Identification of Structure Activity Relationships in Primary Screening Data of High-Throughput Screening Assays

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

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

von
Alexander Dietmar Böcker-Felbek
aus Laupheim

Frankfurt am Main, 2006
(D30)

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. Holger Stark

Datum der Disputation: 17. April 2007, 11:00 Uhr

Danksagung

An dieser Stelle gilt mein Dank Professor Dr. Gisbert Schneider, für seine vorbildliche und stimulierende wissenschaftliche Betreuung. Er hat mich in allen Phasen dieser Arbeit gefördert und stand mir stets mit anregenden Diskussionen zur Seite. Ich danke ihm für die schöne und lehrreiche Zeit, die ich mit ihm verbringen durfte.

Ebenso geht mein Dank an Dr. Andreas Teckentrup von der Firma Boehringer Ingelheim Pharma GmbH & Co. KG. Sein wissenschaftlicher Rat war mir immer von größter Wichtigkeit, und seine herausragende wissenschaftliche Betreuung hat mir geholfen, die vielen Hürden dieser Arbeit erfolgreich zu meistern. Vielen Dank.

Boehringer Ingelheim Pharma GmbH & Co. KG möchte ich danken für die großzügige Finanzierung meiner Arbeit. Besonders hervorheben möchte ich Dr. Ulrike Küfner-Mühl und Dr. Herbert Köppen, die mir ihr Vertrauen entgegengebracht haben und mir einen Platz in ihrem Team zur Verfügung gestellt haben.

Mein besonderer Dank geht an Prof. Dr. Holger Stark für seine wissenschaftliche Unterstützung, seine hervorragenden Anregungen und sein mir entgegengebrachtes Vertrauen beim virtuellen Screening nach Dopaminrezeptorantagonisten. Mein Dank gilt Britta Sasse für die pharmakologische Testung der Liganden am Dopamin D_2- und D_3-Rezeptor, für ihre Geduld mit mir und für ihre wissenschaftliche Leistung mit der sie das gemeinsame wissenschaftliche Projekt bereichert und belebt hat.

Mein herzlichster Dank gilt meinen Biberacher Kollegen Dr. Bernd Beck, Michael Bieler, Dr. Thomas Fox, Dr. Sandra Handschuh, Dr. Kai Hasselbach, Kerstin Höhfeld, Dr. Anne Joergensen, Johannes Koppe, Dr. Jan Kriegl, Dr. Uta Lessel, Dr. Peter Stauffert, Dr. Christofer Tautermann, Dr. Bernd Wellenzohn, Dr. Oliver Wissdorf und den unzähligen Diplomanden und Praktikanten. Es war mir eine große Ehre in einem Team solch exzellenter Wissenschaftler zu forschen. An dieser Stelle geht mein Dank auch an die Gruppe High-Throughput Screening für die Bereitstellung der HTS-Ergebnisse und das Näherbringen der Versuche.

Mein besonderer Dank geht an meine Frankfurter Kollegen Dr. Evgeny Byvatov, Norbert Dichter, Uli Fechner, Lutz Franke, Alireza Givehchi, Tina Grabowski, Michael Meissner, Manuel Nietert, Tobias Noeske, Dr. Steffen Renner, Brigitte Scheidemantel-Geiß, Michael Schmucker, Dr. Petra Schneider und Andreas Schüller, die mir stets mit Rat und Tat zur Seite gestanden haben und mich so liebevoll während meiner Frankfurter Zeit in ihren Kreis aufgenommen haben. Besonders hervorheben möchte ich Svetlana Derksen und Elena Schmidt für deren herausragende Leistung bei der Implementierung des hierarchischen *k*-means Algorithmus. Ich möchte auch dem Beilstein-Institut zur Förderung der Chemischen Wissenschaften meinen Dank aussprechen.

Inniger Dank geht an meine Familie, meine Freunde und das Biberacher Laufteam, die mir stets eine große Stütze während dieser Arbeit waren.

List of Publications

Publications:

Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS Data Mining Approaches. QSAR Comb. Sci. **2004**, *23*, 207-213.

Böcker, A., Derksen, S., Schmidt, E., Teckentrup, A., Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. J. Chem. Inf. Model. **2005**, *45*, 807-815.

Böcker, A., Schneider, G., Teckentrup, A. NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and its Application in Virtual Screening. J. Chem. Inf. Model. **2006**, *46*, 2220-2229.

Böcker, A., Sasse, B. C., Nietert, M., Stark, H. and Schneider, G. GPCR Targeted Library Design: Novel Dopamine D_3 Receptor Ligands, Chemmedchem, **2007**, accepted.

Renner, S., Noeske, T., Böcker, A., Schmucker, M., Parsons, C. G., Weil, T., Schneider, G. Scaffold-Hopping by 3D-Pharmacophores and Neural Network Ensembles. Angew. Chem. Int. Ed Engl., **2007**, accepted.

Oral Presentations:

Böcker, A., Schneider, G., Teckentrup, A. Hierarchical Clustering of Huge Compound Libraries: Theory and Application. 19. Darmstädter Molecular Modelling Workshop, **May 2005**, Erlangen, Germany.

Poster Presentations:

Böcker, A., Teckentrup, A., Schneider, G. NIPALSTREE: A New Approach for Clustering Large Data Sets. *18. Darmstädter Molecular Modelling Workshop*, **May 2004**, Erlangen, Germany.

Böcker, A., Teckentrup, A., Schneider, G. NIPALSTREE: A New Approach for Clustering Large Data Sets. 6th. *Boehringer Ingelheim Structural Research and Computational Chemistry Meeting,* **June 2004**, Biberach, Germany.

Böcker, A., Schneider, G., Teckentrup, A. Hierarchical Clustering of Huge Compound Libraries: Interactive SAR Analyses. *19. Darmstädter Molecular Modelling Workshop*, **May 2005**, Erlangen, Germany.

Böcker, A., Schneider, G., Teckentrup, A. Hierarchical Clustering of Large Compound Libraries: Theory and Application. *1. German Conference on Chemoinformatics*, **November 2005**, Goslar, Germany.

Renner, S., Noeske, T., Böcker, A., Weil, T., Schneider, G. Combining Supervised and Unsupervised Neural Networks for the Identification of Novel Scaffolds of Metabotropic Glutamate Receptor 5 (mGluR5) Modulators. *1. German Conference on Chemoinformatics*, **November 2005**, Goslar, Germany.

Sasse, B.C., Böcker, A., Schneider, G., Stark, H. Classification Techniques in Chemoinformatics: Lead Identification Strategies for Dopamine D_3 Receptor Ligands. *Frontiers in Medicinal Chemistry,* **March 2006**, Frankfurt am Main, Germany.

List of Abbreviations

+H	Positive Hydrophobic
1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
4D	Four Dimensional
ACD	Advanced Chemical Directory
ACE	Angiotensin Converting Enzyme
ADME-T	Absorption, Distribution, Metabolism, Excretion - Toxicity
AH	Acceptor Hydrophobic
Ala	Alanine
ALK	Activin-like Kinase
AMP	Adenosine-Mono-Phosphate
ANN	Artificial Neural Network
Asp	Aspartic Acid
ATP	Adenosine Tri-Phosphate
BRANN	Bayesian Regularized Artificial Neutral Network
CATS	Chemically Advanced Template Search
CHO	Chinese Hamster Ovary
CNS	Central Nervous System
COBRA	Collection Of Bioactive Reference Analogues
CoMFA	Comparative Molecular Field Analysis
COPD	Chronic Obstructive Pulmonary Disease
COX	Cyclo-Oxygenase
CPU	Central Processing Unit
CR	Confirmation Rate
CTL	Control
CV	Cross Validation
DA	Dopamine
DMSO	Di-Methyl-Sulf-Oxide
DPP	Dipeptidyl Peptidase
DTT	Di-Thio-Treitole
EC_{50}	Concentration of a ligand that is required for a 50% effect

EF	Enrichment Factor
EGTA	Ethyleneglycotetraacetic Acid
EMT	Epithel-Mesenchym Transition
FLIPR	Fluorometric Imaging Plate Reader
FN	False-Negatives
FNR	False-Negative Rate
FP	False-Positives
FPR	False-Positive Rate
FRET	Fluorescence Resonance Energy Transfer
GABA	Gamma Amino Butyric Acid
Glu	Glutamic acid
Gly	Glycine
GOLD	Genetic Optimization for Ligand Docking
GPCR	G-Protein Coupled Receptor
GUI	Graphical User Interface
HH	Hydrophobic Hydrophobic
His	Histamine
HIV	Human Immunodeficiency Virus
HOBT	Hydroxybenzotriazole
HTS	High-Throughput Screening
IU	IUPAC Units
IC_{50}	Concentration of a ligand that is required for a 50% inhibition
ICE	Interleukin 1 Cleaving Enzyme
KBD	Kullback Leibler Distance
K_{D}	Equilibrium constant for dissociation
K_{i}	Binding constant
logD	Logarithm of the octanole water distribution coefficient
logP	Logarithm of the octanole water partition coefficient
LSC	LEADseeker Count
Lys	Lysine
MAPK	Mitogene Activated Protein Kinase
MB	Megabyte
MCS	Maximum Common Substructure
MDDR	MDL Drug Data Report

MMFF	Merck Molecular Force Field
MMP	Matrix Metallo-Protease
MOE	Molecular Operation Environment
NBE	New Biological Entity
NCE	New Chemical Entity
NIPALS	Non-linear Iterative Partial Least Squares
NLCA	Non-Linear Component Analysis
NMR	Nuclear Magnetic Resonance
NP	Non-Polynomial
PATTY	Programmable Atom Typer
PBS	Phosphate Buffered Saline
PC	Principal Component
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PEOE	Partial Equalization of Orbital Electro-negativity
$\mathrm{p}K_\mathrm{a}$	Logarithm of the Acid constant K_a
$\mathrm{p}K_\mathrm{i}$	Logarithm of the binding constant K_i
PLS	Partial Least Squares
PPP	Potential Pharmacophore Points
QSAR	Quantitative Structure Activity Relationships
QSPR	Quantitative Structure Property Relationship
RAM	Random Access Memory
RBF	Radial Basis Function
RMSE	Root Mean Square Error
rpm	Rotation Per Minute
R-SMAD	Receptor SMAD
SAR	Structure Activity Relationship
SCA	Stochastic Cluster Analysis
SD	Standard Deviation
SE	Shannon Entropy
SMR	Sum of Molar Refractivity
SOM	Self Organizing Map
SPA	Scintillator Proximity Assay
s*SE*	scaled Shannon Entropy

SVL	Support Vector Language
SVM	Support Vector Machine
SVR	Support Vector-Based Regression
TGF	Transforming Growth Factor
TM	Transmembrane
TN	True-Negatives
TP	True-Positives
TRET	Time Resolved Energy Transfer
Tyr	Tyrosine
UFS	Unsupervised Forward Selection
UK	United Kingdom
VIP	Variable Importance Plot
VLA	Very Late Antigen
VS	Virtual Screening
VSA	Van der Waals Surface Area

Table of Contents

1 Introduction 1
1.1 The Drug Discovery Process 2
1.2 Lead Identification Strategies 3
1.2.1 Virtual Screening 3
1.2.2 High-Throughput Screening 6
1.2.3 Computational Chemistry in the Lead Identification Phase 11
1.3 Transforming Growth Factor-β Type I Receptor Inhibitors 19
1.3.1 Biological Relevance 19
1.3.2 Disease Implication of the TGF-β Signalling Cascade 20
1.4 Dopamine D_3 Receptor Ligands 23
1.4.1 The Dopaminergic Pathways 24
1.4.2 Therapeutic Relevance of Selective Dopamine D_3 Receptors 26
1.4.3 Dopamine D_3 Receptor Agonists, Partial Agonists and Antagonists 28
1.5 Scope of the Thesis 32

1.5 Scope of the Thesis 32

2 Experimental Techniques 33
2.1 High Throughput Screening Techniques 34
2.1.1 TGF-β Type I Receptor 34
2.2 Dopamine Receptor Binding Studies 36
3 Computational Techniques 38
3.1 Data Sets 39
3.1.1 COBRA 39
3.1.2 MDDR 39
3.1.3 SPECS Catalogue 40
3.1.4 HTS Data 40
3.1.5 Dopamine Data 42
3.1.6 Fisher’s Iris Data 42
3.2 Data Preparation Strategies 43
3.2.1 Data Filtering 43
3.2.2 Removal of Counter Ions 44
3.2.3 Neutralization and Charge Assignment 45
3.2.4 Conformer Generation 45

3.2.5 Ontology assignment ... 46
3.3 Molecular Similarity, Descriptors and Descriptor Selection ... 46
3.3.1 CATS 2D ... 47
3.3.2 CATS 3D ... 48
3.3.3 MOE 2D ... 48
3.3.4 Daylight Fingerprints ... 48
3.3.5 Descriptor Preparation and Selection ... 49
3.3.6 The Similarity Concept in Virtual Screening ... 51
3.4 Unsupervised Classification Techniques ... 53
3.4.1 Molecular Scaffold Analysis ... 53
3.4.2 Maximum Diversity Selection ... 54
3.4.3 Principle Component Analysis ... 55
3.4.4 Non-linear Component Analysis ... 56
3.4.5 Clustering Techniques ... 57
3.5 Supervised Classification and Regression Techniques ... 58
3.5.1 Partial Least Squares ... 59
3.5.2 Support Vector-based Regression ... 59
3.5.3 Bayesian Regularized Artificial Neural Networks ... 61
3.5.4 Validation of Classification and Regression Techniques ... 65
3.6 The Pharmacophore Concept ... 67
3.7 Docking Techniques ... 68
3.7.1 GOLD Docking ... 69
4 Results and Discussion ... 71
4.1 Algorithm Development ... 72
4.1.1 Data Sets ... 72
4.1.2 NIPALSTREE ... 72
4.1.3 Outlier PCA ... 77
4.1.4 Hierarchical k-means ... 78
4.1.5 The Stop Threshold Concept ... 79
4.1.6 Statistical Evaluation of Cluster Dendrograms ... 80
4.1.7 Application of NIPALSTREE ... 86
4.1.8 Application of Hierarchical *k*-means ... 91
4.1.9 Comparison of Both Algorithms by Virtual Screening ... 99
4.1.10 Conclusions ... 103

4.2 SAR Analyses in the Cluster Dendrogram 105
4.2.1 General Analysis 106
4.2.2 Terminal Cluster Analyses 112
4.2.3 Relevant Clusters, Selectivity and Specificity 115
4.2.4 Conclusions 124
4.3 Retrospective Analyses of HTS Assays 126
4.3.1 Assay Characterization 126
4.3.2 Not Confirmed Hits 129
4.3.3 False-Negatives 135
4.3.4 Conclusions 144
4.4 Prospective Analysis of Dopamine D_3 Receptor Antagonists 147
4.4.1 Characterization of the Dopamine Data Set 147
4.4.2 Clustering-Based Virtual Screening 149
4.4.3 SAR 152
4.4.4 GOLD Docking 153
4.4.5 Pharmacophore-Based Virtual Screening 155
4.4.6 Regression-Based Activity Prediction 158
4.4.7 Conclusions 164
5 Conclusion and Outlook 166
6 Summary 170
7 Zusammenfassung 173
8 Ausführliche Zusammenfassung 176
9 Literature 182
10 Appendix 209
A Dopamine D_3 Receptor Ligands 210
B Fisher's Iris Data Set 244
C Example Structures of Reactive Functional Groups 246
D Parameter file for SVR 248
E Parameter file for Bayesian regularized artificial neural networks 250
F Parameter file for GOLD 255
11 Curriculum Vitae 257

1 Introduction

1.1 The Drug Discovery Process

The drug discovery process for new chemical entities (NCE) comprises four main sections, namely target identification, lead identification, lead optimization and clinical development as illustrated by a pipeline [Bleicher et al., 2003] (Figure 1.1). New biological entities (NBE), such as monoclonal antibodies, vaccines and protein drugs, are a complementation to NCE, and have been successfully introduced into the market [Adams & Weiner, 2005, Sodoyer & Laffly, 2005]. Given the fact that the later stages of the drug discovery process are characterized by a high attrition rate [Bleicher et al., 2003] the aim of the first three phases is to enter varies NCEs into the clinical phase with a higher rate and improved quality. Since the rate is determined by the slowest process in the pipeline (the bottleneck), the challenge is to expand these bottlenecks allowing more candidate molecules to pass these stages. Computational chemistry comprises techniques supporting the drug discovery process in the lead identification and optimization phase. The presented work concentrates on the integration of these methods into the lead identification phase.

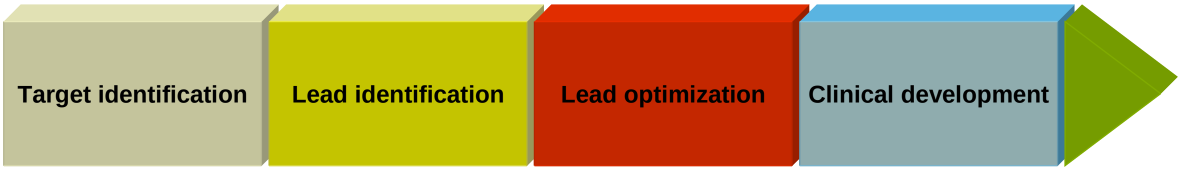


Figure 1.1 The drug discovery process illustrated by a pipeline. Four sections, target identification, lead identification, lead optimization and clinical phase are distinguished.

Initially, new targets are identified (mainly proteins) whose target-selective modulation translates into high therapeutic effect and minimal in vivo side effects [Egner et al., 2005; Knowles & Gromo, 2003; Drews, 2000]. In the lead identification phase leads or lead series are identified for a target linked with a distinct disease. A lead is defined as “a prototypical chemical structure or series of structures that demonstrate activity and selectivity in a pharmacological or biochemical relevant screen” [Bleicher et al., 2003]. High-throughput screening (HTS) [Oldenburg et al., 2001] and virtual screening (VS) [Böhm & Schneider, 2000, Bajorath, 2002] present the two main lead identification strategies and correspond to each other. To define leads or lead series within the identified active molecules selection criteria are applied best covering suitable molecular properties, favourable pharmacodynamics (potency, selectivity and efficacy), acceptable pharmacokinetic properties, chemical optimization potential and patentability [Steinmeyer, 2006]. A promising approach in the lead selection process is computational chemistry which is defined as a discipline using mathematical methods for the calculation of molecular properties or for the simulation of molecular behaviour [Wermuth et al., 1998]. Computational chemistry techniques provide a clustering of data helping to define lead series [Böcker et al., 2003] and offer the ability to create models in order to identify false-negatives and false-positives in the data [Harper et al., 2001; Glick et al., 2004]. An extended integration of computational chemistry in the lead identification process with the aim to derive first crude structure-activity relationships (SAR) in the data, is assumed to be worthwhile for a further rationalization of the selection process.

1.2 Lead Identification Strategies

In the drug discovery process for NCEs several alternative lead identification strategies exist. The two most prominent are HTS and VS. In HTS complete compound repositories or a subset thereof are tested in a miniaturized and automated biochemical or cell-based assay to determine hits against a certain target. To cope with false-negatives emerging from HTS or to identify alternative hits in external vendor catalogues or virtual libraries, VS is applied. Both techniques have the aim to identify starting structures (hits) which might be translated into leads with novel scaffolds. Both screening methods have some characteristics in common. At first, both try to minimize false-negatives and false-positives. This is sometimes referred to as robustness of a technique. Further they usually result in a limited number of hits in a large quantity of non-hits. The imbalance of the number of hits and non-hits and the noise resulting from false-positives and false-negatives are the main challenges for the application of computational chemistry techniques to understand the SAR in the data [Schreyer et al., 2004].

1.2.1 Virtual Screening

Virtual screening techniques can be separated into structure-based and ligand-based screening methods [Bajorath, 2002]. The first method employs the target structure for screening and the second uses information derived from known ligands. They are founded on the similarity principle stating that similar molecules exhibit similar biological effects [Johnson & Maggiora, 1990; Martin et al., 2002]. Another discrimination of VS methods can be done according to the dimensionality. Methods using a one-dimensional (1D), two-dimensional (2D), three-dimensional (3D) or four-dimensional (4D) description can be defined [Bleicher et al., 2003]. One-dimensional techniques comprise methods for compound filtering according to unwanted (reactive) fragments, pharmacokinetic ADME-T properties (Absorption, Distribution, Metabolism, Excretion and Toxicity), and drug-like or lead-like criteria [Van der Waterbeemd & Gifford, 2003; Muegge, 2003]. These methods can be applied as pre-filter for large data sets. In 2D or 3D methods, screening is based on the 2D or 3D representation of the molecules, respectively. The 3D search methods are more complex compared to the 2D methods since in addition to the topology of the molecules stereoisomer, tautomer and conformer representations have to be considered [Kitchen et al., 2004]. Despite of the spatial arrangement of the structures in 2D or 3D, the representation of the molecules as numerical or bit-string representations describing e.g. the occurrence of substructure elements, potential pharmacophore points or physicochemical properties plays a crucial role to leave a molecular scaffold and identify novel compounds with novel scaffolds [Renner & Schneider, 2006]. The identification of alternative scaffolds is of importance since they can result in alternative lead structures which are synthetically easier to access, bear more suited ADME-T or pharmacodynamic properties or are not covered by intellectual properties. Examples of 2D and 3D virtual screening methods are similarity or substructure search methods [Willett, 2005], binary classification methods like recursive partitioning [Rusinko et al., 1999], naïve Bayes' classification [Xia et al., 2004] or Support vector machine based classification [Byvatov et al., 2003] and 3D pharmacophore methods [Güner, 1999; Renner & Schneider, 2004]. In light of this, predictions based on a 2D representation of molecules were shown to perform equally well or better than the comparable predictions based on a 3D representation [Bajorath, 2002; Zhang & Muegge, 2006]. This might be a consequence of additional degrees of freedom in the 3D representation caused by translational, rotational and conformational flexibility of the molecules. 4D techniques represent molecular docking methods, which employ in addition to the 3D representation of the ligands, the target receptor for selecting suitable ligands [Kitchen et al., 2004]. The docking process can be divided into the geometric process of posing the conformational representations of a ligand into the binding pocket and into measuring the interaction strength (scoring). Whereas solutions for the posing can be obtained lying within 2 Å root mean square deviations from the same molecule in the crystal structure, the scoring process is not accurate enough for reliable affinity predictions [Warren et al., 2006]. Reasons are that scoring functions consider mostly enthalpic effects of ligand binding, whereas entropic effects like desolvatation of ligand and receptor are not well integrated. Further, both proteins and ligands are flexible and might mutually induce a different 3D representation, further increasing the complexity of the problem. Despite of that, 4D techniques have been successfully employed for virtual screening [Kitchen et al., 2004].

A third separation of virtual screening methods is into unsupervised and supervised classification or regression methods. The first method classifies compounds only according to their inherent properties. Examples are clustering and partitioning methods [Böcker et al., 2004]. In contrast to that, supervised classification techniques train models based on a predefined classification. The created models are then employed for classifying new compounds. A large variety of such classification methods exist. They can be subdivided into methods coping with linear or non-linear correlation in the data. The advantage of non-linear classification methods is that correlations can be identified which might remain undiscovered using a linear classifier. However these methods translate into so called “black box” models which are difficult to interpret. Examples of linear classification techniques are recursive partitioning [Rusinko et al., 1999] or naïve Bayes’ classification [Xia et al., 2004]. Examples of non-linear classification methods are neural networks [Schneider, 2000] or SVM based classification [Byvatov et al., 2003]. In various applications SVM were shown to outperform other classification methods [e.g. Byvatov et al., 2003, Glick et al., 2006]. At this point it is important to consider that all classification models are intended to allow valid predictions only for the covered chemical data space. An extrapolation is not possible [Sheridan et al., 2004; Polanski et al., 2005].

The effectiveness of a virtual screening method can be assessed in two ways, retrospective and prospective. A retrospective application specifies how many known actives are retrieved from a database in combination with compounds of unknown activity. In contrast, a prospective application is the application to a database of compounds with unknown activity. It includes the ordering and experimental testing of the identified hits. Figure 1.2 shows an example of a prospective virtual screening campaign. Prior to searching, molecule sets (e.g. vendor catalogues) are filtered to either eliminate compounds with undesired properties or retain molecules containing privileged substructure motives [Oprea & Matter, 2004]. Since different screening techniques and different descriptor sets were shown to identify different hit classes [Shanmugasundaram et al. 2005], several such methods can be applied for one virtual screening task. The obtained result lists can be filtered to receive compounds bearing novel scaffold [Saeh et al., 2005; Renner & Schneider, 2006]. This might include the rejection of compounds protected by patents or too similar to known actives. In most applications result lists exceed the maximum number of compounds that can be reasonably handled by experimental screening. Consequently, the result lists have to be further narrowed down. This might be achieved by prioritizing compounds obtained with different methods (ensemble prediction) [Svetnik et al., 2005; Breiman, 1996; Merkwirth et al., 2004] or by ‘cherry picking’ or by creating a maximum diverse representation of the resulting set [Reynolds et al., 2001].

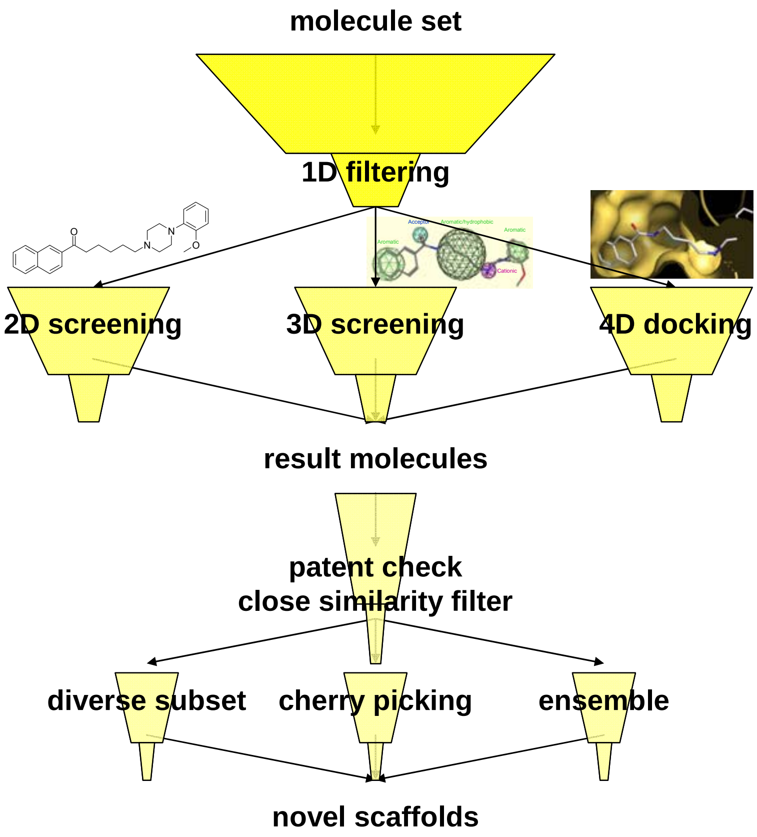


Figure 1.2 Workflow of a virtual screening campaign. After filtering compounds with undesired properties, different virtual screening techniques are applied. Resulting compounds are then filtered according to innovative potential. From the remaining list, either an ensemble of all results is selected or a cherry picking is performed or a maximum diverse subset is created.

1.2.2 High-Throughput Screening

In HTS either the complete compound pool of a pharmaceutical company (full screen) or a subset thereof (focused screen) or subsets in sequential order (sequential screen) is experimentally tested for affinity towards a certain receptor in an automated, miniaturized and cost-efficient way [Schnecke & Boström, 2006]. It is possible is to measure 100,000 and more compounds a day to identify novel lead compounds [Hertzberg & Pope, 2000]. The assay is set up that a robust separation of hits from non-hits is achieved, which corresponds to a high signal to noise or signal to background ratio [Bronson et al., 2001]. HTS has been successfully applied in lead identification campaigns [Golebiowski et al., 2001; Golebiowski et al., 2003]. It has to be pointed out that HTS is not always the matter of choice. HTS assays are cost-intensive; for 1,000,000 million compounds a global cost between $500,000 and $1,000,000 is estimated [Davies et al., 2006]. And HTS assays show hit rates below 2.5% which is exemplified according to the three HTS assays presented in this work (Table 3.1).

In HTS compounds are transferred from a dimethylsulfoxide (DMSO) stock solution into a micro-titer assay plate with 384 or 1,536 and more wells per plate. Additional reagents are added including target protein in case of a biochemical assay or cells for a cell-based assay. Following an incubation period the response signal is measured and converted into a percent inhibition or fold stimulation. To cope with measurement errors compounds judged as hits are further confirmed by determining the IC_{50} (EC_{50}) value, i.e. the molar concentration of a compound that is required for 50% inhibition (effect). In addition to establishing the biological assay itself in HTS, parameters have to be further adopted allowing an automated and miniaturized screening. These parameters range from assay criteria like compound concentration, enzyme/receptor/cell concentration over assay conditions like incubation time, temperature or pH value to screening parameters like appropriate detection (colorimetric, fluorescence, luminescence or radiometric signals), readout, liquid handling, plate handling and compound handling devices.

HTS assays have been categorized into homogeneous and heterogeneous assays [Walters & Namchuk, 2003]. Heterogeneous assays are multi-step assays including washing, filtration or transfer steps. In contrast homogeneous assays perform all steps in one mixture. For HTS the latter assays are preferred since they are easier to automate and less cost-intensive. However signal-to-background separation is more difficult.

When setting up a HTS assay it has to be decided whether a cellular assay or a biochemical assay has to be performed. Cellular assays employ the complete cells for testing instead of the isolated target. They have the advantage, that a functional characterization of the molecules can be obtained. Further additional properties like cellular toxicity can be simultaneously addressed. Finally, targets requiring additional (unknown) co-activators, co-repressors and other factors might not be addressable by biochemical assays [Walters & Namchuk, 2003; Johnston & Jonston, 2002]. In contrast, biochemical assays show less data scattering and are easier to follow since only one target is assessed. It allows identifying more structural classes since screening can be performed at higher concentrations and pharmacokinetic properties are not considered [Walters & Namchuk, 2003].

HTS assays have been categorized according to their measurement principles into homogeneous fluorescence methods, assays with radiometric readout and cell-based assays [Hertzberg & Pope, 2000]. For each type a variety of commercial solutions are available. For an explanation the principles of currently used assays are introduced. A more detailed description of assay techniques and their commercial solutions can be found in Seethala and Fernandes [Seethala & Fernandes, 2001].

Homogeneous fluorescence methods utilize a fluorophore which absorbs a photon. The fluorophore is excited form the ground state *G* to the singlet excited state *E* (typically in 10^{-15} s). By fluorescence emission the fluorophore falls back into the ground state (typically in 10^{-12} s). The emission can be non-radiatively transmitted from the donor state *E* to a second acceptor fluorophore by intermolecular long range dipole-dipole coupling. The acceptor fluorophore is excited to state *E'*. The transmission can be direct (typically in 10^{-9} s) or via an intermediate electronic state (typically in micro- or milliseconds). By fluorescence emission the second fluorophore falls back into the ground state *G'*. The process is illustrated as a Jablonski diagram in Figure 1.3 [Pope et al., 1999; Clegg, 1995].

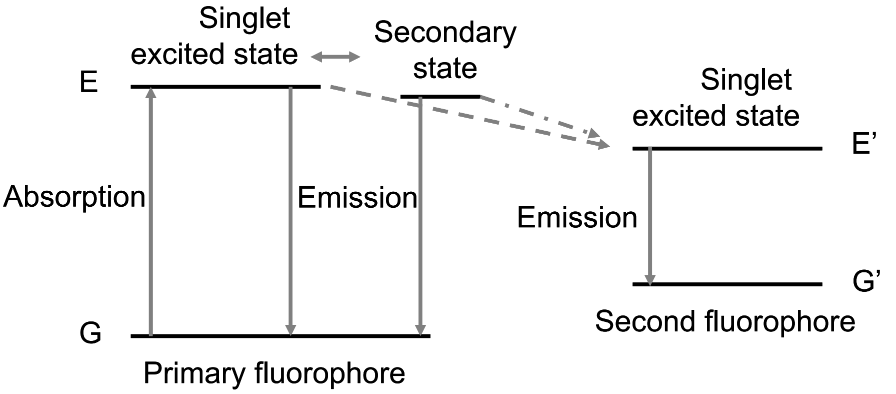


Figure 1.3 Jablonski diagrams of fluorescence absorption and emission. A primary fluorophore absorbs a photon and is excited from the ground state *G* to the singlet excited state *E*. By energy emission it falls back to *G*. The singlet excited state can be transformed to a secondary excited state. By intermolecular dipole-dipole coupling from the singlet excited state or the secondary state energy can be transferred to a second fluorophore. The second fluorophore is excited from the ground state *G'* to the singlet excited state *E'*.

The behaviour of absorption and emission is employed for HTS in three different ways. Fluorescence polarization/anisotropy follows the fact that a fluorophore (or a fluorescence-labelled ligand) which is excited with polarized light emits the light polarized. If the fluorophore is free in solution (rotational diffusion) the measured fluorescence appears depolarized (unbound ligand). In contrast if the fluorophore is bound to a receptor the fluorescence emission is polarized (bound ligand). The amount of polarized light can be measured quantitatively [Hertzberg & Pope, 2000; Pope et al., 1999].

Time resolved energy transfer (TRET, transfer via a second excited state) and fluorescence resonance energy transfer (FRET, prompt transfer) are based on the Förster theory stating that the efficiency of energy transmission from the excited state of a donor fluorophore to an acceptor fluorophore is dependent on the sixth power of the distance *R* between donor and acceptor (Eq. 1.1) [Clegg, 1995].

$$E = \frac{R_o^6}{R_o^6 + R^6} \tag{1.1}$$

R_0 defines the distance between acceptor and donor allowing 50% of the energy to be transferred non-radiatively. In practice Lanthanides (e.g. Europium cryptates coupled to a ligand) have shown to be promising donor fluorophores whereas allophycocyanin coupled to a target is an efficient acceptor fluorophore. Light is absorbed from the donor fluorophore at a certain wavelength transferred to the acceptor fluorophore and emitted at another wavelength. The maximum distance which can translate into an energy transfer lies in the low nanometre range [Seethala & Fernandes, 2001]. Only bound fluorophores are measured quantitatively.

A third homogeneous fluorescence method is fluorescence correlation spectroscopy. It uses confocal detection of small volumes (femtoliter) where only a few fluorescent molecules are present in combination with measuring the fluctuation of the emitted fluorescence (i.e. deviation from the average fluorescence intensity). By using autocorrelation algorithms different properties/behaviours of the fluorophore can be detected. A fluorophore bound to a receptor shows a quantifiable different fluorescence fluctuation compared to the unbound fluorophore in solution. The method is used in combination with FRET or fluorescence polarization [Eggelin et al., 2003; Pope et al., 1999].

HTS methods with radiometric readout immobilize the target onto a solid surface (e.g. a bead or a plate surface) which contains a scintillator (e.g. yttrium silicate). A ligand is labelled with a β-emitting radioisotope like ^{3}H or ^{35}S. The mean path length of the β-particle is 1.5 µm or 66 µm respectively. Scintillation molecules lying in this distance range absorb the energy of the β-particle and emit it proportionally as light (chemiluminescence). Consequently, only if the molecule is bound to the target it is close enough to the scintillator and energy is transferred. If a second molecule is present binding to the target, the radio-labelled ligand is displaced and no energy is transferred. Since the emission is proportional to the amount of absorbed energy, quantitative displacements can be measured [Seethala, 2001]. An example of a *scintillator proximity assay* is shown in Figure 1.4.

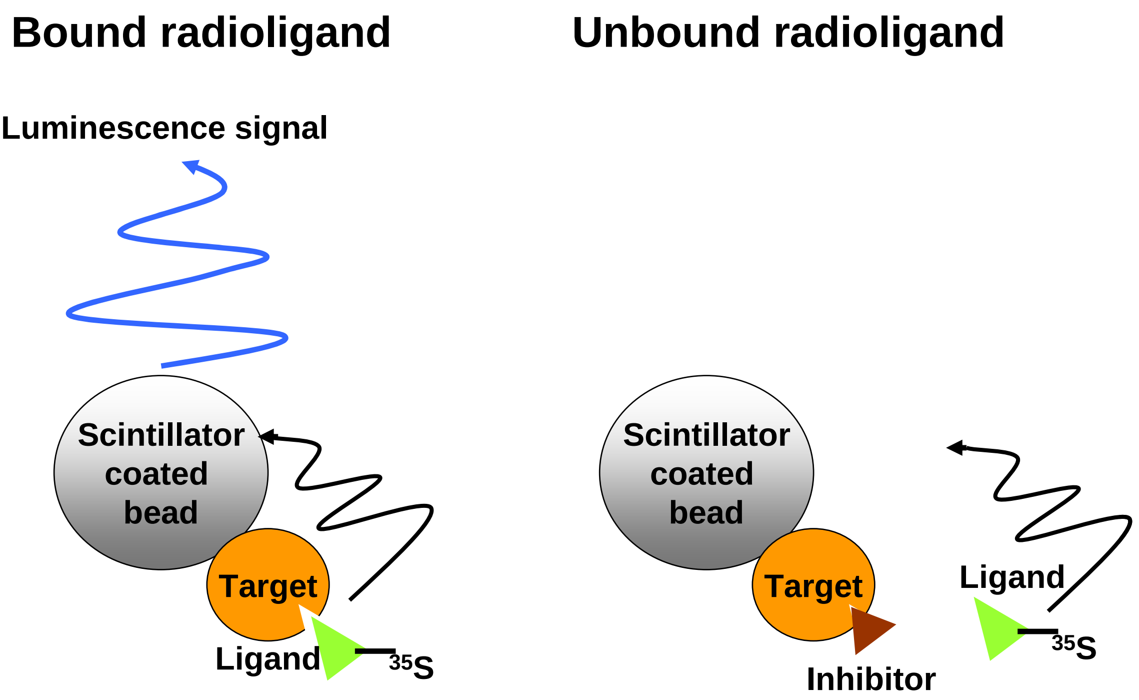


Figure 1.4 Example of a homogeneous HTS assay based on radioactive signal detection.

Cell-based HTS assays are performed with a wide spectrum of different methods. Examples are the measurement of calcium release by G-protein coupled receptors via a calcium sensitive fluorophore, reporter gene assays and confocal imaging platforms for cellular and sub-cellular imaging [Hertzberg & Pope, 2000]. Reporter gene assays measure the stimulation or the inhibition of a target indirectly by a reporter target like firefly luciferase. The reporter needs the outcome of the reaction of the first target to catalyze its own reaction. A quantifiable signal like chemiluminescence is measured. The advantage of the method is that the signal of the first reaction is amplified which allows miniaturization and separation of signal and background [Johnston & Johnston, 2002]. Cellular imaging is an upcoming new technology measuring cellular events with confocal microscopy. The target of interest is tagged with a fluorophore (e.g. red fluorescent protein) and expressed in a cell. Cellular events like agonist-induced translocation of a nuclear hormone receptor into the nucleus, inhibition of viral entry into a cell or inhibition of cell growth have been quantified [Lang et al., 2006]. By using different fluorophores in one experiment several events can be measured simultaneously (e.g. selectivity against a second target) [Lang et al., 2006].

1.2.3 Computational Chemistry in the Lead Identification Phase

A variety of computational methods have become an integral part of the lead identification phase. On overview is given in Figure 1.5. Library design is applied with the aim to create chemical repositories with suitable pharmacokinetic and drug-like properties [Bleicher et al., 2003]. Screening libraries are designed to have a mostly homogenous representation of the company's (drug-like) chemical space showing no singletons or drastically overrepresented regions [Nilakantan & Nunn, 2003]. Finally by employing e.g. knowledge about privileged substructures for a receptor or a receptor family it is tried to create targeted or focused libraries [Klabunde & Hessler, 2002; Bissantz et al., 2005]. Summarizing these methods try to increase the likelihood of identifying hits in HTS.

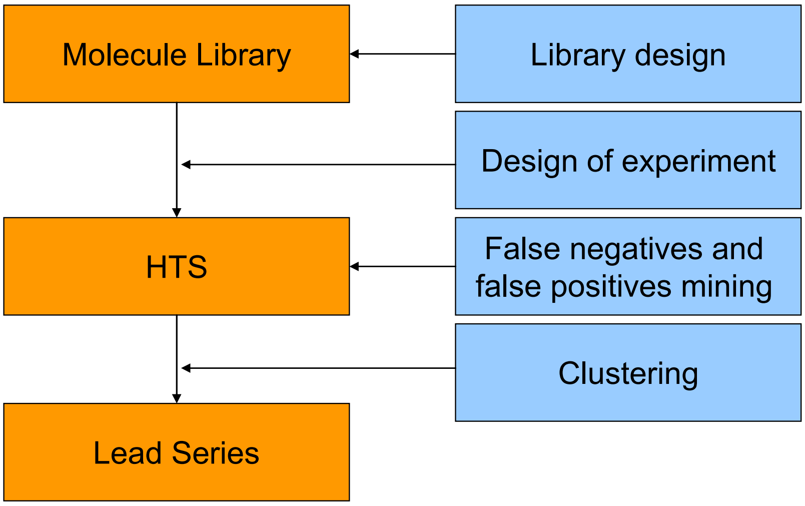


Figure 1.5 Implication of computational chemistry techniques in the HTS process.

After performing a HTS assay, clustering or partitioning techniques are employed to decipher lead series in the screening data [Böcker et al., 2004]. Such series are prioritized according to predicted or experimentally determined molecular, pharmacokinetic and pharmacodynamic properties. Suitable property ranges have been suggested by Steinmeyer et al. and may include a molecular weight between 200 and 500 Dalton, a logarithm of the octanole water participation coefficient between -1 and +5, solubility in water above 5 mg/L, number of hydrogen bond donors below 5, number of hydrogen bond acceptors below 10, caco2 cell permeability above 100 $cms^{-1}x10^{-7}$, human or rat liver microsome stability between 50% and 80% resistance after 30 minutes, no measurable cytochrome P450 interaction, 10 fold selectivity towards related targets, nanomolar potency, cellular activity and many more [Steinmeyer et al., 2006]. Applying these multi-property criteria to the identified leads results in only a few which survive. Consequently there is a strong need for backup series which can be identified by mining the HTS data for false-negatives [Harper et al., 2001, Polgàr et al., 2005, Glick, et al., 2005]. On the other hand the lead profiling is a work- and cost-intense process. The number of false-positives entering the profiling should be as low as possible. According to this it is tried to derive first simple SAR from HTS assays, helping to mine the HTS data for false-positives [Roche et al., 2002, Rishton, 2003].

Mining of False-Positives and False-Negatives in HTS

The theoretical background for determining false-positives (FP) and false-negatives (FN) in HTS has been characterized by Zhang et al. [Zhang et al., 2000]. Activity is measured as percent control (% CTL) which corresponds to the degree of inhibition or stimulation. The obtained % CTL values for the compounds of a library can be approximated by a function *f(υ)*. In a simple case the values are Gaussian distributed (1.2).

$$f(\nu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\nu-\mu_s}{\sigma_s}\right)^2} \tag{1.2}$$

υ represents a discrete % CTL variable, whereas μ_s and σ_s are the corresponding mean and standard deviation, respectively.

A HTS assay is characterized by the measurement error. Assuming a constant error, the measurement error can as well be characterized by a Gaussian function *f(ω)* (1.3).

$$f(\omega) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\omega-\mu_c}{\sigma_c}\right)^2} \tag{1.3}$$

ω represents again a discrete % CTL variable, whereas μ_c and σ_c are the mean and standard deviation, respectively. In the assay a compound scores as a primary hit if its measured % CTL value falls beyond a certain % CTL threshold Θ. The assay is set up allowing Θ to be defined as a % CTL value being several σ_s units away from the average μ_s.

A compound is judged as a confirmed hit if in a second measurement a % CTL value is determined which is again equal to or beyond Θ. Assuming the assay was defined measuring the degree of inhibition, the confirmation probability is estimated by a probability function *P(υ)*. This function depends on σ_s, σ_c and Θ (Eq. 1.4 and Eq. 1.5),

$$P(v) = \frac{1}{\sqrt{2\pi}} \int_{V}^{\infty} e^{-\frac{1}{2}\left(\frac{v-\mu_c}{\sigma_c}\right)^2} \tag{1.4}$$

where V is determined according to equation 1.5,

$$V = \frac{\Theta - \left(\frac{v-\mu_s}{\sigma_s}\right)}{\sigma_c}. \tag{1.5}$$

That means, that the farer Θ is defined from the mean μ_s of the population (i.e. σ_s units) and the farer the % CTL value of the primary hit lies from Θ (i.e. σ_c units) the higher is the probability to confirm the primary hit. For the tested library the confirmation rate (CR, i.e. the number of confirmed hits divided by the number of primary hits) is proportional to the probability of the primary hits to be confirmed multiplied by the frequency of compounds at that activity (1.6).

$$CR = \frac{\int_{-\infty}^{\Theta} f(v)P(v)dv}{\int_{-\infty}^{\Theta} f(v)dv} \tag{1.6}$$

The false-positives represent those primary hits which are not confirmed. They are mostly referred to as not confirmed hits. The false-positive rate (FPR) for the tested library is given by 1.7

$$FPR = 1 - CR. \tag{1.7}$$

False-negatives are all hits which are missed during the primary measurement. The false-negative rate (FNR) is defined as the number of missed hits divided by the overall number of hits (i.e. missed hits and confirmed hits). It can be estimated according to equation 1.8.

$$FNR = \frac{\int_{\Theta}^{\infty} f(v)P(v)dv}{\int_{\Theta}^{\infty} f(v)P(v)dv + \int_{-\infty}^{\Theta} f(v)P(v)dv}, \tag{1.8}$$

where *f(v)* and *P(v)* are defined according to equations 1.2 and 1.4 respectively. That means, that the farer Θ is defined from the mean μ_s of the population (i.e. σ_s units) and the farer the % CTL value of a non-hit lies from Θ (i.e. σ_c units), the lower the probability that the non-hit is false-negative. Schematically the outcome of a HTS assay is illustrated in Figure 1.6.

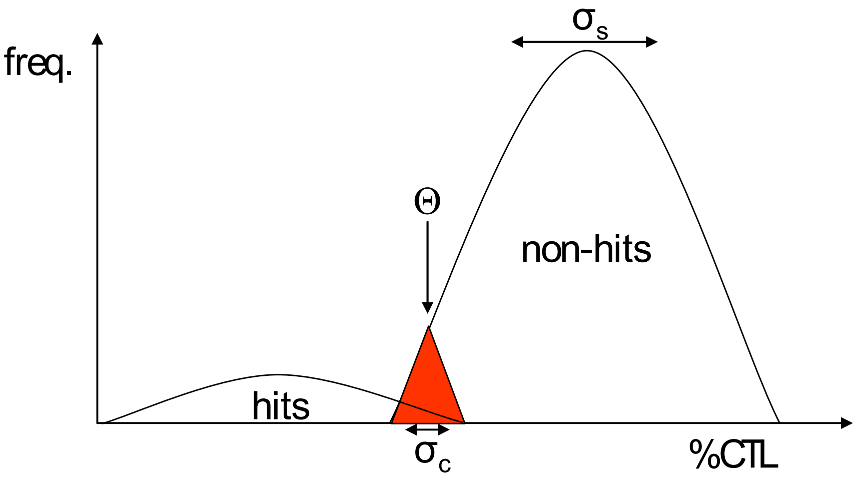


Figure 1.6 Schematic histogram-like representation of the % CTL values of a HTS screen. σ_s represents the standard deviation over all % CTL values whereas σ_c represents the standard deviation of the measurement error. Θ indicated the % CTL value defining hits from non-hits. The red-shaded area represents the % CTL region where both hits and non-hits occur.

The x-axis represents the measured % CTL values whereas the y-axis defines the frequency of compounds having the corresponding % CTL value. Two distributions are shown defining hits and non-hits. σ_s, the standard deviation of the measured % CTL values, is indicated by the large black arrow above the non-hits. σ_c, the standard deviation of the measurement error, is shown as small arrow below the red shaded region. In addition to that Θ, the % CTL threshold defining hits and non-hits is present. The red shaded area marks the % CTL region, where both hits and non-hits are present. The size of this region is characterized by σ_s, σ_c and Θ and defines the number of not confirmed hits and false-negatives. It can significantly contribute to the so-called noise in HTS data. At this point it has to be pointed out that noise is not only arising from measurement errors, but can be a consequence of systematic false-positives. These false-positives have been summarized as reactive compounds, “promiscuous inhibitors” and “frequent hitters” [Rishton, 2003, Rishton, 1997; Roche et al., 2002]. Promiscuous inhibitors are “compounds measured as inhibitory hit but turn out to act noncompetitively, show no meaningful SAR and little target selectivity” [McGovern et al., 2002]. Frequent hitters have been described as non-specific compounds (e.g. promiscuous inhibitors) or as compounds interfering with the assay method [Roche et al., 2002]. Systematic false-positives can only be identified in follow-up experiments and not by confirmation measurements under the same HTS assay conditions.

From the theoretical consideration about the outcome of a HTS assay it is evident that false-negatives and not confirmed hits occur. The not confirmed hits are cost - and work - intense in follow-up characterization. In contrast, false-negatives represent a loss of chemical knowledge. In the worst case this can translate into the loss of a lead structure or lead series. Given the few lead series resulting from HTS [Steinmeyer et al., 2006] and the high attrition rate in later stages of the drug discovery process [Bleicher et al., 2003] it is worthwhile to virtually screen the % CTL data for undiscovered hits. To achieve this, two main strategies have been followed; the identification of false-negatives by hit directed nearest neighbour searching [Shanmugasundaram et al. 2005] and the identification of false-negatives by building classification models [van Rhee et al., 2001; Harper et al., 2001; Engels et al., 2002; Glick et al., 2004; Glick et al., 2005]. The first method describes the tested compound library according to different descriptors and performs a similarity searching around the identified hits using different similarity metrics and coefficients. The second approach calculates a classification model predicting hits and non-hits. For that, the tested library is divided into a training set and a test set. A model is created based on the training set. The model is then applied to the test set in order to identify false-negatives. The challenge is that the classification method has to be able to cope with different amounts of noise and with highly unbalanced data sets (i.e. mostly non-hits). Three example applications shall be given and the conclusions given so far. Harper et al. employed binary kernel discrimination for the creation of classification models of a HTS data set with over 100,000 data points and a hit rate of 2.2%. They utilized randomly selected training sets with 500 and 5000 data points to predict false-negatives in the remaining test sets [Harper et al., 2001]. They pointed out that one single method cannot fully describe the SAR in HTS data and consequently the application of different methods might be necessary [Harper et al., 2001]. In light of this Glick et al. compared recursive partitioning, naïve Bayesian classifiers and support vector machines for classifying four different HTS assays [Glick et al., 2005]. They employed randomly selected training sets with less than 5,000 data points and an average hit rate of 5%. The test sets contained more than 170,000 entries. In the retrospective examination all methods were capable of identifying false-negatives. However the support vector machines outperformed the other methods [Glick et al., 2005]. Van Rhee et al analyzed a HTS assay with recursive partitioning in a similar way. They pointed out that the application of methods which can deal with non-linear correlations is important for HTS since different lead classes might be present which can represent different (uncorrelated) binding modes for a target [van Rhee et al., 2001]. From the applications published till now it is clear that virtual screening techniques have the capacity to identify false-negatives in HTS data [van Rhee et al., 2001; Harper et al., 2001; Engels et al., 2002; Glick et al., 2005]. However current applications employ randomly selected or maximum diverse training sets with less than 10,000 compounds whereas test sets exceed 100,000 entries. This anticipates that only rough global models are created missing a large proportion of the false-negatives in prospective applications. The training of such models on large sets with more than 10,000 data points has not been shown. Furthermore, to the best knowledge of the author, only confirmed hits have been used for the analysis. The usage of primary screening data for the prediction of false-negatives has yet to be addressed.

Clustering and Partitioning of HTS Data

Approaches to decipher lead series in HTS data are partitioning and clustering. Whereas clustering techniques group compounds according to distances in the descriptor space, partitioning techniques assign descriptor space coordinates to form compound groups. Partitioning techniques can be subdivided into supervised and unsupervised algorithms. Examples of the latter approach are cell-based partitioning algorithms [Agrafiotis & Rassokhin, 2002; Jamois et al., 2000; Pearlman & Smith., 1999; Xue & Bajorath, 2000]. The methods divide the descriptor space into hyper-rectangular regions and determine the occupancy of the obtained cells. A difficulty of applying the methods is to find an appropriate grid resolution. However cell-based approaches are available which help choosing a grid resolution by measuring diversity and space coverage of the cells, e.g. by entropy-, Chi2-, or fractal approaches [Agrafiotis & Rassokhin, 2002; Jamois et al., 2000]. A supervised partitioning technique is recursive partitioning [Young & Hawkins, 1995; Hawkins et al., 1997; Rusinko et al., 1999]. It builds up a binary decision dendrogram, by recursively selecting at each partitioning step the descriptor best separating hits from non-hits. The method has become popular for HTS data analysis, as it leads to easily interpretable results; it is fast and thus applicable to large data sets [Rusinko et al., 1995]. In this context several extensions of recursive partitioning have been introduced creating multiple binary decision dendrograms and providing an ensemble prediction [Svetnik et al., 2005; Breiman, 1996].

Data clustering methods are applied to the resulting hits to decipher lead series present in the data. Numerous clustering methods exist, and are usually separated into hierarchical methods like Ward's clustering [Ward, 1963, Brown & Martin, 1996] and non-hierarchical methods like Jarvis-Patrick [Jarvis & Patrick, 1973; Willett et al., 1986; Doman et al., 1996, Menard et al., 1998], *k*-means [Duda et al., 2001; Holliday et al., 2004], self-organizing maps [Kohonen, 1982; Schneider & Wrede, 1998; Teckentrup et al., 2004], or Bayesian unsupervised clustering [Jain et al., 1999]. In this context the superiority of non-hierarchical Jarvis-Patrick-clustering over other non-hierarchical clustering methods [Willett et al., 1986] and the superiority of hierarchical Ward's clustering over the Jarvis-Patrick method [Brown & Martin, 1996] have been claimed. Despite of these descriptor-based or fingerprint based clustering algorithms constantly new methods are being introduced, e.g. to build up a phylogeny-like tree employing maximum common substructures [Nicolaou et al., 2002], or to cluster a data set according to the frequency of substructure elements [Richon, 2000; Roberts et al., 2000] or to cluster a data set according to maximum overlapping substructures [Stahl et al., 2005].

To prioritize lead series resulting from HTS it is worthwhile to analyze the lead series in context of the occurrence of hits and non-hits [Schreyer et al., 2004]. This is illustrated in Figure 1.7 where a data set is schematically clustered in a 2D descriptor space. Putative HTS hits are shown in red whereas non-hits are shown in blue. Different conclusions can be extracted. The cluster with five hits and one non-hit might provide a promising lead series, whereas the cluster containing two hits and five non-hits shows a "steep" and less promising SAR not tolerating chemical exploration of the compound series. The presence of singleton can be judged and it can be distinguished between hits being hit singletons (i.e. hits in clusters with non-hits) and true hit singletons (i.e. hits without neighbours) [Stahl & Mauser, 2005]. For the second group of singletons no conclusion can be drawn. However chemical exploration of the hits might reveal a back-up lead. Finally, simple conclusions can be extracted from the clustering whether a hit has a high likelihood to be false-positive (i.e. a hit surrounded by non-hits) or a non-hit might be false-negative (i.e. a non-hit surrounded by hits).

When working with large data sets with more than several thousand entries a hierarchical clustering is desirable since it creates an interpretable relationship between the clusters. It allows navigating in the data and provides both a coarse-grained and fine-grained view on the data. The limitation of most current hierarchical clustering algorithms is that they have at least squared running time and memory requirement, which renders them unfeasible for data sets exceeding 20,000 entries and thus for HTS data. Only recently first algorithms have been introduced allowing such large scale data analyses [Barnard et al., 2004; Sultan et al., 2002].

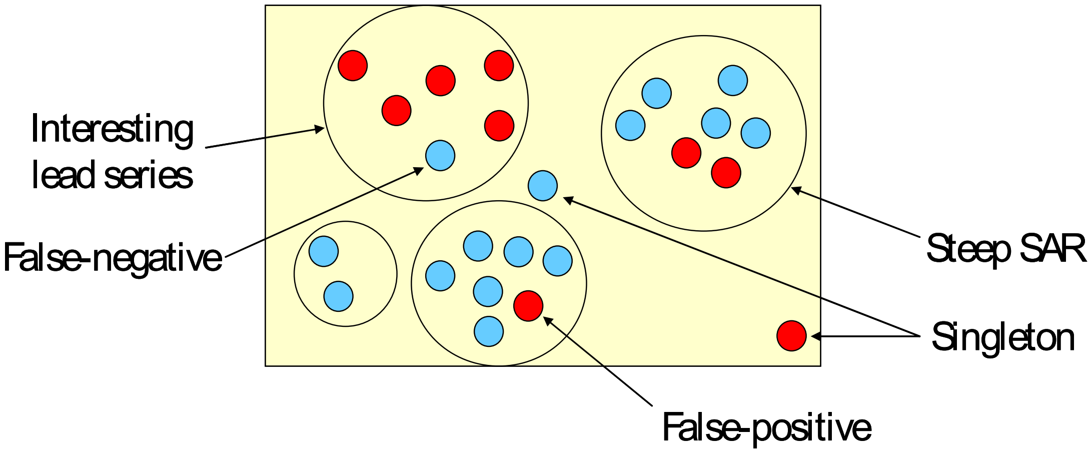


Figure 1.7 Clustering of a data set comprising hits (red) and non-hits (blue) in two dimensions. Clusters are indicated by black circles. Different types of information can be extracted. (i) A cluster containing many hits might provide an interesting lead series. (ii) A cluster with only a few hits and many non-hits might indicate a steep SAR which does not permit chemical exploration of the compound series. (iii) Data points can be judged as singletons. (iv) A hit in a cluster with many non-hits might be false-positive. (v) A non-hit surrounded by many hits might be false-negative.

In the present work a clustering based approach was developed to analyze primary screening data of HTS assays where more than 500,000 molecules have been tested. At first the complete data set was clustered and later the % CTL values were assigned to the compounds in the clusters. The composition of the clusters with hits and non-hits was employed to extract rules identifying false-negatives, not confirmed hits, singletons and clusters enriched with hits. The approach was retrospectively evaluated according to identifying false-negatives and not confirmed hits in the primary screening data of three different HTS assays. One was involved in inhibiting the transforming growth factor-β type I receptor. In a second step the clustering based approach was applied to a prospective virtual screen to identify novel dopamine D_3 receptor ligands. The obtained results were compared to other virtual screening techniques namely pharmacophore based screening, docking and regression based activity prediction. In 1.3 and 1.4 both, transforming growth factor-β type I receptor inhibitors and dopamine D_3 receptor ligands are introduced.

1.3 Transforming Growth Factor-β Type I Receptor Inhibitors

1.3.1 Biological Relevance

The transforming growth factor (TGF)-β receptor family comprises transmembrane receptors with cytoplasmic serine and threonine kinase activity [Kraus, 2001]. TGF-βs have a key-impact on cell proliferation, differentiation and migration of epithelial, endothelial and haematopoetic cell lineages thereby controlling the establishment of the body plan and tissue differentiation [Narasimha & Leptin, 2000]. Despite the tightly regulated role during development, the mediation of immune-responses, wound healing and hematopoesis [Ge et al., 2004], TGF-βs are known tumour suppressors, tumour promoters [Subramanian et al., 2004; Yingling et al., 2004; Arteaga, 2006] and responsible for various types of fibroses upon over-stimulation of the immune system [Kalluri & Neilson, 2003; Flanders, 2004; Blobe et al., 2000]. Figure 1.8 shows the assumed main TGF-β based signal transduction pathway.

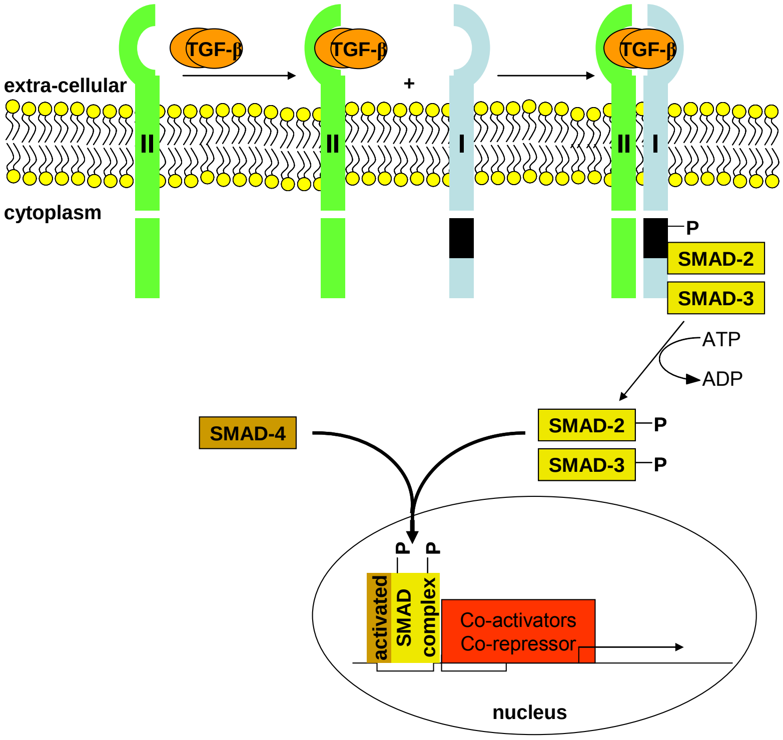


Figure 1.8 Schematic pathway of TGF-β signal transduction. TGF-β binds as dimer to the TGF-β type II receptor. The type II receptor dimerizes with TGF-β type I receptor, phosphorylates the Type I receptor and converts it into the active form. The type I receptor phosphorylates either SMAD-2 or SMAD-3 proteins, which form as a dimer together with SMAD-4 the activated SMAD complex. The SMAD complex binds to specific promoter elements and induces or represses transcription in combination with additional co-activators or co-repressors.

One of the members of the TGF-β family (cytokines comprising TGF-βs, bone morphogenetic proteins and activins) binds as dimer to the TGF-β type II receptor on the extra-cellular side. Subsequently the TGF-β type I receptor *ALK5* (activin-like kinase) is recruited into the complex forming a heterotetrameric complex of two type I and two type II receptors. *ALK5* is phosphorylated by the type II receptor at a glycine-serine region on the cytoplasmic side and converted into the active state. The C-terminal phosphorylation translates into signal transduction from *ALK5* to the receptor associated SMAD (R-SMAD) proteins 2 or 3 [Huse et al., 2001]. The phosphorylated and thus activated R-SMADs oligomerize with SMAD4 whereby two R-SMADs interact with one SMAD4 protein. The activated SMAD complex is translocated into the nucleus where it sequence-specifically binds to promotor elements, interacts with additional co-activators or co-repressors and induces or represses context-dependent gene transcription. TGF-β was also shown to activate other SMAD independent signalling cascades including the p38-MAPK pathways. The simple TGF-β signalling cascade is transformed into a complex cell-type, time and context dependent differential gene expression profile based on the levels of expression of the TGF-β receptor complex, the SMAD proteins, cooperating transcription factors and the activation state of competing signalling cascades [Krauss, 2001; Derynck & Zhang, 2003; Yingling et al., 2004, Massague & Wotton, 2000].

1.3.2 Disease Implication of the TGF-β Signalling Cascade

The TGF-β signalling cascade is implicated in two types of diseases, cancer [Arteaga, 2006] and fibrosis [Flanders, 2004]. Fibrosis results either from an over-stimulated immune response in the wound healing processes of injured tissue or from chronic inflammation. As a consequence, tissue is loosing elasticity, which ultimately can lead to loss of function of corresponding inner organs like lung, liver or kidney. Currently there is no effective treatment of fibrosis available [Sauer et al., 2005]. The TGF-β signalling cascade employing SMAD3 as a cellular transcription factor is one key pathway controlling the inflammation or the wound healing processes. It is implicated in recruiting inflammatory cells and fibroblasts into the injured tissue and in stimulating the recruited cells to produce and accumulate extracellular matrix proteins (e.g. different types of collagens, laminine or nectine). Further the proliferation of fibroblasts and their transdifferentiation into myo-fibroblasts as well as epithelia-to-mesenchymal transition (EMT) of epithelial cells into fibroblasts is stimulated by TGF-β. This allows the accumulation of fibrotic tissue [Flanders, 2004].

TGF-β has a dual role in cancer biology. In animal models it was shown to act both as a tumour suppressor and as a tumour promoter [e.g. Siegel et al., 2003; Kang et al., 2005]. Since TGF-β is a strong inhibitor of cell proliferation of epithelial, endothelial and hematopoetic cell lineages it explains its role as early stages tumour suppressor [Yingling et al., 2004]. However in later stages of tumourgenesis TGF-β was shown to be implicated into the promotion of EMT. This allows cells to adopt mesenchymal characteristics, become motile, leave the epithelium and form metastasis [Kalluri & Neilson, 2003; Kang & Massage, 2004]. Further TGF-β acts autocrine and paracrine on the progression of a tumour after EMT by inducing angiogenesis, facilitating tumour cell invasion and/or metastasis, and inhibiting anti-tumour immunity [Ge et al., 2004]. The dual role of TGF-β in cancer requires establishing a therapeutic index controlling the point where the beneficial effect of antagonizing TGF-β signalling and thereby altering tumour progression overwhelms its tumour suppressive role [Arteaga, 2006; Yingling et al., 2004].

For the treatment of cancer or fibrosis by inhibiting the TGF-β cascade two strategies exists. One is the creation of NBEs like monoclonal anti-bodies or anti-sense RNA targeting one of the TGF-β cytokines. At the moment several NBEs are in clinical phase I-III, seem to be well tolerated and show promising results [Yingling et al., 2004]. Given the great success of the c-Ableson kinase inhibitor Gleevec in the treatment of chronic myelogenous leukaemia [Capdeville et al., 2002], a second strategy is the creation of small organic molecule inhibitors of TGF-β type I receptor (*ALK-5*) kinase. Since all kinase enzymes are targeted via their adenosine tri-phosphate (ATP) binding site, for the development of novel kinase inhibitors specificity has to be addressed as early as possible.

Researchers at GlaxoSmithKline, Lilly and Biogen have investigated in identifying selective inhibitors for *ALK-5* targeting the ATP binding site [Singh et al., 2004; Yingling et al., 2004]. In a HTS campaign at GlaxoSmithKline imidazole derivatives were identified as potent inhibitors. Since these inhibitors had been originally designed as p38 kinase inhibitors the imidazole compounds were further optimized towards the triarylimidazole derivative **1** (IC_{50} = 94 nM) showing no measurable p38 binding [Callahan et al., 2001].

Cellular activity of the closely related compound **2** (IC_{50} = 47 nM) was obtained and a measurable binding to other kinase enzymes was only detected for the *ALK* kinase enzymes *ALK-4* and *ALK-7* [Byfield et al., 2004]. In presumably the same HTS campaign two additional structural classes resulted and were optimized towards the pyrazole derivative **3** (IC_{50} = 25 nM) and the thiazole derivative **4** (IC_{50} = 23 nM). Both compounds show a comparably high cellular activity and especially **3** was proven to have no measurable affinity towards a panel of different kinase enzymes [Gellibert et al., 2004]. A comparable pyrazole derivative was identified in a virtual screening campaign at Biogen leading to the ATP competitive inhibitor **5** (IC_{50} of 27 nM) [Singh et al., 2003]. Furthermore the pyrazole containing compounds **6** [Sawyer et al., 2003] and **7** [Li et al., 2006] were identified and further optimized at Lilly by virtual screening and HTS, respectively. Both compounds show cellular activity in the low nanomolar range and selectivity towards p38. For compound **7**, selectivity was further achieved towards Mixed Lineage Kinase 7 [Li et al., 2006]. The crystal structures of **3**, **5** and **6** in complex with *ALK-5* have been solved and a comparable binding behaviour was obtained [Sing et al., 2003; Gellibert et al., 2004]. Figure 1.9 exemplifies the co-crystallization of **5** with *ALK-5* (PDB code: 1PY5). The distal nitrogen in the quinoline moiety acts as hydrogen bond acceptor for the backbone nitrogen of histidine 283 in the ATP binding site. Further, one of the nitrogens of the pyrazole structure lies in close proximity to lysine 232 in the active site. The second nitrogen forms a hydrogen bond to the carboxyl group in aspartic acid 351. Finally the 2-pyridyl nitrogen forms a water-mediated network of hydrogen bonds to glutamic acid 245, aspartic acid 351 and tyrosine 249 in form of a tetrahedral complex. Despite of the presented compounds inhibiting *ALK-5*, additional patens have been filed [Yingling et al., 2004]. However until now only limited literature data is available for these compounds [Uhl et al., 2004].

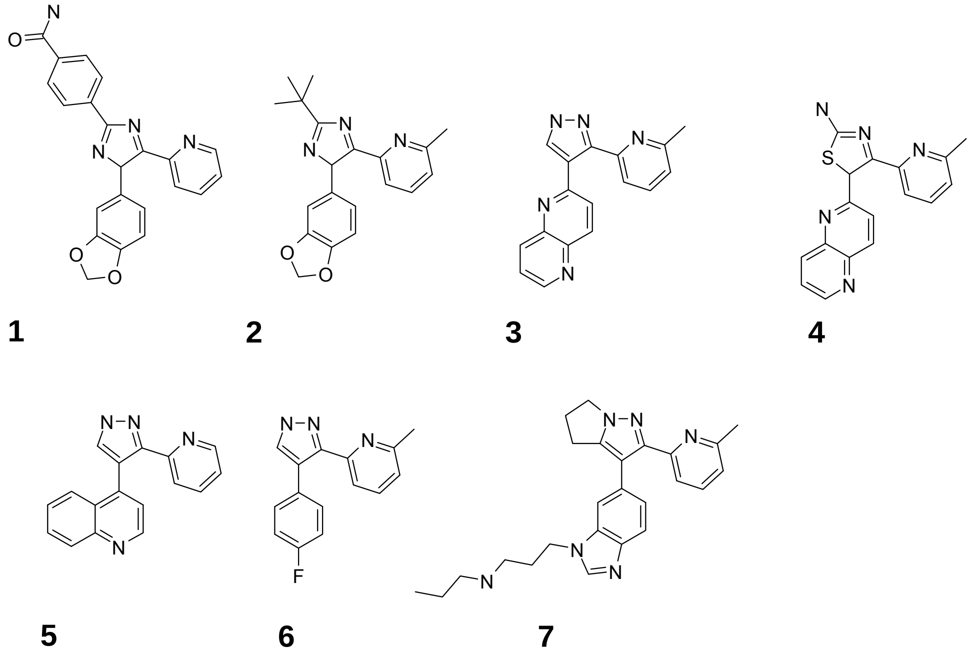


Chart 1.1 *ALK-5* inhibitors targeting the ATP binding site.

ALK-5 is implicated in two severe and deathly diseases, cancer and fibrosis. Until present no drugs for fibroses are on the market. Consequently inhibiting *ALK-5* is a worthwhile strategy for targeting both fibrosis and cancer. All current inhibitors show a strikingly similar scaffold and it is not clear whether they will make it on the market. Hence the identification of novel inhibitors by VS or HTS is important. Since kinase inhibitors are targeting the common ATP binding site, it is a problem to achieve specificity. This makes the identification of new backup series e.g. by mining false-negatives in HTS even more important.

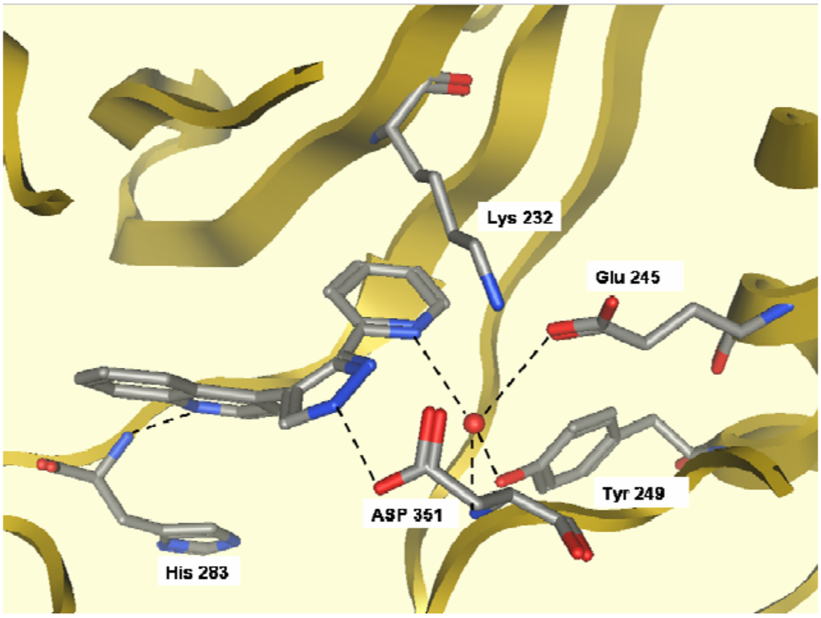


Figure 1.9 Complex of **5** and *ALK-5* (pdb: 1PY5). Hydrogen bonds are present between the quinoline nitrogen and the backbone NH of histidine 283 (His 283), one of the pyrazole nitrogens and aspartic acid 351 (Asp 351) and a water mediated network between 2-pyridyl and Asp 351, tyrosine 249 (Tyr 249) and glutamic acid (Glu 245). Further close proximity of the ligand to lysine 232 (Lys 232) was observed.

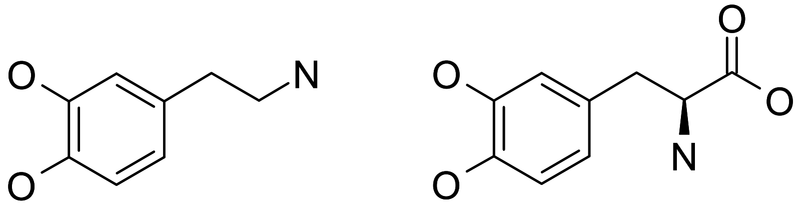
1.4 Dopamine D_3 Receptor Ligands

G-protein-coupled receptors (GPCRs) represent the largest family of membrane-embedded signalling receptors and play a role in a variety of physiological and pathophysiological processes [Hill, 2006; Klabunde & Evers, 2005]. As a common motive, they share seven trans-membrane (TM) helices as well as the intra-cellular signal transduction to one of the heterotrimeric G-proteins. GPCRs can mediate both environmental stimuli such as order, light and taste and internal stimuli by recognizing a diverse set of ligands comprising ions, biogenic amines, nucleosides, peptides, proteins and even light [Becker et al., 2003]. Given the presence of different isoforms of the three G-protein subunits, Gα, β and γ, the occurrence of different co-activators and co-repressors in a cell and the co-stimulation of additional signalling pathways a cell-type-, time- and context-specific biological response may emerge [Ellis, 2004, Malbon, 2005]. Approximately 50% of all launched drugs target only ~30 members of the GPCR family with annual worldwide sales exceeding US $30 billion in 2001 [Klabunde & Hessler, 2002, Wise et al., 2002]. The presence of 210 additional receptors for which the natural ligand is known and 160 ‘orphan receptors’ identified by the human genome project [Venter et al., 2001] in combination with an increasing evidence showing their implication in a wide variety of diseases renders this protein family to one of the most important pharmaceutical targets [Becker et al., 2003].

Based on the similarity of the amino acid sequences, three main subfamilies of GPCRs (rhodopsin-like (A), glucagon-receptor-like (B) and the metabotropic glutamate receptors (C)) are known with family A being the largest, functionally and structurally best characterized. One subfamily of the rhodopsin-like family forms the biogenic amine binding GPCRs. Within this subfamily the dopamine receptors are potent targets for the treatment of schizophrenia, Parkinson’s disease and drug abuse [Joyce, 2001]. Due to the distinct location of dopamine D_3 receptors in limbic brain areas, the dopamine D_3 receptor is assumed to play a pivotal role in these neurological and psychiatric disorders. Recognition of high affine and selective ligands employing different types of lead identification strategies for this dopamine receptor subtype could improve the therapeutical treatment with less adverse side effects [Schwartz et al., 2000].

1.4.1 The Dopaminergic Pathways

Dopamine (DA) (**8**, 2-(3,4-dihydroxyphenyl)ethanamine, Chart 1.2) [Carlsson et al., 1958] is the predominant catecholamine neurotransmitter in the mammalian central nervous system (CNS). Dopamine is biosynthesized via its precursor levodopa (**9**, L-3,4-dihydroxyphenylalanine, Chart 1.2) in the CNS by three major groups of neurons present in the midbrain and in the hypothalamus [Elsworth & Roth, 1997].



8 (dopamine, 2-(3,4-dihydroxyphenyl)ethanamine)

9 (levodopa, L-3,4-dihydroxyphenylalanine

Chart 1.2 Dopamine and its precursor levodopa.

The first group of neurons in the midbrain originates from the retrorubral area (A8) whereas the second and third groups originate from the substantia nigra pars compacta (A9) and the ventral tegmental area (A10), respectively. Neurons located in the hypothalamus are nominated as the A12 group in the nucleus infundibularis. According to the projection of the dopaminergic axons four different dopaminergic pathways with distinct functionality are distinguished. The nigrostriatal pathway comprises neurons originating from the substantia nigra pars compacta and retrorubral area projecting to the dorsal striatum. It is implicated in the control of movement and in Parkinson’s disease [Smith& Kieval, 2000]. Motivated behaviour results from the mesolimbic pathway, where neurons originate from the ventral tegmental area and project to the limbic areas of the nucleus accumbens, the corpus amygdaloideum, and the hippocampus [Diaz et al., 2000]. In the mesocortical pathway the neurons from the ventral tegmental area project to cortical areas of the medial, prefrontal, cingulate and entorhinal cortex responsible for aspects of learning and memory. Both latter pathways are linked to reward and schizophrenia [Diaz et al., 2000, Sokoloff et al., 2006]. Finally, the tuberoinfundibular pathway originating from the hypothalamus projects to the eminentia mediana and the intermediate lobe of the pituitary. It is implicated in inhibiting the prolactin synthesis [Smith & Kieval, 2000]. An overview of the pathways is shown in Figure 1.10.

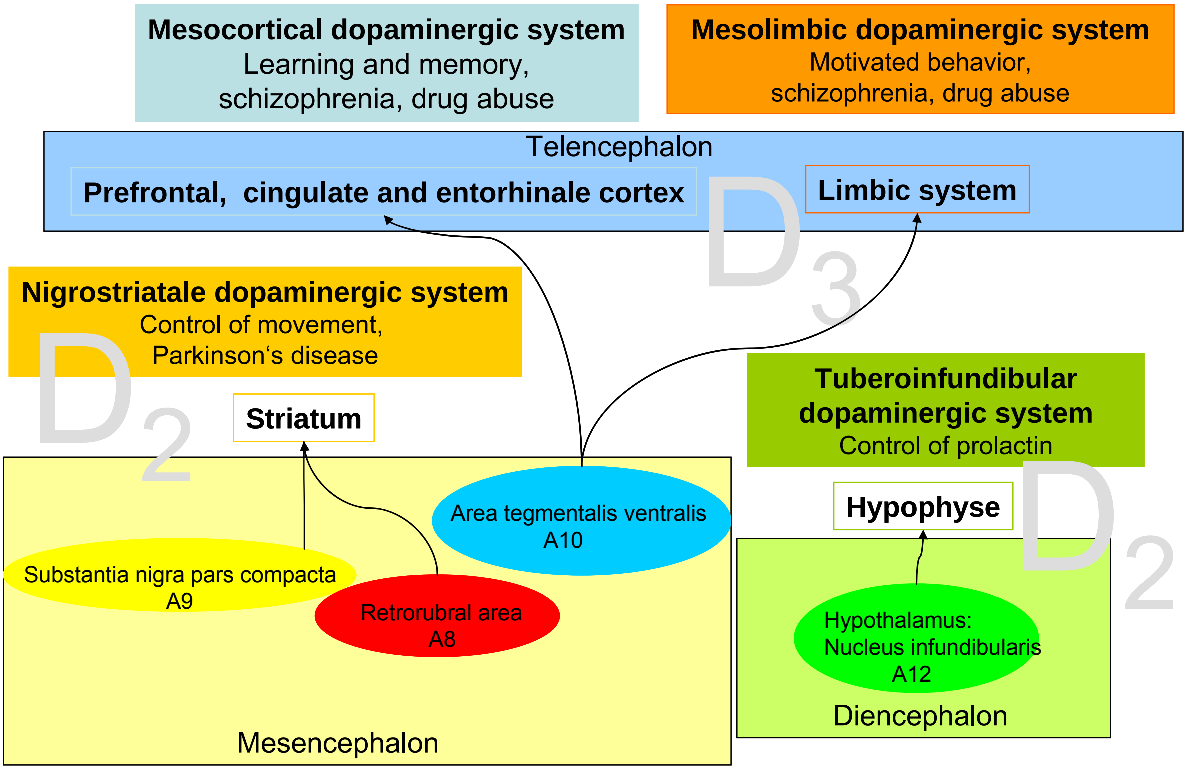


Figure 1.10 Dopaminergic pathways in the mammalian brain.

Dopamine mediates a variety of functions by signalling through the dopamine receptor family of GPCRs. This includes locomotor activity, cognition, emotion, motivated behaviour, positive reinforcement, food intake and endocrine regulation [Missale et al., 1998; Nieoullon & Coquerel, 2003]. Dysregulation of dopaminergic transmission is implicated in neurological and psychiatric disorders such as Parkinson's disease, schizophrenia, drug addiction, Huntington's disease, attention deficit hyperactivity disorder and Tourette syndrome [LeFoll et al., 2005; Joyce, 2001; Emilien et al., 1999]. The dopamine receptors are classified into two different subfamilies, the D_1-like receptors comprising dopamine D_1 and D_5 receptors and the D_2-like receptors consisting of dopamine D_2, D_3 and D_4 receptors [Missale et al., 1998]. The receptors within one subfamily are characterized by a high sequence similarity of 82% between dopamine D_1 and D_5 receptors, 76% between dopamine D_2 and D_3 receptors and 54% between dopamine D_2 and D_4 receptors in the TM domains [Marsden, 2006]. This renders the development of selective ligands for one of the receptors a challenge.

The dopamine D_1-like receptors couple functionally to $G\alpha_s/G\alpha_{olf}$ proteins, activate adenylyl cyclase and increase the production of the second messenger cyclic adenosine-3′,5′-monophosphate (cAMP) whereas the dopamine D_2-like receptors couple to $G\alpha_i/_o$, inhibit adenylyl cyclase and down-regulate the cAMP concentration [Neve et al., 2004]. It has been recognized that dopamine D_2 receptors occur with high density in the caudate putamen and the substantia nigra, responsible for motor activity whereas the dopamine D_3 receptors are predominately present in the ventral striatum (Limbic system) responsible for cognition and motivation [Levesque et al., 1992; Gurevich & Joyce, 1999]. Consequently dopamine D_3 receptors play a pivotal role in pathological processes including schizophrenia, drug abuse and Parkinson's disease whereas dopamine D_2 receptors, besides their therapeutic benefit in Parkinson's disease, are connected to the occurrence of adverse side effects [Schwartz et al., 2000].

In the periphery dopamine was shown to modulate cardiovascular and renal functions, hormone release, vascular tone and gastrointestinal motility mediating its effect through dopamine D_1, D_2, D_3, D_4 and D_5 receptors.

1.4.2 Therapeutic Relevance of Selective Dopamine D_3 Receptors

As mentioned above, an imbalance of the dopaminergic system is implicated in several neurological and psychiatric disorders. Mainly Parkinson's disease, schizophrenia and reinforcing effects of drug abuse are of current interest for developing selective dopamine D_3 receptor ligands. The mental disorder schizophrenia is characterized by positive symptoms including hallucinations, delusions, and bizarre behaviour, and negative symptoms such as diminished affect, loss of motivation and the inability to experience pleasure. The “hyperdopaminergic hypothesis” of schizophrenia assumes that the positive symptoms result from an overrepresentation of dopamine in the limbic system [Willner, 1997]. Non-selective antagonists of dopamine D_2 receptors (i.e. first generation antipsychotic drugs), such as haloperidol, are capable of releasing the positive symptoms. However the antagonism in the dorsal striatum results in extrapyramidal side-effect. Since the dopamine D_3 receptor is highly concentrated in the limbic area of the striatum, which plays a key role in schizophrenia, selective dopamine D_3 receptor antagonists are assumed to possess antipsychotic effects and might prevent extrapyramidal side-effects [Leriche et al., 2004].

Parkinson’s disease is characterized by a progressive loss of dopaminergic neuron in the *substantia nigra pars compacta*. It translates into movement disorders like rigidity, tremor, akinesia or bradykinesia. Further progression of the disease involves the mesolimbic dopaminergic system and results in learning and memory deficiencies. The key drugs in the treatment of Parkinson’s disease are either the dopamine precursor levodopa **9** (chart 1.2) or dopamine D_3-receptor preferring agonists such as pramipexole (Table 1.1) [Mierau et al., 1995; Kushida, 2006]. Despite the immediate benefit of levodopa it was recognized that the long-term administration results in dyskinesia [Bezard et al., 2001]. However the dyskinesia is less strong for the dopamine agonists compared to levodopa [Jenner, 2003]. Further it was shown that the co-administration of dopamine D_3 receptor partial agonists and levodopa relieves the symptoms of dyskinesia while maintaining the clinical benefit [Bezard et al., 2003]. Consequently the development of both selective dopamine D_3 receptor agonists and partial agonists is of great interest.

Abused drugs like cocaine or heroin produce reward and reinforcement effects and may lead to addictive effects. Especially for cocaine no pharmacological treatment is available yet. It was shown that dopamine levels are elevated in the nucleus accumbens upon consumption of a drug [Newman et al., 2005]. This region is implicated in reward and reinforcement effects of a drug. Further dopamine D_3 receptor mRNA levels were shown to be higher in human post-mortem studies of brains obtained from cocaine addicts compared to non-addicts of the same age [Segal et al., 1997]. Taken together, results indicate that dopamine D_3 receptor antagonists or partial agonists might offer a therapeutic option for the treatment of drug abuse. Current animal models confirm this hypothesis. However proof of concept in humans is still missing [Newman et al., 2005].

1.4.3 Dopamine D_3 Receptor Agonists, Partial Agonists and Antagonists

Dopamine D_3 receptors are assumed to possess a high therapeutic potential for the treatment of neuropsychiatric disorders [Sokoloff, 2006]. A difficulty in the development of these ligands is to achieve selectivity against its homologue, the dopamine D_2 receptor, and against other (clinically relevant) receptors of the aminergic GPCR family [Klabunde & Evers, 2005]. Since no 3D structure of the dopamine receptor or homologue receptors is known a structure-based drug design is not possible. Ligands have to be optimized to penetrate the blood-brain barrier and exert their action in the specific brain areas like the nucleus accumbens.

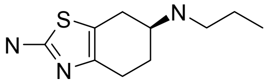
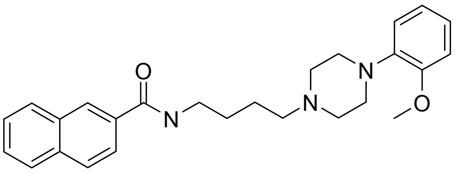
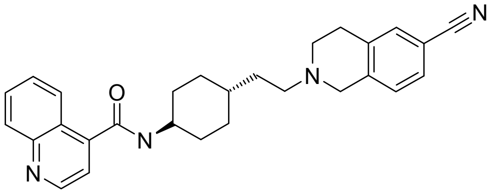
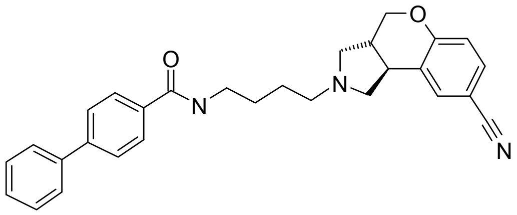
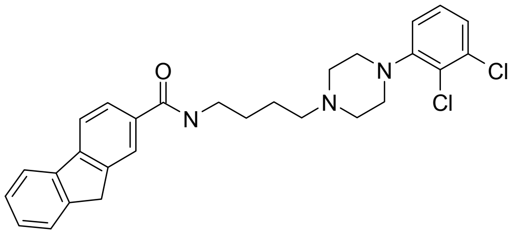
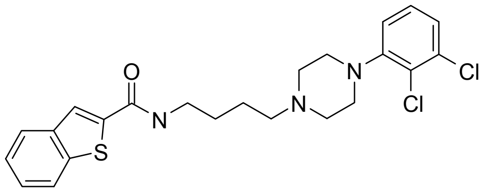
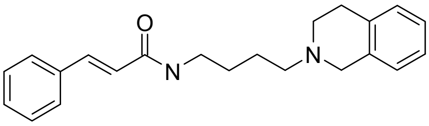
Despite of these difficulties selective agonists, partial agonists and antagonists have been identified. Some of the partial agonists and antagonists have entered the clinical phases whereas a few agonists are already on the market. A representative subset of compounds with corresponding affinities at the dopamine receptors D_1-D_5 is shown in Table 1.1.

The benzthiazole derivative Pramipexole [Mierau et al., 1995] is a dopamine receptor agonist and is successfully applied for the treatment of symptoms of Parkinson's disease and restless legs syndrome. It shows highest affinity for dopamine D_3 receptors and only low affinity at adrenoceptors and 5-HT receptors [Mierau et al., 1995; Kushida, 2006].

The selective partial agonists and antagonists, BP 897 [Pilla et al., 1999], S33084 [DuBuffet et al., 1999], SB 277011 [Stemp et al., 2000], NGB 2904 [Yuan et al., 1998], ST 198 [Bezard et. 2003], FAUC 365 [Bettinetti et al., 2002] (Table 1.1), possess high affinities for the human dopamine D_3 receptors in the low nanomolar or even subnanomolar range. The author stresses to mention that for the same ligand, depending on the assay type, the assay conditions and origin of cloning of dopamine receptors from various species, different selectivity ratios against dopamine D_2 receptors have been reported. They ranged from low 16-fold selectivity to 7200-fold [Newman et al., 2005]. According to this the measured K_i values and selectivity ratios have to be treated with care. Nevertheless clinical data and results obtained from animal models indicate that these compounds fulfil their promises for the treatment of the above mentioned diseases [Joyce and Millan, 2005].

The identified potent and selective dopamine D_3 receptor antagonists and partial agonists contain strikingly similar scaffolds and only a few crucial modifications have been identified being tolerated at dopamine D_3 receptors but not at dopamine D_2 receptors. The ligands can be divided into four different sections, an aryl residue, an amide moiety, a spacer region and a positively charged amine residue (Figure 1.11).

Table 1.1 Representative subset of dopamine receptor agonists, partial agonists and antagonists together with their corresponding affinities at dopamine receptors D_1-D_5.

Compound		K_i **[nM]**				
		D_1	**D_2**	**D_3**	**D_4**	**D_5**
Agonist						
Pramipexole (10)[1]		-	3.3	0.5	3.9, 6.3	
Partial Agonist						
BP 897 (11)[2]		3000	61	0.9	300	-
Antagonist						
SB 277011 (12)[2]		>1000	1030	11	>1000	>1000
S 33084 (13)[2]		500	32	0.3	2000	1300
NGB 2904 (14)[2]		>10000	217	1.4	>5000	>10000
FAUC 365 (15)[3]		8800	2600	0.50	340	-
ST 198 (16)[2]		25000	780	12	5000	-

[1] Mierau, J. et al., [Mierau et al., 1995], [2] Sokoloff et al. [Sokoloff et al., 2006], [3] Bettinetti et al. [Bettinetti et al., 2002].

Binding data suggest that a distance of 6 to 7 Å between the amide oxygen and the positively charged nitrogen is responsible for D_3 selectivity over D_2 [Hackling et al., 2003]. The spacer region has to be extended and linear however aromatic and hydrophobic substitutions are tolerated in the spacer regions [Hackling et al., 2003]. The SAR of the amine rest is “steep” since only azacyclic aryl ring systems with specific substitution pattern (e.g. 2,3-dichloro-phenyl, 2-methoxy-phenyl or p-cyano-phenyl) are tolerated and provide selectivity. Favourable groups for the aryl residue are extended by bi- and tricyclic aryl rings or conjugated olefinic phenyl rings. Substitutions of the aryl ring system with heteroatoms are, to a limited extend, possible [Newman et al., 2005].

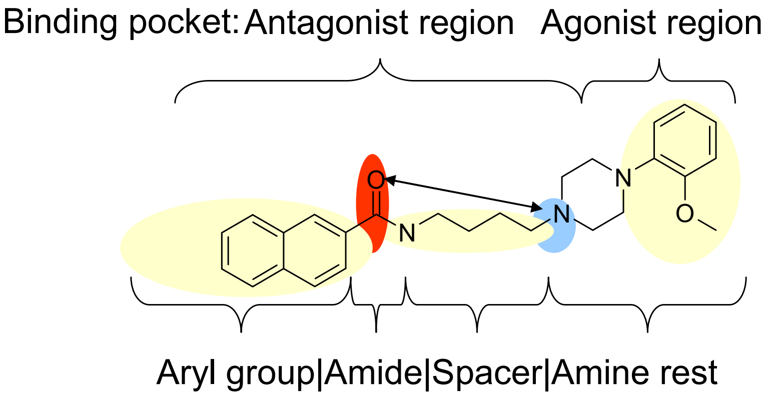

Binding pocket: Antagonist region
Agonist region
O
N
N
N
O
Aryl group|Amide|Spacer|Amine rest

Figure 1.11 SAR of dopamine D_3 receptor antagonists. As example structure BP 897 is shown separated according to two schemes: (i) a section binding to the antagonist part of the binding pocket and a section binding to the agonist part. (ii) Four different sections consisting of an aryl group, an amide moiety, a spacer region and an amine rest. Favourable pharmacophore points (aromatic or hydrophobic = yellow, acceptor = red, cationic or donor = blue) were projected onto BP 897.

The lack of structural diversity within the ligands and the need to obtain more selective dopamine D_3 receptor ligands with favourable pharmacokinetic properties renders this receptor an ongoing challenge and worthwhile example for the application of various kinds of computational methods that help explain the SAR within the molecules and identify novel ligands. One strategy was to model the 3D structure of the dopamine D_2 like receptors based on the crystal structure of rhodopsin. The models were later employed for virtual screening and for explaining the interaction of the ligands with the receptor. It was recognized that different models were necessary to explain the SAR of agonist and antagonists [Bissantz et al., 2003; Klabunde & Evers, 2005]. Further, a conserved binding pocket was proposed either within the D_2 like receptors [Boeckler et al., 2005] or the catecholamine receptors [Xhaard et al., 2006]. According to Klabunde and Evers [Klabunde & Evers, 2005], dopamine D_2

receptors form the agonist binding pocket with trans-membrane helices 3, 5 and 6, whereas helices 1, 2 and 7 were suggested to harbour a lipophilic binding site for antagonists. As a key interaction an aspartic acid in trans-membrane helix 3 (Asp 110) was identified in all studies [Xhaard et al., 2006; Boeckler et al., 2005; Varady et al., 2003]. The created homology models were successfully applied in context of virtual screening and significant enrichment in dopamine D_3 receptor ligands was obtained [Varady et al., 2003].

Various ligand based methods were applied in this context. Pharmacophore models were created for dopamine D_2 and D_3 receptors. A distance of 6 to 7 Å between the amide oxygen and the positively charged nitrogen was employed for D_3 selectivity over D_2 [Hackling et al., 2003]. CoMFA and CoMSIA methods were employed to help transform the 3D homology model from the antagonist to the agonist state [Boeckler et al., 2005]. In addition, 3D QSAR was applied to identify enantiomeric representations of a set of dopamine D_3 receptor agonists and explain important features within the molecules [Elsner et al., 2005]. Finally, an active learning approach using iterative application of SVMs was successfully employed to identify novel dopamine D_3 receptor antagonists [Byvatov et al., 2005].

1.5 Scope of the Thesis

The scope of the thesis was to identify SAR in the primary screening data of HTS assays. The strategy was to hierarchically cluster the compounds, assign the primary screening data to the clusters and employ the clusters in combination with their relationship to each other to derive models helping to identify false-negatives, not confirmed hits, singletons and clusters enriched with hits. The thesis was performed in a four-step process comprising (i) the development of the clustering approach, (ii) the development of a graphical user interface for working with the clustering results, (iii) the retrospective application of the clustering approach to primary screening data of HTS assays and (iv) the prospective application of the clustering approach in combination with alternative chemoinformatic methods.

i) Primary screening HTS data are large. Consequently, the aim was to develop a new and cost-efficient hierarchical clustering algorithm, namely NIPALSTREE being able to cope with large data sets. The second aim was to adopt a known cost-efficient hierarchical clustering algorithm, the hierarchical *k*-means algorithm. The goal was to evaluate both algorithms according to small data sets and compare both clustering algorithms with each other in context of retrospective virtual screening applications.

ii) Clustering large data sets translates into large result lists. This requires new ways of data handling. The aim was to develop a graphical user interface, which allows the display of and the navigation in the data. The second aim was to enrich the graphical user interface with functionalities helping analyse SAR in terminal clusters and singletons. The third aim was to incorporate results of a variety of HTS assays into the clusters and provide tools dealing with hit enrichment, selectivity and specificity.

iii) The primary screening data of three HTS assays were provided for a retrospective analysis. One of the assays was performed for finding novel inhibitors of the TGF-β receptor kinase type I. The aim was to analyze the clustering approach for identifying not confirmed hits and false-negatives in the data. To minimize false-positives, the aim was to combine the clustering-based data mining with a supervised classification.

iv) To identify novel dopamine D_3 receptor ligands the goal was to apply the clustering approach in a prospective virtual screen. The aim was to extend the approach by combining results with docking, pharmacophore-based modelling and regression-based activity prediction. The results of each of the methods should be analyzed for the capacity of helping understand the SAR of the newly identified hits.

2 Experimental Techniques

2.1 High Throughput Screening Techniques

HTS comprise highly automated techniques measuring several thousands up to a few million compounds for their binding capacity at a certain target protein. These techniques follow a common scheme: compounds (or natural product mixtures) are transferred from a dimethylsulfoxide (DMSO) stock solution into a microtiter assay plate (384 wells per plate to 1536 wells per plate). Additional reagents are added including target protein (biochemical assay) or cells (cell-based assay). Following an incubation period the (usually amplified) response signal is measured and converted in a percent inhibition or fold stimulation. Only single point measurements are performed with one defined compound concentration and the biological response is projected in a narrow data range between 0% and 100%. A user-defined % CTL threshold is employed to define hits and non-hits. The hits are then confirmed in additional two-point measurements employing the same assay [Bajorath, 2004].

Establishing a HTS assays is a multi-property optimization in terms of compound concentration, enzyme/receptor/cell concentration and additional assay conditions like incubation time, temperature or pH value. The quality of a HTS assay is determined by its ability to distinguish hits from non-hits, which corresponds to a high signal to noise ratio [Oldenburg et al., 2001]. A widely accepted quality measure is the *Z’* factor (Eq. 2.1).

$$Z' = 1 - \frac{3 \cdot (\sigma_{signal} + \sigma_{background})}{| M_{signal} - M_{background} |} \ , \qquad (2.1)$$

with σ_{signal} and $\sigma_{\text{background}}$ being standard deviations for signal (positive control e.g. uninhibited enzyme reaction) and background (negative control e.g. substrate without receptor) and M_{signal} and $M_{\text{background}}$ the corresponding mean values. *Z’* is projected between 0 and 1 with higher values determining a higher quality [Zhang et al., 1999].

2.1.1 TGF-β Type I Receptor

The HTS assay was performed from the group headed by Dr. Frank Büttner in the Department of Lead Identification at Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany. The low-volume 384-well plates (white) were purchased from Greiner [Greiner Bio-one Inc., Longwood, USA]. The kinase-Glo reagent was purchased from Promega [Promega Corporation, Madison, USA], adenosine-3-phosphate (ATP) was from Sigma [Sigma-Aldrich, Taufkirchen, Germany]. The His-tagged transforming growth factor beta type I receptor (His-TGF-bR1.WT-Xa162-end) was expressed in a Baculo virus system and prepared in the laboratory of Dr. John Park Dept. of Pulmonary Research (Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany). The protein was obtained frozen and stored at -80°. All other materials were of highest grade commercially available.

In the 384-well plates, 3 µl of the test compound diluted in water (bidest., final concentration of compound 5 µg/ml; DMSO 1%) were mixed with 3 µl of the TGF-β type I receptor (diluted to achieve a final concentration of 0.17µg/ml in buffer 1) followed by an incubation of 15 minutes at room temperature. After this step, 3 µl of ATP (diluted in buffer 2, final 200 nM) were added. The plates were then incubated at room temperature for 4 h. After this step, 9 µl of the Kinase-Glo Reagent were added, followed by an incubation time of 15 minutes at room temperature. After this incubation period, the plates were counted in a LEADseeker device [Amersham Biosciences, Freiburg, Germany].

Assay buffer 1:	Assay buffer 2:
50 mM Tris	50 mM Tris
50 mM NaCl	1 mM Na_2VO_4
0.1mM EGTA	0.1mM EGTA
1 mM DTT	1 mM DTT
10% Glycerine	1mM $MgCl_2$
0.1% Triton X-100	10mM $MnCl_2$
adjusted to pH 7.5	adjusted to pH 7.5

Each assay microtitter plate contained wells without TGF-β type I receptor (high values, 100% CTL) and wells only with TGF-β type I receptor (low values, 0% CTL). The analysis of the data was performed by calculation of the percentage of ATP consumption of TGF-β type I receptor in the presence of the test compound (sample) compared to the consumption of ATP in the absence of enzyme (Eq. 2.2).

$$\%CTL = 100\% - \frac{(LSC(sample) - LSC(low\ value)) \cdot 100\%}{LSC(high\ value) - LSC(low\ value)} \,, \qquad (2.2)$$

with LSC being LEADseeker counts. An inhibitor of the TGF-β type I receptor kinase will give values between 100% CTL (no inhibition) and 0% CTL (complete inhibition). Values of more than 100% CTL are normally related to compound-specific physico-chemical properties (e.g. solubility, fluorescence etc.) or indirect biochemical effects such as allosteric regulation. The assay principle is as follows: TGF-β type I receptor kinase reaction will consume ATP. The Beetle Luciferase in the kinase-Glo Reagent needs as well ATP for its activity. An active kinase translates in a low Luciferase signal since less ATP is left for the Luciferase reaction. An inactive (inhibited) kinase translates in a high Luciferase signal since no ATP was consumed by the kinase [Singh et al., 2004].

2.2 Dopamine Receptor Binding Studies

Dopamine receptor binding studies were performed by Britta Sasse at the Johann Wolfgang Goethe-University, Frankfurt, Germany. CHO-$D_{2(short)}$ cells, expressing the recombinant human $D_{2(short)}$ dopamine receptor gene [Hayes et al., 1992], were grown in Dulbecco`s modified Eagle`s medium/F12 (1:1) mixture supplemented with 2 mM glutamine, 10% fetal bovine serum, and 100 I.U./mL penicillin G, 100 µg/mL streptomycin in an atmosphere of 5% CO_2 at 37°C [Gibco^{TM}, Karlsruhe, Germany]. Human dopamine D_3 receptors stably expressed in CHO cells as previously described by Sokoloff *et al.* [Sokoloff et al., 1992] were used. The cell line was cultured in Dulbecco`s modified Eagle's medium supplemented with 2 mM glutamine, and 10% dialyzed fetal bovine serum, and were grown in an atmosphere of 5% CO_2 at 37°C [Gibco^{TM}, Karlsruhe, Germany]. Human dopamine $D_{2(short)}$ and D_3 receptor expressing cell lines were grown to confluence. The medium was removed, and the cells were washed with 10 mL PBS buffer (140 mM NaCl, 3 mM KCl, 1.5 mM KH_2PO_4, 8 mM Na_2HPO_4, pH 7.4) at 4°C. After removing the wash buffer, the cells were scraped from the flasks into 15 mL of ice-cold medium, and centrifuged at 3,000 rpm for 10 min at 4°C. After centrifugation the medium was removed and the cell membranes resuspended in ice-cold Tris-HCl buffer containing 5 mM $MgCl_2$, pH 7.4 and disrupted with a Polytron and centrifuged at 20,000 rpm for 30 min at 4°C. The pellets were resuspended by sonification in ice-cold Tris-HCl buffer (containing 5 mM $MgCl_2$, pH 7.4); membrane aliquots were stored at -70°C. Determination of membrane protein was carried out by the method of Bradford. Cell membranes containing human $D_{2(short)}$ and D_3 receptors from CHO cells were thawed, rehomogenized with ultra sonic waves at 4°C in Tris-HCl, pH 7.4 containing 120 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$ and 1 mM $MgCl_2$ (incubation buffer), and incubated with 0.2 nM [^{3}H]spiperone (106 Ci·mmol^{-1}) [Amersham Biosciences, Freiburg, Germany], and drug diluted in incubation buffer. Nonspecific binding was determined in the presence of 10 µM

BP 897 (prepared in the same laboratory) [Pilla et al., 1999]. Incubations were run at 25°C for 120 min, and terminated by rapid filtration through PerkinElmer GF/B glass fibre filters [PerkinElmer Life Sciences, Rodgau, Germany] coated with 0.3% polyethylenimine [Sigma-Aldrich, Taufkirchen, Germany] using an Inotech cell harvester [Inotech AG, Dottikon, Switzerland]. Unbound radioligand was removed with four washes of 300 µL of ice-cold 50 mM Tris-HCl buffer, pH 7.4, containing 120 mM NaCl. The filters were soaked in 9 mL Beta plate scintillation and counted using a PerkinElmer MicroBeta®Trilux scintillation counter [PerkinElmer Life Sciences, Rodgau, Germany]. Competition binding data were analyzed by the software GraphPad Prism™ (2000, version 3.02) [GraphPad Software Inc., San Diego, CA, USA], using non-linear least squares fit. For detailed screening the compounds have been tested at seven concentrations in triplicate carrying out two to four separate binding experiments for human dopamine $D_{2(\mathrm{short})}$ and for human dopamine D_3 receptors and expressed as mean ± standard deviation (SD). K_i values were calculated from the IC_{50} values according to Cheng-Prusoff equation (Eq. 2.3) [Cheng & Prusoff, 1973].

$$K_\mathrm{i} = \frac{IC_{50}}{1 + \dfrac{L}{K_D}} \quad , \tag{2.3}$$

with L being the concentration of the competing radio ligand ($[^3H]$spiperone) and K_D being the equilibrium dissociation constant of the radioligand affinity. The IC_{50} value represents the concentration of the unlabeled compound displacing 50% of the bound radioligand from the receptor. K_i is the concentration of the ligand binding to half of the receptor at equilibrium in the absence of competitors.

3 Computational Techniques

Different computational techniques were applied to virtual screening and “mining” of HTS data. They can be subdivided into own method developments and methods already described in literature. This section contains only the latter methods whereas newly developed methods are presented in the *Results and Discussion* section. At first a literature survey of the techniques is given, mainly focusing on their implication in early stages of the drug discovery process. Then the application of the method in the present work is defined. The section is structured according to a “typical” chemoinformatic workflow setting, starting with the specification of the used data sets, followed by molecule preparation, pre-filtering, descriptor calculation and descriptor selection and ending with the application of different methods to the prepared data. The applications are further subdivided into unsupervised methods (no incorporation of *a priori* knowledge of measured data), supervised methods (incorporation of *a priori* knowledge of measured data), 3D pharmacophore modelling, and receptor based docking. This separation mirrors the different methods in terms of computation speed and target specificity, starting from rough molecule filtering over application of different alignment-free and descriptor-based supervised and unsupervised methods in 2D to more elaborate precise methods using 3D conformer data sets like pharmacophore modelling or docking (4D) [Bleicher et al., 2003].

3.1 Data Sets

3.1.1 COBRA

COBRA (Collection of Bioactive Reference Analogues) is a constantly updated small high quality data set containing 5,375 pharmacologically active molecules taken from the literature in version 3.1 [Schneider & Schneider, 2003]. Despite the 2D molecular representation, the data set additionally provided information about receptor class, name and subtype of the target and the indication field for each entry. Version 3.1 of the COBRA data set contained only drug-like molecules, whose structures were desalted and formally neutral.

3.1.2 MDDR

The MDDR database (version of August 2003) contained 141,692 biologically relevant molecules taken from patent literature, scientific journals, and meeting reports [Elsevier MDL, San Leandro, CA.]. Each entry contained a 2D molecular structure field, an activity class field and a corresponding activity class index (note that a molecule can be assigned to multiple activity classes). The MDDR was prepared using the steps described in further detail in 3.2: (i) Entries lacking structural information were removed leaving 139,037 compounds. (ii) Counter ions were removed using a statistical in-house approach implemented in Kensington Discovery Edition [InforSense Ltd., London] at Boehringer Ingelheim, and (iii) structures were neutralized using SciTegic Pipeline Pilot [SciTegic, San Diego, CA]. For 453 entries the counter ion was undistinguishable. These entries were removed. Since only small organic drug-like molecules were of interest several drug-likeness filters were applied to the MDDR [Böcker et al. 2005]. Applying all filtering steps to the MDDR the database was reduced to 109,528 entries.

3.1.3 SPECS Catalogue

The SPECS catalogue (version June 2003) is a vendor database consisting of 229,658 small organic molecules which can be purchased to build up diverse screening libraries for HTS and lead discovery programs [SPECS, Delft, The Netherlands]. The used version contains only desalted and neutralized compounds.

3.1.4 HTS Data

At Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany more than 40 HTS assays were available each having more than 350,000 data points measured. The different assays were separated according to the disease area (cardiovascular, metabolic, central nervous system, respiratory or oncology), target type (receptor or enzyme), assay type (functional assay, enzyme activity assay, binding assay, expression assay or transcription assay), assay technology (radioactive, non-radioactive), screening technology (FLIPR, Alpha screen, FRET, SPA LEADseeker, Luciferase reporter assay) and according to the mode of action (activator, inhibitor, antagonist, agonist, modulator). For the present work three different high-quality assays were selected. Selection criteria were based on the number of data points, the quality of the assay (hit rate and confirmation rate) and two assay criteria: an inhibitory assay and no cellular assays. Due to proprietary rights of Boehringer Ingelheim the first two assays are termed assay A and assay B, whereas the third assay was performed against the human TGF β type I receptor (*ALK5*). The main facts of all three assays are summarized in Table 3.1.

Table 3.1 Employed HTS assays.

	Assay A	Assay B	*ALK5*
Indication	Metabolic Syndrome	Respiratory disease	Respiratory disease
Assay technology	Alpha screen non-radioactive	SPA LEADseeker radioactive	Luciferase non-radioactive
Tested compounds*	664,878	549,525	738,861
Primary hits**	2,028	11,853	11,284
Hit rate	0.3 %	2.2%	1.5%
Confirmed hits	1,541	10,775	9,581
Confirmation rate	76%	91%	85%

* Compounds having no structure and redundant compounds were removed

** Primary hits were removed, not tested in confirmation

3.1.4.1 Assay A

Assay A has its implication in the treatment of metabolic syndrome [Ruderman & Prentki, 2004; Curtis et al., 2005; Gronemeyer et al., 2004]. As assay technology a non-radioactive inhibitory alpha screen was performed testing 804,586 different samples [von Leoprechting et al., 2004]. The mean result was 105.5% CTL with standard deviation of 11.4%. Based on an upper threshold of 50% CTL, 2,476 hits occurred representing a hit rate of 0.3%. 2,475 compounds entered confirmation measurements where 1,921 primary hits were confirmed (confirmation rate = 78%). Eliminating all compounds from consideration having no structure specified 2,028 primary hit remained and a total of 664,878 data points. Of these molecules 1,541 remained as confirmed hits.

3.1.4.2 Assay B

Assay B is an enzyme activity assay based on the radioactive SPA LEADseeker technology [Amersham Biosciences, Freiburg, Germany]. It has its implication in the field of respiratory disease namely the treatment of asthma and COPD [Barnes, 2002; Barnes, 2004]. In total 688,738 different compounds were tested leading to a mean result of 87.9% CTL and a standard deviation of 8.7%. The hit threshold was set to 20% CTL leaving an impressive number of 28,239 primary hits (i.e. a hit rate of 4.1 %). Due to the high hit rate a maximum divers set of 12,918 compounds entered confirmation whereof 11,633 true hits occurred having a mean activity below 28.7% CTL (i.e. a confirmation rate of 90.1%). After filtering redundant compounds and compounds having no structure 11,853 primary hit remained and a total of 549,525 measured data points. The hits not tested in confirmation were discarded from the analyses. 10,775 primary hits were confirmed.

3.1.4.3 TGF-β Type I Receptor

The inhibitory assay against TGF-β Type 1 receptor is explained in more detail in 2.1.1 [Yingling et al., 2004]. It is characterized by a Z' factor of 0.7 - 0.8 (Eq. 2.1), showing the high quality of the assay in terms of signal to noise ratio [Zhang et al., 1999]. TGF-β type I receptor is assumed to have various implications in fibrosis remodelling and was performed to identify new drugs for COPD or asthma [Barnes, 2002; Barnes, 2004]. 868,276 different compounds were tested in the primary screen resulting in a mean % CTL of 104.1% and a standard deviation of 4.3%. Based on a hit threshold of 50% CTL, 15,936 hits were obtained (1.8% hit rate). 15,289 compounds entered confirmation measurement and 13,419 true hits resulted based on a threshold of 54% CTL. After filtering redundant compounds and compounds having no structure specified 11,284 primary hit remained and a total of 738,861 measured data points. The hits not tested in confirmation were discarded from the analyses. Of these hits 9,581 hits were confirmed.

3.1.5 Dopamine Data

The sets consist of 472 compounds containing K_i values for dopamine D_2 and D_3 receptors [Missale et al., 1998] with a K_i value of 1 mM as maximum for both receptors. The molecules mostly belong to the class of analogues of BP 897, a clinical phase two dopamine D_3 partial agonist (Table 1.1) [Joyce & Millan, 2005, LeFoll et al., 2005]. The affinity of the compounds is spread from 0.33 nM to 1 mM for dopamine D_3 receptors and from 1.6. nM to 1 mM for dopamine D_2 receptors. For 386 compounds K_i values at both receptors were below 1 mM. Consequently, selectivity ratios of D_2/D_3 or D_3/D_2 were calculated only for these entries. The molecules are present in Appendix A.

3.1.6 Fisher's Iris Data

Fisher's Iris data set consisted of 150 random samples of flowers from the Iris species *setosa*, *versicolor*, and *virginica*. For each species there were 50 observations for sepal length, sepal width, petal length, and petal width in cm, yielding a four-dimensional descriptor space [Fisher, 1936]. The data set is listed in Appendix B.

3.2 Data Preparation Strategies

One essential step in computational chemistry is data preparation. This incorporates (i) pre-filtering of molecules and (ii) molecule preparation. Pre-filtering of molecules can have several reasons, for example to obtain only drug- or lead-like compounds [Muegge, 2003] or to design focussed or targeted libraries [Balakin et al., 2002; Balakin et al., 2003; Lang et al., 2002]. A further reason for pre-filtering is that the accuracy of classification and regression techniques is sensitive to outliers. Hence, extreme outlying molecules should be avoided [Verma & Hansch, 2005].

The type of preparation for a molecule data set highly depends on the application. Especially docking, pharmacophore and QSAR applications require extended preparations due to the following reasons: The strength of H-bond varies greatly from 2-15 fold affinity increase for neutral bonding to up to a 3,000 fold affinity increase for charged bonding [Davis & Teague, 1999]. According to this molecules have to be charged correctly. For docking and 3D pharmacophore applications different tautomeric, stereoisomeric and conformeric representations of a molecule have to be considered. Depending on the number of tautomeric and chiral centres and the flexibility of the molecule, many degrees of freedom result and can lead to "combinatorial explosion" [Kitchen et al., 2004]. To avoid this a possibility is to restrict to the energetically preferred tautomer, reject molecules having more than a predefined number of chiral centres and focus either on one or on all energetically preferred but distinct conformers. A detailed review about conformer generation strategies can be found elsewhere [Leach, 1996]. The following sections describe the applied molecule filtering and molecule preparation steps.

3.2.1 Data Filtering

In the analyses of the MDDR only small organic molecules having drug-like properties were used. Two filtering steps were applied, one based on key words describing the therapeutic implication of the molecules and one based on molecular properties of the molecules:

Key-word Based Filtering

A molecule was filtered out if the key words mentioned below were assigned to the compound. If another key word was additionally present (e.g. kinase inhibitor) no filtering was performed:

blood supplements, vaccines, monoclonal antibodies, molecules for cancer immune therapy, chemo-preventives, chemo-protectives, molecules for gene therapy, radio sensitizer, diagnostic agents, antidotes, antibiotics and antineoplastica.

Property Based Filtering

- Molecules were filtered having a molecular weight below 150 Dalton or above 1,000 Dalton. The molecular weight descriptor implemented in the MOE program package was employed [Chemical Computing Group, Montreal, Canada].
- Molecules were removed containing reactive functional groups [Hann et al. 1999]. Example structures are shown in Appendix C: carbazide, acid anhydride, pentafluorophenylester, paranitrophenylester, hydroxybenzotriazole (HOBT)-ester, triflate, Lawson's reagent, phosphor-amide, aromatic azide, beta-carbonyl-quaternary-nitrogen, acyl-hydrazide, cationic carbon/ chlorine/ iodine/ phosphor/ sulphur, phosphorane, chloramidine, nitroso, phosphor or sulphur halide, carbodiimide, isonitrile, triacyloxime, cyanohydrin, acyl cyanide, sulfonylcyanide, cyanophosphonate, azocyan-amide, azoalkanal, acid halide, peroxide. The program FILTER (version 2.0) [OpenEye Scientific Software, Santa Fe, USA] was used. The functional groups are named after the corresponding annotation in the program.
- Molecules were rejected bearing more than six halogen atoms. The program FILTER was employed (version 2.0) [OpenEye Scientific Software, Santa Fe, USA].
- Molecules were discarded not containing a least one carbon atom and one nitrogen/ oxygen/ sulphur atom. A pearl script was used provided by Dr. Bernd Beck [Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany].
- Molecules were removed containing additional elements than H, C, N, O, F, P, S, Cl, Br, and I. Pipeline Pilot (version 5.1) [SciTegic, San Diego, USA] was used.

3.2.2 Removal of Counter Ions

The simplest strategy to remove counter ions is to reject the smaller fragment. However this is prone to errors since e.g. inorganic fatty acids might be larger than the actual active molecule or both the putative biologically active molecule and the putative counter ion are "drug-like" and have a similar molecular weight. An extended strategy is to remove counter ions and solvents based on catalogues of known counter ions and solvents, and only if the putative counter ion is unknown to remove the smaller molecule. Note that for the latter entries wrong assignments can occur. In this study a statistical approach was developed to identify counter ions not based on rejection of the smaller molecule. The basic idea behind the strategy is that counter ions occur more often in the data set than the active molecules. Therefore the occurrence of each molecule was counted. Two user-defined rules were applied to distinguish active molecules from counter ions/solvents: (i) the active molecule occurs less than six times in the database and the counter ion occurs more than 20 times. (ii) The counter ion occurs more than 50 times in the database irrespective of the occurrence of the active molecule. Applying both rules to the MDDR 453 entries remained undefined and were discarded. The statistical approach was implemented in Kensington discovery edition (version 1.9) [InforSense Ltd., London].

3.2.3 Neutralization and Charge Assignment

To avoid erroneous charge assignments for the ligand-based clustering and virtual screening approaches all molecules were kept neutral. Neutralization was based on catalogues of known basic and acidic groups and was performed using Pipeline Pilot (version 5.1) [SciTegic, San Diego, USA]. The remaining molecules bearing a positive or negative charge were checked manually and if necessary neutralized using MOE-SVL scripts for unrecognized cases [Chemical Computing Group (CCG), Montreal, Canada].

To ionize molecules at a certain pH value various algorithms have been described. They can be categorized into methods based on quantum mechanics *ab initio* calculations, density functional theory, semi-empirical quantum mechanics, comparative molecular field analysis or methods creating models based on known acids and bases [Xing et al., 2003]. For large scale applications like virtual screening only the latter software tools are applicable. They are either based on expert systems looking for known acidic or basic groups (e.g. the PATTY type functionality [Bush & Sheridan, 1993]) or on models predicting the pK_a of a molecule at a certain pH value like ACD/pK_a DB [ACD Inc., Toronto, Canada]. For the docking and pharmacophore approaches the charging routine of MOE was employed. It is based on an extended PATTY type catalogue [Bush & Sheridan, 1993].

3.2.4 Conformer Generation

For the similarity searching (see 3.3.6) and the classification and regression techniques (see 3.5) a single conformation was created for each molecule using CORINA (version 3.2) [Molecular Networks GmbH, Erlangen Germany] with default parameters. For the pharmacophore searching (see 3.6), sampling of the conformer space is necessary for successful screening. Conformers were created using the stochastic search algorithm in MOE: molecules are divided into overlapping fragments. For the fragments a stochastic conformational search is performed. This is followed by a Merck molecular force field (MMFF94x) based energy minimization [Halgren, 1996]. The fragments are assembled using a rigid body superposition. Conformers are removed if clashes or undesirable group conformations occur.

The search algorithm was performed with the default parameter setting of the MOE program package [Chemical Computing Group, Montreal, Canada]; i.e. a maximum strain energy of 4 kcal/mol and a maximum of 250 conformers per molecule. To avoid combinatorial explosion caused by flexible molecules or molecules with many chiral centres, compounds having the following properties were rejected prior to conformational sampling: number of rotatable bonds > 7, single bond chain length > 6, chrial centres > 4, unconstrained chiral centres > 3, number of rings > 8. The different tautomers were not considered.

3.2.5 Ontology assignment

When creating a model for e.g. a targeted library, a first step is to select all molecules interacting with members of the target family. The selection is based on the names of the proteins. Schuffenhauer et al. extended this process and derived a hierarchical ontological activity description for the MDDR molecules, e.g. an angiotensin cleaving enzyme inhibitor is classified after the EC convention into enzymes (root class), hydrolases (subclass 1), peptidases (subclass 2) etc. [Schuffenhauer et al., 2002]. In total, a relation was created for compounds targeting GPCRs, ligand-gated ion channels, nuclear hormone receptors and enzymes. This system was captured and extended for the used MDDR version. It allows a coarse-grained or a fine-grained view on the MDDR molecules. In total an ontological description was assignable for 59,173 molecules.

3.3 Molecular Similarity, Descriptors and Descriptor Selection

A generally accepted hypothesis is that structurally similar molecules have a higher chance to exhibit a similar biological activity profile [Johnson & Maggiora, 1990, Martin et al., 2002]. This *Similarity Principle* is the foundation for a successful application of similarity searching, classification or regression methods. A key feature of the molecular similarity concept is the description of the chemical space. These descriptors can be categorized according to their data representation (e.g. bit strings, numerical values, vector representations); dimensionality (1D, 2D or 3D), their chemical information content (e.g. structural descriptors, physicochemical descriptors or pharmacophore representations) or whether they are alignment-free or not [Bajorath, 2002; Böcker et al., 2004]. In the following only alignment-free descriptors are described, whereas 3.6 describes the alignment-dependent pharmacophore concept. The second key feature of the similarity concept is the definition of a scheme allowing to measure similarity [Willett et al., 1998, Martin, 2001]. Various such schemes have been proposed ranging from simple numerical metrics (e.g. Euclidean or Manhattan metric) over schemes coping well with fingerprints (e.g. Tanimoto dissimilarity in combination with Daylight Fingerprints [Daylight Chemical Information Systems, Inc. Los Altos, USA]) to more sophisticated methods giving the descriptors an additional weighting. Calculating hundreds of descriptors for a data set may require descriptor selection exemplified by the following reasons: (i) saving of disc space, (ii) avoiding feature over-representation due to too many correlated descriptors, (iii) avoiding model over-training by having more descriptors than data points, (iv) identifying descriptors helping to understand SAR in the data. Different strategies are used at present to guide descriptor selection ranging from unsupervised procedures trying to identify redundant or non-relevant features [Whitley et al., 2000] to supervised procedures employing classification techniques to select the descriptors relevant for describing a certain activity. Since the latter is NP-complete (non-deterministic in polynomial time), machine learning techniques like genetic algorithms [Wegner et al., 2004, Hoffman et al., 2000], Particle Swarms [Agrafiotis & Cedeno, 2002] or Artificial Ants [Izrailev & Agrafiotis, 2001] have proven to be useful. In the present study only unsupervised approaches were followed.

3.3.1 CATS 2D

The correlation vector descriptor CATS 2D (150 dimensions) is based on potential pharmacophore points (PPPs). Atoms are assigned to five different PPPs (hydrogen donor, hydrogen acceptor, ionisable or positively charged, ionisable or negatively charged and lipophilic) and correlated with the respective distance counted in bond lengths (ranging from zero to nine bonds) [Schneider et al., 1999].

The PPP are defined as follows: hydrogen-bond donors correspond to oxygen atoms of OH groups and nitrogen atoms of NH- or NH_2-groups. Hydrogen-bond acceptors correspond to 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 a primary amino-group. Negatively charged or ionizable atoms correspond to atoms with a negative charge and carbon, sulphur or phosphorous atoms of a COOH-, SOOH-, or POOH-group, respectively. Lipophilic atoms were chlorine, bromine, iodine, sulphur atoms adjacent to exactly two carbon atoms, and carbon atoms adjacent only to carbon atoms. Applying this definition, atoms were assigned to no, one or two PPP-types.

The CATS descriptor was calculated with the program *speedcatsdotcom* by Uli Fechner (version 1.02, University of Frankfurt, Frankfurt, Germany). Scaling was done with the parameter –d 3, which corresponds to a normalization of a PPP pair to its respective occurrence in the descriptor [Fechner et al., 2003].

3.3.2 CATS 3D

The correlation vector descriptor CATS 3D (420 dimensions) is, in contrast to the topological CATS 2D descriptor, based on potential pharmacophore points (PPP) in 3D space. Atoms are assigned to six different PPPs (hydrogen donor, hydrogen acceptor, cationic, anionic, polar and hydrophobic) and correlated with the respective distance counted in ranges of 1 Å (ranging from zero to 20 Å) [Fechner et al., 2003].

For CATS 3D the modified PATTY atom-types available with the pH4_aType function in MOE were used. Scaling was performed by normalization of PPP pairs to their respective occurrence in the descriptor. Prior to descriptor calculation the 3D structure of the compounds was either calculated employing CORINA [Molecular Networks GmbH, Erlangen Germany] or MOE MMFF94x based energy minimization. The CATS3D descriptor was calculated with the program *cats3d_db* written in SVL in MOE by Dr. Steffen Renner (University of Frankfurt, Frankfurt, Germany).

3.3.3 MOE 2D

The MOE 2D descriptor set contains 146 descriptors describing physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrices, and pharmacophore features (http://www.chemcomp.com/journal/descr.htm) [Chemical Computing Group, Montreal, Canada].

3.3.4 Daylight Fingerprints

The Daylight fingerprint descriptor is a bit-string representation of the presence (1) or absence (0) of a certain structure defined by the Morgan algorithm (i.e. all single atoms, two atom combinations, three atom combinations etc. are identified and hashed on a fingerprint of defined length) [Morgan, 1965]. As fingerprint length a bit-string of 1,024 bits was used. For further detail on fingerprint calculation see URL: www.daylight.com.

3.3.5 Descriptor Preparation and Selection

Mean Centring and Scaling to Unit Variance

Descriptors can cover different data ranges (e.g. molecular weight and logP). For further analyses, like similarity calculations, PCA or PLS descriptors are required to be projected onto the same data range. A probate projection technique is to centre a descriptor according to its mean and later scale it onto unit variance [Otto, 1998]. It offers the additional possibility to detect descriptors with low variance. Such descriptors are not relevant and can be deleted. In the present work descriptors having a standard deviation less than 0.0005 were discarded. It has to be mentioned that this scaling routine is only defined for descriptor showing a normal distribution. Further outliers have a high impact on the scaling and by projection onto unit variance a loss of information occurs. Different alternatives are described in literature like normalization, Pareto scaling [Eriksson, 2001] or variable stability scaling [Keun et al., 2003]. The mean x_{mean} for a descriptor column **x** with *n* entries was calculated according to equation 3.1.

$$x_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{3.1}$$

The standard deviation σ for a descriptor column **x** was calculated according to equation 3.2.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{mean})^2} \,. \tag{3.2}$$

The scaling of a descriptor value x_i was performed according to equation 3.3.

$$x_i' = \frac{x_i - x_{mean}}{\sigma} \quad . \tag{3.3}$$

Entropy based Descriptor Selection

A measure of information content is Shannon entropy (SE) [Shannon, 1948], which was adopted by Bajorath and co-workers to quantify a descriptor's information content [Godden & Bajorath, 2000, Godden & Bajorath, 2001]. For this work descriptor were rejected having low information content. The Shannon Entropy is defined by Equation 3.4.

$$SE = -\sum_{i} p_i \log_2 p_i \ , \tag{3.4}$$

with p_i giving the probability of the number of data entries c_i within a data range i (Eq. 3.5).

$$p_i = \frac{c_i}{\sum c_i} \quad . \tag{3.5}$$

As proposed by Bajorath and co-workers each descriptor value range was subdivided into N_i = 100 equidistant data ranges i ("bins"). Descriptors having only a constant value over the whole data set have no statistical variance and were discarded prior to the calculation of *SE*. To make the *SE* independent of the number of bins, the obtained values were normalized by the logarithm to base two of the amount of N_i (Eq. 3.6).

$$sSE = \frac{SE}{\log_2(N_i)} \ . \tag{3.6}$$

Bajorath and co-workers defined descriptors having a scaled SE (sSE) equal to or less than 0.3 as "information-poor". In the presented work these descriptors were discarded.

Redundancy-based Descriptor Selection

To remove redundant dimensions from a descriptor set, Unsupervised Forward Selection (UFS) was performed as published by Whitely and co-workers [Whitley et al., 2000]. Starting with the two least correlated descriptors, this method builds up a descriptor space by choosing the next descriptor **x** having the lowest multiple correlation coefficient R^2 to the currently selected l descriptors (Eq. 3.7).

$$R^2 = \left| \sum_{k=1}^{l} (\mathbf{x} \cdot \mathbf{c}_k) \mathbf{c}_k \right| . \tag{3.7}$$

{$\mathbf{c}_1$, ..., $\mathbf{c}_l$} represents the orthogonal basis for the subspace spanned by the selected l descriptors. **x** is added to the selected descriptor set whereas its orthogonal projection is obtained by $\mathbf{c}_x = \mathbf{Y}/|\mathbf{Y}|$ where **Y** is defined by equation 3.8

$$\mathbf{Y} = \mathbf{x} - \sum_{k=1}^{l} (\mathbf{x} \cdot \mathbf{c}_k) \mathbf{c}_k . \tag{3.8}$$

The algorithm is performed until a predefined threshold for R^2 is reached [Whitley et al., 2000].

3.3.6 The Similarity Concept in Virtual Screening

In similarity searching nearest neighbours of a query molecule are searched in a target data set. This is achieved by describing all molecules with a set of descriptors or fingerprints. The fingerprints and descriptors are employed to calculate the distance between the query and the target. After searching a sorted list of nearest neighbours to the query is created. Many different distance metrics or similarity coefficients are available for searching [Willett et al., 1998, Willett, 2005]. In the present work the metrics or coefficients listed in Table 3.2 were implemented and used for numerical descriptors. Additionally, the Tanimoto dissimilarity coefficient T shown in Equation 3.9 was used in combination with Daylight fingerprints.

$$T(A,B) = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}} \quad , \tag{3.9}$$

where N_A and N_B are the number of bits set in the bit strings of molecules A and B, respectively, and $N_{A\&B}$ is the number of bits that are set in both.

To judge the outcome of a similarity search the percentage of examined data is plotted against the percentage of known actives in the list and enrichment curves are obtained. If a defined percentage of molecules was selected an enrichment factor is calculated according to Equation 3.10.

$$EF_{i,c} = \frac{\frac{N_{i,c}}{N_c}}{\frac{N_i}{N}} , \tag{3.10}$$

with $N_{i,c}$ being the number of screened entries i belonging to class c, N_i being the total number of screened entries i, N_c being the total number of entries of class c in the data set and N being the overall number of entries. $EF > 1$ indicates that more compounds belonging to the activity class c have been retrieved than expected from an equal distribution.

Table 3.2 Similarity metrics and coefficients for numerical descriptors (adapted from Willet et al., 1998).

Name	Equation*	Data range
Manhattan distance	$D(A,B) = \sum_{i=1}^{n} \lvert x_{i,A} - x_{i,B} \rvert$	∞ to 0
Euclidean distance	$D(A,B) = \sqrt{\sum_{i=1}^{n} (x_{i,A} - x_{i,B})^2}$	∞ to 0
Soergel distance	$D(A,B) = \frac{\sum_{i=1}^{n} \lvert x_{i,A} - x_{i,B} \rvert}{\sum_{i=1}^{n} \max(x_{i,A}, x_{i,B})}$	1 to 0
Tanimoto coefficient	$C(A,B) = \frac{\sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}}{\sum_{i=1}^{n} (x_{i,A})^2 + \sum_{i=1}^{n} (x_{i,B})^2 - \sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}}$	-0.333 to 1
Dice coefficient	$C(A,B) = \frac{2 \cdot \sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}}{\sum_{i=1}^{n} (x_{i,A})^2 + \sum_{i=1}^{n} (x_{i,B})^2}$	-1 to 1
Cosine coefficient	$C(A,B) = \frac{\sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}}{\sqrt{\sum_{i=1}^{n} (x_{i,A})^2 \cdot \sum_{i=1}^{n} (x_{i,B})^2}}$	-1 to 1

*$x_{i,A}$ represents the value descriptor i of molecule A and $x_{i,B}$ represents the value descriptor i of molecule B, respectively

3.4 Unsupervised Classification Techniques

Unsupervised classification techniques embrace methods using only inherent properties of a data set (e.g. molecular descriptors) and no *a priori* knowledge of measured data for classification. One large group of methods represent clustering and unsupervised partitioning approaches, whereby clustering techniques group compounds according to distances in the descriptor space and partitioning techniques assign descriptor space coordinates to form compound groups. Other mostly graph based methods classify molecules according to e.g. maximum common substructures.

In context of the presented work unsupervised techniques were used for data classification, projection, visualization and maximum diverse subset selection. These methods are presented in the following. Two (unsupervised) hierarchical clustering algorithms, NIPALSTREE and hierarchical *k*-means, were implemented. Both algorithms represent the heart of this work. Consequently it was decided to present them in the *Results and Discussion* section.

3.4.1 Molecular Scaffold Analysis

An approach to understand common features present in drug molecules was published by Bemis and Murcko [Bemis & Murcko, 1996, Bemis & Murcko, 1999]. Commercially available drugs were separated into their largest connected ring system and side chains. For the ring systems all atoms were converted into sp^3 hybridized carbon atoms creating a molecular framework/scaffold. This is illustrated in Figure 3.1 for temocapril, a known angiotensin converting enzyme (ACE) inhibitor [Acharya et al., 2003]. In the original study the scaffolds and the largest connected ring systems were further analyzed. The 32 most abundant scaffolds represented half of the data indicating a low diversity among the current drugs [Bemis & Murcko, 1996]. In the present work the aim was to analyze the developed clustering approach for its capacity to retrospectively identify false-negatives bearing novel scaffolds in the primary screening hits of assay A, assay B and the assay against TGF-β type I receptor. The primary screening hits were converted into their corresponding scaffolds with a program written in SVL in MOE by Kristina Grabowski (University of Frankfurt, Frankfurt, Germany). The three most occurring scaffolds from each assay were selected and defined as false-negatives (i.e. they were defined as non-hits).

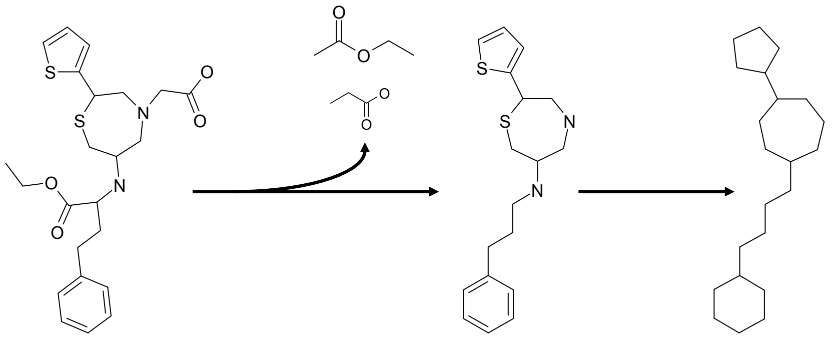


Figure 3.1 Molecular scaffold extraction of temocapril: The largest connected ring system is extracted followed by a conversion of all atoms into sp^3-hybridized carbon atoms.

3.4.2 Maximum Diversity Selection

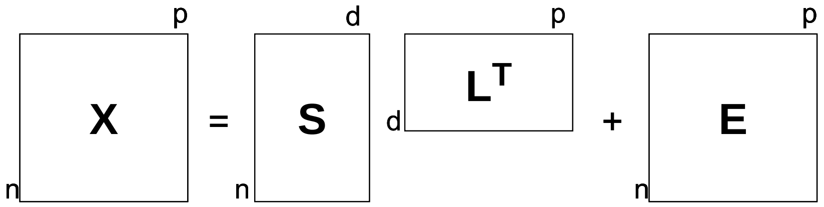
Various situations can occur requiring reducing a data set in size. This can be a virtual screening hit list too large for experimental testing or a computationally expensive method requiring a limited set of compounds. One rational selection approach is the creation of a maximum diverse subset. Algorithms for subset selection range from maximum dissimilarity selection techniques like MaxMin [Schmuker at al., 2004] over cluster-based or partition-based selection [Agrafiotis & Rassokhin, 2002] to techniques employing heuristics like simulated annealing [Reynolds et al., 2001]. In the present work the Stochastic Cluster Analysis (SCA) algorithm [Reynolds et al., 2001] was implemented belonging to the class of maximum dissimilarity selection algorithms:

Step 0: Describe compounds by numerical descriptors (e.g. 3.3.1 – 3.3.3).
Step 1: Define a distance threshold *T*, the desired number of compounds *N* to select and the distance metrics or similarity coefficient from section 3.3.6.
Step 2: Randomly select a compound from the data set and add it to the result list, if the distance to the previously selected molecules does not exceed *T*.
Step 3: Perform step 2 until *N* is reached.

N can only be reached if a suitable threshold *T* was chosen. If *T* is too low, a random selection takes place. In the present work *T* was determined that exactly *N* entries were selected.

3.4.3 Principle Component Analysis

Principle component analysis (PCA) is an orthogonal transformation of a *p*-dimensional data matrix **X** with *n* entries. The *d* new coordinate values are termed principle components. A principle component represents the portion of the original data matrix by one dimension **S**, which explains most of its variance with **L** as a new coordinate system. **L** has the dimensionality of the original data matrix, however each dimension is assigned a weight and a direction to obtain an uncorrelated coordinate system; i.e. **L** reflects the importance of each dimension to explain the variance in the data set. Each principle component contributes cumulatively to explain the variance in **X** and usually a small number of principle components are sufficient to explain most of the variance. The unexplained proportion of **X** remains in the residual matrix **E** (Eq. 3.11). PCA is employed for dimensionality reduction and to visualize a data set in two or three dimensions. It allows getting an overview of the clusters or outliers present in the data [Otto, 1998; Eriksson et al., 2001].



(3.11)

In the present work the NIPALS (non-linear iterative partial least squares) algorithm [Miyashita et al., 1990] was implemented for PCA calculations. It consists of the following steps:

Step 0: Define the number of principle components. Mean centre descriptors and scale to unit variance (3.3.5). Discard a descriptor column if its standard deviation underscores a threshold Θ. In the present study 0.0005 was chosen for Θ [Whitley et al., 2000].
Step 1: Employ the first row as initial loading vector **L**.
Step 2: Project the data matrix **X** onto **S** using **L**: $\mathbf{S} = \mathbf{XL}$
Step 3: Calculate new **L** using **S**: $\mathbf{L}^T = \mathbf{S}^T\mathbf{X}$. (^{T} means transpose).
Step 4: Normalize **L** to length 1: $\mathbf{L} = \mathbf{L}/|\mathbf{L}|$
Step 5: Project **X** onto a new score vector **S** employing **L**: $\mathbf{S} = \mathbf{XL}$
Step 6: Calculate the difference D between previous and new **S**: $D = \mathbf{S}_{\text{old}} - \mathbf{S}_{\text{new}}$. If D exceeds a threshold Φ, return to step 3. In the present study 0.0005 was chosen for Φ.
Step 7: Remove principle component from **X**: $\mathbf{X} = \mathbf{X} - \mathbf{SL}^T$.

Step 8: Continue with step 1 until the predefined number of principle components is reached.

3.4.4 Non-linear Component Analysis

If complex relations occur in data set, a linear transformation employing PCA might be not well suited. Instead a non-linear component analysis (NLCA) can be performed trying to elucidate higher order correlations. One method for NLCA is an encoder network which is exemplified in Figure 3.2 [Duda et al., 2001; Schneider & Wrede 1998].

Encoder networks operate by presenting a *d* dimensional descriptor matrix to both the input layer und the output layer with *d* units each. Both layers are fully connected by three hidden layers, two with non-linear units and one central with linear units. In the present study training was performed using a (*1, λ*) evolution strategy with a sum-squared error minimization fitness function for weight adoption, five non-linear units in each hidden layer and three units in the central hidden layer. By this architecture encoder networks learn to reproduce the input patterns at the output layer (auto-encode) by the internal representation in the hidden units. After training the central units represent the new coordinate system of the non-linear principle components. NLPCA was performed using the program ChemSpaceShuttle by Alireza Givehchi (University of Frankfurt, Frankfurt, Germany) [Givehchi et al., 2003].

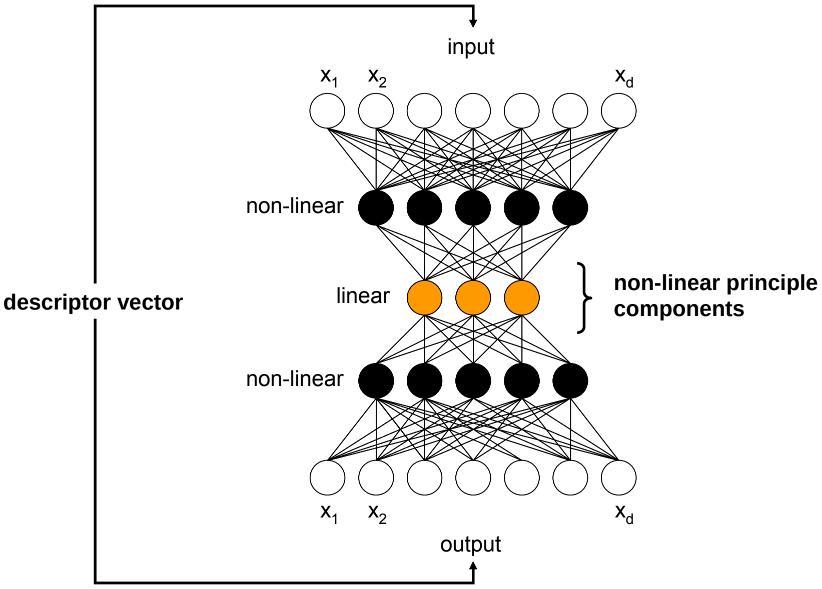


Figure 3.2 Topology of an encoder network for performing a non-linear principle component analysis. The descriptor matrix with *d* descriptors is fed into the output and input layer (white circles). Hidden neurons are represented by black and orange circles. After training the orange circles represent the non-linear principle components.

3.4.5 Clustering Techniques

Self Organizing Map

The self organizing map (SOM) technique [Kohonen 1982, Schneider & Wrede, 1998] is a non-linear clustering technique, allowing visualizing a *d*-dimensional data set as a two-dimensional (or more) map containing *n* clusters (or fields). Employing Kohonen's algorithm [Kohonen, 1982] a topology preserving projection of the *d*-dimensional space is obtained. Each cluster is represented by a neuron c_i with *d* weights. In a distance dependant and time dependant iterative optimization procedure neuron weights are adopted best representing the data set [Teckentrup et al., 2004].

In the present work the Euclidean distance was used, the number of optimization cycles was defined as ten times the number of data points, the neighbourhood weighting was set to half of the maximum edge of the map and the learning step size was set to one. To avoid boundary problems a toroidal topology was employed for the maps. SOM calculation was performed using the program *som_create* by Gisbert Schneider. SOM visualization was performed using the program *som_show* by Gisbert Schneider (University of Frankfurt, Frankfurt, Germany).

Phylogenetic-Like Tree Clustering

The phylogenetic-like tree clustering approach is a hierarchical hybrid clustering method and it consists of the following steps [Nicolaou et al., 2002]:

Step 0: Molecules are described by fingerprint descriptors like MACCS keys [MDL Information System Inc., San Leandro, USA].

Step 1: A SOM is used to cluster molecules in a tree node based on their chemical descriptors.

Step 2: SOM clusters are selected containing only molecules of high similarity.

Step 3: Maximum common substructure (MCS) are extracted from these clusters.

Step 4: Expert rules are applied to evaluate the substructures and to eliminate all that do not constitute a significant gain in knowledge.

Step 5: Each newly identified MCS is used to search the data set for molecule matching the substructure. The resulting list forms a new tree node.

Step 6: The algorithm is repeated from step 1 for the new node.

Results of the clustering are classes, subclasses and singletons. In the present work the similarity threshold defining a cluster was set to obtain homogeneous clusters. Further a redundancy threshold was set to construct diverse and unique classes. Only classes were selected since they are assumed to correspond soonest to “lead classes”. The clustering was performed using the program ClassPharmer (version 3.2) [BioReason Inc., Santa Fe, USA].

Ward’s Clustering

Ward’s hierarchical clustering algorithm is an *agglomerative* clustering method maximising the inter-cluster variance whilst minimizing the intra-cluster variance [Ward, 1963]. In the present work the reciprocal nearest neighbour version of the algorithm was employed [Murtagh, 1985] provided by the Kensington Discovery Edition (version 1.9) [InforSense Ltd., London]. It consists of the following steps:

Step 0: Define a descriptor space.
Step 1: Find nearest neighbours by calculating the Euclidean distance (Table 3.2) between the data points.
Step 2: Trace path of unvisited nearest neighbours until a reciprocal nearest neighbour is found.
Step 3: Merge reciprocal nearest neighbour pair to a new data point (the centroid).
Step 4: Repeat 1 to 3 until one unvisited point remains.
Step 5: Sort reciprocal nearest neighbours by increasing Euclidean distance to each other.

3.5 Supervised Classification and Regression Techniques

Supervised techniques embrace all methods using additional *a priori* knowledge of measured data for model building. They can be subdivided into *classification* methods, which predict the belonging to a certain group (actives or not actives), and *regression* methods, which try to create functional dependencies e.g. between structure and activity. Both concepts have become standard in the field of early drug discovery [Kubinyi, 1993; Höltje & Sippl, 2001; Böhm et al., 2002]. In the presented work at first the supervised methods are introduced employed for classification and regression. In a final chapter detailed introduction will be given to statistical validation of the created models.

3.5.1 Partial Least Squares

Partial least squares (PLS) can be described as the regression extension of principle component analysis (PCA) [Wold, 1966]. Instead of describing the maximum variation in the "measured" data (**X**, e.g. a descriptor set), which is the case for PCA, PLS attempts to derive latent variables, analogues to principle components, which maximize the covariation between the "measured" data **X** and the "response" variables (**Y**, e.g. pIC_{50} values). Result of the PLS analysis are de-correlated scoring vectors for the **X** and **Y** matrix (new dimensions, size = number of rows in **X** and **Y**), loading vectors for the **X** matrix (size = number of columns in **X**) and weighting vectors for **X** and **Y**. The weighting vectors and loading vectors serve to predict **Y** vectors for new molecules. Using the so called "variable influence of projection" (VIP), PLS offers the additional possibility to identify the importance of descriptors for both **X** and **Y** matrix. VIP represents a weighted sum of squares of the PLS weights, employing the amount of explained **Y**-variance of each descriptor.

In the present work, prior to PLS calculation descriptor columns were mean centred and scaled to unit variance. Descriptors with a standard deviation < 0.0005 were discarded. Quality of the model was assessed by calculating Q^2 (goodness of prediction) and R^2 (goodness of fit), whereas Q^2 was determined by seven-fold cross-validation using randomly selected partitions of equal size (software default). The number of latent components was determined, as the point showing an optimal balance between R^2 and Q^2. PLS was performed using SIMCA-P+ 10.5 [Umetrics Ab, Umea, Sweden, Eriksson et al., 2001].

3.5.2 Support Vector-based Regression

Support vector-based regression (SVR) has its foundation in statistical learning theory and belongs to the Kernel-based optimization functions. One SVR type is ε-SVR [Smola & Schölkopf, 1998, Liu et al., 2006; Zhou et al., 2006]. For a data set **X,** with entries **x** and measured values **y** (note that the entries **x** are represented by a set of descriptors), it has the objective to find a function $f(\mathbf{x})$ that has at most ε deviation from the measured values **y** for all entries **x**, and at the same time is as flat as possible. This is described for a linear function by $f(\mathbf{x})=(\mathbf{w}, \mathbf{x})+\mathbf{b}$, with $(\mathbf{w}, \mathbf{x})$ being the dot product in **X** and **w** as small as possible to ensure flatness. $f(\mathbf{x})$ is approximated, that all points ($\boldsymbol{x}$, **y**) are predicted with at least ε precision (margin, Figure 3.3). The regression function $f(\mathbf{x})$ is obtained by solving the constrained quadratic optimization problem and is given by equation 3.12.

$$\mathrm{f}(x) = \sum_{i=1}^{l} \alpha_i \left\langle x_i^{sv} \cdot x \right\rangle + b \ , \qquad (3.12)$$

with α_i being Lagrange multipliers, $\mathbf{x}_i^{sv}$ being the support vectors at margin ε and **b** being a vector with constant values. Analogously to the "soft margin" loss function in support vector machine-based binary classification slack variables ζ_j are introduced to cope with otherwise unsolvable constraints in the optimization problem (Figure 3.3). The trade-off between the flatness of f and the amount of slack variables is regulated by a global parameter C.

To extend the approach to solving non-linear problems, the data points are mapped into a high dimensional space, where a linear function is established. This corresponds to evaluating kernel *k* functions at location $k(\mathbf{x}_i, \mathbf{x})$. In the present study the radial basis function (RBF) kernel was employed (Eq. 3.13) [Schölkopf et al., 1997; Smola & Schölkopf, 1998; Zhou et al., 2006].

$$k(x_i, x) = \exp(-\gamma \|x_i - x\|^2), \gamma \in R. \qquad (3.13)$$

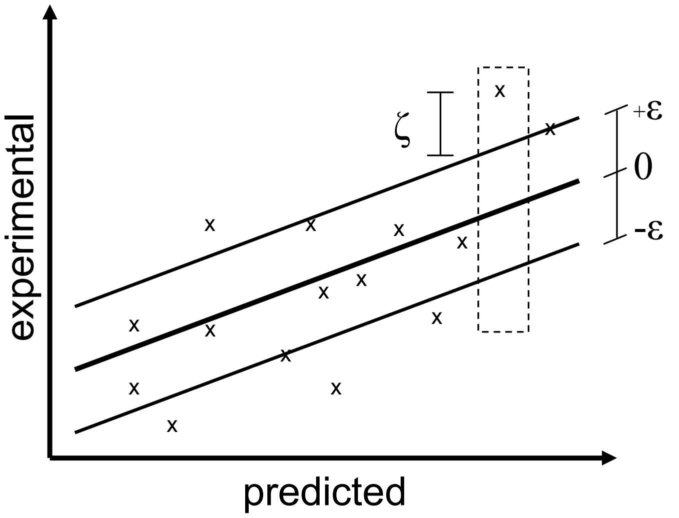


Figure 3.3 Principle of a linear support vector-based regression. By solving the quadratic optimization problem it is tried to find the function f(x)=(w, x)+b approximating all points (x, y) with at most ε precision (margin). Additional slack variables ζ can be introduced to cope with otherwise unsolvable constraints in the optimization problem.

The SVM software package LIB-SVM 2.5 was used [Chang & Lin, 2001]. $\log_{10}(K_i)$ values of the corresponding data set were used as **y** vectors. The descriptors representing the data points

x of the training set were scaled to the interval [-1,1] prior to calculation. The test data and if present the external validation data were scaled according to the scaling factors of the training set. Parameter pairs C and γ for SVR training were systematically examined starting from $\log_2(C) = -2$ to $\log_2(C) = 15$ and $\log_2(\gamma) = -15$ to $\log_2(\gamma) = 2$ with a step size of 0.2 [Kriegl et al., 2005]. The SVR model was selected best predicting the y variables of the test set in terms of R^2. As additional objective function the squared correlation coefficient Q^2 was used, obtained by 7-fold cross validation on ten different random splits of the training set (i.e. an average Q^2 value of 10 individual Q^2 values was created). ε was set to 0.1. The parameter settings are defined in the parameter file shown in Appendix D.

3.5.3 Bayesian Regularized Artificial Neural Networks

Artificial neural networks (ANN) were shown to model classification and regression problems effectively [Manallack & Livingstone, 1999; Schneider, 2000]. As an example a fully connected feed forward network topology is presented in Figure 3.4.

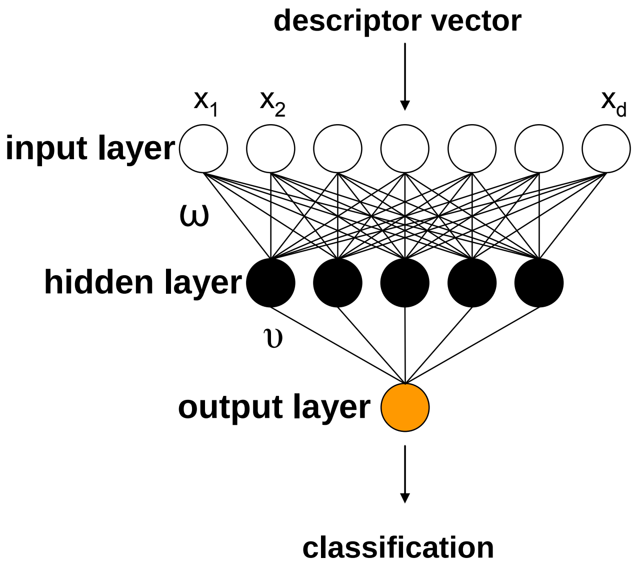


Figure 3.4 Fully connected feed forward artificial neural network for classifying a molecule characterized by a descriptor vector. The input layer consists of as many units as descriptors are present. The hidden layer consists of five units and the output layer contains one unit. Input to hidden layer connections are characterized by weights ω and hidden to output layer connections by weights υ.

The input layer contains as many units as descriptors are used. The output unit contains only one unit sufficient for classification tasks. By introducing a hidden layer non-linear relation in the data can be modelled. Function approximation for f(x) (Eq. 3.14) is achieved by systematically adopting weights ω_{ij} from input layer to hidden layer and υ_j from hidden layer to output layer to minimize a classification error E.

$$\mathrm{f}(x) = g\left(\sum_{j=1}^{m} \upsilon_j h\left(\sum_{i=1}^{n} \omega_{ij} x_i - \zeta_j \right) - \zeta_{out} \right) \tag{3.14}$$

ζ_j and ζ_{out} are bias values for the hidden layer and the output layer respectively. For *g* and *h* different activation functions are used like the logistic sigmoidal function tanh (Eq. 3.15).

$$\tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \tag{3.15}$$

Training an ANN can translate into over-fitting and over-training. Over-fitting occurs when the network topology is too complex and too many weights are present for optimization. An approach to control the effective complexity is regularization which introduces an additional penalization term Ω to the error function E (Eq. 3.16),

$$E^{'} = E + \beta \Omega \ . \tag{3.16}$$

In this context β controls the extent to which Ω influences the form of the solution. A simple form of regularizer is weight decay and is achieved by calculating the sum of squares of all adaptive parameters in the ANN. Over-training occurs when too many optimization cycles were performed and the network models not only regularities but noise. To circumvent this, the prediction for an external test set is monitored. ANNs optimize a single set of parameters, i.e. one solution is identified out of a complex solution landscape. For a detailed description of the discussed neural networks see [Bishop, 1995].

A different ANN approach, Bayesian regularized artificial neural networks (BRANN), was employed in this work. It has been claimed to be less sensitive to over-training and over-fitting [Ajay et al., 1998; Burden & Winkler, 1999; Bruneau, 2001]. Instead of initializing and optimizing a network with a single set of network parameters (weights, biases and offsets) a network is initialized with all possible combinations of parameter values. Bayesian inference integrates the prediction of the network over all parameter values, whereas each parameter set combination is weighted by its posterior probability established from the training data. No error minimization is used and a separation of the input data into training and validation set is not necessary. Further since all network parameters are used for prediction, regularization favouring less complex solutions is an inherent property of BRANN. Bayesian inference initializes a prior probability distribution over the weights *P(w)*. By employing the descriptor vectors ($\mathbf{x}^{\mathbf{1}}$,..., $\mathbf{x}^{\mathbf{N}}$) and the corresponding response vectors ($\mathbf{y}^{\mathbf{1}}$,..., $\mathbf{y}^{\mathbf{N}}$) the posterior probability for each model is determined using Bayes' theorem (Eq. 3.17),

$$P(w \mid x, y) = \frac{P(w)P(y \mid x, w)}{P(y \mid x)}. \tag{3.17}$$

Bayesian analysis allows predicting $\mathbf{y}^{\mathbf{N+1}}$ for a descriptor vector $\mathbf{x}^{\mathbf{N+1}}$ by solving the integral (Eq. 3.18),

$$\mathbf{y}^{\mathbf{N+1}} = \int F_w(\mathbf{x}^{\mathbf{N+1}})P(w \mid x, y)dw, \tag{3.18}$$

where $F_w(\mathbf{x}^{\mathbf{N+1}})$ is the output of the network. In the present work the BRANN software provided by Neal was employed (http://www.cs.toronto.edu/~radford/fbm.software.html). It uses as priors *w* for the weights, biases and offsets independent Gaussian distributions with mean μ_0 of zero and variance σ (3.19).

$$f(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w-\mu_0}{\sigma}\right)^2} \tag{3.19}$$

Since appropriate values for σ are unknown in the beginning, they are defined by a gamma distribution $f(\sigma,\mu,\alpha)$ with mean μ and shape α and gamma function $\Gamma(\alpha)$ (Eq. 3.20). These gamma distributions are termed hyperparameters.

$$f(\sigma, \mu, \alpha) = \sigma^{\alpha-1} \cdot \frac{\alpha \cdot e^{-\frac{\sigma\alpha}{\mu}}}{\mu \cdot \Gamma(\alpha)}, \tag{3.20}$$

For the priors of the weights of the input to hidden layer and the weights of the output to hidden layer the mean μ is controlled by another gamma distribution with mean μ_2 and shape α. The reason for using the three level priors is, that by integration, the hyperparameters are adopted employing the concept of marginalization: the elimination of unwanted variables by integration. In the software the mean μ_2 and shape α for the weights and the mean μ and shape α for the offsets and biases are defined by the user. Additionally, for the response vectors ($\mathbf{y}^1$,..., $\mathbf{y}^N$) a noise parameter is determined which is treated as two-level prior with Gaussian distribution, where the variance σ is defined by a gamma distribution with mean μ and shape α. In the present work input to hidden weights were defined by $\alpha = 0.5$ and $\mu_2 = 0.05$. The bias for the hidden weights was specified by $\alpha = 0.5$ and $\mu = 0.05$. The hidden to output weights were defined by $\alpha = 0.5$ and $\mu_2 = 0.05$. The output bias was specified by $\mu = 100$. The noise for the response vector was defined by $\mu = 0.05$. All other parameters were not set (see the parameter file listed in Appendix E). Hyperparameters are sampled by the software.

The integrals in equation 3.18 are difficult to evaluate due to the degrees of freedom. A hybrid Markov chain Monte Carlo method was used by the software for the integration [Neal, 1994]. Using a metropolis algorithm a sequence of vectors is generated that form an ergodic Markov chain. It is assumed that this chain reaches a stationary (equilibrium) distribution. From N Markov chains present in the stationary condition equation 3.18 can be solved by calculating the averages (Eq. 3.21).

$$\mathbf{y}^{\mathbf{m}+1} = \frac{1}{N} \sum_{t=I}^{I+N-1} F_w(\mathbf{x}^{\mathbf{m}+1}), \tag{3.21}$$

where I specifies the initial values until the Markov chain reaches the stationary distribution. The energy function E(w) which is minimized by the Metropolis algorithm is defined by equation 3.22

$$E(w) = -\log P(w \mid x, y). \tag{3.22}$$

The Markov chain was initialized with a trajectory of 100 so called leapfrog steps, an average window size and a step-size adjustment of 0.2. The sampling was performed with the provided persistent hybrid Monte Carlo method, a trajectory of 1000 leapfrog steps, a step size adjustment of 0.3, a heat-bath decay of 0.3 and an average window size of 10. 1000 iterations were sampled whereas the last 200 iterations were employed for calculating the average (Eq. 3.21, 800 iterations were used to reach the equilibrium). The Monte Carlo settings are defined in the parameter file listed in Appendix E. For the analysis a fully connected feed-forward network with ten units in the hidden layer and one unit in the output layer was used. Note that suitable values defining the gamma distribution of the hyperparameters and the hybrid Monte Carlo algorithm were systematically evaluated by Dr. Jan Kriegl (Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany). The BRANN software was provided by Radford M. Neal [Neal, 1994].

3.5.4 Validation of Classification and Regression Techniques

In the present work three different types of data were distinguish for the regression and classification approaches, training data, test data and validation data. Training data were data points used for creating a model, test data were data points known to the model but not employed for training, and validation data were data points unknown to the model. The data sets are defined in 4.3.2, 4.3.3 and 4.4.5. In the following the statistical measures applied for the classification techniques and the regression techniques are shown. The measures were applied to all three types of data.

Validation of Classification Techniques

Classification techniques were characterized by a contingency table (or confusion matrix),

		Actual	
		-	+
Predicted	-	True-negatives (TN)	False-negatives (FN)
	+	False-positives (FP)	True-positives (TP)

containing the number of true positives (TP), false-positives (FP), false-negatives (FN) and true negatives (TN). From this table the measures summarized in Table 3.3 were calculated. In all cases a value between 0 and 1 is obtained. Sensitivity (or recall) is the correctly classified proportion of positives, whereas specificity (precision) is the correctly classified proportion of negatives.

For predictions both a high recall and a high precision is wanted. The false-negative rate is the proportion of incorrectly classified negatives, whereas the false-positive rate is the proportion of incorrectly classified positives. Both measures were required to be as low as possible. A “naïve” classification measure is accuracy since it does not take chance predictions into account. The κ index and the squared Matthews correlation coefficient R_{MCC}^2 circumvent this by taking the model’s improvement in prediction over chance into account [Chohan et al., 2005]. The R_{MCC}^2, the accuracy and the κ index range from 0 to 1, whereas a value of one represents a perfect prediction and a value of zero represents no correct prediction at all.

Table 3.3 Accuracy and error measures that can be calculated from a contingency table.

Measure	Formula
Accuracy	$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$
Sensitivity (recall)	$sensitivity = \frac{TP}{TP + FN}$
Specificity (precision)	$specificity = \frac{TN}{TN + FP}$
False-positive rate	$false\ positive\ rate = \frac{FP}{FP + TN}$
False-negative rate	$false\ negative\ rate = \frac{FN}{FN + TP}$
Kappa: κ	$\kappa = \frac{(TP + TN) - \frac{(TP + FN)\cdot(TP + FP) + (FP + TN)\cdot(FN + TN)}{FN + FP + TP + TN}}{(FN + FP + TP + TN) - \frac{(TP + FN)\cdot(TP + FP) + (FP + TN)\cdot(FN + TN)}{FN + FP + TP + TN}}$
Matthews R^2_{MCC}	$\mathrm{R}_{\mathrm{MCC}}{}^{2} = \frac{(TP \cdot TN - FP \cdot FN)^2}{(TP + FP)\cdot(TP + FN)\cdot(TN + FP)\cdot(TN + FN)}$

Validation of Regression Techniques

Regression techniques were evaluated according to the goodness of prediction, R^2, the goodness of fit, Q^2 and the root mean square error (RMSE). R^2 was calculated according to Equation 3.23.

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.23)$$

with $\bar{x}$ and $\bar{y}$ being the mean values of measured values x_i and predicted values y_i, respectively. Result of R^2 range from 0 to 1, whereas a value of one represents a perfect prediction and a value of zero represents no correlation at all.

In the present study Q^2 was obtained by leave-group-out cross validation (CV). In an iterative process a group from the training set is left out, a new model is trained for the reduced training set and the activity predicted for this group is then used for Q^2 calculation (Eq. 3.24).

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{3.24}$$

with $\overline{y}$ being the mean values of measured values y_i and x_i being the predicted values. Q^2 ranges from -∞ to 1 with a value of 1 resembling a perfectly robust model. In the present work Q^2 was obtained by 7-fold cross-validation. The indices defining the groups were randomly selected. For the SVR approach the cross-validation was repeated ten times (i.e. 10 different assignments of random indices) and the individual Q^2 values were averaged.

RMSE is a measure to calculate the deviation between predicted values x_i and measured values y_i and should be as low as possible (Eq. 3.25).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \quad . \tag{3.25}$$

3.6 The Pharmacophore Concept

A pharmacophore is defined as an “ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interaction with a specific biological target structure and to trigger its biological response” [Wermuth et al., 1998]. The steric and electronic features are usually defined as “hydrophobic”, “aromatic”, hydrogen-bond donor”, “cationic”, “hydrogen-bond acceptor” and “anionic” chemical features in 3D, are not restricted to single atoms and can possess directionalities. Since the interaction with the receptor is only assumed, the correct specification of such a feature is “potential pharmacophore point” (PPP). To cope with additional receptor constraints inclusion or exclusion volume spheres are possible [Güner, 1999].

A pharmacophore model is created by assigning features to a set of ligands assumed to bind to the same receptor in a similar binding mode. If the coordinates of the ligands are not known molecule alignment tools try to best represent the molecules according to their spatial PPPs arrangement [Kristam et al., 2005; Klabunde & Evers, 2005].

A pharmacophore search consists of the following steps: (i) definition of the 3D pharmacophore model by molecule alignment and feature assignment, (ii) conformer generation of the screening data set and (iii) pharmacophore searching. In the present work the functionalities implemented in the MOE program package were used for all three steps.

(i) Pharmacophore definition: as starting structures for the molecule alignment, conformations were used, obtained by docking into or extracted from the 3D structure of the corresponding receptor (see 3.7). The molecule alignment was refined employing MMFF94x force field [Halgren 1996] based energy minimization. The alignment itself was based on Gaussian feature density overlap calculations. In addition to the standard feature parameters the charge feature was assigned a weight of 1. As pharmacophore features the undirected pharmacophore type descriptions for hydrogen-bond donor, hydrogen-bond acceptor, cationic, anionic, aromatic ring centre and hydrophobic region were used [Lin, 2004]. The size of the potential pharmacophore points was defined manually.

(ii) Conformers of the screening data set were created as described in 3.2.4. For all conformers acids were deprotonated, bases were protonated (see 3.2.3) and undirected pharmacophore features were calculated prior to searching.

(iii) The searching was performed on the conformer data set employing the model(s). Only the best scoring conformer was kept of each molecule.

3.7 Docking Techniques

Docking comprises methods simulating the mutual recognition of small molecules (the ligands) and their macromolecular biological targets (the receptors, mostly proteins). The docking process can be decomposed into addressing two separate problems: posing and scoring. Posing is the determination whether a ligand conformation fits into the receptor binding pocket by geometric complementarity, whereas scoring is the determination of the binding affinity of the conformation in the binding pocket (chemical complementarity). Scoring is employed to rank the order of the ligand poses with the aim to separate correct from incorrect poses [Gohlke & Klebe, 2002; Taylor et al., 2002; Kitchen et al., 2004]. It was demonstrated that docking programs are suited to solve the geometric problem of posing and near native ligand-receptor poses can be obtained. However available scoring functions suited for fast docking approaches are not able to accurately predict binding affinities, which has a direct impact on the later rank ordering [Warren et al., 2006].

Scoring functions can be subdivided into three classes, force-field based scoring functions using mainly van der Waals and electrostatic energy terms, empirical scoring functions employing experimentally determined binding affinities for parameterization of ligand receptor interactions and knowledge-based scoring functions using as well experimentally determined binding affinities to define simple atom-type interaction-pair potentials.

Ligand-receptor interactions are defined by the standard Gibb's free energy of binding $\Delta G°$ which is composed of enthalpic ($\Delta H°$) and entropic ($\Delta S°$) portions (Eq. 3.26).

$$\Delta G° = \Delta H° - T\Delta S° \qquad (3.26)$$

T refers to the absolute temperature. The limitation of currently applied scoring functions is that both terms contribute to the binding affinity and that entropic terms at both the receptor and ligand are hardly addressed (e.g. desolvation effects) [Gohlke & Klebe, 2002]. A second limitation of docking approaches is that both receptor and ligand are flexible targets. For ligands an extended conformational analysis is performed. However the receptor flexibility is only partly included by employing molecular dynamics, rotamer libraries and protein ensembles [Carlson, 2002; Teague, 2003; Kitchen et al., 2004].

Despite of that a correct preparation of both the ligand set (conformers, tautomers, stereoisomers, charges) and the receptor (charges, water molecules in the active site, quality check of the crystal structure) is of importance. If no crystal or NMR structure of the receptor is available, homology models can be created based on the structure of closely related receptors [Hillisch et al., 2004].

3.7.1 GOLD Docking

Docking calculations were performed using the program GOLD (version 2.2) [The Cambridge Crystallographic Data Centre, Cambridge, UK]. GOLD is based on a genetic algorithm optimizing the position of the ligand in the protein binding pocket, the dihedrals of the ligand rotational bonds, the ligand ring geometries and the dihedrals of the protein OH groups and NH_3^+ groups [Jones et al., 1997; Verdonk et al., 2003]. The scoring of the ligand is performed employing hydrogen bonding fitting points and hydrophobic/aromatic fitting points on both the ligand and the protein cavity. As scoring functions GOLD Score (force-field-based) and ChemScore (empirical) were available. In the present study GOLD Score was employed with default parameters. The protein was provided as PDB file whereas the binding pocket was defined by the x, y and z coordinates of the pocket centre in combination with a radius of 12 Å. The docking was performed in the highest accuracy mode. The parameters for the genetic algorithm were left at default. The early termination option was switched off and the ten best binding modes together with the score values were kept for further analyses. An example configuration file is present in Appendix F.

4 Results and Discussion

4.1 Algorithm Development

Two divisive hierarchical clustering algorithms were developed, NIPALSTREE and hierarchical *k*-means for the clustering of large data sets. NIPALSTREE projects a data set onto one dimension using PCA. The data set is sorted according to the scoring vector and split at the median position into two subset. The algorithm is applied recursively onto the subsets. The hierarchical *k*-means recursively separates a data set into two clusters using the *k*-means algorithm. A measure is introduced helping to identify a similarity threshold defining terminal clusters. The statistical evaluation of the hierarchical cluster dendrograms is characterized. Both algorithms are validated and compared to each other according to different example sets.

4.1.1 Data Sets

Four data sets were used to validate both clustering methods, Fisher's Iris data set [Fisher, 1936], two molecule data sets – COBRA [Schneider & Schneider, 2003] and MDDR [Elsevier MDL, San Leandro, USA] - and the SPECS catalogue [SPECS, Delft, The Netherlands] (see 3.1). The MDDR data set was prepared as described in section 3.1.2. For MDDR both CATS 2D and MOE 2D descriptors were calculated. For COBRA and a data set comprising COBRA, MDDR and the SPECS catalogue only MOE 2D descriptors were calculated. Descriptors were mean centred and scaled to unit variance. Finally descriptors were selected based on Shannon entropy and UFS (see section 3.3.5). The data set names, the employed UFS R^2 thresholds and the results of the descriptor selection steps for the different data sets are summarized in Table 4.1.1.

4.1.2 NIPALSTREE

Many hierarchical clustering techniques exhibit quadratic complexity [Jain et al., 1999] rendering them unsuited for large data sets. To reduce this complexity a possibility is to project a descriptor matrix onto one dimension and convert the clustering problem into a sorting problem which scales with $O(n \cdot \log \cdot n)$, with n being the number of data points (e.g. molecules). Two such projection techniques are PCA (see 3.4.3) or NLCA (see 3.4.4) [Otto, 1998]. In a preliminary experiment the MDDRMOE099 data set was reduced to 59,173 entries showing affinity to either an enzyme, a GPCR, a nuclear hormone receptor or a ligand-gated ion channel (see 3.2.5) [Schuffenhauer et al., 2002]. The first three principle components were calculated using PCA and NLCA. The obtained scoring vectors were plotted against each other. Data points were coloured according to their belonging to the four target classes. Visual inspection of the 3D representations revealed that both PCA and NLCA were capable of projecting the different target classes into mostly separate regions of the 3D map. The conclusion was that a separation of different ligand classes employing numerical descriptors and PCA is possible.

Table 4.1.1 Final data sets after descriptor pruning and similarity threshold calculation

	COBRA		MDDR				SPECS + COBRA + MDDR
Data set size	5,375		109,528				344,561
Descriptor set	MOE2D		CATS2D		MOE2D		MOE2D
Original number of descriptors	147		150		147		147
Entropy-based pruning[a]	111		45		110		111
UFS R^2 threshold	0.8	0.99	0.8	0.99	0.8	0.99	0.99
UFS based pruning	22	53	31	45	24	56	60
Final data set name	COBRA 08	COBRA 099	MDDR CATS08	MDDR CATS099	MDDR MOE08	MDDR MOE099	
Threshold Θ[b]	4.0	5.6	4.6	5.2	3.6	5.6	5.2

[a] Descriptors having a standardized Shannon entropy below 0.3 were removed. [b] Calculated similarity threshold for the clustering. The Euclidean metric was employed.

The results encouraged to further adopt PCA for the clustering algorithm as follows: a d-dimensional descriptor matrix is projected onto the first PC. Based on the scoring vector **S**, the descriptor matrix is sorted in ascending order and split at the median position. Two equally large descriptor sets – from now on termed “left” and “right” sub-matrix – are created. This is repeated for the new subsets until the maximum distance between the entries in a sub-matrix underscores a predefined similarity threshold Θ (for an estimation of Θ see 4.1.5). The algorithm was applied to MDDRMOE099. The clustering was fast, needed less than 64 megabyte (MB) random access memory (RAM) and translated into clusters which were enriched with GPCR ligands, enzymes inhibitors, nuclear hormone receptor ligands and ligand-gated ion channel ligands.

Two principal shortcomings of the method should be mentioned: one is that topological errors may occur performing the projection. Clusters which exist in d-dimensional space may be distributed over a broad data range in the first PC. These clusters would be torn apart. A second shortcoming is that splitting at the median can lead to a separation of similar entries lying directly at the median position. The relevance of the first problem can be assessed for each individual case by performing additional similarity searches. With that, related sub-matrices can be detected (see 4.2.1). The second shortcoming led to the current version of the algorithm: To build up a hierarchical dendrogram, the following steps are performed recursively on the data set:

Step 0: Define a distance threshold Θ. In the present study the Euclidian metric was employed. However all similarity metrics and coefficients listed in Table 3.2 can be used.

Step 1: Create a copy of the current descriptor matrix. The PCA projection is performed on this data matrix. The original data matrix is sorted according to the values in the scoring vector (heapsort was employed as sorting method) and the splitting procedure is initiated with the original data matrix (Figure 4.1.1).

Step 2: Generate clusters by splitting. The splitting procedure is illustrated schematically in Figure 4.1.1 B. Three clusters (black, white and grey circles) are represented by one descriptor. Splitting at the median position pulls the white cluster apart. Starting from the median position the left and right neighbouring entries are examined step-wisely to find a better split point i (left side) or j (right side). The point j is used as new split point since no cluster is separated and the resulting left and right data sets are of comparable size. The procedure is shown as algorithmic flow chart in Figure 4.1.1 A. It starts by setting a pointer to the median entry. The split point is by definition assigned to the right side. Starting from the median position in the original data matrix, the Euclidian distance to the left neighbouring entry is calculated. If the distance falls below the threshold Θ, the neighbouring entry is defined as new temporary split point. This is performed iteratively until the left end of the matrix is reached or Θ is exceeded. The procedure is initiated analogously for the right side. For both processes the number of comparisons is counted. If both stepping procedures have reached the end of the data matrix, Θ is decreased and the procedure is reinitiated. If *Θ* reaches zero no splitting is performed. If the number of comparisons for the left side is larger then for the right side, the splitting position is set to the right temporary split point and vice versa. *Θ* is set back to the original value and the left and right sub-matrices are created.

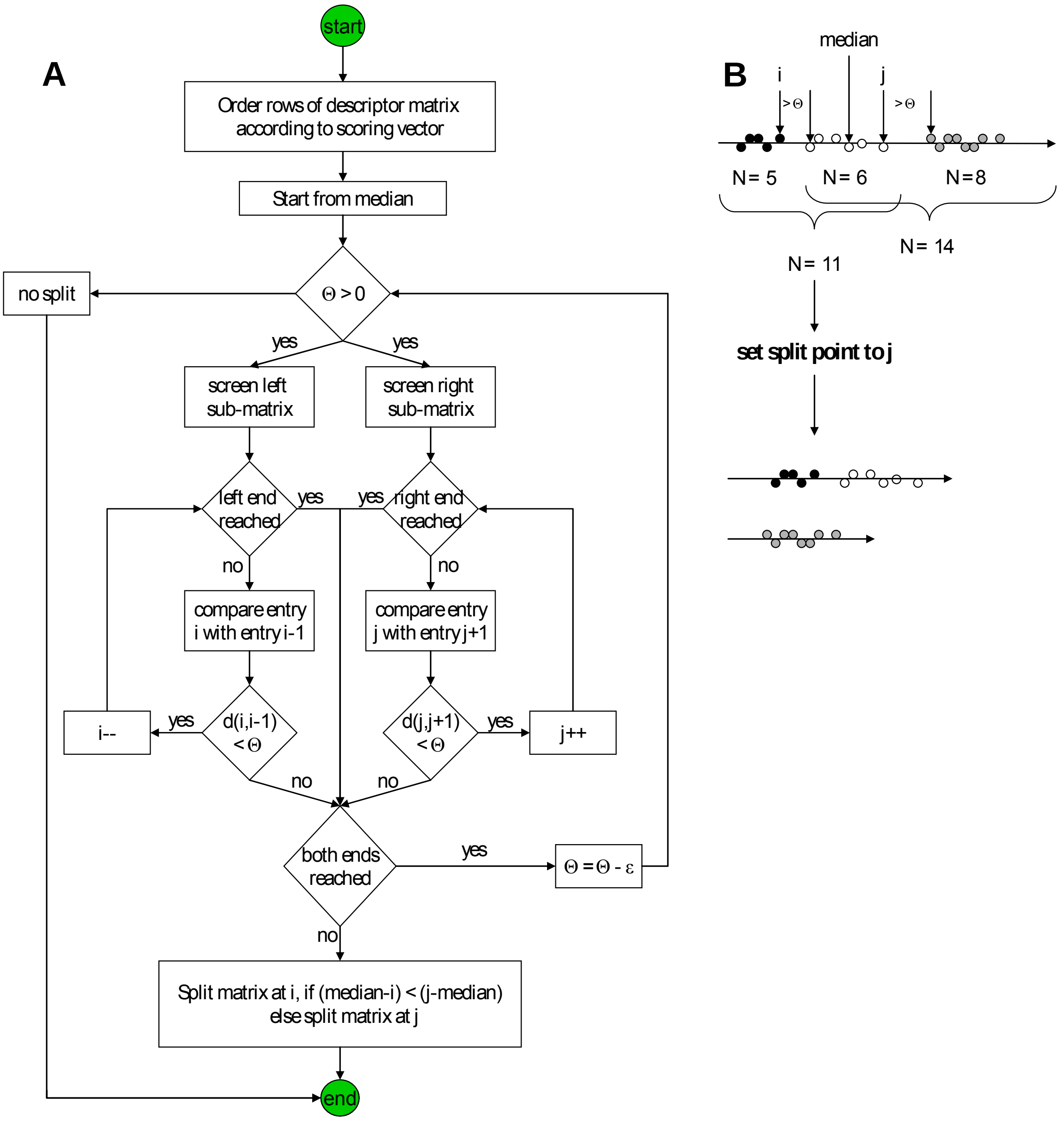


Figure 4.1.1 Concept of wandering neighbour. **A.** Algorithmic flow chart. Abbreviations: d(i,i-1) = (Euclidian) distance between molecule i and i-1, d(j,j+1) = (Euclidian) distance between molecule j and j+1. Θ = similarity threshold, ε = parameter used for systematically lowering Θ. **B.** Schematic illustration of the concept: three clusters are represented in a one-dimensional data set (black, white and grey). Splitting at the median position pulls the white cluster apart. Starting from the median position the left and right neighbouring entries are examined step-wisely to deduce a better split point i (left side) or j (right side). The point j is used as new split point since no clusters are separated and the resulting left and right data sets are of comparable size (N = 11 left and N = 8 right) compared to splitting at i (N = 14 left and N = 5 right).

Step 3: Check the maximum distance within a new subset. If the maximum (Euclidian) distance between the entries in one subset does exceed the predefined threshold Θ, the algorithm restarts with the subsets at Step 1. Otherwise no splitting of this matrix is performed.

The algorithm separates the data according to the first PC. This makes it monothetic in nature, like recursive partitioning [Rusinko et al., 1999]. However in contrast to recursive partitioning, NIPALSTREE is an unsupervised classification technique separating a data set only according to its inherent properties present in the loading vector.

The same similarity threshold Θ was used for both the wandering neighbour procedure (now referred to as Θ_2) and the termination criterion (now referred to as Θ_1). Although it might be intuitive to set both thresholds to the same value, the following pre-examinations were performed: Θ_1 was determined as termination threshold as described in 4.1.5. For this Θ_1 value the algorithm was performed using Θ_2 values within a specified data range. The number of singletons and the number of terminal clusters was calculated, normalized and plotted in one graph (Figure 4.1.2). Results show on the x-axis the Θ_2 value and on the y-axis the corresponding scaled number of singletons (blue) and clusters (magenta). For all analyzed data sets a minimum was obtained for the singleton curve. At approximately the same Θ_2 value the cluster curve converged to a minimum. These minima were observed at Θ_2 values which were similar to the determined Θ_1 values. The aim of the wandering neighbour procedure was to avoid distortion of clusters at the median position. “False” singletons should be minimized and fewer but larger and more homogeneous clusters should occur. The obtained results were in good agreement with the expectation and justify the usage of the same value for Θ_1 and Θ_2.

NIPLASTREE was able to cluster large data sets with feasible run time behaviour and space requirements. The reading, clustering and displaying of 404,148 molecules with 60 descriptors took 39 minutes on a Linux workstation employing a 3.2 GHz Intel Xeon Processor, and required less than 2 GB of memory. For a comparison, the same clustering using the hierarchical *k*-means algorithm (see 4.1.4) was 6 times faster and employed 20% less memory.

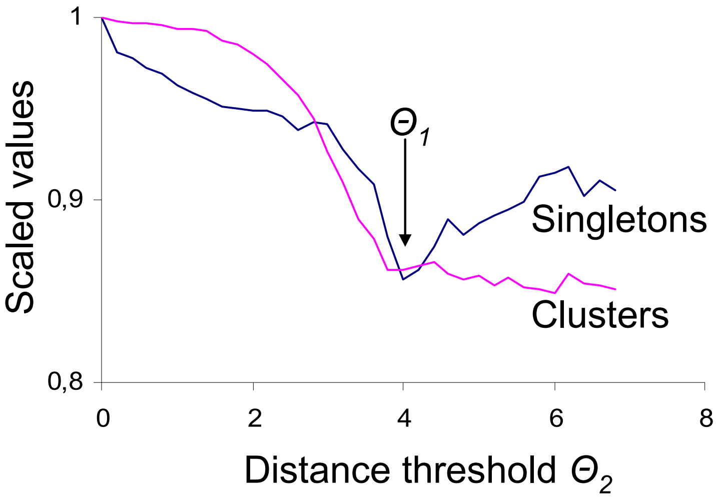


Figure 4.1.2 Identification of a value for the wandering threshold Θ_2. Clustering was performed with NIPALSTREE and COBRA08. Θ_1 was set to 4.0.

4.1.3 Outlier PCA

A possible deficiency of NIPALSTREE is that by the projection of the descriptor matrix onto one dimension clusters in *d*-dimensional space can be distributed over a broad data range in the first PC. To circumvent this putative deficiency, a variant of NIPALSTREE, Outlier PCA, was implemented as a non-hierarchical clustering method. It employs the property of PCA to project outliers to the left and right end of the scoring vectors [Oprea & Gottfries, 2001]:

Step 0: Define a distance threshold Θ and one of the similarity metrics or coefficients listed in Table 3.2.

Step 1: Create a copy of the current descriptor matrix. The PCA projection is performed on this data matrix. The original data matrix is sorted according to the values in the scoring vector and the splitting procedure is initiated with the original data matrix.

Step 2: Generate clusters by splitting. The splitting procedure starts by setting a pointer to the first and last entry of the data matrix. The distance is calculated between the entry at the pointer position and the neighbouring entry in the descriptor matrix. If the distance underscores Θ, the neighbouring entry is defined as new temporary split point. This is performed iteratively until the opposite end of the matrix is reached or the threshold Θ is exceeded.

Step 3: Check for each split point: if both opposite ends have been reached clustering is ready. Otherwise the entries from the beginning of the matrix to the left temporary split point and the entries from the end of the matrix to the right temporary split point define new clusters. The part amid defines the new sub-matrix. The algorithm restarts with the sub-matrix at Step 1.

The advantage of the algorithm is that the space requirement scales linear with the number of data points (*N*). However the time complexity scales quadratic. It renders the algorithm unfeasible for large data sets. A second disadvantage of outlier PCA is that no hierarchy is created between the data points. This complicates display and navigation in the clusters. The algorithm is based on the assumption that at every iteration step clusters are present at the ends of the scoring vector. This needs not be the case.

4.1.4 Hierarchical k-means

The *k*-means algorithm represents a non-hierarchical clustering technique [Jain et al., 1999]. It requires $O(k \cdot n)$ computation time and space, with *n* being the number of data points and *k* being the number of clusters. The *k*-means algorithm randomly selects *k* data points as initial cluster centroids (step 1). *k* clusters are formed by assigning each data point to its nearest centroid (step 2). New virtual centroids are calculated for each cluster (step 3). The second and third steps are iterated until a predefined number of iterations is reached or the clusters do not change anymore. Although the algorithm was shown to produce reliable results, there are several features which have to be dealt with care [Jain et al., 1999]:

(i) The number of cluster centroids *k* has to be predefined. For a large *k* the resulting clusters tend to be small and exclusive, whereas for a low *k*, clusters tend to be large and heterogeneous. The optimal choice of *k* depends on the inherent structure of the data set and the aim of the particular study.
(ii) Due to the nature of the algorithm, a hierarchical relationship between the clusters is not assigned. This can complicate later analysis.

To address this, a modified form of the *k*-means algorithm was implemented, forming *k* clusters at each level of a hierarchical dendrogram [Sultan et al., 2002; Barnard et al., 2004; Böcker et al., 2005]. The advantage is that for a hierarchical clustering no predefinition of the number of clusters is required. In contrast it has to be defined until which distance threshold molecules are treated as similar and should be fused to form one cluster. Consequently the same threshold has to be determined as for the NIPALSTREE algorithm.

The basic steps of the modified algorithm are:

> Step 0: Define *k*. For a binary dendrogram, $k = 2$. Specify a distance threshold Θ. In the present study, the Euclidian metric was employed to define "distance".
>
> Step 1: Perform data clustering employing the *k*-means algorithm. *k* child clusters are created and the data set is partitioned according to the *k*-means algorithm.
>
> Step 2: Check for each cluster: If the maximum distance between the data point exceeds the threshold Θ, repeat Step 1 for this cluster. Otherwise terminate.

It should be noticed, that the hierarchical *k*-means is a technique using a randomization step during the initialization of the centroid vectors. Identical dendrograms need not necessarily result from multiple runs on the same data set. To avoid this, approaches like a maximum dissimilarity selection of the initial cluster centroids might be used. This requires additional calculation steps increasing the time complexity. To avoid this no such pre-selection was performed. The algorithm was implemented in such a way that *k* can be defined by the user. In the presented work *k* was always set to 2.

Hierarchical *k*-means was able to cluster large data sets with feasible run time behaviour and space requirements. Compared to NIPALSTREE, the reading, clustering, and displaying of the same 404,148 molecules with 60 descriptors took only 383 seconds on a Linux workstation employing a 3.2 GHz Intel Xeon Processor.

4.1.5 The Stop Threshold Concept

Many hierarchical clustering algorithms use criteria like the homogeneity or the heterogeneity of the resulting clusters to assess the quality of the algorithm [Everitt et al., 2001]. These criteria are also used to determine, whether clusters containing similar molecules should be merged. The difficulty is that these techniques require a definition of "similarity". Both the NIPALSTREE and the hierarchical *k*-means algorithm face the same problem since the threshold Θ has to be defined as maximum allowed distance between the entries in a terminal cluster. For both clustering algorithms a method was proposed helping to find such a threshold value [Böcker et al., 2005; Böcker et al., 2006]:

For a specified distance range the number of singletons, the number of clusters, and the sum of the maximum distances in each terminal cluster, D_{max} (Eq. 4.1) are calculated, scaled to [0,1] and plotted in one graph (Figure 4.1.3). D_{max} can be interpreted as the sum of the cluster radii of the *n* terminal clusters.

$$D_{\max} = \sum_i \max_j \left(D(x_{i,j}, c_i) \right) \quad with\, 1 \le j \le N_i , \tag{4.1}$$

where D is a distance, $x_{i,j}$ represents data points that are members of the same terminal cluster i, c_i represents the centroid of cluster i, and N_i is the number of members of cluster i. For D all distance metrics or coefficients from Table 3.2 can be selected. Note that for Tanimoto, dice and cosine coefficients the complement, i.e. 1- $D(x_{i,j},c_i)$, is summed up for all j.

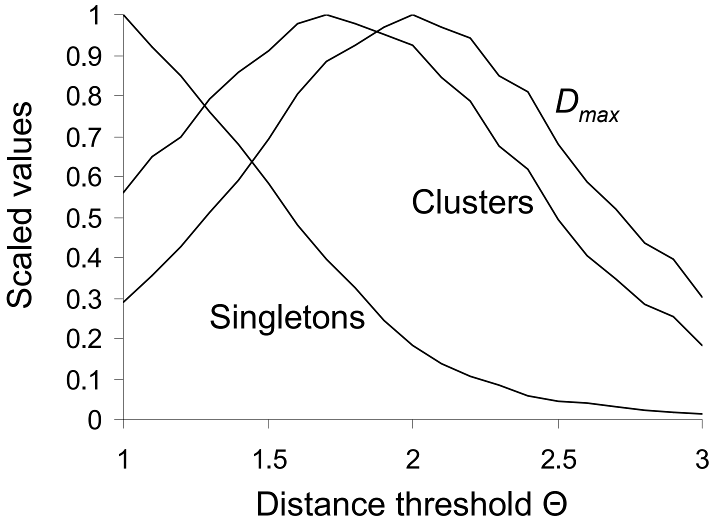


Figure 4.1.3 Identification of the distance threshold Θ. Clustering was performed using the hierarchical *k*-means, Euclidean metric, COBRA08 and different distance thresholds. The number of terminal clusters, singletons and D_{max} was calculated.

In all clustering examples shown in this study the Θ value that led to the maximal D_{max} value was used (Figure 4.1.3). Last row of Table 4.1.1 summarizes the results obtained for the different data sets. D_{max} represents a point where the data set in the dendrogram best adopts the predefined similarity threshold and a maximum dense packing of the terminal clusters is obtained. This maximum dense packing is assumed to represent the point showing a maximum in homogeneity and a minimum in heterogeneity.

4.1.6 Statistical Evaluation of Cluster Dendrograms

Cluster dendrograms can be analysed according to different perspectives. At first it can be distinguished between an unsupervised and a supervised analysis. The latter is only possible if additional biological data like % CTL, IC_{50} or K_i values have been projected onto the emerging clusters. Secondly, a cluster in the dendrogram can be analysed isolated or context-dependant. Thirdly, different parts of a dendrogram can be examined like a dendrogram level or a complete branch (that is, from the root cluster to a terminal cluster). Fourthly a dendrogram can be assessed in its entirety. Measures for all different views have been implemented and can be accessed via a GUI. Here the measures are introduced. Based on a user-defined threshold compounds with biological response values (% CTL, IC_{50} or K_i values) are assigned to the classes "hit" and "non-hit" (and "not-defined"). If not stated otherwise, these names are used for describing classes.

Cluster Evaluation

(i) Clusters are evaluated according to simple measures like the number of data points in a cluster or the percentage of class hit entries in a cluster or the percentage of class hit entries inherited from the father cluster. If more than one biological response column has been loaded (e.g. % CTL values from different HTS), the percentage overlap between the hit classes in a cluster is calculated.

(ii) A cluster is analysed according to the enrichment factor (*EF*, equation 3.10). An $EF > 1$ indicates that more compounds belonging to a class have been clustered than expected from an equal distribution. The *EF* value depends on the size of the cluster under consideration: on upper dendrogram levels, where clusters are large, *EF* values are small. On lower dendrogram levels they can get large without statistical relevance.

(iii) Clusters are evaluated according to impurity measures. The Gini impurity and entropy impurity are both measures judging the clustering of a class c with respect to all additional classes in the cluster. The entropy impurity *Entropy(c)* for class c is calculated according to equation 4.2 [Duda et al., 2001].

$$Entropy(c) = -(p(i = c)\ln(p(i = c)) + p(i \neq c)\ln(p(i \neq c))), \tag{4.2}$$

and the Gini impurity *Gini(c)* for class c is calculated according to equation 4.3.

$$Gini(c) = 1 - (p(i = c)^2 + p(i \neq c)^2), \tag{4.3}$$

with *p(i)* being the percentage of entries in a cluster belonging to a class. A value close to zero indicates for both measures a pure cluster with respect to class c.

(iv) The Pearson correlation coefficient (PCC) provides a measure analysing correlations (a relationship) between the *n* entries $x_{i,n}$ and $x_{j,n}$ of two vectors **i** and **j**. It is calculated according to equation 4.4.

$$PCC_{i,j} = \frac{1}{n}\sum^{n}\left[\frac{(x_{i,n} - \overline{x_i})}{s_i} \cdot \frac{(x_{j,n} - \overline{x_j})}{s_j}\right], \tag{4.4}$$

with $\overline{x}_i$ and $\overline{x}_j$ being the means, and s_i and s_j being the standard deviations of vectors **i** and **j**, respectively. In case of hierarchical *k*-means the PCC is calculated between the centroid vector of a cluster and the centroid vector of its parent cluster. For NIPALSTREE it is calculated between the loading vector of the cluster under consideration and the loading vector of the parent cluster. The PCC is only defined if both vectors are of same length. This is not always true, since with the NIPALSTREE algorithm in each cluster descriptors are rejected having a standard deviation < 0.0005. The PCC ranges from +1 (perfect correlation) over 0 (no correlation) to -1 (perfect inverse correlation). The NIPALSTREE algorithm separates a data set of a cluster into equally large proportions. Since PCA is used for sorting the data, the separation resembles mostly a split in the middle of the variance. Consequently the PCC is expected to be close to zero. Otherwise if the PCC is close to 1 or -1 the parent cluster and the actual cluster are expected to contain similar compounds. For the hierarchical *k*-means a (virtual) representative of a subset (the parent cluster centroid) is compared to a representative of a smaller subset within the subset (the actual cluster centroid). A correlation (PCC close to 1 or -1) between them is always expected.

(v) For the NIPALSTREE algorithm PCA calculation results in the scoring and the loading vector. Values in the loading vector define the importance of the descriptors and their direction to cumulatively explain the variance in a cluster. This capability was extended helping to identify descriptors describing the difference between the entries of the actual cluster and the entries in the brother cluster: the *d* entries in the actual loading vector are selected whose absolutes exceed a threshold ζ. That means that only descriptor weightings are employed explaining some variance. In the present study 0.1 was used for ζ. Note that by focussing only on the absolute values the direction is ignored. The second step is characterized by calculating the ratio *R* between the *d* absolute values in the loading vectors of the actual cluster and the brother cluster (Eq. 4.5).

$$R = \frac{|\,loading_{i,actual}\,|}{|\,loading_{i,brother}\,|} \quad 1 \le i \le d \ , \tag{4.5}$$

Equally weighted descriptors in both clusters have an *R* value of one and are ignored. In contrast differing descriptor weightings translate into a high or a low *R* value.

For the hierarchical *k*-means algorithm the relation *R* for *d* descriptors is calculated according to equation 4.6

$$R = \frac{|\,centroid_{i,actual} - centroid_{i,parent}\,|}{|\,centroid_{i,brother} - centroid_{i,parent}\,|} \quad 1 \le i \le d \,, \tag{4.6}$$

which represents the ratio of the change of the mean descriptor value going from the father cluster to the actual cluster compared to the corresponding change in the brother cluster. Again a minimum or a maximum *R* value indicates a large difference between the descriptor in both clusters.

Dendrogram Evaluation

A generalized view on the distribution of molecules belonging to a hit class c in the cluster dendrogram was obtained using the following analysis: for a dendrogram level the average enrichment factor (*EF*) is calculated for all *n* (*n* = number of clusters) *EF*s of class *c*, which are larger or equal to one (Eq. 4.7).

$$Average\,EF = \frac{1}{n}\sum_{i=1}^{n} EF_{i,c}; \quad EF \ge 1 \ . \tag{4.7}$$

The average *EF* is calculated for all dendrogram levels, where the number of clusters is less than or equals to the number of molecules belonging to class *c*. On higher dendrogram levels, artificially large enrichment factors can bias the average and are thus avoided [Böcker et al., 2006].

Two impurity measures offering a generalized view on the distribution of molecules belonging to a class *c,* have been implemented, SE and Kullback-Leibler distance (*KBD*) [Duda et al., 2001]. The *SE* is defined according to equation 3.4 (see 3.3.5). Here p_i represents the number of class c members c_i in a cluster divided by the total number of class members on a dendrogram level (equation 3.5). To make the *SE* independent of the number of clusters N_i

on a dendrogram level, the obtained values are normalized by the logarithm to base two of N_i (equation 3.6). A scaled *SE* close to zero represents a highly ordered distribution of class *c* members on a dendrogram level. This resembles a distinct clustering of the class.

The *KBD* is calculated according to equation 4.8.

$$KBD = \sum_i p_{obs(i)} \log_2 \frac{p_{obs(i)}}{p_{\exp(i)}}, \tag{4.8}$$

with $p_{obs(i)}$ being the observed number of class c members c_i in a cluster divided by the total number of class members on a dendrogram level. Assuming a random separation of class c in the father cluster, $p_{exp(i)}$ explains the expected number of class c members c_i in a cluster divided by the total number of class members on a dendrogram level. To obtain values independent of the number of clusters N_i, the *KBD* is standardized according to equation 4.9.

$$sKBD = \frac{KBD}{\log_2 N_i}. \tag{4.9}$$

The *KBD* tends to be high if class c is not separated into equally large proportions. A high *KBD* value indicates a dendrogram level where a separation of class c from the rest occurs. Likewise to the average *EF* the Shannon entropy and Kullback-Leibler distances are only calculated for dendrogram levels, where the number of clusters is less or equals to the number of class *c* molecules. Figure 4.1.4 shows an example of the average EF, Shannon entropy and Kullback-Leibler distance obtained for ACE inhibitors (hit) and the remainder (non-hits) of COBRA099. The hierarchical *k*-means algorithm was employed.

All three curves provide a different view on the same dendrogram. With exception of dendrogram level 2, the average *EF* shows a constant rising for class hit stepping down the dendrogram hierarchy. It indicates a systematic separation of this class from the rest. Class non-hit shows no enrichment at all. The Shannon entropy curve for class hit shows a constant decrease in impurity. It can be understood as a systematic clustering of this class in the dendrogram. From dendrogram level 2 ongoing, the *KBD* curve for class hit exhibited values above 0.1. These *KBD* values (with respect to class non-hit) indicate a separation of this class from the rest. It supports the results obtained with *SE* and average *EF*. On the uppermost dendrogram level COBRA099 was separated into clusters containing 73% (left cluster) and 27% (right cluster) of the data set. This unbalanced separation has a direct influence on both impurity measures of class non-hit. It is indicated by a low value in the Shannon entropy plot and the high value in the Kullback-Leibler plot. It cannot be seen in the average *EF* plot. The average *EF* is an intuitive and interpretable measure. However both the *SE* and *KBD* provide extra information and are thus valuable.

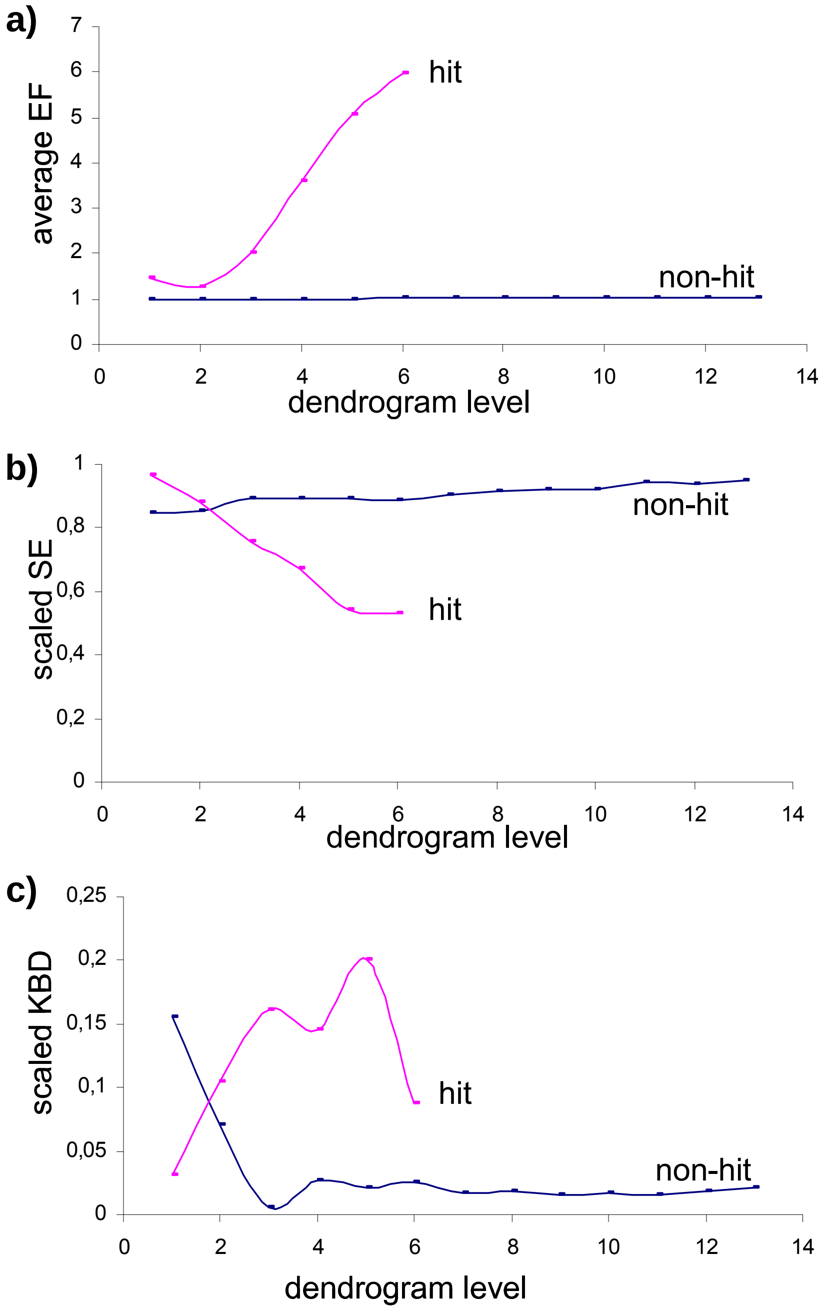


Figure 4.1.4 Average *EF* (a), scaled Shannon entropy (b) and scaled Kullback-Leibler distance (c) values obtained for ACE inhibitors (hit) and the rest (non-hits) of COBRA099 for each dendrogram level. The dendrogram was obtained with the hierarchical *k*-means.

A measure of the balance S of a dendrogram is calculated for each dendrogram level according to equation 4.10.

$$S(level) = \frac{\sum_{n} (l(n)_{obs} + 1)}{\sum_{m} (l(m)_{\exp} + 1)}, \tag{4.10}$$

with n being the number of non terminal clusters in the dendrogram until the dendrogram level is reached and m being the number of non terminal clusters of an optimally split dendrogram. *l(n)* or *l(m)* defines the shortest internal path lengths of clusters n or m to the root cluster. A value of one resembles a balanced dendrogram whereas a value near zero resembles a linear connected list. Figure 4.1.5 shows an example of S values obtained for COBRA099 in combination with the hierarchical k-means (blue curve) or NIPALSTREE (magenta curve). The x-axis represents the dendrogram level and the y-axis the corresponding S value. On level six the dendrogram obtained with the hierarchical k-means algorithm gets unbalanced. A same observation was made for the NIPALSTREE algorithm on level eight.

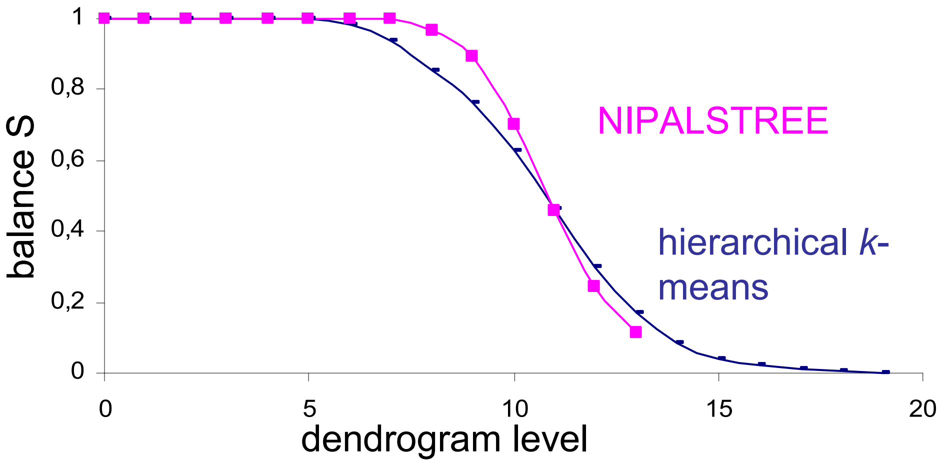


Figure 4.1.5 S values obtained for different dendrogram levels employing COBRA099 and the hierarchical k-means algorithm (blue curve) or the NIPALSTREE algorithm (magenta curve).

4.1.7 Application of NIPALSTREE

NIPALSTREE was applied to Fisher's Iris data set. A Θ value of 1.12 was used and was determined as described in 4.1.5. Figure 4.1.6 shows the projection of the data set on the first two principal components (A). The calculated hierarchical cluster dendrogram is shown in B.

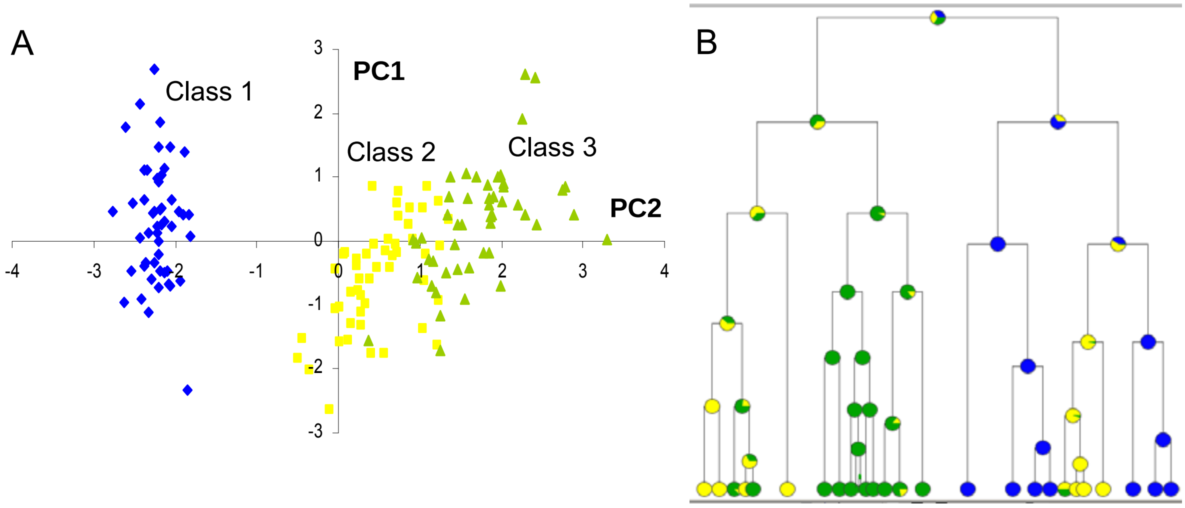


Figure 4.1.6 Clustering of Fisher's Iris data set. A: Score plot of the data according to the first two principle components (explained variance > 95%). The different data classes are coloured in blue (1), yellow (2) and green (3). B: Binary dendrogram (Θ = 1.12) obtained with the NIPALSTREE algorithm. Each dendrogram cluster is represented by a pie chart showing its relative class composition.

Data points in A were coloured according to their belonging to classes 1 (blue), 2 (yellow) or 3 (green). In the binary dendrogram each cluster is shown as a pie chart showing its relative class composition. Obviously the dendrogram representation is in agreement with the PC-projection (note that individual horizontal branches have to be rotated by 180° in the dendrogram). Class 1 occupies a distinct region in the PC plot. In the dendrogram pure class 1 containing branches were observed from dendrogram level three ongoing. Class 2 and class 3 show both overlapping and pure regions in both plots. As can be seen in the Figure 4.1.6 B on the first dendrogram level a proportion of class 2 entries was assigned to the right side. On dendrogram level two these entries were separated from class 1. This observation shows a disadvantage of the algorithm: the separation of a data set into two equally large partitions can force local data densities to be torn apart. However, on subsequent dendrogram levels, these data result in pure but smaller clusters.

The NIPALSTREE algorithm was applied to COBRA08 and COBRA099. As one example, angiotensin converting enzyme (ACE) inhibitors were selected. ACE, a zinc dependent metalloprotease, plays a central role in the angiotensin-renin system. ACE cleaves the decapeptide angiotensin I into the vasopressor angiotensin II. ACE inhibitors have been used for treatment of cardiovascular diseases, including high blood pressure, heart failure and kidney failure [Acharya et al., 2003].

COBRA contained 48 molecules categorized as ACE inhibitors. They can be grouped into four structural classes (class representatives are shown in Figure 4.1.7, **17-20**) and a few "outliers", which cannot easily be assigned to any of the classes (four example structures are shown in Figure 4.1.7, **21-24**). Seven molecules of class 1 (represented by **17**), eleven molecules of class 2 (represented by **18**), twelve molecules of class 3 (represented by **19**) and six molecules of class 4 (represented by **20**) were present. Both descriptor versions of COBRA were clustered using the NIPALSTREE algorithm. Figure 4.1.8 shows the average *EF*s for each dendrogram level, both COBRA versions and the ACE inhibitors and non-ACE inhibitors. As expected the non-ACE inhibitors showed an average *EF* of one (i.e. no enrichment) in the dendrogram. In contrast ACE inhibitors were characterized by constantly increasing average *EFs* for both COBRA versions on deeper dendrogram levels. Going from the uppermost cluster to clusters located on deeper levels, ACE inhibitors were enriched employing the clustering algorithm and both descriptor sets.

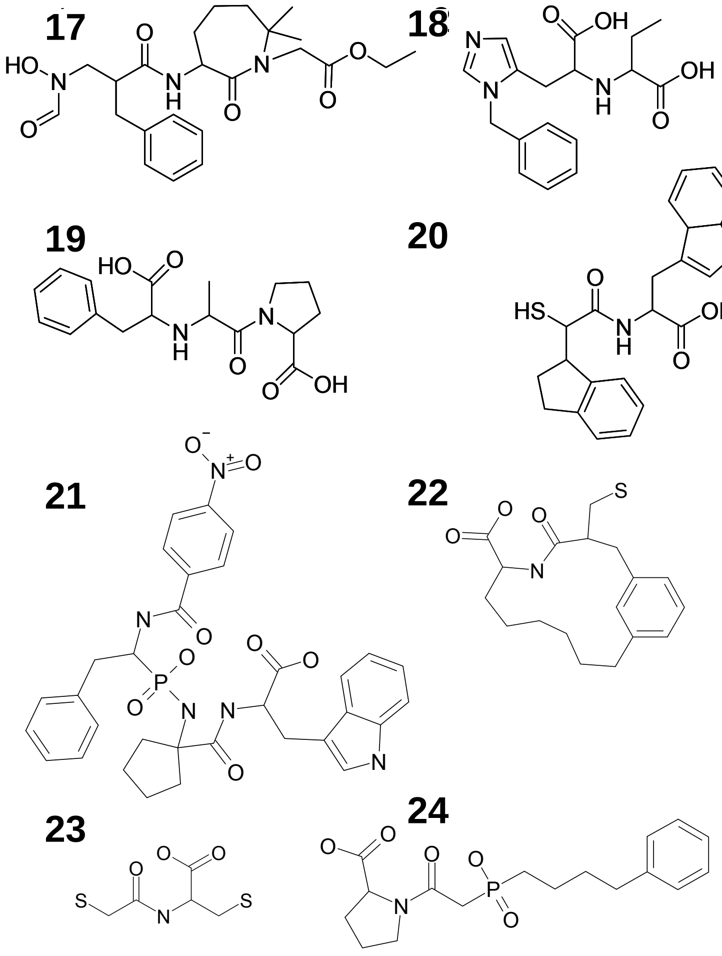


Figure 4.1.7 Four class representatives (**17-20**) and four examples of outliers (**21-24**) of ACE inhibitors in the COBRA data set.

The separation of the four ACE inhibitor classes (Figure 4.1.7, represented by **17-20**) in the dendrograms was analysed in detail. On the root level "outliers" were assigned to the right sub-dendrogram. The four classes were completely assigned to the left sub-dendrogram. On dendrogram level four employing COBRA099 and level five employing COBRA08 class 2 was the first one that was separated from the other classes. Class 2 molecules differ from the other classes in having an imidazole substructure element. An amide moiety is completely missing. The loading vectors of the class 2 containing cluster and its brother cluster were analysed as described in 4.1.6 (Eq. 4.5) The aim was to identify discriminating descriptors between class 2 and the other ACE classes.

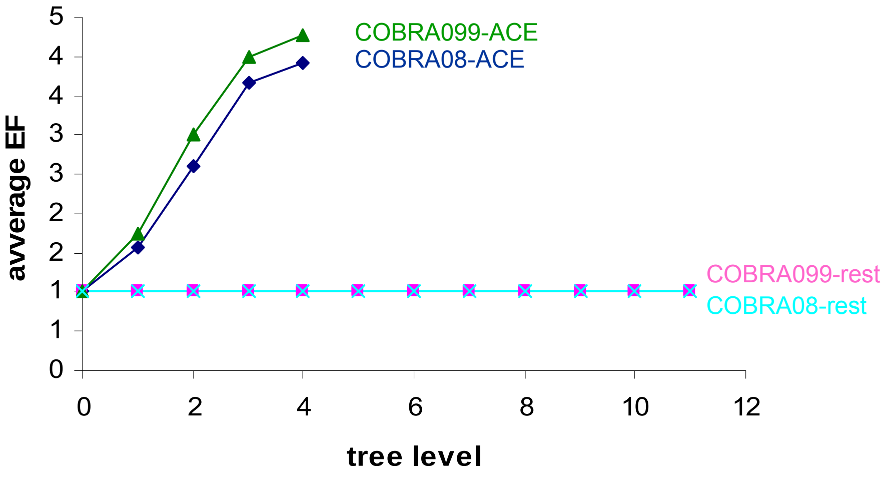


Figure 4.1.8 Average enrichment factors for each dendrogram level obtained for ACE inhibitors and non-ACE inhibitors (rest) of COBRA08 and COBRA099. Green: COBRA099 ACE inhibitors; dark blue: COBRA08 ACE inhibitors; magenta: COBRA099 non-ACE inhibitors; light blue: COBRA08 non-ACE inhibitors.

For COBRA08 high *R* values were present for PEOE_VSA-0, PEOE_VSA-4 and SMR_VSA1 and for COBRA099 for PEOE_VSA-4 and SlogP_VSA1. The corresponding original descriptor values of class 2 and classes 1, 3 and 4 were analysed and the resulting means and standard deviations are shown in Table 4.1.2. Marked differences were obtained for class 2 compared to the other ACE classes. It shows that by analysing loading vectors in class separating dendrogram clusters important descriptor can be identified.

Table 4.1.2 Mean descriptor values and standard deviations obtained for ACE inhibitors of class 2 and classes 1, 3 and 4.

	PEOE_VSA-0 [Å^2]	PEOE_VSO-4 [Å^2]	SlogP_VSA1 [Å^2]	SMR_VSA1 [Å^2]
Class 2	32.9 ± 15.1	5.68	35.7 ± 3.3	18.7
Class 1, 3, 4	51.8 ± 15.9	0	21.6 ± 4.8	3.49 ± 2.9

The terminal clusters obtained with both descriptor sets were examined in detail. Since here the interest lies on separating structural classes, clusters were judged according to the number of identified class members and the enrichment factors of the corresponding classes. The results are summarized in Table 4.1.3

Table 4.1.3 Terminal clusters enriched with ACE inhibitors of COBRA099 or COBRA08.

Data set	ACE class in cluster	Number of class entries in cluster	Cluster size	EF of ACE class for the cluster
COBRA099	2	4	5	391
COBRA099	2	4	9	217
COBRA099	3	4	6	299
COBRA099	3	4	5	358
COBRA099	1	3	5	461
COBRA08	3	5	7	320
COBRA08	2	4	7	279
COBRA08	3	3	8	168

The terminal clusters obtained for COBRA099 showed two clusters enriched with class 3, one cluster enriched with class 1, and two clusters enriched with class 2 ACE inhibitors. Class 4 ACE inhibitors occurred mainly as singletons. The terminal clusters obtained for COBRA08 showed two clusters enriched with class 3 and one cluster enriched with class 2 ACE inhibitors. Class 1 and class 4 inhibitors occurred mainly as singletons. This separation of classes of similar molecules reflects a characteristic of the NIPALSTREE algorithm, which is a consequence of keeping the use of internal memory as low as possible: *via* the projection of a *d*-dimensional space onto one dimension mapping errors occur. As a consequence, closely related molecules appear as singletons in different parts of the created dendrogram. The problem can be fixed by performing additional similarity searches around the terminal clusters. Any metric listed in Table 3.2 can be employed for that.

The results show that a clustering of ACE inhibitors can be obtained employing the NIPALSTREE algorithm. Although differences exist, clustering both descriptor sets led to a comparable separation of the molecules in the dendrogram. No clear descriptor preference can be given. For the terminal clusters in COBRA099 which were enriched with ACE inhibitors, a co-clustering of six additional protease inhibitors and five molecules binding to other receptor classes was observed. For COBRA08, seven additional protease inhibitors were found and five molecules binding to other receptors. Since protease inhibitors usually share a peptidomimetic backbone, the co-clustering of other protease inhibitors was expected. It shows the applicability of the clustering method to targeted library design.

As mentioned in 4.1.3 a variant of the NIPALSTREE algorithm was implemented: outlier PCA. COBRA099 was clustered with the algorithm. Only clusters were analysed containing at least three ACE inhibitors. Three clusters were identified. Cluster one contained five ACE inhibitors of class 2, cluster two contained four compounds of class 3 and cluster three contained three molecules of class 1. The remaining ACE inhibitors occurred mainly as singletons. Outlier PCA shows (as negative aspects) a squared run time and does not create a hierarchical relationship between the clusters. This complicates later analyses. For the ACE inhibitors outlier PCA did not result in superior clustering of the ACE classes with respect to the corresponding NIPALSTREE analysis (Table 4.1.3). A reason for the observed lower validity might be the false assumption that clusters in a data set are projected to the left and right end of the first scoring vector employing the algorithm. According to the results outlier PCA was deprioritized for further analyses.

4.1.8 Application of Hierarchical *k*-means

Similar to the NIPLASTREE algorithm in a first experiment, Fisher's Iris data set was used to test the performance of the hierarchical *k*-means algorithm. Figure 4.1.9 a shows the projection of the data on the first two principal components. The binary *k*-means dendrogram is shown in b (no termination criterion was specified). It is evident that the dendrogram representation is in agreement with the PC-projection. Three distinct classes are shown which occupy different branches of the dendrogram. Three entries of class 2 were assigned to the right side of the dendrogram which is dominated by class 1 examples. These points are indicated by little arrows in Figure 4.1.9. The observation reveals a disadvantage of the algorithm: the number of clusters on a dendrogram level, *k*, forces the data space to be split into *k* sub-regions on each dendrogram level. Local data densities lying at an interface region between two such sub-regions bear the danger of being torn apart. As can be seen in Figure

4.1.9 b, after the third split of the Iris data set, the three "outliers" form a pure cluster. The outcome of this preliminary experiment is promising, since although data points were assigned to the "wrong" side of the dendrogram, in the end pure but smaller clusters were obtained. It should be kept in mind that an optimum solution exists for such a problem. When dealing with large data sets only algorithms can be used which try to reach the optimum. When comparing the dendrogram obtained with the hierarchical *k*-means to the dendrogram obtained with NIPALSTREE, it is evident that the hierarchical *k*-means algorithm shows a more pronounced separation of the three classes. Especially the separation of class 3 from the other two classes shows a clear advantage of the hierarchical *k*-means algorithm: in a cluster the data set needs not to be split into equally large subsets. Outlying clusters can be separated early from the rest of the data set.

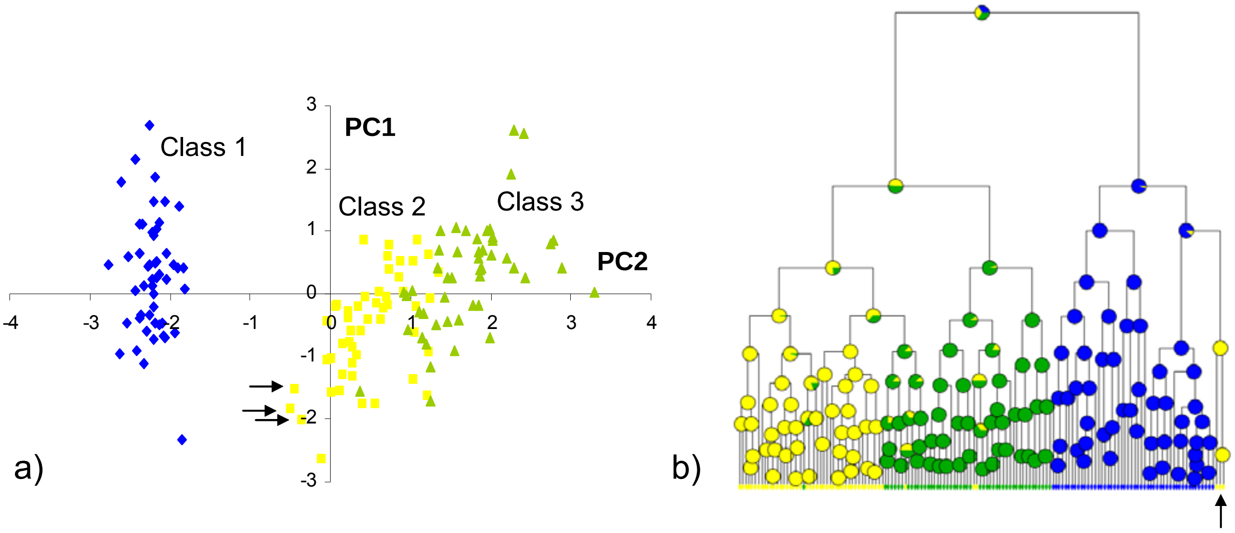


Figure 4.1.9 Clustering of Fisher's Iris data set. a) Score plot of the data according to the first two principle components (explained variance > 95%). The different data classes are coloured in blue (1), yellow (2) and green (3). b) Binary dendrogram ($k = 2$, $\Theta = 0$). Each dendrogram cluster is represented by a pie chart showing its relative class composition. Arrows indicate the location of three class 2 data points in the PC-plot (a) and the *k*-means dendrogram (b).

The hierarchical *k*-means clustering was applied to MDDR and COBRA. Two examples were selected to be discussed in more detail, (i) caspase 1 (interleukin 1 cleaving enzyme; ICE, EC number 3.4.22.36) inhibitors [Talanian et al., 2000; Braddock & Quinn, 2004] from COBRA, and (ii) glucocorticoid receptor ligands [Brody et al., 1998; Norman et al., 2004] from MDDR.

ICE inhibitors prevent IL1 cleavage, which plays a major role in a wide range of inflammatory and autoimmune diseases, like rheumatoid arthritis, osteoarthritis, chronic obstructive pulmonary disease, and asthma. ICE belongs to the family of cysteine proteases and specifically cleaves Asp116-Ala117 and Asp27-Gly28. Inhibitors typically mimic this residue motif. In COBRA 39 ICE inhibitors were present.

Glucocorticoid receptors bind glucocorticoids and induce gene transcription. This leads to catabolic reactions in extrahepatic tissues, anabolic reactions in the liver, immune-suppressive reactions in the lymphatic system and under stress to elevated cortisol levels having inflammation blocking effects. Due to the various functions of glucocorticoids, drugs that bind to glucocorticoid receptors have implications in a lot of therapeutic areas, e.g. rheumatic disease or allergic reactions [Brody et al., 1998]. In the MDDR 91 glucocorticoid receptor ligands were present.

COBRA Clustering (ICE Inhibitors).

Both versions of COBRA (COBRA08 and COBRA099, Table 4.1) were clustered using the hierarchical *k*-means algorithm. With the exception of one outlier, for both data sets all ICE inhibitors (N = 39) were assigned to one side of the dendrogram on the first dendrogram level. On subsequent dendrogram levels the inhibitors were separated into individual branches. Comparing the emerging clusters obtained with the different descriptor sets, different groupings of ICE inhibitors were observed. This is expected since the larger descriptor set (COBRA099) should emphasize properties in a different way than the smaller one. The average *EF*, *SE* and *KBD* diagrams for ICE inhibitors in COBRA08 (ICE_08) and COBRA099 (ICE_099) were analysed (Figure 4.1.10).

The average *EF* curve for ICE_099 shows constantly increasing average enrichment factors. An exception was dendrogram level three. For ICE_08 this rising was only observed for the first three dendrogram levels. When analysing the ICE curves obtained by *SE* or *KBD* and both COBRA versions, opposite results were obtained. Here both descriptor sets performed equally well on dendrogram level one indicated by a high value in the *KBD* diagram and a low value in the *SE* diagram. On subsequent dendrogram levels the *SE* values for ICE_08 were lower and the *KBD* values were higher with respect to corresponding values obtained for ICE_099. It indicates a purer and thus superior clustering for ICE_08. This opposite behaviour of *SE* and *KBD* compared to the average *EF* can be explained by analysing the cluster sizes on the different dendrogram levels. For COBRA099 on the first dendrogram level both ICE inhibitors and the complete data set showed an unbalanced separation. In contrast for COBRA08 only ICE inhibitors showed this unbalanced separation. It translated into a lower average *EF* value.

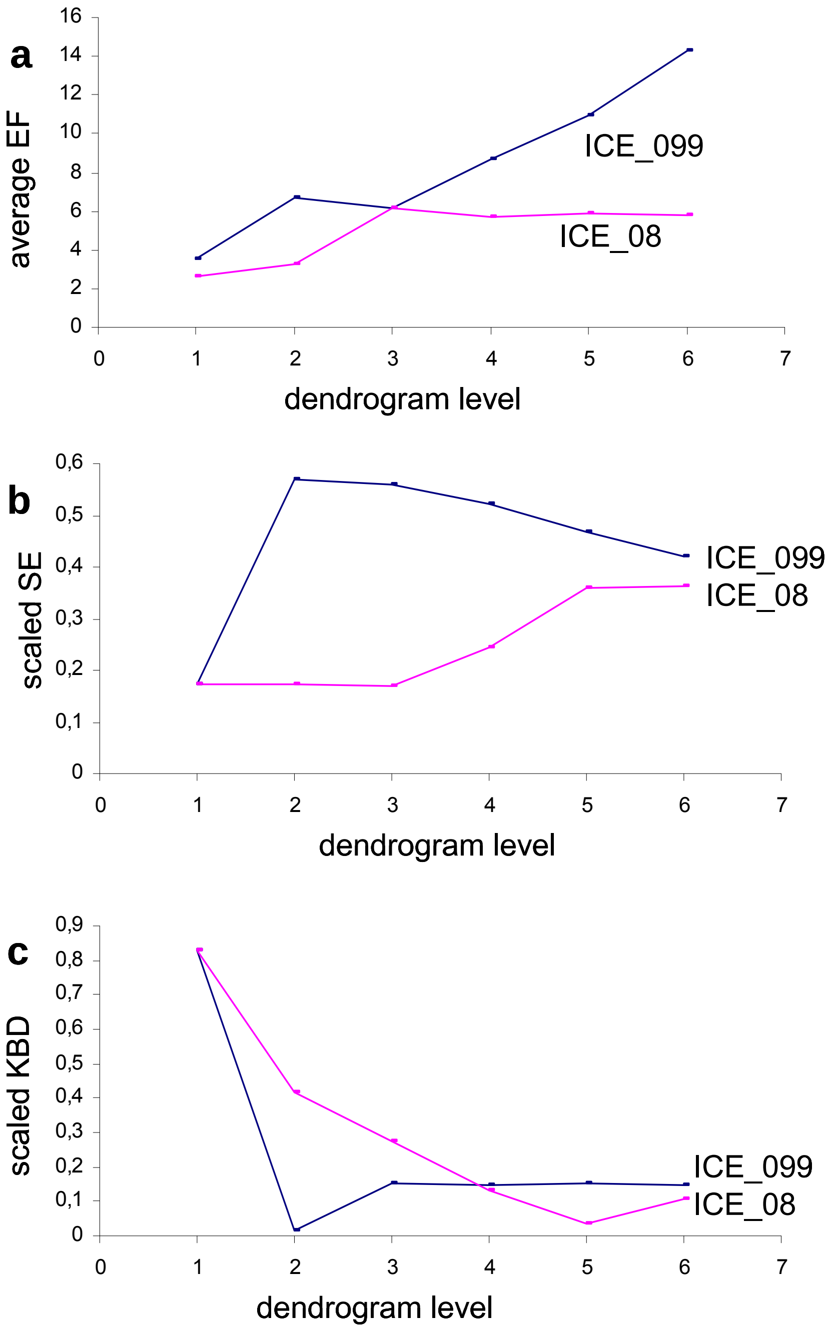


Figure 4.1.10 Average *EF* (a), *SE* (b) and *KBD* distance (c) for dendrograms obtained for ICE inhibitors in COBRA099 (ICE_099) and COBRA08 (ICE_08). The hierarchical *k*-means algorithm was employed.

Both *SE* and *KBD* are based on the number of cluster and the analysed class. The overall data set size has no influence. Consequently no difference can be seen on the first dendrogram level for ICE_08 or ICE_099. On dendrogram level one ICE inhibitors in the COBRA099 dendrogram were separated into comparably large partitions. For COBRA08 again an unbalanced split of ICE inhibitors occurred. This separation is not visible in the average *EF* values. Contradictory both *SE* and *KBD* account for this purer clustering and give valuable additional information. *SE*, *KBD* and average *EF* try to monitor the distribution of a class on a dendrogram level in one value. This simplification allows only draw rough conclusions. The different points of views are necessary to get a deeper understanding.

Focussing on the results obtained for the COBRA08 set, three large clusters emerged containing mainly ICE inhibitors. Cluster one consisted of 15 entries, with nine ICE Inhibitors (EF = 82.7), cluster two comprised 12 entries, with six ICE inhibitors (68.9) and cluster three consisted of 14 entries, with nine ICE Inhibitors (EF = 88.6). The molecules in the three clusters, which were not defined as ICE inhibitors, fall into two classes. Class one consists of other protease inhibitors, like Matrix metalloproteinase inhibitors [Baker et al., 2002], human rhinovirus 3C protease inhibitors [Johnson et al., 2002] or Hepatitis C Virus NS3 protease inhibitors [Goudreau et al., 2004]. This is not surprising, since small organic protease inhibitors try to mimic peptide-sequences [Böhm et al., 2002]. These peptide sequences can be in turn similar to each other. Class two consists of $\alpha_4\beta_1$ intregrin (also known as very late antigen-4, VLA-4) antagonists, which have potential for the treatment of allergic disease like asthma and other chronic inflammatory diseases [Lin et al., 2004]. Although both ICE inhibitors and VLA-4 antagonists play a role in treatment of allergic diseases no interconnection is known to the author. Regarding the ICE inhibitors in the three clusters, two out of the three clusters were composed of structurally similar molecules. The third cluster contained compounds less related to the first two clusters. Representative structures are shown in Figure 4.1.11. Small individual clusters were also observed with only one or two ICE inhibitors. These might have resulted from unsuitable cluster boundaries or be a consequence of shortcomings of the chosen descriptor set.

The representatives of the three large clusters (A, B and C) contain mutual substructure elements that were found in identical or only slightly different form in all other structures of the clusters. A peptidic moiety (Figure 4.1.11, yellow) represents a substructure motif that is present in all ICE inhibitors: a modified aspartic acid, alanine, valine, and a peptide bond. Clusters A and B are closer related to each other, which can be explained by the shared ethyl-phenol group (Figure 4.1.11, magenta). They differ in the occurrence of an acetamide group in cluster A (Figure 4.1.11, green) and a propyl-benzene group in cluster B (Figure 4.1.11, green). The more distant cluster C accounts for two unique substructures, a toluene group and a ring closure connecting the alanine and valine residues (Figure 4.1.11, green). Results show that the distribution of ICE inhibitors in the dendrogram can be interpreted from a structural perspective.

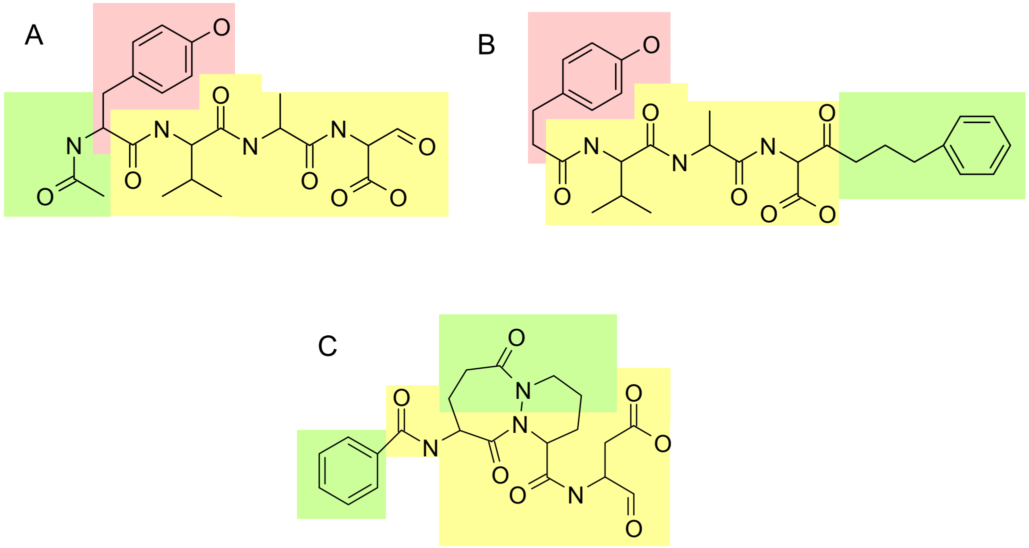


Figure 4.1.11 Representative ICE inhibitors from the three main emerging clusters (A, B, C) obtained for COBRA08 using the hierarchical *k*-means. Clusters A and B contain closely related structures. Yellow: common substructure motif in all three clusters; magenta: common motif in clusters A and B; green: unique motifs.

MDDR Clustering (Glucocorticoid Receptor Ligands).

The MDDR was clustered employing the hierarchical *k*-means algorithm ($k = 2$) in combination with the four descriptor sets listed in Table 4.1.1 and the corresponding calculated stop thresholds. Comparing the different descriptor sets according to their capability to separate glucocorticoid receptor ligands from the rest of MDDR, the MDDRCATS08 descriptor set yielded the best results at the root levels. In total 90% of the Glucocorticoid receptor ligands were assigned to the right side. Glucocorticoid receptor ligands in the MDDR can be divided into three main “lead” classes: 23% class I, 66% class II, and 11% class III (see Figure 4.1.12 A). Judging the different descriptor sets according to their capability to separate the different lead classes, both CATS 2D descriptor sets separated class I from classes II and III on the root level. In contrast, the MDDRMOE099 set separated class III from classes I and II and the MDDRMOE08 descriptor set showed no clear separation at all. It highlights the impact of different descriptor sets on the separation. The terminal clusters containing more than two glucocorticoid receptor ligands were selected form the MDDRCATS08 dendrogram. In total 270 additional molecules were identified. Of these molecules 86.3% have been described as ligands of other nuclear hormone receptors or as ligands being involved in the synthesis of steroid hormones. The remaining 37 molecules could only be assigned to a wide variety of inhibitor classes.

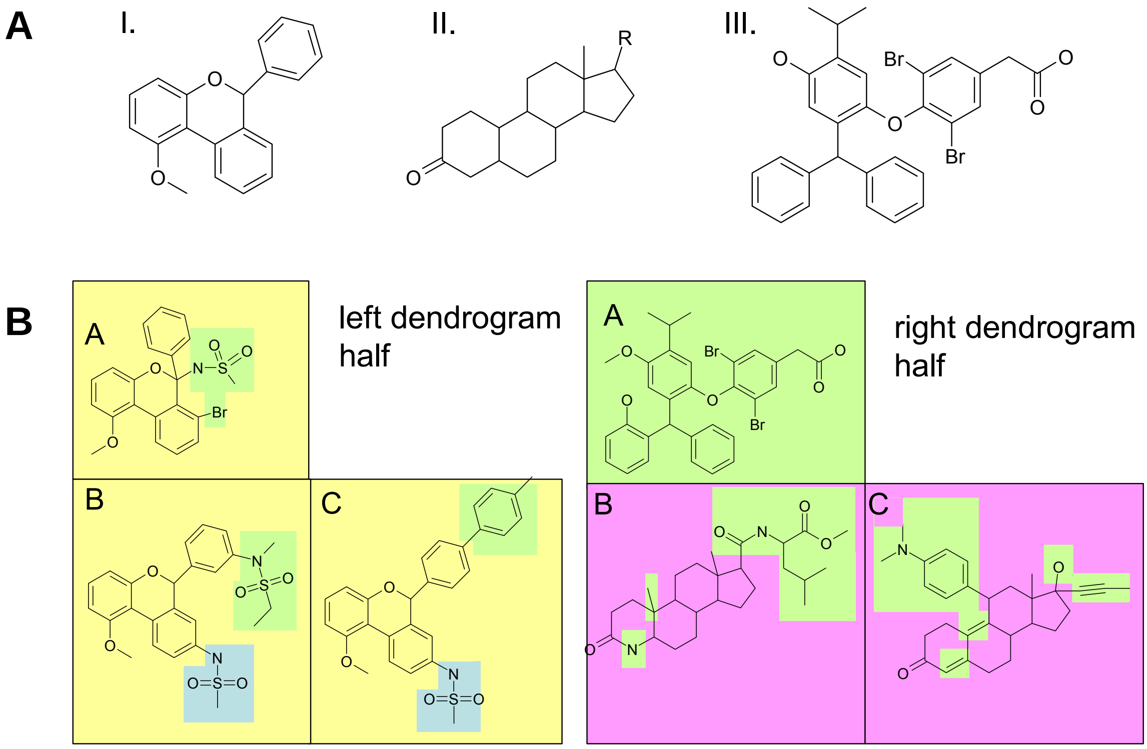


Figure 4.1.12 **A**: The three core structures (I. – III.) of glucocorticoid receptor ligands present in the MDDR database. **B**: Representative glucocorticoid receptor ligands in the three largest terminal clusters (A, B, C) in the left and right sub-dendrogram of dendrogram level 1. Results are shown for MDDRCATS08. On both sides clusters B and C lie in closer proximity to each other compared to cluster A. Yellow: common motif in all three clusters on the left side; blue: common motif in clusters B and C on the left side; magenta: common motif in cluster B and C on the right side; green: unique motives.

The different descriptor sets resulted in a comparable MDDR clustering. For all descriptor sets on the left and right side (from the root) three large clusters appeared containing glucocorticoid receptor ligands. Figure 4.1.12 B shows as an example the results, obtained with the MDDRCATS08 descriptor set. The structures in Figure 4.1.12 B were representatives of the clusters (three on each side of the cluster dendrogram viewed from the root). Cluster "importance" was rated according to the enrichment factor of either class I, II or III (Figure 4.1.12 A), and not according to the enrichment factor of glucocorticoid receptor ligands in general. The enrichment factors in the three class I clusters (Figure 4.1.12 B left dendrogram half) were 2,818, 2,817 and 2,192, whereas enrichment factors in the two class II clusters (Figure 4.1.12 B right dendrogram half, clusters B and C) were 493 and 401 and the enrichment factor of the class III cluster was 5,917 (Figure 4.1.12 B right dendrogram half, cluster A). The compounds were grouped according to their relationship in the dendrogram, so that in both dendrogram halves clusters B and C were in closer proximity to each other compared to cluster A. The three different classes were separated. Class I was only present on the left side. Common motifs in the three clusters are indicated by yellow-coloured substructures. Green colour shows unique substructures across all clusters. Blue fragments represent a common substructure of the related clusters B and C on the left dendrogram half. In clusters B and C of the right dendrogram half class II showed the classical steroid backbone (magenta colour) whereas class III dominated cluster A (Figure 4.1.12). Despite of the separation of the three classes across clusters differently substituted core structures were recognized and grouped. The results show that the different structural classes were separated from each other and enriched in terminal clusters. The clustering mirrors the structural relationship of glucocorticoid receptor ligands and confirms the results obtained for ICE inhibitors in COBRA.

Virtual Screening

To further evaluate the hierarchical *k*-means algorithm, a virtual screening application was designed: The dendrogram can be generated employing a reference data set (e.g., MDDR, COBRA). By projecting new molecules on the dendrogram their potential activity might be predicted by analyzing the co-clustering of new molecules with known actives in terminal clusters. This can also be performed simultaneously by reading-in compounds for which the pharmacological activity is unknown together with reference compounds. It provides a quick and easy way to find new compounds being putatively active. One such study was performed to illustrate the idea: The combined data set of MDDR, COBRA, and the SPECS catalogue was used to build up a binary dendrogram using the hierarchical *k*-means (Table 4.1.1). All known caspase-1 inhibitors from COBRA (N = 39) were marked and the terminal clusters containing these molecules were screened for co-located MDDR caspase-1 inhibitors (N = 188). A challenging question is to look for “scaffold hops” within a cluster. One such pair is given by structures **25** (COBRA) [Edwards, 2003] and **26** (MDDR) [Hagmann et al., 1994] which were grouped together (chart 4.1). Both are known caspase-1 inhibitors with different scaffolds. In this particular cluster, only four molecules were co-located (one from COBRA, three from MDDR). All of them are known cysteine protease inhibitors. Compounds **27** [Cameron et al., 1997] and **28** [Guo et al., 2001] represent the other two molecules from MDDR (chart 4.1). They are both cathepsin L inhibitors [Turk et al., 2000] and share the peptide-like backbone part with the caspase-1 inhibitors. This example demonstrates a possible use of the hierarchical *k*-means for constructing focused screening libraries.

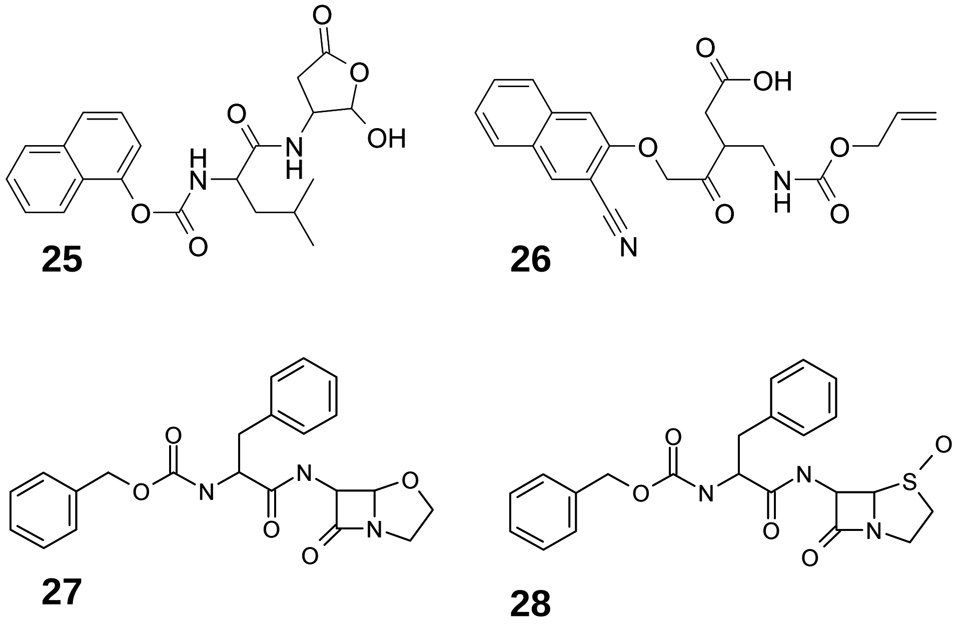


Chart 4.1 Cysteine protease inhibitors

Summarizing, all three clustering examples demonstrated a meaningful automatic grouping of chemical structures by the hierarchical *k*-means approach. The algorithm is feasible for large data sets. The hierarchical nature of the cluster relationships provides a possibility to find SAR in the different clusters if activity data is present.

4.1.9 Comparison of Both Algorithms by Virtual Screening

It was shown that both, NIPALSTREE and hierarchical *k*-means perform a meaningful clustering of large data sets. According to this the question is arising whether one of the algorithms can be given a favour. To answer the question and examine the usefulness of both algorithms in the context of virtual screening the combined data set of the SPECS catalogue, COBRA and MDDR was employed (Table 4.1.1). The SPECS catalogue was engaged to increase the number of molecules for which the activity is unknown. These molecules were treated as “inactive”. With the combined data the hierarchical clustering was performed using both algorithms. The five inhibitor classes listed in Table 4.1.4 were analysed and used as compound labels.

Table 4.1.4 SPECS_COBRA_MDDR: sizes of inhibitor/ligand classes.

Label	N (COBRA)	N (MDDR)
ACE inhibitor	48	494
COX inhibitor	149	1,556
Adrenoceptor ligand	200	542
GABA receptor ligand	85	478
Glucocorticoid receptor ligand	18	91

The quality of both clustering algorithms was evaluated and compared to each other by examining for each dendrogram level what percentage of MDDR entries with a certain label has been co-clustered with COBRA entries bearing the same label. By summing up the cluster sizes of the examined clusters, the screened percentage of the original data set was derived. Enrichment curves for each label were created by plotting the percentage of retrieved MDDR entries bearing the label against the percentage of the screened data set for each dendrogram level. Figure 4.1.13 shows the enrichment curves obtained with the hierarchical *k*-means (Figure 4.1.13 a) and NIPALSTREE (Figure 4.1.13 b) for ACE inhibitors (dark blue curves), COX inhibitors (magenta curves), adrenoceptor ligands (yellow curves), GABA receptor ligands (light blue curves) and glucocorticoid receptor ligands (purple curves). All curves show a steep rising in the lower percentage range (deeper dendrogram levels) for all labels. The markers in the curves correspond to dendrogram levels, with the right-most point being the root of the dendrogram.

To compare both clustering algorithms dendrogram level 11 was selected. On this level the screened data size of each selected screening application was above 3,500 compounds which is a suitable size for further filtering steps, ordering and experimentally testing of the molecules. Enrichment factors were calculated for both algorithms and the disjunctions and conjunctions of the result lists of both algorithms. Table 4.1.5 shows the obtained enrichment factors. In parenthesis the number of retrieved MDDR entries is shown, interacting with the examined receptor class.

Table 4.5 Enrichment factors obtained with the clustering algorithms on dendrogram level 11.

[a]	ACE (N = 494)	COX (N = 1,556)	Adrenoceptor (N= 542)	Glucocorticoid receptor (N = 91)	GABA-receptor (N = 478)
Hierarchical *k*-means[b]	31.2 (N = 246)	11 (N = 761)	8.99 (N = 298)	27.3 (N = 27)	7.97 (N = 110)
NIPALSTREE[b]	16.2 (N = 188)	6.16 (N = 625)	6.14 (N = 270)	18.7 (N = 17)	3.51 (N = 84)
Hierarchical *k*-means + NIPALSTREE disjunction[b]	17.6 (N = 306)	6.42 (N = 980)	5.93 (N = 394)	15.6 (N = 30)	4.86 (N = 165)
Hierarchical *k*-means + NIPALSTREE conjunction[b]	54 (N = 128)	22.6 (N = 406)	16.3 (N = 174)	98 (N = 14)	7.77 (N = 29)

[a] In parenthesis the total number of MDDR ligands in the data set is shown.

[b] In parenthesis the number of MDDR ligands is shown retrieved with the corresponding method.

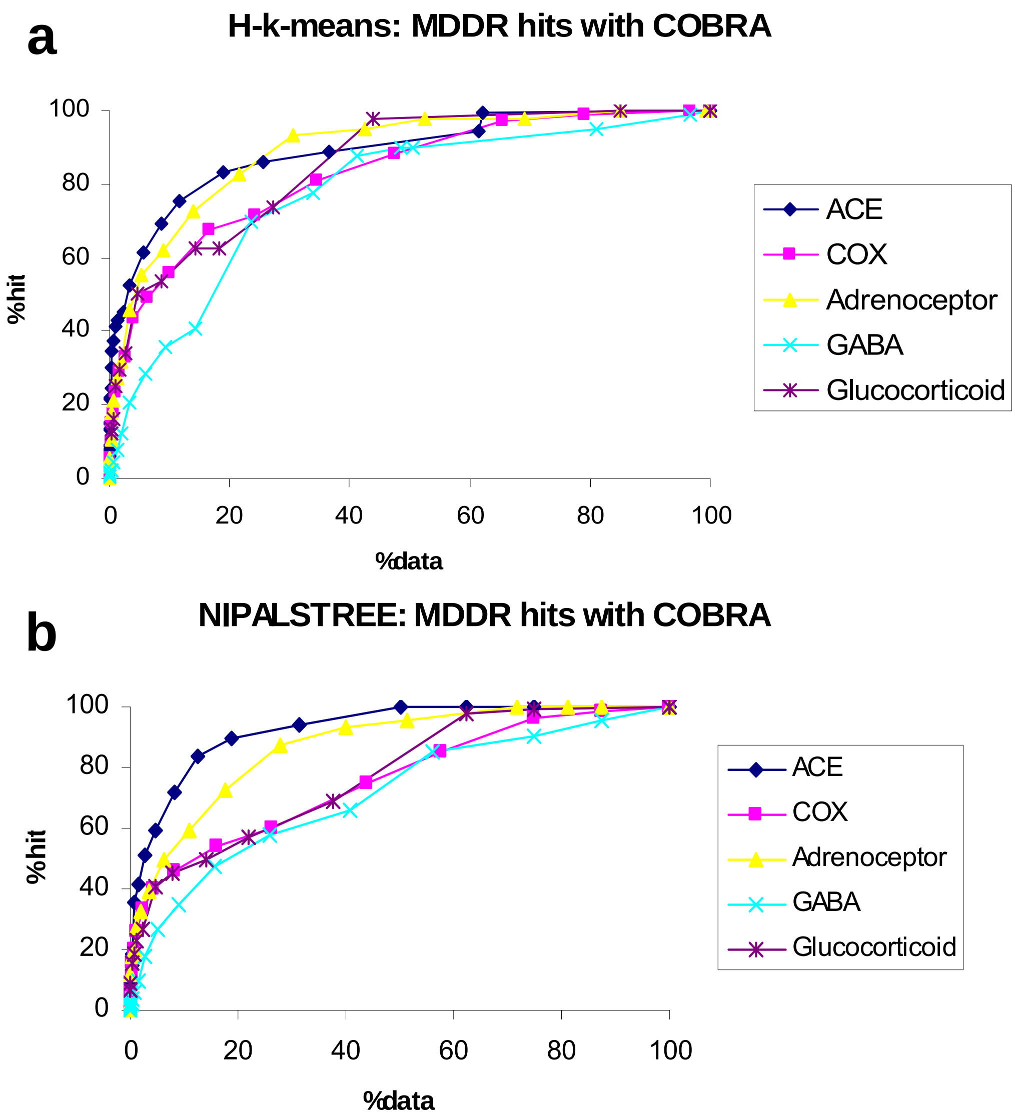


Figure 4.1.13 Clustering of the SPECS_COBRA_MDDR data set using NIPALSTREE (a) and hierarchical *k*-means (b) and specific labels (see colour panels). On a dendrogram level the number of data points in clusters containing COBRA entries with a specific label, is translated into the percentage of virtually screened compounds. The number of co-clustered MDDR entries having the same label, is converted into the percentage of retrieved hits. Points in the diagram correspond to dendrogram levels.

All calculated *EF*s were higher for the hierarchical *k*-means algorithm compared to the corresponding *EF*s from NIPALSTREE. The result shows that for the listed examples the hierarchical *k*-means algorithms outperformed the NIPALSTREE algorithm. With the exception of GABA receptor ligands, the conjunctive combination of both algorithms translated into two-fold higher enrichment factors. Reducing the data space by intersecting the results of the two algorithms can have the consequences of loosing structural classes. To test how many structural classes get lost by combining both algorithms the distribution of ACE inhibitors on level 11 was re-analysed. The employed ACE inhibitor sets are presented in Table 4.1.6.

To identify the structural classes present in the data sets, a phylogenetic-like tree clustering analysis was performed (see 3.4.5). This program is a hierarchical cluster analysis tool extracting maximum common substructures of a given data set. The resulting classes are assumed to correspond closest to lead structure classes. Table 4.1.6 shows, that after combining results of both algorithms (set 3), 15 ClassPharmer classes do no longer occur with respect to set 1 (hierarchical *k*-means) or five classes with respect to set 2 (NIPALSTREE). To estimate whether the not occurring classes represent a loss of real "lead" classes, Daylight Fingerprints were calculated for all MCS of the classes of set 1, 2 and 3. For each rejected MCS of set 1 or 2 the most similar MCS in set 3 was identified using Tanimoto similarity calculations.

Table 4.1.6: MDDR ACE inhibitor sets

Data set [a]	Final data set name	Number of MDDR ACE Inhibitors	ClassPharmer Classes
Hierarchical *k*-means	Set 1	246	39
NIPALSTREE	Set 2	188	29
Hierarchical *k*-means + NIPALSTREE conjunction	Set 3	128	24

[a] Data sets contain only MDDR ACE inhibitors found with both algorithms on dendrogram level eleven

Three representative nearest neighbour pairs are highlighted in Figure 4.1.14. All structures contain the common theme of a carboxylic acid and an amide group separated by an aliphatic carbon atom. Differences exist in the adjacent moieties. The left side of the structure pairs represents a non-matching class of set 1 (hierarchical *k*-means) and the right side the most similar class of set 3. Figure 4.1.14 A shows the case where only minor differences exist between the structural pairs. This occurred for nine of the non-matching classes from data set 1 and three from set 2. Figure 4.1.14 B exemplifies the case where the core "lead" structure with the common theme is still equal in both structure pairs, but the rest groups show differences. This was observed for five of the not matching classes of set 1 and for two from set 2. Figure 4.1.14 C shows the only example where it is assumed that a structural class is lost, since on the left side the hydantoine core structure is replaced by a tetrahydroazepinone on the right side. This structural arrangement is present in four structures of set 2.

In summary the results demonstrate that the conjunctive application of both algorithms reduces the number of compounds to be screened while the number of actual hits is diminished to a lower extent. On dendrogram level eleven of the dendrogram for ACE inhibitors, only one "lead" class is lost. This effect may be different for other ligand classes. However, current results let assume, that both algorithms are likely to produce overlapping clusters and that this is not the case.

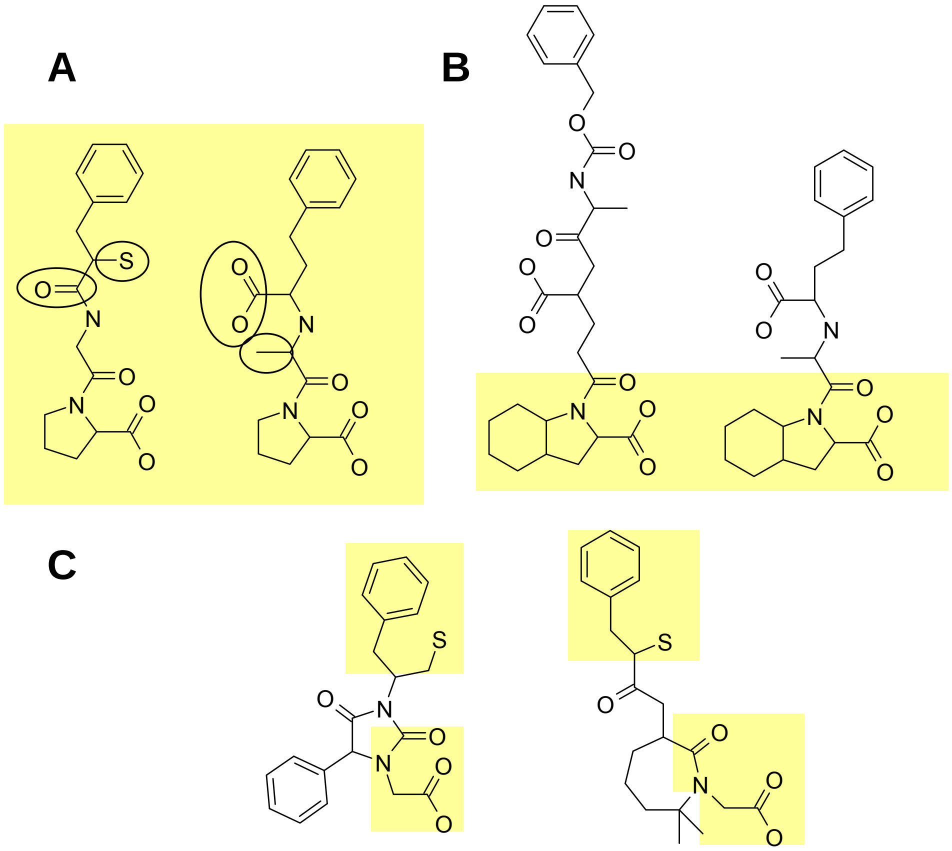


Figure 4.1.14 Example of three most similar MCS pairs obtained by ClassPharmer analysis for set 3 (left side of each structure pair) and set 1 (right side of each structure pair). Similarity was determined by calculating Tanimoto coefficients employing Daylight Fingerprints. Similar structural motifs are highlighted in yellow.

4.1.10 Conclusions

A new hierarchical clustering algorithm, NIPALSTREE, was developed having the capacity to cluster large data sets. The hierarchical *k*-means algorithm was adopted for the same purpose. Both algorithms were able to cluster 400,000 compounds with 60 descriptors in less than one hour on a 3.2 GHz Intel Xeon processor. The memory consumption was below 2 GB. It let to the assumption that both algorithms are capable of clustering several million data points using 64 bit processors in a few days.

The D_{max} calculation was introduced helping identify a stop threshold for the clustering. The maximum D_{max} value is assumed to achieve a packing of maximum density in the terminal clusters. It represents a compromise between maximizing the homogeneity and minimizing the heterogeneity. This theoretical consideration was confirmed for the examined ACE inhibitors, ICE inhibitors and glucocorticoid receptor ligands since both a grouping of structurally similar ligands and a separation from the remaining data set was obtained in terminal clusters.

Both clustering methods were validated and compared to each other according to several examples. Using different descriptor sets both algorithms showed structurally meaningful groupings and were able to enrich different ligand classes in the created dendrograms. Both a coarse-grained and fine-grained view on the data is obtained. The hierarchical *k*-means algorithm seemed to outperform NIPALSTREE in that higher enrichment factors were obtained for a set of different ligand classes. The superiority of the first algorithm might be explained by its polythetic nature compared to the monothetic nature of the latter algorithm. The conjunctive application of both clustering algorithms was shown to improve the *EFs* without loosing a large proportion of the different lead classes present in the data. Further the NIPALSTREE algorithm provides the loading vector for every cluster. It allows direct drawing of conclusions of the importance of different descriptors in a cluster. First insights into SAR of a ligand class can be gained. Consequently, no clear preference can be given for any of the algorithms. A comparison of both clustering algorithms to any other hierarchical clustering algorithm is still missing. It is likely that more calculation intense methods like Wards' clustering or the recently introduced sequential supraparamagnetic clustering [Ott et al., 2004] will provide a grouping of higher quality. These methods exhibit at least quadratic time complexity and space requirement and are not applicable to large data sets.

Several measures were adopted helping to derive structure activity relationships in the dendrogram. By calculating the average enrichment factor, Shannon entropy and Kullback-Leibler distance for the distribution of a class of inhibitors on a dendrogram level, the distribution is judged by one value. By plotting these values for every level in one graph it was possible to judge the clustering of the inhibitors in the dendrogram. Each of the three measures offered a different but complementing view on the separation. A global point of view on the SAR of a class of ligands is obtained. The calculation of the ratio between the values of the loading vector in a cluster and its brother cluster revealed important descriptors distinguishing a hit class from their structurally related non-hits. Although other methods exist identifying such descriptors it shows that the calculated dendrograms provide information to draw conclusions about the SAR in the data.

4.2 SAR Analyses in the Cluster Dendrogram

The strategy of the present work was to cluster a data set and then employ the emerging clusters in combination with biological response values to elucidate SAR in the data. The first section (4.1) has described two hierarchical clustering algorithms allowing clustering of large data sets with up to a few million data points. When working with large data sets, a manual analysis is no longer possible. Instead a guided graphical navigation through chemical space is necessary. Publications addressing this problem exist in computational chemistry and the main strategy was to create integrated program packages allowing to perform a multitude of SAR analyses in one graphical user interface (GUI) [Oellien et al., 2005; Liu et al., 2005; Wild & Blankley, 1999; Kibbey & Calvet, 2005; Gedeck & Willett, 2001; Meyer & Cook, 2000]. Unfortunately these packages have been designed to display clustering results of several thousand molecules and not for millions. According to this a new GUI had to be developed to display the results and perform SAR analyses with the aim to help identifying clusters enriched with actives, singletons or false-negatives and false-positives. For an illustration COBRA099 was employed. A focus was laid on the protease inhibitor classes listed in Table 4.2.1.

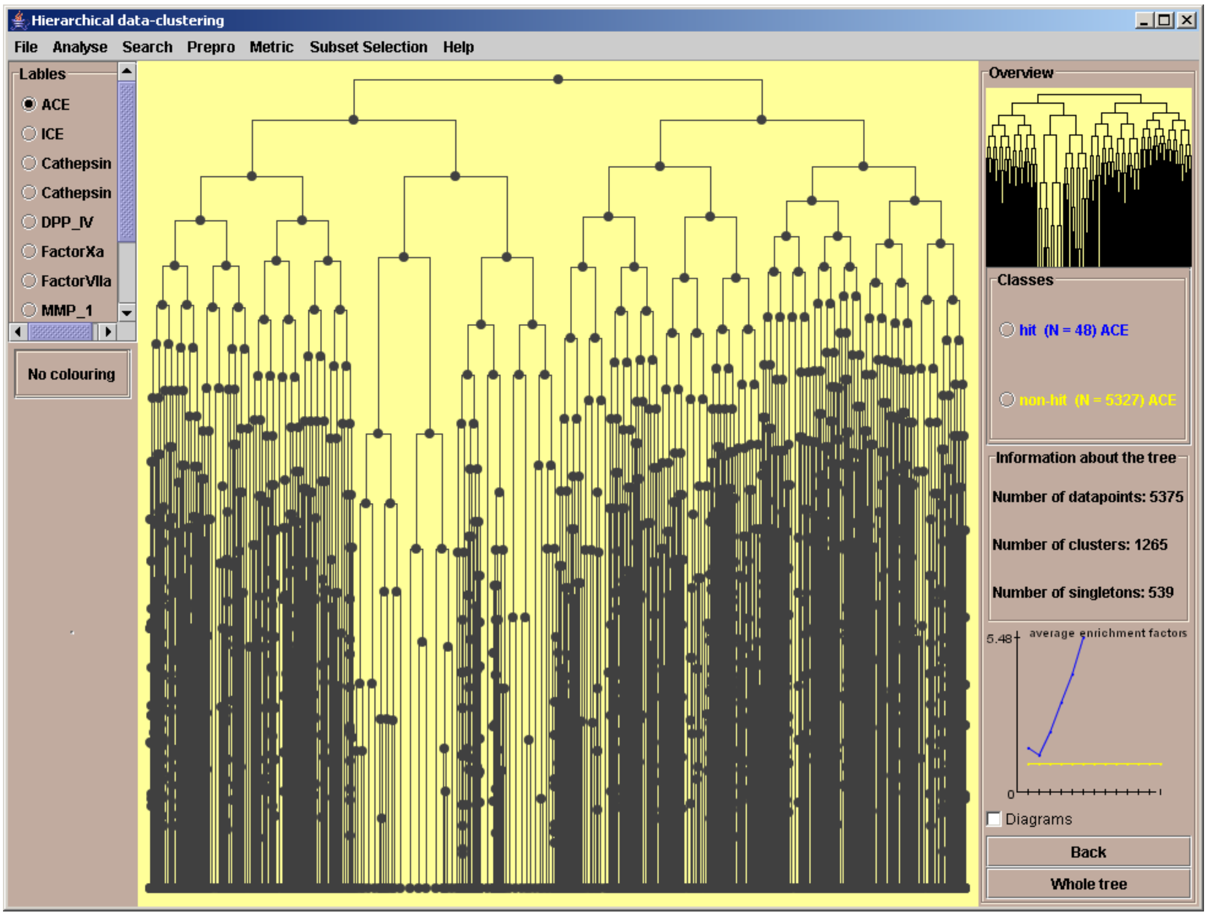
Table 4.2.1 Analyzed protease inhibitors in COBRA099

Protease enzyme	Number of inhibitors	hit rate [%]	Protease type
ACE	48	0.89	Metallo Protease
Collagenase 1 (MMP 1)	24	0.45	Metallo Protease
Cathepsin D	5	0.09	Aspartic Protease
HIV protease	62	1.15	Aspartic Protease
Cathepsin K	24	0.45	Cysteine Protease
ICE	39	0.73	Cysteine Protease
Dipeptidyl peptidase 4 (DPP IV)	25	0.47	Serine Protease
Urokinase	48	0.89	Serine Protease
Thrombin	195	3.63	Serine Protease
Factor VIIa	34	0.63	Serine Protease
Factor Xa	226	4.2	Serine Protease

For each inhibitor class artificial % CTL values were created. Compounds belonging to the inhibitor classes were assigned to a low % CTL value and compounds not belonging to the inhibitor class to a high % CTL value. 4.2.1 introduces the GUI and the navigation in the dendrogram starting from the complete dendrogram over clusters to the molecules in a cluster. 4.2.2 and 4.2.3 explain two types of a guided navigation in the data present in the dendrogram.

4.2.1 General Analysis

COBRA099 was clustered using the hierarchical *k*-means algorithm and the termination threshold specified in Table 4.1.1 (see 4.1.1). The resulting binary cluster dendrogram was reported to an interactive GUI (Figure 4.2.1).



● ACE
○ ICE
○ Cathepsin
○ Cathepsin
○ DPP_IV
○ FactorXa
○ FactorVIIa
○ MMP_1

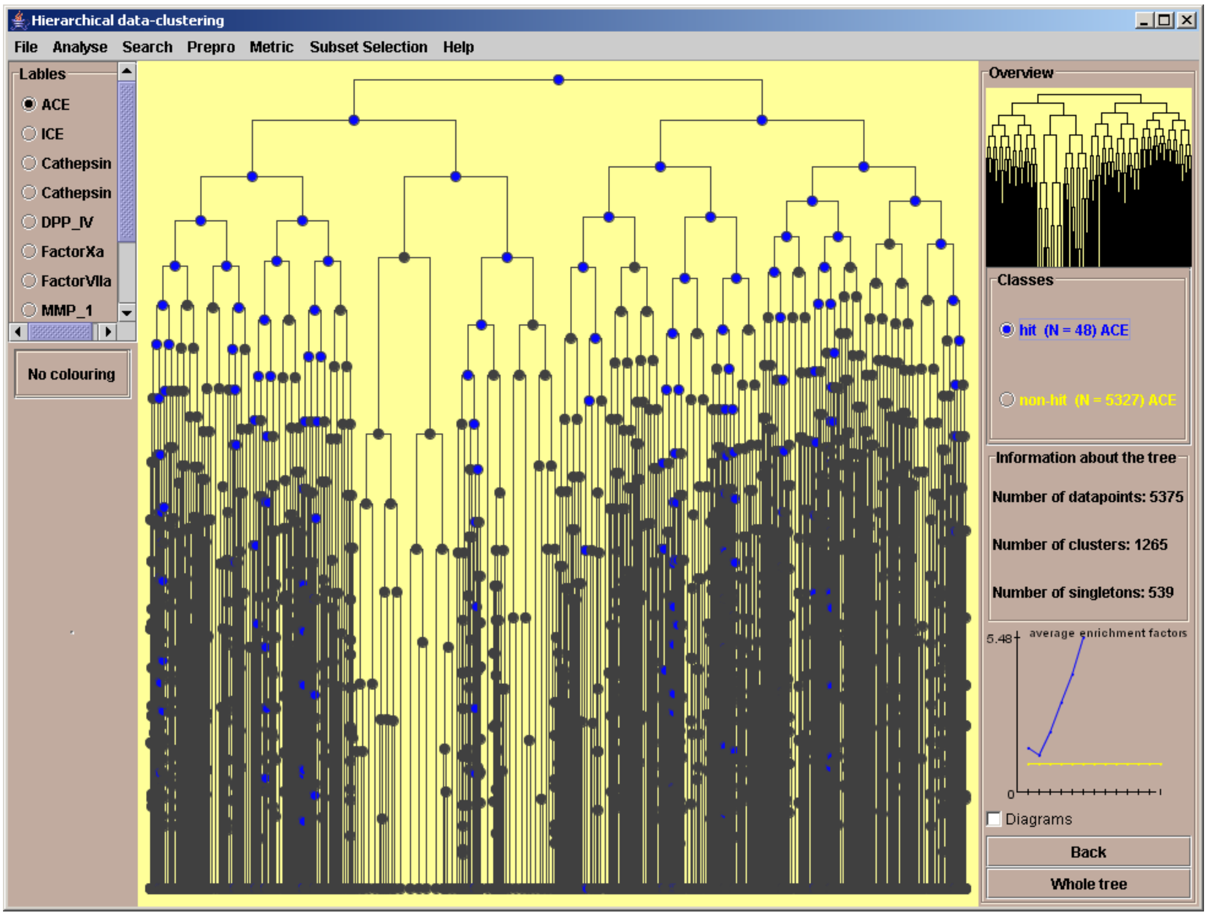
○ hit (N = 48) ACE
○ non-hit (N = 5327) ACE

Number of datapoints: 5375
Number of clusters: 1265
Number of singletons: 539

☐ Diagrams

Figure 4.2.1 Graphical user interface appearing after clustering of COBRA099 employing the hierarchical *k*-means algorithm. Additionally artificial % CTL values of the protease targets in Table 4.2.1 have been loaded.

The GUI is separated into three parts, a main window showing the dendrogram and two side panels on the right and left side. The dendrogram is directed from the uppermost cluster to terminal clusters on the opposite side. The dendrogram clusters are drawn as black circles whereas black lines indicate connections between the clusters. Using the hierarchical *k*-means algorithm the vertical connections were scaled to the diversity of cluster centroids in the sub-dendrogram. In the case of NIPALSTREE they were scaled to the maximum depth of the sub-dendrogram. In the left panel, buttons show the titles of the incorporated (HTS) assays. The button for ACE inhibitors has been activated. Based on a user-defined % CTL threshold the entries were assigned to the two classes “hit” and ”non-hit”. The “classes” field on the right side shows buttons for the ACE classes “hit” (blue) and “non-hit” (yellow). Additionally, the numbers of entries belonging to a class are present in parenthesis. By activating a button in the classes field all clusters in the dendrogram containing at least one member of the class were highlighted with the corresponding colour (Figure 4.2.2).



Lables
- ● ACE
- ○ ICE
- ○ Cathepsin
- ○ Cathepsin
- ○ DPP_IV
- ○ FactorXa
- ○ FactorVIIa
- ○ MMP_1

Classes
- ● hit (N = 48) ACE
- ○ non-hit (N = 5327) ACE

Information about the tree

Number of datapoints: 5375

Number of clusters: 1265

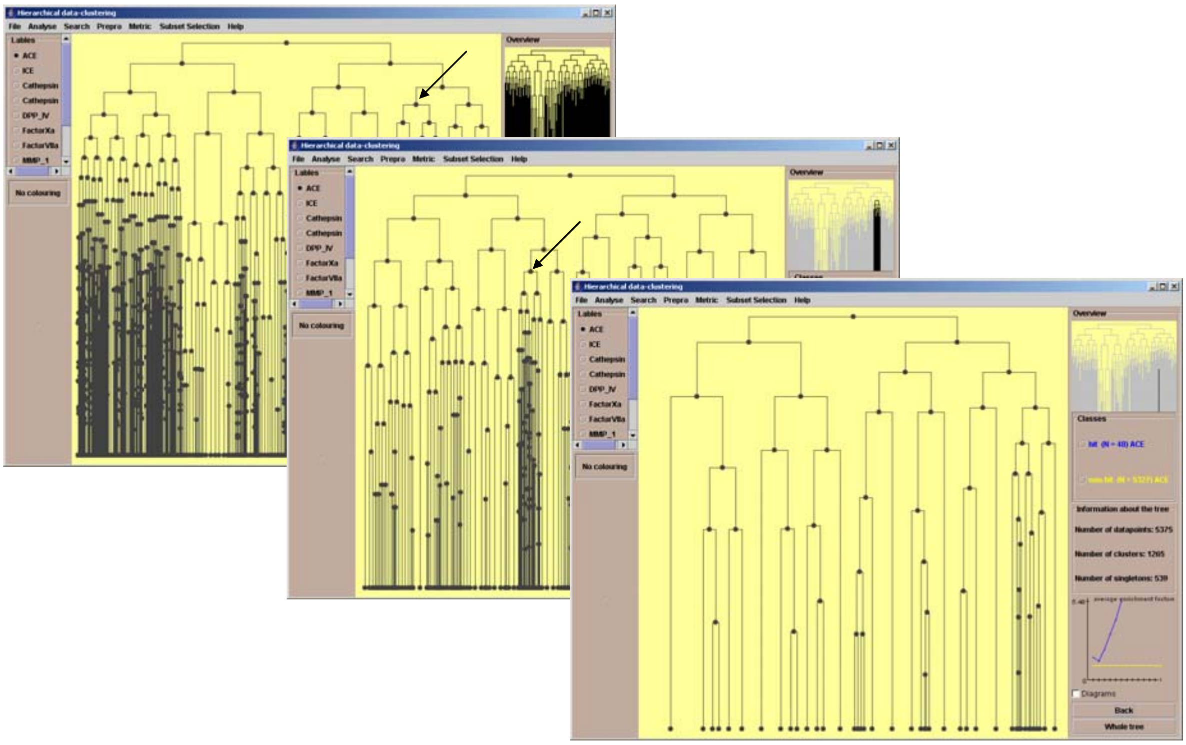
Number of singletons: 539

☐ Diagrams

Figure 4.2.2 Graphical user interface appearing after clustering of COBRA099 employing the hierarchical *k*-means algorithm. % CTL values have been loaded describing the inhibitor classes listed in Table 4.2.1. Based on a user defined % CTL threshold the compounds were assigned to the classes hit and non-hit. Clusters containing at least one ACE inhibitor (hit) were coloured in blue in the dendrogram.

The GUI was implemented so that several buttons on the left side panel can be activated simultaneously. Subsequently all pre-defined classes appear in the classes field. Multiple class buttons can be activated at the same time. Clusters in the dendrogram are then coloured according to the presence of different classes. It gives a first hint on selectivity issues. A second graphical display type has already been shown when describing results obtained for Fisher’s Iris data set, a pie chart representation of clusters in the dendrogram showing their relative class composition (see 4.1.7 and 4.1.8).

The cluster dendrogram in the main window is navigable. A cluster can be selected with the left mouse button and becomes the new root of the displayed dendrogram. To know which part of the dendrogram is displayed the “Overview” field on the right side panel shows the complete dendrogram in miniaturized form. Here, the actually drawn dendrogram is coloured in black. A navigation example is shown in Figure 4.2.3. Using the buttons “Back” and “Whole tree” on the right side, a navigation step back or a step back to show the complete dendrogram can be performed. This type of navigation enables to step in or out of chemical data distributions. It allows a coarse-grained and a fine-grained view on the molecules in the dendrograms. This is illustrated by two ACE inhibitor pairs shown in Figure 4.2.4. Molecules A and B were extracted from a cluster located on an upper dendrogram level. Here clusters were large and different lead structures were presented. Compounds C and D were extracted from a terminal cluster containing analogues of one lead structure.



Information about the tree
Number of datapoints: 5375
Number of clusters: 1265
Number of singletons: 539

Figure 4.2.3 Example of a navigation in the dendrogram obtained for COBRA099 employing the hierarchical *k*-means algorithm. Two focussing steps have been performed. The clusters indicated by the black arrows formed the new roots of the actually displayed dendrogram. The overview window on the right side panel shows the actually displayed dendrogram in black whereas the not visible part is highlighted in light grey.

The right side panel shows a frame titled “average enrichment factors”. It contains the curves for average *EF*s for classes “hit” and “non-hit” (section 4.1.6). Note that the at last selected assay on the left side panel is used for curve calculation. In addition to that both *SE* and *KBD* curves can be accessed by left clicking on this frame.

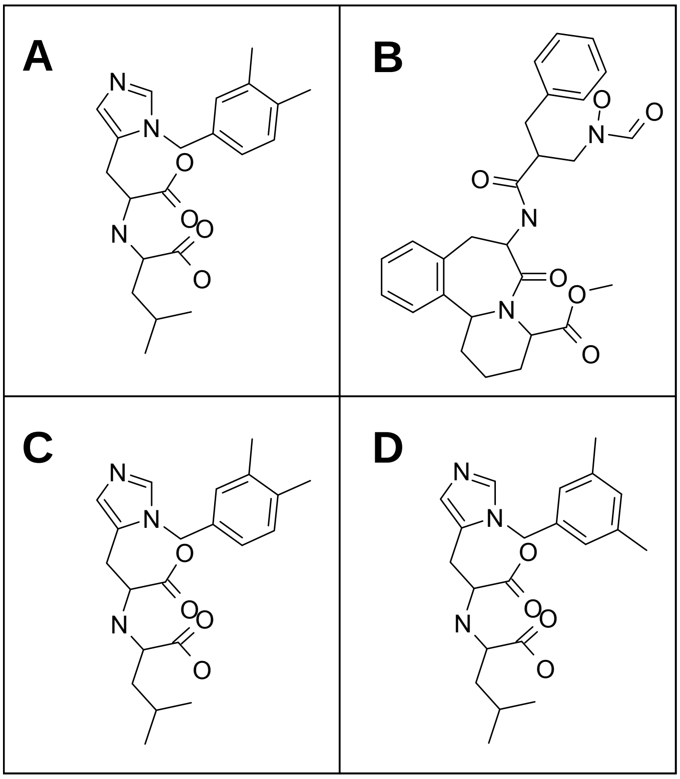
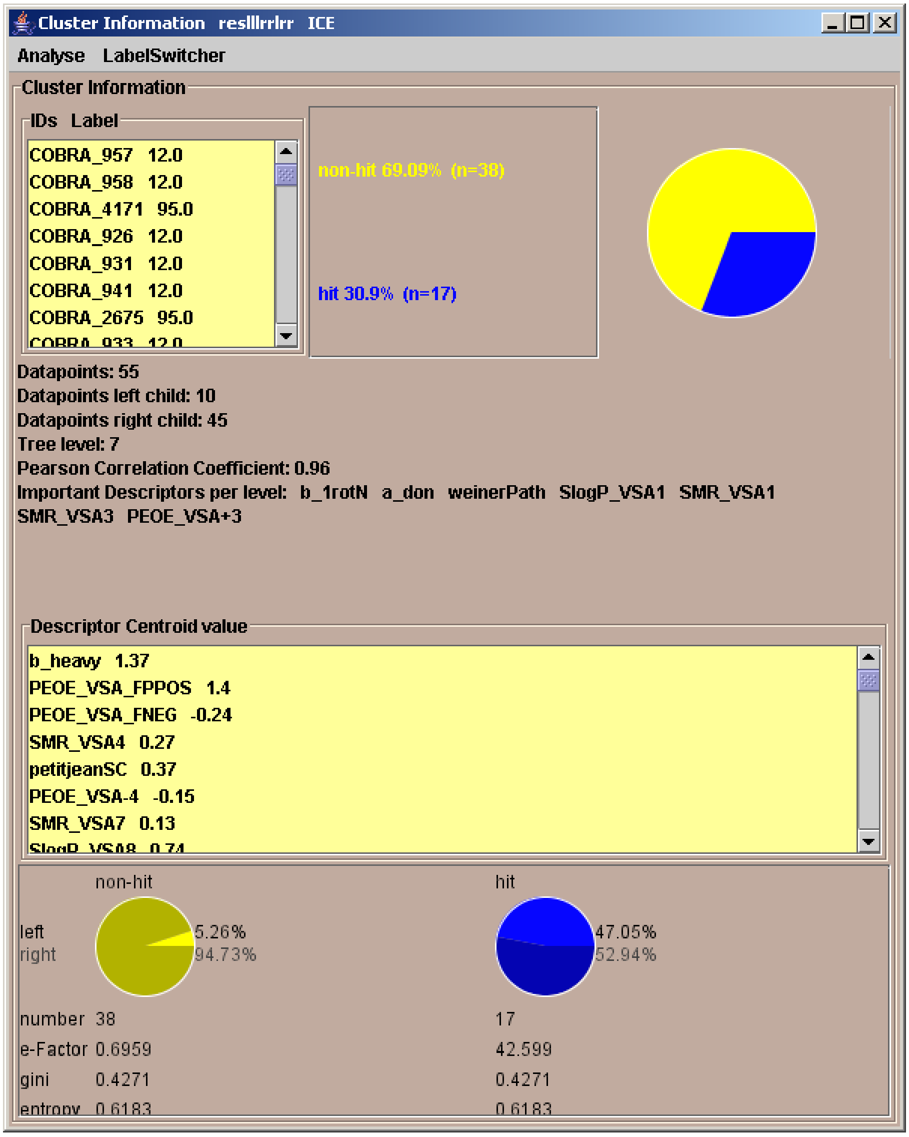


Figure 4.2.4 A-B: ACE inhibitors present in a clusters on an upper dendrogram level. C-D: ACE inhibitors in a terminal cluster.

By right clicking on a cluster in the main window, a new GUI opens. It contains statistical information about the cluster. Only the results of the last selected assay are shown. However, it is possible to access information about the other assays via the menu bar. The GUI is separated into five parts, a list showing the cluster entries combined with their % CTL values, two fields showing the consistency of the classes in the cluster, a general information part, a list showing the cluster centroid (hierarchical *k*-means) or the loading vector (NIPALSTREE) and a field showing statistical measures introduced in 4.1.6. Figure 4.2.5 shows a cluster. It was located on dendrogram level seven. % CTL column values for ICE inhibitors were selected.

The GUI contains an additional menu bar. It allows performing several functionalities for the cluster. One such functionality is to conduct a similarity search around the cluster centroid in the data set using a user-defined threshold. Every similarity metric or coefficient listed in Table 3.2 can be selected. A search might become necessary to cope with misclassifications putatively occurring with both algorithms. Using different metrics or coefficients for searching and clustering offers the chance to have alternative points of views on chemical similarity [Sheridan & Kearsley, 2002]. By applying more coarse grained (wider) similarity thresholds, the chemical space around the cluster is examined. It might translate into identifying new active compounds bearing novel scaffolds.


Cluster Information reslllrrlrr ICE
Analyse
LabelSwitcher
Cluster Information
IDs Label

COBRA_957 12.0
COBRA_958 12.0
COBRA_4171 95.0
COBRA_926 12.0
COBRA_931 12.0
COBRA_941 12.0
COBRA_2675 95.0
COBRA_933 12.0

non-hit 69.09% (n=38)

hit 30.9% (n=17)

Datapoints: 55
Datapoints left child: 10
Datapoints right child: 45
Tree level: 7
Pearson Correlation Coefficient: 0.96
Important Descriptors per level: b_1rotN a_don weinerPath SlogP_VSA1 SMR_VSA1 SMR_VSA3 PEOE_VSA+3

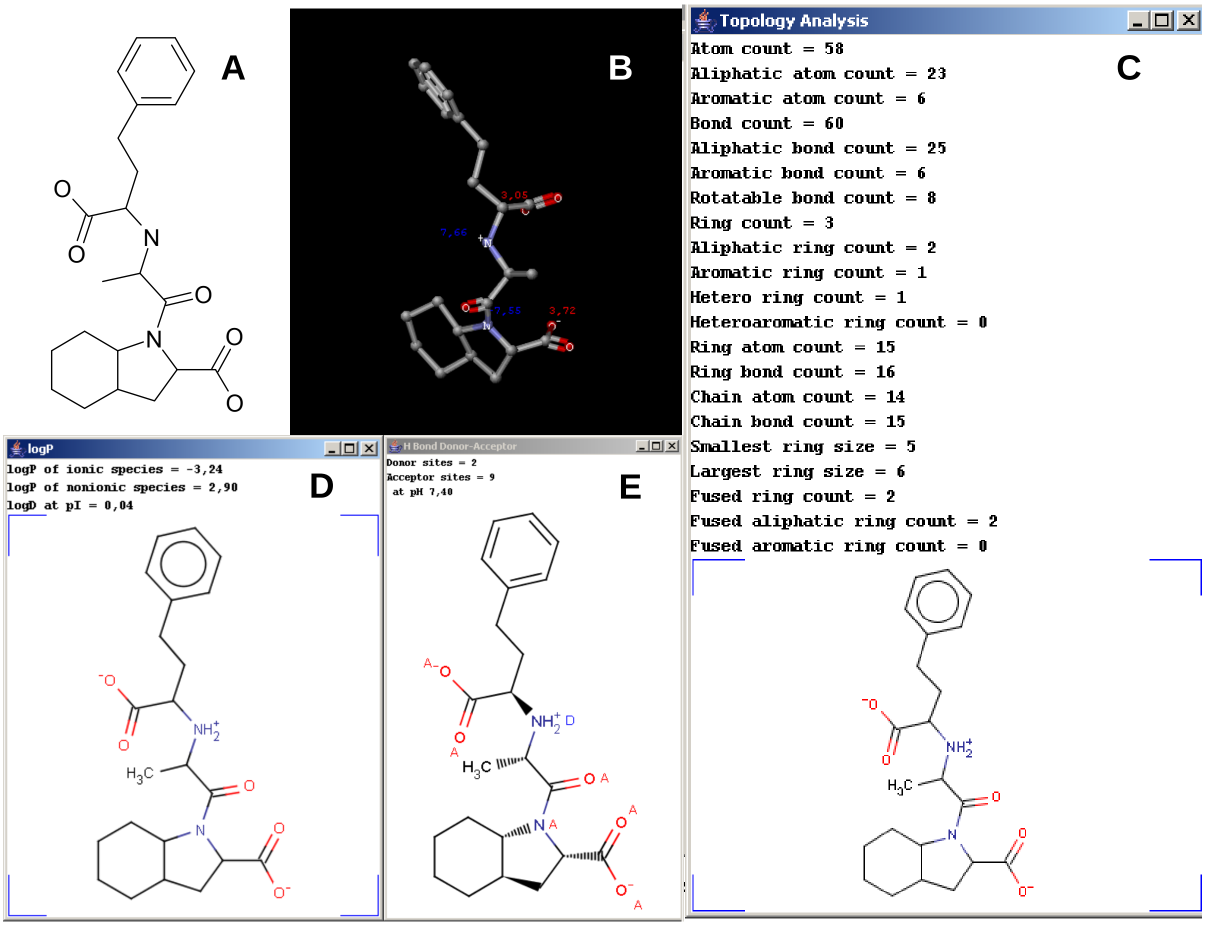
Descriptor Centroid value

b_heavy 1.37
PEOE_VSA_FPPOS 1.4
PEOE_VSA_FNEG -0.24
SMR_VSA4 0.27
petitjeanSC 0.37
PEOE_VSA-4 -0.15
SMR_VSA7 0.13
SlogP_VSA8 0.74

	non-hit	hit
left	5.26%	47.05%
right	94.73%	52.94%
number	38	17
e-Factor	0.6959	42.599
gini	0.4271	0.4271
entropy	0.6183	0.6183

Figure 4.2.5 Example of a GUI showing information about a cluster. % CTL values for ICE inhibitors were selected. Five different parts are displayed, the list showing the identifiers of the entries of the cluster, two fields showing the consistency of the ICE inhibitor classes (hit and non-hit), a general information part, the centroid vector and a field showing statistical information about the classes in the cluster.

A function in the menu bar allows displaying the molecules in the cluster in a scrollable window. For structure drawing Marvin structure viewer (version 3.1) was employed [ChemAxon Ltd., Budapest, Hungary]. It offers the possibility to perform a detailed analysis for a single molecule, ranging from 3D conformer and tautomer generation to property calculations like elemental analysis, topological analysis and Hückel analysis or the prediction of pK_a, logP, logD, polar surface area, polarizability and refractivity. As an example the ACE inhibitor trandolaprilat [Guay, 2003] was analysed (Figure 4.2.6 A).



```
Atom count = 58
Aliphatic atom count = 23
Aromatic atom count = 6
Bond count = 60
Aliphatic bond count = 25
Aromatic bond count = 6
Rotatable bond count = 8
Ring count = 3
Aliphatic ring count = 2
Aromatic ring count = 1
Hetero ring count = 1
Heteroaromatic ring count = 0
Ring atom count = 15
Ring bond count = 16
Chain atom count = 14
Chain bond count = 15
Smallest ring size = 5
Largest ring size = 6
Fused ring count = 2
Fused aliphatic ring count = 2
Fused aromatic ring count = 0
```

```
logP of ionic species = -3,24
logP of nonionic species = 2,90
logD at pI = 0,04
```

```
Donor sites = 2
Acceptor sites = 9
 at pH 7,40
```

Figure 4.2.6 Physicochemical analysis of trandolaprilat (A). B. The lowest free energy conformer of Trandolaprilat was determined and charged according to calculated pK_a values at pH 7.4. C. Topological analysis. D. calculated logP and logD values. E. Trandolaprilat at pH 7.4 in combination with expected donor (D) and acceptor (A) moieties.

As a first step the lowest free energy conformer was created. The pK_a values of the different groups in the molecule were calculated. Figure 4.2.6 B shows the resulting micro-species at the physiological pH of 7.4 in combination with the calculated pK_a values for each group. Both carboxylic groups were deprotonated (pK_a = 3.05 and 3.72), whereas the secondary nitrogen beard a positive charge (pK_a = 7.66). The basic amine and the opposite carboxylic groups are assumed to mutually enhance their basicity and acidity. It explains the lower pK_a for this carboxylic group. The tertiary amide nitrogen was assigned to a negative pK_a value indicating that it cannot take up a proton at any pH. The predicted pK_a properties are in good agreement with literature data [e.g. Williams & Lemke, 2002]. Figure 4.2.6 C shows the topological analysis of trandolaprilat. In D the calculated logP of both the neutral and charged isoform are presented. Figure 4.2.6 E shows the predicted isoform at pH 7.4 in combination with expected donor (D) and acceptor moieties (A).

Performing such detailed analyses for different molecules of a cluster can give a first hint on pharmacokinetic properties. This may help explain the SAR of the molecules or their ADME properties. Other types of analyses might be considered like the extraction of maximum common substructures from a cluster [Stahl & Mauser, 2005] or the alignment of molecules based on CoMFA analysis [Cramer et al., 1988]. Still the implemented or incorporated functionalities already allow a view on either the dendrogram or the cluster or the molecules in a cluster.

4.2.2 Terminal Cluster Analyses

4.2.1 introduced the GUI and the different focusing grades for SAR analyses. Here the guided navigation through terminal clusters is exemplified. When analyzing terminal clusters, three cluster types might be of interest:

(i) *Structural and biological singletons*. The clusters contain only one compound which is assumed to be active. They are outliers in the data set and are avoided in follow-up studies since no (Q)SAR can be directly derived. They might provide a rich source for alternative lead structures.

(ii) *Biological singletons*. The clusters contain several molecules. Only one is assumed to be active. These clusters are deprioritized since the active might be false-positive or possess a “steep” SAR. It is assumed to cause difficulties in the optimisation of this molecule towards higher activity or a better pharmacokinetic profile.

(iii) *Biological clusters*. This cluster type contains several compounds and all or at least a large proportion is assumed to be active. They allow drawing a first conclusion about SAR of the structural class/classes in the cluster.

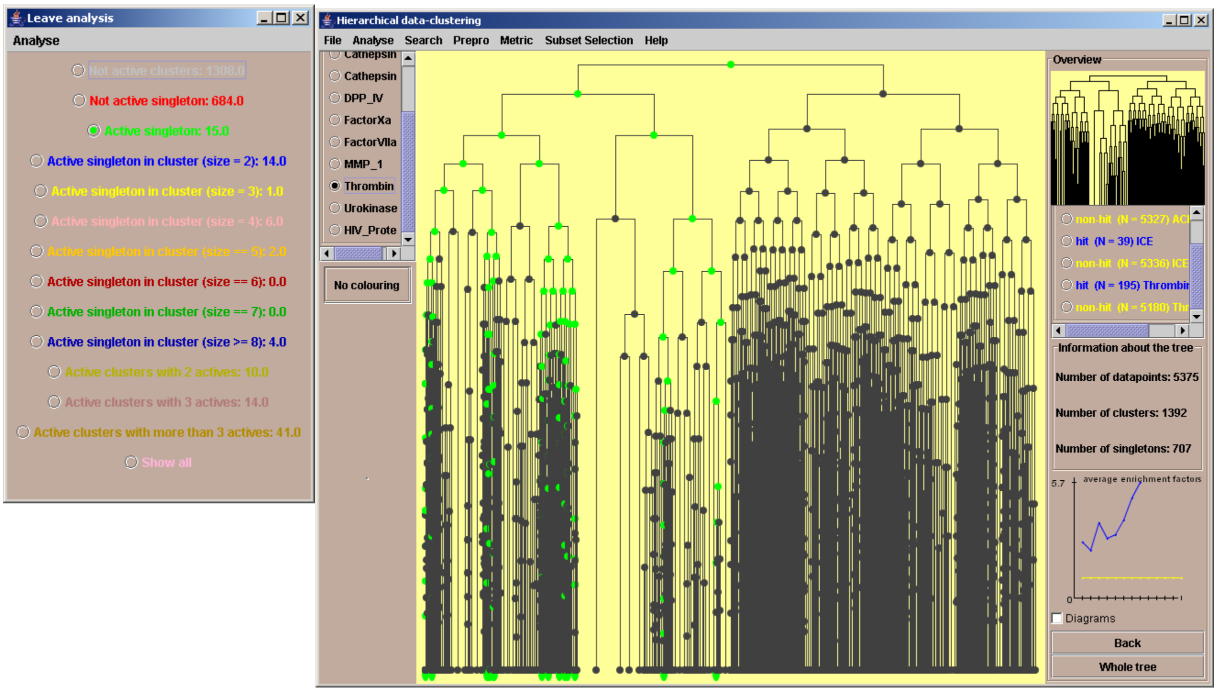
To identify different terminal cluster types a function was implemented counting the number of hits (actives) and non-hits (inactives) in a cluster. If a cluster contains at least one hit an additional similarity search around the cluster centroid is performed. The user-defined threshold (e.g. the estimated termination threshold Θ) and one of the metrics or coefficients of Table 3.2 is used. The final number of hits and non-hits is counted and the cluster is assigned to one of the predefined cluster types listed in Table 4.2.2.

4.2.2 Pre-defined cluster types to separate terminal clusters

Cluster type name	Number of hits	Cluster size
Not active cluster	-	>1
Not active singleton	-	1
Active singleton	1	1
Active singleton in cluster (size = 2)	1	2
Active singleton in cluster (size = 3)	1	3
Active singleton in cluster (size = 4)	1	4
Active singleton in cluster (size = 5)	1	5
Active singleton in cluster (size = 6)	1	6
Active singleton in cluster (size = 7)	1	7
Active singleton in cluster (size > 7)	1	>7
Active cluster with two actives	2	>2
Active cluster with three actives	3	>3

After performing the terminal cluster analysis a new GUI is displayed. It contains selectable buttons for the cluster types. Each button has a specific colour. When activating one of the buttons each terminal cluster belonging to the cluster type is redrawn as an oval with corresponding colour. All non-terminal clusters heading to these clusters are coloured likewise. For further discussion structural thrombin inhibitor singletons from COBRA099 were selected. The “Active singleton” button was activated and the corresponding clusters in the dendrogram were coloured in light-green (Figure 4.2.7).

The clustering of a data set can translate in two types of singletons, true singletons and false singletons. True singletons represent structural singletons whereas false singletons show only minor structural difference to related molecules. The occurrence of the latter singletons in a data set can depend on the used similarity threshold, the similarity metric, the descriptor scaling procedure, the descriptor set used in the application etc. To minimize them different solutions have been proposed like the re-clustering of singletons employing more coarse-grained similarity thresholds [Menard et al., 1998] or “fuzzy” clustering allowing a molecule to be a member of several clusters [Holliday et al., 2004] or the re-clustering of singletons based on maximum common substructure comparisons [Stahl & Mauser, 2005]. When analyzing thrombin inhibitors in the dendrogram only 15 clusters were present judged as structural and biological singletons. These singletons were examined manually by analyzing the parent clusters and screening the data set for the nearest neighbour. Chart 4.2 shows two representative examples. For the putative thrombin singletons **29** and **32** nearest neighbours **30** and **33** in COBRA099 were identified using the Euclidean metric. For **29** the parent cluster was examined. A representative of this cluster is shown in **31**.


Leave analysis
Analyse
○ Not active clusters: 1308.0
○ Not active singleton: 684.0
● Active singleton: 15.0
○ Active singleton in cluster (size = 2): 14.0
○ Active singleton in cluster (size = 3): 1.0
○ Active singleton in cluster (size = 4): 6.0
○ Active singleton in cluster (size == 5): 2.0
○ Active singleton in cluster (size == 6): 0.0
○ Active singleton in cluster (size == 7): 0.0
○ Active singleton in cluster (size >= 8): 4.0
○ Active clusters with 2 actives: 10.0
○ Active clusters with 3 actives: 14.0
○ Active clusters with more than 3 actives: 41.0
○ Show all
Hierarchical data-clustering
File Analyse Search Prepro Metric Subset Selection Help
○ Cathepsin
○ Cathepsin
○ DPP_IV
○ FactorXa
○ FactorVIIa
○ MMP_1
● Thrombin
○ Urokinase
○ HIV_Prote
No colouring
Overview
○ non-hit (N = 5327) ACI
○ hit (N = 39) ICE
○ non-hit (N = 5336) ICE
○ hit (N = 195) Thrombir
○ non-hit (N = 5180) Thr
Information about the tree
5.7
average enrichment factors
0
☐ Diagrams
Back
Whole tree

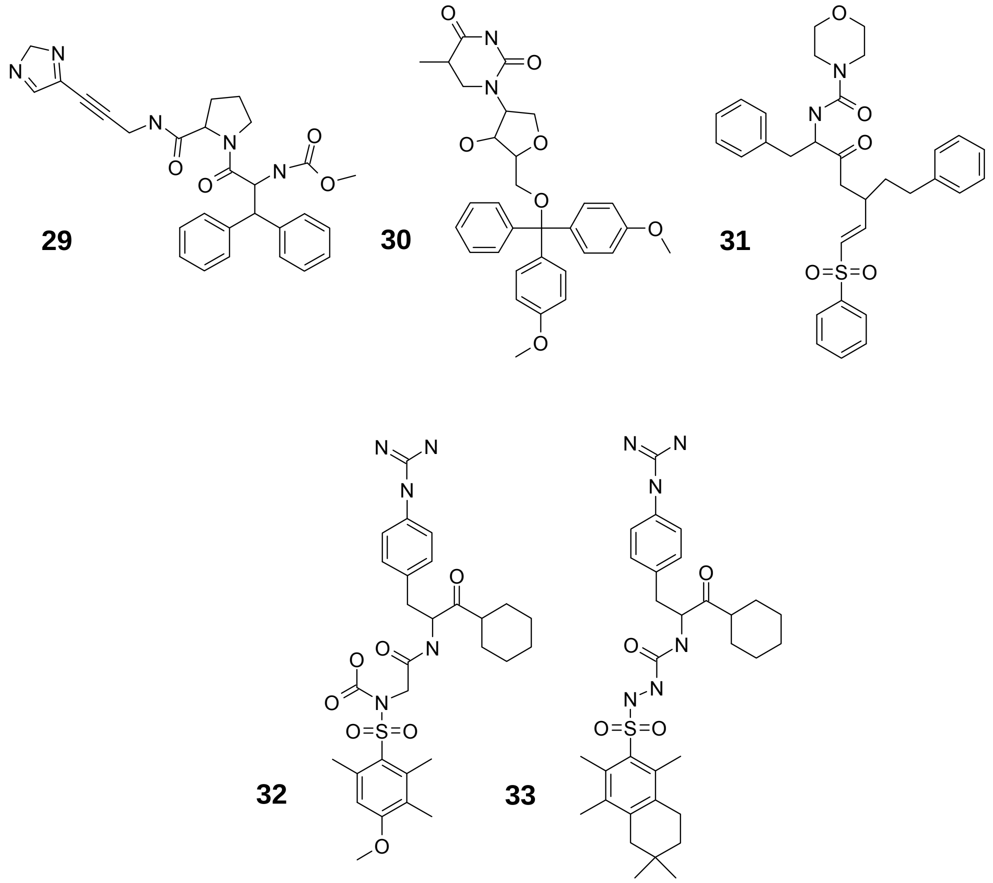
Number of datapoints: 5375

Number of clusters: 1392

Number of singletons: 707

Figure 4.2.7 The right GUI shows the dendrogram obtained by clustering COBRA099 employing the hierarchical k-means algorithm. The left GUI appeared after terminal cluster analysis focussing on thrombin inhibitors. The cluster types are shown as radio button with a specific colour. The “Active singleton” button was activated. In the dendrogram thrombin inhibitor singletons are coloured as light green ovals. All clusters in the dendrogram heading to one of these terminal clusters are coloured correspondingly.

Results indicate that in the first case the thrombin inhibitor **29** represented a real singleton since both the nearest neighbour **30** in the data set and the representative from the parent cluster **31** possessed different molecular scaffolds. In the second case the thrombin singleton **32** and its nearest neighbour **33** were similar. As main differences the methyl-carbamic acid moiety in **32** is replaced by a hydrazine substructure in **33**. False singletons can be a result of the description of the molecules, the employed similarity metric or the descriptor scaling procedure. To further minimize these singletons other types of analysis like MCS alignments or the mapping of potential pharmacophore points might be fruitful. However judging molecules as structural singletons lies in the eye of the beholder and will always require a final manual analysis. With the presented workflow it was possible to reduce the number of singletons to a minimum. The incorporated functionalities assist in quickly identifying these singletons and in the final visual inspection.



29 **30** **31** **32** **33**

Chart 4.2 Thrombin inhibitor singletons.

4.2.3 Relevant Clusters, Selectivity and Specificity

Thus far the focus of the guided dendrogram navigation was on identifying different types of terminal clusters. Here a second navigation type is introduced, helping to identify clusters enriched with actives in the dendrogram. As example application factor Xa inhibitors were selected. They play a central role in the blood clotting cascade, being the point of convergence of intrinsic and extrinsic pathways [Tan et al., 2003]. Consequently factor Xa is an attractive target for the development of new anticoagulants [Frederick et al., 2005; Ueno et al., 2005; Krovat et al., 2005]. For COBRA factor Xa inhibitors (N = 226) a molecular scaffold analysis was performed (see 3.4.1). The occurrence of each scaffold was counted. Scaffolds having more than five entries were visually inspected and if similar combined to a larger cluster. Results of the manual examination were eight different structural classes. Representatives are shown in Chart 4.3.

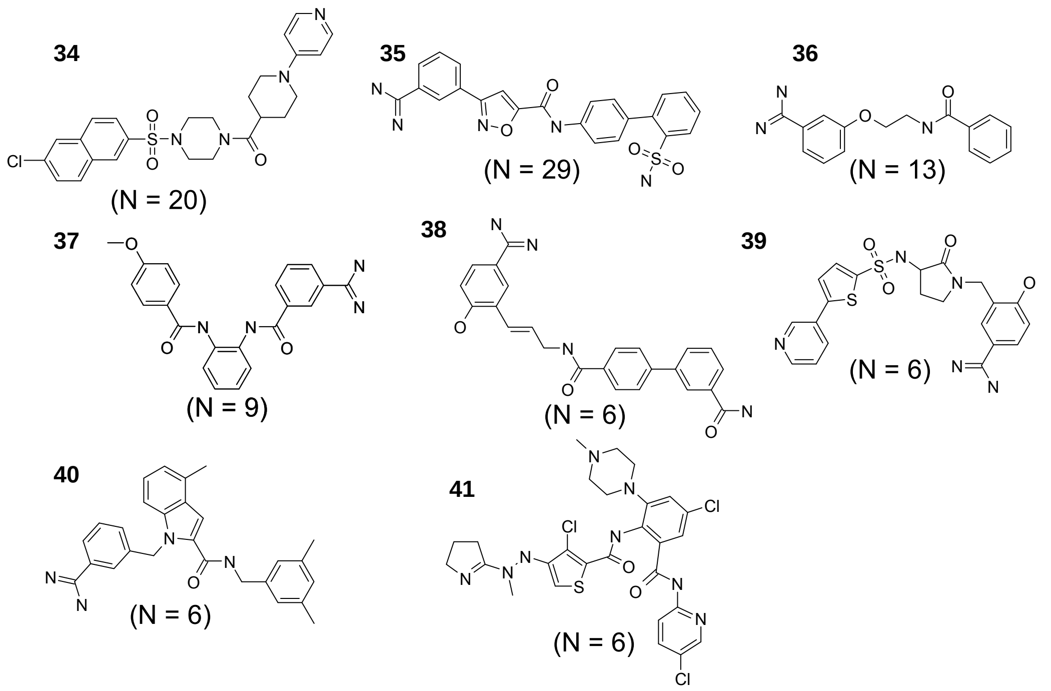


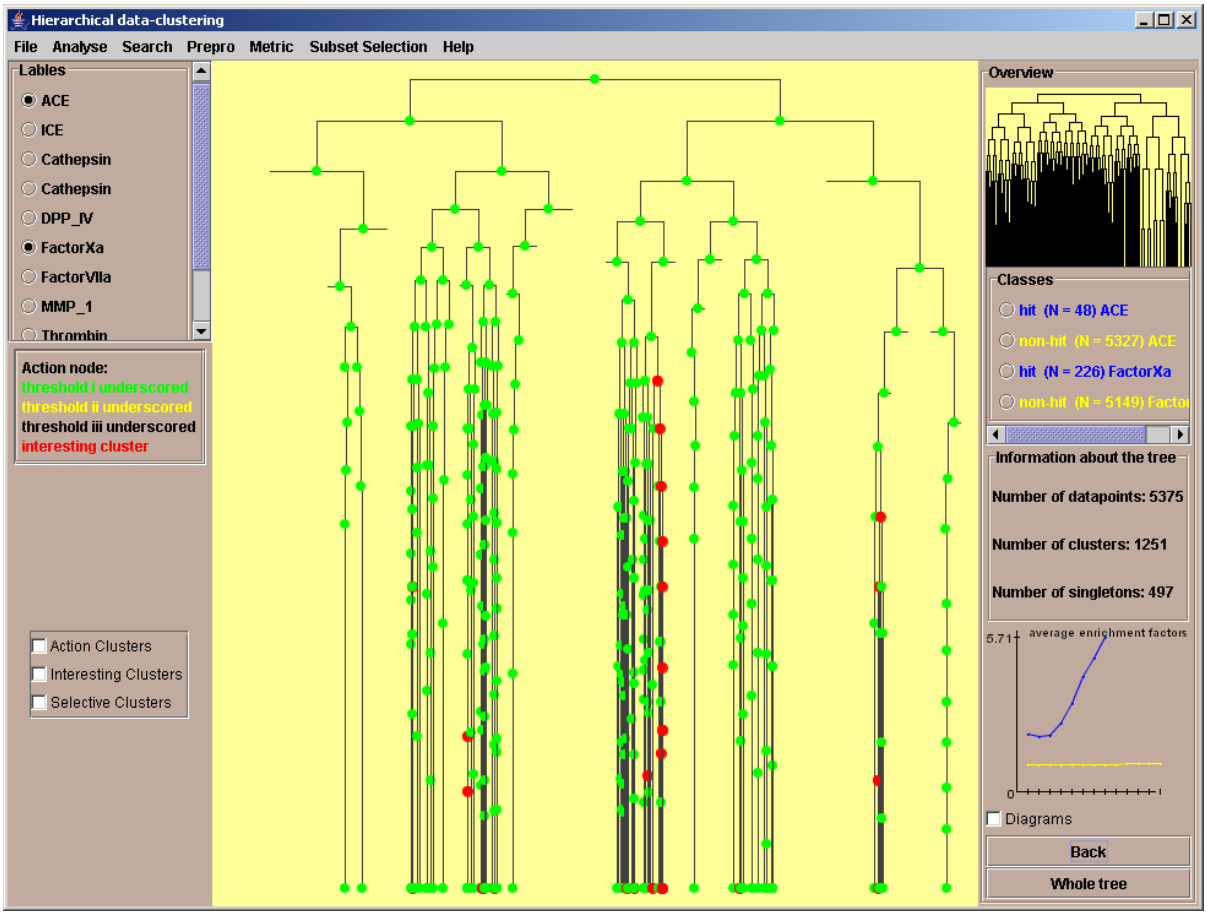
Chart 4.3 Representative factor Xa inhibitors.

To find structural classes of a ligand class in the cluster dendrogram two extreme scenarios might be considered: (i) a structural class can be large, diverse and may consist of both active and inactive compounds. For a detailed understanding of the SAR of this structural class clusters on upper dendrogram levels are of interest. (ii) A structural class can be small and homogenous in activity. Here one is looking for clusters on lower dendrogram levels. Despite of that, if multiple measurements at different targets have been performed for the compounds, selectivity and specificity are of interest.

To identify clusters enriched with active molecules of class *c*, a measure is to colour clusters, whose *EFs* for the ligand class exceeds a user-defined threshold. Problem of this measure is that on upper dendrogram levels clusters are large and *EFs* are low. On lower dendrogram levels enrichment factors can get large without statistical relevance. To obtain a measure identifying clusters enriched with class *c* entries on every dendrogram level with equal weight, the *EF* has to be adjusted to the dendrogram level *k*. Assuming a perfectly balanced binary dendrogram (i.e. at each cluster the data set is separated into equally large partitions) on every dendrogram level 2^k clusters are present. Scaling is achieved by dividing the *EF* for cluster *i* by the logarithm to base two of *k*. By this an *EF*, $bEF_{i,c,k}$ for cluster *i* and class *c* independent of *k* is obtained (Eq. 4.11).

$$bEF_{i,c,k} = \frac{EF_{i,c}}{\log_2 k}. \tag{4.11}$$

bEF values are calculated for clusters located on dendrogram levels, where the number of clusters is less or equals to the number of class c molecules. On higher dendrogram levels, artificially large enrichment factors can bias the calculation [Böcker et al., 2006]. Clusters are coloured red in the dendrogram if the value exceeds a predefined threshold ς. On higher levels clusters are coloured red, whose percentage of class c entries exceeds a predefined threshold τ. In addition to this an activity threshold, $AT1$, is defined. Clusters containing at least one entry having a biological response value exceeding $AT1$ are displayed and coloured green. All other clusters are excluded from display. For the factor Xa inhibitors in the dendrogram obtained for COBRA099 and the hierarchical k-means, values for ς, τ and $AT1$ were defined as 5, 75% and 50% CTL, respectively. The resulting dendrogram is shown in figure 4.2.8.


Hierarchical data-clustering
File Analyse Search Prepro Metric Subset Selection Help
Lables
Overview
Classes
Information about the tree
average enrichment factors
5.71
0
Diagrams
Back
Whole tree

● ACE
○ ICE
○ Cathepsin
○ Cathepsin
○ DPP_IV
● FactorXa
○ FactorVIIa
○ MMP_1
○ Thrombin

Action node:
threshold i underscored
threshold ii underscored
threshold iii underscored
interesting cluster

☐ Action Clusters
☐ Interesting Clusters
☐ Selective Clusters

○ hit (N = 48) ACE
○ non-hit (N = 5327) ACE
○ hit (N = 226) FactorXa
○ non-hit (N = 5149) Factor

Number of datapoints: 5375

Number of clusters: 1251

Number of singletons: 497

☐ Diagrams

Figure 4.2.8 Cluster dendrogram obtained by clustering COBRA_099 with the hierarchical k-means algorithm. Only clusters were displayed in green containing at least one factor Xa inhibitor. In red clusters were shown whose *bEF* for factor Xa inhibitors exceeded the predefined thresholds for ς or τ values.

Obviously when comparing the dendrogram in Figure 4.2.8 with the dendrogram in Figure 4.2.1 the number of displayed clusters is reduced markedly. It indicates that factor Xa inhibitors were clustered in individual branches of the dendrogram. The uppermost clusters coloured red were visually inspected. With the exception of the classes represented by **36** and **38** it was possible to re-identify all structural classes of factor Xa inhibitors. In numbers, 17 out of 20 molecules of the class represented by **34** were identified in two clusters. Only two non-hits were co-located. 27 out of 29 molecules of the class represented by **35** were found in one large cluster. 13 additional factor Xa inhibitors were co-clustered including the six members of the class represented by **39**. 22 non hits occurred. Five out of nine molecules of the class represented by **37** were identified in one cluster. Only hits were present. Four out of six compounds of the class represented by **40** were found in one cluster. No additional non-hit occurred. All five molecules of the class represented by **41** were identified in one cluster. No additional non-hit occurred. However six additional factor Xa inhibitors were co-clustered giving an extended view on the SAR of this structural class.

Two structural classes were missed in the analysis. This might be a result of the descriptor set used, the similarity metric or deficiencies of the clustering algorithm. A possibility to identify these classes might be to lower thresholds ς and τ. However this would increase the number of clusters to examine. For the class represented by **36** the number of molecules in COBRA was counted containing the scaffold. 43 molecules were identified of which 13 were described as factor Xa inhibitors. This large proportion of inactive compounds might serve as an explanation for missing the class with the specified thresholds. Both classes represented by **35** and **41** were clustered in combination with additional factor Xa inhibitors. For the cluster containing analogues of **41** these additional factor Xa inhibitors were missed during the initial scaffold analysis since their scaffold contained an additional ring or a ring of different size. It shows that with the *bEF* approach an extended SAR can be obtained. The identification of clusters consisting of different factor Xa classes (e.g. **35** and **39**) shows the lead hopping potential of this approach. Summarizing, six clusters were analyzed. It was possible to retrospectively identify six out of eight structural classes. It shows the value of the employed measure in combination with hierarchical clustering. When clustering larger data sets with several 100,000 compounds (e.g. molecules from vendor catalogues) and a few known actives two ways of adjusting ς and τ can be thought of. ς and τ can be set so that a high or a low number of co-clustered vendor molecules is retrieved. High amounts of molecules with unknown affinity can later be employed for targeted screening libraries. Low numbers are more suited for virtual screening applications aiming on the identification of new lead compounds.

The cluster containing 38 factor Xa inhibitors (classes represented by **35** and **39**) and 22 additional molecules represents an example worthwhile to addressing selectivity and specificity. The following graphical functionalities have been implemented to analyse a cluster: (i) the display of the number of hits and non-hits of each incorporated assay as a histogram. (ii) The display of enrichment factors of hits and non-hits of the assays as a histogram. (iii) The display of R^2 values as a histogram. These values are calculated between the actually selected activity values and the activity values of the other incorporated results. It should be noticed that this calculation makes only sense if all assay results belong to the same result format (e.g. only K_i values). Further the values should represent a quantitative activity measurement of high quality (that means no qualitative % CTL values, see 4.3.1). (iv) A more robust value is the calculation of percentage overlap between the actually selected hit class and all other hit classes. (v) The display of the molecules in a cluster in combination with all activity values as a histogram [Wilkens et al., 2005]. The cluster with the 38 factor Xa inhibitors and the 22 additional molecules was analysed accordingly. Results of (i), (ii), (iii) and (iv) are shown in Figure 4.2.9. The factor Xa values were selected for calculating R^2 values and the percentage of overlap.

In Figure 4.2.9 in the section titled with “Number” the % CTL values of the 60 compounds in the cluster were converted into the classes hit and non-hit. In total 38 factor Xa inhibitors, five factor VIIa inhibitors, four thrombin inhibitors and one urokinase inhibitor were present (12 undefined compound occurred). All inhibitors target trypsin-like serine proteases and underline that the clustering can be used for constructing focussed screening libraries. The histogram titled with “Enrichment factors” showed that factor VIIa inhibitors were the second most enriched class in this cluster. A phylogenetic sequence alignment employing the blosum62 substitution matrix, a gap start penalty of 7 and a gap extension penalty of 1 [Durbin et al., 1998] revealed that human factor Xa and factor VIIa have the closest similarity to each other with respect to the other target enzymes analysed in this application (Figure 4.2.11 A). This might give an explanation for the observed co-enrichment of both inhibitor classes. The histogram titled with “Correlation to selected label” shows, as expected, a self-correlation of 1 for factor Xa inhibitors. Since an artificial % CTL data set was employed and overlap between hit classes did not exist, all other R^2 values occurred without statistical relevance. This artificial correlation cannot be seen in the histogram titled with “overlap between actives”. It underlines the robustness of this simple measure. Figure 4.2.10 shows the results obtained for (v) as a scrollable window. In A four factor Xa inhibitors were selected and in B four factor VIIa inhibitors. Evidently both molecule groups represent different scaffolds giving a good explanation for affinity towards different receptors. The results show that by employing the *bEFs* approach the dendrogram is reduced to a manageable number of clusters. The implemented functionalities analysing selectivity and specificity highlighted a co-enrichment of factor VIIa and factor Xa inhibitors in a cluster. The combination of molecules in this cluster with the biological results explained the SAR in terms of selectivity towards factor Xa or factor VIIa.

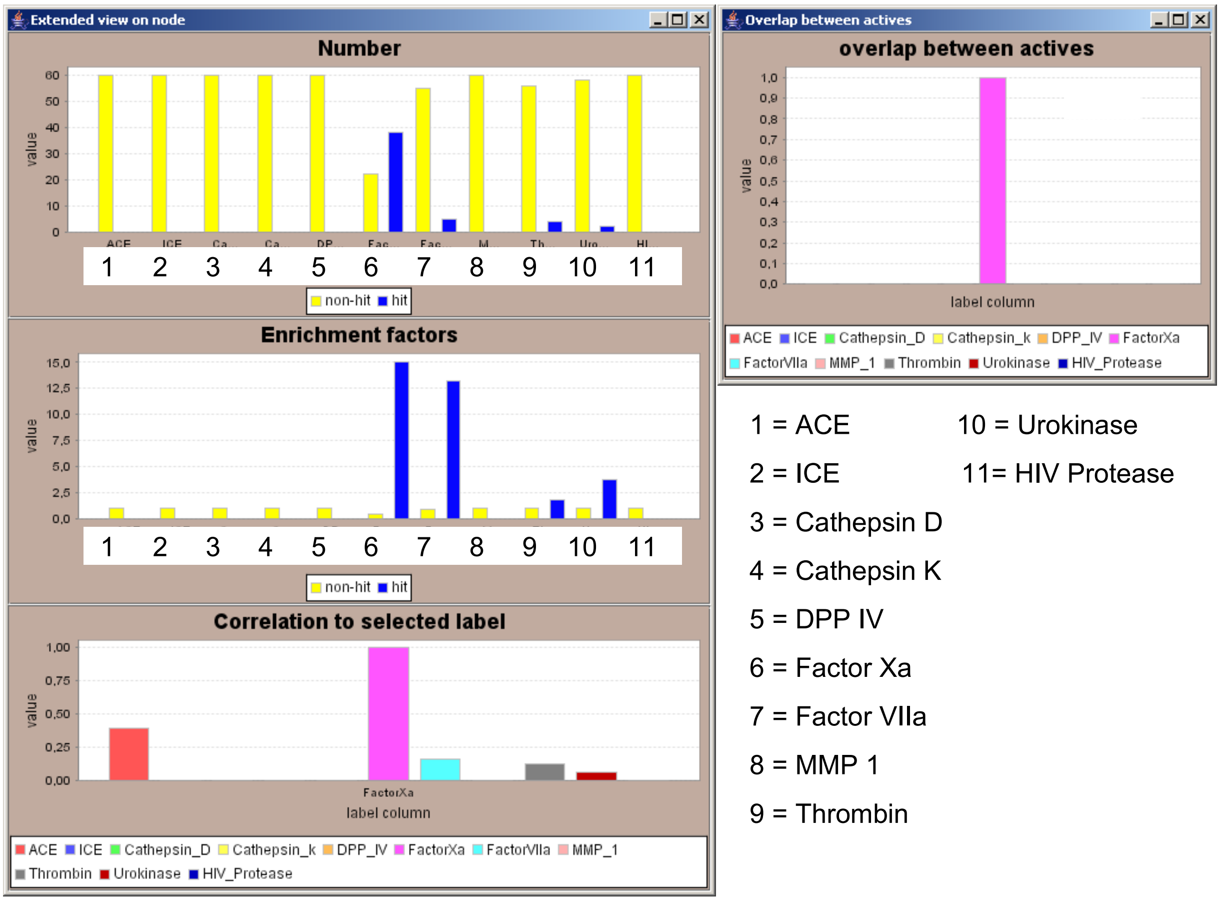


Figure 4.2.9 Results of selectivity and specificity analysis of a cluster. The histogram titled with “Number” shows the number of hits and non-hits in the cluster resulting from different assays. The column titled with “Enrichment factors” shows the equivalent of the first figure in terms of enrichment factors. The histogram titled with “Correlation to selected label” shows R^2 values calculated between the % CTL values of the factor Xa assay and all other assays. The histogram titled with “overlap between actives” shows the percentage of overlap between the hits of the factor Xa assay and the hits of all other assays.

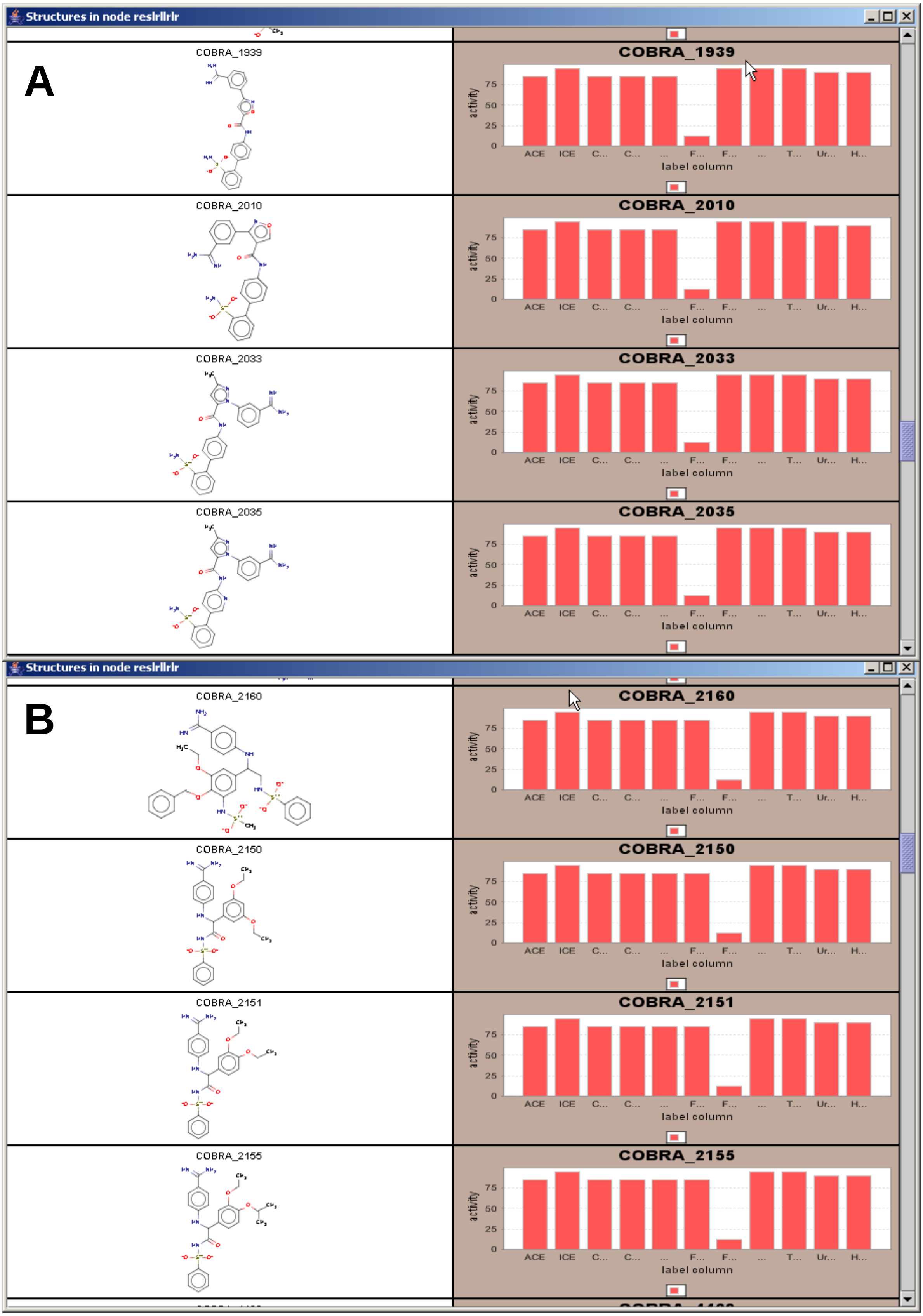


Figure 4.2.10 Display of molecules of a cluster in combination with all activity values. In A factor Xa inhibitors are shown. In B VIIa inhibitors are drawn.

The observed co-clustering of trypsin-like serine protease inhibitors in the dendrogram gave rise to a further specificity analysis. For each % CTL measurement and dendrogram level eight, all 256 clusters on this level were examined. The relative frequencies of the corresponding hits were extracted. For each inhibitor class of Table 4.1.1 a 256 dimensional descriptor was created combining structural information with biological activity. These descriptors were clustered using Ward's hierarchical clustering algorithm [Ward, 1963]. For comparison amino acid sequences of the human target enzymes were extracted from Swiss Prot (http://www.expasy.org/sprot/). Using the phylogenetic multiple sequence alignment algorithm implemented in MOE in combination with the blosum62 substitution matrix, a gap start penalty of 7 and a gap extension penalty of 1 [Durbin et al., 1998] the sequences were aligned. TreeView 1.6.6 was employed for phylogenetic tree display (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html/) [Page, 1996]. The phylogenetic dendrogram (A) and the Ward's clustering of the relative frequencies of the inhibitor classes are shown in Figure 4.2.11.

Both dendrograms show a co-location of factor Xa and factor VIIa. Further thrombin and urokinase are located at least in the same sub-dendrogram in both figures. Consequently a relationship between trypsin-like serine protease amino acid sequence and the structure classes inhibiting the protease enzymes exists. Results were different for the other protease targets. From the EC nomenclature it was anticipated that DPP IV (a serine protease) collocates with the other trypsin-like protease targets. This was only observed in dendrogram B. In the phylogenetic dendrogram DPP IV was clustered together with ACE. This different clustering was seen for all remaining targets. It shows that sequence alignments do not necessarily map the structural inhibitor profile of the enzymes. The clustering of relative frequencies of different inhibitor classes can be understood as ligand-based relationship of the analyzed binding pockets. Such a clustering of binding pockets was defined as "pocketome" analysis [An et al., 2006]. By using alternative information for defining the binding pocket (i.e. sequence information, structural information or ligand information) different views on the pocket relationships were obtained [Pirard & Matter, 2006; An et al., 2005; Oloff et al., 2006]. When searching targets for counter screens, the knowledge of a pocket clustering was assumed to give valuable information [Arnold et al., 2004]. The link between chemical and structural similarity formed a basis for targeted library design [Kellenberger et al., 2006]. Differences in pockets were identified helping define selectivity regions in a protein [Pirard & Matter, 2006]. The here proposed pocketome analysis is unique since no protein information is required. It provides a new, alternative view on the relationship between proteins and is especially useful when no 3D structural information is available for a set of proteins.

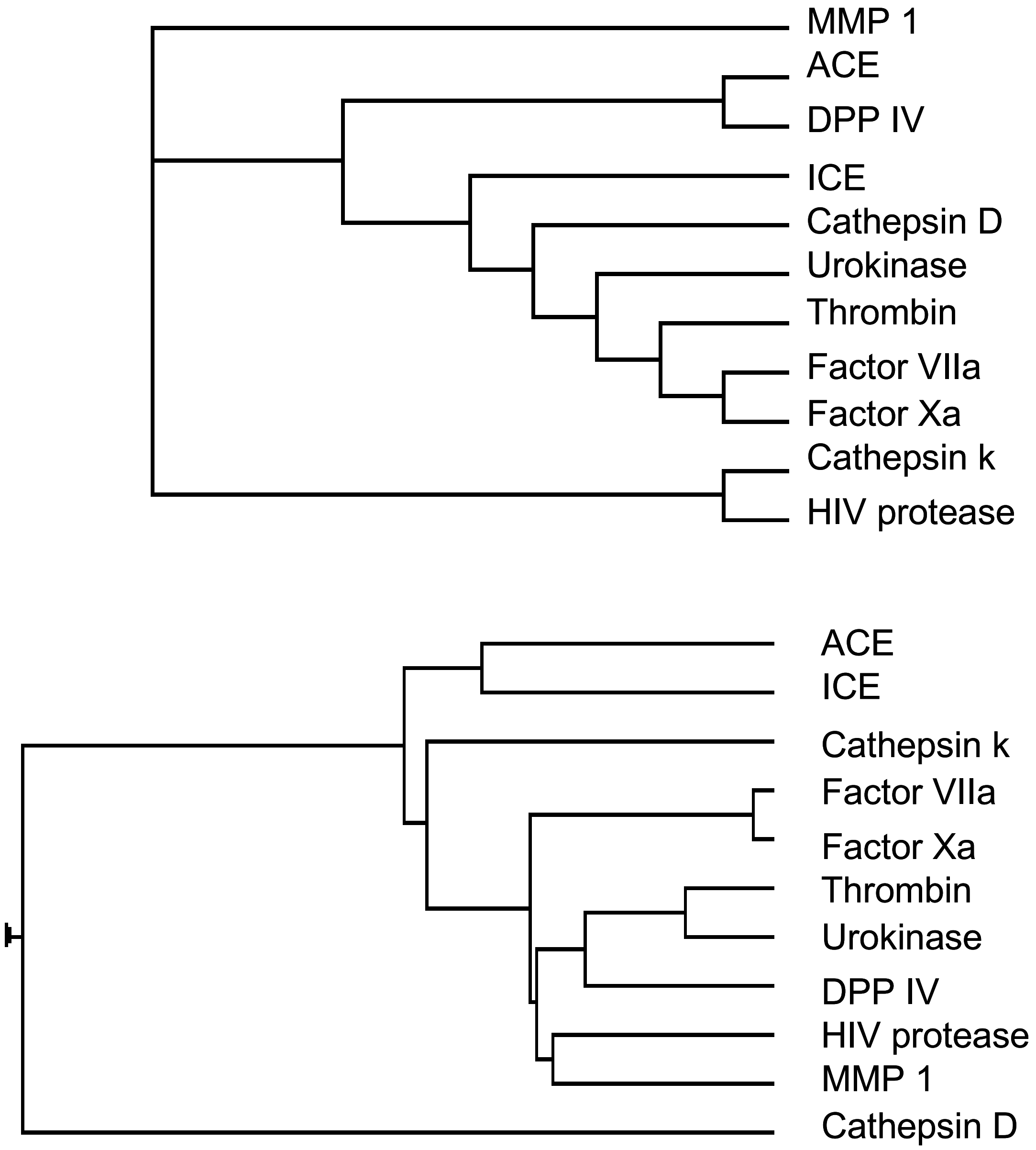


Figure 4.2.11 Hierarchical dendrograms for the inhibitor classes listed in Table 4.1.1. A. Phylogenetic dendrogram resulting from a multiple alignment of the human amino acid sequences of the protease enzymes. B. Ward's clustering of the inhibitor classes. Relative frequencies were employed as descriptors. They were extracted from clusters on dendrogram level 8 of the dendrogram obtained with the hierarchical *k*-means algorithm for COBRA099.

In addition to the presented functionalities several other methods were implemented. To identify selective clusters for one inhibitor class compared to another class (or all other classes) it is possible to divide corresponding *EF*s by each other. If a predefined threshold is exceeded or the ratio falls below a second threshold or the ratio lies in the range between both thresholds clusters are coloured differently. Substructure searching routines have been implemented allowing to identify clusters containing molecules of interest or privileged substructure elements [Schnur et al., 2006]. A projection function has been included allowing to project new molecules on the dendrogram for which e.g. pharmacokinetic measurements have been performed. By measuring the distance of projected molecules to the cluster centroid and comparing the distance to the actual cluster radius, the outlier behaviour of these compounds can be examined.

4.2.4 Conclusions

The hierarchical clustering of large data set requires to have a GUI at hand, which allows the display of and the navigation in the clusters. Such a GUI was developed. It provides functionalities to analyse the dendrogram, a cluster in the dendrogram, the compounds in a cluster or their molecular properties. A variety of measures and functionalities were introduced, which allowed to identify clusters enriched with actives or singletons or to analyse selectivity and specificity. The scaling of the enrichment factor for an inhibitor class to the logarithm of the corresponding dendrogram level was introduced to obtain a value which is independent of the dendrogram level and thereby of the cluster size. It allowed for the simultaneous identification of different lead series by one threshold value. This simplification bears the danger of loosing clusters and other methods employing impurity measures or a chi^2 statistic [Duda et al., 2001] might offer an advantage. However the proposed example of factor Xa inhibitors demonstrates that it was possible to reduce the number of clusters to a manageable size and to retain the majority of lead classes present in the data. The observed co-clustering of other serine protease inhibitors in these clusters highlights the possibility of this measure for constructing focussed screening libraries. An extended view on a lead class is possible and selectivity or specificity issues can be addressed. The identification of “true” singletons is of great value since it might provide a rich source for alternative lead structures. In this context the rating of terminal clusters, employing additional similarity searches, alternative similarity metrics and the constitution of these clusters with hits and non-hits provided a focussed view on these singletons. However the example showed that it was only possible to reduce the number of singletons and a manual analysis was necessary to achieve a final conclusion.

To get an insight into specificity, for a set of different protease inhibitor classes their relative frequencies in the clusters on a dendrogram level were extracted forming a new descriptor set. The inhibitor classes were clustered based on the new descriptor set. Comparison of the obtained hierarchical clustering to a phylogenetic multiple sequence alignment based on the amino acid sequences of the corresponding enzymes showed both a clear overlap as well as differences in the dendrograms. This new clustering allows an alternative ligand-based point of view on the relationship between the binding pockets for a set of proteins. It is useful when no 3D structures of the proteins are available or the binding pockets were shown to be very flexible. It can provide insight into adverse side effects or putative enzymes for which counter screens should be performed [Arnold et al., 2004].

Summarizing the implemented GUI provides functionalities which allow isolating and analysing a set of inhibitors in a large data set. The SAR of this inhibitor class can be viewed from different perspectives. The GUI is assumed to be applicable to the primary screening data of HTS assays.

4.3 Retrospective Analyses of HTS Assays

To apply the hierarchical *k*-means algorithm to SAR extraction in context of large data sets, the primary screening data of three HTS assays were selected. The aim was to develop workflows helping to detect false-negatives and primary screening hits having a high likelihood to translate into not confirmed hits. The introduction summarized the attempts already made in literature (see 1.2.3). The present study differs from them since, to the best knowledge of the author, for the first time the single point primary screening data (prior to confirmation measurement) are used to identify false-negatives and not confirmed hits.

In 4.3.1 the primary screening data of the three HTS assays are characterized. Secondly attempts are shown trying to identify not confirmed hits. In 4.3.3 a false-negative mining is introduced. As clustering techniques only the hierarchical *k*-means algorithm was employed in combination with either CATS 2D, MOE 2D or CATS 3D descriptors. Descriptor scaling and reduction was performed as described in 3.3.5 using a Shannon entropy threshold of 0.3 and a UFS R^2 threshold of 0.99. In all cases the number of descriptors was reduced to 40 to 60 descriptors. This was a prerequisite to run the clustering on a Linux workstation with 3.2 GHz Intel Xeon processor and 4 GB memory. Cluster stop thresholds were calculated according to 4.1.5 using the Euclidean distance metric.

4.3.1 Assay Characterization

An HTS assay can be characterized by the mean activity of the population in combination with its standard deviation σ_s and by the standard deviation of the measurement error σ_c [Zhang et al., 2000]. Ideally parameters like the confirmation rate, the false-positive rate and false-negative rate can be deduced from these values. Three HTS assays, assay A, assay B and the assay against TGF-β type I receptor (see Table 3.1, now referred to as TGF-β HTS) were analyzed. As first step a histogram was created of all % CTL values of the primary screening data. Figure 4.3.1 shows the histogram for the TGF-β HTS assay for the % CTL data range from -5 to 120. A peak defining inactive compounds was observed at a value of 106 % CTL. The standard deviation σ_s was 5.4 % CTL units. The % CTL threshold defining actives of 50 % CTL is more then ten standard deviations away from the observed mean. Hit entries succeeding the threshold were clearly separated from the non-hits. It shows the high quality of the assay. Comparable results were obtained for assay A and B.

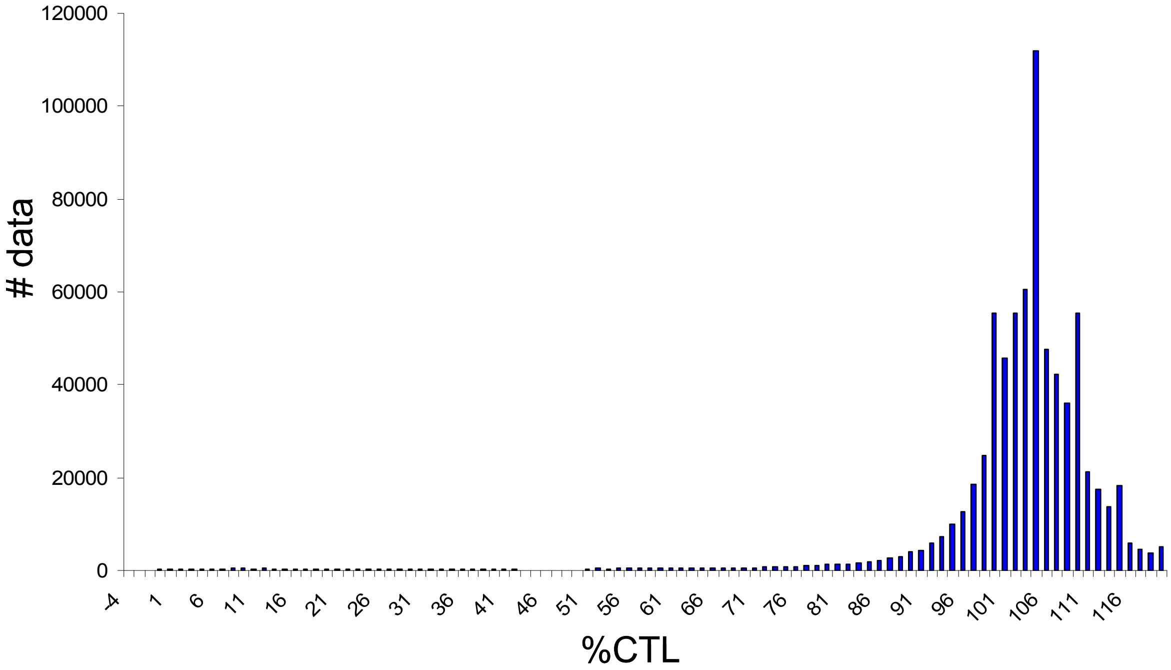


Figure 4.3.1 Histogram of all primary screening % CTL values of TGF-β HTS assay.

In the three HTS assays % CTL values were obtained by single point and single dose measurements. Compounds whose % CTL value fell below the hit threshold were confirmed twice setting the hit threshold one σ_s unit higher. Confirmed hits entered IC_{50} determination. To analyze the correlation between primary screening results and confirmation measurements or IC_{50} values corresponding values were plotted against each other. For illustrating the results assay A was selected. Figure 4.3.2 shows a correlation between primary screening % CTL values and confirmation % CTL values (A) and no correlation between % CTL and IC_{50} values (B, note that average % CTL were employed). Comparable results were obtained for assay B and the TGF-β HTS assay. All three assays showed a dynamic standard deviation σ_c with higher % CTL values translating in a higher σ_c. When constructing a histogram over the % CTL values of the primary screening hits a high bin occupation was observed at % CTL values close to the hit threshold. Not confirmed hits were present in every bin. The confirmed compounds translated into affine hits with IC_{50} values below 20 µM. However, from the missing correlation in the plot of % CTL values against IC_{50} values it is evident that % CTL values do not allow drawing quantitative conclusion about the SAR of the compounds.

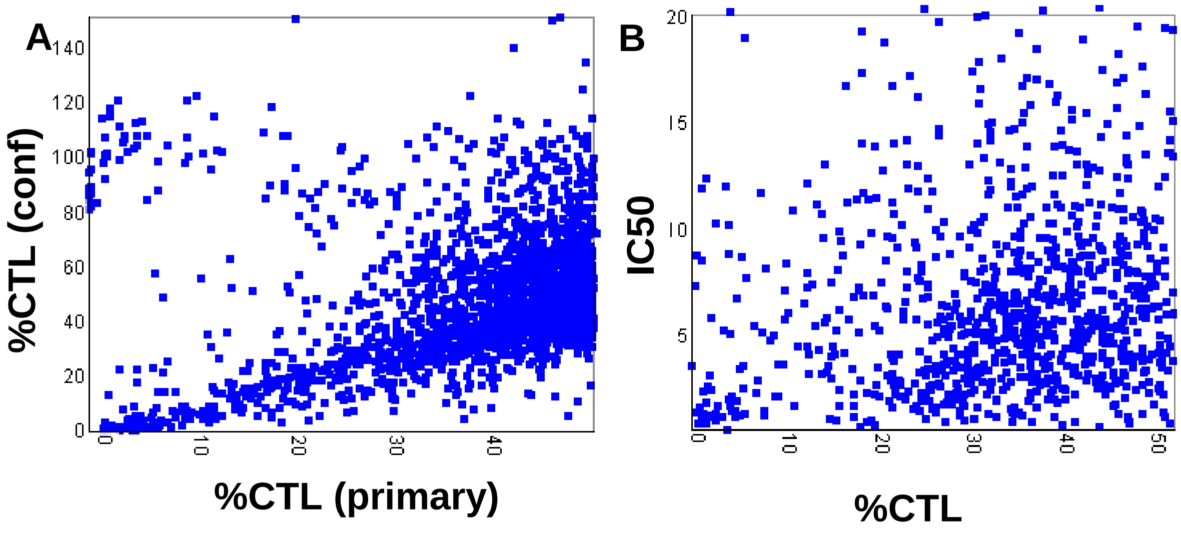


Figure 4.3.2 Analysis of assay A. A. Plot of primary hit % CTL values against % CTL values obtained in confirmation measurement. B. Plot of average % CTL values against IC_{50} values.

To get an insight into the distribution of not confirmed hits histograms were created over the % CTL values of the primary screening hits. Results showed that with the exception of the TGF-β HTS assay the majority of primary screening hits translated into not confirmed hits when the % CTL values were close to the confirmation threshold. For TGF-β HTS assay no such trend was visible. When analyzing histograms of these primary screening hits both the majority of not confirmed hits and confirmed hits showed % CTL values near the primary screening threshold. To draw conclusions enrichment factors were calculated for each histogram bin, defining the not confirmed hits as class c and all primary screening hits as total data set (see 3.3.6). The enrichment factors obtained for the TGF-β HTS assay are shown in Figure 4.3.3.

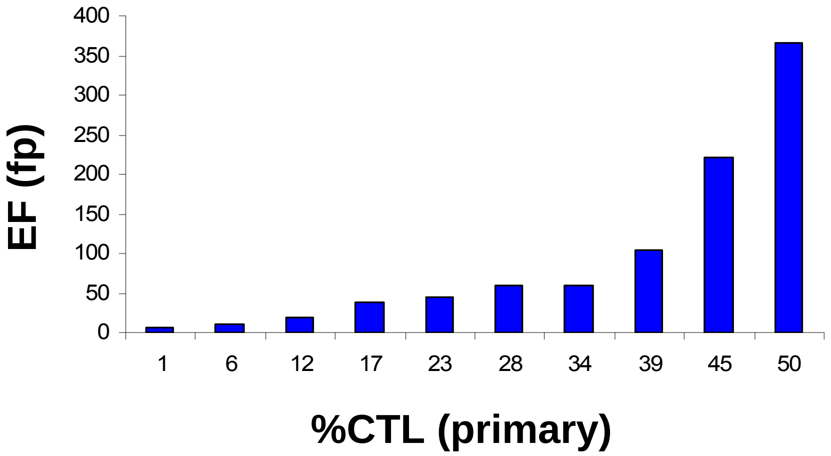


Figure 4.3.3 Enrichment factors obtained for not confirmed hits (EF(fp)) in a predefined data range of the primary screening % CTL values.

Results indicate that with increasing primary screening values the occurrence of not confirmed hits rises with respect to a random distribution. It confirms the theoretical considerations of Zhang et al.,: the closer the % CTL value of a primary screening hit to the hit threshold, the higher is the likelihood, that it translates into a non-confirmed hit [Zhang et al., 2000]. This makes the author confident that the same holds true for false-negatives.

Summarizing, the results show that % CTL thresholds defining hits were selected that a separation of hits from non-hits was achieved. % CTL values showed no correlation to IC_{50} values. It makes the derivation of QSAR with % CTL values impossible. Primary screening hit molecules exhibited IC_{50} values below 20 μM. This shows that it is possible to identified actives with the approach. Compounds having a % CTL value closer to the % CTL threshold showed a higher probability to results as not confirmed hit. The same is assumed for putative false-negatives.

4.3.2 Not Confirmed Hits

To analyze whether not confirmed hits can be predicted from the primary screening data compounds of all three assays were clustered. The idea was to employ the knowledge of the constitution of the terminal clusters with hits and non-hits to derive rules like (i) if a cluster consists only of hits, the occurrence of not confirmed hit is unlikely or (ii) if a cluster is large and contains only one hit, the hit might be a not confirmed hit. Different attempts were undertaken to approach the problem. Only clusters were considered containing at least one hit. To cope with the imperfection of both clustering algorithms additional similarity searches in the data set were performed around the terminal cluster centroids. The pre-calculated stop threshold Θ (see 4.1.5) was used as maximum distance threshold.

To address the identification of not confirmed hits in the cluster dendrograms the cluster size, the number of hits in the cluster and the occurrence of not confirmed hits (1 defines the occurrence of not confirmed hits and 0 the absence) were extracted from the terminal clusters containing hits. The values were plotted against each other using a box plot. Figure 4.3.4 A shows the results obtained for the TGF-β HTS assay, with the x-axis representing the number of hits, the y-axis showing the presence or absence of not confirmed hits in a cluster and the z-axis highlighting the cluster size. No correlation between cluster size, number of hits and not confirmed hits was visible (comparable results were obtained for assay A and B). Figure 4.3.4 B shows a box-whisker plot obtained for the TGF-β HTS assay. The x-axis represents the number of hits in a binned data range. The y-axis shows the proportion of not confirmed hits in that data range. Bars represent the data range between the quartiles of the observed false-positive proportions, whereas white separators and pluses correspond to mean and median values respectively. The whiskers explain a maximum of 1.5 fold the inter quartiles distance. They were always defined by a data point. The blue *x* represent outliers. A relationship between the proportions of not confirmed hits and number of hits in a cluster was not visible. The preliminary results indicate that not confirmed hits cannot be identified by simply focussing on the composition of a cluster with hits and non-hits. Yet additional relations need to be included like the measured affinity or the spatial relation between the clusters in the dendrogram.

A manual analysis of not confirmed hits from assay B was performed in the corresponding cluster dendrograms. The following rules were derived prioritizing hits in a terminal cluster:

(0) Define the hit % CTL threshold Θ. This value is usually defined by the assay performer when the assay is set up.

(1) Perform a similarity search around the cluster centroid in the data set (see 4.2.2).

(2) Consider only hits having a % CTL value above $\Theta\text{-}2\sigma_{s.}$ With that, only clusters are analysed having entries with % CTL values close to Θ. According to Figure 4.3.2 not confirmed hits are enriched in this data range.

(3) Define entries having a % CTL value *i* in: $\Theta < i < \Theta + 2\ \sigma_s$ as hits *n*. The rule copes with the standard deviation of the measurement error σ_c. Entries having % CTL values in this data range might be false-negatives.

(4) The number of primary screening hits and hits *n* is not allowed to be 100%. The rule excludes mainly singletons and small clusters from the analysis.

(5) If *N-score* of the screening hits and hits *n* in a cluster falls below ω, the hits are considered as not confirmed hits. The *N-score* is defined according to equation 4.12 [Krumrine et al., 2005].

$$N - score = sign(X - N \cdot P)\left[-\log\left(\int_X^N B(x, N, P)dx\right)\right], \tag{4.12}$$

with X being the number of hits in a cluster, *N* being the cluster size and *P* being the overall observed hit rate. The integral over the binomial distribution B(*x,N,P*) gives the probability of observing *x* hits in a cluster of size *N* in relation to the overall hit rate. In the present analyses *ω* was set to -0.4. It corresponds to one hit in a cluster of size 20 with a hit rate of 1.5%.

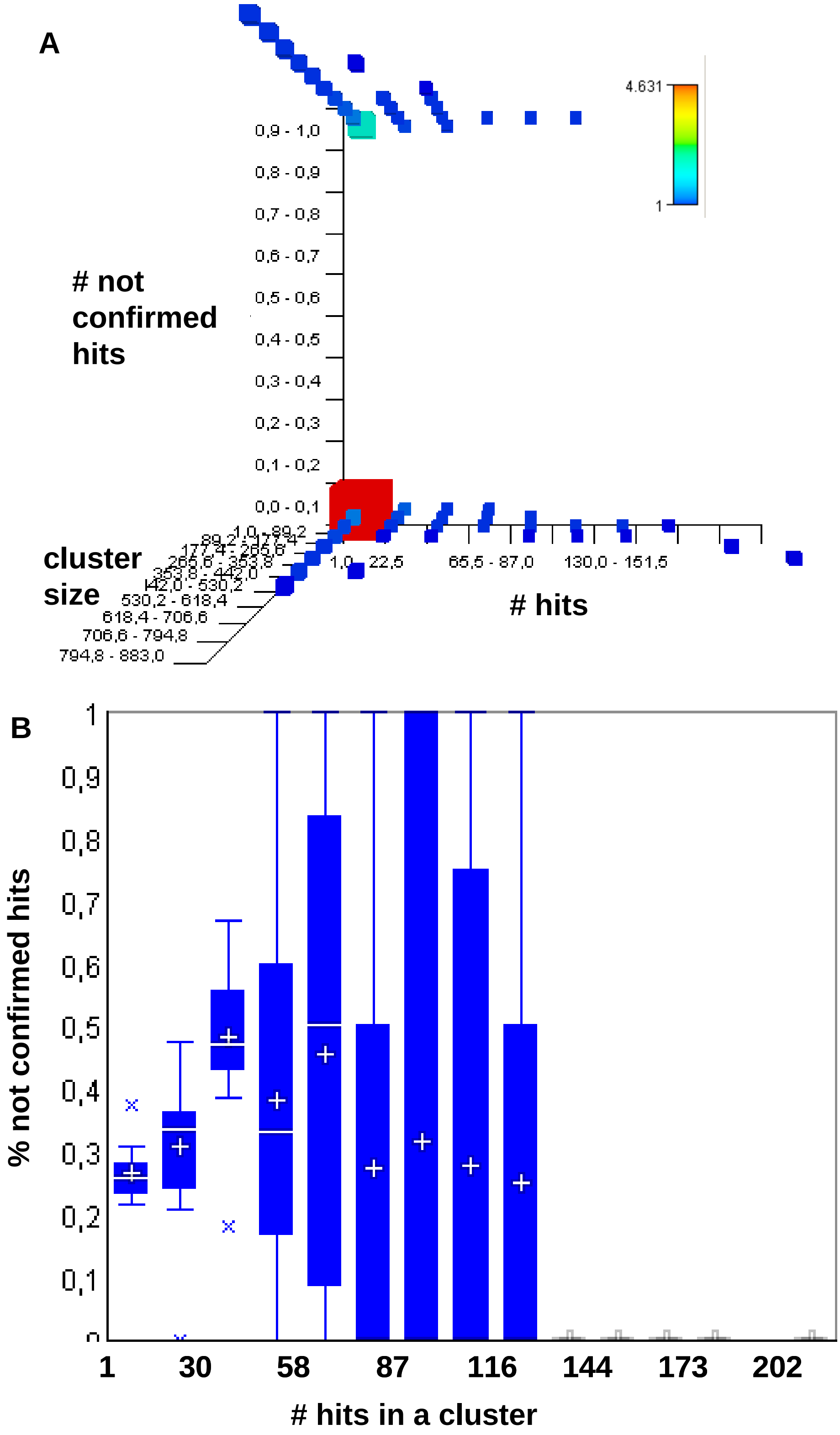


Figure 4.3.4 Analysis of the clustering of the compounds of the TGF-β HTS assay. **A.** Plot of the number of hits in a cluster against the cluster size and the occurrence of not confirmed hits in the cluster. **B.** Box-whiskers plot of the number of hits in a cluster and the proportion of not confirmed hits.

(6) Examine parent clusters recursively. According to this rule, the relation between the cluster and its brother clusters is examined. It corresponds to searching around the cluster centroid with a wider similarity threshold.

(7) No additional new primary screening hits and hits *n* are allowed on upper dendrogram levels. The rule copes with the inclusion of hit singletons or small hit clusters present in a sub-dendrogram. If this occurs no valid conclusion can be drawn.

(8) If the *N-score* of the screening hits and hits *n* in a cluster underscores *ω*, the hits are deprioritized and considered as not confirmed hits.

Dendrograms obtained with compounds of assays A, B and the TGF-β HTS assay were analysed according to the rule catalogue. The MOE 2D descriptor set was employed. Results are summarized in Table 4.3.1.

Table 4.3.1 Results of false-positive prediction

	Assay A	Assay B	TGF-β HTS assay
# primary hits	2,028	11,853	11,284
# confirmed hits	1,541	10,775	9,581
Confirmation rate	76%	91%	85%
Correct predictions	346	777	531
Total predictions	596	2,974	784
% correct predicted	58%	26%	68%
% false-positives identified	71%	72%	34%

For neither of the assays it was possible to obtain overall correct predictions. In case of assay A, assay B and the TGF-β HTS assay 58%, 26% and 67% were correctly predicted as not confirmed hits, respectively. These not-confirmed hits resemble the case where a few hits were co-clustered with many similar non-hits. The data show that not all not-confirmed hits were identified. This is evident from Figure 4.3.4 A since not confirmed singletons and not confirmed hits from clusters with several hits were present. The results let conclude that the rule catalogue has clear limitations for predicting not confirmed hits. The high number of falsely predicted confirmed hits indicates that corresponding molecules were present in clusters with a high proportion of similar non-hits. These hits are difficult for follow-up studies (small chemical modifications translate into loss of activity) and should be deprioritized. Consequently the rule catalogue provides a tool for identifying and rating such hits. This has been already proposed elsewhere [Schreyer et al., 2004].

For both assay A and B it was possible to identify over 70% of all not confirmed hits. For the TGF-β HTS assay only 34% were identified. Supervised classification techniques might offer an alternative to explain the remaining not confirmed hits in the TGF-β HTS assay. To test this, the following experimental setup was created. Dendrograms were created with either MOE 2D, CATS 2D or CATS 3D descriptors. Entries of terminal clusters containing hits were extracted. Clusters which were deprioritized according to the above specified rules were excluded. Based on the % CTL threshold compounds were pre-classified as hits (class 1) or non hits (class 0). Bayesian regularized artificial neural networks (BRANN, see 3.5.3) were trained with the data. The same compounds were then projected through the classification models to re-classify the entries. The basic idea was that the classification techniques are tolerant to noise and that the created models are robust enough to obtain a correct re-classification [Glick et al., 2006]. Results obtained for the TGF-β HTS assay are summarized in Table 4.3.2.

Table 4.3.2 Classification results for not-confirmed hits in TGF-β HTS assay.

		MOE 2D	CATS 2D	CATS 3D
Training data	Not confirmed hits	998	1,069	1,040
	Confirmed hits	9,328	9,555	9,476
	Size	34,153	42,564	37,019
Test data	True negatives	817	893	881
	True positives	5,106	4,975	4,709

The classification results for the test sets show that the not confirmed hits (true negatives) were well predicted with all three descriptor sets. However this is given to a significant loss of confirmed hits (true positives) rendering the method unacceptable for further usage in this context. To get a deeper understanding of the obtained classification, predictions obtained with the MOE 2D descriptor set were analysed in more detail. Contingency tables for the training set and test set are shown in Table 4.3.3. The test set contained only confirmed hits (class 1) and not confirmed hits (class 0), whereas the training set consisted of primary screening hits (class 1) and non-hits (class 0).

Table 4.3.3 Contingency tables for the training and test set for not confirmed hits in TGF-β HTS.

Training set

	Predicted	
Expected	0	1
0	21,953	1,700
1	5,049	5,289

Test set

	Predicted	
Expected	0	1
0	817	181
1	4,218	5,106

Both tables showed for the BRANN models a high specificity and a low sensitivity of around 50% true hits. The employed neural networks do not directly classify compounds into the classes 1 (hit) and 0 (non-hit). Yet instead a value is predicted in the range of 0 and 1. Based on a threshold of 0.5 compounds are assigned to either of the classes. Sensitivity might be increased by setting the classification threshold to a lower value. Figure 4.3.5 shows a histogram over the occurrence of hits and non-hits in the training set at different prediction data ranges.

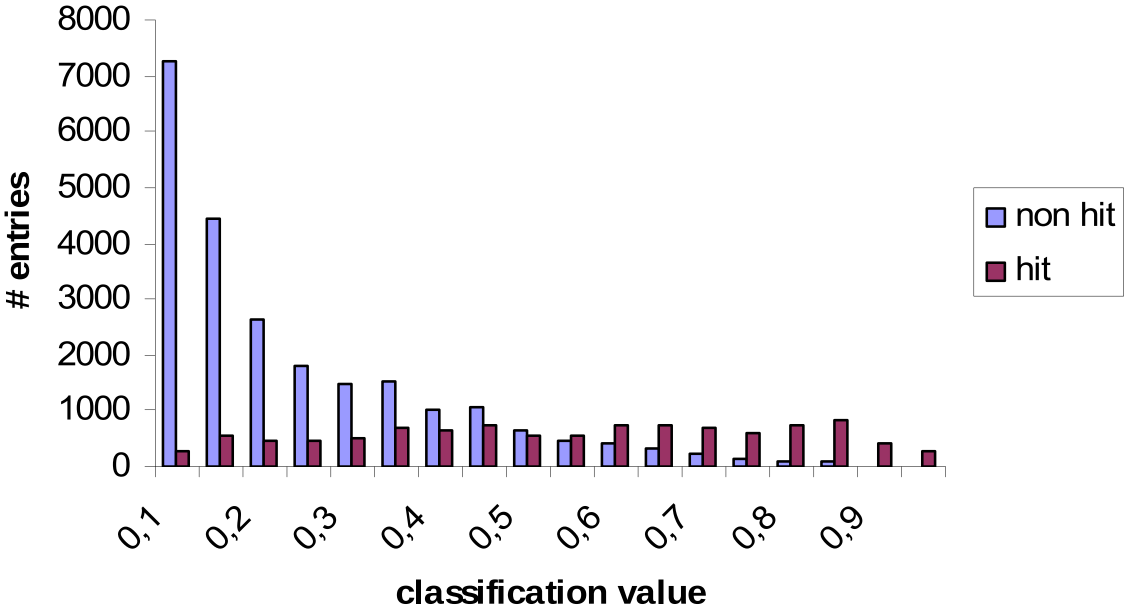


Figure 4.3.5 Histogram over the hits and non-hits of the training set obtained for MOE 2D BRANN model for the TGF-β HTS assay.

The majority of the non-hits were predicted with values clearly below 0.5 confirming the observed high specificity. The hit compounds instead had prediction values covering the complete range between 0 and 1. Same results were obtained when analysing the histograms for the test set (data not shown). In order not to loose a significant amount of confirmed hits the classification threshold has to be defined as low as possible. This in turn minimizes the number of identified not confirmed hits. It might be speculated whether other classification techniques like naïve Bayes' classification or SVM based classification might be better suited. In pre-experiments employing SVMs for classification calculation time was too long for the herein presented data sets (several months). The naïve Bayes' approach on the other hand translated into models with lower specificity and sensitivity in these experiments. This is exemplified by the training data set with MOE 2D descriptors (Table 4.3.2) for the TGF-β HTS assay. The BRANN model showed a sensitivity of 0.5 and a specificity of 0.93. The Bayesian classifier model provided a sensitivity of 0.3 and a specificity of 0.88.

Reasons for the obtained low sensitivity of the calculated BRANN models might have been that the training sets were unbalanced towards the non-hits (2-3 -fold) or that high amounts of noise were present in the data (i.e. systematic false-positives and false-negatives) or that the % CTL threshold defining hits and non-hits was not well chosen. In light of this pre-experiments setting the % CTL threshold one σ_s unit higher or lower for class definition did not further improve prediction accuracy.

Summarizing, the results show that by applying the rule catalogue it was possible to identify not confirmed hits in the cluster dendrogram. False predicted true hits identified by this conservative approach are assumed to be difficult to optimize. They are of no interest in follow up studies. It highlights the ability of the clustering approach to prioritize or de-prioritize hits based on the knowledge of the complete data set. The application of supervised classification techniques for the remaining not confirmed hits resulted in models with high specificity but low sensitivity. They were not suited for further application in this context.

4.3.3 False-Negatives

A prerequisite for judging the potential of a method to identify false-negatives is to know the false-negatives. A close-by idea in a retrospective analysis is to construct these false-negatives artificially. One simple possibility is to randomly define a set of hits as false-negatives. However the more challenging task is the identification of complete false-negative scaffolds thereby examining the scaffold-hopping potential of a method. To analyse whether false-negative scaffolds can be identified in the calculated cluster dendrograms primary screening hits of assays A, B and the TGF-β HTS assay were converted into reduced scaffolds (see 3.4.1). Small scaffolds like cyclohexane or naphthalene were highly abundant and the corresponding molecules did not represent uniform structural classes [Wilkens et al., 2005]. They were eliminated. The occurrence of each remaining scaffold was counted and the three most abundant scaffolds were selected to define false-negative classes. Results for all three assays are summarized in Table 4.3.4. Note that the same dendrograms were used for the three assays as in 4.3.2.

Table 4.3.4 Number of molecules forming false-negative classes 1-3.

	Assay A	Assay B	TGF β HTS
Primary hits	2,028	11,853	11,284
False-negative class 1§	55	546	563
False-negative class 2§	24	269	151
False-negative class 3§	23	196	142

§False-negatives were obtained by converting primary screening hits into reduced scaffolds. The occurrence of each scaffold was counted and the three most abundant scaffolds were selected as false-negative classes.

Only terminal clusters were analysed containing primary screening hits. The assumption was to identify co-clustered false-negatives. To illustrate the false-negative mining procedure results of the TGF-β HTS assay obtained with MOE 2D descriptors and false-negative class 1 are considered in more detail. The final results obtained for all three assays, the three false-negative classes and the three descriptor sets are summarized in Tables 4.3.5 (TGF-β HTS), 4.3.6 (assay A) and 4.3.7 (assay B). A schematic work-flow of the false-negative analysis is present in Figure 4.3.6. By definition the term "screening set" is referred to as the number of non-hits in combination with the number of false-negatives.

By extracting all compounds from terminal clusters containing hits 101 of the 563 false-negatives were identified. In total 39,298 compounds were retrieved (set 1). By prior excluding all clusters from the analysis judged as not confirmed hits according to the rules specified in 4.3.2 the total data size was reduced to 34,870 compounds (set 2). Still 101 false-negatives were identified. It corresponds to an enrichment factor for the false-negative-class of 3.87. Both hierarchical clustering algorithms are not perfect in classification and additional similarity searches around the hit cluster centroids might improve the results. For this the pre-calculated stop threshold Θ was used as maximum distance threshold in combination with Euclidean metric. The same clusters as for set 2 were examined. A high number of 409 compounds of the 563 false-negatives were identified. However the total data size increased likewise to 118,749 entries (set 3). An enrichment factor of 4.52 was obtained which is slightly better than that for set 2.

Summarizing, the results show that the identification of false-negatives in the cluster dendrogram is possible. The number of non-hits in set 2 and set 3 is large. This limits both sets for re-ordering compounds and testing for false-negatives. When analysing BRANN in context of not confirmed hits (4.3.2) specific models were obtained (specificity > 0.9) rendering them well-suited for filtering non-hits. To test whether this capability can be used for predicting false-negatives, BRANN models were trained with set 2 (note that the false-negatives were defined as non-hits). The false-negatives and non-hits were extracted from set 3 and projected through the calculated model. 173 of the 409 false-negatives were correctly predicted. In turn the screening data size was reduced from 108,423 non-hits to 3,288. This is a suitable size for ordering compounds and re-testing them in HTS. By training a supervised classification model with set 2 it was possible to separate non-hits from false-negatives in set 3. This is reflected by the high enrichment factor for the false-negative class of 69.1.

Results obtained for all three assays, the three descriptor set and the three false-negative classes are shown in Tables 4.3.5 - 4.3.7. They confirm the results described for the TGF-β HTS assay obtained with MOE 2D descriptors and false-negative class 1. It was possible to identify the false-negative classes in the dendrogram employing set 2. The only exception was the false-negative class 3 of assay A in combination with the MOE 2D descriptors. Additional similarity searches around the cluster centroid increased the number of identified false-negatives (set 3) and the screening size. No improvement of enrichment factors was achieved.

In 4.1 it was shown that both clustering algorithms are capable of clustering structurally similar compounds in a terminal cluster. Consequently a large proportion of a false-negative class is assumed to be co-clustered. In the present study descriptors were employed focussing on the properties of the molecules instead of the structure itself. By that compounds can be co-clustered having similar properties but different scaffolds. It explains the identification of the false-negative classes in set 1-3. The new hit compounds with novel scaffold might then serve as “seed” structure for further screening.

To reduce the number of co-clustered non-hits and maintain the false-negative classes BRANN models were trained. Projecting test sets obtained for assay B and the TGF-β HTS assay through the calculated models markedly reduced the screening size while maintaining at least a proportion of the false-negative classes. In most cases this led to high enrichment factors for the false-negative classes (see numbers in parentheses in Tables 4.3.5 – 4.3.7). A comparable analysis for assay A resulted in a complete loss of classes in four cases and in four other cases only one or two compounds were maintained.

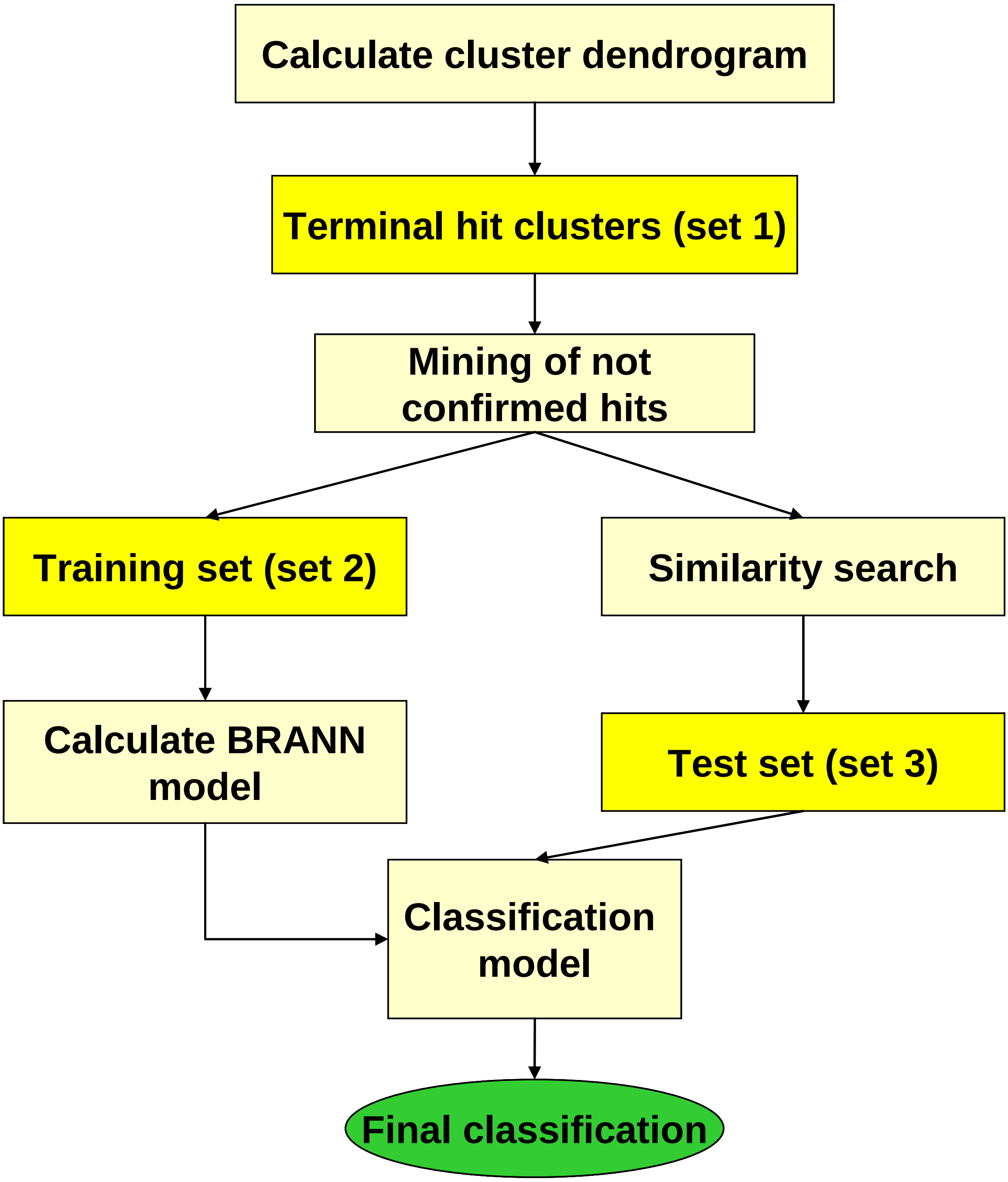


Figure 4.3.6 Workflow of the false-negative mining. Dendrograms are calculated for the compounds of a HTS assay. Only terminal clusters containing hits are considered. They form set 1. A mining of not confirmed hits in the cluster dendrogram is performed (see 4.3.2). Entries of the remaining clusters form the training set (set 2) which is employed for calculating a BRANN classification model. Hits are defined as class 1 and non hits as class 0. For the clusters of set 2 a similarity search in the descriptor matrix is performed. The pre-defined similarity threshold Θ is employed for searching. Resulting entries form the test set 3, which is projected through the BRANN model to obtain a final classification of the compounds.

Table 4.3.5 Results of false-negative mining of TGF-β HTS assay.

			MOE 2D§	CATS 2D§	CATS 3D§
Class 1	Basic	False-negatives	563	563	563
		Size	738,861	738,861	738,861
	Set 2	False-negatives	101 (EF=3.87)	381 (EF=11.9)	176 (EF=6.5)
		Size	34,252	42,199	35,557
	Set 3	False-negatives	409 (EF=4.52)	518 (EF=3.97)	383 (EF=3.52)
		Size	118,749	171,250	142,922
	BRANN	False-negatives	173 (EF=69.1)	51 (EF=19.8)	222 (EF=77.1)
		Size	3,288	3,374	3,780
Class 2	Basic	False-negatives	151	151	151
		Size	738,861	738,861	738,861
	Set 2	False-negatives	95 (EF=13.3)	141 (EF=16.2)	95 (EF=12.6)
		Size	34,870	42,540	37,049
	Set 3	False-negatives	137 (EF=5.59)	151 (EF=4.31)	130 (EF=4.44)
		Size	119,851	171,379	143,344
	BRANN	False-negatives	22 (EF=32.37)	31 (EF=44.8)	37 (EF=48.8)
		Size	3,326	3,387	3,711
Class 3	Basic	False-negatives	142	142	142
		Size	738,861	738,861	738,861
	Set 2	False-negatives	79 (EF=11.8)	119 (EF=14.6)	105 (EF=14.7)
		Size	34,957	42,524	37,198
	Set 3	False-negatives	128 (EF=5.58)	140 (EF=4.25)	129 (EF=4.68)
		Size	119,443	171,376	143,279
	BRANN	False-negatives	66 (EF=87)	39 (EF=58)	45 (EF=64.9)
		Size	3,947	3,502	3,609

§ In parentheses the enrichment factor for the false-negative class is present.

Table 4.3.6 Results of false-negative mining of assay A.

			MOE 2D§	CATS 2D§	CATS 3D§
Class 1	Basic	False-negatives	55	55	55
		Size	664,876	664,876	664,876
	Set 2	False-negatives	4 (EF=6.74)	46 (EF=48)	9 (EF=1.46)
		Size	7,172	11,589	8,089
	Set 3	False-negatives	28 (EF=10.9)	55 (EF=9)	44 (EF=1.01)
		Size	31,195	73,906	57,275
	BRANN	False-negatives	12 (EF=246)	0	2 (EF=3.92)
		Size	590	736	670
Class 2	Basic	False-negatives	24	24	24
		Size	664,876	664,876	664,876
	Set 2	False-negatives	6 (EF=22.8)	14 (EF=33.3)	5 (EF=17.5)
		Size	7,288	11,642	8,267
	Set 3	False-negatives	17 (EF=15.1)	22 (EF=8.26)	8 (EF=4.04)
		Size	31,248	73,770	57,213
	BRANN	False-negatives	2 (EF=73.9)	0	1 (EF=58.1)
		Size	750	1,053	498
Class 3	Basic	False-negatives	23	23	23
		Size	664,876	664,876	664,876
	Set 2	False-negatives	0	1 (EF=2.51)	3 (EF=4.43)
		Size	7,190	11,501	8,189
	Set 3	False-negatives	10 (EF=9.39)	12 (EF=4.73)	14 (EF=3.02)
		Size	30,799	73,343	56,095
	BRANN	False-negatives	0	1 (EF=40.54)	0
		Size	742	713	731

§ In parentheses the enrichment factor for the false-negative class is present.

Table 4.3.7 Results of false-negative mining of assay B.

			MOE 2D[§]	CATS 2D[§]	CATS 3D[§]
Class 1	Basic	False-negatives	546	546	546
		Size	549,619	549,619	549,619
	Set 2	False-negatives	306 (EF=10.3)	175 (EF=4.53)	105 (EF=3.01)
		Size	29,962	38,903	36,460
	Set 3	False-negatives	479 (EF=4.71)	359 (EF=2.28)	148 (EF=1.2)
		Size	102,350	158,300	124,029
	BRANN	False-negatives	4 (EF=1.04.)	69 (EF=16.6)	105 (EF=43.7)
		Size	3,889	4,126	2,421
Class 2	Basic	False-negatives	269	269	269
		Size	549,619	549,619	549,619
	Set 2	False-negatives	131 (EF=9.12)	170 (EF=9.24)	89 (EF=5.4)
		Size	29,363	37,585	36,673
	Set 3	False-negatives	226 (EF=4.58)	232 (EF=3.08)	133 (EF=2.18)
		Size	100,765	153,969	124,414
	BRANN	False-negatives	153 (EF=73.3)	70 (EF=40.9)	94 (EF=78.6)
		Size	4,267	3,496	2,445
Class 3	Basic	False-negatives	196	196	196
		Size	549,619	549,619	549,619
	Set 2	False-negatives	28 (EF=2.69)	75 (EF=5.61)	21 (EF=1.61)
		Size	29,200	37,469	36,578
	Set 3	False-negatives	80 (EF=2.23)	150 (EF=2.73)	38 (EF=0.86)
		Size	100,390	154,216	124,058
	BRANN	False-negatives	35 (EF=24.7)	11 (EF=8.47)	4 (EF=5.01)
		Size	3,981	3,653	2,241

[§] In parentheses the enrichment factor for the false-negative class is present.

Assay A showed a low hit rate of 0.3%. The size of the false-negative classes was small compared to assay B and TGF-β HTS. For the latter two assays screening sizes were reduced to around 3,500 compounds after BRANN classification. For assay A the screening size was reduced to around 700 compounds. A close-by idea to capture more of the false-negatives in assay A is to lower the classification threshold (see 3.5.4) and by that increase the screening size. Figure 4.3.7 shows the false-negative rates and false-positive rates obtained for set 3 and the false-negative class two of assay A in combination with MOE 2D descriptors. The classification threshold was systematically increased form 0 to 1.

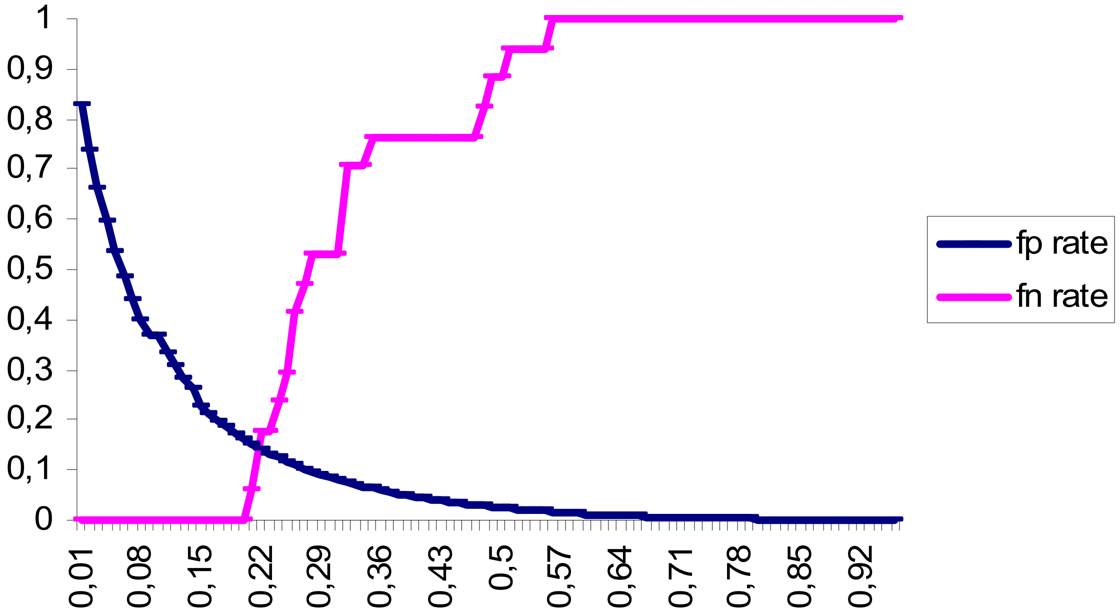


Figure 4.3.7 False-negative rates (magenta) and false-positive rates (blue) obtained for different classification threshold in the range of 0 and 1. BRANN classification results were used obtained for set 3, class 2, assay A and the MOE 2D descriptors.

Results show a crossing of the false-positive rate curve and the false-negative rate curve at a classification threshold of 0.23. At this point 14 out of 17 false-negatives were identified. The screening data size comprised 3,952 compounds. This size is comparable to assays B and TGF-β HTS. Similar curve progressions were obtained in combination with CATS 2D and CATS 3D descriptors. Another possibility to cope with the low hit identification rate of assay A is to combine predictions obtained with all descriptor sets. Entries of the test set of false-negative class 1 of assay A were sorted according to the classification values in descending order. A disjunctive combination of the first 1,000 compounds of all three descriptors was created. In total 15 false-negatives were maintained and the screening size increased to 2,846. Compounds of the MOE 2D data set contributed 12 false-negatives, whereas compounds of the CATS 2D and CATS 3D data sets contributed one and two false-negatives, respectively. The observed almost orthogonal disjunctive combination of result lists shows that with different descriptor sets different molecules were identified. This is a consequence of the descriptors' ability to describe different aspects of the molecules. Thereby different molecules or molecular scaffolds can be retrieved [Fechner et al., 2003; Renner & Schneider, 2006]. Each descriptor set has its unique strength. According to the obtained results a preference cannot be given. Describing the molecules with a combination of all descriptor sets might improve the results but has to be evaluated.

An observation was that result lists, obtained form dendrograms with CATS 2D descriptors were larger. As a consequence more of the false-negatives were identified. Compared to the other descriptor sets for final BRANN classification the training sets were more unbalanced towards the non-hits. This led to less sensitive models. Thus the positive aspect of identifying more false-negatives had a negative impact on the final classification. In light of this a preference might be given for the MOE 2D or the CATS 3D descriptors. The data sets employing the CATS 3D descriptors were more high dimensional (N = 420) compared to the CATS 2D (N = 150) and the MOE 2D descriptors (N = 146). This resulted in an increased calculation time and RAM deployment. Despite of that no clearly superior results were obtained in combination with this descriptor set. Consequently if larger data sets are used and RAM is a limiting factor the lower dimensional descriptor sets might be more useful.

Hert et al. proposed a method called data fusion. The method performs similarity searches around different reference structures in a data set and keeps the best similarity value of the molecules in the data set. The highest scoring entries define the hit list [Hert et al., 2006]. The main difference compared to the present study is that compounds are ranked based on the maximum similarity and not on the outcome of a supervised classification method. Whether this alternative ranking is superior to the proposed method has to be examined.

In 4.1.9 it was shown that the conjunction of result lists obtained with the NIPALSTREE algorithm and the hierarchical *k*-means resulted in a reduction of false-positives. The true hits were maintained. For assay A, instead of performing a final (supervised) classification, the combination of result lists obtained with both algorithms might offer an alternative method. For assay B and the TGF-β HTS assay the result lists were too large. Pre-experiments examining this property matched expectations but a more detailed analysis has to be performed.

The presented results show that the retrospective identification of false-negative scaffolds with the hierarchical *k*-means is possible. By combining results with a final supervised ranking improved enrichment factors can be obtained. This makes the author confident that the same holds true in a prospective screening application.

4.3.4 Conclusions

Three HTS assays were retrospectively analyzed. The histogram in Figure 4.3.1 shows that the assays are setup with the % CTL threshold that a clear separation of hits from non-hits is achieved. The correlation between the % CTL values of the hits of the primary measurement and the confirmation measurement underlines the reproducibility and robustness of the experiments (Figure 4.3.2). The IC_{50} of the confirmed hits were at most below 20 µM. They showed no correlation to the % CTL values. It shows that HTS is suited to identify hits (Figure 4.3.2) [Golebiowski et al., 2001; Golebiowski et al., 2003]. However no quantitative conclusion can be drawn from the % CTL values, most likely because affinity is measured in a narrow data range (e.g. 0% CTL - 50% CTL) whereas IC_{50} values cover several orders of magnitude. Enrichment factors were calculated for not-confirmed hits having a primary screening % CTL value within a certain data range (Figure 4.3.3). They confirmed the theoretical considerations of Zhang et al.,: the closer the % CTL value of a primary screening hit to the hit threshold, the higher is the likelihood, that it translates into a non-confirmed hit [Zhang et al., 2000]. This makes the author confident that the same holds true for false-negatives.

The hierarchical *k*-means algorithm was employed to mine primary screening data of the three HTS assays (prior to confirmation measurement). The aim was to identify not confirmed hits and false-negatives. All entries were extracted from terminal clusters containing hits. No correlation was observed between the cluster size (i.e. similar molecules), the number of hits in a cluster and the proportion of not confirmed hits (Figure 4.3.4). A conservative rule catalogue was developed rating hits in terminal clusters based on the cluster size, the % CTL values of the entries in a cluster, the overall hit rate, the hit rate in the cluster and the environment of a cluster in the dendrogram. The results let conclude that it is possible to identify not confirmed hits with the approach (Table 4.3.1). These hits resemble the case where a few hits were co-clustered with many similar non-hits. The data show that not all not-confirmed hits were predicted correctly (this is evident from Figure 4.3.4) and that confirmed hits were falsely predicted. The latter indicates that several confirmed hits were grouped with a high proportion of similar non-hits. These hits are difficult for follow-up studies and should be de-prioritized. Consequently the rule catalogue provides a powerful tool for rating hits in the data based on the knowledge of non-hits [Schreyer et al., 2004].

Compounds were extracted from terminal clusters containing primary screening hits. These molecules were employed for training BRANN models with the aim to separate hits from non-hits. Applying the models to identify not confirmed hits showed a high specificity of 0.93 but a low sensitivity of 0.5 (i.e. half of the hits are missed). This renders them unsuited for further application. The histogram in Figure 4.3.5 shows that the classification technique predicted hits uniformly in the classification range from 0 to 1. No classification threshold could be set clearly separating hits from non-hits. Although additional supervised techniques, different descriptor sets and different parameter settings have yet to be evaluated. It is assumed that the primary screening data of HTS assays are not suited for predicting these cases.

False-negative classes were retrospectively created using a scaffold based approach. The hierarchical *k*-means algorithm was employed to cluster the data of the three HTS assays. The rule catalogue was applied to de-prioritize hits/clusters. From the remaining terminal clusters containing hits, entries were extracted and analyzed. With one exception it was possible to co-extract and enrich false-negatives. The proposed clustering based method can be compared to nearest-neighbour searching. The results confirm the observation made by Shanmugasundaram et al. that it is possible to retrieve false-negatives from HTS hits by using hit-directed nearest-neighbour searching [Shanmugasundaram et al. 2005]. Although hit-directed similarity searching is computationally more efficient and a comparison of both methods has to be performed, the clustering based approach offers the advantage of providing a hierarchical grouping of the compounds and not a sorted list. This additional information allows the application of the rule catalogue for (de-) prioritizing hits. It provides an alternative grouping of the compounds around virtual centroids. The incorporation of similarity searches for the identification of false-negatives in this study was able to retrieve more false-negatives but it had at most a negative influence on the enrichment. It mirrors the alternative grouping by clustering and underlines the advantage of the clustering based approach.

The hit lists of the cluster-based false-negative mining comprised too many non-hits for reordering and testing. In order to rationalize the selection BRANN models were trained. The non-hits identified with additional similarity searches were projected through the model. For assay B and the TGF-β HTS assay this led to a marked improvement of enrichment factors and to the provision of a suitable amount of data for retesting. For the assay A procedures were demonstrated allowing to achieve the same results. Conclusively, the combined approach of unsupervised classification with a final supervised ranking is a well chosen strategy for identifying false-negative classes.

A question is, whether other classification techniques might be better suited for identifying false-negatives in HTS. This was addressed previously. It was shown that SVM, binary kernel discrimination, recursive partitioning and naïve Bayes' classifier can be applied in a similar context [Harper et al., 2001; Glick et al., 2005; van Rhee et al., 2001]. In a comparative study, SVMs outperformed recursive partitioning and naïve Bayes' classification [Glick et al., 2005]. Different in house studies at Boehringer Ingelheim showed the superiority of SVM and BRANN over PLS for classification. No marked differences in prediction accuracy were obtained when comparing SVM to BRANN [unpublished results]. As an advantage BRANN are computationally more efficient compared to SVM (one week on a single CPU compared to one month on 10 CPUs). Further pre-experiments employing a naïve Bayes' classifier approach of Pipeline Pilot [SciTegic, San Diego, USA] were not able to identify any of the false-negatives. This makes the author confident that the employed BRANN was a well chosen alternative for the classification.

The combined application of an unsupervised classification with a supervised ranking provided high enrichment factors. However supervised classification methods are available which can cope directly with data sets of the size of HTS assays. Examples are the naïve Bayes' classifier approach [Xia et al., 2004] or recursive partitioning [van Rhee et al., 2001]. They might offer an alternative to the proposed approach. However the pre-filtering of non-hits by the unsupervised classification (clustering) creates more focussed training sets. It assumes that more precise and accurate local models are created. As already proposed by Harper et al. a single method is not able to explain the SAR in HTS data [Harper et al., 2001]. In light of this the approach provides a new tool for the chemoinformatic toolbox. It is applicable to large data sets and it provides a rational for identifying false-negatives from primary screening data of HTS assays.

4.4 Prospective Analysis of Dopamine D_3 Receptor Antagonists

A virtual screen for new dopamine D_3 receptor-preferring antagonists (see 1.4) was performed employing both, NIPALSTREE and hierarchical *k*-means. In addition, self- organizing maps (SOM) were used. Molecules were ordered and K_i values were determined at both dopamine D_2 and D_3 receptors. Different computational methods were prospectively evaluated, namely pharmacophore-based virtual screening, docking and regression-based affinity prediction. 4.4 presents the results of the prospective evaluation.

4.4.1 Characterization of the Dopamine Data Set

Starting point of the data analysis was a characterization of the data set (see 3.1.5 and Appendix A) in combination with literature research on antagonists/partial agonists at dopamine D_2 like receptors. The aim was to identify ligands with different scaffolds, high affinity and selectivity for dopamine D_3 receptors. Results of the SAR are summarized in 1.4.3. The second characterization of the data set was regarding the K_i profile at dopamine D_2 and D_3 receptors. 386 compounds were selected having a K_i < 1 mM at both receptor subtypes. Histograms were created for the K_i values (Figure 4.4.1 A and B). The majority of the compounds has a K_i < 60 nM at both receptor subtypes. A more fine-grained histogram focusing on the K_i data range from 0.33 to 60 nM (Figure 4.4.1 C and D.) shows that the data set is shifted towards high affinity binding values below 20 nM for K_i values at dopamine D_3 receptors ($K_\mathrm{i}D_3$). The histogram for K_i values at dopamine D_2 receptors ($K_\mathrm{i}D_2$) shows a broad distribution in the range between 0.33 and 60 nM. When analyzing selectivity ratio histograms of either $K_\mathrm{i}D_3/K_\mathrm{i}D_2$ (dopamine D_2 receptor selectivity, Figure 4.4.1 E.) or $K_\mathrm{i}D_3/K_\mathrm{i}D_2$ (dopamine D_3 receptor selectivity, Figure 4.4.1 F.) with the focus on selectivity ratios above one, it can be seen that more compounds were selective towards dopamine D_3 receptors. Most compounds had only a low selectivity ratio below 10. However in the dopamine D_3 receptor selectivity histogram (Figure 4.4.1 F) a high amount of entries is present showing a selectivity ratio above 10 fold.

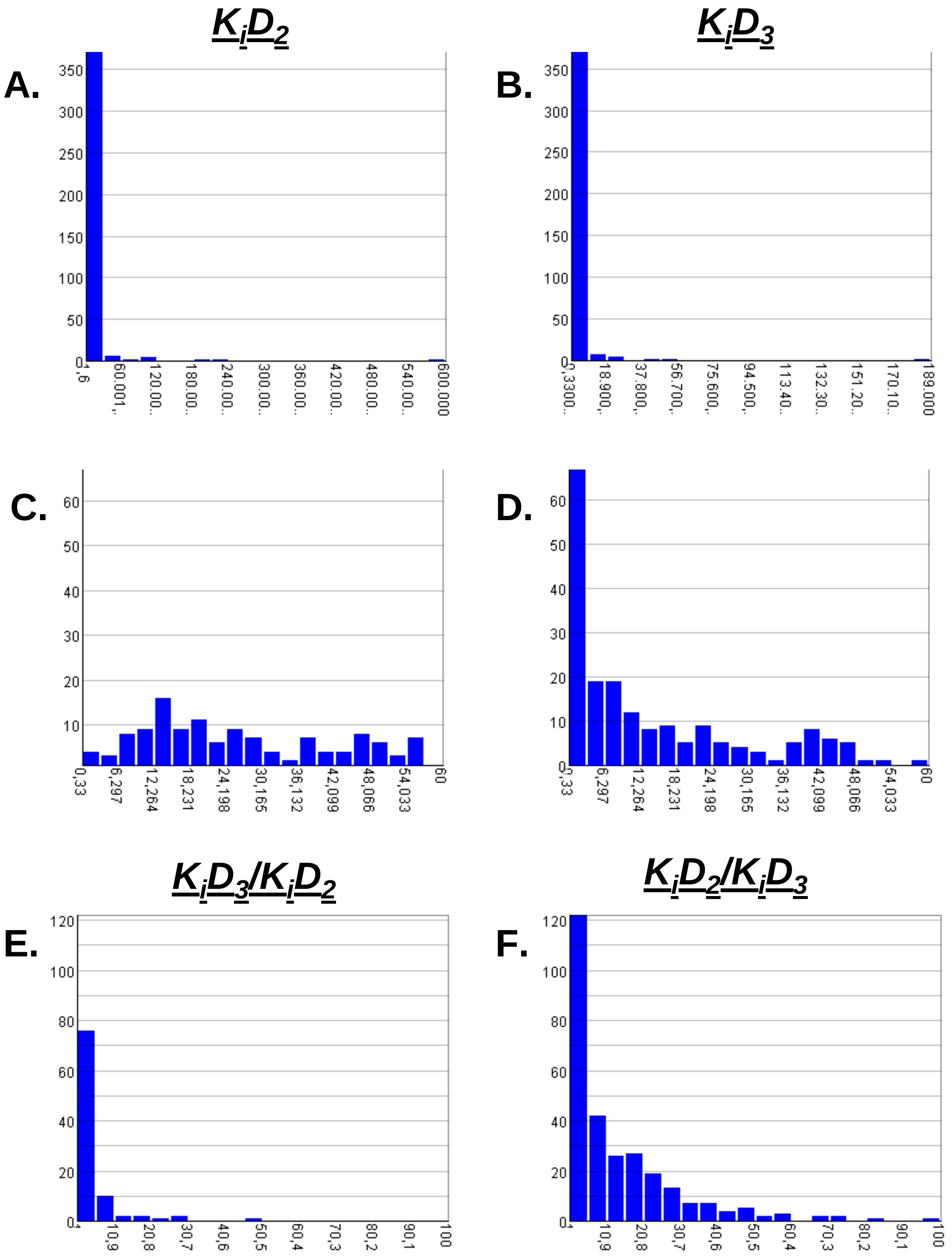


Figure 4.4.1 Histogram analysis of K_i values for dopamine D_2 and D_3 receptors in the dopamine data set (Appendix A). **A**. Histogram of all K_i values for dopamine D_2 receptors. **B**. Histogram of all K_i values for dopamine D_3 receptors. **C**. Histogram of K_i values for dopamine D_2 receptors. Only the data range between 0.3 nM and 60 nM is shown. **D**. Histogram of K_i values for dopamine D_3 receptors. Only the data range between 0.3 nM and 60 nM is shown. **E**. Histogram of the selectivity ratios of K_i values for dopamine D_3 receptors versus K_i values for dopamine D_2 receptors. Only ratios above one are shown. **F**. Histogram of the selectivity ratios of K_i values for dopamine D_2 receptors versus K_i values for dopamine D_3 receptors. Only ratios above one are shown.

4.4.2 Clustering-Based Virtual Screening

For the virtual screening experiments two data sets, DS_MOE and DS_CATS3D, were created (Table 4.4.1) consisting of the dopamine data set and the SPECS catalogue (version of June, 2003). Both descriptor sets were clustered employing the NIPALSTREE algorithm, the hierarchical *k*-means algorithm or a SOM. For both hierarchical clustering algorithms similarity thresholds Θ were determined as stop criterion as described in 4.1.5. The Euclidean metric was employed. The focus was on terminal clusters containing members of the dopamine set (see 4.3.3). Co-clustered SPECS molecules were further analysed. For the NIPALSTREE algorithm 37 additional SPECS molecules were identified and for the hierarchical k-means algorithm 144. The SOM approach was performed as illustrated in Figure 4.4.2.

Table 4.4.1 Data sets used for virtual screening.

	Dopamine D_3 + SPECS	Dopamine D_3 + SPECS
Data size	230,130	230,130
Descriptor set	MOE2D	CATS3D
Original number of descriptor	146	420
UFS R^2-based pruning	110	338
Entropy-based pruning	53	35
Final data set name	DS_MOE	DS_CATS3D
Threshold Θ	2.6	2.1

A SOM was trained with 30x20 neurons using the complete data set (that is, an average of 384 compounds per field). The obtained map was coloured according to the relative frequency of the DS_MOE data set (Figure 4.4.2 B, red over black to blue fields correspond to fields with a high, medium and low frequency respectively) and according to the relative frequency of the 472 dopamine D_3 receptor ligands (Figure 4.4.2 A). A clustering of the latter molecules can be seen, highlighted by the black box. All compounds belonging to these enriched fields were further considered (N = 1,551). A new map of size 15x10 was trained (that is an average of 10 compounds per field). The obtained map was coloured according to the relative frequency of the 1,551 molecules (Figure 4.4.2 D) and according to the relative frequency of the remaining dopamine D_3 receptor ligands (Figure 4.4.2 C). Fields were selected containing at least five dopamine receptor ligands. By that 52 co-clustered SPECS molecules were retrieved.

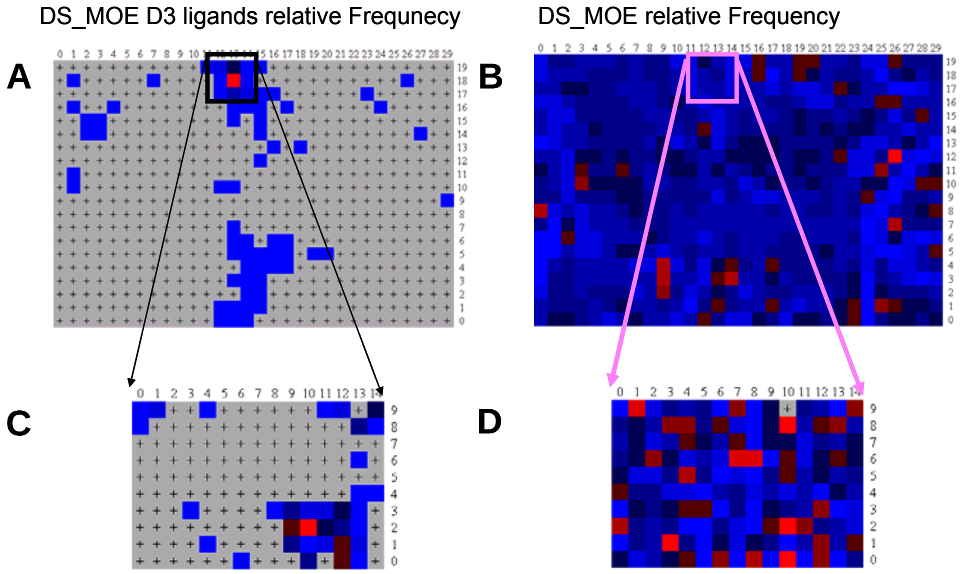


Figure 4.4.2 SOM based virtual screening of DS_MOE. **A, B** A SOM was trained with DS_MOE and coloured according the relative frequency of (**A)** dopamine D_3 ligands of DS_MOE and (**B)** all entries of DS_MOE. Red over black to blue represents a high to low frequency, respectively. The fields highlighted by the black box in **A** were enriched with dopamine D_3 receptor ligands. All corresponding entries in **B**, highlighted by the magenta box, were used for calculating a new SOM shown in **C** and **D**, with **C** being coloured according to relative frequency of dopamine D_3 receptor ligands and **D** being coloured according to relative frequency of all entries.

All obtained SPECS compounds were pooled, which were in total 207 molecules (N = 26 duplicates were avoided). By visual inspection, molecules were eliminated from the list which were too similar to known actives, did not possess a positively charged nitrogen essential for receptor binding [Hackling & Stark, 2002] or did not show drug-like properties [Muegge, 2003]. From the remaining list a maximum diverse subset of 17 molecules was chosen and ordered. The selected molecules have displayed calculated logP values in the range of 2.14 – 5.62 i.e. they possess lipophilicity in the range of central nervous system penetrating drugs [Bodor & Buchwald, 2003]. For this diverse subset of molecules binding affinities were determined by radioligand competition assays at dopamine D_2 and D_3 receptors. Results are listed in Table 4.4.2 for compounds **42-58**.

Table 4.4.2 Dopamine receptor affinities of compounds from the first and second virtual screening cycles.

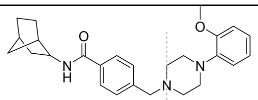
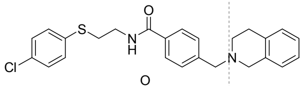
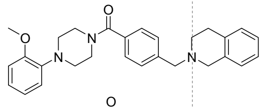
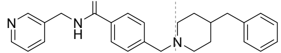
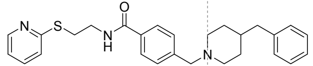
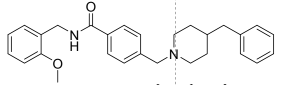
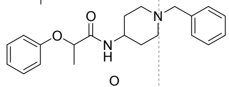
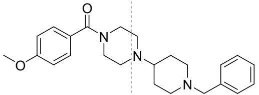
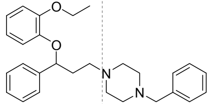
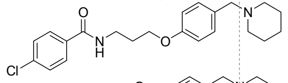
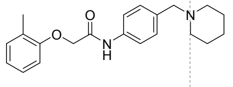
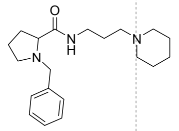
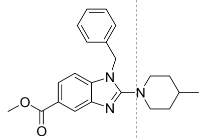
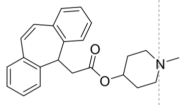
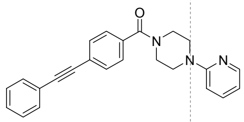
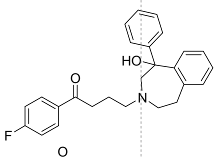
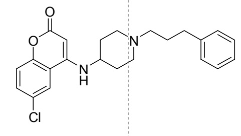
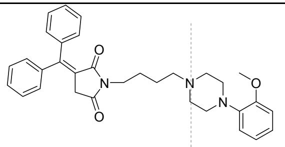
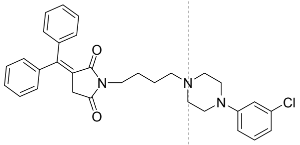
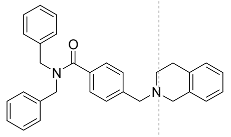
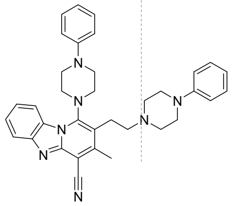
No.	Chemical Structure	log *P*	-log K_i (D_2)^a	-log K_i (D_3)^a	Ratio K_i (D_2/D_3)^a
42		4.18	6.10^b (6.05, 6.14)	7.19 ± 0.05^c	12.6
43		5.62	6.12^b (6.05, 6.18)	6.58^b (6.54, 6.61)	2.9
44		4.77	5.21^b (5.11, 5.31)	6.04^b (5.86, 6.21)	6.4
45		4.18	4.97 ± 0.22^c	5.6^b (5.49, 5.71)	4.6
46		5.39	5.40^b (5.47, 5.32)	6.60^b (6.67, 6.53)	15.9
47		5.39	5.59^b (5.55, 5.63)	6.24^b (6.24, 6.24)	4.5
48		2.93	6.04^b (5.96, 6.11)	6.02^b (6.01, 6.03)	1.0
49		2.72	4.83^b (4.66, 5.00)	5.36^b (5.34, 5.37)	3.6
50		5.43	6.12^b (6.10, 6.13)	6.65 ± 0.31^d	2.9
51		4.18	4.80^b (4.62, 4.97)	5.70^b (5.7, 5.69)	8.5
52		3.58	5.26^b (5.15, 5.36)	5.64^b (5.64, 5.64)	2.5
53		2.14	4.67 ± 0.27^c	5.23^b (5.46, 4.99)	3.4
54		4.93	4.44^b (4.97, 3.91)	5.27^b (5.18, 5.36)	12.3
55		3.84	5.38^b (5.49, 5.26)	6.74 ± 0.13^c	23.1
56		4.56	4.79^b (4.96, 4.62)	5.35^b (5.52, 5.18)	3.7
57		4.68	6.61^b (6.67, 6.55)	5.59^b (5.57, 5.61)	0.1
58		4.24	6.28^b (6.31, 6.24)	6.09^b (5.96, 6.21)	0.6

Table 4.4.2 (continued)

No.	**Chemical Structure**	**log P**	**-log K_i (D_2)^a**	**-log K_i (D_3)^a**	**Ratio K_i (D_2/D_3)^a**
59		4.78	6.80 ± 0.07^d	7.19 ± 0.06^c	2.5
60		5.46	5.87 ± 0.10^c	6.31 ± 0.07^d	2.7
61		6.53	5.87 ± 0.09^c	5.44^b (5.50, 5.39)	0.6
62		7.25	5.36 ± 0.03^c	5.58^b (5.54, 5.62)	1.7

$^a K_i$ values (mean value with standard deviation (SD)) were measured in CHO cells stably expressing hD_{2s} and hD_3 receptors in triplicates by using [^{3}H]spiperone. ^{b}Two independent experiments. ^{c}Three independent experiments. ^{d}Four independent experiments. All compounds were aligned according to the basic nitrogen.

4.4.3 SAR

Defining compounds with a K_i threshold below 1 µM a "hit", nine structures were active at dopamine D_3 receptors and six at dopamine D_2 receptors. Among the molecules five compounds possessed a K_i below 300 nM for dopamine D_3 receptors (**42**, **43**, **46**, **50**, **55**) and one compound (**57**) demonstrated a K_i value of 250 nM for dopamine D_2 receptors. Six molecules (**45**, **49**, **51**, **53**, **54**, **56**) totally lacked of affinity binding for dopamine D_2 receptors (> 10 µM) and demonstrated low affinity binding for dopamine D_3 receptors (2 - 7 µM).

Compound **42** was the top scoring molecule with a K_i value of 65 nM and a 13-fold preference for dopamine D_3 receptors. 14 out of 17 structures showed a preferred selectivity for dopamine D_3 receptors and compound **55** displayed the best selectivity ratio (23-fold) of dopamine D_3 versus D_2. By analysing the structures of these quite encouraging results, a novel promising structural feature, a benzamide moiety, was recognized and described for the first time. This benzamide element, incorporated as a linker in between the aryl moiety and the amine residue, was well tolerated by both dopamine D_2 and D_3 receptors. Benzamides have been described as aryl moieties in the atypical antipsychotics sulpiride and raclopride, showing high affinities for dopamine D_2 and D_3 receptors [Luedke & Mach, 2003].

It was published recently [Hackling et al., 2003] that the exchange of an alkyl chain into a rigid xylene spacer resulted in moderate to good affinity binding for dopamine D_2 and D_3 receptors. In the present study the training set also beard compounds with xylene spacer explaining the newly identified benzamide spacer. Benzamides are synthetically easy accessible and allow to produce a large variety of derivatives using parallel synthesis. By this an extended knowledge about (Q)SAR of the compound class might be easily derived.

Two novel structural features for the aryl moiety were identified, bicycle[2.2.1]heptane (**42**) and an aryl-thioether structure (**43**, **46**). Similarity to well-known drugs for the treatment of neuropsychiatric disorders can be seen in compound **55** and **57**. The former, bearing a dibenzocycloheptadiene residue, is closely related to tricyclic antipsychotic drugs (e.g. Clozapine or Olanzapine) [Härtter & Hiemke, 2002]. Compound **57** is a butyrophenone derivative. It is closely related to haloperidol [Abraham, 2003], a neuroleptic drug. It showed moderate affinity binding and a 10-fold selectivity for D_2 receptors, as seen for haloperidol [Abraham, 2003]. This dopamine D_2 receptor preference has also been noticed for **58**, a chromen-2-one derivative. Compound **48** has the same moderate affinity for both receptor subtypes. It can be explained by its short distance between the amide oxygen and the positively charged nitrogen. Compound **50** (K_i (D_3) = 264 nM) is unusual since the amide residue is completely missing. Instead, two ether oxygen atoms are conjugated with a phenyl ring. Partial charge calculations with the software package Gaussian [Gaussian Inc., Pittsburgh, USA] revealed that both ether oxygen atoms have – due to the aromatic conjugation – a partial charge comparable to the BP 897 amide oxygen (data not shown). This makes the author assume, that one of the ether atoms might overtake the role of the amide oxygen. The results have demonstrated that both hierarchical clustering algorithms and SOM were able to identify new and dopamine D_3 receptor-preferring ligands.

4.4.4 GOLD Docking

Docking techniques are designed to identify correct binding modes of a ligand in the binding pocket of a receptor. By using a scoring function the discrimination between strong binders (hits) and weak binders (non-hits) is assumed to be possible [Kitchen et al., 2004]. The scoring might be used as a post-processing filter for the clustering-based virtual screening. In addition to that an insight into the binding mode of the newly identified compounds can be obtained. This may allow identifying positions in a ligand which can be extended in order to optimize the binding affinity of a compound. Therefore docking was prospectively applied (i.e. prior to compound testing). A 3D homology model of the dopamine D_3 receptor was available due to a previous investigation and shown to produce meaningful results [Byvatov et al., 2005]. This model was employed for docking analyses. Compounds **42** – **58** were docked into the homology model using GOLD (see 3.7.1). Only positive score values were obtained indicating that all compounds fitted into the binding pocket. The Gold scores represented by the best binding mode of each molecule were plotted against the experimentally determined $\log_{10}K_i(D_3)$ values (Figure 4.4.3).

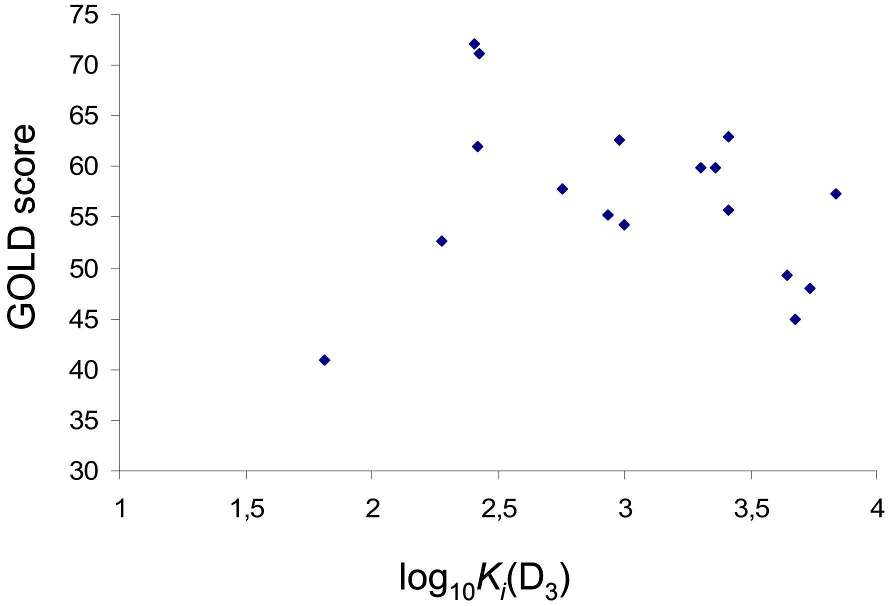


Figure 4.4.3 Plot of determined GOLD scores for the best binding modes of compounds **42** - **58** (*y*-axis) against experimentally determined $\log_{10} K_i(D_3)$ values (*x*-axis).

No correlation was observed ($R^2 = 0.03$), a result already described in literature as a general limitation of current docking algorithms [Warren et al., 2006]. In this context the approach is not suited as a post-processing filter and a manual analysis of the different binding modes is required to identify the putative accurate mode. During this binding mode analyses it has been observed that the ligands bind with their aryl moiety into two alternative binding pockets of the homology model of the dopamine D_3 receptor. This is exemplified in Figure 4.4.4 A for two different binding modes of compound **42** (yellow) and compound **55** (light-grey) and in Figure 4.4.4 B for BP 897 (grey) and the phenylpiperazinebenzoxazinone (green), previously identified in a SVM based virtual screening [Byvatov et al., 2005]. Close proximity of the ligands was observed to aspartic acid 110, phenylalanine 345, phenylalanine 346, serine 192 and threonine 369, which have been claimed to be important interaction partners of dopamine D_3 receptor ligands [Varady et al., 2003, Byvatov et al., 2005, Hackling et al., 2003]. A hydrogen bond was only observed between the positively charged nitrogen of the ligands to aspartic acid 110.

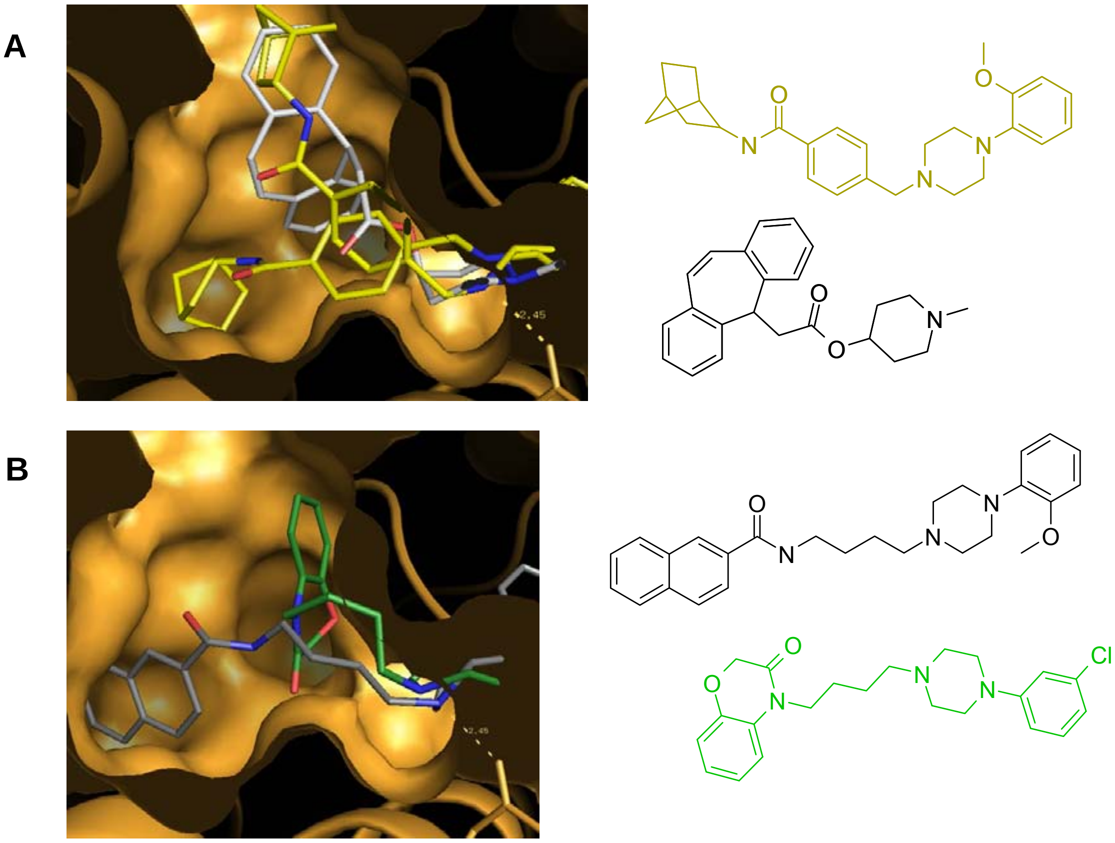


Figure 4.4.4 Two different binding modes of compound **42** (yellow, A**),** molecule **55** (light-grey, A), BP 897 (grey, B), and its morpholino analogue (green, B) docked into the homology model of the dopamine D_3 receptor. The ligands bind with their aryl moiety into alternative parts of the binding pocket.

4.4.5 Pharmacophore-Based Virtual Screening

In a preliminary experiment a dopamine D_3 antagonist pharmacophore model was created reflecting the 2D model in Figure 1.11 (see 1.4.3). It contains an aromatic PPP in the aryl moiety, an acceptor PPP at the position of the oxygen amide, a hydrophobic or aromatic PPP in the spacer region, an essential cationic PPP and an aromatic PPP in the amine rest (Figure 4.4.5). Three of the five potential pharmacophore points had to match for screening. A shape filter in terms of inclusion or exclusion volume was not specified. The model was validated employing two data sets. The first set contained 374 of the 386 dopamine ligands having K_i values at both receptors below 1 mM (the missing 12 compounds did not pass prior drug-like filters) and 1,500 randomly selected SPECS compounds. 27 of the SPECS compounds did not pass prior drug-like filters. 305 dopamine D_3 ligands were correctly predicted and only 6 additional SPECS compounds were identified. The 69 false-negative dopamine D_3 receptor ligands were analogues of a dopamine D_3 receptor agonist (pramipexole), which has been described to require a different pharmacophore model [Klabunde & Evers, 2005]. For further prospective validation the model was applied to compounds **42** - **58**. Setting a $K_i(D_3)$ threshold to 3 μM, 15 compounds were correctly classified and two false-positives occurred. The results suggest that a good antagonist model has been created and is well suited for virtual screening or as a post-processing filter of virtual screening. Both data sets contained dopamine D_3 and D_2 receptor-preferring ligands. They were all identified with the pharmacophore model. Consequently the model is not suited as a selectivity filter.

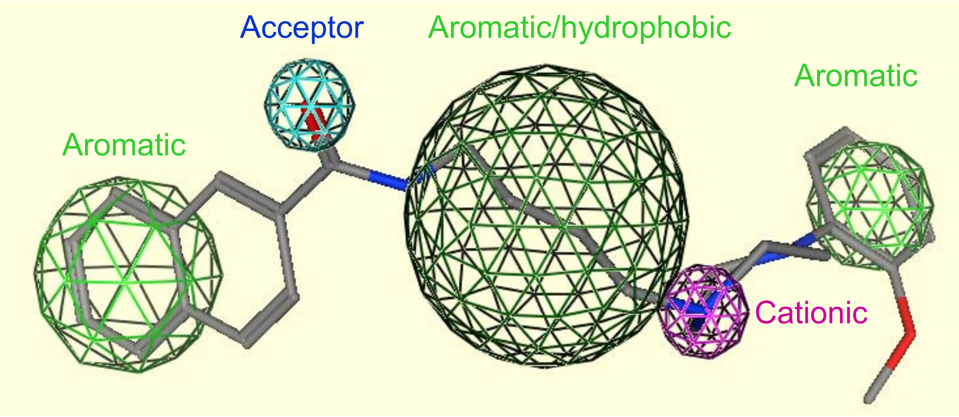


Figure 4.4.5 Dopamine D_3 receptor antagonist pharmacophore model in combination with BP 897.

To identify compounds which bind into both predicted aryl pockets, the pharmacophore model was extended as shown in Figure 4.4.6. In addition to the above described PPPs, an extra aromatic PPP was introduced describing the alternative aryl binding pocket. All PPPs were required as essential with the exception of the acceptor and the pharmacophore in the spacer region. Screening the entire SPECS catalogue 35 molecules were identified obeying the specified rules. To draw conclusions about the effect of the two aryl residues, three compounds remained being only different in the aryl pocket compared to antagonists in the dopamine data set. These molecules and an additional molecule (**62**) were ordered and experimentally tested at dopamine D_3 and D_2 receptors. Results are shown in Table 4.4.2 for compounds **59** - **62**.

Considering the additional aromatic residue for the alternative second binding pocket, only structure **59** and **60**, both containing a benzhydrylidene substituted pyrrolidindione residue, have shown good (**59**) and moderate (**60**) affinities and a slight dopamine D_3 receptor-preference. Consequently, planar rigidized molecules are tolerated at dopamine receptors. Compound **59** displayed a good affinity binding for dopamine D_3 receptors with a K_i value of 65 nM. The more flexible dibenzylcarbamoylbenzyl substituted 1,2,3,4-tetrahydroisoquinoline molecule **61** and the bulky benzimidazole substituted phenylpiperazine compound **62** clearly decreased affinity binding for both receptor subtypes. Too bulky features like in **62** might give a suitable explanation for the loss of affinity binding. To understand the missing affinity of compound **61** (yellow), it was aligned to **59** (red), giving the acceptor and the charged nitrogen a high weight (Figure 4.4.7). As can be seen in the alignment the branching point where both putative aryl elements are split is closer to the acceptor oxygen in **61** compared to **59**. This might give an explanation for the observed differences.

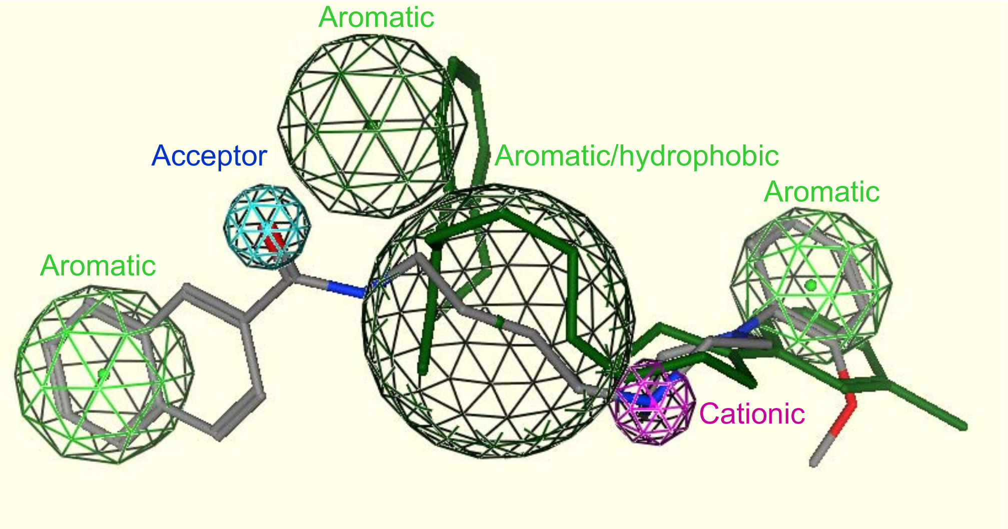


Figure 4.4.6 BP 897 and its morpholino analogue in combination with PPPs used for virtual screening of molecules containing both predicted aryl moieties.

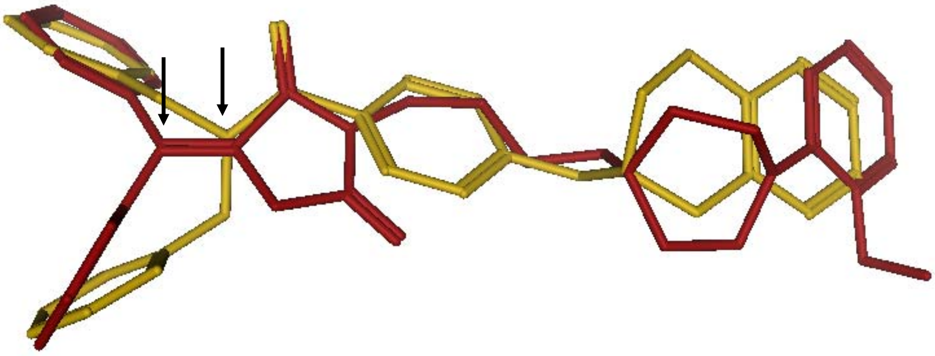


Figure 4.4.7 Molecule alignment of **61** (yellow) and **59** (red). The charge and acceptor moiety was given a higher weight. The black arrows indicate the branching points of the two putative aryl moieties.

In summary the combined application of homology modelling, docking into the homology model for hypothesis generation and pharmacophore based virtual screening translated into one compound out of four showing good affinity at dopamine D_3 receptors. Although the overall results are controversial it shows that the approach is suited to extend the current SAR and find novel lead structures.

4.4.6 Regression-Based Activity Prediction

Docking was not able to predict the affinity of compounds **42** - **58** correctly. It showed its value as idea generator by hinting on a putative alternative binding pocket in the dopamine D_3 receptor. Still a method predicting affinity values for compounds received via virtual screening would be worthwhile since it allows filtering false-positives. Ultimately it might help gain a deeper understanding of the QSAR of the identified compounds. The dopamine data sets included measured K_i values for both dopamine D_2 and D_3 receptors. It offered the possibility to employ these ligands to prospectively predict the affinities for compounds **42** - **58**. To keep the influence of experimental errors as low as possible, a focus was set on a high quality subset of 89 molecules, tested under the same condition as the ordered molecules. The data set was employed with either MOE 2D, CATS 2D or CATS 3D descriptors listed in Table 4.4.3. Descriptors were mean centred and scaled to unit variance. Further descriptors were selected according to UFS (R^2 = 0.99, see 3.3.5). As regression techniques one non-linear method, support vector based regression (SVR, see 3.5.2), and a linear technique, partial least squares (PLS, see 3.5.1) were used. Models were trained on pK_i values. Prior to model calculation a PCA was performed for the training set. Compounds **42** - **58** were projected onto the score vectors. According to “distance to model” calculations the molecules did not show outlier behaviour indicating that they are part of the models applicability domain [Eriksson et al., 2001]. Model qualities were assessed by calculating R^2 (goodness of fit) and Q^2 (goodness of prediction) values.

Table 4.4.3 Dopamine data sets used for regression based affinity predictions.

	Dopamine_MOE2D	Dopamine_CATS2D	Dopamine_CATS3D
Data size	89	89	89
Descriptor set	MOE2D	CATS2D	CATS3D
Original number of descriptors	146	150	420
UFS R^2-based pruning	38	46	75

PLS model training was performed as indicated in 3.5.1 for $pK_i(D_3)$ and $pK_i(D_2)$ values. As a training set all compounds listed in Table 4.4.3 were used. Q^2 values for the three different descriptor sets were obtained by 7-fold cross validation. As a validation set compounds **42** - **58** were employed. Results are shown in Table 4.4.4.

Table 4.4.4 Q^2 and R^2 values obtained by PLS for the dopamine data set

	MOE 2D	CATS 2D	CATS 3D
$R(D_2)^2$ (training set)$^\$$	0.55	0.5	0.74
$Q(D_2)^2$ (training set)$^\$$	0.51	0.41	0.56
$R(D_2)^2$ (validation set)§	0.09	0.01	0.33
$R(D_3)^2$ (training set)$^\$$	0.27	0.33	0.33
$Q(D_3)^2$ (training set)$^\$$	0.16	0.15	0.18
$R(D_3)^2$ (validation set)§	0	0.02	0.01

$^\$$Compounds of Table 4.4.3 were employed. §Compounds **42** - **58** were employed

SVR model training was performed as indicated in 3.5.2 for p$K_i(D_3)$ and p$K_i(D_2)$ values. The data sets listed in Table 4.4.3 were divided into 25% test set and 75% training set either based on a maximum diverse selection (see 3.3) or on random selection. In total three different random sets (Random 1 – Random 3) and one maximum diverse set (SCA) were created. The support vector regression model was selected best predicting the test set (maximum R^2). As a validation set compounds **42** - **58** were employed. Resulting R^2 and Q^2 values are present in Table 4.4.5. In Figure 4.4.8 all R^2 values are represented as a histogram obtained with the methods for the validation set. It contains six sections, one for the PLS models, four for the different SVR models (SCA and random 1 - 3) and one for a majority voting of all models for both receptor subtypes. Beneath each other the p$K_i(D_3)$ (red) and p$K_i(D_2)$ (blue) R^2 histogram bars are located for each descriptor set, starting with MOE 2D descriptors, over CATS 2D descriptors to CATS 3D descriptors.

R^2 of the Training and Test Set and Q^2 of the Training Set

Support vector machines are intended to create models with high R^2 values for the training set [Smola & Schölkopf, 1998]. They reached their aim in the present study since R^2 values were in all cases above 0.8. For further discussion the focus was set on the more meaningful Q^2 of the training set and R^2 of the test and validation sets. For the PLS models both R^2 and Q^2 of the training set and R^2 of the validation set were considered. When analyzing the R^2 and Q^2 values, predictions were clearly better for pK_i (D_2) values. The Q^2 for pK_i (D_2) exceeded, with one exception (PLS with CATS2D) 0.5 indicating a predictive ability for pK_i values at dopamine D_2 receptors. This was different for the pK_i (D_3) models, where both Q^2 and R^2 values were low. Since the data set was shifted towards high affine ligands for the D_3 receptor it might be concluded that pK_i (D_3) models were either sensitive to outliers in the biological response or that the data range in the biological response was not coarse grained enough for robust model training. Neither of the created models allowed a robust prediction for the validation set. For the pK_i (D_3) prediction this was already indicated by the low R^2 and Q^2 values. For the pK_i (D_2) models only the R^2 and Q^2 values of the training set of the PLS models allowed drawing such conclusion.

Effects of the Test Set Selection

For the SVR approaches the training set was selected based on a maximum diversity algorithm and on random selection. To analyse whether any of the selection methods might be given a favour Q^2 of the training sets or R^2 of test and validation set were examined. To date no clear advantage can be given for any of the methods. However for a valid conclusion more examinations would be required employing different diversity selection algorithms, different data sets and differently large test and training sets.

Descriptor Preference for the Model Creation

One point of interest was whether a preference can be given for any of the descriptor sets. With the exception of CATS 3D descriptors in combination with PLS and pK_i (D_2) values no noticeable differences in the R^2 and Q^2 values for the test and training sets were observed for either of the descriptor sets. For the validation set and pK_i (D_2) values the SVR models obtained with CATS 3D or CATS 2D descriptors tended to outperform models obtained with the MOE 2D descriptors. For the PLS models only the CATS 3D descriptors showed this tendency. For pK_i (D_3) prediction in the validation set a preference might be given for the CATS 2D descriptors in combination with SVR.

Model Preference

To analyse whether any of the regression methods might be given a preference, R^2 and Q^2 values were examined for the training and test sets. SVRs outperformed PLS, an observation already described elsewhere [e.g. Zhao et al., 2004; Sorich et al., 2003]. However this was given to the cost of introducing non-linearity. This makes later interpretation more difficult. When judging the methods according their capability of predicting the validation set none of the methods was able to perform overall correct predictions, since R^2 did not exceed 0.34. All methods and all descriptor sets contributed at least one best prediction (data not shown). If a preference should be given for any of the descriptor sets and regression methods, the CATS 2D descriptors in combination with SVR seemed to perform best for pK_i(D_3) and CATS 3D in combination with PLS for pK_i(D_2). A majority voting combining all models leads for the pK_i (D_2) prediction to a R^2 comparable to the best method. For pK_i (D_3) prediction R^2 drops markedly ($R^2 = 0.06$). The reason for this remains elusive to the author.

Table 4.4.5 Q^2 and R^2 values obtained by SVR for the dopamine data set.

	SCA			RANDOM 1			RANDOM 2			RANDOM 3		
	MOE2D	CATS2D	CATS3D	MOE2D	CATS2D	CATS3D	MOE2D	CATS2D	CATS3D	MOE2D	CATS2D	CATS3D
$R(D_2)^2$ (training)[\$]	0.92	0.99	0.93	0.92	0.87	0.98	0.91	0.96	0.99	0.96	0.96	0.89
$Q(D_2)^2$ (training) [\$]	0.66	0.6	0.56	0.64	0.61	0.59	0.63	0.59	0.63	0.66	0.69	0.59
$R(D_2)^2$ (test) [\$]	0.7	0.7	0.67	0.7	0.44	0.68	0.78	0.76	0.65	0.62	0.65	0.47
$R(D_2)^2$ (validation)[§]	0.04	0.09	0.13	0.05	0.27	0.25	0.05	0.15	0.23	0.03	0.27	0.18
$R(D_3)^2$ (training) [\$]	0.91	0.95	0.99	0.9	0.9	0.99	0.99	0.96	0.99	0.97	0.92	0.99
$Q(D_3)^2$ (training) [\$]	0.31	0.42	0.42	0.32	0.4	0.45	0.39	0.49	0.52	0.4	0.48	0.42
$R(D_3)^2$ (test) [\$]	0.58	0.62	0.41	0.62	0.6	0.56	0.41	0.39	0.45	0.39	0.48	0.44
$R(D_3)^2$ (validation) [§]	0.17	0.3	0	0.07	0.28	0.03	0.19	0.34	0.05	0.09	0.15	0.01

[\$] Compounds from Table 4.4.3 were employed. [§] Compounds **42-58** were used.

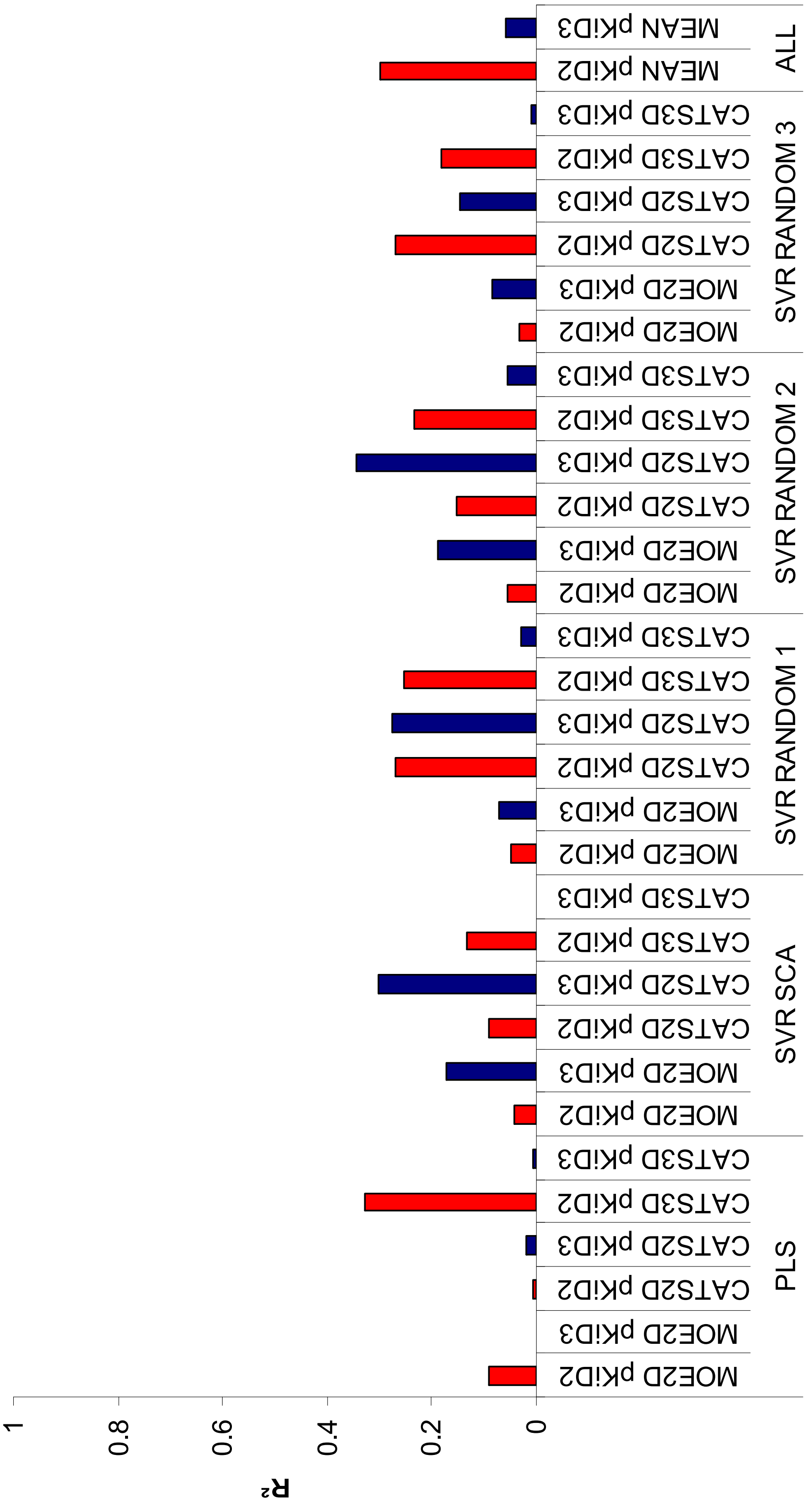


Figure 4.4.8 R^2 values obtained for compounds **42** - **58**. As regression technique PLS and support vector based regression in combination with four different training sets (see 3.8) were used. Besides each other R^2 values are shown for $pK_i(D_2)$ (red) and $pK_i(D_3)$ (blue) obtained for the MOE 2D, CATS 2D and CATS 3D descriptor sets (Table 4.4.3). The last section shows a majority voting of all methods and all descriptor sets.

The top scoring PLS model for p*K*$_i$ (D$_2$) in combination with the CATS3D descriptor set was employed to analyze descriptor importance according to “variable importance plots” (VIP) and coefficients of loadings [Eriksson et al., 2001]. It was possible to confirm observations already presented in literature: (i) a distance of 6 to 7 Å between the acceptor oxygen of the amide and the basic nitrogen is favourable for D$_3$ [Hackling et al., 2003]. This was expressed by a high weighting of the +A07 descriptor (positively charged moiety and acceptor group with a distance of 6-7 Å). (ii) The spacer defines the distance between the amine moiety and aryl rest [Hackling et al., 2003; Newman et al., 2005]. This is expressed by the descriptor HH12 requiring the distance of two hydrophobic residues to be between 11 and 12 Å. (iii) The dopamine data set contained several analogues of pramipexole, which is a dopamine D$_3$ receptor-preferring agonist [Schneider & Mierau, 1987; Biglan & Holloway, 2002; Kushida, 2006] (Table 1.1). In the PLS model descriptors +H04 (positively charged moiety and hydrophobic group with a distance of 3-4 Å) and AH03 (acceptor atom and hydrophobic group with a distance of 2-3 Å) were given a high weight and were present only in this class of analogues. A recent publication described potential pharmacophore points for dopamine D$_3$ receptor agonists including pramipexole [Elsner et al., 2005]. Both identified features, +H04 and AH03 fit nicely into their models. It shows that by analysis of VIP and of the coefficients of loadings it is possible to simultaneously identify important features in different structural classes.

Summarizing results showed that by employing different descriptor sets and different regression techniques robust models can be created. However their application to structurally diverse sets obtained by virtual screening was of limited success. Several reasons might be thought of explaining the low R^2 values for compounds **42 - 58**. One is that the training set was not diverse enough to construct predictive models for this data set. Another is that the training and test sets were optimized for dopamine D$_3$ receptor affinity whereas the validation set contained only five highly affine compounds with *K*$_i$ (D$_3$) below 300 nM. A regression based affinity prediction is usually performed for structural classes during lead optimization. Here it was tried to predict activities for novel and diverse structural classes in lead identification. The conclusion was that in this context an affinity prediction for virtual screening data is not accurate enough to provide reliable results.

4.4.7 Conclusions

In this study, for the first time, NIPALSTREE and the hierarchical *k*-means were prospectively applied to identifying false-negatives. In combination with self organizing maps novel lead candidates for dopamine D_3 receptors were identified. As new structural motive, a benzamide was recognized and described for the first time in context of a linker in between the aryl moiety and the amine residue. It was well tolerated by both dopamine D_2 and D_3 receptors. Two novel structural features for the aryl moiety were identified, bicycle[2.2.1]heptane (**42**) and an aryl-thioether structure (**43**, **46**). This confirms the observation of the retrospective HTS analyses that false-negatives with novel scaffolds can be identified employing the clustering techniques.

The clustering based virtual screening was performed employing a data set containing at most dopamine D_3 receptor-preferring ligands (see Figure 4.1.1 F). Not all newly identified compounds were active and even less were selective towards dopamine D_3 receptors. Different attempts were tried, either to colour the calculated SOMs or the resulting cluster dendrograms according to the selectivity of $K\text{i}(D_3)/K\text{i}(D_2)$ or vice versa. It was not possible to identified clusters enriched with selective compounds (data not shown). Consequently an unsupervised classification of the ligands based on the descriptors used in this study is not sufficient to explain selectivity. This is supported by the fact that the regression methods trained on selectivity and applied to compounds **42** – **58** were not able to address selectivity correctly ($R^2 < 0.05$, data not shown). It is assumed that more complex models are required employing exact 3D structures of the ligands (bound to the respective receptor and unbound) and quantum chemical descriptors/calculation to enlighten new discriminating molecular features for both receptors. Until now these methods are too time- and space-consuming for early phase screening applications.

Docking studies were committed employing a dopamine D_3 receptor homology model and compounds **42** – **58**. The approach was not able to accurately distinguish active from not active molecules. Possible reasons might be the limited set of docked compounds not allowing to draw statistically relevant conclusions, the potential error in the homology model or the putative lack of parameterization of the scoring function for this receptor. The clear advantage of docking is the visualization of ligands in the binding pocket leading in the present study to the identification of two putative binding pockets for the aryl moiety of dopamine D_3 receptor antagonists. Although experimental results obtained for molecules simultaneously requiring both aryl binding pockets are controversial, it shows that docking analyses have great impact as an idea generator and are thus irreplaceable in the drug discovery process.

A pharmacophore model was constructed for dopamine D_3 receptor ligands and prospectively and retrospectively evaluated**.** The results showed that a robust antagonist model was created filtering hits from non-hits. To enlighten the role of the different binding modes, the pharmacophore model was extended simultaneously requiring both binding modes. After screening the SPECS catalogue four molecules remained. The best compound possessed a K_i value of 65 nM at dopamine D_3 receptors. The data successfully shows how to translate new ideas, made by docking, into a model allowing to extend the current SAR.

Two regression techniques, PLS and SVR, were employed with three different descriptor sets (MOE 2D, CATS 2D and CATS 3D). Predictive models were obtained for dopamine D_2 and D_3 pK_i values. This is underlined by the VIP analysis for the PLS model with CATS 3D descriptors and pK_i (D_2) values. It was possible to identify descriptors explaining SAR already made in literature for dopamine D_3 antagonists and agonists [Hackling et al., 2003, Newman et al., 2005, Elsner et al., 2005]. The prospective application of the created models to compounds **42** – **58** was not able to perform overall correct predictions. Even so outlier behaviour of the compounds was checked prior to the prediction. The conclusion is that an activity prediction for diverse and novel (virtual) screening data is not accurate enough to provide reliable results in this context.

An objective of the study was to examine the applicability of virtual screening methods in early stages of drug discovery process for the generation of structurally new leads. The clustering approaches provided new compounds, the docking served as an idea generator for a pharmacophore model providing a new insight into SAR and the regression methods were able to identify important molecular features. It shows that different methods are necessary to explain the SAR in the data [Harper et al., 2001]. By strategic combination of the techniques a successful finding of novel "lead" candidates for the dopamine D_3 receptor was possible. To clarify whether both hypothesized aryl pockets exist and whether they can be used to design ligands with more suited pharmacokinetic and pharmacodynamic properties has to be addressed by chemical synthesis of structural analogues. This "lead optimization" might be supported by de novo design techniques allowing a further fine tuning [Fechner & Schneider, 2005].

5 Conclusion and Outlook

The scope of the thesis was to identify SAR in the primary screening data of HTS assays. By hierarchical clustering of the compounds, assigning the primary screening data to the clusters and employing the clusters in combination with their relationship to each other, models should be derived which identify false-negatives, not confirmed hits, singletons and clusters enriched with hits. For this purpose two hierarchical clustering algorithms, NIPALSTREE and hierarchical *k*-means, have been developed. A GUI was implemented for working with the clustering results. Both retrospective and prospective applications of the clustering approach were performed. A combination of clustering with different computational methods was committed to obtain an extended SAR. What were the key findings and what were the main conclusions? Are primary screening HTS data suited for extraction of SAR? Is the clustering-based approach applicable to find SAR in HTS?

The two hierarchical clustering algorithms, NIPALSTREE and hierarchical k-means were developed. They demonstrated the successful application and clustered large data sets with more than 700,000 data points. Both clustering methods were validated and compared to each other. The hierarchical *k*-means algorithm seemed to outperform NIPALSTREE. The conjunctive application of both clustering algorithms was able to improve *EFs* without loosing high proportions of lead classes. The NIPALSTREE algorithm provides the loading vector which allows drawing a conclusion about the importance of descriptors. First insights into SAR were gained. According to this no clear preference can be given for any of the algorithms. Future investigations might aim at the extension of the hierarchical *k*-means to cope with fingerprints [Engels et al., 2006], the re-clustering of the clustering results using a maximum common substructure approach [Nicolaou et al., 2002] and the comparison of both algorithms to other hierarchical clustering algorithms. However these algorithms are at most characterized by a quadratic complexity and are not applicable to data sets with more than 100,000 data points.

A GUI was developed allowing the display of and the navigation in the clustering results. It provides functionalities to analyse the dendrogram, a cluster in the dendrogram and the molecules in a cluster. Measures were introduced to identify clusters enriched with actives, to characterize singletons and to analyse selectivity and specificity. The scaling of *EF*s for an inhibitor class to the logarithm of the dendrogram level was able to simultaneously retrieve and separate the different structural classes of the inhibitor class, irrespective of the dendrogram level. By analysing co-clustered molecules an extended SAR was obtained. It offers the possibility to construct focussed screening libraries. Singletons in the dendrogram were investigated as a source of alternative lead structures. It was possible to reduce the number of singletons by additional similarity searching and nearest neighbour analysis. However for a final conclusion a manual analysis was necessary. Relative frequencies of different protease inhibitor classes were extracted from clusters of a dendrogram level. The values were clustered and compared to a phylogenetic multiple sequence alignment of the corresponding proteins. Both overlap and differences in the dendrograms were observed. This new clustering provides an alternative ligand-based view on the relationships of the binding pockets of a set of proteins. A deeper insight into SAR of an inhibitor class was obtained, providing hints about adverse side effects or related enzymes for which counter screens should be performed [Arnold et al., 2004]. The GUI has reached a state where it is possible to draw a conclusion about the SAR in the clusters. A variety of extensions of the GUI might be considered in the future. This might include approaches to further visualize SAR like R-group decomposition [Kibbey & Calvet, 2005] or the supporting of the singleton analysis with a maximum common substructure comparison [Stahl & Mauser, 2005].

Three HTS assays were analyzed in a retrospective study. Comparing the % CTL values to the IC_{50} values it is obvious that no quantitative conclusion can be drawn from % CTL values. It was possible to confirm the theoretical considerations of Zhang et al.,: the closer the % CTL value of a primary screening hit to the hit threshold, the higher is the likelihood, that it translates into a not confirmed hit [Zhang et al., 2000]. A valid conclusion is that the same holds true for false-negatives. The hierarchical *k*-means was used to cluster the data and analyze its capacity to retrospectively identify not confirmed hits and false-negative scaffolds. No correlation was observed between cluster size (i.e. between similar entries), the number of hits in a cluster and the proportion of not confirmed hits. A rule catalogue was implemented rating hits in terminal clusters based on the cluster size, the % CTL values of the entries in a cluster, the overall hit rate, the hit rate in the cluster and the environment of a cluster in the dendrogram. With this approach it was possible to identify not confirmed hits. Further the method provides a way of rating hits in context of non-hits [Schreyer et al., 2004]. This is a unique ability of the hierarchical clustering. False-negative scaffolds were created and the data were clustered. Compounds were extracted from terminal clusters containing hits. It was possible to co-extract and enrich false-negatives. To minimize the number of false-positives in the extracted lists, BRANN classification models were trained with the data. Applying the models clearly improved the enrichment factors of the false-negative scaffolds. It can be concluded that the combined approach of unsupervised pre-classification with a final supervised ranking is a well chosen strategy to identifying and enriching false-negatives in HTS assays. As an outlook the approach bears the opportunity to construct a local classification model for each cluster. By integration of all models it might be possible to obtain overall higher prediction accuracies. This has been already demonstrated for smaller data sets by combining self organizing maps with supervised neural networks [Spycher et al., 2005; Gini et al., 2004].

NIPALSTREE, hierarchical *k*-means and SOM were prospectively applied to identify novel lead candidates for dopamine D_3 receptors. It was possible to retrieve compounds with novel scaffolds and low nanomolar binding affinity (65 nM for compound **42**) confirming the retrospective HTS analysis. Different computational methods were examined for their applicability to generate structurally new leads and explain SAR. Both strength and limitations of each technique were observed. Docking studies were performed. The visual inspection of the binding modes revealed the hypothesis of two alternative binding pockets for the aryl moiety of dopamine D_3 receptor antagonists. However it was not possible to distinguish “hits” from “non-hits”. A pharmacophore model was created which simultaneously required both aryl moieties. Virtual screening by applying this model identified a nanomolar hit (65 nM) corroborating the hypothesis. SVR and PLS were examined. It was possible to create predictive models for dopamine D_2 and D_3 pK_i values. A VIP analysis was performed for one PLS model. Descriptors explaining SAR were identified [Hackling et al., 2003, Newman et al., 2005, Elsner et al., 2005]. The prospective application of the models to diverse and novel virtual screening data was not able to predict activity correctly. The data clearly shows that different methods are able to explain different parts of the SAR [Harper et al., 2001]. Key to success is their combined application employing the strength of each method.

Is the proposed clustering based approach suited to identify SAR in the data and, is it possible to identify SAR in the primary screening data of HTS assays? The presented data shows that SAR can be identified in HTS with the clustering algorithms. They can be used to identify false-negatives. The data also suggest that limitations are present resulting from both HTS data and the clustering approach. From the data obtained with the BRANN classification and the dopamine D_3 receptor virtual screening it is concluded that each method has its own strength and that each method provides a different view on the SAR in HTS. This mirrors a multi-disciplinary approach viewing the data from different perspectives. Such multi-disciplinary settings with computational chemistry as one integral part are necessary to successfully pass the hurdles of the early phase drug discovery and decipher new lead series.

6 Summary

The aim of the thesis was to identify structure activity relationships (SAR) in the primary screening data of high-throughput screening (HTS) assays. The strategy was to perform a hierarchical clustering of the molecules, assign the primary screening data to the created clusters and derive models from the clusters. The models should serve to identify singletons, clusters enriched with actives, not confirmed hits and false-negatives. Two hierarchical clustering algorithms, NIPALSTREE and hierarchical *k*-means have been developed and adapted for this purpose, respectively. A graphical user interface (GUI) has been implemented to extract SAR from the clustering results. Retrospective and prospective applications of the clustering approach were performed. SAR models were created by combining the clustering results with different chemoinformatic methods.

NIPALSTREE projects a data set onto one dimension using principle component analysis. The data set is sorted according to the scoring vector and split at the median position into two subsets. The algorithm is applied recursively onto the subsets. The hierarchical *k*-means recursively separates a data set into two clusters using the *k*-means algorithm. Both algorithms are capable of clustering large data sets with more than a million data points. They were validated and compared to each other on the basis of different structural classes. NIPALSTREE provided with the loading vectors first insights into SAR whereas the hierarchical *k*-means yielded superior results.

A GUI was developed allowing the display of and the navigation in the clustering results. Functionalities were integrated to analyse the clusters in the dendrogram, molecules in a cluster, and physicochemical properties of a molecule. Measures were developed to identify clusters enriched with actives, to characterize singletons and to analyse selectivity and specificity. Different protease inhibitors of the COBRA database were examined using the hierarchical *k*-means algorithm. Supported by similarity searches and nearest neighbour analyses thrombin inhibitor singletons were quickly isolated and displayed in the dendrogram. By scaling enrichment factors to the logarithm of the dendrogram level, clusters enriched with different structural classes of factor Xa inhibitors were simultaneously identified. The observed co-clustering of other protease inhibitors provided a deeper insight into selectivity and specificity and shows the utility of the approach for constructing focussed screening libraries. Specificity was analyzed by extracting and clustering relative frequencies of the protease inhibitors from the clusters of dendrogram level 7. A unique ligand based point of view on the pocketome of the protease enzymes was obtained.

To identify not confirmed hits and false-negatives in the primary screening data of HTS assays, three assays were retrospectively analysed with the hierarchical *k*-means algorithm. A

rule catalogue was developed judging hits in terminal clusters based on the cluster size, the percent control values of the entries in a cluster, the overall hit rate, the hit rate in the cluster and the environment of a cluster in the dendrogram. It resulted in the identification of a high proportion of not confirmed hits and provided for each hit a rating in context of related non-hits. This allows prioritizing compounds for follow-up studies. Non-hits and hits were retrieved from terminal clusters containing hits. Molecules bearing false-negative scaffolds were co-extracted and enriched. To minimize the number of false-positives in the extracted lists, Bayesian regularized artificial neutral network classification models were trained with the data. Applying the models marked improvement of enrichment factors for the false-negatives was obtained. It proofs the scaffold-hopping potential of the approach.

NIPALSTREE, the hierarchical *k*-means algorithm and self-organising maps were prospectively applied to identify novel lead candidates for dopamine D_3 receptors. Compounds with novel scaffolds and low nanomolar binding affinity (65 nM, compound **42**) were identified. To provide a deeper insight into the SAR of these molecules, different alternative computational methods were employed. Support vector-based regression and partial least squares were examined. Predictive models for dopamine D_2 and D_3 receptor binding affinity values were obtained. Important features explaining SAR were extracted from the models. The prospective application of the models to the diverse and novel virtual screening data was of limited success only. Docking studies were performed using a homology model of the dopamine D_3 receptor. The visual inspection of the binding modes resulted in the hypothesis of two alternative binding pockets for the aryl moiety of dopamine D_3 receptor antagonists. A pharmacophore model was created simultaneously requiring both aryl moieties. Virtual screening with the model identified a nanomolar hit (65 nM, compound **59**) corroborating the hypothesis of the two binding pockets and providing a new lead structure for dopamine D_3 receptors.

The presented data shows that the combined approach of hierarchically clustering a data set in combination with the subsequent usage of the clusters for model generation is suited to extract SAR from screening data. The models are successful in identifying singletons, clusters enriched with actives, not confirmed hits and false-negative scaffolds.

7 Zusammenfassung

Das Ziel der Arbeit war es, Struktur-Aktivitätsbeziehungen (SAR) in primären Screeningdaten von Hochdurchsatzscreening (HTS)- Assays zu finden. Als Strategie sollten die Moleküle hierarchisch geclustert werden, die primären Screeningdaten den gebildeten Clustern zugeordnet und Modelle aus den Clustern abgeleitet werden. Die Modelle sollten das Auffinden von Singletons, mit Hits angereicherter Cluster, nicht bestätigter Hits und falsch Negativer ermöglichen. Zu diesem Zweck wurden zwei hierarchische Clusteralgorithmen, NIPALSTREE und hierarchischer *k*-means, entwickelt bzw. angepasst. Eine graphische Benutzeroberfläche (GUI) wurde implementiert, um SAR aus den Ergebnissen der Clusterung abzuleiten. Retrospektive und prospektive Anwendungen wurden mit den Clusteransätzen verfolgt. SAR Modelle wurden durch Verwendung der Ergebnisse der Clusterung mit verschiedenen chemoinformatischen Verfahren erstellt.

NIPALSTREE projiziert mit Hilfe der Hauptkomponentenanalyse einen Datensatz auf eine Dimension. Der Datensatz wird anhand des Scoringvektors sortiert und, basierend auf dem Median, in zwei Teilmengen aufgetrennt. Der Algorithmus wird rekursiv auf die neu gebildeten Mengen angewandt. Der hierarchische *k*-means Algorithmus trennt, basierend auf dem *k*-means Algorithmus, einen Datensatz rekursiv in zwei Cluster auf. Beide Algorithmen sind in der Lage, große Datenmengen mit mehr als einer Million Datenpunkte zu clustern. Sie wurden anhand verschiedener Strukturklassen validiert und miteinander verglichen. NIPALSTREE erbrachte mit dem Loadingvektor erste Einblicke in die SAR, wohingegen der hierarchische *k*-means zu besseren Ergebnissen führte.

Eine GUI wurde entwickelt, die es erlaubt, die Clusterergebnisse darzustellen und darin zu navigieren. Funktionalitäten wurden bereitgestellt, um die Cluster im Dendrogramm, die Moleküle eines Clusters und die physikochemischen Eigenschaften eines Moleküls zu analysieren. Verfahren wurden entwickelt, um mit Hits angereicherte Cluster zu finden, Singletons zu charakterisieren und Selektivität und Spezifität zu analysieren. Verschiedene Proteaseinhibitoren aus der COBRA-Datenbank wurden mit dem hierarchischen *k*-means Algorithmus näher betrachtet. Mit Hilfe von Ähnlichkeitssuchen und nächsten Nachbaranalysen wurden Thrombininhibitorsingletons im Dendrogram in kürzester Zeit isoliert und dargestellt. Cluster, die mit verschiedenen Strukturklassen von Faktor-Xa-Inhibitoren angereichert waren, wurden, durch Skalierung des Anreicherungsfaktors auf den Logarithmus der Dendrogrammebene, gleichzeitig im Dendrogramm identifiziert. Eine Clusterung der Faktor-Xa-Inhibitoren mit anderen Proteaseinhibitoren wurde beobachtet. Sie erbrachte einen vertieften Einblick in Selektivität und Spezifität und zeigt die Anwendbarkeit des Ansatzes zur Erstellung fokussierter Screeningbibliotheken. Durch Extrahierung und Clusterung der relativen Anteile der Proteaseinhibitoren aus den Clustern von Dendrogrammebene sieben wurde die Spezifität der Proteaseinhibitoren analysiert. Eine spezifische, Liganden basierte Betrachtung des Pocketoms der Proteaseenzyme wurde erhalten.

Um nicht bestätigte Hits und falsch Negative in den primären Screening Daten von HTS Assays zu finden, wurden drei Assays in Retrospektive mit dem hierarchischen *k*-means analysiert. Ein Regelwerk wurde entwickelt, welches Hits anhand der Clustergröße, des Prozent-Kontrollwertes der Einträge eines Clusters, der Gesamthitrate, der Hitrate in einem Cluster und der Umgebung des Clusters im Dendrogramm bewertet. Das Regelwerk führte zum Auffindung eines großen Anteils nicht bestätigter Hits. Zudem wurde für jeden Hit eine Bewertung im Kontext verwandter Nichthits erhalten. Dies erlaubt ein Priorisieren von Molekülen für Folgeuntersuchungen. Nichthits und Hits wurden aus Endcluster, die Hits enthielten, extrahiert. Moleküle mit falsch negativen Molekülgrundgerüsten wurden koextrahiert und angereichert. Um falsch Positive in den extrahierten Listen zu minimieren, wurden Bayesische regularisierte neuronale Klassifizierungsnetze mit den Daten trainiert. Die Anwendung der Modelle ergab eine deutliche Verbesserung der Anreicherungsfaktoren der falsch Negativen. Es zeigt, dass die Methode in der Lage ist, einen Molekülgrundgerüstwechsel durchzuführen.

NIPALSTREE, der hierarchische *k*-means und selbst organisierende Karten wurden prospektiv angewandt, um neue Leitstrukturkandidaten für Dopamin-D_3-Rezeptoren zu finden. Moleküle mit neuen Molekülgrundgerüsten und Bindungsaffinitäten im niedrigen nanomolaren Bereich wurden gefunden (65 nM für Molekül **42**). Um einen tieferen Einblick in die SAR dieser Moleküle zu erhalten, wurden verschiede Computerverfahren verwendet. Supportvektorregression und PLS („partial least squares“) wurden untersucht. Es war möglich, voraussagende Modelle für Dopamin-D_2 und D_3 Bindungsaffinitäten zu erstellen. Die SAR erklärende Moleküleigenschaften konnten aus den Modellen extrahiert werden. Die prospektive Anwendung der Modelle auf die diversen und neuen virtuellen Screeningdaten war nur von begrenztem Erfolg. Dockingstudien wurden mit einem Homologiemodell des Dopamin-D_3-Rezeptors durchgeführt. Die visuelle Begutachtung der Bindemoden führte zur Hypothese zweier alternativer Bindetaschen für den Aryl-Rest von Dopamin-D_3-Rezeptorantagonisten. Ein Pharmakophormodell wurde erstellt, welches beide Aryl-Reste gleichzeitig benötigt. Ein virtuelles Screening mit dem Modell identifizierte einen nanomolaren Hit (65 nM für Molekül **59**), welcher die Hypothese unterstützt und eine neue Leitstruktur für Dopamin-D_3-Rezeptoren darstellt.

Die vorgestellten Daten zeigen, dass der kombinierte Ansatz aus hierarchischer Clusterung und anschließender Verwendung der Cluster zur Modellerstellung, SAR in HTS-Daten findet. Die Modelle sind geeignet zum Auffinden von Singletons, mit Hits angereichter Cluster, nicht bestätigter Hits und falsch negativer Molekülgrundgerüste.

8 Ausführliche Zusammenfassung

Hochdurchsatzscreening (HTS) - Assays werden standardmäßig angewendet, um neue Leitstrukturen in Substanzbibliotheken mit mehr als einer Million Molekülen zu finden. Das Ergebnis eines Hochdurchsatzscreens ist eine Vielzahl von Aktivitätsdaten, die geordnet und analysiert werden müssen. Das Ziel der Arbeit war es Struktur-Aktivitätsbeziehungen (SAR) in den primären Screeningdaten der Assays zu finden. Die Strategie war es, alle Moleküle hierarchisch zu clustern, die primären Screeningdaten den Clustern zuzuordnen und die Cluster und deren Beziehung zueinander zu verwenden um daraus Modelle abzuleiten. Die Modelle sollen das Finden falsch Negativer, nicht bestätigter Hits, erster Leitstrukturklassen und Singletons in den Daten ermöglichen. Zu diesem Zweck wurde ein neuer hierarchischer Clusteralgorithmus, NIPLASTREE, entwickelt und ein zweiter Algorithmus, der hierarchisch *k*-means, angepasst. Eine graphische Benutzeroberfläche wurde implementiert, um die Ergebnisse der Clusterung darzustellen und daraus erste SAR abzuleiten. Beide Clusteralgorithmen wurden retrospektiv und prospektiv evaluiert. SAR-Modelle wurden durch Verbindung der Clusterergebnisse mit Dockingstudien, Pharmakophorsuchen und Klassifzierung- bzw. Regressionsmethoden erstellt.

Der NIPALSTREE Algorithmus projiziert unter Verwendung der Hauptkomponentenanalyse einen Datensatz auf eine Dimension. Der Datensatz wird anhand des Scoringvektors sortiert und am Median in zwei Teilmengen aufgetrennt. Der Algorithmus wird rekursiv auf die neu gebildeten Mengen angewandt. Der hierarchische *k*-means Algorithmus trennt mit Hilfe des *k*-means Algorithmus einen Datensatz in zwei Cluster auf. Durch rekursive Anwendung des Algorithmus auf die neu gebildeten Teilmengen wird eine hierarchische Clusterung erreicht. Die Algorithmen sind in der Lage, große Datenmengen mit mehr als einer Million Datenpunkte hierarchisch zu clustern.

Für beide Algorithmen wurde die Berechnung des D_{max} Wertes eingeführt. Die Clusterung wird dabei mit verschiedenen Ähnlichkeitsschwellenwerten in einem Datenintervall durchgeführt. Cluster, deren Mitglieder Distanzen zueinander aufweisen, die den Schwellenwert unterschreiten, werden nicht mehr weiter aufgetrennt. Für jeden Schwellenwert werden die Clusterradien der Endcluster aufsummiert. D_{max} repräsentiert den Schwellenwert, bei dem ein Maximum der Summe erreicht wird. An diesem Punkt wird eine maximal dichte Packung der Endcluster erreicht. Dies entspricht dem Schwellenwert, bei dem maximale Homogenität und minimale Heterogenität der Moleküle in den Clustern erreicht wird.

Eine Vielzahl an Bewertungsschemata wurden entwickelt bzw. angepasst, um bei der Ableitung von SAR im Dendrogramm zu helfen. Dies beinhaltete Formeln zur Bewertung eines Clusters, Teilen des hierarchischen Clusterdendrogramms und des gesamten Dendrogramms. Um die Verteilung von „angiotensin converting enzyme“- und „interleukin-1 cleaving enzyme“-Inhibitoren aus der COBRA-Datenbank auf einer Dendrogrammebene zu bewerten wurden der durchschnittliche Anreicherungsfaktor, die Shannon Entropie und die Kullback-Leibler Distanz berechnet. Durch Darstellen der berechneten Werte für jede Dendorgrammebene war es möglich, die Clusterung der Liganden im Dendrogramm zu beurteilen. Jede der drei Bewertungsfunktionen erlaubte eine unterschiedliche und sich gegenseitig ergänzende Sichtweise auf die Auftrennung der Inhibitoren. Sie geben einen direkten Hinweis auf die Ordnung der Liganden im Dendrogramm. Dadurch werden erste Schlüsse über das Vorhandensein von Leitstrukturserien in den Daten erhalten.

NIPALSTREE stellt für jeden Cluster den Loadingvektor zur Verfügung. Um Deskriptoren zu identifizieren, die Moleküle eines Clusters von strukturell verwandten Molekülen des Nachbarclusters unterscheiden, wurde das Verhältnis zwischen den Deskriptorgewichtungen in den Loadingvektoren berechnet. Deskriptoren, für die Extremwerte erreicht wurden, boten eine gute Erklärung für die Unterscheidung und ermöglichten erste Einblicke in die SAR.

Beide Clusterverfahren wurden anhand mehrerer Datensätze validiert und miteinander verglichen. Als Moleküldatenbanken wurden die COBRA Datenbank, die MDDR Datenbank und eine kombinierte Datenbank aus MDDR, COBRA und dem SPECS Substanzkatalog verwendet. Die Clusterung wurde basierend auf den CATS 2D Deskriptoren und den MOE 2D Deskriptoren durchgeführt. Beide Algorithmen führten zu hierarchischen Clusterungen, die strukturell interpretierbar waren. Verschiedene Strukturklassen wurden voneinander im Dendrogramm getrennt und systematisch angereichert. Der hierarchische *k*-means Algorithmus erbrachte bessere Ergebnisse als NIPLASTREE, da für eine Auswahl verschiedener Ligandklassen höhere Anreicherungsfaktoren erhalten wurden. Die Überlegenheit des ersten Algorithmus kann durch seine polythetische Funktionsweise im Gegensatz zur monothetischen Funktionsweise des letzteren erklärt werden. Eine Schnittmengenbildung der Ergebnislisten, die mit beiden Algorithmen für die Ligandklassen erhalten wurde, führte zu einer deutlichen Verbesserung der Anreicherungsfaktoren. Am Beispiel von „angiotensin converting enzyme“-Inhibitoren wurde gezeigt, dass durch die Schnittmengenbildung der Verlust an Leitstrukturklassen minimal ist. Beide Algorithmen geben somit eine unterschiedliche und sich gegenseitig ergänzende Sicht auf die Daten. Aus diesem Grund ist es nicht möglich eine Präferenz für einen der beiden Algorithmen zu geben.

Um mit den Ergebnissen der hierarchischen Clusterung arbeiten zu können, wurde eine graphische Benutzeroberfläche entwickelt, die es erlaubt, die Clusterergebnisse darzustellen, in den Clustern zu navigieren und experimentell ermittelte Aktivitäten den Molekülen zuzuordnen. Eine Vielzahl an Funktionen wurden integriert, um das Dendrogramm, die Cluster im Dendrogramm, die Moleküle eines Clusters und die physikochemischen Eigenschaften der Moleküle zu analysieren. Verfahren wurden entwickelt, um mit Hits angereicherte Cluster zu finden, Singletons zu charakterisieren und die Selektivität bzw. Spezifität einer Hitklasse zu analysieren. Verschiedene Proteaseinhibitoren aus der COBRA Datenbank wurden exemplarisch mit Hilfe des hierarchischen *k*-means Algorithmus untersucht. Singletons, die eine Quelle alternativer Leitstrukturen darstellen, wurden am Beispiel von Thrombininhibitoren im Dendrogramm charakterisiert. Durch Verwendung zusätzlicher Ähnlichkeitssuchen und nächster Nachbaranalysen wurden die Singletons im Dendrogramm in kürzester Zeit isoliert und dargestellt. Eine abschließende, visuelle Begutachtung wurde dadurch ermöglicht. Die Anwendung erlaubt das Identifizieren alternativer Leitstrukturen.

Das Clustern aller Moleküle, die in einem HTS Assay getestet wurden, ermöglichen das Analysieren der Hits im Kontext der Nichthits. Dabei ist das Ziel, die verschiedenen Leitstrukturserien in den Daten zu identifizieren und zu bewerten. Acht Strukturklassen von Faktor-Xa-Inhibitoren aus der COBRA Datenbank wurden hierfür betrachtet. Die Anreicherungsfaktoren der Faktor-Xa-Inhibitoren wurden in den Clustern auf den Logarithmus der Dendrogrammebene skaliert. Cluster wurden ausgewählt, sobald der skalierter Anreicherungsfaktor einen Schwellenwert von fünf überschritt. Das Dendrogramm wurde dadurch auf sechs Cluster reduziert, in denen sechs Strukturklassen gefunden wurden. Das gemeinsame Auftreten einiger Strukturklassen erbrachte einen vertieften Einblick in die SAR. Eine Clusterung von Faktor-Xa-Inhibitoren mit anderen Serinproteaseinhibitoren wurde beobachtet. Dadurch wurden erste Analysen hinsichtlich Selektivität und Spezifität möglich. Diese Resultate zeigen die Anwendbarkeit der Methode zur Erstellung fokussierter Screening-Bibliotheken.

Um einen Einblick in Spezifität zu erhalten, wurde die Inhibitoren von elf verschiedenen Proteaseenzymen aus der COBRA Datenbank analysiert. Der relative Anteil der Inhibitoren in den Clustern der Dendrogrammebene sieben wurde extrahiert. Dadurch wurde ein neuer Satz Deskriptoren für jedes Proteaseenzym erhalten. Basierend auf diesen Deskriptoren wurden die Proteaseenzyme hierarchisch geclustert. Ein Vergleich der Clusterung mit einem phylogenetischen Stammbaum der Aminosäuresequenzen der Enzyme zeigte sowohl Übereinstimmungen als auch Unterschiede. Diese neue Clusterung bietet somit eine spezifische, Liganden basierte Betrachtung des „Pocketoms“ der Proteaseenzyme. Sie hat großen Nutzen bei der Auswahl von Targets für Selektivitätsscreens, beim Erstellen fokussierter Screeningbibliotheken und bei der Analyse von Bindungstaschen.

Der hierarchische *k*-means Algorithmus wurde verwendet, um die primären Screeningdaten dreier HTS Assays zu untersuchen. Einer der Assays war gegen den transformierenden Wachstumsfaktor-β-Rezeptor Typ I gerichtet. Das Ziel war es, die Assays zu charakterisieren und nicht-bestätigte Hits und falsch Negative zu identifizieren. Folgende Assaycharakteristika wurden beobachtet: die Assayparameter waren so eingestellt, dass anhand der Prozent-Kontrollwerte eine deutliche Trennung von Hits und Nichthits erreicht wurde. Eine Korrelation zwischen IC_{50}-Werten und Prozent-Kontrollwerten wurde nicht beobachtet. Nicht bestätigte Hits traten gehäuft am Prozent-Kontrollschwellenwert auf, der zur Unterscheidung von Hits von Nichthits diente. Eine Korrelation zwischen Clustergröße, Anzahl Hits in einem Cluster und dem Vorhandensein nicht bestätigter Hits wurde nicht beobachtet.

Zum Auffinden nicht bestätigter Hits wurde ein Regelwerk entwickelt. Es bewertet Hits in den Endclustern anhand der Clustergröße, des Prozent-Kontrollwertes der Einträge eines Clusters, der Gesamthitrate, der Hitrate in einem Cluster und der Umgebung des Clusters im Dendrogramm. Die Anwendung des Regelwerkes auf die primären Screeningergebnisse der drei HTS Assays führte zur Identifizierung eines großen Anteils nicht bestätigter Hits. Zudem wurde für jeden Hit eine Bewertung im Kontext verwandter Nichthits erhalten. Dies erlaubt ein Priorisieren von Molekülen für Folgeuntersuchungen. Das Regelwerk wurde zusätzlich mit Bayesischen regularisierten neuronalen Klassifizierungsnetzen kombiniert. Die Netze wurden auf die Trennung von Hits und Nichthits trainiert. Modelle mit hoher Spezifität und niedriger Sensitivität wurden erhalten. Ihr Einsatz für die Vorhersage nicht bestätigter Hits ist daher limitiert. Dafür ermöglichen sie eine weitere Priorisierung der Moleküle für Folgeuntersuchungen.

Falsch negative Leitstrukturserien mit neuen Molekülgrundgerüsten wurden in den drei HTS Assays retrospektiv generiert. Die modifizierten primären Screeningdaten wurden hierarchischen Clusterungen zugeordnet, die mit MOE 2D, CATS 2D und CATS 3D Deskriptoren und dem hierarchischen *k*-means erhalten wurden. Nichthits und Hits wurden aus den Endclustern, die Hits enthielten, extrahiert. Falsch negative Moleküle wurden koextrahiert und angereichert. Da unterschiedliche falsch Negative mit den verschiedenen Deskriptoren gefunden wurden, kann keine Deskriptor-Präferenz gegeben werden. Um falsch Positive in den extrahierten Listen zu minimieren, wurden Bayesische regularisierte neuronale Klassifizierungsnetze mit den Daten trainiert. Die Anwendung der Modelle auf die durch zusätzliche Ähnlichkeitssuchen angereicherten Ergebnislisten ergab eine deutliche

Verbesserung der Anreicherungsfaktoren der falsch Negativen. Der kombinierte Ansatz ist in der Lage, einen Molekülgrundgerüstwechsel durchzuführen.

NIPALSTREE, der hierarchische *k*-means und selbst organisierende Karten wurden prospektiv angewandt, um neue Leitstrukturkandidaten für Dopamin-D_3-Rezeptoren zu finden. Der SPECS Substanzkatalog wurde mit einem Moleküldatensatz bekannter Dopamin-D_3-Rezeptorliganden kombiniert und unter Verwendung der MOE 2D und CATS 3D Deskriptoren geclustert. Substanzen des SPECS Katalogs wurden aus Endclustern extrahiert, die bekannte Dopamin-D_3-Rezeptorliganden enthielten. Moleküle mit Bindungsaffinitäten im niedrigen nanomolaren Bereich wurden erhalten (z.B. K_i = 65 nM für Molekül **42**). Als neue Molekülgrundgerüstelemente wurden Benzamide als Verbindungselement im Molekül (**42**-**47**, **61**), ein Arylthioether-Rest (**43**, **46**) und ein Bicyclo[2.2.1]heptan-Rest (**42**) gefunden. Um einen vertieften Einblick in die SAR der Moleküle zu erhalten, wurden die Ergebnisse der Clusterung mit verschiedenen chemoinformatischen Verfahren kombiniert: Supportvektor basierte Regression und „partial least squares“ (PLS) wurden verwendet. Das Training wurde mit bekannten Dopamin-D_3-Rezeptorliganden und den dazugehörigen Bindungsaffinitäten an Dopamin-D_2 und -D_3-Rezeptoren durchgeführt. MOE 2D, CATS 2D und CATS 3D Deskriptoren wurden verwendet. Voraussagende Modelle für Dopamin-D_2 und -D_3-Rezeptorbindungsaffinitäten wurden erhalten. SAR erklärende Moleküleigenschaften konnten aus den Modellen extrahiert werden. Die prospektive Anwendung der Modelle auf die diversen und neuen virtuellen Screeningdaten war nur von begrenztem Erfolg. Dockingstudien wurden mit einem Homologiemodell des Dopamin-D_3-Rezeptors durchgeführt. Die visuelle Begutachtung der Liganden-Bindemoden führte zur Hypothese zweier alternativer Bindungstaschen für den Aryl-Rest von Dopamin-D_3-Rezeptorantagonisten/Partialagonisten. Ein Pharmakophormodell wurde erstellt, welches beide Aryl-Reste gleichzeitig benötigt. Ein virtuelles Screening mit dem Modell identifizierte einen nanomolaren Hit (K_i = 65 nM für Molekül **59**). Dieser unterstützt die Hypothese und stellt einen neuen Leitstrukturkandidaten für Dopamin-D_3-Rezeptoren dar.

Die vorgestellten Daten zeigen, dass der kombinierte Ansatz aus hierarchischer Clusterung und anschließender Verwendung der Cluster zur Modellerstellung erfolgreich anwendbar ist zur Identifizierung von SAR in primären Screeningdaten von HTS Assays. Die Modelle sind geeignet zum Auffinden von Singletons, mit Hits angereichter Cluster, nicht bestätigter Hits und falsch negativen Molekülgrundgerüsten.

9 Literature

Abraham, D. J. *Burger's Medicinal Chemistry and Drug Discovery 6th Edition.* John Wiley & Sons Inc. Haboken, USA, 2003.

ACD Inc. Toronto, Canada. http://www.acdlabs.com/contact.html

Acharya, K. R.; Sturrock, E. D.; Riordan, J. F.; and Ehlers, M. R. Ace revisited: a new target for structure-based drug design. Nat. Rev. Drug Discov. **2003,** *2*, 891-902.

Adams, G. P.; Weiner, L. M. Monoclonal Antibody Therapy of Cancer. Nature Biotechnol. **2005**, *9*, 1147-1157.

Agrafiotis, D. K. Cedeno, W. Feature Selection for Structure – Activity Correlations Using Binary Particle Swarms. J. Med. Chem., **2002**, *45*, 1098-1107.

Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial Informatics in the Post-Genomics Era. Nat. Rev. Drug Disc. **2002,** *1*, 337-346.

Agrafiotis, D. K.; Rassokhin, D. N. A Fractal Approach for Selecting an Appropriate Bin Size for Cell-Based Diversity Estimation. J. Chem. Inf. Comput. Sci. **2002,** *42*, 117-122.

Ajay, A.; Walters, P. W.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? J. Med. Chem., **1998**, *41*, 3314-3324.

Amersham Biosciences Inc., Freiburg, Germany

An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelops. Mol. Cell. Proteomics **2005**, *4*, 752-761.

Arnold, J. R.; Burdick, K. W.; Pegg, S. C. H.; Toba, S.; Lamb, M. L.; Kuntz, I. D. SitePrint: Three-Dimensional Pharmacophore Descriptors Derived from the Protein Binding Site for Family Based Active Site Analysis, Classification, and Drug Design. J. Chem. Inf. Comput. Sci. **2004**, *44*, 2190-2198.

Arteaga, C. L. Inhibition of TGF β Signalling in Cancer Therapy. Curr. Opin. Gen. Devel. **2006**, *16*, 30-37.

Bajorath, J. *Chemoinformatics Concepts, Methods and Tools for Drug Discovery* Humana Press, Totowa, USA, 2004.

Bajorath, J. Integration of Virtual and High-Throughput Screening. Nat. Rev. Drug Disc. **2002**, *1*, 882-894.

Baker, A. H.; Edwards, D. R.; and Murphy, G. Metalloproteinase Inhibitors: Biological Actions and Therapeutic Opportunities. *J. Cell Sci.* **2002,** *115*, 3719-3727.

Balakin, K. V.; Lang, S. A:; Skorenko, A. V.; Tkachenko, S. E. J. Chem. Comp. Group **2003,** *43*.

Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Structure-Based versus Property-Based Approaches in the Design of G-Protein-Coupled Receptor-Targeted Libraries J. Chem. Inf. Comput. Sci. **2002**, *42* 1332-1342.

Barnard, J. M.; Downs, G. M.; Wild, D. J.; and Wright, P. M. Better Clusters Faster. Third Joint Sheffield Conference on Chemoinformatics **2004**.

Becker, O. M.; Shacham, S.; Marantz, Y.; Noiman, S. Modelling the 3D Structure of GPCRs: Advances and Application to Drug Discovery. Curr. Opin. Drug Discov. Devel. **2003**, *6*, 353-361.

Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. **1996**, *39*, 2887-2893.

Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 2. Side Chains. J. Med. Chem. **1999**, *42*, 5095-5099.

Bettinetti, L.; Schlotter, K.; Hubner, H.; Gmeiner, P. Interactive SAR Studies: Rational Discovery of Super-Potent and Highly Selective Dopamine D3 Receptor Antagonists and Partial Agonists. J. Med. Chem. **2002**, *45*, 4594-4597.

Bezard, E.; Brotchie, J.M.; Gross, C.E. Pathophysiology of Levodopa-Induced Dyskinesia: Potential for New Therapies. Nat Rev Neurosci **2001**, *2*, 577-588.

Bezard, E.; Ferry, S.; Mach, U.; Stark, H.; Leriche, L.; Boraud, T.; Gross, C.; Sokoloff, P. Attenuation of Levodopa-Induced Dyskinesia by Normalizing Dopamine D3 Receptor Function. Nat. Med. **2003**, *9*, 762-767.

Bibello, J. A. The Agony and Ecstasy of "OMIC" Technologies in Drug Development. Curr. Mol. Med. **2005**, *5*, 39-52.

Biglan, K. M.; Holloway, R. G. A Review of Pramipexole and its Clinical Utility in Parkinson's Disease. Expert Opin. Pharmacother. **2002**, *3*, 197-210.

Bioreason Inc., Santa Fe, USA http://www.bioreason.com/ .

Bishop, C. M. *Neural Networks for Pattern Recognition* Oxford University Press, Oxford, Great Britain, 1995.

Bissantz, C.; Schalon, C.; Guba, W.; Stahl, M. Focused Library Design in GPCR Projects on the Example of 5-HT$_{2c}$ Agonists: Comparison of Structure-Based Virtual Screening with Ligand-Based Search Methods. Proteins **2005**, *61*, 938-952.

Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. II. Are Homology Models of G-Protein Coupled Receptors Suitable Targets? Proteins **2003**, *50*, 5-25.

Bleicher, K. H.; Böhm, H. J.; Müller, K.; and Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. Nat. Rev. Drug Discov. **2003,** *2*, 369-378.

Blobe, G. C.; Schiemann, W. P.; Lodish, H. F. Role of Transforming Growth Factor β in Human Disease. N. Engl. J. Med. **2000**, *342*, 1350-1358.

Blundell, T. L.; Jhoti, H.; Abell, C. High-Throughput Crystallography for Lead Discovery in Drug Design. Nat. Rev. Drug Disc. **2002**, *1*, 45-54.

Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. J. Chem. Inf. Model. **2005,** *45*, 807-815.

Böcker, A.; Schneider G,; Teckentrup, A.; Schneider, G. NIPALSTREE A new Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening. J. Chem. Inf. Model. **2006,** *46*, 2220-2229.

Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS Data Mining Approaches. QSAR Comb. Sci. **2004**, 23, 207-213.

Boeckler, F.; Lanig, H.; Gmeiner, P. Modeling the Similarity and Divergence of Dopamine D2-like Receptors and Identification of Validated Ligand-Receptor Complexes. J. Med. Chem. **2005**, *48*, 694-709.

Boeckler, F.; Ohnmacht, U.; Lehmann, T.; Utz, W.; Hübner, H.; Gmeiner, P. CoMFA and CoMSIA Investigation Revealing Novel Insight into the Binding Modes of Dopamine D3 Receptor Agonists. J. Med. Chem. **2005**, *48*, 2493-2508.

Böhm, H. J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg, Germany, 2002.

Böhm, H. J.; Schneider, G. *Virtual Screening for Bioactive Molecules* Wiley-VCH, Weinheim, Germany, 2000.

Bodor, N.; Buchwald, P. Brain-Targeted Drug Delivery: Experiences to Date. Am. J. Drug Targ. **2003**, *1*, 13-26.

Braddock, M.; Quinn, A. Targeting IL-1 in Inflammatory Disease: New Opportunities for Therapeutic Intervention. *Nat. Rev. Drug Discov.* **2004,** *3*, 330-340.

Branes, P. J. New Drugs for Asthma. Nat. Rev. Drug Disc. **2004**, *3*, 831-844.

Branes, P. J. New Treatments for COPD. Nat. Rev. Drug Disc. **2002**, *1*, 437-446.

Breiman, L. Bagging Predictors. Machine Learning, **1996**, *24*, 123-140.

Brody, T. M.; Larner, J.; Minneman, K. P. *Human Pharmacology. Molecular to Clinical.* C. V. Mosby 1998.

Bronson, D.; Hentz, N.; Janzen, W. P.; Lister, M. D.; Menke, K.; Wegrzyn, J.; Sittampalam, G. S. Basic Considerations in Designing High-Throughput Screening Assays. 5-30 in Seethala, R.; Fernandes, P. B. (eds.) *Handbook of Drug Screening* Marcel Dekker, Inc., New York, USA, 2001.

Brown, R. D.; Martin, Y. C. 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.

Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. J. Chem. Inf. Comput. Sci. **2001**, *41*, 1605-1616.

Buchwald, P.; Bodor, N. Computer-Aided Drug Design: The Role of Quantitative Structure-Property, Structure-Activity and Structure-Metabolism Relationships (QSPR, QSAR, QSMR). Drug. Future **2002**, *27*, 577-588.

Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. J. Med. Chem. **1999**, *42*, 3183-3187.

Bush, B. L.; Sheridan, R. P.; PATTY: A Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. J. Chem. Inf. Comput. Sci. **1993,** *33,* 756-762.

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

Byvatov, E.; Sasse, B. C.; Stark, H.; Schneider, G. From Virtual to Real Screening for D3 Dopamine Receptor Ligands. Chembiochem **2005**, *6*, 997-999.

Byfield, S. D.; Major, C.; Laping, N. J.; Roberts, A. B. SB-505124 Is a Selective Inhibitor of Transforming Growth Factor-β Type I Receptors ALK4, ALK5 and ALK7. Mol. Pharmacol. **2004**, *65*, 744-752.

Callahan, J. F.; Burgess, J. L.; Fornwald, J. A.; Gaster, L. M.; Harking, J. D.; Harrington, F. P.; Heer, J.; Kwon, C.; Lehr, R.; Mathur, A.; Olson, B. A.; Weinstock, J.; Laping, N. J. Identification of Novel Inhibitors of the Transforming Growth Factor β1 (TGF-β1) Type 1 Receptor (ALK5). J. Med. Chem. **2002**, *45*, 999-1001.

Cameron, A.; Guo, D.; Kaleta, J.; Menard, R.; Micetich, R. G.; Purisima, E.; Zhou, N. E. [WO 9738008]. **1997.**

Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A. Glivec (ST 571, Imatinib), a Rationally Developed, Targeted Anticancer Drug. Nat. Rev. Drug Disc. **2002**, *1*, 493-502.

Carlson H. A. Protein Flexibility and Drug Design: How to Hit a Moving Target. Curr. Opin. Chem. Biol., **2002**, *6*, 447-452.

Carlsson, A.; Lindquist, M.; Magnusson, T.; Waldeck, B. On the Presence of 3-Hydroxytyramine in Brain. Science **1958**, *127*, 471.

Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines, **2001**. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

ChemAxon Ltd., Budapest, Hungary http://www.chemaxon.com/

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

Cheng, Y.; Prusoff, W. H. Relationship Between the Inhibition Constant (K1) and the Concentration of Inhibitor which Causes 50 Percent Inhibition (I50) of an Enzymatic Reaction. Biochem. Pharmacol., **1973**, *22*, 3099-3108.

Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries. J. Med. Chem. **2005**, *48*, 5154-5161.

Clark, T. Quantum Mechanics in Gasteiger, J. (Ed.) *Handbook of Chemoinformatics. From Data to Knowledge* Wiley-VCH, Weinheim, Germany, 2003.

Clegg, R. M. Fluorescence Resonance Energy Transfer. Curr. Opin. Biotechnol. **1995**, *6*, 103-110.

Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect pf Shape on Binding of Steroids to Carrier Proteins. J. Am. Chem. Soc., **1988**, *110*, 5959-5967.

Curtis, R.; Geesaman, B. J.; DiStefano, P. S. Ageing and Metabolism: Drug Discovery Opportunities. Nat. Rev. Drug Disc. **2004**, *4*, 569-580.

Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. Angew. Chem., **1999**, *6*, 736-749.

Davies, J. W.; Glick, M.; Jenkins, J. L. Streamlining Lead Discovery by Aligning in Silico and High-Throughput Screening. Curr. Opin. Chem. Biol. **2006**, *10*, 343-35.

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

Diaz, J.; Pilon, C.; LeFoll, B.; Gros, C.; Triller, A.; Schwartz, J.-C.; Sokoloff, P. Dopamine D_3 Receptors Expressed by All Mesencephalic Dopamine Neuros. J Neurosci **2000**, *20*, 8677-8684.

DuBuffet, T.; Newman-Tancredi, A.; Cussac, D.; Audinot, V.; Loutz, A.; Millan, M. J.; Lavielle, G. Novel Benzopyrano[3,4-c]pyrrole Derivatives as Potent and Selective Dopamine D3 Receptor Antagonist. Bioorg. Med. Chem. Lett. **1999**, *9*, 2059-2064.

Derynck, R.; Zhang, Y. E. Smad-Dependent and Smad-Independent Pathways in TGF-β Family Signalling. Nature **2003**, *425*, 577-584.

Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N. ; Patlewicz, G. ; Niemela, J. ; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. J. Chem. Inf. Comput. Sci. **2005,** *45*, 839-849.

Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; and Spangler, D. P. Algorithm5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories. J. Chem. Inf. Comput. Sci. **1996,** *36*, 1195-1204.

Drews, J. Drug Discovery: A Historical Perspective. Science **2000**, *287*, 1960-1964.

Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification Second Edition*. John Wiley & Sons, Inc., New York, USA. 2001.

Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis.* Cambridge University Press. Cambridge, United Kingdom, 1998.

Edwards, P. Combinatorial Chemistry. Drug Discov. Today **2003,** *8*, 326-327.

Egner, U.; Krätschmar, J.; Kreft, B.; Pohlenz, H.-D.; Schneider, M. The Target Discovery Process. Chembiochem **2005**, *6*, 468-479.

Eggelin, C.; Rand, L.; Ullmann, D.; Jäger, S. Highly Sensitive Fluorescence Detection Technology Currently Available for HTS. Drug Disc. Today **2003**, *8*, 632-641.

Ellis, C. The State of GPCR Research in 2004. Nat. Rev. Drug Discov. **2004**, *3*, 577-626.

Elsner, J.; Boeckler, F.; Heinemann, F. W.; Hübner, H.; Gmeiner, P. Pharmacophore-Guided Drug Discovery Investigations Leading to Bioactive 5-Aminotetrahydropyrazolopyridines. Implications for the Binding Mode of Heterocyclic Dopamine D3 Receptor Agonists. J. Med. Chem. **2005**, *48*, 5771-5779.

Elsworth, J. D.; Roth, R. H. Dopamine Synthesis, Uptake, Metabolism, and Receptors: Relevance to Gene Therapy of Parkinson's Therapy. Exp Neurol **1997**, *144*, 4-9.

Emilien, G.; Maloteaux, J. M.; Geurts, M.; Hoogenberg, K.; Cragg, S. Dopamine Receptors-Physiological Understanding to Therapeutic Intervention Potential. Pharmacol. Ther. **1999**, *84*, 133-156.

Engels, M. F.; Wouters, L.; Verbeeck, R.; Vanhoof, G. Outlier Mining in High Throughput Screening Experiments. J. Biomol. Screen. **2002**, *7*, 341-351.

Engels, M. F.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S.; Agrafiotis, D. K. A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. J. Chem. Inf. Model. **2006**, *46*, 2651-2660.

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

Ertl, P.; Mühlbacher, J.; Rhode, B.; Selzer, P. Web-based Cheminformatics and Molecular Property Prediction Tools Supporting Drug Design and Development at Novartis. SAR QSAR Environ. Res. **2003**, *14*, 321-328.

Ertl, P; Selzer, P.; Mühlbacher, J. Web-based Cheminformatics Tools Deployed via Corporate Intranets. Drug Discovery Today **2004**, *2*, 201-207.

Everitt, B. S.; Landau, S.; Leese, M. *Cluster Analysis.* Arnold, London, Great Britain, 2001.

Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual Screening of Biogenic Amine-Binding G-Protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. J. Med. Chem. **2005**, *48*, 5448-5465.

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

Fisher, R. A. The Use of Multiple Measurements in Axonomic Problems. Annals of Eugenics **1936,** *7*, 179-188.

Flanders, K. C. Smad3 as a Mediator of the Fibrotic Response. Int. J. Exp- Path. **2004**, *85*, 47-64.

Frédérick, R.; Robert, S.; Charlier, C.; Ruyck, J.; Wouters, J.; Pirotte, B. ; Masereel, B. ; Pochet, L. 3,6-Disubstituted Coumarins as Mechanism-Based Inhibitors of Thrombin and Factor Xa. J. Med. Chem. **2005**, *48*, 7592-7603.

Gaussian Inc., Pittsburgh, USA.

Ge R.; Rajeev, V.; Subramanian, G.; Reiss, K. A.; Liu, D.; Higgins, L.; Joly, A.; Dugar, S.; Charkravarty, J.; Henson, M.; McEnroe, G.; Schreiner, G.; Reiss, M. Selective Inhibitors of Type I Receptor Kinase Block Cellular Transforming Growth Factor- β Signalling. Biochem. Pharmacol. **2004**, *68*, 41-50.

Gedeck, P.; Willett, P. Visual and Computational Analysis of Structure-Activity Relationships in High-Throughput Screening Data. Curr. Opin. Chem. Biol. **2001**, *5*, 389-395.

Gellibert, F.; Woolven, J.; Fouchet, M.-F.; Mathews, N.; Goodland, H.; Lovegrove, V.; Laroze, A.; Nguyen, V.-L.; Sautet, S.; Wang, R.; Janson, C.; Smith, W.; Krysa, G.; Boullay, V.; deGouville, A.-C.; Huet, S.; Hartley, D. Identification of 1,5-Naphthydrine Derivatives as a Novel Series of Potent and Selective TGF-β Type I Receptor Inhibitors. J. Med. Chem. **2004**, *47*, 4494-4506.

Gibco Inc., Karlsruhe, Germany.

Gini, G.; Craciun, M. V.; König, C.; Benfenati, E. Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity. J. Chem. Inf. Comput. Sci. **2004**, *44*, 1897-1902.

Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G. ChemSpaceShuttle: A Tool for Data Mining in Drug Discovery by Classification, Projection, and 3D Visualization. QSAR Comb. Sci. **2003**, *22*, 549-559.

Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers. J. Chem. Inf. Model., **2006**, *46*, 193-200.

Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naive Bayes Classifier. J. Biomol. Screen. **2004**, *9*, 32-36.

Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a sensitive measure of differences in database variability of molecular descriptors. J. Chem. Inf. Comput. Sci. **2001,** *41*, 1060-1066.

Godden, J. W.; Bajorath, J. Shannon entropy--a novel concept in molecular descriptor and diversity analysis. J. Mol. Graph. Model. **2000,** *18*, 73-76.

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

Golebiowski, A.; Klopfenstein, S. R.; Portlock, D. E. Lead Compounds Discovered from Libraries: Part 2. Curr. Opin. Chem. Biol. **2003**, *7*, 308-325.

Golebiowski, A.; Klopfenstein, S. R.; Portlock, D. E. Lead Compounds Discovered from Libraries. Curr. Opin. Chem. Biol. **2001**, *5*, 273-284.

Goudreau, N.; Cameron, D. R.; Bonneau, P.; Gorys, V.; Plouffe, C.; Poirier, M.; Lamarre, D.; Llinas-Brunet, M. NMR Structural Characterization of Peptide Inhibitors Bound to the Hepatitis C Virus NS3 Protease: Design of a New P2 Substituent. J. Med. Chem. **2004,** *47*, 123-132.

GraphPad Software Inc., San Diego, USA.

Greiner Bio-one Inc., Longwood, USA.

Gronemeyer, H.; Gustafsson, J. Å.; Laudet V. Principles for Modulation of the Nuclear Receptor Superfamily. Nat. Rev. Drug Disc. **2004**, *3*, 950-964.

Guay, D. R. Trandolapril: A Newer Angiotensin-Converting Enzyme Inhibitor. Clin. Ther. **2003**, *25*, 713-775.

Güner, O. F. *Pharmacophore Perception, Development, and Use in Drug Design.* International University Line, La Jolla, USA, 1999.

Guo, D.; Micetich, R. G.; Singh, R.; Zhou, N. E.; Zhou, N. E. [US 6232305]. **2001**.

Gurevich, E. V.; Joyce, J. N. Distribution of Dopamine D3 Receptor Expressing Neurons in the Human Forebrain: Comparison with D2 Receptor Expressing Neurons. Neuropsychopharmacology **1999**, *20*, 60-80.

Hackling, A. E.; Stark, H. Dopamine D_3 Receptor Ligands with Antagonist Properties. Chembiochem, **2002**, *3*, 946-961.

Hackling, A.; Ghosh, R.; Perachon, S.; Mann, A.; Höltje, H.-D.; Wermuth, C. G.; Schwartz, J.-C.; Sippl, W.; Sokoloff, P.; Stark, H. N-(-(4-(2-Methoxyphenyl)piperazin-1-yl)alkyl)carboxamides as Dopamine D2 and D3 Receptor Ligands J. Med. Chem., **2003**; *46*; 3883-3899.

Hagmann, W. K.; MacCoss, M.; Mjalli, A. M.; Zhao, J. J. [WO9505192]. **1994**.

Halgren, T. A. The Merck Molecular Force Field. J. Comp. Chem. **1996,** *17*, 490-641.

Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening. J. Chem. Inf. Comput. Sci. **1999,** *39*, 897-902.

Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. J. Chem. Inf. Comput. Sci. **2001**, *41*, 1295-1300.

Härtter, S. & Hiemke, C. Pharmakokinetik, Interaktionspotential und TDM. Pharmazie in unserer Zeit **2002**, 546-557.

Hawkins, D. M. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. **2004,** *44*, 1-12.

Hawkins, D. M.; Young, S. S.; Rusinko III, A. Analysis of a Large Structure-Activity Data Set Using Recursive Partitioning. Quant. Struct. Act. Relat. **1997**, *16*, 296-302.

Hayes, G.; Biden, T. J.; Selbie, L. A.; Shine, J. Structural Subtypes of the Dopamine D_2 Receptor are Functionally Distinct: Expression of the Cloned D2A and D2B Subtypes in a Heterologous Cell Line. Mol. Endocrinol. **1992**, *6*, 920-926.

Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. J. Chem. Inf. Model. **2006,** *46,* 462-470.

Hertzberg, R. P.; Pope, A. J. High-Throughput Screening: New Technology for the 21st Century. Curr. Opin. Chem. Biol. **2000**, *4*, 445-451.

Hill, S. J. G-Protein Coupled Receptors: Past, Present and Future. Br. J. Pharmacol. **2006**, *147*, S27-S37.

Hillisch, A.; Pineda, L. F.; Hilgenfeld, R. Utility of Homology Models in the Drug Discovery Process. Drug Discovery Today, **2004**, *9*, 659-669.

Höltje, H. D.; Sippl, W. *Rational Approaches to Drug Design*, Prous Science, Barcelona, Spain, 2001.

Hoffman, B. T.; Kopatic, T.; Katz, J. L. Newman, A. H. 2D QSAR Modelling and Preliminary Database Searching for Dopamine Transporter Inhibitors Using Genetic Algorithm Variable Selection of MolconnZ Descriptors. J. Med. Chem. **2000**, *43*, 4151-4159.

Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M.; Mahfouf, M.; Lawson, K.; and Mullier, G. Clustering Files of Chemical Structures Using the Fuzzy *k*-Means Clustering Method . J. Chem. Inf. Comput. Sci. **2004,** *44*, 894-902.

Hopkins, A. L.; Groom, C. R: The Drugable Genome. Nat. Rev. Drug Disc. **2002**, *1*, 727-730.

Huse, M.; Muir, T. W.; Xu, L.; Chen, Y.-G.; Kuriyan, J.; Massague, J.; The TGFß Receptor Activation Process: An Inhibitor- to Substrate-Binding Switch. Mol. Cell **2001**, *8*, 671-682.

InforSense Ltd., London, U.K: http://www.inforsense.com/

Inotech AG, Dottikon, Switzerland.

Izrailev, S.; Agrafiotis, D. K. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. J. Chem. Inf. Comput. Sci. **2001**, *41*, 176-180.

Jacoby, E.; Bredel, M. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. Nat. Rev. Genet. **2004**, *5*, 262-275.

Jacoby, E.; Schuffenhauer, A.; Floersheim, P. Chemogenomics Knowledge-Based Strategies in Drug Discovery. Drug News Perspect. **2003**, *16*, 93-102.

Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. J. Chem. Inf. Comput. Sci. **2000**, *40*, 63-70.

Jain, A. k.; Murty, M. N.; Flynn, P. J.; Data Clustering: A Review. ACM Computing Surveys **1999,** *31*, 265-323.

Jansen, J. M.; Martin, E. J. Target-Biased Scoring Approaches and Expert Systems in Structure-Based Virtual Screening. Curr. Opin. Chem. Biol., **2004**, *8*, 359-364.

Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. IEEE Trans. Comput. **1973,** *22*, 1025-1034.

Jenner, P. Dopamine Agonists, Receptor Selectivity and Dyskinesia Induction in Parkinson's Disease. Curr. Opin. Neurol. **2003**, *16*, S3-S7.

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

Johnson, T. O.; Hua, Y.; Luu, H. T.; Brown, E. L.; Chan, F.; Chu, S. S.; Dragovich, P. S.; Eastman, B. W.; Ferre, R. A.; Fuhrman, S. A.; Hendrickson, T. F.; Maldonado, F. C.; Matthews, D. A.; Meador, J. W., III; Patrick, A. K.; Reich, S. H.; Skalitzky, D. J.; Worland, S. T.; Yang, M.; Zalman, L. S. Structure-Based Design of a Parallel Synthetic Array Directed Toward the Discovery of Irreversible Inhibitors of Human Rhinovirus 3C Protease. J. Med. Chem. **2002,** *45*, 2016-2023.

Johnston, P. A.; Johnston, P. A. Cellular Platforms for HTS: Three Case Studies. Drug Discovery Today **2002**, *7*, 353-363.

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

Joyce, J. N.; Millan, M. J. Dopamine D_3 Receptor Antagonists as Therapeutic Agents. Drug Discov. Today, **2005**, *10,* 917-925.

Joyce, J. N. Dopamine D3 receptor as a Therapeutic Target for Antipsychotic and Antiparkinsonian Drugs. Pharmacol. Ther. **2001**, *90*, 231-59.

Kalluri, R.; Neilson, E. G. Epithelial-Mesenchymal Transition and Its Implications for Fibrosis. J. Clin. Invest. **2003**, *112*, 1776-1784.

Kang, Y.; Massague, J. Epithelial-Mesenchymal Transition: Twist in Development and Metastasis. Cell **2004**, *118*, 277-279.

Kang, Y.; Wei, H.; Tulley, S.; Gupta, G. P.; Serganova, I.; Chen, C.-R.; Manova-Todorova, K.; Blasberg, R.; Gerald, W. L.; Massagué, J. Breast Cancer Bone Metastasis Mediated by the Smad Tumour Suppressor Pathway. Proc. Nat. Acad. Sci. **2005**, *102*, 13909-13914.

Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. Sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank. J. Chem. Inf. Model. **2006**, *46*, 717-727.

Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Improved Analysis of Multivariate Data by Variable Stability

Scaling: Application to NMR-based Metabolic Profiling. Analytica Chimica Acta, **2003,** *490*, 265-276.

Kibbey, C.; Calvet, A. Molecular Property Explorer: A Novel Approach to Visualizing SAR Using Tree-Maps and Heatmaps. J. Chem. Inf. Comput. Sci. **2005**, *45*, 523-532.

Kiechle, F. L.; Zhang, X.; Holland-Staley, C. A: The –Omics Era and Its Impact. Arch. Pathol. Lab. Med. **2004**, *128*, 1337-1345.

Kitano, H. Computational Systems Biology. Nature **2002**, *420*, 206-210.

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

Klabunde, T.; Evers, A. GPCR Antitarget Modelling: Pharmacophore Models for Biogenic Amine Binding GPCRs to Avoid GPCR-Mediated Side Effect. Chembiochem, **2005**, *6*, 876-889.

Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. Chembiochem, **2002**, *3*, 928-944.

Knowles, J; Gromo, G. Target Selection in Drug Discovery. Nat. Rev. Drug Disc. **2003,** *2*, 63-69.

Kohonen, T. Self-organized Formation of Topologically Correct Feature Maps. Biol. Cybern. **1982**, *43*, 59-69.

Korfmacher, W. A: Lead Optimization Strategies as Part of a Drug Metabolism Environment. Curr. Opin. Drug Discov. Devel. **2003**, *6*, 481-485.

Kraus, G. *Biochemistry of Signal Transduction and Regulation. Second edition.* Wiley-VCH, Weinheim, Germany, 2001.

Kriegl, J. M.; Arnold, T.; Beck, B.; Fox, T. A Support Vector Machine Approach to Classify Human Cytochrome P450 3A4 Inhibitors J. Comput. Aided Mol. Des. **2005,** *19,* 189-201.

Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. A Comparison of Conformational Analysis Techniques to Generate Pharmacophore Hypotheses Using Catalyst. J. Chem. Inf. Model. **2005**, *45*, 461-476.

Krovat, E. M.; Frühwirth, K. H.; Langer, T. Pharmacophore Identification, in Silico Screening, and Virtual Library Design for Inhibitors of the Human Factor Xa. J. Chem. Inf. Model. **2005**, *45*, 146-159.

Krumrine, J. R.; Maynard, A. T.; Lerman, C. L. Statistical Tools for Virtual Screening. J. Med. Chem. **2005**, *48*, 7477-7481.

Kubinyi, H. *3D QSAR in Drug Design*, ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993.

Kushida, C. A. Pramipexole for the Treatment of Restless Legs Syndrome. Expert Opin. Pharmacother. **2006**, *7*, 441-451.

Lang, P.; Yeow, K.; Nichols, A.; Scheer, A. Cellular Imaging in Drug Discovery. Nat. Rev. Drug Disc. **2006,** *5*, 343-356.

Laffly, E.; Sodoyer, R. Monoclonal and Recombinant Antibodies, 30 Years After. Human Antibodies **2005**, *14*, 33-55.

Lang, S. A.; Kozyukov, A. V.; Balakin, K. V.; Skorenko, A. V.; Ivashchenko, A. A.; Savchuk, N. P.; Classification Scheme for the Design of Serine Protease Targeted Compound Libraries. J. Comput. Aided Mol. Des. **2002,** *16* 803-807.

LeFoll, B.; Goldberg, S. R.; Sokoloff, P. The Dopamine D_3 Receptor and Drug Dependence: Effects on Reward or Beyond? Neuropharmacology, **2005**, *49*, 525-541.

Leach, A. R. *Molecular Modelling Principles and Applications.* Addison Wesley Longman Limited Harlow, England 1996**.**

Leach, A.R.; Gillet, V.J.; *An Introduction to Chemoinformatics.* Kluver Academic Publisher, Dordrecht, The Netherlands, 2003

Lengauer, C.; Diaz, L. A.; Saha, S. Cancer Drug Discovery Through Collaboration. Nat. Rev. Drug Disc. **2005**, *4*, 375-380.

Leriche, L.; Diaz, J.; Sokoloff, P. Dopamine and Glutamate Dysfunctions in Schizophrenia: Role of the Dopamine D_3 Receptor. Neurotox. Res. **2004**, *6*, 63-71.

Levesque, D. Diaz, J.; Pilon, C.; Matres, M. P. ; Giros, B. ; Souil, E.; Schott, D. ; Morgrat, J. L.; Schwartz, J. C.; Sokoloff, P. Identification, Characterization, and Localization of the Dopamine D3 Receptor in Rat Brain Using 7-[3H]hydroxy-N,N-di-n-propyl-2-aminotetralin. Proc. Nat. Acad. Sci. **1992**, *89*, 8155-8159.

Li, H.-Y.; Wang, Y.; Heap, C. R.; King, C.-H. R.; Mundla, S. R.; Voss, M.; Clawson, D. K.; Yan, L.; Campbell, R. M.; Anderson, B. D.; Wagner, J. R.; Britt, K.; Lu, K. X.; McMillen, W. T.; Yingling, J. M. Dihydropyrrolopyrazole Transforming Growth Factor- β Type I Receptor Kinase Domain Inhibitors: A Novel Benzimidazole Series with Selectivity versus Transforming Growth Factor-β Type II Receptor Kinase and Mixed Lineage Kinase-7. J. Med. Chem. **2006**, *49*, 2138-2142.

Lin, A. Overview of Pharmacophore Applications in MOE. J. Chem. Comp. Group, **2004**, http://www.chemcomp.com/journal/ph4.htm.

Lin, L. S.; Lanza, T. J.; Castonguay, L. A.; Kamenecka, T.; McCauley, E.; Van Riper, G.; Egger, L. A.; Mumford, R. A.; Tong, X.; MacCoss, M.; Schmidt, J. A.; Hagmann, W. K. Bioisosteric Replacement of Anilide with Benzoxazole: Potent and Orally Bioavailable Antagonists of VLA-4. Bioorg. Med. Chem. Lett. **2004,** *14*, 2331-2334.

Lipinski, C. A.; Lombardo, F.; Dominy, B.W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Del. Rev. **1997**, *23*, 3-25.

Liu, K.; Feng, J.; Young, S. S. PowerMV: A Software for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. J. Chem. Inf. Comput. Sci. **2005**, *45*, 515-522.

Liu, W.; Meng, X.; Xu, Q.; Flower, D. R.; Li, T. Quantitative Prediction of Mouse Class I MHC Peptide Binding Affinity Using Support Vector Machine Regression (SVR) Models. BMC Bioinformatics **2006**, *7:182.*

Lottspeich, F.; Zorbas, H. *Bioanalytik* Spektum Akademischer Verlag, Heidelberg, 1998.

Lowrie, J.F.; Delisle, R. K.; Hobbs, D. W.; Diller, D. J. The different strategies for designing GPCR and kinase targeted libraries. Comb. Chem. High Throughput Screen., **2004**, *7*, 495-510.

Luedtke, R. R.; Mach, R. H. Progress in Developing D_3 Dopamine Receptor Ligands as Potential Therapeutic Agents for Neurological and Neuropsychiatric Disorders. Curr. Pharm. Des. **2003**, *9*, 643-671.

Mach, U. R.; Hackling, A. E.; Perachon, S.; Ferry, S.; Wermuth, C. G.; Schwartz, J.; Sokoloff, P.; Stark, H. Development of Novel 1,2,3,4-Tetrahydroisoquinoline Derivatives and Closely Related Compounds as Potent and Selective Dopamine D3 Receptor Ligands. Chembiochem **2004**, *5*, 508-518.

Malbon, C. C. G Proteins in Development. Nature Rev. Mol. Cell Biol. **2005**, *6*, 689-701.

Manallack, D. T.; Livingstone, D. J. Neural Networks in Drug Discovery: Have They Lived up to Their Promise. Eur. J. Med. Chem. **1999**, *34*, 195-208.

Marsden, C. A. Dopamine: The Rewarding Years. Br. J. Pharmacol. **2006**, *147*, 136-144.

Martin, Y. C. Diverse Viewpoints on Computational Aspects of Molecular Diversity. J Comb. Chem., **2001**, *3*, 231-250.

Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? J. Med. Chem. **2002**, *45*, 4350-4358.

Massagué, J.; Wotton, D. Transcriptional Control by the TGF β/Smad Signalling System. EMBO J. **2000**, *19*, 1745-1754.

McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. J. Med. Chem. **2002**, *45*, 1712-1722.

MDL Information System Inc., San Leandro, USA. http://www.mdl.com/

Menard, P. R.; Lewis, R. A.; and Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. J. Chem. Inf. Comput. Sci. **1998,** *38*, 497-505.

Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble Methods for Classification in Cheminformatics. J. Chem. Inf. Comput. Sci. **2004,** *44*, 1971-1978.

Meyer, R. D.; Cook, D. Visualization of Data. Curr. Opin. Biotechnol., **2000**, *11*, 89-96.

Mierau, J.; Schneider, F. J.; Ensinger, H. A.; Chio, C. L.; Lajiness, M. E.; Huff, R. M. Pramipexole Binding and Activation of Cloned and Expressed Dopamine D2, D3 and D4 Receptors. Eur. J. Pharmacol. **1995**, *290*, 29-36.

Missale, C.; Nash, R.; Robinson, S. W.; Jaber, M.; Caron, M. G. Dopamine Receptors: From Structure to Function. Physiol. Rev., **1998**, *78*, 189-225.

Miyashita, Y.; Itozawa, T.; Katsumi, H.; Sasaki, S.-I. Short Communication Comments on the NIPALS Algorithm. J. Chemom. **1990**, *4*, 97-100.

Molecular Networks GmbH, Erlangen Germany. http://www.mol-net.de

Morgan, H.L. The Generation of Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Services. J. Chem. Doc., **1965**, *5*, 107-113.

Muegge, I. Selection Criteria for Drug-like Compounds, Med. Res. Rev. **2003,** *23* 302-321.

Murtagh, F. *Multidimensional Clustering Algorithms. COMPSTAT Lectures 4*, Physica Verlag, Vienna, Austria, 1985.

Narasmha, M.; Leptin, M. Cell Movements During Gastrulation: Come in and Be Induced. Trends Cell Biol. **2000**, *10*, 169-172.

Neal, R. M. An Improved Acceptance Procedure for the Hybrid Monte-Carlo Algorithm. J. Comput. Phys. **1994**, *111*, 194-203.

Neve, K. A.; Seamans, J. K.; Trantham-Davidson, H. Dopamine Receptor Signalling. J. Recept. Signal Transduct. Res. **2004**, *24*, 165-205.

Newman, A. H.; Grundt, P.; Nader, M. A. Dopamine D3 Receptor Partial Agonists and Antagonists as Potential Drug Abuse Therapeutic Agents. J. Med. Chem. **2005**, *48*, 3663-3679.

Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; and Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. J. Chem. Inf. Comput. Sci. **2002,** *42*, 1069-1079.

Nieoullon, A.; Coquerel, A. Dopamine: a Key Regulator to Adapt Action, Emotion, Motivation and Cognition. Curr. Opin. In Neurol. **2003**, *16*, S3-S9.

Nilakantan, R.; Nunn, D. S. A Fresh Look at Pharmaceutical Screening Library Design. Drug Discovery Today **2003**, *8*, 668-672.

Norman, A. W.; Mizwicki, M. T.; Norman, D. P. Steroid-Hormone Rapid Actions, Membrane Receptors and a Conformational Ensemble Model. Nat. Rev. Drug Discov. **2004,** *3*, 27-41.

Oellin, F.; Ihlenfeldt, W.-D.; Gasteiger, J. InfVis- Platform-Independent Visual Data Mining of Multidimensional Chemical data Sets. J. Chem. Inf. Comput. Sci. **2005**, *45*, 1456-1467.

Oldenburg, K. R.; Kariv, I.; Zhang, J., Chung, T. D. Y.; Lin, S. Assay Miniaturization: Developing Technologies and Assay Formats. 525-562 in Seethala, R.; Fernandes, P. B. (eds.) *Handbook of Drug Screening* Marcel Dekker, Inc., New York, USA, 2001.

Olesen, P. H. The Use of Bioisosteric Groups in Lead Optimization. Curr. Opin. Drug Discov. Devel. **2001**, *4*, 471-478.

Oloff, S.; Zhang, S.; Sukumar, N.; Breneman, C.; Tropsha, A. Chemometric Analysis of Ligand Receptor Complementarity: Identifying Complementary Ligands Based on Receptor Information (CoLiBRI). J. Chem. Inf. Model. **2006**, *46*, 844-851.

OpenEye Scientific Software, Santa Fe, USA. http://www.eyesopen.com/

Oprea, T. I.; Gottfries. J. Chemography: The Art of Navigating in Chemical Space. J. Comb. Chem. **2001**, *3*, 157-166.

Oprea, T. I.; Matter, H. Integrating Virtual Screening in Lead Discovery. Curr. Opin. Chem. Biol. **2004**, *8*, 349-358.

Ott, T.; Kern, A.; Schuffenhauer, A.; Popov, M.; Acklin, P.; Jacoby, E.; Stoop, R. Sequential Supraparamagnetic Clustering for Unbiased Classification of High-Dimensional Chemical Data. J. Chem. Inf. Comput. Sci. **2004**, *44*, 1358-1364.

Otto, M. *Chemometrics. Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH, Weinheim, Germany, 1998.

Page, R. D. M. Treeview: An Application to Display Phylogenetic Trees on Personal Computers. Computer Applications in the Bioscience **1996**, *12*, 357-358.

Parker, C. N.; Schreyer, S. K. Application of Chemoinformatics to High-Throughput Screening. 85-110 in Bajorath, J. (eds.) *Chemoinformatics Concepts, Methods and Tools for Drug Discovery* Humana Press, Totowa, USA, 2004.

Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. Chem. Rev. **1996**, *96*, 3147-3176.

Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. J. Chem. Inf. Comput. Sci. **1999**, *39*, 28-35.

PerkinElmer Life Sciences Inc., Rodgau, Germany.

Pilla, M.; Perachon, S.; Sautel, F.; Garrido, F. ; Mann, A. ; Wermuth, C. G.; Schwartz, J. C.; Everitt, B. J.; Sokoloff, P. Selective Inhibition of Cocaine-Seeking Behaviour by a Partial Dopamine D_3 Receptor Agonist. Nature **1999**, *400*, 371-375.

Pirard, B.; Matter, H. Matrix Metalloproteinase Target Family Landscape: A Chemometrical Approach to Ligand Selectivity Based on Protein Binding Site Analysis. J. Med. Chem. **2006**, *49*, 51-69.

Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Modelling Robust QSAR. J. Chem. Inf. Comput. Sci., **2006**, *46*, 2310-2318.

Polgàr, T.; Baki, A.; Szendrei, G. I.; Keseru, G. M. Comparative Virtual and Experimental High-Throughput Screening for Glycogen Synthase Kinase-3β Inhibitors. J. Med. Chem. **2005**, *48*, 7946-7959.

Pope, A. J.; Haupts, U. M.; Moore, K. J. Homogeneous Fluorescence Readouts for Miniaturized High-Throughput Screening: Theory and Practice. Drug Disc. Today **1999,** *4*, 350-362.

Promega Corporation, Madison, USA.

Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. Nat. Rev. Drug Disc. **2004**, *3*, 660-672.

Renner, S.; Schneider, G. Fuzzy Pharmacophore Models from Molecular Alignments for Correlation-Vector-Based Virtual Screening. J. Med. Chem. **2004**, *47*, 4653-4664.

Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. Chemmedchem **2006**, *1*, 181-185.

Reynolds, C. H.; Tropsha, A.; Pfahler, L. B.; Druker, R.; Charkravorty, S.; Ethiraj, G.; Zheng, W. Diversity and Coverage of Structural Sublibraries Using SAGE and SCA Algorithms. J. Chem. Inf. Comput. Sci. **2001,** *41*, 1470-1477.

Richon, A. LeadScope: data visualization for large volumes of chemical and biological screening data. J. Mol. Graph. Model. **2000,** *18*, 76-79.

Rishton, G. M. Nonleadlikeness and Leadlikeness in Biochemical Screening. Drug Discovery Today **2003**, *8*, 86-96.

Rishton, G. M. Reactive Compounds and in Vitro False-positives in HTS. Drug Discovery Today **1997**, *2*, 382-384.

Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; and Blower, P. E., Jr. LeadScope: software for exploring large sets of screening data. J. Chem. Inf. Comput. Sci. **2000,** *40*, 1302-1314.

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

Ruderman, N.; Prentki, M., AMP Kinase and Malonyl-CoA: Targets for Therapy of the Metabolic Syndrome Nat. Rev. Drug Disc. **2004**, *3*, 340-351.

Rusinko III,A.; Farmen,M.W.; Lambert,C.G.; Brown,P.L.; Young, S.S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. J. Chem. Inf. Comput. Sci. **1999,** *39*, 1017-1026.

Saeh, J. C.; Lyne, P. D.; Takasaki, B: k:, Cosgrove, D. A: Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. J. Chem. Inf. Model. **2005**, *45*, 1122-1133.

Sauer, B.; Schaefer-Korting, M.; Kleuser, B. Fibrose. Deutsche Apotheker Zeitung **2005**, *38*, 73-78.

Sawyer, J. S.; Anderson, B. D.; Beight, D. W.; Campbell, R. M.; Jones, M. L.; Herron, D. K.; Lampe, J. W.; McCowan, J. R.; McMillen, W. T.; Mort, N.; Parson, S.; Smith, E. C. R.; Vieth, M.; Weir, L. C.; Yan, L.; Zhang, F.; Yingling, J. M. Synthesis and Activity of New Aryl- and Heteroaryl-Substituted Pyrazole Inhibitors of the Transforming Growth Factor-β Type I Receptor Kinase Domain. J. Med. Chem. **2003**, *46*, 3953-3956.

Schmidt, C. Metabolomics Takes Its Place as Latest Up-and-Coming “Omic” Science. J. Nat. Cancer Inst. **2004**, *96*, 732-734.

Schmuker, M.; Givehchi, A.; Schneider, G. Impact of Different Software Implementations on the Performance of the Maxmin Method for Diverse Subset Collection. Mol. Divers. **2004**, *8*, 421-425.

Schnecke, V.; Boström, J. Computational Chemistry-Driven Decision Making in Lead Generation. Drug Discovery Today **2006**, *11*, 43-50.

Schneider, C. S.; Mierau, J. Dopamine Autoreceptor Agonists: Resolution and Pharmacological Activity of 2,6-Diamnotetrahydrobenzothiazole and an Aminothiazole Analogue of Apomorphine. J. Med. Chem. **1987**, *30*, 494-498.

Schneider, G. Neural Networks are Useful Tools for Drug Design. Neural Netw. **2000**, *13*, 15-16.

Schneider, G.; Böhm, H.-J. Virtual Screening and Fast Automated Docking Methods. Drug Discovery Today, **2002,** *7*, 64-70.

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

Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. Prog. Biophys. Mol. Biol. **1998,** *70*, 175-222.

Schneider, P.; Schneider, G. Collection of Bioactive Reference Compounds for Focused Library Design. QSAR Comb. Sci. **2003**, 22, 713-718.

Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? J. Med. Chem. **2006**, *49*, 2000-2009.

Schölkopf, B.; Sung, K.; Burges, C.; Girosi, F.; Niyogi, P.; Poggio, T.; Vapnik, V. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. IEEE Transact. Signal Proc. **1997**, *45*, 2758–2765.

Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. Data Shaving: A Focused Screening Approach. J. Chem. Inf. Comput. Sci. **2004**, *44*, 470-479.

Schuffenhauer A.; Zimmermann, J.; Stoop, R.; van der Vyver, J. J.; Lecchini, S.; Jacoby, E.; An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design. J. Chem. Inf. Comput. Sci. **2002,** *42,* 947-955.

Schwartz, J.-C.; Diaz, J.; Pilon, C.; Sokoloff, P. Possible Implications of the Dopamine D3 Receptor in Schizophrenia and in Antipsychotic Drug Actions. Brain Res. Rev. **2000**, *31*, 277-287.

SciTegic, San Diego, CA. http://www.scitegic.com/

Seethala, R. Homogeneous Assays for High-Throughput and Ultrahigh-Throughput Screening. 69-128 in Seethala, R.; Fernandes, P. B. (eds.) *Handbook of Drug Screening* Marcel Dekker, Inc., New York, USA, 2001.

Seethala, R.; Fernandes, P. B. (eds.) *Handbook of Drug Screening* Marcel Dekker, Inc., New York, USA, 2001.

Segal, D. M.; Moraes, C. T.; Mash, D. C. Up-Regulation of D_3 Dopamine Receptor mRNA in the Nucleus Accumbens of Human Cocaine Fatalities. Mol. Brain Res. **1997**, *45*, 335-339.

Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-Directed Nearest-Neighbour Searching. J. Med. Chem. **2005**, *48*, 240-248.

Shannon, C. E. A mathematical theory of communication. Bell System Technical Journal **1948,** *27*, 379-423.

Sheridan , R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. 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.

Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. 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.

Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? Drug Disc. Today, **2002**, *7*, 903-911.

Showell, G. A.; Mills, J. S. Chemistry Challenges in Lead Optimization: Silicon Isosteres in Drug Discovery. Drug Discovery Today **2003**, *8*, 551-556.

Siegel, P. M.; Shu, W.; Cardiff, R. D.; Muller, W. J.; Massagué, J. Transforming Growth Factor β Signalling Impairs Neu-induced Mammary Tumorgenesis while Promoting Pulmonary Metastasis. Proc. Nat. Acad. Sci. **2002**, *100*, 8430-8435.

Sigma-Aldrich Inc., Taufkirchen, Germany.

Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W.-C.; Pontz, T.; Corbley, M. J.; Cheung, H.-K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Successful Shape-Based Virtual Screening: The Discovery of a Potent Inhibitor of the Type I TGFß Receptor Kinase (TβRI). Bioorg. Med. Chem. Lett. **2003**, *13*, 4355-4359.

Singh, P.; Harden, B. J; Lillywithe, B. J.; Broad, P. M. Identification of Kinase Inhibitors by an ATP Depletion Method. Assay Drug Dev. Technol. **2004,** *2*, 161-169.

Smith, Y.; Kieval, J. Z. Anatomy of the Dopamine System in the Basal Ganglia. Trends in Neurosciences **2000**, *23*, S28-S33.

Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. NeuroCOLT Technical Report Series, **1998**.

Sokoloff, P.; Andrieux, M.; Besancon, R.; Pilon, C.; Martes, M. P.; Giros, B.; Schwartz, J.C. Pharmacology of Human Dopamine D_3 Receptor Expressed in a Mammalian Cell Line: Comparison with D_2 Receptor. Eur J Pharmacol., **1992**, *225*, 331-337

Sokoloff, P.; Diaz, J.; LeFoll, B.; Guillin, O.; Leriche, L.; Bezard, E.; Gross, C. The Dopamine D_3 Receptor: A Therapeutic Target for the Treatment of Neuropsychiatric Disorders. CNS Neurol. Disord. Drug Targets, **2006**, *5*, 25-43.

Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of Linear and Non-linear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. J. Chem. Inf. Comput. Sci. **2003**, *43*, 2019-2024.

SPECS, Delft, The Netherlands. http://ww.specs.net/

Spycher, S.; Pellegrini, E.; Gasteiger, J. Use of Structure Descriptors to Discriminate Between Modes of Toxic Action of Phenols. J. Chem. Inf. Model. **2005**, *45*, 200-208.

Stahl, M.; Mauser, H.; Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. J. Chem. Inf. Comput. Sci. **2005,** *45*, 542-548.

Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A Robust Clustering Method for Chemical Structures. J. Med. Chem. **2005**, *48,* 4358-4366.

Steinmeyer, A. The Hit-to-Lead Process at Schering AG: Strategic Aspects. Chemmedchem **2006**, *1*, 31-36.

Stemp, G.; Ashmeade, T.; Branch, C. L.; Hadley, M. S.; Hunter, A. J.; Johnson, C. N.; Nash, D. J.; Thewlis, K. M.; Vong, A. K.; Austin, N. E.; Jeffrey, P.; Avenell, K. Y.; Boyfield, I.; Hagan, J. J.; Middlemiss, D. N.; Reavill, C.; Riley, G. J.; Routledge, C.; Wood, M. Design and Synthesis of Trans-N-[4-[2-(6-cyano-1,2,3,4-tetrahydroisoquinolin-2-yl)ethyl]cyclohexyl]-4-quinolinecarboxamide (SB-277011): A Potent and Selective Dopamine D(3) Receptor Antagonist with High Oral Bioavailability and CNS Penetration in the Rat. J. Med. Chem. **2000**, *43*, 1878-1885.

Strange, P. G. The Energetics of Ligand Binding at Catecholamine Receptors. Trends Pharmacol. Sci. **1996,** *17*, 238-244.

Subramanian, G.; Schwarz, R. E.; Higgins, L.; McEnroe, G.; Chakravarty, S.; Dugar, S.; Reiss, M. Targeting Endogenous Transforming Growth Factor β Receptor Signalling in SMAD4-Deficient Human Pancreatic Carcinoma Cells Inhibits Their Invasive Phenotype. Cancer Res. **2004**, *64*, 5200-5211.

Sultan, M.; Wigle, D. A.; Cumbaa, C. A.; Maziarz, M.; Glasgow, J.; Tsao, M. S.; and Jurisica, I. Binary Tree-Structured Vector Quantization Approach to Clustering and Visualizing Microarray Data. Bioinformatics. **2002,** *18*, 111-119.

Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modelling. J. Chem. Inf. Comput. Sci. **2005,** *45*, 786-799.

Talanian, R. V.; Brady, K. D.; Cryns, V. L. Caspases as Targets for Anti-Inflammatory and Anti-Apoptotic Drug Discovery. *J. Med. Chem.* **2000,** *43*, 3351-3371.

Tan, K. T.; Makin, A.; Lip, G. YH. Factor X Inhibitors. Expert Opin. Investig. Drugs **2003**, *12*, 799-804.

Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. J. Comput. Aided Mol. Des. **2002**, *16*, 151-166.

Teague, S. J. Implication of Protein Flexibility for Drug Discovery. Nat. Rev. Drug Disc. **2003**, *2*, 527-541.

Teckentrup, T.; Briem, H.; Gasteiger, J. 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.

The Cambridge Crystallographic Data Centre, Cambridge, UK.

Turk, B.; an Turk, D.; and Turk, V. Lysosomal cysteine proteases: more than scavengers. Biochim. Biophys. Acta **2000,** *1477*, 98-111.

Ueno, H.; Yokota, K.; Hoshi, J.; Yasue, K.; Hayashi, M.; Hase, Y.; Uchida, I.; Aisaka, K.; Katoh, S.; Cho, H. Synthesis and Structure-Activity Relationships of Novel Selective Factor Xa Inhibitors with a Tetrahydroisoqinoline Ring. J. Med. Chem. **2005**, *48*, 3586-3604.

Uhl, M.; Aulwurm, S.; Wischhusen, J.; Weiler, M; Ma, J. Y.; Amirez, R.; Mangadu, R.; Liu, Y.-W. ; Platten, M.; Herrlinger, U.; Murphy, A.; Wong, D. H.; Wick, W.; Higgins, L. S.; Weller, M. SD-208, a Novel Transforming Growth Factor β Receptor I Kinase Inhibitor, Inhibits Growth and Invasiveness and Enhances Immunogenicity of Murine and Human Glioma Cells In vitro and In vivo. Cancer Research **2004**, *64*, 7954-7961.

Umetrics AB, Umea Sweden.

Van de Waterbeemd, H.; Gifford, E. ADMET In Silico Modelling: Towards Prediction Paradise? Nat. Rev. Drug Disc. **2003**, *2*, 192-204.

Van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning. J. Comb. Chem. **2001**, *3*, 267-277.

Varady, J.; Wu, X.; Fang, X.; Min, J.; Hu, Z.; Levant, B.; Wang, S. Molecular Modeling of the Three-Dimensional Structure of Dopamine 3 (D3) Subtype Receptor: Discovery of Novel and Potent D3 Ligands Through a Hybrid Pharmacophore- and Structure-Based Database Searching Approach. J. Med. Chem. **2003**, *46*, 4377-4392.

Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.: Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florae, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. The Sequence of the Human Genome. Science **2001,** *291*, 1304-1351.

Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. Proteins, **2003,** *52*, 609-623.

Verma, R. P.; Hansch, C.; An Approach Towards the Problem of Outliers in QSAR. Bioorg. Med. Chem. **2005**, *13*, 4597-4621.

Von Leoprechting, A.; Kumpf, R.; Menzel, S.; Reulle, D.; Griebel, R.; Valler, M. J.; Büttner, F. H. Miniaturization and Validation of a High-Throughput Serine Kinase Assay Using the Alpha Screen Platform. J. Biomol. Screen. **2004**, *9*, 719-725.

Walters, W. P.; Namchuk, M. Designing Screens: How to Make your Hits a Hit. Nat. Rev. Drug Disc. **2003,** *2*, 259-266.

Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening – an Overview Drug Discov. Today **1998,** *3* 160-178.

Ward, J. H. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **1963,** *58*, 236-244.

Warren, G. L.; Andrews, C. W.; Capelli,A.; Clarke, B.; LaLonde, J; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. J. Med. Chem. **2006**, *49*, 5912-5931.

Wegner, J. K.; Fröhlich, H.; Zell, A. Feature Selection based Classification Models. 1. Theory and GA-SEC Algorithm. J. Chem. Inf. Comput. Sci. **2004**, *44*, 921-930.

Wermuth, C. G.; Gannelin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms used in Medicinal Chemistry. Pure & Appl. Chem. **1998**, *70*, 1129-1143.

Wess, G. Challenges for Medicinal Chemistry. Drug Discovery Today **1996**, *1*, 529-532.

Whitley, D. C.; Ford, M. G.; and Livingstone, D. J. Unsupervised forward selection: a method for eliminating redundant variables. J. Chem. Inf. Comput. Sci. **2000,** *40*, 1160-1168.

Wild, D. J.; Blankley, C. J. VisualiSAR: A Web-Based Application for Clustering, Structure Browsing, and Structure-Activity Relationship Study. J. Mol. Graphics Modell. **1999**, *17*, 85-89.

Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. J. Med. Chem. **2005**, *48*, 3182-3193.

Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. J. Med. Chem., **2005**, *48*, 4183-4199.

Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. **1998**, *38*, 986-996.

Willett, P.; Winterman, V.; and Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. J. Chem. Inf. Comput. Sci. **1986,** *26*, 109-118.

Williams, D. A.; Lemke, T. L. *Foye's Principles of Medicinal Chemistry.* Lippincott Williams & Wilkins, Baltimore, USA, 2002.

Willner, P. The Dopamine Hypothesis of Schizophrenia: Current Status, Future Prospects. Int. Clin. Psychopharmacol. **1997**, *12*, 297-308.

Wise, A.; Gearing, K.; Rees, S. Target validation of G-protein coupled receptors. Drug Discov Today **2002**, *7*, 235-246.

Wold, H. *Estimation of Principal Components and Related Models by Iterative Least Squares, Multivariate Analysis.* Academic Press, New York, USA, 1966.

Xhaard, H.; Rantanen, V.-V.; Nyrönen, T.; Johnson, M. S. Molecular Evolution of Adrenoceptors and Dopamine Receptors: Implications for the Binding of Catecholamines. J. Med. Chem. **2006**, *49*, 1706-1719.

Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. J. Med. Chem. **2004**, *47*, 4463-4470.

Xing, L.; Glen, R. C.; Clark, R. D.; Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. J.Chem.Inf.Comput.Sci. **2003,** *43* 870-879.

Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principle Component Analysis Identified by a Genetic Algorithm. J. Chem. Inf. Comput. Sci. **2000**, *40*, 801-809.

Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. J. Chem. Inf. Model. **2005**, *45*, 1784-1790.

Yingling, J. M.; Blanchard, K. L.; Sawyer, J. S.; Development of TGF-β Signalling Inhibitors for Cancer Therapy. Nat. Rev. Drug Disc. **2004**, *3*, 1011-1022.

Young, S. S; Hawkins, D. M. Analysis of a 29 Full Factorial Chemical Library. J. Med. Chem. **1995**, *38*, 2784-2788.

Yuan, J.; Chen, X.; Brodbeck, R.; Primus, R.; Braun, J.; Wasley, J. W. F.; Thurkauf, A. NGB 2904 and NGB 2849: Two Highly Selective Dopamine D-3 Receptor Antagonists. Bioorg. Med. Chem. Lett. **1998**, *8*, 2715-2718.

Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J. Biomol. Screen., **1999**, *4*, 67-73.

Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R. Confirmation of Primary Active Substances from High Throughput Screening of Chemical and Biological Populations: A Statistical Approach and Practical Considerations. J. Com. Chem. **2000**, *2*, 258-265.

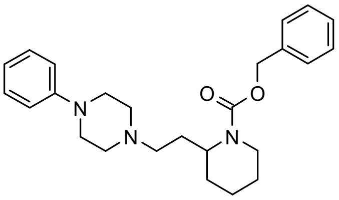
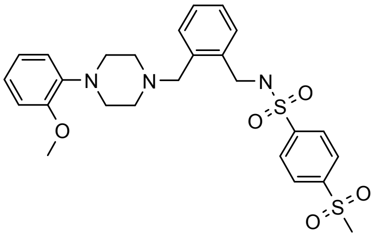
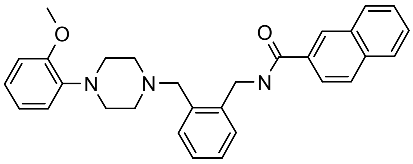
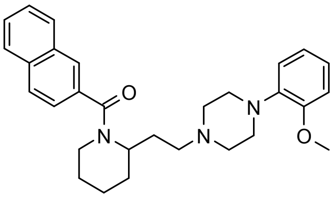
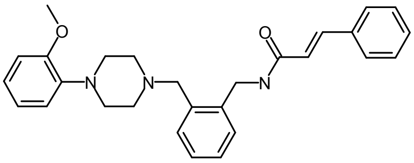
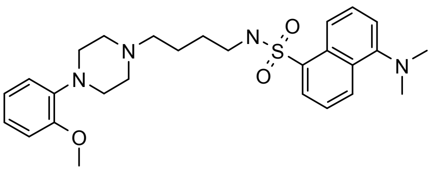
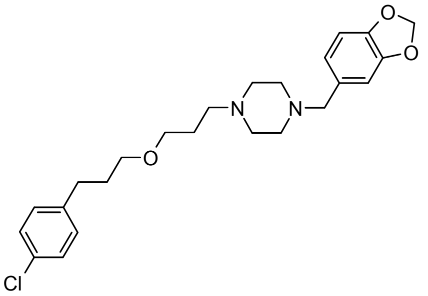
Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. J. Med. Chem. **2006**, *49*, 1536-1548.

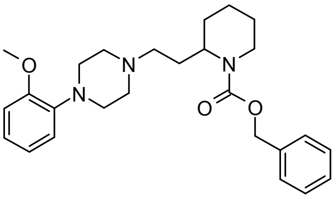
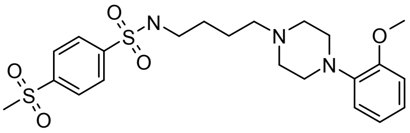
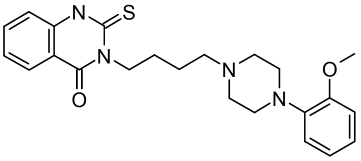
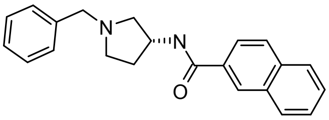
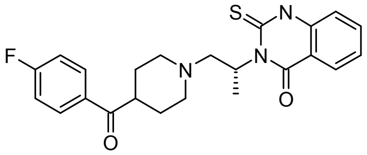
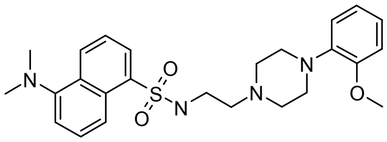
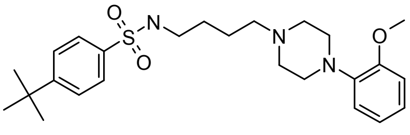
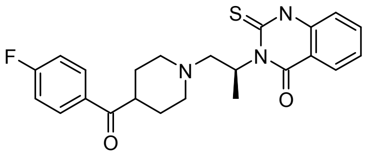
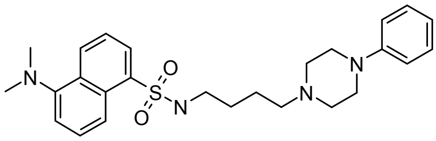
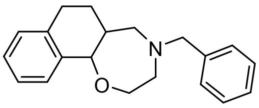
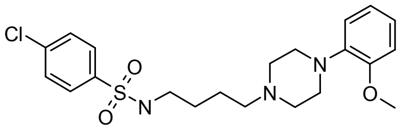
Zhao, C. Y.; Zhang, R. S.; Liu, H. X.; Xue, C. X.; Zhao, S. G.; Zhou, X. F.; Liu, M. C.; Fan, B. T. Diagnosing Anorexia Based on Partial Least Squares, Back Propagation Neural Networks, and Support Vector Machines. J. Chem. Inf. Comput. Sci. **2004**, *44*, 2040-2046.

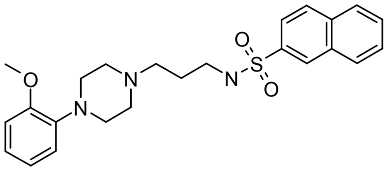
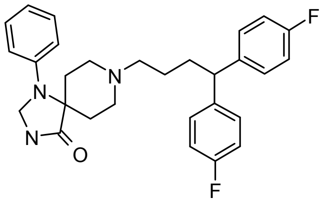
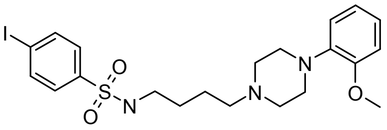
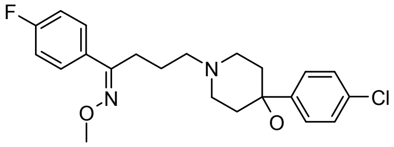
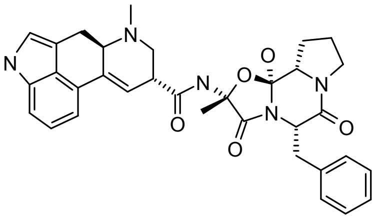
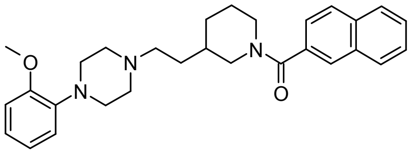
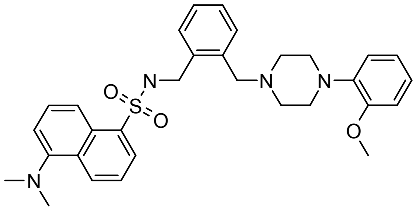
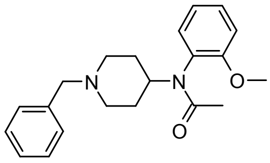
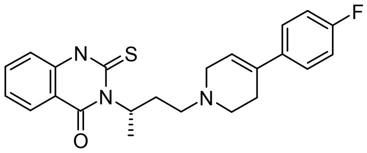
Zhou, Y. A.; Jiang, J. H.; Lin, W. Q.; Zou, H. Y.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Boosting Support Vector Regression in QSAR Studies of Bioactives of Chemical Compounds. Eur. J. Pharm. Sci. **2006**, *28*, 344-353.

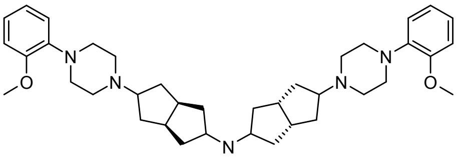
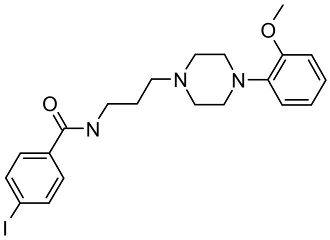
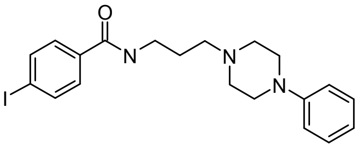
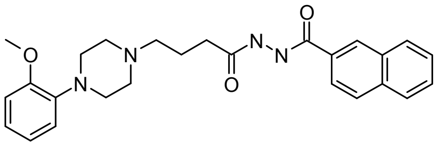
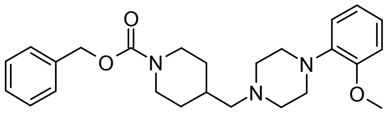
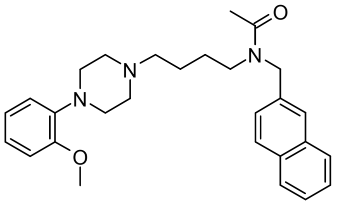
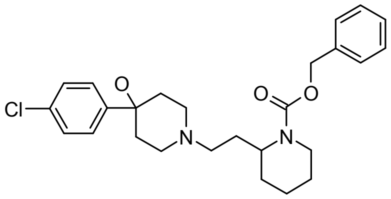
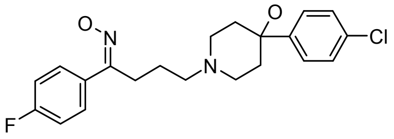
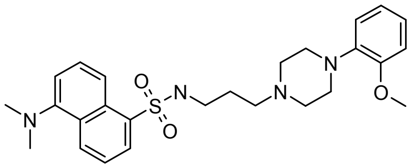
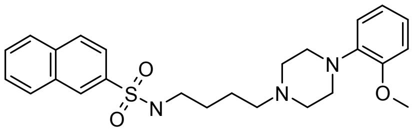
10 Appendix

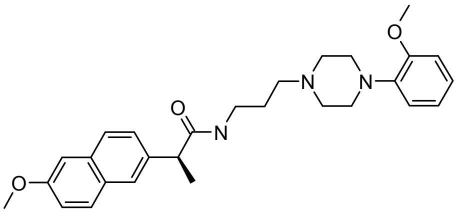
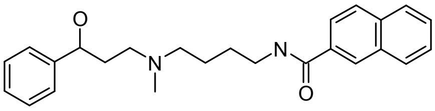
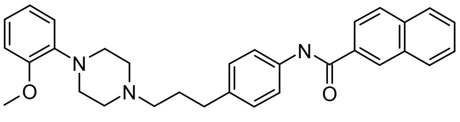
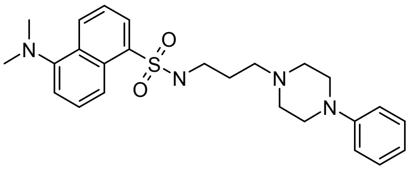
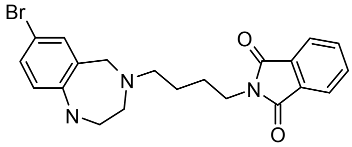
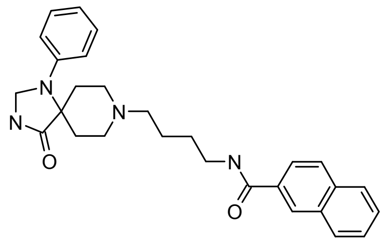
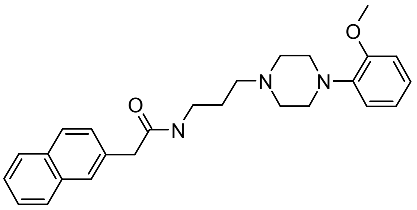
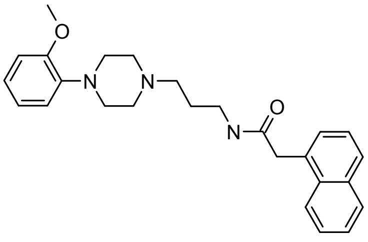
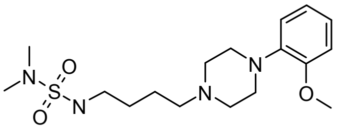
A Dopamine D_3 Receptor Ligands

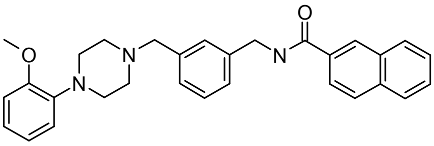
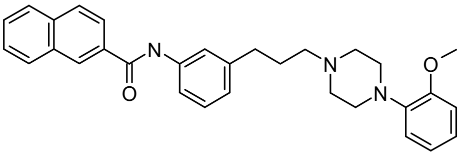
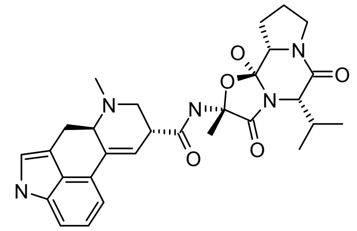
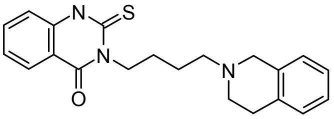
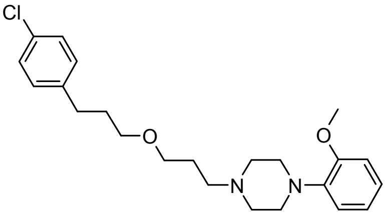
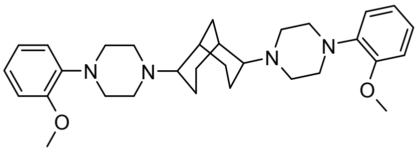
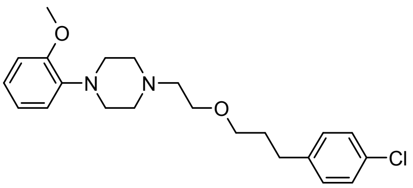
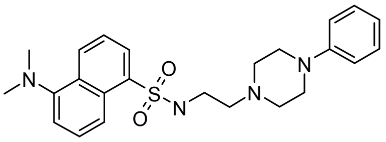
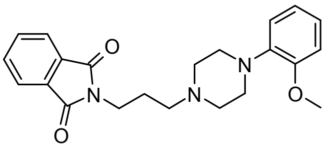
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST380		14.13	667.7	0.02
ST333		28.8	852	0.03
ST177		28.4	750	0.04
ST348		20.2	517.2	0.04
ST239		18	325	0.06
ST148		14	250	0.06
ST374		1306	15683	0.08

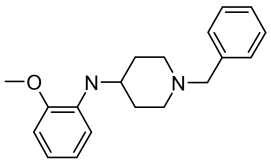
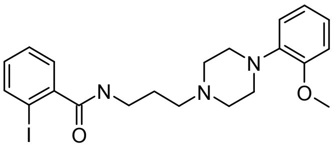
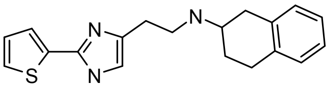
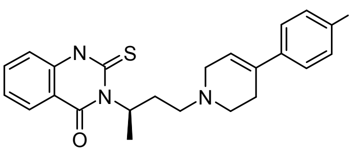
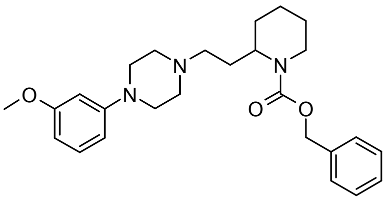
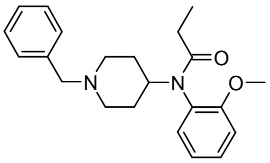
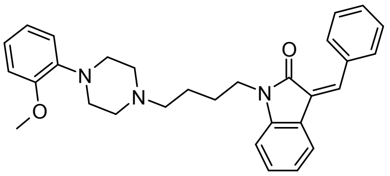
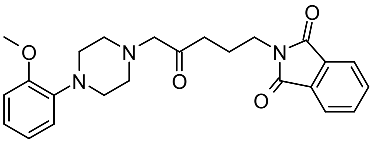
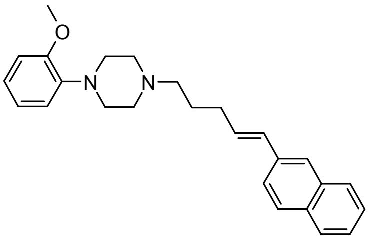
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST393		21.9	254.9	0.09
ST331		2.05	20.8	0.1
EX1*		5.35	44.9	0.12
ST292		3.3	25	0.132
ST112		89.1	664	0.13
ST330		136	956	0.14
ST204		185	1300	0.14
ST316		18.1	120.1	0.15
ST329		7.4	47.4	0.16
ST222		190	1200	0.16
ST224		250	1500	0.17
ST332		6.4	36.2	0.18

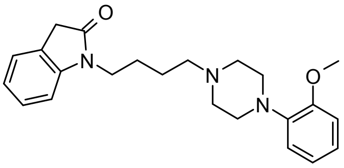
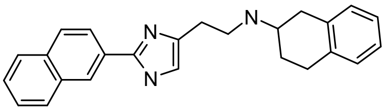
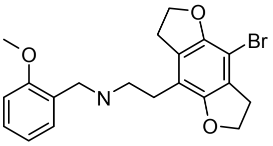
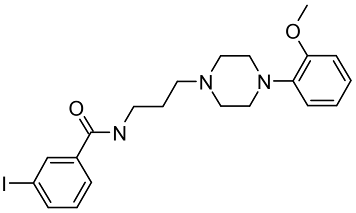
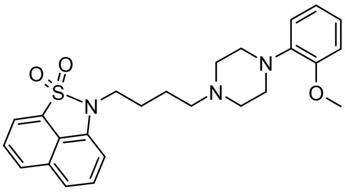
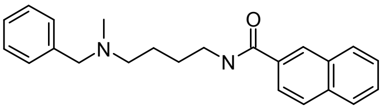
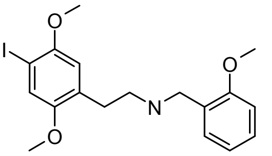
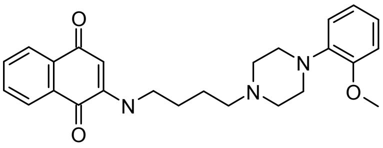
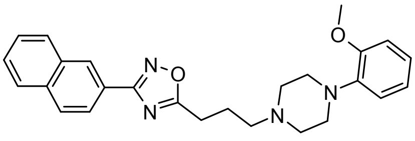
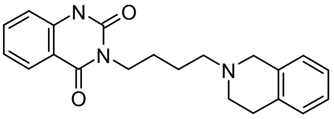
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST72		129	723	0.18
ST271		3	16.8	0.18
ST318		8.14	40.8	0.2
ST58		6.3	30.8	0.2
Ergotamine		1.6	7.5	0.21
ST349		33	138.4	0.24
ST240		250	1000	0.25
EX2*		26.8	106	0.25
ST355		1560	6025	0.26
ST327		154	561	0.27

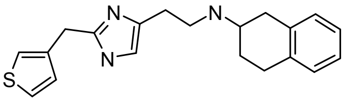
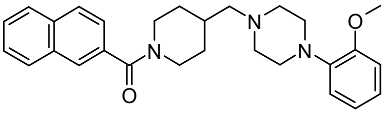
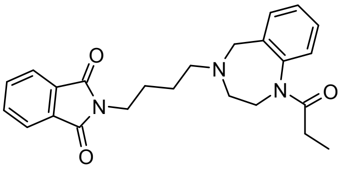
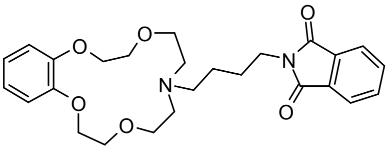
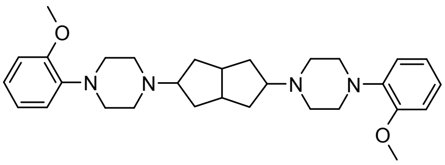
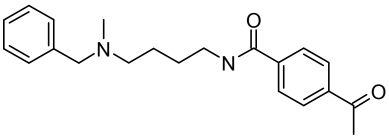
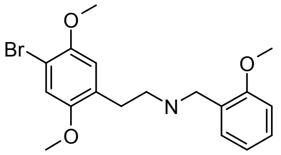
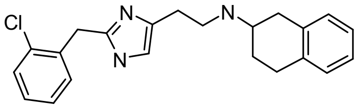
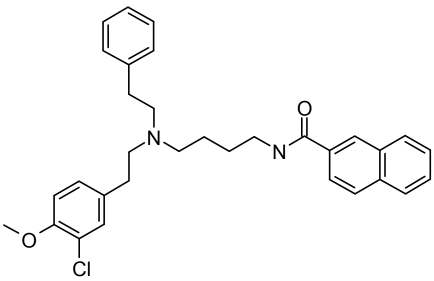
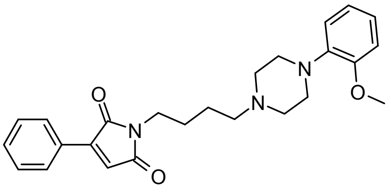
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST207		103	358	0.29
ST68		117	396	0.3
ST62		336	1132	0.3
ST136		15.7	52	0.3
ST345		47.7	156	0.31
ST106		24.3	75.7	0.32
ST381		1667	4990	0.33
ST51		5.2	15	0.35
ST213		105	300	0.35
ST73		85	240	0.35

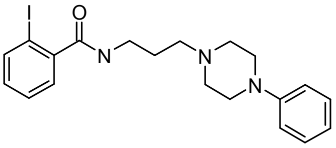
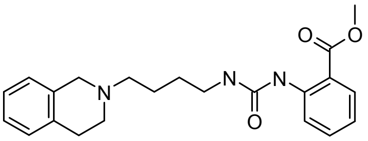
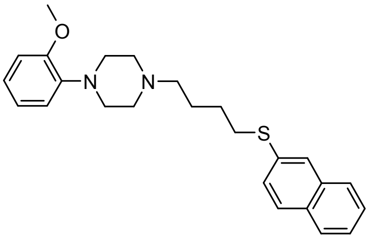
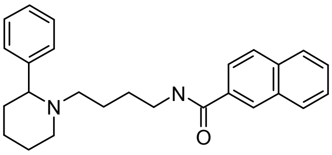
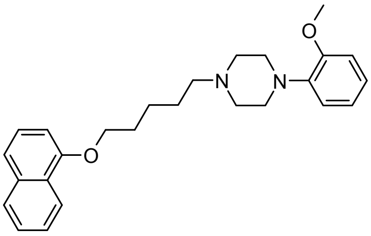
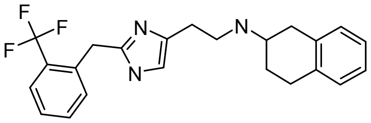
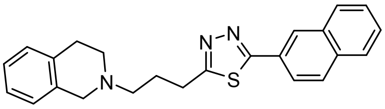
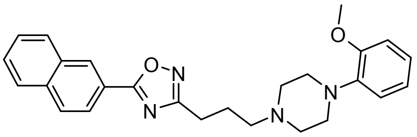
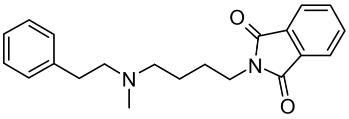
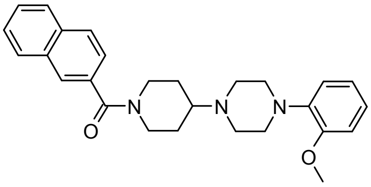
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST102		94.58	262	0.36
ST322		1504	4007	0.38
ST351		99.2	261.8	0.38
ST203		460	1200	0.38
ST360		2030	5230	0.39
ST270		150	380	0.4
ST91		56	140	0.4
ST93		44	108	0.41
ST210		46	110	0.42

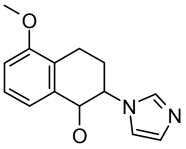
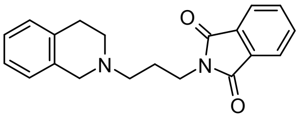
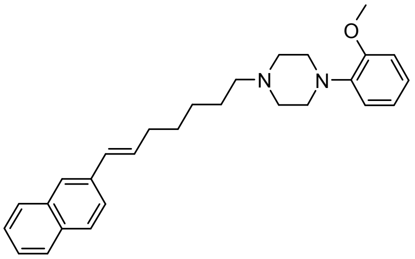
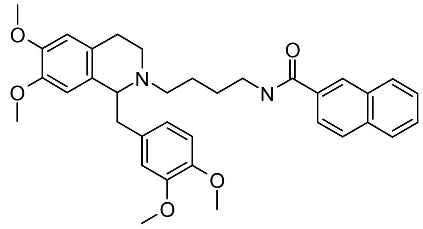
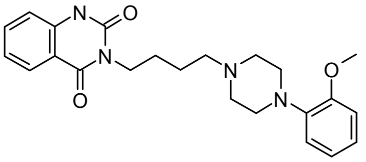
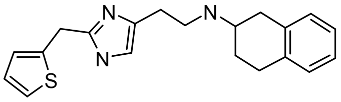
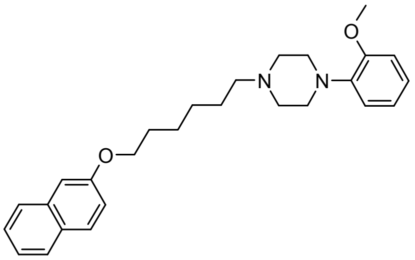
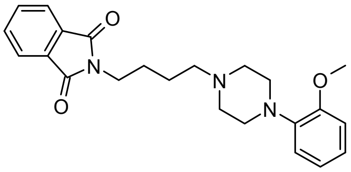
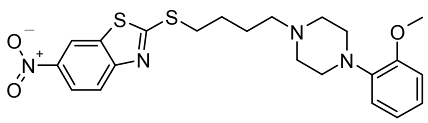
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST176		45	100	0.45
ST350		32.4	71.7	0.45
Ergovaline		250	540	0.46
ST293		37	77	0.48
ST363		19.5	40	0.49
EX3*		6600	13300	0.5
ST208		1100	2200	0.5
ST382		430.5	838	0.51
ST202		2500	4700	0.53
ST88		300	560	0.54

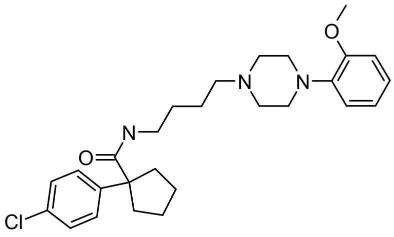
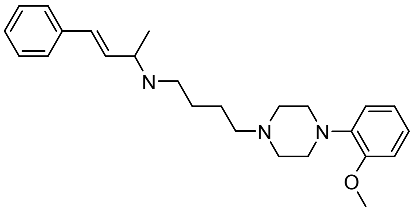
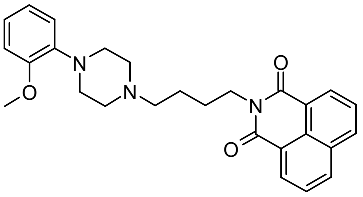
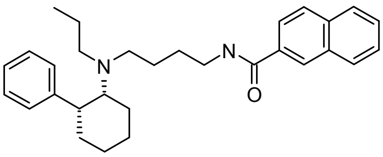
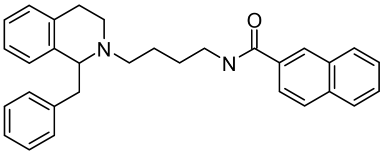
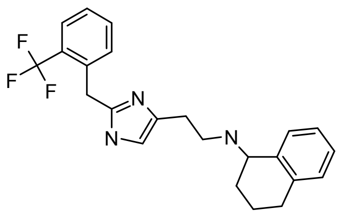
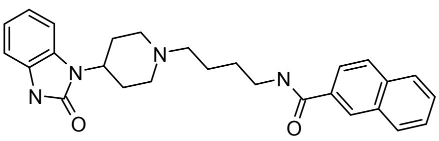
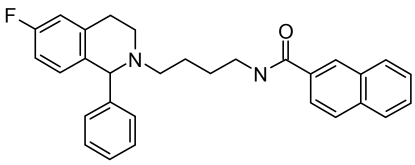
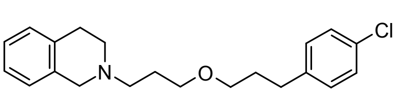
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST344		193	356	0.54
EX4*		217.4	383.7	0.57
EX5*		13400	23200	0.58
ST66		126	213	0.59
ST117		200	330	0.61
ST328		88	145	0.61
ST398		187	303.1	0.62
EX6*		51	82.48	0.62
ST356		1270	2030	0.63
ST273		170	270	0.63
ST103		2106	3312	0.64
ST172		450	700	0.64

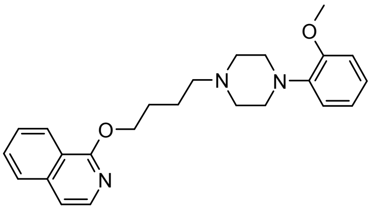
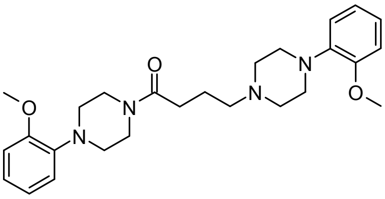
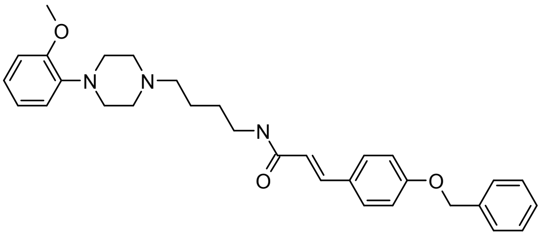
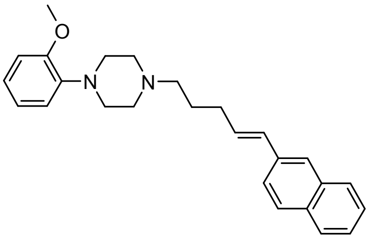
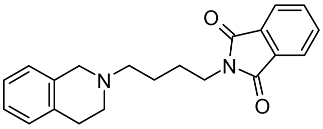
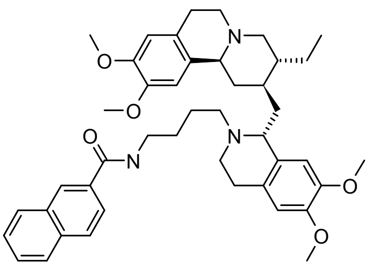
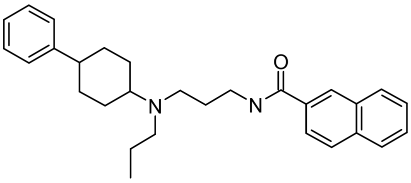
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST272		47	73	0.64
EX7*		9000	13156	0.68
ST116		660	930	0.71
EX8*		2000	2700	0.74
ST326		2307	3019	0.76
ST67		235	300	0.78
ST87		20	24	0.83
ST107		5000	6000	0.83
EX9*		40000	48000	0.83
ST325		1725	2027	0.85
ST137		7	8	0.88
ST143		700	800	0.88
ST297		37	42	0.88

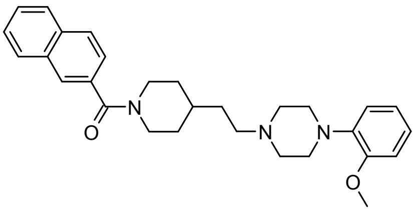
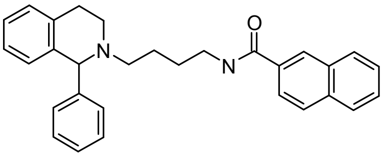
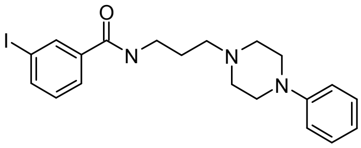
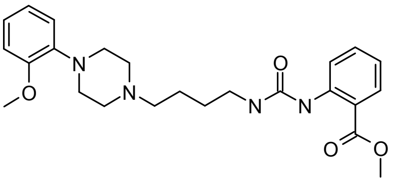
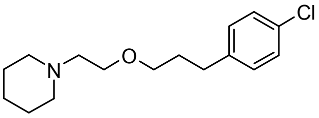
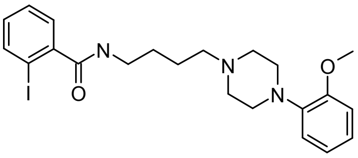
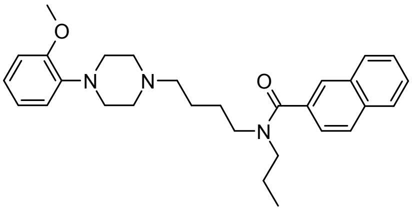
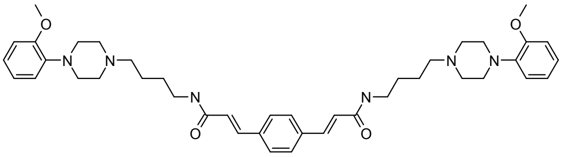
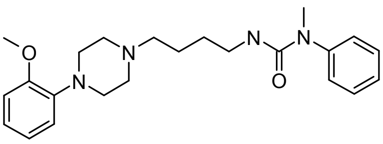
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST119		470	530	0.89
ST346		195	219	0.89
ST359		16720	18580	0.9
ST361		16721	18580	0.899946
ST206		630	700	0.9
ST110		8400	9300	0.9
ST324		1614	1786	0.9
ST120		330	360	0.92
ST209		1300	1400	0.93
EX10*		43	46	0.93
EX11*		873.7	905.9	0.96
ST288		330	320	1.03

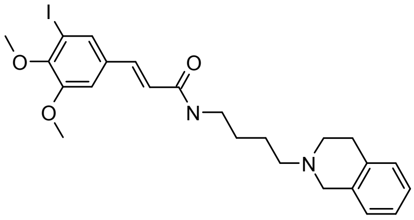
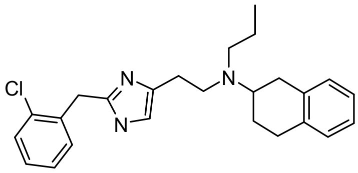
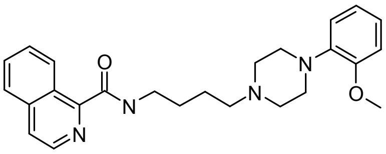
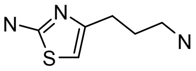
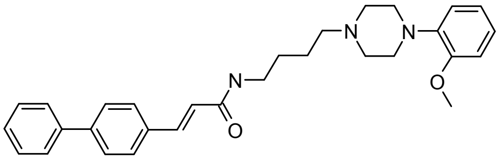
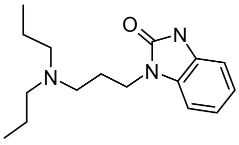
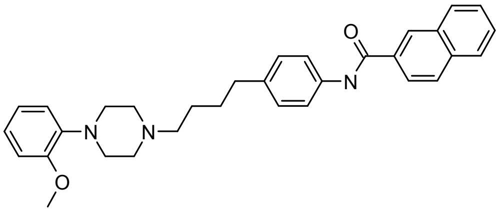
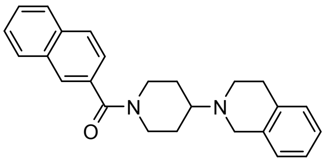
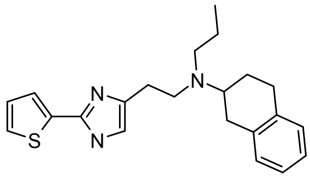
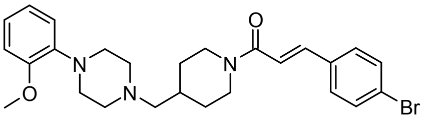
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST60		1000	949	1.05
ST296		70	65	1.08
ST74		814	755	1.08
ST262		1300	1200	1.08
ST78		445	403	1.1
ST121		510	460	1.11
ST142		4500	4000	1.13
ST139		250	220	1.14
ST90		1500	1300	1.15
ST133		1100	950	1.16

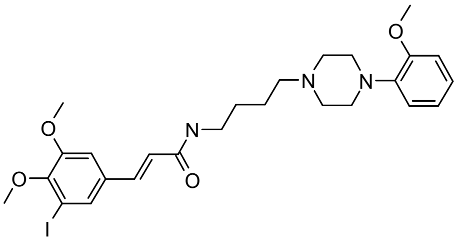
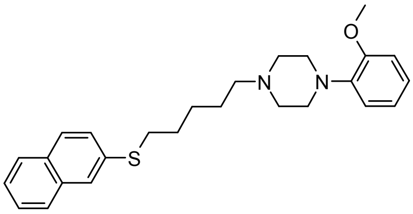
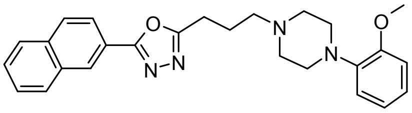
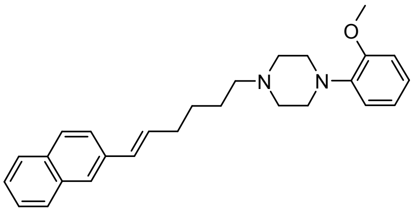
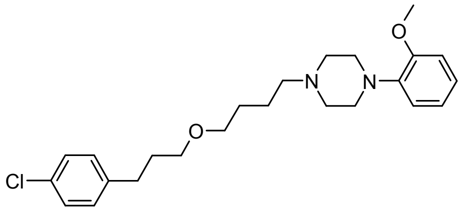
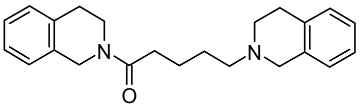
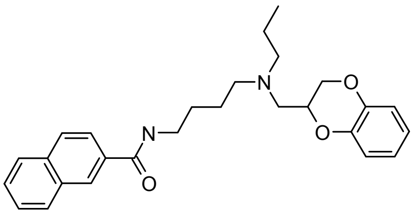
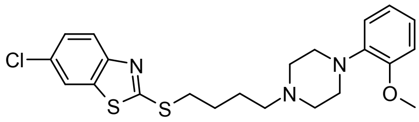
ID	**Structure**	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST1		219000	189000	1.16
ST123		6500	5600	1.16
ST179		167	140	1.19
ST157		3000	2500	1.2
ST295		12.1	10	1.21
ST118		670	540	1.24
EX12*		266	214	1.24
ST76		245	196	1.25
EX13*		5570	4251	1.31
ST84		50	38	1.32
ST170		93	70	1.33

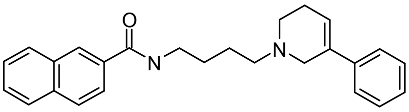
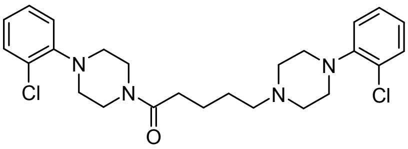
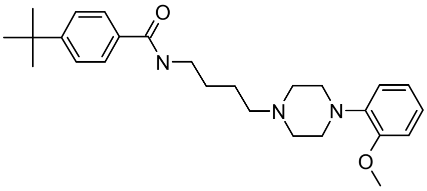
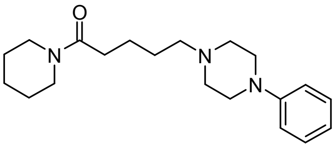
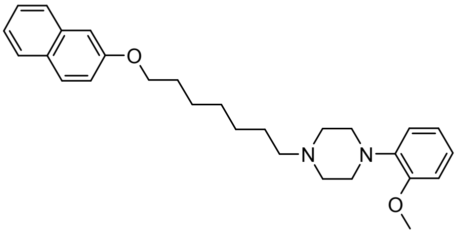
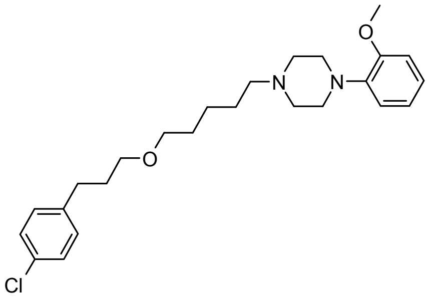
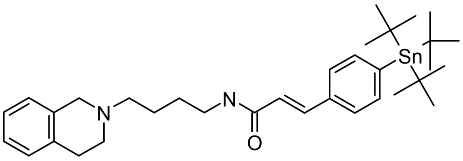
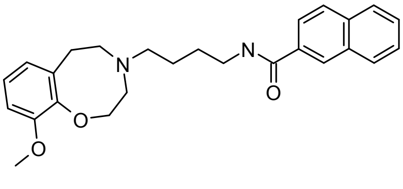
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST150		56	42	1.33
ST255		293	214	1.37
ST86		40	29	1.38
ST266		2500	1790	1.4
ST160		700	500	1.4
ST122		1530	1050	1.46
ST165		300	200	1.5
ST225		8400	5600	1.5
ST362		1131	726	1.56

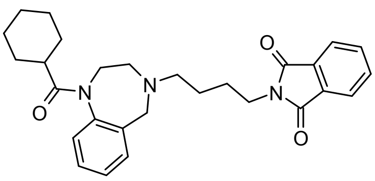
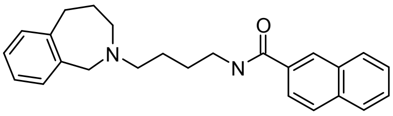
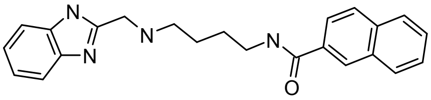
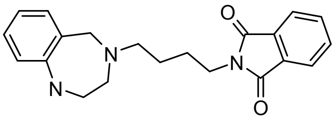
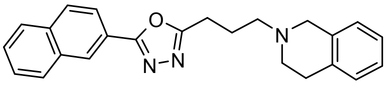
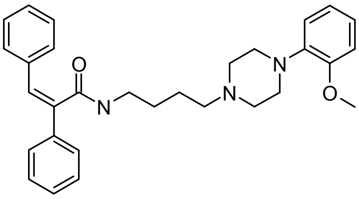
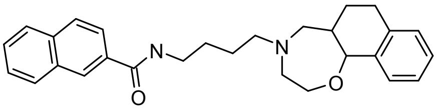
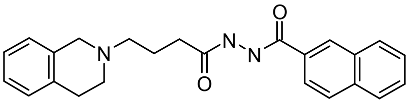
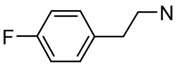
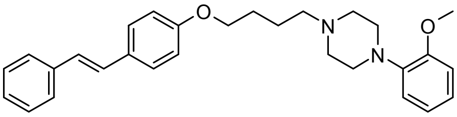
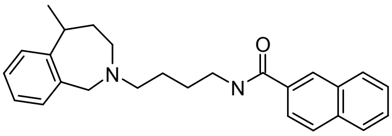
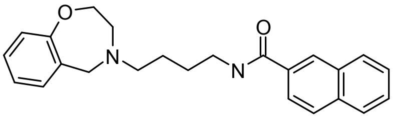
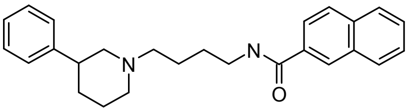
ID	Structure	K$_i$D$_2$[nM]	K$_i$D$_3$ [nM]	D$_2$/D$_3$
ST212		7.4	4.7	1.58
ST306		72	44	1.64
ST230		32	19.5	1.64
ST171		250	150	1.67
ST124		400	233	1.72
ST201		690	400	1.73
ST132		2000	1150	1.74
EX14*		28000	16000	1.75

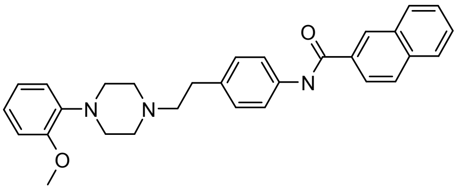
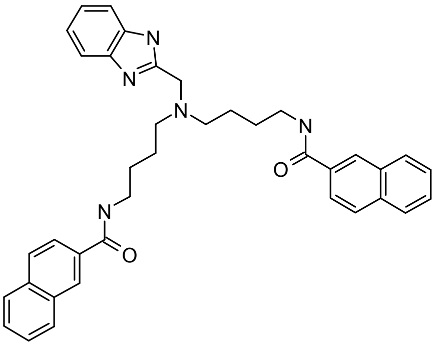
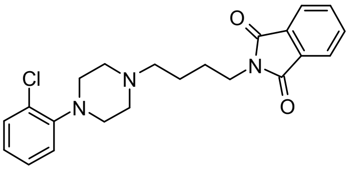
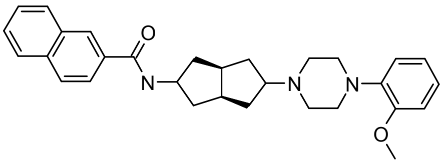
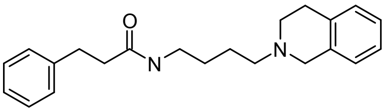
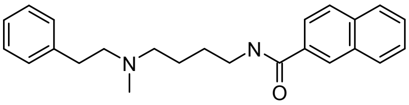
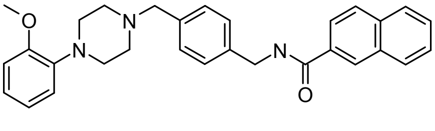
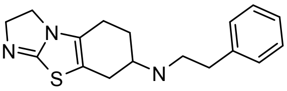
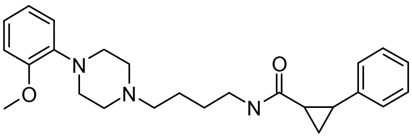
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST347		128.6	72.1	1.78
ST156		4500	2500	1.8
ST61		3583	1980	1.81
ST294		25	13	1.92
ST365		1523	785	1.94
ST69		15.2	7.8	1.95
ST105		47.14	24	1.96
ST235		7.2	3.6	2
ST149		56	28	2
EX15*		600	300	2

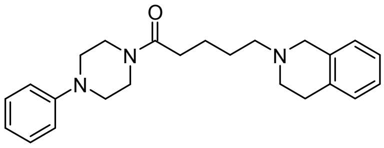
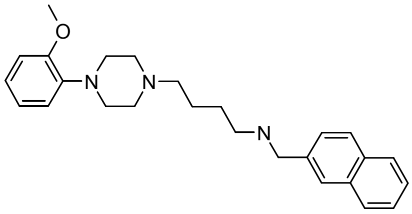
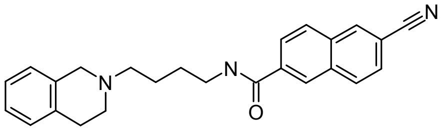
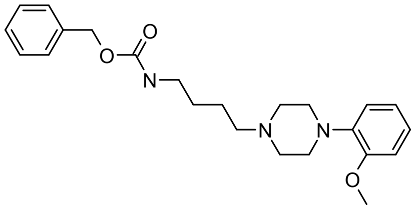
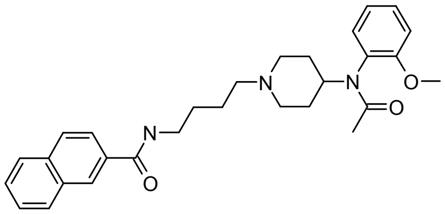
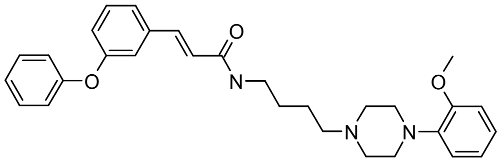
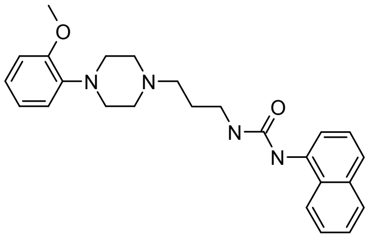
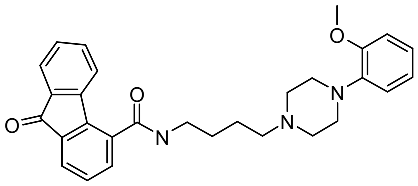
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST287		800	400	2
ST216		300	146	2.06
ST100		20.87	9.62	2.17
ST3		48000	21000	2.29
ST231		37	16	2.31
ST33		26300	11290	2.33
ST200		15.5	6.6	2.35
ST134		3300	1400	2.36
EX16*		119.7	50.63	2.36
ST211		300	126	2.38
EX17*		4947	2030	2.44
ST352		54.3	21.9	2.48

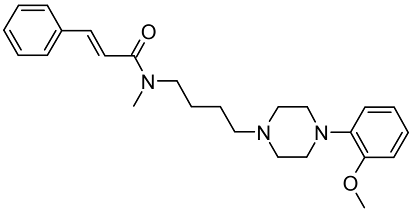
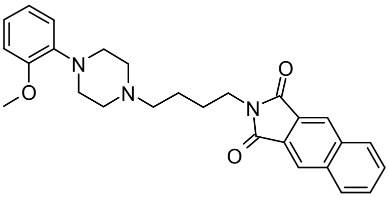
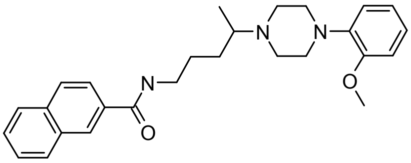
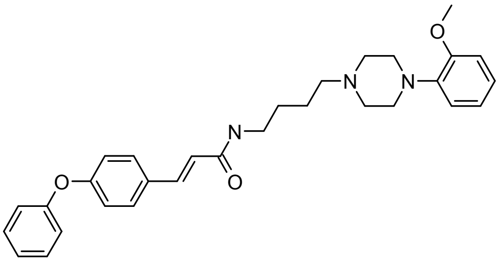
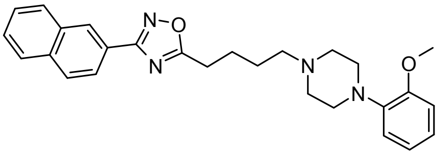
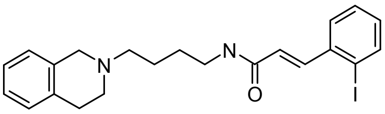
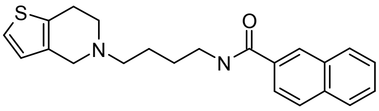
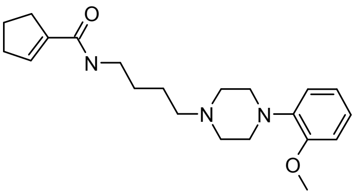
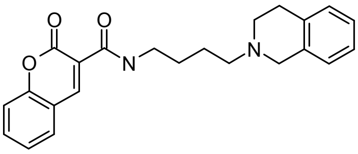
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
EX18*		41.9	16.8	2.5
ST286		38.5	14.9	2.58
EX19*		11.3	4.4	2.6
EX20*		2200	840	2.62
EX21*		10.98	4.07	2.7
ST75		605	220	2.75
EX22*		1895	684	2.77
ST140		283	100	2.83
ST178		105	37	2.84
ST372		170.1	59.8	2.85
ST366		503	174	2.89
ST238		1000	335	2.99
ST153		300	100	3

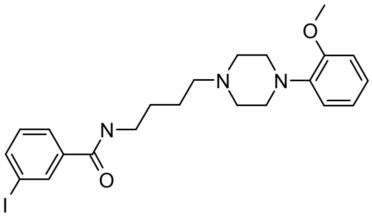
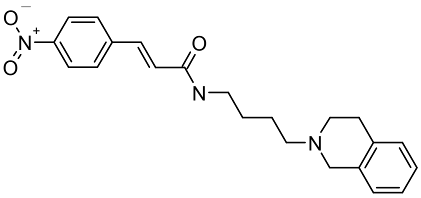
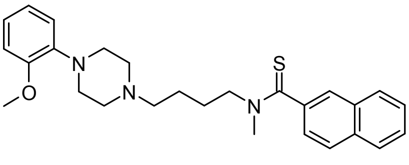
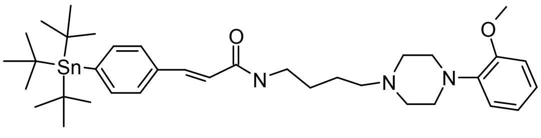
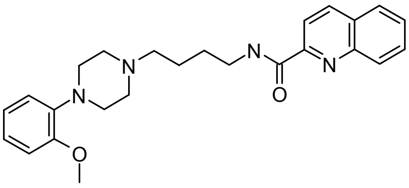
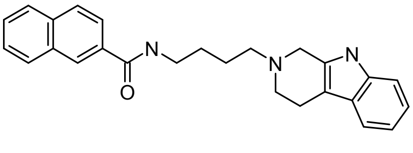
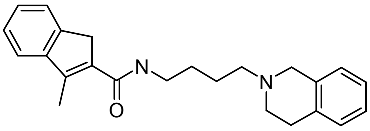
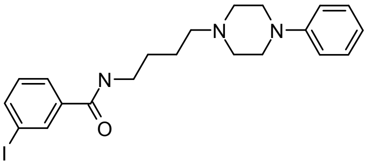
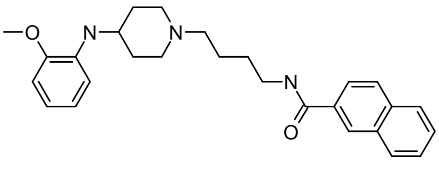
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
EX23*		290	96	3.02
ST111		540	177	3.05
EX24*		65500	21300	3.08
EX25*		19000	5900	3.22
ST399		807.6	247.2	3.27
ST315		26.5	7.99	3.32
ST397		328.8	98.4	3.34
ST77		221	65	3.4
ST400		219.1	61.4	3.57
ST301		910	255	3.57
ST342		2846	797	3.57

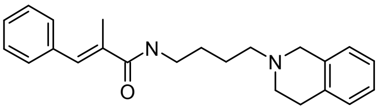
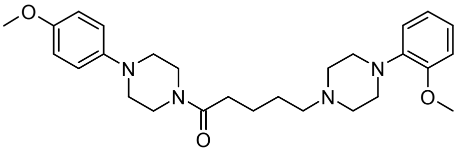
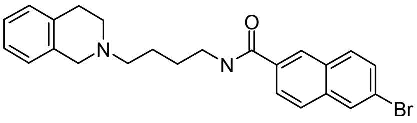
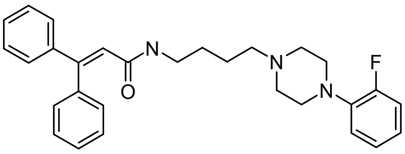
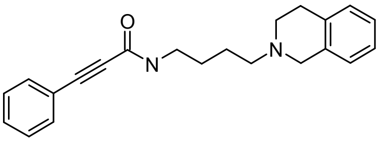
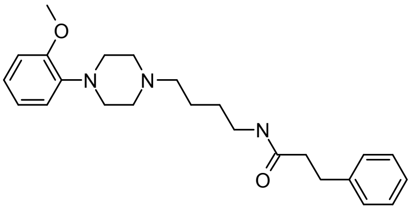
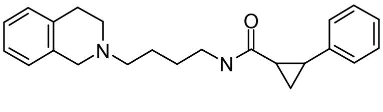
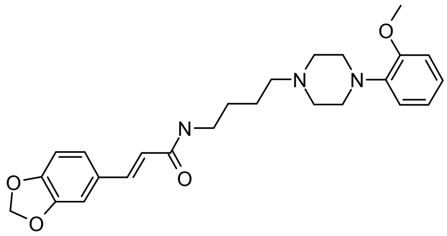
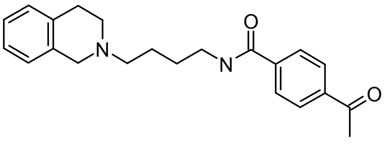
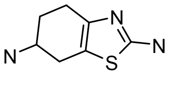
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST391		4465	1233	3.62
ST299		4000	1100	3.64
ST154		5500	1500	3.67
ST358		17280	4525	3.82
ST141		6500	1700	3.82
ST194		67	17.5	3.83
ST223		1000	260	3.85
ST135		5000	1300	3.85
ST4		105000	27000	3.89
ST256		160	41	3.9
EX26*		33.9	8.54	4
ST313		4200	1016	4.13
EX27*		34900	8400	4.16
ST340		3630	831	4.37
ST166		350	80	4.38

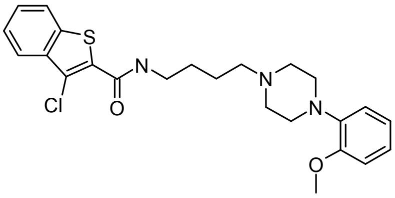
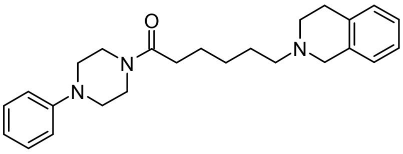
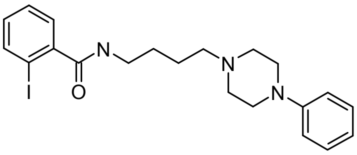
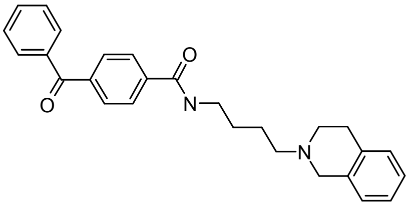
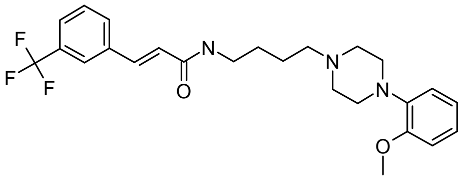
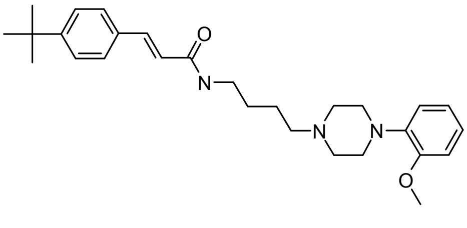
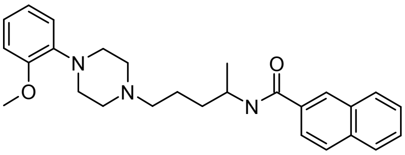
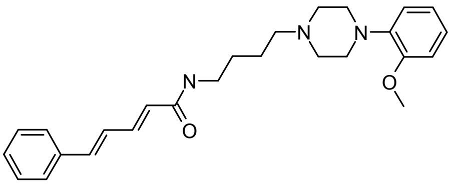
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST175		37.6	8.2	4.59
ST155		3000	650	4.62
ST128		104	22.1	4.71
ST205		178	37.2	4.79
ST275		910	190	4.79
EX28*		120.1	24.8	4.84
EX29*		38.25	7.72	4.9
ST108		612	123	4.98
ST81		200	40	5
ST31		96900	19300	5.02
EX30*		13.22	2.62	5.05
ST253		81	15.5	5.23

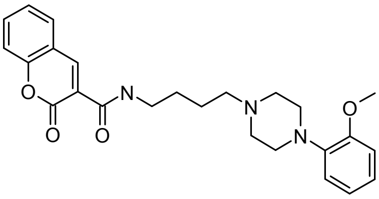
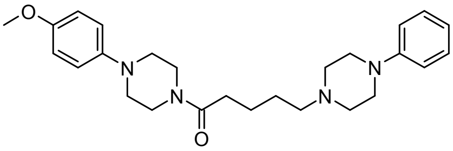
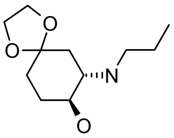
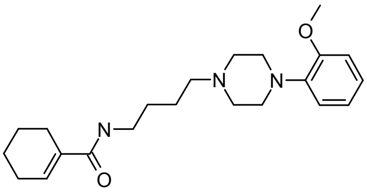
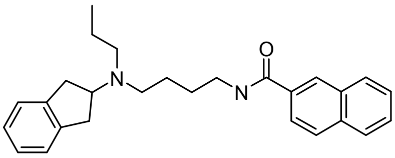
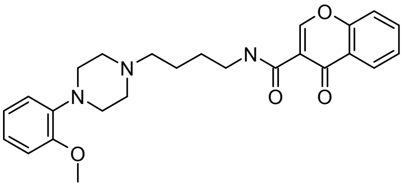
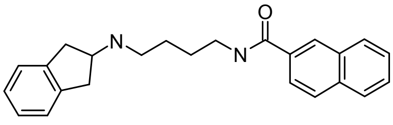
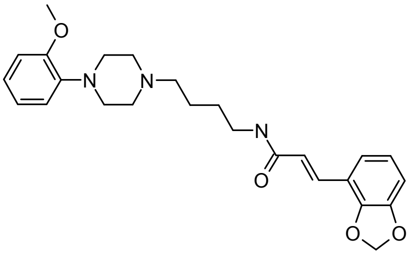
ID	Structure	K$_i$D$_2$[nM]	K$_i$D$_3$ [nM]	D$_2$/D$_3$
ST367		3089	581	5.32
ST83		150	28	5.36
EX31*		118.14	21.48	5.5
ST221		260	47	5.53
ST377		54.24	9.59	5.66
EX32*		94	16.6	5.66
ST357		13350	2340	5.71
ST229		114	19.5	5.85
ST94		16	2.7	5.93
ST101		17.99	2.96	6.08
EX33*		75.49	12.4	6.09

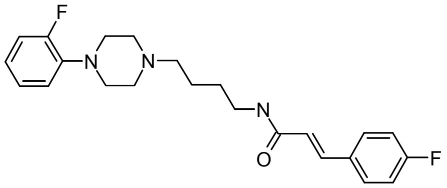
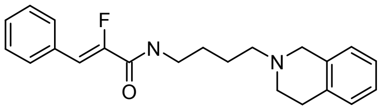
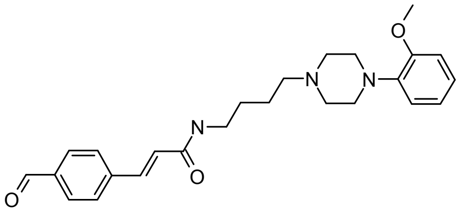
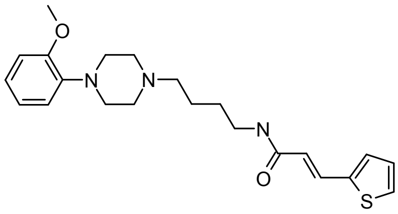
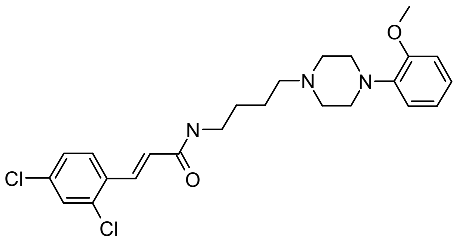
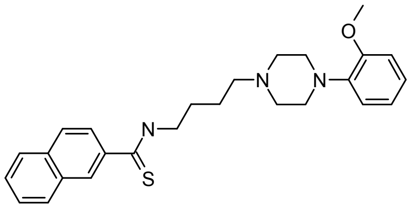
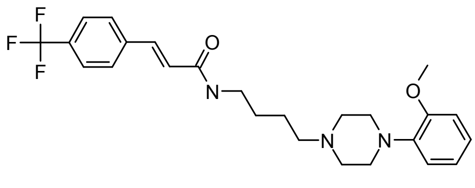
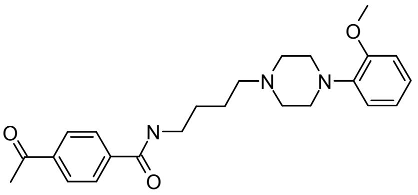
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST276		40.4	6.6	6.12
ST85		145	23.3	6.22
EX34*		22.65	3.62	6.26
ST214		88	14	6.29
EX35*		460	73	6.3
ST228		65.2	10.1	6.46
EX36*		135	20.4	6.62
ST302		213	32	6.66
ST285		1000	150	6.67
ST289		1200	180	6.67
ST354		28.5	4.1	6.95
ST146		650	93	6.99

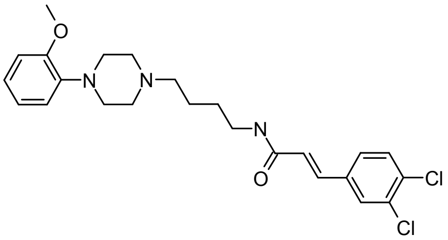
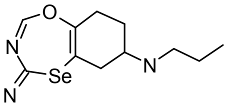
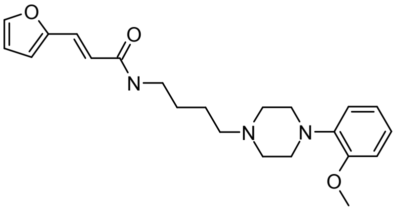
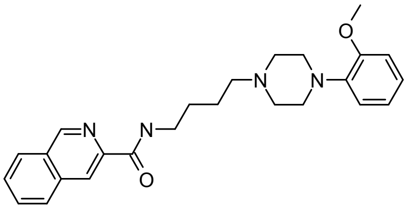
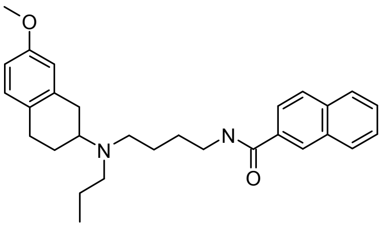
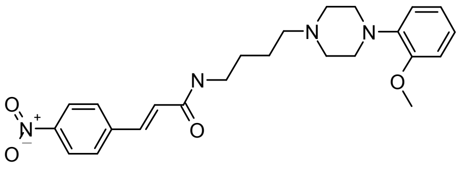
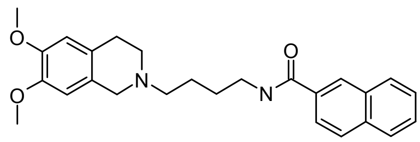
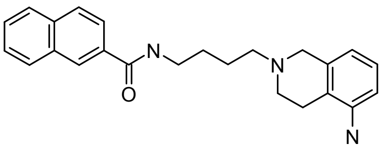
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST70		68	9.2	7.39
ST185		290	38	7.63
ST158		52	6.7	7.76
EX37*		1554	198	7.8
ST300		729	92	7.92
ST98		53.2	6.6	8.06
EX38*		13	1.6	8.2
ST115		610	74	8.24
ST291		800	93	8.6
ST64		377	43	8.77
EX39*		3809	433	8.8
ST343		163	18.4	8.86
EX40*		10000	1100	9.09

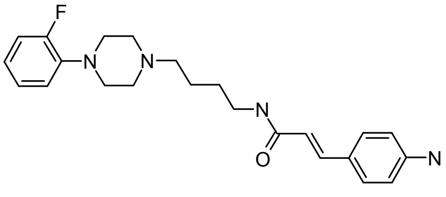
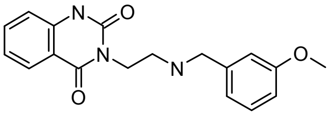
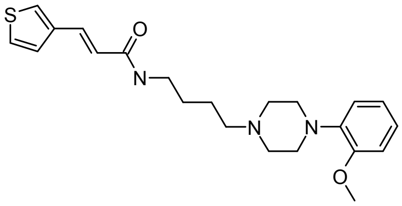
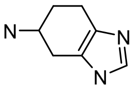
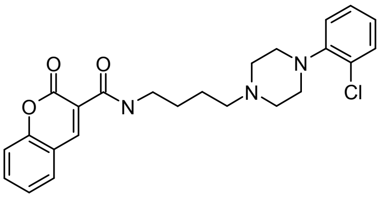
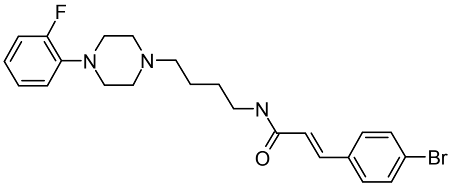
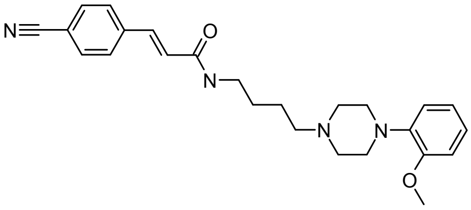
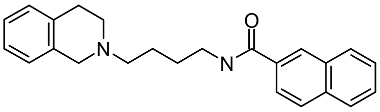
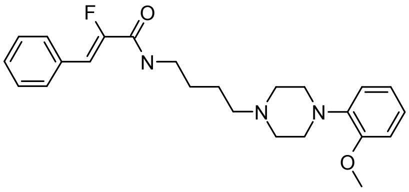
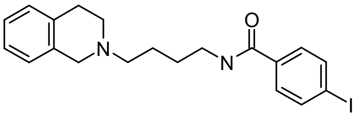
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST311		238	26	9.15
ST369		747	80	9.34
EX41*		47.66	5.07	9.41
EX42*		5.01	0.53	9.5
ST199		1000	105	9.52
ST321		211	21.8	9.68
ST312		165	17	9.71
EX43*		218.2	21.67	10.1
ST274		48.7	4.8	10.15
ST278		1600	157.7	10.15
ST232		10.3	1	10.3
ST125		500	48	10.42
ST36		22000	2100	10.48

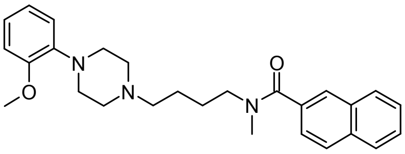
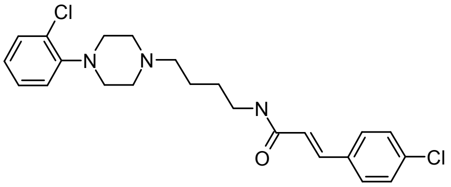
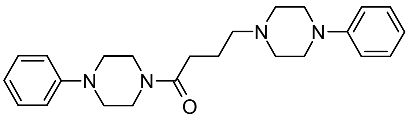
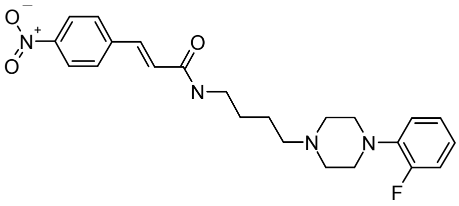
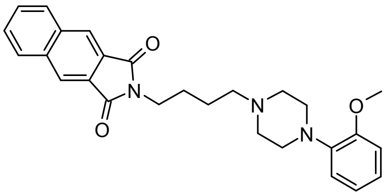
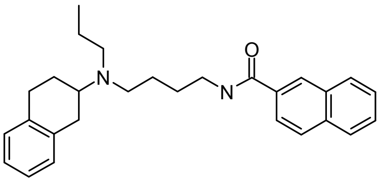
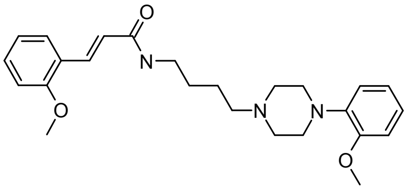
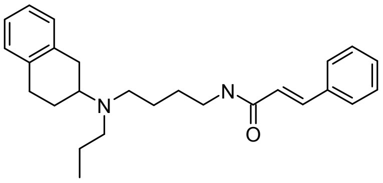
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST254		71.9	6.7	10.73
ST379		1436.4	132	10.89
ST63		295	26	11.35
ST127		700	60.5	11.57
EX44*		48.6	4.2	11.6
ST227		50.2	4.1	12.24
EX45*		17.3	1.4	12.4
ST317		128	10.2	12.549
ST217		19	1.5	12.67
ST236		19	1.5	12.67

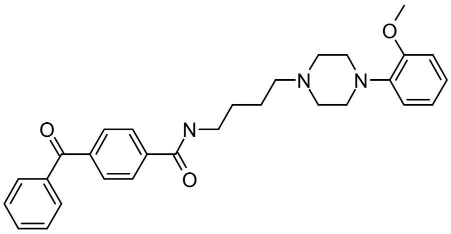
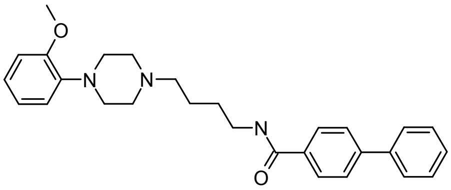
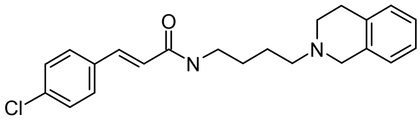
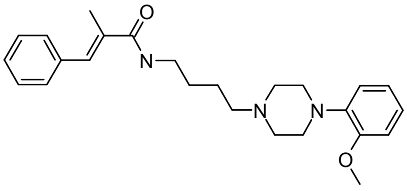
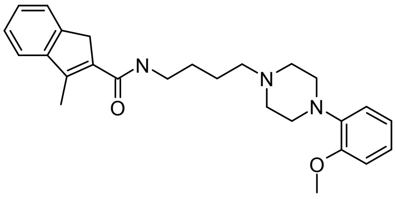
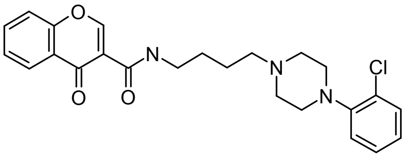
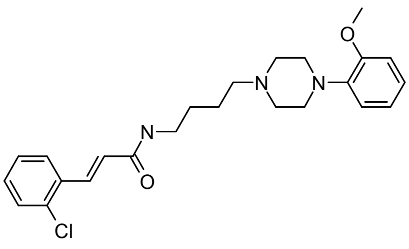
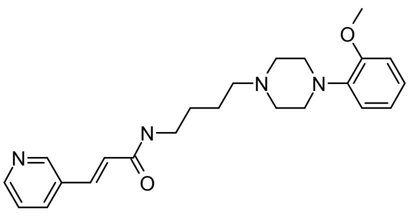
ID	**Structure**	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST144		23	1.8	12.78
ST368		890	69	12.9
ST56		600000	46000	13.04
ST353		22.9	1.72	13.31
ST182		78	5.7	13.68
EX46*		26.7	1.9	14
ST147		47	3.3	14.24
ST163		1000	70	14.29
EX47*		89.71	6.22	14.43
ST335		8	0.55	14.55
EX48*		4456	305.6	14.58
EX49*		21.6	1.5	14.7
EX50*		14.87	1.01	14.74
EX51*		14.18	0.96	14.77
EX52*		3650	242	15.08

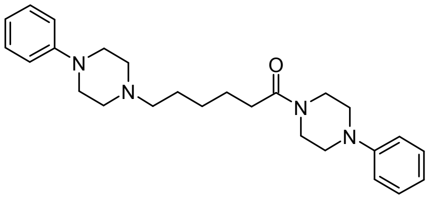
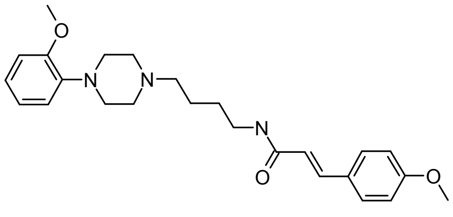
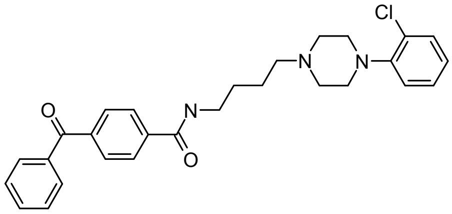
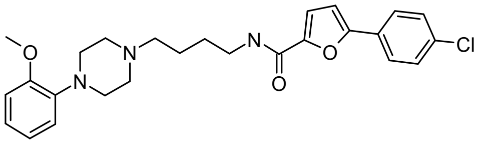
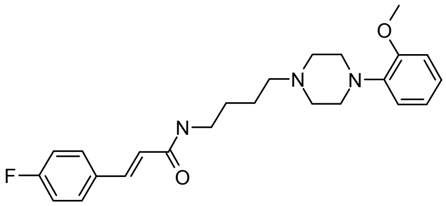
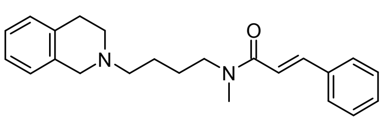
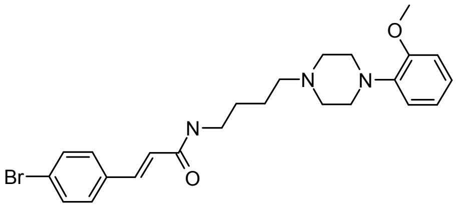
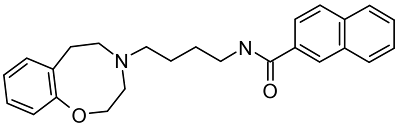
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST320		25.6	1.67	15.33
ST304		189	12.1	15.62
ST226		11.5	0.73	15.75
EX53*		3105	193	16.09
ST233		9.5	0.59	16.1
EX54*		100000	6200	16.13
ST193		21	1.28	16.41
ST138		47	2.8	16.79
ST189		36.79	2.14	17.19
EX55*		50000	2900	17.24
EX56*		12.69	0.72	17.5
ST96		17.5	1	17.5

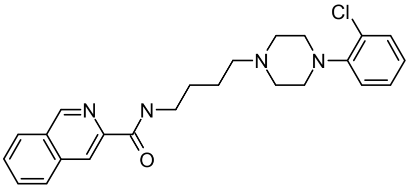
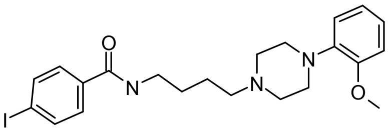
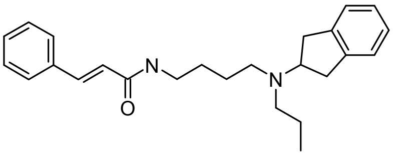
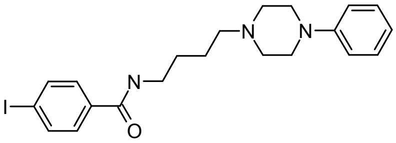
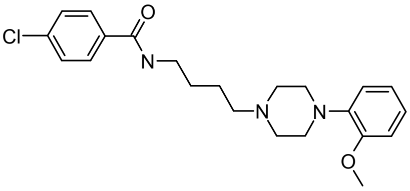
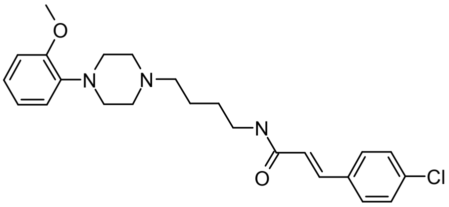
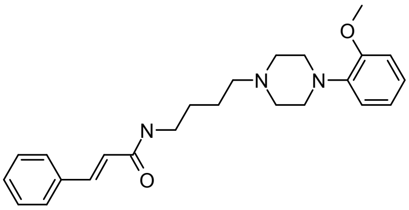
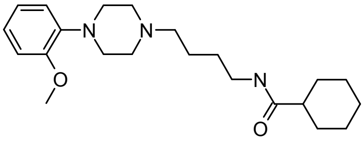
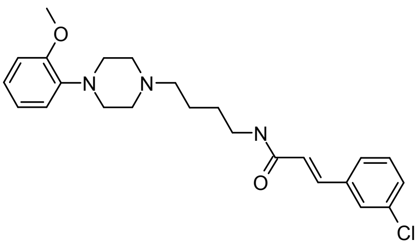
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
EX57*		2540	86.7	17.51
ST192		23	1.31	17.56
ST21		3628	204	17.78
ST334		11.1	0.6	18.5
EX58*		106	5.7	18.5
ST99		27.46	1.48	18.55
ST162		300	16	18.75
ST190		14	0.74	18.92
EX59*		837	39.9	19
ST174		840	44	19.09
EX60*		13.51	0.71	19.1
ST264		830	42.7	19.44
EX61*		14	0.71	19.8

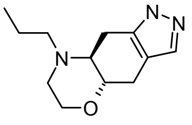
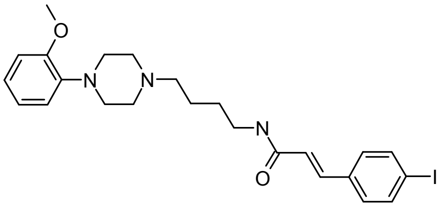
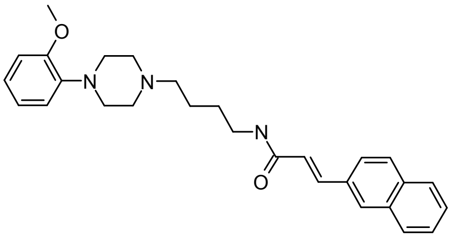
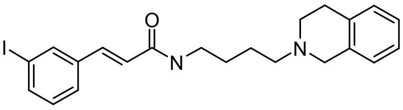
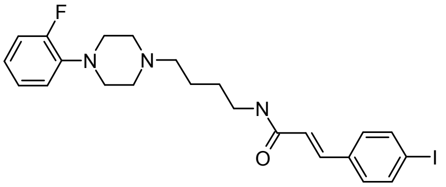
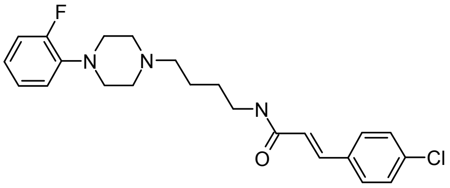
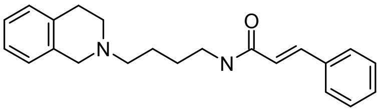
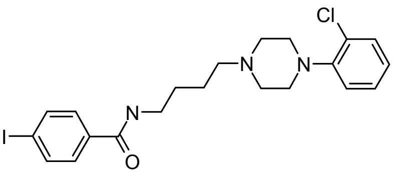
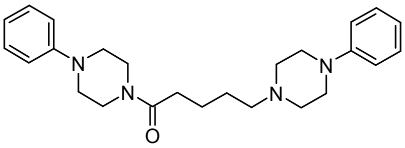
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST339		17.44	0.88	19.82
ST323		186863	9291	20.11
ST234		13	0.64	20.31
ST37		16000	780	20.51
EX62*		901	43.8	20.6
ST145		63	3	21
EX63*		52	2.5	21.2
ST337		170	7.93	21.44
ST197		10	0.46	21.74
ST80		2200	98	22.45
ST303		27	1.2	22.5
ST126		630	28	22.5

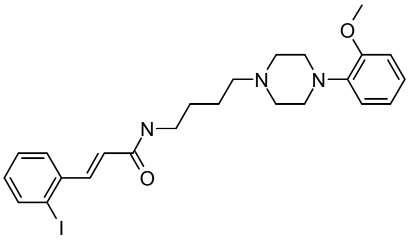
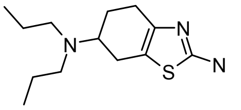
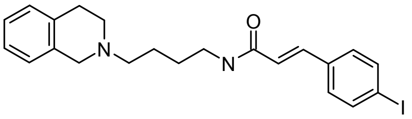
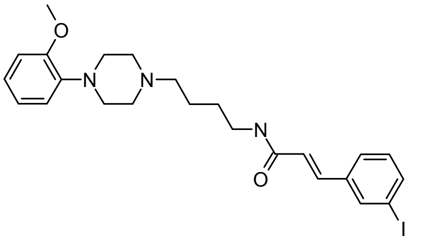
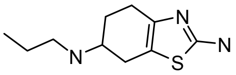
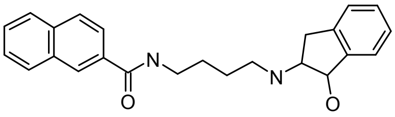
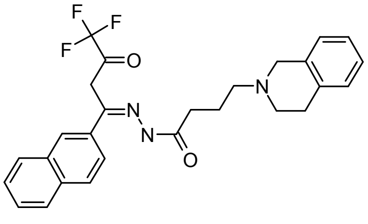
ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
EX64*		7463	327.5	22.8
ST104		46.62	2.04	22.85
EX65*		18.4	0.8	23.1
EX66*		313	13.5	23.2
ST195		420	18	23.33
ST305		2569	109	23.57
ST338		57	2.41	23.65
ST151		140	5.9	23.73
ST161		200	8.1	24.69
ST220		14	0.56	25
ST259		350	14	25

ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST95		63	2.5	25.2
ST82		75	2.85	26.32
ST184		293	11	26.64
ST310		30	1.1	27.27
ST290		50	1.8	27.78
ST169		40	1.4	28.57
ST196		15	0.52	28.85
ST279		14.5	0.5	29

ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST378		290	9.89	29.32
EX67*		61.89	2.11	29.34
EX68*		152.5	5.16	29.58
EX69*		2744	92.75	29.6
EX70*		223	7.5	29.7
EX71*		1388	45.9	30.2
ST219		12	0.38	31.58
ST130		190	6	31.67
ST152		200	6.3	31.75
ST187		19.7	0.61	32.3
ST277		920	26	35.39
ST188		24.5	0.69	35.51
ST341		4023	113	35.6

ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST131		100	2.74	36.5
ST71		146	3.9	37.44
ST260		326	8.5	38.35
ST65		389	10	38.9
ST167		19.5	0.5	39
ST168		24.2	0.61	39.67
EX72*		89.2	2.25	39.7
ST186		13.54	0.33	41.03
ST376		29.73	0.71	41.87
ST191		20	0.46	43.48

ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
ST54		100000	2300	43.48
ST280		29.3	0.64	45.78
ST218		32	0.68	47.06
EX73*		53000	1115	47.53
EX74*		70	1.4	50
ST284		1700	34	50
ST319		166.1	3.2	51.91
ST336		190	3.44	55.23
EX75*		43.7	0.74	58.8
ST198		720	12	60
EX76*		8361	138	60
ST129		190	2.8	67.86
EX77*		1446	21.3	67.9
ST314		467	6.5	71.85

ID	Structure	K_iD_2[nM]	K_iD_3 [nM]	D_2/D_3
EX78*		7514	101.97	73.69
ST282		56	0.67	83.58
EX79*		386	3.9	98.97
EX80*		4500	42	107.14
ST32		1446	13.1	110.38
EX81*		8360	73.4	114
EX82*		65	0.56	116.07
ST283		1500	12.2	122.95
ST281		150	1.21	123.97
ST19		1325	9.6	138.02
ST164		1000	2.5	400
ST173		18000	45000	2.5

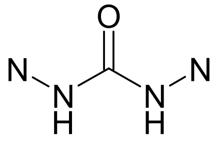
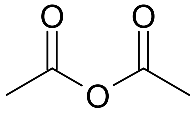
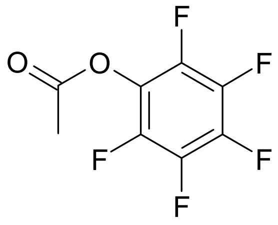
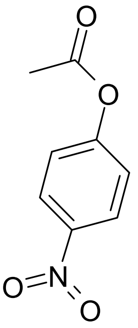
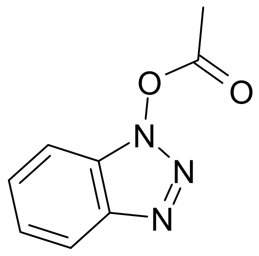
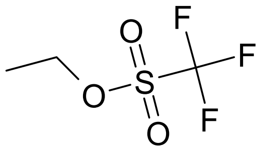
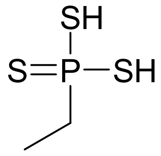
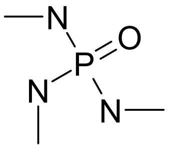
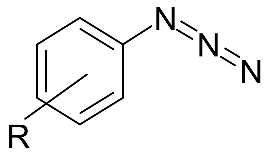
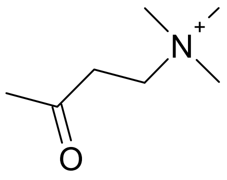
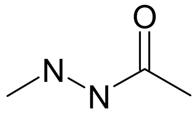
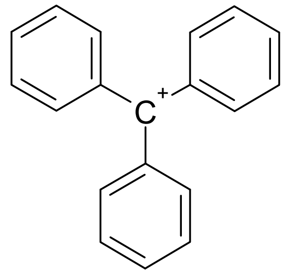
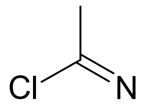
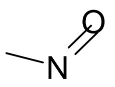
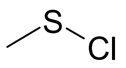
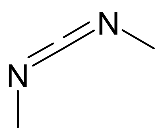
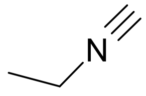
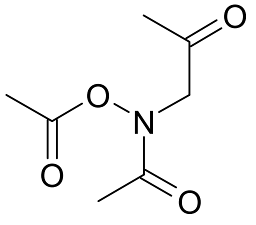
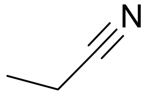
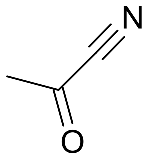
* Structure not shown due to proprietary reasons.

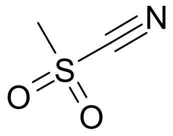
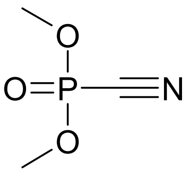
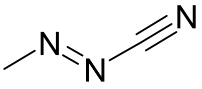
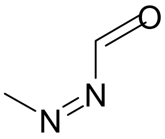
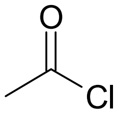
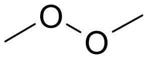
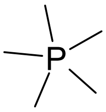
B Fisher’s Iris Data Set

ID	Sep_len	Sep_wid	Pet_len	Pet_wid
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1
14	4.3	3	1.1	0.1
15	5.8	4	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.5	0.4
23	4.6	3.6	1	0.2
24	5.1	3.3	1.7	0.5
25	4.8	3.4	1.9	0.2
26	5	3	1.6	0.2
27	5	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2
29	5.2	3.4	1.4	0.2
30	4.7	3.2	1.6	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.2
36	5	3.2	1.2	0.2
37	5.5	3.5	1.3	0.2
38	4.9	3.6	1.4	0.1
39	4.4	3	1.3	0.2
40	5.1	3.4	1.5	0.2
41	5	3.5	1.3	0.3
42	4.5	2.3	1.3	0.3
43	4.4	3.2	1.3	0.2
44	5	3.5	1.6	0.6
45	5.1	3.8	1.9	0.4
46	4.8	3	1.4	0.3
47	5.1	3.8	1.6	0.2
48	4.6	3.2	1.4	0.2
49	5.3	3.7	1.5	0.2
50	5	3.3	1.4	0.2
51	7	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4	1.3
55	6.5	2.8	4.6	1.5
56	5.7	2.8	4.5	1.3
57	6.3	3.3	4.7	1.6
58	4.9	2.4	3.3	1
59	6.6	2.9	4.6	1.3
60	5.2	2.7	3.9	1.4
61	5	2	3.5	1
62	5.9	3	4.2	1.5
63	6	2.2	4	1
64	6.1	2.9	4.7	1.4
65	5.6	2.9	3.6	1.3
66	6.7	3.1	4.4	1.4
67	5.6	3	4.5	1.5
68	5.8	2.7	4.1	1
69	6.2	2.2	4.5	1.5
70	5.6	2.5	3.9	1.1
71	5.9	3.2	4.8	1.8
72	6.1	2.8	4	1.3
73	6.3	2.5	4.9	1.5
74	6.1	2.8	4.7	1.2
75	6.4	2.9	4.3	1.3
76	6.6	3	4.4	1.4
77	6.8	2.8	4.8	1.4
78	6.7	3	5	1.7
79	6	2.9	4.5	1.5
80	5.7	2.6	3.5	1
81	5.5	2.4	3.8	1.1
82	5.5	2.4	3.7	1
83	5.8	2.7	3.9	1.2
84	6	2.7	5.1	1.6
85	5.4	3	4.5	1.5
86	6	3.4	4.5	1.6
87	6.7	3.1	4.7	1.5
88	6.3	2.3	4.4	1.3
89	5.6	3	4.1	1.3
90	5.5	2.5	4	1.3
91	5.5	2.6	4.4	1.2
92	6.1	3	4.6	1.4
93	5.8	2.6	4	1.2
94	5	2.3	3.3	1
95	5.6	2.7	4.2	1.3
96	5.7	3	4.2	1.2
97	5.7	2.9	4.2	1.3
98	6.2	2.9	4.3	1.3
99	5.1	2.5	3	1.1
100	5.7	2.8	4.1	1.3
101	6.3	3.3	6	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3	5.8	2.2
106	7.6	3	6.6	2.1
107	4.9	2.5	4.5	1.7
108	7.3	2.9	6.3	1.8
109	6.7	2.5	5.8	1.8
110	7.2	3.6	6.1	2.5
111	6.5	3.2	5.1	2
112	6.4	2.7	5.3	1.9
113	6.8	3	5.5	2.1
114	5.7	2.5	5	2

115	5.8	2.8	5.1	2.4
116	6.4	3.2	5.3	2.3
117	6.5	3	5.5	1.8
118	7.7	3.8	6.7	2.2
119	7.7	2.6	6.9	2.3
120	6	2.2	5	1.5
121	6.9	3.2	5.7	2.3
122	5.6	2.8	4.9	2
123	7.7	2.8	6.7	2
124	6.3	2.7	4.9	1.8
125	6.7	3.3	5.7	2.1
126	7.2	3.2	6	1.8
127	6.2	2.8	4.8	1.8
128	6.1	3	4.9	1.8
129	6.4	2.8	5.6	2.1
130	7.2	3	5.8	1.6
131	7.4	2.8	6.1	1.9
132	7.9	3.8	6.4	2
133	6.4	2.8	5.6	2.2
134	6.3	2.8	5.1	1.5
135	6.1	2.6	5.6	1.4
136	7.7	3	6.1	2.3
137	6.3	3.4	5.6	2.4
138	6.4	3.1	5.5	1.8
139	6	3	4.8	1.8
140	6.9	3.1	5.4	2.1
141	6.7	3.1	5.6	2.4
142	6.9	3.1	5.1	2.3
143	5.8	2.7	5.1	1.9
144	6.8	3.2	5.9	2.3
145	6.7	3.3	5.7	2.5
146	6.7	3	5.2	2.3
147	6.3	2.5	5	1.9
148	6.5	3	5.2	2
149	6.2	3.4	5.4	2.3
150	5.9	3	5.1	1.8

C Example Structures of Reactive Functional Groups

Example structures of reactive functional groups			
Carbazide	Acid anhydride	Para-fluoro-phenyl ester	Para-nitro-phenyl ester
			
HOBT ester	Triflate	Lawesson’s reagent	Phosphor amide
			
Aromatic azide	Beta-carbonyl-quaternary nitrogen	Acyl-hydrazine	Cationic C/Cl/I/P/S
			
Chloramidine	Nitroso	P or S halide	Carbodiimide
			
Isonitrile	Triacycloxime	Cyanohydrins	Acyl cyanide
			

Example structures of reactive functional groups			
Sulfonyl cyanide	Cyano phosphonate	Azo-cyan-amide	Azoalkanal
			
Acid halide	Peroxide	Phosphoane	
			

D Parameter file for SVR

```
# parameter file for LibSVM
# -------------------------------------------------------------------------
# file info
working directory               :
file with training data         :
file with test data             :
file with validation data       :
file with cv results            : data.opti
file with top models            : data.top
log file                        : data.log
# -------------------------------------------------------------------------
# data manipulations
x statistics (yes/no, bins)     : yes, 100
transform y (no)                :
scale x data    (box)           : box
lower bound for attributes (-1) :
upper bound for attributes (1)  :
scale y data    (no)            :
lower bound for y (-1)          :
upper bound for y (1)           :
classify data (no)              : no
separators (transformed ys)     :
labels (-1/+1)                  :
# -------------------------------------------------------------------------
# SVM specification
SVM-type (C-SVC)                : eps-SVR
   C-SVC                (classification)
   nu-SVC               (classification)
   one-class SVM        (classification)
   eps-SVR              (regression)
   nu-SVR               (regression)
kernel type (RBF)               :
   linear                       : K(u,v) = u*v
   polynomial                   : K(u, v) = (u*v+coef0)^degree
   RBF, radial basis function   : K(u, v) = exp(-gamma*|u-v|^2)
   sigmoid                      : K(u, v) = tanh(gamma*u*v+coef0)
# -------------------------------------------------------------------------
# model selection (note: gridsearch only for C-SVC, eps-SVR, RBF kernel)
optimization ? (yes)            :
start log2C (-5)                :
end log2C (15)                  :
step log2C (2)                  : 0.2
start log2gamma (3)             :
end log2gamma (-15)             :
step log2gamma (2)              : -0.2
cross validation mode (5)       : 7
model selection mode (best)     : auto, 2.5
number of top models (-1)       :
worst cv result that is allowed :
number of random partitions     : 10
number of random y shuffles (-1):
cv criterion (err/acc/corr)     : corr
search mode (stupid / permute)  : shuffle
compute mode (n locals / rlogin): 12
# -------------------------------------------------------------------------
# SVM parameters (if not chosen via optimization procedure)
degree (3)                      : 5
gamma (1/attributes)            :
```

```
coef0 (0)                       :
cost parameter (1)              :
nu (0.5)                        :
epsilon in eps-SVR (0.1)        : 0.1
```

E Parameter file for Bayesian regularized artificial neural networks

```
#--------------------------------------------------------
# PARAMETER FILE FOR THE TRAINING OF
# BAYESIAN REGULARIZED NEURAL NETS USING
# N E A L 'S   F B M   S O F T W A R E
#
#--------------------------------------------------------
# Location of output files
#--------------------------------------------------------
Working directory:
#--------------------------------------------------------
# Data
# use full paths only !
#--------------------------------------------------------
Training data:
Test data:
#--------------------------------------------------------
# Data pre-treatment
# Exclude X variables with a stdev lower than a predefined threshold
# Scaling
# options: uv -     unit variance and mean centering
#          box -    scale to [-1, 1]
#          pareto - mean centering and 1/sqrt(stdev)
#          center - mean centering
# Transformation of Y variables
# options: log10, log
# Classification: specify boundaries as comma-separated list
# Class labels: are generated automatically according to the
#               class boundaries given above
#               -> integer list: 0,1,2,3....
#--------------------------------------------------------
X threshold : 0.0005
X scaling: uv
#Y scaling: uv
Classification: 0.5

#--------------------------------------------------------
# Net architecture
#--------------------------------------------------------
Units in hidden layer: 10
#--------------------------------------------------------
#  priors for groups of weights, biases, and offsets:
#
#  ti [ ih bh th { hh ih bh th } ] { ho } io bo  [ / { ah } ao]
#
#  ti = offsets of input units
#  ih = input hidden weights
#  bh = hidden bias
#  th = hidden unit offsets
#  hh = hidden to hidden weights
#  ho = hidden-output weights
#  io = input output weights
#  bo = output bias
#  ah =
#  ao =
#
#--------------------------------------------------------
#  [x]Width[:[alpha-group][:[alpha-sub-group][:[alpha parameter]]]][!]
#
```

```
# examples for the prior specification for input hidden weights (can be
empty):
# ih x:                 yes/no
# ih width:       0.05
# ih alpha-group: 0.5
# ih alpha-sub-group:
# ih alpha parameter:
#
#--------------------------------------------------------
#  prior for offsets of input units (ti)
#--------------------------------------------------------
ti x:
ti width:
ti alpha-group:
ti alpha-subgroup:
ti alpha parameter:
#--------------------------------------------------------
#  prior for the input hidden weights (ih)
#--------------------------------------------------------
ih x:
ih width:         0.05
ih alpha-group:         0.5
ih alpha-subgroup:
ih alpha parameter:
#--------------------------------------------------------
#  prior for hidden bias (bh)
#  (value x makes no sense)
#--------------------------------------------------------
bh width:         0.05
bh alpha-group:         0.5
bh alpha-subgroup:
bh alpha parameter:
#--------------------------------------------------------
#  prior for hidden unit offsets (th)
#--------------------------------------------------------
th x:
th width:
th alpha-group:
th alpha-subgroup:
th alpha parameter:

#--------------------------------------------------------
#  prior for hidden to output weights (ho)
#--------------------------------------------------------
ho x:             yes
ho width:         0.05
ho alpha-group:         0.5
ho alpha-subgroup:
ho alpha parameter:

#--------------------------------------------------------
#  prior for input output weights (io)
#--------------------------------------------------------
io x:
io width:
io alpha-group:
io alpha-subgroup:
io alpha parameter:

#--------------------------------------------------------
#  prior for output bias (bo)
#  (value x makes no sense)
#--------------------------------------------------------
```

```
bo width:         100
bo alpha-group:
bo alpha-subgroup:
bo alpha parameter:

#---------------------------------------------------------
# Specify model to use for target variables:
#
# Models are available for regression with real-valued targets, logistic
# regression with binary targets, generalized logistic regression with
# targets taking on a finite set of values, and survival analysis with a
# hazard function that may depend on various covariates, and on time.
#
# real noise-prior [ "acf" corr { corr } ]
#      | binary | count | class | survival ...
#
# example:
#   Target value type:        real
#   noise prior width:              0.05
#   noise prior alpha-group:
#   noise prior alpha-sub-group:
#   noise prior alpha parameter:
#
#---------------------------------------------------------
Target value type:            real
Noise prior width:            0.05
Noise prior alpha-group:
Noise prior alpha-sub-group:
Noise prior alpha parameter:

#---------------------------------------------------------
# Training:
# I N I T I A L     P H A S E
#---------------------------------------------------------
#  A value for the fixed hyperparameters can be set here:
#  (can be empty)
#
# example:
#   value of fixed hyperparameters in the initial phase:    0.5
#
#---------------------------------------------------------
Fixed hyperparameters during initial phase:     0.5

#---------------------------------------------------------
#  Each iteration consists of how many repetitions of the
#  initial sample-noise heat bath hybrid Monte Carlo operations?
#
#  example:
#    Number of repetitions in the initial phase:      10
#---------------------------------------------------------
Repetitions during initial phase:   10

#---------------------------------------------------------
#  Initial phase:
#  Hybrid Monte Carlo update with a trajectory of how
#  many leapfrog steps long?
#
#  example:
#    Number of steps in the initial phase:      100
#---------------------------------------------------------
Steps during initial phase:   100

#---------------------------------------------------------
```

```
#  Window size of states at the beginning and end of the
#  trajectory which determines whether a state is accepted
#  or not
#  (can be empty; default = 1, i.e. standard hybrid Monte Carlo)
#
#  example:
#    Window size in the initial phase:    10
#---------------------------------------------------------
Window size during initial phase:   10

#---------------------------------------------------------
#  Step size adjustment in the initial phase:
#
#  example:
#    Step size adjustment in the initial phase: 0.2
#---------------------------------------------------------
Step size adjustment during initial phase:      0.2

#---------------------------------------------------------
#  Training:
#  S A M P L I N G    P H A S E
#---------------------------------------------------------
#  Operations executed in the sampling phase:
#
#  Define the operation type:
#
#  Two possibilities:
#      - Standard hybrid Monte Carlo ("standard")
#      - Persistent Hybrid Monte Carlo ("persistent")
#
#  example:
#    Sampling mode:     standard
#---------------------------------------------------------
Sampling mode:    persistent

#---------------------------------------------------------
#  Each iteration consists of how many repetitions of the
#  sample-noise heat bath hybrid Monte Carlo operations?
#
#  example:
#    Repetitions during sampling phase:   10
#---------------------------------------------------------
#Repetitions during sampling phase: 10

#---------------------------------------------------------
#  Only in case of operation type "persistent", the decay
#  can be specified for the heat bath operation:
#
#  If decay is zero (the default), the current momentum
#  is forgotten, and new values are picked randomly from
#  their distribution.  If decay is non-zero, the momentum
#  variables are multiplied by decay, and Gaussian noise
#  with variance 1-decay^2 is then added.
#
#  example:
#    Heat bath decay:   0.3
#---------------------------------------------------------
Heat bath decay:  0.3

#---------------------------------------------------------
#  Hybrid Monte Carlo Method, sampling phase:
#  How many steps? resp. number of stats along a trajectory?
#
```

```
# example:
#    Number of steps in the sampling phase:     1000
#---------------------------------------------------------
Steps during sampling phase:  1000

#---------------------------------------------------------
#  Hybrid Monte Carlo Method:
#  Window size of states at the beginning and end of the
#  trajectory which determines whether a state is accepted
#  or not
#  (can be empty; default = 1, i.e. standard hybrid Monte Carlo)
#
#  example:
#    Window size in the sampling phase:   10
#---------------------------------------------------------
Window size during sampling phase:  10

#---------------------------------------------------------
#  Hybrid Monte Carlo Method:
#  Step size adjustment in the sampling phase:
#
#  example:
#    Step size adjustment in the sampling phase:      0.4
#---------------------------------------------------------
Step size adjustment during sampling phase:     0.3

#---------------------------------------------------------
# Number of iterations calculated?
#
# example:
# Number of iterations: 400
#---------------------------------------------------------
Number of iterations during sampling phase: 1000

#---------------------------------------------------------
# P R E D I C T I O N
#---------------------------------------------------------
# The last n iterations are used for the prediction of the
# Y-values (calculation of the mean of the last n iterations).
# How many iterations shall be used for the prediction?
# Maximum number provided by the BRANN software is 200.
#
# default: 100
#
# example:
# Number of iterations for the prediction: 100
#---------------------------------------------------------
Number of iterations used for predictions: 200
```

F Parameter file for GOLD

```
  GOLD CONFIGURATION FILE

  generated by gold front end (GOLD v2.2)


  POPULATION
popsiz = 100
select_pressure = 1.100000
n_islands = 5
maxops = 100000
niche_siz = 2

  GENETIC OPERATORS
pt_crosswt = 95
allele_mutatewt = 95
migratewt = 10

  FLOOD FILL
radius = 12
origin = 59.06 9.916 -9.896
do_cavity = 1
floodfill_atom_no = 0
cavity_file = cavity.atoms
floodfill_center = point

  DATA FILES
protein_datafile data.pdb
ligand_data_file data.sdf 10
param_file = DEFAULT
set_ligand_atom_types = 1
set_protein_atom_types = 1
directory = .
tordist_file = DEFAULT
make_subdirs = 0
save_lone_pairs = 1
fit_points_file = fit_pts.mol2
read_fitpts = 0

  FLAGS
display = 0
internal_ligand_h_bonds = 0
n_ligand_bumps = 0
flip_free_corners = 0
flip_amide_bonds = 0
flip_planar_n = 1
flip_pyramidal_n = 0
use_tordist = 1
start_vdw_linear_cutoff = 4
initial_virtual_pt_match_max = 2.5

  TERMINATION
early_termination = 0
n_top_solutions = 3
rms_tolerance = 1.5

  CONSTRAINTS
force_constraints = 0
```

```
  COVALENT BONDING
covalent = 0


  SAVE OPTIONS
save_score_in_file = 1
save_protein_torsions = 1
concatenated_output = results.sdf
clean_up_option delete_all_solutions
clean_up_option delete_redundant_log_files
clean_up_option delete_empty_directories
output_file_format = MACCS

  RUN TYPE
```

11 Curriculum Vitae

PERSONAL DATA

Name	Alexander D. Böcker-Felbek
Date of Birth	September 5, 1976
Place of Birth	Laupheim, Germany

WORK EXPERIENCE

2006 – present	**Research Scientist.** Boehringer Ingelheim Ltd., Laval, Canada.

UNIVERSITY EDUCATION

2003 – 2006	**Performing of Ph.D. Thesis**: “Identification of Structure Activity Relationships in High-Throughput Screening Assays”. Johann Wolfgang Goethe -University, Institute of Organic Chemistry and Chemical Biology, Frankfurt am Main, Germany, and Boehringer Ingelheim Pharma GmbH & CO. KG, Biberach, Germany. Supervisors: Prof. Dr. Gisbert Schneider and Dr. Andreas Teckentrup.
1997 - 2003	**Study of Biology**. Albert-Ludwigs University, Freiburg, Germany.
2003	**Diploma in Biology**.
2002 - 2003	**Performing of Diploma Thesis**: “Searching for Novel Interaction Partners of the C-terminus of TCF4E”. Albert-Ludwigs University, Institute of Molecular Medicine, Freiburg, Germany. Supervisors: Prof. Dr. Rolf Kemler and Prof. Dr. Andreas Hecht.
2002	**Diploma Exams**: Biochemistry, Molecular Biology/ Genetics, Bioinformatics, Computer Science.
1999	**Intermediate Diploma Biology**.

SCHOOL EDUCATION

1996	**Abitur** Hochrhein Gymnasium Waldshut, Germany.