: Detailed analysis of grid-based molecular docking: A case study of CDOCKER - A CHARMm-based MD docking algorithm. *J Comput Chem* **24**:1549-1562, 2003.

Xie HZ, Li LL, Ren JX, Zou J, Yang L, Wei YQ, Yang SY: Pharmacophore modeling study based on known spleen tyrosine kinase inhibitors together with virtual screening for identifying novel inhibitors. *Bioorg Med Chem Lett* **19**:1944-1949, 2009.

Xu J: A New Approach to Finding Natural Chemical Structure Classes. *J Med Chem* **45**:5311–5320, 2002.

Xu W, Chen G, Liew OW, Zuo Z, Jiang H, Zhu W: Novel non-peptide beta-secretase inhibitors derived from structure-based virtual screening and bioassay. *Bioorg Med Chem Lett* **19**:3188-3192, 2009.

Yang CY, Wang R, Wang S: M-Score: a knowledge-based potential scoring function for protein atom mobility. *J Med Chem* **49**:5903-5911, 2006.

Yang JM, Chen CC: GEMDOCK: A generic evolutionary method for molecular docking. *Proteins: Struct Funct Bioinf* **55**:288-304, 2004.

Yao J, Zhang Q, Min J, He J, Yu Z: Novel enoyl-ACP reductase (FabI) potential inhibitors of Escherichia coli from Chinese medicine monomers. *Bioorg Med Chem Lett* **20**:56-59, 2010.

Zhang Q, Muegge I: Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting and Consensus Scoring. *J Med Chem* **49**:1536–1548, 2006.

Zhang QY, Wan J, Xu X, Yang GF, Ren YL, Liu JJ, Wang H, Guo Y: Structure-Based Rational Quest for Potential Novel Inhibitors of Human HMG-CoA Reductase by Combining CoMFA 3D QSAR Modeling and Virtual Screening. *J Med Chem* **9**:131–138, 2007.

Zhao Y, Sanner MF: FLIPDock: Docking flexible ligands into flexible receptors. *Proteins: Struct Funct Bioinf* **68**:726-737, 2007.

Zhu Z, Cuozzo J. High-Throughput Affinity-Based Technologies for Small-Molecule Drug Discovery. *J Biomol Screen* **14**:1157-1164, 2009.

Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP: eHiTS: A new fast, exhaustive flexible ligand docking system. *J Mol Graph Model* **26**:192-212, 2006.

Tuomo Kalliokoski

Accelerating Three-Dimensional Virtual Screening

New Software and Approaches for Computer-Aided Drug Discovery

Computers are routinely used in the modern drug discovery process. In virtual screening, the bioactivity of a compound is predicted in silico. The focus of this study has been in the development of novel rapid virtual screening software and acceleration of current methods. This dissertation describes new approaches for both protein- and ligand-based virtual screening.

UNIVERSITY OF EASTERN FINLAND

Publications of the University of Eastern Finland
Dissertations in Health Sciences

ISBN 978-952-61-0181-1