Analysis of Biological Screening Data and Molecular Selectivity Profiles Using Fingerprints and Mapping Algorithms

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich–Wilhelms–Universität Bonn

vorgelegt von
Ingo Vogt
aus
Neuss

Bonn
Juli 2008

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich–Wilhelms–Universität Bonn.

1. Gutachter: Prof. Dr. rer. nat. Jürgen Bajorath
2. Gutachter: Prof. Dr. med. Joachim L. Schultze
Tag der Promotion: 24. September 2008

Erscheinungsjahr: 2008

Für meine Eltern Rita, Gabi und Norbert, meine Brüder Daniel und Sascha,
meine Großeltern Christel und Walter sowie Gerti und Theo,
meine Tanten Ellen, Jutta und Isabell, meinen Onkel Manfred
und meine Cousinen Yvi und Melissa,
und für Kerstin.

Abstract

The identification of promising drug candidates is a major milestone in the early stages of drug discovery and design. Among the properties that have to be optimized before a drug candidate is admitted to clinical testing, potency and target selectivity are of great interest and can be addressed very early. Unfortunately, optimization–relevant knowledge is often limited, and the analysis of noisy and heterogeneous biological screening data with standard methods like QSAR is hardly feasible. Furthermore, the identification of compounds displaying different selectivity patterns against related targets is a prerequisite for chemical genetics and genomics applications, allowing to specifically interfere with functions of individual members of protein families. In this thesis it is shown that computational methods based on molecular similarity are suitable tools for the analysis of compound potency and target selectivity. Originally developed to facilitate the efficient discovery of active compounds by means of virtual screening of compound libraries, these ligand–based approaches assume that similar molecules are likely to exhibit similar properties and biological activities based on the *similarity property principle*. Given their holistic approach to molecular similarity analysis, ligand–based virtual screening methods can be applied when little or no structure–activity information is available and do not require the knowledge of the target structure.

The methods under investigation cover a wide methodological spectrum and only rely on properties derived from one– and two–dimensional molecular representations, which renders them particularly useful for handling large compound libraries. Using biological screening data, these virtual screening methods are shown to be able to extrapolate from experimental data and preferentially detect potent compounds. Subsequently, extensive benchmark calculations prove that existing 2D molecular fingerprints and dynamic mapping algorithms are suitable tools for the distinction between compounds with differential selectivity profiles. Finally, an advanced dynamic mapping algorithm is introduced that is able to generate target–selective chemical reference spaces by adaptively identifying most–discriminative molecular properties from a set of active compounds. These reference spaces are shown to be of great value for the generation of predictive target–selectivity models by screening a biologically annotated compound library.

Acknowledgments

“Some days even my lucky rocketship underpants won’t help.“
Calvin

Luckily, I have had a group of remarkable companions along my journey, especially during the last stage that spans my three years of being a doctoral candidate. Without their help I might have not been able to stay calm and find my way back all the times I felt like trailing away. Teaming up with others to share a part of the way might not necessarily make you get ahead faster, but it gives us the opportunity to learn something valuable about ourselves and people in general, even if it means taking a detour.

I would like to express my gratitude to my supervisor Prof. Dr. Jürgen Bajorath for his thoughtful guidance, patience and honest encouragement. Every so often he kept me from getting lost in side shows and helped me to re–focus. I also would like to thank Prof. Dr. Joachim L. Schultze for taking the time to be my second referee. Special thanks go to Jeffrey W. Godden who accompanied me during my first year as doctoral candidate with calm and friendly advice. For their multi–faceted and invaluable aid, my dear colleagues deserve all my thanks. I would especially like to mention Jens Auer, who has always helped me with debugging my code and solving any other programming–related problems. I would also like to thank Dagmar Stumpfe and Hany Emary Ali Ahmed with whom I have worked closely and successfully with for more than one year which has taught me some important lessons in teamwork. Finally, I owe many thanks to José Batista, Jens Auer, Tom Crisman, Eugen Lounkine and Martin Vogt for their thorough proof–reading. Tom and Eugen have also been very helpful in tackling important issues from a different perspective.

Last, but not least, I am honestly grateful to my family and my close friends who are always standing by me and giving me encouragement whenever I need it.

“Follow the river ! Zack zack !”

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pharmaceutical research as well as chemical biology critically depend on the identification of small molecules that specifically bind to a target protein, thereby affecting its biological activity in a desired manner. The main goal of chemical biology is to explain the molecular and cellular functions of biological targets with the help of small molecules (Stockwell [2004]). In this context, chemical genetics (Alaimo et al. [2001]; Spring [2005]) seeks to elucidate specific molecular mechanisms by perturbing biological processes using small ligands, whereas chemical genomics (Spring [2005], Bredel and Jacoby [2004]) systematically studies therapeutic target–ligand interactions on a large scale in order to identify new targets and biologically active compounds (Bredel and Jacoby [2004]). However, the identification of small molecules that selectively bind to a given target has already long been a major goal of drug design.

The Drug Design Process

In pharmaceutical research, the process of discovering or designing a new therapeutic agent can roughly be divided into three parts as indicated in *Table 1.1.* In the beginning, possible therapeutic targets associated with a given disease, such as receptors, enzymes, DNA, or RNA have to be identified. Then, small molecules need to be found and optimized that specifically bind to a selected target and it must be demonstrated, that this interaction leads to desired therapeutic effects. Ultimately, such molecules might become clinical candidates. High–throughput–screening (HTS) is a major source for the identification of new hits or leads. HTS can process very large compound libraries in a relatively short period of time thanks to progressing robotic automation and miniaturization. If it is possible

Stage	Steps
Drug discovery	Target identification and validation Lead identification and optimization
Preclinical studies	Laboratory tests to determine the effects *in vitro* and *in vivo* Drug formulation and manufacturing
Clinical studies	Multi–phase study of safety and effectiveness in humans

Table 1.1: Drug discovery and drug development process. Drug candidates have to pass several stages to verify medicinal activity, effectiveness, and safety before being approved. The entire process takes on average ten to 12 years.

to identify lead compounds[1], they serve as a starting point for a successive optimization process aiming at the improvement of pharmacodynamic and –kinetic properties such as potency, target–selectivity, and ADME[2] properties. Drug candidates have to be thoroughly tested *in vitro* and then in non–human organisms *in vivo* in order to verify their efficacy and safety. A new drug is finally approved if it successfully passes at least three clinical trials that prove safety and effectiveness in humans. As the pass through all stages is very time– and cost–intensive[3], optimization of the drug discovery and development process is of great importance. In this context, it is believed that complementing high–throughput discovery technologies with computational approaches is necessary to increase the success rate of drug discovery projects (Jorgensen [2004]; Bajorath [2002]). Thus, there is a substantial interest in the development of computational tools that can aid in, for example, the steps of lead identification and optimization.

Biological screening data obtained from early–stage HTS experiments present a suitable starting point for computational analysis. Furthermore, methods for *virtual screening* (VS) that are able to efficiently screen large compound databases *in silico* and select a limited number of candidate structures for subsequent experimental testing complement HTS in a meaningful way (Bajorath [2002]).

[1] A lead is a prototypical structure with desired biological activity and selectivity (Bleicher et al. [2003])

[2] Acronym for Absorption, Distribution, Metabolism, and Excretion

[3] DiMasi et al. [2003], report average pre–approval costs of US$ 802 million per new drug

Analysis of High–Throughput Screening Data

In drug discovery, high–throughput screening is the most widely used approach to rapidly test large amounts of compounds that potentially modulate a disease–associated target (Macarron [2006]). Systematic methods for the efficient screening of compounds date back to the beginning of the last century when Paul Ehrlich tested more than 600 compounds as possible drugs against syphilis (Ehrlich and Bertheim [1912]). Although the scale of HTS campaigns has consistently increased ever since, having currently arrived at a volume of between one and five millions compounds per screen (Crisman et al. [2007]), the number of approved drugs per year has almost remained constant over the last decade (Bajorath [2002]). This is attributed in part to the approximate nature of HTS results, suffering from a number of systematic difficulties (Good et al. [2000]; Bajorath [2002]; Bleicher et al. [2003]). First of all, the activity threshold applied to distinguish active from inactive compounds is often arbitrarily set, so that, depending on compound library design and drugability[4] of the target, detection of false–positives and false–negatives is likely. Furthermore, Gao et al. [2002] discovered that the accuracy of predictive models based on HTS data analysis is often impaired by boundary effects that arise when compounds with activity close to the threshold are taken into account. This is further rationalized by their finding that these compounds are often more similar to each other than most and least potent hits. Additionally, measurement errors can lead to an incorrect classification of compounds with an activity close to the activity threshold of the assay. In general, biological screening data are noisy and prone to errors arising from different sources (Bajorath [2002]). Non–specific binding events, off–target binding in cell–based assays, toxic effects or promiscuous ligands, so called *frequent hitters*, are responsible for false–positives while degradation of compounds on screening plates, limited purity and low concentrations in compound mixtures can result in false–negatives. In the case of compound mixtures special care has to be taken that pooled compounds do not react in order to avoid flawed measurements.

Although differing in their conceptual origins, high–throughput and virtual screening are highly complementary disciplines in modern early–phase drug discovery programs. However, the success rate of discovery programs has not scaled with the expanded efforts put into high–throughput technologies (Bajorath [2002]). Today, it is increasingly recognized that a combination of experi-

[4] The feasibility of a target to be effectively modulated by a suitable drug candidate (Bleicher et al. [2003])

mental and computational methods early on is beneficial for the overall success of drug discovery and design (Bajorath [2001a]). It has also been found that compounds optimized for potency and selectivity might often not respond well to subsequent modifications to further improve important biophysical, biochemical, and ADME properties (Bleicher et al. [2003]). Therefore, computational filtering of screening libraries to ensure drug–likeness in advance and/or in parallel to potency and selectivity optimization is well suited to reduce the late–stage attrition of drug candidates (Bajorath [2002]). This also means that the identification of multiple hits with significant potency is highly desirable in order to facilitate the parallel multi–property optimization. As has been shown by a number of studies (Rusinko et al. [1999]; Jones-Hertzog et al. [1999]), virtual screening methods can successfully be applied to analyze HTS data in order to generate predictive models of activity that can be used for further focused and/or sequential screening. Once lead compounds are identified, quantitative structure–activity relationship (QSAR) methods are applied to correlate structural features and properties of molecules with their activity (Esposito et al. [2004]). The paradigm of QSAR analysis is to suggest small structural modifications that significantly improve the biological activity of test compounds. Therefore, QSAR analysis requires the presence of discontinuous structure–activity relationships (SARs), but exploring such SARs is also prone to significant errors (Maggiora [2006]).

Given their whole–molecule perspective, methods based on molecular similarity do not make any assumption about pharmacophores or parts of molecules that render them biologically active (Bajorath [2002]) and can thus be applied when little or no SAR information is available. Similarity methods require the presence of continuous SARs, where departures from the structures of active compounds cause gradual changes in biological activity, consistent with the *similarity property principle*, stating that *"similar molecules should have similar biological properties"* (Johnson and Maggiora [1990]). In contrast to QSAR analysis, similarity methods usually do not take differences in compound potency into account. In addition, the qualitative manner in which SARs are explored causes a limitation of similarity methods: newly identified hits are generally much less potent than the reference molecules because one deliberately departs from optimized structural motifs (Bajorath [2002]).

Therefore, virtual screening methods that can be used for potency and selectivity analysis and that additionally are able to cope with hit diversity are of great interest. Importantly, the public availability of biological screening data as provided by PubChem[5] or other initiatives presents a major opportunity for the evaluation of such methods under realistic conditions (Vogt and Bajorath [2007]; Stumpfe et al. [2007]).

Virtual Screening

Computational methods for virtual screening of compound databases can be divided into structure– and ligand–based approaches. Structure–based methods (Shoichet [2004]) try to estimate how good a small molecule binds to a target protein, for example by trying to dock it into the protein's binding site. These methods depend on the availability of three–dimensional protein structure information, whose prediction is in the spotlight of structural genomics. Also, a detailed knowledge of the binding mode is required and affinity–scoring functions remain a crucial issue of structure–based approaches like docking. However, methods that screen compounds by assessing their similarity to already known ligands are still dominant in the field of virtual screening (Bajorath [2002]). The reason for this is that information about known ligands that bind to the target or a closely related one are often easier to obtain than knowledge of the three–dimensional target structure (Bajorath [2001a]).

Aiming at the identification of novel active molecules, molecular similarity analysis was introduced in the early 1990s, drawing upon the formulation of the similarity property principle. Molecular similarity analysis captures information about molecular structure and physicochemical features, like the solubility in polar solvents, with the help of substructures and mathematical models that enable the comparison of molecules and thus quantification of their similarity (Bajorath [2001b]; Livingstone [2000]). Such mathematical models are termed molecular descriptors. Since the beginning of molecular similarity analysis, literally thousands of descriptors have been defined (Todeschini et al. [2000]). According to the type of molecular representation from which they are derived, descriptors are often classified as one–, two–, or three–dimensional. *Table 1.2* shows three examples of molecular representations of different dimensionality and provides examples of molecular descriptors. The molecular formula gives the counts of all present atoms and can thus be used, for example, to determine the molecular mass of the molecule or detect the presence of certain elements. The two–dimensional molecular structure can be represented intuitively in a molecular graph as shown in

[5] The PubChem Project
`http://pubchem.ncbi.nlm.nih.gov`

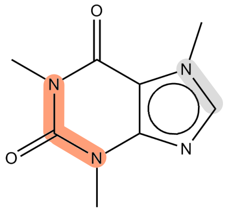
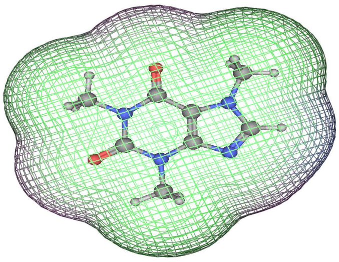
Representation		Descriptors
$C_8H_{10}N_4O_2$	⟶	molecular weight: 194.2 u number of heavy atoms: 14
	⟶	number of rings: 2 logP(o/w): -0.604 MACCS keys: 65, 77
	⟶	van der Waals volume: 175 $Å^3$ van der Waals surface area: 203 $Å^2$

Table 1.2: Representations and molecular descriptors. Depicted are one–, two–, and three–dimensional molecular representations together with several molecular descriptors. The classification of descriptors is not always strict, for instance, molecular surfaces can also be approximated from 2D representations. The highlighted substructures of the 2D molecular graph correspond to the indicated MACCS keys, which account for the presence of structural fragments, as discussed in the text.

Figure 1.2, where atoms correspond to nodes and bonds to edges. Alternatively, SMILES and InChI strings (Weininger [1988]; Stein et al. [2003]) were designed to encode the two–dimensional structure of a molecule in a one–dimensional character string. These representations permit the determination of two–dimensional features as aromatic rings or connectivity patterns and physicochemical properties like solubility. 3D descriptors are able to describe molecular properties such as the van der Waals volume or the electrostatic interaction energy, that depend on the conformation of a molecule in three–dimensional space. As another example, 3D pharmacophores represent spatial arrangements of steric and electrostatic features that are essential for bioactivity.

A survey of the spectrum of virtual screening methods is provided in *Figure 1.1*, ranging from 3D structure–based approaches like docking to 2D and 3D ligand–based methods based on appropriate molecular representations and descriptors. Ligand–based virtual screening (LBVS) methods can essentially be separated into two methodologically different approaches: *similarity searching* and

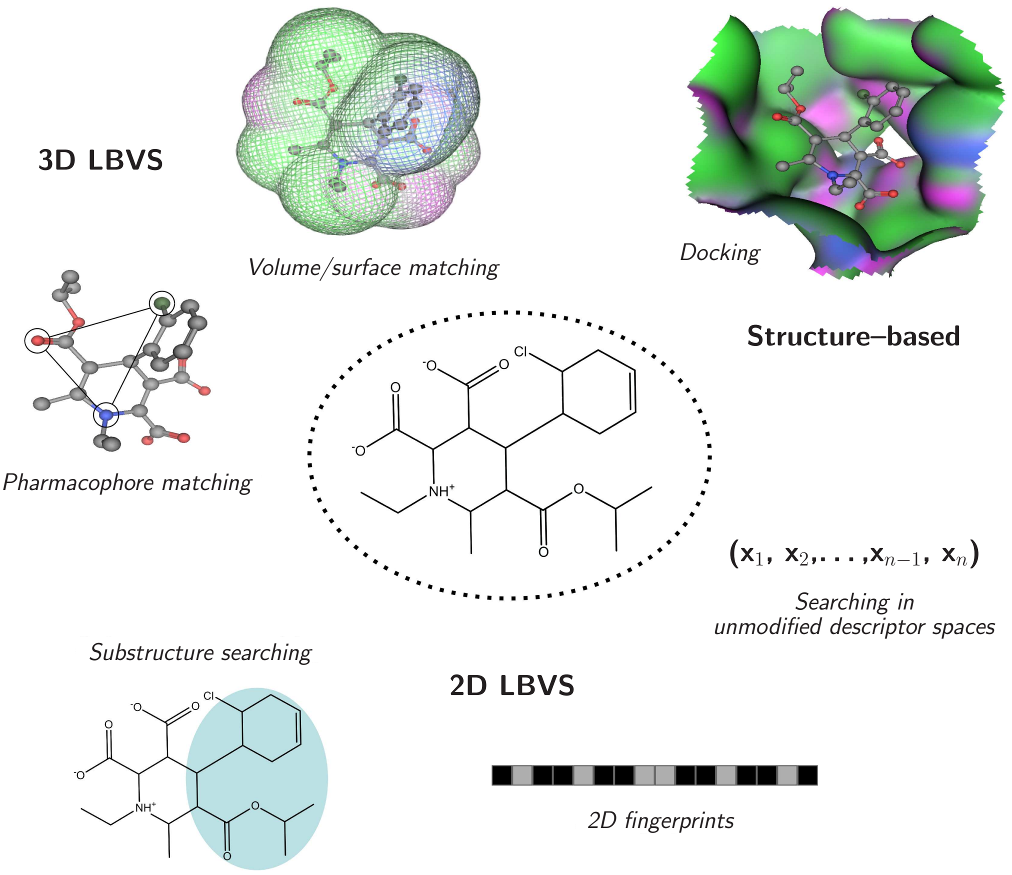


Figure 1.1: Survey of virtual screening methods and tools. Virtual compound databases can be screened by either docking compounds into the target's binding site or comparing them to known ligands. In ligand–based methods, two– and three–dimensional representations can be used to determine molecular features like substructures, connectivity patterns or physicochemical properties that can be encoded and employed for similarity analysis in various ways.

compound classification (Stahura and Bajorath [2005]). Methods used for similarity searching calculate a quantitative measure of similarity of a compound to one or more active compounds, termed *template* or reference compounds, whereas classification methods assess compounds qualitatively with respect to the properties derived from a set of multiple templates. Crucial to all of these methods is the appropriate choice of descriptors that constitute the chemical reference space into which compounds are projected (Bajorath [2001b]; Agrafiotis et al. [2002]). A widely followed scheme is the generation of low–dimensional and/or orthogonal references spaces, thereby minimizing correlation between descriptors of the reference space that is generally believed to unfavorably bias similarity analysis. Usually, generation of the reference space takes place before similarity analysis is carried out, but for some methods like recursive partitioning and dynamic mapping algorithms (Friedman [1977]; Eckert et al. [2006]), as introduced later on, chemical reference spaces are produced by the methods themselves. These methods are able to solve the descriptor–selection problem based on their ability to assess descriptor specificity in the given experimental context.

Molecular fingerprints

Molecular fingerprints are widely used descriptors in molecular similarity searching and encode information about the molecular structure and the physicochemical properties of molecules as bit strings (Bajorath [2001b], Bajorath [2002]). Molecular information encoded in fingerprints can be derived from arbitrary molecular representations such as 3D conformational models, 2D graphs, or even 1D representations like the molecular formula. Based on the way bit positions are associated with molecular features, one can further distinguish fingerprints into keyed and hashed designs. In keyed fingerprints, each bit accounts for the presence or absence of a given feature, such as a substructure, or, alternatively, whether or not the value of a property descriptor lies within a certain range. Different concepts have been suggested for the encoding of numerical descriptors into fingerprints (Xue et al. [2003]; Eckert and Bajorath [2006a]). In Xue et al. [2003], the statistical median of the screening database is used to transform the descriptor values of compounds into a dichotomous variable. Thus, if the descriptor value of a molecule is less than or equal to the database median, the corresponding bit is set to 0 - and to 1 if it is greater. The concept of a keyed fingerprint (Barnard and Downs [1997]) is illustrated in *Figure 1.2.* A popular 2D keyed fingerprint is the MACCS keys fingerprint based on a subset of 166 MDL substructures.[6]

Other examples of keyed fingerprints include designs that are based on connectivity patterns of atom types or pharmacophoric features. Via analysis of the environment of heavy atoms, a fingerprint termed MOLPRINT 2D has been created that consists of count vectors reporting the number of atom types at a given distance from the reference atom (Bender et al. [2004a]). Furthermore, patterns of 3D pharmacophores (*Figure 1.1*) are usually recorded in 3D keyed fingerprints that can contain several millions of bits based on systematic conformational enumeration (Bradley et al. [2000]). On the other hand, 2D pharmacophore fingerprint

[6] MACCS structural keys. MDL Information Systems Inc., San Leandro, CA, USA.

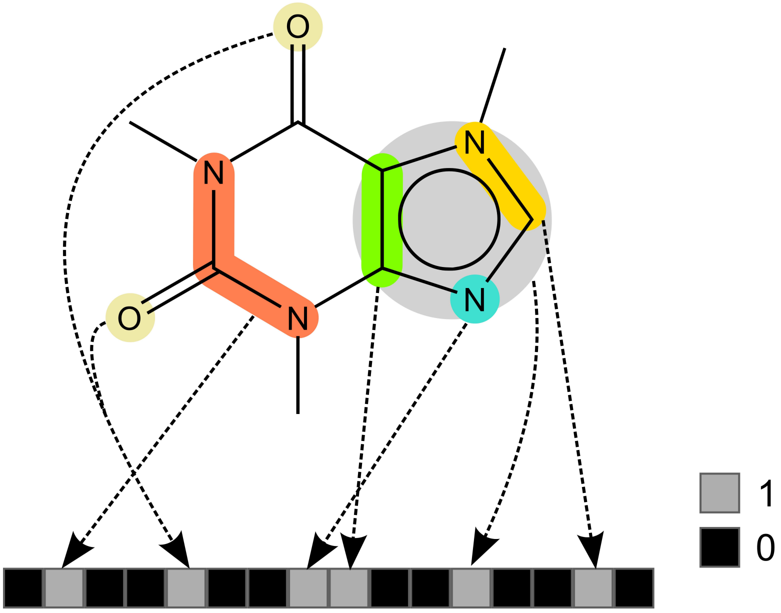


Figure 1.2: Example of a 2D keyed fingerprint. Given a set of 16 fragments, a fingerprint containing the same number of bits can be generated for a given molecule. The value of a bit is set to 1 if the fragment is present; otherwise, it is set to 0. Color–coded fragments are taken from the subset of 166 publicly available MACCS keys. Their arrangement in this example does not correspond to their position in the MACCS fingerprint used in this thesis.

designs can be created by ignoring the 3D information and replacing the spatial distances with bond distances. For instance, TGD, as implemented in MOE[7], is an example of a 2–point 2D pharmacophore fingerprint that generates pharmacophore patterns from seven atoms types such as hydrogen bond donor and acceptor, combined with their binned graph distances. Different from keyed fingerprint designs, hashed fingerprints map feature patterns to bit sets that overlap such that the presence of a certain feature is only given with some probability (James and Weininger [2008]). Daylight fingerprints are an example of 2D hashed fingerprint designs that capture connectivity pathways in molecules (James and Weininger [2008]).

The basic principle of similarity analysis on the basis of molecular fingerprints is to quantify the degree of overlap between the fingerprints of two molecules. In the past, a large variety of similarity or distance metrics have been devised (Holliday et al. [2002]) that use the information about identical and different bit

[7] Molecular Operating Environment. Chemical Computing Group, Montreal, Canada. `http://www.chemcomp.com`

settings to calculate a single value indicating the degree of similarity or dissimilarity between two molecules. In similarity searching, the screening database can then be ranked in the order of decreasing similarity to the template molecule(s). According to the similarity property principle, compounds with similar biological activity are likely to be enriched among top–ranked compounds so that one can select an arbitrary number of top–scoring compounds for activity evaluation.

Mapping algorithms for ligand–based virtual screening

An integral part of many chemoinformatics approaches, and in particular of most molecular similarity–based methods, is the generation of chemical reference spaces (Johnson and Maggiora [1990]; Stahura and Bajorath [2003, 2005]). Many methods in library design (Martin [2001]; Schnur et al. [2004]) and compound classification (Stahura and Bajorath [2003, 2005]) apply arrays of molecular property descriptors (Todeschini et al. [2000]; Livingstone [2000]) to construct spaces of chemical features that provide a basis for the analysis of molecular similarity relationships or compound diversity. In context of virtual compound screening (Stahura and Bajorath [2005]; Bajorath [2002]) and target–focused library design (Schnur et al. [2004]), the relevance of the chosen chemical space representations for the evaluation and prediction of biological activity is of paramount importance (Bajorath [2002]; Lipinski and Hopkins [2004]). This is because these methods cannot succeed if selected descriptors do not respond to activity–determining molecular features.

A number of attempts have been made to rationalize feature selection for the design of chemical reference spaces and to ensure their appropriateness for the problems under investigation. For instance, by introducing the *receptor–relevant subspace* concept, Pearlman and Smith [1999] attempted to study compounds in chemical reference spaces formed by complex orthogonal descriptors that combine chemical features generally known to be important for mediating specific receptor–ligand interactions. This concept assumes that compounds, that preferentially populate certain subspaces and cluster along selected descriptor axes, are likely to share similar biological activity. Therefore, so–generated reference spaces are generally relevant for the study of target–ligand interactions. Taking another approach, Agrafiotis et al. [2002] generated low–dimensional reference spaces by selection of those descriptors from high–dimensional space representations that are responsible for the feature variance within a set of compounds. Also, partitioning algorithms as described in Xue and Bajorath [2002] systematically search for descriptor combinations that group classes of active compounds and make them distinguishable from others. Furthermore, descriptor assessment based on information theoretic concepts has been successfully applied to guide for effective descriptor selection (Godden and Bajorath [2002]).

Recently, methods termed *mapping algorithms* have been developed that facilitate the design of activity–class directed chemical reference spaces by selecting descriptor combinations that have systematically different settings in different sets of compounds. Until now, four mapping algorithms have been introduced including Dynamic Mapping of Consensus Positions (DMC, Godden et al. [2004a]), Mapping to Activity–class specific Descriptor value ranges (MAD, Eckert and Bajorath [2006b]), Dynamic MAD (DynaMAD, Eckert et al. [2006]), and Continuous–Adaptive DynaMAD (CA–DynaMAD, Vogt and Bajorath [2008]).

Operating on binary transformed molecular descriptors (Xue et al. [2003]), DMC generates consensus bit strings that reflect preferential bit settings among a set of template compounds. These consensus bit strings are created in a stepwise manner, and each subsequent consensus bit string allows for more bit variability. During this process termed *dimension extension*, the number of consensus bit settings increases. After each dimension extension step, the bit strings of compounds from the screening database are *mapped* to the consensus bit string, that is, it is examined whether or not the bit string of a database compound matches the consensus bit string. If this is not the case, the compound is discarded; otherwise, it is retained for the next dimension extension step. As shown by Godden et al. [2004b], it is possible to direct this mapping process towards the identification of more potent compounds.

Based on the foundations of descriptor–specificity assessment in MAD, DynaMAD was developed for generating and navigating high–dimensional chemical space representations and efficient processing of very large compound databases. The method automatically selects descriptors from basis sets using a descriptor scoring function that calculates the probability of a database compound to map the descriptor value range of a set of reference molecules (typically a set of active compounds). According to their score, descriptors are assigned to dimension extension levels in order of decreasing reference–set specificity. Analogous to DMC, the mapping process iteratively maps database compounds to the value ranges of descriptors of the current dimension extension level and retains only those compounds that match all value ranges.

Being the latest mapping algorithm, CA–DynaMAD constitutes an advancement to DynaMAD in shifting the notion of reference–set specificity of descriptors. After selection of the current top–scored descriptor and subsequent compound mapping, all other descriptors are re–evaluated on the basis of the remaining database compounds. In DynaMAD, descriptor scores allow to draw conclusions about activity–specificity for each descriptor individually and independent of other descriptors. In CA–DynaMAD, however, scores for all but the first selected descriptor are dependent on the set of previously chosen descriptors. Hence, reference space generation is focused on efficiently separating the database compounds from the templates. The result is that chemical reference spaces created with CA–DynaMAD tend to be of significantly lower dimensionality than those created with DynaMAD. At the same time, they are at least equally if not more activity–specific with respect to the compound database. Thus, the chance of missing compounds in the database that are active but structurally distinct to the template set is reduced.

Thesis Outline

The investigations that are presented in this thesis attempt to answer a number of questions that focus on the aspects of ligand potency and selectivity and on the question to what extent 2D LBVS methods are suitable for their analysis:

Question 1 *Are 2D virtual screening methods capable of discriminating between active compounds with different potency for a given target?*

Question 2 *Given two sets of active compounds binding to two related targets with differing target–selectivity, can conventional two–dimensional molecular fingerprints enrich compounds with desired selectivity in reasonable sized selection sets?*

Question 3 *Is it possible to distinguish sets of compounds that are selective for one target from compounds that are active against the target family?*

Question 4 *Can one create target–selective chemical reference spaces suitable for model building as well as for large–scale virtual screening?*

For this purpose, a wide range of methods are examined that are based on different molecular features of variable complexity derived from 1D and 2D molecular representations. These methods are used to conduct virtual screening experiments aiming at the detection of active compounds with desired potency or target–selectivity using carefully assembled ligand databases and large compound collections. The inclusion of biological screening data makes it possible to validate the obtained results and evaluate their potential for practical applications.

Chapter 2 is concerned with the methodological aspects of this study. Initially, it is explained how ligand potency can be measured and how conclusions about target–selectivity can be derived. Then descriptions of the methods under investigation are provided, covering state–of–the–art as well as recently published approaches for molecular similarity analysis. Finally, an improvement of an existing dynamic mapping algorithm is introduced that is well–suited for the generation of target–selective chemical reference spaces.

In *Chapter 3*, four consecutive studies are presented in detail that were conducted in order to answer the aforementioned questions. By analyzing biological screening data it is shown that similarity search methods are able to distinguish potent ligands from less potent ones. Based on these findings, the two following studies document that compound mapping algorithms and 2D molecular fingerprints of varying complexity show promising performance in selectivity search calculations. The results of the final study substantiate the utility of chemical similarity–based approaches for analysis of ligand selectivity by generating target–selective reference spaces.

Finally, *Chapter 4* summarizes the results of these studies and presents conclusions.

Chapter 2

Methods

This chapter introduces the computational methods that are applied in the studies reported herein. Initially, it is explained how the potency of protein ligands can be experimentally measured and how target–selectivity can be derived from potency differences. The methodological spectrum of virtual screening methods investigated in this thesis covers several 2D molecular fingerprint designs, dynamic mapping algorithms, and standard classification methods like clustering and recursive partitioning.

2.1 Compound Potency and Target Selectivity

Often, high–throughput screening is applied to identify potential drug candidates in early–stage drug discovery programs, particularly when little or no knowledge about the structure of the target is available. In the context of drug discovery, an assay is a test for modulatory activity with respect to a biological or biochemical mechanism exerted by an active compound. In an automated fashion, HTS runs a screen of an assay against a large compound library in a short time period.

Many therapeutic drugs are small molecule ligands that act as enzyme inhibitors. Upon binding, the enzyme's ability to bind substrate[1] is lowered and/or its catalytic activity is decreased. Enzyme activators on the other hand, increase an enzyme's catalytic activity. In addition to enzymes, receptors are another important class of therapeutic targets (Overington et al. [2006]). The activity of receptors is regulated by binding interactions with agonists, inverse–agonists, and antagonists. Binding of antagonists to the active or an allosteric binding

[1] A chemical entity that is altered by an enzyme

site prevents agonist–induced receptor response, while inverse–agonists cause the opposite receptor response upon binding to the active site.

For testing compound activity with high–throughput screening, solutions with a defined concentration of compounds[2] are filled into the wells of microtiter plates. Depending on the type of assay, cells or enzymes are added to each well and after a predefined incubation time it is automatically measured if a desired effect has occurred. Typically, each compound is tested at a single concentration in the primary screen, so that the outcome is of a qualitative nature. That is, a compound is either classified as being active or inactive, given their ability to exert a minimum degree of target modulation. Active compounds yielded by the primary screen, so called initial hits, are then subjected to a subsequent assay for dose–response confirmation. By testing the initial hits several times at varying concentrations it is possible to calculate a measure of compound potency.

The term *potency* is generally used to refer to a quantitative measurement of a compound's ability to interfere with the function of its target and is closely associated with the terms IC_{50}, EC_{50}, and K_i. For competition binding and functional antagonist assays IC_{50} is the most common measure of potency, EC_{50} for activator/agonist assays. In the following, IC_{50} and K_i will be explained with respect to competitive enzyme inhibitors. However, these concepts analogously apply to enzyme activators and receptor agonists.

If an enzyme inhibitor competes with the enzyme's substrate to bind to its active site, it is termed a *competitive inhibitor*, and its potency can be experimentally measured in a competition or displacement assay. In this type of assay, a fixed concentration of a labeled substrate is used as a reference to determine the potency with which the unlabeled inhibitor competes for the binding. For this purpose, the specific binding of the labeled ligand is monitored at different concentrations of the unlabeled inhibitor. Afterwards, results are fitted to a logistic function[3], which is then used to determine the concentration of the unlabeled inhibitor at which the binding of the labeled ligand is half maximal (see *Figure 2.1*) (Motulsky and Christopoulos [2003]). This concentration is called the IC_{50} of the unlabeled drug, and depends on its own affinity for the target and on the binding affinity and concentration of the labeled ligand. The affinity of the unlabeled ligand for the target is expressed as its equilibrium dissociation constant K_i, which can be calculated with *Equation 2.1* (Cheng and Prusoff [1973]).

[2] Pure or a mixture of several compounds

[3] A logistic function or logistic curve models the S–curve of growth of some set P.

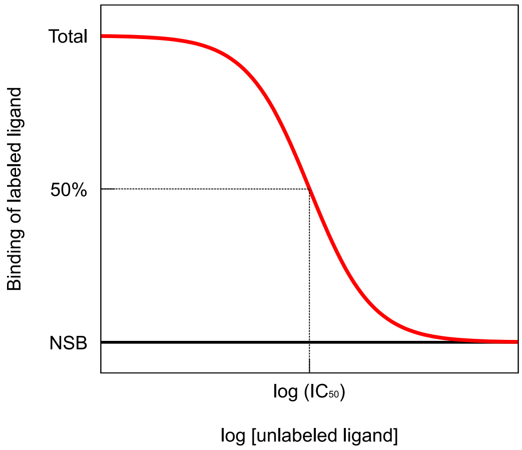


Figure 2.1: Competitive binding curve. The concentration of the unlabeled ligand at which the labeled ligand binds to half of the available active sites is called IC$_{50}$. NSB stands for *non–specific binding* and refers, for example, to binding to cell membranes.

$$K_i = \frac{IC_{50}}{1 + \frac{[labeled\ ligand]}{K_m}} \tag{2.1}$$

In *Equation 2.1*, K$_m$ is the Michaelis constant of the labeled ligand, its concentration at half maximal enzyme reaction rate in the absence of a competitive inhibitor (approx. K$_d$ under certain conditions). The K$_i$ value is the concentration of the unlabeled ligand at which it will bind to half of the binding sites at equilibrium, in the absence of other competitors (Motulsky and Christopoulos [2003]). Therefore, the lower the K$_i$ value is for an competitive ligand, the higher its affinity is for the target. Based on the nature of the IC$_{50}$ value, it should not be mistaken as a direct measure of affinity, but rather as a measure of a ligand's ability to interfere with labeled ligand/substrate binding to the active site (Motulsky and Christopoulos [2003]). However, as applied concentrations of the labeled ligand are usually at or closely below its K$_d$ one can assume that IC$_{50}$ values tend to be a low multiple of the K$_i$ values. In summary, K$_i$ and IC$_{50}$ values define a compound's ability to bind to its target in different contexts but are nevertheless interrelated as shown by Cheng and Prusoff [1973] and can thus both be regarded as comparable quantifications of compound potency.

If one needs to measure the selectivity of an active compound for a given target with respect to other targets, it is necessary to obtain information about its potency for each of the targets first. Then, the pairwise selectivity sel_i^{AB} of a compound i for target A over target B can be calculated according to *Equation 2.2*.

$$sel_i^{AB} = \frac{Pot_i^B}{Pot_i^A} \tag{2.2}$$

Pot_i^A and Pot_i^B are either the K_i or IC_{50} values of compound i for targets A and B. When $sel_i^{AB} > 1$, the compound shows selectivity for target A over target B. Usually, in order to allow a meaningful analysis of compound selectivity, one might want to apply threshold values for this measure of target–selectivity. Therefore, selectivity thresholds of 50– or 100–fold were applied in all reported target–selectivity related studies herein. Throughout this thesis, a set of active compounds, which are selective for target A over target B, is denoted by A/B.

2.2 2D Fingerprints

One of the most popular class of tools for similarity searching are molecular fingerprints, which encode information about molecular properties and structure in sets of features, mainly as ordered bit strings. Although such sets or strings are themselves 1D, they are able to capture chemical features from various higher dimensional molecular representations (2D graph, 3D surfaces, 3D pharmacophores) and thus enable computationally efficient similarity analysis in large compound databases. Its advantage over compound classification methods is that the knowledge of only one known bioactive compound is sufficient. This compounds serves as search template, while classification methods such as machine learning or Bayes classification depend on the availability of multiple template structures (Bajorath [2002]). Nevertheless, many studies have shown that fingerprint performance in similarity search calculations further increases in the majority of cases when using multiple reference compounds (Hert et al. [2006]).

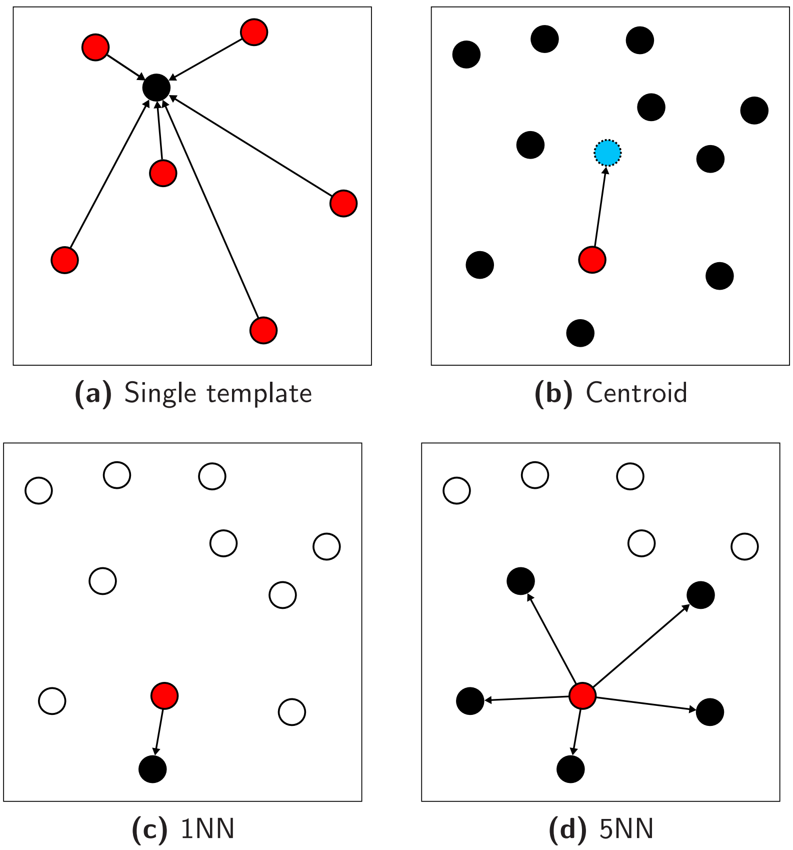
For similarity searching with molecular fingerprints one also needs to quantify the overlap of two fingerprints. A variety of such functions called similarity metrics exist (Holliday et al. [2002]), of which three of the most widely used ones are listed in *Table 2.1.* These metrics try to relate the number of commonly set bits (that is, set to 1) to the number of set bits in each of the bit strings and return similarity values between 0 and 1, where 1 denotes a perfect match. It should be noted that binary fingerprints bear no information about how often a feature is present in a molecule. If a bit is set to 1 it only means that the corresponding feature is found at least once. A understanding of this is crucial for the correct interpretation of results of similarity analysis based on fingerprints. For instance, if the Tanimoto

Metric	Range	Formula
Cosine	$[0, 1]$	$\frac{c}{\sqrt{a+b}}$
Dice	$[0, 1]$	$\frac{2 \cdot c}{a+b}$
Tanimoto	$[0, 1]$	$\frac{c}{a+b-c}$

Table 2.1: Similarity metrics for dichotomous variables. Three examples of common functions (also called *coefficients*) used to measure fingerprint similarity. The Tanimoto coefficient is also known as Jaccard coefficient. Given two fingerprint bit strings A and B, a and b are the number of bits set to 1 in the corresponding strings and c is the number of bits set to 1 in both strings. There also exist formulations for continuous variables (Willett et al. [1998]).

coefficient (Tc) calculated from the MACCS fingerprints of two compounds is 1, the only valid conclusion is that the bit strings are identical, but not necessarily the molecules themselves. To overcome this potential problem, molecular holograms have also been introduced that record the number of feature occurrences in an integer string (Flower [1998]).

When searching with only one template structure, all compounds from the screening database are compared to it and ranked in decreasing order of similarity values. If multiple templates are available there are several different search strategies, including the centroid (Schuffenhauer et al. [2003]) and nearest neighbor approaches (Schuffenhauer et al. [2003], Hert et al. [2004]). The centroid approach involves the generation of an averaged fingerprint (see *Figure 2.2b*) whose similarity to database compound fingerprints has to be measured by utilizing a metric suited for continuous variables, for example, the formulation of the Tc for continuous variables (Willett et al. [1998]). This technique, as well as related approaches such as fingerprint scaling (Xue et al. [2001]), emphasizes features that are specific for a set of active compounds and thus potentially correlated with biological activity. As a data fusion technique, the nearest neighbor approach merges pairwise similarity values between a database compound and the k most similar reference compounds by averaging. The examples shown in *Figures 2.2c* and *2.2d* visualize nearest neighbor strategies with $k = 1$ and $k = 5$ (1NN and 5NN, respectively).


(a) Single template

(b) Centroid

(c) 1NN

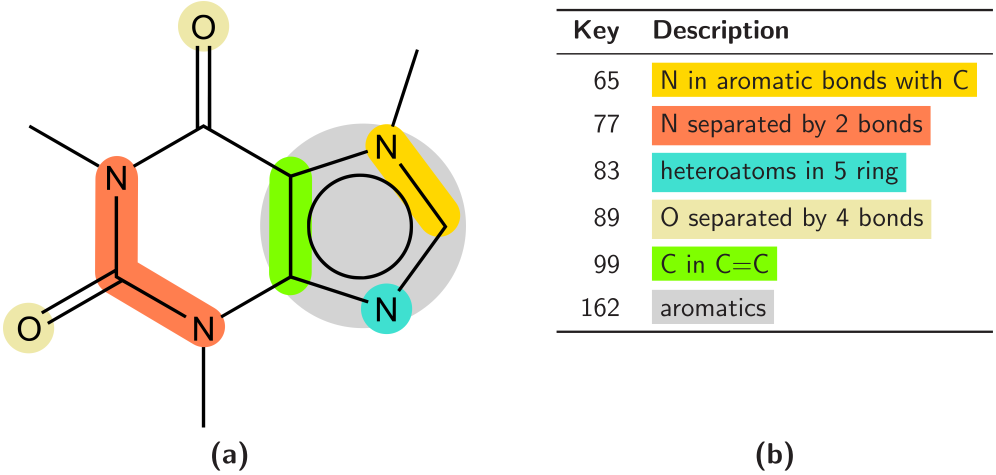
(d) 5NN

Figure 2.2: Similarity search strategies. Red dots represent database compounds and black dots and circles represent template compounds that are included in the calculation of the similarity value or not. The blue dot indicates the centroid position calculated from all template fingerprints.

2.2.1 MACCS

Probably one of the best known and widely used structural 2D fingerprints is based on a subset of MDL MACCS structural keys[4]. It monitors the presence of 166 small topological structural fragments none of which considers information about stereochemistry (McGregor and Pallai [1997]). *Figure 2.3* shows a compound containing substructural features that are detected by six MACCS keys. See *Appendix Table A.2* for a complete listing of the 166 structural keys used to generate the MACCS fingerprint.

[4] Symyx Software, San Ramon, CA, USA.



Key	Description
65	N in aromatic bonds with C
77	N separated by 2 bonds
83	heteroatoms in 5 ring
89	O separated by 4 bonds
99	C in C=C
162	aromatics

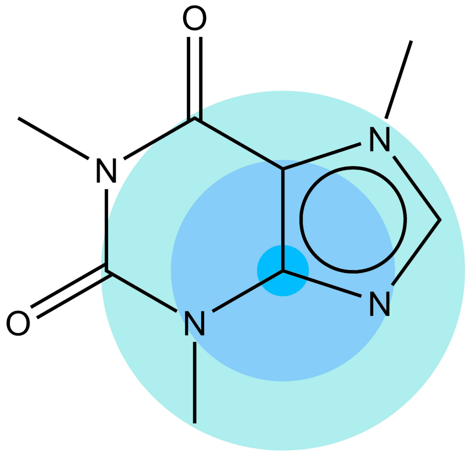
Figure 2.3: MACCS substructural keys. In (a) a caffeine molecule is shown with six highlighted substructural features that correspond to the MACCS keys listed in (b). The binary MACCS fingerprint detects the absence or presence of those features in a compound, indicated by bit settings of 0 or 1, respectively.

2.2.2 MOLPRINT 2D

Calculated from the 2D connectivity table, MOLPRINT 2D (Bender et al. [2004a,b]) represents molecules by a set of atom environments, each of which reports the occurrence of SYBYL atom types (Clark et al. [1989]) up to a given distance from the center atom. In this thesis, atom environments up to a distance of two bonds are considered for the generation of the atom environments, as recommended by Bender et al. [2004a]. *Figure 2.4* illustrates the concept of atom environments as implemented in the MOLPRINT 2D fingerprint. The fingerprint itself is not a bit string but consists of a set of strings generated from the count vectors, as shown in *Figure 2.4b*. The size of the set is correlated to the number of heavy atoms in a compound, and up to 2^{50} unique atom environments are theoretically possible (from the combinatorial point of view).

2.2.3 TGT

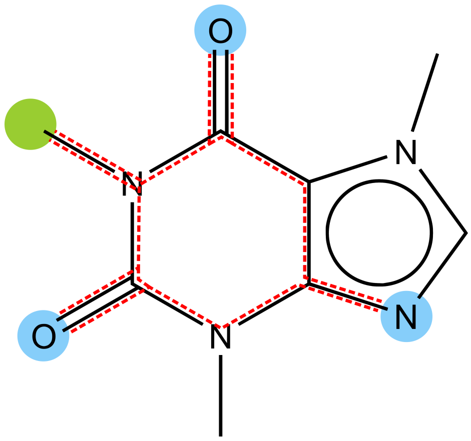
The Typed Graph Triangle fingerprint implemented in MOE calculates 3–point pharmacophore features from a 2D graph representation of a molecule. All atoms are classified according to a set of four atom types: *hydrogen bond donor or base*, *hydrogen bond acceptor or acid*, *both hydrogen bond donor and acceptor*, and *hydrophobic*. From this set of typed atoms all possible atom triplets are generated and the graph distances (that is, the number of bonds in the shortest



Layer	Atom types
0	C.ar
1	C.ar, N.2, N.am
2	C.2, N.pl3, C.2, C.3, C.2

(a) **(b)**

Figure 2.4: Illustration of atom environments as used in MOLPRINT 2D. (a) For each heavy atom, all atoms, classified according to SYBYL atom types, are reported up to a distance of two bonds. The count vectors listed in (b) are combined into a single atom feature, and the set of all such features present in a compound constitutes its corresponding MOLPRINT 2D fingerprint.



Typed graph triangles
(Don, Don, Don, 4, 4, 4)
(Don, Don, Hyd, 4, 3, 3)
(Don, Don, Hyd, 4, 5-9, 3)

(a) **(b)**

Figure 2.5: Example of three different typed graph triangles. In (a), four atoms are highlighted according to their atom type in blue (hydrogen bond donor, *Don*) and green (hydrophobic, *Hyd*). Combined with their bond distance intervals, these four atom types can used to generate three unique, symmetry–free TGT features as shown in (b).

path connecting two atoms) between all pairs are determined and binned into six distance categories {1, 2, 3, 4, 5-9, 10-}. This information is then coded as feature tuples and their presence is then recorded in the fingerprint. In total, the TGT fingerprint accounts for 1 704 unique 3–point pharmacophore features.

2.2.4 MP–MFP

As a hybrid design, the MP–MFP fingerprint (Xue et al. [2003]) originally combines 110 selected MACCS keys and 61 binary transformed molecular property descriptors, where the selection of descriptors and structural features was determined by the analysis of their information content in a large compound database. For molecular descriptors, their information content was measured on the basis of Shannon entropy analysis (Godden et al. [2000], Godden and Bajorath [2002]) in a large compound database, and those with high Shannon entropy and thus high information content were subsequently subjected to correlation analysis. Elimination of highly correlated descriptors with less information content resulted in a set of 61 molecular descriptors (Xue et al. [2003]). More than half of these descriptors carry implicit 3D information by approximating molecular surface areas from 2D representations onto which various physicochemical properties are mapped (Labute [2000]). Based on the median partitioning approach of Godden et al. [2003], selected descriptors were binary encoded based on the statistical medians in the screening database. Thus, if a compound's descriptor value is above the database median, the corresponding bit is set to 1, or 0 if it is not. In addition, 110 MACCS keys were chosen that displayed relative bit frequencies between 10% and 90%, omitting those substructural keys having low discriminatory power. The current design is summarized in *Table 2.2*, which was obtained by adapting the MACCS keys selection to the statistics of the ZINC6 database. As both bit settings (0 and 1) of the binary encoded property descriptors have an equivalent information content, a new similarity coefficient was defined based on the Tc for dichotomous variables (see *Table 2.3*). As formulated in *Equation 2.3*,

$$avTc = \frac{Tc_1 + Tc_0}{2} \tag{2.3}$$

the so called average Tc ($avTc$) is defined as average of the two Tanimoto coefficients Tc_1 and Tc_0, measuring the ratio of coincident 1s and 0s, respectively.

Descriptor class	# Descriptors
Complex surface area descriptors	31
Atom and bond counts	16
van der Waals surface descriptors	6
Topological descriptors	4
Partial charge descriptors	3
Physicochemical descriptors	1

(a)

MACCS keys
19, 32, 33, 36, 38, 42, 50-55, 57-62, 64-67, 69, 71-100, 102-149, 151-162, 164

(b)

Table 2.2: Constitution of the MP–MFP fingerprint. For the presented studies the MP–MFP design was adapted to the ZINC6 database, combining (a) 61 2D molecular property descriptors with (b) 113 MACCS keys into a fingerprint consisting of 174 bits. See *Appendix Chapter A* for details on the used molecular descriptors and MACCS keys.

2.2.5 PDR–FP

The ultimate goal in ligand–based virtual screening is the identification of compounds with diverse structures and similar activity to template compounds, often referred to as *lead hopping*. As a consequence, virtual screening methods applied for lead hopping should not be entirely based on molecular representations that over–emphasize structural similarity such as structural keys. The design of the recently developed PDR–FP fingerprint (Eckert and Bajorath [2006a]) follows this idea and is based on extensive and careful analysis of 2D molecular property descriptors and their relevance for bio–activity.

For this purpose, the DynaMAD scoring function (see *Equation 2.16*) was used to assess general activity class–specificity of 184 1D and 2D property descriptors implemented in MOE. This analysis resulted in the selection of a set of 93 descriptors (Eckert and Bajorath [2006a]). The generation of PDR–FP depends on the screening database statistics of these 93 descriptors. The descriptor value ranges are divided into non–overlapping intervals so that the same number of database compounds fall into each of them. Descriptor value ranges can be binary encoded by associating each interval with a single bit, which results in a fingerprint size of 500 bits in accordance with the binning scheme applied by Eckert and Bajorath [2006a]. Then, bit strings are generated by mapping the calculated descriptor values to the descriptor intervals. Bits associated with intervals into which a compounds descriptor values fall are set to 1, all other bits are set to 0. Hence, there are always exactly 93 bits set to 1, which makes PDR–FP size–independent. For the analysis of compound similarity in the presence of multiple reference compounds, an *activity–oriented search string* is created in a second step that reflects the descriptor value distributions of the template set relative to the distribution of the screening database. As the binning of descriptor value ranges into intervals directly depends on the value distributions of the screening database, the concentration of template compounds in only a very small number of intervals indicates activity–specificity. By summing up all bit frequencies from the search string that correspond to bits set on in the fingerprint of a database compound and normalizing by the sum of maximum bit frequencies for all descriptors, compound similarity between the screening database and multiple reference structures is expressed by the similarity coefficient given in *Equation 2.4*.

$$SV = \frac{\sum_{i=1}^{500} x_i y_i}{NF} \tag{2.4}$$

In this equation, NF is the normalization factor. The maximum value of 1 is achieved if all set bits in a fingerprint coincide with all bits that have the highest frequency among all template compounds. If the descriptor values of the database compound do not match any interval whose search string value is above zero, the compound is assigned the lowest PDR coefficient of 0.

2.3 Mapping Algorithms

The mapping algorithms described in the following sections are distinct from conventional ligand–based virtual screening methods and are especially designed for multiple template screens. Common to these approaches is the determination of activity–specific consensus positions in chemical space, whether represented as fingerprints or unmodified descriptor values. A consensus position is defined as a set of features that fulfill pre–selected activity–specific requirements. Compounds from the screening database are mapped to a consensus position and the result serves as basis for similarity analysis. In DMC and DynaMAD, this procedure is iteratively applied by assigning features to different layers that correspond to consensus positions of varying specificity. Distantly related to other approaches such as cell–based partitioning, mapping algorithms do not require the computation of pairwise distances between the multiple reference compounds and the screening database. Thus, computational costs are generally lower than for similarity search methods.

DMC, MAD, and DynaMAD create chemical reference spaces from independent descriptor contributions. By contrast, CA–DynaMAD refines the initial consensus position in a stepwise manner by adding only descriptors that maximize the separation from the reference database.

2.3.1 DMC

DMC (*D*ynamic *M*apping of *C*onsensus positions) is a mapping algorithm that seeks to identify consensus positions of active compounds in simplified descriptor spaces of stepwise increasing dimensionality. These chemical reference spaces are generated from 1D and 2D molecular descriptors that have been simplified in advance by binary transformation. Initially, the statistical medians of descriptor value distributions in the screening database are determined. Then, the position of each active compound in the chemical space created by all descriptors is binary transformed. If the descriptor value of a compound is larger than the database median, the corresponding bit is set to 1; if the compound's descriptor value is less than or equal to the median, it is set to 0. Once all active compounds are assigned a descriptor bit string, the descriptor scoring function given in *Equation 2.5* is applied using the mean $\overline{b_j}$ of the bit settings to indicate bit variability at position j inside the activity class.

$$score_{\mathrm{DMC}} = |0.5 - \overline{b_j}| \tag{2.5}$$

If all bit settings are identical, the descriptor values of the template compounds fall on the same side of the database median and top scores of 0.5 are achieved. This indicates a potentially class–specific feature. On the other hand, descriptors are assigned the minimal score of 0 if they show maximal variability, that is, half of the descriptor values of the templates are less than or equal to the median, and the other half is above. Based on the descriptor scores, consensus positions permitting stepwise increasing variability can be determined. The initial consensus position, permitting no variability, is defined by a descriptor vector composed of descriptors with a score of 0.5. By allowing increasing bit setting variability, consensus positions are generated that no longer require identical descriptor settings for all active compounds. According to their amount of tolerated variability, consensus positions are assigned to different layers. Layer 0 defines the initial consensus position and starts with the descriptors having a score of 0.5. Subsequently, the permitted variability is increased by a certain value, which is determined using *Equation 2.6*.

$$Stepwidth_{\mathrm{DMC}} = \frac{1}{\#actives} \tag{2.6}$$

For example, if the activity class consists of 10 compounds, all descriptors with a score of 0.4 are assigned to layer 1, those with a score of 0.3 to layer 2, and so on. Only descriptors with a score of 0 are not assigned to any layers, because they show no class–specificity. Thus, DMC implements dimension extension by allowing class–size dependent bit setting variability per extension step. *Figure 2.6* illustrates the generation of consensus positions in DMC and compares it to POT–DMC.

Finally, starting with the initial consensus position, the mapping process begins. During mapping it is examined if the bit string of a database compound coincides with all bit settings of the current consensus position. If it does, the compound is retained for the next mapping step. Because consensus positions that allow increasing bit variability include more descriptors, proceeding to the next step and mapping the remaining compounds to the next consensus position is termed *dimension extension*. The dimension extension process is continued until a specified number of database compounds remain.

2.3.2 POT–DMC

In order to emphasize the contribution of highly potent active compounds over less potent ones, a scaling function was devised and incorporated into DMC. Prior to descriptor scoring, the potency–scaled DMC algorithm (POT–DMC, Godden et al. [2004b]) calculates for every active template compound i the logarithmic scaling factor SF according to *Equation 2.7*, based on the comparison of its potency, pot_i, to the lowest one, pot_{min}.

$$SF_i = \log_{10}(pot_{min}) - \log_{10}(pot_i) + 1 \tag{2.7}$$

As the potencies of active compounds can span several orders of magnitude, the use of logarithmic scaling functions avoids dominance of the most potent templates by linear scaling. The lowest scaling factor possible is 1. Thus, this scaling function ensures that even less potent compounds are considered and the general structural information is taken into account. The scaling factors are applied when bit frequencies are determined, namely by summing and normalizing scaled bit values according to *Equation 2.8.*

$$\overline{b'_j} = \frac{\sum_{i=1}^{n} SF_i \times b_{ij}}{\sum_{i=1}^{n} SF_i} \tag{2.8}$$

SF_i and b_{ij} are the scaling factor and the value of bit j of active compound i, respectively. Subsequently, these potency–scaled bit frequencies are used to calculate consensus position as described. As shown, potency scaling effectively means that during calculation of bit frequencies compounds are counted multiple times based on their scaling factor, the definition of the stepwidth in POT–DMC has to be adapted, as shown in *Equation 2.9.*

$$Stepwidth_{\text{POT-DMC}} = \frac{1}{\sum_{i=1}^{n} SF_i} \tag{2.9}$$

Whereas the initial consensus position at layer 0 remains unaffected compared to DMC, the consensus positions of successive dimension extension are influenced by the weighted compound contributions.

Figure 2.6 shows the key concepts of DMC and POT–DMC and elucidates the differences. This example shows four template compounds $C_1 - C_4$ with different potency, which are represented by fingerprints consisting of eight binary encoded descriptors. Based on the potencies given as IC_{50} values, the scaling factors can be computed with the help of the logarithmic scaling function given in *Equation 2.7*; their range lies between 1 for compound C_3 with the lowest potency and 3.6 for compound C_2, which is 400–fold more potent. In the lower left corner, the resulting scores for each method are reported. When comparing DMC and POT–DMC scores of each bit position the effect of scaling becomes apparent. POT–DMC scores for bit positions two, four, and seven are significantly higher than DMC, while the score for bit position six is lower. These differences in scoring directly translate to the generation of consensus positions. For POT–DMC, bit position six is set to zero only in the last dimension extension (POT–DMC consensus position CP_4), while bits two and four are already set in consensus positions one to three, resulting in differences in the composition of chemical reference space.

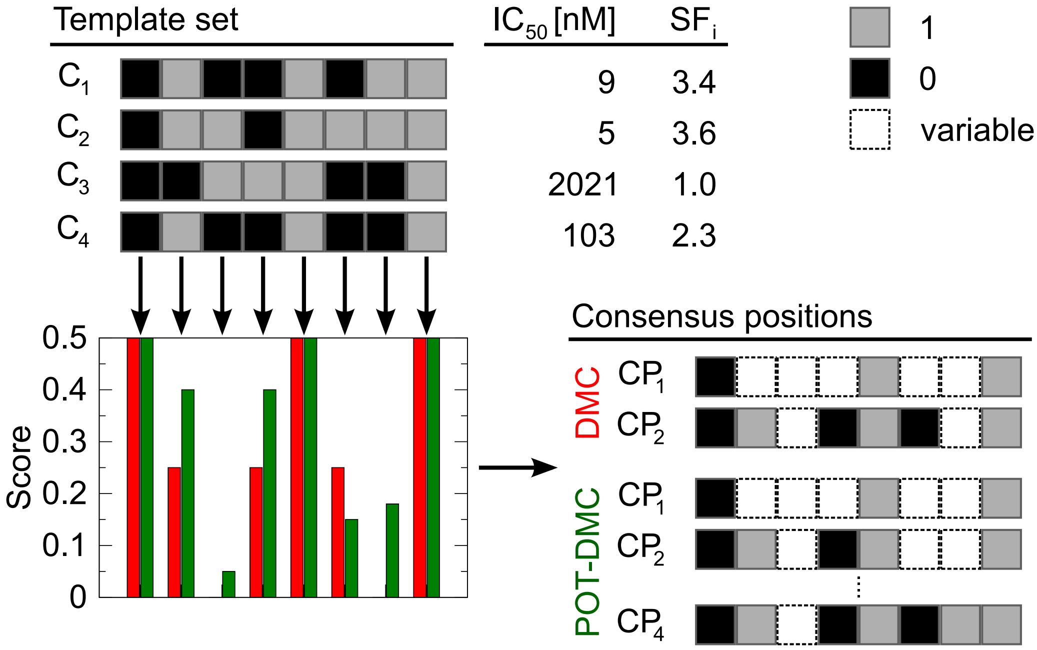


Figure 2.6: Comparison of DMC and POT–DMC. In the top left corner four examples of compound bit strings of length eight are shown. Bit frequency analysis reveals three high–, three medium–, and two non–specific bit settings, as is reflected by the DMC scores reported in red. Therefore, DMC generates two consensus positions as illustrated in the lower right corner. The use of the scaling function in POT–DMC increases the contributions of potent compounds and results in different descriptor scores (reported in green). Consequently, consensus positions of DMC and POT–DMC can differ.

Remarkably, bit seven, which is one of two descriptors that seemingly do not display activity–specificity, is assigned a significantly higher score by POT–DMC (as opposed to being assigned a score of 0 by DMC), because its value is identical for both highly potent compounds C_1 and C_2. Thus, the corresponding descriptor setting contributes to CP_4 of POT–DMC, while it remains variable in all DMC consensus positions.

For all DMC and POT–DMC calculations reported herein, 155 1D and 2D molecular property descriptors implemented in MOE were used and binary transformed with respect to the screening database, as described above.

2.3.3 MAD

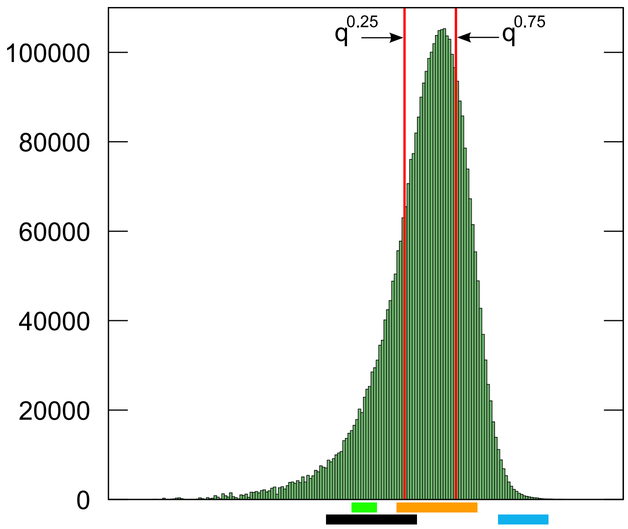


Figure 2.7: Overlap profiles for MAD descriptor scoring. This example shows the value distribution of a descriptor in a large compound database and how the value ranges of four different active classes, represented by the colored bars at the bottom, overlap with the three subranges of the database. The value ranges indicated by black and orange bars are examples for partial and complete overlap with the central range of the database, thus showing less activity specificity. The value ranges indicated by green and blue bars show higher specificity due to small range sizes and location in the upper or lower database subrange, where a maximum of 25% of the database can match these ranges.

Further pursuing the concept of generation of activity–specific chemical reference spaces, the MAD algorithm (*M*apping of *A*ctivity–specific *D*escriptor value ranges, Eckert and Bajorath [2006b]) focuses on the comparison of descriptor value distributions of activity classes and screening databases. Descriptors are scored depending on the degree of overlap of their activity class value ranges with the value distribution of the database used for screening.

In MAD, descriptor value distributions inside the screening database are represented by their 25%– and 75%–quantiles, $q^{0.25}$ and $q^{0.75}$, which devide the database value range into three subranges, whereas for activity classes the entire value ranges are used as defined by the minimum and maximum values, $classMin$ and $classMax$. The implemented scoring scheme distinguishes between three overlapping profiles defined by *Equation 2.10* and illustrated in *Figure 2.7*.

$$score_{MAD} = \begin{cases} \frac{q^{75}-q^{25}}{sizeRange} & \text{if } classMin \geq q^{25} \text{ OR } classMax \leq q^{75} \\ 2 \cdot \frac{q^{25}-dbMin}{sizeRange} & \text{if } classMax < q^{25} \\ 2 \cdot \frac{dbMax-q^{75}}{sizeRange} & \text{if } classMin > q^{75} \end{cases} \tag{2.10}$$

In case of *sizeRange* $= 0$, all active compounds have the same descriptor value, and while this is a likely sign of activity specificity, the corrections in *Equations 2.11 - 2.14* apply.

$$\delta^d_{db} = \frac{\sigma^d_{db}}{200} \tag{2.11}$$

$$classMin_{new} = classMin_{old} - \delta^d_{db} \tag{2.12}$$

$$classMax_{new} = classMax_{old} + \delta^d_{db} \tag{2.13}$$

$$sizeRange_{new} = 2 \cdot \delta^d_{db} = \frac{\sigma^d_{db}}{100} \tag{2.14}$$

In *Equations 2.11* to *2.14*, σ^d_{db} is the standard deviation of the descriptor inside the database. A score that is greater than 1 generally indicates that less than half of the database compounds fall within the value range of the active class. In particular, when the value range has no overlap with the central range, the score is greater than 2 and at most 25% of the database match (Eckert and Bajorath [2006b]). Hence it is possible to select descriptors that show activity class specificity to generate a tailored chemical reference space. Experiments in Eckert and Bajorath [2006b] have shown that a score threshold of 1 is already sufficient to produce meaningful results. On the basis of selected descriptors, the simple similarity measure shown in *Equation 2.15* captures how good a compound matches all value ranges, where D is the total number of selected descriptors and M the number of descriptors whose value ranges are matched by the given compound.

$$s_{MAD} = \frac{M}{D} \tag{2.15}$$

Thus, MAD similarity scores can adopt values between 0 and 1; larger values suggest higher similarity to the active compounds.

2.3.4 DynaMAD

As a further refinement of the concepts developed in DMC and MAD, DynaMAD (Eckert et al. [2006]) combines scoring of descriptor value ranges with dynamic extension of chemical reference spaces. Instead of using the value range size and the overlapping profile to estimate how activity–specific a given descriptor is, it directly measures the probability with which database compounds map to its value range. Similar to DMC, descriptor scores are then used to build chemical reference spaces of stepwise increasing dimensionality, until a desired termination criterion is met.

The underlying notion of descriptor specificity is that the fewer database compounds are likely to map the value range spanned by the minimum and maximum value ($classMin$ and $classMax$) of the activity class, the more specific this value range is. Therefore, the descriptor scoring function of DynaMAD was designed to capture mapping probabilities of database compounds as shown in *Equation 2.16*.

$$score_{\text{DynaMAD}} = (1 - P(classMin \leq X \leq classMax)) \cdot 100 \tag{2.16}$$

The mapping probability of discrete descriptor values is defined according to *Equation 2.17*.

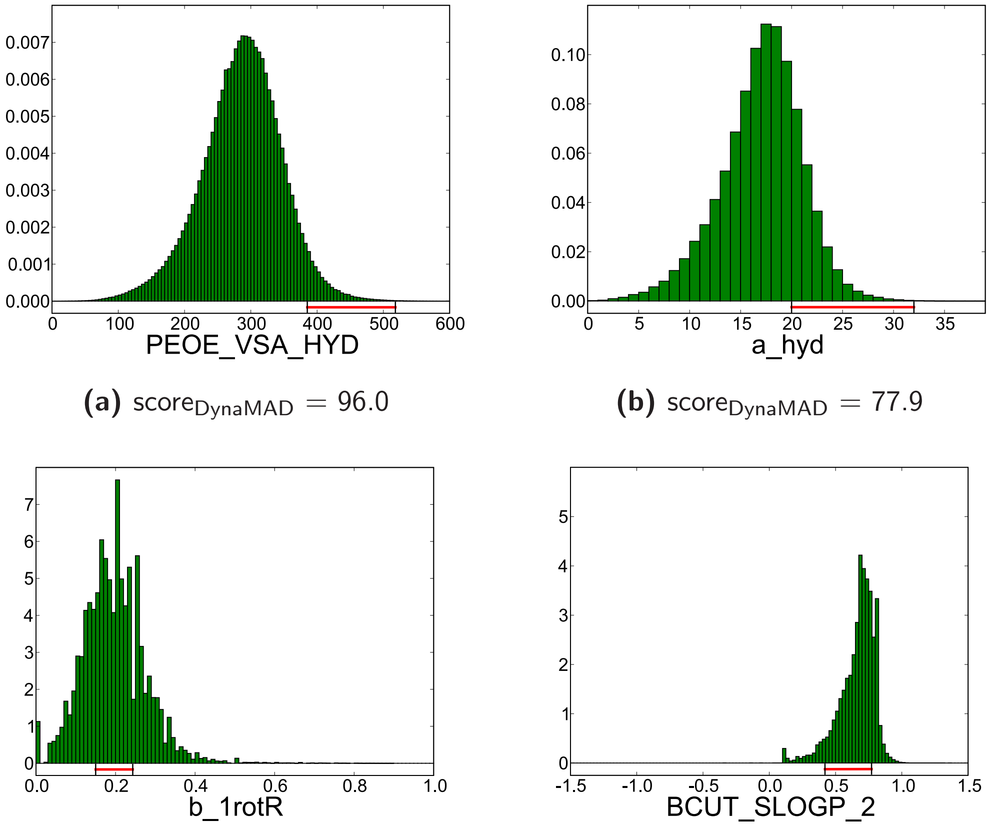
$$P(classMin \leq X \leq classMax) = \sum_{classMin \leq x_i \leq classMax} P(X = x_i) \tag{2.17}$$

Continuous descriptor values are combined into mini ranges $[x_{i1}, x_{i2}]$ and the calculation of mapping probabilities follows *Equation 2.18* (Eckert et al. [2006]).

$$P(classMin \leq X \leq classMax) = \sum_{classMin \leq x_{i2}; classMax \leq x_{i1}} P(X \in [x_{i1}, x_{i2}]) \tag{2.18}$$

Scores obtained by application of *Equation 2.16* range between 0 and 100 and give the percentage of database compounds that do not map the respective descriptor value range. Hence, high scores correspond to high descriptor specificity as illustrated in *Figure 2.8*.

Then, similar to DMC, descriptors are assigned to different mapping layers based on their degree of activity–class–specificity. Therefore, the entire scoring



(a) $\text{score}_{\text{DynaMAD}} = 96.0$

(b) $\text{score}_{\text{DynaMAD}} = 77.9$

(c) $\text{score}_{\text{DynaMAD}} = 50.8$

(d) $\text{score}_{\text{DynaMAD}} = 26.1$

Figure 2.8: Descriptor specificity and DynaMAD scores. Shown are four different value distributions of molecular descriptors of ∼3.7 million ZINC7 compounds and, indicated by red bars, the value ranges of a set of 37 dopamine D3 selective antagonists (Stumpfe et al. [2007]). From (a) to (d) descriptor specificity decreases as mirrored by the DynaMAD scores.

range is divided into equally sized intervals, and all descriptors whose score fall into the same scoring interval define one layer. In the original implementation, 20 scoring intervals of size five were chosen, designating the top–scoring range as layer 0, and representing descriptors with scores ≥ 95. As layer 0 contains the most specific descriptors, it is the most appropriate layer to start the mapping process with. Database compounds are mapped to all value ranges of descriptors contained in this layer, and qualify for the next mapping step upon successful matching to all of these ranges. Thus, dynamic dimension extension in DynaMAD is achieved through the consecutive addition of descriptor layers to which database compounds are mapped. This process is continued until a predefined termination criterion is met, for example, a specified number of compounds remain.

In summary, DynaMAD is able to generate and efficiently navigate high–dimensional chemical space representations due to its automated descriptor selection scheme. This scheme is based on the evaluation of independent descriptor value range specificity for a given class of active compounds and a screening database.

2.4 Design of CA–DynaMAD

Given the ability of DynaMAD to generate compound class–directed chemical space representations, the question arises whether it might be possible to systematically design target–selective chemical spaces. In the context of similarity searching, assessing compound selectivity is more complicated than searching for active compounds because it is required to distinguish chemically related compounds from each other that have differential activities against multiple members of a target family. Consequently, chemical reference spaces for such tasks must be designed in previously unexplored ways. However, for applications in chemical biology and medicinal chemistry, computational methods that are capable of analyzing and predicting ligand selectivity profiles within target families are highly attractive (Bajorath [2008]). By developing new schemes for scoring and dimension extension, a second generation DynaMAD–like mapping algorithm was created termed Continuous–Adaptive DynaMAD (CA–DynaMAD, Vogt et al. [2008]), that allows for better control of dimension extension and explicitly considers coherences among descriptors. To achieve this, the algorithm evaluates one descriptor per step and performs descriptor value range analysis on the basis of the continuously updated (size–reduced) background compound database.

Continuous Dimension Extension

In order to overcome stringency–related problems and ineffective descriptor selection as occasionally observed in DynaMAD, the dimension extension process is refined in CA–DynaMAD. Instead of assigning multiple descriptors to a single dimension extension step, descriptors are added one–by–one in order of decreasing scores, thereby continuously increasing dimensionality of the reference space. In case multiple descriptors have the same score, the descriptor with the lowest correlation to all previously selected descriptors is chosen. The correlation of descriptor i with n previously chosen ones is calculated as an average of the sum of absolute values of Pearson's product–momentum correlation coefficients r as stated by *Equation 2.19*.

$$corr_i = \frac{1}{n} \sum_{j=1}^{n} |r_{ij}| \qquad (2.19)$$

In *Equation 2.19*, pairwise descriptor correlation r_{ij} can be computed from the screening database or from another adequate set of compounds. For the current implementation of CA–DynaMAD the MDDR, version 2005.2, was analyzed.

Adaptive Descriptor Scoring

Descriptor scoring as introduced in DynaMAD requires the comparison of activity class value ranges and database distributions of these descriptors. In the case of DynaMAD, the descriptor distributions are calculated only once for the initial screening database, which typically contains a large number of compounds ($\sim$ 3.7 million in the case of used ZINC7 version, see *Appendix Chapter B.2* for further details). These distributions are then used to calculate fixed descriptor scores according to *Equation 2.16*. During the ensuing mapping process, the number of compounds is continuously reduced, thereby changing the descriptor value distributions, in particular during the first dimension extension steps when large numbers of compounds are discarded. Thus, descriptors proposed by DynaMAD might not be an optimal choice at any dimension extension level, except the first. Therefore, CA–DynaMAD utilizes an adaptive scoring scheme that repeats the descriptor scoring process after each dimension extension step with recalculated mapping probabilities for the size–reduced database, subsequently updating descriptor scores.

The example calculations presented in *Table 2.3* illustrate the effects of adaptive descriptor scoring used in CA–DynaMAD compared to the fixed scoring scheme of DynaMAD. As a test system, 37 selective D3 antagonists were used as templates and 10 000 ZINC7 compounds were randomly selected as a screening database. Then, five mapping steps were carried out, with fixed descriptor scoring like in DynaMAD or with adaptive scoring. As can be seen in *Table 2.3*, the first mapping step is identical for both scoring schemes. When comparing the remaining mapping steps, the differences become apparent. First of all, the mapping method based on adaptive scoring retains approximately ten times fewer decoys from the screening database after five dimension extensions. In addition, comparison of the selected descriptors reveals that the generated reference spaces are

Step	Descriptor	$score_{fixed}$	$score_{adapt}$	ZINC7$_{10k}$
1	GCUT_SLOGP_0	77.9	77.9	2 207
2	BCUT_SLOGP_0	76.5	6.4	2 065
3	GCUT_SMR_0	72.3	22.1	1 608
4	PEOE_VSA_FPNEG	70.4	49.4	813
5	BCUT_SMR_0	68.1	44.4	452

(a) Fixed scoring

Step	Descriptor	$score_{adapt}$	$score_{fixed}$	ZINC7$_{10k}$
1	GCUT_SLOGP_0	77.9	77.9	2 207
2	PEOE_VSA+0	60.1	58.0	881
3	GCUT_SLOGP_3	72.4	64.5	243
4	BCUT_PEOE_3	59.3	66.6	99
5	PEOE_VSA_FPNEG	54.5	70.4	45

(b) Adaptive scoring

Table 2.3: Fixed versus adaptive descriptor scoring. Given a set of 37 D3 antagonists and a random sample of 10 000 compounds from ZINC7, five mapping steps are reported for fixed and adaptive descriptor scoring, adding the best–scoring descriptor per step. $score_{fixed}$ and $score_{adapt}$ report the calculated fixed scores and the adapted scores, respectively. For comparison, scores according to both scoring schemes are reported for each descriptor.

significantly distinct from each other, only sharing two descriptors. Furthermore, analysis of the scores reveal that the descriptors with high constant scores lose specificity when mapping progresses and are less effective in deselecting decoys than descriptors chosen on the basis of adaptive scores.

2.5 Reference Methods

This section briefly presents methods that are used for reference calculations throughout the studies reported herein, including standard classification methods like clustering and partitioning as well as a method especially designed for molecular similarity analysis.

2.5.1 Cluster Analysis

In general, the classification of a set of objects into different subsets according to a measure of similarity is termed *clustering*. Cluster analysis is a widely used

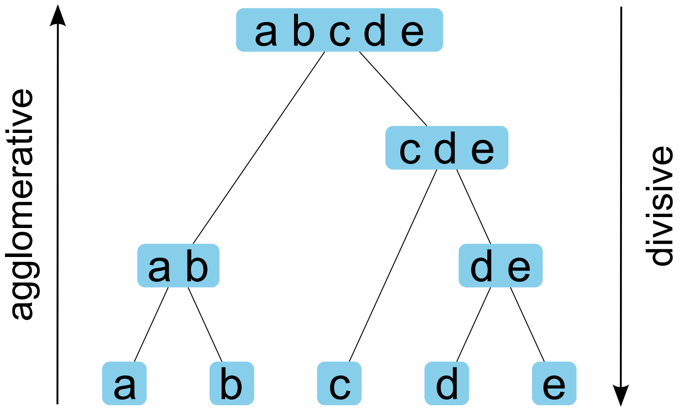


Figure 2.9: Hierarchical clustering. Divisive hierarchical clustering methods start from a single cluster, the root of the dendrogram, which is successively divided into smaller clusters. Agglomerative methods start with each object in a separate cluster, also called a singleton, and then successively merge clusters until all objects are placed into a single cluster.

technique and applied in a variety of different fields such as data mining, pattern recognition, bio– and chemoinformatics. Methods for data clustering can be divided into two main groups based on their general approach of cluster–generation. *Hierarchical* clustering algorithms generate a succession of clusters, where newly built clusters are derived from their predecessors and the established hierarchy can be represented using a *dendrogram*. As is shown in *Figure 2.9*, there exist two different types of hierarchical clustering approaches: *divisive* and *agglomerative*. Divisive hierarchical clustering starts with all objects in one cluster and then divides the clusters in a stepwise manner such that the distance between the newly created clusters is maximized, proceeding until all objects are partitioned into clusters of size 1. By contrast, agglomerative hierarchical clustering starts with all objects in separate clusters, subsequently combining at each step the pair of clusters with the shortest distance, until all objects are placed into a single cluster. Linkage methods describe how the distance d_{AB} between clusters a and b is determined. Three examples of common linkage methods are presented in *Table 2.4*. An alternative approach of hierarchical cluster analysis was proposed by Ward (Ward [1963]) which is not based on a distance measure between, but on statistical variance within the clusters.

Contrary to this, *non–hierarchical* or *partitional* clustering algorithms estab-

Single Linkage	$d_{AB} = \min_{i,j} d(x_i, y_j)$	shortest distance between members of clusters A and B, where $x_i \in A$ and $y_i \in B$
Complete Linkage	$d_{AB} = \max_{i,j} d(x_i, y_j)$	greatest distance between members of clusters A and B
Average Linkage	$d_{AB} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} d(x_i, y_j)$	average distance between all members of clusters A and B

Table 2.4: Linkage methods for clustering. Shown are three common methods to determine the degree of association between clusters.

lish a partitioning of the objects without hierarchical relationships among the clusters, and there exist several approaches on how clusters can be generated. As an example, k–means clustering is a *relocation method* that initially generates a pre–defined number of clusters and then iteratively re–assigns objects to these clusters in order to improve the quality of the clustering. Another approach of non–hierarchical clustering is based on the analysis of nearest neighbors, for which the algorithm introduced by Jarvis–Patrick is a widely known example.

The following two subsections introduce the methods by Ward and Jarvis–Patrick as examples of hierarchical and non–hierarchical clustering algorithms.

Ward's Clustering

Ward's clustering (Ward [1963]) is an agglomerative clustering algorithm that seeks to minimize the loss of information resulting from joining two clusters, where information loss is defined as an error–sum–of–squares criterion (ESS). At each step of the clustering procedure, all possible cluster unions are evaluated and those with minimum information loss are finally applied.

With a simple example of univariate data, *Figure 2.10* explains the concept of Ward's clustering. Given a set of values (2, 6, 5, 6, 2, 2, 0, 0, 0, 2) associated with ten objects, combining all ten objects in one cluster would result in an ESS of 50.5 (*Figure 2.10a*). On the other hand, clustering the objects according to their values into four clusters as depicted in *Figure 2.10b* would result in an ESS of 0. In other words, Ward's method forms clusters so that the cluster variance is minimized.

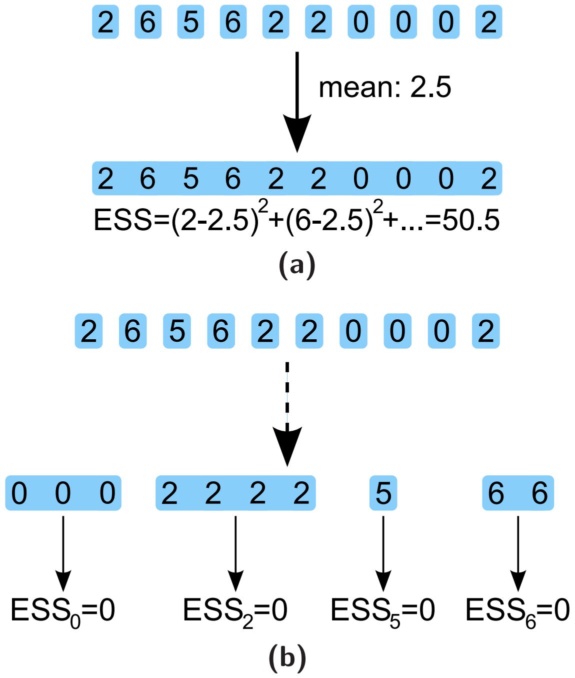


Figure 2.10: Ward's clustering. (a) Sub–optimal clustering. (b) Optimal clustering. The overall error–sum–of–squares is the sum of error–sum–of–squares of each individual cluster.

Jarvis–Patrick Clustering

Jarvis–Patrick clustering (Jarvis and Patrick [1973]) is a non–hierarchical clustering method that is based on the nearest–neighbor principle. Apart from a suitable distance measure, it requires two parameters J and K. For each object, the J nearest neighbors are determined. Two objects will then be placed into the same cluster if they appear in each other's nearest–neighbor lists and have at least K other nearest neighbors in common. *Figure 2.11* illustrates the concept of the Jarvis–Patrick clustering with an example of 12 objects that are partitioned into two clusters. It should be noted that by following this algorithm not all objects in the same cluster necessarily have at least K neighbors in common. If there are three objects l, m and n, and l and m have K neighbors in common like m and n, all three objects are placed into the same cluster even though l and n might have less than K common neighbors.

The Jarvis–Patrick clustering calculations are reported in *Chapter 3.3* and use the MACCS fingerprint as molecular representation and the pairwise Tc similarity as

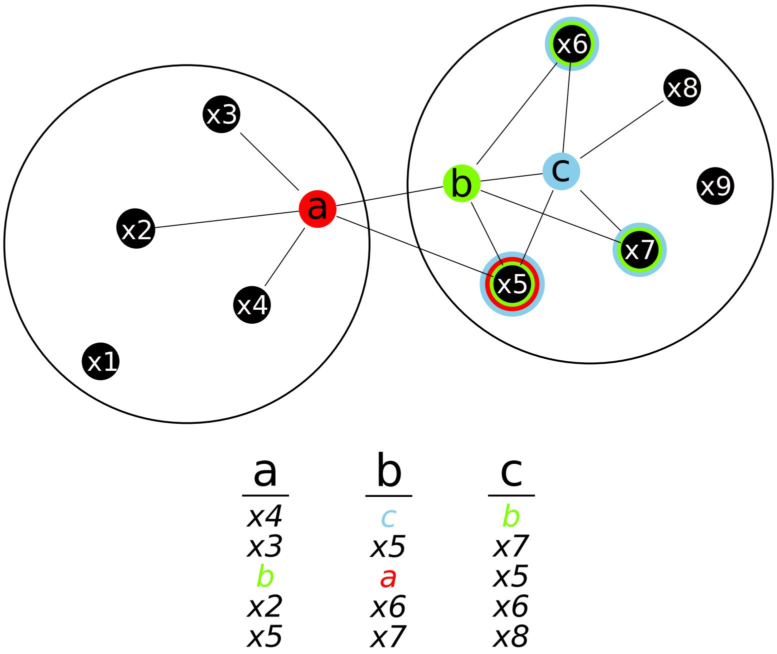


Figure 2.11: Jarvis–Patrick clustering. Based on the values of the input parameters, $J = 5$ and $K = 3$, two clusters are formed. Optimization of clustering results can be achieved by systematically varying J and K.

a measure of association.

2.5.2 Recursive Partitioning

The recursive partitioning (RP) technique belongs to the group of tree–based methods for statistical data analysis whose primary goal is to act as an expert system for classifications (Friedman [1977]; Rusinko et al. [1999]). However, it has also been used to identify important variables associated with object properties of interest (Zhang et al. [2001]). Although a number of different RP implementations exist, they share the basic concept of how a decision tree is created for classification.

Figure 2.12 explains the recursive partitioning process. Suppose there is a training set t_1 consisting of 38 observations of which 22 belong to class A and 16 to class B. These observations are represented by two properties x and y. Based on a measure of purity, a split of the training set according to any of the two properties is sought that results in an increase of purity in the terminal nodes, called leaves. The simplest purity measure is the absolute purity of terminal nodes with respect to classes A and B. Determination of class labels for the leaves as shown here is

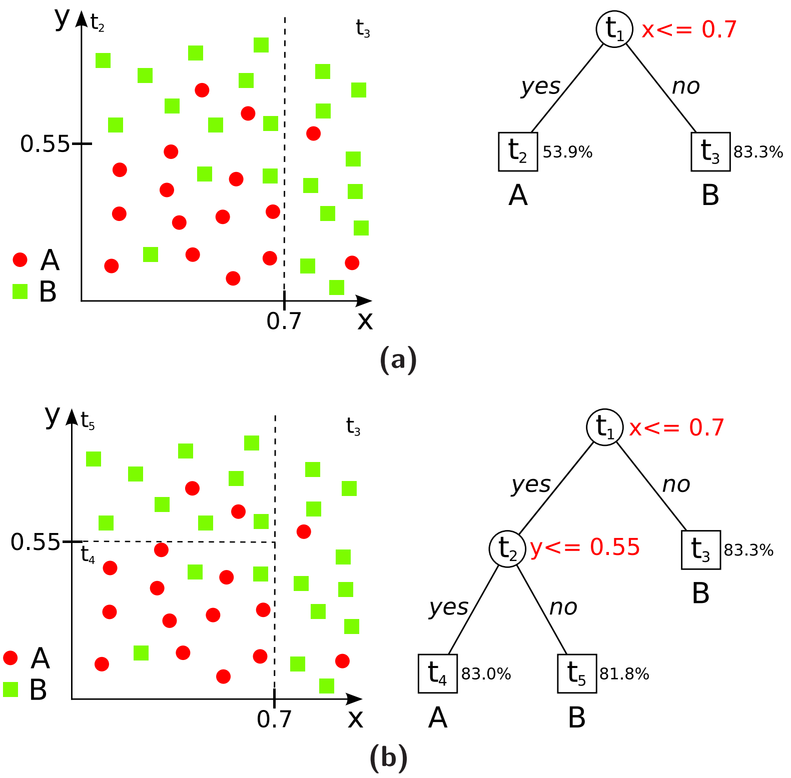


Figure 2.12: Recursive partitioning. Two classes of observations, *A* and *B*, are recursively split into subsets based on the two properties *x* and *y*. Each split leads to an overall increased purity of decision tree leaves, which correspond to the newly generated subsets.

a majority voting method where the most prominent class in a leaf determines the class assignment. As shown in *Figure 2.12a*, the best split initially is achieved when asking the question:

Is x *less or equal to 0.7?*

In a recursive manner, each of the two newly created subsets t_2 and t_3 is subsequently analyzed to determine, if further splits are possible that improve terminal node purity. In this case, only splitting subset t_2 at $y = 0.55$ is feasible, which leads to the generation of the subsets t_4 and t_5 with a purity of 83 and 81.8%, respectively. Essentially, this procedure can be carried out until each leaf contains only one observation, but in this case the decision tree would be overfitted and less suited for class prediction. Thus, a stopping criterion for the recursive process needs to be chosen, for example, a maximum number of observations in leaves or a purity threshold. In case of the example of *Figure 2.12*, a minimum leaf purity of 80% was applied.

2.5.3 DACCS

The DACCS method (*D*istance in *A*ctivity–*C*entered *C*hemical *S*pace, Godden and Bajorath [2006]) is designed to operate in high–dimensional but unmodified descriptor spaces and takes contributions of many arbitrarily chosen descriptors into account. As illustrated in *Figure 2.13*, DACCS centers chemical reference space through a scaling procedure on a subspace populated by a set of reference molecules. It then approximates this subspace as an orthogonal system by calculation of Euclidian–like distances from the center of the subspace to all compounds in a database. Thus, DACCS calculations generate a distance–based ranking of database compounds from the centroid of an active subspace. Given a set of known active reference compounds, the distance in scaled chemical space from the center of the "active subspace" is calculated for database compounds as in *Equation 2.20*.

$$d_{DACCS} = \sqrt{\sum_{i}^{d} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2} \tag{2.20}$$

In *Equation 2.20*, x_i is the value of one of the d descriptors for a compound x, μ_i the mean value of descriptor i for a set of active template compounds, and σ_i is the standard deviation of the descriptor values for the templates. If this standard deviation is zero, then the standard deviation of the entire compound population is used instead. If both values are zero, the descriptor is omitted from the calculation (which would apply to a descriptor having the same value for all active and database compounds).

In POT–DACCS, the logarithmic potency scaling function initially devised for POT–DMC (see *Equation 2.7*) becomes effective when the means and standard deviations are calculated for centering the active subspace. Under scaling conditions, descriptor values of potent compounds have a greater effect on the means and standard deviations than weakly potent compounds. Thus, the descriptor statistics change through the weight put on the values of more potent compounds. Essentially, this causes these values to be counted several times more than the ones of weakly potent molecules. As a consequence, the active subspace is shifted towards the positions of potent compounds.

1) Add active compounds to database

2) Add active compounds to database

3) Compute the distance from the centroid for all compounds

4) Use the distance as a measure of similarity to generate a selection set

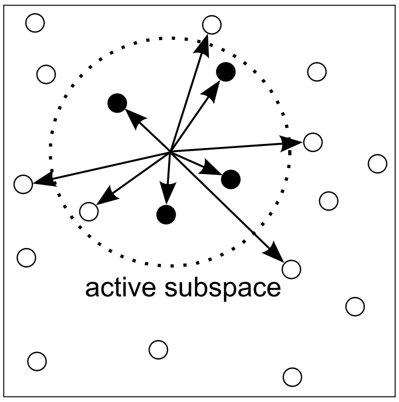


Figure 2.13: The DACCS algorithm. Black dots represent a subset of reference molecules used to calculate the centroid of an active subspace (dotted circle) and open dots database compounds located close to or within the subspace. Arrows indicate Euclidian–like distances of compounds from the centroid.

Chapter 3

Structure–Selectivity Relationships

In this chapter, four studies are presented that investigate the suitability of 2D ligand–based virtual screening methods to analyze structure–selectivity relationships. In particular, multiple 2D molecular fingerprints and a new class of 2D LBVS methods termed mapping algorithms are applied to study their ability to distinguish between active compounds with differential target selectivity profiles. Initially, it is shown that similarity search and classification methods are able to capture potency–relevant molecular features and can be directed to preferentially detect potent compounds based on the analysis of biological screening data. Then, extensive benchmark calculations with five different fingerprint designs, two mappings algorithms, and several reference methods are conducted, aiming at the identification of compounds with desired selectivity. Again, biological screening data are included in order to validate the results under more realistic VS conditions. The final study establishes that the most advanced mapping algorithm can be used to generate chemical reference spaces useful for the generation of predictive models for target selectivity.

3.1 Analysis of HTS Data

In this section, the analysis of biological screening data obtained from a cathepsin B HTS experiment is presented. This analysis was conducted with the aid of two conceptionally distinct approaches, which can be extended to incorporate potency information of template compounds by inclusion of a scaling function. The results suggest that LBVS methods are suitable for potency–directed compound analysis drawing upon knowledge gained from HTS data.

Experimental Setup

The HTS data set was generated at the Penn Center for Molecular Discovery at the University of Pennsylvania[1] using an assay for inhibitors of cathepsin B, a cysteine protease in lysosomes, and made public through PubChem. It consists of 63 332 compounds including 40 hits with IC_{50} values ranging from 46 nM to 46 μM. Examples of hits are shown in *Figure 3.1.* The structural resemblance of active compounds was evaluated on the basis of average pairwise Tanimoto coefficients calculated for MACCS fingerprints. The similarity of active and inactive molecules was assessed using single template search calculations with MACCS keys fingerprint for each known active against the remainder of the database.

For VS trials, 30 sets of ten active compounds each were randomly selected as reference compounds while the remaining 30 active compounds were added to the set of inactives. The only preselection requirement was that the reference compounds had to cover the experimentally observed potency range in order to ensure realistic potency scaling. Each set of reference compounds was individually used as input for scaled and nonscaled DACCS and DMC calculations. For search calculations, a set of 155 1D and 2D molecular descriptors was calculated with MOE. For DMC and POT–DMC calculations, descriptor medians were calculated for the HTS data set.

Characterization of HTS Data

Initially, the distribution of hits in the HTS set and the similarity of active and inactive compounds was analyzed. Three compounds were active below 100 nM and five additional ones below 1 μM; 17 hits had potency in the range between 1 and 10 μM, and the remaining 15 hits were active above 10 μM concentration. Thus, the potency distribution of active molecules was continuous but relatively narrow, with most compounds being active in the low micromolar range. The mean and median potencies were 13.2 μM and 6.9 μM, respectively. Highly potent (that is, low nanomolar) compounds are rarely found in primary screening data sets. However, despite moderate average potency, the cathepsin B set contained a total of eight sub–micromolar hits including three highly active compounds. *Figure 3.2*

[1] The Penn Center for Molecular Discovery (PCMD).
`http://www.seas.upenn.edu/~pcmd`

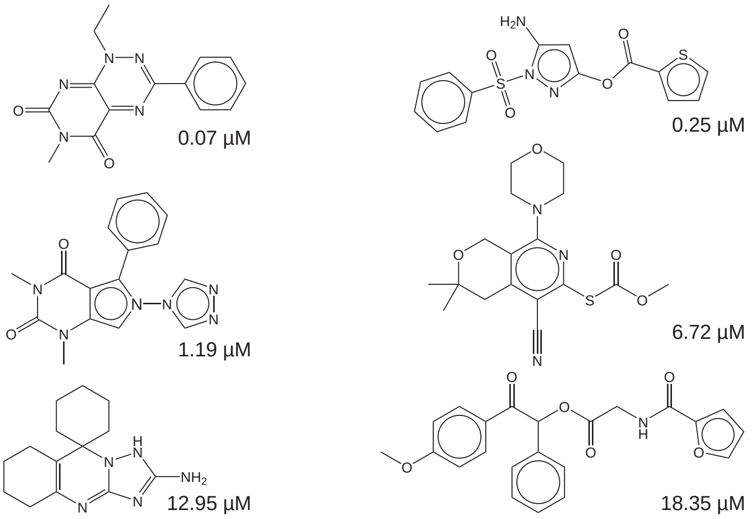


Figure 3.1: Representative cathepsin B HTS hits. Most of the cathepsin B inhibitors have a unique core structure and are of limited structural complexity.

reports the results of pairwise MACCS Tc comparisons of potent compounds. In the Tc matrix, off–diagonal red patterns indicate significant similarity between hits. The figure reveals that similarity between hits was limited to relatively few (of 780 possible) compound pairs. Many comparisons yielded Tc values of 0.5 or less, which indicates dissimilar compounds in terms of MACCS structural keys. Similarity was greatest among seven potent (top–left red cluster) and seven weakly active (lower right) compounds. The two most active compounds were similar to each other, but also to a significantly less potent molecule, and compounds in the second top–left cluster were also found to display distinct similarity to five weakly active ones in total. Therefore, compound potency within this set did not clearly correlate with compound similarity, as can be observed from the representative structures shown in *Figure 3.1.* Active compounds were not confined to one or two analog series, which would have also been reflected in a higher intraset similarity.

Table 3.1 reports the results of MACCS similarity search calculations within the screening set using each hit individually as a query. A considerable number of inactive data set compounds were found to be structurally similar or very similar to hits, including nearly identical compounds. At a MACCS Tc level of 0.90 or greater, nearly 200 matches of hits to inactive compounds were detected. At a MACCS Tc threshold value of ∼0.85, the structural resemblance of compounds was clearly visible, and nearly 700 matches were detected at this level. These

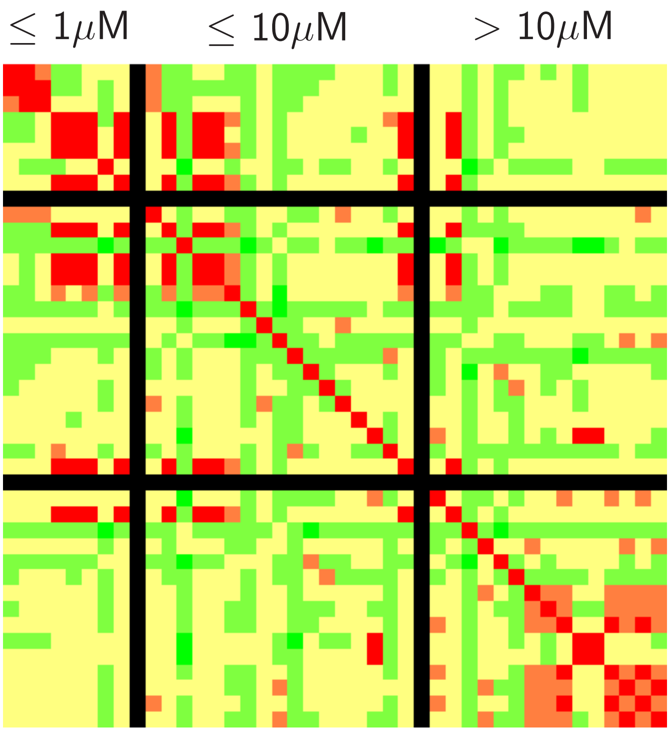


Figure 3.2: Similarity analysis of HTS hits. Structural analysis of cathepsin B HTS hits using a matrix of pairwise MACCS Tc similarity values, ordered by increasing IC_{50} values. Tc values are reported in matrix–format for systematic pairwise compound comparisons. The values are color–coded according to following scheme: dark green 0-0.2, light green 0.2-0.4, yellow 0.4-0.6, orange 0.6-0.8, red 0.8-1. Among all potency subranges as indicated, significant compound diversity can be observed, while at the same time a subset of similar compounds spans the entire potency range.

findings nicely illustrate that Tc values alone are not a reliable indicator of similar activity (Martin et al. [2002]), and that very similar structures can either be active or inactive (Kubinyi [1998]). The latter point is well–known in medicinal chemistry and analog design and presents an intrinsic limitation of methods that evaluate similarity from a whole–molecule perspective (Bajorath [2002]; Bender and Glen [2004]). In conclusion, significant structural diversity among hits is observed on one hand, while on the other hand these active compounds are structurally similar to a considerable number of inactive compounds.

IC_{50} [nM]	1.00	(1.00, 0.95]	(0.95, 0.90]	(0.90, 0.85]	(0.85, 0.80]
46.1	0	0	0	1	7
71.0	0	0	0	0	11
72.2	0	0	0	1	14
246.6	0	5	5	14	43
434.6	0	2	7	8	16
691.9	0	5	8	19	105
844.5	0	0	0	0	3
923.5	0	5	7	16	92
1 185.1	1	3	22	47	64
1 260.8	0	2	6	8	14
1 639.5	0	0	0	0	0
1 749.1	0	5	7	12	75
1 990.0	0	6	4	17	71
2 087.4	1	0	5	0	1
2 120.0	0	0	0	0	0
2 247.2	1	0	6	53	217
3 170.9	0	0	6	16	62
4 169.7	0	0	0	1	1
6 355.6	0	0	1	3	8
6 714.5	0	1	10	16	29
7 113.8	0	0	0	2	16
8 563.1	0	0	3	19	102
8 926.9	0	0	1	17	38
9 388.7	0	0	0	1	13
9 562.6	2	5	10	37	132
11 457.8	0	0	0	1	4
12 265.1	2	6	12	56	144
12 947.1	0	0	1	1	0
14 196.4	0	0	0	3	23
18 349.1	0	1	3	12	34
19 686.4	0	0	0	2	3
28 132.2	0	0	2	5	48
33 855.4	2	1	1	1	5
37 190.0	0	0	0	2	20
38 470.6	1	0	0	1	4
39 986.8	0	0	1	1	19
44 181.1	0	0	1	2	27
44 577.0	0	1	3	56	237
44 782.6	0	0	0	3	29
46 161.2	0	0	3	44	160
Sum	**10**	**48**	**135**	**498**	**1 891**

Table 3.1: Single template search calculations with MACCS fingerprint. Reported are the number of inactive compounds within a Tc similarity range between 0.8 and 1. Hits are sorted by decreasing potency (IC_{50}).

DMC and DACCS Calculations

In order to determine whether active molecules could be retrieved from the HTS set irrespective of their potency levels, individual nonscaled search calculations using different sets of reference compounds were analyzed. Results obtained for DMC and DACCS are reported in *Tables 3.2* and *3.3*, respectively. For DMC, active compounds were found at the final dimension extension levels. In a few cases, the selection sets contained a large number of inactive compounds together with two or three hits, for example, in trials 13 and 21 in *Table 3.2*. For practical VS applications, such calculations would not be suitable because too many database compounds would need to be evaluated. However, most calculations produced reasonably sized selection sets, and in nine trials, one or two inhibitors were recovered together with only approximately ten or even fewer inactive compounds. The majority of DACCS calculations reported in *Table 3.3* also detected inhibitors. Five trials failed to produce active molecules among the top-scoring 100 compounds, but 14 calculations revealed between one and six hits among the top 50 compounds with, on average, 3.5 inhibitors per trial. The remaining calculations produced between one and six active molecules among the top 100 compounds. In general, the results of DMC and DACCS calculations were much influenced by the composition of the reference molecule sets, which has also been observed in VS benchmark calculations (Bajorath [2002]).It has to be considered that, different from benchmarking, hit and recovery rates are not a primary measure of success for VS applications because compound recall cannot be determined in practical screens. Furthermore, the ability to identify novel active molecules in small VS selection sets is much more relevant than actual hit rates, which represent a more suitable measure for large–scale screening campaigns. Since both DMC and DACCS calculations consistently retrieved active molecules from the HTS data sets, it was possible to explore the key question of whether or not increasingly potent inhibitors would be detected under the conditions of potency scaling.

Comparison of DMC and POT–DMC

The comparison of DMC and POT–DMC search calculations revealed that both methods detected an overall comparable number of hits as reported in *Table 3.2*. However, POT–DMC recognized fewer inactive molecules in 17 of 30 calculations, which considerably reduced the size of the selection sets. In a number of cases, the reduction in the number of inactive molecules was quite dramatic, for example,

Trial	DMC	POT–DMC
1	5/435	1/221
2	2/196	3/39
		2/12
3	2/4	3/24
	1/2	2/9
4	1/41	1/34
5	2/97	2/36
	1/15	1/3
6	2/24	4/9
		2/4
7	1/11	1/2
8	2/663	2/92
		1/19
9	1/13	2/36
		1/4
10	8/749	2/285
11	1/223	1/160
12	2/269	1/70
13	2/5 320	2/1 669
14	1/93	1/7
15	1/23	2/29
	1/4	1/4
16	1/466	1/9
17	3/189	3/82
	2/4	2/9
18	2/98	2/493
19	1/7	2/37
		1/2
20	1/49	2/14
21	3/780	2/152
22	2/64	3/8
23	1/62	1/3
24	3/36	2/4
	2/4	2/2
25	1/170	2/5
26	1/20	3/55
		1/2
27	2/80	2/60
		1/4
28	1/61	2/36
	1/8	1/2
29	1/60	3/23
		1/12
30	2/49	4/47
		2/12

Table 3.2: Hits in DMC and POT–DMC compound selection sets. Results of 30 individual search calculations with DMC and POT–DMC are summarized. For each calculation, hits identified at the final or second to last dimension extension level are reported. For example, "2/24" means that the compound selection set contained two cathepsin B inhibitors and 22 inactive compounds.

in trials 8, 21, or 25. Thus, potency scaling increased the specificity of the calculations. Next, the potency distribution of the identified inhibitors was analyzed. POT–DMC detected on average 1.3 hits with potency <1 μM in selection sets of fewer than 100 compounds, whereas DMC detected on average 0.6 of these hits. The POT–DMC retrieval rate was considerable because only four potential hits with <1 μM potency were available in the screening set. POT–DMC also detected approximately twice as many hits as DMC with a potency of up to 10 μM but did not detect hits with >10 μM potency. Thus, there was a consistent trend of POT–DMC calculations to enrich selection sets with more potent compounds than DMC.

	DACCS		**POT–DACCS**	
Trial	**S50**	**S100**	**S50**	**S100**
1	0	1	3	4
2	0	0	0	0
3	4	5	5	5
4	0	0	2	3
5	5	6	5	5
6	1	1	1	4
7	0	0	0	0
8	6	6	5	5
9	0	0	1	1
10	2	4	3	4
11	4	4	6	6
12	0	1	0	0
13	5	6	6	6
14	0	2	1	1
15	0	0	1	2
16	0	0	1	1
17	2	3	3	3
18	3	4	4	5
19	4	5	5	5
20	1	1	0	0
21	0	1	1	2
22	0	1	1	1
23	6	6	5	5
24	0	0	1	1
25	0	0	0	0
26	1	1	3	3
27	5	5	5	5
28	0	2	1	2
29	0	1	0	0
30	0	0	0	0

Table 3.3: Hits in DACCS and POT–DACCS compound selection sets. For individual DACCS and POT–DACCS calculations, the number of hits contained in selection sets of 50 (*S50*) or 100 (*S100*) database compounds are reported.

Underlying effects could be studied by analyzing the mapping pathways of hits with different potencies to consensus bit strings during dimension extension, as reported in *Figure 3.3*. For DMC, there were only little differences in the mapping characteristics of active compounds, although some potent compounds reached higher dimension extension levels than the weakly potent molecules. By contrast, for POT–DMC, significant changes were observed. Here, weak hits were eliminated earlier during the search than in DMC calculations, but the most potent hits were found in much higher dimension levels than in nonscaled calculations, which resulted in a distinct separation of more and less potent compounds during POT–DMC calculations. In both cases, search strings consisting of a small number of consensus bits produced comparable results. The reason for this is that at the beginning of the calculations, only low bit variability is permitted, which results in a suppression of the scaling effect. The observation that potent hits were retained under increasingly stringent search conditions during POT–DMC calculations can be explained when considering the fact that POT–DMC generates consensus bit settings that selectively favor highly potent molecules.

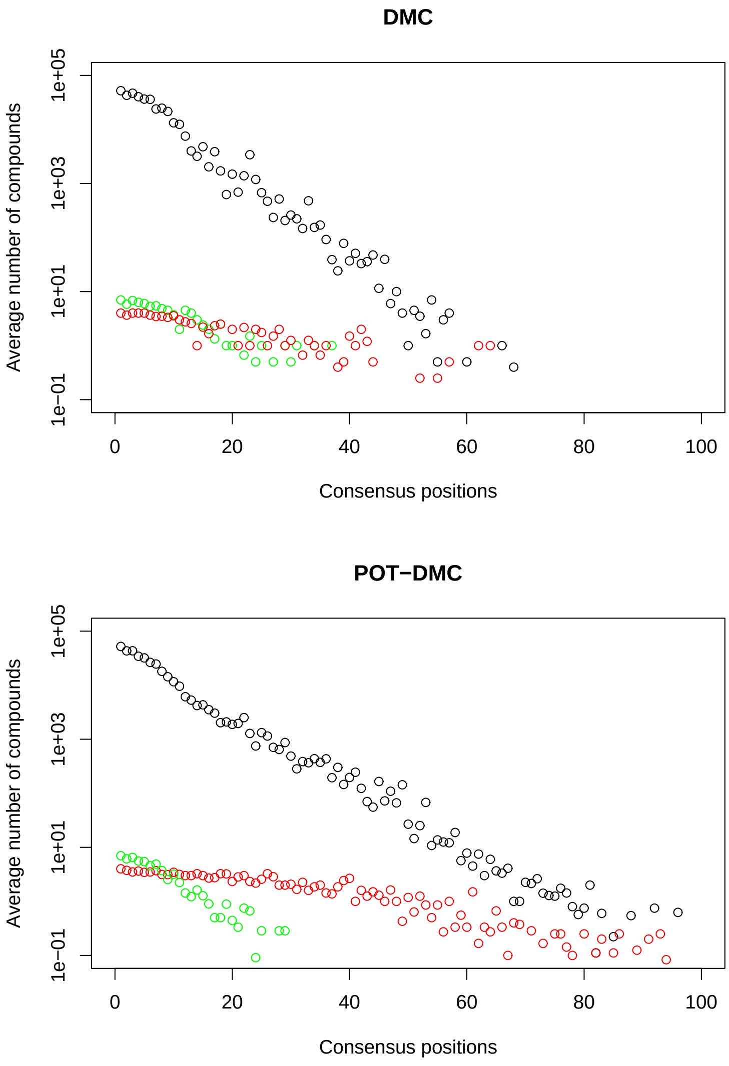


Figure 3.3: Potency distribution of hits at increasing dimension extension levels. Shown is a comparison of compounds matching search strings of increasing length in DMC and POT–DMC calculations. *Consensus positions* reports the number of accepted descriptor bits. During dimension extension, the number of descriptor consensus bits increases. Black circles represent database compounds, green circles hits with >10 μM potency, and red circles hits with <1 μM potency.

DACCS versus POT–DACCS

Next, the effects of POT–DACCS were investigated. This method is algorithmically distinct from POT–DMC since it produces a distance–based ranking of database compounds. As shown in *Table 3.3*, POT–DACCS calculations identified more hits than DACCS in approximately half of the trials. In all of the trials that produced hits, POT–DACCS achieved higher average potency for retrieved compounds than DACCS. For example, among the top scoring 100 data set compounds, POT–DACCS recognized on average one of the four most potent hits available in the database, whereas DACCS recognized on average only ∼0.3. Similar to POT–DMC, POT–DACCS displayed a tendency to deselect weakly active hits relative to DACCS and generally did not rank them among the top scoring 100 compounds.

In contrast to POT–DMC, compound ranking also needs to be taken into account when evaluating POT–DACCS. On the basis of systematic rank versus potency comparisons, two effects were found to be responsible for the enrichment. POT–DACCS ranked potent hits more highly than DACCS and assigned lower ranks to weakly potent hits, which resulted in an enrichment of most potent hits at higher rank positions, as shown in *Table 3.4*. Overall, the average rank position of the most potent hits within the top scoring 100 compounds was 25 for DACCS and 19 for POT–DACCS, and for hits with potency between 1 μM and 10 μM, the average rank was 23 for DACCS and 20 for POT–DACCS. For hits with potency >10 μM, an average rank of 34 was determined for DACCS, whereas POT–DACCS did not select weakly potent hits among the top 100 compounds. These findings confirmed that POT–DACCS assigned lower ranks to weakly active hits than DACCS and higher ranks to increasingly potent hits. Given the findings discussed above, it was attempted to illustrate the effects of potency scaling in an intuitive manner. Since DACCS operates in unmodified descriptor spaces, a

Potency [μM]	DACCS	POT–DACCS
< 1	25.4	19.4
$1 - 10$	23.2	19.8
> 10	33.7	–

Table 3.4: Comparison of DACCS and POT–DACCS. Reported are the average ranks of active compounds among the top 100 compounds.

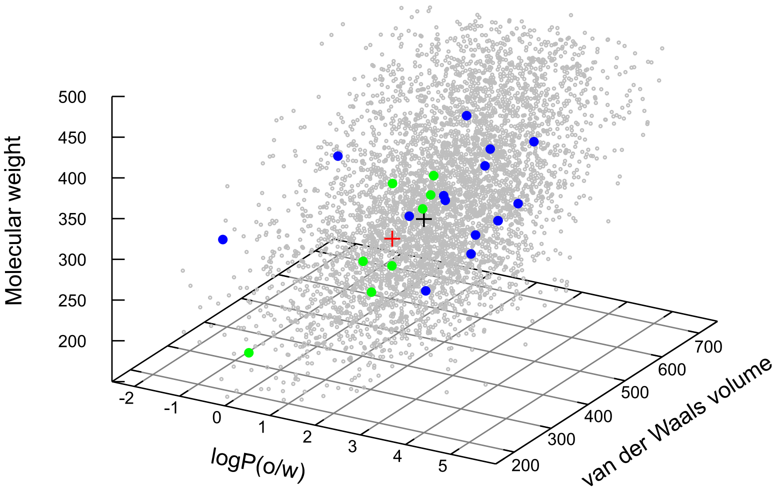


Figure 3.4: Compound distributions and centroids. A sample of the cathepsin B HTS data set was projected into a three–dimensional descriptor space. Hits with potency <1 μM are represented as green dots, hits with potency >10 μM as blue dots, and inactive compounds as small gray dots. Average centroid positions for all DACCS (black cross) and POT–DACCS calculations (red) are mapped.

systematic search was carried out for combinations of three property descriptors that would produce a notable separation of more and less potent hits in a three–dimensional representation. An example is shown in *Figure 3.4*. Here, mapping a random sample of 5 000 inactive compounds from the HTS data set into a space constituted by three intuitive descriptors (logP(o/w), molecular weight, and approximated van der Waals volume) produced a visible separation of more and less potent inhibitors. Of course, this descriptor combination would not be sufficient for VS because it did not separate active from inactive compounds. However, despite its limited resolution, this reference space made it possible to visualize the effects of potency scaling on the location of the centroid position. *Figure 3.4* reveals that potency scaling shifted the DACCS centroid position away from weakly potent hits into a region predominantly occupied by molecules having higher potency, which rationalizes why POT–DACCS calculations detected more potent inhibitors. Therefore, at least in part the effects of potency scaling can be attributed to re–centering of the activity–dependent subspaces. It is reasonable to assume that similar changes in centroid positions also occur in high–dimensional space representations when potency scaling is applied.

3.2 Similarity Searching for Selective Ligands

After demonstrating that 2D VS methods are sensitive to molecular features that relate to compound potency, the next step was to investigate whether such methods could also be used to distinguish between active compounds with different target selectivity. Given an especially designed database of compounds with selective activity against members of different protein families, similarity search calculations were conducted with five different 2D molecular fingerprints. The analysis of the results presented in this section shows that 2D molecular fingerprints are suited to differentiate between compounds with different target selectivity. Moreover, the fingerprints exhibited comparable performance, irrespective of fingerprint design and complexity.

Experimental Setup

The molecular selectivity benchmark system (Stumpfe et al. [2007]) consists of a total of 558 compounds selective against 13 individual targets belonging to three different families: biogenic amine G protein–coupled receptors (GPCRs), papain–like thiol proteases, and chymotrypsin–like serine proteases. The compounds are divided into 26 selectivity sets containing varying numbers of compounds, as summarized in *Table 3.5*. Compounds in each of the selectivity sets are selective for one target over another closely related one. Therefore, each set represents an individual test case. Importantly, the selectivity sets were assembled to cover different scaffold–selectivity relationships: they contain compounds having similar scaffolds but different selectivity and also compounds having diverse scaffolds yet similar selectivity. Compounds within a selectivity set often originate from different sources or independent investigations, which further contributes to intra–set diversity of compounds and scaffolds, as can be observed from the MACCS Tc similarity matrices shown in *Appendix Figures B.1, B.5* and *B.6*. Further details are given in *Appendix B*.

Dependent on the sources of selective compounds, selectivity sets were added to one of two different background databases for similarity searching. Compound data for the cathepsins B, L, and S (in part) were taken from high–throughput screening experiments and, therefore, a background database of confirmed inactive compounds was assembled from high–throughput screens for inhibitors of cathepsin B, L, and S, respectively. It consists of 55 055 compounds showing no measurable activity against any of these targets (see *Appendix Chapter B.2*). Ligands of GPCRs and inhibitors of serine proteases were generally taken from

Target family	Target	Selectivity set	Compounds
GPCRs	D1	D1/D2	31
		D1/D4	13
	D2	D2/D1	26
		D2/D4	9
		D2/5HT1a	11
	D3	D3/D4	12
	D4	D4/D1	20
		D4/D2	64
		D4/D3	33
	5HT1a	5HT1a/D2	24
		5HT1a/Alpha1	46
	Alpha1	Alpha1/5HT1a	27
Papain–like proteases	Cat B	Cat B/Cat L	23
		Cat B/Cat S	23
	Cat L	Cat L/Cat B	33
		Cat L/Cat S	33
	Cat S	Cat S/Cat B	23
		Cat S/Cat L	23
		Cat S/Cat K	20
	Cat K	Cat K/Cat S	25
Chymotrypsin–like proteases	Thrombin	Thrombin/Trypsin	35
		Thrombin/Factor Xa	52
	Trypsin	Trypsin/Thrombin	20
		Trypsin/Factor Xa	25
	Factor Xa	Factor Xa/Thrombin	49
		Factor Xa/Trypsin	48

Table 3.5: Benchmark system for pairwise selectivity. Compounds in each set are selective for one target over another family member. For example, the designation D1/D2 means that the compounds in this set are selective for dopaminergic receptor subtype D1 over D2.

the scientific literature. These compounds were added to a subset of the ZINC6 database (Irwin and Shoichet [2005]), consisting of ∼1.44 million compounds considered as decoys. Five 2D fingerprints were chosen for this study that represent different designs and, in addition, have significantly different complexity. *Table 3.6* gives a short overview of selected fingerprints that are described in greater detail in *Chapter 2*.

Fingerprint	Design
MACCS	166 structural fragment keys
MOLPRINT 2D	atom environments, no bit representation
MP–MFP	174 bits, hybrid of selected binary transformed property descriptors and MACCS keys
PDR–FP	500 bits, 93 value–range encoded property descriptors
TGT	1 704 unique 3–point 2D pharmacophore features

Table 3.6: **Selected 2D molecular fingerprints.**

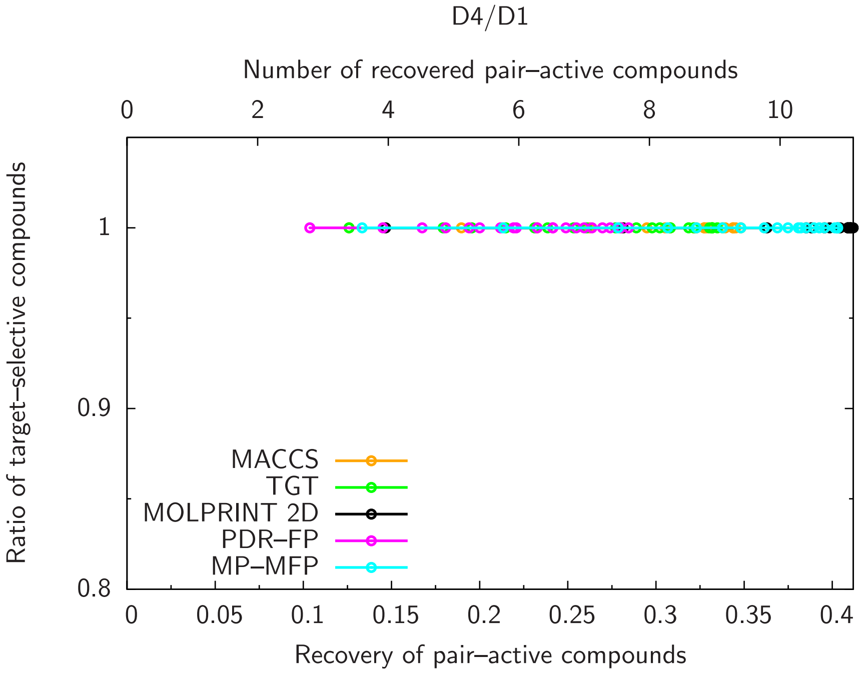
Test Calculations

To evaluate different fingerprint designs for their ability to detect selective compounds, multiple reference compounds were used and a nearest neighbor approach for similarity scoring was applied (as described in *Chapter 2.2*). This type of similarity search protocol has been shown to frequently perform best in comparisons with alternative search strategies (Hert et al. [2004]; Willett [2006]). For each selectivity set and fingerprint, 25 independent search trials were carried out with randomly selected sets of reference molecules and the results were averaged. For selectivity sets containing more than 15 compounds, ten reference molecules were chosen. In the remaining sets, half of the compounds in each set were used as reference molecules and the other compounds were added to the background database as potential hits. For reference sets of ten molecules, five nearest neighbors (5NN) were considered when determining the similarity score for each database compound; for reference sets of fewer than ten molecules, three nearest neighbors (3NN) were chosen. For MACCS, TGT, MOLPRINT 2D, and MP–MFP, the Tanimoto coefficient was chosen as similarity measure. For PDR–FP, similarity is expressed by means of a specifically developed dot product metric (see *Chapter 2.2.5*).

Pairs of selectivity sets provided the basis for analyzing the potential of similarity searching to detect selective compounds. For each target pair (A and B), a two–step analysis was carried out. First, subsets of compounds selective for A over B (set A/B) were taken as reference molecules and searched against a background database containing the remaining selective A/B compounds and also compounds belonging to set B/A. Then, the calculations were repeated using subsets of B/A compounds as reference molecules after adding the remaining B/A and all A/B compounds to the database. This procedure made it possible to evaluate the search results at different 'selectivity levels': (i) 'target–selective' molecules had to belong to the same selectivity set as the reference molecules, that is, compounds selective for the second target are considered false–positives; and (ii) 'pair–active' compounds belong to both selectivity sets (B/A and A/B). Thus, for each target in a pair of targets, pair–active compounds are defined as the sum of target–selective and false–positive compounds. Although the identification of target–selective molecules is the ultimate goal of selectivity searching, the ability to detect pair–active compounds also provides valuable information. This is the case because pair–active compounds are active at the level of target families. A search calculation should also identify compounds that are active at the level of target families and distinguish them from irrelevant database compounds. Thus, the identification of target–selective molecules and the ratio of target–selective over pair–active compounds represent meaningful measures for the evaluation of search calculations.

Selectivity Search Data

As outlined above, systematic search calculations were carried out for all 26 selectivity sets and five fingerprints. In the following, representative results are discussed in *Figures 3.5* and *3.6* that reflect major trends revealed in this analysis. All remaining results are provided in *Appendix Figures C.1* and *C.2*. Furthermore, *Appendix Table C.1* reports hit and recovery rates for all fingerprints and selectivity sets and selection sets consisting of the top–scoring 100 database compounds. For each calculation, maximally possible hit rates are reported in *Appendix Table C.1*. Maximum hit rates are calculated by dividing the number of selective compounds that are available in the database as potential hits by the selection set size. As stated above, for selectivity sets containing more than 15 compounds, ten reference molecules were used and the remaining compounds were added as potential hits to the database. For example, the D1/D2 set contains 31 compounds and, in this case, the maximum hit rate for target–selective compounds is 21% (see *Appendix Table C.1*). The D2/D1 set contains 26 compounds and the corresponding maximum hit rate is 16%. Accordingly, for the target pair (D1 and D2), the maximum possible hit rate for pair–active compounds is 47%. Over all selectivity sets, theoretically possible hit rates for target–selective molecules range from 4% to 59%. These limits are important to consider when putting the search performance into perspective. *Table 3.7* reports average search results for each fingerprint and the three target families studied here.

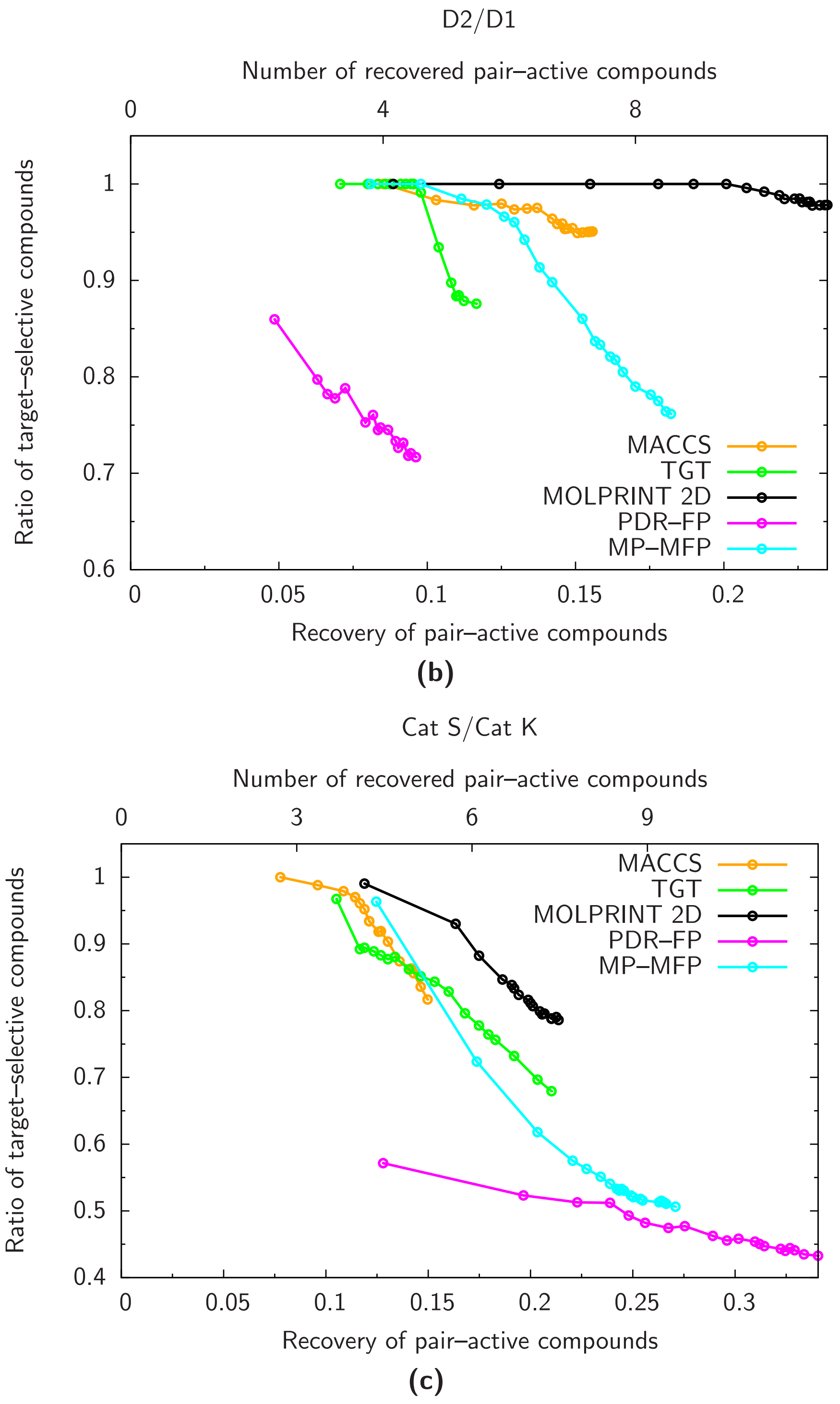


(a)

Figure 3.5: Retrieval of selective compounds. In (a)-(c), the average recovery of hits in selection sets of increasing size, ranging from 5 to 100 database compounds, is reported. The leftmost data points correspond to the smallest selection set size of five compounds; then, set sizes increase in increments of five compounds. The graphical representations monitor the retrieval of target–selective versus pair–active compounds, as defined in the text. The total number of recovered pair–active compounds is reported (top horizontal axis) and also the rate (bottom). The vertical axis reports the ratio of target–selective over pair–active compounds.

Detection of Target–Selective Compounds

Figure 3.5 monitors the recovery of target–selective versus pair–active compounds in database selection sets of increasing size and *Figure 3.6* reports average numbers of correctly identified target–selective and pair–active compounds. Regardless of the targets, a few general trends are clearly evident. All fingerprints are capable of detecting target–selective compounds and in a number of cases (for example, *Figure 3.5a* and *Appendix Figure C.1d*) their performance is almost indistinguishable. For other selectivity sets (for example, *Figure 3.5b*, *3.5c*, and *Appendix Figure C.1i*), in part significant differences in search performance are observed. In a number of cases, selection sets of increasing size only enrich target–selective but not false–positive active compounds (for example, *Figure 3.5a* and *Appendix*



(b)

(c)

Figure 3.5: Continued.

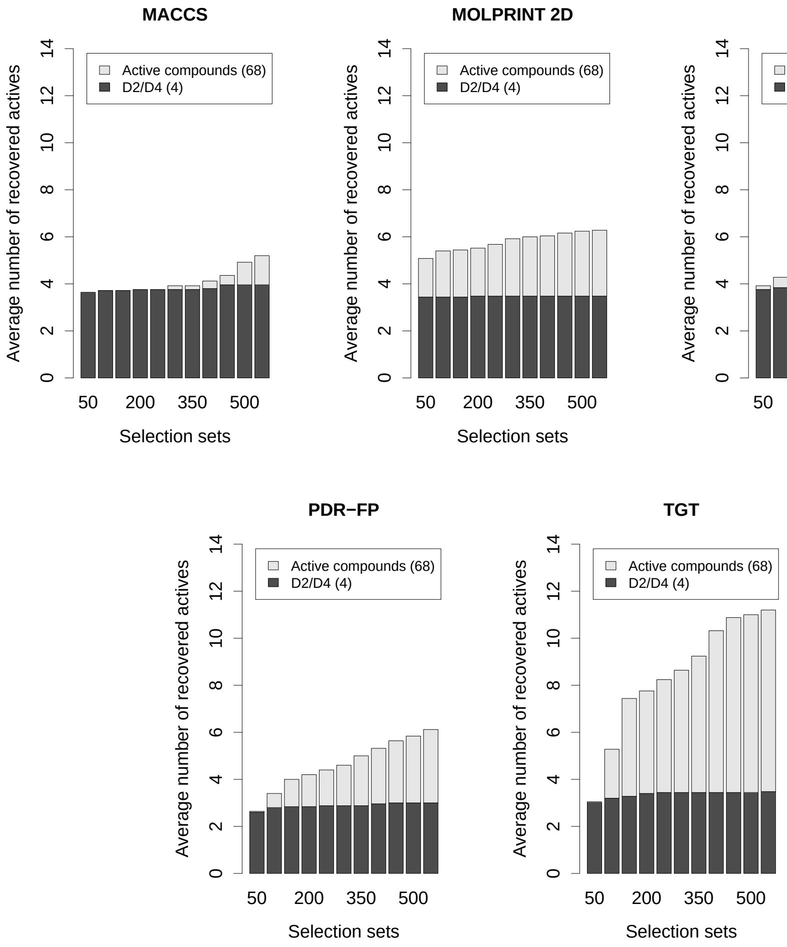
Figure C.1k), which is reflected by the fact that the ratio of target–selective compounds remains at the 100% level. In other cases, false–positive active compounds are gradually enriched as selection sets increase in size (for example, *Figure 3.5c* and *Appendix Figure C.1h*). Thus, the curves in *Figure 3.5* point downwards. However, when there is a notable tendency to enrich false–positive active compounds in selection sets, target–selective compounds are already enriched in small selection sets. These findings suggest that limiting selection set size is likely to direct search calculations towards exclusive detection of selective compounds. Thus, if maximizing compound recall is not a primary objective, focusing on only small numbers of top–scoring database compounds is a preferred search strategy for the identification of target–selective compounds. This is relevant for practical applications, since the novelty or individual quality of chosen compounds is often a primary objective, rather than the total number of active or selective compounds one might be able to find.

Target– versus Pair–Active Compounds

Figure 3.6 compares the ratio of target–selective versus pair–active compounds for different fingerprints and reveals that in most cases, significantly more selective than false–positive active compounds are found, which is an encouraging result. Furthermore, differences between fingerprints are only small for the majority of selectivity sets. However, there are exceptions across all three target families. For example, for the D2/D4 set (*Figure 3.6a*), TGT displays a significant tendency to detect false–positive actives, in contrast to MACCS. However, both fingerprints show comparable performance on target–selective compounds. For sets Cat K/Cat S and Cat S/Cat K (*Figure 3.6b* and *Appendix Figure C.2h*), essentially all fingerprints show a notable tendency to detect false–positive actives. In contrast, for the trypsin/thrombin set (*Figure 3.6c* and *Appendix Figure C.2i*), only PDR–FP has a high false–positive rate. Thus, fingerprints show differences in the ratio of pair–active versus target–selective compounds on a case–by–case basis, but overall differences are surprisingly small and target–selective molecules are much more frequently detected than false–positive active compounds.

Differences between Target Families

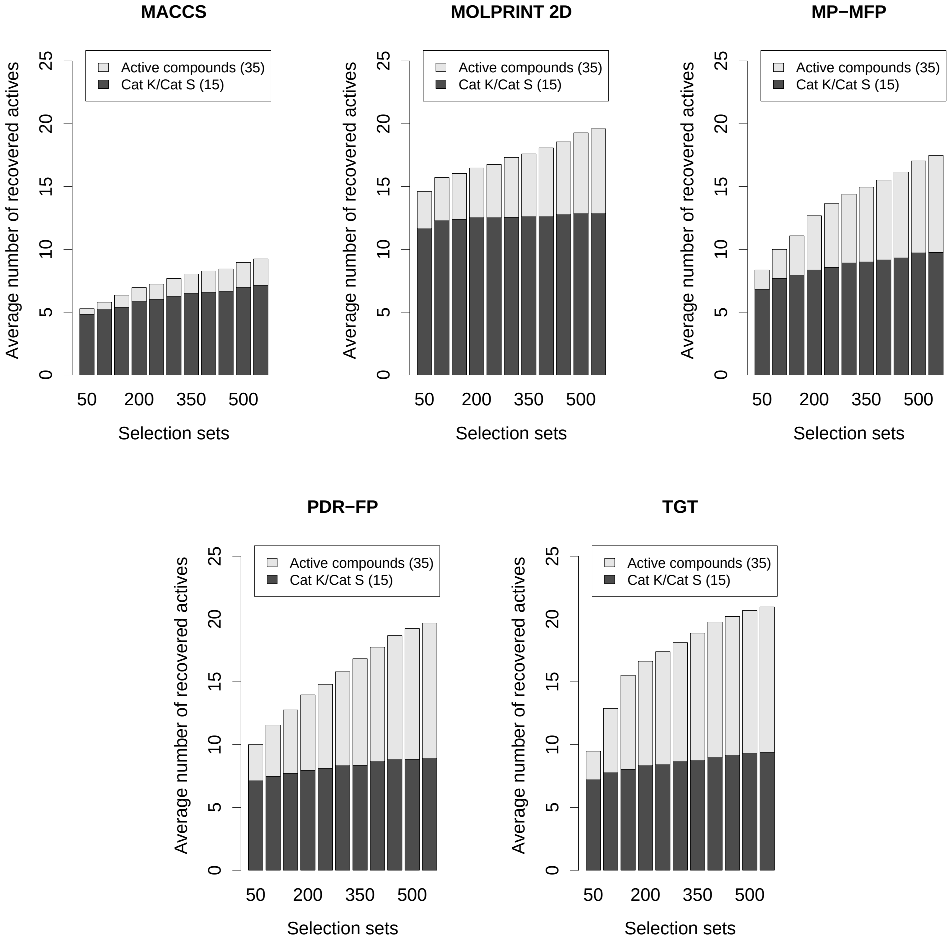
As shown in *Appendix B.1*, selectivity sets for all three target families differ in their degree of intra–set structural diversity, although the GPCR ligands tend to be on average more similar to each other than the protease inhibitors (Stumpfe



(a)

Figure 3.6: Ratio of false–positive active versus target–selective compounds. The bar graphs shown in (a)-(c) report the relative amounts of false–positive active (light gray) and target–selective (dark gray) compounds for each fingerprint method as an average number of these compounds in differently sized selection sets ranging from 50 to 550 compounds. According to the definition in the text, the sum of false–positive active and target–selective molecules are pair–active compounds. Total numbers of pair–active compounds are reported in parentheses.

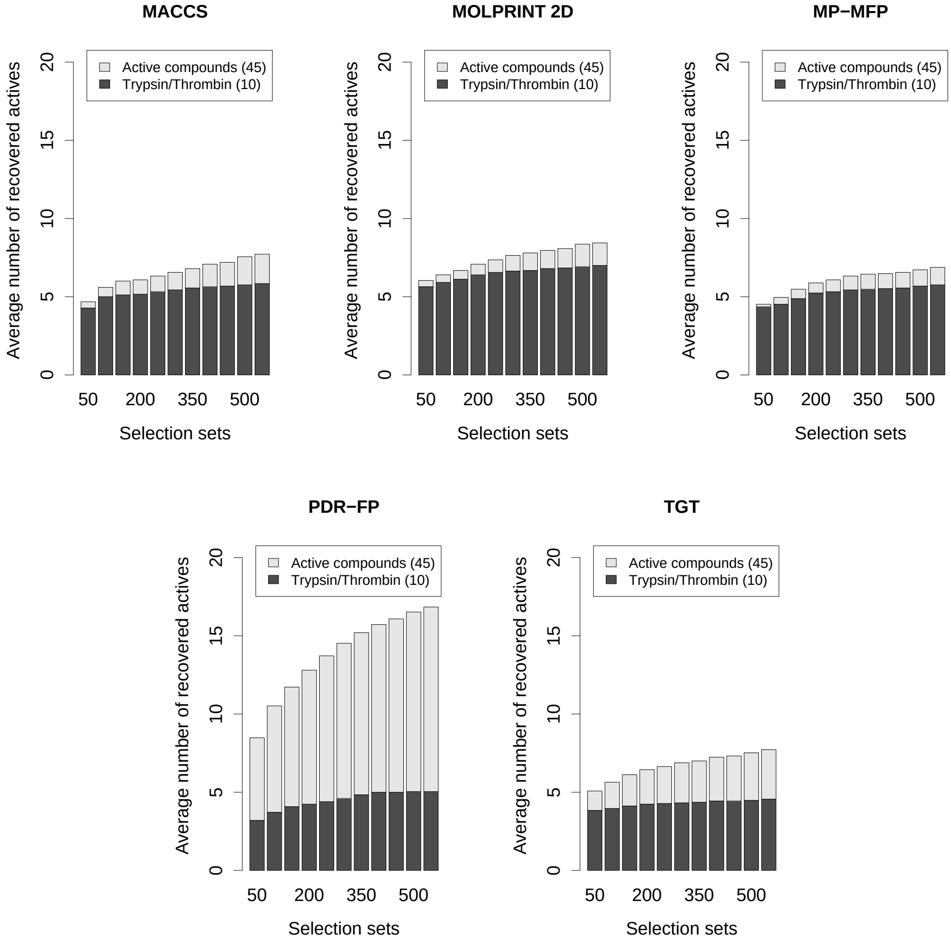
et al. [2007]). As expected, cathepsin inhibitors obtained from high–throughput screening data sets were structurally most diverse. At least two structural factors are expected to complicate selectivity analysis: (i) if compounds belonging to different selectivity sets have distinct structural similarity (as is the case for some of the GPCR sets), it might be difficult to distinguish between them; (ii) if selec-



(b)

Figure 3.6: Continued.

tive compounds show significant intra–set structural diversity (as is the case for some of the cathepsin inhibitors), it might be difficult to recognize them as being similar. However, the results in *Figure 3.5* demonstrate that selectivity search calculations are not systematically affected by general differences in intra–set and inter–set structural resemblance. For example, calculations on a number of GPCR sets and also on the Cat B/Cat S and Cat S/Cat B sets (*Figure C.2f* and *C.2g*) exclusively produce selective compounds. On the other hand, the trypsin/thrombin set generally produces more false–positives than the thrombin/trypsin set (*Figure 3.6c* and *C.2i*). Thus, the results of selectivity searching using fingerprints do not simply correlate with differences in the structural resemblance of compounds;



(c)

Figure 3.6: Continued.

compound class–selective features are indicated to play an important role.

Fingerprint Search Performance

Table 3.7 summarizes the average search performance of the different fingerprints for the three target families. For selection sets of 100 database compounds, hit rates for target–selective compounds are generally below 10% for GPCR ligands and cathepsin inhibitors and above 10% for serine protease inhibitors. However, it must be taken into account that theoretically possible hit rates for target–selective molecules range from only 4% to 59% (see *Appendix Table C.1*), with an average maximum hit rate of approximately 20% over all selectivity sets. Furthermore,

	Biogenic amine GPCRs		**Papain–like proteases**		**Chymotrypsin–like proteases**	
	PAC	TSC	PAC	TSC	PAC	TSC
MACCS						
HR	9.4	8.2	4.8	4.6	11.6	10.5
RR	19.6	55.6	12.4	35.1	16.3	39.1
TGT						
HR	7.6	7.1	6.2	5.1	11.6	11.0
RR	17.6	49.2	16.1	38.6	16.7	40.4
MOLPRINT 2D						
HR	11.4	10.6	7.0	6.4	15.5	14.4
RR	26.4	69.4	18.3	46.7	22.4	52.0
PDR–FP						
HR	6.3	5.9	6.0	4.5	19.2	12.4
RR	15.2	44.5	15.8	34.0	29.2	47.5
MP–MFP						
HR	9.7	9.1	6.0	5.1	11.5	10.7
RR	22.7	61.2	15.6	38.4	15.8	38.9

Table 3.7: Average fingerprint performance. Comparison of family–based averaged similarity search performance of the studied five fingerprints, evaluated according to hit and recovery rates of pair–active (*PAC*) and target–selective (*TSC*) compounds, given as percentage.

comparable numbers of target–selective compounds are often found in selection sets of much smaller size (*Figure 3.5*). As illustrated in *Figure 3.6*, only small numbers of false–positive active compounds are detected in most cases. In addition, *Table 3.7* shows that recovery rates for selection sets of the top 100 scored compounds are significant, ranging from approximately 30% to 70% for target–selective molecules. The fingerprints were chosen because they represent different design strategies and have in part very different complexity. However, there is no apparent relationship between fingerprint complexity and search performance. In some cases, different fingerprints achieve very similar results. In others, performance varies depending on the selectivity set. Furthermore, for target pairs, fingerprint performance is frequently seen to vary depending on the search direction. Individual differences in fingerprint search performance often occur when selective compounds are structurally heterogeneous, for example, in the case of cathepsin inhibitors. Despite these compound set–dependent variations, differences in selectivity search performance between the fingerprints tested here are not dramatic. The two fingerprints of lowest complexity, MACCS and MP–MFP, produce significant recovery rates comparable or higher than those obtained with more complex designs. Overall, MOLPRINT 2D gives highest hit rates on target–selective compounds and PDR–FP lowest (see also *Appendix Table C.1*). These findings contrast the outcome of similarity search calculations on increasingly structurally diverse compound activity classes where PDR–FP was often found to perform best (Eckert and Bajorath [2006a]; Tovar et al. [2007]). Therefore, the results of activity– and selectivity–oriented similarity searching do not correspond to each other. This means that molecular similarity relationships are likely to differ within compound activity classes and selectivity sets and that recognizing compounds with similar activity against a single target or detecting molecules with different target selectivity challenges search methods in different ways. Clearly, a characteristic feature of the presented search calculations is that target–selective compounds are typically found in small database selection sets. Another important observation has been that in about half of the test cases recognition of false–positive actives does not significantly increase when larger sets of database compounds are selected. Thus, these calculations have notable specificity.

3.3 Analysis of Differential Selectivity Profiles

Based on the findings that 2D fingerprints are capable of differentiating between compounds from pairwise selectivity sets, the study presented in this section investigated the performance of a wider spectrum of computational methods to study different aspects of target selectivity. Compounds belonging to the same selectivity sets are now selective for one target over one or multiple other related other ones, thereby allowing the analysis of selectivity in terms of an entire target family. A systematic computational analysis of structure–selectivity relationships was carried out on a refined biogenic amine GPCR dataset with different classification methods and the results were compared compared to two previously studied 2D fingerprints.

Calculation Setup

The 267 antagonists contained in the GPCR$_f$ database cover seven different receptors belonging to three amine receptor subfamilies and display different selectivity profiles for one receptor over several others, as is shown in *Table 3.8*.

For fingerprint searching, recursive partitioning (RP), and DynaMAD, 25 different reference sets from each selectivity set were randomly selected and 25 independent calculations were carried out. The results were averaged. Reference sets always consisted of half of the molecules in each selectivity set. The remaining antagonists were added as potential hits to a subset of the ZINC7 database, containing ∼3.7 million compounds. In each DynaMAD calculation, dimension extension was continued over 20 dimension extension steps and hit and recovery rates were monitored.

RP models were trained using the reference sets in the presence of different sets of 500 randomly selected database molecules and used to search the background database. Non–hierarchical Jarvis–Patrick (JP) and hierarchical–agglomerative Ward's clustering (WA) were applied to these selectivity sets in the absence of background database compounds because these clustering methods are computationally expensive and it is difficult to use them for the analysis of large databases. Furthermore, similarity search calculations using the MACCS and MOLPRINT 2D fingerprints were performed on the basis of the Tanimoto similarity coefficient and using multiple reference compounds in combination with the 5NN nearest neighbor search strategy. For Ward's clustering and RP, the same pool of descriptors was used as for DynaMAD. Calculations for JP were carried out using the MACCS keys fingerprint.

Performance of Clustering Methods

In order to estimate the degree of difficulty of distinguishing the structure–selectivity relationships in the antagonist database, it was attempted to separate the selectivity sets in the absence of background database compounds using standard clustering methods. The 267 antagonists were clustered using the JP and WA algorithms together with MACCS keys and the pool of 155 1D/2D descriptors, respectively. Smaller descriptor sets were also selected and evaluated on the basis of feature significance and contingency analysis carried out with MOE. JP clustering did not produce meaningful results. Different similarity and overlap settings produced between ca. 90 to 100 clusters dominated by singletons and containing no more than three compounds. WA hierarchical clustering, that has often shown better performance in the classification of molecules than non–hierarchical clustering (Brown and Martin [1996]), produced better results that are summarized in *Appendix Table C.2.* Here, different descriptor sets produced comparable cluster compositions. For example, eight clusters at clustering level three contained between eight and 53 compounds and were 33% - 100% pure. However, seven of eight clusters combined antagonists from multiple selectivity sets. Thus, even in the absence of database compounds, clustering of selectivity

Selective for	Over	Size
5HT1a	5HT2a, Alpha1, D2	53
5HT2a	5HT1a, Alpha1, D1, D2	21
Alpha1	5HT1a, 5HT2a, D1, D2, D3, D4	26
D1	D2, D4	33
D2	5HT1a, D1, D3, D4	25
D3	D1, D2, D4	37
D4	Alpha1, D1, D2, D3	72

Table 3.8: Composition of the GPCR$_f$ database. This database contains seven compound classes of selective GPCR antagonists. To be included in a compound selectivity set, an antagonist is required to have at least 50–fold higher potency for one biogenic amine GPCR over one or more others. For a detailed description see *Appendix B.1.1*.

sets was difficult suggesting that differentiating the underlying selectivity profiles on the basis of molecular structure would be a non–trivial exercise.

Analysis of Similarity Search Calculations

Compared to a conventional virtual screen for compounds having similar activity, searching for target–selective compounds presents additional challenges. As stated above, this is the case because GPCR antagonists must not only be distinguished from database decoys but also separated on the basis of their selectivity profiles. Each of the selectivity sets consists of target–selective antagonists and the combination of these sets can be regarded as an array of biogenic amine GPCR family–selective compounds. Thus, selectivity searching, as described herein, should ideally maximize the recall of target–selective and minimize the recall of family–selective and background database compounds. Preferential detection of family–selective compounds is in principle also a positive result, similar to a conventional search for active compounds, but does not account for selectivity at the level of individual targets, which represents the primary goal of this investigation.

As already shown in the last section, 2D fingerprints displayed a tendency to preferentially recognize selective molecules over database compounds when using multiple selective reference compounds. Therefore, two fingerprints of different design and complexity, MACCS and MOLPRINT 2D, were tested on the newly assembled selectivity sets. *Table 3.9* reports the results obtained for MACCS and MOLPRINT 2D calculations. Both fingerprints successfully retrieved target–selective molecules within the 100 top–scoring database compounds. Depending

Target	TSC	FSC	ZINC7	HR [%]	RR [%]
5HT1a	11.6	1.0	87.4	11.6	43.1
5HT2a	5.6	0.1	94.3	5.6	51.3
Alpha1	7.9	0.0	92.1	7.9	60.6
D1	10.6	0.0	89.4	10.6	62.6
D2	6.0	0.0	94.0	6.0	45.9
D3	13.6	0.4	86.1	13.6	71.4
D4	12.3	0.4	87.4	12.3	34.1

(a) MACCS

Target	TSC	FSC	ZINC7	HR [%]	RR [%]
5HT1a	23.9	9.8	66.3	23.9	88.6
5HT2a	8.2	2.2	89.6	8.2	74.2
Alpha1	10.8	17.4	71.8	10.8	83.1
D1	16.0	1.0	83.0	16.0	93.9
D2	11.7	12.6	75.7	11.7	89.9
D3	18.1	4.8	77.2	18.1	95.2
D4	29.2	3.2	67.5	29.2	81.2

(b) MOLPRINT 2D

Table 3.9: Average results of 5NN MACCS and MOLPRINT 2D similarity search trials. *TSC* and *FSC* stand for target– and family–selective compounds, respectively. For TSC, FSC, and ZINC7, the average number of compounds among the top–scoring 100 compounds per trial is reported. Hit and recovery rates are calculated for target–selective compounds.

on the selectivity set, MACCS achieved hit rates of up to 13.6% and recovery rates of 71.4%. MOLPRINT 2D achieved hit rates of up to 29.2% and had consistently high recovery rates ranging from greater than 70% to 95%. However, it can be observed from *Table 3.9* that MOLPRINT 2D recognized considerable numbers of family–selective antagonists, whereas MACCS essentially detected none. For two classes, Alpha1 and D2, MOLPRINT 2D detected more family– than target–selective compounds. Thus, although MOLPRINT 2D produced consistently higher recall than MACCS, it was much less capable of differentiating between target– and family–selective antagonists. This is an interesting finding in light of the low complexity of MACCS, which suggests that structural keys are particularly attractive descriptors for selectivity analysis.

Recursive Partitioning

Next, RP was investigated as a statistical compound classification technique that can effectively process large numbers of descriptors and database compounds, in contrast to standard clustering methods. For each set, 25 independently trained RP models were used to search for antagonists in ZINC7. The best models produced top recall rates between 23% and 82% for the selectivity sets, but selected also between 209 and 2 171 ZINC7 compounds, which resulted in maximal hit rates between 0.4% and 3.3%. Average hit rates for all 25 models were lower than 1% for six of seven selectivity sets. Thus, in general, RP models selected too many database compounds, which might in part be attributed to the relatively small number of selective reference compounds available for model building.

Dynamic Compound Mapping

Then, the performance of compound mapping was evaluated. DynaMAD was found to successfully discriminate between selective molecules and database compounds and approached the overall performance of fingerprint searching. As shown in *Table 3.10*, hit and recall rates of 5.2% to 21.1% and 21.5% to 45.9%, respectively, were observed. DynaMAD hit and recall rates were lower than for MOLPRINT 2D, but DynaMAD hit rates were overall higher than for MACCS. DynaMAD generally also recognized significantly more target– than family–selective antagonist, except for D2 where more family–selective molecules

Target	TSC	FSC	ZINC7	HR [%]	RR [%]	DEL	Descr
5HT1a	8.5	0.1	85.4	9.0	31.4	9	47.6
5HT2a	2.4	0.0	26.3	8.2	21.5	6	25.8
Alpha1	3.6	0.0	66.0	5.2	28.0	5	24.4
D1	7.8	1.5	77.6	9.0	45.9	7	20.7
D2	3.9	8.9	48.9	6.4	30.2	6	17.7
D3	8.0	5.4	24.4	21.1	41.9	1	23.3
D4	12.5	1.0	61.8	16.6	34.8	14	90.1

Table 3.10: Average results of DynaMAD trials. *TSC* and *FSC* stand for target– and family–selective compounds, respectively. For TSC, FSC, and ZINC7, the average number of compounds per run is reported. Hit and recovery rate are calculated for target–selective compounds. Results are reported for the dimension extension level (*DEL*) yielding a total of 100 or fewer database compounds. *Descr* gives the number of descriptors at the reported DEL, which is equivalent to the dimensionality of the chemical reference space for this mapping step.

were recognized. Overall, the ability of DynaMAD calculations to discriminate target– and family–selective compounds was intermediate between MACCS and MOLPRINT 2D. In contrast to fingerprint search calculations, DynaMAD does not produce a ranking of database compounds according to similarity scores. Rather, mapping statistics at different dimension extension levels determine the number of database compounds that successfully match selective descriptor value ranges. Therefore, dimension extension levels were chosen as reference points for calculation of hit and recovery rates that retained fewer than 100 database molecules, as shown in *Table 3.10*.

However, *Appendix Table C.3* reports the corresponding results over all 20 dimension extension steps and *Figure 3.7* shows a graphical representation of hit rates over all dimension extension levels. As can be seen, for five of seven selectivity sets, DynaMAD calculations displayed a strong tendency to deselect database compounds in increasingly high–dimensional descriptor spaces, that is, under increasing mapping stringency, and preferentially retained selective molecules, thereby producing hit rates approaching 100%. As reported in *Appendix Table C.3*, compounds retained until the end of dimension extension were mostly, or exclusively, target–selective. These observations suggest a search strategy for practical applications when recall rates are not a primary measure of success and it is more important to identify attractive new hits. Under these conditions, DynaMAD calculations should be continued until only very few database compounds remain. These might then have a high probability to contain molecules with desired selectivity.

3.4 Design of Target–Selective Descriptor Spaces

As shown in the previous two sections, both DynaMAD and state–of–the–art fingerprints display a tendency to detect target–selective compounds in screening databases when selective reference molecules are used. These findings suggested that selectivity differences between molecules that are active against related targets might be predictable using currently available similarity–based methods. Assessing compound selectivity is more demanding than normal similarity searching because chemically related compounds with differential target activities have to be distinguished, and thus the design of adequate chemical reference spaces is of crucial importance. In order to investigate the feasibility of target–selective chemical space design, a second generation DynaMAD–like mapping algorithm termed CA–DynaMAD was developed and applied to the GPCR$_f$ database (see

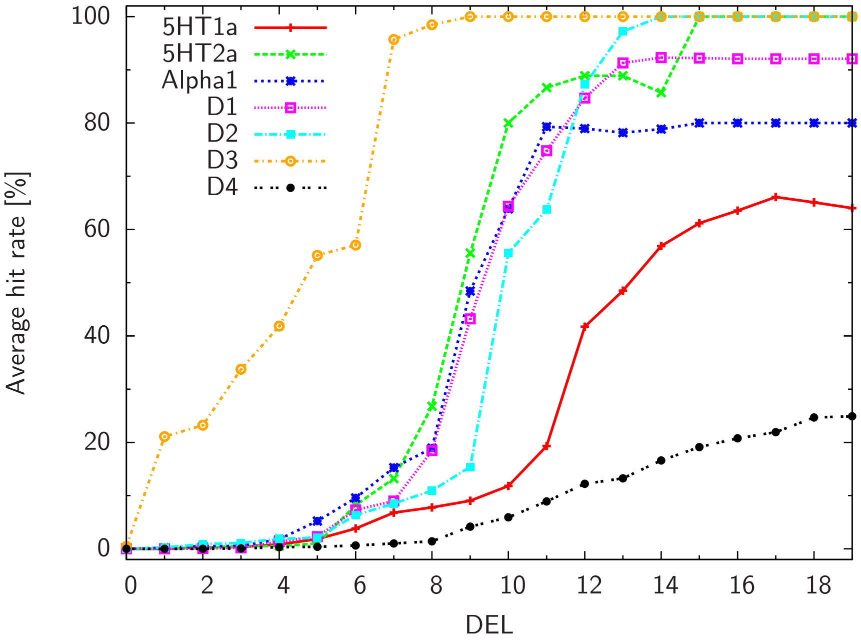


Figure 3.7: Average performance of DynaMAD. For each selectivity set, average hit rates are monitored over all 20 dimension extension levels (*DEL*).

Appendix B.1.1 for a detailed description), containing molecules with experimentally determined binding profiles against multiple closely related biogenic amine GPCRs. Additionally, the performance of CA–DynaMAD and DynaMAD was compared and analyzed.

Benchmark Calculations

For the comparison of DynaMAD and CA–DynaMAD, systematic test calculations were carried out as follows. Each selectivity set was 100 times randomly divided in half. One subset of the selective compounds was used as a reference set, while the remaining compounds from the second subset were added to the the ZINC7 database as potential selective hits. In total, 100 independent search calculations were carried out for each selectivity set and each calculation was terminated when 100 or fewer compounds in total remained.

Flexible Design of Target–Selective Spaces

In order to identify descriptor combinations that distinguish a selectivity set from all others, each selectivity set was used once as a reference set and the remaining six sets as test compounds. CA–DynaMAD calculations were carried out until the maximum number of these GPCR test compounds were eliminated through addition of high–scoring reference set descriptors. The resulting descriptor combinations form target–selective chemical reference spaces. Subsequently, the MDDR database (see *Appendix Chapter B.2*) was mapped to the previously generated target–selective descriptor spaces and the number of MDDR compounds matching all selectivity set value ranges was determined. For each selectivity set, dimension extension was then continued until 100 or 50 molecules remained. The ACTIVITY and ACTION fields in the entries of all compounds in MDDR selection sets were inspected to determine whether antagonists and target–selective hits were identified. Prior to this analysis, selectivity set compounds also present in the MDDR were removed from the database.

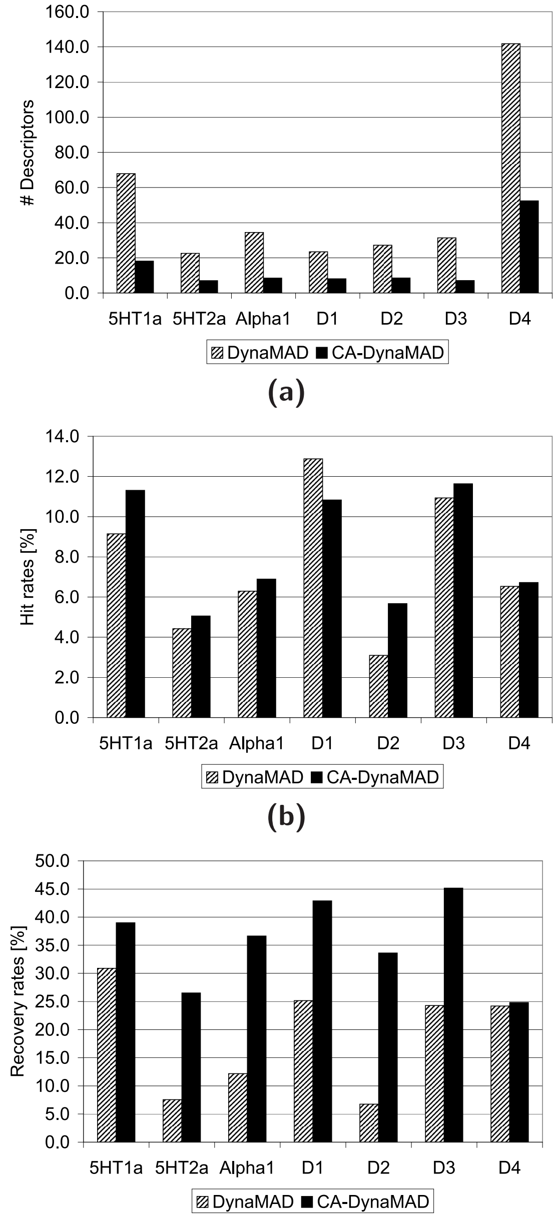
Comparison of DynaMAD and CA–DynaMAD

For the comparison of CA–DynaMAD and DynaMAD, test calculations were carried out on the seven compound selectivity sets. Akin to typical virtual screening benchmark calculations, subsets of target–selective compounds were used as reference sets to search for the remaining selective molecules added to a background database. For each selectivity set, a total of 100 independent search calculations were carried out with varying reference set composition. Thus, a total of 1 400 search calculations were performed. In 1 376 of these trials, final database selection sets of 100 or fewer compounds contained correctly identified target–selective molecules. DynaMAD calculations failed to recover selective compounds in 23 individual trials (four, six, and 13 for Alpha1, 5HT2a, and D2, respectively), whereas CA–DynaMAD failed only in a single of its 700 trials (for Alpha1). A comparison of average results is shown in *Figure 3.8*. The effects of using a continuous dimension extension and adaptive descriptor scoring on the dimensionality of the final space representations can be observed in *Figure 3.8a*. For compound mapping using CA–DynaMAD, significantly fewer descriptors were required than for DynaMAD, on average, only about a third. *Figure 3.8b* illustrates that hit rates achieved with CA–DynaMAD and DynaMAD were overall comparable, despite significant differences in the number of descriptors that were required. However, as shown in *Figure 3.8c*, the recovery rates of selective compounds using CA–DynaMAD were consistently higher. On average, about 25% more target–selective compounds were recovered with CA–DynaMAD than with DynaMAD. This can be rationalized by the achieved gradual reduction of database size through continuous dimension extension and adaptive descriptor scoring, thus reducing the number of potential hits lost during dimension extension. This is the case because descriptors that discriminate most effectively between reference and database compounds are re–determined at each step. Thus, the test calculations confirmed the former expectations and the desired improvements of CA–DynaMAD over the original DynaMAD implementation.

Generation of Target–Selective Descriptor Spaces

Next, the design of target–selective space representations using CA–DynaMAD was investigated. In these calculations, compound mapping was carried out using each complete selectivity set as a reference until the maximum number of compounds belonging to the other selectivity sets was eliminated. For five of seven selectivity sets, all other GPCR antagonists were eliminated, thus producing entirely "pure" target–selective reference spaces. In the case of D4, a single 5HT1a compound remained and vice versa. Thus, even for selectivity sets D4 and 5HT1a, impurities were minute. Interestingly, only 29 of the 155 descriptors in our basis set contributed to target–selective space representations. A summary of these descriptors is provided in *Table 3.11.* Preferred descriptors included complex and orthogonal designs of the BCUT (Pearlman and Smith [1998]) and VSA (Labute [2004]) types, a number of topological indices, and even simple descriptors such as the fraction of rotatable bonds or the number of double bonds in a molecule. Of these 29 descriptors, 23 only contributed to one of seven target–selective space representations and four to two. The balabanJ index and the SlogP_VSA4 descriptor occurred three and four times, respectively. Thus, only a small number of property descriptors was required to successfully differentiate between GPCR antagonists having different selectivity and the majority of these descriptors uniquely contributed to different reference spaces. *Table 3.12* reports the CA–DynaMAD mapping statistics for each of the seven selectivity sets. As can be seen, only three to eight descriptors were required to build target–selective space representations. During dimension extension, GPCR antagonists belonging to other selectivity sets were gradually eliminated. However, the top–scoring one or two descriptors made the most significant contributions. In five cases, more than 70% of other GPCR antagonists did not match the selectivity reference set value ranges of the first descriptors. Thus, the respective descriptor value ranges were already signatures of GPCR antagonist selectivity. For selectivity sets D2 and D3, the top–scoring descriptors (BCUT_PEOE_0 and KierA3, respectively) deselected more than 90% of the other antagonists.

Taken together, these findings illustrate the ability of CA–DynaMAD calcu-



(a)

(b)

(c)

Figure 3.8: Comparison of DynaMAD and CA–DynaMAD. Average results of 100 randomized search calculations are reported for each selectivity set. In (a), the number of selected descriptors for DynaMAD and CA–DynaMAD calculations is given; (b) and (c) report hit and recovery rates at the final dimension extension steps, respectively.

Type	Descriptor	Explanation
Adjacency and distance matrix descriptors	balabanJ	Balaban's topological connectivity index
	BCUT_PEOE_0 BCUT_PEOE_3 BCUT_SLOGP_1 BCUT_SLOGP_3 BCUT_SMR_0	BCUT–type descriptors for PEOE partial charges, SlogP or molar refractivity (SMR)
	GCUT_PEOE_0 GCUT_SLOGP_0 GCUT_SMR_0	GCUT descriptors use graph distances instead of bond order information (like BCUTs)
	VDistEq	Distance matrix index
Atom and bond counts	b_1rotR	Fraction of rotatable bonds
	b_double	Number of double bonds
	lip_don	Number of OH and NH groups
Connectivity and shape indices	chi1	Connectivity index
	KierA3	Shape index
Partial charge descriptors	PEOE_VSA+0 PEOE_VSA+1	Fractional polar van der Waals surface area (vdW_sa)
	PEOE_VSA_FPPOS	Fractional positive (vdW_sa)
	PEOE_VSA_HYD	Total hydrophobic (vdW_sa)
Molecular surface descriptors	vsa_acc	Approximate sum of vdW surface areas of hydrogen bond acceptors
	vsa_pol	Approximate sum of vdW surface areas of polar atoms
Subdivided surface area descriptors	SlogP_VSA1 SlogP_VSA4 SlogP_VSA5 SlogP_VSA7	Combined with SlogP
	SMR_VSA1 SMR_VSA2 SMR_VSA4 SMR_VSA6	Combined with molar refractivity (SMR)

Table 3.11: Descriptors for target–selective chemical spaces. Descriptors are abbreviated and explained according to the Molecular Operating Environment and its documentation material.

lations to identify preferred descriptors for target–selective chemical space design and compound classification.

Space Representations for Database Searching

Next, ~160 000 MDDR compounds were mapped to the selectivity set value ranges of the descriptors forming target–selective reference spaces. The results are reported in *Table 3.13*. Only ~1 000 to 7 000 MDDR compounds, depending on the selectivity set, matched the descriptor value ranges of the target–selective spaces (on average, ~3 000 compounds). Thus, although these space representations were low–dimensional and derived only on the basis of a total of 267 GPCR antagonists, they were already highly selective, accepting on average only less than 2% of the MDDR, that contains ~11 000 GPCR ligands (see *Appendix B.2*).

The dimensionality of the target–selective spaces was subsequently extended in order to eliminate increasing numbers of MDDR compounds from them and determine whether new target–selective compounds could be identified. Dimension extension using CA–DynaMAD was continued until only small database selection sets remained. It should be noted that dimension extension does not modify the original selectivity of these spaces. All target–selective compounds remain selected because they represent the reference set for continued dimension extension, and all other GPCR antagonists used for space design remain deselected. However, each additional dimension extension step eliminates a subset of database compounds. The results for the last dimension extension steps are reported in *Table 3.14*. For six selectivity sets (except D4), database selection sets contained 50 or fewer MDDR compounds. For D4, 113 compounds remained because no discriminatory descriptors were available to further reduce the number of database compounds. In order to deselect nearly all MDDR compounds, the dimensionality of the original reference spaces needed to be in part significantly increased. As reported in *Table 3.14*, addition of between five (D1) and 51 (D4) descriptors was required. This was the case because random database compounds typically have much broader value ranges than selectivity sets, which generally reduces the discriminatory power of individual descriptors. Consequently, more descriptors were required. However, through adaptive descriptor scoring, highly selective chemical references spaces were obtained for each of the seven selectivity sets that eliminated almost all MDDR compounds.

Step	Descriptor	Deselection [%]	Purity [%]
1	chi1	54.7	35.3
2	balabanJ	20.6	50.0
3	SMR_VSA1	11.2	64.6
4	SlogP_VSA1	6.1	76.8
5	PEOE_VSA+0	3.3	85.5
6	GCUT_PEOE_0	1.4	89.8
7	VDistEq	1.4	94.6
8	SlogP_VSA4	0.9	98.2

(a) 5HT1a

Step	Descriptor	Deselection [%]	Purity [%]
1	BCUT_SLOP_3	76.4	26.6
2	PEOE_VSA+1	11.8	42.0
3	SlogP_VSA7	5.3	56.8
4	BCUT_SLOGP_1	3.3	72.4
5	GCUT_SMR_0	2.9	95.5
6	SMR_VSA2	0.4	100.0

(b) 5HT2a

Step	Descriptor	Deselection [%]	Purity [%]
1	vsa_acc	85.5	42.6
2	SMR_VSA4	12.4	83.9
3	b_1rotR	1.7	96.3
4	BCUT_PEOE_3	0.4	100.0

(c) Alpha1

Step	Descriptor	Deselection [%]	Purity [%]
1	BCUT_SMR_0	74.8	35.9
2	SlogP_VSA5	17.5	64.7
3	lip_don	4.3	80.5
4	SMR_VSA6	2.6	94.3
5	balabanJ	0.4	97.1
6	SlogP_VSA4	0.4	100.0

(d) D1

Table 3.12: Generation of target–selective chemical space representations. *Deselection* reports the percentage of the total number of GPCR antagonists with different selectivity that were eliminated through the addition of each descriptor. *Purity* gives the cumulative percentage of target–selective antagonists among all compounds passing the dimension extension step.

Step	Descriptor	Deselection [%]	Purity [%]
1	BCUT_PEOE_0	91.3	54.4
2	b_double	6.2	80.7
3	vsa_pol	2.5	100.0

(e) D2

Step	Descriptor	Deselection [%]	Purity [%]
1	KierA3	93.5	71.2
2	GCUT_SLOGP_0	4.8	90.2
3	SlogP_VSA4	1.7	100.0

(f) D3

Step	Descriptor	Deselection [%]	Purity [%]
1	PEOE_VSA_HYD	50.8	42.9
2	BCUT_SLOGP_3	22.6	58.1
3	GCUT_SMR_0	14.4	75.0
4	balabanJ	6.7	86.8
5	b_1rotR	3.1	93.6
6	SMR_VSA4	1.0	96.0
7	SlogP_VSA4	0.5	97.3
8	PEOE_VSA_FFPOS	0.5	98.6

(g) D4

Table 3.12: Continued.

Identification of Selective GPCR Antagonists

These small MDDR selection sets where then analyzed in order to determine whether the CA–DynaMAD calculations identified selective GPCR antagonists. Therefore, the biological annotation fields of all selected MDDR compounds were studied. The results are summarized in *Table 3.15*. For each selectivity set, a number of antagonists (between four and 44) with confirmed activity against the selection set target were correctly detected but only for a total of 14 of these compounds selectivity information was available. However, 13 of these 14 compounds belonging to six selectivity sets were found to have correct selectivity profiles. Only a single false–positive compound was found for D2 (with selectivity for 5HT1a over D2 and Alpha1). Thus, reference spaces generated with CA–DynaMAD were successfully applied to identify target–selective antagonists in the MDDR, suggesting that these space representations might have considerable potential for practical applications.

Class	No. of descriptors	No. of MDDR compounds
5HT1a	8	1 096
5HT2a	6	1 397
Alpha1	4	7 418
D1	6	2 339
D2	3	2 651
D3	3	4 153
D4	8	1 914

Table 3.13: MDDR compounds in low–dimensional target–selective GPCR descriptor spaces. Reported are the number of MDDR compounds that successfully mapped to each of the initially created target–selective reference spaces.

Class	Dimensionality	No. of MDDR compounds
5HT1a	30 (+22)	49
5HT2a	14 (+8)	50
Alpha1	12 (+8)	45
D1	11 (+5)	47
D2	9 (+6)	44
D3	10 (+7)	48
D4	59 (+51)	113

Table 3.14: MDDR compounds in high–dimensional target–selective descriptor spaces. The numbers in parentheses report the increase in dimensionality relative to the original low–dimensional target–selective spaces.

Class	Identified antagonists	Confirmed selective antagonists
5HT1a	22 (49)	1 [5HT1a over D1, D2, and 5HT2] 2 [5HT1a over Alpha1]
5HT2a	6 (50)	1 [5HT2a over Alpha1 and D2] 1 [5HT2a over 5HT1a, D4, and Alpha1]
Alpha1	9 (45)	1 [Alpha1 over D2 and 5HT2]
D1	11 (47)	1 [D1 "over other GPCRs"] 1 [D1 over D2]
D2	4 (44)	1 [D2 over D3 and D4] *1 [for 5HT1a over D2 and Alpha1]*
D3	5 (48)	–
D4	44 (113)	3 [D4 over D2] 1 [D4 over 5HT1a and D2]

Table 3.15: Selective GPCR antagonists found in the MDDR. In the MDDR entries, ACTIVITY fields were inspected to determine the number of correctly identified GPCR antagonists (*identified antagonists*) among the total number of selected MDDR compounds (in parentheses) and ACTION fields to obtain selectivity information (*confirmed selective antagonists*). The numbers of confirmed selective antagonists with different selectivity profiles are given in bold and the selectivity profiles are reported in brackets. The profile of the only antagonist with incorrect selectivity that was detected (for D2) is highlighted in italics. For example, the row for 5HT1a needs to be interpreted as follows: 22 of 49 selected MDDR compounds were designated 5HT1a antagonists (ACTIVITY) and for three of those, selectivity information was provided (ACTION).

Chapter 4

Summary

In this thesis, ligand potency and selectivity were computationally analyzed using 2D molecular fingerprints and mapping algorithms. Biological screening data were used to evaluate and confirm the suitability of 2D LBVS methods for discriminating between active compounds with different potency. Then, two studies established the general ability of 2D virtual screening methods to distinguish between active compounds with different target–selectivity profiles at the target–, sub–family– and family level. A dynamic mapping algorithm was designed capable of generating target–selective chemical reference spaces.

Chapter 3.1 reports the application of potency–directed similarity searching to retrieve most potent hits from biological screening data. An approach is described to incorporate compound potency as a parameter in similarity search calculations in two conceptually different virtual screening methods. In both cases, a logarithmic scaling function is used that assigns scaling factors to active compounds according to their relative potency. In order to assess their potential for potency–directed similarity searching, virtual screening calculations have been conducted using a database of experimentally confirmed active and inactive compounds originating from an high–throughput screen for cathepsin B inhibitors. The structural diversity among HTS hits spanning a relatively narrow and continuous potency range and the similarity of active and inactive molecules made this screening set a rather challenging test case for potency–scaled similarity analysis. However, it provided a much more realistic and better defined benchmarking situation. In general, narrow potency distributions and diverse active compounds are difficult distinguish in potency–oriented molecular similarity calculations. By contrast, the presence of discontinuous potency distributions over several orders of magnitude and structural homogeneous subsets of highly and weakly potent compounds makes it easy to direct search calculations toward the recognition of the most potent database hits through potency scaling procedures. Given the features of the cathepsin B HTS set as described, it was possible to ask two major questions in this study: (a) can active molecules be recovered from this experimental set through application of DMC and DACCS, and if so, (b) is it possible to enrich selections with potent database compounds by incorporating ligand potency information into these methods? In other words, can similarity calculations be directed toward the recognition of potent hits when mining experimental screening data?

Although the biological screening data were challenging for the virtual screening methods, DMC and DACCS calculations consistently retrieved active molecules, in part, in very small selection sets. Both POT–DMC and POT–DACCS calculations favored the recognition of potent hits and displayed a tendency to deselect weakly potent ones, while retrieving similar or larger numbers of hits from the HTS data set compared to nonscaled calculations. Thus, it can be concluded that potency scaling, as implemented in POT–DMC and POT–DACCS, can successfully extrapolate from the features of potent reference molecules and direct LBVS calculations toward the recognition of potent hits. Molecular similarity methods are typically qualitative in nature, and POT–DMC and POT–DACCS are the first similarity methods that explicitly consider relative compound potency during database searching. Since DMC and DACCS are methodologically distinct, it is conceivable that potency scaling could also be applied to similarity search tools or compound classification methods other than these two algorithms.

In *Chapter 3.2*, a systematic study is presented that assesses the ability of existing state–of–the–art 2D molecular fingerprints to distinguish between ligands with different target selectivity. The selected fingerprints represent designs of varying complexity and are based on structural fragments, connectivity patterns, molecular property descriptors, or combinations thereof. The calculations analyzed compounds selective for 13 targets belonging to the three families of biogenic amine GPCRs, papain and chymotrypsin proteases, defining a set of 26 classes of compounds with selectivity for one target over another one.

Although selectivity search performance has only been moderate in some of the test cases studied here, an interesting finding has been that essentially all fingerprints tested displayed the potential to detect selective compounds in small database selection sets, regardless of their design and complexity. The results of systematic similarity search calculations reveal that 2D fingerprints are capable of identifying compounds having different selectivity against closely related target proteins, although fingerprints were originally not developed for such applications. Selectivity search performance displayed a substantial compound class–dependence, as has often been observed in similarity searching for active compounds. Thus, it would be very difficult to predict which type of fingerprint to use for a particular selectivity search application. A suitable strategy would be to apply different fingerprints in parallel, which is further supported by the observation that essentially all fingerprints evaluated in this study had predictive value. These findings also suggest that it might not necessarily be required to develop conceptually novel computational methods to aid in chemical genetics and genomics applications. Rather, existing methods might be adapted for such tasks. In addition, the results indicate that different molecular similarity relationships determine compound activity and target selectivity.

The finding that 2D fingerprints of markedly different complexity frequently yield comparable performance in selectivity–oriented similarity searching expands the scope and application radius of already established similarity–based methods in chemoinformatics. As six of the compound classes originated from HTS data, the corresponding experiments could be conducted under more realistic conditions by using the confirmed inactives as screening database. Thus, it can be anticipated that the findings of this study also hold true in real–world applications.

Based on the results reported in *Chapter 3.2*, the potential of fingerprint searching and different compound classification methods to identify and distinguish between GPCR antagonists displaying seven different target selectivity profiles was explored as reported in *Chapter 3.3*. A refined database of selective antagonists of biogenic amine GPCRs was used that enabled the study of target selectivity with respect to an entire protein family.

The MACCS and MOLPRINT 2D fingerprints, which already had proven their value for the identification of selective compounds in the previous study, were selected and added to a set of compound classification methods, covering standard algorithms like recursive partitioning as well as the recently introduced DynaMAD approach.

Prior to virtual screening benchmark experiments, two clustering methods were applied to group the GPCR antagonists according to their selectivity profiles. While JP clustering failed by mainly producing singletons, Ward's clustering provided better grouping but with limited cluster purity, suggesting that discriminating between compounds with different selectivity profiles based on molecular structure would not be trivial. RP calculations achieved acceptable recall rates but failed to effectively distinguish active from database compounds. The analysis of the remaining methods revealed that the MACCS and MOLPRINT 2D fingerprints and DynaMAD detected selective GPCR antagonists for most of the test sets among reasonably sized selection sets. MACCS and DynaMAD displayed the strongest tendency to preferentially identify target– over family–selective compounds, while MOLPRINT 2D produced the overall highest recall of family– and target–selective antagonists. This suggests that compound mapping in high–dimensional chemical space representations has the potential to detect subtle differences in closely related molecules that are selectivity determinants. Moreover, it can be concluded that structural fragment–type descriptors are attractive tools for the analysis of structure–selectivity relationships.

Finally, *Chapter 3.4* reports on the design of target–selective chemical reference spaces generated by an advancement of the DynaMAD algorithm. The CA–DynaMAD algorithm iteratively increases the dimensionality of reference spaces in a controlled manner by evaluating a single molecular property descriptor per iteration. The aim of this study was to evaluate whether the CA–DynaMAD algorithm could be used to generate focused reference spaces facilitating the identification of active or selective compounds.

As could be shown in comparative VS calculations with DynaMAD, the improved descriptor scoring and dimension extension function of CA–DynaMAD led to an increase in compound recall while reducing the number of descriptor variables. Going beyond virtual screening, CA–DynaMAD was successfully applied to design target–selective reference spaces for antagonists of seven biogenic amine GPCRs. In a first step, the algorithm was applied only to the database of selective GPCR antagonists. For all seven antagonist classes, target–selective reference spaces were designed such that nearly all antagonists could be correctly classified. Subsequently, these reference spaces were used to successfully identify selective compounds from a large biologically annotated database.

The design of appropriate chemical reference spaces is a key requirement for many applications in chemoinformatics and computer–aided medicinal chemistry. Although the computational analysis and prediction of target selectivity is more complicated than distinguishing active from inactive compounds, highly resolved target–selective spaces that effectively deselected database compounds were successfully derived and used to identify other target–selective GPCR antagonists. Therefore, the generation of target–selective reference spaces using CA–DynaMAD complements and extends currently available approaches to chemical space design. The utility of CA–DynaMAD to identify individual descriptors that discriminate between antagonists with different selectivity has also been demonstrated. Such descriptors can be used in many other computational applications, for example, QSAR analysis. Taken together, these findings suggest that the design of "selectivity spaces" should merit further investigation and have significant potential for practical applications.

Bibliography

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

Alaimo, P. J.; Shogren-Knaak, M. A.; Shokat, K. M. Chemical genetic approaches for the elucidation of signaling pathways. *Curr. Opin. Chem. Biol.* **2001** 5, 360–367.

Bajorath, J. Rational drug discovery revisited: Interfacing experimental programs with bio– and chemo–informatics. *Drug Discov. Today* **2001**a 6, 989–995.

Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Chem. Inf. Comput. Sci.* **2001**b 41, 233–245.

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

Bajorath, J. Computational analysis of ligand relationships within target families *Curr. Opin. Chem. Biol.* **2008** 12, 352–358.

Balaban, A. T. Highly discriminating distance–based topological index. *Chem. Phys. Lett.* **1982** 89, 399–404.

Barnard, J.; Downs, G. Chemical fragment generation and clustering software. *Chem. Inf. Comput. Sci.* **1997** 37, 141–142.

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

Bender, A.; Mussa, H.; Glen, R.; Reiling, S. Molecular similarity searching using atom environments, information–based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**a 44, 170–178.

Bender, A.; Mussa, H.; Glen, R.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**b 44, 1708–1718.

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

Bradley, E.; Beroza, P.; Penzotti, J.; Grootenhuis, P.; Spellmeyer, D.; Miller, J. A rapid computational method for lead evolution: description and application to α_1–adrenergic antagonists. *J. Med. Chem.* **2000** 43, 2770–2774.

Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Rev. Genet.* **2004** 5, 262–275.

Brown, R.; Martin, Y. 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.

Cheng, Y.-C.; Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **1973** 22, 3099–3108.

Clark, M.; Cramer, R. D.; Opdenbosch, N. V. Validation of the general purpose Tripos 5. 2 force field. *J. Comput. Chem.* **1989** 10, 982–1012.

Crisman, T. J.; Jenkins, J. L.; Parker, C. N.; Hill, W. A. G.; Bender, A.; Deng, Z.; Nettles, J. H.; Davies, J. W.; Glick, M. ”Plate Cherry Picking”: A novel semi–sequential screening paradigm for cheaper, faster, information–rich compound selection. *J. Biomol. Screen.* **2007** 12, 320–327.

DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003** 22, 151–185.

Eckert, H.; Bajorath, J. Design and evaluation of a novel class–directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.* **2006**a 46, 2515–2526.

Eckert, H.; Bajorath, J. Determination and mapping of activity–specific descriptor value ranges for the identification of active compounds. *J. Med. Chem.* **2006**b 49, 2284–2293.

Eckert, H.; Vogt, I.; Bajorath, J. Mapping algorithms for molecular similarity analysis and ligand–based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006** 46, 1623–1634.

Ehrlich, P.; Bertheim, A. Über das salzsaure 3,3'–Diamino–4,4'–dioxyarsenobenzol und seine nächsten Verwandten. *Ber. Dtsch. Chem. Ges.* **1912** 45, 756–766.

Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure–activity relationship paradigm. *Methods Mol. Biol.* **2004** 275, 131–214.

Flower, D. On the properties of bit string–based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998** 38, 379–386.

Friedman, J. H. A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Comput.* **1977** 26, 404–408.

Gao, H.; Lajiness, M. S.; Drie, J. V. Enhancement of binary QSAR analysis by a GA–based variable selection method. *J. Mol. Graphics Modell.* **2002** 20, 259–268.

Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges. *Tetrahedron* **1980** 36, 3219–3228.

Godden, J.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE–DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002** 42, 87–93.

Godden, J.; Bajorath, J. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.* **2006** 46, 1094–1097.

Godden, J.; Furr, J.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003** 43, 182–188.

Godden, J.; Furr, J.; Xue, L.; Stahura, F.; Bajorath, J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary–transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**a 44, 21–29.

Godden, J.; Stahura, F.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000** 40, 796–800.

Godden, J.; Stahura, F.; Bajorath, J. POT–DMC: a virtual screening method for the identification of potent hits. *J. Med. Chem.* **2004**b 47, 5608–5611.

Good, A.; Krystek, S.; Mason, J. High–throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov. Today* **2000** 5, 61–69.

Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure–property modeling. *Rev. Comput. Chem.* **1991** 2, 367–422.

Hert, J.; Willett, P.; Wilton, D.; 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.

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

Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of inter–molecular similarity and dissimilarity using 2D fragment bit–strings. *Comb. Chem. High Throughput Screen.* **2002** 5, 155–166.

Irwin, J.; Shoichet, B. ZINC – A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005** 45, 177–182.

James, C. A.; Weininger, D. *Daylight theory manual.* Daylight Chemical Information Systems, Inc., Irvine, CA., **2008**.

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

Johnson, M.; Maggiora, G. M., (Eds.) *Concepts and application of molecular similarity.* John Wiley & Sons, **1990**.

Jones-Hertzog, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. Use of recursive partitioning in the sequential screening of G–protein–coupled receptors. *J. Pharmacol. Toxicol. Methods* **1999** 42, 207–15.

Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004** 303, 1813–1818.

Klabunde, T.; Evers, A. GPCR antitarget modeling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR–mediated side effects. *ChemBioChem* **2005** 6, 876–889.

Klabunde, T.; Hessler, G. Drug design strategies for targeting G–protein–coupled receptors. *ChemBioChem* **2002** 3, 928–944.

Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998** 9–11, 225–252.

Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000** 18, 464–477.

Labute, P. Derivation and applications of molecular descriptors based on approximate surface area. *Methods Mol. Biol.* **2004** 275, 261–278.

Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004** 432, 855–861.

Livingstone The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000** 40, 195–209.

Macarron, R. Critical review of the role of HTS in drug discovery. *Drug Discov. Today* **2006** 11, 277–279.

Maggiora, G. M. On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* **2006** 46, 1535.

Martin, Y.; Kofron, J.; Traphagen, L. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002** 45, 4350–4358.

Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001** 3, 231–250.

McGregor, M.; Pallai, P. Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997** 37, 443–448.

Motulsky, H. J.; Christopoulos, A. *Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting.* GraphPad Software Inc. San Diego, CA **2003**. http://www.graphpad.com.

Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **2006** 5, 993–996.

Pearlman, R.; Smith, K. Metric validation and the receptor–relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999** 39, 28–35.

Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Design* **1998** 9–11, 339–353.

Rusinko, A.; Farmen, M.; Lambert, C.; Brown, P.; Young, S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999** 39, 1017–1026.

Schnur, D.; Beno, B. R.; Good, A.; Tebben, A. Approaches to target class combinatorial library design. *Meth. Mol. Biol.* **2004** 275, 355–378.

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

Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004** 432, 862–865.

Spring, D. R. Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev.* **2005** 34, 472–482.

Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003** 10, 707–715.

Stahura, F. L.; Bajorath, J. New methodologies for ligand–based virtual screening. *Curr. Pharm. Des.* **2005** 11, 1189–1202.

Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An open standard for chemical structure representation: The IUPAC chemical identifier. In *Proceedings of the 2003 International Chemical Information Conference* Collier, H., (Ed.) Infonortics, Tetbury, Nimes, France, **2003** 131–143.

Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004** 432, 846–854.

Stumpfe, D.; Ahmed, H. E. A.; Vogt, I.; Bajorath, J. Methods for computer–aided chemical biology. Part 1: Design of a benchmark system for the evaluation of compound selectivity. *Chem. Biol. Drug Des.* **2007** 70, 182–194.

Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H., (Eds.) *Handbook of molecular descriptors.* Methods and Principles in Medicinal Chemistry Wiley, New York, **2000**.

Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D fingerprint methods for multiple–template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2007** 2, 208–217.

Vogt, I.; Ahmed, H.; Auer, J.; Bajorath, J. Exploring structure–selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Diversity* **2008** 12, 25–40.

Vogt, I.; Bajorath, J. Analysis of a high–throughput screening data set using potency–scaled molecular similarity algorithms. *J. Chem. Inf. Model.* **2007** 47, 367–375.

Vogt, I.; Bajorath, J. Design and exploration of target–selective chemical space representations. *J. Chem. Inf. Model.* **2008** *in press.*

Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **1963** 58, 236–244.

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988** 28, 31–36.

Wildman, S.; Crippen, G. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inform. Comput. Sci.* **1999** 39, 868–873.

Willett, P. Similarity–based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006** 11, 1046–1053.

Willett, P.; Barnard, J.; Downs, G. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998** 38, 983–996.

Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002** 42, 757–764.

Xue, L.; Godden, J.; Stahura, F.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003** 43, 1151–1157.

Xue, L.; Stahura, F.; Godden, J.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001** 41, 746–753.

Zhang, H.; Yu, C.-Y.; Singer, B.; Xiong, M. Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Nat. Acad. Sci.* **2001** 98, 6730–6735.

Symbols and Abbreviations

μ	Arithmetic mean
σ	Standard deviation
5HT1a/2a	5–hydroxytryptamine (serotonin) receptors 1a/2a
ADME	Absorption, distribution, metabolism, and excretion
Alpha1	Adrenaline receptor alpha1
CA–DynaMAD	Continuous adaptive DynaMAD
Cat B, L, S, K	Cathepsin B, L, S, and K
D1-4	Dopamine receptors 1-4
DACCS	Distance in activity–centered chemical space
DEL	Dimension extension level
Descr	Descriptors
DMC	Dynamic mapping of consensus positions
DynaMAD	Dynamic mapping of activity–selective descriptor value ranges
ESS	Error sum of squares
FSC	Family–selective compounds
GPCR	G Protein–coupled receptor
HR	Hit rate
HTS	High–throughput screening
IC_{50}	Inhibition concentration 50%, concentration of ligand that decreases enzyme activity by 50%
InChI	IUPAC international chemical identifier
JP	Jarvis–Patrick clustering
K_i	Equilibrium dissociation constant, concentration of unlabeled ligand at which it binds to half of the binding sites at equilibrium in absence of competitors
LBVS	Ligand–based virtual screening
MACCS	Molecular access system
maxHR	Maximum possible hit rate

MDDR	MDL drug data report
MOE	Molecular Operating Environment
MP–MFP	Median partitioning minifingerprint
PAC	Pair–active compounds
QSAR	Quantitative structure–activity relationship
RP	Recursive partitioning
RR	Recovery rate
SAR	Structure–activity relationship
SF	Scaling factor
SMILES	Simplified molecular input line entry specification
Tc	Tanimoto coefficient
TGT	Typed graph triangle
TSC	Target–selective compounds
VS	Virtual screening
WA	Ward's clustering

Appendices

Appendix A

Details on Molecular Representations

A.1 Molecular Property Descriptors

The 1D and 2D molecular property descriptors used in the presented studies were calculated using the Molecular Operating Environment (MOE). These 155 descriptors can be divided into seven classes each of which contains between 12 and 33 descriptors as reported in *Table A.1.* The *adjacency and distance matrix descriptors* are calculated from the connection table of a molecule and include Balaban's topological connectivity index (Balaban [1982]) and BCUT–type descriptors (Pearlman and Smith [1998]) for PEOE partial charges (Gasteiger and Marsili [1980]), SlogP or molar refractivity (Wildman and Crippen [1999]). The

Type	**Number of descriptors**
Adjacency and distance matrix descriptors	33
Atom and bond counts	33
Partial charge descriptors	30
Subdivided surface area descriptors	18
Physicochemical properties	13
Connectivity and shape indices	16
Molecular surface descriptors	12

Table A.1: Subset of 155 1D and 2D molecular property descriptors implemented in MOE.

occurrence of specific atom or bond types such as the number of rotatable bonds is monitored by a set of *atom and bond counts.* Based on the method of partial equalization of orbital electronegativities (PEOE) (Gasteiger and Marsili [1980]), the *partial charge descriptors* like the total hydrophobic van der Waals surface area account for the partial charge of each atom. *Physicochemical property descriptors* include molecular density, molecular weight and logP(o/w). *Subdivided surface area descriptors* as introduced by Labute [2004] combine physicochemical properties with the approximate accessible van der Waals surface area. The *connectivity and shape indices* from Kier & Hall (Hall and Kier [1991]) are topological descriptors intended to model different aspects of molecular shape. Finally, *molecular surface descriptors* capture contributions of different atom types like hydrogen bond donors or hydrogen bond acceptors to the van der Waals surface area.

A.2 MDL MACCS keys

Developed for the purpose of substructure searching, the MDL MACCS keys were introduced in 1979 and consist of 960 proprietary and 166 publicly available keys. In this thesis, a fingerprint based on the public key set was used for similarity searching and clustering. For all experiments presented in this thesis, MACCS fingerprints have been calculated with MOE. *Table A.2* lists all 166 keys as defined and documented in MOE.

Key	Description
1	isotopes
2	atoms with atomic number $>$ 103
3	group IVA, VA and VIA periods 4–6
4	Actinides
5	group IIIB, IVB elements
6	Lanthanides
7	group VB, VIB, VIIB elements
8	heteroatoms in 4–membered rings
9	group VIIIB elements
10	alkaline earth elements
11	atoms in 4–membered ring
12	group IB, IIB elements
13	N connected to 1 O and 2 C
14	S atoms in S-S groups
15	C connected to 3 O
16	heteroatoms in 3–membered rings
17	C in CC triple bonds
18	group IIIA elements
19	atoms in 7 ring
20	silicon atoms
21	C = bonded to C and 3 heavy atoms
22	atoms in 3 ring
23	C bonded 1 N and 2 O
24	O-N single bonds
25	C bonded to at least 3 N atoms
26	C in 3 ring bonds and a double bond
27	iodine atoms
28	XCH2X, where X<>C
29	phosphorous atoms
30	non–C Q4 bonded to $\geq$ 3 C
31	halogens connected to non–carbons
32	S bonded to an N and a C
33	S atoms bonded to N
34	CH2= units
35	alkali (group IA) elements
36	S atoms in rings
37	C bonded to $\geq$ 1 O & $\geq$2 N
38	C bonded $\geq$ 2 N and 1 C
39	S atoms bonded to 3 O
40	S single bonded to OQ2
41	N in CN
42	fluorine atoms
43	X-H heteroatoms 2 bonds from another
44	other elements
45	N atoms adjacent to -C=C
46	bromine atoms
47	S two bonds from an N
48	non–C bonded to $\geq$ 3 O
49	charged atoms
50	C in C=C bonded to $\geq$ 3 C
51	S bonded to a C and an O
52	N bonded to N
53	QH 4 bonds from another QH
54	QH 3 bonds from another QH
55	S bonded to $\geq$2 O
56	N bonded to $\geq$ 2 O and $\geq$ 1 C
57	O in rings
58	S bonded to $\geq$2 non–carbon atoms
59	non–aromatic S-[a]
60	[S+]-[O-]
61	SQ3
62	non–ring bonds that connect rings
63	N atoms in double bonds with O
64	non–ring S attached to a ring
65	N in aromatic bonds with C
66	CX4 bonded to $\geq$3 carbons
67	S attached to heteroatoms
68	QH bonded to another QH
69	QH bonded to another Q
70	N bonded to two non–C heavy atoms

Table A.2: **Publicly available set of 166 MACCS keys.**

Key	Description
71	N bonded to O
72	O separated by 3 bonds
73	S in double/charge separated bonds
74	dimethyl substituted atoms
75	N non–ring bonded to a ring
76	C in C=C bonded to $\geq$ 3 heavy atoms
77	N separated by 2 bonds
78	N double bonded to C
79	N separated by 3 bonds
80	N separated by 4 bonds
81	S attached to Q $\geq$ 3 atoms
82	heteratoms attached to a CH2
83	heteroatoms in 5 ring
84	NH2 groups
85	N bonded to $\geq$ 3 C
86	CH2 or CH3 separated by non–C
87	halogens bonded to any ring
88	sulfurs
89	O separated by 4 bonds
90	het. 3 bonds from a CH2
91	het. 4 bonds from a CH2
92	C bonded to $\geq$1 N, $\geq$1 C & $\geq$ 1 O
93	methylated heteroatoms
94	N bonded to non–C
95	O 3 bonds from an N
96	atoms in 5–rings
97	O 4 bonds from an N
98	het. in 6–ring
99	C in C=C
100	N attached to CH2
101	atoms in 8–ring or higher
102	O bonded to non–C heavy atoms
103	chlorine atoms
104	hets. 2 bonds from a CH2
105	hets. ring bonded to a 3–ring bond X
106	X bonded to $\geq$ 3 non–C
107	XQ>3 bonded to at least 1 halogen
108	CH3 4 bonds from a CH2
109	O attached to CH2
110	O 1 C from an N
111	N 2 bonds from a CH2
112	atoms with coordination number $\geq$ 4
113	O in non–aromatic bonds to an [a]
114	CH3 attached to CH2
115	CH3 2 bonds from a CH2
116	CH3 3 bonds from a CH2
117	N 2 bonds from an O
118	(key(147)-1 if key(147)>1; else 0)
119	N in double bonds
120	(key(137)-1 if key(137)>1; else 0)
121	N in rings
122	N with coordination number $\geq$3
123	O separated by 1 C
124	het-het bonds
125	Is AROMATIC RING > 1?
126	non–ring O bonded to 2 heavy atoms
127	(key(143)-1 if key(143)>1; else 0)
128	CH2s separated by 4 bonds
129	CH2s separated by 3 bonds
130	(key(124)-1 if key(124)>1; else 0)
131	(het atoms with H)

Table A.2: Continued.

Key	Description
132	O 2 bonds from CH2
133	N non–ring bonded to a ring
134	halogens
135	N in a non–aromatic bond with [a]
136	Bit: is there more than 1 O=
137	Total ring HETEROCYCLE atoms
138	(key(153)-1 if key(153)>1; else 0)
139	OH groups
140	(key(164)-3 if key(164)>3; else 0)
141	(key(160)-2 if key(160)>2; else 0)
142	(key(161)-2 if key(161)>1; else 0)
143	non–ring O connected to a ring
144	atoms separated by (!:):(!:)
145	6M RING > 1
146	Key(164)-2 if key(164)>2; else 0
147	CH2 attached to CH2
148	non–C with coordination number ≥3
149	(key(160)-1 if key(160)>1; else 0)
150	X separated by (!r)-r-(!r)
151	NH
152	C bonded to ≥2 C and 1 O
153	non–carbons attached to CH2
154	O in C=O
155	non–ring CH2
156	XN where coord. of X≥3
157	O in C-O single bonds
158	N in C-N single bonds
159	Key(164)-1 if key(164)>1; else 0
160	CH3 groups
161	N
162	aromatics
163	atoms in 6 rings
164	oxygens
165	ring atoms
166	Is there more than 1 fragment?

Table A.2: Continued.

Appendix B

Compound Databases

B.1 Classes of Selective Compounds

B.1.1 Biogenic Amine GPCR Antagonists

In this thesis, two molecular benchmark systems consisting of antagonists of biogenic amine binding G protein–coupled receptors (GPCRs) are used to study compounds with different selectivity profiles and evaluate 2D similarity methods for their potential to detect selectivity differences between these closely related molecules.

GPCRs form a large protein superfamily that plays an important role in many physiological and patho–physiological processes and, currently, presents the largest group of pharmaceutical targets (Klabunde and Hessler [2002]; Overington et al. [2006]). In particular, the subfamily of biogenic amine binding GPCRs has provided attractive drug targets for the treatment of various neurological and cardiological diseases (Overington et al. [2006]). In computer–aided drug discovery, these GPCRs have traditionally been major targets for ligand design based on pharmacophore analysis (Klabunde and Evers [2005]). As summarized in *Table B.1*, targeted receptors belong to three major amine receptor subfamilies, the dopaminergic, serotonergic, and adrenergic GPCR subfamilies. The databases contain only competitive and reversible antagonists or partial agonists and generally share a canonical molecular organization where arylpiperazine or piperidine moieties are connected through alkyl or alkenyl spacers to heteroaromatic systems (Stumpfe et al. [2007]). Variations in these structural motifs are responsible for differences in ligand selectivity. For computational approaches, distinguishing between compounds having limited structural diversity, but significant differences

Superfamily	**Family**	**Subfamily**	**Targets**
Rhodopsin–like GPCRs	Biogenic amine GPCRs	Adrenergic	Alpha1
		Dopaminergic	D1, D2, D3, D4
		Serotonergic	5HT1a, 5HT2a

Table B.1: **Biogenic amine G protein–coupled receptor.**

in selectivity is thought to be a challenging task. The two databases of selective biogenic amine GPCR antagonists, GPCR$_{pw}$ and GPCR$_{f}$, differ in the way how selectivity profiles are captured and how individual selectivity sets are constituted.

GPCR$_{pw}$ This database was introduced by Stumpfe et al. [2007] and contains 248 compounds with antagonistic activity against at least two of six targets belonging to the family of biogenic amine GPCRs. Based on differences in their potency against multiple receptors, these compounds are grouped into 12 classes of antagonists with selectivity for one receptor over another one. Thus, the GPCR$_{pw}$ permits to study differential selectivity considering two targets at once on the basis of pairwise selectivity sets. A compound is considered selective for a given target if its K$_i$ or IC$_{50}$ value is at least 50–fold lower than for targets within the same subfamily. As shown in *Table B.2*, sets from this database contain between nine and 64 compounds, spanning a considerable range in target selectivity. The intra– and inter–set MACCS Tc similarity is visualized in *Figure B.1* and the analysis reveals a generally higher degree of similarity compared to the protease inhibitor sets described further below. However, comparison of selectivity sets for target pairs shows rather different diversity patterns (for example, *Figures B.1a* and *B.1d*). Similarly, the scaffold diversity measured by the ratio of chemical scaffolds over the total number of compounds per set varies significantly.

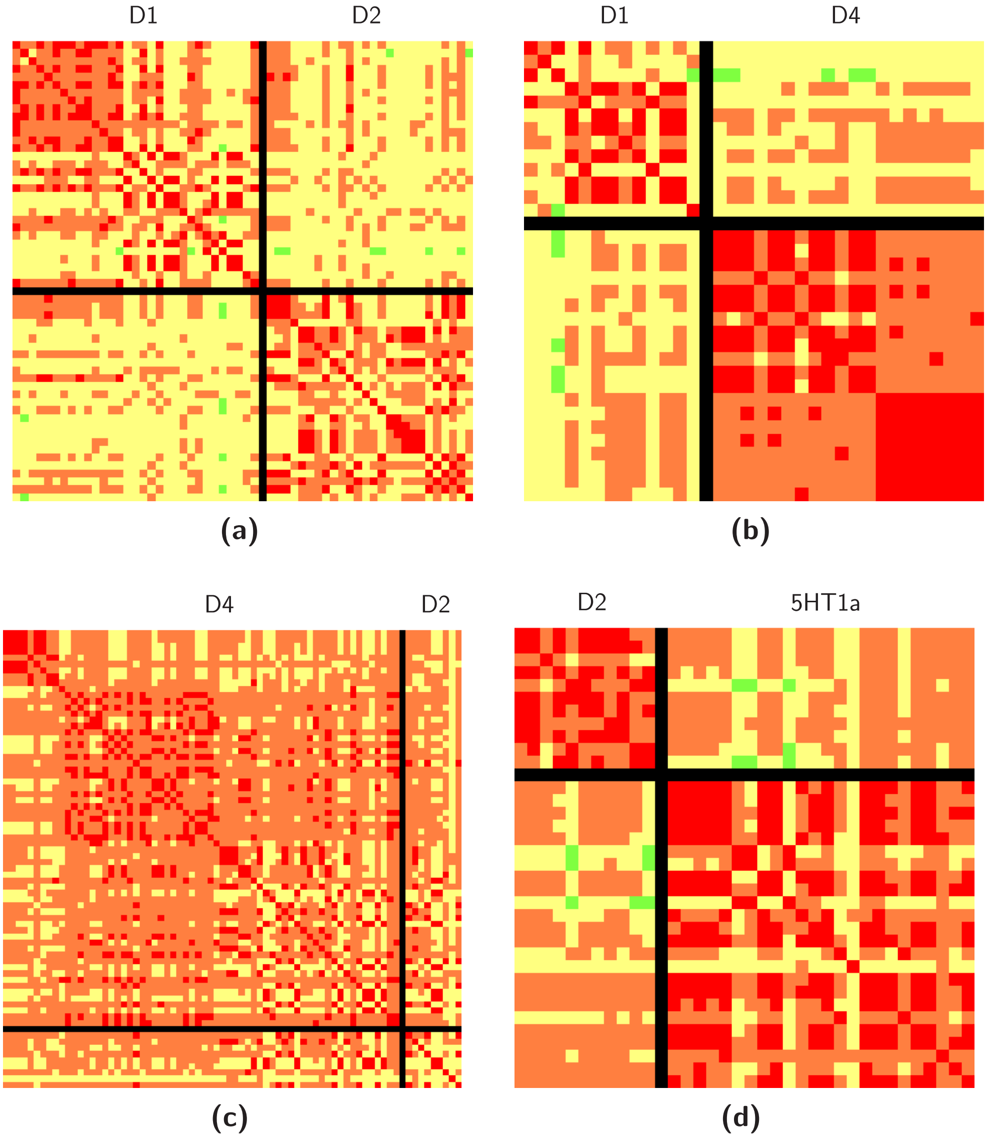
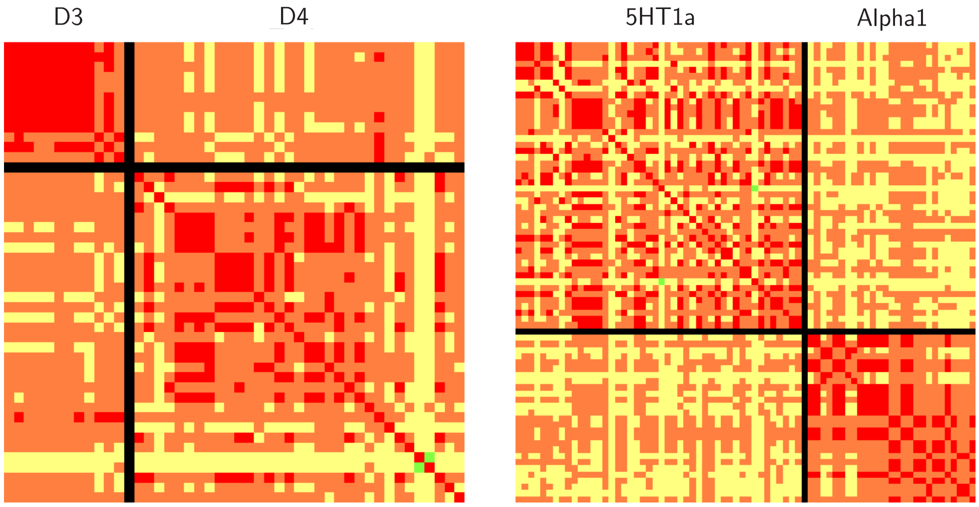


Figure B.1: MACCS Tc analysis of GPCR$_{pw}$ selectivity sets. MACCS Tc values are reported in matrix format for systematic intra–set and inter–set pairwise compound comparisons in (a) - (f). The values are color–coded according to following scheme: dark green 0-0.2, light green 0.2-0.4, yellow 0.4-0.6, orange 0.6-0.8, red 0.8-1. Compound sets with "bidirectional" selectivity are compared. For example, the D1 versus D2 matrix in (a) compares D1 compounds selective over D2 and vice versa.



(e)

(f)

Figure B.1: Continued.

Selective for	Over	Size	Selectivity	Scaffold Diversity
5HT1a	Alpha1	46	55 - 10 000	0.70
	D2	24	58 - 10 000	0.88
Alpha1	5HT1a	27	59 - 20 000	0.67
D1	D2	31	59 - 10 084	0.55
	D4	13	65 - 4 761	0.39
D2	D1	26	67 - 18 310	0.54
	D4	9	50 - 834	0.33
	5HT1a	11	50 - 1 628	0.27
D3	D4	12	51 - 1 609	0.92
D4	D1	20	190 - 30 769	0.45
	D2	64	59 - 17 600	0.33
	D3	33	52 - 15 000	0.64

Table B.2: Pairwise selective antagonist classes in GPCR$_{pw}$. *Selectivity* gives the range of potency ratios that are a quantitative measure of selectivity for the corresponding target. *Scaffold Diversity* reports the ratio of chemical scaffolds over the total number of compounds per set. Chemical scaffolds were calculated by deleting all non–ring substituents except linkers between ring systems.

GPCR$_f$ The second biogenic amine GPCR database extends GPCR$_{pw}$ and shifts the analysis of differential compound selectivity profiles from related target–pairs to an entire target family. It contains 267 selective antagonists distributed among seven receptor selectivity sets. All compounds have at least 50-fold higher potency for one biogenic amine GPCR over one or more others. *Table B.3* shows that antagonists in the selectivity sets cover large selectivity ranges, spanning several orders of magnitude. *Figure B.2* illustrates a part of the structural spectrum of antagonists in GPCR$_f$ with different selectivity. The structures of antagonists displaying different selectivity profiles are often topologically similar but each of the selectivity sets contains compounds representing multiple scaffolds. There are differences in molecular size and rigidity of chemotypes within and between selectivity sets. Intra– and inter–set structural similarity is generally comparable, as revealed by systematic pairwise MACCS Tc compound comparisons reported in *Table B.4*, supplemented by *Figure B.3*. Overall, the antagonists represent a rather continuous structural spectrum with varying and in part overlapping selectivity relationships. These characteristics make it generally difficult to systematically distinguish between compounds having different selectivity and hence these selectivity sets present difficult test cases for differentiating the underlying structure–selectivity relationships.

Selective for	Over	Size	Selectivity	Scaffold diversity
5HT1a	5HT2a, Alpha1, D2	53	65 - 476 190	0.53
5HT2a	5HT1a, Alpha1, D1, D2	21	67 - 49 122	0.38
Alpha1	5HT1a, 5HT2a, D1, D2, D3, D4	26	59 - 62 500	0.81
D1	D2, D4	33	55 - 10 084	0.42
D2	5HT1a, D1, D3, D4	25	73 - 18 310	0.36
D3	D1, D2, D4	37	51 - 17 600	0.70
D4	Alpha1, D1, D2, D3	72	54 - 28 000	0.38

Table B.3: Selective GPCR antagonist classes in GPCR$_f$. *Selectivity* gives the range of potency ratios that are a quantitative measure of selectivity for the corresponding target. *Scaffold Diversity* reports the ratio of chemical scaffolds over the total number of compounds per set. Chemical scaffolds were calculated by deleting all non–ring substituents except linkers between ring systems.

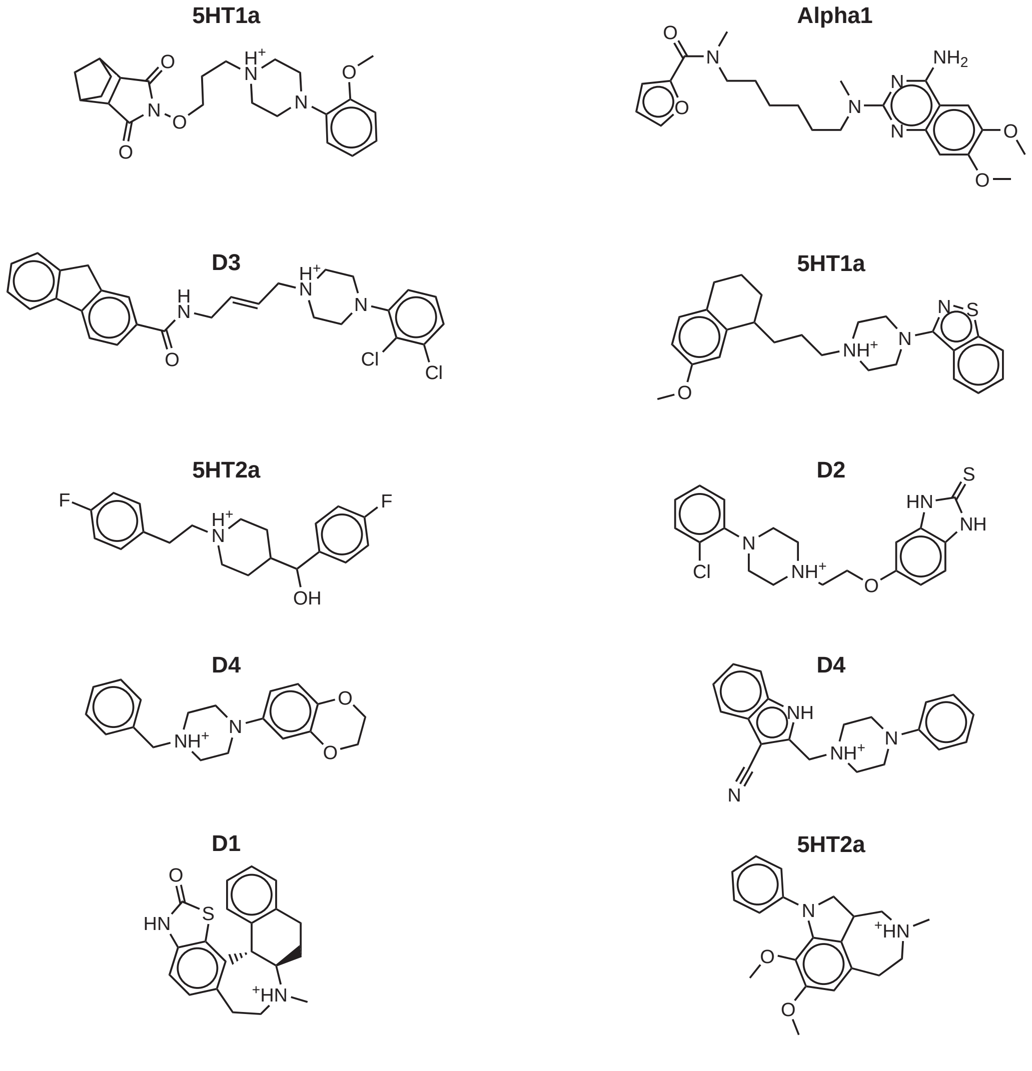


Figure B.2: Representative structures of selective GPCR antagonists from GPCR$_{pw}$. At the top left and right, the most selective 5HT1a and Alpha1 antagonists are shown, respectively. In addition, antagonists from other selectivity sets with varying degrees of similarity to these highly selective compounds are shown. This representative overview illustrates a part of the structural spectrum of GPCR antagonists with different selectivity.

Selectivity set		5HT1a	5HT2a	Alpha1	D1	D2	D3	D4
5HT1a	*Min*	**0.39**	0.35	0.36	0.28	0.34	0.37	0.34
	Max	**1.00**	0.82	1.00	0.74	0.80	0.85	0.97
	Mean	**0.64**	0.53	0.58	0.51	0.58	0.53	0.58
5HT2a	*Min*	0.35	**0.37**	0.34	0.29	0.37	0.33	0.34
	Max	0.82	**1.00**	0.85	0.72	0.78	0.76	0.79
	Mean	0.53	**0.59**	0.52	0.48	0.54	0.49	0.53
Alpha1	*Min*	0.36	0.34	**0.40**	0.32	0.39	0.34	0.40
	Max	1.00	0.85	**1.00**	0.80	0.84	0.85	0.89
	Mean	0.58	0.52	**0.67**	0.56	0.55	0.58	0.59
D1	*Min*	0.28	0.29	0.32	**0.31**	0.33	0.31	0.30
	Max	0.74	0.72	0.80	**1.00**	0.81	0.70	0.71
	Mean	0.51	0.48	0.56	**0.64**	0.50	0.50	0.52
D2	*Min*	0.34	0.37	0.39	0.33	**0.40**	0.37	0.37
	Max	0.80	0.78	0.84	0.81	**1.00**	0.77	0.82
	Mean	0.58	0.54	0.55	0.50	**0.62**	0.50	0.55
D3	*Min*	0.37	0.34	0.40	0.30	0.37	**0.38**	0.34
	Max	0.85	0.79	0.89	0.71	0.82	**1.00**	0.83
	Mean	0.53	0.53	0.59	0.52	0.55	**0.61**	0.54
D4	*Min*	0.34	0.34	0.40	0.30	0.37	0.34	**0.38**
	Max	0.97	0.79	0.89	0.71	0.82	0.83	**1.00**
	Mean	0.58	0.53	0.59	0.52	0.55	0.54	**0.63**

Table B.4: Intra– and inter–set compound similarity of GPCR$_f$ set measured by pairwise MACCS Tc values. Intra–set similarity values are given in bold.

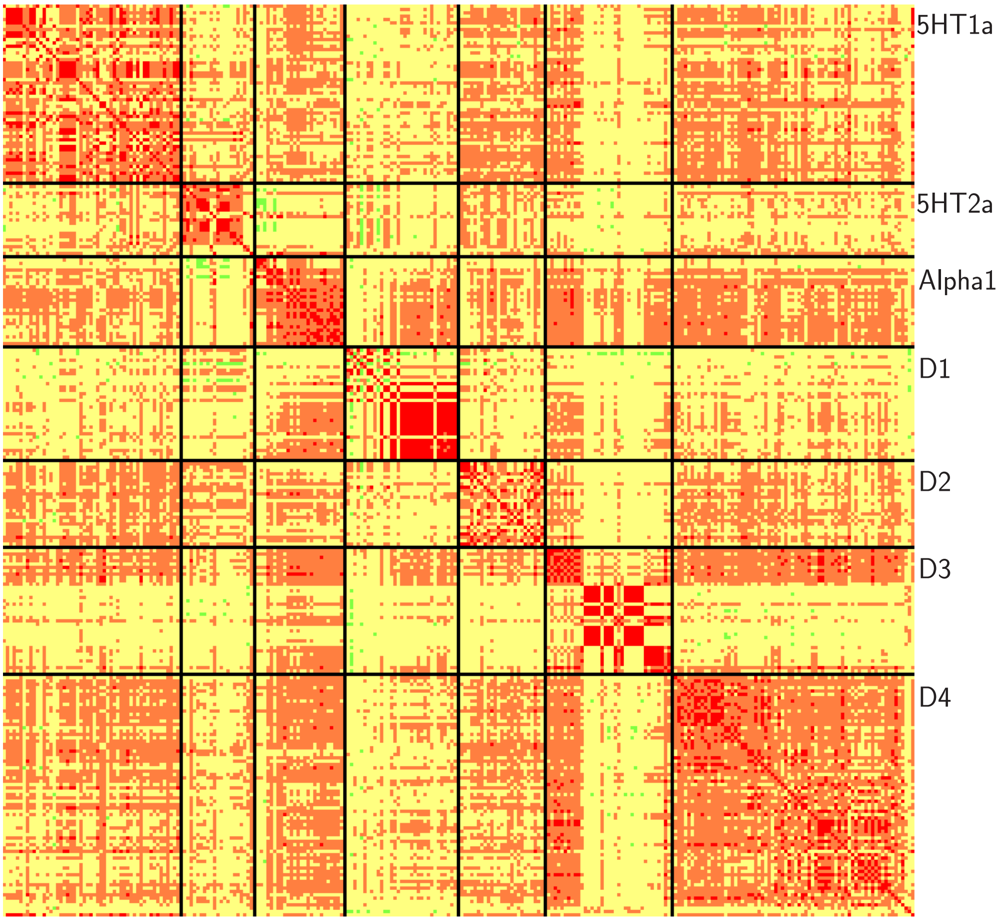


Figure B.3: MACCS Tc matrix of GPCR$_f$. MACCS Tc values are reported in matrix format for systematic intra–set and inter–set pairwise compound comparisons. The values are color–coded according to following scheme: dark green 0-0.2, light green 0.2-0.4, yellow 0.4-0.6, orange 0.6-0.8, red 0.8-1.

B.1.2 Selective Cathepsin Inhibitors

For four members belonging to the C1 papain–like family, the lysosomal cysteine proteases cathepsins B, L, S, and K, a number of active compound classes could be built from publicly available sources. For cathepsins B, L, and S, biological screening data are available from three different HTS campaigns carried out at the Penn Center for Molecular Discovery at the University of Pennsylvania[1], made public through PubChem[2]. In addition, 45 selective inhibitors for cathepsin S and K were gathered from literature and database sources as described in Stumpfe et al. [2007].

Superfamily	Family	Subfamily	Targets
Papain–like thiol proteases	C1 Papain	Cat B–like	Cat B
		Cat L–like	Cat L, S, K

Table B.5: **Lysosomal cysteine proteases.**

The cathepsin HTS data set used for the study of selectivity profiles was obtained by merging the results of the three screens for inhibitors of cathepsins B, L and S, respectively. As shown in *Table B.9*, the intersection of these three different assays consists of 55 134 compounds including 79 hits with confirmed IC_{50} below 45 μM against only one of the targets. *Figure B.4* enables the assessment of pairwise–similarity among confirmed HTS hits and inactives. It can clearly be seen that the similarity between active compounds is comparable to the similarity between actives and inactives, as would be expected from HTS results. As reported in *Table B.6*, the scaffold diversity is overall greater than for the prior described GPCR classes.

As shown in *Figure B.5*, inter–set similarity of cathepsin HTS hits is quite heterogeneous. Intra–set similarity of cathepsin B and L is overall lower than for cathepsin S HTS hits. Comparison of compound sets gathered from literature sources shows different patterns of slightly higher similarity (see *Figure B.5b*). Overall, the cathepsin L set is structurally most diverse.

[1] The Penn Center for Molecular Discovery (PCMD).
`http://www.seas.upenn.edu/~pcmd/`

[2] PubChem BioAssay Summaries AID 453, 460, and 501
`http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=453`
`http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=460`
`http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=501`

Selective for	Over	Size	Selectivity	Scaffold Diversity
Cat B	Cat L	23	HTS specific	0.70
	Cat S	23	HTS specific	0.70
Cat L	Cat B	33	HTS specific	0.91
	Cat S	33	HTS specific	0.91
Cat S	Cat B	23	HTS specific	0.78
	Cat L	23	HTS specific	0.78
	Cat K	20	65 - 119 298	0.95
Cat K	Cat S	25	50 - 1 000	0.88

Table B.6: Pairwise–selective cathepsin sets. *Selectivity* reports potency ratios for compounds from the literature. *HTS specific* means that active compounds come from HTS data and showed measurable activity against only one target. *Scaffold diversity* reports the ratio of chemical scaffolds over the total number of compounds per set.

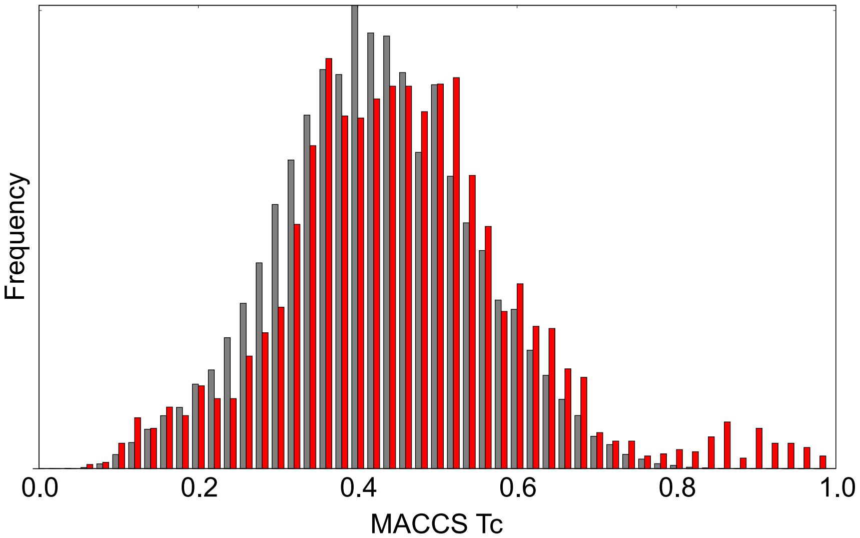


Figure B.4: Cat BLS HTS Tc similarity. Histogram of pairwise Tc similarity based on MACCS keys fingerprint. Number of occurrences of MACCS Tc values among actives and between actives and inactives are colored in red and gray, respectively.

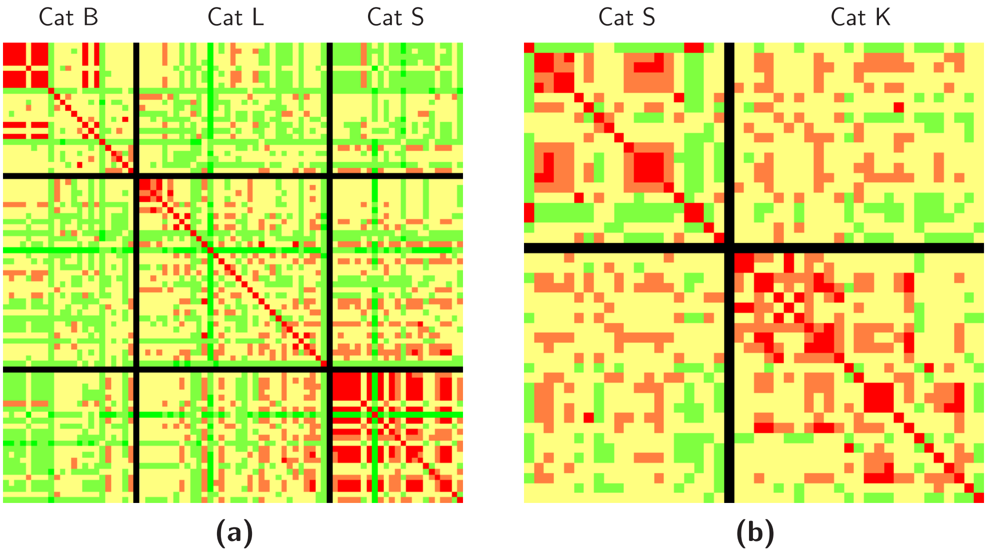


Figure B.5: MACCS Tc matrices of cysteine protease inhibitors. MACCS Tc values are reported in matrix format for systematic intra–set and inter–set pairwise compound comparisons in (a) and (b). The values are color–coded according to following scheme: dark green 0-0.2, light green 0.2-0.4, yellow 0.4-0.6, orange 0.6-0.8, red 0.8-1.

B.1.3 Serine Inhibitor Classes

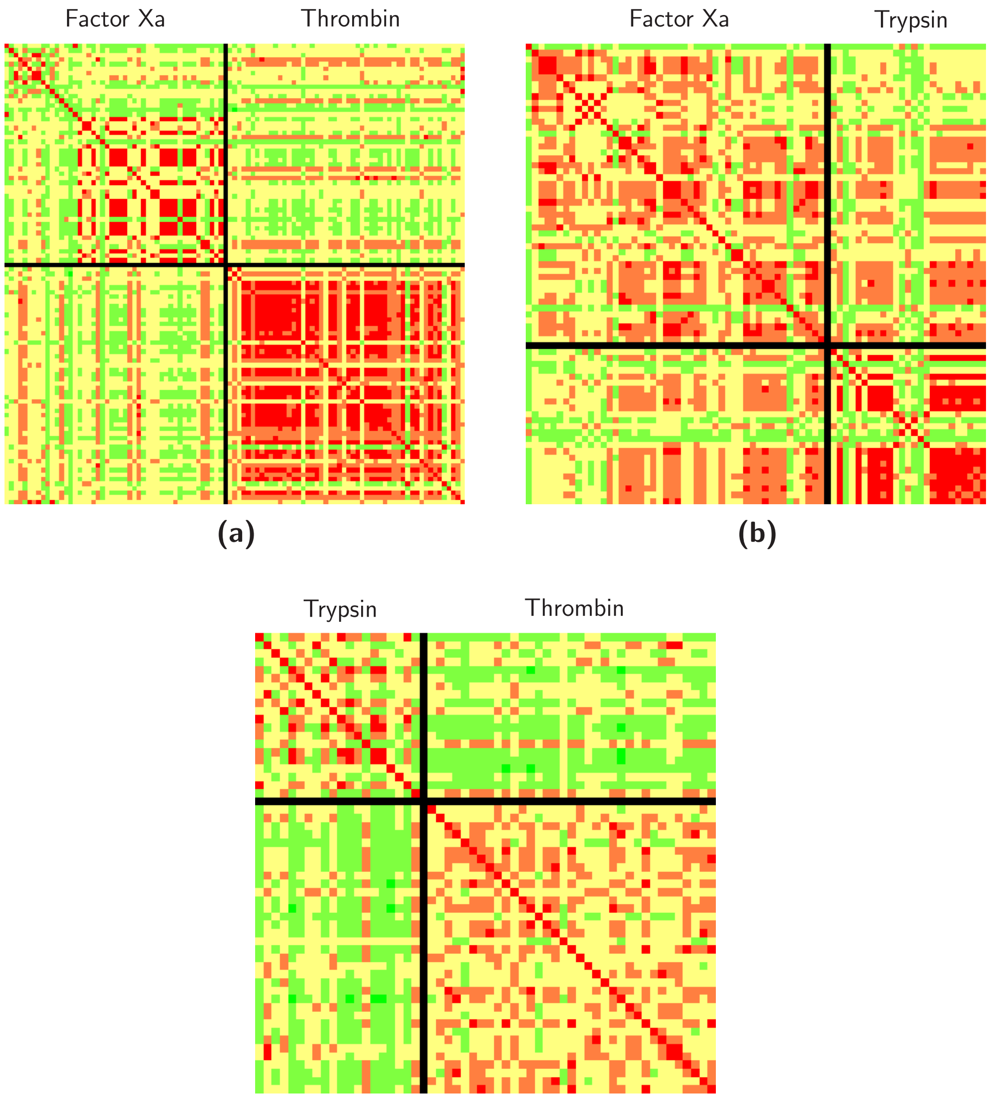
As shown in *Table B.7*, the second database of selective protease inhibitors targets three members of the S1 chymotrypsin family, namely trypsin, thrombin, and factor Xa. The six pairwise selectivity sets contain between 20 to 52 compounds with significant selectivity for one target over another one and originate from a considerable variety of chemical scaffolds (*Table B.8*). The significantly different diversity patterns for selectivity sets and their comparison in *Figure B.6* demonstrate that there is no obvious correlation between structural similarity or diversity and compound selectivity. Both thrombin and trypsin inhibitors that are selective over factor Xa have significant intra–set similarity. By contrast, thrombin inhibitors selective over trypsin are structurally heterogeneous. Structurally similar compounds, but also structurally diverse compounds, are found to be target–selective and compounds having opposite selectivity are often related by different degrees of structural diversity.

Superfamily	**Family**	**Subfamily**	**Targets**
Chymotrypsin–like serine proteases	S1 Chymotrypsin	Trypsin–like	Thrombin, Trypsin, Factor Xa

Table B.7: **Trypsin–like serine proteases.**

Selective for	**Over**	**Size**	**Selectivity**	**Scaffold Diversity**
Thrombin	Trypsin	35	103 - 122 059	1.00
	Factor Xa	52	100 - 490 000	0.64
Trypsin	Thrombin	20	100 - 11 428	0.65
	Factor Xa	25	100 - 100 000	0.72
Factor Xa	Thrombin	48	104 - 200 000	0.55
	Trypsin	49	100 - 400 000	0.79

Table B.8: Pairwise selective serine inhibitor classes. *Selectivity* gives the range of potency ratios that are a quantitative measure of selectivity for the corresponding target. *Scaffold Diversity* reports the ratio of chemical scaffolds over the total number of compounds per set.



(a)

(b)

(c)

Figure B.6: MACCS Tc matrices of serine protease inhibitors. MACCS Tc values are reported in matrix format for systematic intra–set and inter–set pairwise compound comparisons in (a) - (c). The values are color–coded according to following scheme: dark green 0-0.2, light green 0.2-0.4, yellow 0.4-0.6, orange 0.6-0.8, red 0.8-1.

B.2 Screening Databases

To verify and benchmark methods for ligand–based virtual screening, the availability of an adequate screening database is of crucial importance. Ideally, such a database is biologically annotated, that is, information on biological activity in general or with respect to a given target is available for each compound. However, as such information must be experimentally determined, benchmark experiments aiming at differentiating active from inactive compounds often employ databases of up to several millions of drug–like compounds for which no biological activity information is given, but which are considered inactive. In the following, the screening databases employed for the presented studies are listed, ranging from experimentally confirmed HTS data for selected targets to databases of compounds without or with incomplete activity information.

B.2.1 Inactive HTS Compounds

From the biological screening data of three HTS campaigns for inhibitors of cathepsins B, L, and S, conducted at the Penn Center for Molecular Discovery at the University of Pennsylvania[3], made public through PubChem[4], the confirmed inactive compounds were extracted. As all three campaigns used a largely overlapping screening database, it was possible to generate a database of 55 055 compounds inactive ($IC_{50} \geq 50\mu M$) against all three targets. This database was used for the study of selectivity profiles reported in *Chapter 3.2*. The number of confirmed inactive compounds is reported in *Table B.9*.

Target	No. of compounds
Cathepsin B	63 292
Cathepsin L	57 773
Cathepsin S	61 995
intersection	55 055

Table B.9: Inactive HTS compounds. All compounds in the reported sets have an IC_{50} value above 50 μM against the respective target.

[3] The Penn Center for Molecular Discovery (PCMD).
`http://www.seas.upenn.edu/~pcmd/`

[4] The PubChem Project
`http://pubchem.ncbi.nlm.nih.gov/`

B.2.2 MDL Drug Data Report

The MDL Drug Data Report[5] (MDDR) is a biologically annotated compound database covering the patent literature, journals, and congresses. The herein reported experiments use version 2005.2, containing 161 845 compounds with unique 2D structure with a wide variety of biological targets. According to the biological activity annotation, there are ∼5.9k GPCR antagonists and ∼5.2k GPCR agonists, approximately 3.7% and 3.2% of the entire database, respectively.

B.2.3 ZINC

During the course of the studies reported in this thesis two different subsets of the ZINC database (Irwin and Shoichet [2005]), provided by the Shoichet Laboratory at the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF)[6], were used as databases of assumed inactive compounds for virtual screening experiments. The first database contains 1 442 389 compounds with unique 2D structure from the drug–like set of ZINC6. The second database is a subset of the current version ZINC7 and contains 3 695 953 compounds with unique 2D structure. It was generated by applying a reactivity filter and a broad Lipinski rule filter as defined in *Table B.10*.

Property	Range
Molecular weight	0 - 600
logP	-2 - 6
Hydrogen bond donors	1 - 10
Hydrogen bond acceptors	1 - 10
Rotatable bonds	0 - 18

Table B.10: **Lipinski–like filter applied to the ZINC7 database.**

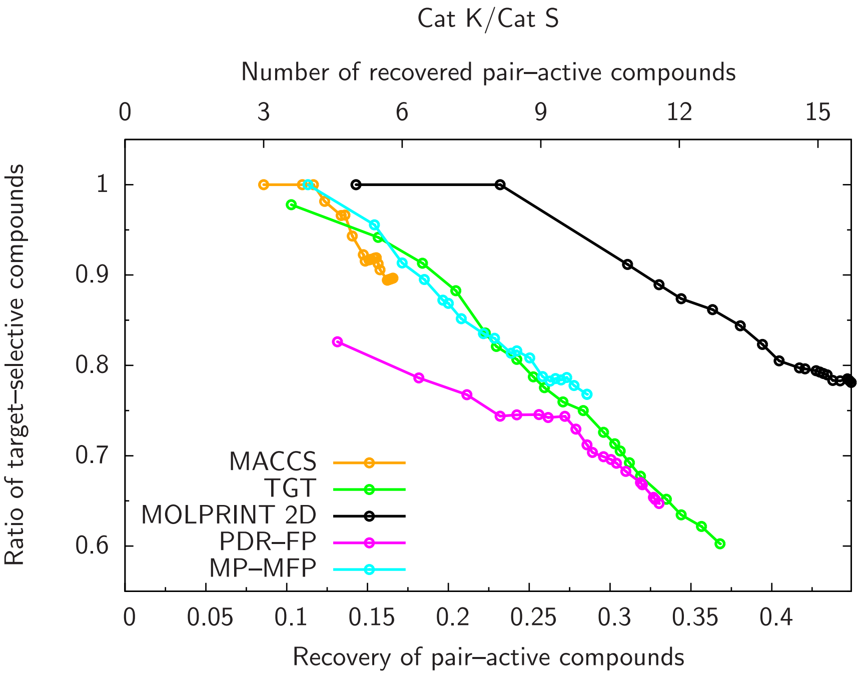
[5] MDL Drug Data Report, Symyx Software: San Ramon, CA.
http://www.mdli.com/products/knowledge/drug_data_report/

[6] ZINC - a free database of commercially available compounds for virtual screening
http://zinc.docking.org/

Appendix C

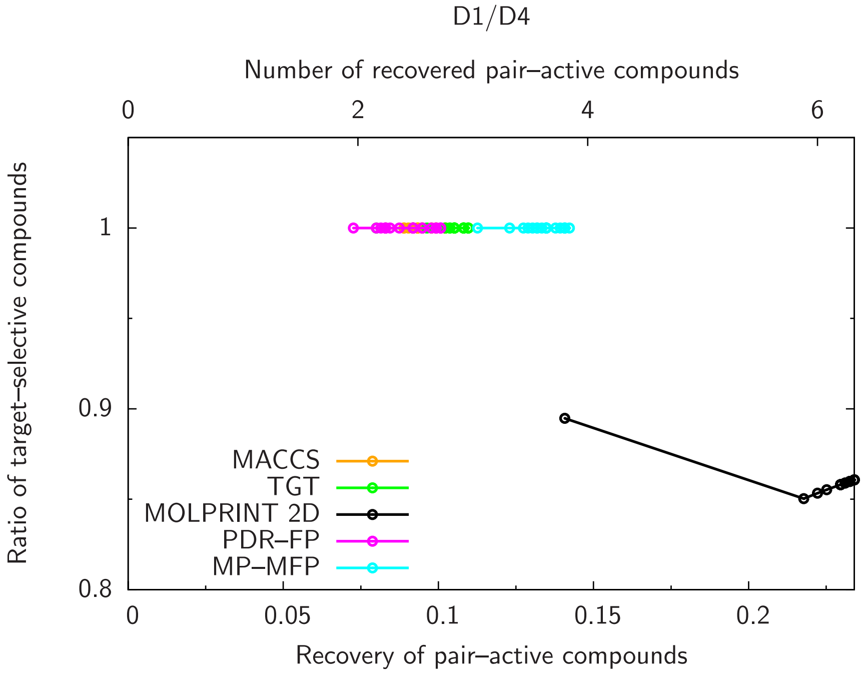
Calculation Data

C.1 Similarity Search Results

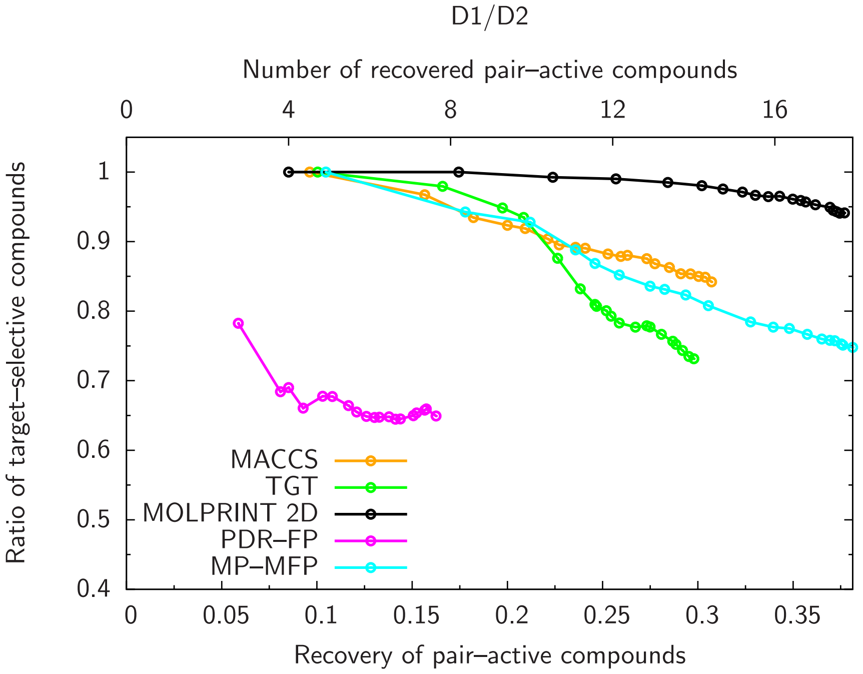


(a)

Figure C.1: Retrieval of selective compounds. In (a)-(l), the average recovery of hits in selection sets of increasing size, ranging from 5 to 100 database compounds, is reported. The leftmost data points correspond to the smallest selection set size of five compounds; then, set sizes increase in increments of five compounds. The graphical representations monitor the retrieval of target–selective versus pair–active compounds, as defined in the text. The total number of recovered pair–active compounds is reported (top horizontal axis) and also the rate (bottom). The vertical axis reports the ratio of target–selective over pair–active compounds.

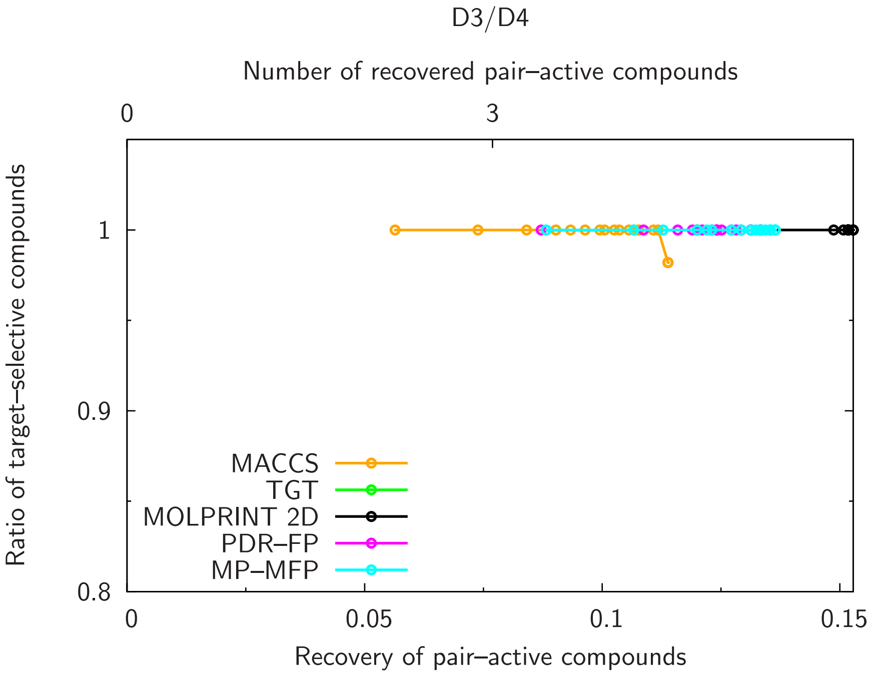


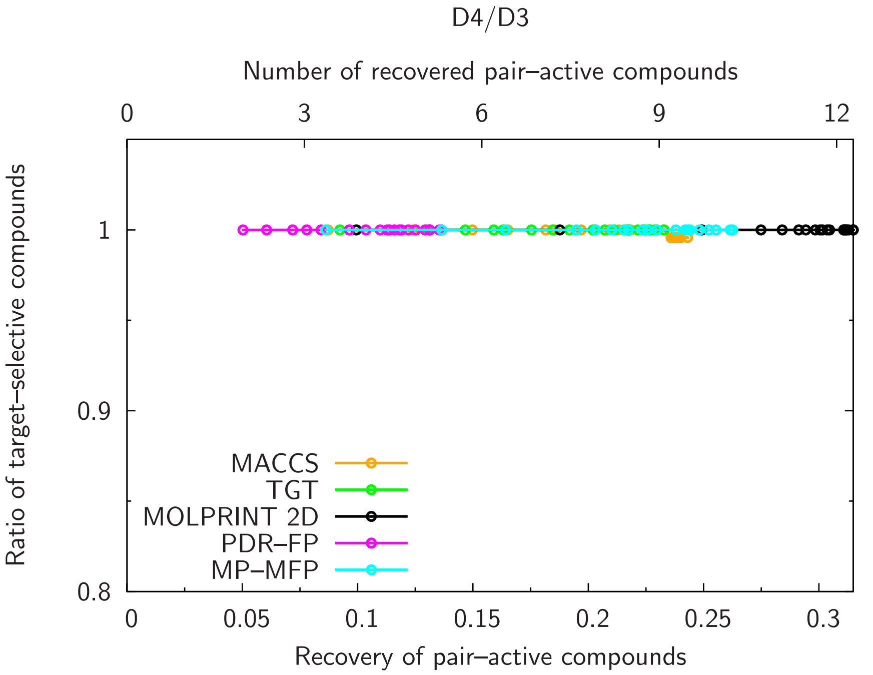
(b)



(c)

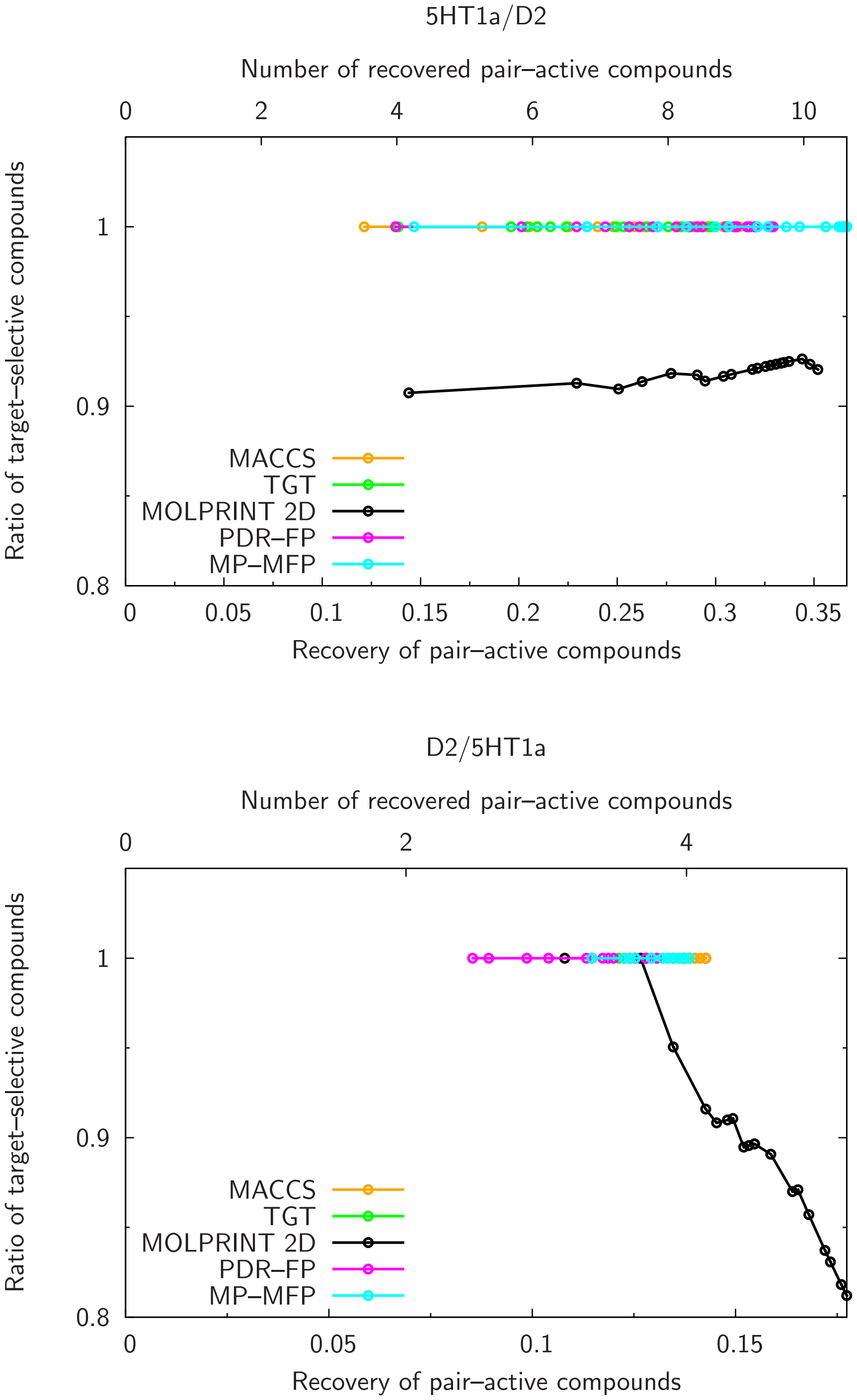
Figure C.1: Continued.





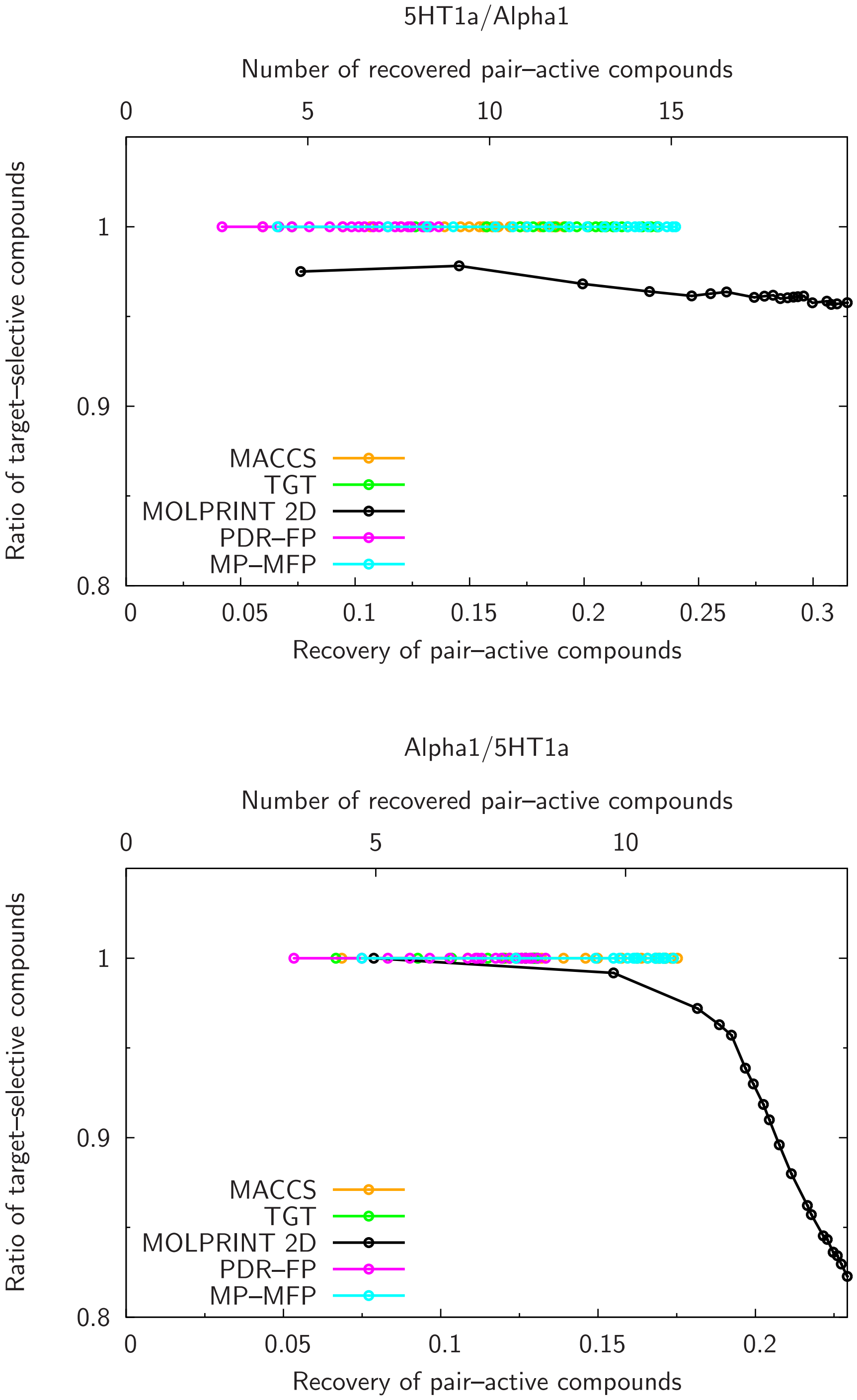
(d)

Figure C.1: Continued.



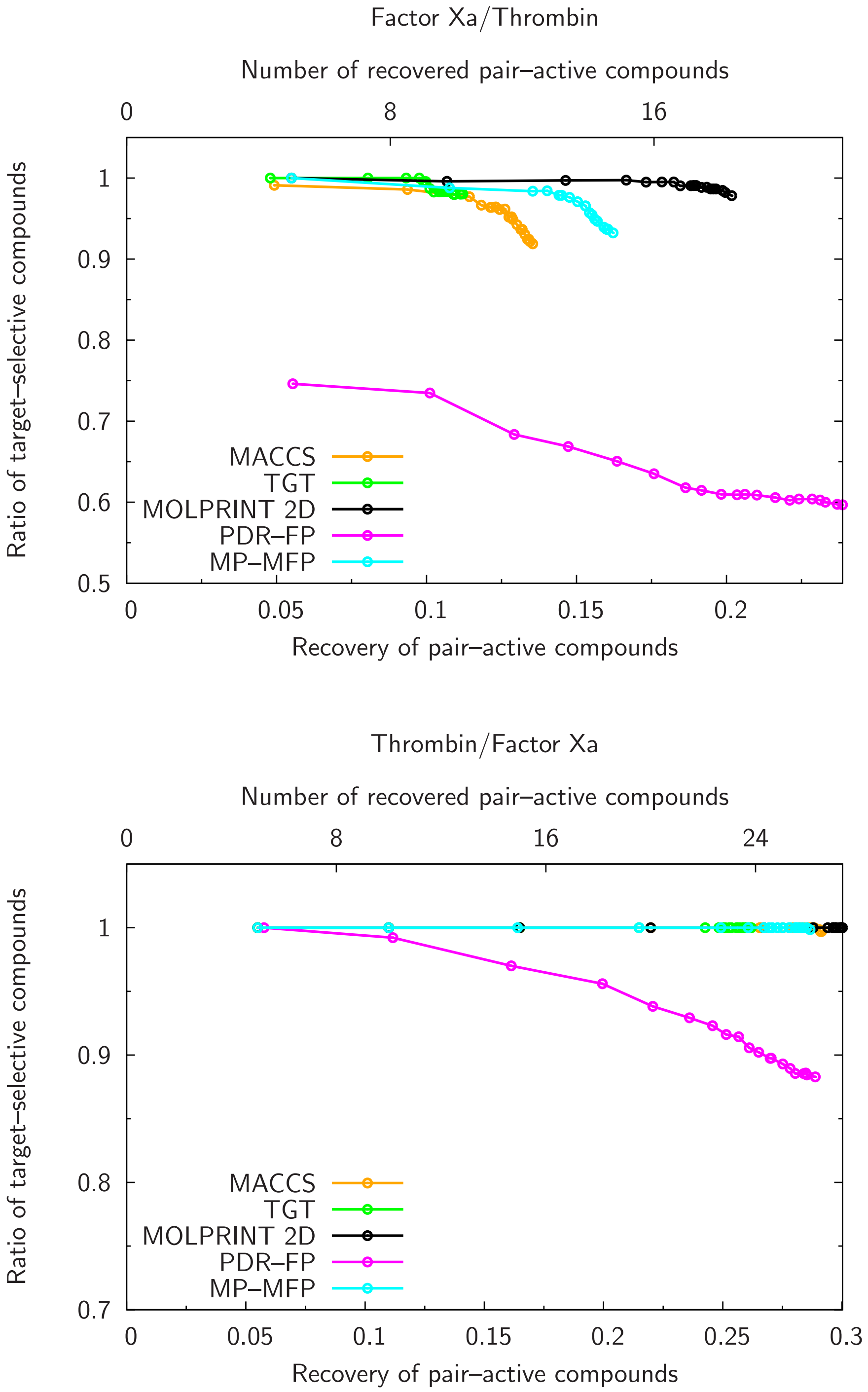
(e)

Figure C.1: Continued.



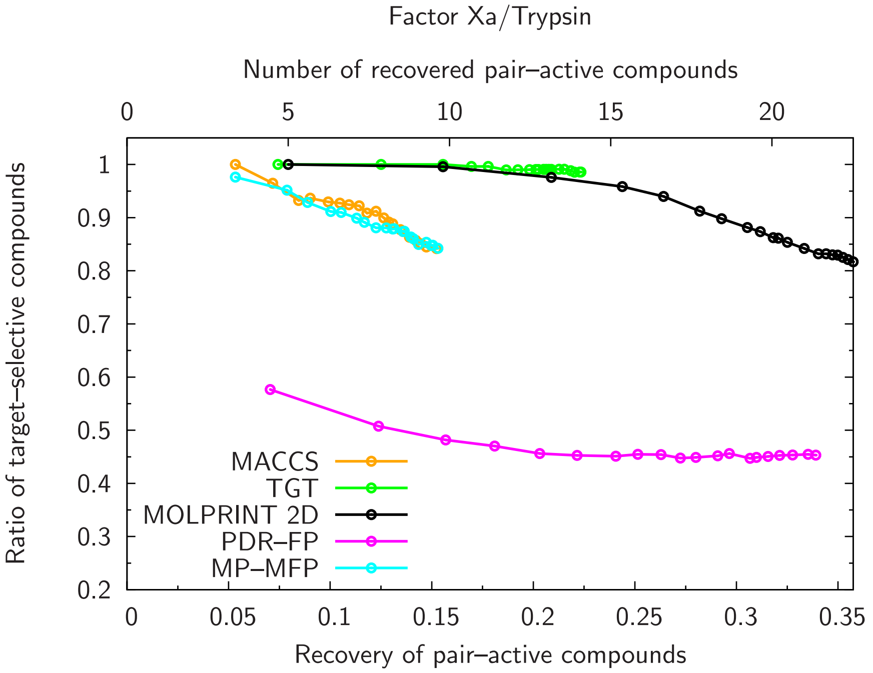
(f)

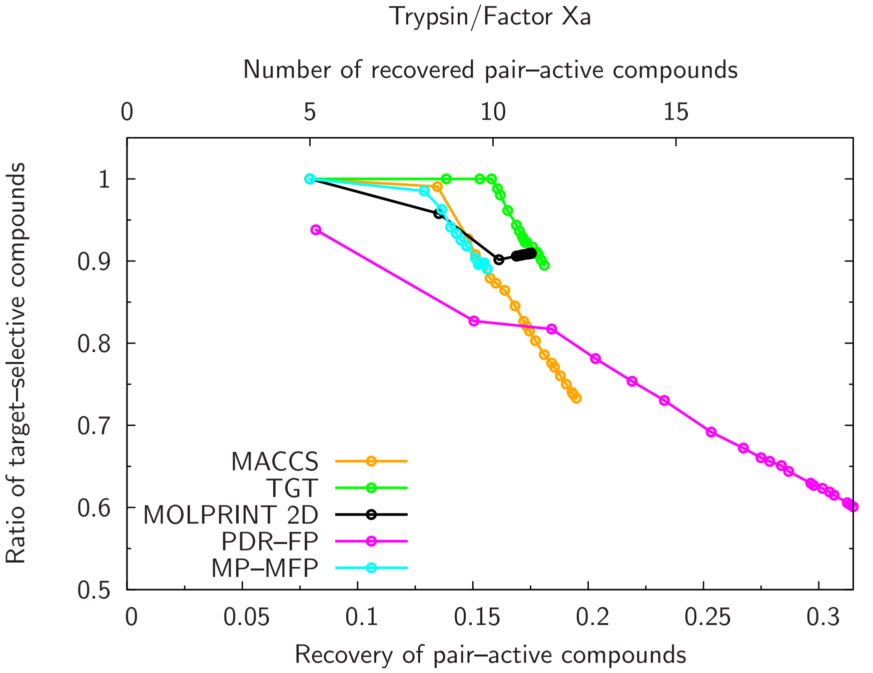
Figure C.1: Continued.



(g)

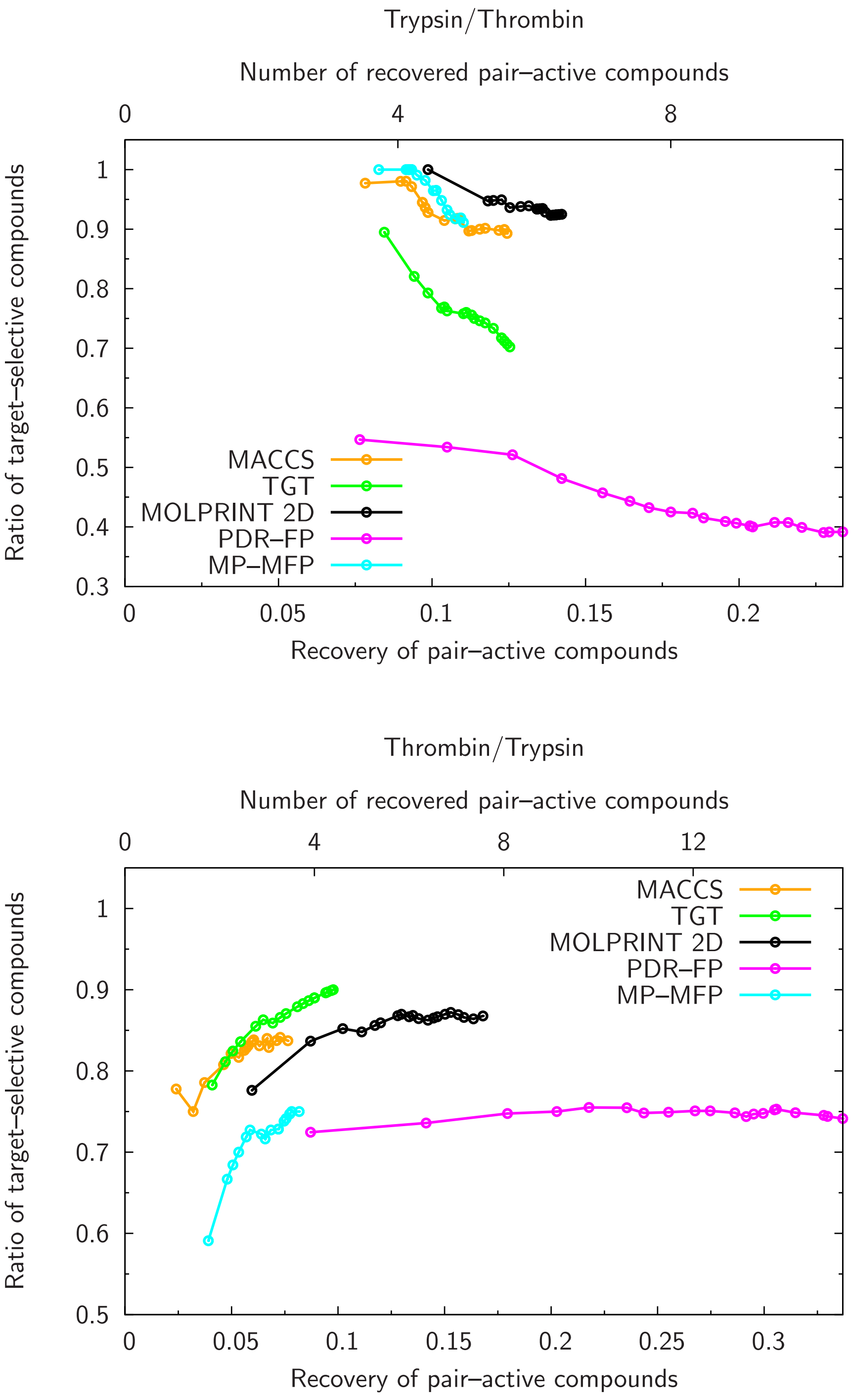
Figure C.1: Continued.





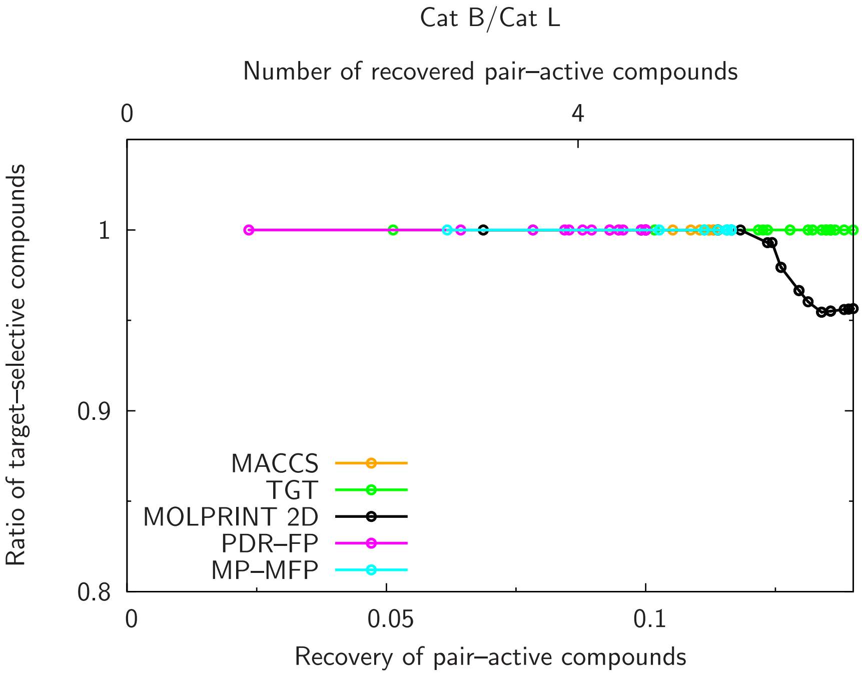
(h)

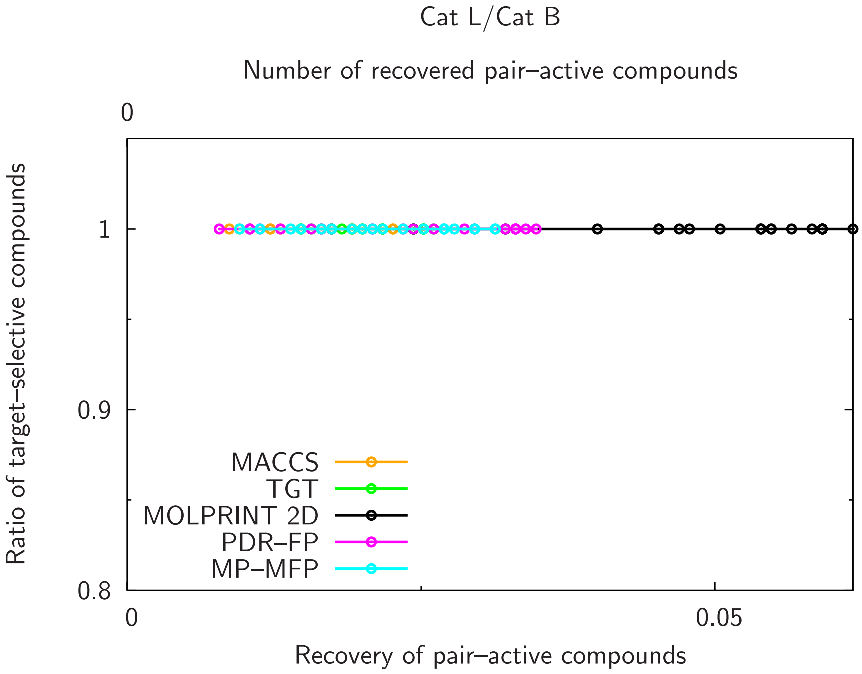
Figure C.1: Continued.



(i)

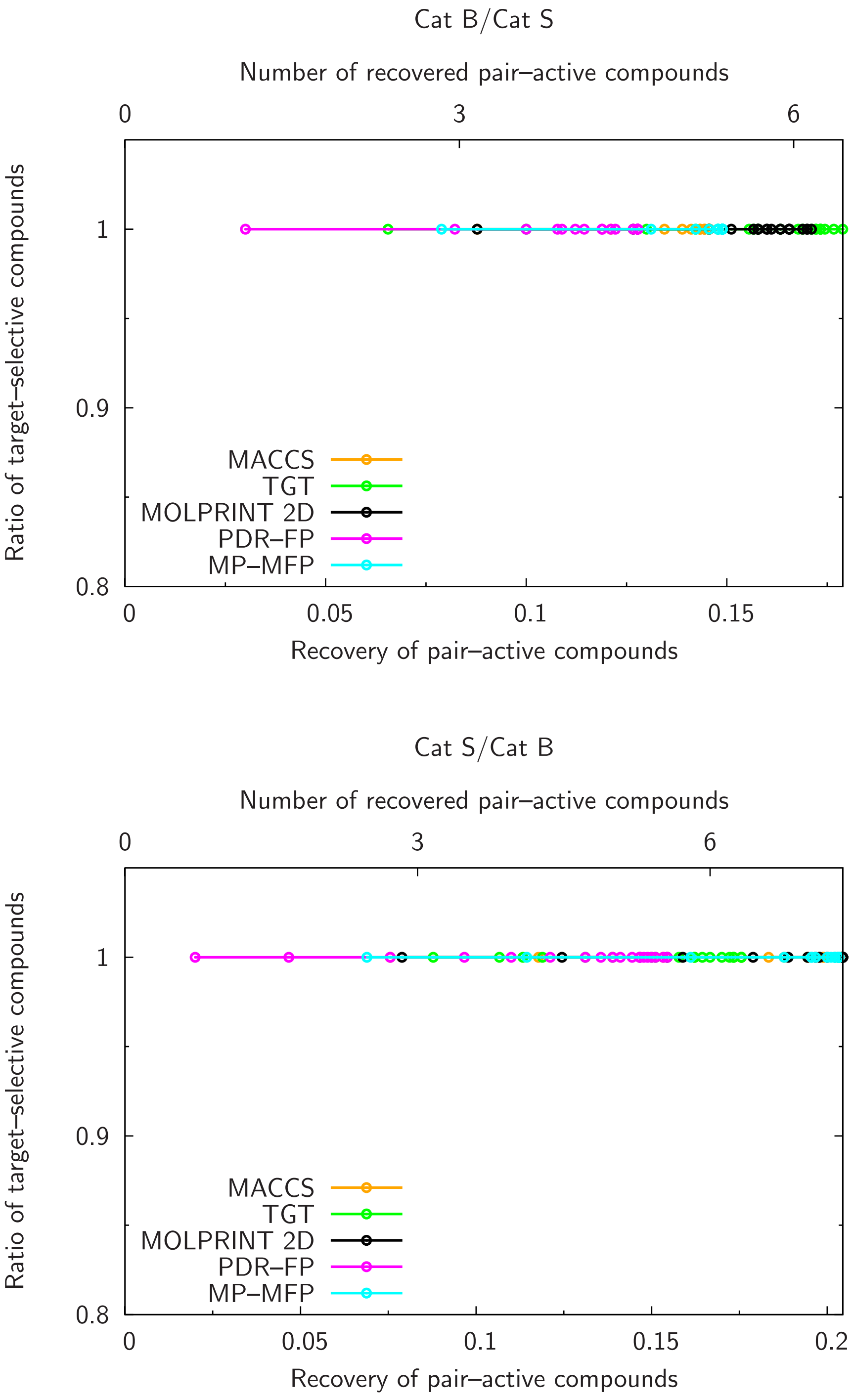
Figure C.1: Continued.





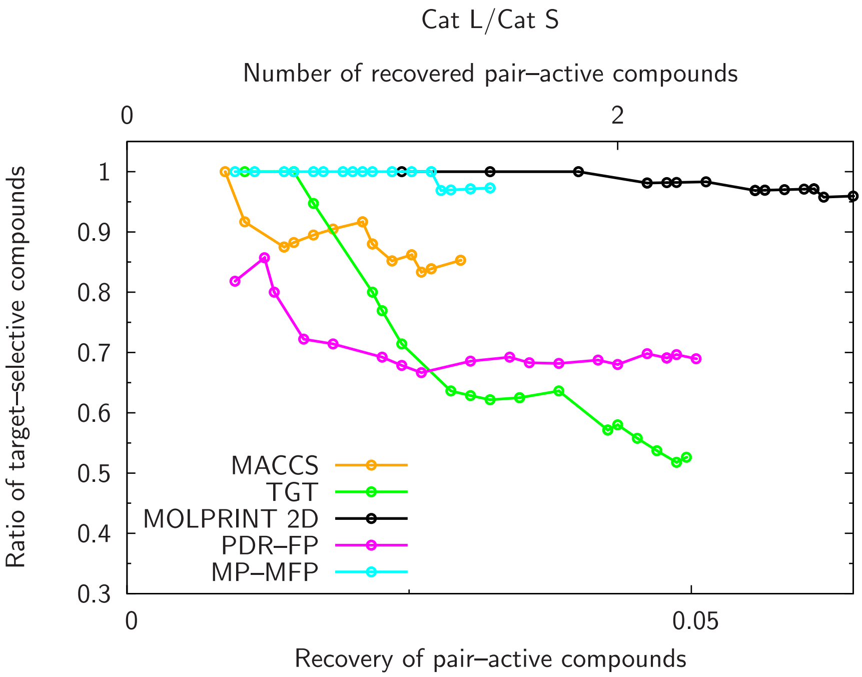
(j)

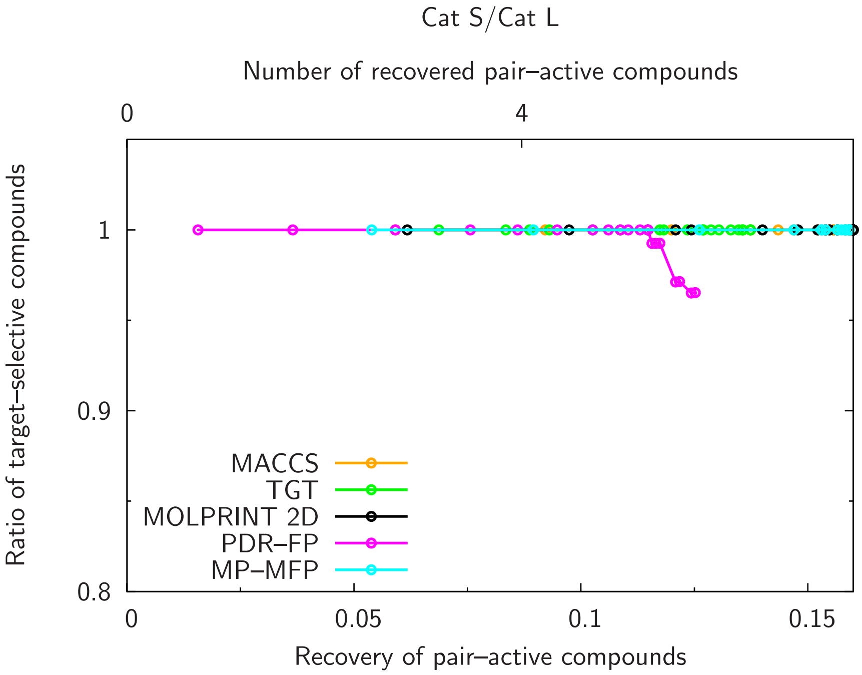
Figure C.1: Continued.



(k)

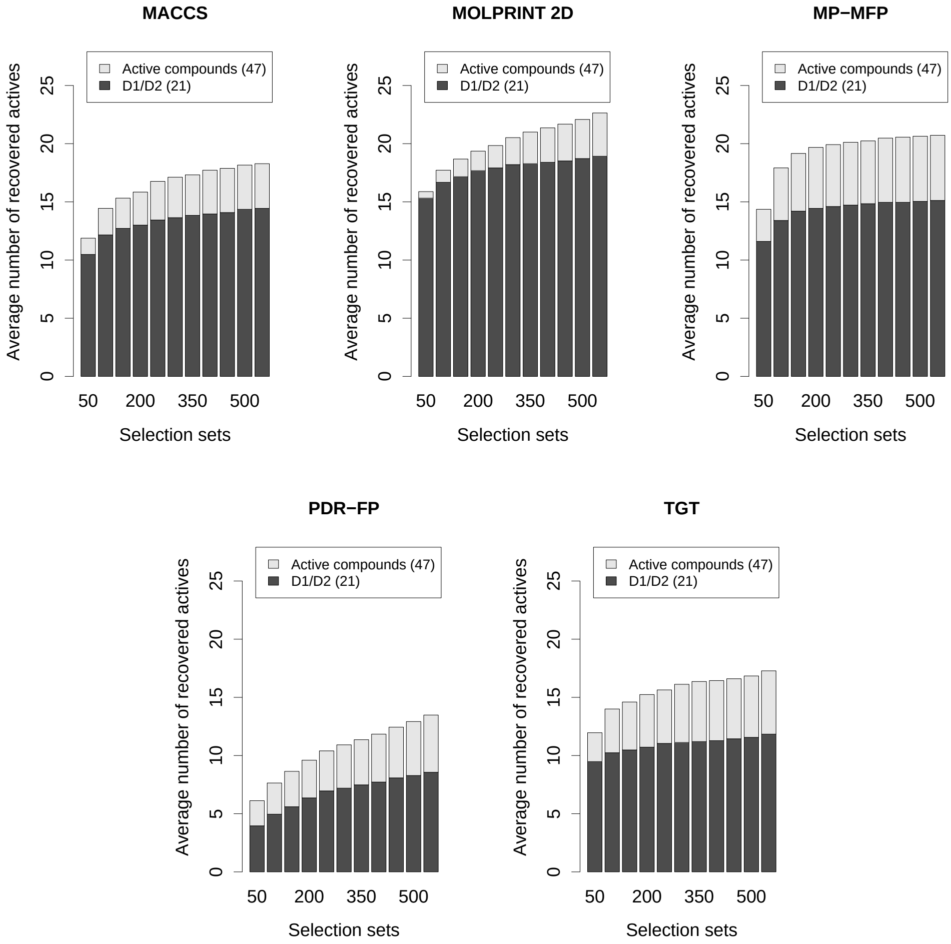
Figure C.1: Continued.





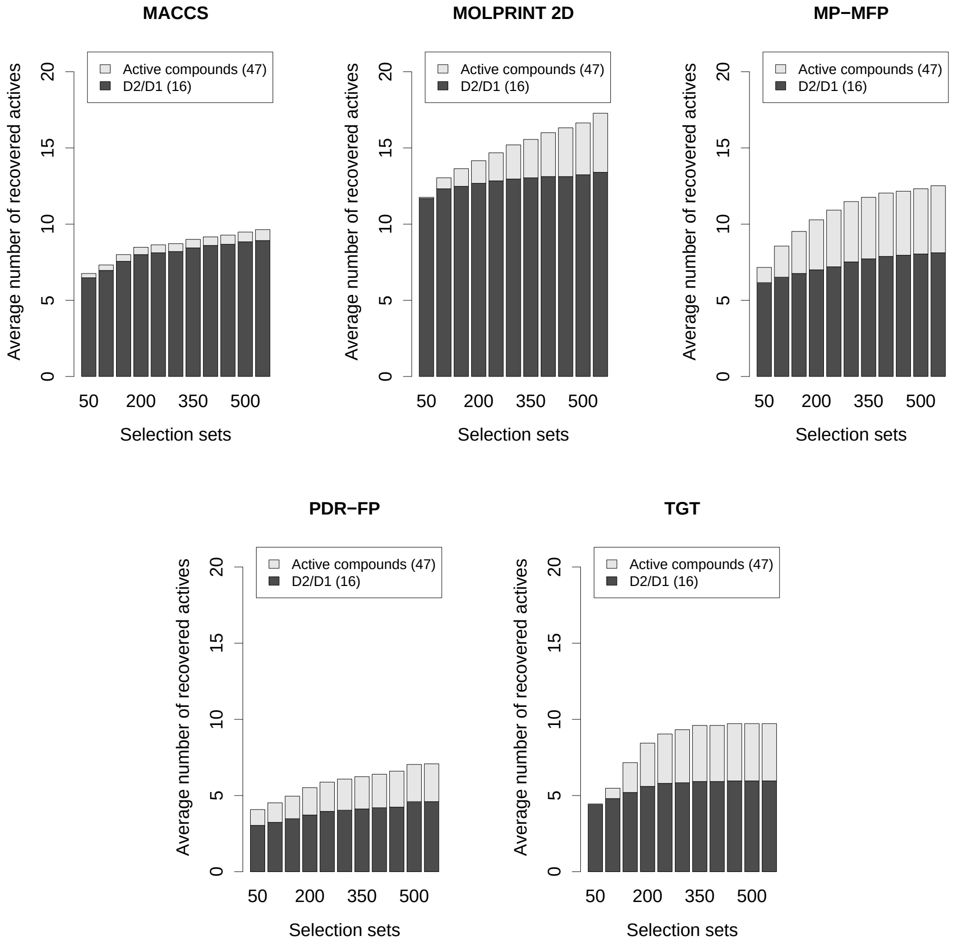
(l)

Figure C.1: Continued.



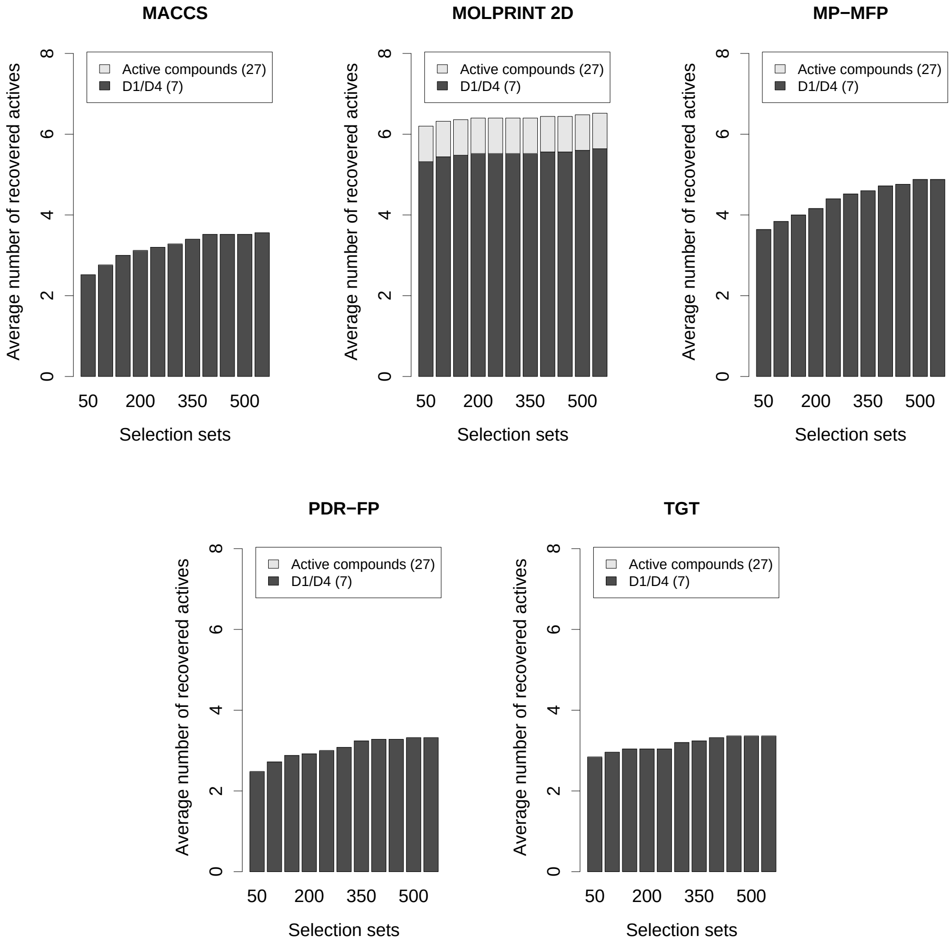
(a)

Figure C.2: Ratio of false–positive active versus target–selective compounds. The bar graphs shown in (a)-(w) report the relative amounts of false–positive active (light gray) and target–selective (dark gray) compounds for each fingerprint method as an average number of these compounds in differently sized selection sets ranging from 50 to 550 compounds. According to the definition in the text, the sum of false–positive active and target–selective molecules are pair–active compounds. Total numbers of pair–active compounds are reported in parentheses.



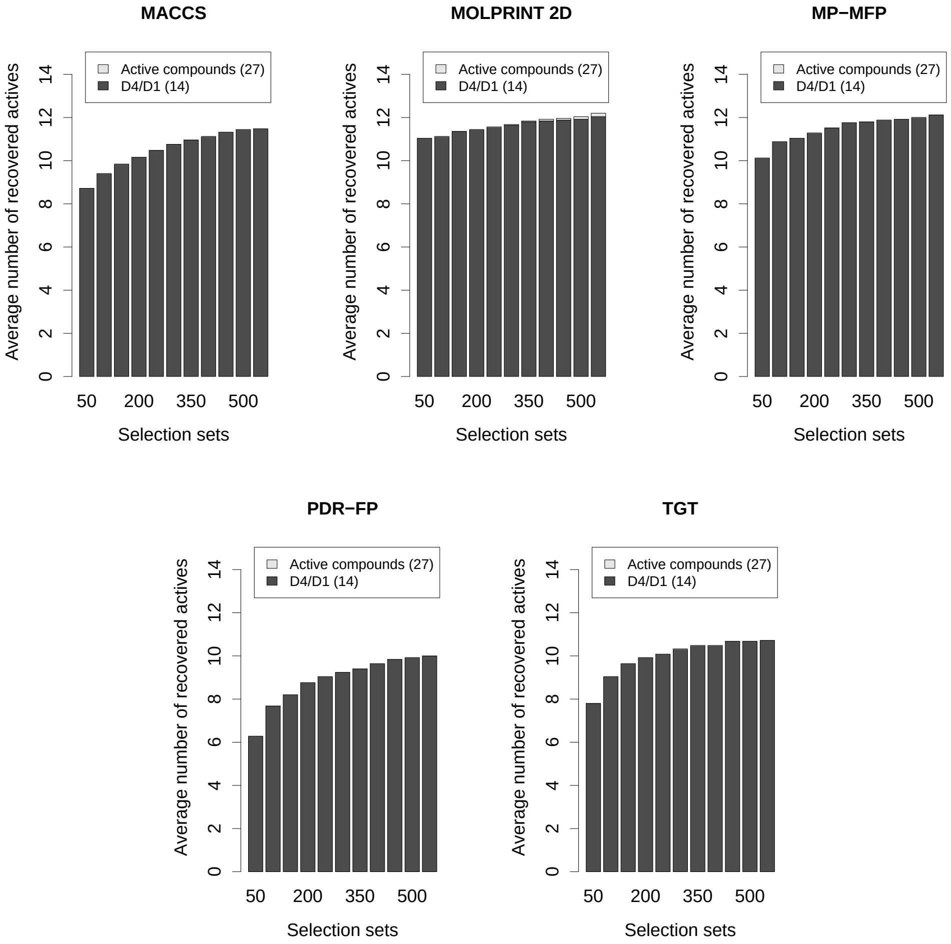
(b)

Figure C.2: Continued.



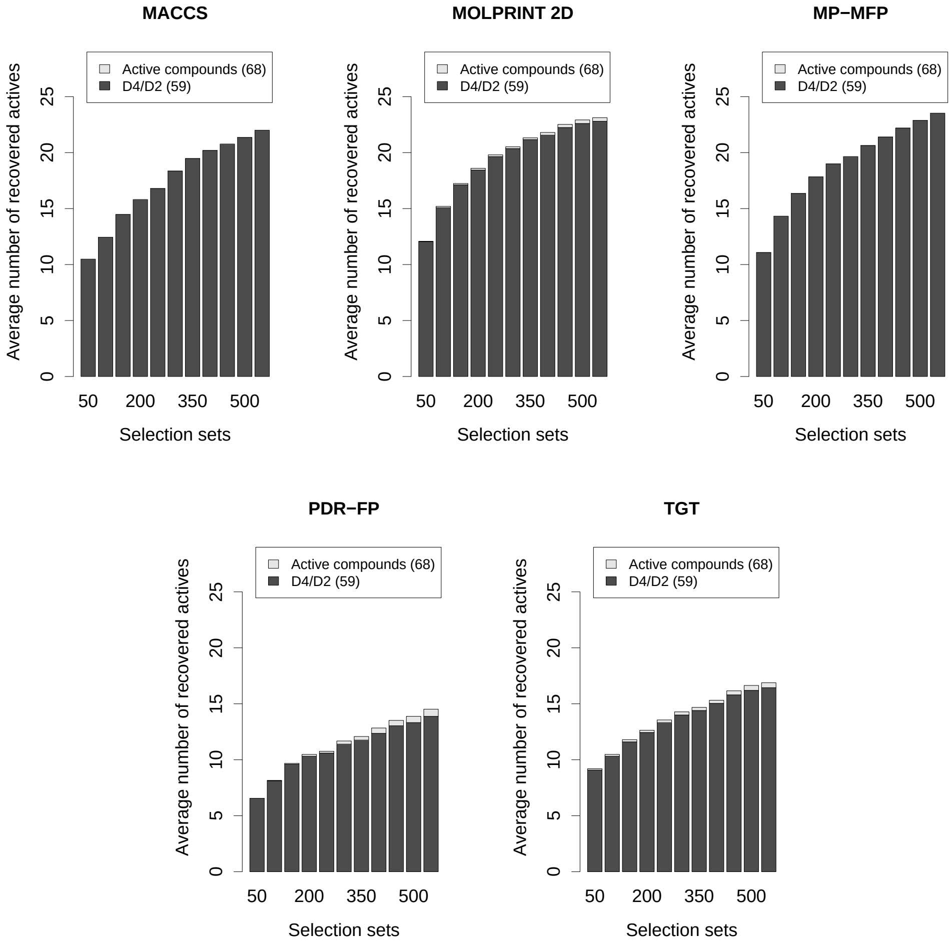
(c)

Figure C.2: Continued.



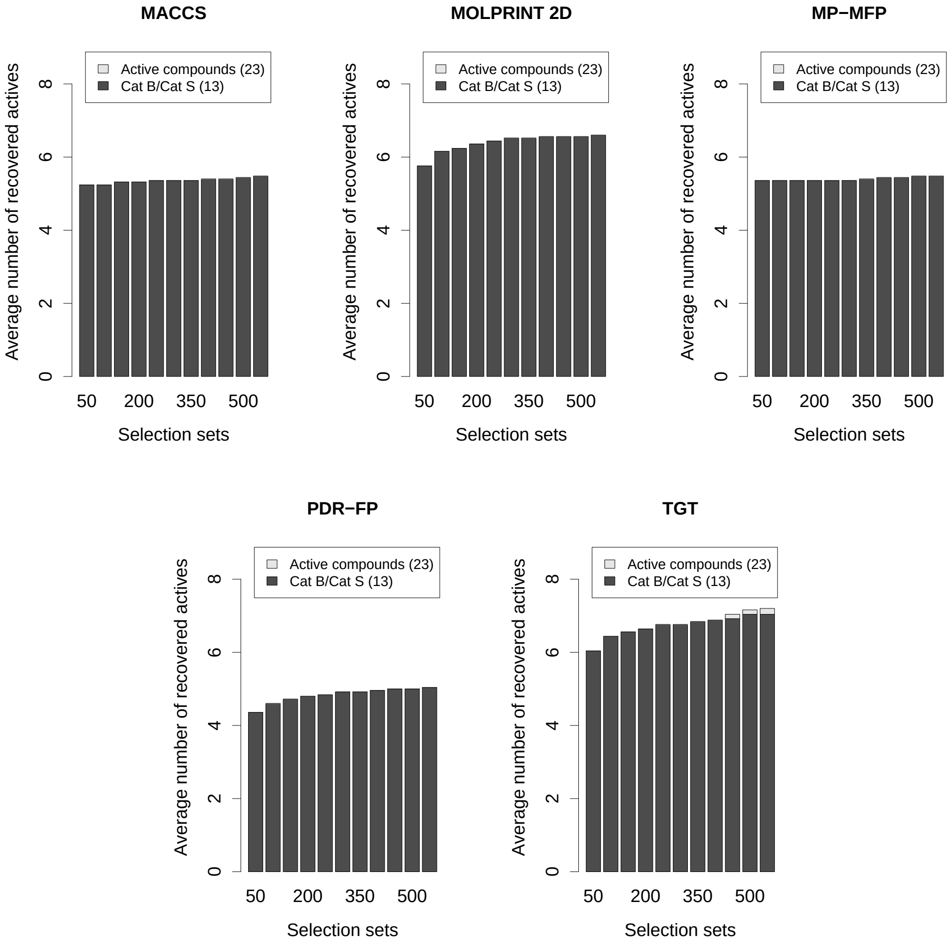
(d)

Figure C.2: Continued.



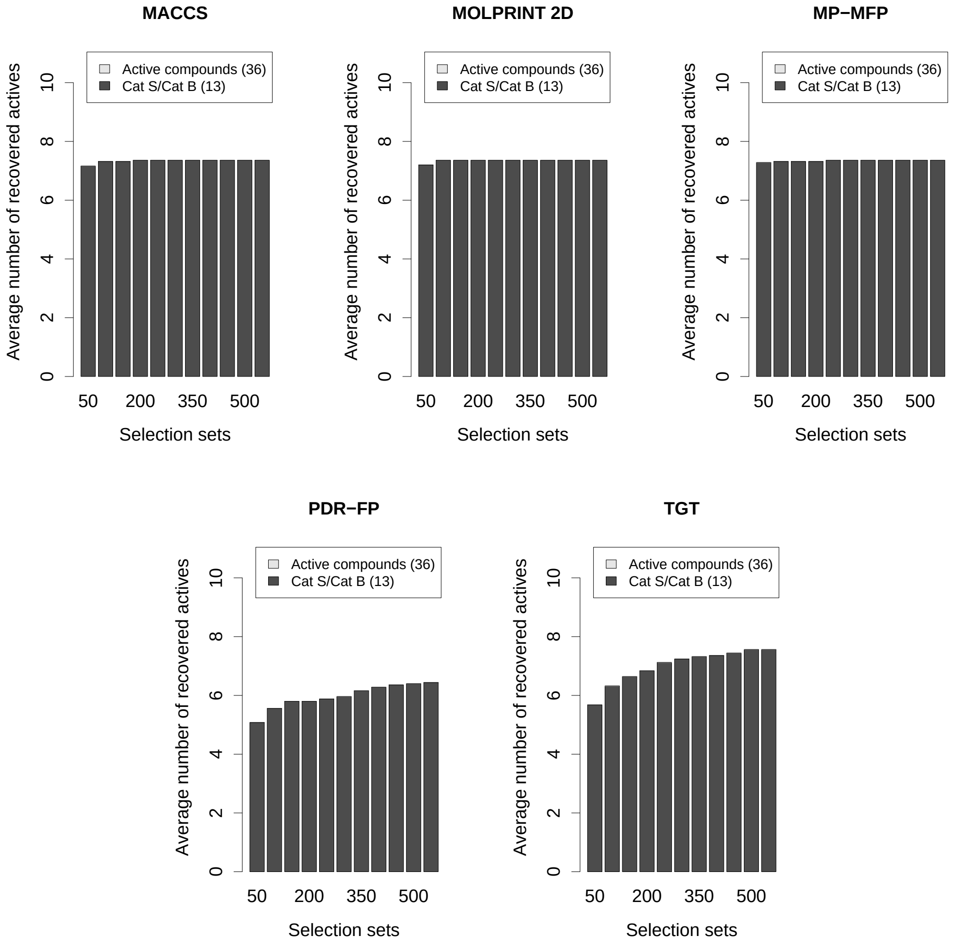
(e)

Figure C.2: Continued.



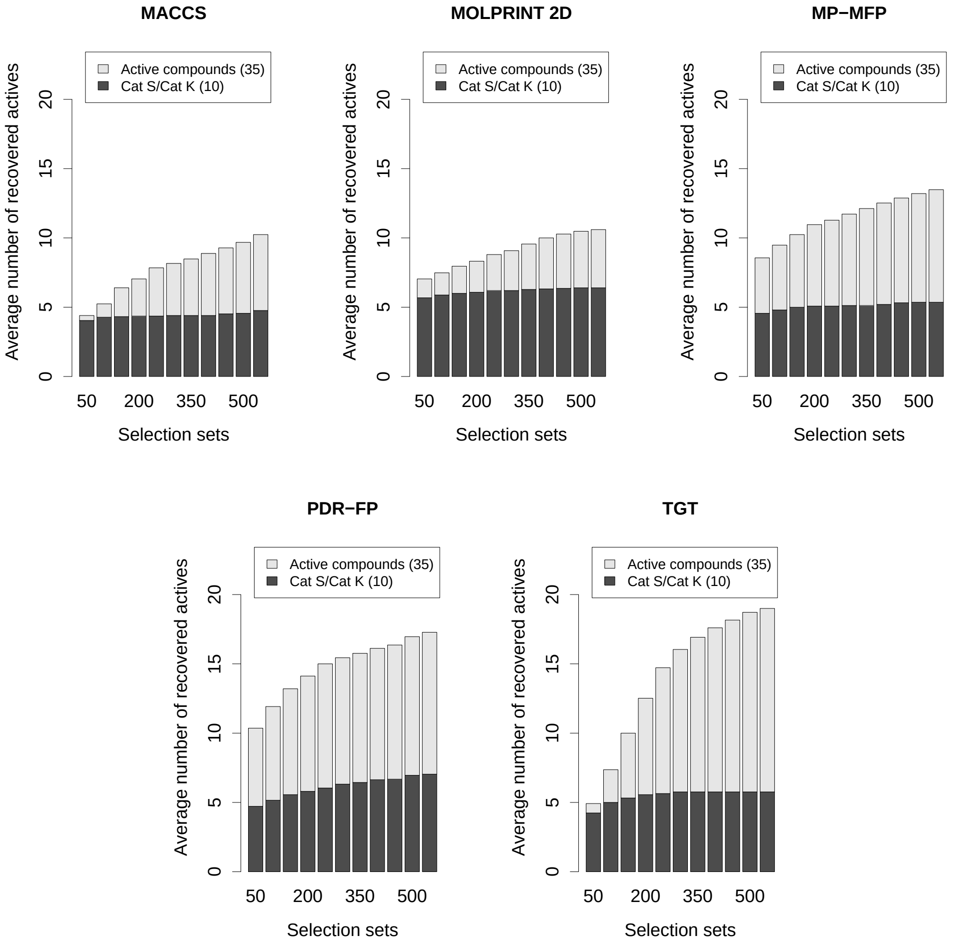
(f)

Figure C.2: Continued.



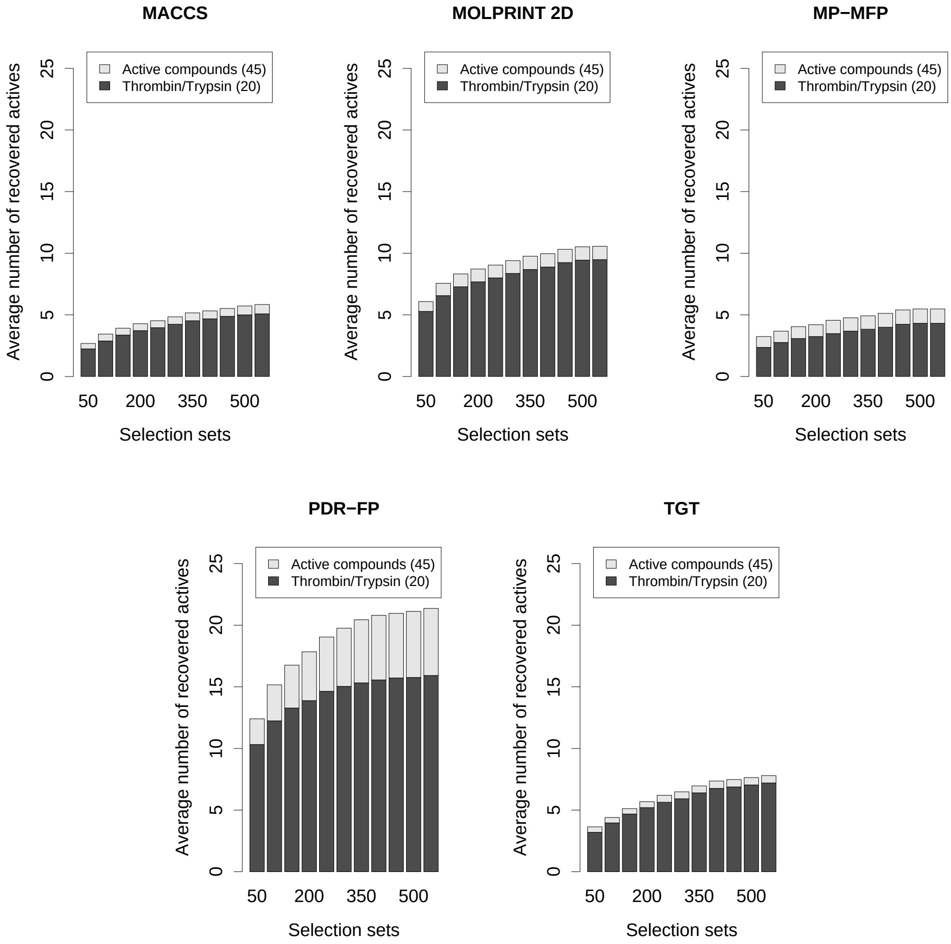
(g)

Figure C.2: Continued.



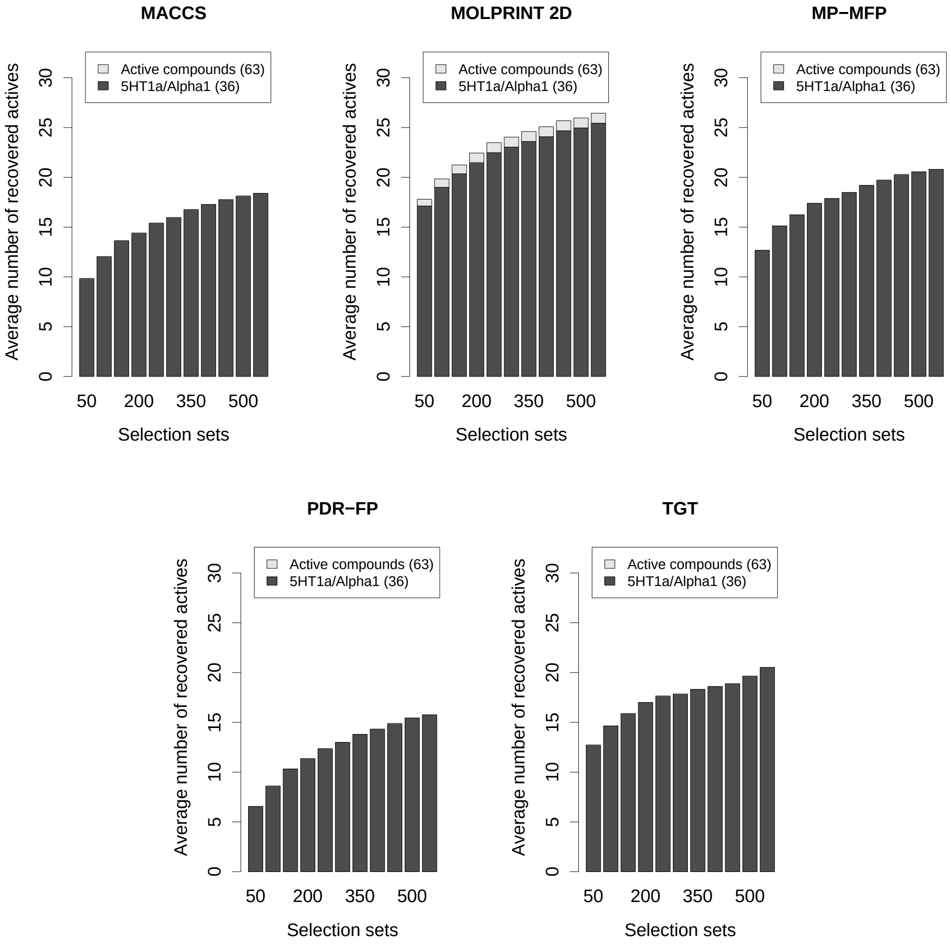
(h)

Figure C.2: Continued.



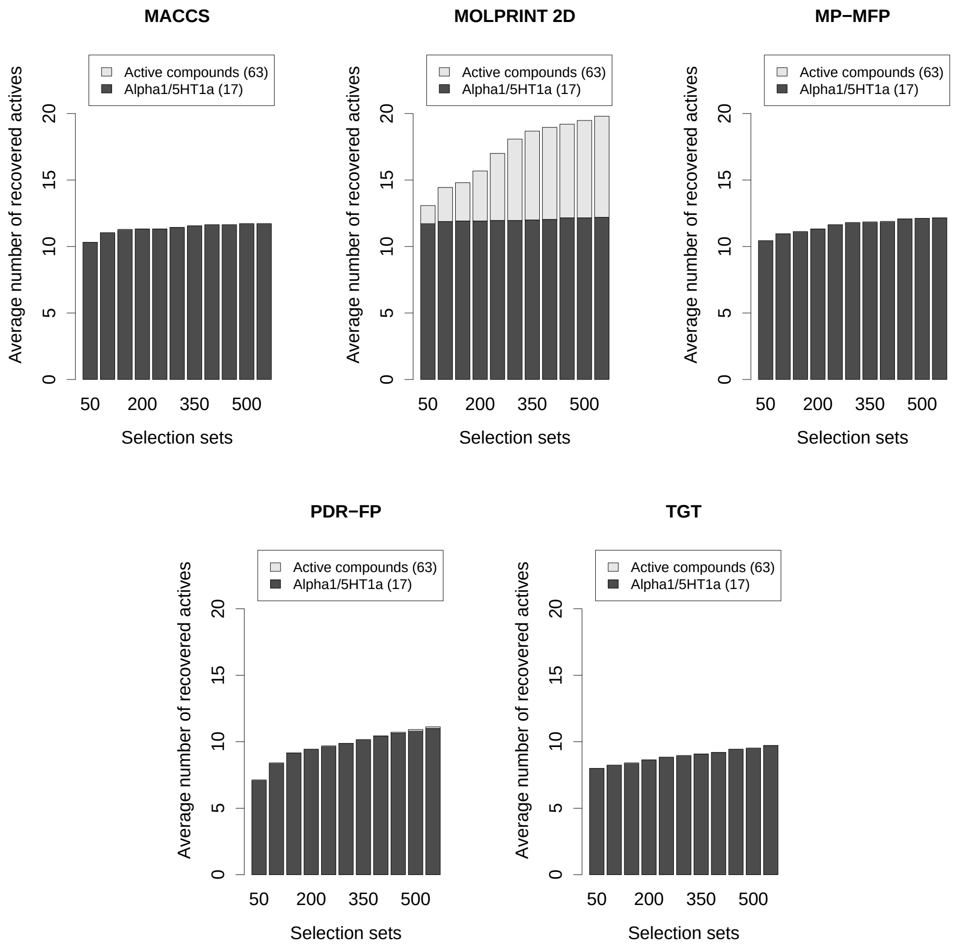
(i)

Figure C.2: Continued.



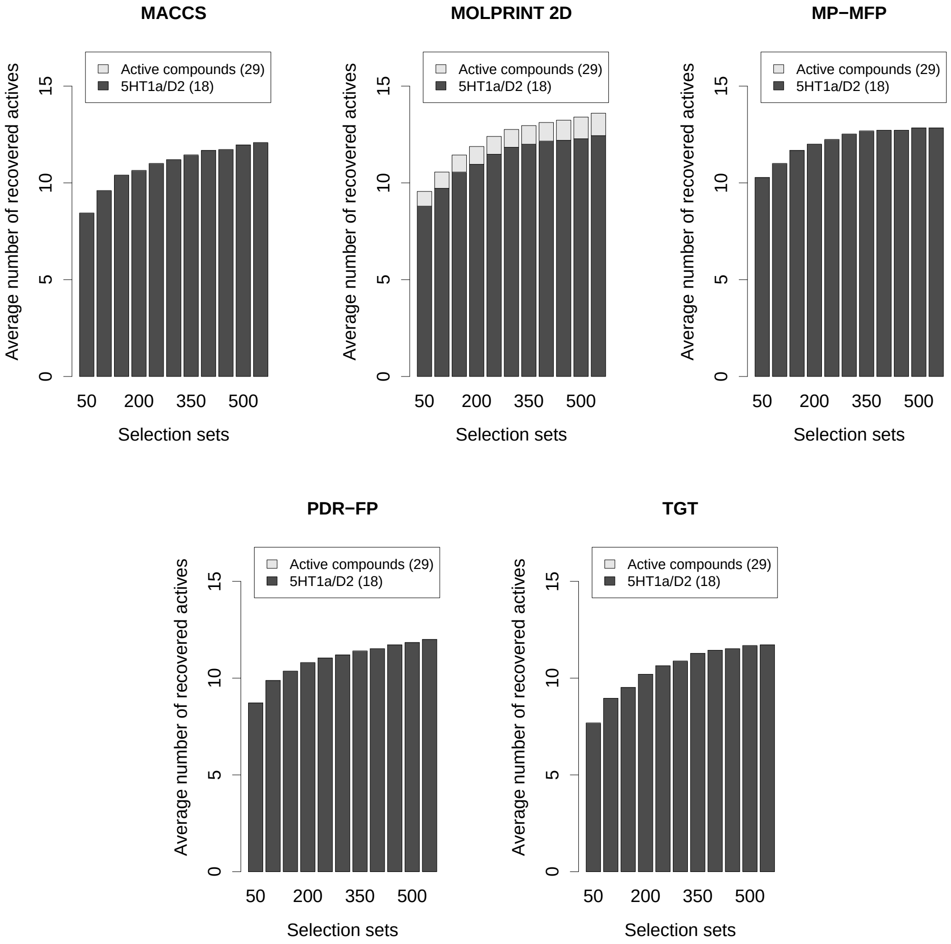
(j)

Figure C.2: Continued.



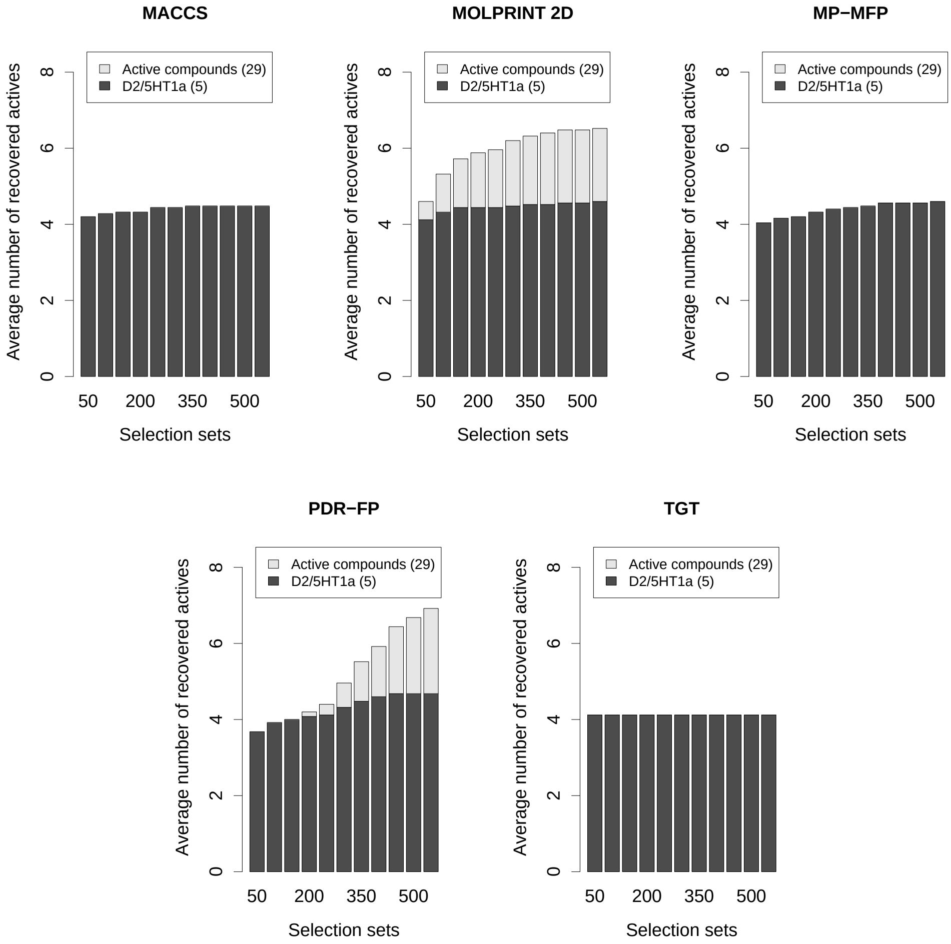
(k)

Figure C.2: Continued.



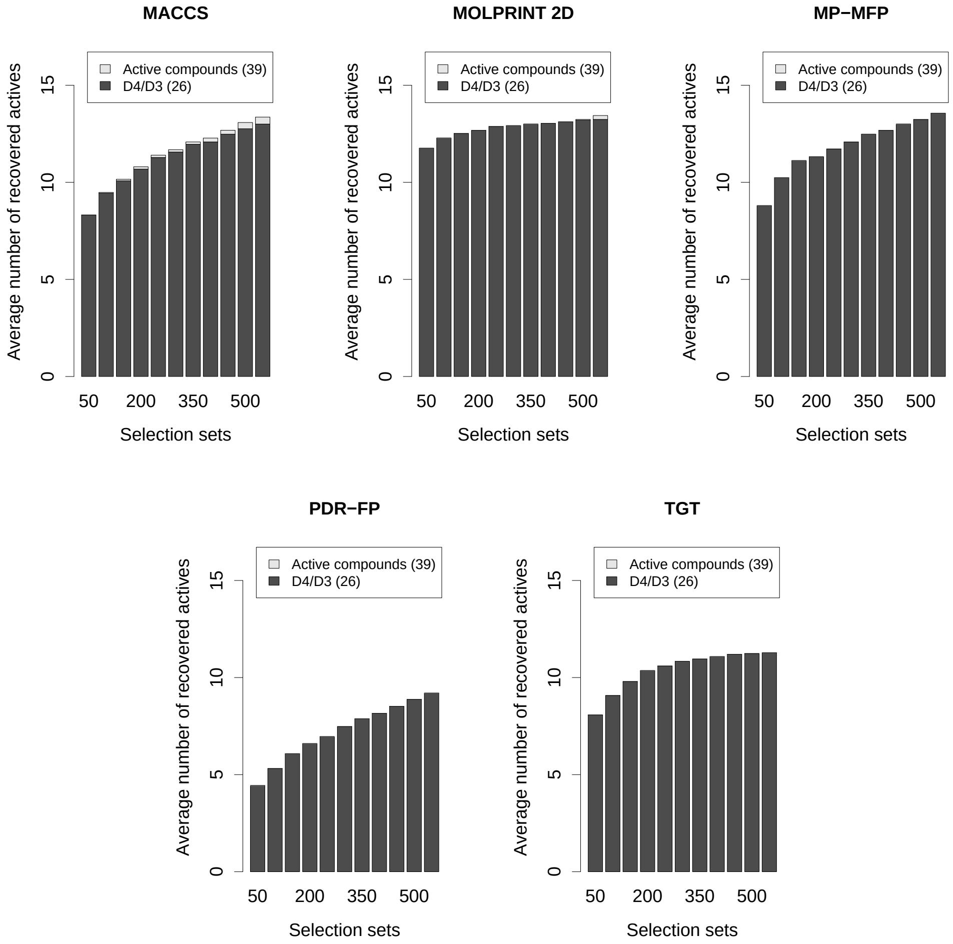
(I)

Figure C.2: Continued.



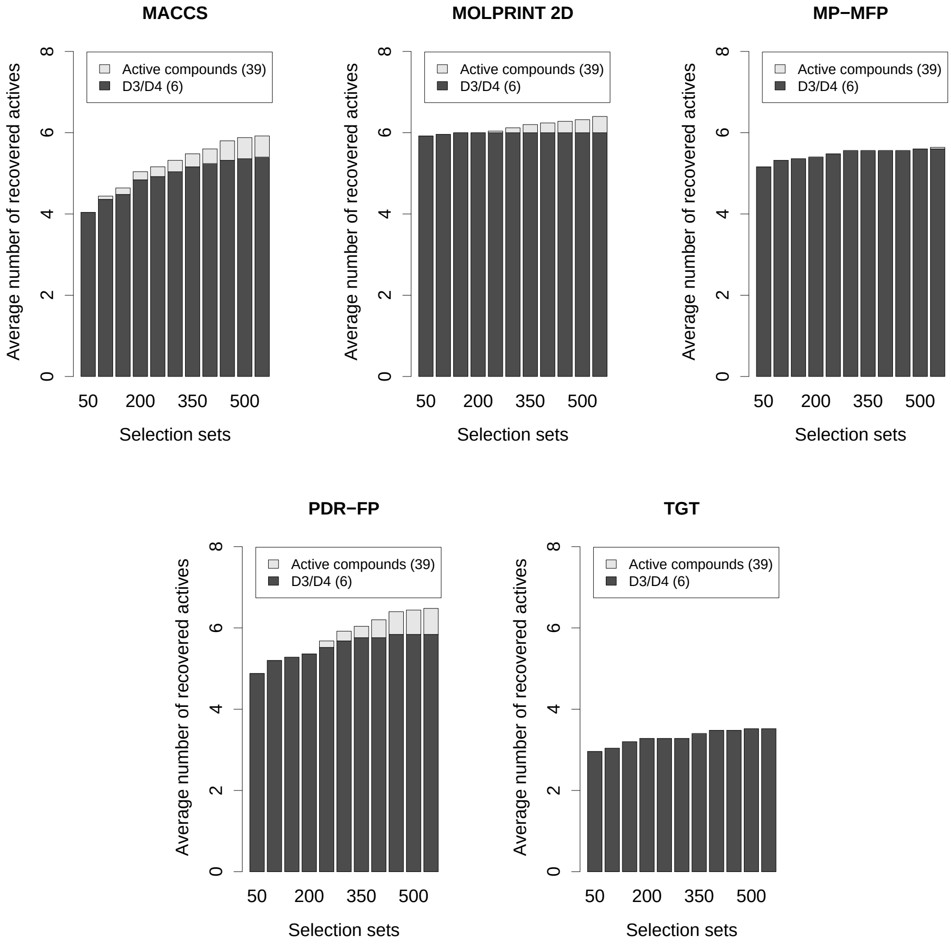
(m)

Figure C.2: Continued.



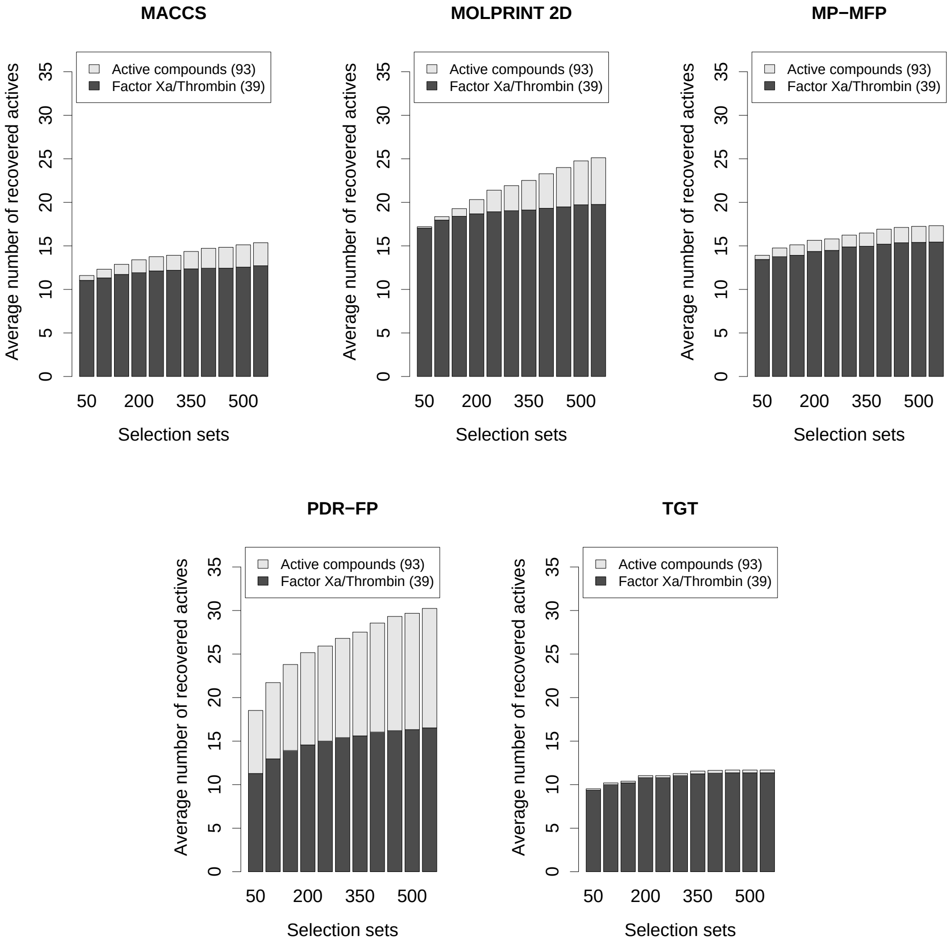
(n)

Figure C.2: Continued.



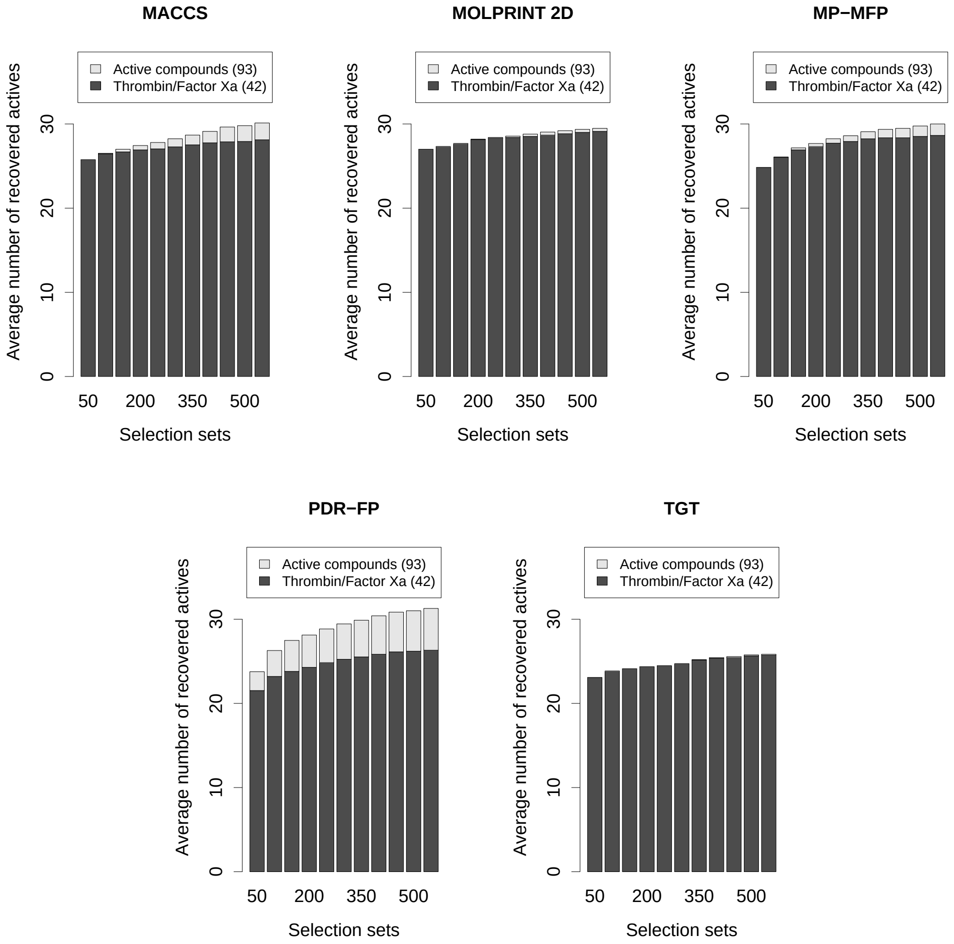
(o)

Figure C.2: Continued.



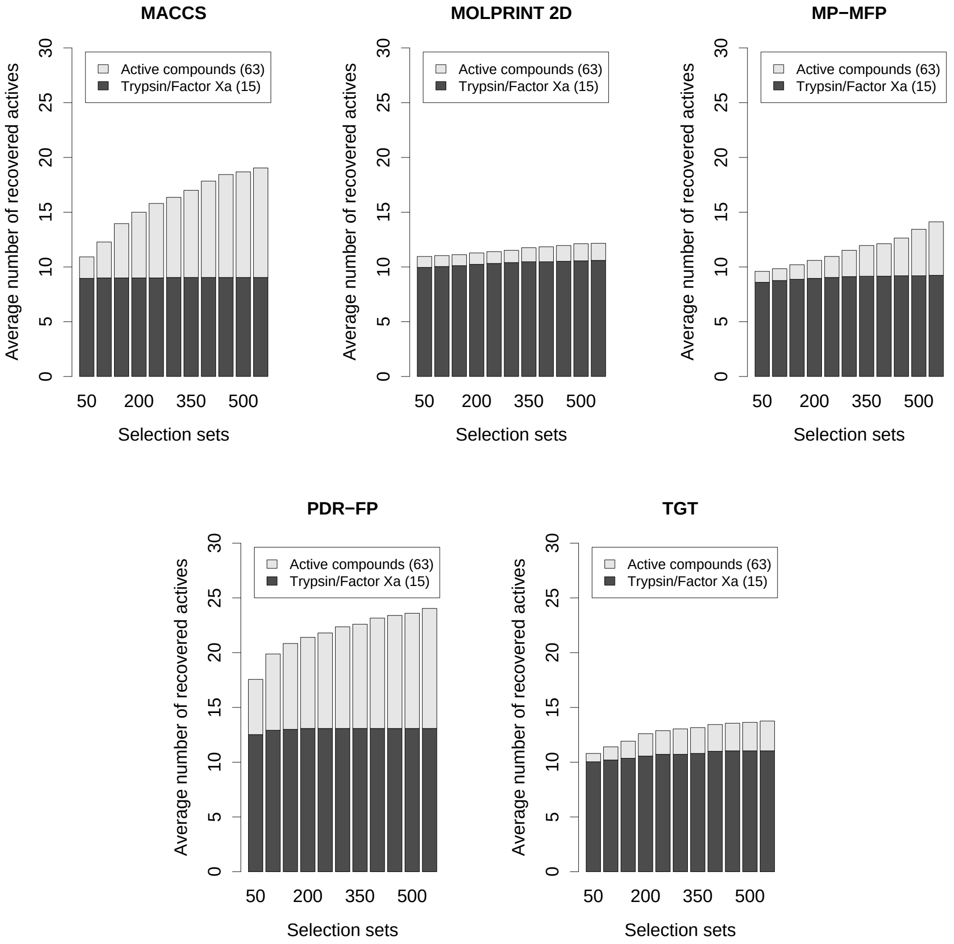
(p)

Figure C.2: Continued.



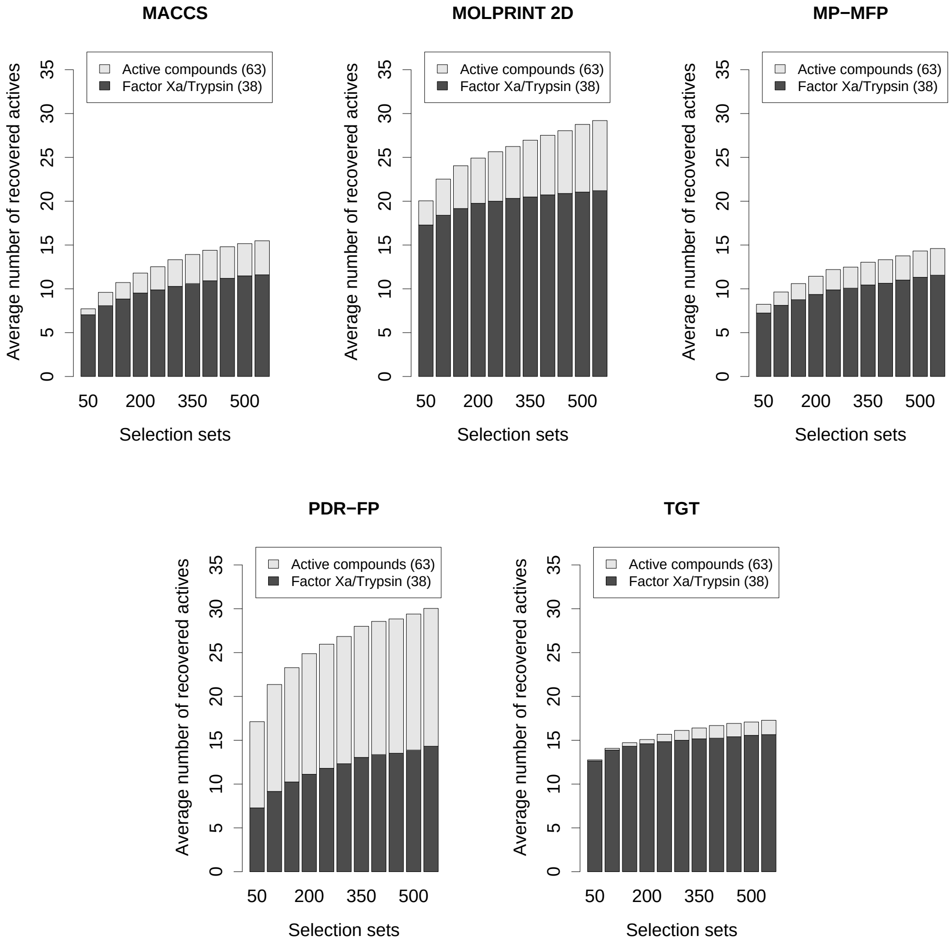
(q)

Figure C.2: Continued.



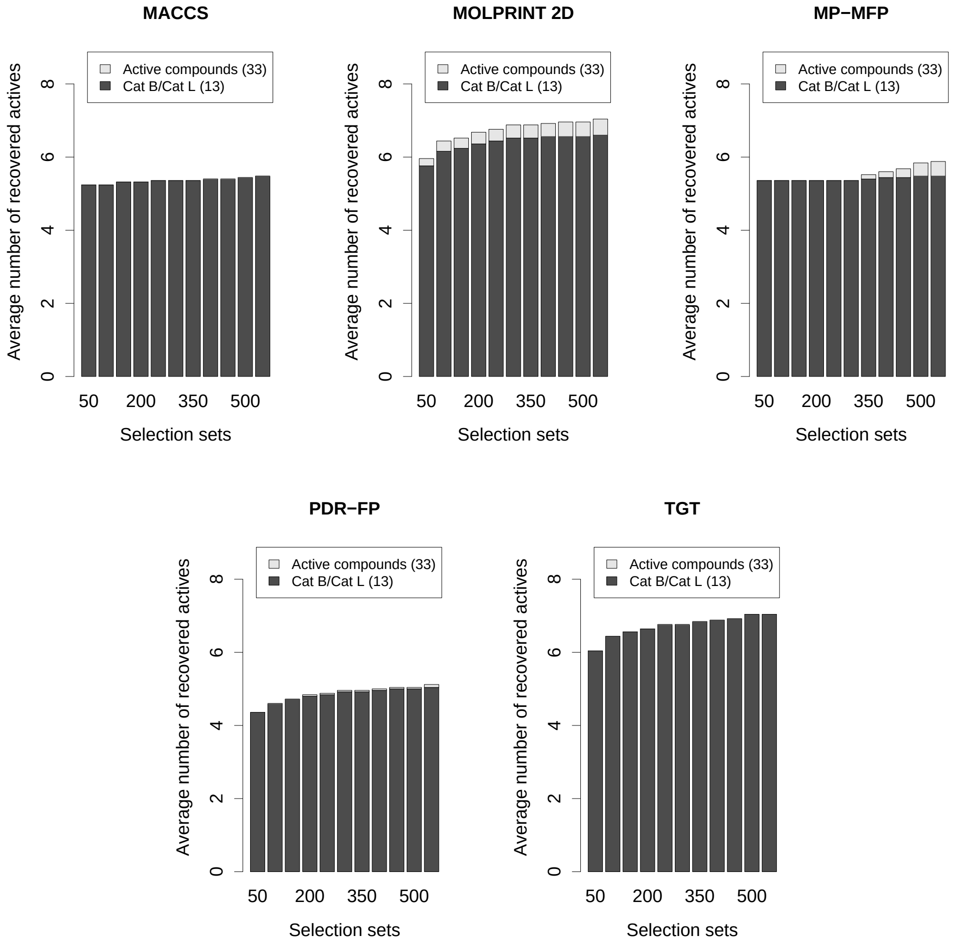
(r)

Figure C.2: Continued.



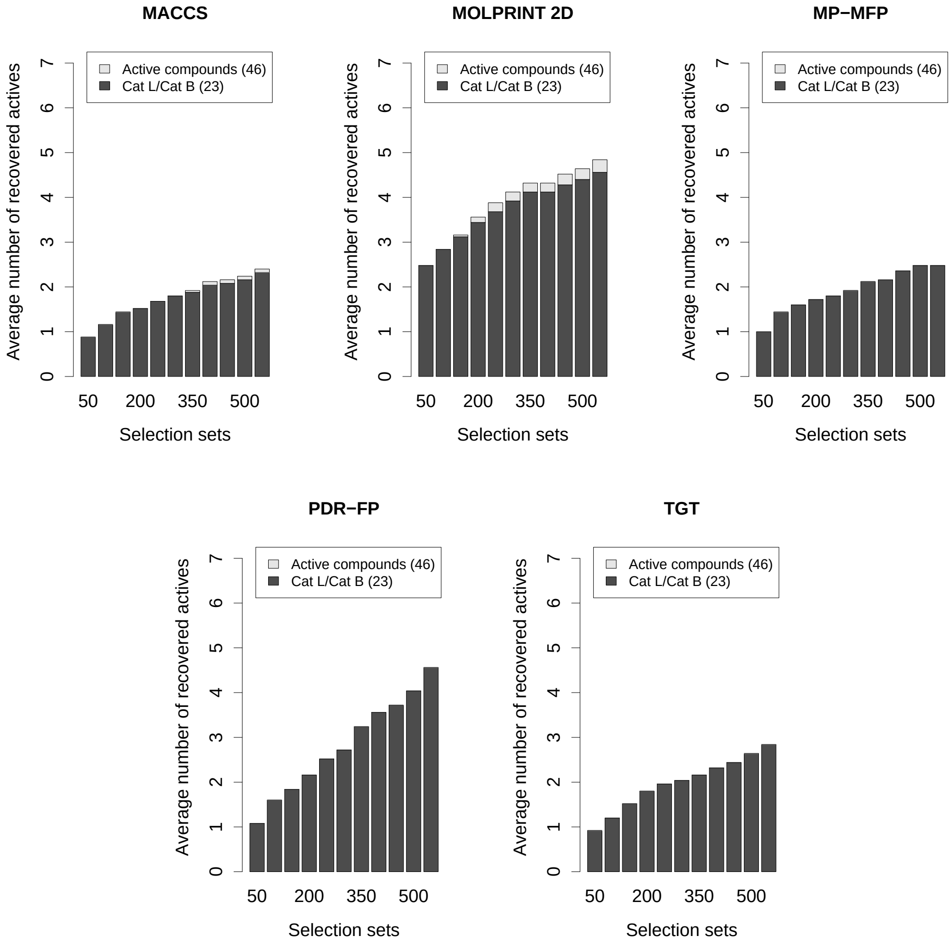
(s)

Figure C.2: Continued.



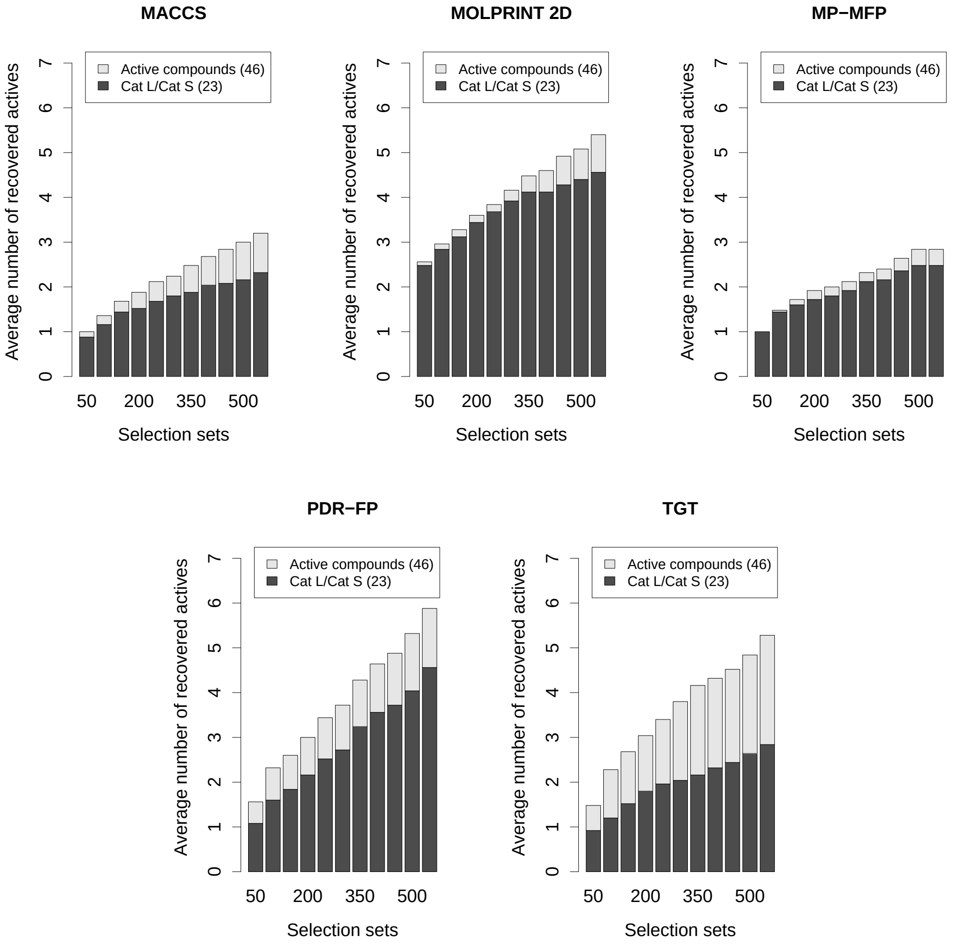
(t)

Figure C.2: Continued.



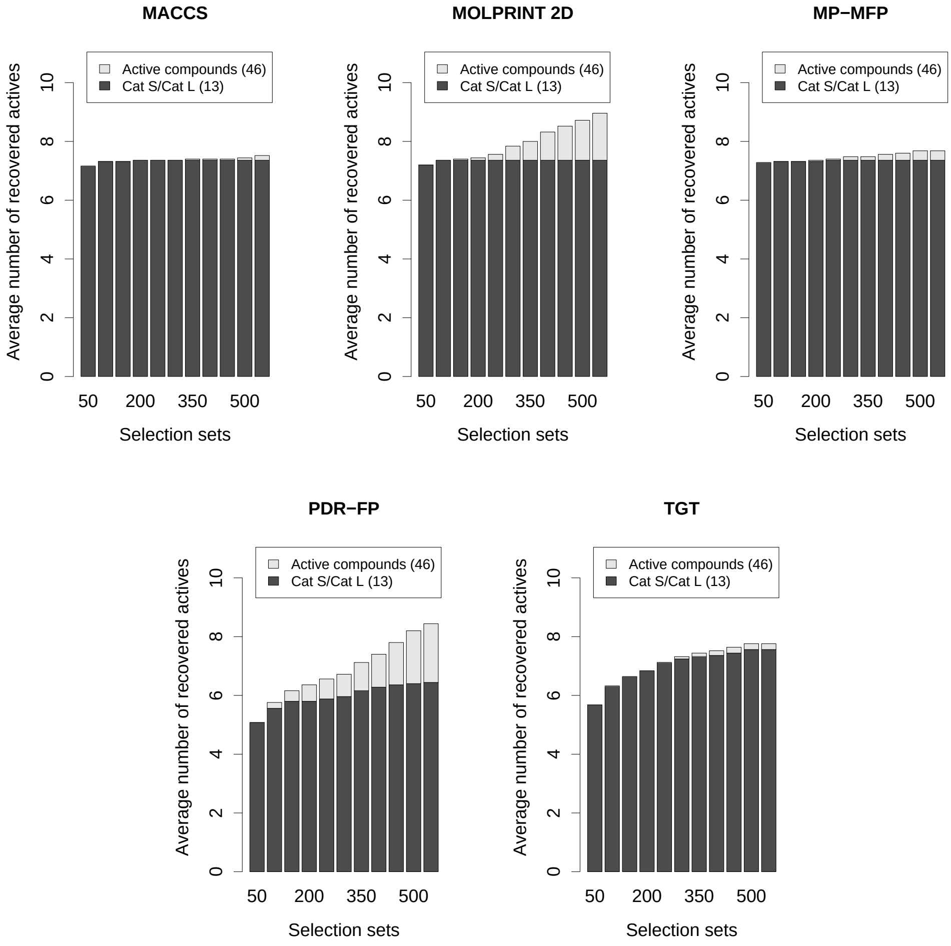
(u)

Figure C.2: Continued.



(v)

Figure C.2: Continued.



(w)

Figure C.2: Continued.

MACCS	PAC [%]			TSC [%]		
	maxHR	HR	RR	maxHR	HR	RR
D1/D2	47.0	14.4	30.7	21.0	12.2	57.9
D2/D1	47.0	7.3	15.6	16.0	7.0	43.5
D1/D4	27.0	2.8	10.2	7.0	2.8	39.4
D4/D1	27.0	9.4	34.8	14.0	9.4	67.1
D2/D4	68.0	3.7	5.4	4.0	3.7	93.0
D4/D2	68.0	12.4	18.3	59.0	12.4	21.1
D3/D4	39.0	4.4	11.4	6.0	4.4	72.7
D4/D3	39.0	9.5	24.3	27.0	9.4	35.0
D2/5HT1a	29.0	4.3	14.8	5.0	4.3	85.6
5HT1a/D2	29.0	9.6	33.1	18.0	9.6	53.3
Alpha1/5HT1a	63.0	11.0	17.5	17.0	11.0	65.0
5HT1a/Alpha1	63.0	12.0	19.1	36.0	12.0	33.4
Cat B/Cat L	46.0	5.2	11.4	13.0	5.2	40.3
Cat L/Cat B	46.0	1.2	2.5	23.0	1.2	5.0
Cat B/Cat S	36.0	5.2	14.6	13.0	5.2	40.3
Cat S/Cat B	36.0	7.3	20.3	13.0	7.3	56.3
Cat L/Cat S	46.0	1.4	3.0	23.0	1.2	5.0
Cat S/Cat L	46.0	7.3	15.9	13.0	7.3	56.3
Cat K/Cat S	35.0	5.8	16.6	15.0	5.2	34.6
Cat S/Cat K	35.0	5.2	15.0	10.0	4.3	42.8
Thrombin/Factor Xa	91.0	26.5	29.1	42.0	26.4	63.0
Factor Xa/Thrombin	91.0	12.3	13.5	39.0	11.3	29.0
Trypsin/Factor Xa	63.0	12.3	19.5	15.0	9.0	60.0
Factor Xa/Trypsin	63.0	9.6	15.2	38.0	8.1	21.3
Trypsin/Thrombin	45.0	5.6	12.4	10.0	5.0	50.0
Thrombin/Trypsin	45.0	3.4	7.6	25.0	2.9	11.5

(a)

Table C.1: Search results for all fingerprints and selectivity sets. (a) – (e) report for each fingerprint the average hit (*HR*) and recovery rates (*RR*) of pair–active (*PAC*) and target–selective (*TSC*) compounds among the 100 top scoring database compounds. The *maxHR* column reports the theoretically possible maximum hit rate for each selectivity set, given that the number of potential hits in the background database is always smaller than 100.

MOLPRINT 2D	**PAC** [%]			**TSC** [%]		
	maxHR	**HR**	**RR**	**maxHR**	**HR**	**RR**
D1/D2	47.0	17.7	37.7	21.0	16.7	79.4
D2/D1	47.0	13.0	27.7	16.0	12.3	77.0
D1/D4	27.0	6.3	23.4	7.0	5.4	77.7
D4/D1	27.0	11.1	41.2	14.0	11.1	79.4
D2/D4	68.0	5.4	8.0	4.0	3.4	86.0
D4/D2	68.0	15.2	22.4	59.0	15.1	25.6
D3/D4	39.0	6.0	15.3	6.0	6.0	99.3
D4/D3	39.0	12.3	31.5	27.0	12.3	45.5
D2/5HT1a	29.0	5.3	18.3	5.0	4.3	86.4
5HT1a/D2	29.0	10.6	36.4	18.0	9.7	54.0
Alpha1/5HT1a	63.0	14.4	22.9	17.0	11.9	69.9
5HT1a/Alpha1	63.0	19.8	31.5	36.0	19.0	52.8
Cat B/Cat L	46.0	6.4	14.0	13.0	6.2	47.4
Cat L/Cat B	46.0	2.8	6.2	23.0	2.8	12.3
Cat B/Cat S	36.0	6.2	17.1	13.0	6.2	47.4
Cat S/Cat B	36.0	7.4	20.4	13.0	7.4	56.6
Cat L/Cat S	46.0	3.0	6.4	23.0	2.8	12.3
Cat S/Cat L	46.0	7.4	16.0	13.0	7.4	56.6
Cat K/Cat S	35.0	15.7	44.9	15.0	12.3	81.9
Cat S/Cat K	35.0	7.5	21.4	10.0	5.9	58.8
Thrombin/Factor Xa	91.0	27.3	30.0	42.0	27.3	65.0
Factor Xa/Thrombin	91.0	18.4	20.2	39.0	18.0	46.1
Trypsin/Factor Xa	63.0	11.0	17.5	15.0	10.0	66.9
Factor Xa/Trypsin	63.0	22.5	35.7	38.0	18.4	48.4
Trypsin/Thrombin	45.0	6.4	14.2	10.0	5.9	59.2
Thrombin/Trypsin	45.0	7.6	16.8	25.0	6.6	26.2

(b)

Table C.1: Continued.

MP–MFP	**PAC** [%]			**TSC** [%]		
	maxHR	**HR**	**RR**	**maxHR**	**HR**	**RR**
D1/D2	47.0	17.9	38.1	21.0	13.4	63.8
D2/D1	47.0	8.6	18.2	16.0	6.5	40.8
D1/D4	27.0	3.8	14.2	7.0	3.8	54.9
D4/D1	27.0	10.9	40.3	14.0	10.9	77.7
D2/D4	68.0	4.3	6.3	4.0	3.8	96.0
D4/D2	68.0	14.3	21.1	59.0	14.3	24.3
D3/D4	39.0	5.3	13.6	6.0	5.3	88.7
D4/D3	39.0	10.2	26.3	27.0	10.2	37.9
D2/5HT1a	29.0	4.2	14.3	5.0	4.2	83.2
5HT1a/D2	29.0	11.0	37.9	18.0	11.0	61.1
Alpha1/5HT1a	63.0	11.0	17.4	17.0	11.0	64.5
5HT1a/Alpha1	63.0	15.1	24.0	36.0	15.0	42.0
Cat B/Cat L	46.0	5.4	11.7	13.0	5.4	41.2
Cat L/Cat B	46.0	1.4	3.1	23.0	1.4	6.3
Cat B/Cat S	36.0	5.4	14.9	13.0	5.4	41.2
Cat S/Cat B	36.0	7.3	20.3	13.0	7.3	56.3
Cat L/Cat S	46.0	1.5	3.2	23.0	1.4	6.3
Cat S/Cat L	46.0	7.3	15.9	13.0	7.3	56.3
Cat K/Cat S	35.0	10.0	28.6	15.0	7.7	51.2
Cat S/Cat K	35.0	9.5	27.1	10.0	4.8	48.0
Thrombin/Factor Xa	91.0	26.1	28.7	42.0	26.0	62.0
Factor Xa/Thrombin	91.0	14.8	16.2	39.0	13.8	35.3
Trypsin/Factor Xa	63.0	9.8	15.6	15.0	8.8	58.4
Factor Xa/Trypsin	63.0	9.6	15.3	38.0	8.1	21.4
Trypsin/Thrombin	45.0	5.0	11.02	10.0	4.5	45.2
Thrombin/Trypsin	45.0	3.7	8.2	25.0	2.8	11.0

(c)

Table C.1: Continued.

PDR–FP	PAC [%]			TSC [%]		
	maxHR	HR	RR	maxHR	HR	RR
D1/D2	47.0	7.6	16.3	21.0	5.0	23.6
D2/D1	47.0	4.5	9.6	16.0	3.2	20.3
D1/D4	27.0	2.7	10.1	7.0	2.7	38.9
D4/D1	27.0	7.7	28.4	14.0	7.7	54.9
D2/D4	68.0	3.4	5.0	4.0	2.8	70.0
D4/D2	68.0	8.2	12.0	59.0	8.1	13.8
D3/D4	39.0	5.2	13.3	6.0	5.2	86.6
D4/D3	39.0	5.3	13.6	27.0	5.3	19.7
D2/5HT1a	29.0	3.9	13.5	5.0	3.9	78.4
5HT1a/D2	29.0	9.9	34.1	18.0	9.9	54.9
Alpha1/5HT1a	63.0	8.4	13.3	17.0	8.4	49.4
5HT1a/Alpha1	63.0	8.6	13.7	36.0	8.6	23.9
Cat B/Cat L	46.0	4.6	10.0	13.0	4.6	35.4
Cat L/Cat B	46.0	1.6	3.5	23.0	1.6	7.0
Cat B/Cat S	36.0	4.6	12.8	13.0	4.6	35.4
Cat S/Cat B	36.0	5.6	15.4	13.0	5.6	42.8
Cat L/Cat S	46.0	2.3	5.0	23.0	1.6	7.0
Cat S/Cat L	46.0	5.8	12.5	13.0	5.6	42.8
Cat K/Cat S	35.0	11.6	33.0	15.0	7.5	49.9
Cat S/Cat K	35.0	11.9	34.0	10.0	5.2	51.6
Thrombin/Factor Xa	91.0	26.3	28.9	42.0	23.2	55.2
Factor Xa/Thrombin	91.0	21.7	23.9	39.0	13.0	33.2
Trypsin/Factor Xa	63.0	19.9	31.6	15.0	12.9	86.1
Factor Xa/Trypsin	63.0	21.4	33.9	38.0	9.2	24.1
Trypsin/Thrombin	45.0	10.5	23.4	10.0	3.7	37.2
Thrombin/Trypsin	45.0	15.2	33.7	25.0	12.2	49.0

(d)

Table C.1: Continued.

TGT	**PAC** [%]			**TSC** [%]		
	maxHR	**HR**	**RR**	**maxHR**	**HR**	**RR**
D1/D2	47.0	14.0	29.8	21.0	10.2	48.8
D2/D1	47.0	5.5	11.7	16.0	4.8	30.0
D1/D4	27.0	3.0	11.0	7.0	3.0	42.3
D4/D1	27.0	9.0	33.5	14.0	9.0	64.6
D2/D4	68.0	5.3	7.8	4.0	3.2	80.0
D4/D2	68.0	10.5	15.4	59.0	10.3	17.5
D3/D4	39.0	3.0	7.8	6.0	3.0	50.6
D4/D3	39.0	9.0	23.3	27.0	9.0	33.6
D2/5HT1a	29.0	4.1	14.2	5.0	4.1	82.4
5HT1a/D2	29.0	9.0	30.9	18.0	9.0	49.8
Alpha1/5HT1a	63.0	8.2	13.1	17.0	8.2	48.5
5HT1a/Alpha1	63.0	14.6	23.2	36.0	14.6	40.6
Cat B/Cat L	46.0	6.4	14.0	13.0	6.4	49.6
Cat L/Cat B	46.0	1.2	2.6	23.0	1.2	5.2
Cat B/Cat S	36.0	6.4	17.9	13.0	6.4	49.5
Cat S/Cat B	36.0	6.3	17.6	13.0	6.3	48.6
Cat L/Cat S	46.0	2.3	5.0	23.0	1.2	5.2
Cat S/Cat L	46.0	6.3	13.7	13.0	6.3	48.6
Cat K/Cat S	35.0	12.9	36.8	15.0	7.8	51.7
Cat S/Cat K	35.0	7.3	21.0	10.0	5.0	50.0
Thrombin/Factor Xa	91.0	23.8	26.2	42.0	23.8	56.8
Factor Xa/Thrombin	91.0	10.2	11.2	39.0	10.0	25.6
Trypsin/Factor Xa	63.0	11.4	18.1	15.0	10.2	68.0
Factor Xa/Trypsin	63.0	14.1	22.3	38.0	13.9	36.5
Trypsin/Thrombin	45.0	5.6	12.5	10.0	4.0	39.6
Thrombin/Trypsin	45.0	4.4	9.8	25.0	4.0	15.8

(e)

Table C.1: Continued.

C.2 Ward's Clustering and DynaMAD

Level	Size	Small	5HT1a	5HT2a	Alpha1	D1	D2	D3	D4	Purity [%]
0	267	0	53	21	26	33	25	37	72	27
1	157	0	34	17	3	7	22	0	64	41
	110		19	4	23	16	3	37	8	34
2	54	0	1	4	0	14	8	0	27	50
	103		33	13	3	3	14	0	37	36
	24		0	0	3	1	0	20	0	83
	86		19	4	20	15	3	17	8	23
3	26	0	1	1	0	0	0	0	24	92
	28		0	3	0	14	8	0	3	50
	50		17	7	2	0	10	0	14	34
	53		16	6	1	3	4	0	23	43
	16		0	0	3	1	0	12	0	75
	8		0	0	0	0	0	8	0	100
	49		16	3	2	12	2	7	7	33
	37		3	1	18	3	1	10	1	49
4	10	1	1	0	0	0	0	0	9	90
	16		0	1	0	0	0	0	15	94
	9		0	2	0	2	2	0	3	33
	19		0	1	0	12	6	0	0	63
	18		3	4	1	0	7	0	3	39
	32		14	3	1	0	3	0	11	44
	32		10	4	1	3	4	0	10	31
	21		6	2	0	0	0	0	13	62
	6		0	0	2	0	0	4	0	67
	6		0	0	1	0	0	5	0	83
	5		0	0	0	0	0	5	0	100
	25		10	3	0	6	2	0	4	40
	16		6	0	1	0	0	7	2	44
	7		0	0	3	1	0	3	0	43
	10		0	1	7	0	1	1	0	70
5	10	4	1	0	0	0	0	0	9	90
	16		0	1	0	0	0	0	15	94
	9		0	2	0	2	2	0	3	33
	9		0	0	0	7	2	0	0	78
	7		0	0	0	4	3	0	0	57
	18		3	4	1	0	7	0	3	39
	14		2	0	1	0	3	0	8	57
	7		4	3	0	0	0	0	0	57
	8		8	0	0	0	0	0	0	100
	7		0	0	1	3	3	0	0	43
	6		6	0	0	0	0	0	0	100
	19		4	4	0	0	1	0	10	53
	4		0	1	0	0	0	0	3	75
	17		6	1	0	0	0	0	10	59
	6		0	0	2	0	0	4	0	67

Table C.2 – continued on next page

Table C.2 – continued from previous page

Level	Size	Small	5HT1a	5HT2a	Alpha1	D1	D2	D3	D4	Purity [%]
	4		0	0	0	1	0	3	0	75
	6		0	0	1	0	0	5	0	83
	5		0	0	0	0	0	5	0	100
	7		0	0	0	5	0	0	2	71
	6		3	3	0	0	0	0	0	50
	2		7	0	0	1	2	0	2	58
	8		0	0	1	6	0	0	1	75
	13		4	0	0	0	0	7	2	54
	9		0	0	7	0	0	2	0	78
	7		0	0	3	1	0	3	0	43
	10		0	1	7	0	1	1	0	70
	4		0	0	0	0	0	4	0	100
	7		3	0	1	2	0	0	1	43

Table C.2: Results of Ward's clustering. *Level* gives the clustering level in the cluster tree, and *Size* is the number of compounds in each cluster. *Small* reports the number of clusters containing one to three compounds. *Purity* gives the percentage of compounds of the most frequently occurring selectivity set in each cluster.

DEL	Descr	TSC	FSC	ZINC	HR [%]	RR [%]
5HT1a						
0	0.2	26.0	198.4	3 114 327.1	0.0	96.4
1	2.8	22.5	70.2	134 494.4	0.0	83.3
2	5.4	20.2	35.6	7 372.5	0.3	75.0
3	7.6	18.6	29.2	4 991.2	0.4	68.9
4	13.0	16.0	10.8	1 898.8	0.8	59.1
5	20.5	14.0	3.8	737.4	1.8	51.7
6	28.9	12.0	0.6	300.4	3.8	44.3
7	34.1	10.9	0.1	149.8	6.8	40.4
8	40.4	9.4	0.1	111.9	7.8	35.0
9	**47.6**	**8.5**	**0.1**	**85.4**	**9.0**	**31.4**
10	54.8	7.8	0.1	58.2	11.8	28.9
11	68.4	7.0	0.1	29.3	19.3	26.1
12	79.6	6.4	0.1	8.8	41.7	23.6
13	92.5	5.8	0.1	6.0	48.5	21.3
14	105.2	5.1	0.1	3.8	56.9	19.0
15	115.8	4.6	0.1	2.8	61.2	17.0
16	126.1	4.6	0.1	2.6	63.5	17.0
17	131.8	4.6	0.1	2.3	66.1	17.0
18	144.6	4.4	0.1	2.3	65.1	16.3
19	155.0	4.2	0.1	2.3	64.0	15.6
5HT2a						
0	0.9	9.0	130.7	1 512 766.0	0.0	81.8
1	3.7	6.0	32.9	23 970.9	0.0	54.2
2	8.6	4.9	3.0	2 515.4	0.2	44.4
3	12.4	4.2	0.4	792.4	0.5	37.8
4	15.0	3.6	0.4	719.1	0.5	32.7
5	18.9	3.2	0.2	308.6	1.0	28.7
6	**25.8**	**2.4**	**0.0**	**26.3**	**8.2**	**21.5**
7	31.7	1.5	0.0	10.0	13.2	13.8
8	37.0	1.2	0.0	3.3	26.8	10.9
9	47.5	0.8	0.0	0.6	55.6	7.3
10	56.2	0.6	0.0	0.2	80.0	5.8
11	66.6	0.5	0.0	0.1	86.7	4.7

Table C.3 – continued on next page

Table C.3 – continued from previous page

DEL	Descr	TSC	FSC	ZINC	HR [%]	RR [%]
12	79.1	0.3	0.0	0.0	88.9	2.9
13	90.6	0.3	0.0	0.0	88.9	2.9
14	101.8	0.2	0.0	0.0	85.7	2.2
15	113.4	0.1	0.0	0.0	100.0	1.1
16	120.9	0.1	0.0	0.0	100.0	1.1
17	129.8	0.1	0.0	0.0	100.0	1.1
18	142.4	0.1	0.0	0.0	100.0	1.1
19	155.0	0.1	0.0	0.0	100.0	1.1
Alpha1						
0	1.0	11.0	159.6	1 960 117.4	0.0	84.9
1	6.3	8.2	5.3	4 162.7	0.2	62.8
2	9.2	6.8	3.4	1 366.5	0.5	52.6
3	11.8	6.0	2.3	1 100.4	0.5	46.5
4	17.4	4.8	0.8	258.1	1.8	36.6
5	**24.4**	**3.6**	**0.0**	**66.0**	**5.2**	**28.0**
6	33.7	2.7	0.0	25.7	9.6	20.9
7	45.7	2.1	0.0	11.6	15.2	16.0
8	55.4	2.0	0.0	8.4	18.9	15.1
9	67.3	1.9	0.0	2.0	48.5	14.5
10	72.9	1.8	0.0	1.0	63.9	14.2
11	84.4	1.8	0.0	0.5	9.3	14.2
12	93.6	1.8	0.0	0.5	78.9	13.9
13	101.5	1.7	0.0	0.5	78.2	13.2
14	109.7	1.6	0.0	0.4	78.8	12.6
15	120.7	1.6	0.0	0.4	80.0	12.3
16	128.0	1.6	0.0	0.4	80.0	12.3
17	132.0	1.6	0.0	0.4	80.0	12.3
18	144.3	1.6	0.0	0.4	80.0	12.3
19	155.0	1.6	0.0	0.4	80.0	12.3
D1						
0	0.1	16.4	219.8	3 400 431.7	0.0	96.2
1	1.8	13.7	107.0	199 947.8	0.0	80.5
2	4.4	11.8	63.7	21 176.2	0.1	69.7

Table C.3 – continued on next page

Table C.3 – continued from previous page

DEL	Descr	TSC	FSC	ZINC	HR [%]	RR [%]
3	7.7	10.1	21.1	5 311.1	0.2	59.5
4	10.6	9.8	10.5	702.2	1.4	57.4
5	12.2	9.1	5.1	383.3	2.3	53.4
6	17.3	8.6	2.3	106.5	7.3	50.4
7	**20.7**	**7.8**	**1.5**	**77.6**	**9.0**	**45.9**
8	26.1	6.7	0.8	28.8	18.5	39.3
9	32.3	5.8	0.2	7.4	43.2	34.4
10	37.7	4.5	0.0	2.4	64.4	26.4
11	47.3	3.7	0.0	1.2	74.8	21.7
12	58.0	2.9	0.0	0.5	84.7	16.9
13	68.0	2.5	0.0	0.2	91.3	14.8
14	82.6	2.4	0.0	0.2	92.3	14.1
15	100.4	2.4	0.0	0.2	92.2	13.9
16	112.9	2.3	0.0	0.2	92.1	13.7
17	128.4	2.3	0.0	0.2	92.1	13.7
18	139.7	2.3	0.0	0.2	92.1	13.7
19	155.0	2.3	0.0	0.2	92.1	13.7
D2						
0	1.0	11.1	78.8	105 315.0	0.0	85.2
1	3.7	8.4	50.9	3 102.7	0.3	64.3
2	6.9	6.0	23.6	647.4	0.9	46.5
3	10.1	5.2	22.3	436.4	1.1	39.7
4	13.0	4.9	17.2	239.5	1.9	37.9
5	14.2	4.5	14.8	196.6	2.1	34.5
6	**17.7**	**3.9**	**8.9**	**48.9**	**6.4**	**30.2**
7	27.9	2.8	3.1	26.6	8.5	21.2
8	32.6	2.6	2.4	18.4	10.9	19.7
9	39.1	2.1	1.8	9.9	15.4	16.3
10	49.8	1.8	0.4	1.1	55.6	13.9
11	59.4	1.8	0.0	1.0	63.8	13.5
12	71.6	1.6	0.0	0.2	87.2	12.6
13	83.5	1.4	0.0	0.0	97.2	10.8
14	97.5	1.2	0.0	0.0	100.0	8.9

Table C.3 – continued on next page

Table C.3 – continued from previous page

DEL	Descr	TSC	FSC	ZINC	HR [%]	RR [%]
15	110.3	1.1	0.0	0.0	100.0	8.3
16	121.5	1.1	0.0	0.0	100.0	8.3
17	129.9	1.0	0.0	0.0	100.0	7.7
18	143.1	0.8	0.0	0.0	100.0	6.5
19	155.0	0.8	0.0	0.0	100.0	6.2
D3						
0	6.6	10.7	9.1	2 254.6	0.5	56.2
1	**23.3**	**8.0**	**5.4**	**24.4**	**21.1**	**41.9**
2	36.4	6.8	3.2	19.4	23.2	36.0
3	45.6	4.6	2.8	6.3	33.7	24.4
4	52.0	4.3	1.8	4.2	41.9	22.7
5	55.8	4.1	0.8	2.5	55.1	21.5
6	57.9	4.0	0.8	2.2	57.1	21.3
7	61.4	3.6	0.0	0.2	95.7	18.7
8	65.6	2.6	0.0	0.0	98.5	13.5
9	70.4	2.0	0.0	0.0	100.0	10.7
10	76.3	1.7	0.0	0.0	100.0	9.1
11	82.2	1.7	0.0	0.0	100.0	8.8
12	89.0	1.7	0.0	0.0	100.0	8.8
13	94.4	1.7	0.0	0.0	100.0	8.8
14	101.9	1.4	0.0	0.0	100.0	7.4
15	112.2	1.2	0.0	0.0	100.0	6.3
16	122.2	0.9	0.0	0.0	100.0	4.6
17	132.0	0.9	0.0	0.0	100.0	4.6
18	42.6	0.9	0.0	0.0	100.0	4.6
19	155.0	0.8	0.0	0.0	100.0	4.4
D4						
0	0.0	36.0	195.0	3 684 443.0	0.0	100.0
1	0.4	34.6	147.8	2 472 909.9	0.0	96.0
2	2.2	31.5	81.9	437 082.9	0.0	87.4
3	4.7	29.5	64.5	33 396.5	0.1	81.9
4	8.9	28.0	19.9	9 112.9	0.3	77.9
5	10.2	27.3	14.7	7 277.4	0.4	75.8

Table C.3 – continued on next page

Table C.3 – continued from previous page

DEL	Descr	TSC	FSC	ZINC	HR [%]	RR [%]
6	14.3	26.2	9.3	4 042.6	0.6	72.7
7	17.4	24.5	7.2	2 399.8	1.0	68.0
8	22.0	22.5	4.2	1 562.9	1.4	62.6
9	27.7	20.7	3.0	471.1	4.2	57.4
10	34.1	18.3	2.8	288.4	5.9	50.9
11	44.5	16.6	2.2	168.2	8.9	46.2
12	60.2	14.8	1.5	105.2	12.2	41.2
13	71.9	14.0	1.5	89.9	13.2	38.8
14	**90.1**	**12.5**	**1.0**	**61.8**	**16.6**	**34.8**
15	107.2	11.4	0.9	47.4	19.1	31.7
16	123.6	10.4	0.7	39.0	20.8	28.9
17	129.2	9.7	0.7	33.9	21.9	27.0
18	143.5	9.6	0.7	28.7	24.7	26.8
19	155.0	9.5	0.7	28.0	24.9	26.4

Table C.3: Average DynaMAD results over all dimension extension levels and selectivity sets. Results are reported for all dimension extension levels (*DEL*). Rows with dimension extension levels producing fewer than 100 compounds are given in bold. *Descr* reports the number of descriptors at the corresponding DEL, which is equivalent to the dimensionality of the chemical reference space at that mapping step. *TSC* and *FSC* stand for target– and family–selective compounds, respectively. Hit and recovery rates are calculated for target–selective compounds.