Development and application of fast fuzzy pharmacophore-based virtual screening methods for scaffold hopping

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich
Biochemie, Chemie und Pharmazie
der Johann Wolfgang Goethe–Universität
in Frankfurt am Main

von
Steffen Renner
aus Freiburg im Breisgau

Frankfurt 2006
(DF1)

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

Dekan:	Prof. Dr. Harald Schwalbe
erster Gutachter:	Prof. Dr. Gisbert Schneider
zweiter Gutachter:	Prof. Dr. Ernst Egert
Datum der Disputation:	4. Mai 2006

Danksagung

Ich möchte mich an dieser Stelle herzlich bei Prof. Dr. Gisbert Schneider bedanken, für eine sehr schöne, interessante und lehrreiche Zeit, die ich in seiner Arbeitsgruppe verbringen durfte.

Weiterer Dank gilt der Merz Pharmaceuticals GmbH für die Vergabe eines Stipendiums, besonders meiner dortigen Betreuerin Dr. Tanja Weil für eine sehr angenehme und fruchtbare Zusammenarbeit.

Der Arbeitsgruppe danke ich für die schöne Zeit in Frankfurt, viel Hilfe und manchen Programmcode. Vielen Dank vor allem an Uli Fechner, Michael Schmucker, Lutz Franke, Evgeny Byvatov, Tobias Noeske, Alexander Böcker, Andreas Schüller, Michael Meissner, Tina Grabowski, Manuel Nietert, Swetlana Derksen, Petra Schneider, Norbert Dichter, Alireza Givehchi und Brigitte Scheidemantel-Geiß. Vielen Dank auch an Obdulia Rabal und Yusuf Tanrikulu für die Fortführung meiner Arbeit an SQUID. An dieser Stelle geht auch Dank an das Beilstein-Institut zur Förderung der Chemischen Wissenschaften, das die „Beilstein Stiftungsprofessur“ erst ermöglicht hat.

Bei Tobias Noeske, Dr. Mirko Hechenberger und Dr. Chris Parsons möchte ich mich für die experimentellen Messungen im mGluR Projekt bedanken. Dank gilt in diesem Zusammenhang auch Dr. Gabriele Costantio für die angenehme Zusammenarbeit bei der Homologie-Modellierung der mGluRs, auch wenn es dieses Projekt leider nicht mehr rechtzeitig in die Doktorarbeit geschafft hat.

Für die gemeinsame Arbeit an den Inhibitoren der TAR-RNA möchte ich Prof. Dr. Michael Göbel, Verena Ludwig, Oliver Boden und Dr. Ute Scheffer danken.

Prof. Dr. Rolf Marschalek und Jens Rabenstein möchte ich für die Zusammenarbeit an den Taspase1 Liganden danken.

Prof. Dr. Johnny Gasteiger und Dr. Christof Schwab danke ich für die interessante Zusammenarbeit in der Untersuchung der Konformations-Abhängigkeit meiner Deskriptoren.

Besonders möchte ich mich noch bei meiner Frau Katharina, bei meinen Eltern und meinen Schwiegereltern bedanken, die mit ihrer großen Unterstützung einen großen Anteil am Gelingen dieser Arbeit hatten. Der größte Dank geht aber an meine Tochter Lena Maria, die noch gar nicht weiß, was eine Doktorarbeit ist und so viel zur effizienten Fertigstellung derselben beigetragen hat, wenn mal Zeit dafür war.

Publications resulting from this thesis:

Contributions to scientific journals and books:

(1) Fechner, U., Franke, L., Renner, S., Schneider, P., and Schneider, G. (2003) Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des. 17*, 687-698.

(2) Paetz, J., Fechner, U., Franke, L., Renner, S., Schneider, P., and Schneider, G. (2004) Pharmacophore feature selection with a neuro-fuzzy system. In: *Proceedings of the 4th Europ. Symp. on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (EUNITE 2004)*, (Ed. Elite-Foundation), Aachen, Mainz Wissenschaftsverlag, pp. 179-184.

(3) Renner, S., and Schneider, G. (2004) Fuzzy pharmacophore models from molecular alignments for correlation-vector based virtual screening. *J. Med. Chem. 47*, 4653-4664.

(4) Renner, S., Noeske, T., Parsons, C. P., Schneider, P., Weil, T., and Schneider, G. (2005) New allosteric modulators of metabotropic glutamate receptor 5 (mGluR5) found by ligand-based virtual screening. *ChemBioChem*, *6*, 620-625.

(5) Renner, S., Ludwig, V., Boden, O., Scheffer, U., Göbel, M., and Schneider, G. (2005) New inhibitors of the Tat-TAR RNA interaction found with a "fuzzy" pharmacophore model, *ChemBioChem*, *6*, 1119-1125.

(6) Schneider, G., Renner, S., and Fechner, U. (2005) Navigation in chemical space based on correlation-vector representations of molecules. *Proceedings of the Beilstein-Workshop 2004* (Kettner, K., Hicks, M.; Eds.), Logos Verlag, Berlin.

(7) Renner, S., Fechner, U., and Schneider, G. (2006) Alignment-free pharmacophore patterns – A correlation-vector approach. In: Pharmacophores and Pharmacophore Searches (Langer, T. and Hoffmann, E.; Eds), Wiley-VCH, Weinheim, 49-79.

(8) Renner, S., and Schneider, G. (2006) Scaffold-hopping potential of ligand-based similarity concepts, *ChemMedChem, 1,* 181-185.

(9) Schüller, A.P., Fechner, U., Renner, S., Franke, L., Weber, L., and Schneider, G. (2006) A pseudo-ligand approach to virtual screening, *Comb. Chem. High-Throughput Screen , 9,* 359-364.

(10) Renner, S., Schwab, C., Gasteiger, J., and Schneider, G. (2006) Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors, *J. Chem. Inf. Model., 46*, 2324-2332.

Talks:

(1) Renner, S. (2003) Enhancing MOE's Pharmacophore Capabilities, *MOE Usergroup Meeting,* Frankfurt am Main.

(2) Renner, S. (2003) Fuzzy Pharmacophore Models for Rapid Virtual Screening, *Bayer Workshop „Computational Chemistry"*, Wuppertal.

(3) Renner, S., (2005) Successful Virtual Screening for Tat-TAR Interaction Inhibitors with a Fuzzy Pharmacophore Model, *7th International Conference on Chemical Structures*, Noordwijkerhout, Niederlande.

Posters:

(1) Renner, S., Fechner, U., and Schneider, G. (2004) Correlation Vector Approaches for Ligand-Based Similarity Searching. *3rd Joint Sheffield Conference on Cheminformatics*, Sheffield.

(2) Parsons, C.G., Noeske, T., Bauer, T., Renner, S., Schneider, G., and Weil, T. (2004) Non-competitive antagonists of mGluR5 – Assay development and molecular modeling investigations, *Neuroscience*, San Diego.

(3) Renner, S., and Schneider, G. (2005) Correlation vector-based virtual screening approaches for scaffold hopping, *MGMS Annual International Meeting*, Dublin.

(4) Renner, S., Noeske, T., Böcker, A., Weil, T., and Schneider, G. (2005) Combining supervised and unsupervised neural networks for the identification of novel scaffolds of metabotropic glutamate receptor 5 (mGluR5) modulators, *1st German Conference on Chemoinformatics*, Goslar.

(5) Hechenberger, M., Renner, S., Baude, A., Schneider, G., Parsons, C. G., and Weil, T. (2005) Characterization of the MPEP-binding site by computer modeling and functional expression of mGluR5 mutants, *Neuroscience*, Washington.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ACE	Angiotensin converting enzyme
ACHE	Acetylcholinesterase
ADMET	Absorption, distribution, metabolism, excretion, toxicity
ANN	Artificial neural network
ATS	Autocorrelation of a topological structure
BACE	Beta-amyloid converting enzyme
BLAST	Basic local alignment search tool
CAII	Carbonic anhydrase II
CATS	Chemically advanced template search
CATS3D	Chemically advanced template search 3D
cc	Matthews correlation coefficient
COBRA	Collection of bioactive reference analogues
COX2	Cyclooxygenase 2
CRD	Cystein-rich domain
CRF	Corticotropin releasing factor
CV	Correlation-vector
D_1	Manhattan distance
D_2	Euclidean distance
DNA	Desoxyribonucleic acid
DPP	Dipeptidyl-peptidase
ef	Enrichment factor
ELA	Elastase
ES	Evolutionary strategy
FEPOPS	Feature point pharmacophores
FMN	Flavin mononucleotide
FRET	Fluorescence resonance energy transfer
FXA	Factor Xa
GPCR	G-protein coupled receptor
GRIND	Grid independent descriptors
HE	Heptahelical domain
HIV	Human imunodeficiency virus
HIVP	Human imunodeficiency virus protease
HTS	High-throughput screening
IC_{50}	Concentration of a drug that is required for 50 % inhibition
IUPAC	International Union of Pure and Applied Chemistry
K_i	Binding constant
LFD	Local feature density
M5vsCO	mGluR5 vs. COBRA
M5vsCOpca	Principle components of M5vsCO
M5vsM1	mGluR5 vs. mGluR1
M5vsM1pca	Principle components of M5vsM1
MACCS	Molecular access system
Meqi	Molecular equivalence indices
mGluR	Metabotropic glutamate receptor
MLL	Mixed lineage leukemia
MMP	Matrix metalloprotease

MOE	Molecular operating environment
MSE	Mean square error
NEU	Nauraminidase
NIPALS	Nonlinear iterative partial least squares
NK	Neurokinine receptors
NMR	Nuclear magnetic resonance
NUC	Nuclear receptors
PCA	Principle component analysis
PDB	Protein data bank
PLS	Projection to latent structures *or* partial least squares
PPAR	Peroxisome proliferators-activated receptor
PPP	Potential pharmacophore point
PSA	Polar surface area
PTK-CSRC	Protein tyrosine kinase c-src
PTP1B	Protein tyrosine phosphatase 1b
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationship
ReSc	Reduced scaffold
RNA	Ribonucleic acid
RMSD	Root mean square deviation
SAR	Structure activity relationship
Sc	Scaffold
SE	Shannon entropy
SOM	Self-organizing map
sSE	Scaled shannon entropy
SURFCATS	Surface CATS
SQUID	Sophisticated quantification of interaction distributions
STRO1	Stromelysin 1
SVL	Scientific vector language
TAR	Transactivation response element
THR	Thrombin
TPP	Thiamine pyrophosphate
UTPA	Urokinase type plasminogen activator
VFTM	Venus flytrap module
w	Conservation weight

Contents

1 INTRODUCTION 1

1.1 Scope of the thesis 1

1.2 The drug discovery process 3

1.3 Chemoinformatics in the drug discovery process 3

1.4 Virtual screening 4

1.5 Molecular similarity 6

1.6 Scaffold hopping 7

1.7 The pharmacophore concept 8

1.8 Representation of molecules 9

1.9 Autocorrelation descriptors 11

1.10 Retrospective and prospective screening 12

1.11 Artificial neural networks in virtual screening – machine learning based on molecular representations 13

1.12 Incorporating receptor structure information into virtual screening 14

1.13 The metabotropic glutamate receptor 5 (mGluR5) 15

1.14 RNA drug design and the Tat-TAR RNA interaction 18

1.15 Taspase1 20

2 COMPUTATIONAL METHODS 22

2.1 Correlation-vector based descriptors 22
2.1.1 CATS 22
2.1.2 CATS3D 24
2.1.3 SURFCATS 25

2.2 Descriptor vector based virtual screening 26
2.2.1 Retrospective screening evaluation 27

2.3 SQUID 28
2.3.1 Calculation of the SQUID pharmacophore model 29
2.3.2 Virtual Screening 32

2.4 Methods of Section 4.1: Influence of similarity metrics and descriptor vector scaling on CATS3D retrospective screening 33

2.5 Methods of Section 4.2: Impact of conformational flexibility on CATS3D virtual screening 34

2.6 Methods of Section 4.3: Virtual screening and scaffold hopping efficiency of alignment-free pharmacophore pair descriptors 37

2.7 Methods of Section 4.4: Prospective screening for mGlurR5 allosteric modulators with CATS3D 38

2.8 Methods of Section 4.5: Prospective screening for mGluR5 allosteric modulators with an artificial neural network approach based on CATS3D representations 39

2.9 Methods of Section 4.6: Retrospective evaluation of SQUID fuzzy pharmacophore models 46

2.10 Methods of Section 4.7: Prospective screening for inhibitors of the Tat-TAR RNA interaction with a SQUID fuzzy pharmacophore model and CATS3D 47

2.11 Methods of Section 4.8: Prospective screening for taspase1 inhibitors with a receptor-derived pharmacophore model 49

3 EXPERIMENTAL SECTION 52

3.1 Determination of IC_{50} values for mGluR 52

3.2 Determination of IC_{50} values for TAR-RNA 55

4 MAIN SECTION 57

4.1 Influence of similarity metrics and descriptor vector scaling on CATS3D retrospective screening 58
4.1.1 Conclusion 60

4.2 Impact of conformational flexibility on CATS3D virtual screening 62
4.2.1 Calculation of conformations for the PDBbind dataset and the COBRA database 64
4.2.2 Reproducing the crystal-structure conformations of reference ligands 66
4.2.3 Retrospective screening 68
4.2.4 Conclusion 71

4.3 Virtual screening and scaffold hopping efficiency of alignment-free pharmacophore pair descriptors 73
4.3.1 Conclusion 82

4.4 Prospective screening for mGlurR5 allosteric modulators with CATS3D 83
4.4.1 Conclusion 90

4.5 Prospective screening for mGluR5 allosteric modulators with an artificial neural network approach based on CATS3D representations 91
4.5.1 Training of feedforward ANNs 93
4.5.2 Prediction of allosteric mGluR5 modulators 99
4.5.3 Selection of a representative subset by SOMs 100
4.5.4 Binding assay results 104
4.5.5 Conclusions 107

4.6 Retrospective evaluation of SQUID fuzzy pharmacophore models 108
4.6.1 Pharmacophore model of COX-2 ligands 110
4.6.2 Retrospective screening for COX-2 inhibitors 112
4.6.3 Pharmacophore model of thrombin ligands 117
4.6.4 Retrospective screening for thrombin inhibitors 119
4.6.5 Method performance 121
4.6.6 Conclusion 124

4.7 Prospective screening for inhibitors of the Tat-TAR RNA interaction with a SQUID fuzzy pharmacophore model and CATS3D 126
4.7.1 Calculation of an alignment of reference compounds 128
4.7.2 Calculation of pharmacophores and virtual screening 130
4.7.3 FRET determination of the inhibition constants 131

4.7.4 Conclusions 133

4.8 Prospective screening for Taspase1 inhibitors with a receptor-derived pharmacophore model 135
4.8.1 Conclusions 144

5 SUMMARY 146

5.1 Summary 146

5.2 Zusammenfassung 151

6 APPENDIX 157

6.1 Enrichment factors of activity classes from Section 4.3 157

6.2 Protein report from MOE for the taspase1 homology model from Section 4.8 161

7 REFERENCES 163

1 Introduction

1.1 Scope of the thesis

Rationalization of the drug discovery process is crucial to be prepared for future challenges in human health care [Tollman *et al.*, 2001]. Technological developments like combinatorial synthesis and high-throughput screening (HTS) had a large impact on the drug discovery process: hundreds of thousands up to millions of molecules can be tested today for a single target [Bajorath, 2002]. Despite this large increase in assay capacities such techniques have not led to an increased number of approved new chemical entities per year [Xu & Agrafiotis, 2002]. One reason for this failure might be grounded in the focus on large numbers of tested molecules instead of high quality experiments, i.e. testing the right molecules. Computer based methods might provide a means to rationalize these experiments incorporating the challenges provided by the high-throughput experiments [Agrafiotis *et al.*, 2002; Bleicher *et al.*, 2003; Bajorath, 2002].

Computational methods for the compilation of molecule-libraries for pharmacological screening are called virtual screening methods [Böhme & Schneider, 2000]. Using such methods one can restrict pharmacological screening to molecules with a high probability of being active instead of testing all molecules accessible. Within the scope of this thesis virtual

screening methods were developed, evaluated and applied with the aim to contribute to the rationalization of the drug discovery process.

Virtual screening can either be applied by knowledge of the receptor structure or of active ligands [Böhm & Schneider, 2000]. The focus of this work was on ligand-based virtual screening methods for "scaffold hopping" [Schneider *et al.*, 1999]: the ability to retrieve molecules that have a different topology compared to known active molecules. In other words we were interested in methods that were able to retrieve the non-obvious hits from the vast chemical space.

The CATS method is such an approach based on an alignment-free topological pharmacophore pair description of molecules [Schneider *et al.*, 1999]. Molecules with similar CATS descriptors are likely to evoke similar biological responses. Since the binding of a ligand to a receptor is a three-dimensional interaction, a three-dimensional extension of such descriptors is an attractive approach and might improve the ability of the descriptor to find isofunctional molecules.

The first goal of this thesis was to develop and evaluate novel alignment-free pharmacophore pair based descriptors for virtual screening, based on the three-dimensional conformation of a molecule. Therefore the CATS approach was extended to a three-dimensional pharmacophore pair descriptor (CATS3D) and a molecular surface-based descriptor (SURFCATS). These methods were evaluated and optimized by the following retrospective screening experiments:

- Comparison of different similarity metrics and scaling methods
- Dependence on the correct "receptor bound" conformation
- Comparison of the enrichment performance and "scaffold hopping" capability with CATS and MACCS substructure keys
- Combination of CATS3D with artificial neural networks

The second goal of the thesis was to develop and evaluate a three-dimensional "fuzzy" pharmacophore model method for virtual screening. The fuzzy description of molecules should result in a more general pharmacophore representation which might be favorable to retrieve isofunctional molecules with new scaffolds. The resulting approach was compared to existing virtual screening methods.

The last goal was to apply the developed virtual screening methods prospectively to retrieve novel inhibitors for the TAR RNA, the metabotropic glutamate receptor 5 and taspase1. In quest of this goal the methods developed in this thesis were employed for a prospective evaluation. Taspase1 could not serve as test case for the ligand-based methods

since no inhibitors were known. For this project a homology model derived pharmacophore hypothesis was used for virtual screening, complementing the methods developed in this thesis.

1.2 The drug discovery process

The drug discovery and -development process can be illustrated by a value chain (Figure 1.1) [Bleicher *et al.*, 2003]. The initial step in this process is to identify a target (mostly a protein) that is associated with a disease state under consideration and which can be modulated to alter this state. Having identified a target, first hits have to be found, i.e. molecules which possess a minimum biological activity. This can be achieved by high-throughput screening (HTS) of large libraries of molecules or by modification of endogenous or competitor's ligands available from literature or patents. The next step is the lead generation where the initial hits are refined into leads or lead series, variants of prototypical molecules with a unique core structure, showing high in-vitro activity, selectivity and initial structure activity relationships (SAR). Lead optimization includes further optimization of activity, selectivity and of ADME (absorption, distribution, metabolism and excretion) and toxicity properties to obtain molecules appropriate for the clinical trials.

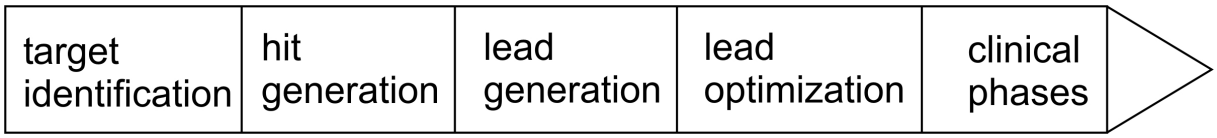


Figure 1.1 Drug discovery value chain.

1.3 Chemoinformatics in the drug discovery process

The name "chemoinformatics" was introduced in 1998 for computational methods used for improved decision making in the drug discovery process [Brown, 1998]. According to the book "Chemoinformatics" by Gillet and Leach [Leach & Gillet, 2003] chemoinformatics methods include the handling of chemical libraries, calculating the similarity and diversity of compounds, clustering, predictions of properties and structure activity relationships. From the viewpoint of drug design, computational techniques like docking [Kitchen *et al.*, 2004], homology-modeling [Hillisch *et al.*, 2004], molecular mechanics [Karplus & McCammon, 2002], quantum chemistry calculations [Clark, 2003] or sequence alignment [Durbin *et al.*,

1998] are of relevance, too. These latter techniques are mostly assigned to the fields of computational chemistry or bioinformatics.

Applications of computational approaches were reported for each step in the drug discovery process. Designing libraries for HTS can be rationalized by chemoinformatics methods [Schneider, 2002; Bajorath, 2002] incorporating "chemogenomics" strategies [Schuffenhauer *et al.*, 2003] or ADME and "drug-likeness" considerations [Lipinski *et al.*, 1997; Ajay *et al.*, 1998]. In lead optimization incorporation of computational models for quantitative structure activity relationships (QSAR) or the incorporation of the receptor structure facilitates the rational improvement of ligands [Kubinyi, 1993; Hansch *et al.*, 1995; Kitchen *et al.*, 2004].

1.4 Virtual screening

The number of chemically feasible molecules which could be in principle used as drug candidates has been estimated to be 10^{100} [Walters *et al.*, 1998], which is larger than the number of atoms in the universe. This number has two main consequences: first, it should be possible to find a ligand with appropriate characteristics for each biological macromolecule. Second, it is absolutely impossible to test all these ligands experimentally.

Virtual screening provides a means to enlarge the number of molecules which can be tested for some desired property by several orders of magnitude [Xu & Agrafiotis, 2002]. Even though computational prediction of properties will probably never replace biochemical measurements, it is much faster. In this way large amounts of molecules can be excluded prior to pharmacological experiments to avoid a waste of resources for molecules which have a high probability of not being active.

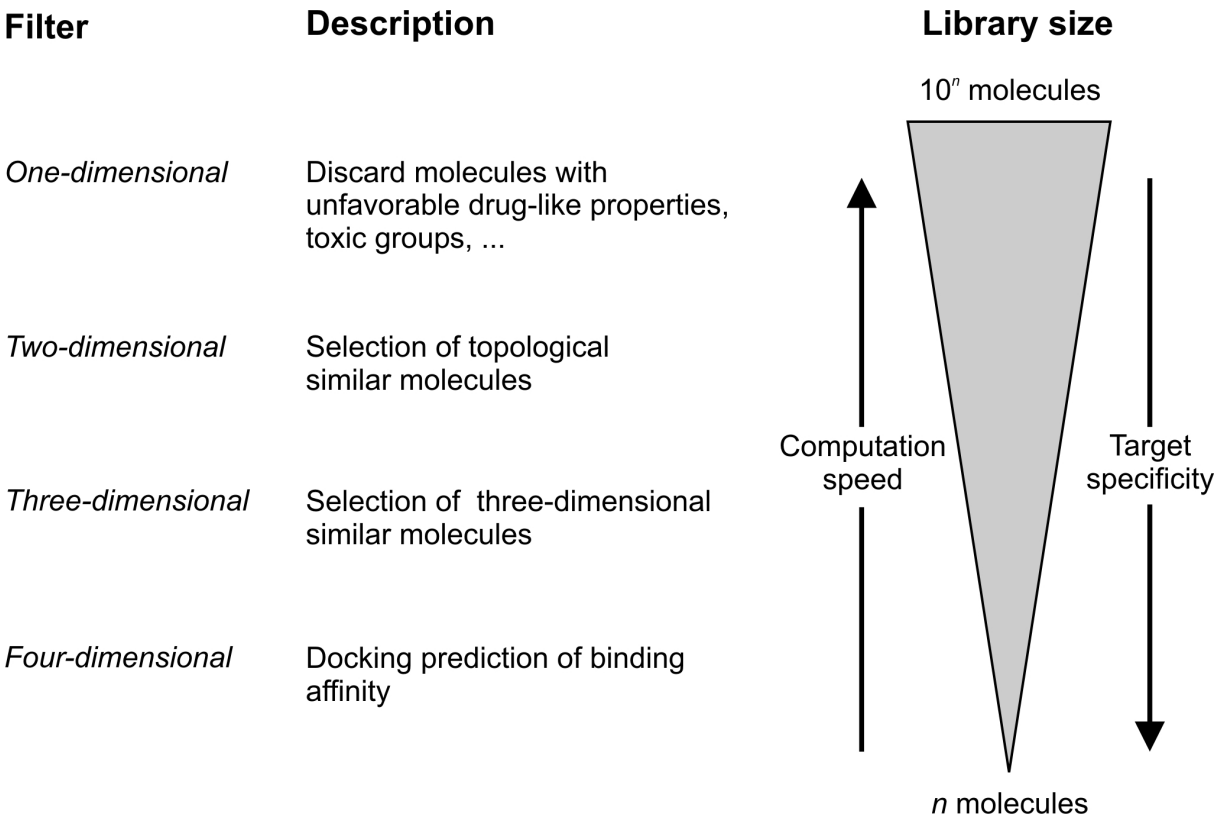


Figure 1.2 Hierarchical virtual screening. Virtual screening campaigns are often organized hierarchically. First simple and computationally fast filters are applied to remove undesired molecules. Subsequent methods are increasingly accurate, more problem specific and often computationally slower.

In virtual screening campaigns a hierarchical sequence of increasingly complex and specific methods is often applied [Bleicher *et al*., 2003; Böhm & Schneider, 2000] (Figure 1.2). Starting from a database of molecules (real or virtual molecules) the first step of a hierarchical virtual screening is to eliminate all molecules which have undesired properties. These properties could be reactive or toxic groups or a violation of the Lipinski "rule of five" [Lipinski *et al.*, 1997] which was suggested as a rule of thumb assessing the potential oral bioavailability of a molecule. Another possible approach is to predict the "drug-likeness" or the "lead-likeness" of molecules to consider only molecules which possess some general properties derived from the analysis of known drugs or lead molecules [Ayay *et al.*, 1998; Sadowski & Kubinyi, 1998; Byvatov *et al.*, 2003; Teague *et al.*, 1999]. Subsequently with smaller libraries increasingly more target specific and computational demanding approaches can be applied. These methods range from similarity searching [Willett *et al.*, 1998] based on topological or three-dimensional descriptions of molecules to three-dimensional

pharmacophore searching [Güner, 2000] and docking methods which also incorporate receptor information [Kitchen *et al.*, 2004].

1.5 Molecular similarity

Many virtual screening methods search for molecules which are similar to a reference molecule of known activity. This approach is called “similarity searching” [Willett *et al.*, 1998]. Given a suitable definition of similarity it has been demonstrated that similar molecules have a higher chance of exhibiting a similar biological activity than dissimilar molecules [Brown & Martin, 1996; Martin *et al.*, 2002]. A quality criterion for a similarity searching method is the “neighborhood behavior” [Patterson *et al.*, 1996]. A similarity measure satisfies the neighborhood behavior criterion if modifications of a molecule, which lead to small changes in the molecular descriptor result in small changes in the activity and modifications which lead to large changes in the descriptor result in larger changes in the activity.

No single method is best-suited for all targets and all small molecules. Molecular similarity is dependent on the context of the ligands chemotypes and the receptor [Schneider & So, 2003; Bender & Glen, 2004]. Different representations of molecules focus on different aspects of molecules and for different ligand-receptor complexes there are different interactions which are important for ligand-binding. Employing a variety of different descriptors increases the probability to have an appropriate molecular encoding suitable for a problem under consideration [Sheridan & Kearsley, 2002].

One difficulty for molecular similarity considerations is that the “fitness-landscape” of molecules in drug discovery projects is often found to be multimodal, i.e. there are multiple local optima found [Schneider & So, 2003]. The “fitness-landscape” is the relation of a molecular descriptor (the landscape) with a desired property (the fitness), which can be e.g. the binding affinity, selectivity or metabolic stability. An ideal “fitness-landscape” would be smooth with respect to the neighborhood of molecules. In such a “fitness-landscape” similar molecules would exhibit similar properties. “Fitness-landscapes” representing the QSAR of molecules in drug discovery projects are believed to be jagged [Maggiora *et al.*, 2004]. Maggiora compared ideal fitness-landscapes with the hills of Kansas and realistic fitness-landscapes with the Bryce Canyon (Figure 1.3) [Maggiora *et al.*, 2004].

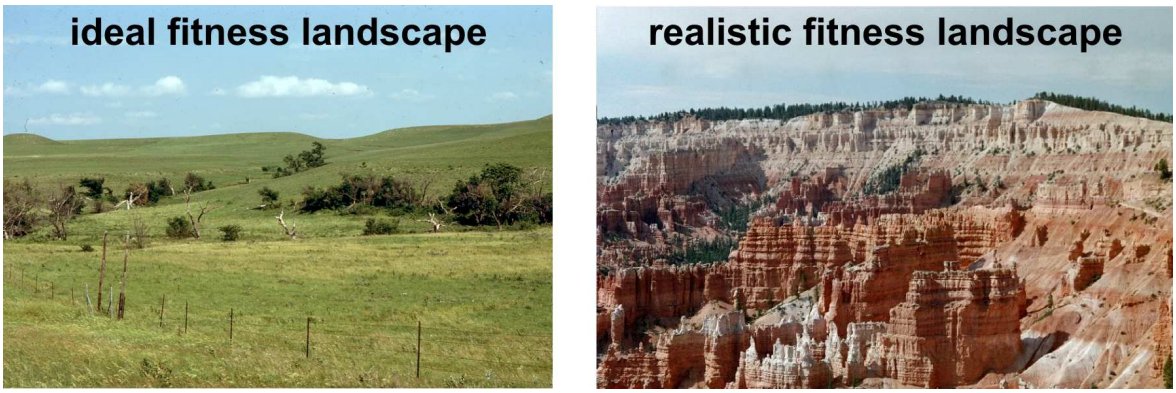


Figure 1.3 Fitness-landscapes in drug discovery projects. Ideal fitness-landscapes are smooth and show few local optima. This would support the rational optimization of molecules. Realistic fitness-landscapes are assumed to be jagged and filled with local optima, which can render the rational optimization of ligands impossible.

1.6 Scaffold hopping

A naïve approach for similarity searching is to compare the molecular connection tables to assess the similarity between two molecules. Such approaches were reported for searching the maximum common substructure between two molecules [Raymond & Willett, 2002]. If a structural element is known to be associated with activity, other molecules containing this substructure can be retrieved and tested for activity [Barnard, 1993]. A drawback of such methods is the lack of the ability to retrieve molecules with largely different topologies. This ability is called “Scaffold hopping” [Schneider *et al.*, 1999]. Two molecules are considered to have different scaffolds if they have different topologies [Böhm *et al.*, 2004]. This idea is based on the concept that drug-like molecules are built up from a scaffold (framework) and side-chains (Figure 1.4) [Bemis & Murcko, 1996].

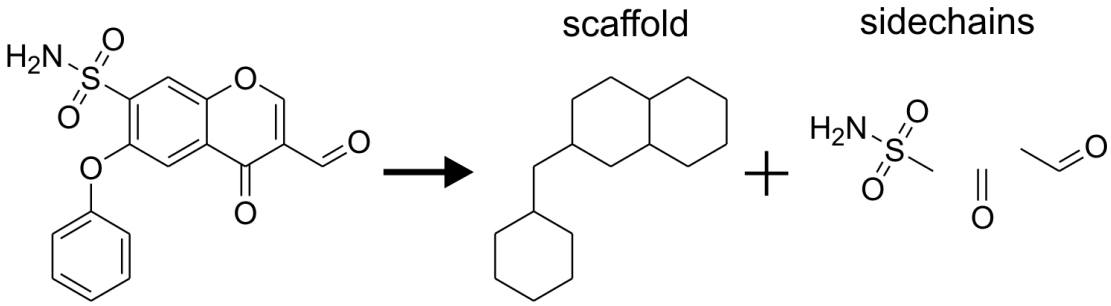


Figure 1.4. The atoms of a molecule can be separated into scaffold and sidechains. The scaffold determines the pharmacological properties, which can be obtained by variation of the sidechains.

Scaffold hopping is one of the most challenging goals in virtual screening. Ideal virtual screening methods would not only find a maximum number but also a maximum diverse set of active compounds from a given chemical subspace. There are several reasons for seeking a set of diverse structures. Diverse structures offer the medicinal chemist a choice in terms of chemical accessibility and prospects for lead optimization. Multiple leads ("backup" leads) lower the chance of drug development attrition in case of undesirable ADMET properties [Jenkins *et al.*, 2004]. Furthermore, the creation of intellectual property is facilitated.

Different virtual screening concepts have been proposed for scaffold-hopping [Böhm *et al.*, 2004]. These include three-dimensional pharmacophore models [Good & Mason, 1996; DeEsch *et al.*, 2001], pseudoreceptors [Lloyd *et al.*, 2004], protein structure-based *de novo* design [Schneider & Fechner, 2005; Stahl *et al.*, 2002], and ligand-based similarity searching [Willett *et al.*, 1998]. In contrast to the former methods similarity searching is based on the comparison of descriptor vectors rather than on the alignment of molecules to a reference and can thus be applied efficiently for large datasets [Willett *et al.*, 1998].

From the viewpoint of a "fitness-landscape" the scaffold defines the region of the landscape that is accessible using different sidechains. Different scaffolds might have some overlapping regions in the "fitness-landscape", but also some regions which might not be accessible by other scaffolds. This behavior is especially attractive if multiobjective "fitness-landscapes" are considered [Gillett *et al.*, 2002]. Drugs have to satisfy many objectives like tight binding, selectivity or acceptable ADMET properties. Different scaffolds provide a higher chance of finding molecules that can access acceptable regions in the "fitness-landscape" for all these objectives.

1.7 The pharmacophore concept

It has long been recognized that some fragments of chemical molecules can be mutually exchanged without much affecting the biological activity. Such fragments are called bioisosteric groups [Patani & LaVoie, 1996]. Bioisosteric groups mediate identical or similar interactions with the receptor.

Ligand receptor interactions can be clustered into three general groups: "hydrophobic", "polar positive" and "polar negative" [Horvath *et al.*, 2004]. These groups can be further broken down into "hydrophobic-alkyl", "aromatic", "hydrogen-bond donor",

“cation”, “hydrogen-bond acceptor” and “anion”. Various definitions and combinations of these groups have been reported as pharmacophore atom-types in literature [Güner, 2000; Bush & Sheridan, 1993; Pickett, 2003]. An object with an associated pharmacophoric type is called a potential pharmacophore point (PPP). PPPs can represent atoms or larger fragments of a molecule.

Based on the spatial arrangement of PPPs of a ligand, a pharmacophore hypothesis can be derived. According to the IUPAC definition 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]. Accordingly to define a pharmacophore, prior knowledge about the importance of the PPPs of a molecule is needed. A pharmacophore can be derived from structure-activity data or from conserved features within a set of ligands. Receptor-based pharmacophores have also been reported [Pickett, 2003].

The most widely used application of pharmacophores is to search for new molecules comprising the pharmacophore. These molecules are expected to have a similar biological effect. If the relevant pharmacophore pattern is not known, one can also utilize the distribution of PPPs of molecules for similarity searching: Molecules that have similar distributions of PPPs are likely to have a similar activity (neighborhood principle).

The description of a ligand-receptor interaction by pharmacophores is a crude simplification which does not consider effects like entropy or solvatation. Also some groups like fluorine, which can interact like hydrophobic groups and as hydrogen-bond acceptors [Böhm *et al.*, 2004] are hard to model correctly by pharmacophore types.

1.8 Representation of molecules

The way the structure of a molecule is encoded has a major influence on the way how molecules can be compared. Molecules can be represented either by a full connection table or by sets of substructures that are present or absent in the molecule. The first representation is more detailed, but to establish a similarity calculation, molecules have to be aligned or a maximum common sub-graph between to molecules has to be calculated [Labute *et al.*, 2001; Willett *et al.*, 1998]. This can be a time-consuming procedure, especially for the three-dimensional alignment of flexible molecules. The comparison of the presence and absence of substructures can be computed more efficient. Such methods are called “alignment-free”. Substructure similarity can be calculated on the basis of predefined substructure dictionaries

(e.g. the MACCS keys [MDL Information Systems]) or on the basis of molecule-specific generated substructures (e.g. Daylight fingerprints [Daylight Chemical Information Systems]). Substructures can be defined as exact chemical fragments (e.g. MACCS), or loosely defined like pairs, triplets or quartets of PPPs [Schneider *et al.*, 1999; Good & Kuntz, 1995; Mason *et al.*, 1999]. Substructure descriptions like the MACCS keys contain only local descriptions neglecting the overall topology of the molecule.

Both kinds of molecular encodings are found combined with two-dimensional and three-dimensional representation of molecules. Two-dimensional topological representations of molecules have the advantage that the time-consuming calculation of three-dimensional conformations for the molecules can be avoided. The stereochemistry of molecules can also be left aside. On the other hand the binding event is a three-dimensional interaction between ligand and receptor. So it should be advantageous to include such information. The naïve assumption about the three-dimensional conformation of a molecule in the binding-pocket would be that the conformation of the molecule with the lowest internal energy would be the most likely to be found in the receptor. However it has been shown that the "bioactive" conformation, i.e. the conformation of a molecule bound to the receptor, does not necessarily correspond either to the global torsion-angle energy minimum or to a torsion-angle energy minimum at all [Nicklaus *et al.*, 1995; Boström *et al.*, 1998; Perola & Charifson, 2004]. In practice, this renders the task of finding the "bioactive" conformation of a molecule to the computational demanding task of presenting a large number of low-energy conformations. While it is clear that methods, which are based on the explicit three-dimensional alignment of molecules, strictly rely on the presence of a fitting conformation, alignment-free descriptors have produced reasonable results using only a small set of conformations or even a single conformation [Sheridan *et al.*, 1996; Brown & Martin, 1996].

A step further away from the atomic representation of molecules is the description of molecules based on their molecular surface. Since the interaction between ligand and receptor is mediated by the molecular surfaces, surface-based descriptions are thought to be more general than atom based descriptions [Wagener *et al.*, 1995; Zamora *et al.*, 2003, Stiefl & Baumann, 2003; Clark, 2004]. Field-based methods are another way to circumvent an atom-based description of a molecule [Cramer *et al.*, 1988; Klebe *et al.*; 1994; Pastor *et al.*, 2000].

1.9 Autocorrelation descriptors

Spatial autocorrelation is a quantitative measure for the probability to find objects of defined properties within a distance of interest [Wagener *et al.*, 1995; Todeschini & Consonni, 2000]. The idea of a molecular descriptor based on the autocorrelation concept was first introduced into the field of cheminformatics by Morau and Broto in 1980 [Moreau & Broto, 1980] with the ATS (*A*utocorrelation of a *T*opological *S*tructure) descriptor. For this approach the atoms of a molecule were represented by properties like atomic mass or partial charge. The distance between atoms was measured as the number of bonds between the respective atoms (topological distance).

The ATS descriptor for a given topological distance *d* is calculated by:

$$ATS_d = \sum_{i=1}^{A}\sum_{i=1}^{A}\delta_{ij,d}(w_i w_j), \qquad \text{(Eq. 1.1)}$$

where *w* is the atomic property, *A* is the number of atoms in the molecule, $\delta_{ij,d}$ (Kronecker delta) evaluates to 1 for all pairs of atoms with distance *d.*

To obtain the full descriptor the ATS autocorrelation is calculated over all defined distances and concatenated to a vector {ATS_0, ATS_1, ATS_2, … , ATS_D}, where *D* is the maximum distance considered. Moreau, Broto and Vandycke were also the first who applied this approach to the three-dimensional conformation of a molecule [Moreau *et al.*, 1984]. For the three-dimensional approach the topological distance was replaced by the spatial Euclidean distance between two atoms. Pairs of atoms were clustered into groups with distances falling into predefined distance ranges (bins). All atom pairs within one bin were treated as having the same distance. Gasteiger extended this approach to the spatial autocorrelation of the partial charges calculated for surface points [Wagener *et al.*, 1995]. The resulting vector values were normalized by dividing the raw counts by the number of atom pairs in each distance range.

In 2000 Pastor and coworkers [Pastor *et al.*, 2000] presented GRIND (*Gr*id-*In*dependent *D*escriptors), an approach very similar to the autocorrelation descriptors. The GRIND descriptor is calculated from force field-based interaction energies calculated for GRID [Goodford, 1985] points surrounding a molecule. Instead of summing up all products of interaction energies for pairs of GRID points within a distance range, only the most favorable energy contribution is stored for each distance range. Given a descriptor vector, pairs of grid points can be identified that are responsible for each descriptor value. Such a

trace back from the descriptor to the underlying pairs of grid points is not amenable to other autocorrelation approaches.

In 1985 Carhart [Carhart *et al.*, 1985] introduced a topological atom pair descriptor using atom-types instead of atom property values. Each atom is assigned to one atom-type class instead of an atom property value. Atom-types are defined by their element, the number of neighboring non-hydrogen atoms and their number of π-electrons. The employment of these atom-types led to a further distinction of chemical elements according to the atom environment. Binary values are assigned to each atom, i.e. an atom does or does not have a specific atom-type. Consequently and in contrast to the Moreau-Broto approach, the resulting autocorrelation vector for an atom-type is equivalent to a histogram counting the frequencies of the atom pairs of the considered atom-type over the different atom-atom distances. Calculation of the autocorrelation between pairs of atoms of different atom-types is called crosscorrelation. The final Carhart descriptor vector consists of the autocorrelation vectors for all atom-types and the crosscorrelation vectors of all pairs of different atom-types.

In 1996 Sheridan and coworkers [Sheridan *et al.*, 1996] were the first to use pharmacophoric atom-types for an autocorrelation approach. This technique provides a description presumably most relevant to characterize ligand-receptor interactions in a general way, allowing for more different but equally interacting molecules to be identified as similar. In this work Sheridan and coworkers also extended the topological Carhart approach to the three-dimensional conformation of molecules. This approach was soon followed up by a binary representation of such a descriptor [Brown & Martin, 1996]. In 2003 Stiefl and Baumann [Stief & Baumann, 2003] reported an autocorrelation approach using surface points representing pharmacophoric features.

The work of Schneider and coworkers [Schneider *et al.*, 1999] first focused on the applicability of the autocorrelation descriptors, in this case topological pharmacophores, for scaffold hopping. The general description of the atoms with pharmacophore atom-types in combination with the decomposition of molecules into atom-pairs was shown to be especially successful to find new molecules with significant different molecular scaffold, maintaining the desired biological effect.

1.10 Retrospective and prospective screening

The effectiveness of a virtual screening method can be assessed in two ways: retrospective and prospective screening. Given a reference molecule with known biological effect,

retrospective screening quantifies the ability of a method to retrieve molecules with the same biological activity from a database containing molecules with various biological activities [Willett *et al.*, 1998; Hert *et al.*, 2004a; Hert *et al.*, 2004b; Xu & Agrafiotis, 2002]. Several or all molecules of selected classes of biological activities are mutually taken as reference for the screening. For each individual virtual screening experiment, the molecules remaining in the database are ranked according to the similarity or distance to the reference molecule.

A method is considered successful if molecules with the same annotated activity (the "active" molecules) as the reference are statistically better scored than molecules with different annotated activities ("inactive" molecules). A shortcoming of retrospective screening is that it is mostly not known if molecules which are considered inactive for one receptor are true inactives or molecules for which the respective activity has not been tested experimentally. It is likely that the latter situation represents the majority of cases.

The most rigorous test for a virtual screening method is prospective screening. Only in this way it is possible test the ability of a method to find novel active molecules. On the other hand, prospective screening requires much more effort in time and costs and consequently in most cases only a smaller number of experiments can be performed, resulting in a less reliable statistical assessment of the results. In the worst case this could lead to a poor rating of a method which was able to find similar molecules which were inactive due to small unfavorable interactions to the receptor, like a steric clash from a methyl group of the molecule. Consequently it is best to probe a method by both retrospective and prospective screening to obtain a realistic assessment of its performance.

1.11 Artificial neural networks in virtual screening – machine learning based on molecular representations

Artificial neural networks (ANN) had a large impact on recent drug discovery projects [Zupan & Gasteiger, 1999; Schneider, 2000; Terflot & Gasteiger, 2001; Livingstone & Manallack, 2003]. Applications of ANNs are found for classification, prediction, visualization, and clustering. One can distinguish between supervised methods like feedforward networks and unsupervised networks like self-organizing maps (SOM) [Kohonen, 1982]. Supervised methods establish a relationship between a representation of an object of interest (e.g. a molecular descriptor) and an observed response (e.g. a binding affinity or a class affiliation). Unsupervised methods cluster the data based on their representation. One particular implementation (SOM) projects a data distribution from a high-dimensional space (i.e. the

molecular representation) to a lower dimensional space (e.g. two-dimensional for visualization) [Kohonen, 1982].

In chemoinformatics, supervised ANNs are mainly applied in the establishment of quantitative structure activity relationships (QSAR), quantitative structure property relationships (QSPR) or binary classification tasks [Schneider, 2000]. ANNs provide a means to establish in principle any linear or non-linear relationship between descriptor and observed data [Zupan & Gasteiger, 1999]. As a drawback, an ANN behaves like a black box: the modeled relationship between the input variables is difficult to extract [Livingstone & Manallack, 2003]. Applications of supervised neural networks range from general predictions like drug-likeness [Sadowski & Kubinyi, 1998; Ajay *et al.*, 1998] or the identification of frequent hitters [Roche *et al.*, 2002a] to more specific tasks like the prediction of binding to the hERG K^+ channel [Roche *et al.*, 2002b] or to cytochrome P450 [Molnar & Keseru, 2002].

Unsupervised SOMs can be used for the projection of data into lower dimensional space for visualization. This can be utilized for example to evaluate different descriptor-representation of molecules for their suitability to distinguish between different classes of activities [Teckentrup *et al.*, 2004]. Comparison of diversity and coverage of chemical space of chemical databases or combinatorial libraries were also reported for SOMs [Schneider & Schneider, 2003; Anzali *et al.*, 1998]. A trained SOM can also be used for the prediction of class affiliation for new molecules [Schneider *et al.*, 2003; Teckentrup *et al.*, 2004].

1.12 Incorporating receptor structure information into virtual screening

The binding-event is an interaction between ligand and receptor. Using the receptor for virtual screening should enhance the capability for scaffold hopping in comparison to ligand-based methods [Xu & Agrafiotis, 2002; Böhm *et al.*, 2004]. The latter methods are intrinsically biased towards the chemotypes of the reference molecules. Receptor-based ranking of molecules is independent of reference molecules. Structure-based approaches provide a rational basis for the establishment of new interactions between ligand and receptor, not realized in known ligands before. Using ligand-based approaches new interactions can only be found by trial and error. Structure-based virtual screening can suffer from the difficulty in scoring ligand-receptor complexes correctly [Halperin *et al.*, 2002; Schneider & Böhm, 2002], the flexibility of the receptor upon ligand binding [Teague, 2003] and from inaccuracies in protein structural models [Davis & Teague, 2003].

Receptor information can also be exploited for the derivation of pharmacophore models [Pickett, 2003]:

- the alignment of ligands can be calculated on the basis of the receptor [Grüneberg, 2005].
- promising potential pharmacophore points derived from the receptor can be incorporated into or solely used for a pharmacophore model [Wolber & Langer, 2005, Pirard *et al.*, 2005].
- receptor information can be used to disregard molecules, that were regarded as active by other methods, that overlap with receptor atoms [Pickett, 2003].

Following this idea, multiple receptor conformations obtained from molecular dynamics simulations were used to establish a receptor-based dynamic pharmacophore model, which was successfully applied for the prediction of new HIV-1 integrase inhibitors [Carlson *et al.*, 2000].

If no receptor information is available, homology modeling of the receptor structure provides an approach for virtual screening [Hillisch *et al.*, 2004; Bissantz *et al.*, 2003; Grüneberg, 2005; Evers *et al.*, 2003; Evers & Klabunde, 2005]. Homology modeling is based on the fact that the sequence of proteins is less conserved than the structure [Chothia & Lesk, 1986, Andreeva *et al.*, 2004]. Consequently the structure of a protein can be predicted based on the structure of a closely related protein. The quality of the resulting model critically depends on the sequence similarity of the modeled protein to the template structure [Hillisch *et al.*, 2004].

1.13 The metabotropic glutamate receptor 5 (mGluR5)

Glutamate is the major excitatory neurotransmitter in the mammalian central nervous system [Conn & Pin, 1997]. The effect of glutamate is mediated by ionotropic and metabotropic glutamate receptors, via pre- and postsynaptic mechanisms. The long term modulating effect of glutamate is mediated by the metabotropic glutamate receptors [Conn & Pin, 1997]. The family of metabotropic glutamate receptors comprises a set of at least eight subtypes. These can be further clustered into three groups on the basis of sequence similarity, pharmacology and the respective signal transduction mechanism. Group I (mGluR1 and -5) are coupled to the activation of phospholipase C, group II (mGluR2 and -3) and group III (mGluR4, -6, -7, and -8) are negatively coupled to cAMP production [Hermans & Challiss, 2001].

The mGluRs belong to family 3 of the G-protein-coupled receptors (GPCRs) [Bockaert & Pin, 1999; Fredriksson *et al.*, 2003]. Other members of family 3 are the $GABA_B$, Ca^{2+}-sensing, vomeronasal, pheromone and putative taste receptors [Pin *et al.*, 2003]. GPCRs are characterized by a general topology of seven transmembrane helices. Class 3 GPCRs differ from the other GPCR classes by the presence of an additional N-terminal extracellular ligand binding domain, the venus-flytrap module (VFTM), connected to the heptahelical domain (HD) via a cystein rich region (Figure 1.5). Other classes of GPCRs contain ligand binding regions directly within the seven-transmembrane domain. Family 3 GPCRs are found as homodimers or heterodimers [Pin *et al.*, 2003]. Receptor dimerization does also include dimerization of the venus-flytrap modules [Kunishima *et al.*, 2000].

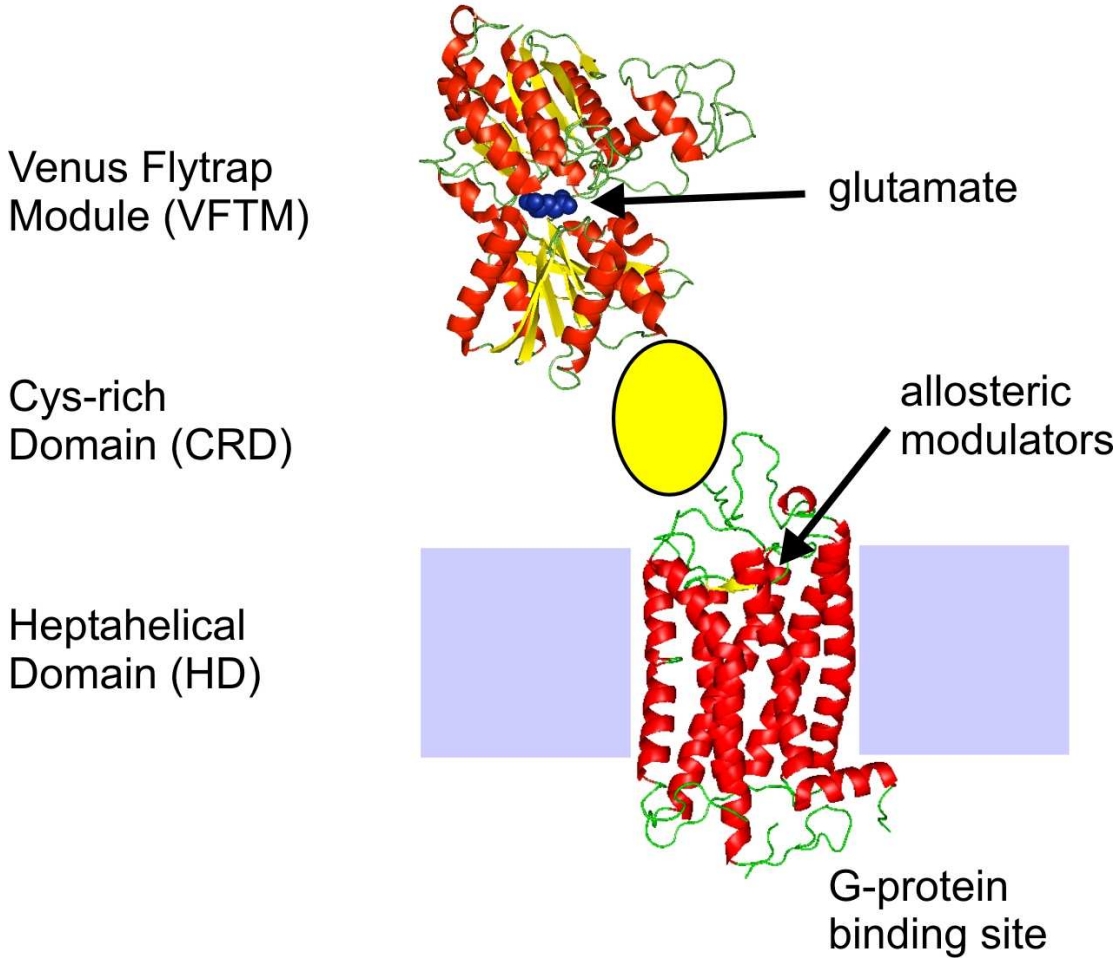


Figure 1.5 Topology of the metabotropic glutamate receptors. The receptor consist of an extracellular venus-flytrap module (PDB code: 1etw) for glutamate binding, and a heptahelical transmembrane domain (PDB code from the bovine rhodopsin structure: 1f88), which are connected by a Cystein-rich domain. Allosteric modulators bind in the transmembrane domain.

The activation mechanism of mGluRs involves several steps. Upon glutamate binding the venus-flytrap domain undergoes a large conformational change resulting in a closure of the venus-flytrap [Kunishima et al., 2000]. This leads to a modified relative orientation of the two VFTMs of a dimer. The activated dimer complex of the two VFTMs is assumed to stabilize the active conformation of the heptahelical domains [Pin *et al.*, 2004].

Like with many GPCRs, a basal "constitutive activity" can be measured without ligand binding [Pin *et al.*, 2004]. Agonists like the natural substrate glutamate stabilize the active state via the VFTM. The "constitutive activity" is not inhibited by competitive antagonists that prevent the VFTM closure [Prezeau *et al.*, 1996]. Consequently neither the active nor the inactive state of the heptahelical domain is predominantly stable.

Recently molecules were reported that bind to an allosteric binding site in the heptahelical domain of mGluR5 [Gasparini *et al.*, 1999; O'Brien *et al.*, 2003], i.e. in the region where the ligand binding site is found in the other classes of GPCRs. These molecules are called "allosteric modulators". Allosteric antagonists which stabilize the inactive state of the heptahelical domain are called "inverse agonists". Such modulators were able to completely inhibit the "constitutive activity" and the effect of agonist binding in the VFTM [Gasparini *et al.*, 1999]. Allosteric modulators which are able to stabilize the active state of the receptor are "positive allosteric modulators". These molecules cannot activate the receptor by themselves, but have a potentiating effect on agonist binding. This effect is assumed to be caused in a stabilizing effect of the opened VFTM on the inactive state of the heptahelical domain [Pin *et al.*, 2004]. If the VFTM is pruned from the heptahelical domain of mGluR5, positive allosteric modulators behave as conventional agonists while inverse agonists shown antagonistic behavior [Goudet *et al.*, 2004].

Group I mGluRs provide a great prospect for pharmaceutical applications. Molecules antagonizing the function of mGluR5 have a potential in prevention of pain and anxiety, and in the treatment of Parkinson's disease [Spooren & Gasparini, 2001; Swanson *et al.*, 2005]. A potential role in the treatment of drug dependence has also been reported [Chiamulera *et al.*, 2001]. Activators or potentiators of group I mGluRs were proposed to be useful in the therapy of schizophrenia and Alzheimer's disease [Pin *et al.*, 2004].

For GPCRs only the crystal structure of rhodopsin in the inactive state has been resolved, so far [Palczewski *et al.*, 2000]. Homology models based on this structure were shown to provide a basis for structure-based virtual screening for GPCR ligands [Evers & Klebe, 2004; Evers & Klabunde, 2005]. Successful applications of homology-model based

virtual screening for family 3 GPCRs have not been reported until now, despite the fact that such models have been published [Pagano et al., 2000; Malherbe *et al.*, 2003]. Many allosteric modulators of mGluR1 and mGluR5 were reported in literature. This also renders mGluR5 a target for ligand-based virtual screening and library design.

1.14 RNA drug design and the Tat-TAR RNA interaction

In recent years it has become clear that RNA is an active multifunctional player of the cell instead of just a passive vehicle for sequence information [Special journal issues on RNA as drug target]. RNA was found to have enzymatic functionality, e.g. the self-splicing intron of the Tetrahymena pre-rRNA [Kruger *et al.*, 1982] or within the ribosome [Noller *et al.*, 1992]. Gene regulatory elements on the mRNA can have an effect on the transcriptional and on translational level. This effect can be mediated by specific RNA-protein interactions and directly by RNA-small molecule interactions [Mandal & Breaker, 2004]. Such small molecules can be the metabolites of the genes under control, e.g. TPP (thiamine pyrophosphate) [Winkler *et al.*, 2002a] or FMN (flavin mononucleotide) [Winkler *et al.*, 2002b]. With the RNAi mechanism, RNA was also found to participate in anti-infective responses [Dykxhoorn *et al.*, 2003].

Together with these functionalities it was found that RNA can fold into complex and well defined three-dimensional structures. Like within proteins, these complex structures provide interfaces for specific intermolecular protein-RNA and small molecule-RNA interactions. These findings have led to a constantly increasing interest in RNA as a potential drug target with a plethora of potential applications [Zaman et al., 2003; Drysdale *et al.*, 2002; Gallego & Varani, 2001; Sucheck & Wong, 2000], and several natural and synthetic small molecules have been reported to interact specifically with RNA [Hermann, 2003].

In principle it is possible to employ the same approaches for RNA drug discovery as for molecules targeting proteins [Hermann, 2000]. One difference can be found in the relative importance of ligand-protein interactions and ligand-RNA interactions. The latter is biased towards electrostatic and stacking interactions in comparison to protein-ligand interactions [Hermann, 2000]. This might raise complications with unspecific binding of small molecules that comprise a large number of positive charges. In addition, very polar or charged ligands bear the danger of low oral bioavailability [Lipinski *et al.*, 1997; Mayer & James, 2004]. Other problems might arise from unspecific stacking of molecules which can lead to toxic effects from DNA-intercalation [Snyder *et al.*, 2004]. Another difference between protein and

RNA targets is the comparably high flexibility of RNA, especially of structures with low structural complexity like stem-loop RNAs [Schroeder *et al.*, 2004]. For these structures it has been found that different ligands result in ligand-receptor complexes with largely different conformations of the RNA structure. A recent publication even reported an RNA sequence which was able to fold into two completely different tertiary structures with two different enzymatic activities [Schultes & Bartel, 2000].

Beside these complications a structure-based automated docking approach including a scoring function optimized for RNA was shown to be useful for finding small and enriched sets of molecules inhibiting the Tat-TAR interaction [Filikov *et al.*, 2000; Lind *et al.*, 2002]. Other studies indicated that the inherent flexibility of RNA structures might limit the applicability of entirely structure-based approaches [Williamson, 2000; Gallego & Varani, 2001; Leulliot & Varani, 2001].

One of the best characterized RNA-based regulatory systems is the transactivation response element (TAR) of the HIV mRNA [Karn, 1999]. Specific binding of the Tat protein to TAR is essential for virus transcription. Without bound Tat protein the elongation of the HIV transcript is early aborted due to a poorly processive RNA polymerase II. Bound Tat recruits a Tat-associated kinase which activates the RNA polymerase. The activated polymerase is able to synthesize the remainder of the HIV transcript [Karn, 1999].

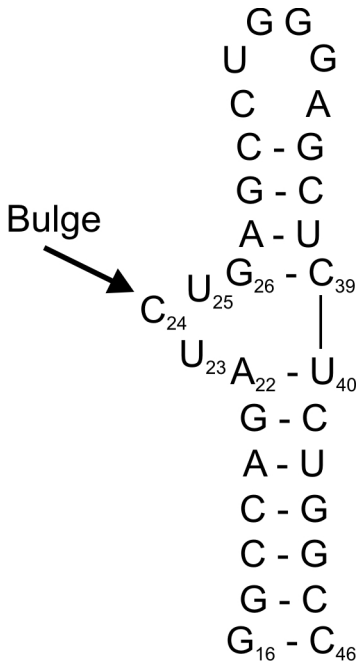


Figure 1.6 TAR RNA regulatory element. TAR RNA consists of two stem loop regions and a bulge of three nucleotides. The bulge is responsible for the specific interaction to the Tat protein, essential for HIV replication.

The TAR RNA represents a potential target for defeating HIV as well as a model system to deepen the understanding of RNA-small molecule interactions and the development of drugs for RNA targets in general. The structure of TAR consists of two rigid double strand stems connected by a flexible bulge of three bases, which provides a specific binding pocket for the Tat protein [Karn, 1999] (Figure 1.6). A variety of molecules have been found that inhibit the Tat-TAR interaction and consequently virus replication [Froeyen & Herdewijn, 2002; Krebs *et al.*, 2003]. Among these molecules are peptidic derivatives of the binding motive from Tat like argininamide, antibiotics like neomycin, and a set of small molecules with non-natural scaffolds. Most classes of bulge-binding ligands, for which structures have been determined, bind in distinct regions and stabilize different conformations of the bulge [Du *et al.*, 2002].

The availability of a small set of RNA-ligand complex NMR structures renders the TAR RNA as an interesting target for ligand and structure based virtual screening. The reported problems in RNA docking make the ligand based approach more attractive at present.

1.15 Taspase1

Taspase1 is a threonine aspartase which catalyzes the proteolytic cleavage of MLL (Mixed-Linage Leukemia) protein, resulting in its activation [Hsieh *et al.*, 2003]. MLL is required for the maintenance of *HOX* gene expression in embryogenesis and hematopoiesis. Chromosome translocations leading to chimera proteins of the N-terminus of MLL and varying translocation partners result in human infant leukemia. This effect is associated with an up-regulation of *HOX* genes. Specific inhibition of taspase1 might present a possibility to treat human infant leukemia [Hsieh *et al.*, 2003].

Taspase1 cleaves MLL directly after an aspartate at two positions with sequences D/GADD and D/GVDD, respectively. An N-terminal threonine acts as an active site nucleophile for the cleavage reaction. Other known threonine proteases are found in the 20S proteasome and the archaea proteasome and the catalytic subunit of the *Escherichia coli* (*E.coli*) HsIV [Hsieh *et al.*, 2003]. These proteases are not structurally related to taspase1.

Taspase1 reveals sequence similarity to glycosylasparaginase and L-asparaginase, which also have an N-terminal threonine involved in the reaction mechanism.

Glycosylasparaginases catalyze the cleavage of N-acetylglucosamine-asparagine to 1-amino-N-acetylglucosamine and aspartate. L-asparaginase catalyzes the conversion of L-asparagine to L-aspartate. All three classes of proteins are translated in an inactive form. Activation occurs by an autoproteolysis step catalyzed by the N-terminal threonine [Hsieh *et al.*, 2003].

Inhibitors have not been reported for taspase1, but crystal structures are available for glycosylasparaginases and L-asparaginases [Oinonen *et al.*, 1995; Prahl *et al.*, 2004]. This renders taspase1 a target for homology-model based drug design.

2 Computational Methods

2.1 Correlation-vector based descriptors

Three types of correlation vector descriptors were applied in this thesis, which all belong to the group of potential pharmacophore point (PPP) pair descriptors: the topological CATS descriptor [Schneider *et al*., 1999], the three-dimensional CATS3D descriptor and the surface-based SURFCATS descriptor (Figure 2.1). Auto- and crosscorrelation between all types of PPPs are transformed into a histogram, counting for the frequencies of the respective pairs of PPPs. The pairs of PPPs are further subdivided into distance "bins" which were topological distances in the two-dimensional case and distance ranges in the three-dimensional case. Each dimension ("bin") of the CATS3D CV was calculated according to Equation 2.1.

$$\mathrm{CV}_d^T = \sum_i \sum_j \delta_{ij,d}^T \,, \qquad \text{(Eq. 2.1)}$$

where *i* and *j* are atom indices, *d* is a distance or a distance range, *T* is the pair of PPP types of atoms *i* and *j*, and δ_d^T (Kronecker delta) evaluates to 1 for all pairs of atoms of type *T* within the distance range *d*.

2.1.1 CATS

The CATS (Chemically advanced template search) descriptor is a topological atom-pair descriptor developed by Schneider and coworkers [Schneider *et al.*, 1999]. The descriptor consists of the frequencies of pairs of PPPs within defined topological distances. Distances were calculated as the shortest paths between two PPPs. PPP-PPP distances were considered from 0 to 10 bonds.

The PPP definition was as follows: Hydrogen-bond donors were oxygen atoms of OH-groups and nitrogen atoms of NH- or NH_2-groups. Hydrogen-bond acceptors were oxygen atoms and nitrogen atoms not adjacent to a hydrogen atom. Positively charged or ionizable atoms were defined as atoms with a positive charge or nitrogen atoms of an NH_2-group.

Negatively charged or ionizable atoms were defined as atoms with a negative charge and carbon, sulfur or phosphorous atoms of a COOH-. SOOH-, or POOH-group. Lipophilic atoms were chlorine, bromine, or iodine, sulfur atoms adjacent to exactly two carbon atoms, and carbon atoms adjacent only to carbon atoms. With this definition atoms were assigned to no, one or two PPP-types. Using 10 topological distances "bins" for each of the 15 combinations of PPPs resulted in a descriptor of 150 dimensions.

The CATS descriptor was calculated with the program *speedcatsdotcom* (version 1.02) by Uli Fechner [Fechner *et al.*, 2003]. Scaling was done with the parameter –d 3, which corresponds to *scaling2* in CATS3D (see Section 2.1.2).

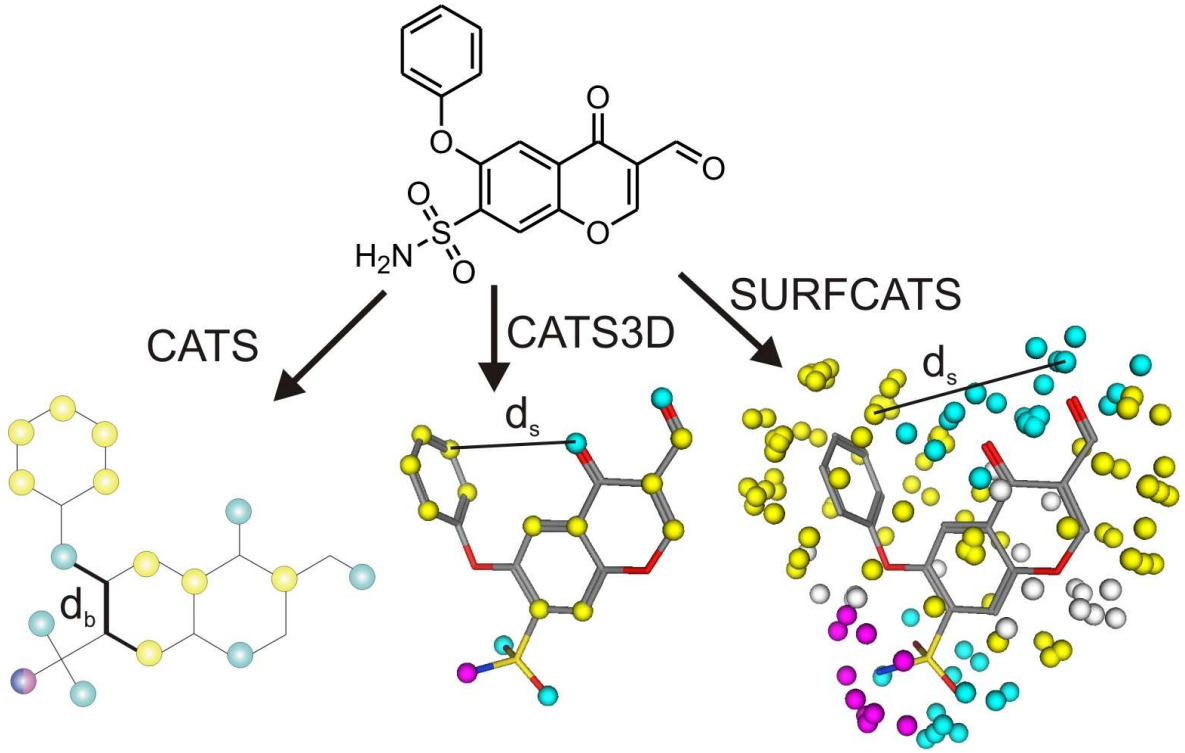


Figure 2.1 The CATS-family of descriptors: CATS, CATS3D and SURFCATS. All descriptors are based on a potential pharmacophore (PPP) type description of the underlying molecule. For each descriptor, pairs of PPPs are transformed into a correlation vector. CATS is calculated from the topological distances of atom-based PPP pairs. For CATS3D the spatial distances between atom-based PPPs are used instead. SURFACTS uses the spatial distances between PPPs on the contact surface of a molecule. Here the PPPs represent the atom-types of the nearest atom to each surface point. Yellow = hydrophobic PPP, cyan = hydrogen-bond acceptor, magenta = hydrogen-bond donor, blue = cation, white = no pharmacophore type assigned.

2.1.2 CATS3D

The CATS3D descriptor is an extension of the CATS descriptor into three-dimensional space. CATS3D was developed and implemented as part of this work. An overview over the CATS3D principle is shown in Figure 2.2.

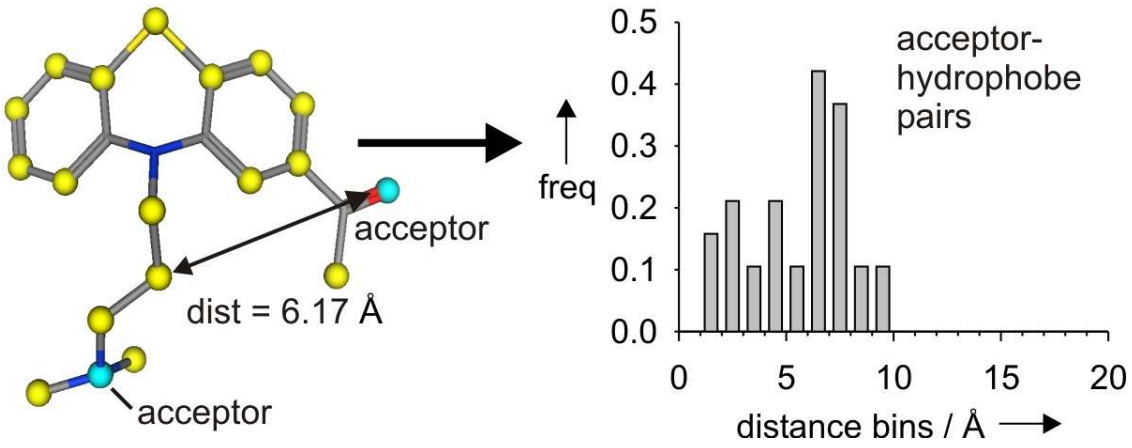


Figure 2.2 Calculation of the CATS3D correlation vector. Atoms are colored according to their pharmacophore atom-type (yellow = hydrophobe, cyan = hydrogen-bond acceptor). Distances are measured between all pairs of atoms, and frequencies of pairs are determined for all pairs of pharmacophoric types and for defined distance ranges ("bins"). As an example, a section of the resulting CV representing hydrogen-bond acceptor – hydrophobe pairs is shown.

The main difference in the correlation vector representation of a 3D conformation in comparison to a topological representation of a molecule is that the distances between the atoms are no longer shortest paths. Instead, Euclidean distances between all atoms were used. Distances between atoms are not restricted to integer values, thus the distances had to be partitioned into a set of distance bins. Several such binning schemes have been proposed [Wagener *et al.*, 1995; Sheridan *et al.*, 1996; Brown & Martin, 1996]. For CATS3D 20 distance bins that cover distances from 0 to 20 Å in steps of 1 Å were employed, i.e. if a pair of PPPs is found with a distance of 6.17 Å it is counted in the bin from 6 to 7 Å. Distances up to 20 Å were considered to include information of most pairs of atoms in the descriptor, even for large ligands.

For CATS3D the modified PATTY atom-types [Bush & Sheridan, 1993] available with the pH4_aType function in MOE (Molecular Operating Environment) [Chemical Computing Group] were used. This function provides six PPP types: cation (+), anion (-),

hydrogen-bond acceptor (A), hydrogen-bond donor (D), polar (P, hydrogen-bond acceptor AND hydrogen-bond donor) and hydrophobic (H). Whereas the topological CATS descriptor allows assignments of more than a single PPP type to one atom, the CATS3D descriptor employs a single PPP type per atom.

Using 20 distance bins for each of the 21 possible combinations of PPP pairs resulted in a descriptor of 420 dimensions. The values of the dimensional were calculated according to Eq. 2 with the difference that each pair of PPPs was only counted once and pairs of PPPs with themselves were not considered.

Three different scaling methods were implemented for the CATS3D descriptor:

- No scaling ("*unscaled*").
- Division by the number of PPPs of a molecule ("*scaling1*").
- Division of each of the 21 possible pairs of PPPs by the added occurrences of the two respective PPPs ("*scaling2*").

Scaling2 was always applied unless otherwise indicated. *Scaling1* is the scaling scheme originally developed for the CATS descriptor.

The CATS3D descriptor was implemented in the software *spacecats*. *Spacecats* was written in the SVL language in MOE [Chemical Computing Group].

Note to the program MOE: all calculations were performed with program versions 2003.02 and 2004.03. To our knowledge there were no differences between the two versions with an impact on the calculations in this thesis. Results obtained with earlier versions are not included in this work due to a major revision in the pharmacophore type definitions.

2.1.3 SURFCATS

The SURFCATS approach is a further extension of the CATS3D concept. The interaction between ligand and receptor is mediated by the surface between the two molecules. Accordingly it might be advantageous to describe molecules by their surface properties.

The surface points for the calculation of SURFCATS were taken from the molecular surface which was calculated with the Gauss-Connolly function in MOE with a spacing of 2 Å. The molecular surface is defined by the inward-facing part of a virtual probe sphere rolling on the van der Waals surface of the molecule [Richards, 1977]. Accordingly this surface

definition represents the contact space between ligand and receptor. Each surface point was assigned to the PPP type of the nearest atom. Like with CATS3D 20 equal distance bins were used from 0 to 20 Å with a stepsize of 1 Å. The SURFCATS CV was calculated exactly like the CATS3D CV except that surface points were used as PPPs instead of atoms. *Scaling2* was always applied.

The SURFCATS descriptor was implemented in the software *surfcats*. *Surfcats* was written in the SVL language of MOE [Chemical Computing Group].

2.2 Descriptor vector based virtual screening

For descriptor vector-based similarity searching, three distance indices were employed: the Manhattan distance, the Euclidean distance and the Tanimoto similarity coefficient. The first two metrics express distances, i.e. similar molecules have distances lower than dissimilar molecules. For similarity metrics this relation is inverted. To avoid confusion the term "similarity" will be used for both similarity and distance metrics. The definitions of the metrics are given in Table 2.1. Since all CATS derived descriptors contain non-binary data-values, the continuous version of the Tanimoto coefficient was applied. This version of the Tanimoto coefficient gives identical results for binary-data. A more detailed description of similarity metrics is given in [Willett *et al.*, 1998].

Table 2.1 Equations of similarity metrics for continuous variables. *A* and *B* are vectors (here: molecular descriptor representations), *N* is the total number of vector elements, x_i the value of the vector element *i*, $D_{A,B}$ denotes the distance and $S_{A,B}$ the similarity between objects *A* and *B*. Note that the range of the Tanimoto coefficient is 0 to 1 if all attributes of *A* and *B* are restricted to non-negative values.

Similarity metric	Equation	Range
Manhattan distance	$D_{A,B} = \sum_{i=1}^{N} \lvert x_{iA} - x_{iB} \rvert$	0 to ∞
Euclidean distance	$D_{A,B} = \sqrt{\sum_{i=1}^{N} (x_{iA} - x_{iB})^2}$	0 to ∞
Tanimoto coefficient	$S_{A,B} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA} x_{iB}}$	-0.333 to +1

Virtual screening was employed in two ways, using two programs: *rankIt* by Uli Fechner [Fechner *et al.*, 2003] and *SQUIDscreen*. The workflow of the two programs is illustrated in Figure 2.3. Both programs were designed to rank a database of molecular descriptors according to the similarity to a reference molecular descriptor, applying slightly different virtual screening protocols (Figure 2.3). The output of both programs is a ranked list for each reference molecular descriptor and the respective enrichment factor. *SQUIDscreen* is also able to handle multiple conformations of molecules in the virtual screening database. For this purpose each conformation of a molecule must be encoded separately. For the result *SQUIDscreen* selects the conformation with the best similarity score. Other conformations are discarded from the ranked result list.

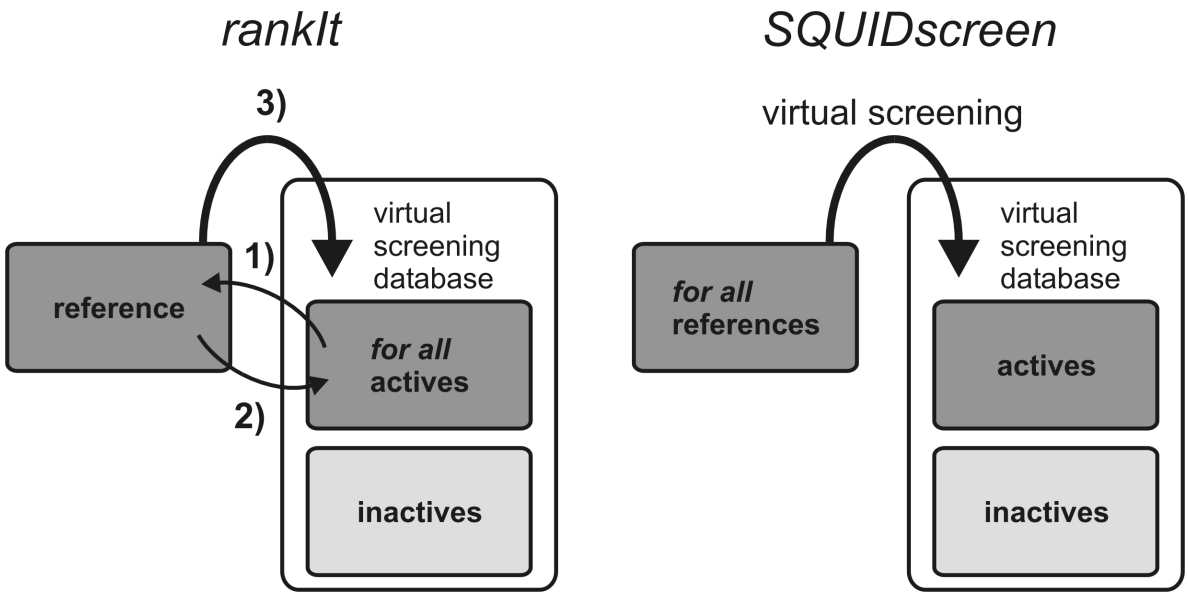


Figure 2.3 Virtual screening protocols of the programs *rankIt* and *SQUIDscreen*. *rankIt* iteratively takes reference molecules from the pool of actives (1) of the virtual screening database, performs virtual screening (2), and returns the reference back into the virtual screening database (3). This procedure is repeated for all active molecules. *SQUIDscreen* operates on distinct sets of reference molecules from the virtual screening database. In *SQUIDscreen* all reference molecules are iteratively submitted to virtual screening.

2.2.1 Retrospective screening evaluation

A quantitative measure for the evaluation of virtual screening results based on the obtained hit-lists is the enrichment factor *ef* [Xu & Agrafiotis, 2002]. This index quantifies the ability of a method to retrieve more active molecules than expected by random. The *ef* is defined as:

$$ef = \left(\frac{F_{act}}{F_{all}} \right) \Big/ \left(\frac{D_{act}}{D_{all}} \right) \qquad \text{(Eq. 2.2)}$$

where F_{act} and D_{act} are the numbers of annotated active molecules in a subset and the whole database, and F_{all} and D_{all} are the total numbers of molecules in the subset and the whole database respectively. An enrichment factor of 1 corresponds to a random distribution of active molecules in the ranked database, thus an effective pharmacophore model results in an *ef* above 1.

Subsets which were considered for the calculation of the *ef* were the first 1% and the first 5% of a ranked hit-list from virtual screening. The usage of a 5% subset of the hit-lists results in statistically more significant results. In real applications it is not always possible to test such large fraction of a database. This is especially important if only small numbers of active molecules are applied.

2.3 SQUID

SQUID Fuzzy Pharmacophore models approximate the spatial distribution of pharmacophoric points in an alignment of molecules by a set of generalized potential pharmacophore points (PPPs) of Gaussian probability densities. Atoms in the alignment comprising the same pharmacophoric features were clustered into PPPs for a more general and “fuzzy” representation of the major characteristics of the alignment. The resolution of the model was defined by the cluster radius, which is the parameter that affects how strict features are clustered into PPPs. The ideal resolution of the pharmacophore model had to be determined separately for each set of aligned ligands.

Each PPP in the pharmacophore model was represented by four attributes. The first attribute was the pharmacophore type of the atoms which are represented by the PPP, the second was the PPP position in 3D space, the third was the standard deviation σ which characterized the width of the distribution of the atoms that were represented by a PPP (in graphical illustrations of SQUID pharmacophore models σ is visualized by the radius of the PPPs). The fourth attribute (the conservation weight *w*) weighted each PPP by its conservation among the molecules of the alignment (in graphical illustrations of SQUID pharmacophore models *w* is visualized by the intensity of the color of a PPP). This was done under the assumption that more conserved features of a set of molecules binding to the same

receptor with comparable affinity are more important for the binding than less conserved features.

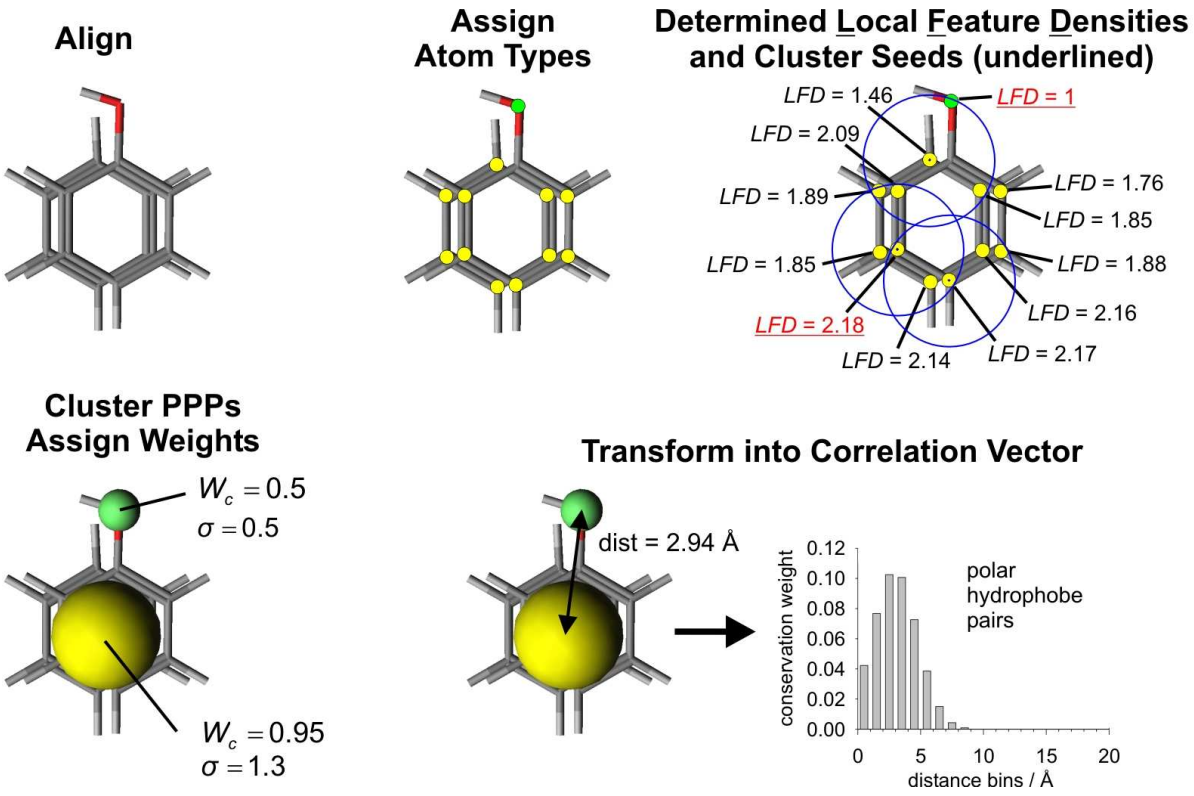


Figure 2.4 Calculation of the SQUID fuzzy pharmacophore correlation vector. Pharmacophore atom-types are assigned to all atoms of a set of aligned molecules (yellow = hydrophobe, green = polar). Maxima in the LFDs (red) are determined to be used as cluster seeds. In this example a cluster radius (r_c) of 1.5 Å was used. Standard deviations (σ) and conservation weights (w) are calculated for each PPP that resulted from the clustering procedure. Finally distances between all pairs of PPPs are measured and the three-dimensional representation is transformed into a correlation vector by equation 2.5. As an example, a section of the resulting CV representing polar – hydrophobe pairs is shown.

2.3.1 Calculation of the SQUID pharmacophore model

A schematic overview of the calculation of a SQUID pharmacophore is given in Figure 2.4. The starting point was an alignment of known active reference compounds. Assignment of pharmacophoric types (cation, anion, hydrogen-bond acceptor, hydrogen-bond donor, polar, or hydrophobic as defined with the pH4_aType function in MOE [Chemical Computing Group]) transformed the alignment into a field of pharmacophoric features. Maxima in the local feature densities (LFD) were used as cluster seeds to cluster the features into PPPs for a

more general representation of the underlying alignment. For each atom *k* of type *t* in the alignment the LFD was calculated by

$$\mathrm{LFD}(\mathrm{atom}_k^t) = \sum_i \max\left\{0{,}1 - \frac{\mathrm{D}_2(\mathrm{atom}_k^t, \mathrm{atom}_i^t)}{r_c}\right\}, \qquad \text{(Eq. 2.3)}$$

where *i* are all atoms of type *t* in the molecular ensemble, D_2 is the Euclidean distance between two atoms and r_c is the cluster radius. Positions of atoms of type *t* for which no other atom of type *t* within r_c was found yielding a higher LFD were taken as cluster seeds for PPPs of type *t*. All atoms were subsequently clustered to their nearest cluster seed of their respective type. The geometric center of the atoms of a cluster was taken as the position of the resulting PPP. The median distance of all atoms contributing to a PPP to the center of the PPP was taken as the value of the standard deviation σ of the PPP. For this value a minimum of 0.5 Å was used. The conservation weights of the PPPs were calculated by

$$w(\mathrm{PPP_k}) = \sum_{i=1}^{m} \min\left\{\frac{1}{m}, \frac{\text{no. atoms from molecule}_\mathrm{i}\ \text{of PPP}_\mathrm{k}}{\text{no. atoms of PPP}_\mathrm{k}}\right\}, \qquad \text{(Eq. 2.4)}$$

where *m* is the number of molecules in the model. This function returns a maximum value of 1 for PPPs representing the same number of atoms from all molecules of the ensemble and a minimum of n^{-1} for PPPs which consist only of atoms of one molecule.

For virtual screening the three-dimensional distribution of PPPs was transformed into a two point PPP-CV (Figure 2.4), arranged exactly like the CATS3D CV. The SQUID CV represents the three-dimensional distribution of Gaussian densities in the form the distribution of pairs of PPPs over the distance bins and over the feature types. The transformation was calculated according to Equation 2.5.

$$\mathrm{CV}_d^T = \frac{1}{\text{no.pairs}(T)} \sum_p \sum_q \frac{1}{2} \delta_{pq}^T \left(\frac{w_p\, w_q}{\sqrt{2\pi}(\sigma_p + \sigma_q)} \exp\left(-\frac{1}{2} \frac{\left(\mathrm{D}_2(p,q) - center_d\right)^2}{\left(\sigma_p + \sigma_q\right)^2} \right) \right), \qquad \text{(Eq. 2.5)}$$

where *p* and *q* are PPPs, *d* is a distance range ("bin"), *T* is the pair of pharmacophoric types of *p* and *q* (e.g. Figure 2.4: *p* = hydrophobic, *q* = polar), *w* are the PPP conservation weights, *σ* is the standard deviation of a PPP, $center_d$ is the center of the distance range *d*, and δ^T (Kronecker delta) evaluates to 1 for all pairs of PPPs of types *T*. D_2 is the Euclidean

distance metric. The factor of 0.5 in the sum avoids double counting of pairs. Pairs of PPPs with themselves were not considered. The values of each dimension were scaled by the total number of possible pairs of PPPs of the two features considered. Finally the CV was scaled to a maximum value of 1, i.e. the largest value in the descriptor was scaled to a value of one and the other values were scaled proportionally. Like the CATS3D descriptor, the SQUID CV consisted of 420 dimensions, representing the same distance bins and pairs of atom-types. The SQUID CV was used to rank molecules encoded with the CATS3D descriptor. When CATS3D was used to encode molecules for SQUID database screening, the final CATS3D descriptor vector was also scaled to a maximum value of 1.

The calculation of the SQUID CV was done with the program *SQUID* which was written in the SVL language of MOE [Chemical Computing Group].

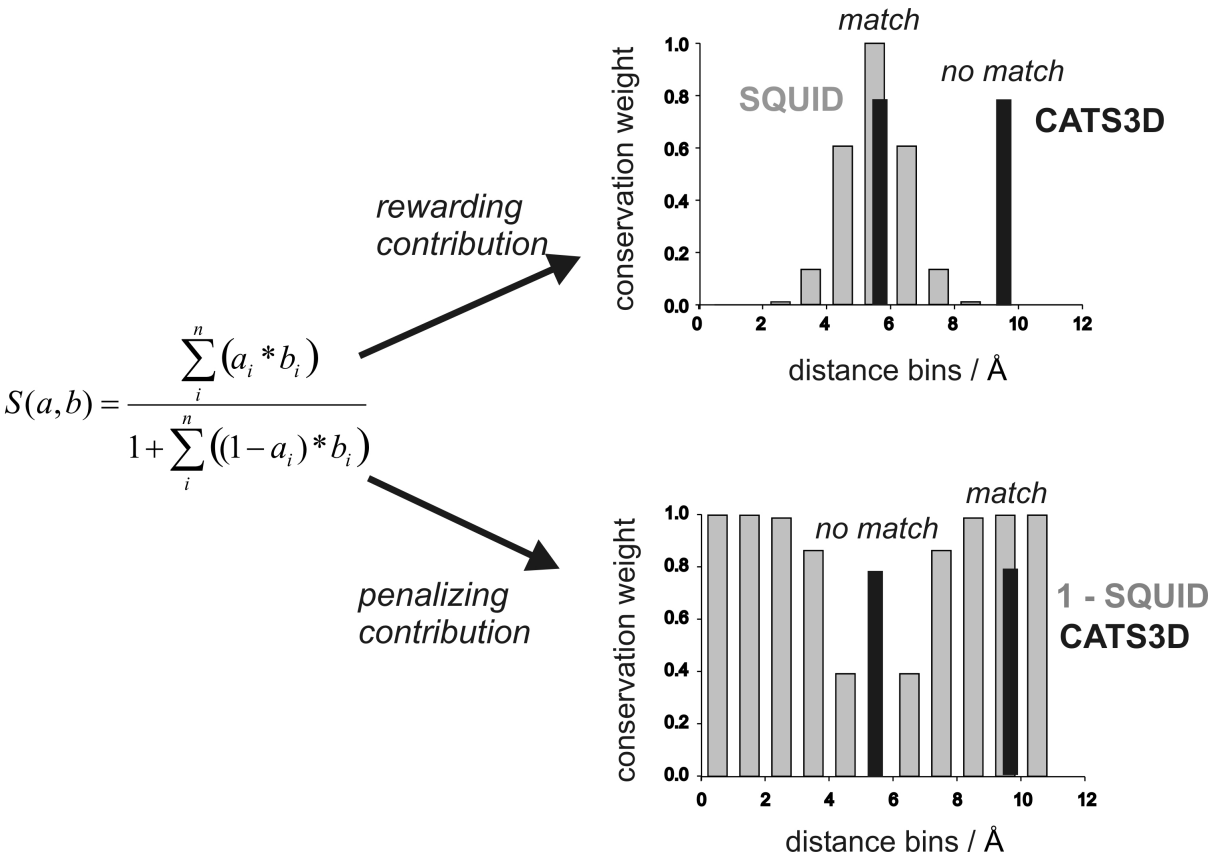


Figure 2.5 The SQUID similarity score. The numerator of the term describes rewarding contributions of the score, i.e. CATS3D dimensions in regions with a high probability in the SQUID correlation vector result in high score (match). CATS3D dimensions in low probability regions have a low impact on that term (no match). The term for penalizing contributions (the denominator) weights CATS3D dimensions by the inverse SQUID vector.

2.3.2 Virtual Screening

For virtual screening the SQUID CV representation was used to weight CATS3D representations of molecules according to their fitness according to their distribution of pharmacophoric features. The SQUID CV and the CATS3D CVs differ significantly in the meaning of their content. The SQUID CV describes a broad range of descriptor areas which are favorable for the desired biological activity, while the CATS3D descriptor contains only a smaller subset of the actual occurrences of atom-pairs in a specific ligand. Consequently, similarity metrics like the Euclidean distance or the Tanimoto index, which are based on the assumption that both descriptors, which are to be compared, represent objects in the same way, cannot be used to assess the activity of the molecules under consideration. To overcome this problem a SQUID similarity score was developed (Eq. 2.6):

$$S_{A,B} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{1 + \sum_{i=1}^{N} (1 - x_{iA}) x_{iB}}, \qquad \text{(Eq. 2.6)}$$

where x_{iA} is the value of the i-th element of the SQUID CV, x_{iB} is the value of the i-th element of a molecule CV and N is the total number of dimensions. The idea of the SQUID similarity is further illustrated in Figure 2.5. The value x_{iA} may be considered as the idealized probability of the presence of features in x_{iB}. This results in high scores for molecules with many features in regions of the query descriptor which have a high probability. To penalize the presence of such atom pairs in regions with a low probability, the denominator weights the presence of atom pairs with the inverted probabilities of the descriptor of the pharmacophore model (a value of 1 was added to the denominator to avoid division by zero and high scores resulting from a very low value in the denominator of the term).

For virtual screening additional weights (“feature-type weights”) were used to weight the importance of each of the pharmacophoric features-types (e.g. hydrophobic or hydrogen-bond donor) in the CV. The sums of the single feature-type weights were used to weight the importance of each pair of feature-types in the CV. The sum of the probabilities in the CV for each pair of features over all distance bins was scaled to the value of the feature-type weights. Finally the whole CV was scaled to a maximum of 1. It was found that a simple optimization by permutation of all combinations of the weight values {0.1, 0.2, 0.3, 0.4, 0.5} for each of the single features and subsequent testing of these weights in virtual screening was sufficient to retrieve good virtual screening results.

Evaluation of the different pharmacophore models obtained from different cluster-radii and feature type weights was done with the program *SQUIDopt* which is based on the workflow of the program *SQUIDscreen* (Figure 2.3). In *SQUIDopt* all pharmacophore model variants (e.g. models from different cluster radii or different feature type weights) serve as references for virtual screening. In this way the different models can be prioritized and the model with the best *ef* value could be used for further virtual screening. In cases where the *ef* was not discriminative enough to favor one or a small set of models a more sensitive measure was used, the enrichment value *ev*:

$$ev = \sum_{i=1}^{100} (101 - i)\ \ ef\,(i\%)\,, \qquad \text{(Eq. 2.7)}$$

where *ef*(*i*%) is the enrichment factor for the first *i*% of the hit-list. This returns the weighted sum of the enrichment factors of the whole database. The smaller the fraction of the database, the higher is the weight for the *ef*.

Virtual screening was performed with the program *SQUIDscreen*, which was written in C++.

2.4 Methods of Section 4.1: Influence of similarity metrics and descriptor vector scaling on CATS3D retrospective screening

Data set

For the retrospective screening experiments the COBRA database (version 2.1; 4705 molecules) [Schneider & Schneider, 2003] of annotated reference molecules from recent scientific literature was employed. Twelve different non-overlapping subsets of COBRA were defined as active molecules (used as query) and the respective remainder of the dataset as inactive molecules. The sets of actives contained molecules that bind to the angiotensin converting enzyme (ACE, 44 compounds), cyclooxygenase 2 (COX2, 93), corticotropin releasing factor (CRF antagonists, 63), dipeptidyl-peptidase IV (DPP, 25), G-protein coupled receptors (GPCR, 1642), human immunodeficiency virus protease (HIVP, 58), matrix metalloprotease (MMP, 77), neurokinin receptors (NK, 188), nuclear receptors (NUC, 211),

peroxisome proliferator-activated receptor (PPAR, 35), beta-amyloid converting enzyme (BACE, 44) and thrombin (THR, 188).

For all COBRA molecules hydrogens were added with CLIFF and single 3D conformations were calculated with CORINA (version 2.64) [CORINA]. CATS3D descriptors were calculated using the three scaling schemes *no-scaling*, *scaling1* and *scaling2*.

Virtual screening

For all 12 activity classes of the COBRA database, and for the three scaling schemes retrospective screening experiments were performed with the program *rankIt*, using the Manhattan distance, the Euclidean distance and the Tanimoto similarity. The relative performance of the different parameter sets was assessed by enrichment factors.

2.5 Methods of Section 4.2: Impact of conformational flexibility on CATS3D virtual screening

Data set

The PDBbind database [Wang *et al.*, 2004] (version 2002) served as a reference set of high-quality crystal structures of receptor-bound ligands for the virtual screening experiments. For retrospective screening we used the COBRA database [Schneider & Schneider, 2003] (version 3.12) consisting of 5,376 annotated ligands compiled from scientific literature. The ligands of the PDBbind database were grouped according to their target annotation. All clusters containing less then five ligands were removed. Clusters were also removed for which no ligands were found in the COBRA database with the same target annotation as in PDBbind. From multiple incidences of identical ligands all but the one with the best resolution were removed. The final set of reference ligands consisted of 11 groups (“activity classes”) with a total number of 177 ligands. The final set of ligands with the corresponding PDB identifier is given in Table 2.2.

The corresponding set of “active” ligands in the COBRA database contained 674 molecules, which means that the COBRA database contained 4,702 additional ligands that were not considered as “active” in either of the 11 activity classes. The final set of annotated activity classes and their abbreviations were: acetylcholinesterase (ACHE, 6 compounds from PDBbind, 13 compounds from COBRA, overlap: 0), carbonic anhydrase II (CAII, 30, 25, 2),

elastase (ELA, 8, 8, 0), factor Xa (FXA, 5, 226, 5), HIV-protease (HIVP, 58, 61, 8), neuraminidase (NEU, 8, 28, 1), protein tyrosine kinase c-src (PTK-CSRC, 7, 16, 0), protein tyrosine phosphatase 1b (PTP1B, 5, 36, 0), stromelysin 1 (STRO1, 7, 19, 0), thrombin (THR, 32, 194, 10), and urokinase type plasminogen activator (UTPA, 11, 48, 3). Since we were not interested in the absolute performance of the method, but in the relative performance using different degrees of conformational information, we did not remove ligands that were present in both databases ("overlap").

Table 2.2. Ligands from the PDBbind dataset selected as reference molecules for virtual screening.

Activity class	**PDB identifier**
Acetylcholinesterase	1e66, 1gpk, 1gpn, 1h22, 1h23, 1vot
Carbonic anhydrase II	1a42, 1avn, 1bcd, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1bzm, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1cny, 1g45, 1g48, 1g4j, 1g4o, 1g52, 1h4n, 1if7, 1if8, 1okl, 1okn, 1ydb
Elastase	1bma, 1ela, 1elb, 1elc, 1eld, 1ele, 1inc, 7est
Factor Xa	1ezq, 1f0r, 1f0s, 1fjs, 1ksn, 1xka
HIV-protease	1a30, 1a94, 1aaq, 1ajv, 1ajx, 1b6j, 1b6k, 1b6l, 1b6n, 1b6o, 1b6p, 1bdq, 1bwa, 1bwb, 1c70, 1d4k, 1d4l, 1d4y, 1dmp, 1g2k, 1g35, 1hbv, 1hih, 1hiv, 1hos, 1hpo, 1hps, 1hpv, 1hpx, 1hsh, 1htf, 1htg, 1hvh, 1hvi, 1hvj, 1hvl, 1hvr, 1hwr, 1hxw, 1izh, 1k6p, 1k6t, 1k6v, 1mtr, 1ody, 1ohr, 1pro, 1qbs, 1qbu, 1sbg, 2bpv, 2bpy, 3aid, 4hvp, 5hvp, 7hvp, 7upj, 8hvp
Neuraminidase	1f8c, 1f8d, 1f8e, 2qwb, 2qwc, 2qwe, 2qwf, 2qwg
Protein tyrosine kinase c-src	1a07, 1a08, 1a09, 1a1b, 1a1c, 1a1e, 1is0
Protein tyrosine phosphatase 1b	1c83, 1c84, 1c87, 1c88, 1ecv
Stromelysin 1	1b8y, 1caq, 1ciz, 1hfs, 1sln, 1usn, 2usn
Thrombin	1d3d, 1d3p, 1d4p, 1d6w, 1d9i, 1etr, 1ets, 1ett, 1g37, 1ghv, 1ghy, 1gi4, 1gj5, 1kts, 1qbv, 1tmt, 1tom, 1uvt, 7kme, 1a4w, 1bcu, 1bhx, 1c1u, 1c1v, 1c4u, 1c4v, 1c5n, 1c5o, 1fpc, 1jwt, 1k21, 1k22
Urokinase type plasminogen activator	1f5k, 1f5l, 1gi7, 1gi8, 1gi9, 1gj7, 1gj8, 1gj9, 1gja, 1gjc, 1gjd,

Calculation of conformations

Single three-dimensional conformations were calculated with CORINA [Sadowski *et al.*, 1994] and multiple three-dimensional conformations were calculated with ROTATE (version 1.15) [Schwab, 2003], based on the CORINA conformations. Conformations were calculated for the selected reference molecules from the PDBbind database and all molecules from the COBRA database. For each database single conformations were calculated with CORINA. To restrict the number of possible output conformations from ROTATE only the five most central rotatable bonds were subjected to torsion angle variation, and conformations with an internal (symbolic) energy of more than 100 kJ/mol above the lowest-energy conformation were rejected. The resulting conformations were further clustered in torsion angle space to obtain only representative conformations. To obtain databases of different conformational resolutions (i.e. different numbers of conformations) different thresholds of 120° (resulting database further referred to as R1), 60° (R2) and 45° (R3) were applied. CATS3D descriptors were calculated for all four COBRA databases with different conformations and the PDBbind crystal structure conformations using *scaling2*.

Superposition and calculation of the RMSD

Rigid body superimposition of molecules was performed to compare two conformations of one molecule. The similarity of two conformations was quantified by the RMSD (root mean square deviation) value of Cartesian atom positions. This was done with the program *Match3d* by Jens Sadowski. *Match3d* takes into account the symmetry of nondistinguishable but differently numbered groups (e.g. the two oxygen atoms in a carboxylate group) for the calculation and thereby avoids artificially introduced high RMSD values. Only non-hydrogen atoms were considered for the calculation.

Virtual screening

The crystal structure conformations of the 11 ligand classes were used as references for retrospective screening of the COBRA database versions with different numbers of conformations, using the program *SQUIDscreen* with the Manhattan distance. The relative performance of the different amounts of conformations was assessed by enrichment factors.

2.6 Methods of Section 4.3: Virtual screening and scaffold hopping efficiency of alignment-free pharmacophore pair descriptors

Data set

For the retrospective screening experiments the COBRA database (version 2.1) [Schneider & Schneider, 2003] was employed using the same activity classes as in Section 4.1, except that the two very general classes G-protein coupled receptors (GPCR) and nuclear receptors (NUC) were discarded for the experiments. For the virtual screening experiments hydrogens were added with CLIFF and single 3D conformations were calculated with CORINA (version 2.64) [Sadowski, 1994]. CATS, CATS3D and SURFCATS descriptors were calculated with *scaling2*. The MACCS keys were calculated with MOE [Chemical Computing Group].

Molecular equivalence numbers

Molecular equivalence indices [Xu & Johnson, 2001; Xu & Johnson, 2002] were used to identify identical scaffolds in molecular databases. The calculations were done with the program Meqi (*M*olecular *eq*uivalence *i*ndices) [Pannanugget Consulting]. Meqi reduces the molecular representation to the scaffold of a molecule and calculates an equivalence number with a modified version of the Morgan algorithm [Morgan, 1965]. For the calculation of equivalence numbers all molecules were preprocessed in the following way in Meqi: First all hydrogens were stripped off of the molecules. Second, all atoms were transformed into carbons with the command “Vertex-labeling.list: C ?”. Third, all bonds were transformed to single bonds with the command “Edge-labeling list: 1 1 2 3 4”.

Two different definitions of scaffolds were used for the equivalence number calculation: cyclic system (scaffold) and reduced cyclic system (reduced scaffold) (Figure 2.6). Scaffolds represent the molecule without sidechains, indifferent for types of atoms and bonds. Scaffolds are chosen with the “Subgraphs: CyclicSystem” button. Reduced representations are characterized by a simplifying representation of rings, which does not further discriminate between rings comprised of different numbers of heavy atoms. Conjugated systems with different numbers of rings are not considered as identical. Reduced representations were obtained with the command “Topology: Reduced”. Exact representations of rings were used with “Topology: Unchanged”. Other parameters of the program were held

constant for all calculations: “Embedding: Unembed”, “Components: Group”, “Attachment type: RingSys”.

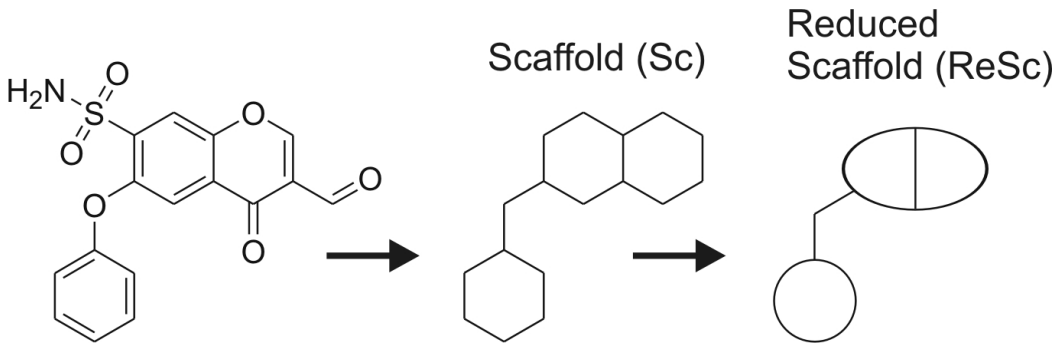


Figure 2.6 Definition of cyclic system “Scaffold” (Sc) and reduced cyclic system “Reduced Scaffold” (ReSc). In this work we defined the scaffold of a molecule as the side-chain depleted molecular graph without annotation of atom-types. A reduced scaffold is a more general representation which does not discriminate between rings consisting of different numbers of heavy atoms, but systems containing different numbers of rings are still not considered being equal.

Virtual screening

For all 10 activity classes of the COBRA database, and for the four molecular descriptors CATS, CATS3D, SURFACTS and the MACCS keys retrospective screening experiments were performed with the program *rankIt*, using the Manhattan distance, the Euclidean distance and the Tanimoto similarity. The relative performance of the different parameter sets was assessed by enrichment factors. To compare the enrichment of scaffolds and reduced scaffolds, enrichment factors were calculated from the first occurrences of each unique scaffold and reduced scaffold in the set of active molecules.

2.7 Methods of Section 4.4: Prospective screening for mGlurR5 allosteric modulators with CATS3D

Data set

A set of seven allosteric inhibitors of mGluR5 with reported low nanomolar activity was compiled from scientific and patent literature [Gasparini *et al.*, 1999; Mutel *et al.*, 2002; Cosford *et al.*, 2003; Gasparini *et al*., 2003] as reference compounds in virtual screening.

For prospective screening the Asinex Gold compound collection [ASINEX] (version april 2003) was used, which contained 194,563 molecules. As a pre-screening filter we selected the 20,000 most “drug-like” compounds [Schneider & Schneider, 2004] in the same manner as described previously for the SPECS database. 3D-conformations of the screening compounds were calculated in MOE using the MMFF94 force field [Halgren, 1996]. The results were restricted to a maximum of 20 lowest energy conformations per molecule. CATS3D descriptors were calculated with the *scaling2* option.

For the analysis of the virtual screening results CATS descriptors were calculated with the program *speedcatsdotcom* [Fechner *et al.*, 2003] with default parameters and the MACCS keys were calculated with MOE [Chemical Computing Group].

Alignment of reference molecules

To form a hypothesis about receptor-bound 3D-conformations of the reference molecules the flexible alignment tool of MOE was used with default parameters and the MMFF94 forcefield [Halgren, 1996]. Ligands were successively aligned, starting with the most rigid molecule to the most flexible molecule.

Virtual screening

Prospective screening was performed with each of the reference molecules with *SQUIDscreen* using the Manhattan distance.

2.8 Methods of Section 4.5: Prospective screening for mGluR5 allosteric modulators with an artificial neural network approach based on CATS3D representations

Data sets

For neural network training 68 mGluR5 allosteric antagonists from literature, patents and from unpublished results of Merz Pharmaceuticals, and 158 allosteric antagonists of mGluR1 from patents and literature were used. Molecules that were not active on either mGluR5 or mGluR1 were compiled from the COBRA database (version 3.12) [Schneider & Schneider,

2003]. From the COBRA database all molecules were removed with a substring “mGluR” in the identifier.

For all molecules single 3D conformations were calculated with CORINA [Sadowski, 1994]. CATS3D descriptors were calculated with *scaling2*. MACCS keys were calculated with MOE [Chemical Computing Group].

Maximal diverse subset selection

Maximal diverse subsets were selected with the MaxMin algorithm [Kennard & Stone, 1996]. The algorithm starts with an initial molecule as the subset selection. Successively, the molecule from the remaining molecules, which is most dissimilar to the already selected molecules, is added to the selected molecules. The procedure stops, when the desired number of molecules is selected. For subset selection the program *MaxMinSelection* [Schmucker *et al.*, 2004] by Michael Schmucker was used, employing the Euclidean distance for dissimilarity assessment. An extended version of the program by Uli Fechner was used which enables the initialization with a randomly selected molecule in the selected.

Shannon entropy based variable selection

Selection of variables is important for predictive QSAR results if not all variables in a descriptor, e.g. all 420 CATS3D dimensions, contain information which is related with the prediction problem. Other variables might not show much or any variance and are though not useful for predictions either. We used Shannon entropy based variable selection, which is based on the Shannon entropy concept formulated by Shannon in 1963 [Shannon, 1963]. This concept was shown to be successful in descriptor selection for classification and QSAR applications [Stahura et al., 2000; Godden & Bajorath, 2003]. The Shannon entropy is a measure for the distribution of a variable over a range of values. If all possible states of a variable are equally populated the Shannon entropy is at maximum. If only a single state is populated, the variable has a minimum of entropy. Variables with larger values for the Shannon entropy are preferred over variables with lower entropy.

The Shannon entropy is defined by Equation 2.8:

$$SE = -\sum_{i=1}^{N} p_i \log_2 p_i \,, \qquad \text{where} \quad p_i = c_i \Big/ \sum_{i=1}^{N} c_i \quad . \qquad \text{(Eq. 2.8)}$$

p_i is the probability of observing a particular descriptor value, falling into a bin *i*. For continuous variables, the range of values of the descriptor is partitioned into *N* equal sized bins. c_i is the number of instances having a descriptor value falling into bin *i*.
In the formulation of Eq. 2.8 the Shannon entropy is dependent on the number of bins. A bin number independent formulation is the scaled Shannon entropy (Eq. 2.9)

$$sSE = SE / \log_2 N \; . \qquad \text{(Eq. 2.9)}$$

The range of *sSE* is from 0 to 1. For our studies we used $N = 100$, defined from the minimum to the maximum value. Variables were selected with a $sSE \geq 0.3$.

Autoscaling

Autoscaling was used as a pre-procession step for the principle component analysis. With autoscaling variables are scaled by their standard deviation, leading to data with zero mean and unit variance. In this way differences between variables resulting from different value ranges and different size ranges are eliminated. A scaled variable x^* is obtained by Equation 2.10:

$$x_{ik}^* = \frac{x_{ik} - \bar{x}_k}{s_k} \quad , \qquad \text{where} \quad s_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} \left(x_{ik} - \bar{x}_k \right)^2} \; . \qquad \text{(Eq. 2.10)}$$

x_{ik} is the value of the k^{th} dimension of molecule *i*, $\bar{x}_k$ is the mean value of all x_{ik} and s_k is the standard deviation. Autoscaling results in data vectors scaled to length $\sqrt{n-1}$.

Principle component analysis (PCA)

Principle component analysis is a method to obtain uncorrelated variables. Correlated variables of a descriptor introduce a bias for these descriptor variables, which can deteriorate the performance of prediction methods. PCA can also be used for the visualization of high dimensional data in a two- or three-dimensional coordinate system.

Uncorrelated variables are obtained by a linear projection from an original *m*-dimensional space X into a lower *d*-dimensional space S by S = XL. The projection is defined by the loadings matrix L^T which contains *d* vectors of *m* coefficients. The matrix containing the *d* new coordinates or variables for each molecular object is called the scoring matrix S.

The principle components (PC) represent the new coordinate system of the projected variables. The first PC coordinate axis is directed parallel to the maximum variance of the distribution of the data points in the original space. Accordingly the first PC explains most of the variance in the data. The second PC is orthogonal to the first PC and explains most of the remaining variance of the data. *m* PCs explain the full variance of the data. An efficient algorithm for the calculation of the PCs is the NIPALS algorithm [Wold, 1966; Wold, 1975], which was utilized here.

The eigenvalue of a PC is the variance which is explained by the PC. The eigenvalue is calculated by the sum of the squared loadings of the PC. To obtain a small set of relevant uncorrelated variables, only PCs with eigenvalues ≥ 1 were selected.

PCA transforms were calculated with the program nipals by Alexander Böcker.

Feed-forward artificial neural networks

The most widespread architecture of ANNs is multilayered feed-forward networks. The non-linear behavior of multilayered feed-forward neural networks enables ANNs to learn in principle any relationship between input and output. For our studies we used three layered fully-connected networks with an input layer, a hidden layer and an output layer (Figure 2.7).

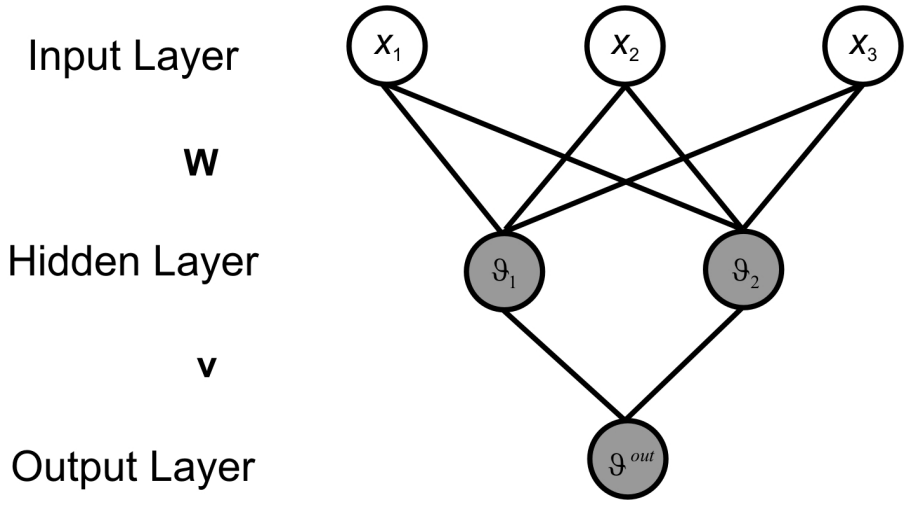


Figure 2.7 Three layered feed-forward artificial neural network. The input layer consists of as many neurons as the dimensionality of the input data. The optimal number of hidden neurons has to be determined experimentally. The hidden layer and the output layer consist of sigmoidal neurons. **w** and **v** are the weights of the input to the hidden neurons and from the hidden neurons to the output neuron, respectively. ϑ_j are the bias values from the hidden layer and ϑ^{out} is the bias of the output neuron.

The output of a three-layered ANN can be expressed by Equation 2.11:

$$\mathrm{f}(x) = \mathrm{Sigm}\left(\sum_{j=1}^{m} v_j \mathrm{Sigm}\left(\sum_{i=1}^{n} w_{ij} x_i - \vartheta_j \right) - \vartheta^{out} \right), \qquad \text{(Eq. 2.11)}$$

where w_{ij} are the weights from the input to the hidden neurons, v_j are the weight from the hidden to the output neurons, ϑ_j are the bias values from the hidden layer and ϑ^{out} is the bias of the output neuron. Sigm (Eq. 2.12) is the sigmoidal function

$$\mathrm{Sigm}(x) = \frac{1}{1+\mathrm{e}^{-x}}. \qquad \text{(Eq. 2.13)}$$

The bias and weight values were determined in a training procedure employing a (1, λ) evolution strategy [Schneider & So, 2003]. Evolutionary strategies (ES) are assumed to be favorable in comparison to gradient based optimization methods like the backpropagation algorithm for optimization tasks in complex multimodal fitness-landscapes like found within drug discovery projects [Schneider & So, 2003]. A (1, λ) evolution strategy selects from a pool of samples only the fittest (the first parameter in (1, λ)), that is used as a parent for the generation of λ offsprings. The parent dies after reproduction. This is assumed to avoid the selection of local minima solutions. The ES starts with a random set of weights and bias values and generates a set of children with Gaussian distributed variations. At the beginning of the optimization process, the width of the Gaussian distribution (the step -size σ) is large to facilitate the search for the approximate location of the global optimum. As the algorithm proceeds to the optimum σ becomes smaller. This is realized by inclusion of the σ values into the evolutionary optimization.

The evaluation of prediction accuracy of the candidate ANNs in the training was done by the mean square error (*MSE*) function (Eq. 2.14)

$$MSE = \frac{1}{S} \sum_{i=1}^{S} \left(output_i^{actual} - output_i^{desired} \right)^2, \qquad \text{(Eq. 2.14)}$$

where *S* is the number of data samples. The *MSE* quantifies the distance between the predicted values to the desired values from the training samples.

The accuracy of classification tasks was assessed by the Matthews correlation coefficient (Eq 2.15):

$$cc = \frac{PN - OU}{\sqrt{(N+U)(N+O)(P+U)(P+O)}}, \qquad \text{(Eq. 2.15)}$$

where P is the number of positive correct prediction, N is the number of negative correct predictions, O is the number of false-positives (overpredictions), and U is the number of false-negatives (underpredictions). The results of the *cc* range from -1 to 1. A *cc* of 1 means a perfect prediction, a *cc* of 0 corresponds to a random prediction and a *cc* of -1 means a 100% wrong prediction. A threshold of 0.5 was used to classify objects as active or inactive for the calculation of the *cc*.

A problem in neural network training is overfitting of the ANN to the training data. This results in a loss of generalization for the prediction of new data. The ANN has learnt the examples but no rules to separate the class of actives from inactives. To avoid overfitting the original dataset is randomly split into training and test data sets. The training set is used for the training of the ANN and the test set is used to supervise the generalization ability of the ANN. The training is stopped when the prediction accuracy of the test set starts to decrease after an initial phase of improvement while the prediction of the training set still improves. The random split into training and test data followed by training is repeated several times to obtained statistically significant stopping criteria. This procedure is called cross-validation.

The training of ANNs was done with the program *profi* by Gisbert Schneider [Schneider & Wrede, 1993]. Ten times cross-validation was applied splitting the data in equally sized fractions of 50% / 50%. For the evolution strategy 500 solutions per generation, an initial step size of $\sigma = 1$ and a minimal step size of $\sigma = 0.001$, with a reset step-size of 0.01 were used. The reset step size is the minimum value of σ for each new generated child in the evolution strategy. For the training of classification tasks active molecules were marked with a target value of 1 and inactives with a target value of 0.

Using a consensus score obtained from an ensemble of neural networks has been shown to improve the quality of predictions in comparison with a single ANN [So & Karplus, 1996; Kauffman & Jurs, 2000]. Accordingly average values from the scores of multiple neural networks were used for the prediction of properties.

Self organizing maps (SOMs)

SOMs represent a class of unsupervised neural networks which are mainly applied for clustering, feature extraction and topology preserving projections [Schneider & So, 2003]. SOMs consist of a single layer of neurons that have the same dimensionality as the data vectors. The two-dimensional distribution of neurons resembles the distribution of data points in the original high dimensional space. There is no predicted output value for an input object but a winner neuron which is most similar to the input object. Each neuron represents a prototype vector for the data objects which are most similar to this neuron. The field within which data points are assigned to a neuron is called the receptive field of this neuron. A neuron is activated if a data points falls into its receptive field. The preservation of the data topology is achieved by the definition of a topology in the neuron layer. This layer is fitted onto the original data distribution, preserving the original topology of the data. To avoid boundary problems the maps have a toroidal topology.

The training of SOMs was done according to Kohonen [Kohonen, 1982]:

(1) Initialize the map with $N = N_1 * N_2$ neurons c_i with reference vectors w_{c_i} chosen randomly from the distribution p(ζ) of training patterns. Initialize connections between the neurons to form a rectangular N_1 x N_2 grid. The time parameter t is set to 0.

(2) Randomly select a training pattern ζ from p(ζ) as input signal.

(3) Determine the winner neuron with the smallest Euclidean distance to the input signal.

(4) Adapt each neuron in the SOM to the training pattern.

(5) Increase the time parameter: $t = t + 1$.

(6) If t < tmax continue with step (2), else terminate.

The adaptation of a neuron r to a training pattern ζ is done according to:

$$w_r = w_r + \Delta w_r \text{ , where}$$

$$\Delta w_r = \varepsilon(t) h_{rs} (\zeta - w_r).$$

$$h_{rs} = \exp\left(\frac{-D_1(r,s)^2}{2\sigma(t)^2} \right)$$

is the Gaussian neighborhood function around the winner neuron s and D_1 is the Manhattan distance. The time dependent standard deviation is defined as

$$\sigma(t) = \sigma_{initial}\left(\sigma_{final} / \sigma_{initial}\right)^{t/t_{\max}},$$

and the time dependent learning rate is defined as

$$\varepsilon(t) = \varepsilon_{initial}\left(\varepsilon_{final} / \varepsilon_{initial}\right)^{t/t_{\max}}.$$

The training of SOMs was calculated with the program *som_create* by Gisbert Schneider using 5 x 5 = 25 neurons, $t_{\max}$ = 85000, σ_{initial} = 5 and $\varepsilon_{\text{initial}}$ = 1, and $\varepsilon_{\text{initial}}$ and σ_{initial} = ($t_{\max}$ / number of training patterns). Visualization of SOMS was done with the program *som_show* by Gisbert Schneider.

2.9 Methods of Section 4.6: Retrospective evaluation of SQUID fuzzy pharmacophore models

Data set

Pharmacophore models were calculated for COX-2 and thrombin on the basis of molecules which were reported in pharmacophore models for the respective targets [Palomer *et al.*, 2002; Patel *et al.*, 2002]. For calculation of a COX-2 pharmacophore model, the crystal structures of COX-2 with the specific inhibitor SC-558 (1CX2) and the structures of COX-2 with the unspecific inhibitors flurbiprofen (3PGH) and indomethacin (4COX) were used to model a template alignment for the flexible alignment of the specific COX-2 inhibitors rofecoxib and molecule 5 (M5) from Palomer *et al.* [Palomer *et al.*, 2002]. For calculation of the thrombin pharmacophore model the crystal structures with PDB codes 1C4V, 1D4P, 1D6W, 1D9I, 1DWD, 1FPC and 1TOM were used [Patel *et al.*, 2002].

For retrospective screening we used the COBRA database [Schneider & Schneider, 2003] (version 2.1). Two versions were calculated: one database with single conformations was calculated with CORINA [Sadowski, 1994] and one database of up to 50 energy minimized conformations was calculated with MOE [Chemical Computing Group] using the MMFF94 forcefield [Halgren, 1996]. For retrospective screening the molecules that were used for the pharmacophore model generation were removed from the datasets. The resulting datasets consisted of 92 active molecules and 4611 inactive molecules for COX-2, and 188

actives and 4517 inactive compounds for thrombin. The CATS3D descriptor for the COBRA sets was calculated using *scaling2*.

Alignment of reference molecules

Alignments of inhibitors were either obtained by the flexible alignment tool in MOE with default parameters and the MMFF94 forcefield [Halgren, 1996] or the homology align tool in MOE using default parameters.

Virtual screening

Virtual screening with SQUID pharmacophore models was performed with the program *SQUIDscreen*. CATS3D similarity searching was also performed with *SQUIDscreen* using the Euclidean distance. MOE pharmacophore models were calculated using the PCH_ALL atom-type scheme, which consists of atom-types for cationic, anionic, hydrogen-bond donor, hydrogen-bond acceptor, aromatic ring centers, and hydrophobic interactions.

2.10 Methods of Section 4.7: Prospective screening for inhibitors of the Tat-TAR RNA interaction with a SQUID fuzzy pharmacophore model and CATS3D

Data set

Two reference inhibitors for the Tat TAR interaction were taken from literature: acetylpromazine [Lind *et al.*, 2002] in the receptor bound conformation from the NMR structure 1LVJ [Du2002] that served as a template for the flexible alignment of CGP40336A [Hamy *et al.*, 1998].

For the optimization of the “feature-type weights” the two reference ligands, used for the pharmacophore calculation, were used for retrospective screening in the COBRA database [Schneider & Schneider, 2003] (version 3.12). Up to 20 low energy conformations were calculated with MOE [Chemical Computing Group] for each of the molecules in this database using the MMFF94 forcefield [Halgren, 1996].

For prospective screening the SPECS database [SPECS] (june 2003 version) with 229,658 molecules was used. To obtain higher quality results and to restrict the calculation of

3D conformations the 20,000 most druglike molecules were selected, as predicted by an artificial neural network approach [Schneider & Schneider, 2004]. For each of these molecules multiple conformations were calculated in MOE like for the COBRA molecules.

Calculation of drug-likeness score

"Drug-likeness" was calculated according to a procedure described in [Schneider & Schneider, 2004]. Three parameters were used for the calculation: i) the output ("score") of an artificial neural network that was trained to distinguish between "drugs" and "nondrugs", based on CATS representations of molecules, ii) predicted aqueous solubility [Engkvist & Wrede, 2002], and iii) calculated polar surface area (PSA) (ASA_P option from MOE). Subsequent principal component analysis of this three-dimensional "drug-likeness" space was performed to obtain uncorrelated variables. A ranking of compounds was performed on the basis of their distance to "optimal" variable values (i.e., high drug-likeness score; high solubility value; PSA < 140 Å2). A detailed description of this procedure is given in [Schneider & Schneider, 2004].

Alignment of reference molecules

For the alignments of the known reference Tat-TAR interaction inhibitors ligands the flexible alignment tool in MOE was used with default parameters and the MMFF94 forcefield [Chemical Computing Group].

Virtual screening

Virtual screening with SQUID pharmacophore models was performed with the program *SQUIDscreen*. CATS3D similarity searching was performed with *SQUIDscreen* using the Manhattan distance.

2.11 Methods of Section 4.8: Prospective screening for taspase1 inhibitors with a receptor-derived pharmacophore model

Data set

The sequence of the human taspase1 was obtained from swiss-prot (entry Q9H6P5). Homologous crystal structures that were used as template for the homology models calculation were 1T3M, 2GAW and 1APZ.

For prospective screening the SPECS database [SPECS] (june 2003 version) with 229,658 molecules was used. To obtain higher quality results and to restrict the calculation of 3D conformations the database was filtered according to the Lipinski “rule of five” [Lipinski *et al.*, 1997] and additional target specific filters prior to conformation calculation. For each of these remaining molecules up to 20 low energy conformations were calculated in MOE like using the MMFF94 forcefield [Halgren, 1996]. Finally all bases were protonated and all acids deprotonated with the MOE database/wash function.

BLAST search

The BLAST [Altschul *et al.*, 1997] search was performed using the sequence of taspase1 as query. The BLOSUM62 matrix with a gap opening penalty of 11 and a gap extension penalty of 1 was applied.

Homology modeling

Homology models were calculated with MOE [Chemical Computing Group]. Sequence and structure based alignments were calculated with the Homology/Align function in MOE, using the default values (blosum62 substitution matrix with a gap start panelty of 7 and gap extention panelty of 1). The visualization of the alignments was done with the program CHROMA [Goodstadt & Ponting, 2001].

Ten models were calculated based on the alignment. The coordinates of the final model were calculated as the average of the atom coordinates of the intermediate models. Refinement of the model was done by minimizing the sidechains of the models (backbone atoms were held fixed) with the MMFF94xx forcefield including solvatation to a RMS

gradient of 0.1. Minimization was done using chiral constraints, i.e. the chirality of the molecule was held fixed. Results were controlled with the protein report function of MOE.

Docking

Docking calculations were computed with the program GOLD [Jones *et al.*, 1997]. For GOLD, the genetic algorithm parameters were used with the standard default settings. Chemscore was applied as fitness function [Eldridge *et al.*, 1997]. Early termination was disabled.

Database filtering

The SPECS database was filtered according to the Lipinski "rule of five" [Lipinski *et al.*, 1997] and target specific filters, based on the MOE descriptors in Table 2.3. Molecules were discarded from the SPECS database which satisfied one of the criteria based on an extended version of the "rule of five": > 500 Da, logP > 5, > 5 hydrogen-bond donors, > 10 hydrogen-bond acceptors, > 10 rotatable bonds. Since the inhibitors were thought to depend on one acidic group, all molecules with less than one acidic group were removed. Molecules with Br, I, B, P, S- and nitro groups and sulfat as only single acidic group were also removed

Table 2.3. MOE Descriptors

Descriptor name	Description
Weight	Molecular mass
logP(o/w)	logarithm of the octanol / water partition coefficient
a_don	number of hydrogen-bond donors
a_acc	number of hydrogen-bond acceptors
b_rotN	number of rotatable bonds
a_nBr	number of bromine atoms
a_nI	number of iodine atoms
a_nP	number of phosphorus atoms

Virtual screening

Virtual screening was performed with the MOE pharmacophore search tool.

3 Experimental Section

Note: The following methods and experimental procedures were not applied by the author of the thesis. Inclusion of the experimental details is for the purpose of completeness of the scientific results.

3.1 Determination of IC_{50} values for mGluR

Materials

[3H]-MPEP was obtained from Tocris Cookson (Bristol, UK). MPEP was synthesized for in-house use as a reference compound according to [Gasparini *et al.*, 1999; Sashida *et al.*, 1988]. Test compounds were purchased as dry powder from ASINEX Ltd. (Moscow, Russia). The ASINEX Gold Collection Database was provided by ASINEX Ltd. [3H]-MRZ 3415 was synthesized by Amersham Biosciences (Buckinghamshire, UK). MRZ 3415 was synthesized for in-house use as a reference compound by the Latvian Institute of Organic Synthesis (Riga, Latvia).

Membrane preparation

Male Sprague Dawly Rats (approx. 200-250 g) were anaesthetized and decapitated. Forebrains were removed and homogenized (Ultra Turrax, 8 strokes, 600 rpm) in 0.32 M Sucrose. The suspension was centrifuged at 1,500 g for 4 min. using a Centrikon T-2050 Ultracentrifuge (Tegimenta AG, Rotkreuz, Switzerland). Supernatant was removed and centrifuged at 20,800 g for 20 min. The resulting pellet was re-suspended in ice-cold distilled water and centrifuged at 7,600 g for 20 min. Supernatant and loosely associated flocculent membrane material (buffy coat) were removed by gentle trituration of the pellet and centrifuged at 75,000 g for 20 min. Supernatant was discarded and the membrane pellet was resuspended by sonication in Tris-Buffer (5 mM, pH 7.4) and afterwards centrifuged at 75,000 g for 20 min. The last step was repeated twice and membranes were re-suspended in Tris-Buffer (50 mM, pH 7.5).

The concentration of protein was determined by the Lowry protein assay with bovine serum albumin as a standard. Membranes were stored frozen at –24°C, thawed on the day of the assay and washed again four times at 75,000g for 20 min.

All centrifugation steps were carried out at 4°C.

[3H]-MPEP binding

After thawing, membranes were washed four times with ice-cold binding buffer containing 50 mM Tris-HCl, pH 7.5. Binding assays were performed at room temperature in duplicate using fixed concentrations of test compound (10 µM). The assay was incubated for 1 h in the presence of radiotracer (5 nM) and membranes (1.2 mg/ml) and non-specific binding was estimated using 10 µM MPEP. Binding was terminated by rapid filtration through GF 52 glass-fiber filters (Schleicher&Schuell, Dassel, Germany) using a 1225 Sampling Manifold (Millipore GmbH, Eschborn, Germany). Filters were washed twice with ice-cold assay-buffer and transferred to scintillation vials. After addition of Ultima-GoldTM MV (Packard Bioscience, Groningen, The Netherlands) radioactivity collected on the filters was counted in a 1500 Tri-Carb Packard Scintillation Counter.

[3H]-MRZ 3415 Binding

After thawing, membranes were washed four times with ice-cold binding buffer containing 50 mM Tris-HCl, pH 7.5. Binding assays were performed at room temperature in quadruplicate on 96-well format using fixed concentrations of test compound (10 µM). The assay was incubated for 1 h in the presence of radiotracer (1 nM) and membranes (0.8 mg/ml) and non-specific binding was estimated using 10 µM MRZ 3415. Directly after transferring the reaction volume onto a 96-well multiscreen plate with glass fiber filter 0.22 µm (Millipore GmbH, Eschborn, Germany) binding was terminated by rapid filtration using a multiscreen vacuum manifold (Millipore GmbH, Eschborn, Germany). Afterwards, filters were washed four times with ice-cold assay-buffer and Ultima-GoldTM MV Scintillation Cocktail (Packard Bioscience, Groningen, The Netherlands) was added. After 14 h – 16 h radioactivity was counted in a MicroBeta®Trilux (Perkin Elmer Life Sciences GmbH, Rodgau-Jügesheim, Germany).

Solubility Determination

40 µl of the stock-solution (10 mM, dimethyl sulfoxyd (DMSO) as solvent) of each test compound were diluted with 1.96 ml DMSO to a final concentration of 200 µM. 100 µl of this solution were diluted by addition of 1.99 ml of a solvent consisting of methanol and deionized water (1:1). The resulting solution *A* has a concentration of 10 µM of the test compound containing 5% DMSO. Solution *B* was prepared in the same manner but using Tris-buffer 50 mM, pH 7.5 as solvent instead of the methanol/deionized water mixture.

To determine peaks of the different solutions a HP Series 1100 HPLC device with diode array detector (Agilent Technologies, Waldbronn, Germany) was used. Both solutions flew separately through a SymmetryTM C18 Column (Waters Corporation, Milford, MA) with a average pressure of 190 atmosphere. The resulting peaks of both solutions were compared at a wavelength where the "area under the curve" (AUC) of the peak of solution *A* and solution *B* respectively displayed a maximum. The AUC of solution *A* was defined as 100%-value. Thus, the solubility of each test compound was determined as follows:

$$\text{Solubility}[\%] = \frac{AUC_{solution\ B}}{AUC_{solution\ A}} * 100 \qquad \text{(Eq. 3.1)}$$

IC_{50}-value Estimation

IC_{50} values were estimated from the % of control values from the scintillation assay with a four parameter logistic equation. If both the radio-ligand and the competitor reversibly bind to the same binding site, binding at equilibrium follows equation 3.2.

$$y = \frac{100\%}{1 + (\frac{x}{IC_{50}})^{s}} \ . \qquad \text{(Eq. 3.2)}$$

If *s* is assumed to be 1 equation 3.2 can be reformulated to:

$$IC_{50} = \frac{x}{(\frac{100\%}{y} - 1)} \ , \qquad \text{(Eq. 3.3)}$$

where *s* = slope factor = 1;

x = concentration of test compound [µM] in the assay;

y = result of the binding assay for the test compound [% of control].

K_i-values were calculated from the IC_{50}-values by the Cheng-Prussof Equation [Cheng & Prussof, 1973].

$$K_i = \frac{IC_{50}}{1 + \frac{L}{K_d}} \quad , \qquad \text{(Eq. 3.4)}$$

where L corresponds to the radio-ligand concentration and K_d to its dissociation constant.

3.2 Determination of *IC₅₀* values for TAR-RNA

Materials

Argininamide was purchased from *Sigma Chemical Corp*. (St. Louis, USA). The molecules resulting from virtual screening were purchased from *SPECS* (Delftechpark, The Netherlands) as 10 mM stock solutions in DMSO, and diluted for binding assays with DEPC- treated water to 1 mM or 100 µM, respectively. Fluorescence based binding assays[20] were performed in 96 well microplates at 37°C. Reader: *FluoStar Galaxy* (*BMG Labtechnologies*, Offenburg, Germany), excitation wavelength 540 nm, emission wavelength 590 nm. Microplates: *Corning* 6860, black, non binding surface. The dye labeled $Tat_{49\text{-}57}$–sequence fluoresceine–AAARKKRRQRRRAAAC–rhodamine (1 µM stock solution) was purchased from *Thermo Electron Corporation* (Ulm, Germany). Oligonucleotides were obtained from *Biospring* (Frankfurt, Germany).

In vitro transcription

Equimolar amounts of T7-primer (5‘-TAATACGACTCACTATAG-3‘) and TAR template (5‘-GGCCAGAGAGCTCCCAGGCTCAGATCTGGCCCTATAGTGAGTCGTATTA-3‘) were mixed in TE buffer (10 mM Tris-HCl, 1 mM EDTA; pH 7.4) to give a final concentration of 50 pmol / µL in a volume of 100 µL. After heating to 90 °C for 5 minutes, the reaction was allowed to cool down slowly to room temperature. All *in vitro* transcriptions were performed with T7 polymerase containing RiboMax™ Large Scale RNA Production

Systems Kit (#P1300; *Promega*, Mannheim, Germany) as described by the manufacturer. Subsequent to transcription the DNA template was removed as follows: After heating the transcription mixture to 95 °C for 5 minutes it was chilled immediately on ice. 10 µL RQ1 DNase buffer (*Promega*) and 20 µL RQ1 RNase-free DNase (20 U, *Promega*) were added and the mix was incubated for 30 minutes at 37 °C. Following phenol / chloroform extraction, RNA was precipitated with 3 volumes of ethanol in the presence of 0.3 M sodium acetate (pH 5.2). The RNA was desalted on a NAP^{TM} column (*Amersham Biosciences*, Freiburg, Germany). After lyophilisation, the RNA pellet was redissolved in DEPC treated water to a concentration of 100 µM (stock solution) or 1 µM (final dilution), respectively.

FRET assay

The following stock solutions were used in the assay: labeled Tat-peptide (1 µM), TAR-RNA (1 µM), TK buffer (500 mM Tris-HCl, 200 mM KCl, 0.1% Triton-X 100, pH 7.4). The final volume per well was 100 µL. The fluorescence of pure Tat peptide was determined first: 10 µL stock solution of Tat and 10 µL TK buffer were filled up with DEPC treated water to 100 µL. 10 µL of Tat solution, 10 µL of TAR solution (each 1 µM), 10 µL TK buffer, and 70 µL DEPC treated water were then mixed in a second well to measure the emission of the Tat-TAR complex. Having established the numbers for free and for bound peptide, single point measurements of potential inhibitors were carried out at concentrations of 1000, 100, and 10 µM by using 10 µL of the stock solutions (10 mM, 1 mM, and 100 µM). RNA and peptide concentrations were kept constant at 100 nM in each well (10 µL Tat, 10 µL TAR, 10 µL TK buffer, 10 µL inhibitor, and 60 µL DEPC treated water). Addition of DMSO strongly increases the fluorescence of rhodamine independently from peptide-RNA binding. To eliminate this effect, samples of Tat and of Tat-TAR (each 100 nM) were also measured in the presence of 10 %, 1 %, or 0.1 % DMSO. Dividing these numbers by the value obtained in pure water generated the correction factors. For compounds which showed an inhibitory effect in the preliminary test, complete titration curves were determined from 11 data points. The molecular concentration at which the fitted titration curve intersected with the mean value of the fluorescence counts of the Tat-TAR complex and uncomplexed Tat was used as the IC_{50} value of a molecule.

4 Main Section

The main section is organized in the following way: Sections 1-3 cover the retrospective evaluation of pharmacophore pair descriptors (CATS, CATS3D and SURFCATS) with respect to similarity metrics, scaling, multiple conformations and scaffold hopping. Sections 4 and 5 cover the SQUID fuzzy pharmacophore model approach, including the evaluation of the method and a prospective virtual screening for Tat-TAR inhibitors. Section 6 and 7 report prospective virtual screening experiments for allosteric antagonists of the metabotropic glutamate receptor 5 using CATS3D similarity searching and an artificial neural network approach. Section 8 addresses the prospective virtual screening with a ligand- and binding-site-based pharmacophore model of taspase1.

4.1 Influence of similarity metrics and descriptor vector scaling on CATS3D retrospective screening

The CATS3D descriptor is a three-dimensional extension of the topological pharmacophore-pair CATS descriptor developed by Schneider [Schneider *et al.*, 1999] for ligand based virtual screening. Several parameters can influence the effectiveness of virtual screening. Among these are the set of reference molecules, the molecular descriptor and the similarity metric. We wanted to test if there are some general optimal settings for virtual screening with the CATS3D descriptor. In detail we wanted to test the influence of different similarity indices, namely the Manhattan distance, the Euclidean distance and the Tanimoto similarity coefficient (Table 2.1). Further we were interested in the effect of different scaling schemes on the performance of the CATS3D descriptor. Three different scaling schemes were tested:

- No scaling ("*no-scaling*").
- Division by the number of PPPs of a molecule ("*scaling1*").
- Division of each of the 21 possible pairs of PPPs by the added occurrences of the two respective PPPs ("scaling2").

No-scaling corresponds to the histogram of pairs of PPPs in a molecule. *Scaling1* was the original scaling scheme reported for the CATS descriptor [Schneider *et al.*, 1999]. The aim of *scaling1* is to reduce dissimilarities of molecules based on different molecular size. *Scaling2* is an extension of scaling1, first reported for CATS3D [Fechner *et al.*, 2003]. The aim of *scaling2* is to reduce the bias of highly populated types of PPPs (mainly the hydrophobic PPPs) on the similarity between molecules.

For the retrospective screening experiments we employed the COBRA database (version 2.1) [Schneider & Schneider, 2003] of annotated reference molecules from recent scientific literature. Twelve different datasets were compiled from the COBRA database. These non-overlapping subsets were defined as active molecules (used as query) and the respective remainder of the dataset as inactive molecules. The sets of actives contained molecules that bind to angiotensin converting enzyme (ACE, 44 compounds), cyclooxygenase 2 (COX2, 93), corticotropin releasing factor (CRF antagonists, 63), dipeptidyl-peptidase IV (DPP, 25), G-protein coupled receptors (GPCR, 1642), human immunodeficiency virus protease (HIVP, 58), matrix metalloprotease (MMP, 77), neurokinin receptors (NK, 188), nuclear receptors (NUC, 211), peroxisome proliferator-activated receptor (PPAR, 35), beta-amyloid converting enzyme (BACE, 44), and thrombin (THR, 188). For the virtual screening experiments single CORINA [Sadowski *et al.*, 1994] 3D conformations were used.

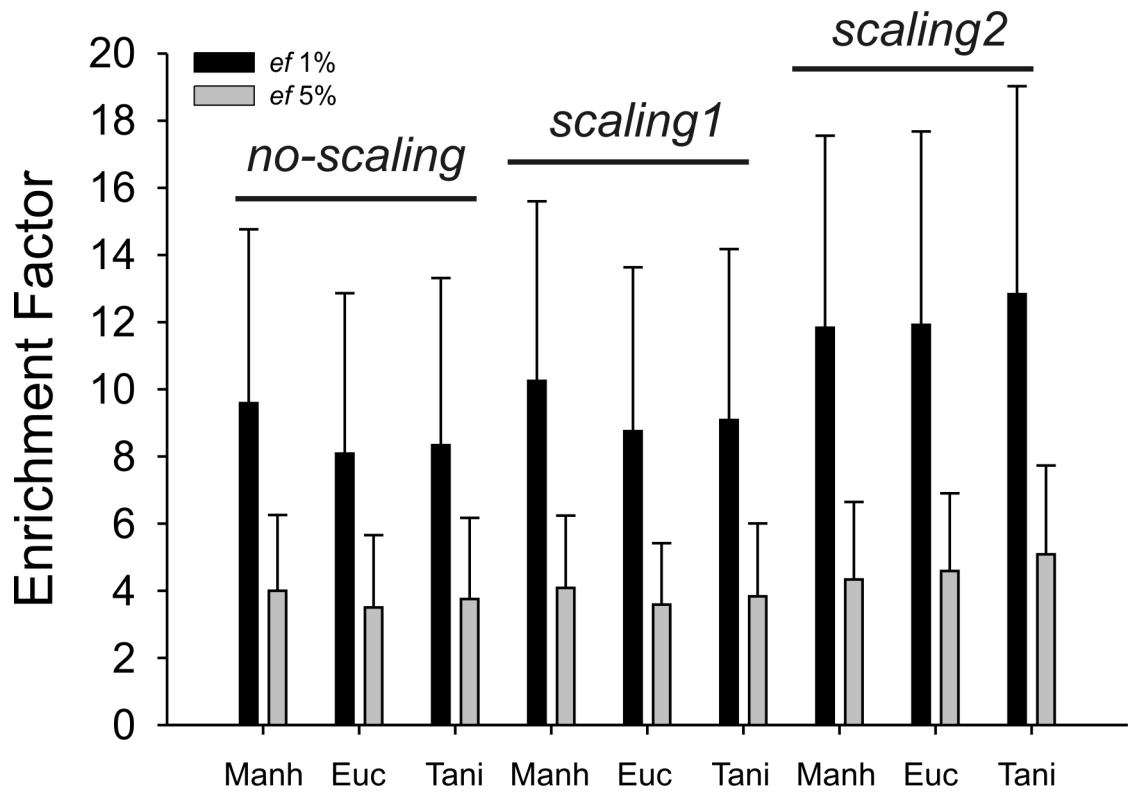


Figure 4.1. Average of the enrichment factors over the 12 activity classes for the different scaling methods and similarity metrics from retrospective screening of the COBRA database. Blue bars denote 1% enrichment and red bars 5% enrichment factors.

Each of the molecules from the subsets was employed as query for one virtual screening experiment. Averages of the enrichment factors over all twelve classes are shown in Figure 4.1 for the first 1% and 5% of the hit-lists. Apparently the performance of the three scaling schemes was *scaling2* > *scaling1* > *no-scaling*, independent of the similarity metric applied. The performance of the similarity metrics showed no clear ranking. For *no-scaling* and *scaling1* the Manhattan distance was found to be best. For *scaling2* the Tanimoto coefficient was the best performing similarity metric. The differences between the similarity metrics were significantly smaller than between the scaling schemes. The standard deviations of the average *ef* values were found to be up to 64 % of the mean values (*ef*(5%) for *no-scaling* with Tanimoto in Figure 4.1). Accordingly the results have to be taken with care. To assess the significance of the average *efs* we further investigated the results of the individual target classes.

The enrichment factors for all classes are given in Table 4.1. For all classes expect GPCR significant *ef* values were found. GPCR is a very general class comprising many different receptors. Thus a lack of significant enrichment is not surprising. NUC, another

general class, was successful in retrospective screening. The major trend found for the average *ef* values was confirmed. Though large standard deviations were also found in the single activity classes the major trend of the average ef values seems to be confirmed. For almost all classes the best *ef* values for 1% of the hit-list were found with *scaling2*. Only for HIVP with Manhattan distance and *no-scaling* or *scaling1*, and for NK with Manhattan distance and the Tanimoto coefficient with *no-scaling* resulted in better or equal *ef* values than for the respective screenings using *scaling2*. For 5% of the hit-list more examples were found with equal or better *ef* values using scaling schemes other than *scaling2*. For the Tanimoto coefficient, this was only found for *no-scaling* NK.

4.1.1 Conclusion

The impact of different scaling schemes (*no-scaling*, *scaling1*, and *scaling2*) and different similarity metrics (Manhattan distance, Euclidean distance, and Tanimoto similarity) on virtual screening with the CATS3D descriptor was investigated with retrospective screening in ten target classes of the COBRA database. The results suggest a general preference for *scaling2* (scaling by the added occurrences of the PPP pairs). *Scaling2* was found to be best for most of the target classes. Accordingly for all further experiments this scaling scheme was applied. Using *scaling2* the Tanimoto coefficient was found to be best. The differences between the similarity metrics were low in comparison to the differences between the scaling schemes. Therefore one could also suggest the use of the Manhattan distance for the screening of large datasets, since the Manhattan distance is the fastest of the three applied similarity metrics.

Table 4.1. Retrospective screening results for CATS3D using different scaling schemes and similarity metrics. Three scaling schemes (*no-scaling*, *scaling1*, and *scaling2*) and the three similarity metrics Manhattan distance (Manh), Euclidean distance (Euc), and the Tanimoto similarity (Tani) were applied. Values in brackets are standard deviations.

	no-scaling			*scaling1*			*scaling2*		
% DB	**Manh**	**Euc**	**Tani**	**Manh**	**Euc**	**Tani**	**Manh**	**Euc**	**Tani**
ACE									
1	12 (11)	12 (10)	12 (10)	14 (11)	12 (10)	13 (10)	16 (11)	17 (11)	20 (13)
5	4 (3)	4 (3)	4 (4)	5 (3)	5 (3)	5 (3)	6 (3)	6 (4)	7 (4)
COX2									
1	17 (11)	16 (11)	17 (12)	18 (11)	18 (12)	19 (12)	21 (12)	20 (13)	22 (13)
5	7 (4)	6 (3)	7 (4)	8 (4)	7 (4)	7 (4)	9 (4)	8 (4)	9 (4)
CRF									
1	19 (11)	15 (10)	15 (9)	21 (11)	15 (9)	15 (8)	22 (10)	21 (10)	20 (11)
5	8 (3)	6 (3)	6 (3)	9 (3)	7 (3)	7 (3)	10 (3)	10 (3)	10 (3)
DPP									
1	13 (8)	9 (6)	9 (7)	13 (9)	11 (8)	11 (8)	16 (11)	16 (12)	15 (13)
5	4 (2)	4 (2)	4 (2)	5 (2)	4 (2)	3 (2)	5 (3)	5 (3)	4 (3)
GPCR									
1	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)
5	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
HIVP									
1	13 (8)	13 (8)	14 (8)	13 (8)	13 (7)	14 (7)	12 (8)	15 (9)	20 (10)
5	7 (4)	7 (4)	8 (4)	6 (3)	7 (3)	8 (3)	5 (3)	7 (4)	9 (4)
MMP									
1	7 (5)	5 (4)	5 (4)	7 (5)	5 (4)	6 (4)	10 (7)	11 (7)	13 (8)
5	3 (2)	3 (2)	3 (2)	3 (2)	3 (2)	3 (2)	4 (3)	4 (2)	5 (3)
NK									
1	13 (8)	11 (7)	12 (7)	12 (8)	11 (7)	12 (8)	11 (7)	12 (8)	15 (8)
5	7 (4)	6 (3)	7 (4)	6 (3)	6 (3)	7 (4)	5 (3)	6 (3)	8 (3)
NUC									
1	7 (6)	6 (5)	7 (5)	7 (6)	7 (5)	7 (5)	8 (6)	7 (5)	8 (6)
5	4 (3)	3 (2)	4 (2)	4 (3)	3 (2)	4 (2)	4 (3)	4 (3)	5 (3)
PPAR									
1	7 (5)	5 (4)	4 (5)	7 (6)	5 (5)	6 (5)	9 (7)	9 (8)	8 (8)
5	3 (2)	2 (2)	2 (2)	3 (2)	2 (2)	2 (2)	3 (2)	3 (2)	3 (3)
BACE									
1	7 (5)	5 (3)	4 (2)	8 (6)	6 (4)	5 (4)	12 (10)	12 (10)	11 (9)
5	2 (2)	2 (1)	2 (1)	3 (2)	2 (1)	2 (1)	3 (2)	3 (3)	4 (3)
THR									
1	6 (4)	5 (3)	5 (3)	6 (4)	6 (4)	6 (4)	7 (5)	8 (5)	9 (5)
5	3 (2)	3 (1)	3 (1)	3 (2)	3 (2)	3 (2)	3 (2)	4 (2)	5 (2)
average									
1%	10.1 (5.1)	8.6 (4.7)	8.8 (5.1)	10.7 (5.4)	9.2 (4.9)	9.6 (5.1)	12.2 (5.8)	12.4 (5.6)	13.5 (6.2)
5%	4.5 (2.3)	4.0 (2.2)	4.3 (2.4)	4.7 (2.3)	4.1 (2.0)	4.3 (2.2)	4.8 (2.5)	5.1 (2.4)	5.7 (2.7)

4.2 Impact of conformational flexibility on CATS3D virtual screening

Virtual screening methods like docking or three-dimensional pharmacophore searching rely on the on the "bioactive" conformation of molecules to asses the biological effect of a molecule. Three-dimensional pharmacophore correlation vector methods have been shown to produce reasonable results using only a small set of conformations or even a single conformation per molecule [Sheridan1 *et al.*, 1996, Brown & Martin, 1996, Section 4.1].

While it is a comparably easy task for small and rigid ligands with only few rotatable bonds to sample the conformational space exhaustively, there are still practical limits in the number of conformations that can be handled efficiently due to the exponential explosion of the number of potential conformations with an increasing number of rotatable bonds [Schwab, 2003]. Accordingly three-dimensional methods which rely only moderately on the presence of an exact fitting conformer would be interesting for virtual screening.

In the present study we examined the influence of the incorporation of different amounts of multiple conformations on the ability of the CATS3D approach to find isofunctional molecules in a retrospective screening experiment. Therefore reference molecules from co-crystal structures were used as queries for the retrospective virtual screening experiments. Different numbers of conformations were calculated for the virtual screening database. We compared the effect of using the different virtual screening libraries.

The PDBbind database [Wang *et al.*, 2004] (version 2002) served as a reference set of high-quality crystal structures of receptor-bound ligands for the virtual screening experiments. For retrospective screening we used the COBRA database [Schneider & Schneider, 2003] (version 3.12) consisting of 5,376 annotated ligands compiled from scientific literature. The ligands of the PDBbind database were grouped according to their target annotation. All clusters containing less then five ligands were removed. Clusters were also removed for which no ligands were found in the COBRA database with the same target annotation as in PDBbind. From multiple incidences of identical ligands all but the one with the best resolution were removed. The final set of reference ligands consisted of 11 groups ("activity classes") with a total number of 177 ligands. The corresponding set of "active" ligands in the COBRA database contained 674 molecules, which means that the COBRA database contained 4,702 additional ligands that were not considered as "active" in either of the 11 activity classes. The final set of annotated activity classes and their abbreviations were: acetylcholinesterase (ACHE, 6 compounds from PDBbind, 13 compounds from COBRA,

overlap: 0), carbonic anhydrase II (CAII, 30, 25, 2), elastase (ELA, 8, 8, 0), factor Xa (FXA, 5, 226, 5), HIV-protease (HIVP, 58, 61, 8), neuraminidase (NEU, 8, 28, 1), protein tyrosine kinase c-src (PTK-CSRC, 7, 16, 0), protein tyrosine phosphatase 1b (PTP1B, 5, 36, 0), stromelysin 1 (STRO1, 7, 19, 0), thrombin (THR, 32, 194, 10), and urokinase type plasminogen activator (UTPA, 11, 48, 3). Since we were not interested in the absolute performance of the method, but in the relative performance using different degrees of conformational information, we did not remove ligands that were present in both databases ("overlap"). An overview over the average number of rotatable bonds and the average molecular weights of the activity classes is given in Table 4.2. Before further procession of the data all molecules were neutralized with a script written in the SVL-language of MOE [Chemical Computing Group].

Table 4.2. Average number of rotatable bonds and molecular weight of the activity classes. Values in brackets are standard deviations.

	PDBbind		COBRA	
Activity class	Rotatable bonds	Molecular weight	Rotatable bonds	Molecular weight
ACHE	6.7 (5.8)	334 (116)	8.2 (4.5)	253 (81)
CAII	7.2 (3.5)	321 (84)	7.3 (3.0)	366 (100)
ELA	16.4 (3.2)	545 (60)	10.9 (4.2)	431 (126)
FXA	11.0 (5.3)	435 (29)	12.0 (5.7)	489 (82)
HIVP	21.7 (9.7)	637 (116)	19.3 (6.2)	614 (116)
NEU	12.9 (1.4)	305 (20)	12.6 (7.7)	320 (130)
PTK-CSRC	24.9 (3.4)	557 (64)	7.6 (3.2)	444 (80)
PTP1B	6.0 (0.0)	277 (34)	10.2 (6.4)	464 (150)
STRO1	12.9 (6.4)	487 (108)	17.1 (5.6)	489 (106)
THR	10.7 (5.0)	423 (125)	15.4 (5.1)	500 (107)
UTPA	6.6 (1.9)	294 (86)	10.2 (5.8)	165 (116)

4.2.1 Calculation of conformations for the PDBbind dataset and the COBRA database

Conformations were calculated for the selected reference molecules from the PDBbind database and all molecules from the COBRA database. For each database single conformations were calculated with CORINA. To restrict the number of possible output conformations from ROTATE only the five most central rotatable bonds were subjected to torsion angle variation, and conformations with an internal (symbolic) energy of more than 100 kJ/mol above the lowest-energy conformation were rejected. The resulting conformations were classified after the calculation in torsion angle space by applying different thresholds to further reduce the number of conformers.

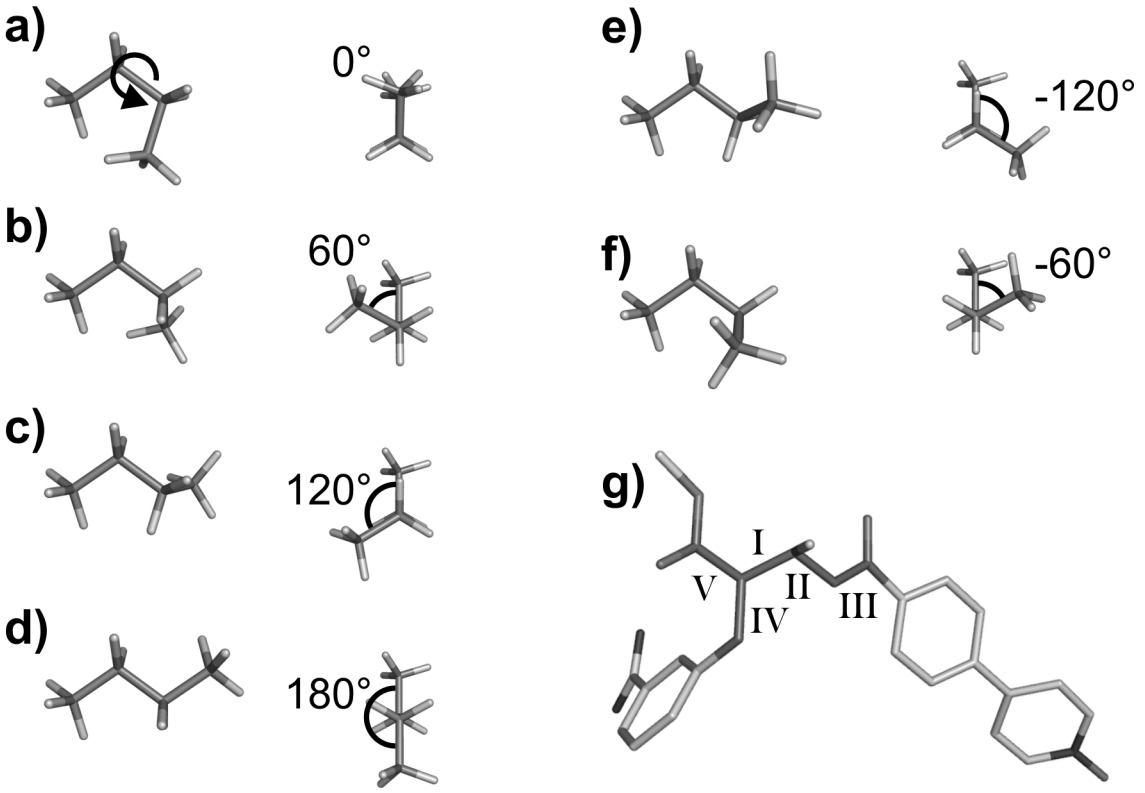


Figure 4.2. The torsion angle. a) – f) Example of the torsion angle variation of the central rotatable bond of butane in steps of 60°. b), d) and f) correspond to minima in the torsion energy; a), c) and f) correspond to unfavorable states with maxima in the torsion angle energy. g) illustrates the five innermost rotatable bonds of the Factor Xa inhibitor Fxv673 (Roman letters at the bonds in dark grey), that were used for the conformation generation with ROTRATE.

Table 4.3. Average number of conformations per molecule calculated for 11 activity classes. Multiple conformations were calculated with the following thresholds for the classification in torsion angle space: R1: 120 °, R2: 60° and R3: 45°.

	PDBbind			COBRA		
Activity class	R1	R2	R3	R1	R2	R3
ACHE	1.7	10.7	19.2	1.5	10.7	20.8
CAII	3.5	18.9	31.3	3.1	22.0	38.9
ELA	3.8	34.9	51.5	2.8	27.0	54.9
FXA	3.8	38.4	69.0	3.9	32.8	59.4
HIVP	2.9	30.1	56.4	2.8	28.9	51.4
NEU	2.0	24.5	45.1	2.4	19.0	35.2
PTK-CSRC	2.6	26.7	48.0	4.0	21.9	41.4
PTP1B	4.0	14.6	25.6	3.2	25.4	46.9
STRO1	3.4	35.4	64.9	3.2	31.0	55.6
THR	3.5	26.5	49.6	3.4	34.3	63.6
UTPA	3.3	11.0	19.2	3.3	22.8	48.0
Average	3.1	24.7	43.6	3.1	25.1	46.9

For the final classification we used torsion angle thresholds of 120° (resulting database further referred to as R1), 60° (R2) and 45° (R3). Table 4.3 gives an overview over the average number of conformations that were calculated per molecule for the different activity classes of both datasets. On average approximately three conformations were generated for each molecule in the R1 datasets, roughly 25 conformations in the R2 and about 45 conformations per molecule in the R3 datasets. For some of the activity classes (e.g. PTP1B, UTPA) the number of conformations differed significantly between the reference dataset and the COBRA database. Since the number of possible conformations is mainly determined by the number bonds which were rotated, this difference indicates that the topological similarity between the entries of these classes was low.

Table 4.4. Average best RMSD of the calculated conformations to the reference conformations of the PDBbind molecules. Improvements are given for usage of multiple conformations in comparison to the RMSD obtained with a single CORINA conformation. The improvement *I* (Rx, x=1,2,3) was calculated by RMSD (CORINA conformation) / RMSD (best Rx conformation). Values in brackets are standard deviations.

	RMSD in Å				Improvement over CORINA		
Activity class	CORINA	R1	R2	R3	I (R1)	I (R2)	I (R3)
ACHE	1.5 (1.8)	1.1 (1.2)	0.7 (0.7)	0.7 (0.8)	1.4	2.1	2.0
CAII	1.1 (0.4)	0.9 (0.4)	0.8 (0.4)	0.8 (0.4)	1.2	1.4	1.5
ELA	2.2 (0.4)	1.6 (0.3)	1.4 (0.4)	1.3 (0.4)	1.4	1.6	1.7
FXA	2.0 (0.4)	1.7 (0.4)	1.0 (0.2)	0.8 (0.2)	1.2	2.1	2.5
HIVP	3.1 (0.8)	2.6 (0.9)	2.2 (0.7)	2.1 (0.7)	1.2	1.4	1.5
NEU	1.0 (0.6)	1.2 (0.4)	0.8 (0.5)	0.8 (0.5)	0.9	1.3	1.3
PTK-CSRC	2.2 (0.5)	2.2 (0.2)	1.8 (0.1)	1.8 (0.1)	1.0	1.2	1.2
PTP1B	0.9 (0.2)	0.5 (0.1)	0.4 (0.1)	0.4 (0.0)	1.6	2.1	2.4
STRO1	2.1 (0.8)	1.5 (0.7)	1.2 (0.6)	1.2 (0.6)	1.4	1.7	1.8
THR	1.9 (0.9)	1.4 (0.8)	1.2 (0.7)	1.1 (0.7)	1.4	1.6	1.7
UTPA	0.9 (0.5)	0.5 (0.4)	0.4 (0.3)	0.4 (0.2)	1.7	2.3	2.4
Average	1.7 (0.7)	1.4 (0.6)	1.1 (0.6)	1.0 (0.5)	1.3 (0.2)	1.7 (0.4)	1.8 (0.5)

4.2.2 Reproducing the crystal-structure conformations of reference ligands

In order to assess the reproduction of the receptor-bound conformations of the PDBbind reference ligands we calculated the RMSD value of all generated conformations to their corresponding experimentally determined geometry. The results of the calculation are shown in Table 4.4. In a recent publication [Boström, 2002] an RMSD of less than 0.5 Å to the reference conformation was considered as a successfully reproduced conformation. According to this threshold, only for two activity classes (PTP1B, UTPA) the bioactive conformation could be reproduced, even with the R3 database containing the largest number of calculated conformations. Applying a less stringent RMSD criterion of 1 Å, the CORINA conformations

already reproduced the bioactive conformation for three of the eleven activity classes (NEU, PTP1B, UTPA). For six activity classes the bioactive conformation could be reproduced in the R3 database. The best RMSD values were found for PTP1B and UTPA, the two classes with the minimum of rotatable bonds of 6 and 6.5 on average (Table 4.2), using the maximum of conformations. Only for two classes RMSD values higher than 1.3 Å were obtained: for HIVP and PTK-CSRC, the two classes with the largest number of rotatable bonds (21.7 for HIVP, 24.9 for PTK-CSRC). Interestingly, the largest improvement using more conformations could be obtained for Factor Xa inhibitors (FXA) which have 11 rotatable bonds on average (Table 4.2). For UTPA and PTP1B the second- and the third-best improvement were found. The smallest improvement was obtained for PTK-CSRC, which is probably caused by the fact that not all rotors were processed for the generation of multiple conformations and these two classes had the most additional bonds that were not rotated.

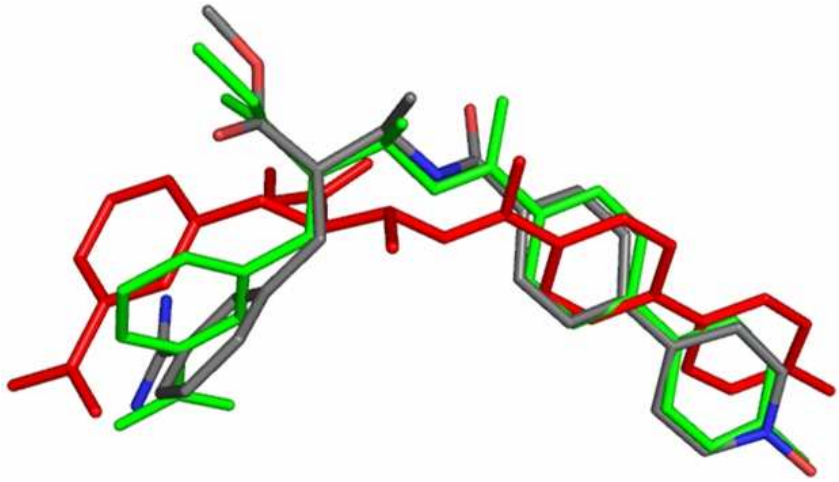


Figure 4.3. Superposition of the CORINA conformation (red) and the best R3 conformation (green) of the Factor Xa inhibitor Fxv673 (PDB code 1KSN) to the reference conformation from the crystal structure.

An example of the Factor Xa inhibitor Fxv673 (PDB code 1KSN) CORINA conformation (red) and the best R3 conformation (green) superimposed onto the reference is shown in Figure 4.3. The bound ligand conformation has a central kink that is not found in the geometrically more stretched CORINA conformation (RMSD = 2.3 Å). The best ROTATE conformation reproduced the kink which resulted in an improved RMSD of 1.1 Å. Summarizing, using more conformations resulted in a lower RMSD, and in most cases conformations were found close to the receptor-bound conformation.

4.2.3 Retrospective screening

In order to determine the impact of multiple conformations on 3D similarity searching the presumable "bioactive" conformations of the reference ligands selected from the PDBbind database were used to screen the COBRA database for ligands with similar biological activity. The results of the retrospective screening experiments are compiled in Table 4.5. Most reference classes were able to significantly enrich the first percent of the ranked database with molecules from the same activity class. Surprisingly, for PTK-CSRC and PTP1B no actives at all were found in the top 1% of the ranked database. In comparison to the other activity classes significant differences in the average number of rotatable bonds and the average molecular weight of the reference molecules and the molecules from the COBRA database can be found (Table 4.2). This indicates that the ligand sets from the two databases differed from each other and were therefore not considered as "similar" by the virtual screening method.

For probing the impact of multiple conformations for similarity searching with CAT3D correlation vectors we were interested in the improvement of using multiple conformations over single conformations and not just in the overall performance of each class. Interestingly, while significant improvement was observed for several of the activity classes, on average no significant improvement in the enrichment factor was observed when multiple conformations were incorporated. The largest improvement was observed for FXA and THR yielding an enrichment factor of 1.8 for R2 and R3, respectively. For the other activity classes much smaller improvements were detected. For ACHE even a significant deterioration was observed. In all cases no large difference in the *ef* between R2 and R3 was observed.

Furthermore, no obvious correlation between the improvement of the RMSD from Table 4.4 and the improvement in similarity searching (Table 4.5) was found. Figure 4.4 shows the plots of the enrichment factors versus the best RMSD values to the receptor-bound (bioactive) conformation found in the various conformational ensembles (single CORINA conformation, R1, R2, and R3) for the different activity classes. For example for UTPA, for which the RMSD could be largely improved for the PDBbind dataset, the usage of multiple conformations for COBRA led only to a small improvement for R3. On the other hand, FXA resulted in the largest improvement in both RMSD and in enrichment. HIVP and STRO1, the two classes with the most rotatable bonds in the COBRA dataset, showed nearly no improvement for R3. In R1 both classes even showed a small deterioration in the *ef*. This is

likely to be due to the rotation of only the five innermost rotatable bonds in the molecules. This limitation seems to prevent the reproduction of the crystal structure, i.e. the presumably "bioactive" conformation. This substantiates the observation in the previous experiment where the two classes with more than 20 rotatable bonds resulted in the two largest RMSD values (HIVP and PTK-CSRC in Table 4.4). Regarding the *ef* values, both HIVP and STRO1 performed well, even with a single conformation. In contrast, THR, which ranked third in the number of rotatable bonds in the COBRA database, improved significantly with more conformations (Table 4.5, Figure 4.3). For ELA with an average number of 11.9 rotatable bonds in the COBRA database (Table 4.2), the enrichment did not increase although the RMSD to the receptor-bound conformation was lowered from 2.2 Å (CORINA single conformation) to 1.3 Å (R3 database).

Table 4.5. Result of the retrospective screening of the COBRA database with the PDBbind reference structures. Enrichment factors were calculated for the first percent of the ranked databases. The improvement *I* (Rx, x=1,2,3) was calculated by *ef* (best Rx conformation) / *ef* (CORINA conformation. Values in brackets are standard deviations.

	Enrichment factor				Improvement over CORINA		
Activity class	CORINA	R1	R2	R3	I (R1)	I (R2)	I (R3)
ACHE	5.1 (4.0)	2.5 (3.9)	1.3 (3.1)	1.3 (3.1)	0.5	0.3	0.3
CAII	3.8 (4.0)	4.5 (3.9)	4.6 (4.0)	4.6 (4.7)	1.2	1.2	1.2
ELA	1.6 (4.4)	1.6 (4.4)	1.6 (4.4)	1.6 (4.4)	1.0	1.0	1.0
FXA	4.8 (2.4)	7.0 (2.3)	8.5 (2.3)	8.7 (2.5)	1.5	1.8	1.8
HIVP	12.2 (11.8)	11.4 (11.2)	13.2 (13.1)	13.3 (13.3)	0.9	1.1	1.1
NEU	22.2 (10.5)	21.3 (11.0)	23.5 (10.0)	22.1 (10.3)	1.0	1.1	1.0
PTK-CSRC	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	-	-	-
PTP1B	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	-	-	-
STRO1	9.0 (8.9)	6.7 (7.2)	8.2 (9.0)	9.0 (10.7)	0.7	0.9	1.0
THR	2.9 (3.7)	4.0 (4.8)	5.2 (5.8)	5.3 (5.8)	1.4	1.8	1.8
UTPA	4.9 (5.8)	6.6 (8.4)	6.2 (8.9)	6.0 (8.9)	1.3	1.3	1.2
Average	6.0 (6.5)	6.0 (6.1)	6.6 (6.9)	6.5 (6.6)	1.1 (0.3)	1.1 (0.4)	1.1 (0.4)

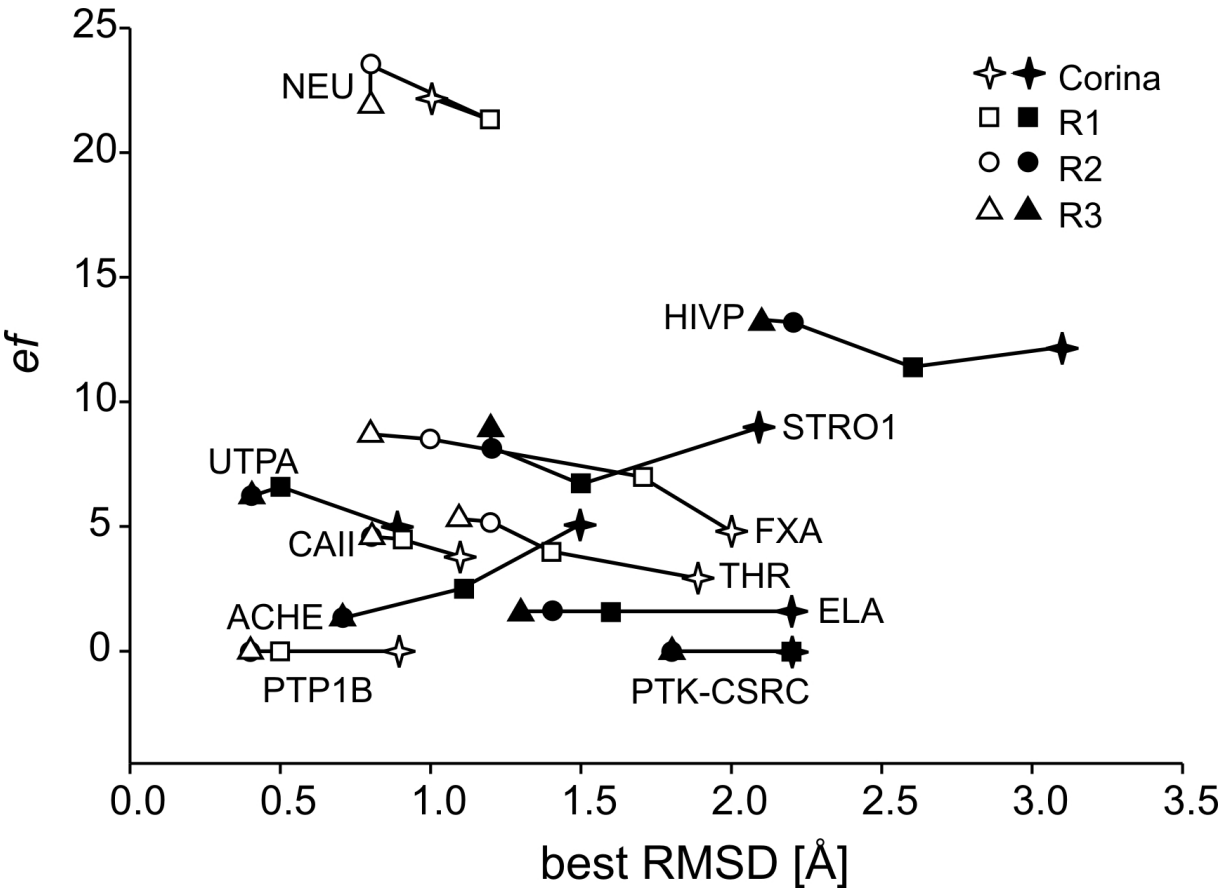


Figure 4.4. Enrichment factors, ef (cf. Table 4.5), versus the best RMSD values, obtained with the four conformational ensembles (CORINA, R1, R2, and R3; cf. Table 4.4) for each of the activity classes. ACHE: acetylcholinesterase, CAII: carbonic anhydrase II, ELA: elastase, FXA: factor Xa, HIVP: HIV-protease, NEU: neuraminidase, PTK-CSRC: protein tyrosine kinase c-src, PTP1B: protein tyrosine phosphatase 1b, STRO: stromelysin 1, THR: thrombin, UTPA: urokinase type plasminogen activator.

To find an explanation for the low impact of multiple conformations on similarity searching, we further investigated the Manhattan distances of the molecules obtained from different conformational samplings to the reference molecules. In Table 4.6 the average Manhattan distances from the best scoring conformations of all active molecules from the COBRA database to the reference molecules are given. Only an average improvement of 1.1 was found using R3 in comparison to the COBRA conformations. For comparison the average Manhattan distances of the 10 best scoring inactives from each virtual screening experiment to the respective reference molecular descriptor are given in Table 4.7. Again an average improvement of 1.1 was found using R3 instead of single COBRA conformations. For ACHE no improvement was found for the active molecules but a small improvement of 1.1 (I(R3))

was found for the inactives. This explains the decreased enrichment factor for ACHE using multiple conformations.

Table 4.6. Average Manhattan distances of the actives from the COBRA database to the reference molecules from the PDBbind database. The improvement *I* (Rx, x=1,2,3) was calculated by average distance (Rx) / average distance (CORINA). Values in brackets are standard deviations.

	Average Manhattan distance to the reference molecules				Improvement over CORINA		
Activity class	CORINA	R1	R2	R3	I (R1)	I (R2)	I (R3)
ACHE	9.5 (3.8)	9.4 (3.8)	9.2 (3.8)	9.2 (3.8)	1.0	1.0	1.0
CAII	10.3 (3.2)	10.0 (3.1)	9.7 (3.1)	9.6 (3.1)	1.0	1.1	1.1
ELA	10.7 (2.4)	10.3 (2.2)	10.0 (2.2)	9.9 (2.2)	1.0	1.1	1.1
FXA	12.4 (3.5)	11.9 (3.5)	11.4 (3.4)	11.3 (3.5)	1.0	1.1	1.1
HIVP	15.3 (4.4)	14.9 (4.4)	14.2 (4.5)	14.1 (4.5)	1.0	1.1	1.1
NEU	9.9 (2.8)	9.6 (2.7)	9.3 (2.7)	9.2 (2.7)	1.0	1.1	1.1
PTK-CSRC	13.2 (2.7)	12.6 (2.6)	12.0 (2.5)	11.9 (2.5)	1.0	1.1	1.1
PTP1B	14.4 (3.9)	13.7 (4.0)	13.1 (4.1)	13.0 (4.1)	1.1	1.1	1.1
STRO1	19.8 (4.8)	19.0 (4.8)	18.7 (4.8)	18.7 (4.8)	1.0	1.1	1.1
THR	12.7 (3.5)	12.3 (3.3)	12.0 (3.2)	12.0 (3.1)	1.0	1.1	1.1
UTPA	11.1 (4.0)	10.8 (3.9)	10.5 (3.8)	10.5 (3.8)	1.0	1.1	1.1
Average	12.7 (3.0)	12.2 (2.8)	11.8 (2.8)	11.8 (2.8)	1.0 (0.0)	1.1 (0.0)	1.1 (0.0)

4.2.4 Conclusion

Investigating the impact of multiple conformations on 3D similarity searching with CATS3D, it was demonstrated that using only a single conformation per molecule already resulted in significant enrichment of actives. This observation was also made for ligand classes with many rotatable bonds. On average these results did not significantly improve using multiple conformations. Nevertheless, for some classes of molecules considerable improvement in the enrichment of active molecules was observed. Furthermore, no clear correlation between the

improvement of the enrichment factor and the improvement of the RMSD to the bioactive conformation could be derived when screening with CATS3D correlation vectors. It was found that the Manhattan distance of single conformations did not change significantly using more conformations. This was observed for molecules with the same activity as well as for false-positives. Since the calculation of multiple conformations is computationally expensive it seems to be preferential to use single conformations for large databases, e.g. virtual combinatorial libraries, for a first-pass virtual screen. Single conformations can be computed efficiently with CORINA, even for large databases. In a second screening round, e.g. with smaller databases or flexible ligands, it can be worthwhile considering multiple conformations.

Table 4.7. Average Manhattan distances of the best inactives from the COBRA database to the references molecules from the PDBbind database. The 10 best scoring inactive molecules from each retrospective screening experiment were used as inactives. The improvement *I* (Rx, x=1,2,3) was calculated by average distance (Rx) / average distance (CORINA). Values in brackets are standard deviations.

	Average Manhattan distance to the reference molecules				Improvement over CORINA		
Activity class	CORINA	R1	R2	R3	I (R1)	I (R2)	I (R3)
ACHE	5.7 (2.7)	5.3 (2.6)	5.1 (2.6)	5.0 (2.6)	1.1	1.1	1.1
CAII	5.5 (1.9)	5.3 (1.9)	5.1 (1.9)	5.1 (1.9)	1.0	1.1	1.1
ELA	6.2 (1.5)	5.8 (1.4)	5.4 (1.4)	5.3 (1.4)	1.1	1.1	1.2
FXA	7.8 (2.2)	7.6 (2.1)	7.3 (2.1)	7.2 (2.2)	1.0	1.1	1.1
HIVP	10.5 (3.4)	10.1 (3.3)	9.6 (3.2)	9.5 (3.2)	1.0	1.1	1.1
NEU	7.2 (1.1)	7.0 (1.0)	6.5 (1.1)	6.4 (1.1)	1.0	1.1	1.1
PTK-CSRC	7.9 (2.0)	7.5 (2.1)	7.1 (1.9)	7.0 (1.9)	1.1	1.1	1.1
PTP1B	7.5 (2.6)	7.2 (2.6)	6.8 (2.5)	6.7 (2.4)	1.0	1.1	1.1
STRO1	14.8 (4.0)	14.4 (3.9)	13.8 (4.0)	13.7 (3.9)	1.0	1.1	1.1
THR	6.3 (0.6)	6.2 (0.6)	6.0 (0.6)	5.9 (0.6)	1.0	1.1	1.1
UTPA	5.1 (1.5)	4.9 (1.4)	4.7 (1.4)	4.7 (1.4)	1.0	1.1	1.1
Average	7.7 (2.8)	7.4 (2.7)	7.0 (2.6)	7.0 (2.6)	1.0 (0.0)	1.1 (0.0)	1.1 (0.0)

4.3 Virtual screening and scaffold hopping efficiency of alignment-free pharmacophore pair descriptors

Manipulating living systems at the molecular level requires profound knowledge of the variability of small molecule effectors that provoke a particular cellular response. Medicinal chemistry relies on libraries of molecular probes that can be rationally designed to contain a desired degree of chemotype diversity. Despite great advances in the field of virtual screening and rational compound library design, “scaffold-hopping” remains a challenging goal [Schneider & Fechner, 2005]. The concept of scaffold-hopping aims at finding isofunctional but structurally dissimilar molecular entities [Schneider *et al.*, 1999, Schneider *et al*., 2000; Böhm *et al.*, 2004, Jenkins *et al.*, 2004]. Ideal screening methods that perform successful scaffold-hops would not only find a maximum number but also a maximally diverse set of active compounds from a given chemical subspace. Only until recently, the focus in the development and evaluation of virtual screening methods has often been purely on the retrieval of large numbers of “active” molecules -- irrespective of the number of retrieved chemotypes. This has led to the impression that methods employing a low level of abstraction from the molecular structure, e.g. substructure fingerprints, are among the most efficient ligand-based virtual screening methods [Brown & Martin, 1996; Hert *et al.*, 2004b]. In contrast to substructure-based molecular descriptors, pharmacophore models and physicochemical metrics represent a comparably high level of abstraction from chemical structure. Consequently, such methods have been employed for screening library design relying on their scaffold-hopping potential [Schneider *et al.*, 1999; Schneider *et al*., 2000; Matter, 1997; Nærum *et al.*, 2002]. In this study we compared the scaffold-hopping efficiency of similarity searching with topological, three-dimensional and molecular surface-based pharmacophore pair descriptors and a substructure fingerprint method.

Similarity searching is founded on the similarity principle which states that similar molecules exhibit similar biological effects [Johnson & Maggiora, 1990]. A straightforward approach for similarity searching is to compare the connection tables to assess the similarity between two molecules. Such methods include substructure fingerprints like the MACCS keys [MDL Information Systems] which are based on exact chemical substructures. Substructure matching approaches were reported to be among the most successful for virtual screening [Brown & Martin, 1996; Hert *et al.*, 2004b]. The classification of intermolecular interactions into general pharmacophore types provides a means to obtain a more general description of the underlying chemotypes of molecules [Schneider *et al.*, 2000; Mason &

Good, 2001]. Three such descriptors were employed in this work: the topological CATS descriptor [Schneider *et al*. 1999; Fechner *et al.*, 2003], the three-dimensional CATS3D descriptor, and the molecular surface based SURFCATS descriptor.

Molecular representations that are grounded on three-dimensional conformations like molecular surface-based descriptors are independent from the molecular connectivity and should have a favorable scaffold-hopping potential [Bender & Glen, 2004; Clark, 2004]. For comparison with a conceptually different descriptor the MACCS keys were used as implemented in MOE.

To assess the degree of scaffold-hopping, one must define the term “scaffold”. Here, we followed the concept of Xu and Johnson employing the software suite Meqi [Pannanugget Consulting L. L. C.], which has recently been devised for the analysis of chemical libraries [Xu & Johnson, 2001; Xu & Johnson, 2002]. Two different definitions of a scaffold were applied: cyclic system (“Scaffold”, Sc) and reduced cyclic system (“Reduced Scaffold, ReSc) (Figure 2.6). In Meqi, each molecular topology is specified by a particular *m*olecular *eq*uivalence *i*ndex (meqi) which is used to distinguish between different scaffolds and reduced scaffolds.

Ligands from ten different target classes from the COBRA database [Schneider & Schneider, 2003] of annotated ligands were used as reference for retrospective virtual screening: angiotensin converting enzyme (ACE, 44 compounds, 28 scaffolds, 17 reduced scaffolds), cyclooxygenase 2 (COX2, 94, 27, 14), corticotropin releasing factor (CRF antagonists, 63, 33, 23), dipeptidyl-peptidase IV (DPP, 25, 13, 7), human immunodeficiency virus protease (HIVP, 58, 46, 31), matrix metalloproteinase (MMP, 77, 47, 19), neurokinin receptors (NK, 118, 65, 49), peroxisome proliferator-activated receptor (PPAR, 35, 29, 17), beta-amyloid converting enzyme (BACE, 44, 13, 12), and thrombin (THR, 188, 100, 36). According to the number of scaffolds and reduced scaffolds in relation to the number of molecules the datasets range from sets with a low scaffold diversity (e.g. COX2) to sets with a large relative scaffold diversity (e.g. PPAR, HIVP). The complete COBRA database contained 1,628 different scaffolds and 637 distinct reduced scaffolds. For retrospective screening each molecule from each target class was taken iteratively as the reference molecule for a virtual screening experiment, where all other molecules were ranked according to their similarity to the reference molecule. For quantification of “similarity” three similarity indices were employed: Manhattan distance, Euclidean distance, and Tanimoto similarity (Table 2.1).

To summarize: For the retrospective screening experiments we employed ten different datasets, four descriptors (CATS, CATS3D, SURFCATS, MACCS), and three molecular representations (atomic, scaffold, and reduced scaffold representation).

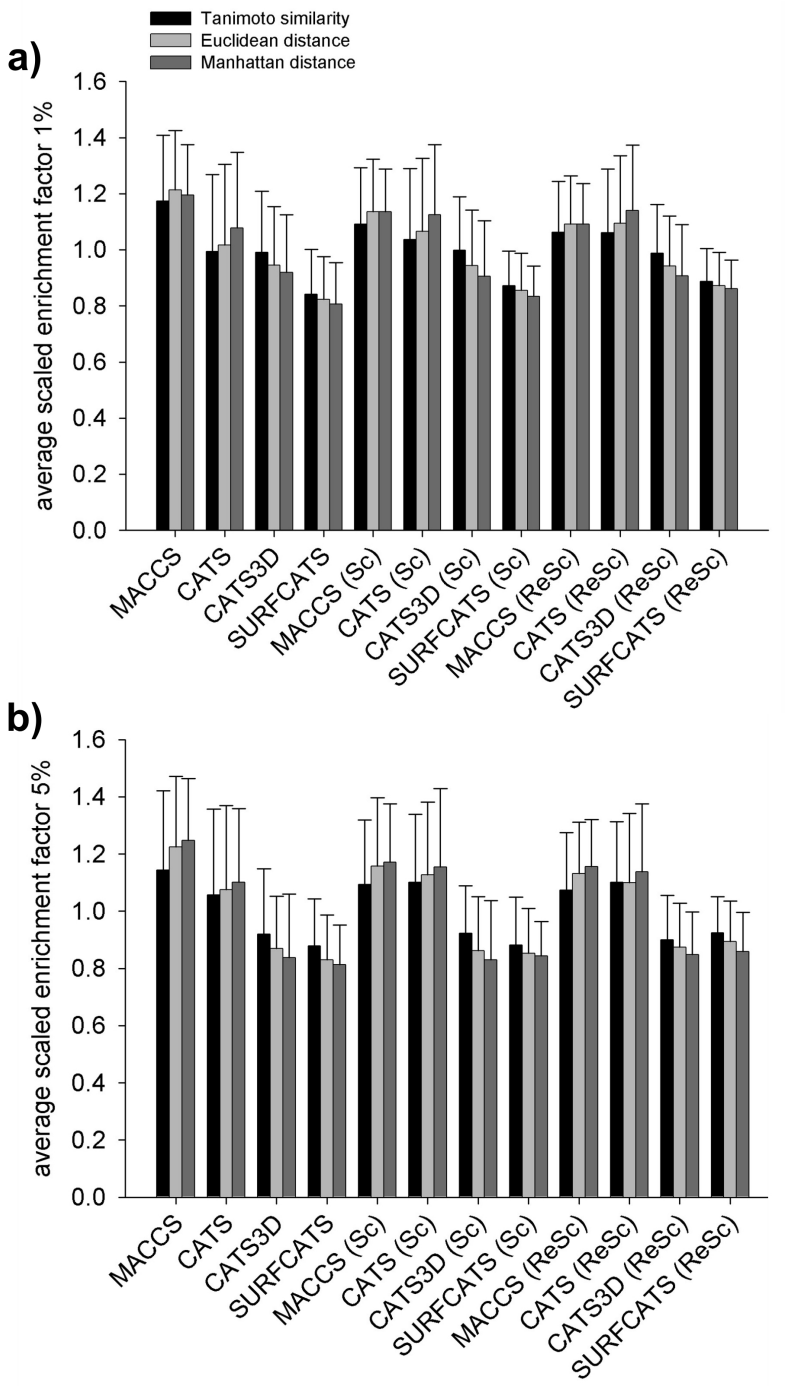


Figure 4.5. Average relative performance for the first 5% over 10 ligand classes from the COBRA database. Comparison of the performance of MACCS, CATS, CATS3D and SURFCATS for molecules, scaffolds (Sc) and reduced scaffolds (ReSc). Three similarity metrics were applied: the Tanimoto similarity, the Euclidean distance and the Manhattan distance.

The average relative performance of the four methods for the first 5% of the database over the ten activity classes is summarized in Figure 4.5. The relative performance of one particular method within one activity class was defined as the *ef* yielded with this method divided by the average *ef* of the four methods (using the same similarity index). The influence of different similarity indices on the overall enrichment was low, for most parts indistinguishable within the standard deviations. For all molecular representations the order of the methods in terms of the enrichment factors for the top 5% of the hit-lists was found to be MACCS > CATS > CATS3D = SURFCATS when looking at the average values only. With regard to the enrichment of scaffolds and reduced scaffolds CATS, CATS3D and SURFCATS slightly improved in comparison to the MACCS keys.

Table 4.8. Enrichment factors of different molecular representations ("Molecules", "Scaffolds", "Reduced Scaffolds") over the activity classes. *ef* values are given for the first 1% and 5% of the hit-lists. The Tanimoto coefficient was used to rank the molecules.

	Molecules			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE				
1	22 (11)	23 (13)	20 (13)	21 (15)
5	9 (4)	11 (5)	7 (4)	8 (4)
COX2				
1	27 (17)	14 (9)	22 (13)	19 (11)
5	11 (6)	5 (3)	9 (4)	8 (4)
CRF				
1	28 (15)	13 (8)	20 (11)	16 (10)
5	12 (4)	7 (3)	10 (3)	9 (3)
DPP				
1	21 (14)	12 (9)	15 (13)	13 (10)
5	6 (4)	4 (4)	4 (3)	3 (2)
HIVP				
1	14 (7)	24 (11)	19 (10)	20 (11)
5	6 (2)	11 (3)	9 (4)	9 (4)
MMP				
1	13 (9)	12 (7)	13 (8)	12 (9)
5	5 (3)	5 (2)	5 (3)	5 (3)
NK				
1	9 (6)	8 (4)	15 (8)	9 (6)
5	5 (2)	5 (2)	7 (3)	5 (3)
PPAR				
1	17 (17)	17 (12)	8 (8)	10 (8)
5	5 (4)	6 (3)	3 (3)	4 (2)
BACE				
1	13 (10)	12 (10)	11 (9)	6 (5)
5	6 (4)	4 (3)	4 (3)	3 (2)
THR				
1	12 (6)	14 (7)	9 (5)	8 (5)
5	6 (2)	9 (4)	5 (2)	5 (3)

((continued below))

	Scaffolds				Reduced Scaffolds			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE								
1	22 (11)	27 (12)	21 (11)	20 (10)	25 (12)	29 (12)	23 (10)	22 (8)
5	9 (3)	12 (3)	7 (2)	8 (3)	9 (3)	11 (3)	8 (3)	8 (2)
COX2								
1	27 (15)	16 (8)	23 (9)	21 (10)	33 (15)	20 (10)	25 (10)	26 (11)
5	10 (4)	5 (2)	8 (3)	8 (3)	10 (3)	6 (2)	8 (2)	9 (3)
CRF								
1	24 (12)	16 (10)	22 (12)	17 (10)	28 (13)	19 (11)	23 (11)	19 (12)
5	11 (4)	7 (4)	10 (3)	9 (3)	11 (4)	7 (3)	9 (3)	8 (3)
DPP-IV								
1	21 (13)	12 (8)	16 (12)	14 (12)	24 (12)	18 (11)	23 (15)	20 (16)
5	7 (4)	6 (5)	4 (3)	4 (2)	9 (5)	7 (4)	6 (4)	5 (3)
HIVP								
1	15 (8)	26 (13)	22 (12)	23 (13)	21 (11)	34 (15)	28 (15)	31 (16)
5	6 (3)	11 (4)	10 (4)	10 (4)	8 (3)	13 (4)	11 (4)	11 (4)
MMP								
1	17 (11)	15 (9)	16 (11)	16 (12)	24 (14)	24 (12)	24 (13)	23 (13)
5	6 (3)	6 (3)	6 (3)	7 (4)	8 (3)	8 (3)	8 (3)	8 (4)
NK								
1	10 (6)	9 (4)	16 (8)	10 (6)	12 (6)	11 (5)	16 (8)	11 (7)
5	5 (2)	5 (2)	8 (3)	6 (3)	6 (2)	6 (2)	8 (3)	6 (3)
PPAR								
1	16 (14)	17 (11)	8 (8)	10 (8)	19 (16)	23 (14)	10 (9)	14 (10)
5	5 (3)	6 (3)	3 (2)	4 (2)	7 (4)	8 (3)	4 (2)	5 (2)
BACE								
1	14 (9)	14 (10)	12 (7)	8 (5)	15 (9)	16 (11)	13 (8)	9 (6)
5	4 (3)	4 (2)	4 (2)	3 (2)	4 (3)	5 (3)	4 (2)	3 (2)
THR								
1	15 (6)	19 (9)	12 (7)	11 (7)	19 (8)	28 (12)	19 (9)	18 (9)
5	7 (2)	10 (4)	6 (3)	6 (3)	8 (3)	11 (4)	8 (3)	7 (3)

An explanation for the high performance of the MACCS keys in scaffold enrichment might be that the connectivity of the substructures is not accounted for by this descriptor. This can lead to an effective retrieval of molecules with slightly different scaffolds but similar side-chain decoration. Does this finding justify the conclusion that substructure fingerprints are best-suited for the purpose of scaffold-hopping? To find an answer to this question, a more detailed analysis was performed looking on the enrichment of the individual activity classes. We calculated *ef* values for all ten different classes, yielded with the Tanimoto coefficient (Table 4.8; results for the Manhattan distance and the Euclidean distance can be found in Appendix 6.1). None of the descriptors performed generally superior to the other descriptors within the error bars. Judging from the average values only, MACCS performed best for COX2, CRF, and DPP for full molecules, scaffolds and reduced scaffolds. CATS performed best for ACE, HIVP and THR, and CATS3D for NK in all molecular representations. SURFCATS was not found to be best for any one class. However, each descriptor of the CATS family was found to be better than the other family members for some ligand classes. This underlines the dependence of the descriptor performance on the screening

database. In other words, there is no *globally* best descriptor. It has to be stressed that this interpretation has limited relevance due to the large standard deviations and represents trends only. Further investigations with additional descriptors and metrics, and larger high-quality drug databases will be needed to scrutinize these findings.

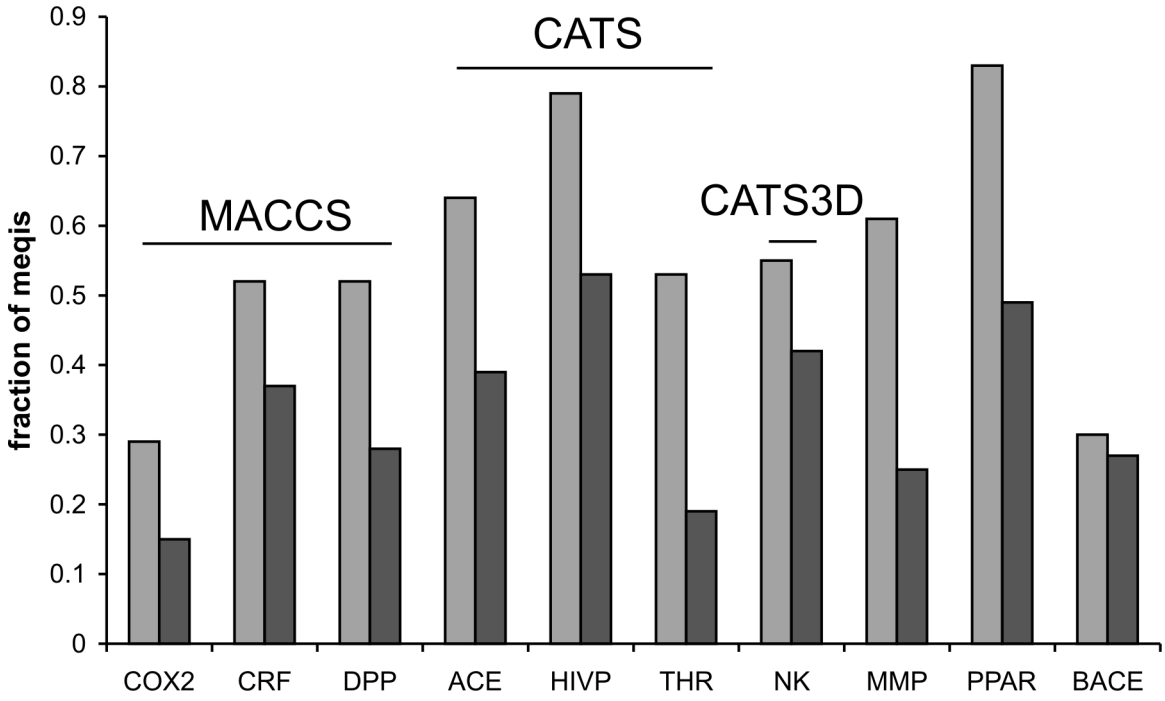


Figure 4.6. Scaffold diversity of the ligand classes. The diversity is given by the number of scaffolds (light gray) or reduced scaffolds (dark gray) relative to the number of molecules in a data set. With enrichment factors for the first 5% MACCS performed best fort he classes COX2, CRF and DPP, CATS performed best for the classes ACE, HIVP and THR and CATS3D performed best for NK.

Figure 4.6 shows the fraction of scaffolds and reduced scaffolds found in the ten ligand classes. For the classes preferred by MACCS the average fraction of scaffolds was 0.44 (± 0.13) and the average fraction of reduced scaffolds was 0.27 (± 0.11). For CATS the fractions were 0.65 (± 0.13) and 0.37 (± 0.17), and for CATS3D 0.55 and 0.42, respectively. One might speculate that MACCS performed best in classes with low numbers of different topologies, i.e. low scaffold diversity. CATS and CATS3D performed best in classes revealing a high degree of scaffold diversity. We conclude that pharmacophore descriptors might be more suited for designing diverse compound libraries compared to substructure fingerprints. Still one must be aware that these results are comparable within the error margin.

In an earlier publication we reported that different descriptors are often found to retrieve different molecules, despite having equal enrichment factors [Fechner *et al.*, 2003]. In the present study we witness a similar situation: descriptors complement each other in the retrieval of different scaffolds and reduced scaffolds (Table 4.9).

Table 4.9. Overlap of the results for pairs of descriptors in the first 5% of the hit-list. Shown are the average numbers over all ten classes of retrieved scaffold representations which were found by both methods. The numbers on the diagonal (shown in bold) are the average numbers of scaffolds found with the respective descriptor. The employed similarity index was the Tanimoto coefficient.

Scaffold representations

Descriptor	MACCS	CATS	CATS3D	SURFCATS
MACCS	**13.8**			
CATS	8.6	**15.4**		
CATS3D	8.2	9.3	**13.2**	
SURFCATS	7.7	8.9	9.8	**12.9**

Reduced Scaffold representations

Descriptor	MACCS	CATS	CATS3D	SURFCATS
MACCS	**8.9**			
CATS	6.1	**9.8**		
CATS3D	5.8	6.5	**8.7**	
SURFCATS	5.3	6.1	6.5	**8.1**

Two of the virtual hit-lists were further investigated: the results for the COX-2 inhibitors celecoxib (Figure 4.7) and rofecoxib (Figure 4.8). For each scaffold class, the best-ranking hits were surveyed. Although the two reference molecules share a common reduced scaffold different scaffold classes were retrieved on different ranking positions. Again, the four similarity searching methods differed in their ability to retrieve diverse scaffold which results in a complementation of the methods. This outcome is remarkable especially because of the striking relatedness of the query structures.

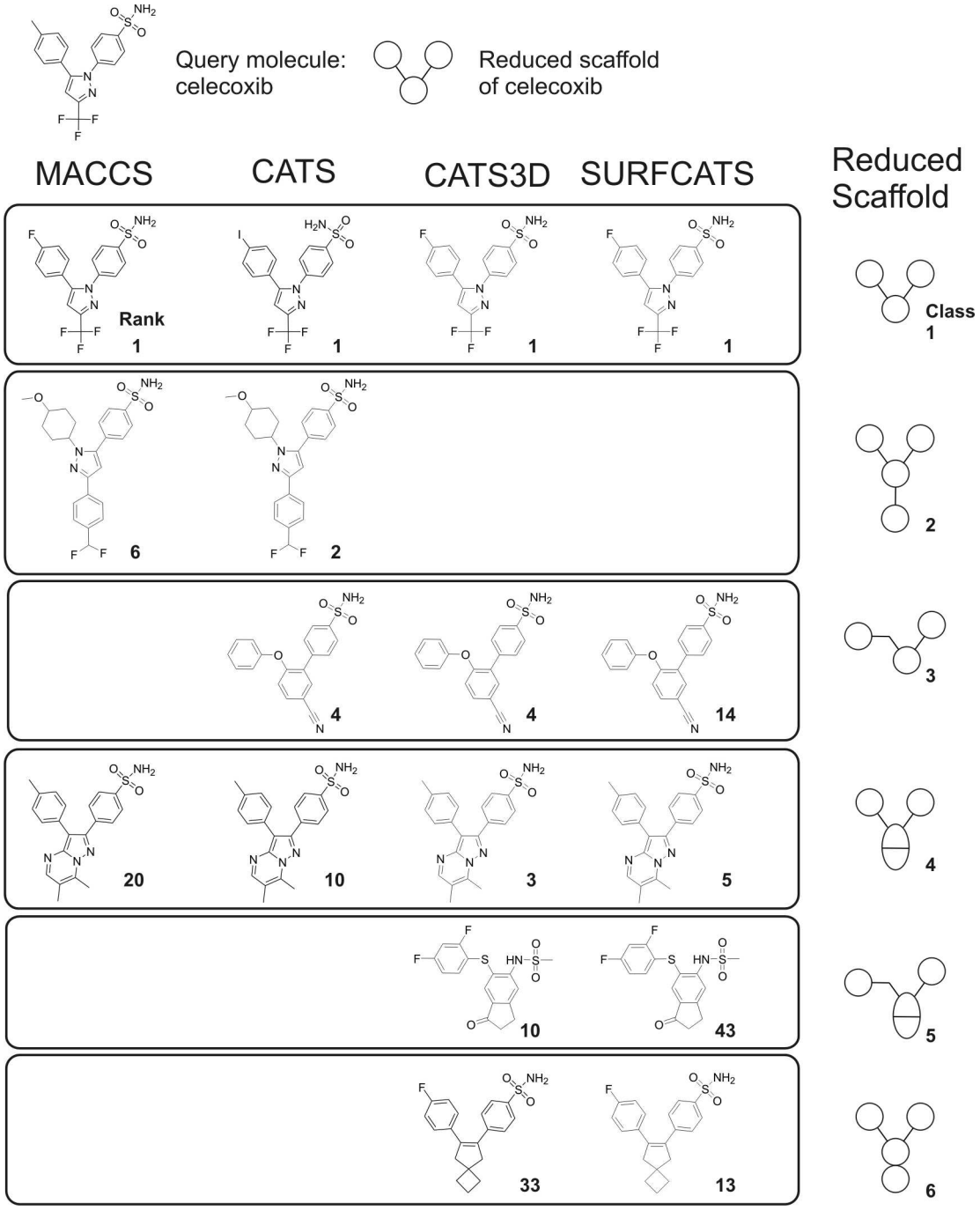


Figure 4.7. Best hits for each reduced scaffold obtained with celecoxib. For each descriptor the best scored molecule in each reduced scaffold class is shown that was retrieved in the first 1 % of the ranked database.

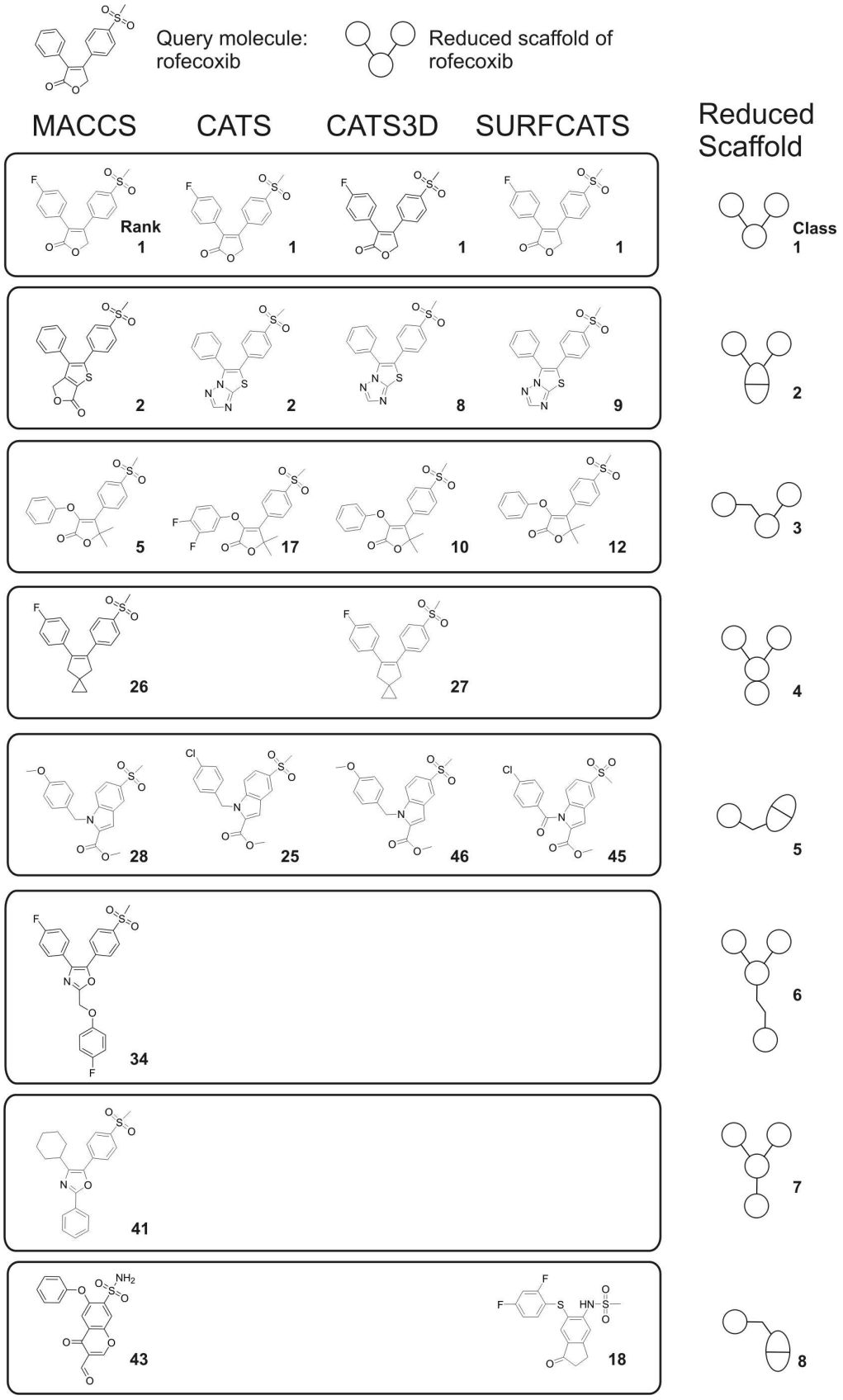


Figure 4.8. Best hits for each reduced scaffold obtained with rofecoxib. For each descriptor the best scored molecule in each reduced scaffold class is shown that was retrieved in the first 1 % of the ranked database.

The two reduced scaffolds that were found exclusively with the MACCS keys for rofecoxib (ReSc classes 6 and 7) reflect that MACCS keys include no direct information of the size of the retrieved molecules. These molecules might have been rejected by the other methods due to their large size. Large reduced scaffolds were also found with CATS for celecoxib (ReSc class 2), which might have resulted from the restriction of the descriptor to a maximal path length of 10 bonds. Such a cut-off might be inappropriate for a database with potentially long ligands and respective pharmacophores, such as those annotated to HIVP, MMP, and PPAR – particularly in prospective screens

4.3.1 Conclusion

Concluding, we found that both substructure fingerprints (MACCS) and pharmacophore-pair descriptors (CATS) are suited for retrospective scaffold retrieval. For more diverse ligand classes the pharmacophore-based CATS descriptors slightly outperformed substructure (MACCS) keys as an average trend. The fact that structurally focused collections of pharmacologically active compounds are typically employed for retrospective screening studies might explain the often found high performance of substructure keys or related descriptors. Our results suggest that for the particular purpose of scaffold-hopping a reasonable strategy might be to use more generalizing molecular representations like pharmacophore descriptors. The use of several complementing methods can also be recommended for the purpose of scaffold hopping. We hope that our study will stimulate further investigations on this important topic of medicinal chemistry.

4.4 Prospective screening for mGlurR5 allosteric modulators with CATS3D

Allosteric modulators for the metabotropic glutamate receptor 5 are a promising class of molecules for addressing several disorders of the central nervous system [Hermans & Challiss, 2001]. Being part of the pharmaceutical interesting class of GPCRs, for which rare receptor structure information is available, mGluR5 is an ideal target to test ligand based virtual screening approaches.

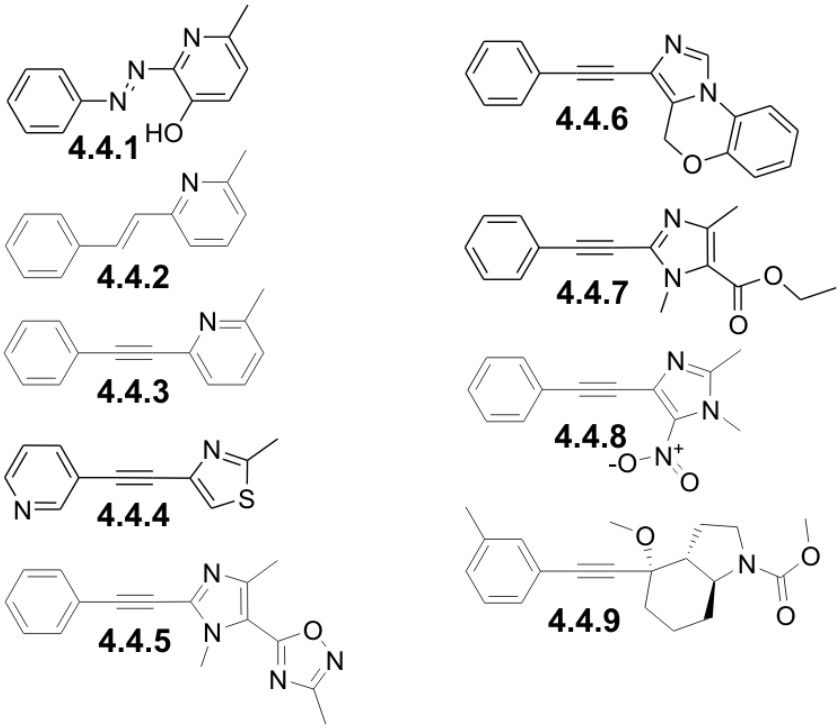


Figure 4.9. Allosteric mGluR5 antagonists.

First selective allosteric antagonists for mGluR5, SIB-1751 (**4.4.1**) and SIB-1893 (**4.4.2**), were published in 1999 [Varney *et al.*, 1999]. SIB-1751 was identified by high-throughput screening (HTS), and SIB-1893 resulted from a UNITY search for analogues [Varney *et al.*, 1999]. In phosphoinositol (PI) hydrolysis assays the two molecules revealed IC_{50} values of 3.1 µM and 2.3 µM, respectively. Chemical variation of SIB-1893 resulted in the much more potent highly selective mGluR5 antagonist 2-methyl-6-(phenylethynyl)-pyridine (MPEP, **4.4.3**, Figure 4.9) with an IC_{50} of 36 nM in PI hydrolysis assays [Gasparini *et al.*, 1999]. Several MPEP-analogues (**4.4.4**–**4.4.9**, Figure 4.9) [**4:** Cosford *et al.*, 2003; **5**-**8**:

Mutel *et al.*, 2002; **9**: Gasparini *et al.*, 2003] with reported low nanomolar activity have been published in the scientific and patent literature since then. Nonetheless, the mode of action of these ligands is not completely understood. Recent publications of MPEP and MPEP derivatives also reported off-target activity [**4:** Cosford *et al.*, 2003; **5**-**8**: Mutel *et al.*, 2002; **9**: Gasparini *et al.*, 2003] and a short plasma half life [Anderson *et al.*, 2003]. In particular, the latter could be attributed to potential metabolic instability of the ethynyl linker.

Pharmacophore-based similarity searching has been proven to be suited for finding new ligands which exhibit similar biological activity but are based on a different chemical scaffold [Böhm & Schneider, 2000]. Using a set of known specific allosteric antagonists of mGluR5 (**4.4.3**-**4.4.9**) [Gasparini *et al.*, 1999; Cosford *et al.*, 2003; Mutel *et al.*, 2002; Gasparini *et al.*, 2003], which were compiled from scientific and patent literature, as a query we applied a hierarchical, ligand-based virtual screening approach to identify novel compounds accomplishing mGluR5 modulation. First, a "drug-likeness" estimation by an artificial neural network system was employed for prescreening to focus only on molecules with a predicted "drug-like" structure [Schneider & Schneider, 2004]. For subsequent similarity searching we used the CATS3D descriptor.

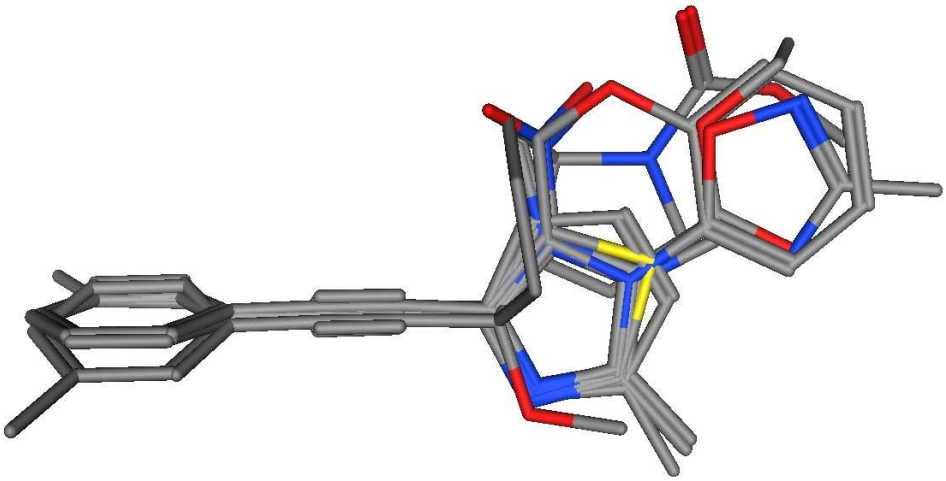


Figure 4.10. Flexible pharmacophore-based alignment of reference molecules **4.4.3**-**4.4.9**. Red: oxygen; blue: nitrogen; yellow: sulfur; gray: carbon.

To form a hypothesis about receptor-bound 3D-conformations of **4.4.3**-**4.4.9** we used the flexible alignment tool of MOE (Figure 4.10). Ligands were successively aligned from **4.4.3** to **4.4.9** and conformations were chosen based on existing knowledge among the best ranked results. Molecule **4.4.9** was manually adjusted to fit to the alignment, i.e. the angle

between the two planes of the ring systems was set to 90° (Merz, unpublished results). These individual 3D conformations served as query structures for CATS3D similarity searching.

In search for new ligands we virtually screened the Asinex Gold compound collection, version of April 2003 [ASINEX], which contained 194,563 molecules. As a pre-screening filter we selected the 20,000 most "drug-like" compounds as described previously [Schneider & Schneider, 2004]. The result of this procedure can be seen, e.g., for the neural network prediction: the average drug-likeness score of the complete Asinex Gold collection according to the artificial neural network was 0.36 (σ = 0.28), for our screening set the score was 0.60 (σ = 0.23) (higher values indicate more "drug-like" molecules).

3D-conformations of the screening compounds were calculated in MOE using the MMFF94 force field. The results were restricted to a maximum of 20 lowest energy conformations per molecule. Similarity between a database entry and a reference molecule was expressed by the Manhattan distance. Separate similarity searches were performed with each of the molecules **4.4.3**-**4.4.9**, and 27 of the top-scoring molecules (Figure 4.11) were selected for experimental testing. Molecules were chosen which had low Manhattan distances to one of the reference molecules and which were not too similar to the previously selected molecules, as judged by visual inspection from a medicinal chemistry perspective (Table 4.10).

Table 4.10. Results of virtual screening and the binding assays

Molecule no.	Most similar reference molecule	Rank (CATS3D)	Virtual Screening			Binding Assay		
			CATS3D Manhattan distance	CATS2D Manhattan distance	MACCS Tanimoto similarity	K_i mGluR5 (µM)	K_i mGluR1 (µM)	Selectivity (Ki mGluR1 / Ki mGluR5)
10	3	1	0.68	2.85	0.21	24	> 100	> 4.2
11	3	4	0.88	2.2	0.2	> 100	63	< 0.6
12	3	5	0.94	5.03	0.17	> 100	41	< 0.4
13	3	6	0.95	3.79	0.22	> 100	> 100	1
14	3	7	1.02	2.64	0.17	> 100	> 100	1
15	3	17	1.12	3.06	0.24	> 100	> 100	1
16	4	1	1.52	2.54	0.35	> 100	> 100	1
17	4	3	1.67	5.27	0.22	> 100	> 100	1
18	4	4	1.67	2.34	0.34	> 100	> 100	1
19	4	6	1.73	1.88	0.25	> 100	> 100	1
20	5	3	2.14	1.79	0.42	> 100	> 100	1
21	5	7	2.22	1.79	0.36	> 100	> 100	1
22	5	38	2.52	2.66	0.38	41	64	1.6
23	6	5	1.41	2.23	0.48	33	61	1.8
24	6	6	1.45	1.91	0.31	12	17	1.5
25	7	2	1.55	2.69	0.38	35	> 100	> 2.9
26	7	3	1.56	2.41	0.39	> 100	> 100	1
27	7	5	1.6	2.62	0.53	> 100	14	< 0.14
28	8	2	0.79	5.49	0.38	> 100	> 100	1
29	8	7	0.91	5.37	0.24	> 100	> 100	1
30	8	9	1	5.37	0.31	40	> 100	2.54
31	8	12	1.14	4.81	0.28	> 100	> 100	1
32	8	36	1.3	5.33	0.2	14	45	3.2
33	9	1	1.49	2.19	0.46	63	> 100	> 1.6
34	9	2	1.54	1.94	0.45	38	> 100	> 2.7
35	9	5	1.59	2.59	0.46	> 100	> 100	1
36	9	7	1.64	6.63	0.46	> 100	> 100	1

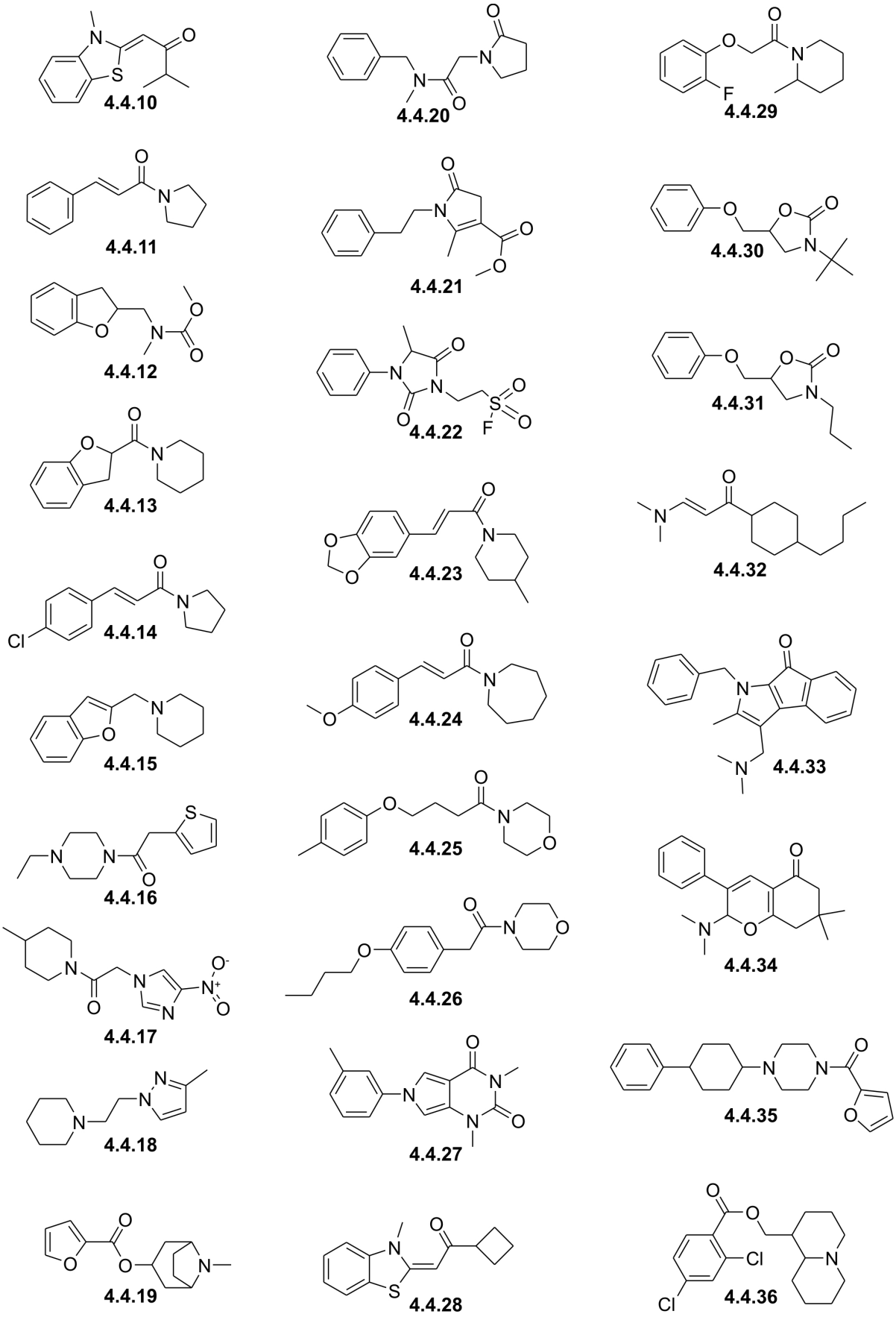


Figure 4.11. Molecules selected from CATS3D virtual screening.

To estimate the degree of "scaffold-hopping" we compared the average distance of each molecule **4.4.10**-**4.4.36** to its respective nearest reference (<D_{lib}>) compound with the average distance between the reference molecules **4.4.3**-**4.4.9** (<D_{ref}>). Three such indices were employed: the CATS3D Manhattan distance, the topological CATS2D Manhattan distance, and the substructure-based MACCS key Tanimoto similarity from MOE. While the average CATS3D distance of the library compounds to their reference molecules was significantly smaller in comparison to the average distance between the reference molecules (<D_{lib}> = 1.41 (± 0.45); <D_{ref}> = 2.66 (± 0.89)), <D_{lib}> was only marginally smaller than <D_{ref}> for CATS2D (3.31 (± 1.48) vs. 3.6 (± 1.4)). With the MACCS keys <D_{lib}> was smaller than <D_{ref}> (0.33 (± 0.11) vs. 0.39 (± 0.15)), indicating a greater similarity among the reference set than between the virtual screening hits and the reference molecules. This demonstrates that the compiled library contains scaffolds which are different from the references (as estimated by MACCS substructure fingerprints) but are still considered isofunctional by the CATS pharmacophore approaches.

In vitro binding studies for mGluR5 were performed on the basis of a [^{3}H]MPEP displacement assay. Estimates of K_i values for the ligands were made from measurements at a fixed concentration of 10 µM. Selectivity of the ligands versus mGluR1, the most similar receptor to mGluR5 within the mGluR family, was assessed by a displacement assay with the Merz proprietary selective mGluR1 antagonist MRZ 3415. Nine molecules (**4.4.10**, **4.4.22**, **4.4.23**, **4.4.24**, **4.4.25**, **4.4.30**, **4.4.32**, **4.4.33**, **4.4.34**) exhibited a K_i value below 70 µM for mGluR5 (Table 4.10), with structure **4.4.10** being the most selective inhibitor. With our assay system we determined a K_i of 12.5 nM for MPEP on mGluR5.

The predicted rank-order of the tested library compounds does not correlate with binding affinity (Table 4.10). It is evident that the Manhattan distance, which was used for compound prioritization, does not distinguish between molecular attributes that are relevant or irrelevant for a particular receptor-ligand interaction. Furthermore, the small list of virtual hits that was compiled for each reference molecule prevents a sound statistical evaluation.

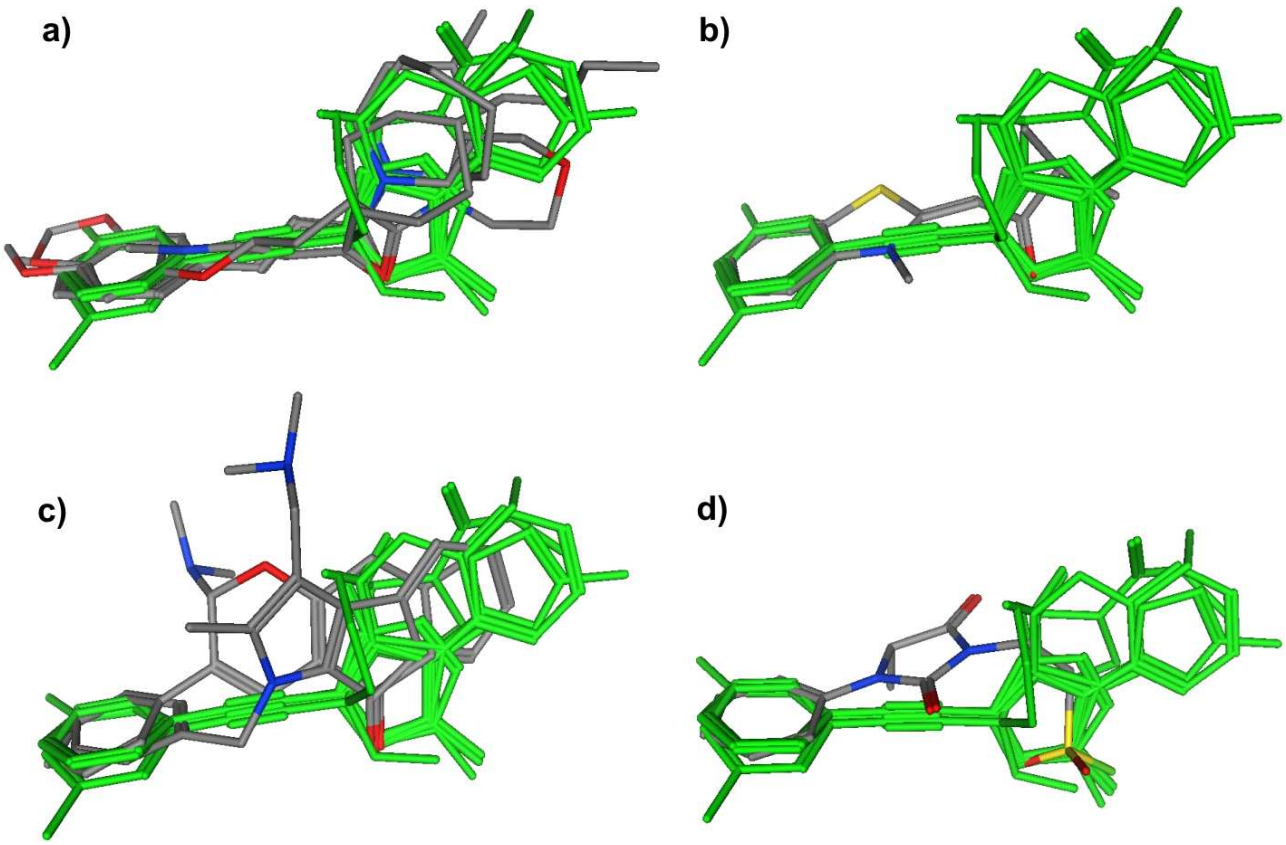


Figure 4.12. Flexible alignment of the most potent found mGluR5 modulators to the alignment of reference molecules **4.4.3-4.4.9** (green). Alignments are shown for (a) **4.4.23**, **4.4.24**, **4.4.25**, **4.4.32**, (b) **4.4.10**, (c) **4.4.33**, **4.4.34**, (d) **4.4.22**.

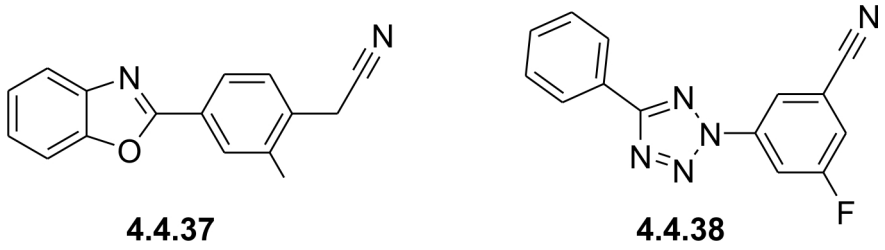
The best found nine molecules were aligned to the reference molecule alignment with the MOE flexible alignment tool (Figure 4.12). **4.4.23**, **4.4.24**, **4.4.25**, **4.4.32** fitted well into the reference alignment (Figure 4.12a) with the keto-group of each molecule superposed onto the pyridine nitrogen as a hydrogen-bond acceptor substitute, and the various linker moieties aligned to the triple bond linkers of the MPEP derivatives. For **4.4.30** a comparable binding mode might be anticipated, which was not found by the flexible alignment since MOE did not recognize the oxazolidine oxygen of **4.4.30** as a potential hydrogen-bond acceptor. Based on the alignment it cannot be decided whether molecules **4.4.10**, **4.4.33**, **4.4.34** and **4.4.22** (Figure 4.12b-d) were actually aligned in a reasonable fashion. For these molecules large substructure elements were placed in the MPEP linker region which we assume to bind to into a narrow part of the receptor binding pocket. To our surprise **4.4.28** and **4.4.31** -- analogs of ligands **4.4.10** and **4.4.30** – showed to be inactive. For both molecules this effect might be explained by steric restrictions in the receptor.

The selectivity of the hits was low. Compound **4.4.27** was even found to be a potent and selective binder of mGluR1. This might indicate the existence of similar binding pockets in both receptor subtypes. Overlap of the binding pockets for antagonists of both receptor subtypes has already been shown [Pagano *et al.*, 2000]. Similar binding pockets are further supported by the weaker mGluR1 selective binder **4.4.11**, which is similar to **4.4.23**, **4.4.24** and **4.4.25.** These are more selective towards mGluR5. Compound **4.4.14** was inactive in both mGluR1 and mGluR5 binding studies, although it might be regarded as a close analogue of **4.4.11**. A higher selectivity of the compounds might be achieved by incorporation of selective molecules acting on mGluR1 in the virtual screening procedure. These might be used as an anti-target in additional similarity searching experiments. Molecules with a high similarity to mGluR5 ligands and a low similarity to mGluR1 ligands might exhibit a better selectivity profile.

A challenging goal of pharmacophore-based similarity searching is "scaffold-hopping". This aim was clearly met in this study. Isofunctional alternatives to the MPEP scaffold were found, which provide several starting points for lead structure development. As an important outcome, the metabolically unstable triple-bond linker present in the MPEP-derived reference molecules is substituted by various alternatives in the compounds that were selected by virtual screening. Noteworthy, the double-bond linker of **4.4.23**, **4.4.24**, and **4.4.32** is structurally identical to the one present in SIB-1893 and similar to the linker type of SIB-1757, both of which were not present in the reference collection (Figure 4.9). Some of the tested compounds (**4.4.12**, **4.4.13**, **4.4.15**) resemble structural similarity to the recently reported mGluR5 antagonist **4.4.37** [Wang *et al.*, 2004] (Figure 4.13), that was found by HTS. This further underlines the ability of the CATS3D approach to find isofunctional but structurally different scaffolds. Molecule **4.4.38**, a recently reported mGluR5 antagonist with a tetrazole linker (Figure 4.13) [Roppe *et al.*, 2004], shows that more voluminous groups like in **4.4.22** might also be allowed in the linker region, assuming an identical binding mode. The novel scaffolds of compounds **4.4.33** and **4.4.34** present a promising opportunity for straightforward combinatorial design with the aim to significantly improve binding behavior.

One possible reason for the low selectivity of **4.4.23**, **4.4.24**, **4.4.25** and **4.4.32** might be due to the replacement of the SIB-1893 pyridine by a keto-group. While the hydrogen-bond acceptor functionality of the pyridine is maintained, the substitution results in a loss of possible steric and stacking interactions. These findings indicate that the receptor subtype selectivity of MPEP-like mGluR5 antagonists might be based on steric or π-π stacking interactions mediated by the pyridine ring. Reference molecule **4.4.9**, which lacks an aromatic

ring, supports the hypothesis that a defined steric interaction in the region of the MPEP pyridine might be sufficient for selectivity.



4.4.37 **4.4.38**

Figure 4.13. Recently reported mGluR5 antagonists with new scaffolds.

4.4.1 Conclusion

Summarizing, it has been demonstrated that pharmacophore-based similarity searching can lead to novel, isofunctional molecular scaffolds that provide a basis for lead structure development. The target was an allosteric binding site of a pharmacologically challenging GPCR. Although homology-based models of the MPEP binding pocket have been published recently [Pagano *et al.*, 2000, Malherbe *et al.*, 2003], successful virtual screening exploiting this information has not been reported until now. The entirely ligand-based CATS3D approach can thus be seen as a working alternative to more demanding structure-based design techniques with the main aim to develop novel lead series.

4.5 Prospective screening for mGluR5 allosteric modulators with an artificial neural network approach based on CATS3D representations

Artificial neural networks (ANN) are an attractive tool for the identification of molecules with a desired biological activity. In this section we used an ensemble of ANNs and self organizing maps (SOMs) to find new specific and diverse allosteric antagonists of mGluR5. The following setup was employed (Figure 4.14):

1. 10 ANNs were trained on the prediction of mGluR5 activity.
2. Two ANNs were trained on the selectivity against mGluR1.
3. Self organizing maps were used to select representative subsets of the predicted virtual hits for pharmacological characterization.

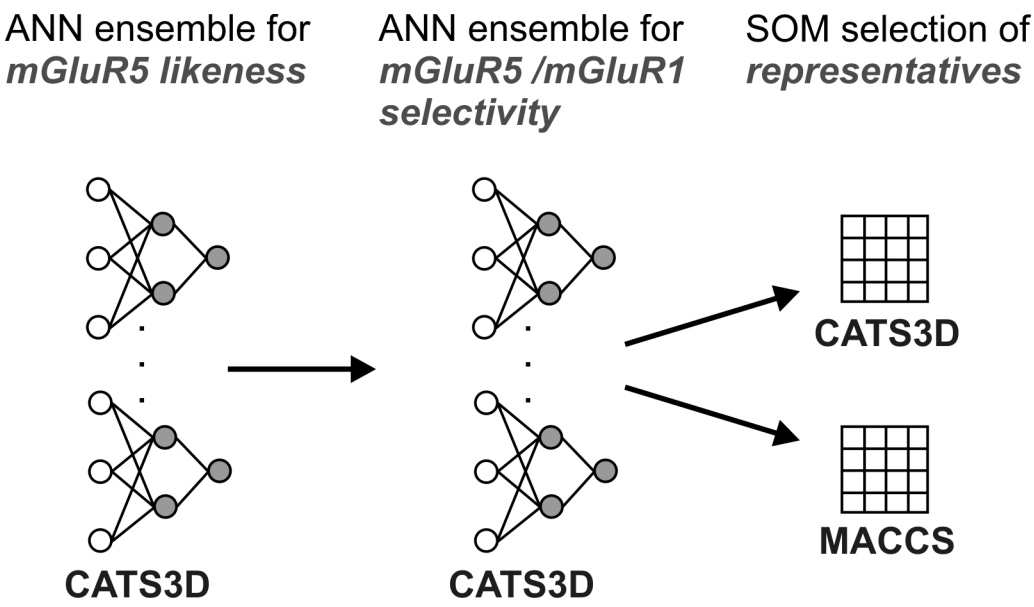


Figure 4.14. Combination of supervised and unsupervised artificial neural networks for the compilation of a focused diverse mGluR5 library. The focus is realized with ensembles of supervised feed-forward networks for the prediction of general “mGluR5-likeness” and “mGluR5 vs. mGluR1 selectivity”. Diversity is obtained with unsupervised self organizing map (SOM) selection of representative subsets of the best fraction of hits from the previous steps.

Many allosteric antagonists of mGluR5 have been described in patents and in scientific literature. However the structural classes of ligands are very dissimilar and it is

unclear how the classes are related to each other from the view of the binding mode in the receptor. The availability of many different active molecules is a genuine starting point for a machine learning algorithm. On the other hand, potential different binding modes and different sub-pockets in the binding site of these ligands might hamper such an approach, where all ligands are considered as active. The origin of the activity data from different assay systems as well as the absence of an obvious alignment of the ligands renders a classification method more appropriate than an approach for the prediction of inhibition constants. Artificial neural networks (ANN) have been shown to be successful in complex classification tasks in drug discovery related projects. Our aim was to create a prediction tool to identify new molecules with specific mGluR5 allosteric modulator activity. Using CATS3D as molecular descriptor, the inherent suitability of this descriptor for scaffold hopping should further support this goal.

For this approach we combined supervised and unsupervised ANNs. First, ensembles of supervised ANNs were trained to separate molecules which possess mGluR5 allosteric antagonist activity ("mGluR5-likeness") from molecules without that property and from molecules with mGluR1 allosteric antagonist activity. Second, unsupervised ANNs, self organizing maps (SOMs), were used to cluster the best scoring molecules and to retrieve representative subsets for experimental testing. Recently an approach was reported combining self organizing maps (SOMs) with feed-forward neural networks [Gini *et al*., 2004]. In these studies SOMs were used as a pre-processing tool to cluster similar molecules. For each of the clusters separate neural networks were trained. These methods obtained an improved prediction accuracy of activities of molecules since neural networks were trained on similarly acting molecules in comparison to a single ANN trained on all molecules [Gini *et al*., 2004]. This approach is similar to the approach of counterpropagation networks [Zupan & Gasteiger, 1999]. Counterpropagation networks consist of one SOM layer, that is trained unsupervised, and an additional output layer for the prediction of observables, that is trained in a supervised manner. However for such an approach sufficiently large datasets are crucial for a successful training of the large number of neural networks. In comparison to our approach this combination of unsupervised and supervised neural networks results in a set of local models with the aim of the highest possible prediction accuracy. Our approach results in a global model with the aim to identify properties of active molecules and to find novel structural clusters which were not identified before.

In a recent article it was stated that the similarity of molecules with predicted properties to the training set is a good indicator for the accuracy of the prediction [Sheridan *et*

al., 2004]. We were interested if this relation was also found for our ANN SOM combination approach, namely if there were more molecules retrieved from SOM neurons containing molecules from the training set in comparison to neurons without training molecules, or not.

4.5.1 Training of feedforward ANNs

Neural networks were trained on two classification tasks. One set of ANNs was trained on the distinction between mGluR5 allosteric antagonists (further referred to as "actives") from other molecules (further referred to as "inactives"). Another set of ANNs was trained on the distinction between actives and mGluR1 allosteric antagonists (further referred to as "side-actives"), the most similar receptor to mGluR5. The training set for the actives consisted of 68 mGluR5 allosteric antagonists from literature, patents and from unpublished molecules from Merz Pharmaceuticals GmbH (Frankfurt). The side-actives set consisted of 158 allosteric antagonists of mGluR1 from patents and literature. Inactives were compiled from the COBRA database. The training procedure of ANNs requires approximately equally sized fractions of molecules form two classes. To obtain a reasonable sampling of the molecules of the COBRA database, five different training sets of 100 molecules were compiled using the MaxMin algorithm [Kennard & Stone, 1996] for maximal diverse subset selection. The dissimilarity was calculated based on the CATS3D descriptor.

For the training of neural networks with the aim to discriminate between actives and inactives all five COBRA subsets were merged with the set of 68 actives, resulting in five data sets of 168 molecules. For all five sets all variables from the CATS3D descriptor with a scaled Shannon entropy (Eq. 2.8) of less than 0.3 were eliminated, leading to 75 to 79 remaining variables. The resulting datasets are further referred to with M5vsCO1 (*m*GluR*5 vs*. *CO*BRA set *1*), M5vsCO2, M5vsCO3, M5vsCO4 and M5vsCO5. Table 4.11 gives an overview over the selected variables for the data sets. The selection differed only in few cases for variables describing larger distances. Variables including cation- and anion-interactions were not selected since all molecules were neutralized before descriptor calculation.

Training with uncorrelated variables can result in improved prediction quality [Schneider & So, 2003]. To test this hypothesis for our classification tasks we calculated uncorrelated versions of M5vsCO1 to M5vsCO5. All variables were autoscaled and a principle component analysis (PCA) was performed. All principle components with eigenvalues above or equal to 1 were used for further calculations. This resulted in ten principle components for each of the data sets. The five resulting data sets with uncorrelated variables are further referred to as M5vsCO1pca, M5vsCO2pca, …, M5vsCO5pca. The

percentages of explained variance of the first two principle components were 69.8 % and 11.1 %, 70.3 % and 10.9 %, 69.2 % and 11.6 %, 70.2 % and 10.7 %, and 70.0 % and 11.0 %, respectively. Accordingly in all five data sets more than 80 % of the variance is explained by the first two principle components. This indicated that the variables of the CATS3D descriptor were highly correlated for the description of these data sets. Projections of actives and inactives using the first 2 principle components revealed that the CATS3D description seemed to be appropriate to separate active molecules from inactive molecules (Figure 4.15).

Table 4.11. Variables selected by scaled Shannon entropy for the different data sets. The variables are coded in the following way: e.g. PA 3-11 means that all polar – hydrogen bond-acceptor bins with distance ranges from 2 to 3, 3 to 4, … , 10 to 11 Å are selected. P = polar, A = hydrogen-bond acceptor, D = hydrogen-bond donor, H = hydrophobic.

selected descriptors	M5vsCO1	M5vsCO2	M5vsCO3	M5vsCO4	M5vsCO5	M5vsM1
PA	3-11	3-12	3-10	3-12	3-10	
PH	2-15	2-16	2-15	2-15	2-15	
DA	3-9	3-9	3-7	3-9	3-9	
DH	2-13	2-13	2-14	2-14	2-14	3-7
AA	3-8	3-9	3-8	3-10	3-8	
AH	2-15	2-15	2-14	2-15	2-15	2-13
HH	2-17	2-17	2-17	2-17	2-17	2-14
total number	77	79	75	79	77	29

For the training of neural networks to discriminate between actives and side-actives the datasets of 68 actives and 158 side-actives were merged. Variables with a scaled Shannon entropy below 0.3 were eliminated, resulting in 29 remaining variables. This dataset is further referred to as M5vsM1 (*m*GluR*5 vs. m*GluR*1*). Due to the smaller variation of the molecules in this data set in comparison to the COBRA subsets, a smaller number of variables were selected having entropies above 0.3. Details on the selected variables are shown in Table 4.11. Uncorrelated variables were obtained by autoscaling and subsequent PCA. The resulting dataset is further referred to as M5vsM1pca. Seven principle components were found with

eigenvalues above 1. The first PC explained 59.8 % and the second PC explained 17.4 % of the variance in the data.

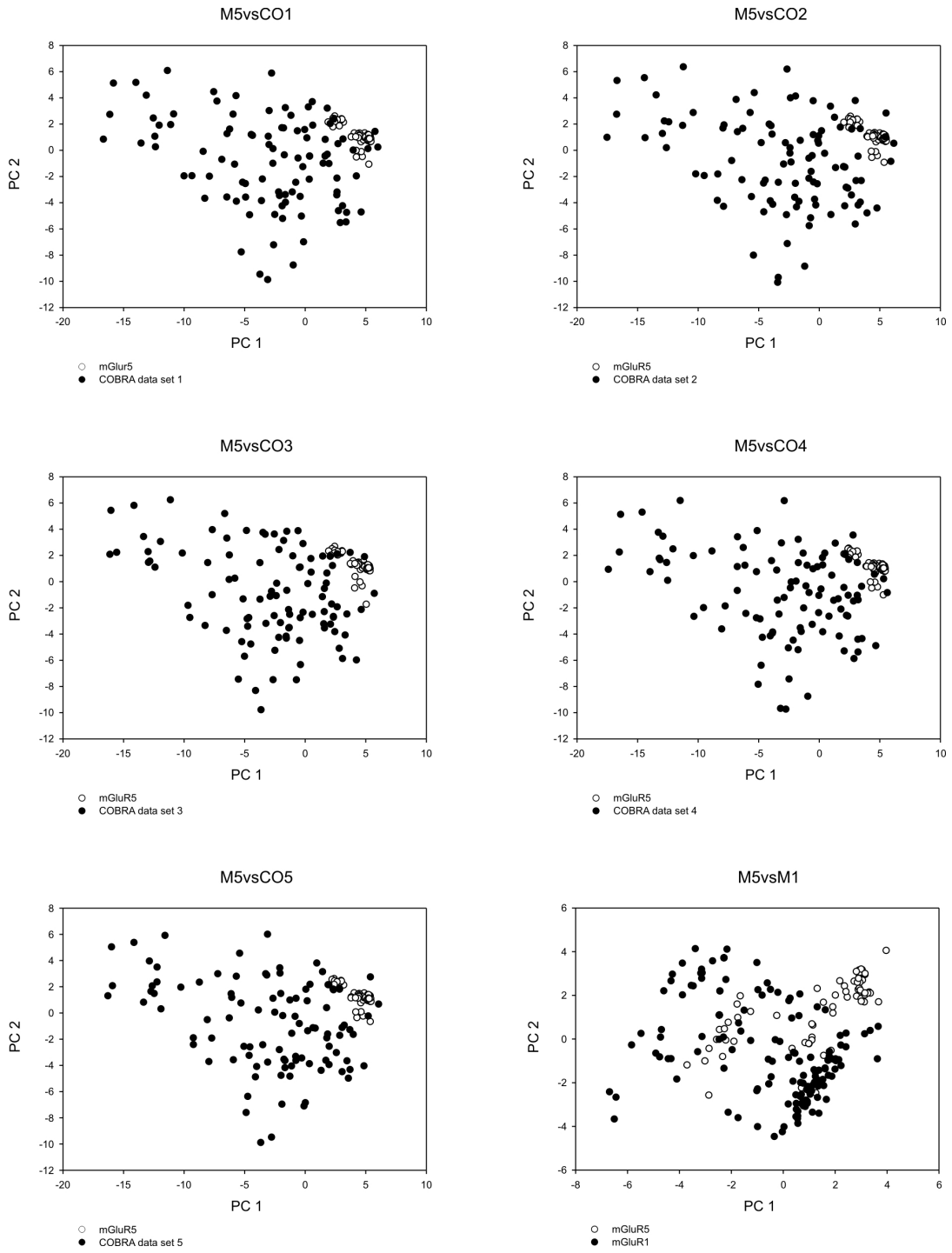


Figure 4.15. Principle component projection of the data sets used for the training of the supervised neural networks. Five datasets were used for "mGluR5-likeness" prediction (M5vsCO1, …, M5vsCO5) and one dataset was used for the "mGluR5 vs. mGluR1 selectivity" prediction (M5vsM1). White dots represent actives and black dots represent inactives or side-actives.

The projection of the first two principle components is shown in Figure 4.15. It is apparent that mGluR5 and mGluR1 were not easily separable by a linear function. However both subsets form several clusters and should therefore be separable by a non-linear classifier like ANN. On the other hand some of the regions overlap which supports the results from Section 4.4, that selectivity between the two classes is determined by small variations of the ligands, which are not yet clearly understood.

Table 4.12. Results of the 10-fold cross-validation for M5vsCO1 to M5vsCO5. The average *cc* values for the training- and test-sets were calculated after 100 steps of training. Selected nets are printed in bold. Standard deviations are given in brackets.

	M5vsCO1		M5vsCO2		M5vsCO3		M5vsCO4		M5vsCO5	
no. hidden neurons	train	test	train	test	train	test	train	test	train	test
1	0.98 (0.03)	0.79 (0.06)	0.97 (0.04)	0.78 (0.05)	0.98 (0.03)	0.80 (0.09)	0.99 (0.02)	0.83 (0.06)	1 (0)	0.84 (0.05)
2	0.99 (0.01)	0.84 (0.06)	1 (0)	0.85 (0.06)	0.99 (0.01)	0.87 (0.04)	1 (0.01)	0.86 (0.03)	1 (0)	0.83 (0.07)
3	1 (0)	0.83 (0.07)	1 (0)	0.86 (0.04)	1 (0)	0.86 (0.07)	1 (0.01)	0.84 (0.08)	1 (0)	0.85 (0.07)
4	1 (0)	0.86 (0.06)	1 (0)	0.87 (0.08)	1 (0)	0.88 (0.05)	1 (0)	0.86 (0.07)	1 (0)	0.88 (0.04)
5	1 (0)	0.86 (0.06)	1 (0)	0.87 (0.06)	1 (0)	0.85 (0.05)	1 (0)	0.88 (0.04)	1 (0)	0.85 (0.05)
6	1 (0)	0.86 (0.06)	1 (0)	0.83 (0.04)	1 (0)	0.85 (0.05)	1 (0)	0.86 (0.05)	1 (0)	0.89 (0.03)
7	1 (0)	0.87 (0.05)	1 (0)	0.84 (0.07)	1 (0)	0.91 (0.03)	1 (0)	0.84 (0.08)	1 (0)	0.88 (0.05)
8	1 (0)	0.86 (0.05)	1 (0)	0.85 (0.06)	1 (0)	0.88 (0.04)	1 (0)	0.84 (0.05)	1 (0)	0.88 (0.05)
9	1 (0)	0.87 (0.03)	**1** (0)	**0.88** (0.04)	1 (0)	0.89 (0.03)	1 (0)	0.83 (0.06)	1 (0)	0.87 (0.06)
10	**1** (0)	**0.88** (0.02)	1 (0)	0.86 (0.06)	**1** (0)	**0.91** (0.04)	1 (0)	0.85 (0.05)	1 (0)	0.86 (0.07)
11	1 (0)	0.85 (0.05)	1 (0)	0.85 (0.06)	1 (0)	0.89 (0.06)	**1** (0)	**0.88** (0.05)	1 (0)	0.89 (0.03)
12	1 (0)	0.85 (0.03)	1 (0)	0.87 (0.07)	1 (0)	0.89 (0.04)	1 (0)	0.84 (0.06)	**1** (0)	**0.90** (0.04)
13	1 (0)	0.85 (0.05)	1 (0)	0.86 (0.06)	1 (0)	0.90 (0.04)	1 (0)	0.85 (0.05)	1 (0)	0.89 (0.04)

All twelve described data sets were employed for the training of neural networks. The optimal number of hidden neurons and the optimal number of training steps had to be determined for each of the twelve datasets. Several numbers of hidden neurons were tested for each of the datasets with ten-fold cross-validation. Cross-validation was employed by random division of the datasets into equal sized fractions of training and test data. This procedure was

repeated ten times to obtain an estimation of network performance. A number of 100 steps were used for the training. In preliminary training experiments 100 steps seemed to be a reasonable compromise between a sufficiently large number of training steps to extract the underlying structure activity relationship and overtraining of the neural networks, which prevents generalization of the predictions.

Table 4.13. Results of the 10 fold cross-validation for M5vsCO1pca to M5vsCO5p. The average *cc* values for the training- and test-sets were calculated after 100 steps of training. Selected nets are printed in bold. Standard deviations are given in brackets.

	M5vsCO1pca		M5vsCO2pca		M5vsCO3pca		M5vsCO4pca		M5vsCO5pca	
no. hidden neurons	train	test	train	test	train	test	train	test	train	test
1	0.95 (0.02)	0.79 (0.05)	0.97 (0.02)	0.84 (0.04)	0.97 (0.02)	0.82 (0.05)	0.97 (0.05)	0.78 (0.05)	0.97 (0.03)	0.80 (0.06)
2	0.99 (0.02)	0.83 (0.08)	**0.98** (0.02)	**0.87** (0.07)	1 (0)	0.85 (0.06)	0.99 (0.02)	0.82 (0.05)	**1** (0.01)	**0.87** (0.06)
3	1 (0.01)	0.81 (0.08)	0.99 (0.02)	0.83 (0.07)	1 (0.01)	0.86 (0.06)	**1** (0)	**0.85** (0.06)	0.99 (0.01)	0.82 (0.06)
4	1 (0)	0.84 (0.05)	1 (0)	0.87 (0.06)	1 (0)	0.88 (0.03)	1 (0.01)	0.84 (0.06)	1 (0)	0.83 (0.06)
5	1 (0.01)	0.82 (0.07)	1 (0.01)	0.87 (0.04)	1 (0)	0.90 (0.03)	1 (0)	0.85 (0.05)	1 (0)	0.85 (0.06)
6	1 (0)	0.86 (0.04)	1 (0)	0.86 (0.06)	1 (0)	0.91 (0.03)	1 (0)	0.84 (0.04)	1 (0)	0.87 (0.06)
7	**1** (0)	**0.89** (0.03)	1 (0)	0.84 (0.06)	**1** (0)	**0.92** (0.05)	1 (0)	0.85 (0.05)	1 (0)	0.86 (0.06)
8	1 (0)	0.85 (0.05)	1 (0.01)	0.87 (0.04)	1 (0)	0.91 (0.04)	1 (0)	0.84 (0.05)	1 (0)	0.85 (0.05)

Table 4.14. Results of the 10 fold cross-validation for M5vsM1 and M5vsM1pca. The average *cc* values for the training- and test-sets were calculated after 100 steps of training. Selected nets are printed in bold. Standard deviations are given in brackets.

	M5vsM1		M5vsM1pca	
no. hidden neurons	train	test	train	test
1	0.99 (0.02)	0.87 (0.05)	0.96 (0.02)	0.8 (0.05)
2	**1** (0.01)	**0.88** (0.03)	0.98 (0.02)	0.81 (0.06)
3	1 (0.01)	0.89 (0.05)	0.96 (0.02)	0.84 (0.07)
4	1 (0.01)	0.88 (0.04)	0.98 (0.02)	0.83 (0.05)
5	**0.99** (0.01)	**0.91** (0.02)	0.99 (0.01)	0.78 (0.06)
6	1 (0.01)	0.88 (0.04)	0.99 (0.02)	0.79 (0.08)
7	1 (0.01)	0.88 (0.05)		
8	0.98 (0.02)	0.9 (0.04)		
9	0.99 (0.01)	0.89 (0.05)		
10	0.99 (0.01)	0.9 (0.04)		

The evaluation of different number of hidden neurons for the ANN training is given in detail in Tables 4.12, 4.13, and 4.14. In all active/inactive classification networks the test data was 100% correctly predicted employing more than two or three hidden neurons. For the actives/side-actives networks this was not true. Using larger numbers of hidden neurons did not always lead to 100% prediction accuracy. This effect might be grounded on neural network training using too few training steps or on the complexity of the separation task.

ANNs were selected with the best Matthews *cc* in the training and the test data. In the case that more than one network obtained equal best *cc* values, the net with the lowest number of hidden neurons was selected. Employing the best found number of hidden neurons, successful predictions of the test sets were obtained with Matthews coefficients equal or larger than 0.84 for all test data sets. For the actives/inactives separation only small differences in the test data prediction accuracy were found between the raw descriptor values and the uncorrelated variables (best found Matthews *cc* = 0.91 vs. 0.92 for the test data). For the actives/side-actives separation, the raw descriptor values performed better than the uncorrelated variables (best found Matthews *cc* = 0.91 vs. 0.84 for the test data). This might be caused by the difference of the classification task. A general classification task applied to relatively easily separable data like with the actives/inactives classification might profit or at least not be hampered by a more general data presentation with uncorrelated variables. For the specific and complex separation of actives and side-actives details of the descriptors might have played a role, which were lost in the uncorrelated variables. Additionally, more descriptor variables automatically lead to a better separation between data classes, but also favors overfitting of the descriptor values to the observed data.

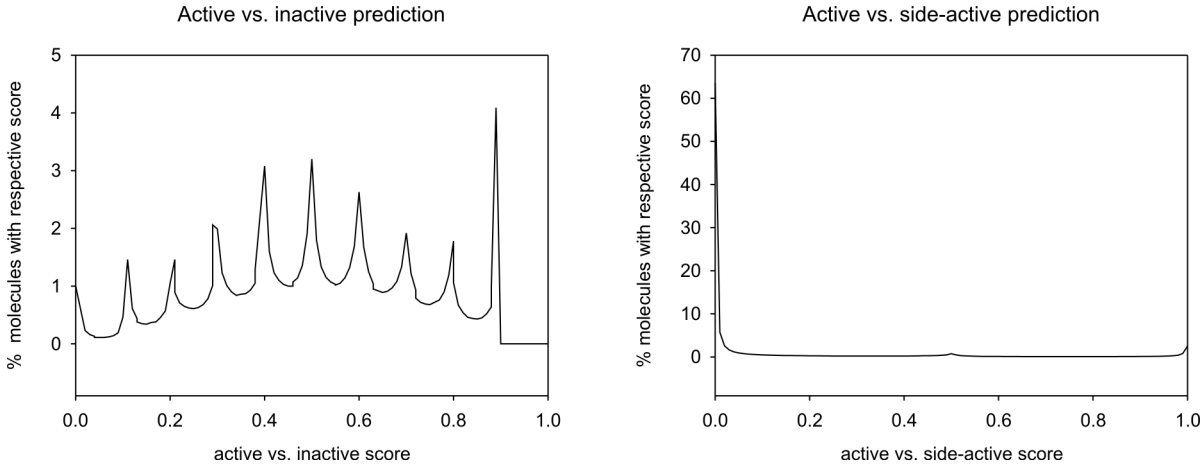


Figure 4.16. Frequency of predicted score values for the molecules of the Enamine database.

For each data set for actives/inactives classification the parameters of the best performing model with the minimum number of hidden neurons were selected for the final training. The selected nets are printed in bold in Tables 4.12, 4.13 and 4.14. For the actives/side-actives classification ANNs with uncorrelated variables were not used. Instead two nets with two and five hidden neurons were selected with the original variables. Final neural networks were trained using the full data sets with the given number of hidden neurons.

4.5.2 Prediction of allosteric mGluR5 modulators

For prospective virtual screening we used the database of the commercial molecule-supplier Enamine [Enamine], which consisted of 1,022,483 molecules. For each neural network the molecules were processed like the training data: the same variables were selected as for the training data. When necessary, autoscaling and PCA were applied using the means, standard deviations and transformation matrices from the training data sets. Consensus scores were obtained by calculation of the average values of the ten ANNs for the actives-inactives classification and for the two ANNs for the actives-side-actives classification. A histogram of the two distributions of the score values is shown in Figure 4.16. Interestingly the scores for the actives/inactives classification did not exceed a value of 0.89 for the Enamine dataset. This was an effect of the ensemble neural network average score that was applied for the prediction. One of the trained ANNs (M5vsCO5) did not predict any of the Enamine molecules as active, despite the fact that this network performed best in training. This might be an indicator that the Enamine dataset might not be appropriate for the screening for mGluR5 allosteric antagonists. An alternative to average scores to find a decision based on ensembles of neural networks is the jury decision: a compound is considered as active if the majority of neural networks consider the compound as active. In this work a more stringent criterion was used: a unanimous decision was needed to result in a maximum score of 1. This strategy was applied due to the fact that a large fraction of the Enamine dataset was considered as active, by most of the networks (Figure 4.16).

Figure 4.16 shows the effect of the consensus scoring by average score values. The number of peaks found in the scores reflects the number of individual score-values used for the average score. For the actives/side-actives classification three peaks were found: a first large peak of many molecules where both nets agree that the molecules are more likely side-actives than actives, a small peak at 0.5 where the two nets do not agree, and another small

peak where both nets agree that the molecules belong to the actives class. For the actives/inactives classification there should be eleven peaks, but since one ANN did not predict any molecule as active, only ten peaks were found.

From the actives/inactives prediction we selected all molecules with an average score larger than 0.885. This value was found to be a reasonable compromise between the presumed diversity of the hits and the number of obtained hits. We assumed to find structurally more diverse molecules by this strategy in comparison to using the top scoring molecules alone. This resulted in a selection of 41,663 molecules. From the actives/side-active classification we selected all molecules with an average score above 0.99, which resulted in an additional 32,099 molecules. The union of these sets gave a set of 8,403 molecules which were considered as our focused library for mGluR5 allosteric antagonists.

4.5.3 Selection of a representative subset by SOMs

The obtained focused library was further analyzed by self organizing maps (SOM). SOMs provide a topology-preserving projection of data from a high-dimensional space into a low dimensional space. The resulting maps also define clusters in the data and provide representative and diverse subsets of the original data. Two different SOMs were trained: one SOM based on the CATS3D descriptor and one SOM based on MACCS keys. In this way we wanted to analyze our focused library by two different objectives. One objective was to get an overview over the distribution of diverse sets of scaffolds (based on the MACCS keys) and the other objective was to get an overview over the different pharmacophores in the library (based on CATS3D). CATS3D representations of molecules consist of a comparably large number of dimensions (420) in relation to the two-dimensional SOM projection. To facilitate SOM training the discrepancy between the two variable spaces was reduced by the scaled Shannon entropy. Only CATS3D-dimensions with a scaled Shannon entropy above 0.3 were used for SOM training. To estimate the overlap of the library with the chemical space covered by the known mGluR5 allosteric antagonists, all mGluR5 actives from the training set were included in the SOM training. Two SOMs with 5 x 5 neurons were trained. The resulting SOMs are shown in Figure 4.17.

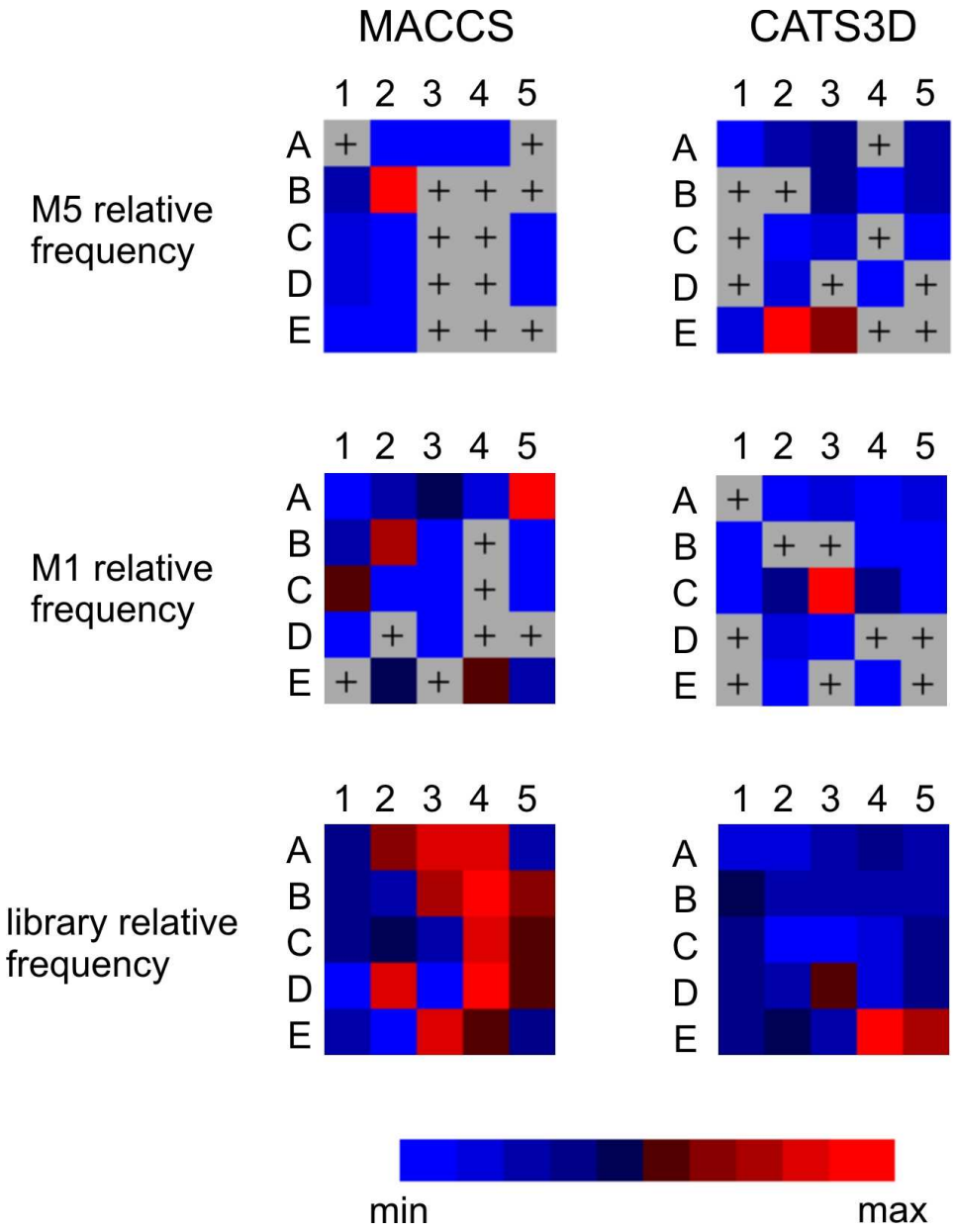


Figure 4.17. Self organizing maps (SOMs) of the best predicted Enamine molecules. SOMs were calculated using MACCS keys and CATS3D descriptors. Shown are the distribution of the mGluR training set, the mGluR1 training set and the frequency of the predicted hit molecules, projected on the trained SOM.

In the MACCS SOM 13 of 25 neurons and in the CATS3D SOM 15 of 25 neurons contained known actives. This indicates that the predicted molecules broadly covered the chemical space of the known mGluR5 allosteric antagonists. All neurons without active molecules from the training set were directly neighboring a neuron containing active molecules within its receptive field. In the MACCS SOM the neurons containing active molecules from the training set built a cluster with a single neuron containing most of the

molecules. In contrast, in the CATS3D SOM the neurons containing active molecules were loosely distributed over the map with more neurons containing a larger fraction of active molecules. The results from these two projections indicate that the ANN approach was able to predict novel scaffolds and chemotypes which were similarly distributed between the CATS3D representations, but are found outside of the cluster of known actives in the MACCS SOM. A projection of the side-actives onto the trained map revealed that these molecules were distributed broadly over the map. In the CATS3D SOM 16 neurons were activated by mGluR1 antagonists including 10 neurons that also contained mGluR5 antagonists. For the MACCS SOM 18 neurons were found with mGluR1 antagonists overlapping with 10 neurons with mGluR5 antagonists. The SOMs were not able to define a full separation of actives and side-actives on the basis of unsupervised learning. This might be grounded on the selection of inappropriate descriptors for that task and on the similarity of the two classes of ligands (Figure 4.17). The relative frequency of the selected library is shown in Figure 4.17. According to the MACCS keys, the library was broadly distributed over the map. With CATS3D most of the library compounds were found in a small set of two neighboring neurons. Interestingly both of these neurons did not contain any known active reference. These results indicate the presence of large sets of analogues in the Enamine dataset which introduced a bias in the molecules selection. By selection of representative compounds for the biochemical verification of the compounds this bias was circumvented. For experimental screening for new allosteric modulators of mGluR5, all molecules which were nearest to the neuron centroids were selected. The respective molecules are shown in Figure 4.18 and Figure 4.19. The representative molecules from both SOMs show a similar topology in comparison to known mGluR5 allosteric antagonists (see Section 4.4). Most molecules consist of two ring systems connected by a linker of 3 or more bonds length. These characteristics were found in molecules from neurons with and without known active references and in the representative molecules of both SOMs. Some of the selected molecules included charged groups like the nitro group. Charged groups were not accounted for in the CATS3D descriptor after the variable selection procedure. Especially in the MACCS SOM these molecules were found mostly outside of the neurons containing the known active molecules.

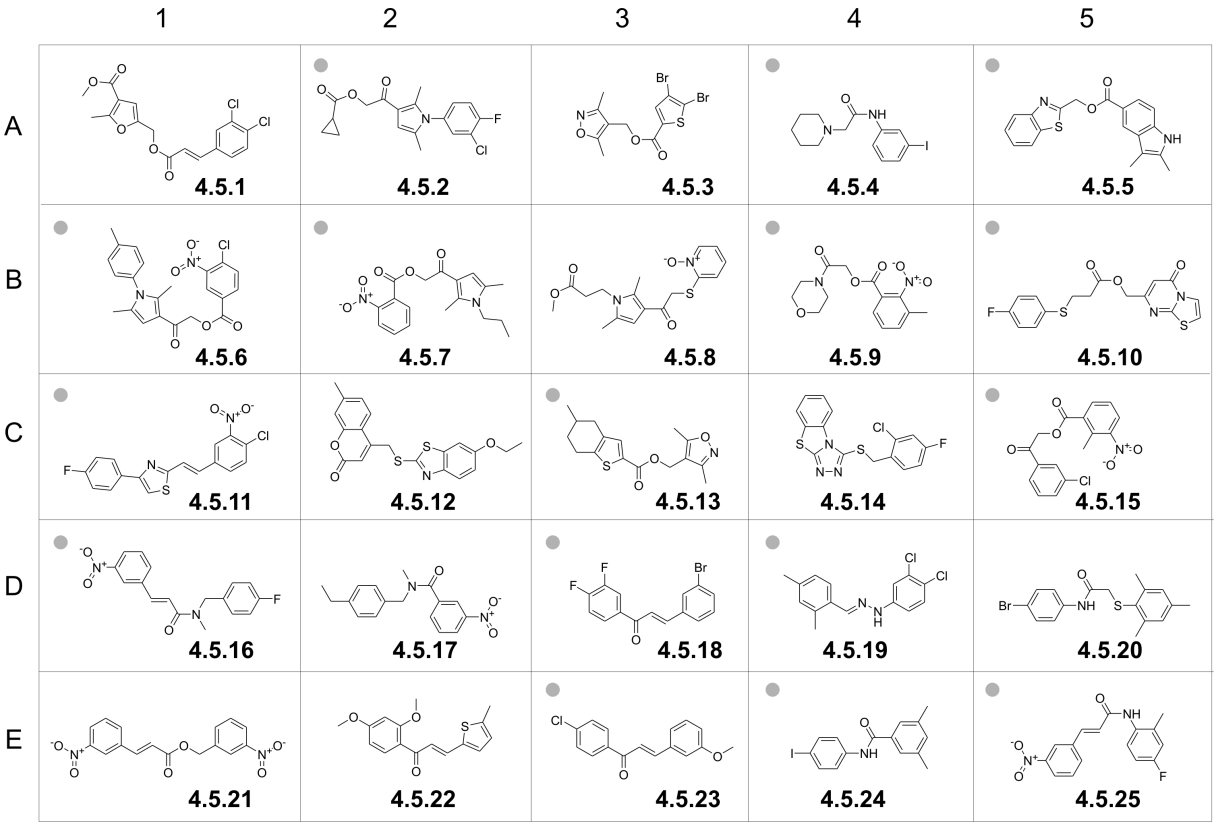


Figure 4.18. Representative molecules selected from the CATS3D SOM. Grey dots indicate molecules that were tested in the binding assay.

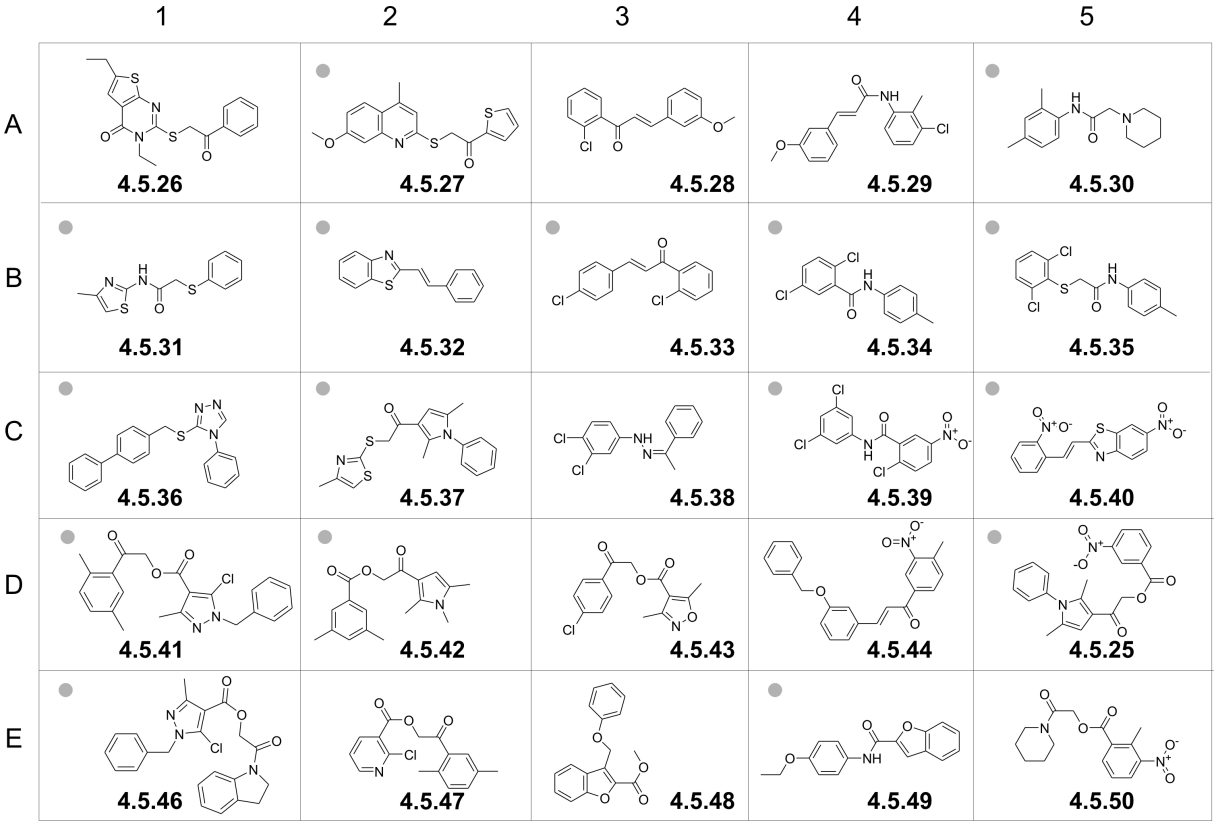


Figure 4.19. Representative molecules selected from the MACCS SOM. Grey dots indicate molecules that were tested in the binding assay.

4.5.4 Binding assay results

The molecules selected by the SOMs, that were available (32 of 50 molecules) were purchased from Enamine [Enamine] and tested in an mGluR5 binding assay. The hits and some of the inactive molecules from the mGluR5 assay were also tested in an mGluR1 binding assay to assess the selectivity of the molecules. The results of the assay are shown in Table 4.15. Three of the 16 tested molecules from the CATS3D SOM and two of the 16 molecules tested from the MACCS SOM showed mGluR5 binding.

The best binding ligand for mGluR5 was found with the CATS3D SOM in neuron e3: **4.5.23** with a K_i of 21 µM. This ligand is structurally similar to ligand **4.4.24** (K_i = 12 µM) found with CATS3D similarity searching in Section 4.4, though the mGluR5 hits from Section 4.4 were not included in the training data. The K_i values from Section 4.4 were determined considering the fraction of solvated ligand under the assay conditions. This parameter was not determined in this section. For **4.4.24** 60 % of the molecule was found in solution. Assuming a comparable solubility for **4.5.23** than for **4.4.24**, similar K_i values were found for the two molecules. The best hit from the MACCS SOM was found in neuron b3: **4.5.33** with a K_i of 33 µM, which is also similar to **4.5.23** and **4.4.24**. One apparent difference of **4.5.33** to **4.5.23** and **4.4.24** is the lack of a hydroxyl-group substituent that is present in the other two molecules at the benzene ring distant to the linker oxygen. Regarding the lower K_i of **4.5.33** in comparison to **4.5.23** and **4.4.24**, the hydroxyl group might provide a favorable interaction with the receptor. A part of the effect of the hydroxyl-group might also be addressed to a lower solubility of **4.5.33**.

The selectivity of the molecules for mGluR5 over mGluR1 was low: nine of the 12 molecules tested from the CATS3D SOM and one of the molecules tested from the MACCS SOM were also found to bind to the mGluR1. Thus more molecules were found binding to mGluR1 than to mGluR5. The molecule with the highest binding affinity was also found for mGluR1 (**4.5.13** with a K_i of 8 µM for mGluR1 and a K_i of 38 µM for MgluR5). Similar results were observed in Section 4.4 where additional potent ligands for mGluR1 were found with CATS3D similarity searching using specific mGluR5 allosteric antagonists as query molecules. This might reveal that the employed molecular representation with CATS3D was not appropriate for this particular task. While CATS3D was designed for scaffold hopping, and accordingly fuzzy representation of the molecules the selectivity between mGluR5 and mGluR1 seems to require subtle differences in the ligands, due to the high similarity of the binding pockets [Pagano *et al*., 2000].

Many of the molecules that were active on at least one of the receptors mGluR1 or mGluR5 were found to have amide groups (**4.5.24**, **4.5.35**, **4.5.49**) or ester groups (**4.5.5**, **4.5.7**, **4.5.10**, **4.5.13**) as linkers. These linkers were not found in the training data and might be considered as alternative chemotypes for the linker part of the molecules.

One of the most challenging goals of virtual screening is to retrieve novel scaffolds and chemotypes with the desired biological activity. In this approach all molecules predicted as active were clustered by SOMs. Retrieving active molecules from clusters of molecules that did not contain active training samples is a way to find molecules different from the training chemotypes. In the SOMs different clusters are represented by different neurons. With our approach we were able to find novel active chemotypes from SOM neurons that did not contain any of the known reference molecules. This was found for mGluR5 and for mGluR1, using both molecular descriptors MACCS and CATS3D for the SOM calculation. For the molecules tested from the MACCS SOM all molecules that were active on mGluR5 were found in neurons that did not contain any known mGluR5 antagonist and the molecule that was active on mGluR1 was found in a neuron that did not contain any of the mGluR1 training molecules. For the CATS3D SOM molecules were found to be active that were from neurons not containing reference molecules from the training set (M5: **4.5.7**, M1: **4.5.16**, **4.5.19**, **4.5.23**). However the most active molecules were found in neurons with reference molecules: the best identified mGluR5 antagonist **4.5.23** was found in the neuron containing the second largest number of reference molecules and the best mGluR1 antagonist **4.5.13** was found in the neuron with the most mGluR1 references. These results are in agreement with the findings of Sheridan et al. [Sheridan *et al*., 2004], that best predictions are obtained for molecules similar to the training data. In contrast to their findings we also found actives with lower activities in neurons not containing training molecules. This ability might have resulted from the strategy of training a global model for mGluR5 instead of a set of local models. However the task of mGluR5/mGluR1 selectivity might have been better represented using a set of local models.

Table 4.15. Results from the mGluR5 and mGluR1 binding assays.

Molecule (neuron)	K_i mGluR5 (µM)	K_i mGluR1 (µM)	Selectivity (IC_{50} mGluR1 / IC_{50} mGluR5)
CATS3D-SOM			
4.5.2 (a2)	> 100		
4.5.4 (a4)	> 100	73	< 0.7
4.5.5 (a5)	> 100	46	< 0.5
4.5.6 (b1)	> 100	> 100	1
4.5.7 (b2)	44	51	1.2
4.5.9 (b4)	> 100		
4.5.10 (b5)	> 100	69	< 0.7
4.5.11 (c1)	> 100	> 100	1
4.5.13 (c3)	38	8	0.2
4.5.15 (c5)	> 100		
4.5.16 (d1)	> 100	45	< 0.5
4.5.18 (d3)	> 100		
4.5.19 (d4)	> 100	55	< 0.6
4.5.23 (e3)	21	64	3.1
4.5.24 (e4)	> 100	56	< 0.6
4.5.25 (e5)	> 100	> 100	1
MACCS-SOM			
4.5.30 (a5)	> 100		
4.5.31 (b1)	> 100		
4.5.33 (b3)	33	> 100	> 3.0
4.5.34 (b4)	> 100	69	< 0.7
4.5.35 (b5)	> 100		
4.5.36 (c1)	> 100		
4.5.37 (c2)	> 100		
4.5.39 (c4)	> 100	> 100	1
4.5.40 (c5)	> 100	> 100	1
4.5.41 (d1)	> 100		
4.5.42 (d2)	> 100		
4.5.45 (d5)	> 100		
4.5.46 (e1)	> 100		
4.5.47 (e2)	> 100		
4.5.48 (e3)	> 100		
4.5.49 (e4)	59	> 100	> 1.7

4.5.5 Conclusions

Using an artificial neural network approach we retrieved novel chemotypes for allosteric modulators of mGluR5. We used a combination of feedforward neural networks trained on the separation of mGluR5 allosteric antagonists from molecules without that activity and trained on the separation of allosteric antagonists of mGluR5 and mGluR1. A representative set of molecules for biochemical testing was compiled using unsupervised SOMs. Novel mGluR5 antagonists with a best K_i value of 21 were found. We were able to retrieve new active molecules from regions in the SOMs that contained molecules from the training set and from regions that did not contain these molecules. Thus our method was able to correctly predict molecules as active that were not similar to the reference molecules. This ability might have resulted from the training of a set of global model based on all molecules from the heterogeneous training set of mGluR5 antagonists instead of a set of local models. Prediction of the selectivity of ligands was not successful. This property might have been better predicted with a set of local models using less general molecular descriptors. Thus the combination of the CATS3D descriptor with neural networks might be best suited for the purpose of scaffold hopping, but not for the purpose of ligand optimization.

4.6 Retrospective evaluation of SQUID fuzzy pharmacophore models

Conventional similarity searching (like with the CATS family methods) employs a single query molecule for each virtual screening run. In contrast, ensemble-based pharmacophore searching [Güner, 2000] (for three-dimensional substructures, e.g. like with Catalyst [Greene *et al*., 1994]) incorporates information from multiple active molecules. Using information from multiple reference molecules has also been shown to improve alignment-free descriptor vector based virtual screening [Xu *et al*., 2001; Hert *et al*., 2004a; Hert *et al*., 2004b]. However there is the limitation that conserved features in the alignment-free descriptor space are not necessarily conserved in a three-dimensional alignment of ligands.

Traditional pharmacophore searching approaches define a query as a substructure. For regions in molecules not covered by the substructure no preference is assigned. This can lead to the effect that many hits contain large or undesired structural elements in the undefined regions. Excluded volumes can compensate for a part of the problem by preventing the selection of molecules that are too large for the binding pocket [Güner, 2000].

Using pairs, triplets, or even quartets of atoms as PPPs is one possibility for the construction of a CV descriptor. An extension to this approach is to use pairs of larger and more general objects, which might result in a more generalized and abstract description of the molecule.

The SQUID (Sophisticated Quantification of Interaction Distributions) fuzzy pharmacophore is an approach that was designed to tackle the above mentioned topics. In SQUID pairs of Gaussian probability densities are used for the descriptor calculation. The Gaussians represent clusters of atoms comprising the same pharmacophoric feature within an alignment of several active reference molecules. The incorporation of multiple aligned ligands within the SQUID approach resembles conceptual similarity to the traditional idea of a pharmacophore model [Güner, 2000]. Based on an alignment of active molecules, tolerances for the features are usually estimated to compensate for ligand and receptor flexibility. Pharmacophoric features that are present in many of the reference molecules result in a high probability, and features which are sparse in the underlying molecules result in a low probability. In this way all features of the reference molecules are included in the model and not just the most conserved substructure. Tolerances of the features, which are considered by this approach, might be better represented by Gaussian densities than by rigid spheres. For the resulting fuzzy pharmacophore models different degrees of fuzziness can be defined, e.g. the

model can be very generalizing or more restricted to the underlying distribution of atoms from the alignment. The fuzziness can be affected by the cluster radius, a variable which determines the radius within which atoms are clustered into PPPs.

For virtual screening the three-dimensional spatial distribution of Gaussian densities is transformed into a two-point correlation vector representation which describes the same probability density for the presence of atom pairs, comprising defined pharmacophoric features. This representation is independent from translation and rotation which makes rapid database screening possible without the necessity to explicitly align the molecules, which can be a limiting step for the screening of large databases. This renders the fuzzy pharmacophore CV useful for ranking 3D pharmacophore-based CV representations of molecules, namely CATS3D descriptors of molecules. Consequently SQUID can be characterized as a hybrid approach between conventional pharmacophore searching, similarity searching and fuzzy modeling.

The goal of this study was to evaluate the pharmacophore model perception and virtual screening ability of the SQUID fuzzy pharmacophore models. The ability of SQUID pharmacophore models to find important interaction points was tested for known reference pharmacophore models from literature. The effectiveness in virtual screening was compared with CATS3D similarity searching and traditional pharmacophore searching with MOE [Chemical Computing Group]. An optimization procedure of feature-type weight was necessary in model calculation. The robustness of this optimization was evaluated, too.

For the evaluation study we selected pharmacophore models for cyclooxygenase 2 (COX-2) and thrombin from literature [Palomer *et al*., 2002; Patel *et al*., 2002]. Both targets are well characterized in the literature and crystal structures of the receptors with bound inhibitors are available. This was important since our method depends on a meaningful alignment of ligands. Large sets of ligands for both targets are known, which is essential for statistical significant results. Ligands from both activities differ largely in size and molecular interactions. COX-2 inhibitors are known to be a class of lower diversity (see section scaffold hopping) while thrombin inhibitors show a higher diversity in chemotypes and scaffolds (see section scaffolds). Using these two references ligand classes the scaffold hopping capability of SQUID could also be assessed.

For retrospective screening we used the COBRA database [Schneider & Schneider, 2003] (version 2.1). Two versions were calculated: one database with single conformations was calculated with CORINA [Sadowski *et al.*, 1994] and one database of up to 50 energy minimized conformations was calculated with MOE [Chemical Computing Group]. For

retrospective screening molecule that were used for pharmacophore model generation were removed from the datasets. The resulting datasets consisted of 92 active molecules and 4611 inactive molecules for COX-2, and 188 actives and 4517 inactive compounds for thrombin.

4.6.1 Pharmacophore model of COX-2 ligands

Palomer *et al.* [Palomer *et al*., 2002] derived a pharmacophore model for COX-2 inhibitors on the basis of five specific inhibitors SC-558 (**4.6.1**), rofecoxib (**4.6.2**), DFU, celecoxib, and a molecule which they termed "molecule 5" (M5, **4.6.3**). For calculation of a 3D structural alignment of these ligands they used a template alignment of all COX-2 ligands, for which there was a crystal structure of the ligand-receptor complex available. Crystal structures were at hand for SC-558 (1CX2) and the two unspecific inhibitors flurbiprofen (3PGH) and indomethacin (4COX). The alignment of these molecules was performed by superposition of their protein structures. The remaining ligands were aligned to the template alignment with the program Catalyst [Greene *et al*., 1994]. This approach was taken as a reference for the development of a pharmacophore model with our own program SQUID. The molecules DFU and celecoxib were not included in the SQUID pharmacophore model, because they are close analogs of rofecoxib and SC-558. The 2D structures of the remaining molecules are shown in Figure 4.20. Crystal structures 1CX2, 3PGH and 4COX were aligned with the homology alignment tool of MOE [Chemical Computing Group]. Rofecoxib and M5 were aligned to this template alignment with the flexible alignment tool of MOE. First, rofecoxib was aligned to the fixed template alignment. Then, M5 was aligned to the fixed alignment resulted from the previous step. For the final alignment the unspecific inhibitors were removed. The resulting alignment of COX-2 inhibitors is shown in Figure 4.21. In accordance with the model of Palomer *et al.* the crucial pharmacophore features of these molecules are the sulfonyl group and the two aromatic six-membered rings [Palomer *et al*., 2002]. The aromatic rings close to the sulfonyl group, further referred to as "ring A", are nearly parallel to each other in the model. The angles between the planes of the distant aromatic rings, further referred to as "ring B", seem to be less constrained. The least conserved region of the model is the linker region between the two aromatic ring centers.

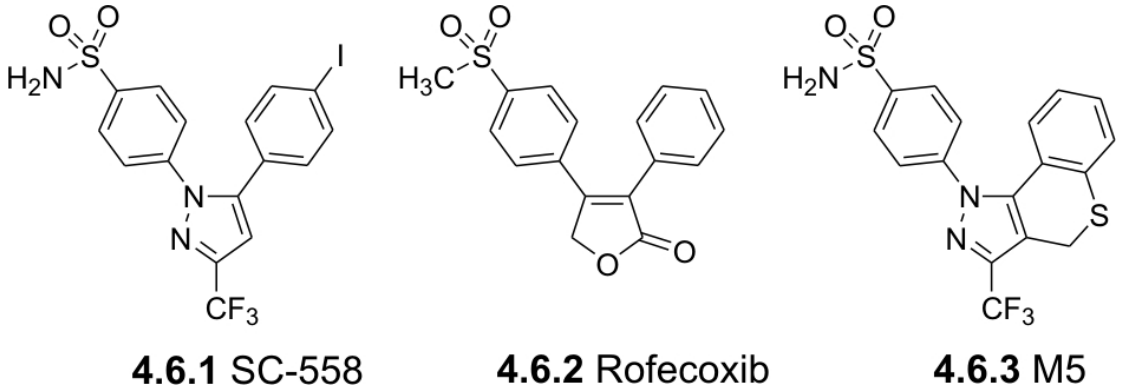


Figure 4.20. Reference COX-2 inhibitors used for the calculation of the SQUID pharmacophore model.

SQUID pharmacophore models were calculated with cluster radii from 0.5 Å to 3.5 Å in steps of 0.1 Å. A sample set of these pharmacophore models is shown in Figure 4.22. The models consisted of only three generalized interaction types: hydrogen-bond donors, hydrogen-bond acceptors, and hydrophobic interactions. The model resulting from 1 Å cluster radius is the most detailed one. Here atoms in close proximity are combined to PPPs, which results in a low abstraction from the chemical scaffolds. In contrast to all other models shown, the preferred angle between the two aromatic rings A and B are preserved in this model. The models resulting from 1.5 and 2.0 Å exhibit a higher degree of generalization from molecular structure. Many atoms, especially in the regions of the aromatic rings A and B, were combined to form large PPPs, covering several atoms from each of the molecules. Up to 2.0 Å only hydrophobic atoms were combined. The models from the cluster radii 2.5 Å and 3.0 Å still represent the overall shape of the molecular alignment with three hydrophobic PPPs, but in the 3.5 Å model the shape of the alignment is only marginally visible. In all models with a cluster radius up to 2.0 Å the sulfonyl group is represented by two highly conserved hydrogen-bond donor PPPs, one hydrogen-bond donor PPP, and one hydrophobic PPP. In the models resulting from cluster radii greater than 2.0 Å all oxygen atoms of the sulfonyl group are represented by a single large PPP. Moreover, the hydrophobic PPP vanished since the methyl group was assigned to the PPP of ring A.

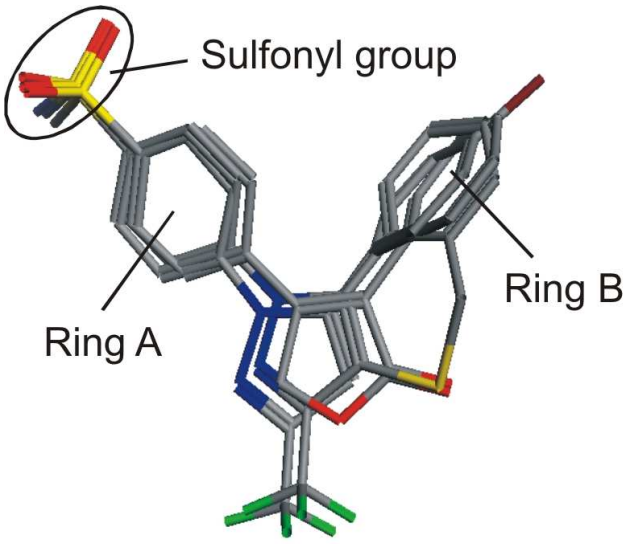


Figure 4.21. Three-dimensional alignment of the COX-2 inhibitors. Rofecoxib and M5 were aligned to the crystal structure conformation of SC-558 bound to COX-2. Essential interactions for specific COX-2 inhibitors are the aromatic rings A and B and the sulfonyl group.

4.6.2 Retrospective screening for COX-2 inhibitors

As the results of retrospective screening were sensitive to the feature-type weights (data not shown), a restrained exhaustive search for the optimization of these weights is part of the model creation procedure. For every calculated model, each of the feature-type weights for features present in the pharmacophore model was varied from 0.1 to 0.5 in steps of 0.1, which resulted in 125 different weighting schemes for the COX-2 pharmacophore models. Each of the resulting descriptors was evaluated by retrospective screening. To obtain statistically more significant results, five different subsets of the COBRA database were created. For each of the subsets 50% of actives and 50% of inactives were randomly chosen from the original database for retrospective screening.

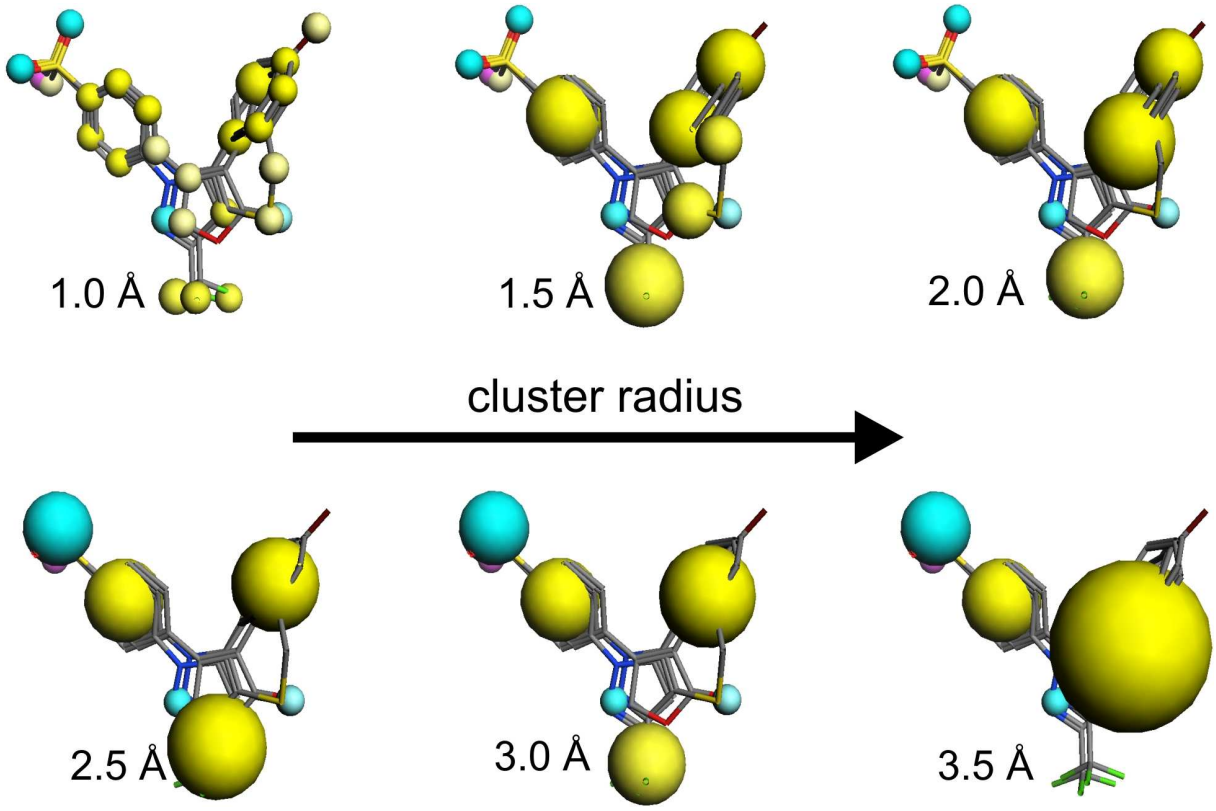


Figure 4.22. SQUID fuzzy pharmacophore models for COX-2 calculated from different cluster radii. The Gaussian PPPs of SQUID are represented by spheres. The radius of a sphere denotes the standard deviation of the PPP and the intensity of the color illustrates the conservation weight of the PPP. Yellow = hydrophobic, cyan = hydrogen-bond acceptor, magenta = hydrogen-bond donor.

The results of the optimization procedure are shown in Figure 4.23. For each model calculated with a different cluster radius the average enrichment factors for the first 1% and 5% of the 5 ranked databases obtained with the best found weighting scheme are shown. The highest average enrichment factor of 39 for the first 1% of the database was obtained with the model calculated with a cluster radius of 1.4 Å and feature-type weights of 0.1 for hydrogen-bond donors, 0.4 for hydrogen-bond acceptors and 0.3 for hydrophobic interactions.

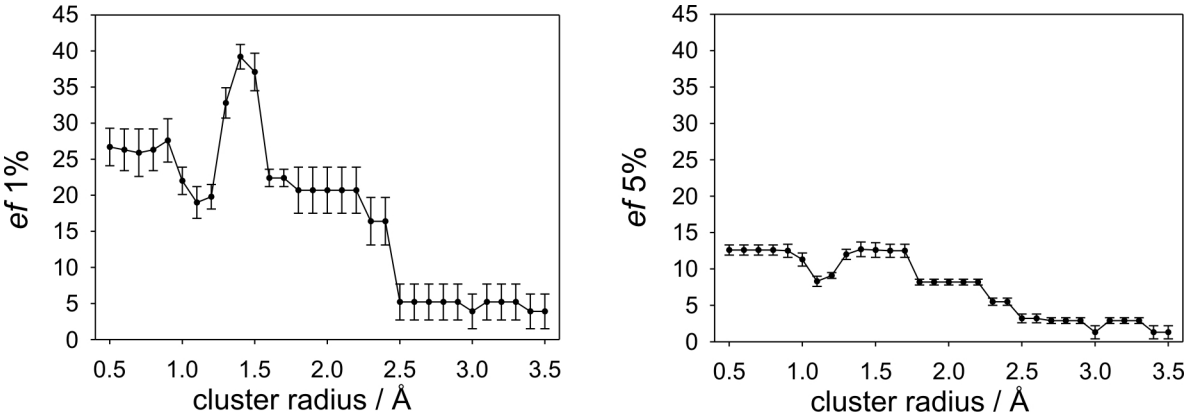


Figure 4.23. Enrichment factors for the first 1% and 5% from retrospective screening with COX-2 pharmacophore models with cluster radii from 0.5 Å to 3.5 Å. For each cluster radius the result from the best found feature type weights from the optimization procedure is shown. The shown enrichment factors are average values from screening of five randomly selected subsets of the COBRA database.

As it could be anticipated, the standard deviations of the enrichment factors were the highest for the first 1% and decreased for the first 5%. Nevertheless according to their standard deviations the enrichment factors for the first 1% of the database still seem to be appropriate for an evaluation of our pharmacophore models. Both curves exhibit the same general characteristics for different cluster radii, although the differences between the models vanish more and more considering the enrichment of the first 5% of the database (Figure 4.23).

Considering the performance of the models for the enrichment in the first 1% of the database, large enrichment factors could be obtained for all models with a cluster radius from 0.5 Å to 2.4 Å. As can be seen in Figure 4.22 these models only differ in the description of the hydrophobic interactions, while models with 2.5 Å and greater cluster radii differ from the other models in the description of the oxygen atoms of the sulfonyl group. The models with a large cluster radius use a single PPP for the description of these atoms while the models with small cluster radius use two PPPs. It seems that a single PPP for the description of these oxygen atoms is not sufficient for a reasonably performing pharmacophore model. The models from 0.5 Å to 2.4 Å can be divided into 4 groups. The pharmacophore models of the first group from 0.5 Å to 0.9 Å with enrichment factors of roughly 27 consist only of PPPs merging atoms from different molecules within close spatial proximity, e.g., all aromatic rings are described by six PPPs. From 1.0 Å to 1.2 Å a minimum in the performance of the models was observed. In these models ring A is represented by six PPPs, and ring B is represented by

four or five PPPs, which might not be an adequate number for the description of an aromatic six-ring.

The three best performing models were obtained with cluster radii of 1.3 Å to 1.5 Å. Both models from 1.4 Å and 1.5 Å describe ring A with a single PPP and ring B with three and two PPPs respectively. Like within the poorly performing models employing cluster radii from 1.0 Å to 1.2 Å, in the model obtained with a cluster radius of 1.3 Å ring B is represented by four PPPs, but ring A is represented by three PPPs. The larger tolerances of the three PPPs of ring A might have compensated the unfavorable description of ring B. Within the models from 1.6 Å to 2.4 Å the hydrophobic interactions are represented by a decreasing number of five to three hydrophobic PPPs.

For comparison, a pharmacophore model was calculated including the two additional COX-2 inhibitors DFU and celecoxib from the model of Palomer *et al*. [Palomer *et al*., 2002]. A slightly better *ef* for the first 1% of the database (*ef* = 40) was obtained with a model calculated with a cluster radius of 1.5 Å and feature-type weights of 0.2 for hydrogen-bond donors, 0.5 for hydrogen-bond acceptors and 0.5 for hydrophobic interactions (data not shown).

To test if our approach for the optimization of feature-type weights is also valid in situations with significantly fewer reference molecules we repeated the optimization procedure with only the molecules from the pharmacophore model as reference molecules for assessment of the enrichment capabilities of the SQUID models. For all models with cluster radii from 0.5 Å to 2.4 Å several weighting schemes were found that ranked two of the three reference molecules into the first percent of the database. In no case all three molecules were found in the first percent. Ranking of all models according to Eq. 2.7 resulted in four similarly top scoring 1.4 Å models with different weighting schemes. Among these models the previously found best working model was found, with feature-type weights of 0.1 for hydrogen-bond donors, 0.4 for hydrogen-bond acceptors and 0.3 for hydrophobic interactions. The worst of the other three models still resulted in an *ef* of 34 screening the database with the 92 COX-2 inhibitors.

For comparison the maximum *ef* value for COX-2 (all 47 molecules of the first 1% are COX-2 inhibitors) would be 51. Accordingly at least 34 times more molecules were found than expected from a random selection of molecules and at least two-thirds of the COX-2 inhibitors which could be found at all in the first 1% were retrieved with the SQUID fuzzy pharmacophore models. However one has to take care that the actual values of the *ef* cannot be compared between different sets of molecules because the *ef* depends on the total number

of active molecules in the virtual screening database and thus on the a priory probability to find a hit. A low a priori probability for actives results in a larger increase in the *ef* value for each retrieved active molecule than a high a priori probability.

To compare our method with another established method we performed retrospective screenings with the molecules from which the pharmacophore models were calculated. For this approach we encoded these molecules with the CATS3D descriptor, but without scaling the descriptor to a maximum of 1. The database molecules were scored by the Euclidean distance to the query molecule and the database was sorted according to the calculated distances to the query molecule. A comparison of the results of the similarity search with the results obtained from the best SQUID model is shown in Figure 4.24. Rofecoxib performed best in comparison to the other two COX-2 inhibitors. This might be a consequence of its comparably small size. The pharmacophore model performed better than rofecoxib for the first 15% of the database. With the SQUID approach 75% of the active COX-2 inhibitors were ranked into the first 6% of the database. In comparison, rofecoxib retrieved 75% of the actives among the top 16% of the ranked database. Interestingly, the performance of the pharmacophore model decreased significantly for the last 25% of the active molecules in comparison to the COX-2 inhibitors.

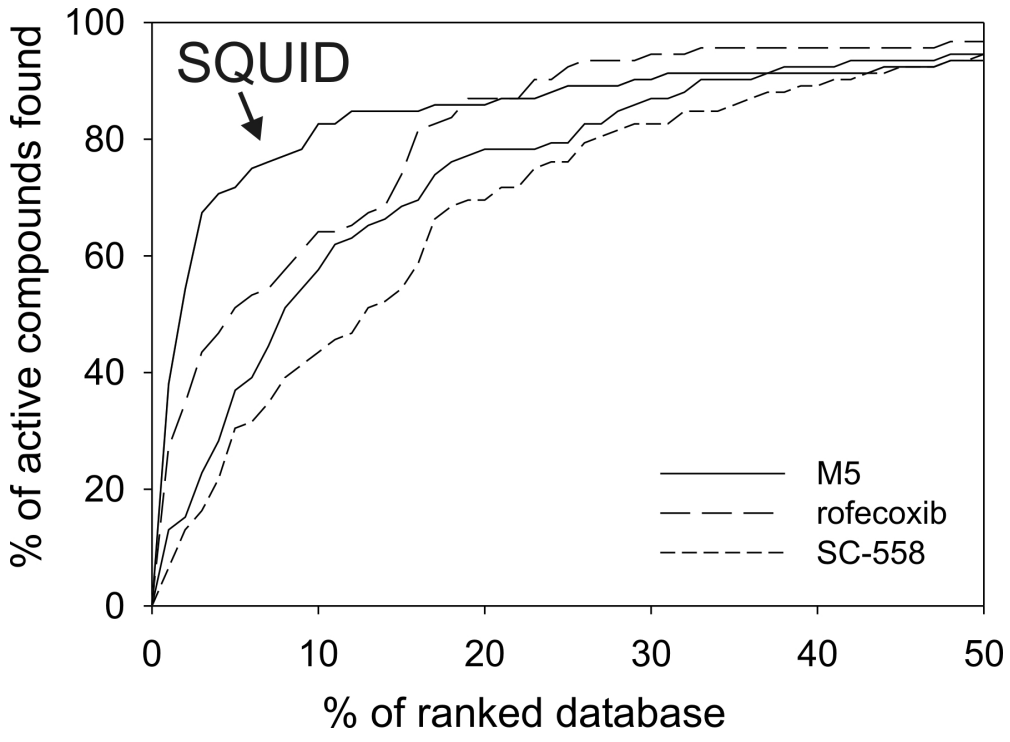


Figure 4.24. Comparison of the enrichment curves of the best COX-2 SQUID model with CATS3D similarity searching using the COX-2 inhibitors from model calculation.

4.6.3 Pharmacophore model of thrombin ligands

A diverse set of seven non-covalent, non-peptidic thrombin inhibitors was adopted from Patel *et al.* [Patel *et al.*, 2002]. The 2D structures of these molecules are shown in Figure 4.25. All ligands were aligned by superposition of the protein structures with the homology alignment tool of MOE. The resulting alignment of the thrombin inhibitors is shown in Figure 4.26. According to Patel and coworkers the major interactions are B, H1, H2 and H3, where B is a basic interaction which interacts with the carboxylic group of Asp189. H1, H2 and H3 are hydrophobic interactions. Less conserved interactions are D1 and A1, where D1 is a hydrogen-bond donor and A1 is a hydrogen-bond acceptor. SQUID pharmacophore models were calculated from the 3D alignment with cluster radii from 0.5 Å to 3.5 Å within steps of 0.1 Å.

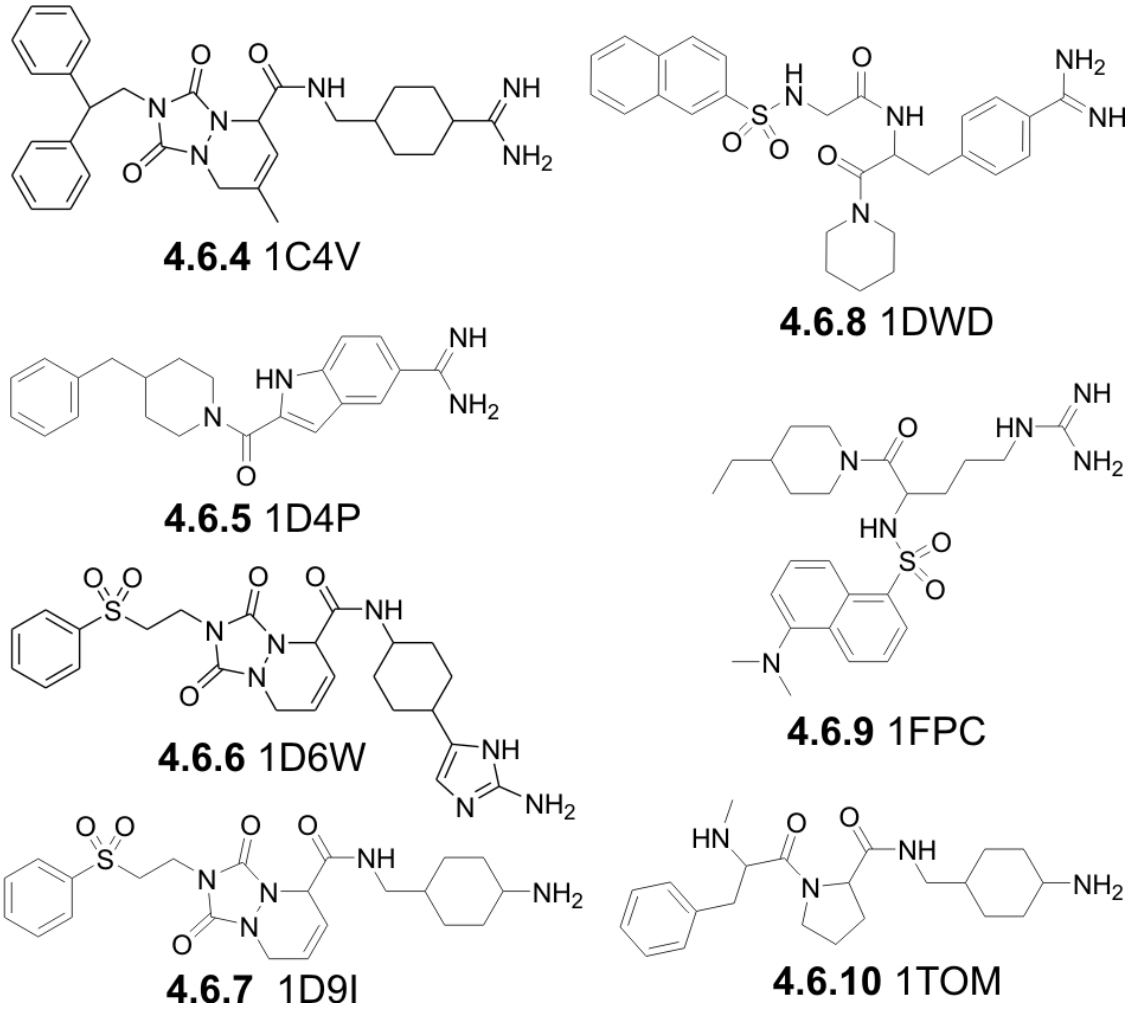


Figure 4.25. Reference thrombin inhibitors used for the calculation of the SQUID pharmacophore model. The names beneath the molecules are the pdb identifiers of the protein structures from which the conformations of these molecules were extracted.

A sample set of the resulting models is shown in Figure 4.27. Four generalized interaction types were found in the ligands based on the ph4_aType function of MOE: hydrogen-bond acceptor, hydrogen-bond donor, polar and hydrophobic. Since all ligands were presented in neutralized state, interaction B was not identified as cationic feature, instead it was represented by hydrogen-bond donor and polar interactions and an additional hydrogen-bond acceptor. In the 1.0 Å and 1.5 Å models the description of the three hydrophobic interactions H1, H2 and H3 is very detailed using a large number of PPPs. With a cluster radius of 2.0 Å only four PPPs are left. In the models with cluster radii of 2.5 Å, 3.0 Å and 3.5 Å these hydrophobic interactions are represented by only three PPPs. Both A1 and D1 are structurally conserved features in the alignment. All appropriate atoms from the different molecules lie in near proximity to each other. A1 is represented by a small conserved PPP in all models except for the 3.5 Å model, where it is represented by a large PPP, including other hydrogen-bond acceptors. D1 is also represented by a small conserved PPP except for the models with 3.0 Å and 3.5 Å cluster radius.

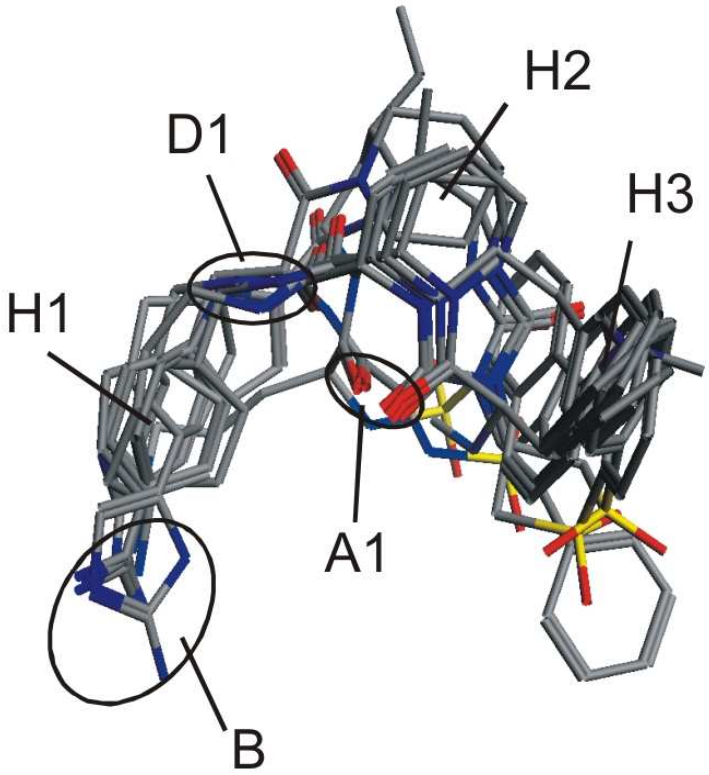


Figure 4.26. Three-dimensional alignment of thrombin inhibitors. The molecules were aligned by superposition of their appropriate protein structures. Essential interactions with the receptor are: B is a basic interaction, H1, H2, and H3 are hydrophobic interactions, A1 is a hydrogen-bond acceptor and D1 is a hydrogen-bond donor.

4.6.4 Retrospective screening for thrombin inhibitors

For retrospective screening with the SQUID pharmacophore models obtained from the alignment of thrombin inhibitors the same procedure for feature-type weight optimization was applied as for the screening for COX-2 inhibitors. For the thrombin optimization 625 weighting schemes had to be evaluated per model.

The results of the optimization procedure are shown in Figure 4.28. The best average enrichment factor of 18 for the first 1% of the database was obtained with the model calculated with a cluster radius of 2.0 Å and feature-type weights of 0.4 for polar, 0.5 for hydrogen-bond donors, 0.3 for hydrogen-bond acceptors and 0.5 for hydrophobic interactions.

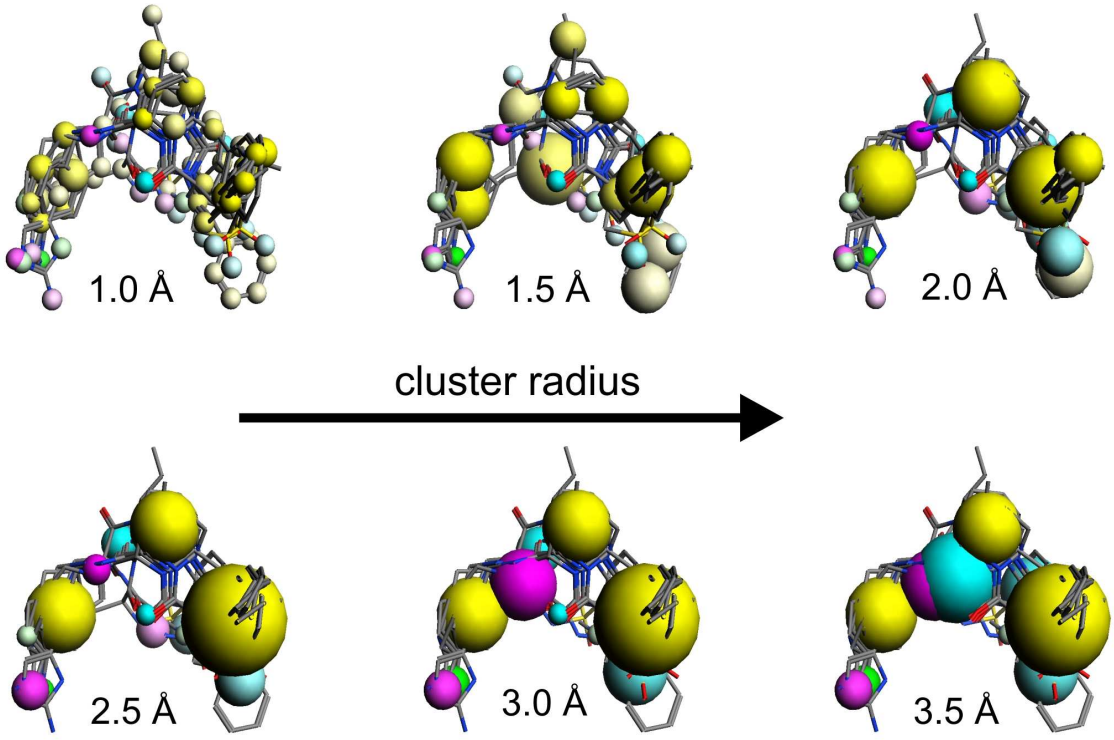


Figure 4.27. SQUID fuzzy pharmacophore models for thrombin calculated from different cluster radii. The Gaussian PPPs of SQUID are represented by spheres. The radius of a sphere denotes the standard deviation of the PPP and the intensity of the color illustrates the conservation weight of the PPP. Yellow = hydrophobic, cyan = hydrogen-bond acceptor, magenta = hydrogen-bond donor, green = polar.

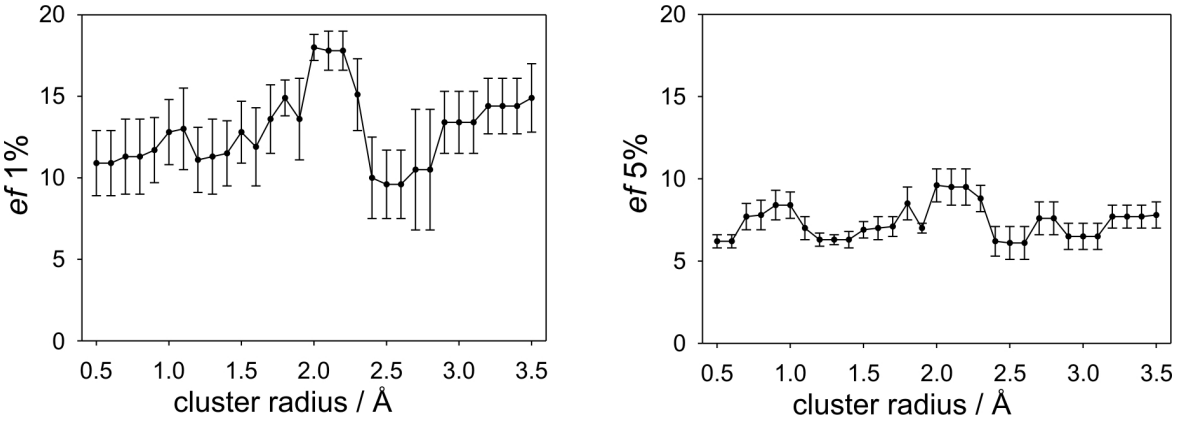


Figure 4.28. Enrichment factors for the first 1% and 5% from retrospective screening with thrombin pharmacophore models with cluster radii from 0.5 Å to 3.5 Å. For each cluster radius the result from the best found feature type weights from the optimization procedure is shown. The shown enrichment factors are average values from screening of five randomly selected subsets of the COBRA database.

We detected two peaks for each of the enrichment factors, one for models with a high degree of generalization with cluster radii from 2.0 Å to 2.2 Å, and one for models with a lower degree of generalization with cluster radii of 1.0 Å and 1.1 Å. Interestingly, models with cluster radii greater than 2.8 Å performed very well, too. As can be seen in Figure 4.27, the model with a cluster radius of 1.0 Å from the first peak mainly clustered atoms within near proximity into PPPs, while already favoring conserved atoms. The model with a cluster radius of 2.0 Å from the second peak represents the features with a drastically diminished overall number of PPPs. In particular the three hydrophobic interactions are represented by four PPPs, in contrast to all other models with a smaller cluster radius. The models resulting from cluster radii larger than 2.8 Å consist mostly of PPPs with large tolerances, but unlike for COX-2, these PPPs represent the shape of the molecular alignment very well.

Like for COX-2 the optimization procedure was repeated with only the molecules from the pharmacophore model as reference molecules. For many models weighting schemes were found which ranked two of the seven reference molecules into the first 1% of the database. In no case more molecules were found in the first 1%. Ranking of all models according to Eq. 2.7 resulted in the previously found best working 2.0 Å model with feature-type weights of 0.4 for polar interactions, 0.5 for hydrogen-bond donors, 0.4 for hydrogen-bond acceptors and 0.5 for hydrophobic interactions.

The result of the 2.0 Å SQUID model was compared with results from retrospective screening with CATS3D descriptors calculated from the molecules used for the calculation of the pharmacophore model. Enrichment curves are shown in Figure 4.29. Major differences were observed in the performance of the individual thrombin inhibitors. The inhibitors from the crystal structures 1FPC and 1DWD performed best. The three inhibitors from structures 1D4P, 1D9I, and 1TOM performed even worse than a random distribution of active molecules within some regions of the ranked database. The SQUID pharmacophore model performed better than the most successful similarity search for the first 40% of the database. 50% of the active molecules were ranked into the first 6% of the database by the pharmacophore model in comparison to the best thrombin inhibitor from 1DWD, which ranked 50% of the active molecules into the first 13% of the database.

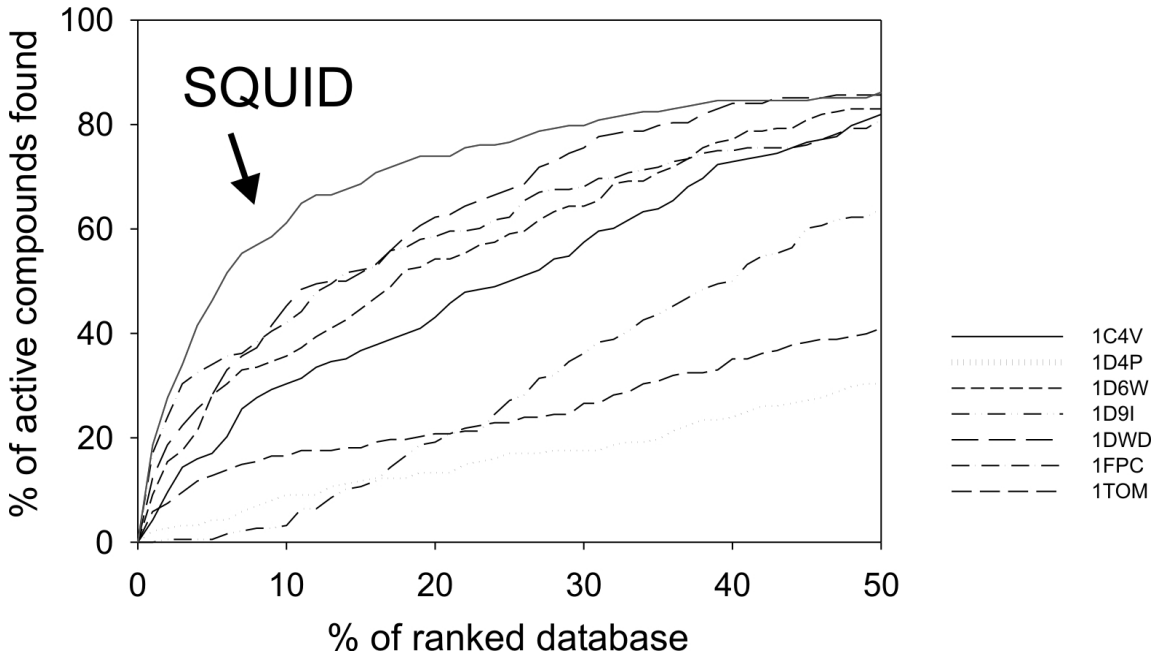


Figure 4.29. Comparison of the enrichment curves of the best thrombin SQUID model with CATS3D similarity searching using the thrombin inhibitors from the model calculation.

4.6.5 Method performance

For an additional comparison of the SQUID pharmacophore model with an established method we calculated pharmacophore models from the two alignments of COX-2 and thrombin reference compounds with the pharmacophore tool of MOE [Chemical Computing Group]. For both models we used the atom-type scheme PCH_ALL which consists of atom-types for cationic, anionic, hydrogen-bond donor, hydrogen-bond acceptor, aromatic ring

centers and hydrophobic interactions. In contrast to SQUID, one PPP in MOE can describe multiple atom-types, which can be combined by logic operators. As a starting point for the alignments pharmacophore models were calculated automatically with the consensus pharmacophore function using MOE default parameters. This function clusters features into PPPs which are more conserved than a threshold value. For the threshold 50 % conservation was used. Retrospective screening with these first pharmacophore models was very slow and the program even failed to screen the whole database due to limitations of the software. As a consequence, we modified the models manually by removing PPPs which were not among the key features of the pharmacophore models published by Palomer *et al.* [Palomer *et al*., 2002] or Patel *et al.* [Patel *et al*., 2002] respectively (Figure 4.21, Figure 4.26). For the thrombin model the radii and the positions of the PPPs for H1, H2 and D1 were manually adjusted for a more accurate representation of the underlying ring structures and the cluster of hydrogen-bond donors. Additional multiple features of the PPPs were also removed. The resulting MOE pharmacophore models are shown in Figure 4.30. Both models were evaluated by retrospective screening of the COBRA database.

With the MOE COX-2 model (Figure 4.30a) we retrieved 84 matching molecules among which we found 49 (58 %) of the known COX-2 inhibitors. In comparison, the COX-2 SQUID model found 47 (56 %) active molecules in the first 84 compounds from the ranked database. Reinsertion of a PPP from the first MOE model, which represents the central five-ring of the COX-2 inhibitors by an acceptor, aromatic or hydrophobic interaction, resulted in 48 actives (91 %) out of 53 matches. Within the first 53 molecules of the ranked database the SQUID pharmacophore model retrieved only 38 (72 %) active compounds. A comparison of the actives found by MOE and SQUID showed that the overlap was only 25 molecules, i.e., that both methods complement each other. SQUID retrieved additional 13 actives which were missed by the refined MOE model.

With the MOE thrombin model (Figure 4.30b) we retrieved 5 actives (31 %) among 16 matches, in comparison to the SQUID model which retrieved 13 actives (81 %) among the first 16 molecules of the sorted database. Retrospective screening with the partial match option of the MOE pharmacophore search function requiring only six of the seven PPPs as matching criterion resulted in 489 matches including 87 (18 %) thrombin inhibitors. With SQUID 119 actives (24 %) were found among the first 489 molecules of the ranked database. The two sets of actives have 64 molecules in common. Again, we conclude that the two pharmacophore searching approaches complement each other.

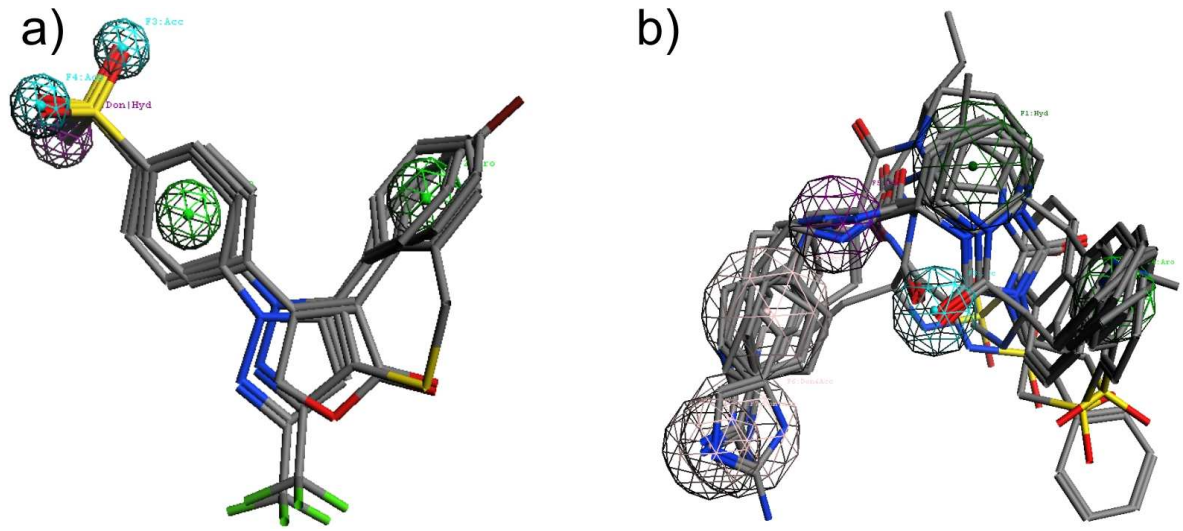


Figure 4.30. MOE Pharmacophore models for COX-2 (a) and thrombin (b). In the COX-2 pharmacophore model the rings A and B are represented by two aromatic ring center PPPs, and the sulfonyl group is represented by a PPP for a donor or hydrophobic interaction. In the thrombin pharmacophore model the hydrophobic interactions H1 and H2 are represented by hydrophobic PPPs while H3 is represented by an aromatic PPP. For A1 a hydrogen-bond acceptor PPP and for D1 a hydrogen-bond donor PPP was found. The basic interaction B was represented by two PPPs, one for hydrogen-bond acceptor or hydrogen-bond donor and one for hydrogen-bond acceptor and hydrogen-bond donor.

To gain further confidence in our approach we took a look at the two top-scoring non-active molecules from each of the best pharmacophore models for COX-2 and thrombin (Figure 4.31). Molecules **4.6.11** [Woo *et al*., 1998] and **4.6.12** [Supuran *et al*., 2003] were found with the COX-2 pharmacophore model with 1.4 Å cluster radius, and molecules **4.6.13** [Marlowe *et al*., 2000] and **4.6.14** [Rudolf *et al*., 1994] were found with the thrombin pharmacophore model with 2.0 Å cluster radius. Ethoxzolamide (**4.6.12**) is an inhibitor of carbonic anhydrase. Also, it has been shown recently that celecoxib is a nanomolar inhibitor of carbonic anhydrase [Weber *et al*., 2004]. EMATE (**4.6.11**) is an inhibitor of estrone sulfatase, and a nanomolar inhibitory effect of EMATE on carbonic anhydrase activity has been reported [Ho *et al*., 2003]. This indicates that both "non-active" molecules share common features with the COX-2 inhibitors from the pharmacophore model.

Molecule **4.6.13** (BOC-D-Arg-Pro-Arg) is an inhibitor of Factor Xa for which nanomolar inhibition of thrombin has been reported [Ho *et al*., 2003]. It thus represents a real hit. BIBP3226 (**4.6.14**) is an antagonist of the neuropeptide Y_1 receptor. To our knowledge thrombin activity has not been tested for this molecule.

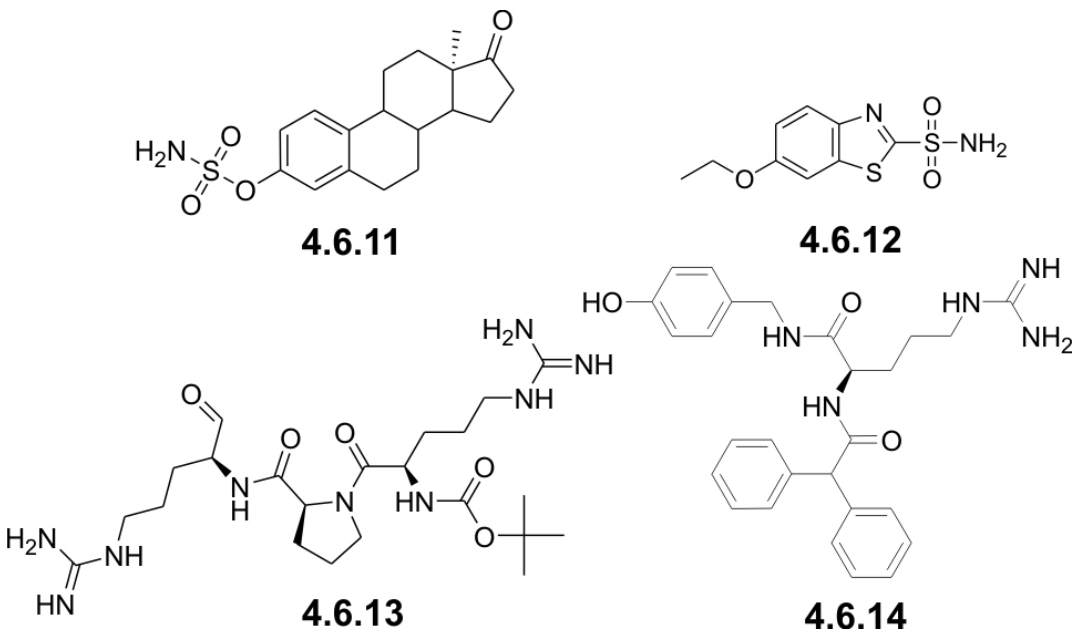


Figure 4.31. Best scoring false-positive hits found with the SQUID fuzzy pharmacophore models. Compounds **4.6.11** and **4.6.12** were found with the COX-2 pharmacophore model with 1.4 Å cluster radius. Compounds **4.6.13** and **4.6.14** were found with the thrombin model derived form 2.0 Å cluster radius.

4.6.6 Conclusion

We challenged our SQUID approach using inhibitors of COX-2 and thrombin. For both classes COX-2 and thrombin SQUID pharmacophore models were able to find an appropriate representation of important pharmacophoric interactions. The optimization procedure was found to be robust in cross-validation using five different randomly sampled subsets of 50 % of the COBRA database. Using only the molecules from the pharmacophore model as references for the optimization resulted in identical (COX-2) or near identical (thrombin) results as for using all active molecules as references. The best retrospective screening results for COX-2 were obtained with the model resulting from a cluster radius of 1.4 Å, yielding an enrichment factor of 39 for the first 1% of the ranked database. For thrombin, the best results for the enrichment in the first 1% of the database were obtained with the model resulting from a cluster radius of 2.0 Å, yielding *ef* = 18. For both targets, the best models outperformed retrospective screening by CATS3D similarity searching. This showed that - independent from the overall enrichment and thus independent of the explicit selection of active molecules - the pharmacophore model outperformed conventional similarity searching. In comparison to conventional pharmacophore searching with MOE, SQUID identified additional actives and thus complements existing methods. We demonstrated that the SQUID pharmacophore model

approach provides a potentially useful new method for virtual screening. The inherent fuzzy description of the molecules should support the goal of ‘scaffold hopping’, especially with higher degrees of fuzziness.

4.7 Prospective screening for inhibitors of the Tat-TAR RNA interaction with a SQUID fuzzy pharmacophore model and CATS3D

RNA is a relatively new target to be tackled deliberately in drug discovery projects. Molecules inhibiting the interaction between the TAR RNA and the Tat protein might be useful to defeat HIV. The first inhibitor found was argininamide (**4.7.1**), a derivative of the arginine which is responsible for specific binding of Tat to TAR [Tao & Frankel, 1992]. So far, only structure-based virtual screening has been reported for TAR, where an automated docking approach including a scoring function optimized for RNA led to the identification of acetylpromazine (**4.7.2**) and chlorpromazine (**4.7.3**) [Lind *et al*., 2002].

An alternative for structure-based virtual screening are ligand-based approaches [Schneider & Böhm, 2002]. Especially methods including the active-analog idea of pharmacophores have been shown to be suited for scaffold-hopping [Schneider *et al*., 1999]. Pharmacophore based similarity searching which was originally developed to identify protein-ligands might be robust enough to identify new RNA ligands without altering the definitions of the pharmacophoric interactions towards RNA specific interactions.

The goal of this study was to enhance the evaluation CATS3D and SQUID with prospective virtual screening experiments. Further the applicability of our alignment-free pharmacophore based virtual screening approaches should be tested for RNA targets. The SPECS compound set [SPECS] containing 229,658 screening compounds was virtually screened for potential inhibitors of the Tat-TAR interaction. Virtual screening consisted of three steps: i) calculation of a “drug-likeness” score by an artificial neural network as a prescreening step, ii) CATS3D pharmacophore similarity searching, and iii) SQUID pharmacophore similarity searching based on the flexible alignment of known active reference molecules. Steps ii) and iii) were performed independently for the 20,000 most “druglike” compounds.

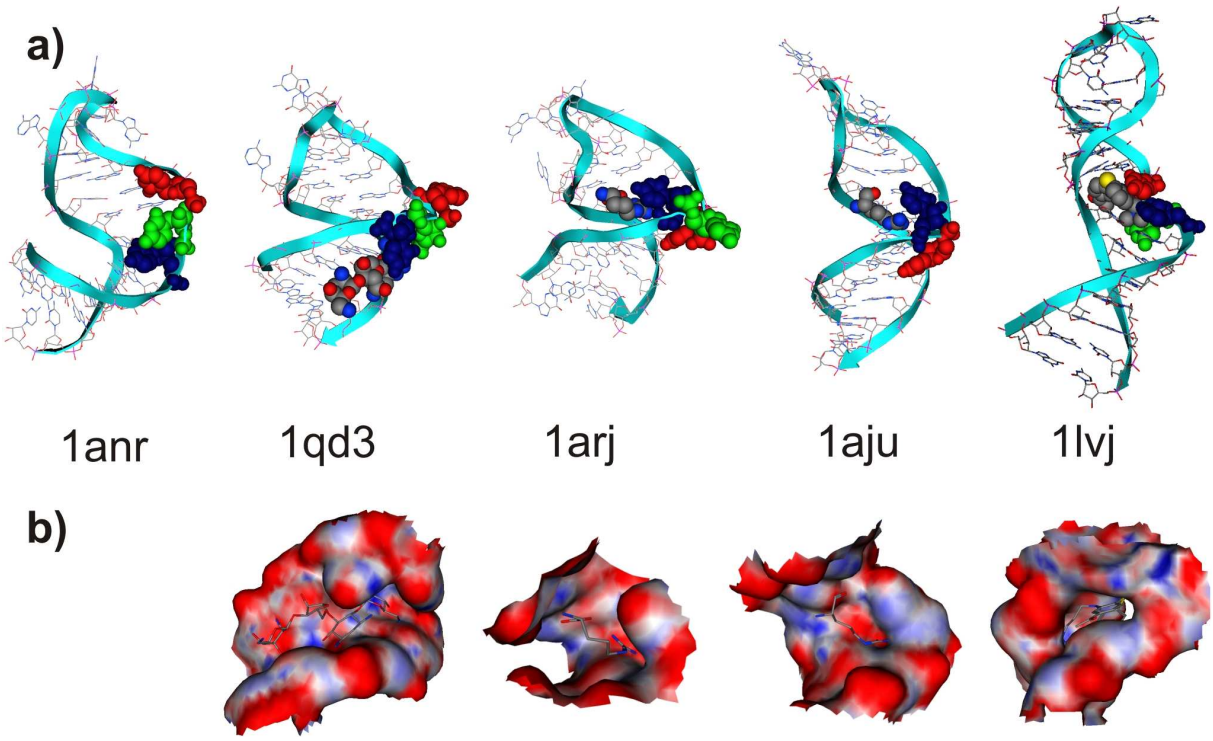


Figure 4.32. NMR structures of TAR RNA. **a)** TAR with bulge binders: 1anr (free TAR), 1qd3 (TAR complexed with neomycin), 1arj (arginine), 1aju (argininamide), 1lvj (acetylpromazine).). The bulge nucleotides are represented in space filling: U23 (blue), C24 (green), and U25 (red). All structures are from HIV-1 TAR except 1aju (HIV-2, C24 is missing). **b)** Binding pockets of the TAR ligands. Surface representation of the binding sites, mapped by electrostatic partial charges (red = negative partial charge, blue = positive partial charge).

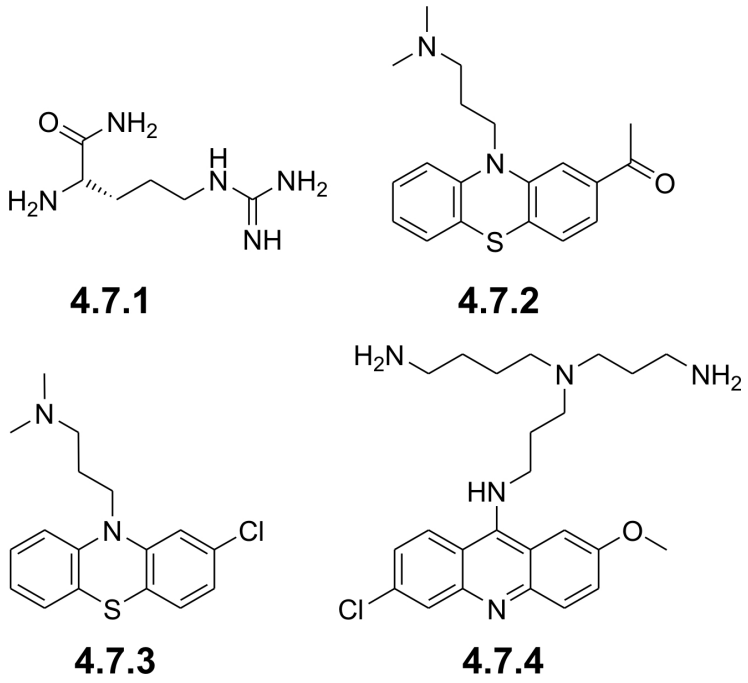


Figure 4.33. TAR-Tat interaction inhibitors. Argininamide **4.7.1**, acetylpromazine **4.7.2**, chlorpromazine **4.7.3**, CGP40336A **4.7.4**.

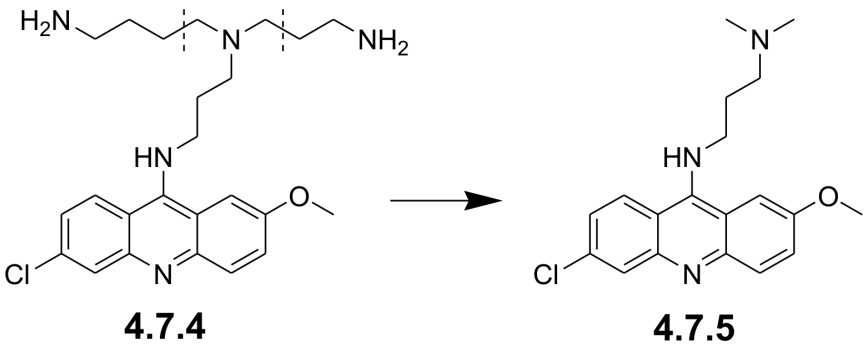


Figure 4.34. Modification of ligand **4.7.4** for the alignment.

4.7.1 Calculation of an alignment of reference compounds

Several NMR structures of the TAR-RNA with bound inhibitors are publicly available (Figure 4.32). Different ligands address different binding-sites and stabilize different conformations of the bulge (Figure 4.32a). Acetylpromazine (**4.7.2**) is bound in a deep binding-site mediated by a combination of stacking and charged interactions whereas the other ligands are bound in shallow binding-sites dominated by charged interactions (Figure 4.32b). Because of the seemingly more druglike relation of ligand-receptor interactions in the acetylpromazine binding-site in comparison to the other sites we decided to design ligands for the former binding-site. Acetylpromazine (**4.7.2**) [Lind *et al*., 2002] and CGP40336A (**4.7.4**) [Hamy *et al*., 1998] (Figure 4.33) were chosen as reference ligands from literature with reported nanomolar IC_{50} values. For both molecules binding to the bulge had been experimentally verified, however detailed structural data was not available for **4.7.4**. As **4.7.4** contains a ring system -- which might be involved in stacking interactions like in **4.7.2** -- and a charged flexible part -- which might interact similar to a potential charge-π interaction of **4.7.2** with C24 [Du *et al*., 2002] --, we assumed that **4.7.4** could have a comparable binding mode as **4.7.2**. For calculation of a SQUID model the two ligands had to be aligned to each other. One possibility would be to dock the reference ligands into the TAR binding pocket; the other possibility is to perform a flexible ligand-based alignment. Since we were not able to reproduce the experimentally determined TAR-bound conformation of acetylpromazine **4.7.2** within the binding pocket using either MOE [Chemical Computing Group] docking or the AUTODOCK approach [Morris *et al*., 1998] (results not shown), we decided to align CGP40336A **4.7.4** to the NMR conformation of **4.7.2** by help of the flexible alignment tool of MOE. Interestingly, fruitless attempts to reproduce the NMR structure of **4.7.2** complexed to

TAR RNA was also reported by Detering and Varani who successfully reproduced many other RNA-ligand complexes using AUTODOCK, but failed to reproduce the acetylpromazine binding mode with an RMSD value below 2 Å [Detering & Varani, 2004]. Their study supports our decision to follow the ligand-based alignment approach. For the alignment calculation we used the first NMR model of the Protein Database entry (PDB code: 1LVJ) [Du *et al*., 2002]. Since it was not possible to predict reasonable conformations of the aliphatic amino groups of **4.7.4** based on flexible alignment alone, we decided to cut off these groups and use molecule **4.7.5** instead (Figure 4.34) for the alignment and virtual screening. The top scoring solutions of the flexible alignment were visually inspected, and we selected the conformation where the ligand appeared to fit best into the receptor (Figure 4.35). Stacking and polar interactions of **4.7.4** occupy the same parts of the binding pocket as acetylpromazine **4.7.2**, so we think that a reasonable starting solution was found.

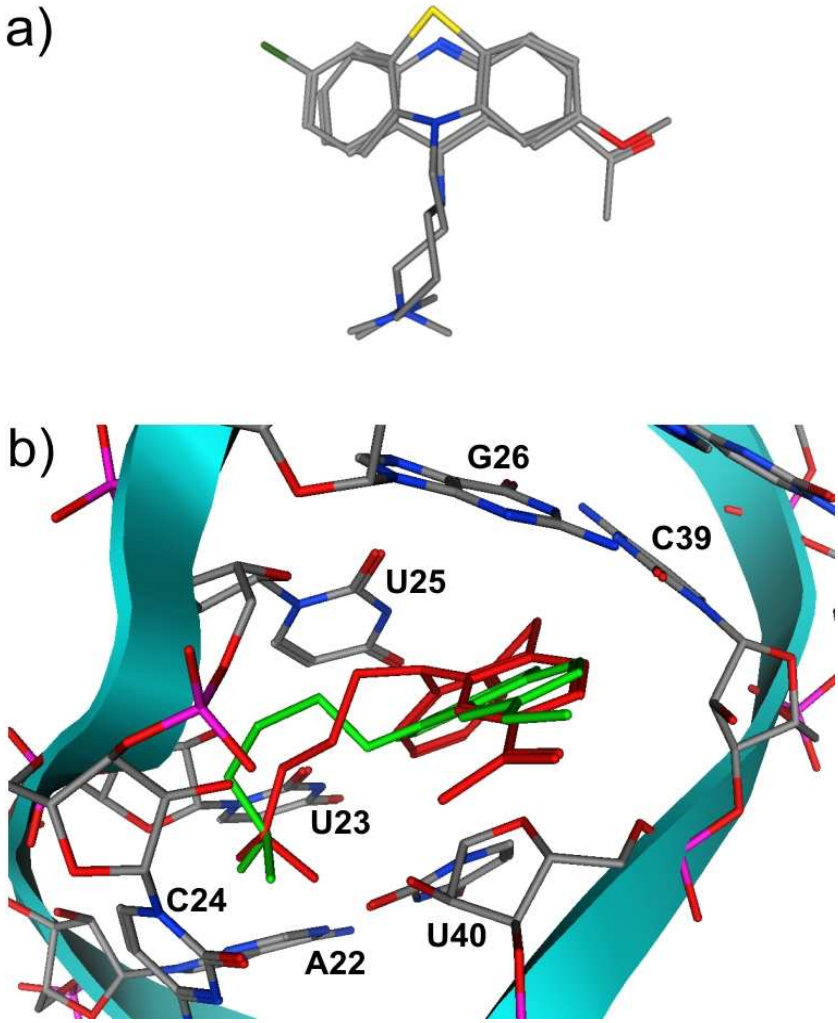


Figure 4.35. Alignment of **4.7.5** to the NMR conformation of **4.7.2** (a) (PDB-code: 1LVJ). The alignment shown in the binding pocket of TAR (b) with **4.7.2** in red and **4.7.5** in green.

4.7.2 Calculation of pharmacophores and virtual screening

For virtual screening with CATS3D we calculated the CATS3D descriptor from those conformations of the reference molecules that resulted from the flexible alignment. For screening with the SQUID pharmacophore model the best resolution of the model, i.e. the optimal PPP cluster radius, and the best weights for the different features had to be determined. The performance of the different parameter sets was determined by their ability to rank the two molecules from the pharmacophore model to top positions in comparison to molecules from the COBRA reference dataset [Schneider & Schneider, 2003] (version 3.12) of bioactive molecules, as described earlier in Section 4.6.

For the optimization cluster radii from 0.5 to 3.0 Å in steps of 0.1 Å were applied. Feature type weights were applied from 0.1 to 0.5 in steps of 0.1 for hydrogen-bond acceptors, hydrogen-bond donors and hydrophobic interactions. This resulted in 125 combinations of feature type weights explored for each of the 26 cluster-radii. For all cluster radii models were found which ranked at least one of the query compounds into the first 1% of the hit-list. Equal best results were obtained with cluster-radii of 1.4, 1.5, and 1.6 Å: the same eight combinations of feature type weights were found for each model ranking both query compounds into the first 1% of the database. The eight combinations were: {(0.1 for hydrogen-bond donors, 0.2 for hydrogen-bond acceptors, 0.3 for hydrophobes), (0.1, 0.3, 0.3), (0.1, 0.3, 0.4), (0.1, 0.4, 0.4), (0.1, 0.4, 0.5), (0.1, 0.5, 0.5), (0.2, 0.4, 0.5), (0.2, 0.5, 0.5)}. For virtual screening we chose the intermediate model with cluster radius = 1.5 Å and weights of (0.1, 0.3, 0.4). The selected pharmacophore model is shown in Figure 4.36.

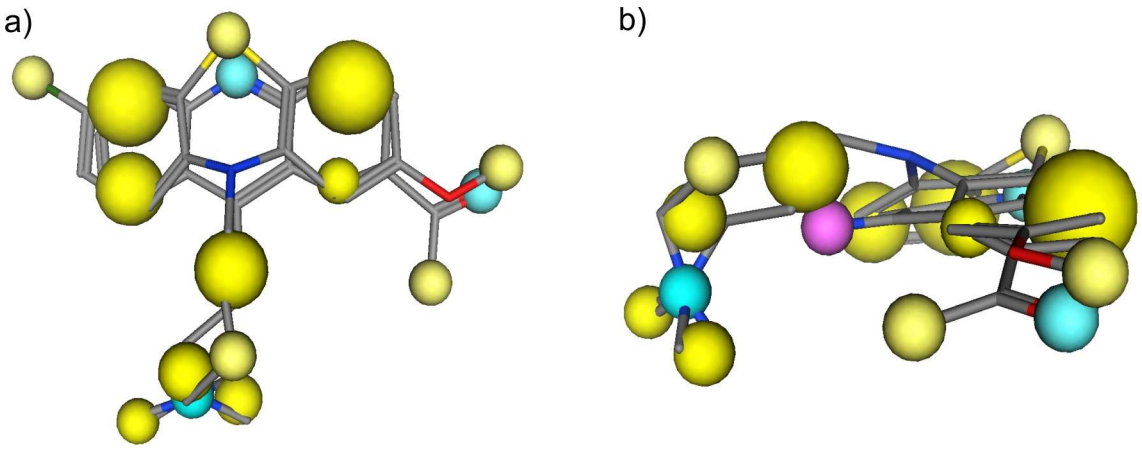


Figure 4.36. SQUID fuzzy pharmacophore model derived from **4.7.2** and **4.7.5** in top-view (a) and side-view (b). The spheres represent the Gaussian PPPs of SQUID. The radius of a sphere denotes the standard deviation of the PPP and the intensity of the color illustrates the conservation weight of the PPP. Yellow = hydrophobic, cyan = hydrogen-bond acceptor, magenta = hydrogen-bond donor.

Three virtual screening experiments were performed with different queries: i) + ii) the two CATS3D CVs which were calculated from molecules **4.7.2** and **4.7.5**, and iii) the CV from the optimized SQUID pharmacophore model. From the results the top scoring database molecules were visually inspected, and a set of 19 molecules (10 molecules from SQUID and 10 molecules from CATS3D, one molecule overlap) was selected for experimental testing (Figure 4.37). To estimate the degree of "scaffold-hopping" of the retrieved molecules the average MACCS Tanimoto similarity of the hits to the respective most similar reference molecules was calculated. For SQUID this similarity was found to be 0.52 ± 0.13 and for CATS3D 0.53 ± 0.11. For comparison the MACCS Tanimoto similarity between the two reference molecules was found to be 0.61. According to this criterion the chemotypes retrieved were more dissimilar to the reference chemotypes than the references to themselves.

4.7.3 FRET determination of the inhibition constants

All 19 molecules were tested for their potency in a Tat-TAR inhibition assay (These experiments were performed by Verena Ludwig and Ute Scheffer in collaboration with the group of Prof. Göbel, Frankfurt). As reference we determined the IC_{50} values of argininamide, acetylpromazine and chlorpromazine -- three inhibitors from the literature with reported values of K_i ~ 1 mM for argininamide [Tao & Frankel, 1992], and IC_{50} < 1 µM for acetylpromazine and chlorpromazine . IC_{50} values in our assay were 1.4 mM for argininamide and 500 µM for acetylpromazine and chlorpromazine [Lind *et al*., 2002]. The strong discrepancy in the IC_{50} for acetylpromazine and chlorpromazine compared to the reported values is in accordance with a recently published article which reported a discrepancy in the same order of magnitude for acetylpromazine (K_D = 270 µM compared to IC_{50} ~ 1 µM, as previously stated) [Mayer & James, 2004]. As a first prescreening of the compounds we performed single-point measurements of the inhibition potency using three fixed concentrations of 10, 100 and 1000 µM of the candidate molecule. Molecules **4.7.14** (hit form SQUID) [Tugusheva *et al*., 1998] and **4.7.21** (hit from CATS3D with reference molecule **4.7.5**) [Shanazarov *et al*., 1989] (Figure 4.37) showed a stronger inhibition than argininamide in the single point measurements. Multipoint measurements yielded IC_{50} values of 46 µM and 500 µM for **4.7.14** and **4.7.21**, respectively.

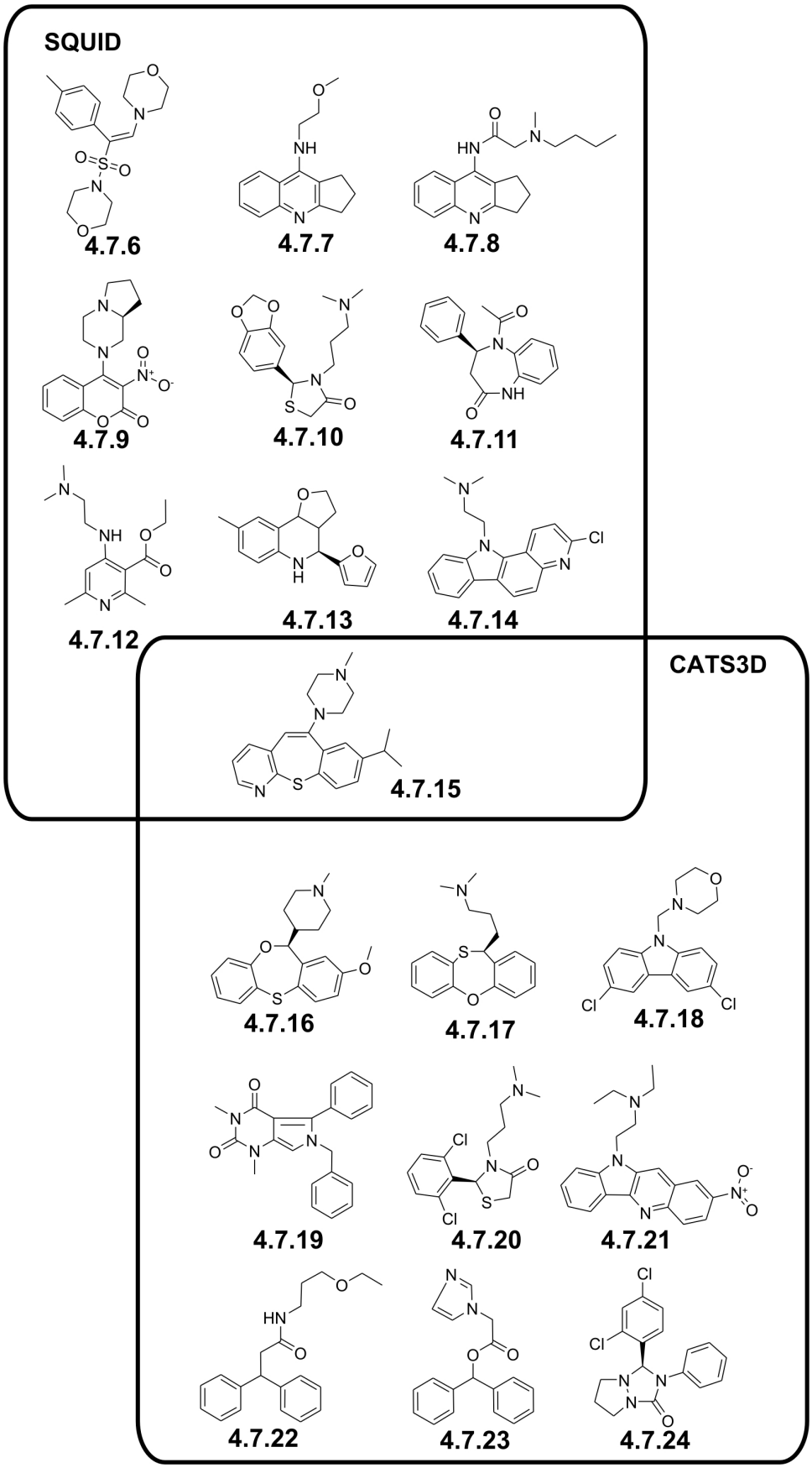


Figure 4.37. Molecules selected from SQUID and CATS3D virtual screening.

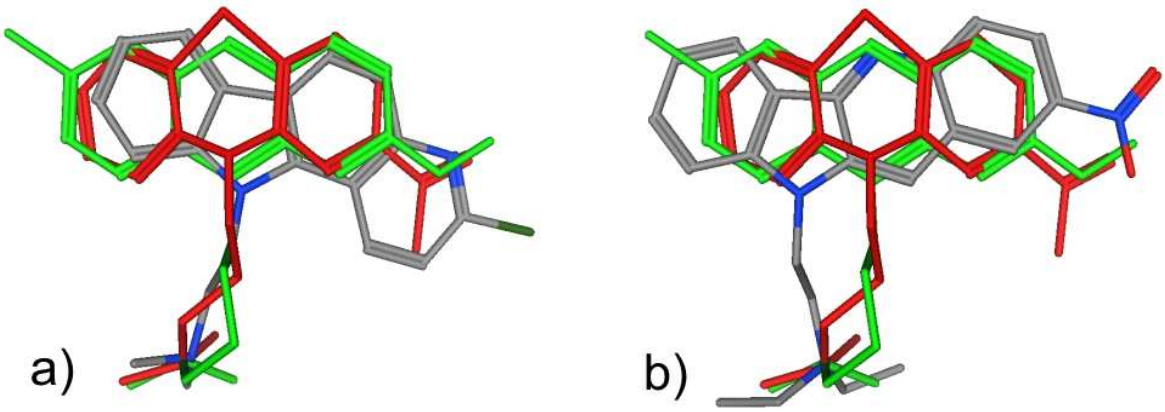


Figure 4.38. Flexible alignment of **4.7.14** (a) and **4.7.21** (b) to the aligned reference molecules **4.7.2** (red) and **4.7.5** (green).

The two ligand-based pharmacophore methods were able to perform "scaffold-hopping", retrieving isofunctional but slightly different molecular scaffolds from the SPECS catalogue. Both new ligands contain a central structure consisting of three rings with an aliphatic amide side-chain, like the reference compounds. An additional aromatic ring is present at different positions in both molecules, extending the original ring systems to four concatenated rings. Flexible alignments of **4.7.14** and **4.7.21** (Figure 4.38) revealed that **4.7.14** fits better to the reference alignment than **4.7.21.** Also, the aliphatic amide side-chain of **4.7.14** was closer aligned to the corresponding side-chains of the references. The nitrogen of the additional pyridine ring of **4.7.14** was positioned directly above the potential hydrogen-bond acceptors of the reference molecules. In both **4.7.14** and **4.7.21** the additional ring might be used for more favorable stacking interactions with the receptor. In **4.7.21** this potentially favorable effect might have been compensated by steric stress due to an unfavorable orientation of the ring or the amide side-chain. Still the IC_{50} value is comparable to acetylpromazine and chlorpromazine.

4.7.4 Conclusions

In this study we presented the application of two ligand-based virtual screening approaches for the compilation of a small focused library containing potential TAR RNA ligands. Among the 19 molecules tested we found two molecules which were able to inhibit the Tat-TAR interaction in a FRET assay. The SQUID fuzzy pharmacophore approach yielded the most

potent molecule with an improved activity of one order of magnitude compared to acetylpromazine **4.7.2** or chlorpromazine **4.7.3**. This could be an effect of incorporating information from multiple active reference molecules into the pharmacophore-based search for new TAR ligands.

Ligand-based approaches provide a complementary concept to structure based design, which might be hampered by the large inherent flexibility of RNA targets. Though it has been shown that specific parameterization of scoring functions is not essential for ligand docking to RNA it is still significantly slower than a ligand-based approach [Detering & Varani, 2004]. It has been demonstrated that ligand-based pharmacophore approaches are capable of finding new RNA ligands. Although the best molecule resulted in a moderate IC_{50} of only 46 µM in the FRET assay this molecule might provide a starting point for further optimization. Certainly, other assay types will be needed to confirm and further scrutinize these findings. The new inhibitors might not represent ideal candidates for starting a lead optimization project. Additional experiments will have to be performed addressing the question which role the additional ring system actually plays for RNA recognition and binding affinity. Furthermore, structures **4.7.14** and **4.7.21** might be intercalating agents and exhibit unspecific binding to both RNA and DNA targets due to the planar ring systems and relatively high lipophilicity. Such issues could also be addressed in a different setting of the virtual screening approach. For example, to obtain selectivity towards RNA, known DNA-binders and intercalators might be used as negative examples for similarity searching. This tactic is currently pursued in our laboratory.

Irrespective of the outcome of such analyses, both ligand-based methods have proved to be useful for finding new molecules within the activity range of known reference compounds. Notably both approaches were originally developed for protein ligands, but they also seem to be applicable to virtual screening for RNA ligands. To our knowledge this study presents the first inhibitors of an RNA-protein complex found by ligand-based virtual screening.

4.8 Prospective screening for taspase1 inhibitors with a receptor-derived pharmacophore model

When no ligand and no receptor structure information are available for virtual screening alternative approaches have to be applied. One possibility is to predict the protein structure with a homology model and utilize the predicted structure for virtual screening [Hillisch *et al*., 2004; Bissantz *et al*., 2003; Grüneberg, 2005; Evers & Klebe, 2004; Evers & Klabunde, 2005]. Homology models are better described as a good approximation of the real protein structure than as a high accuracy replicate. Hence, high-throughput docking studies can be misleading, when based on homology models alone. An alternative to this approach are receptor-derived pharmacophore models [Wolber & Langer, 2005; Pirard *et al*., 2005]. Such an approach was used to search for a first inhibitor of human taspase1. Taspase1 is a threonine protease [Hsieh *et al*., 2003] and hence the problem of finding an inhibitor for taspase1 involves the problem of scaffold-hopping form a peptide substrate to a drug-like molecule inhibitor. This is a comparably complex task for ligand-based pharmacophore-descriptor approaches like the CATS descriptor that are often hampered by the many potentially interacting groups in peptides [Sheridan *et al*., 2001]. Thus a pharmacophore model focusing on small numbers of relevant interactions might be favorable for this task.

The sequence of the human taspase1 from the swiss-prot entry Q9H6P5 was used for a BLAST [Altschul *et al*., 1997] search for related protein structures from the PDB database [Berman *et al*., 2000]. Protein structures were selected with a significant similarity in both subunits of taspase1. Mutant proteins were discarded. An overview of the PDB structures which were finally selected is given in Table 4.16.

1T3M [Prahl *et al*., 2004] is an isoaspartyl peptidase with an additional L-asparaginase activity [Hejazi *et al*., 2002; Borek *et al*., 2004] (Figure 4.39). 2GAW [Guo *et al*., 1998] and 1APZ [Oinonen *et al*., 1995] have a glycosylasparaginase activity and also an L-asparaginase activity [Noronoski *et al*., 1997, Tarentino & Plummer, 1993] (Figure 4.39). All activities include the hydrolysis of a beta-N amide linking an aspartate and varying substituents. In taspase1 there is also an amide bond hydrolyzed (Figure 4.39): the peptide bond between aspartate and glycin. Though the glycin is not bound to the beta amide of an asparagine, the sidechain carboxyl group of this aspartate might interact similarly to the free carboxyl group of the asparagine of the other enzymes. For all enzymes, isoaspartyl peptidase, glycosylasparaginases and taspase1 is has been demonstrated that they undergo

autoproteolysis as an activation step, which is mediated by the same reaction centre as used for the enzymatic activity [Xu *et al*., 1999, Hsieh *et al*., 2003]. Accordingly all actives sites have a proteolytic activity.

Table 4.16. Selected reference protein structures for homology modeling of taspase1. Identities were determined according to the BLAST alignment. (NT = N-terminal subunit; CT = C-terminal subunit).

PBD code		1T3M	2GAW	1APZ
% sequence identity	NT:	42 / 151 (27%)	29 / 87 (33%)	22 / 48 (45%)
	CT:	35 / 70 (50%)	22 / 68 (32%)	29 / 98 (29%)
e-value	NT:	$4e^{-9}$	$4e^{-5}$	$3e^{-5}$
	CT:	$7e^{-12}$	$4e^{-6}$	$2e^{-5}$
crystal structure resolution		1.65 Å	2.2 Å	2.3 Å
enzymatic function		isoaspartyl peptidase / L-asparaginase	glycosyl-asparaginase	glycosyl-asparaginase
organism		*Eschrichia coli*	*Flavobacterium meningosepticum*	*Human*

The structure 1APZ is a co-crystal structure of glycosylasparaginase with the reaction product aspartate. Mutagenesis experiments of residues near to the bound aspartate identified a set of eight amino-acids essential for the catalytic activity [Liu *et al*., 1998]. An overview of the spatial orientation of these sidechains with respect to the bound aspartate is given in Figure 4.40. T152 is the key functionality providing the nucleophile for the hydrolase reaction. The hydroxyl of T170 contributes to the reaction rate. D183 and R180 bind to aspartate via hydrogen or ionic bonds to the alpha amino- and the alpha carboxy-group. W11 is involved in the regulation of the enzyme reaction rate. S50, D66 and T203 were also shown to be important for the enzymatic activity. A ninth important residue is revealed by the crystal structure 1APZ: G203 which shows a hydrogen bonding interaction with the aspartate mediated via the back-bone oxygen of the glycine.

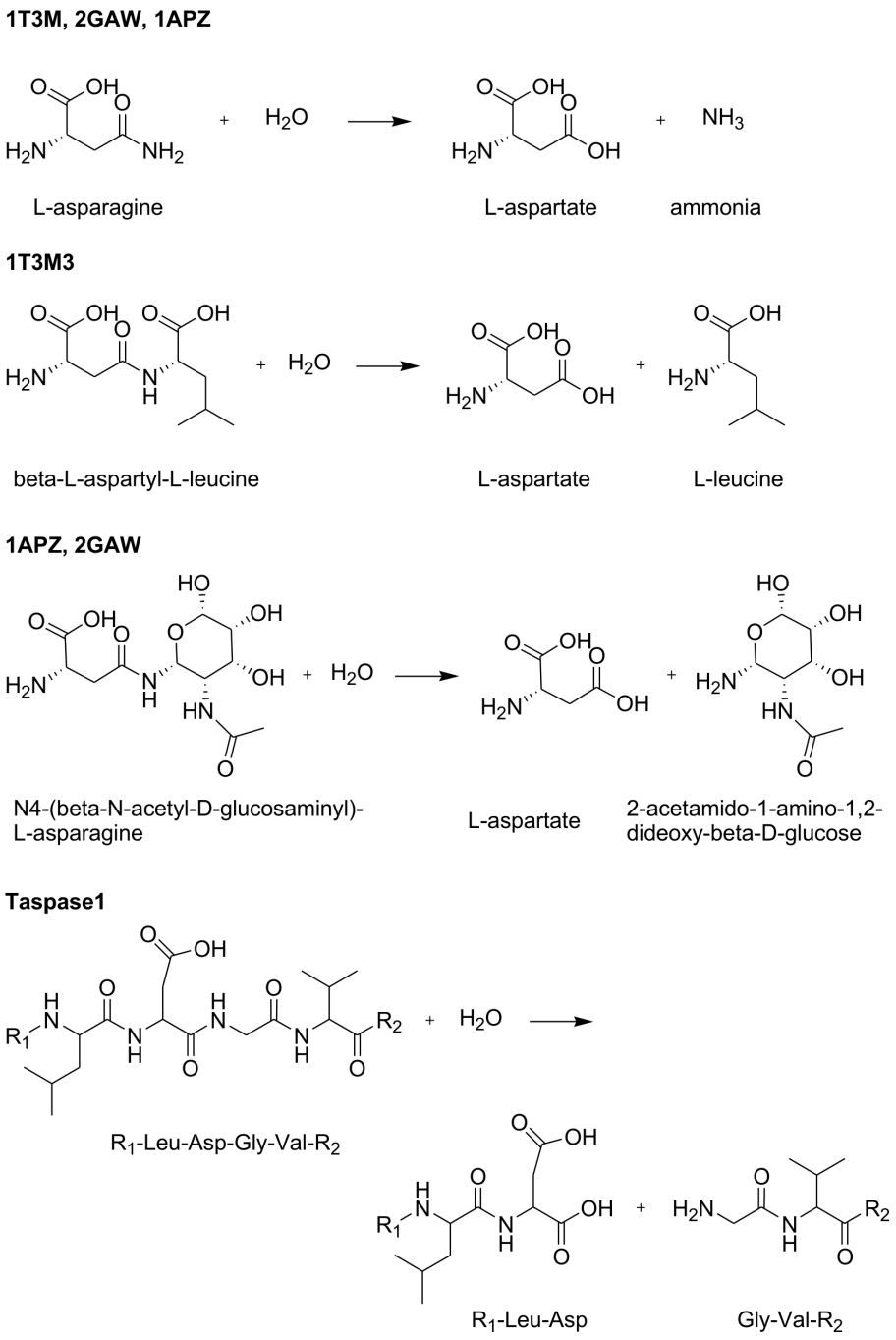


Figure 4.39. Reactions catalyzed by the isoaspartyl peptidase (PDB code 1T3M) and the glycosylasparaginases (PDB code 2GAW and 1APZ) in comparison to taspase1.

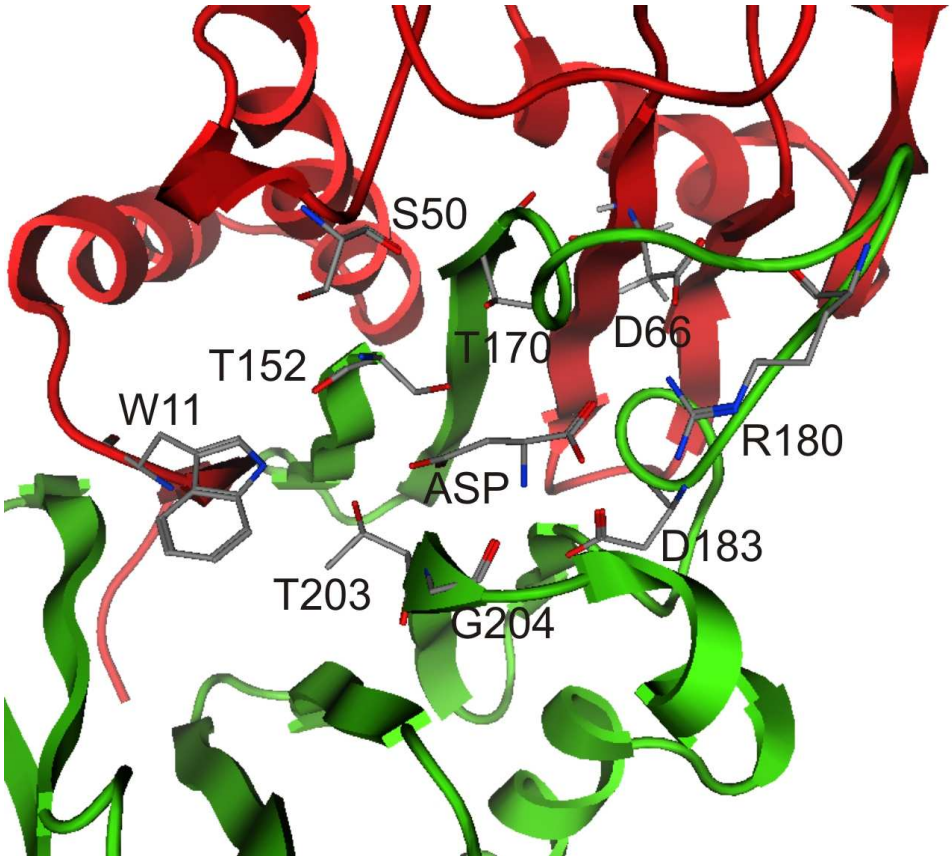
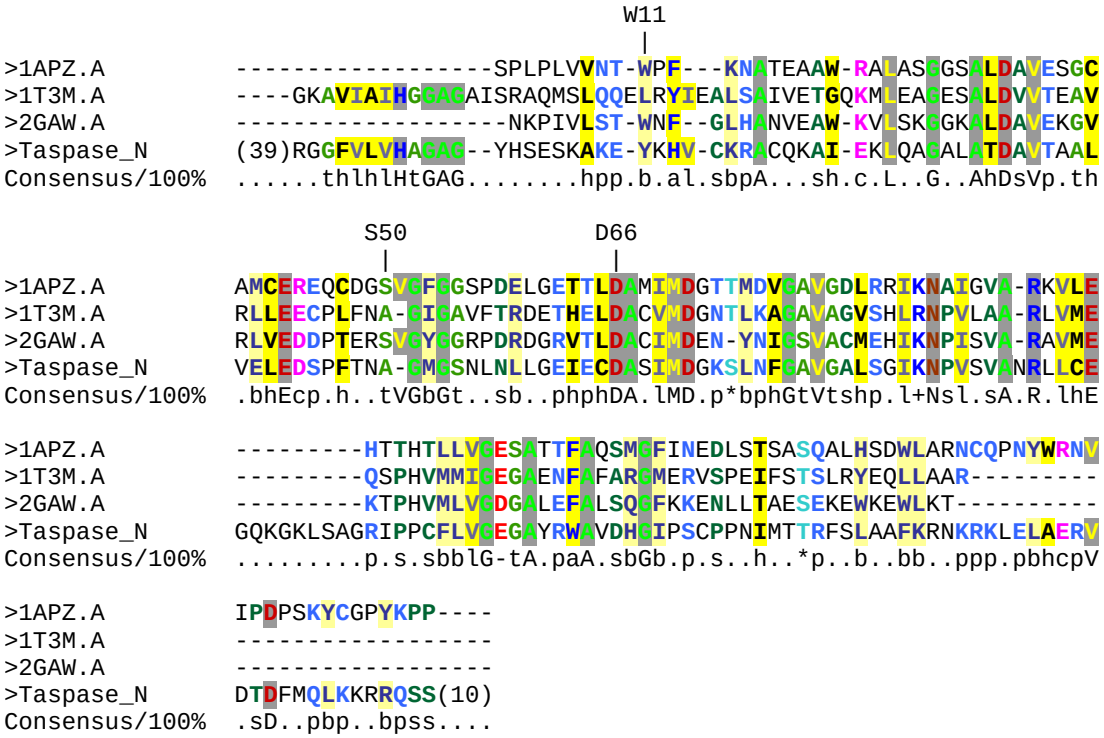


Figure 4.40. Aspartate bound in the active site of 1APZ. Essential residues in the binding pocket are indicated.

To construct a template for the calculation of a homology model of taspase1 all three crystal structures were aligned based on the structure of the enzymes with the homology align tool in MOE [Chemical Computing Group]. This structural alignment was used as a template to align the taspase1 sequence. The resulting alignment is shown in Figure 4.41. All residues considered essential for the enzymatic activity in 1APZ [Liu *et al*., 1998], except for W11, were fully conserved (D66, T152, R180, G204) or replaced by isofunctional amino-acids (S50A, T170S, D183Q, T203S). The full conservation of the reaction center T152, the carboxyl interaction partner R180, the hydrogen-bond acceptor interaction of G204, and D66 reflects the similar reaction of the enzymes and underlines the possibility of a similar binding mode of the ligands.

N-terminus



C-terminus

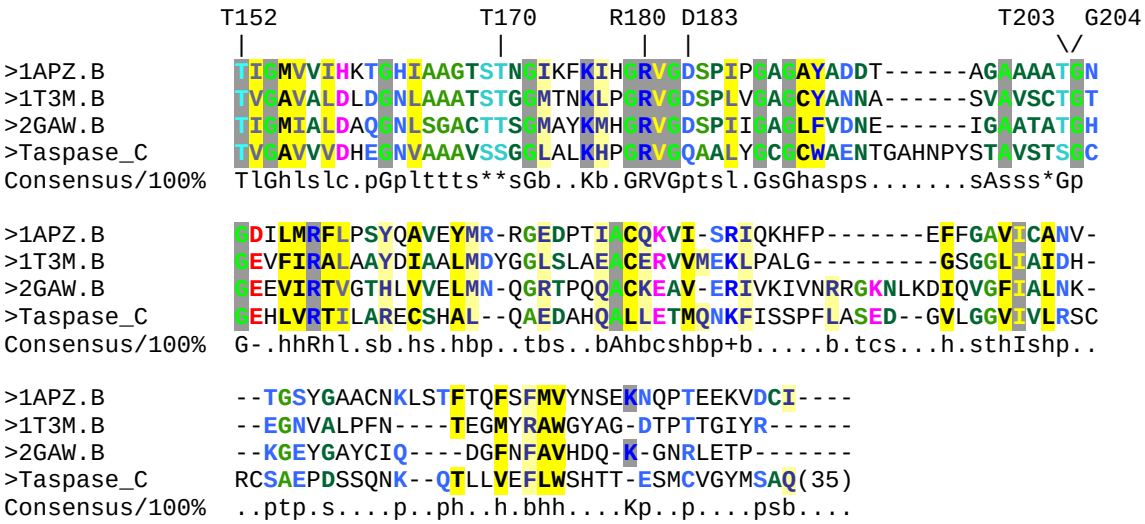


Figure 4.41. Alignment of the C-terminal and the N-terminal taspase1 sequences to the structural alignment of 1T3M, 1APZ and 2GAW. Consensus symbols other than residue letters are: - = negative, * = ser/thr, | = aliphatic, + = positive, t = tiny, a = aromatic, c = charged, s = small, p = polar, b = big, h = hydrophobic. Essential residues for catalytic activity in glycosylasparaginase are marked according to the numbering of [Liu et al., 1998]. Amino acid symbols at alignment positions that are 100% conserved over proteins that have a residue at the respective positions were colored. Conservation was first considered on the level of residue identity. For non-identical residues conservation was further considered on the level of similar biochemical properties: red = negative, cyan = S/T, grey highlighted yellow = aliphatic, dark blue = positive, light green = tiny, dark blue highlighted yellow = aromatic, pink = charged, dark green = small, light blue = polar, light blue highlighted yellow = big, black highlighted yellow = hydrophobic.

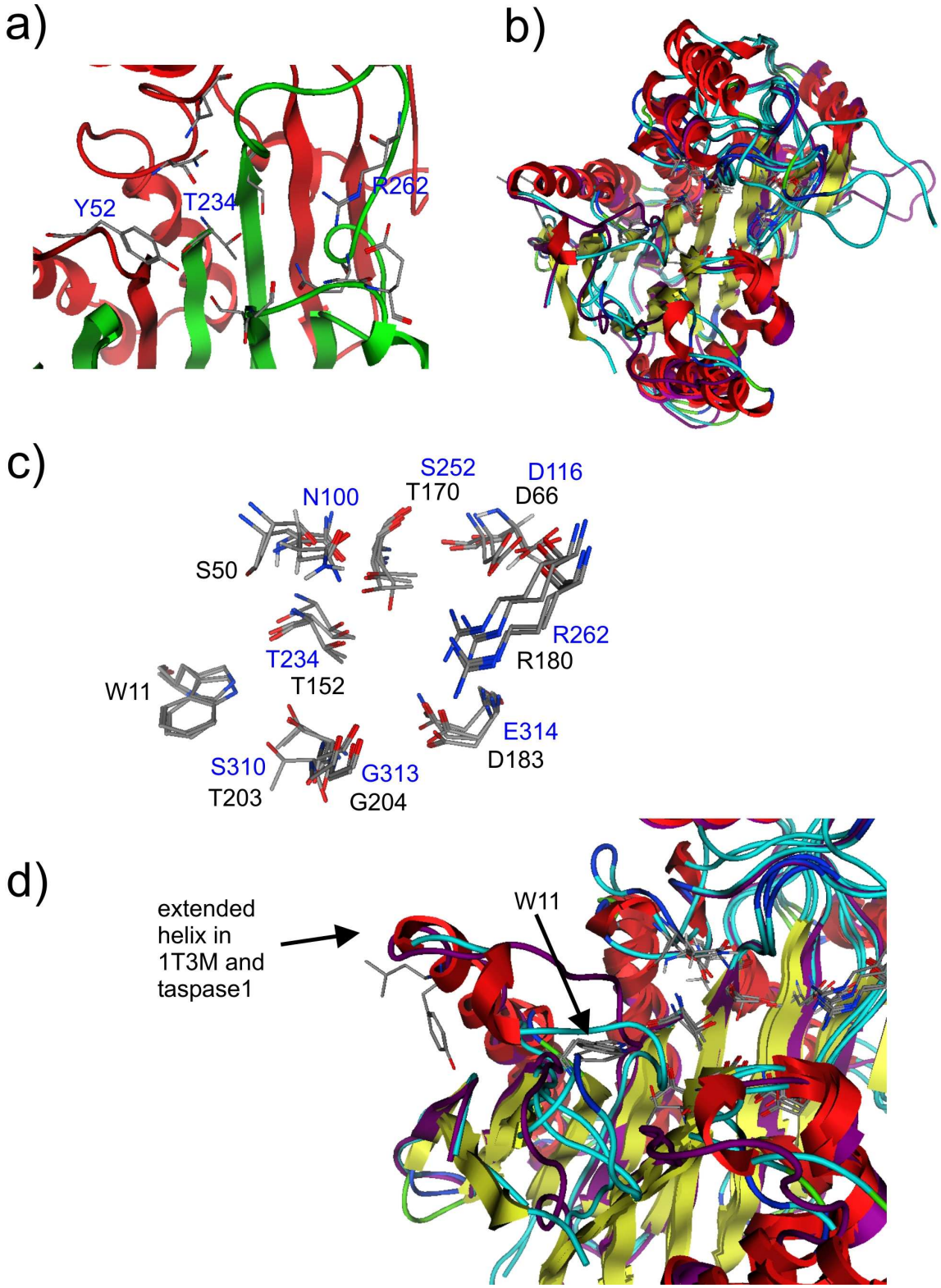


Figure 4.42. Evaluation of the taspase1 homology model. a) Binding pocket of taspase1 with essential residues. Y52 was the only binding-site residue flagged in the MOE protein report to have unusual backbone torsion angles. b) Structure alignment of the taspase1 homology model with the reference structures 1T3M, 2GAW and 1APZ. Comparison of the protein structures. c) Comparison of the active site residues (black. essential residues in 1APZ, blue: aligned residues in taspase1). d) Comparison of the loops after helix one. Large differences in the protein structures are found in this region.

A homology model was calculated for taspase1 with MOE using the alignment from Figure 4.41. The Cartesian average of ten homology models was used for further experiments. The model was minimized including solvatation terms and chiral constraints with the MMFF94xx forcefield. The protein structure quality of the model was controlled with the protein report function in MOE (see Appendix 6.2). Dihedral torsion deviations were mostly observed in loop regions distant from the binding site. Most relevant for the model might be an omega torsion angle deviation of Y52, which is directed into the binding pocket (Figure 4.41a). Since this deviation only affected the back-bone it might have no effect on the binding site geometry. It was left unchanged.

To assess the similarity of the taspase1 homology model to the template structures, the calculated structure was aligned with the reference structures with the homology alignment tool of MOE. The calculated structure of taspase1 fitted well to the reference structures (Figure 4.41b): according to our model, the protein core with the beta-sheet and the flanking alpha-helices was structurally conserved over the enzymes. Differences were found in the loop connecting the secondary structure elements. All essential residues, except for W11, were found to be aligned (Figure 4.41c). After the new structural alignment including the taspase1 model the S50 position of the glycosylasparaginases was aligned to glutamine in taspase1 and isoaspartyl peptidase, which seems more reasonable than the former alignment to the directly neighboring alanines: serine and glutamine are both capable to perform hydrogen-bonding interactions while alanine is not able to do so. For the W11 position no structural alignment was found. A reason for this can be seen in Figure 4.41d. Both in the isoaspartyl peptidase and in taspase1 the loop where W11 is located was found displaced due to an extended alpha helix, connected to the loop. Both, the former “aligned” leucin from isoaspartyl peptidase and the tyrosine from taspase1 did not fit to the tryptophanes (Figure 4.41d). The extended helix and the connected loop of taspase1 also fitted poorly to the respective structural element of 1T3M. Accordingly this region of the model seems to be the least reliable with respect to the binding site. However the remaining part of the binding site covering the essential residues except W11 was conserved between the enzymes and thus provided a reliable structural basis for the understanding of the taspase1 activity and virtual screening for inhibitors.

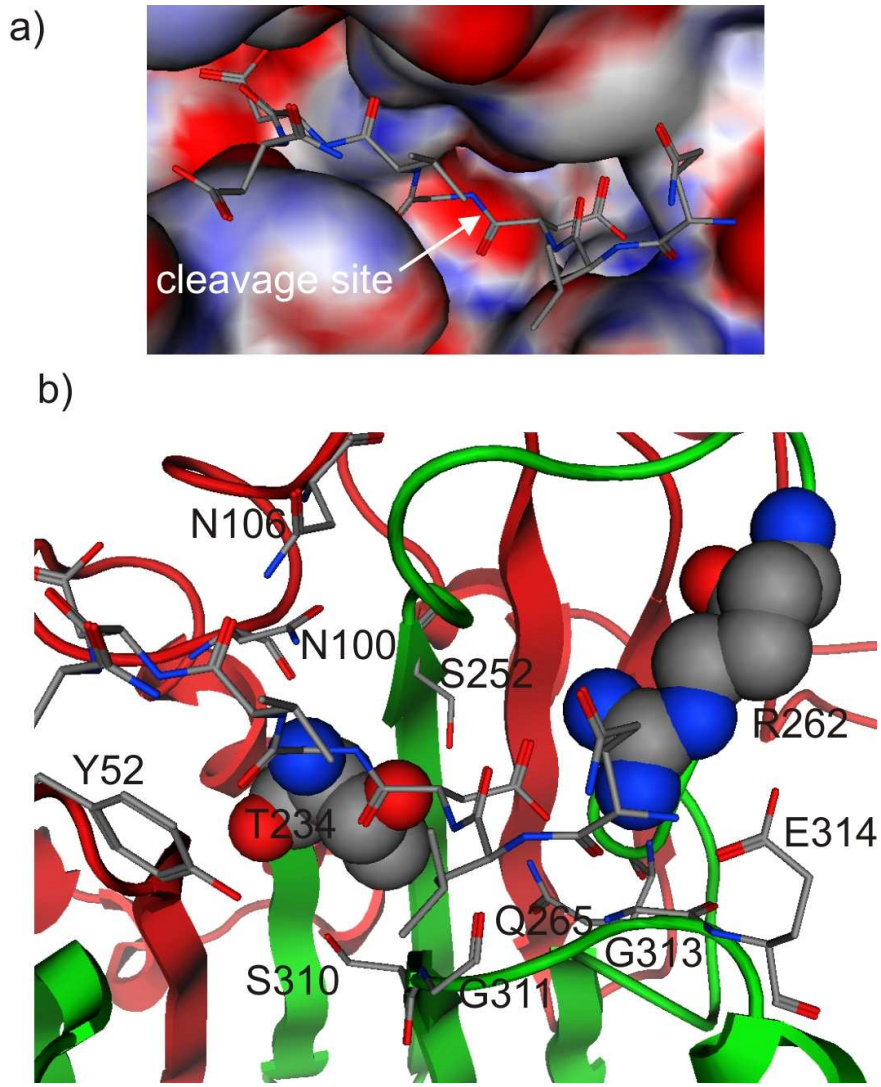


Figure 4.42. Docking solution for the peptide QLDGVDD in the taspase1 active site. a) Surface representation of the active site, colored according to the electrostatic potential. b) Potential interacting residues of taspase1 with the peptide.

To test whether the cleaved sequence of the MLL protein can bind in a comparable manner to the substrates of the reference proteins we applied a docking approach. For the docking we used the peptide QLDGVDD. To avoid unwanted charged interactions the C-terminus of the peptide was amidated for the docking experiments. The N-terminal alpha-nitrogen of threonin was set positively charged. The best docking solution is shown in Figure 4.42. As expected the peptide bond between aspartate and glycine was directly situated above the reactive T234. The sidechain carboxylate of the peptide aspartate was found to interact with R262. The absence of an amino group beside this carboxy-group was compensated by the mutation of aspartate to glutamine at position 265. A hydrogen bonding interaction

between G311 and the backbone nitrogen of the cleaved aspartate in the peptide was also found.

The interpretations of the docking solution have to be taken with care: it is comparably more difficult to retrieve the correct receptor bound conformation of a peptide than for small molecules by docking calculations [Liu *et al.*, 2004]. Peptides are much more flexible than drug-like molecules and thus provide a much larger set of possible docking solutions that have to be ranked by a scoring function. The scoring of the docking solutions is already a non-trivial task for drug-like molecules [Kitchen *et al.*, 2004] and might be even more complicated for peptides. Given that the position of the peptide bond to be cleaved was placed directly above the nucleophile and the carboxylic group of the aspartate at the cleavage-site was found to interact with R262, the found docking solution seems to provide a reasonable starting hypothesis for the ligand binding-mode.

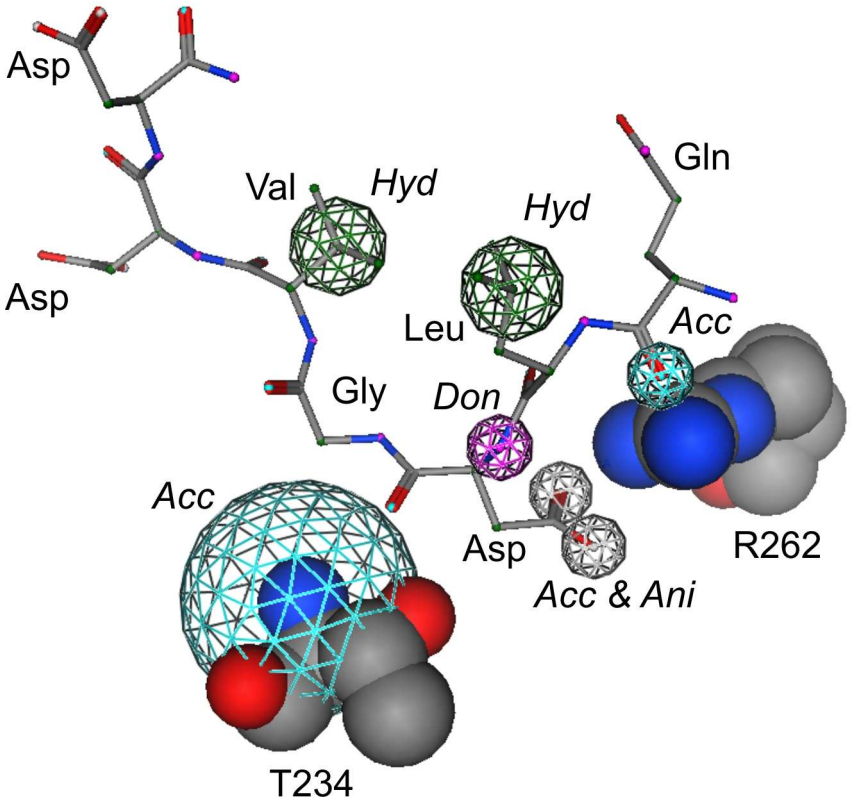


Figure 4.43. Pharmacophore hypothesis derived from the docked peptide and interactions with the receptor. The hydrogen-bond acceptor (Acc) interaction with T234 and the hydrogen-bond acceptor AND anion (Acc & Ani) interaction were defined essential for virtual screening. Hyd = hydrophobe, Don = hydrogen-bond donor.

Using the bound peptide and the receptor structure a pharmacophore model (Figure 4.43) was created for virtual screening for ligands inhibiting the taspase1 reaction. Three interactions were defined as essential, i.e. interactions which had to be satisfied by a molecule to be considered as hit. Essential interactions were: a hydrogen-bond acceptor function within a radius of 3 Å around the N-terminal nitrogen of T234 and two (hydrogen-bond acceptor AND anion) functions within a radius of 0.8 Å around the two oxygens form the aspartate carboxyl-group interacting with R262. Other hydrogen bonding interactions were defined at the positions of the two substrate interaction partners of G311 and G313. Two hydrophobic interactions were defined within a radius of 1.2 Å around the beta and gamma carbon atom of the hydrophobic sidechains of valine and leucine of the substrate, respectively. All protein atoms were defined as excluded spheres. To be considered as a hit, a molecule had to satisfy at least four of the pharmacophore points.

We screened the SPECS database of compounds (june 2003 version). For the virtual screening experiment the database was preprocessed in the following way with MOE:

1) Acids and bases were set charged.
2) All molecules were discarded with: > 500 Da, logP > 5, > 5 hydrogen-bond donors, > 10 hydrogen-bond acceptors, > 10 rotatable bonds.
3) Molecules lacking acidic groups were removed.
4) Molecules with Br, I, B, P, S- and nitro groups and sulfate as only single acidic group were removed.

For the remaining 8,018 molecules initial conformations were calculated with CORINA [Sadowski *et al.*, 1994] and up to 20 low energy conformations were calculated with MOE.

376 drug-like and non-peptidic molecules satisfied the pharmacophore. This list was manually reduced to 85 compounds removing too similar molecules or molecules with unreasonably appearance. These molecules were purchased for experimental testing. Unfortunately at the time of finishing this thesis no results from the assays about the potency of the selected potential ligands were available.

4.8.1 Conclusions

A pharmacophore model derived from a receptor homology model and a binding mode hypothesis were used to virtually screen for the first inhibitors for human taspase1. Yet with the assay system still in development one can already state that our approach was successful

in retrieving 85 diverse drug-like and non-peptidic molecules, that satisfied the pharmacophore and looked reasonable to exhibit at least slight activity according to the medicinal chemistry experience of the authors. While this does not tell much about the actual activity of the compounds one can already resume that the approach seemed to produce in principle meaningful results and thus represents an alternative to entirely ligands based or structure-based approaches. The approach seems to be attractive for scaffold hopping from peptides to drug-like molecules and in combination with homology model approximations of the receptor structure.

5 Summary

5.1 Summary

The goal of this thesis was the development, evaluation and application of novel virtual screening approaches for the rational compilation of high quality pharmacological screening libraries. The criteria for a high quality were a high probability of the selected molecules to be active compared to randomly selected molecules and diversity in the retrieved chemotypes of the selected molecules to be prepared for the attrition of single lead structures. For the latter criterion the virtual screening approach had to perform “scaffold hopping”. The first molecular descriptor that was explicitly reported for that purpose was the topological pharmacophore CATS descriptor, representing a correlation vector (CV) of all pharmacophore points in a molecule. The representation is alignment-free and thus renders fast screening of large databases feasible.

In a first series of experiments the CATS descriptor was conceptually extended to the three-dimensional pharmacophore-pair CATS3D descriptor and the molecular surface based SURFCATS descriptor. The scaling of the CATS3D descriptor, the combination of CATS3D with different similarity metrics and the dependence of the CATS3D descriptor on the three-dimensional conformations of the molecules in the virtual screening database were evaluated in retrospective screening experiments. The “scaffold hopping” capabilities of CATS3D and SURFCATS were compared to CATS and the substructure fingerprint MACCS keys. Prospective virtual screening with CATS3D similarity searching was applied for the TAR RNA and the metabotropic glutamate receptor 5 (mGlur5). A combination of supervised and unsupervised neural networks trained on CATS3D descriptors was applied prospectively to compile a focused but still diverse library of mGluR5 modulators. In a second series of experiments the SQUID fuzzy pharmacophore model method was developed, that was aimed to provide a more general query for virtual screening than the CATS family descriptors. A prospective application of the fuzzy pharmacophore models was performed for TAR RNA ligands. In a last experiment a structure-/ligand-based pharmacophore model was developed for taspase1 based on a homology model of the enzyme. This model was applied prospectively for the screening for the first inhibitors of taspase1.

The effect of different similarity metrics (*Euc*: Euclidean distance, *Manh*: Manhattan distance and *Tani*: Tanimoto similarity) and different scaling methods (*unscaled*, *scaling1*: scaling by the number of atoms, and *scaling2*: scaling by the added incidences of potential pharmacophore points of atom pairs) on CATS3D similarity searching was evaluated in retrospective virtual screening experiments. 12 target classes of the COBRA database of annotated ligands from recent scientific literature were used for that purpose. *Scaling2*, a new development for the CATS3D descriptor, was shown to perform best on average in combination with all three similarity metrics (enrichment factor *ef* (1%): *Manh* = 11.8 ± 4.3, *Euc* = 11.9 ± 4.6, *Tani* = 12.8 ± 5.1). The Tanimoto coefficient was found to perform best with the new scaling method. Using the other scaling methods the Manhattan distance performed best (*ef* (1%): *unscaled*: *Manh* = 9.6 ± 4.0, *Euc* = 8.1 ± 3.5, *Tani* = 8.3 ± 3.8; *scaling1*: *Manh* = 10.3 ± 4.1, *Euc* = 8.8 ± 3.6, *Tani* = 9.1 ± 3.8).

Since CATS3D is independent of an alignment, the dependence of a "receptor relevant" conformation might also be weaker compared to other methods like docking. Using such methods might be a possibility to overcome problems like protein flexibility or the computational expensive calculation of many conformers. To test this hypothesis, co-crystal structures of 11 target classes served as queries for virtual screening of the COBRA database. Different numbers of conformations were calculated for the COBRA database. Using only a single conformation already resulted in a significant enrichment of isofunctional molecules on average (*ef* (1%) = 6.0 ± 6.5). This observation was also made for ligand classes with many rotatable bonds (e.g. HIV-protease: 19.3 ± 6.2 rotatable bonds in COBRA, *ef* (1%) = 12.2 ± 11.8). On average only an improvement from using the maximum number of conformations (on average 37 conformations / molecule) to using single conformations of 1.1 fold was found. It was found that using more conformations actives and inactives equally became more similar to the reference compounds according to the CATS3D representations. Applying the same parameters as before to calculate conformations for the crystal structure ligands resulted in an average Cartesian RMSD of the single conformations to the crystal structure conformations of 1.7 ± 0.7 Å. For the maximum number of conformations, the RMSD decreased to 1.0 ± 0.5 Å (1.8 fold improvement on average).

To assess the virtual screening performance and the scaffold hopping potential of CATS3D and SURFACATS, these descriptors were compared to CATS and the MACCS keys, a fingerprint based on exact chemical substructures. Retrospective screening of ten classes of the COBRA database was performed. According to the average enrichment factors the MACCS keys performed best (*ef* (1%): MACCS = 17.4 ± 6.4, CATS = 14.6 ± 5.4,

CATS3D = 13.9 ± 4.9, SURFCATS = 12.2 ± 5.5). The classes, where MACCS performed best, consisted of a lower average fraction of different scaffolds relative to the number of molecules (0.44 ± 0.13), than the classes, where CATS performed best (0.65 ± 0.13). CATS3D was the best performing method for only a single target class with an intermediate fraction of scaffolds (0.55). SURFCATS was not found to perform best for a single class. These results indicate that CATS and the CATS3D descriptors might be better suited to find novel scaffolds than the MACCS keys. All methods were also shown to complement each other by retrieving scaffolds that were not found by the other methods.

A prospective evaluation of CATS3D similarity searching was done for metabotropic glutamate receptor 5 (mGluR5) allosteric modulators. Seven known antagonists of mGluR5 with sub-micromolar IC_{50} were used as reference ligands for virtual screening of the 20,000 most drug-like compounds – as predicted by an artificial neural network approach – of the Asinex vendor database (194,563 compounds). Eight of 29 virtual screening hits were found with a K_i below 50 µM in a binding assay. Most of the ligands were only moderately specific for mGluR5 (maximum of > 4.2 fold selectivity) relative to mGluR1, the most similar receptor to mGluR5. One ligand exhibited even a better K_i for mGluR1 than for mGluR5 (mGluR5: K_i > 100 µM, mGluR1: K_i = 14 µM). All hits had different scaffolds than the reference molecules. It was demonstrated that the compiled library contained molecules that were different from the reference structures – as estimated by MACCS substructure fingerprints – but were still considered isofunctional by both CATS and CATS3D pharmacophore approaches.

Artificial neural networks (ANN) provide an alternative to similarity searching in virtual screening, with the advantage that they incorporate knowledge from a learning procedure. A combination of artificial neural networks for the compilation of a focused but still structurally diverse screening library was employed prospectively for mGluR5. Ensembles of neural networks were trained on CATS3D representations of the training data for the prediction of “mGluR5-likeness” and for “mGluR5/mGluR1 selectivity”, the most similar receptor to mGluR5, yielding Matthews *cc* between 0.88 and 0.92 as well as 0.88 and 0.91 respectively. The best 8,403 hits (the focused library: the intersection of the best hits from both prediction tasks) from virtually ranking the Enamine vendor database (ca. 1,000,000 molecules), were further analyzed by two self-organizing maps (SOMs), trained on CATS3D descriptors and on MACCS substructure fingerprints. A diverse and representative subset of the hits was obtained by selecting the most similar molecules to each SOM neuron. Binding studies of the selected compounds (16 molecules from each map) gave that three of

the molecules from the CATS3D SOM and two of the molecules from the MACCS SOM showed mGluR5 binding. The best hit with a K_i of 21 µM was found in the CATS3D SOM. The selectivity of the compounds for mGluR5 over mGluR1 was low. Since the binding pockets in the two receptors are similar the general CATS3D representation might not have been appropriate for the prediction of selectivity. In both SOMs new active molecules were found in neurons that did not contain molecules from the training set, i. e. the approach was able to enter new areas of chemical space with respect to mGluR5. The combination of supervised and unsupervised neural networks and CATS3D seemed to be suited for the retrieval of dissimilar molecules with the same class of biological activity, rather than for the optimization of molecules with respect to activity or selectivity.

A new virtual screening approach was developed with the SQUID (*S*ophisticated *Qu*antification of *I*nteraction *D*istributions) fuzzy pharmacophore method. In SQUID pairs of Gaussian probability densities are used for the construction of a CV descriptor. The Gaussians represent clusters of atoms comprising the same pharmacophoric feature within an alignment of several active reference molecules. The fuzzy representation of the molecules should enhance the performance in scaffold hopping. Pharmacophore models with different degrees of fuzziness (resolution) can be defined which might be an appropriate means to compensate for ligand and receptor flexibility. For virtual screening the 3D distribution of Gaussian densities is transformed into a two-point correlation vector representation which describes the probability density for the presence of atom-pairs, comprising defined pharmacophoric features. The fuzzy pharmacophore CV was used to rank CATS3D representations of molecules. The approach was validated by retrospective screening for cyclooxygenase 2 (COX-2) and thrombin ligands. A variety of models with different degrees of fuzziness were calculated and tested for both classes of molecules. Best performance was obtained with pharmacophore models reflecting an intermediate degree of fuzziness. Appropriately weighted fuzzy pharmacophore models performed better in retrospective screening than CATS3D similarity searching using single query molecules, for both COX-2 and thrombin (*ef* (1%): COX-2: SQUID = 39.2., best CATS3D result = 26.6; Thrombin: SQUID = 18.0, best CATS3D result = 16.7). The new pharmacophore method was shown to complement MOE pharmacophore models.

SQUID fuzzy pharmacophore and CATS3D virtual screening were applied prospectively to retrieve novel scaffolds of RNA binding molecules, inhibiting the Tat-TAR interaction. A pharmacophore model was built up from one ligand (acetylpromazine, IC_{50} = 500 µM) and a fragment of another known ligand (CGP40336A), which was assumed to bind

with a comparable binding mode as acetylpromazine. The fragment was flexible aligned to the TAR bound NMR conformation of acetylpromazine. Using an optimized SQUID pharmacophore model the 20,000 most druglike molecules from the SPECS database (229,658 compounds) were screened for Tat-TAR ligands. Both reference inhibitors were also applied for CATS3D similarity searching. A set of 19 molecules from the SQUID and CATS3D results was selected for experimental testing. In a fluorescence resonance energy transfer (FRET) assay the best SQUID hit showed an IC_{50} value of 46 µM, which represents an approximately tenfold improvement over the reference acetylpromazine. The best hit from CATS3D similarity searching showed an IC_{50} comparable to acetylpromazine (IC_{50} = 500 µM). Both hits contained different molecular scaffolds than the reference molecules.

Structure-based pharmacophores provide an alternative to ligand-based approaches, with the advantage that no ligands have to be known in advance and no topological bias is introduced. The latter is e.g. favorable for hopping from peptide-like substrates to drug-like molecules. A homology model of the threonine aspartase taspase1 was calculated based on the crystal structures of a homologous isoaspartyl peptidase. Docking studies of the substrate with GOLD identified a binding mode where the cleaved bond was situated directly above the reactive N-terminal threonine. The predicted enzyme-substrate complex was used to derive a pharmacophore model for virtual screening for novel taspase1 inhibitors. 85 molecules were identified from virtual screening with the pharmacophore model as potential taspase1-inhibitors, however biochemical data was not available before the end of this thesis.

In summary this thesis demonstrated the successful development, improvement and application of pharmacophore-based virtual screening methods for the compilation of molecule-libraries for early phase drug development. The highest potential of such methods seemed to be in scaffold hopping, the non-trivial task of finding different molecules with the same biological activity.

5.2 Zusammenfassung

Ziel dieser Arbeit war die Entwicklung, Untersuchung und Anwendung von neuen virtuellen Screening-Verfahren für den rationalen Entwurf hoch-qualitativer Molekül-Datenbanken für das pharmakologische Screening. Anforderung für eine hohe Qualität waren eine hohe a priori Wahrscheinlichkeit für das Vorhandensein aktiver Moleküle im Vergleich zu zufällig zusammengestellten Bibliotheken, sowie das Vorhandensein einer Vielfalt unterschiedlicher Grundstrukturen unter den selektierten Molekülen, um gegen den Ausfall einzelner Leitstrukturen in der weiteren Entwicklung abgesichert zu sein. Notwendig für die letztere Eigenschaft ist die Fähigkeit eines Verfahrens zum „Grundgerüst-Springen“. Der erste Molekül-Deskriptor, der explizit für das „Grundgerüst-Springen“ eingesetzt wurde war der CATS Deskriptor – ein topologischer Korrelations-Vektor („*correlation vector*“, CV) über alle Pharmakophor-Punkte eines Moleküls. Der Vergleich von Molekülen über den CATS Deskriptor geschieht ohne eine Überlagerung der Moleküle, was den effizienten Einsatz solcher Verfahren für sehr große Molekül-Datenbanken ermöglicht.

In einer ersten Serie von Versuchen wurde der CATS Deskriptor erweitert zu dem dreidimensionalen CATS3D Deskriptor und dem auf der Molekül-Oberfläche basierten SURFCATS Deskriptor. In retrospektiven Studien wurde für diese Deskriptoren der Einfluss verschiedener Skalierungs-Methoden, die Kombination mit unterschiedlichen Ähnlichkeits-Metriken und die Auswirkung verschiedener dreidimensionaler Konformationen untersucht. Weiter wurden das Potential der entwickelten Deskriptoren CATS3D und SURFCATS im „Grundgerüst-Springen“ mit CATS und dem Substruktur-Fingerprint MACCS keys verglichen. Prospektive Anwendungen der CATS3D Ähnlichkeitssuche wurden für die TAR-RNA und den metabotropen Glutamat Rezeptor 5 (mGluR5) durchgeführt. Eine Kombination von überwachten und unüberwachten neuronalen Netzen wurde prospektiv für die Zusammenstellung einer fokussierten aber dennoch diversen Bibliothek von mGluR5 Modulatoren eingesetzt. In einer zweiten Reihe von Versuchen wurde der SQUID Fuzzy Pharmakophor Ansatz entwickelt, mit dem Ziel zu einer noch generelleren Molekül-Beschreibung als mit den Deskriptoren aus der CATS Familie zu gelangen. Eine prospektive Anwendung der „Fuzzy Pharmakophor“ Methode wurde für die TAR-RNA durchgeführt. In einem letzten Versuch wurde für Taspase1 ein Struktur-/Liganden-basiertes Pharmakophor-Modell auf der Grundlage eines Homologie-Modells des Enzyms entwickelt. Dieses wurde für das prospektive Screening nach Taspase1-Inhibitoren eingesetzt.

Der Einfluss verschiedener Ähnlichkeits-Metriken (*Euk*: Euklidische Distanz, *Manh:* Manhattan Distanz, *Tani:* Tanimoto Ähnlichkeit) und verschiedener Skalierungs-Methoden

(*Ohne-Skalierung*, *Skalierung1*: Skalierung aller Werte nach der Anzahl Atome, *Skalierung2*: Skalierung der Werte eines Paares von Pharmakophor-Punkten entsprechend der Summe aller Pharmakophor-Punkte mit denselben Pharmakophor-Typen) auf die Ähnlichkeits-Suche mit CATS3D wurde in retrospektiven virtuellen Screening Experimenten untersucht. Für diesen Zweck wurden 12 verschiedene Klassen von Rezeptoren und Enzymen aus der COBRA Datenbank von annotierten Liganden aus der jüngeren wissenschaftlichen Literatur eingesetzt. *Skalierung2*, eine neue Entwicklung für CATS3D, zeigte im Durchschnitt die beste Performanz in Kombination mit allen drei Ähnlichkeits-Metriken (Anreicherungs-Faktor *ef* (1%): *Manh* = 11,8 ± 4,3; *Euk* = 11,9 ± 4,6; *Tani* = 12,8 ± 5,1). Die Kombination von *Skalierung2* mit dem Tanimoto Ähnlichkeits-Koeffizienten lieferte die besten Ergebnisse. In Kombination mit den anderen Skalierungen brachte die Manhattan Distanz die besten Ergebnisse (*ef* (1%): *Ohne-Skalierung*: *Manh* = 9,6 ± 4,0; *Euk* = 8,1 ± 3,5; *Tani* = 8,3 ± 3,8; *Skalierung1*: *Manh* = 10,3 ± 4,1; *Euk* = 8,8 ± 3,6; *Tani* = 9,1 ± 3,8).

Da die CATS3D Ähnlichkeits-Suche unabhängig von der Überlagerung einzelner Moleküle ist, könnte ebenfalls eine gewisse Unabhängigkeit von der vorhandenen 3D Konformation bestehen. Eine solche Unabhängigkeit wäre interessant um die zeitaufwendige Berechnung multipler Konformationen zu umgehen. Um diese Hypothese zu untersuchen wurden Co-Kristalle von Liganden aus 11 Klassen von Rezeptoren und Enzymen ausgewählt, um als Anfrage-Strukturen im virtuellen Screening in der COBRA Datenbank zu dienen. Verschiedene Versionen der COBRA Datenbank mit unterschiedlicher Anzahl Konformationen wurden berechnet. Bereits mit einer einzigen Konformation pro Molekül konnte im Mittel eine deutliche Anreicherung an aktiven Molekülen beobachte werden (*ef* (1%) = 6,0 ± 6,5). Diese Beobachtung beinhaltete auch Klassen von Molekülen mit vielen rotierbaren Bindungen. (z.B. HIV-Protease: 19,3 ± 6,2 rotierbare Bindungen in COBRA, *ef* (1%) = 12,2 ± 11,8). Im Mittel konnten dazu bei Verwendung der maximalen Anzahl Konformationen (durchschnittlich 37 Konformationen / Molekül) nur eine Verbesserung von 1.1 festgestellt werden. Nach der CATS3D Ähnlichkeit wurden die inaktiven Moleküle im gleichen Maß ähnlicher zu den Referenzen als die aktiven Moleküle. Zum Vergleich konnte durch Verwendung multipler statt einzelner Konformationen eine 1,8-fache Verbesserung des RMSD zu den Konformationen aus den Kristall-Struktur Konformationen erreicht werden (einzelne Konformationen: 1,7 ± 0,7 Å; max. Konformationen: 1,0 ± 0,5 Å).

Um die Leistungsfähigkeit von CATS3D und SURFCATS im virtuellen Screening und im Grundgerüst-Springen zu beurteilen, wurden diese Deskriptoren mit CATS und den MACCS keys, einem Fingerprint basierend auf exakten chemischen Substrukturen,

verglichen. Für die retrospektive Analyse wurden 10 Klassen von Rezeptoren und Enzymen aus der COBRA Datenbank ausgewählt. Nach den mittleren Anreicherungs-Faktoren ergaben sich für MACCS die besten Resultate (*ef* (1%): MACCS = 17,4 ± 6,4; CATS = 14,6 ± 5,4; CATS3D = 13,9 ± 4,9; SURFCATS = 12,2 ± 5,5). Es zeigte sich, dass die Klassen, in denen MACCS die besten Ergebnisse erzielen konnte, einen geringen gemittelten Anteil von verschiedenen Grundgerüsten aufwiesen im Verhältnis zu der Anzahl an Molekülen (0,44 ± 0,13) als die Klassen, in denen CATS am besten war (0,65 ± 0,13). CATS3D war nur in einer Klasse mit einem mittleren Anteil von Grundgerüsten (0,55) die beste Methode. SURFCATS war für keine Klasse besser als alle anderen Methoden. Diese Ergebnisse deuten darauf hin, dass Methoden wie CATS und CATS3D besser geeignet sind, um neue Grundgerüste zu finden. Es konnte weiter gezeigt werden, dass sich die Methoden einander ergänzen, dass also mit jeder Methode Grundgerüste gefunden werden konnten, die mit keiner der anderen Methoden gefunden werden konnten.

Eine prospektive Anwendung wurde für CATS3D in der Suche nach neuen allosterischen Modulatoren des metabotropen Glutamat Rezeptors 5 (mGluR5) durchgeführt. Sieben bekannte allosterische mGluR5 Antagonisten mit sub-mikromolaren IC_{50} Werten wurde als Referenzen eingesetzt. Das virtuelle Screening wurde auf den 20.000 von einem künstlichen neuronalen Netz als am wirkstoff-artigsten vorhergesagten Molekülen der Asinex Datenbank (194.563 Moleküle) durchgeführt. Acht der 29 gefundenen Hits aus dem virtuellen Screening zeigten K_i Werte unter 50 µM in einem Bindungs-Assay. Die Mehrheit der Liganden zeigte nur eine geringe Selektivität (Maximum > 4,2-fach) gegenüber mGluR1, dem ähnlichsten Rezeptor zu mGluR5. Einer der Liganden zeigte einen besseren K_i für mGluR1 als für mGluR5 (mGluR5: K_i > 100 µM, mGluR1: K_i = 14 µM). Alle gefundenen Moleküle zeigten verschiedene Grundgerüste als die Referenz Moleküle. Es konnte gezeigt werden, dass die zusammengestellte Bibliothek von den MACCS keys als unterschiedlich zu den Referenz Strukturen betrachtet wurden, von CATS und CATS3D aber noch als isofunktional betracht wurden.

Künstliche neuronal Netze („*artificial neural net*“, ANN) bieten eine Alternative zur Ähnlichkeits-Suche im virtuellen Screening mit dem Vorteil, dass in einer Serie von Liganden enthaltenes implizites Wissen über eine Lernprozedur in ein Modell integrierte werden kann. Eine Kombination von ANNs für die Zusammenstellung einer fokussierten aber dennoch diversen Molekül-Bibliothek wurde prospektiv für die Suche nach mGluR5 Antagonisten eingesetzt. Gruppen von ANNs wurden auf den Basis von CATS3D Repräsentationen für die Vorhersage von „mGluR5-artigkeit“ und „mGluR5/mGluR1 Selektivität“ trainiert. Dabei

ergaben sich Matthews *cc* zwischen 0,88 und 0,92 sowie zwischen 0,88 und 0,91. Die besten 8.403 Hits (die Schnittmenge der besten Hits aus beiden Vorhersagen) aus einem virtuellen Screening der Enamine Datenbank (ca. 1.000.000 Moleküle) ergab die fokussierte Bibliothek. Diese wurde weiter mit Selbstorganisierten Karten („*self organizing maps*“, SOM) analysiert, die auf CATS3D und MACCS key Repräsentationen trainiert wurden. Eine diverse und repräsentative Untermenge der Moleküle wurde gewonnen, indem die jeweils nächsten Moleküle zu jedem der Neuronen der Karten ausgewählt wurden. Bindungsstudien der selektierten Moleküle (16 von jeder der Karten) ergaben, dass drei Moleküle aus der CATS3D SOM und zwei der Moleküle aus der MACCS SOM mGluR5 Bindung zeigten. Der beste Hit mit einem K_i von 21 µM wurde über die CATS3D SOM gefunden. Die Selektivität der gefundenen Moleküle gegenüber mGluR1 war wiederum gering. Da sich die Bindungstaschen der beiden Rezeptoren sehr ähnlich sind, könnte die verallgemeinernde Beschreibung der Moleküle mit CATS3D nicht geeignet für eine solche Vorhersage gewesen sein. In beiden SOMs wurden neue aktive Moleküle in Neuronen gefunden, in denen sich keine der bekannten Inhibitoren befanden, d.h. es wurden mit diesem Ansatz neue chemische Bereiche auf der SOM für mGluR5 beschrieben. Die Verbindung von überwachten und unüberwachten neuronalen Netzen mit CATS3D scheint am besten geeignet zu sein, um Moleküle mit unterschiedlicher Struktur, aber gleicher Aktivitätsklasse aufzufinden. Die Optimierung auf höhere Aktivität oder Selektivität schien weniger geeignet zu sein.

Mit dem SQUID (*S*ophisticated *Qu*antification of *I*nteraction *D*istributions) Fuzzy Pharmakophor Modell wurde ein neuer Ansatz für das virtuelle Screening entwickelt. In SQUID werden Paare von Gauß-Wahrscheinlichkeits-Dichten für die Konstruktion eines Korrelations-Vektors eingesetzt. Die Gauß-Dichten repräsentieren Gruppen von Atomen desselben Pharmakophor-Typs in einer Überlagerung mehrerer aktiver Referenz-Moleküle. Die unscharfe Repräsentation der Moleküle sollte das Springen zwischen Grundgerüsten erleichtern. Der Ansatz ermöglicht die Definition von Pharmakophor-Modellen verschiedener Unschärfe oder Auflösung, was eventuell eine Möglichkeit darstellt, die Flexibilität von Ligand und Rezeptor zu berücksichtigen. Für das virtuelle Screening wird die dreidimensionale Verteilung der Gauß-Dichten in einen 2-Punkt CV transformiert, der die Wahrscheinlichkeit für die Anwesendheit von Paaren von Pharmakophor-Punkten beschreibt. Der Fuzzy Pharmakophor CV wurde eingesetzt um CATS3D Repräsentationen zu bewerten. Evaluiert wurde die Methode durch retrospektives Screening nach COX-2 und Thrombin Inhibitoren. Eine Serie von Modellen mit verschiedener Auflösung wurde für beide Molekülklassen getestet. Die besten Ergebnisse wurden in beiden Fallen mit Modellen mit

mittlerer Auflösung erzielt. Geeignet gewichtete Pharmakophor-Modelle erzielten bessere Resultate als CATS3D Ähnlichkeits-Suche mit den einzelnen Molekülen aus den Pharmakophor-Modellen (*ef* (1%): COX-2: SQUID = 39,2; bestes CATS3D Resultat = 26,6; Thrombin: SQUID = 18,0; bestes CATS3D Resultat = 16,7). Es konnte weiter gezeigt werden, dass die neue Methode MOE Pharmakophor Modelle in den gefundenen Molekülen ergänzt.

Der SQUID Fuzzy Pharmakophor Ansatz sowie CATS3D wurden prospektiv eingesetzt für die Suche nach neuen Grundgerüsten für RNA bindendende Inhibitoren der Tat-TAR Interaktion. Ein SQUID Modell wurde auf der Grundlage von einem Liganden (Acetylpromazin, IC_{50} = 500 µM) und dem Fragment eines weiteren bekannten Liganden (CGP40336A) berechnet, von dem ein zum Acetylpromazin vergleichbarer Bindungsmodus angenommen wurde. Das Fragment wurde flexibel an die TAR-RNA gebundene NMR Konformation des Acetylpromazins aligned. Mit einem optimierten SQUID Modell wurden die 20.000 wirkstoffartigsten Moleküle der SPECS Datenbank (229.658 Moleküle) virtuell nach TAR-RNA Liganden durchsucht. Mit beiden Referenz Inhibitoren wurden zum Vergleich auch CATS3D Suchen durchgeführt. 19 Moleküle aus den Hits von SQUID und CATS3D wurden für einen FRET Assay ausgewählt. Der beste Hit von SQUID zeigte einen IC_{50} Wert von 46 µM, was eine ca. 10-fache Verbesserung im Verhältnis zu Acetylpromazin darstellt. Der beste CATS3D Hit war vergleichbar mit Acetylpromazin (IC_{50} = 500 µM). Beide gefundenen Moleküle zeigten unterschiedliche Grundgerüste im Verhältnis zu den Referenz Molekülen.

Struktur-basierte Pharmakophor Modelle stellen eine Alternative zu Liganden-basierten Ansätzen dar, mit dem Vorteil, dass keine bekannten Liganden benötigt werden und somit keine Beeinflussung hin zu bekannten Strukturen in das Modell hineingebracht wird. Die letztere Eigenschaft sollte günstig für das Grundgerüst Springen sein. Ein Homologie Modell der Threonin Aspartase Taspase1, eine Protease, wurde auf der Grundlage der Kristallstruktur einer homologen Isoaspartyl-Peptidase berechnet. Über Docking Studien mit dem Programm GOLD wurde ein Bindungsmodus des natürlichen Substrats identifiziert, in dem die gespaltene Peptidbindung direkt über dem reaktiven N-terminalen Threonin angeordnet lag. Der vorhergesagte Enzym-Substrat-Komplex wurde herangezogen, um ein Pharmakophor-Modell zu für das virtuelle Screening nach neuen Taspase1 Inhibitoren zu entwickeln. 85 Moleküle wurden aus der SPECS Datenbank als potentielle Inhibitoren identifiziert, jedoch fehlten bei Fertigstellung dieser Arbeit noch gesicherte experimentelle Daten über die pharmakologischen Eigenschaften der gefundenen Moleküle.

Zusammenfassend wurde in dieser Arbeit die erfolgreiche Entwicklung, Verbesserung und Anwendung von Pharmakophor-basierten virtuellen Screening Methoden für den Entwurf von Molekül-Bibliotheken für die frühe Wirkstoff-Entwicklung gezeigt. Das Potential dieser Methoden schien besonders im Grundgerüst-Springen zu liegen, also in der nicht-trivialen Identifikation von unterschiedlichen Molekülen mit gleicher biologischer Aktivität.

6 Appendix

6.1 Enrichment factors of activity classes from Section 4.3

Enrichment factors of different molecular representations ("Molecules", "Scaffolds", "Reduced Scaffolds") over the activity classes. *ef* values are given for the first 1% and 5% of the hit-lists. The Manhattan distance was applied as similarity metric.

	Molecules			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE				
1	25 (12)	27 (17)	16 (11)	18 (13)
5	10 (4)	10 (5)	6 (3)	6 (3)
COX2				
1	30 (16)	16 (10)	21 (12)	19 (11)
5	13 (5)	6 (3)	8 (4)	8 (4)
CRF				
1	25 (14)	14 (9)	22 (10)	16 (9)
5	11 (4)	7 (3)	10 (3)	8 (3)
DPP				
1	21 (14)	14 (11)	16 (11)	13 (10)
5	7 (3)	5 (3)	5 (3)	4 (2)
HIVP				
1	13 (7)	21 (11)	12 (8)	15 (10)
5	5 (2)	9 (4)	5 (3)	6 (4)
MMP				
1	13 (8)	12 (7)	10 (7)	12 (9)
5	5 (3)	5 (2)	4 (2)	5 (3)
NK				
1	9 (6)	9 (4)	11 (7)	7 (5)
5	5 (2)	5 (2)	5 (3)	4 (2)
PPAR				
1	18 (15)	19 (12)	8 (7)	9 (8)
5	6 (4)	7 (3)	3 (2)	3 (2)
BACE				
1	14 (11)	12 (10)	12 (10)	7 (5)
5	6 (4)	4 (3)	3 (2)	2 (2)
THR				
1	12 (6)	14 (7)	7 (5)	7 (5)
5	6 (2)	8 (4)	3 (2)	4 (3)

((continued below))

	Scaffolds				Reduced Scaffolds			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE								
1	25 (12)	30 (16)	15 (10)	16 (9)	29 (13)	33 (14)	18 (10)	19 (9)
5	11 (4)	11 (4)	6 (2)	6 (2)	11 (3)	11 (4)	6 (2)	6 (2)
COX2								
1	28 (12)	17 (8)	21 (8)	19 (9)	33 (12)	22 (10)	26 (10)	24 (11)
5	11 (4)	6 (2)	8 (2)	8 (3)	12 (3)	7 (2)	9 (2)	9 (3)
CRF								
1	24 (12)	18 (11)	23 (10)	17 (10)	28 (13)	20 (11)	25 (10)	18 (11)
5	10 (4)	8 (3)	10 (3)	8 (3)	11 (4)	8 (3)	10 (3)	8 (3)
DPP-IV								
1	22 (13)	15 (10)	17 (10)	15 (11)	25 (12)	22 (14)	23 (14)	21 (16)
5	8 (3)	6 (3)	5 (2)	4 (2)	9 (5)	7 (3)	6 (3)	6 (3)
HIVP								
1	14 (8)	23 (12)	14 (9)	17 (11)	19 (12)	31 (15)	18 (12)	23 (15)
5	6 (3)	10 (4)	5 (3)	7 (4)	7 (3)	11 (4)	7 (4)	8 (4)
MMP								
1	16 (11)	15 (9)	13 (10)	16 (12)	23 (12)	23 (12)	21 (13)	22 (14)
5	6 (3)	6 (3)	4 (3)	6 (4)	8 (3)	8 (3)	6 (4)	7 (4)
NK								
1	10 (6)	10 (4)	12 (7)	8 (5)	11 (6)	11 (5)	13 (7)	9 (6)
5	5 (2)	5 (2)	5 (3)	4 (2)	5 (2)	6 (2)	5 (3)	5 (2)
PPAR								
1	16 (13)	19 (11)	9 (7)	10 (9)	20 (15)	24 (14)	11 (8)	13 (11)
5	5 (3)	7 (3)	3 (2)	3 (2)	7 (4)	8 (4)	4 (2)	4 (3)
BACE								
1	14 (10)	12 (8)	11 (6)	8 (5)	15 (11)	14 (9)	12 (7)	9 (5)
5	4 (2)	4 (2)	3 (2)	3 (2)	5 (3)	4 (2)	3 (2)	3 (2)
THR								
1	14 (6)	18 (10)	9 (6)	10 (7)	18 (8)	27 (13)	14 (8)	17 (9)
5	6 (2)	8 (4)	4 (3)	5 (3)	8 (3)	10 (4)	5 (3)	6 (3)

Enrichment factors of different molecular representations (“Molecules”, “Scaffolds”, “Reduced Scaffolds”) over the activity classes. *ef* values are given for the first 1% and 5% of the hit-lists. The Euclidean distance was applied as similarity metric.

	Molecules			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE				
1	25 (12)	22 (12)	17 (11)	20 (14)
5	10 (4)	10 (5)	6 (4)	7 (4)
COX2				
1	30 (16)	14 (9)	20 (13)	18 (11)
5	13 (5)	5 (3)	8 (4)	8 (4)
CRF				
1	25 (14)	12 (8)	20 (10)	16 (10)
5	11 (4)	7 (3)	10 (3)	8 (3)
DPP				
1	21 (14)	13 (10)	16 (12)	13 (11)
5	7 (3)	4 (3)	5 (3)	4 (2)
HIVP				
1	13 (7)	22 (11)	15 (9)	17 (11)
5	5 (2)	10 (3)	7 (4)	7 (4)
MMP				
1	13 (8)	11 (6)	11 (7)	11 (9)
5	5 (3)	5 (2)	4 (2)	5 (3)
NK				
1	9 (6)	8 (4)	12 (8)	8 (5)
5	5 (2)	5 (2)	6 (3)	5 (3)
PPAR				
1	18 (15)	18 (12)	9 (8)	10 (7)
5	6 (4)	7 (3)	3 (2)	3 (2)
BACE				
1	14 (11)	12 (10)	12 (10)	6 (5)
5	6 (4)	4 (3)	3 (3)	3 (2)
THR				
1	12 (6)	14 (7)	8 (5)	7 (5)
5	6 (2)	8 (4)	4 (2)	4 (3)

((continued below))

	Scaffolds				Reduced Scaffolds			
% DB	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**	**MACCS**	**CATS**	**CATS3D**	**SURFCATS**
ACE								
1	25 (12)	24 (13)	17 (10)	19 (10)	29 (13)	27 (13)	20 (10)	21 (8)
5	11 (4)	11 (4)	6 (3)	7 (3)	11 (3)	11 (3)	7 (3)	8 (2)
COX2								
1	28 (12)	16 (9)	21 (9)	20 (9)	33 (12)	20 (10)	25 (10)	26 (11)
5	11 (4)	5 (2)	8 (3)	8 (3)	12 (3)	6 (2)	9 (2)	9 (3)
CRF								
1	24 (12)	15 (10)	21 (10)	17 (10)	28 (13)	18 (10)	22 (9)	18 (11)
5	10 (4)	7 (4)	10 (3)	8 (3)	11 (4)	7 (3)	9 (3)	8 (3)
DPP-IV								
1	22 (13)	13 (10)	18 (11)	14 (12)	25 (12)	18 (12)	24 (14)	20 (17)
5	8 (3)	5 (4)	5 (3)	4 (2)	9 (5)	6 (4)	6 (3)	5 (3)
HIVP								
1	14 (8)	24 (12)	16 (10)	20 (13)	19 (12)	31 (15)	22 (13)	27 (17)
5	6 (3)	10 (4)	7 (4)	8 (4)	7 (3)	12 (4)	8 (4)	10 (4)
MMP								
1	16 (11)	15 (8)	14 (10)	15 (12)	23 (12)	23 (11)	21 (13)	22 (13)
5	6 (3)	6 (2)	5 (3)	6 (4)	8 (3)	8 (3)	6 (3)	8 (3)
NK								
1	10 (6)	9 (4)	13 (8)	9 (6)	11 (6)	11 (5)	14 (8)	10 (6)
5	5 (2)	5 (2)	6 (3)	5 (3)	5 (2)	6 (2)	6 (3)	5 (3)
PPAR								
1	16 (13)	18 (11)	9 (8)	10 (8)	20 (15)	24 (14)	11 (9)	13 (10)
5	5 (3)	6 (3)	3 (2)	3 (2)	7 (4)	8 (4)	4 (2)	4 (2)
BACE								
1	14 (10)	14 (9)	11 (6)	8 (5)	15 (11)	16 (10)	11 (7)	8 (5)
5	4 (2)	4 (2)	3 (2)	3 (1)	5 (3)	4 (2)	3 (2)	3 (2)
THR								
1	14 (6)	18 (9)	9 (6)	10 (6)	18 (8)	27 (12)	15 (8)	16 (8)
5	6 (2)	9 (4)	4 (3)	5 (3)	8 (3)	10 (4)	6 (3)	6 (3)

6.2 Protein report from MOE for the taspase1 homology model from Section 4.8

```
Protein Report
Wed Jan 05 14:22:19 2005 (MOE 2004.03, MOE-ProEval 2002.01)

Options:
--------

Z-Score Threshold               : 5
VDW Contact Threshold           : 70
Write Outliers Only             : TRUE
Contacts Within Chains Only     : TRUE

Topics:
--------

Dihedrals                       : TRUE
Bond Angles                     : TRUE
Bond Lengths                    : TRUE
Contacts                        : TRUE

Protein Report: Dihedrals
Chain/Residue       phi     psi    omega   chi1    chi2     zeta
-------------------------------------------------------------------------------
-----
>1  TaspaseASN49     106.7<  -48.3< -176.1  -158.5    12.3    26.5
>1  TaspaseGLU103    106.8<  -53.7<  114.2< -179.9   179.0    31.2
>1  TaspaseASP104    -77.4   132.9  -115.2< -138.5   -30.0    29.8
>1  TaspaseLEU107   -158.8   146.7  -145.3<   10.3  -172.3    39.4
>1  TaspaseGLY108   -120.7   171.2   137.3<   -       -       -
>1  TaspaseARG117     99.3< -126.4<  175.4   -65.8   -90.0    24.5
>1  TaspaseLYS128   -107.2    99.7  -143.0<   53.1  -109.5    27.9
>2  TaspaseTYR52     143.7   167.6   150.1<  -86.4    19.8    21.1
>2  TaspaseHIS53    -123.3  -101.4  -146.9< -170.8    80.5    32.5
>2  TaspaseSER54      42.5   -24.2  -151.0< -154.0    -       27.7
>2  TaspaseGLU55     176.3    31.2  -121.4<  159.8  -178.7    34.7
>2  TaspaseILE73  T -139.0    62.0   -23.0<  -94.9   -71.2    29.7    cis
>2  TaspaseGLY150   -123.9   138.5   149.5<   -       -       -
>2  TaspaseGLN151    -82.2   116.6   104.5<  -89.6    50.8    31.8
>2  TaspaseLYS152    -78.7   120.4    63.2<   43.9   163.2    32.5    cis
>2  TaspaseLYS154   -164.1< -106.8< -157.2   -24.4    75.2    33.9
>2  TaspaseLEU155     18.2   -75.0   143.7< -143.7    76.7    29.0
>2  TaspaseALA157    -89.5    61.3   149.7<   -       -       32.9
>2  TaspaseARG159    146.3<  -37.7<   59.3< -167.0    47.2    60.2<   cis
>2  TaspasePRO162    -66.2   113.3   135.5<   22.6   -33.8    38.5
>2  TaspaseCYS163     72.0<  -51.4<   15.3<  -83.3    -       27.4    cis
>2  TaspasePHE164    175.7   107.7  -148.2<  177.0   -67.5    34.8
```

Protein Report: Bond Angles

Chain/Residue	C-N-CA	N-CA-C	N-CA-CB	CB-CA-C	CA-C-N	CA-C-O	O-C-N
>1 TaspaseGLN126	120.3	127.6<	118.6	93.0	112.3	122.8	124.9
>1 TaspaseASN127	108.0	142.0<	105.8	103.4	115.5	121.0	123.4
>2 TaspaseILE73	172.1<	109.2	106.2	121.4	110.8	124.3	124.5
>2 TaspaseGLN151	119.3	92.2<	116.0	113.8	117.7	120.5	120.5
>2 TaspaseARG159	117.7	137.5<	99.9	89.4<	117.7	118.1	122.3
>2 TaspaseILE160	110.9	94.8	129.8<	118.4	120.9	117.8	121.2
>2 TaspaseCYS163	137.1<	111.1	120.8	108.7	113.0	121.1	124.7

Protein Report: Bond Lengths

Chain/Residue	N-CA	CA-CB	CA-C	C-O	C-N
>1 TaspaseTYR36	1.460	1.236<	1.528	1.227	1.387
>1 TaspaseALA120	1.452	1.385<	1.522	1.229	1.386
>1 TaspasePHE135	1.462	1.662<	1.528	1.223	1.400
>1 TaspaseLEU136	1.463	1.689<	1.545	1.226	1.398
>2 TaspaseALA50	1.456	1.288<	1.524	1.226	1.389
>2 TaspaseHIS63	1.457	1.706<	1.534	1.226	1.388
>2 TaspaseGLN70	1.467	1.356<	1.529	1.224	1.391
>2 TaspaseGLY153	1.421	-	1.531	1.229	1.424<
>2 TaspaseLYS154	1.479	1.507	1.551	1.228	1.423<
>2 TaspaseILE160	1.438	1.582	1.494	1.222	1.430<

Protein Report: Contacts

Chain/Residue	Atom	Distance	Chain/Residue	Atom

No Items to Report

Protein Report: Summary

% residues in CORE : 78.47

Parameter	Observed		Expected	
	mean	s.d.	mean	s.d.
trans omega :	170.7	10.6	180.0	5.8
C-alpha chirality :	33.4	3.2	33.8	4.2
chi1 - gauche minus :	-66.0	25.1	-66.7	15.0
chi1 - gauche plus :	56.4	19.2	64.1	15.7
chi1 - trans :	195.3	18.1	183.6	16.8
helix phi :	-78.3	21.0	-65.3	11.9
helix psi :	-27.1	15.6	-39.4	11.3
chi1 - pooled s.d. :	-	19.2	-	15.7
proline phi :	-69.3	7.7	-65.4	11.2

7 References

D. Agrafiotis, V. S. Lobanov, F. R. Salemme. Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug. Discov.* **2002**, *1*, 337-346.

A. Ajay, W. P. Walters, M. A. Murcko, Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314-3324.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Ac. Res.* **1997**, *25*, 3389-3402.

J. J. Anderson, M. J. Bradbury, D. R. Giracello, D. F. Chapman, G. Holtz, J. Roppe, C. King, N. D. Cosford, M. A. Varney, In vivo receptor occupancy of mGlu5 receptor antagonists using the novel radioligand [3H]3-methoxy-5-(pyridin-2-ylethynyl)pyridine). *Eur. J. Pharmacol.* **2003**, *473*, 1, 35-40.

A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Cothia, A. G. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data, *Nuc. Ac. Res.* **2004**, *32*, D226-D229.

S. Anzali. J. Gasteiger, U. Holzgrabe, J. Polanski, A. Teckentrup, M. Wagener, The use of self-organizing neural networks in drug design. *Persp. Drug Disc. Design* **1998**, *9-11*, 273-299.

ASINEX Ltd., Moscow, Russia. (http://www.asinex.com)

J. Bajorath, Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882-894.

J. M. Barnard, Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532-538.

G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.

A. Bender, R. C. Glen, Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.

H. .M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nuc. Ac. Res.* **2000**, *28*, 235-242.

C. Bissantz, P. Bernard, M. Hilbert, D. Rognan, Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets?. *PROTEINS: Struct. Funct. Genet.* **2003**, *50*, 5-25.

K. H. Bleicher, H.-J. Böhm, K. Müller, A. I. Alanine, Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2003**, *2*, 369-378.

J. Bockaert, J. P. Pin, Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J*. **1999**, *18*, 1723-1729.

H.-J. Böhm, G. Schneider (Eds.), *Virtual screening for bioactive molecules*, Wiley-VCH: Weinheim, 2000

H.-J. Böhm, A. Flohr, M. Stahl. Scaffold hopping. *Drug Discov. Today: Technologies* **2004**, *1*, 217-224.

H.-J. Böhm, D. Banner, S. Bendels, M. Kansey, B. Kuhn, K. Müller, U. Obst-Sander, M. Stahl, Fluorine in medicinal chemistry. *ChemBioChem* **2004**, *5*, 637-543.

D. Borek, K. Michalska, K. Brzezinski, A. Kisiel, J. Podkowinski, D. T. Bonthron, D. Krowarsch, J. Otlewski, M. Jaskolski. Expression, purification and catalytic activity of Lupinus luteus asparagine beta-amidohydrolase and its Escherichia coli homolog. *Eur J Biochem.* **2004**, *271*, 3215-26.

J. Boström, P.-O. Norrby, T. Liljefors, Conformational energy penalties of protein-bound ligands. *J. Comput. Aided Mol. Des.* **1998**, *12*, 383-396.

J. Boström, Reproducing the Conformation of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2002**, *15*, 1137-1152.

R. D. Brown, Y. C. Martin, Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572-584.

F. K. Brown, Chemoinformatics: what is it and how does it impact drug discovery. *Ann. Rep. Med. Chem.* **1998**, *33*, 375.

B. L. Bush, R. P. Sheridan, PATTY: A programmable atom-typer and language for automatic classification of atoms in molecular databases, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756-762.

E. Byvatov, U. Fechner, J, Sadowski, G. Schneider, Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882-1889.

R. E. Carhart, D. H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: Definitions and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.

H. A. Carlson, K. M. Masukawa, K. Rubins, F. D. Bushman, W. L. Jorgensen, R. D. Lins, J. M. Briggs, J. A. McCammon, Developing a dynamic pharmacophore model for HIV-1 integrase. *J. Med. Chem.* **2000**, *43*, 2100-2114.

Chemical Computing Group, Montreal, Canada (http://www.chemcomp.com)

C. Cheng, W. H. Prussoff, Relationship between the inhibition constant (Ki) and the concentration of inhibitor which causes 50 per cent inhibition (IC50) of an enzymatic reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

C. Chiamulera, M. P. Epping-Jordan, A. Zocchi, C. Marcon, C. Cottiny, S. Tacconi, M. Corsi, F. Orzi, F. Conquet, Reinforcing and locomotor stimulant effects of cocaine are absent in mGluR5 null mutant mice. *Nat. Neurosci.* **2001**, *4*, 873-874

C. Chothia, A. M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.* **1986**, *5*, 823-826.

T. Clark, Does quantum chemistry have a place in Cheminformatics? In *Molecular informatics confronting complexity*. M. G. Hicks, C. Kettner, Eds., Logos Verlag: Berlin, 2003, pp.193-207.

T. Clark, QSAR and QSPR based solely on surface properties?, *J. Mol. Graph. Model.* **2004**, *22*, 519-525.

P. J. Conn, J.-P. Pin, Pharmacology and Functions of Metabotropic Glutamate Receptors *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 205-237.

N. D. Cosford, L. Tehrani, J. Roppe, E. Schweiger, N. D. Smith, J. Anderson, L. Bristow, J. Brodkin, X. Jiang, I. McDonald, S. Rao, M. Washburn, M. A. Varney, 3-[(2-Methyl-1,3-thiazol-4-yl)ethynyl]-pyridine: A Potent and Highly Selective Metabotropic Glutamate Subtype 5 Receptor Antagonist with Anxiolytic Activity. *J. Med. Chem.* **2003**, *46*, 204-206.

R. D. Cramer, III, D. E. Patterson, J. D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.

A. M. Davis, S. J. Teague, G. J. Kleywegt, Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 2718-2736.

Daylight Chemical Information Systems, Inc.: Los Altos, CA, USA. (http://www.daylight.com)

Z. Du, K.E. Lind, T.L. James, Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.* **2002**, *9*, 707-712.

R. Durbin, S. Eddy, A. Krogh, G. Mitchinson, *Biological sequence analysis.* Cambridge University Press: Cambridge, 1998.

I. J. DeEsch, J. E. Mills, T. D. Perkins, G. Romeo, M. Hoffmann, K. Wieland, R. Leurs, W. M. Menge, P. H. Nederkorn, P. M. Dean, H. Timmerman. Development of a pharmacophore model for histamine H3 receptor antagonists, using the newly developed molecular modeling program SLATE. *J. Med. Chem.* **2001**, *44*, 1666-1647.

C. Detering, G. Varani, Validation of Automated Docking Programs for Docking and Database Screening against RNA Drug Targets. *J. Med. Chem.* **2004**, *47*, 4188-4201.

M. .J. Drysdale, G. Lentzen, N. Matassova, A. I. H. Murchie, F. Aboul-ela, M. Afshar, RNA as a drug target. *Prog. Med. Chem.* **2002**, *39*, 73-119.

D. M. Dykxhoorn, C. D. Novina, P. A. Sharp, Killing the messenger: short RNAs that silence gene expression. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 457-467.

M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini R. P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes., *J Comput. Aided Mol. Des*. **1997**, *11*, 425-445.

Enamine Ltd., Kiev, Ukraine. (http://www.enamine.net)

O. Engkvist, P. Wrede, High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247-1249.

L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, *Multi- and Megavariate Data Analysis: Principles and Applications*, Umetrics Academy: Umeå, 2001.

A. Evers, H. Gohlke, G. Klebe, Ligand-supported homology modeling of protein binding-sites using knowledge-based potentials. *J. Mol. Biol.* **2003**, *334*, 327-345.

A. Evers, G. Klebe, Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J. Med. Chem.* **2004**, *47*, 5381-5392.

A. Evers, T. Klabunde, Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the alpha 1A adrenergic receptor. *J. Med. Chem.* **2005**, *48*, 1088-1097.

U. Fechner, L. Franke, S. Renner, P. Schneider, G. Schneider, Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des.* **2003**, *17*, 687-698.

A. V. Filikov, V. Mohan, T. A. Vickers, R. H. Griffey, P. D. Cook, R. A. Abagyan, T. L. James, Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput. Aided Mol. Des.* **2000**, *14*, 593-610.

R. Fredriksson, M. C. Lagerström, L.-G. Lundin, H. B. Schiöth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **2003**, *63*, 1256-1272.

M. Froeyen, P. Herdewijn, RNA as a target for drug design, the example of Tat-TAR interaction. *Curr. Top. Med. Chem.* **2002**, *2*, 1123-1145.

J. Gallego, G. Varani, Targeting rna with small-molecule drugs: therapeutic promise and chemical challenges. *Acc. Chem. Res.* **2001**, *34*, 836-843.

F. Gasparini, K. Lingenhöhl, N. Stoehr, P. J. Flor, M. Heinrich, I. Vranesic, M. Biollaz, H. Allgeier, R. Heckendorn, S. Urwyler, M. A. Varney, E. C. Johnson, S. D. Hess, S. P. Rao, A. I. Sacaan, E. M. Santori, G. Veliçelebi, R. Kuhn, 2-Methyl-6-(phenylethynyl)-pyridine

(MPEP), a potent, selective and systemically active mGlu5 receptor antagonist. *Neuropharmacology* **1999**, *38*, 1493-1503.

F. Gasparini, Y. Auberson, S. Ofner, WO03/047581 **2003**.

V. J. Gillet, W. Khatib, P. Willett, P. J. Fleming, D. V. S. Green, Compbinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375-385.

G. Gini, M. V. Craciun, C. König, E. Benfenati, Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897-1902.

J. Godden, J. Bajorath, An Information-Theoretic Approach to Descriptor Selection for Database Profiling and QSAR Modeling. *QSAR Comb. Sci.* **2003**, *22*, 487-479.

A. C. Good, J. S. Mason. Three-dimensional structure database searches. In *Reviews in Computational Chemistry*, Vol. 7. K. B. Lipkowitz, D. B. Boyd, Eds., VCH Publishers, 1996, pp 67-117.

A. C. Good, I. D. Kuntz, Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J Comput Aided Mol Des* **1995**, *4*, 373.

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

L. Goodstadt, C. P. Ponting, CHROMA: consensus-based colouring of multiple alignments for publications. *Bioinformatics* **2001**, *17*, 845-846.

O. Güner (Ed.), *Pharmacophore perception, development, and use in drug design*, International University Line: La Jolly CA, 2000.

C. Goudet, F. Gaven, J. Kniazeff, C. Vol, J. Liu, M. Cohen-Gonsaud, F. Acher, L. Prézeau, J. P. Pin, Heptahelical domain of metabotropic glutamate receptor 5 behaves like rhodopsin-like receptors. *Proc. Natl. Acad. Sci. U S A* **2004** , *101*, 378-383.

J. Greene, S. Kahn, H. Savoj, P. Sprague, S. Teig, Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1297-1308.

S. Grüneberg, A QSAR model derived from a homology model: a strategy to include structural information in ligand-based design. *QSAR Comb. Sci.* **2005**, *24*, 517-526.

H.-C. Guo, Q. Xu, D. Buckley, C. Guan, Crystal structures of Flavobacterium glycosylasparaginase. An N-terminal nucleophile hydrolase activated by intramolecular proteolysis. *J Biol Chem* **1998**, *273*, 20205-20212.

T. A. Halgren, Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *5-6*, 490-519.

I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of Docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409-443.

F. Hamy, V. Brondani, A. Florsheimer, W. Stark, M.J. Blommers, T. Klimkait, A New Class of HIV-1 Tat Antagonist Acting through Tat-TAR Inhibition. *Biochemistry* **1998**, *37*, 5086-5095.

C. Hansch, A. Leo, D. Hoekman (Eds.), *Exploring QSAR – Hydrophobic, electronic, and steric constants*, American Chemical Society: Washington, 1995.

M. Hejazi, K. Piotukh, J. Mattow, R. Deutzmann, R. Volkmer-Engert, W. Lockau, Isoaspartyl dipeptidase activity of plant-type asparaginases. *Biochem J.* **2002**, *364*, 129-136.

T. Hermann, Chemical function and diversity of small molecule ligands for RNA. *Biopolymers* **2003**, *70*, 4-18.

T. Hermann, Strategies for the Design of Drugs Targeting RNA and RNA-Protein Complexes. *Angew Chem Int Ed Engl.* **2000**, *39*, 1890-1904.

E. Hermans, R. A. J. Challiss, Structural, signalling and regulatory properties of the group I metabotropic glutamate receptors: prototypic family C G-protein-coupled receptors. *Biochem. J.* **2001**, *359*, 465-484.

J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.

J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256.

A. Hillisch, L. F. Pineda, R. Hilgenfeld, Utility of homology models in the drug discovery process. *Drug Discov. Today* **2004**, *9*, 659-669.

Y. T. Ho, A. Purohit, N. Vicker, S. P. Newman, J. J. Robinson, M. P. Lease, D. Ganeshapillai, L. W. Woo, B. V. Potter, M. J. Reed. Inhibition of carbonic anhydrase II by steroidal and nonsteroidal sulphamates. *Biochem. Biophys. Res. Commun*. **2003**, *305*, 909-914.

D. Horvath, B. Mao, R. Gozalbes, F. Barbosa, S. Rogalski, Strength and limitations of pharmacophore-based virtual screening. In *Chemoinformatics in drug discovery*. T. I. Oprea, Ed., Wiley-VCH: Weinheim, 2004, pp.117-140.

J. J. Hsieh, E. H. Cheng, S. L. Korsmeyer, Taspase1: a threonine aspartase required for cleavage of MLL and proper HOX gene expression. *Cell* **2003**, *115*, 293-303.

J. L. Jenkins, M. Glick, J. W. Davies. A 3D method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004**, *47*, 6144-6159.

M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, 1990.

G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748. GOLD, CCDC Software Limited, Cambridge Crystallographic Data Centre, Cambridge, UK.

J. Karn, Tackling Tat. *J. Mol. Biol.* **1999**, *293*, 135-154.

M. Karplus, J. A. McCammon, Molecular dynamics simulation of biomolecules. *Nat Struct. Biol.* **2002**, *9*, 646-652.

G. W. Kauffman, P. C. Jurs, Prediction of inhibition of the sodium ion-proton antiporter by benzoylguaninidine derivatives from molecular structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753-761.

R. W. Kennard, L. A. Stone, Computer aided design of experiments. *Technometrics* **1969**, *11*, 137-148.

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

G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130-4146.

T. Kohonen, Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59-69.

A. Krebs, V. Ludwig, O. Boden, M.-W. Göbel, Targeting the HIV trans-activation responsive region--approaches towards RNA-binding drugs. *Chembiochem* **2003**, *4*, 972-978.

K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, T. R. Cech, Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **1982**, *31*, 147-157.

H. Kubinyi (Ed.), *3D QSAR in Drug Design*, ESCOM: Leiden, 1993.

N. Kunishima, Y. Shimada, Y. Tsuji, T. Sato, M. Yamamoto, T. Kumasaka, S. Nakanishi, H. Jingami, K. Morikawa, Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* **2000**, *407*, 971-977.

P. Labute, C. Williams, M. Feher, E. Sourial, J. M. Schmidt, Flexible alignment of small molecules. J. Med. Chem. 2001, 44, 1483-1490.

A. R. Leach, V. J. Gillet, *An introduction to chemoinformatics*, Kluwer Academic Publishers: Dordrecht, 2003.

N. Leulliot, G. Varani, Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **2001**, *40*, 7947-7956.

K. E. Lind, Z. Du, K. Fujinaga, B. M. Peterlin, T. L. James, Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chem. Biol.* **2002**, *9*, 185-193.

C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Del. Rev. 1997, 23, 3-25.

Y. Liu, C. Guan, N. N. Aronson Jr., Site-directed mutagenesis of essential residues involved in the mechanism of bacterial glycosylasparaginases. *J. Biol. Chem.* **1998**, *273*, 9688-9694.

Z. Liu, B. N. Dorniny, E. I. Shakhnovitch, Structural Mining: Self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. *J. Am. Chem. Soc.* **2004**, *126*, 8515-8528.

D. J. Livingstone, D. T. Manallack, Neural networks in 3D QSAR, *QSAR Comb. Sci.* **2003**, *22*, 510-518.

D. G. Lloyd, C. L. Buenemann, N. P. Todorov, D. T. Manallack, P. M. Dean, Scaffold hopping in de novo design. Ligand generation in the absence of receptor information. *J. Med. Chem.* **2004**, *47*, 493-496.

G. M. Maggiora, V. Shanmugasundaram, M. J. Lajiness, T. N. Doman, M. W- Schultz, A practical strategy for directed compound acquisition. In *Chemoinformatics in drug discovery*. T. I. Oprea, Ed., Wiley-VCH: Weinheim, 2004, pp.315-332.

P. Malherbe, N. Kratochwil, M. T. Zenner, J. Piussi, C. Diener, C. Kratzeisen, C. Fischer, R. H. Porter, Mutational analysis and molecular modeling of the binding pocket of the metabotropic glutamate 5 receptor negative modulator 2-methyl-6-(phenylethynyl)-pyridine. *Mol. Pharmacol.* **2003**, *64*, 823-832.

M. Mandal, R. R. Breaker, Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 451-463.

Y. C. Martin, J. L. Kofron, L. M. Traphagen, Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350-4358.

J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, R. F. Labaudiniere, New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures.*J Med Chem* **1999**, *42*, 3251.

J. S. Mason, A. C. Good, E. J. Martin, 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567-597.

H. Matter, Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219-1229.

Marlowe, C. K., Sinha, U., Gunn, A. C., Scarborough, R. M. Design, synthesis and structure-activity relationship of a series of arginine aldehyde factor Xa inhibitors. Part 1: structures based on the (D)-Arg-Gly-Arg tripeptide sequence. *Bioorg. Med. Chem. Lett.* **2000**, 10, 13-16.

M. Mayer, T. L. James, NMR-based characterization of phenothiazines as a RNA binding scaffold. *J. Am. Chem. Soc.* **2004**, *126*, 4453-4460.

MDL Information Systems Inc.: San Leandro, CA, USA. (http://www.mdl.com)

L. Molnar, G. M. Keseru, A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg. Med. Chem.* **2002**, *12*, 419-421.

G. Moreau, P. Broto, The auto-correlation of a topological structure – a new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359-360.

G. Moreau, P. Broto, C. Vandycke, Molecular structures – perception, auto-correlation descriptor and SAR studies – auto-correlation descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66-70.

H. L. Morgan, The generation of a unique machine description for chemical structures – a technique developed at the chemical abstract service. *J. Chem. Doc.* **1965**, *5*, 107.

G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.

V. Mutel, J. U. Peters, J. Wichmann, WO02/46166, **2002**.

L. Nærum, L. Nørskov-Lauritsen, P. H. Olesen, Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525-1528.

M. C. Nicklaus, S. Wang, J. S. Driscoll, G. W. Milne, Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411-428.

H- F. Noller, V. Hoffarth, L. Zimniak, Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **1992**, *256*, 1416-1419.

T. Noronoaki, I. B. Stoivea, D. D. Petkov, I. Mononen. Recombinant human glycosylasparaginase catalyued hydrolysis of L-asparagine. *FEBS Lett.* **1997**, *412*, 149-152.

J. A. O'Brien, W. Lemaire, T. B. Chen, R. S. Chang, M. A. Jacobson, S. N. Ha, C. W. Lindsley, H. J. Schaffhauser, C. Sur, D. J. Pettibone, P. J. Conn, D. L. Williams Jr., A family of highly selective allosteric modulators of the metabotropic glutamate receptor subtype 5. *Mol. Pharmacol.* **2003**, *64*, 731-740.

C. Oinonen, R. Tikkanen, J. Rouvinen, L. Peltonen, Three-dimensional structure of human lysosomal aspartylglucosaminidase. *Nat. Struct. Biol.* **1995**, *2*, 1102-1108.

A. Pagano, D. Ruegg, S. Litschig, N. Stoehr, C. Stierlin, M. Heinrich, P. Floersheim, L. Prezeau, F. Carroll, J. P. Pin, A. Cambria, I. Vranesic, P. J. Flor, F. Gasparini, R. Kuhn, The Non-competitive Antagonists 2-Methyl-6- (phenylethynyl)pyridine and 7-

Hydroxyiminocyclopropan [b]chromen-1a-carboxylic Acid Ethyl Ester Interact with Overlapping Binding Pockets in the Transmembrane Region of Group I Metabotropic Glutamate Receptors. *J. Biol. Chem.* **2000**, *275*, 33750-33758.

K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, M. Miyano, Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739-745.

A. Palomer, F. Cabre, J. Pascual, J. Campos, M. Trujillo, A. Entrena, M. Gallo, L. Garcia, D. Mauleon, A. Espinosa, Identification of novel cyclooxygenase-2 selective inhibitors using pharmacophore models, *J. Med. Chem.* **2002**, *45*, 1402-1411.

Pannanugget Consulting L. L. C., Kalamazoo, MI, USA. (http://www.pannanugget.com)

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

G. A. Patani, E. J. LaVoie, Bioisosterism: A rational approach ni drug design. *Chem. Rev.* **1996**, *96*, 3147-3176.

Y. Patel, V. J. Gillet, G. Bravi, A. R. Leach, Comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput. Aided Mol. Des.* **2002**, *16*, 653-681.

D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, L. E. Weinberger, Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049-3059.

E. Perola, P. S. Charifson, Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499-2510.

S. Pickett, The biophore concept. In *Protein-ligand interactions.* H.-J. Böhm, G. Schneider, Eds., Wiley-VCH: Weinheim, 2003, pp 73-106.

J.-P. Pin, T. Galvez, L. Prézeau, Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol. Ther.* **2003**, *98*, 325-54.

J.-P. Pin, J. Kniazeff, C. Goudet, A.-S. Bessis, J. Liu, T. Galvez, F. Acher, P. Rondard, L. Prézeau, The actvation mechanism of class-C G-protein coupled receptors. *Biol. Cell* **2004**, *96*, 335-342.

B. Pirard, J. Brendel, S. Peukert, The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. *J. Chem. Inf. Model.* **2005**, *45*, 477-485.

A. Prahl, M. Pazgier, M. Hejazi, W. Lockau, J. Lubkowski, Structure of the isoaspartyl peptidase with L-asparaginase activity from Escherichia coli. *Acta Crystallogr D Biol Crystallogr.* **2004**, *60*, 1173-1176.

L. Prézeau, J. Gomeza, S. Ahern, S. Mary, T. Galvez, J. Bockaert, J.-P. Pin, Changes in the carboxyl-terminal domain of metabotropic glutamate receptor 1 by alternative splicing

generate receptors with differing agonist-independent activity. *Mol. Pharmacol.* **1996**, *49*, 422-429.

J. W. Raymond, P. Willett, Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des*. **2002**, *16*, 521-533.

F. M. Richards, Areas, volumes, packing and protein structure. *Ann. Rev. Biophys. Bioeng.* **1977**, *6*, 151-176.

O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E.-M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137-142.

O. Roche, G. Trube, J. Zuegge, P. Pflimlin, A. Alanine, G. Schneider, A Virtual Screening Method for Prediction of the hERG Potassium Channel Liability of Compound Libraries. *ChemBioChem* **2002**, *3*, 455-459.

J. Roppe, N. D. Smith, D. Huang, L. Tehrani, B. Wang, J. Anderson, J. Brodkin, J. Chung, X. Jiang, C. King, B. Munoz, M. A. Varney, P. Prasit, N. D. Cosford, Discovery of Novel Heteroarylazoles That Are Metabotropic Glutamate Subtype 5 Receptor Antagonists with Anxiolytic Activity. *J. Med. Chem.* **2004**, *47*, 4645-4648.

Rudolf, K., Eberlein, W., Engel, W., Wieland, H. A., Willim, K. D., Entzeroth, M., Wienen, W., Beck-Sickinger, A. G., Doods, H. N. The first highly potent and selective non-peptide neuropeptide Y Y1 receptor antagonist: BIBP3226. *Eur. J. Pharmacol.* **1994,** 271, R11-R13.

J. Sadowski, J. Gasteiger, G. Klebe, Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008. CORINA is available from Molecular Networks GmbH, Erlangen, Germany (http://www.mol-net.com).

J. Sadowski, H. Kubinyi, A soring scheme for discriminating between drugs and nondrugs, *J. Med. Chem.* **1998**, *41*, 3325-3329.

H. Sashida, M. Kato, T. Tsuchiya, Thermal rearrangements of cyclic amine ylides. VIII Intramolecular cyclization of 2-ethenylpyridine. N-ylides into indolizines and cycl[3.2.2]azines. *Chem. Pharm. Bull.* **1988**, *36*, 3826-3832.

M. Schmucker, A. Givehchi, G. Schneider, Impact of different software implementations on the performance of the *Maxmin* method for diverse subset selection. *Mol. Div.* **2004**, *8*, 421-425.

G. Schneider, P. Wrede, PROFI – a tool for the analysis of protein sequence features using a simple artificial neural network. *Protein Seq. Data Anal.* **1993**, *5*, 419-421.

G. Schneider, W. Neidhart, T. Giller, G. Schnmid, "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 2894-2896.

G. Schneider, O. Clément-Chomienne, L. Hilfiger, P. Schneider, S. Kirsch, H.-J. Böhm, W. Neidhard, Virtual screening for bioactive molecules by evolutionary de novo design. *Angew. Chem. Int. Ed.* **2000**, *39*, 4130-4133.

G. Schneider, Neural networks are useful tools for drug design. *Neural Netw.* **2000**, *13*, 15-16.

G. Schneider, H.-J. Böhm, Virtual screening and fast automated docking methods, *Drug Discov. Today* **2002**, *7*, 64-70.

G. Schneider, Trends in combinatorial library design, *Curr. Med. Chem.* **2002**, *9*, 2095-2101.

G. Schneider, S.-S. So, Adaptive systems in drug design. Eurekah.com / Landes Bioscience: Georgetown, USA, 2003.

G. Schneider, M. Nettekoven, Ligand-Based Combinatorial Design of Selective Purinergic Receptor (A2A) Antagonists Using Self-Organizing Maps. *J. Comb. Chem.* **2003**, *5*, 233-237.

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

G. Schneider, P. Schneider, Navigation in Chemical Space: Ligand-based Design of Focused Compound Libraries. In *Chemogenomics in Drug Discovery.* H. Kubinyi, G. Müller, Eds., Wiley-VCH: Weinheim, 2004. pp 341-376.

G. Schneider, U. Fechner, Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649-663.

E. A. Schultes, D. P. Bartel, One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **2000**, *289*, 448-452.

R. Schroeder, A. Barta, K. Semrad, Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol*. **2004**, *5*, 908-919.

A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391-405.

C. H. Schwab, Conformational Analysis and Searching. In *Handbook of Cheminformatics*; J. Gasteiger, Ed.; Wiley-VCH: Weinheim, New-York, 2003; pp 262-301. ROTATE is available from Molecular Networks GmbH, Erlangen, Germany (http://www.mol-net.com).

A. K. Shanazarov, V. G. Granik, N. I. Andreeva, S. M. Golovina, M. D. Mashkovskii, Synthesis and pharmacological activity of pyrano[3,2-a]carbazole and pyrido[3,2-a]carbazole derivatives. *Khimiko-Farmatsevticheskii Zhurnal* **1989**, *23*, 1197-1200.

C. E. Shannon, W. Weaver, *The mathematical theory of communication.* University of Illinois Press: Urbana, USA, 1963.

R.P. Sheridan, S.B. Singh, E.M. Fluder, S.K. Kearsley, Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395-1406.

R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128-136.

R. P. Sheridan, S. K. Kearsley, Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903-911.

R. P. Sheridan, B. P. Feuston, V. N. Maiorov, S. K. Kearsley, Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912-1928.

R.D. Snyder, D. E. Ewing, L. B. Hendry, Evaluation of DNA intercalation potential of pharmaceuticals and other chemicals by cell-based and three-dimensional computational approaches. *Environ. Mol. Mutagen.* **2004**, *44*, 163-173.

Special journal issues on RNA as drug target: a) *Biopolymers* **2003**, *70*, 1-119; b) *Chembiochem* **2003**, *4*, 913-1106 .

S.-S. So, M. Karplus, Evolutionary optimization in quantitative structure activity relationship: an application of genetic neural networks. *J. Med. Chem.* **1996**, *39*, 1521-1530.

Specs, Delft, The Netherlands. (http://www.specs.net)

W. P. Spooren, F. Gasparini, T. E. Salt, R. Kuhn, Novel allosteric antagonists shed light on mglu(5) receptors and CNS disorders. *Trends Pharmacol. Sci.* **2001**, *295*, 1267-1275.

M. Stahl, N. P. Todorov, T. James, H. Mauser, H. J. Böhm, P. M. Dean. A validation study on the practical use of automated de novo design. *J. Comput. Aided Mol. Des.* **2002**, *16*, 459-478.

F. L. Stahura, J. W. Godden, L. Xue, J. Bajorath, Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245-1252.

N. Stiefl, K. Baumann, Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique. *J. Med. Chem.* **2003**, *46*, 1390-1407.

S. J. Sucheck, C. H. Wong, RNA as a target for small molecules. *Curr. Opin. Chem. Biol.* **2000**, *4*, 678-686.

Supuran, C. T., Scozzafava, A., Casini, A. Carbonic anhydrase inhibitors. *Med. Res. Rev.* **2003**, 23, 146-189.

C- J. Swanson, M. Bures, M. P. Johnson, A.-M. Linden, J. A. Monn, D. D. Schoepp, Metabotropic glutamate receptors as novel targets for anxiety and stress disorders. *Nat. Rev. Drug Discov.* **2005**, *4*, 131-144.

J. Tao, A. D. Frankel, Specific binding of arginine to TAR RNA. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 2723-2726.

A. L. Tarentino, T. H. Plummer Jr., The first demonstration of a procariotic glycosylasparaginase. *Biochem. Biophys. Res. Commun.* **1993**, *197*, 179-186.

S. J. Teague, A. M. Davis, P. D. Leeson, T. I. Oprea, The design of leadlike compbinatorial libraries, *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 3743-3748.

S. J. Teague, Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov*. **2003**, *2*, 527-542.

A. Teckentrup, H. Briem, J. Gasteiger, Mining high-throughput screening data of combinatorial libraries: development of a filter to distinguish hits from nonhits. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 626-634.

L. Terfloth, J. Gasteiger, Neural networks and genetic algorithms in drug design. *Drug Disovc. Today*, **2001**, *6*, S102-S108.

R. Todeschini, V. Consonni, *Handbook of MolecularDescriptors*, Wiley-VCH: Weinheim, 2000.

P. Tollman, P. Guy, J. Altshuler, A. Flanagan, M. Steiner. *A revolution in R&D. How genomics and genetics are transforming the biopharmaceutical industry*. The Boston Consulting Group. 2001.

N. Z. Tugusheva, S. Y. Ryabova, N. P. Solov'eva, V. G. Granik, Investigation of indolo[2,1-b]quinolines. *Chem. Heterocyc. Comp*. **1998**, *34*, 216-221.

M. Varney, N. D. Cosford, C. Jachec, S. P. Rao, A. I. Sacaan, F. F. Lin, L. Bleicher, E. Santori, P. J. Flor, H. Allgeier, F. Gasparini, R. Kuhn, S. D. Hess, G. Veliçelebi, E. C. Johnson, SIB-1757 and SIB-1893: Selective, Noncompetitive Antagonists of Metabotropic Glutamate Receptor Type 5. *J. Pharmacol. Exp. Ther*. **1999**, *290*, 170-181.

M. Wagener, J. Sadowski, J. Gasteiger, Aurocorrelation of molecular surface properties for modeling *corticosteroid binding globulin* and cytosolic *Ah* receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.

W. P. Walters, M. T. Stahl, M. A. Murcko, Virtual screening – an overview. *Drug Discov. Today* **1998**, *3*, 160-178.

B. Wang, J. M. Vernier, S. Rao, J. Chung, J. J. Anderson, J. D. Brodkin, X. Jiang, M. F. Gardner, X. Yang, B. Munoz, Discovery of novel modulators of metabotropic glutamate receptor subtype-5. *Bioorg. Med. Chem.* **2004**, *12*, 17-21.

R. Wang, X. Fang, S. Wang, The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *37*, 2977-2980.

Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C. T., Scozzafava, A., Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, 47, 550-557.

C. G. Wermuth, C. R. Gannelin, P. Lindberg, L. A. Mitscher, Glossary of terms used in medicinal chemistry. *Pure & Appl. Chem.* **1998**, *70*, 1129.

P. Willett, J. M. Barnard, G. M. Downs, Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.

J. R. Williamson, Induced fit in RNA-protein recognition. *Nat. Struct. Biol.* **2000**, *7*, 834-837.

W. Winkler, A. Nahvi, R. R. Breaker, Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **2002**, *419*, 952-956.

W. C. Winkler, S. Cohen-Chalamish, R. R. Breaker, An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 15908-15913.

G. Wolber, T. Langer, LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160-169.

H. Wold, Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*; P. R. Krishnaiah, Ed.; Academic Press, New York, 1966, pp 391-420.

H.Wold, Path models with latent variables: The NIPALS approach. In *Quantitative Sociology: International perspectives on mathematical and statistical model building*, H.M. Blalock et al., Eds.; Academic Press, New York, 1975, pp 307-357.

Woo, L. W., Howarth, N. M., Purohit, A., Hejaz, H. A., Reed, M. J., Potter, B. V. Steroidal and nonsteroidal sulfamates as potent inhibitors of steroid sulfatase. *J. Med. Chem.* **1998**, 41, 1068-1083.

Q. Xu, D. Buckley, C. Guan, H.-C. Guo, Structural insight into the mechanism of intramolecular proteolysis. *Cell* **1999**, *98*, 651-661.

Y. Xu, M. Johnson, Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J Chem Inf Comput Sci* **2001**, *41*, 181.

Y. J. Xu, M. Johnson, Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J Chem Inf Comput Sci* **2002**, *42*, 912.

H. Xu, D. K. Agrafiotis, Retrospect and prospect of virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2002**, *2*, 1305-1320.

L. Xue, F. L. Stahura, J. W. Godden, J. Bajorath, Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746-753.

J. R. Zaman, P. J. A. Michiels, C. A. A. van Boeckel, Targeting RNA: new opportunities to address drugless targets. *Drug Discov. Today* **2003**, *8*, 297-306.

I. Zamora, T. I. Oprea, G. Cruciani, M. Pastor, A. L. Ungell, Surface descriptors for protein-ligand affinity prediction. *J. Med. Chem.* **2003**, *46*, 25-33.

J. Zupan, J. Gasteiger, Neural *Networks in Chemistry and Drug Design* (2nd edn.), Wiley-VCH: Weinheim, 1999.

Lebenslauf

Zur Person

Steffen Renner, geboren am 20.12.1975
in Freiburg im Breisgau,
Staatsangehörigkeit: deutsch,
verheiratet, eine Tochter.

Schulische Ausbildung

1982 – 1986	Grundschule Umkirch
1986 - 1995	Wentzinger Gymnasium Freiburg im Breisgau
07/95	Abitur

Hochschulbildung

10/97	**Studium der Biologie** an der Albert-Ludwigs-Universiät, Freiburg im Breisgau
10/99	Vordiplom in Biologie / Hauptstudium mit Fokus auf Biochemie, Genetik, Bioinformatik und Informatik
02/02 – 02/03	**Diplomarbeit** in der Arbeitsgruppe von Prof. Dr. Wolfgang Haehnel Albert-Ludwig-Universität Freiburg im Breisgau *"Computergestützter Entwurf und Synthese einer kombinatorischen Bibliothek von Cytochrom P450 analogen Vier-Helix-Bündel Proteinen"*
04/03 – 06/05	**Promotion** in der Arbeitsgruppe von Prof. Dr. Gisbert Schneider Johann Wolfgang Goethe-Universität Frankfurt am Main *"Development and application of fast pharmacophore-based virtual screening methods"*
04/04 – 06/05	Stipendiat von Merz Pharmaceuticals GmbH Frankfurt am Main
seit 06/05	**Postdoktorale Arbeit** mit Merz Pharmaceuticals GmbH Frankfurt am Main