Computational Methods for the Integration of Biological Activity and Chemical Space

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

vorgelegt von
Eugen Lounkine
aus Moskau

Bonn
2009

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

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath

2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow

Tag der Promotion: 29.10.2009

Erscheinungsjahr: 2009

For my Parents

Abstract

One general aim of medicinal chemistry is the understanding of structure-activity relationships of ligands that bind to biological targets. Advances in combinatorial chemistry and biological screening technologies allow the analysis of ligand-target relationships on a large-scale. However, in order to extract useful information from biological activity data, computational methods are needed that link activity of ligands to their chemical structure.

In this thesis, it is investigated how fragment-type descriptors of molecular structure can be used in order to create a link between activity and chemical ligand space. First, an activity class-dependent hierarchical fragmentation scheme is introduced that generates fragmentation pathways that are aligned using established methodologies for multiple alignment of biological sequences. These alignments are then used to extract consensus fragment sequences that serve as a structural signature for individual biological activity classes.

It is also investigated how defined, chemically intuitive molecular fragments can be organized based on their topological environment and co-occurrence in compounds active against closely related targets. Therefore, the Topological Fragment Index is introduced that quantifies the topological environment complexity of a fragment in a given molecule, and thus goes beyond fragment frequency analysis. Fragment dependencies have been established on the basis of common topological environments, which facilitates the identification of activity class-characteristic fragment dependency pathways that describe fragment relationships beyond structural resemblance.

Because fragments are often dependent on each other in an activity class-specific manner, the importance of defined fragment combinations for similarity searching is further assessed. Therefore, Feature Co-occurrence Networks are introduced that allow the identification of feature cliques characteristic of individual activity classes. Three differently designed molecular fingerprints are compared for their ability to provide such cliques and a clique-based similarity searching strategy is established. For molecule- and activity class-centric fingerprint designs, feature combinations are shown to improve similarity search performance in comparison to standard methods. Moreover, it is demonstrated that individual features can form activity-class specific combinations.

Extending the analysis of feature cliques characteristic of individual activity classes, the distribution of defined fragment combinations among several compound classes acting against closely related targets is assessed. Fragment

Formal Concept Analysis is introduced for flexible mining of complex structure-activity relationships. It allows the interactive assembly of fragment queries that yield fragment combinations characteristic of defined activity and potency profiles. It is shown that pairs and triplets, rather than individual fragments distinguish between different activity profiles. A classifier is built based on these fragment signatures that distinguishes between ligands of closely related targets.

Going beyond activity profiles, compound selectivity is also analyzed. Therefore, Molecular Formal Concept Analysis is introduced for the systematic mining of compound selectivity profiles on a whole-molecule basis. Using this approach, structurally diverse compounds are identified that share a selectivity profile with selected template compounds. Structure-selectivity relationships of obtained compound sets are further analyzed.

Acknowledgments

I like to thank my supervisor Prof. Dr. Jürgen Bajorath for his guidance and help. I also would like to thank Prof. Dr. Michael Gütschow for his willingness to be the co-referent. Special thanks go to Dr. José Batista for his help and advice during the entire project and to Ye Hu and Felix Krüger for their collaboration on individual studies. Finally, I would like to thank all my colleagues from B-IT for the encouraging and friendly working atmosphere and all my friends who have supported me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In medicinal chemistry, small molecules, or *ligands* (usually $\leq$500 Da) are distinguished from macromolecular biological targets that they bind.[1] Ligands can be physiological mediators, like the neurotransmitter acetylcholine, xenobiotics, and drugs. For example, the drug Aspirin inhibits the enzyme cyclooxygenase, which is involved in inflammation processes. This inhibition leads to the anti-inflammatory effect of the drug. Targets are usually proteins, but also include nucleic acids or lipids.[1] A primary goal of medicinal chemistry is the identification and optimization of compounds that bind with high affinity and specificity to defined biological targets in order to induce a specific therapeutic effect.[1,2]

Target-, Ligand-, and Target-Ligand Space

Chemical space is estimated to theoretically contain up to 10^{60} organic molecules.[3,4] In the genomic era, disease-associated proteins are assessed as potential drug targets on a large-scale[4] and chemical library design approaches have increasingly shifted from diversity-oriented to target-focused design strategies.[5] Despite the growth of both target and chemical space, only a subset of the estimated 1,000-3,000 "druggable" targets[6] has been investigated by pharmaceutical industry.[7,8] The emerging interdisciplinary field of *chemogenomics* aims at establishing relationships between all possible targets and ligands. Therefore, target-, ligand-, and target-ligand spaces are defined.[4,9,10]

Proteins are often organized in target space based on their amino acid sequence or structural similarity.[4] However, in chemogenomics-related approaches, the primary focus lies on their binding sites, where the interaction with ligands takes place. Thus, targets can be classified according to binding site characteristics or based on the ligands they share. Since most ligands bind to a subset of related proteins in target space, there is a considerable overlap between these two classification schemes.[4,7]

In ligand space molecules are organized based on their chemical sim-

ilarity. Connectivity tables are a standard computational representation of molecular structure, which are a computationally accessible form of molecular graphs used by chemists to depict compounds. However, connectivity tables are difficult to compare using computational methods, because graph matching algorithms are of high computational complexity.[11] Therefore, compounds are often represented using descriptors that are amenable to fast computational comparison.[4,12,13] These can be classified into 1D, 2D, and 3D descriptors based on the dimensionality of the molecular representation that serves as the basis for their calculation. 1D descriptors can be calculated from the molecular composition formula and include, for example, molecular weight and atom counts. 2D descriptors are calculated from the connectivity table, i. e. the specific way in which atoms are connected. 2D descriptors are of varying complexity and range from substructure counts, over more complex shape indices that summarize bonding patterns, to abstract descriptors that are calculated using matrix operations on the atom adjacency matrix.[11] 3D descriptors are dependent on spacial atom positions and can be used to distinguish different molecular conformations. They include electrostatic field potentials and three-dimensional shape descriptors like solvent-accessible surface area. Another example are pharmacophore descriptors. Pharmacophores represent the specific three-dimensional arrangement of atoms or functional groups that are essential for receptor binding, specifically hydrogen bond donors and acceptors, charged groups involved in electrostatic interactions, and hydrophobic moieties accounting for van-der-Waals interactions with the receptor.[11] Molecular fingerprints are representations that are amenable to fast computational processing. They are usually bit strings where each bit accounts for the presence or absence of a defined structural feature or descriptor range. Descriptors of all dimensionalities and combinations of them can serve as the basis for fingerprint design.[11] Figure 1.1 summarizes the different descriptor types.

Target-ligand space considers the interactions between ligands and targets. A straightforward way to establish such a space utilizes compound-target affinity that is reported in form of binding constants (K_i) or functional effects, e. g. compound concentration at which half of the maximal effect is measured (EC_{50}) or inhibited (IC_{50}).[4] This data can be used to predict ligand affinity based on its affinity to other targets,[14] analyze structure-activity relationships between two targets having common ligands,[15] or predict global pharmacological profiles of compounds.[16] Replacement of affinities with interaction descriptors, e. g. structural interaction fingerprints, is possible.[4,17] These fingerprints capture information about functional groups of a ligand, the amino acids of the protein's binding site, and their specific interactions. Figure 1.2 summarizes the relationships between target, ligand, and target-ligand space.

Using chemogenomic approaches, in a retrospective study, a drug-target network has been established that integrates information about approved drugs,

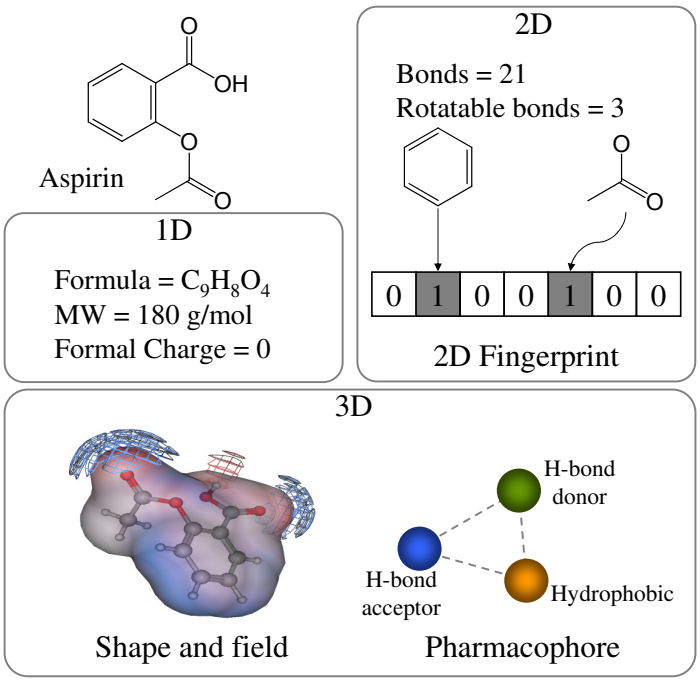


Figure 1.1: Molecular descriptors. Examples of different descriptors including fingerprint representations are shown for Aspirin. Based on the dimensionality of the molecular representation for which they are calculated, descriptors are classified as 1D, 2D, and 3D descriptors.

their targets, the protein-protein interactions between drug targets, and the underlying disease-related genes.[8] The analysis of this network has revealed that most drugs act on well-known protein targets and in many cases do not target proteins that are directly involved in pathogenesis. This trend is evident for established drugs that do not yet target defined disease-associated genes, e. g. oncogenes, which are involved in cancerogenesis.[8] In a related approach, approved drugs and their targets have been organized based on phenotypic side-effect similarities.[18] Using the so defined networks, unexpected drug-target interactions have been predicted.[18]

Structure-Activity Relationships

The projection of molecules into ligand space or target-ligand space, focuses on related yet distinct properties of biologically active compounds. In ligand space, chemical properties of ligands are compared and clusters of molecules identified.[4,13] Projection into target-ligand space does not consider chemical resemblance of ligands, but rather similarity in their biological activity. Thus, from a ligand-centric point of view, *chemical* and *activity* space can be defined for biologically active compounds.

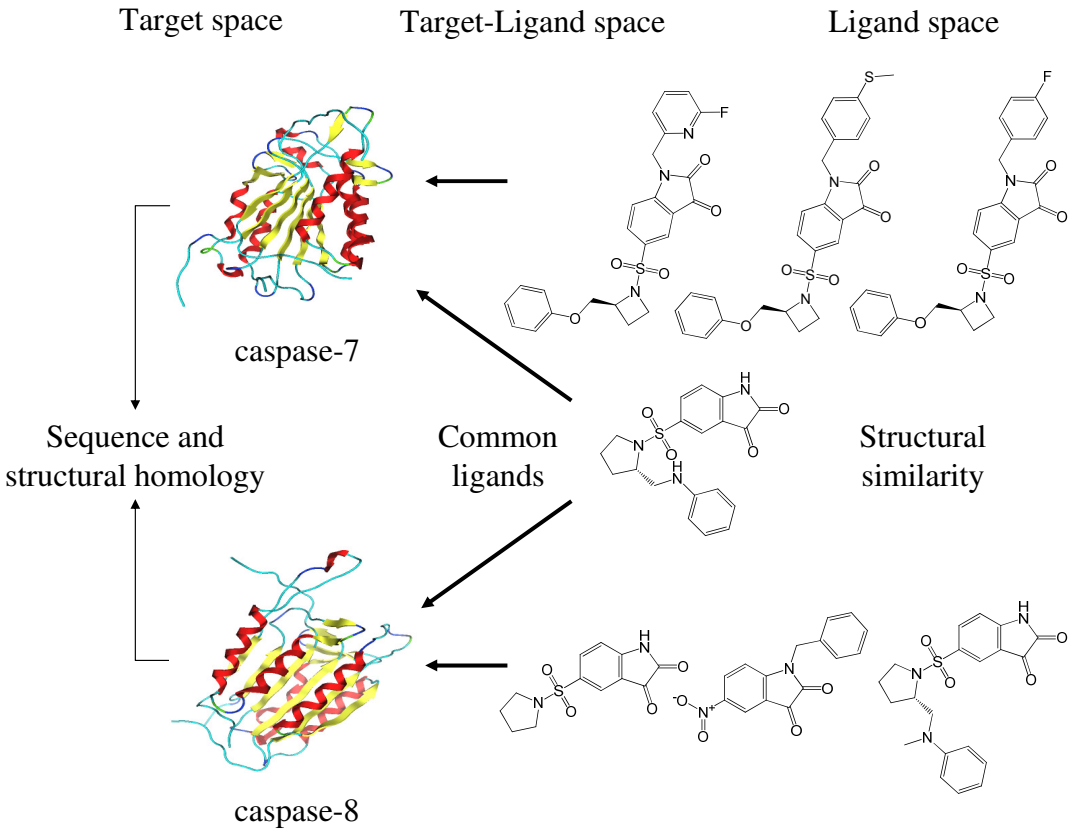


Figure 1.2: Target-, ligand-, and target-ligand space. Caspase-7 and caspase-8 are compared in target space based on their primary sequence and structural homology. On the right, ligands are shown that bind to one or both targets. In ligand space, molecules are organized based on their chemical resemblance. Targets and ligands are projected into target-ligand space where targets can be compared based on shared ligands.

In chemical space, ligands are compared based on their atom bonding patterns and derived properties and a variety of *chemoinformatics* methods have been developed to assess structural similarity of molecules.[12,13] For example, the *Tanimoto coefficient* (Tc)[13] represents a widely accepted metric of molecular similarity based on fingerprint overlap. Distinct from structural similarity, compound similarity in activity space is calculated from affinity profiles against a set of targets.[10,14,19] Activity profile similarity can be used, for example, to predict missing potency values.[4]

Ultimately, the specific physical and chemical properties of a compound define its interaction with a particular macromolecule.[1,20] This notion has lead to the formulation of the *similarity property principle*, which states that structurally similar compounds are likely to have similar biological activity.[21] However, structure-activity relationships (SARs) are often complicated. On the one

hand, structurally distinct compounds can bind to the same target; on the other hand, small structural changes can lead to great differences in biological activity.[22–24] Ligand-target space can thus be viewed as an "activity landscape". Structurally similar compounds with big activity differences constitute activity cliffs, whereas gradual changes in structure leading to only moderate activity differences correspond to smooth regions.[25] Optimization efforts exploit activity cliffs in order to increase compound potency by introducing only small structural changes. Different SAR types, such as continuous and discontinuous SAR have been quantified based on pair-wise molecule comparison in both activity and chemical space.[23,24] For example, the SAR Index (SARI)[23] relates differences in compound activity to the structural similarity of two compounds. Figure 1.3 shows an example of continuous and discontinuous SARs.

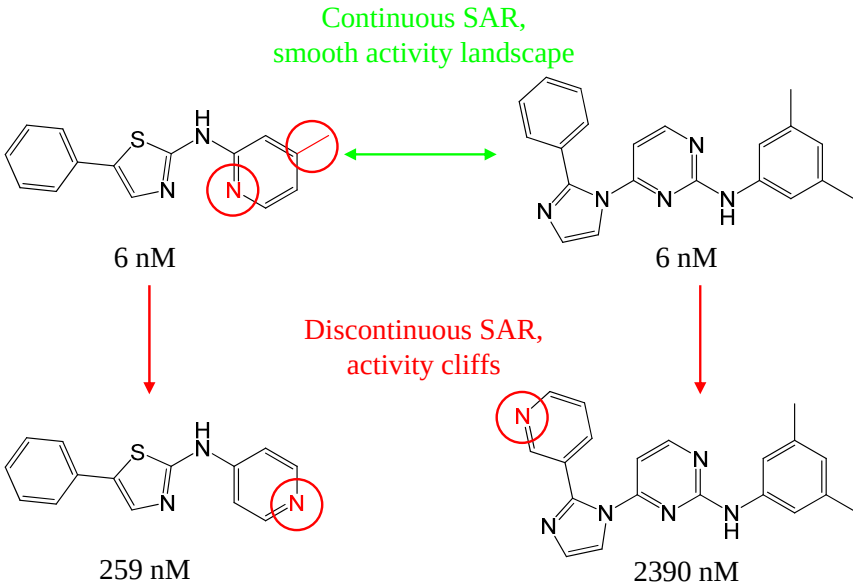


Figure 1.3: Activity cliffs. Four inhibitors of vascular endothelial growth factor receptor (VEGFR-2) tyrosine kinase are shown. The two upper structures are dissimilar, but neighbors in activity space. By contrast, structural analogues (bottom) of each of the ligands show greatly reduced activity, thus constituting an activity cliff. Structural differences between analogues are highlighted in red. The figure is adopted from Peltason and Bajorath.[23]

Recent efforts have focused on systematic exploration and visualization of activity landscapes.[24,26–28] In network like similarity graphs different layers of information are incorporated, including compound potency, pairwise structural similarity and SAR (dis)continuity.[26] This allows the identification of representative compounds for distinct local SARs and the identification of potential compound optimization pathways in biological screening data.[27] Utilizing different measures for the assessment of structural ligand similarity, *consensus*

activity cliffs have been identified that are recurrent for different structural representations of compounds.[28]

Integration of Activity and Chemical Space

For ligand-based identification of novel compounds with a defined biological activity, the integration of chemical and activity space is essential.[22,29] Chemical similarity searching[13] relies on the similarity property principle and assumes that neighbors in chemical space will also be closely related in activity space.[12] In ligand based virtual screening, database compounds are prioritized based on their similarity to active reference molecules. General chemical space designs as well as similarity metrics have been established that can identify active compounds based on reference ligands of the same or related targets.[12,29,30]

Going beyond the similarity property principle, methods have been developed that use structural information of active reference compounds for the design of activity class dependent chemical spaces and similarity metrics.[22] Current methodologies for the integration of compound activity and structural similarity often rely on reference compound sets representing clusters in activity space.[31,32] For example, in Bayes Affinity Fingerprints[31] the assessment of compound similarity is carried out in a two-step process. Descriptors constituting the chemical space are first transformed to a hypothetical affinity profile by calculating the structural similarity of a compound to a wide panel of activity classes. The scores from individual activity class models then constitute a vector that is used to compare molecules. This allows the definition of a global low-dimensional bioactivity space that is characterized by prototypic activity classes representing distinct target families.[31] In a reverse approach, mapping of structural similarity to activity space organizes biological targets based on the structural resemblance of their ligands. These ligand-based target networks can be used to predict off-target affinities of known drugs.[33,34]

A more structure-centered approach to the integration of activity and chemical space is the annotation of structural descriptors with activity information. The notion of privileged substructures allowed linking structural motifs to different target families,[5,35,36] including G Protein Coupled Receptors (GPCRs)[37] and kinases.[38–40] A privileged substructure can be defined as "a substructure/scaffold exhibiting strong preferences for a particular area of the target space (for example, GPCRs) and suitable to orient the design of targeted compound libraries".[38] Furthermore, recurrent structural motifs and combinations of molecular fragments have been identified in known drugs[41,42] and compound libraries.[43,44]

Extending the concept of privileged substructures, individual activity classes can be mined for activity class-characteristic features. This can be done on the basis of predefined or hierarchically generated substructures,[45,46] random

fragment populations,[47,48] or through exhaustive substructure mining.[49] Also, structural motifs have been identified that distinguish compounds with different potency against defined targets.[50]

In summary, searching for bioactive compounds and SAR assessment benefit from the integration of activity and chemical space. This integration can be based on structural similarity of test compounds to active reference molecules. However, the similarity property principle underlying this concept is often not sufficient to fully describe complex SARs. Therefore, molecular descriptors should be designed and evaluated in a compound activity-sensitive manner.[22]

Goals and Approaches

The primal aim of this thesis project was the development of computational methods for the integration of biological activity and chemical ligand spaces.

First, it was assessed how activity class-specific structural information can be used to guide systematic fragmentation of compounds. Therefore, an activity-class directed fragmentation approach was introduced.

Then, the organization of defined structural descriptors was analyzed based on activity criteria. The topological environment of chemically intuitive fragments in molecules was quantified, which allowed their organization in hierarchies based on co-occurrence of fragments in active compounds.

In order to assess the significance of feature combinations for the identification of active compounds in different chemical space designs, activity class-specific feature combinations were systematically extracted from three distinct molecular fingerprints and applied to virtual screening.

Furthermore, the distribution of molecular fragment combinations among compounds with different biological activity and potency was explored. Therefore, Formal Concept Analysis (FCA) was adapted, a data mining technique from information theory.[51] *Fragment Formal Concept Analysis* (FragFCA) has been designed for the mining of molecular fragment combinations specific for defined activity and potency profiles.

Extending this approach to the molecular level, *Molecular Formal Concept Analysis* (MolFCA) has also been developed. The method has been designed to systematically explore activity space with a particular focus on compound potency and selectivity. MolFCA identifies compounds satisfying complex selectivity profiles in activity space. The identified compounds can then be used to assess structure-selectivity relationships.

Thesis Outline

In *Chapter 2*, an overview of currently used molecular fragmentation methods is provided. Four major fragmentation schemes, i. e. systematic / hierarchical, knowledge-based, retrosynthetic, and random are presented. Furthermore, it is described how random fragment populations are mined for substructures that are characteristic of defined activity classes. *Core Trees* are introduced as an activity class-directed hierarchical fragmentation scheme. From Core Trees, fragment pathways are extracted and aligned using a multiple sequence alignment algorithm, yielding *Consensus Fragment Sequences* (CFS) that can be used as signatures for individual activity classes.

In *Chapter 3*, the *Topological Fragment Index* (ToFi) is presented, which assesses the topological environment of defined substructures within active compounds. Using this index, fragments generated based on retrosynthetic criteria have been organized in hierarchies that reflect fragment co-occurrence in compounds with different biological activities. These hierarchies allow the identification of fragment topology clusters that are characteristic of individual activity classes.

Chapter 4 reports the development and application of *Feature Co-occurrence Networks* (FCoN). FCoN are utilized to systematically extract molecular feature cliques of varying size that are characteristic of individual activity classes. Feature cliques are prioritized using information about their distribution in a background database and utilized for virtual screening. Three fingerprint representations of molecules have been assessed for their potential to provide activity class characteristic feature combinations.

In *Chapter 5*, Formal Concept Analysis (FCA) is described and Fragment Formal Concept Analysis (FragFCA) is introduced, which allows mining of molecular fragment combinations that are specific to defined activity and potency profiles of active compounds. FragFCA has been applied to GPCR ligands with partially overlapping activity against seven targets. Furthermore, the design and evaluation of a compound activity classifier based on fragment combinations identified by FragFCA is reported.

Chapter 6 introduces Molecular Formal Concept Analysis (MolFCA) for selectivity profile mining in biologically annotated databases. MolFCA allows systematic exploration of activity space including multiple biological activities. MolFCA queries assess compound potency and selectivity against multiple targets. Compounds are identified in a structurally unbiased manner, yielding diverse molecules. These compound sets can then be assessed for SARs and structure-selectivity relationships (SSRs).

Chapter 2

Molecular Fragmentation Approaches

This chapter provides an overview of four major types of molecular fragmentation approaches: knowledge-based, systematic/hierarchical, retrosynthetic, and random. It further describes how random fragment populations can be mined for *Activity Class Characteristic Substructures* (ACCS) and introduces *Core Trees* as an activity class-directed fragmentation and organization scheme.

2.1 Historical Overview of Fragment Design

Molecular fragments have a long history as structural descriptors. They are chemically intuitive and can be much easier understood than many other more complex mathematical models of chemical structure and properties. Most importantly, however, given the simplicity and intuitive nature of their design, substructures and fragment descriptors are surprisingly powerful in analyzing and predicting SARs. This is very likely the case because these types of descriptors implicitly capture much chemical information.[52–54]

The introduction of molecular fragments as tools for chemical data analysis dates back to the 1950s, when fragment collections were generated on the basis of topological criteria, i.e. by adding layers of bonded atoms to preselected central atoms.[55] These so-called atom-centered fragments were originally applied to estimate physical properties of synthetic molecules such as, for example, P(o/w), the octanol-water partition coefficient, a measure of hydrophobicity.[56] Property estimation was often attempted by addition of known values for substructures forming a molecule. The distribution of atom-centered fragments in chemical databases was first studied in the early 1970's[57] and later on fragment descriptors were used to associate small molecules with biological activities.[20,58] During the same decade, methods were introduced for the systematic generation of sets of atom- or bond-centered fragments that occur with

a certain frequency, or equal frequency, in a database[59,60] During the 1970s, atom- and bond-centered fragments were also first encoded as bit strings,[61,62] and these fingerprint representations of molecular structure have continued to be one of the most widely used descriptor formats for chemical similarity searching to this date, such as, for example, the set of 166 publicly available MACCS structural keys[63] or the BCI standard dictionary (1,052 fragments).[64]

2.2 Molecular Fragmentation Approaches

Four principal approaches to the generation of fragments from 2D molecular representations can be distinguished: knowledge-based, systematic / hierarchical, retrosynthetic, and random. All of these methodologies have in common that they operate on the connectivity tables of molecules. However, the individual fragmentation strategies are distinct from each other and tailored towards different applications.

2.2.1 Knowledge-Based Molecular Fragmentation

Knowledge-based methods make use of chemical and pharmaceutical expertise to design substructures. For example, knowledge-based dictionaries have been introduced for the prediction of ADME[a] properties of bioactive substances[64] or the removal of compounds with reactive or toxic fragments from screening sets.[65–67] Figure 2.1 shows examples of reactive groups that are often avoided in the design of pharmaceutical compound libraries.

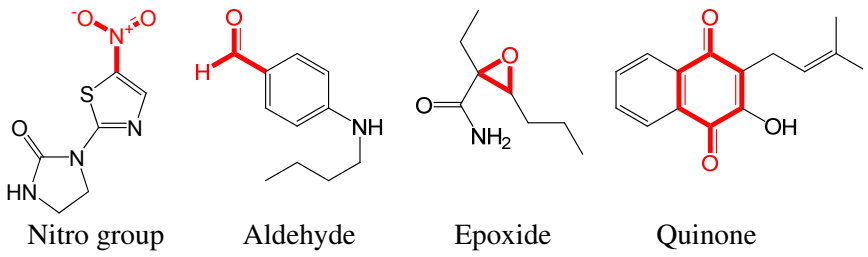


Figure 2.1: Examples of non-drug-like chemical groups. Four compounds are shown that contain chemical entities (highlighted in red) that are undesired in the design of drug-like compounds. The examples are taken from Walters *et al.*[67]

Another example of knowledge-based design is the definition of privileged substructures that are recurrent in compounds active against members of therapeutic target families such as GPCRs[37] or protein kinases.[38–40]

[a] ADME stands for Absorption, Distribution, Metabolism, and Excretion. ADME properties characterize the pharmacokinetics of a compound.

2.2.2 Hierarchical and Systematic Fragmentation

Systematic approaches to generate atom- and bond-centered fragments were one of the origins of molecular fragmentation, as discussed above. Atom-centered fragments are illustrated in Figure 2.2A. They are generated by adding layers of bonded atoms to preselected central atoms. Today this type of fragments serves as the basis for the design of fingerprints that capture strings of layered atom environments and are currently among the state-of-the-art similarity search tools.[68,69] For example, *extended connectivity fingerprints* (ECFP) represent molecules as ensembles of atom-centered fragments that are encoded as integers using a hash function.[59]

Atom pairs represent another pioneering development of fragment-type descriptors that were systematically derived following topological criteria.[70] These descriptors are of the form $AT_i - Dist_{ij} - AT_j$, where $Dist_{ij}$ is the length of the shortest bond path between an atom of type AT_i and another one of type AT_j. Often atom types correspond to the element of the atom. Further refined atom types encode the element as well as the number of attached non-hydrogen atoms and the number of π-bond electrons.[71] A recent study has shown that atom pair descriptors can be utilized for encoding virtual combinatorial libraries (i. e. monomers that can be theoretically combined in various ways) without the necessity of enumerating all virtual compounds.[72] Figure 2.2B shows examples of atom pair descriptors.

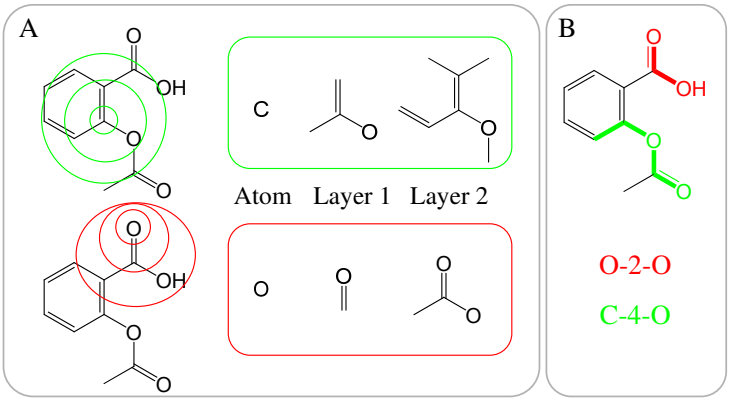


Figure 2.2: Atom-centered fragments and atom pairs. (**A**) Examples of atom-centered fragments of Aspirin are shown that were derived using different numbers of atom layers. (**B**) For Aspirin, examples of atom pairs are provided that correspond to pathways in the molecule.

Hierarchical fragment design strategies introduced by Bemis and Murcko[41,42] focus on structural elements that are thought to be important for drug design and SAR analysis. Molecular graphs are decomposed by distinguishing ring assemblies, linkers connecting these ring assemblies, and side chains. Rings and linkers together form molecular scaffolds that are often regarded as central

building blocks in medicinal chemistry. Different levels of abstraction can be applied to represent core structures. In the context of hierarchical methods, heteroatom-containing core structures without side chains (functional groups) are regarded as scaffolds and molecular frameworks are obtained from scaffolds by replacing all heteroatoms with carbon atoms. Furthermore, bond types can be reduced to single bonds in order to compare molecules on the basis of *carbon skeletons*, which represent the highest level of structural abstraction. Figure 2.3 illustrates the hierarchical fragmentation of a compound and application of different abstraction levels. A refined hierarchical fragmentation scheme has been proposed for the organization of scaffolds present in virtual compound libraries.[45] It disintegrates molecular frameworks and ring assemblies based on a set of elaborate rules. The so obtained "Scaffold Trees" have been annotated with biological activity allowing mining for activity class-prevalent scaffolds.[45,73]

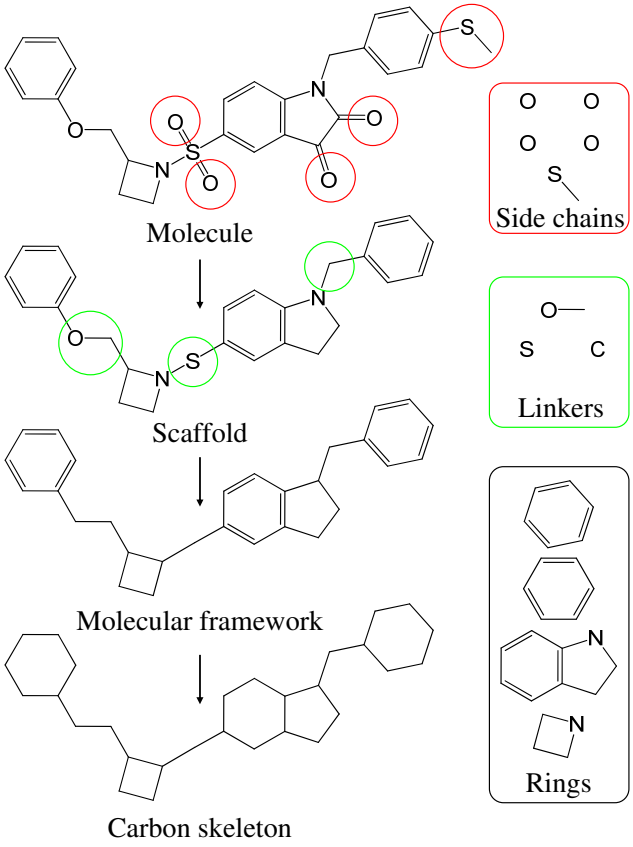


Figure 2.3: Hierarchical fragmentation and scaffold abstraction. Hierarchical fragmentation of an exemplary compound is shown. The scaffold is obtained by deleting all side chains (red). It is decomposed into rings (black) and linkers (green). A molecular framework is obtained by replacing all heteroatoms by carbons and a carbon skeleton is generated by setting all bond orders to single bonds.

2.2.3 Fragmentation Based on Retrosynthetic Criteria

A prominent example of fragmentation design strategies that use retrosynthetic criteria is the *Retrosynthetic Combinatorial Analysis Procedure* (RECAP), which has been introduced in order to provide fragment libraries that are suitable as the basis for the design of combinatorial libraries.[74] Therefore, fragments in RECAP are generated by breaking bonds that are formed by common chemical reactions. The identification of frequently occurring RECAP fragments in compounds with defined biological activities provides guidelines for the generation of combinatorial libraries that are tailored towards an activity of interest. Originally, eleven cleavable bond types have been defined. However, this list has been extended, for example, in the Molecular Operating Environment (MOE[b]). Figure 2.4A reports the original bond type definitions and Figure 2.4B depicts five additional bond types introduced in MOE. The RECAP fragmentation of an exemplary compound is illustrated in Figure 2.4C. In order to specify the attachment points of a fragment, isotope-like labels are used that represent the bond class of individual atoms. Note that RECAP fragmentation does not have to be complete, i. e. only a subset of possible bond cleavages can be applied to generate valid RECAP fragments.

2.2.4 Random Fragmentation Methods

Random fragmentation approaches deliberately depart from knowledge-based or systematic fragmentation schemes. This allows the assessment of chemical information content[75] and mining for activity class characteristic substructures in an unbiased manner, without the need of exhausting fragment enumeration.[47,76]

Brownian processing has been used to assess and quantify the information content of organic molecules. In Brownian processing,[75] so-called tape recordings of random walks (Brownian motion) through molecular graphs are generated. Two types of processing are distinguished: sequential and parallel. In sequential processing, each step is recorded in the order atoms are visited as a code unit. In parallel processing, code units of a given length are extracted from random walks (e.g. "H-C-C-H"). These units represent random substructures of a compound that correspond to topological pathways. Brownian processing is illustrated in Figure 2.5. On the basis of tape recordings, the chemical information content of molecules can be compared by analyzing the frequency with which substructures up to a certain length (e.g. atom pairs, triplets, etc.) occur and by determining their overlap. Complex molecules producing code units of high diversity show high chemical information content, whereas structures that are topologically less complex contain less chemical information.

[b] An overview of the software and databases used in this study is provided in Appendix A.

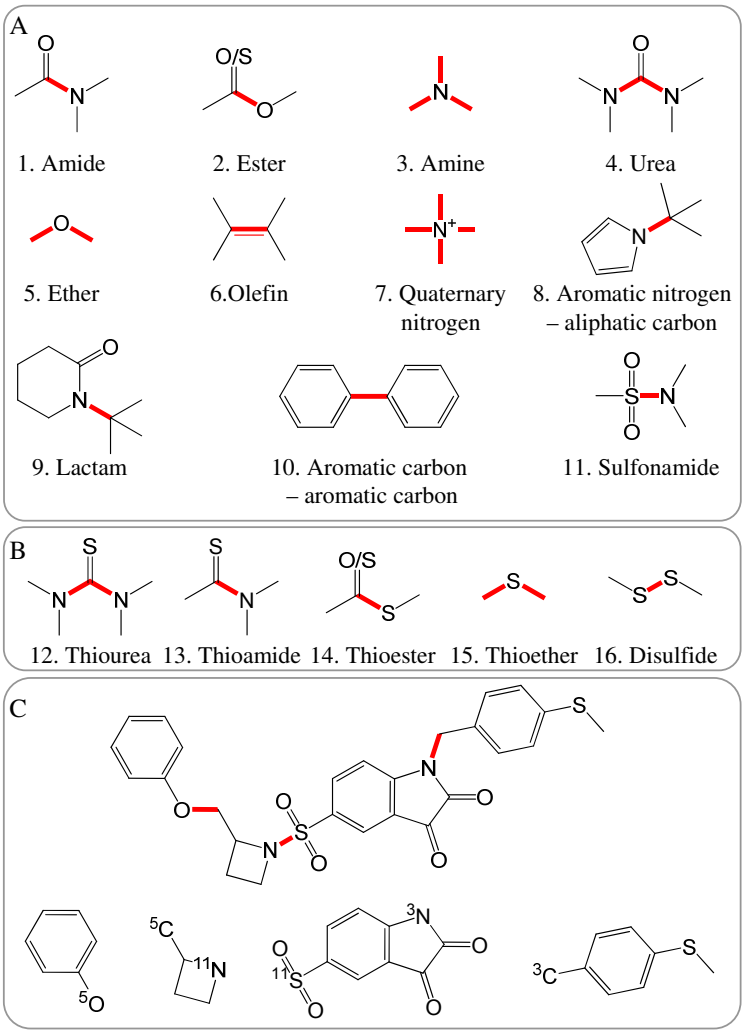


Figure 2.4: RECAP fragmentation. (**A**) The eleven original RECAP bond types are shown. (**B**) shows five additional bond types introduced in MOE. (**C**) An exemplary compound is fragmented by cleaving three bonds (red). Isotope-like labels of atoms identify the RECAP bond type.

A random fragmentation approach termed *MolBlaster* has been introduced[76,77] that generates fragment populations by iteratively deleting a number of randomly chosen bonds in the hydrogen-suppressed molecular graph. Figure 2.6 illustrates the MolBlaster fragmentation protocol. For the computational processing of fragments, they are represented as strings in Simplified Molecular Input Line Entry Specification (SMILES) notation.[78] The SMILES notation encodes the connectivity table as a linear string of element labels (like C, N, O, Cl) and special characters that encode branches and rings. Canonical SMILES[79] strings are used, which has the advantage that identical fragments

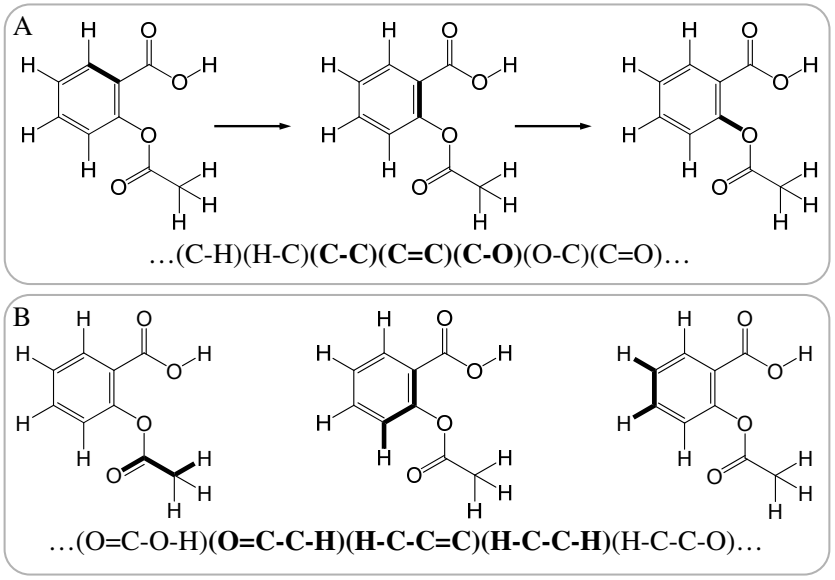


Figure 2.5: Brownian processing. Serial (**A**) and parallel (**B**) Brownian processing of Aspirin is illustrated. While serial processing records atom pairs that are visited in sequence, parallel processing yields independent code units of a predefined length (here four atoms).

can be found by string comparison. MolBlaster allows fragmentation in an unbiased manner, because all bonds have equal probability to be cleaved during one iteration. Fragments generated from multiple iterations are sampled and the fragment frequencies recorded. It has been shown that 2,000 iterations are sufficient to produce stable fragment populations of individual molecules that show only minor fluctuations in fragment composition.[76,77]

It has also been demonstrated that random fragment populations can be used to assess molecular similarity and structure-activity relationships of bioactive compounds based on histogram comparison of fragment distributions.[76] Furthermore, histogram comparison has been applied to database searching, which has revealed that random fragments preferentially occur in given activity classes.[77] Accordingly, fragment populations derived from sets of bioactive compounds have been filtered against a background fragment population of inactive molecules in order to extract *Activity Class Characteristic Substructures* (ACCS). ACCS have been defined as fragments that occur in at least two active compounds but no background molecules.[48,80] Figure 2.7 illustrates the derivation of ACCS.

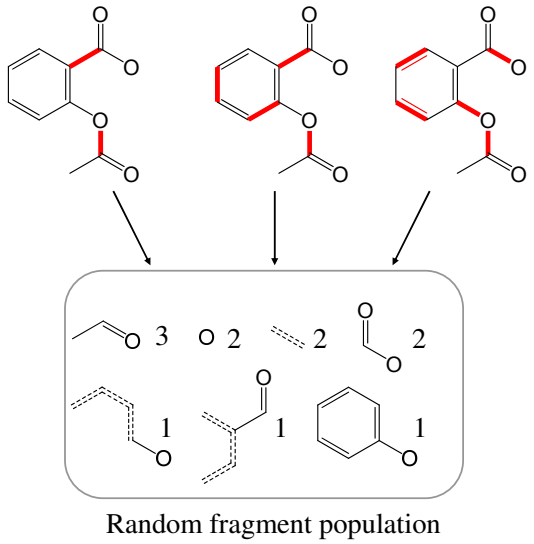


Figure 2.6: MolBlaster fragmentation. Three MolBlaster fragmentation iterations are shown for Aspirin with increasing numbers of bonds deleted. Cleaved bonds are highlighted in red. Dashed bonds represent fragmented aromatic systems. The fragments from all iterations are sampled and fragment frequencies recorded.

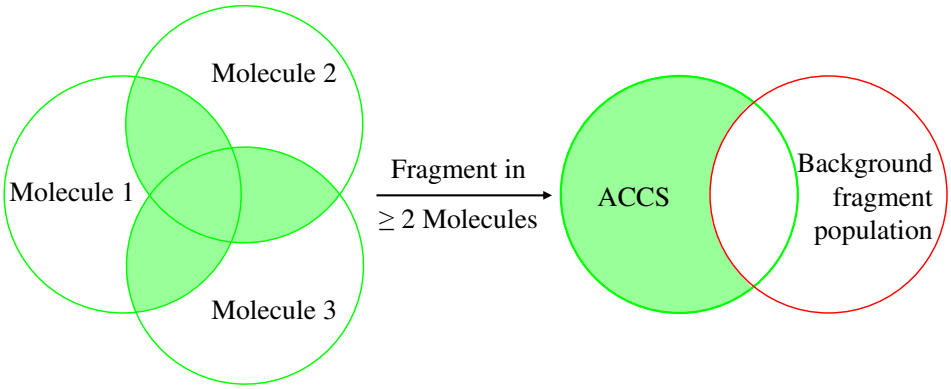


Figure 2.7: Activity class characteristic substructures. Random fragments are generated from a set of active reference compounds. Circles represent random fragment populations. Fragments that occur in at least two of the reference molecules are retained. The resulting population is compared to a background fragment population (red circle).

2.3 Activity Class Directed Fragmentation

This section describes how ACCS isolated from random fragment populations are used to map molecular core structures in an activity class-dependent manner. Mapped cores represent the origin of ACCS extracted from molecules using MolBlaster and discriminate between conserved core regions and variable peripheral parts of active compounds.

Core Trees are introduced that are based on core mapping and represent an activity class-directed hierarchical fragmentation scheme. From core trees, activity class-dependent fragment pathways are extracted that are amenable to sequence alignment. *Consensus Fragment Sequences* (CFS) are then derived from multiple core path alignments that serve as activity class signatures.

2.3.1 Molecular Core Mapping

The structural origin of ACCS has been assessed in a study using molecular maps of ACCS overlap in active compounds.[80] The study has revealed that overlapping ACCS form coherent molecular cores. Core mapping provides a molecular map of structural hot spots that confer activity class characteristic information encoded in random fragment populations. Figure 2.8 illustrates the core mapping procedure. For each ACCS, atom counters of the molecule are increased by one for each successfully mapped atom. After all ACCS have been mapped, *atom match rates* are calculated by dividing the counters by the number of successfully matched ACCS.

The characteristic distribution of atom match rates produced by a set of ACCS yields molecular cores by grouping atoms together that exceed a predefined match rate threshold, e. g. atoms with a match rate ¿ 80%. By applying different thresholds, 10 cumulative cores have been defined for active compounds. Cores can be formed in a systematic way for different activity classes and represent coherent structural entities that are characteristic of individual classes.[80]

2.3.2 Core Trees

The Core Tree methodology comprises five consecutive steps. First, an activity class-directed hierarchical fragmentation scheme is applied that utilizes atom match rate distributions in molecules with a mapped core. This fragmentation scheme separates peripheral parts of molecules from conserved core regions. Then, *Core Trees* are described for the organization of all generated fragments in a tree structure based on iterative hierarchical disintegration of fragments. In Core Trees, fragments are annotated with average match rates of their atoms. This allows the distinction of conserved core fragments from variable parts of the molecule. In a following step, core paths are extracted from Core Trees that

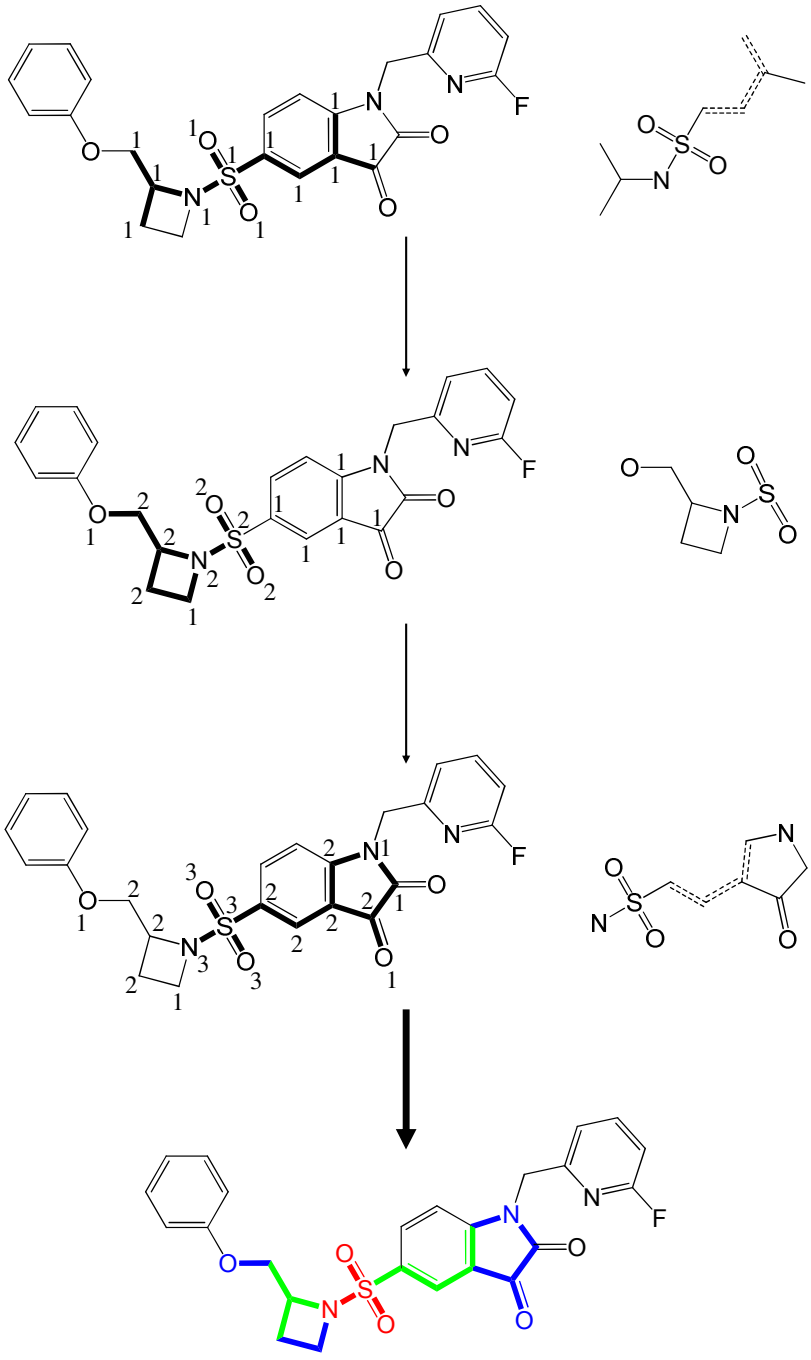


Figure 2.8: Molecular core mapping. The mapping of three exemplary ACCS (right) onto a compound is shown. Dashed double bonds indicate bonds in aromatic rings. For each mapped ACCS, atom counters of matched atoms are increased (small numbers; zero counters are omitted for clarity). Atom match rates are calculated by dividing each counter by the total number of successfully mapped ACCS (here three). Cores are defined as sets of atoms with match rates exceeding a predefined threshold. In this example, three cumulative cores can be distinguished: ¿ 66%, red; ¿ 33%, red and green; ¿ 0%, red, green, and blue (structure at the bottom). Black atoms are not mapped by any ACCS.

represent conserved core regions of the molecule in form of defined fragment sequences. Core paths contain fragments of decreasing specificity and size, but increasing conservation within active compounds, and hence balance fragment specificity and conservation.

In order to compare core paths, methods for the global alignment of biological sequences are applied. Therefore, individual core path fragments are compared using a fragment similarity scoring function. Corresponding fragments are identified and molecular similarity quantified based on core path alignments. Core paths of individual activity classes are then organized in multiple core path alignments. From multiple alignments, Consensus Fragment Sequences (CFS) are derived that combine core path fragments from the entire activity class and organize them in an alignment-specific manner. Thus, in a last step, activity class signatures are extracted in form of CFS.

Data Sets

Core Trees have been generated for a previously published set of 1,025 compounds that was also used to evaluate core mapping.[80] In this set, individual compounds are annotated with biological activity and grouped into 45 activity classes. Random molecular fragment populations were generated by applying the MolBlaster fragmentation procedure with 3,000 iterations per molecule. ACCS have been isolated by filtering against a background database of 2,000 randomly selected and fragmented molecules from ZINC.[c] ACCS have then been mapped onto the compounds they originated from and atom match rates calculated as described above. Cores were successfully mapped for ∼95% of the compounds.

Core-Oriented Molecular Fragmentation

Core mapping of ACCS onto active compounds produces characteristic atom match rate distributions. Atom match rates distinguish between molecular regions that are characteristic of an activity class (high match rates) and regions that are variable (low atom match rates). In order to separate conserved core regions from variable peripheral regions of a molecule, atom match rates are systematically compared. Bonded atoms with high match rate difference (ΔMR) belong to distinct regions of the molecule, while bonds with low ΔMR constitute coherent regions. Figure 2.9 illustrates match rate distributions of a molecule with a mapped core.

Cleavage of bonds with high ΔMR separates core regions of a molecule from peripheral fragments. The resulting fragments are further divided by deleting bonds with maximal ΔMR. Thus, sorting of bonds by decreasing

[c] ZINC is a publicly available database of small weight molecules (see Appendix A).

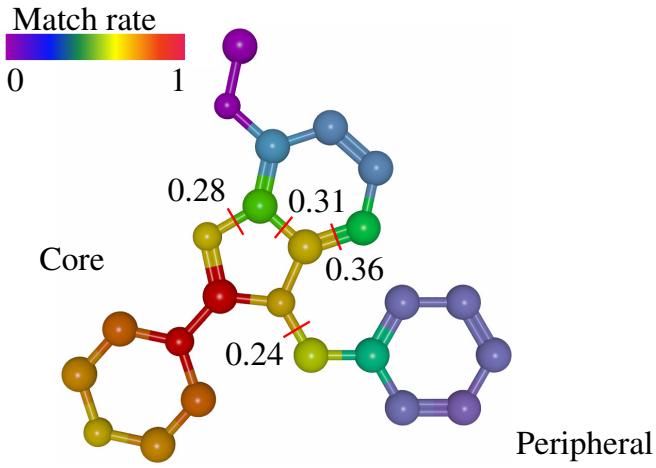


Figure 2.9: Core-based fragmentation. The core mapping for a serotonin receptor antagonist is shown. Match rates are encoded as a color spectrum. The core region is red, whereas peripheral regions are shown in purple. Individual bonds are annotated with match rate differences. Bonds are iteratively cleaved starting with bonds for which match rate differences are maximal. Thus, peripheral fragments are separated from the core region.

ΔMR defines a hierarchical fragmentation scheme. Bonds with high ΔMR are cleaved first, whereas bonds connecting atoms with equal match rates are cleaved last.

Core-oriented molecular fragmentation combines activity class-specific information encoded in random fragment populations with a hierarchical fragmentation scheme. This scheme allows the generation of activity class-directed fragments, which is not possible using conventional hierarchical fragmentation schemes, where bonds are cleaved based on general rules rather than active core information. By contrast, core-based fragmentation distinguishes different molecular regions in an activity class-sensitive manner.

Core Trees

The fragment hierarchy produced by iterative bond deletion is encoded in a tree structure termed the *Core Tree*. If a fragment A is separated into fragments B and C, fragment A is considered the *parent* fragment of its *children* B and C. Core Trees organize all of these parent-child relationships by connecting each parent with its children, as illustrated in Figure 2.10.

The root of a Core Tree is the original molecule and the leaves are individual atoms. The full Core Tree visualizes the entire fragmentation hierarchy and reports all fragments that are generated. In order to distinguish core from peripheral fragments, an average atom match rate MR_{frag} is calculated for each fragment. Core fragments receive a high MR_{frag}, in contrast to peripheral

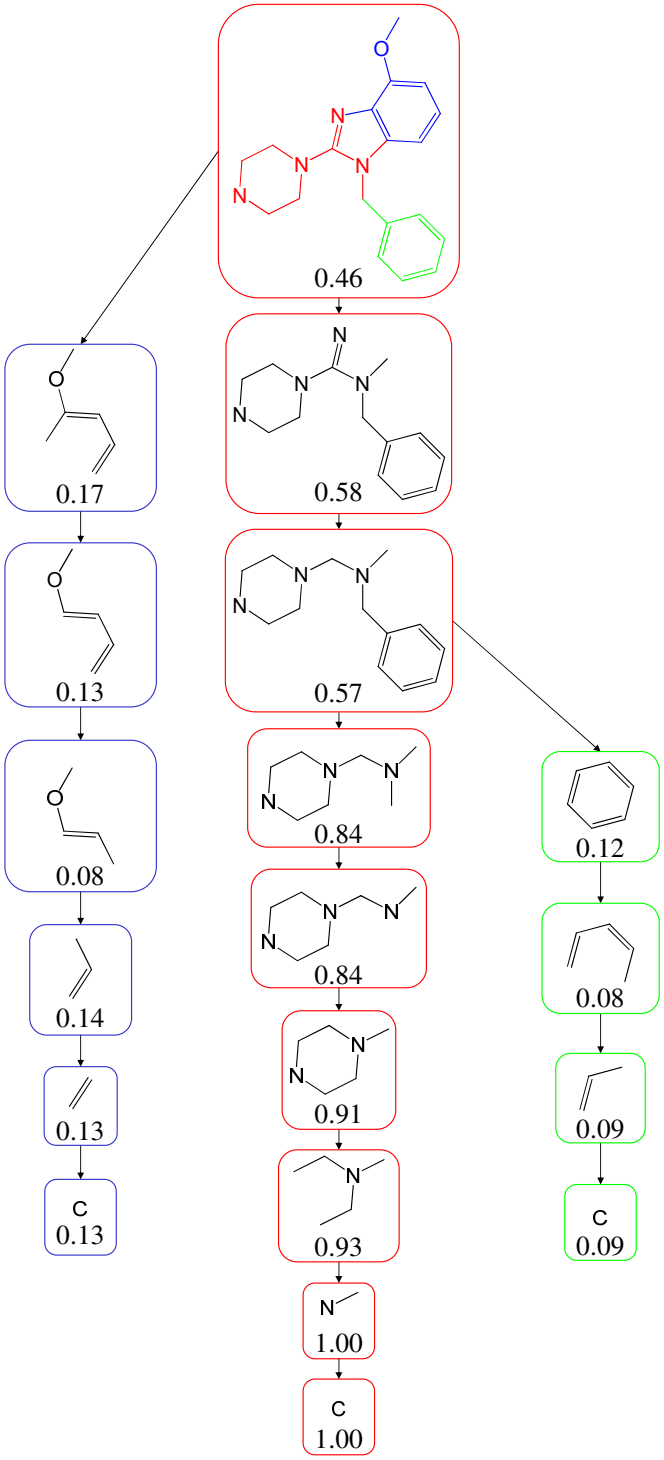


Figure 2.10: Exemplary core tree. Part of the Core Tree for a serotonin receptor antagonist is shown. For clarity, only nodes constituting a path are retained. The numbers report average match rates of each fragment.

fragments, which obtain low average match rates.

This annotation allows the distinction of fragments forming part of the conserved core of active compounds from fragments that describe variable peripheral regions. Thus, in addition to the generation of hierarchical fragments in an activity class-directed manner, Core Trees quantify the structural conservation of individual fragments within an activity class on the basis of atom match rates.

Identification of Core Fragmentation Pathways

In Core Trees, children of individual fragments can have lower or higher MR_{frag} values than their parents. A child has a higher fragment match rate if it describes the core region more closely than its parent, and a lower match rate if it describes peripheral fragments separated from the core. Fragment match rate distributions after separation of peripheral fragments from the core are illustrated in Figure 2.10.

Fragment match rates are used to delineate fragmentation pathways that represent conserved or peripheral regions in molecules. In order to identify these pathways, a $score_{edge}$ is assigned to each edge in the Core Tree. The score is calculated from fragment match rates and the number of atoms (N) of two fragments connected by the edge. If the parent's average match rate is 0, then $score_{edge}$ is also 0. Otherwise, it is calculated as

$$score_{edge} = \frac{N_{child}}{N_{parent}} \sqrt{\frac{MR_{child}}{MR_{parent}}}.$$

Starting at the root node, which corresponds to the molecule, a core path is identified by following maximal edge scores until an atom is reached. In Figure 2.10 the core path (red) contains nine fragments including the molecule and the terminal carbon atom. In analogy to the core path, additional fragmentation pathways representing peripheral regions can be identified. Therefore, the starting fragments are chosen from the remaining nodes in the Core Tree that do not belong to the core path, but are children of core path fragments. In order to ensure that peripheral fragmentation pathways can originate at five-membered rings, fragments with ≥ 5 heavy atoms are chosen as potential starting points.

Core and peripheral paths have been identified for 1,025 molecules spanning 45 activity classes. Figure 2.11 reports the path length distribution for core and peripheral paths. On average, a test molecule was described by one core path and 1.7 peripheral paths. Core paths are generally longer (seven to ten fragments) than peripheral pathways (three to seven fragments).

Core paths encode the conserved regions of test molecules as a sequence of fragments with decreasing specificity for individual compounds, but increas-

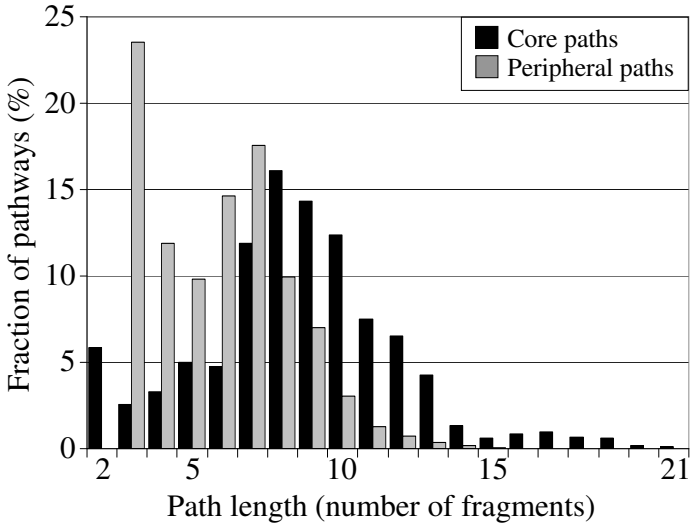


Figure 2.11: Fragmentation pathways. The histogram shows the length distribution of core (black) and peripheral (gray) paths.

ing conservation within the activity class, expressed in higher MR_{frag} values. Most molecule-specific and large fragments (including the molecule) are found at the beginning of the core path, and most conserved and small fragments (including the terminal atom) at the end. Thus, core paths balance fragment specificity and conservation in active compounds.

Core Path Alignment

Fragmentation pathways constitute a defined sequence of molecular fragments. Hence, in order to compare two core pathways, methods for the alignment of biological sequences can be applied. For this purpose, the Needleman-Wunsch algorithm,[81] which is used to globally align protein and DNA sequences, has been adapted for core path alignment. In this implementation, the algorithm uses the dynamic programming method to find the optimal alignment between two fragmentation pathways by comparing individual fragments and introducing gaps. For two sequences with length i and j, an $i \times j$ alignment matrix $\mathbf{A}$ is initialized with multiples of the gap penalty gp:

$$A_{i,0} = i \times gp, A_{0,j} = j \times gp.$$

Affine gaps are used with a gap opening penalty of -5 and a gap extension penalty of -2. Thus, if n is the length of the gap, gp is calculated as

$$gp = -5 + (n-1) \times (-2).$$

The alignment matrix is then filled iteratively by calculating the maximum score depending on the left, upper, and upper left (diagonal) neighbor of each cell and a fragment similarity function S that compares fargment i of the first core path with fragment j of the second core path:

$$A_{i,j} = \max \begin{cases} A_{i-1,j} + gp \\ A_{i-1,j-1} + S(i,j) \\ A_{i,j-1} + gp \end{cases} .$$

The final global alignment score is given by the lower right cell of the alignment matrix. In order to make different pathway alignments comparable, the alignment score is normalized by dividing it by the average of the alignment scores of each pathway aligned with itself.

The fragment similarity function S is applied instead of an amino acid substitution matrix utilized in protein sequence alignment. It is based on unique SMILES strings and compares individual fragments within core paths. The function comprises a fragment size term and a string similarity term, which are added in order to yield the final fragment score. The size term is calculated as

$$sizeterm = \begin{cases} 10 \times \frac{\min(N_a,N_b)}{\max(N_a,N_b)} & : \quad N_a \neq N_b \\ 20 & : \quad N_a = N_b \end{cases} .$$

N_a and N_b are the numbers of atoms in fragments a and b, respectively.

The string similarity is evaluated and scored in three steps: (1) if SMILES strings of two fragments are identical, a string similarity score of 15 is returned; (2) if they are not identical, branches (i. e. parts of the SMILES in parenthesis) are eliminated and if the resulting cropped strings are identical, a score of 5 is returned; (3) if the cropped strings are not identical, but a largest common substring exists, a score of 2 is returned. Otherwise, 0 is returned. This is the case for some small fragments in pathways that terminate at heteroatoms, where a common substring is not always found. Figure 2.12 illustrates string similarity calculation.

Score levels with an approximately three-time change in magnitude (i.e. 2, 5, and 15) have been empirically determined to produce meaningful fragmentation pathway alignments.

Thus, core path alignment allows systematic quantitative comparison of core paths. Because fragmentation pathways from different molecules show overlapping yet distinct substructures, sequence alignment methods are particularly suited for the comparison and alignment of core paths. Corresponding core path fragments of two molecules are found in core path alignments and global alignment scores can be used to quantify molecular similarity.

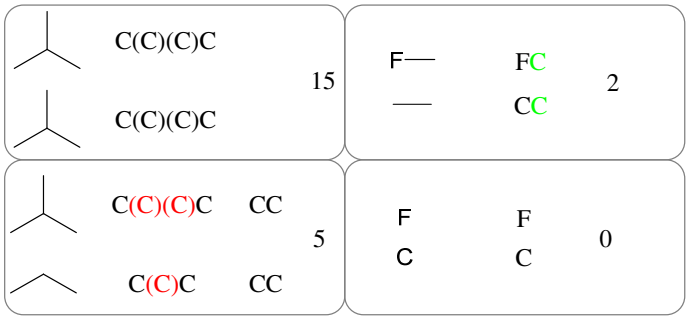


Figure 2.12: Fragment string similarity. Four possible scores for fragment SMILES string similarity are shown. Exemplary fragments are shown with corresponding SMILES strings. Red: branches that are deleted. Green: common substring.

Multiple Alignment and Consensus Fragment Sequence

Extending pair-wise core path alignments, all core paths of an activity class can be aligned in a multiple core path alignment. Therefore, a phylogenetic tree is calculated based on the normalized pair-wise alignment scores using the *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA).[82] This tree resembles a hierarchical clustering of compounds and is used as a guidance for multiple core path alignment. Closely related paths are aligned first, and subsequent paths are aligned to the existing alignment. Therefore, an alignment is treated as one fragment sequence; the scores for each position in multiple alignments are derived as unweighted averages of pair-wise fragment similarity scores.[81]

For multiple pathway alignments, the fragment similarity scoring function is adjusted in order to avoid gaps at the termini: the first and the last fragment (i. e. the molecule and terminal atom of the core path) are each substituted with a placeholder ("*") that yields a very high score of 100 when matched with another "*". After completion of the alignment, the placeholders are replaced by the original fragments. This has the advantage that a multiple core path alignment always starts with a column of molecules and ends with a column of terminal atoms, as shown in Figure 2.13.

For a multiple core path alignment, a *Consensus Fragment Sequence* (CFS) is derived by combining non-redundant fragments at each alignment position. Figure 2.13 illustrates the CFS derivation from a multiple core path alignment. For the 45 classes studied here, the CFS contained between 11 and 215 unique fragments, with an average number of 63 fragments per CFS. Thus, activity classes were described with a comparably small number of fragment-type descriptors. Activity class-relevant information in CFS is encoded in two ways. First, the individual fragments are generated in an activity class-directed manner. Second, the multiple alignment encodes fragment specificity and conservation in form of position-specific sets of unique fragments.

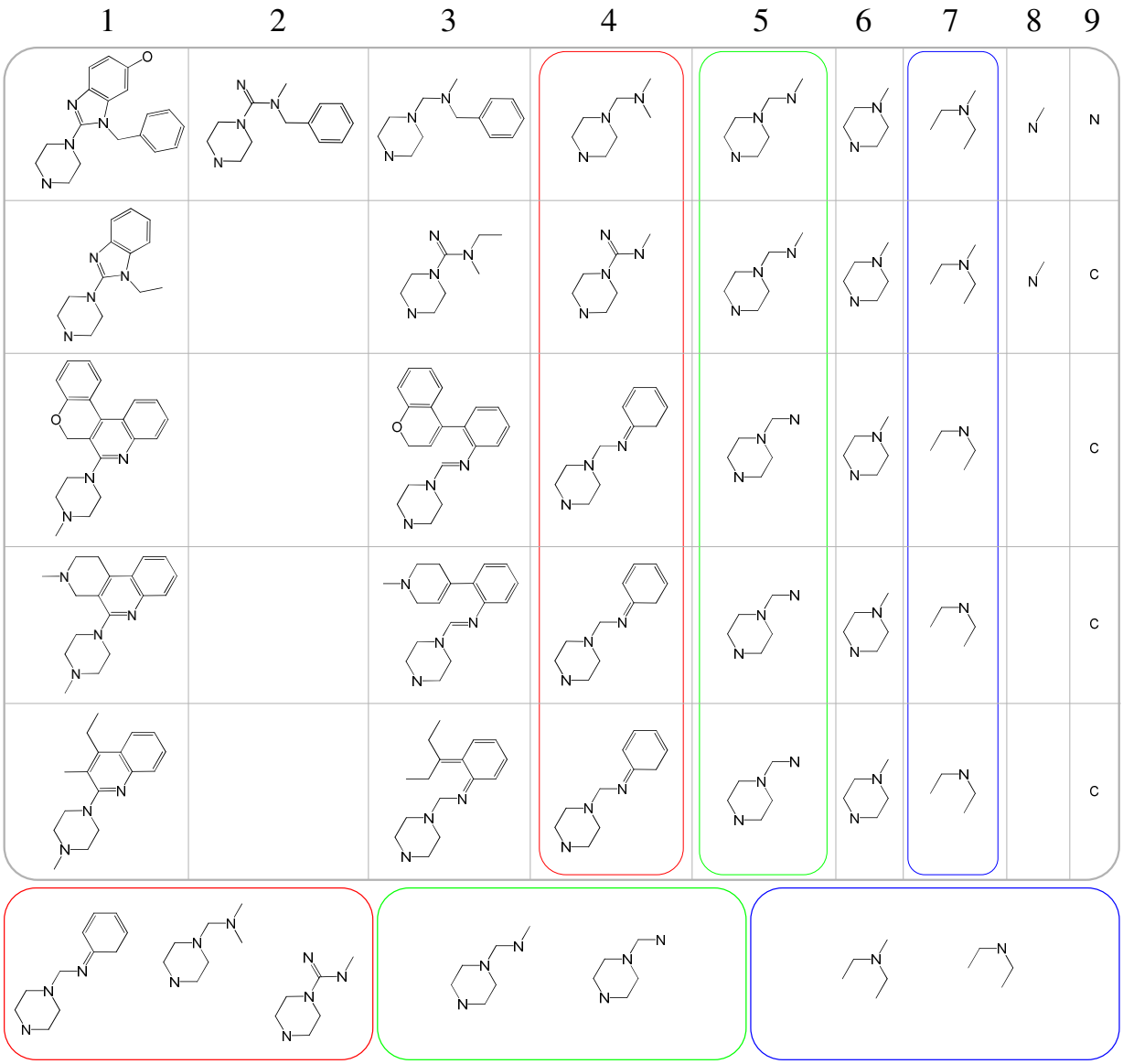


Figure 2.13: Multiple core path alignment. Part of a multiple core path alignment is shown for serotonin receptor antagonists. The top row numbers indicate positions in the alignment. Every core path begins with the whole molecule at position 1 and ends with a terminal atom. The positions 2 and 8 contain gaps for four and three molecules, respectively. Three CFS positions (4, 5, and 7) are shown at the bottom.

Thus, multiple core path alignments allow clustering of molecules based on core path similarity. All core path fragments of an activity class are organized in a Consensus Fragment Sequence that serves as an activity class signature. The CFS provides position-encoded information about the conservation and specificity of all core path fragments derived from an activity class.

2.4 Summary

Different methodologies have been developed for the generation of molecular fragments that serve as molecular descriptors for active compounds. Different fragmentation strategies are tailored towards distinct applications.

An activity class-directed hierarchical fragmentation scheme has been introduced by combining random and hierarchical fragmentation approaches. Core Trees organize the generated fragments and allow the distinction of conserved core fragments from peripheral substructures on the basis of atom match rates resulting from ACCS core mapping. Core Trees go beyond standard hierarchical fragmentation methods, which do not take activity class-characteristic cores into account.

Core paths have been identified that encode most activity class-characteristic molecular information in form of fragment sequences of decreasing specificity. They usually contain the largest number of fragments compared to peripheral fragmentation pathways. Large fragments at the beginning of core pathways are highly specific for individual compounds, but not well-conserved within the activity class. By contrast, small fragments represent highly conserved core regions, but are less specific. Thus, core paths allow balancing activity class conservation and specificity of fragments.

Individual core paths have been compared using methods for the global alignment of biological sequences. Alignment scores quantify molecular similarity and find corresponding core path fragments. It has been shown that all core paths of individual activity classes can be combined in multiple alignments. A multiple core path alignment organizes all core path fragments in a Consensus Fragment Sequence that serves as an activity class signature. Unique fragments are grouped together in position-specific sets that allow the selection of conserved and/or specific fragments.

Thus, the Core Tree methodology extends existing hierarchical fragmentation schemes by activity class-directed fragmentation and fragment organization in consensus sequences that reflect conserved molecular cores.

Chapter 3

Analysis of the Topological Environment of Substructures in Active Compounds

The previous chapter has described different approaches to the generation of molecular fragments and introduced an activity class-directed hierarchical fragmentation scheme. Multiple alignments of fragmentation pathways were then used as activity class signatures. It is also often of interest to analyze activity class-specific properties of known, chemically well-defined substructures like RECAP fragments that are easily interpretable by medicinal chemists.

In this chapter, the *Topological Fragment Index* (ToFI) is introduced for the quantitative assessment of the topological environment of a fragment within active compounds. ToFI extends fragment counts because it also takes into account bond patterns that are relevant for the generation of a given fragment. Moreover, it is applicable to any type of molecular fragments, regardless of how they are derived. Here, ToFI calculations have been applied to RECAP fragments. On the basis of ToFI calculations, fragments can be organized in hierarchies that are based on topological fragment environment and fragment co-occurrence in active compounds and define fragment pathways that are specific for individual activity classes.

First, the ToFI calculation method is described in general. Then, its application to RECAP fragments and formation of dependency graphs is reported. Structural relationships of dependent fragment are assessed and Activity Class Characteristic RECAP Fragments are identified. Furthermore, subgraphs representing fragment topology clusters with characteristic distributions of ACCRF are extracted.

3.1 Topological Fragment Index Method

Libraries of molecular fragments can be mapped onto active compounds to determine their presence or absence or count the number of fragment occurrences. Both presence or absence and counts can serve as the basis for the generation of molecular fingerprints. Figure 3.1 shows a binary (i. e. presence/absence) and count fingerprint of Aspirin.

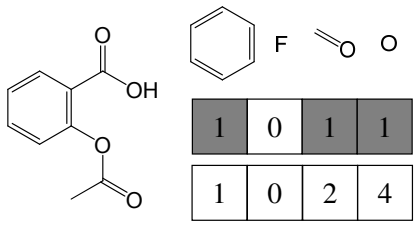


Figure 3.1: Binary fingerprints and counts. Two hypothetical fingerprints reporting the presence of individual fragments or their counts are shown for Aspirin. Fingerprints based on counts have higher information content than binary fingerprints.

The *Topological Fragment Index* (ToFI) extends fragment counts and quantifies the topological environment of a given fragment in a molecule using an integer score. ToFI operates on hydrogen-suppressed molecular graphs, i. e. only bonds between non-hydrogen atoms are considered. In principle, three parameters account for fragment generation and topological environment information:

1. the total number of bonds in the molecule (n),

2. the number of bonds that must be cleaved in order to obtain the fragment (k),

3. the number of bonds that are not permitted to be cleaved because atoms connected by these bonds constitute the fragment (l).

Fragments can often be mapped in different ways onto molecules. ToFI independently assesses fragment instances that differ in matched molecule atoms. First, ToFI values are calculated for each fragment instance:

$$ToFI_{Instance}(n,k,l) = \frac{1}{n+1} \sum_{i=0}^{n-(k+l)} \frac{\binom{n-(k+l)}{i}}{\binom{n}{k+i}}.$$

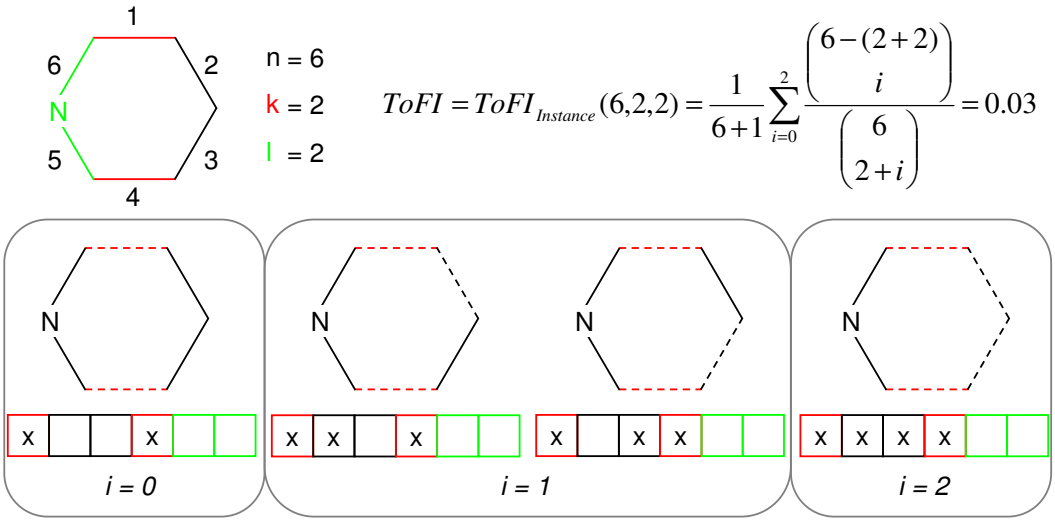
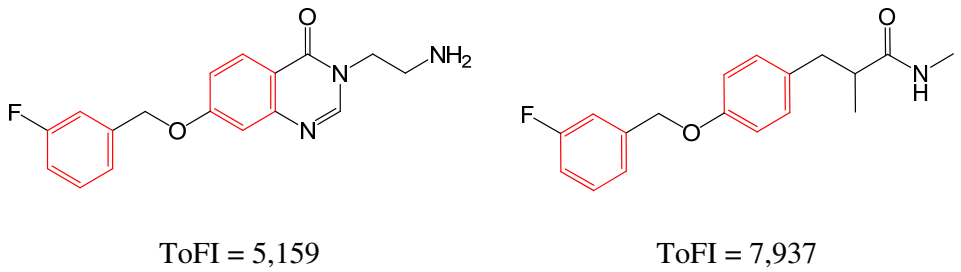


Figure 3.2: ToFI calculation. The ToFI calculation for the extraction of a fragment (green) from piperidine is shown. Bonds that have to be cleaved are drawn in red, whereas bonds that constitute the fragment are shown in green. The bottom panels enumerate bond cleavage patterns leading to separation of the fragment. Small boxes indicate bonds and an "x" is placed into a box if the corresponding bond is cleaved. Red boxes correspond to bonds 1 and 4, i. e. bonds that must be cleaved, and always contain an "x". Green boxes represent bonds 5 and 6 and never contain an "x". The remaining bonds provide four degrees of freedom for choosing bond cleavage patterns leading to fragment generation.

The variable i denotes the number of bonds that are deleted in addition to k, but do not influence the generation of the given fragment instance. Figure 3.2 illustrates the ToFI calculation for a fragment and provides a structural interpretation of the parameters n, k, l, and i.

The final ToFI value is the sum over all ToFI values calculated for individual fragment instances. In order to control computational cost and provide easily interpretable ToFI scores, floating point ToFI values are transformed into integers by multiplying each calculated value with an empirically chosen constant of 10^6 and rounding the resulting value to the nearest integer.

Figure 3.3 illustrates how ToFI distinguishes different topological environments of a fragment that are not captured by fragment counts. The larger the ToFI value becomes, the less complex is the topological environment of the fragment. Fragments with equal counts in different compounds are further distinguished by ToFI, if they occur in distinct topological contexts. In ToFI calculations, the complexity of a fragment is increasing with the number of bonds within the fragment and the number of bonds between fragment atoms and other atoms in the molecule. Therefore, it is also possible to interpret ToFI values as the likelihood for any given fragment to be isolated from source molecules by randomized bond cleavage. According to this interpretation, a



ToFI = 5,159 ToFI = 7,937

Figure 3.3: Exemplary ToFI values. ToFI values for a phenyl fragment (red) in two molecules are given. Fragment counts in both cases are equal (2 fragments). ToFI further distinguishes the mappings based on the topological environment of the fragments.

ToFI value of five means that if the compound was randomly fragmented one million times, the expectation value for its generation is five times.

3.2 Application of ToFI to RECAP Fragments

3.2.1 Data Sets

Eighteen activity classes have been assembled from the MDDR on the basis of MDDR activity indices associated with defined biological targets. The activity classes were grouped into five supersets of ligands binding between two and four closely related targets (for example, dopamine receptors D1-D4). The supersets contained between 252 and 2,267 ligands. Compound numbers and individual activity classes are reported in Table 3.1.

Superset	Biological activity	Cmpds.	MDDR act. indices
CCK	CCK A/B agonists/antagonists	730	42705, 42706, 42712, 42713
Dopamine	Dopamine D1-4 antagonists	1557	07702, 07701, 07703, 07710
MAO	MAO A/B inhibitors	711	08410, 08420
Opioids	$\kappa/\delta/\mu$ agonists		01131, 01132, 01133
PDE	Phosphodiesterase I-IV inhibitors	2567	78415, 78416, 78417, 78418

Table 3.1: Datasets for ToFI calculation and distribution analysis. Five supersets containing a varying number of activity classes against related targets, as defined by "MDDR act. indices", were used for the generation of dependency graphs. "Cmpds." reports the number of compounds in each superset.

3.2.2 RECAP Fragmentation and Mapping

ToFI calculations have been systematically applied to RECAP fragments. In RECAP analysis, fragments are generated by cleaving bonds that are formed

by common chemical reactions. Atoms of RECAP fragments that participate in cleaved bonds are annotated with bond type information in form of isotope-like labels (see Figure 2.4). These RECAP atom types are used to further distinguish fragments that are otherwise identical. A library of 10,246 RECAP fragments was generated from the MDDR using MOE. For substructure mapping purposes, RECAP fragments were encoded using recursive SMARTS strings,[83] which allows for the definition and mapping of chemical atom environments. Table 3.2 on page 36 reports the number of mapped RECAP fragments for each superset.

3.2.3 ToFI Calculation for RECAP Fragments

ToFI values have been calculated for all mapped RECAP fragments. Figure 3.4 shows exemplary ToFI value distributions for two RECAP fragments in all test molecules. Often different ToFI values were obtained for a fragment, hence distinguishing its topological environment in active compounds. Three examples are shown in Figure 3.4 for the top histogram. The fragment count is one for all three molecules. ToFI further distinguishes the different topological environments. Five different ToFI values were observed for the top fragment and nine different values were found for the second fragment.

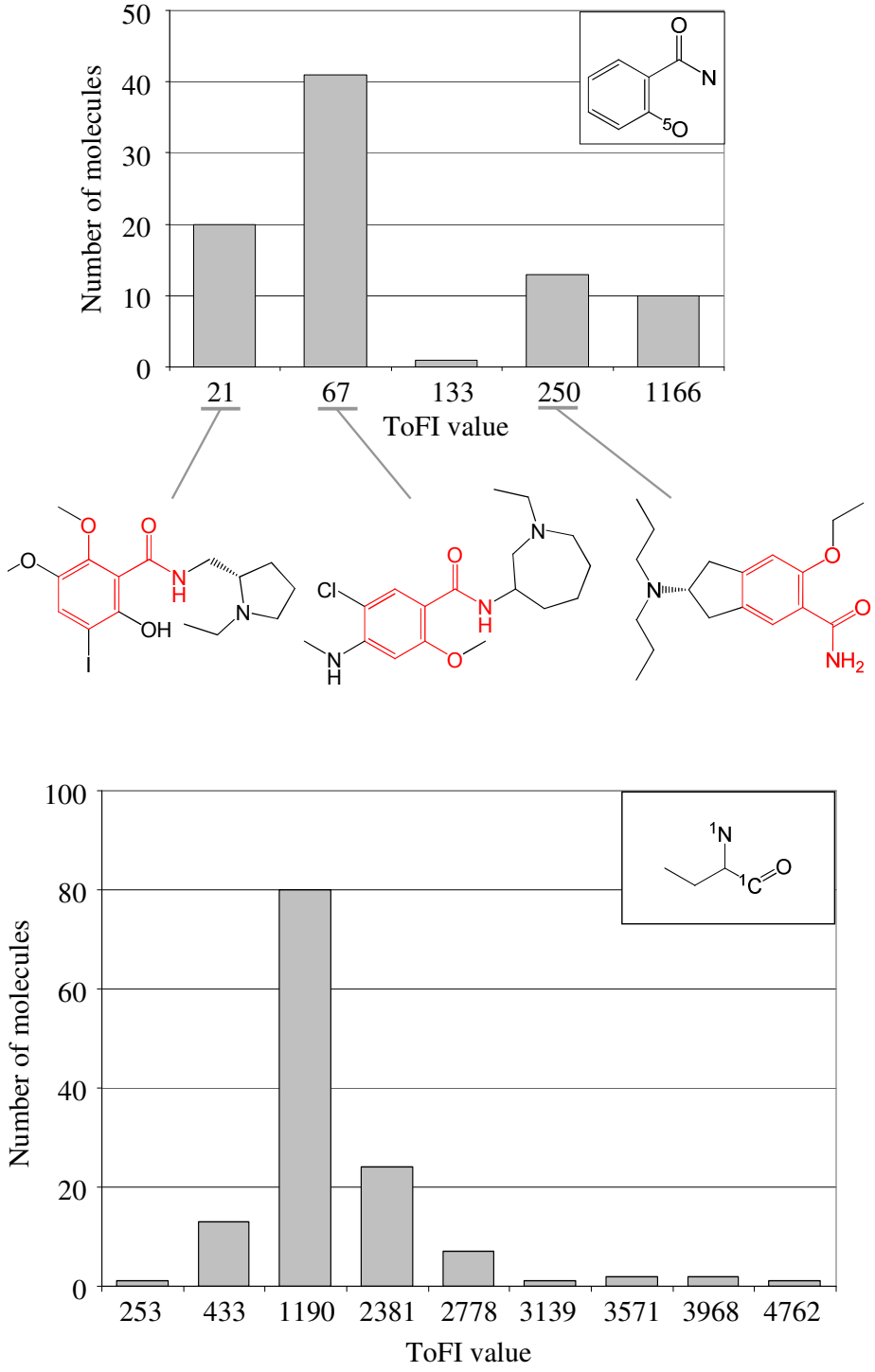


Figure 3.4: RECAP ToFI calculation. The histograms report the distribution of ToFI values for two RECAP fragments. Three exemplary fragment mappings are shown for the first fragment.

3.3 Hierarchical Organization of RECAP Fragments based on ToFI Value Distributions

In the previous chapter, Core Trees were introduced that organize activity class-directed fragments in hierarchies allowing the identification of core paths. Hierarchically generated fragments have also been organized in Scaffold Trees that reflect the fragmentation procedure.[45] In Scaffold Trees, small core fragments constitute the root of each tree and are augmented by other structures until complete molecules are formed. Thus, from each fragment in a Scaffold Tree molecules containing the scaffold can be reached. These hierarchies can be annotated with activity information and used to derive common scaffolds of active compounds, facilitating SAR analysis.[73]

A conceptually distinct approach developed for the organization of random fragments defines hierarchies based on fragment co-occurrence in active compounds, rather than structural criteria. Fragment dependency relationships are encoded in dependency graphs that incorporate activity information of fragments and enable the identification of activity class-specific fragment pathways. This approach has revealed that random fragment populations contain activity class-specific information that is associated with fragment combinations and frequencies of fragment occurrence.[47]

However, dependency graphs have not yet been applied to non-random fragments because the method depends on fragment frequencies, which are typically derived from random fragment populations. These fragment frequencies can be interpreted as an indicator for the complexity of the topological fragment environment in molecules.[76] ToFI allows the extension of this type of hierarchical organization to non-random fragments. This section describes the calculation of dependency graphs based on ToFI values, identification of Activity Class Characteristic RECAP Fragments, and analysis of structural relationships between dependent fragments.

3.3.1 Dependency Graph Calculation

Dependency graphs have been designed to account for fragment co-occurrence in molecules. They report subsets of fragments that only occur together with others, i. e. a fragment A must be present (conditional) for a fragment B to occur (dependent). Fragment dependencies are quantified based on ToFI values using the following formalism. Given N compounds, each fragment is represented as an N-dimensional vector where each component reports the ToFI score for a specific molecule in the data set. For each fragment, its dependency on other fragments is expressed as the ratio of the respective ToFI components. A fragment *dep* only depends on a fragment *cond* if all vector components of *dep* are smaller or equal to the respective components of *cond* and at least one

component of *dep* is smaller than the corresponding component of *cond*. The dependency is then calculated as:

$$dependency(dep, cond) = \delta \sum_{i=1}^{N} \frac{ToFI_i(dep)}{ToFI_i(cond)}.$$

The δ operator summarizes the above stated conditions. Formally, it is defined as:[47]

$$\delta = \begin{cases} 1 & \text{if} \quad \forall\ 1 \leq i \leq N : ToFI_i(dep) \leq ToFI_i(cond) \\ & \text{and} \quad \exists\ 1 \leq i \leq N : ToFI_i(dep) < ToFI_i(cond), \\ 0 & \text{else} \end{cases} .$$

Here, N is the total number of compounds, $\forall$ means "for all" and $\exists$ means "there exists at least one". Dependency relationships are then encoded in a graph representation. Therefore, for each fragment, the set of conditional fragments with maximal dependency value is retained and an edge is drawn between the conditional and dependent fragments. Figure 3.5 illustrates the calculation of dependency graphs.

Molecules are added to the dependency graph as "superfragments". In the ToFI vector of a molecule, the component that corresponds to the molecule is set to 1, whereas all other components are set to 0. This ensures that molecules constitute the termini of dependency pathways. Fragment pathways in the graph that do not terminate at a molecule are deleted because they contain fragments that have low signature character for active compounds.[47] Table 3.2 reports the number of RECAP fragments retained in the dependency graph of each superset.

Superset	Mapped	Graph	ACCRF
CCK	1268	203	59
Dopamine	1789	406	111
MAO	252	148	80
Opioids	1480	231	62
PDE	1836	444	167

Table 3.2: ToFI dataset statistics. Reported are the number of successfully mapped substructures ("Mapped"), and the number of fragments remaining in the dependency graph after deletion of paths that did not terminate at a molecule ("Graph"). "ACCRF" refers to the number of Activity Class Characteristic RECAP fragments found in the graph.

Fragments that are connected in the dependency graph most strongly depend on each other, i. e. they are most similar with respect to their topological context and distribution among active compounds. An edge between a fragment A and fragment B in the dependency graph indicates two relationships: first, fragment A is always present in a molecule, if fragment B is present; second,

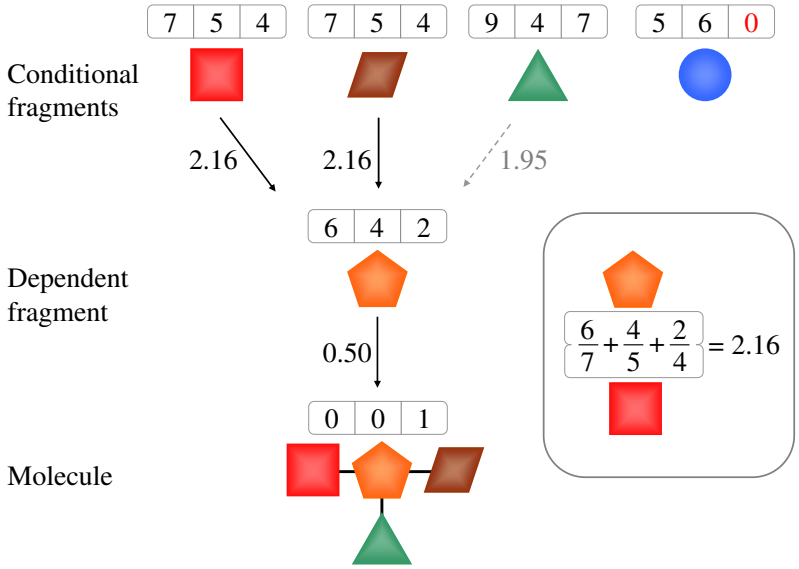


Figure 3.5: Dependency graph calculation. The calculation of the dependency graph is illustrated. Colored shapes correspond to individual fragments that build up a molecule (bottom). The ToFI value vector for three molecules is reported above each fragment. Edges in the graph are annotated with the dependency score. The pentagon fragment depends on three of four fragments. It is independent of the blue circle fragment, because the third component of the circle fragment violates the dependency condition, i. e. it is smaller than the third component of the pentagon fragment. The dependency score shown in gray (i. e. 1.95) is smaller than the maximal value for the pentagon fragment (i. e. 2.16), and, therefore, no edge is drawn in the graph between the triangle and the pentagon fragment. The box on the right illustrates the calculation of the dependency score from ToFI vectors.

of all fragments that co-occur with fragment A, fragment B has the highest likelihood to do so. Fragments that do not depend on any other fragments constitute root nodes in the graph. They are per definition the most abundant fragments with high ToFI scores. Multiple root fragments that are independent of each other can be present in a graph.

3.3.2 Activity Class-Characteristic RECAP Fragments

In ToFI dependency graphs, all fragments are annotated based on their ToFI scores in compounds with different biological activity. Activity Class Characteristic RECAP Fragments (ACCRF) are defined as those fragments that show non-zero ToFI values for compounds of one individual activity class only. Nodes are color coded according to the biological activity of the respective compounds. Table 3.2 reports the number of ACCRF in each superset. Figure 3.6 shows part of the dependency graph for the superset MAO. Dependency graph depictions were generated using the freely available software Tulip.[84]

Dependency graphs facilitate the identification of ACCRF and show

their co-occurrence patterns in active compounds. ACCRF constitute activity class specific fragment dependency pathways because they exhibit ToFI distributions that have signature character for different sets of active compounds. For the supersets MAO, Opioids, and Dopamine ACCRF have been identified for all activity classes. For both CCK and PDE, ACCRF were found for three of four classes (except for CCK B agonists and PDE II inhibitors).

Thus, ToFI dependency graphs allow for the hierarchical organization of chemical space based on reference compounds that are neighbors in activity space. Here, chemical space is defined by RECAP fragments but ToFI is applicable to any type of substructures that can be mapped onto active molecules. Co-occurrence dependencies link the structural information encoded in molecular fragments to activity information represented in form of closely related activity classes.

3.3.3 Fragment Relationships

In order to assess structural resemblance of dependent fragments, edges in the dependency graphs have been annotated with information about substructural relationships between fragments. Therefore, four categories of substructure relationships are distinguished for a pair of a conditional (parent) and a dependent (child) fragment: (1) the parent is a substructure of the child; (2) the child is a substructure of the parent; (3) the fragments are identical but differ in individual RECAP atom types; (4) no substructural relationship exists. Table 3.3 reports the distribution of dependency relationships among these categories.

	CCK	**Dopamine**	**MAO**	**Opioids**	**PDE**	**Average**
No subgraph	67.2	60.2	56.4	63.7	52.4	60.0
Parent in child	25.2	28.4	33.4	27.0	37.6	30.3
Child in parent	0.3	0.7	0.7	0.7	0.2	0.5
Equal	7.2	10.8	9.5	8.6	9.8	9.2

Table 3.3: Substructural relationships of dependent fragments. For each superset, the percentage of edges is reported that connect fragments having four distinct structural relationships: "No subgraph", no sub- or supergraph relationship; "Parent in child", the parent is a subgraph of the child; "Child in parent", the child is a subgraph of the parent; "Equal", fragments differ only in individual RECAP atom types.

On average, 60% of the detected dependencies did not correspond to a structural (sub- or supergraph) relationship, as exemplified in Figure 3.6 for fragments 5 and 13. The second largest subset of fragment dependencies (30%) included fragment pairs where the dependent (child) fragment contained the conditional (parent) fragment as a substructure, i. e. the child was larger than the parent. An example is shown in Figure 3.6 for fragments 13 and 14.

This result shows that fragment relationships in ToFI dependency graphs

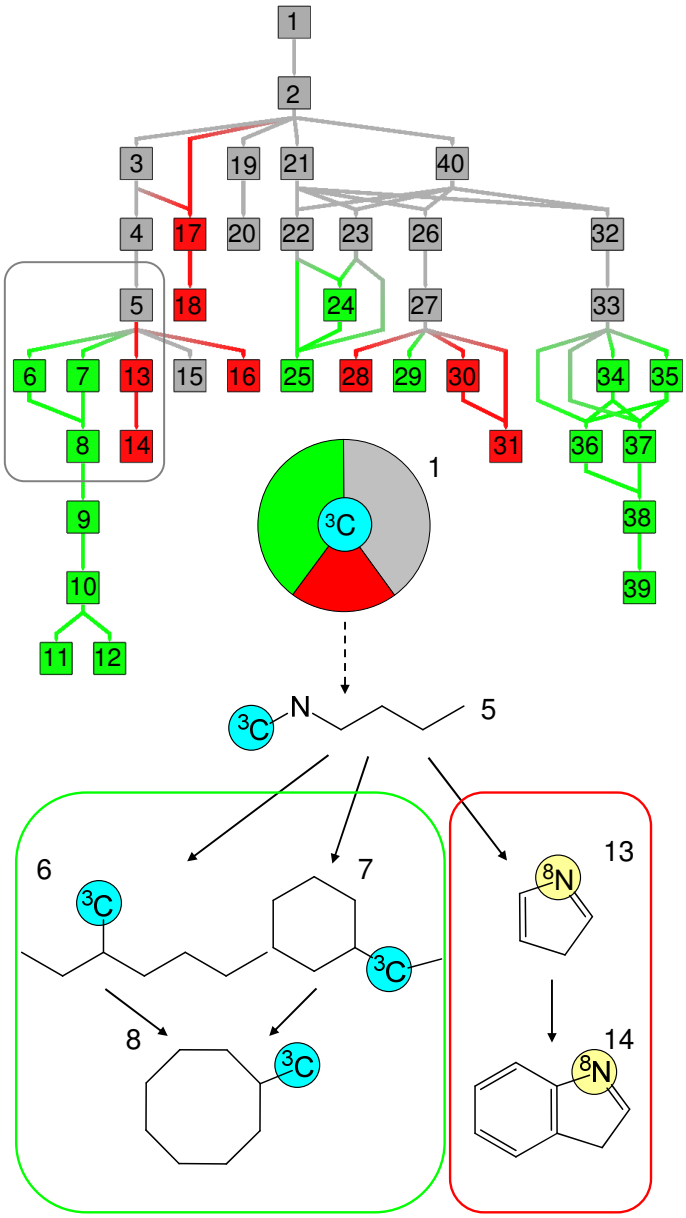


Figure 3.6: Exemplary ToFI dependency subgraph. For the superset "MAO", the subgraph containing all nodes reachable from the root fragment ^{3}C is shown. Numbers are fragment identifiers. The pie chart reports the ACCRF distribution of this subgraph. Grey nodes correspond to generic fragments, i. e. fragments that occur in more than one activity class. Color-coded segments report the distribution of ACCRF (red: MAO A, green: MAO B). At the bottom, fragments are shown that correspond to the boxed region in the subgraph. The isotope-like labels define RECAP atom types.

go beyond structural resemblance of fragments, but instead reflect the distribution of fragments among activity classes based on their topological environment.

3.3.4 Fragment Topology Clusters

ToFI based dependency graphs make it possible to analyze individual subpopulations of RECAP fragments that are dependent on each other. Therefore, subgraphs are extracted on the basis of root fragments. Each root fragment defines one subgraph, i. e. the collection of nodes and edges that are reachable from the root fragment. These subgraphs describe *topology clusters* of RECAP fragments that can overlap with other clusters. Fragments of each topology cluster form fragment dependency pathways originating at the root fragment.

Systematic analysis of ToFI-based fragment dependencies revealed that ToFI fragment hierarchies accurately described the interdependence of RECAP fragments containing atoms with correspondning atom types. Table 3.4 shows that most fragments ($\sim 90\%$) in ToFI topology clusters shared the same chemical environment of individual atoms with their root fragment.

Atom type	CCK	Dopamine	MAO	Opioids	PDE	Average
1	83.8	96.5	95.0	97.6	94.0	93.4
2	100.0	100.0	100.0	100.0	91.7	98.3
3		53.7	91.3	86.8	90.0	80.5
4	100.0		100.0	90.0	93.8	95.9
5	86.1	85.1	78.0	92.3	66.7	81.6
6		78.6	83.3	100.0	100.0	90.5
8				100.0	87.8	93.9
10		100.0			98.4	99.2
11	100.0				75.0	87.5
12					100.0	100.0
15				100.0	84.6	92.3

Table 3.4: RECAP atom type distribution in ToFI topology clusters. For each superset, the percentage of fragments is reported that contain at least one atom having the same RECAP atom type as the corresponding root fragment. Eleven of 16 possible RECAP atom types were present in root fragments.

This result indicates that ToFI value distributions reflect similar topological contexts of fragments that have atoms with the same chemical environment in common.

3.3.5 Distribution of ACCRF in Topology Clusters

Topology clusters have been systematically compared and their overlap quantified based on the number of shared fragments and/or molecules. Furthermore, the distribution of ACCRF in each topology cluster was assessed. Figure 3.7

shows a representative organization graph of topology clusters for the MAO superset. Topology clusters generally shared only few, if any fragments.

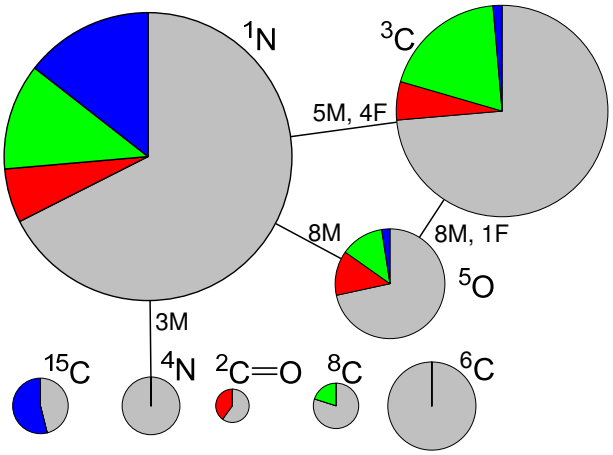


Figure 3.7: ToFI fragment topology clusters. The dependency graph topology for superset "Opioids" is shown. Each pie chart corresponds to a topology cluster. The root fragments are reported next to each pie chart. Grey segments correspond to generic fragments. Colored segments report the distribution of ACCRF (red: κ, green: δ, blue: μ). Pie charts are scaled proportionally to the number of nodes (i. e. fragments and molecules) and connected if the respective subgraphs overlap. Edges are annotated with the number of shared fragments (F) and/or molecules (M).

Different topology clusters show distinct ACCRF distributions, as can be seen in Figure 3.7. In this example, the distribution of ACCRF is similar for the ^3C and the ^5O topology cluster, but differs from the largest ^1N cluster.

The distribution of structurally equivalent fragments that belonged to different ToFI topology clusters was further systematically assessed. For all supersets, such fragments could be found. Moreover, among these fragments, ACCRF could be identified that belonged to different ToFI topology clusters and were characteristic of different activity classes. Figure 3.8 provides examples of such fragments that occur in different MAO activity classes and distinct ToFI topology clusters.

These results demonstrate that ToFI topology clusters capture activity class-specific distribution of RECAP fragments with different atom types. Different topology clusters show distinct ACCRF distributions. Similar fragments that differ only in individual atom types but are characteristic of different activity classes often belong to different topology clusters. Thus, ToFI dependency graphs can be used to organize RECAP fragments in an activity class-dependent manner.

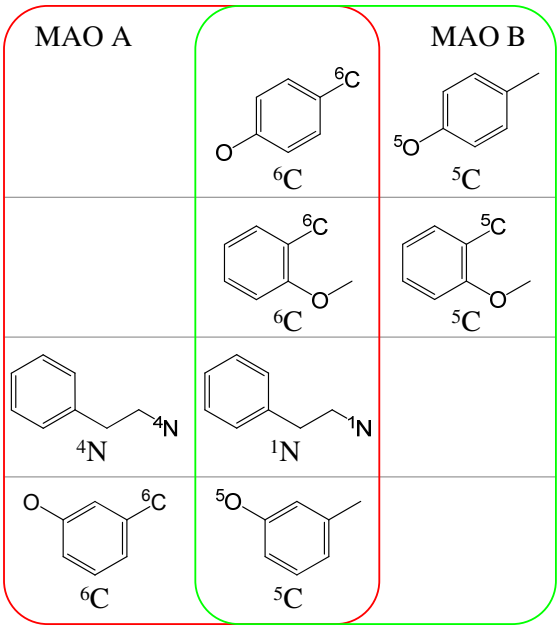


Figure 3.8: ACCRF topology cluster distribution. For the "MAO" superset, four examples of structurally identical fragments are shown that belong to different topology clusters identified by ToFI and differ in their biological activity (red box: MAO A; green box: MAO B). The root fragment of the topology cluster is reported in each cell.

3.4 Summary

Conventional fragment mapping approaches determine the presence of substructures in test compounds and fragment counts, i. e. the number of fragment instances present in a molecule. However, these methods do not account for the topological environment of fragments, which often differs in active compounds.

In order to assess and quantify the topological environment of molecular fragments, the Topological Fragment Index (ToFI) has been introduced. ToFI extends fragment counts in molecules because it incorporates information about the atom bonding patterns constituting the fragment environment within a compound. It has been shown that ToFI values distinguish between different fragment environments.

On the basis of ToFI value distributions, RECAP fragments of active compounds have been organized in dependency graphs that encode topological environment similarity and fragment co-occurrence information. These graphs facilitate the identification of Activity Class Characteristic RECAP Fragments (ACCRF) and encode fragment relationships that go beyond structural resemblance, reflecting activity class-characteristic fragment co-occurrence patterns. Moreover, ToFI fragment dependencies accurately account for inter-dependence

of RECAP fragments with common atom types.

From ToFI dependency graphs, topology clusters of RECAP fragments have been extracted that show characteristic ACCRF profiles and group fragments according to similar topological environment and distribution among active compounds.

Thus, ToFI adds to the repertoire of fragment mapping methods and allows the organization of non-random fragments in activity class-dependent hierarchies.

Chapter 4

Relevance of Feature Combinations for Similarity Searching

Activity class dependent hierarchies of randomly generated fragments, or predefined fragments using ToFI revealed that the occurrence of molecular substructures is often determined by dependency relationships. Thus, fragments might co-occur in an activity class-characteristic manner. Therefore, this chapter extends the analysis of fragment co-occurrence. In particular, the relevance of structural feature combinations for recall of active compounds in similarity searching is assessed.

First, three different molecular fingerprints used in this study are described that take activity class-related information to different degrees into account. In addition, current strategies for similarity searching using multiple reference compounds are discussed.

Feature Co-occurrence Networks (FCoN) are then introduced for the systematic extraction of feature combinations from multiple reference molecules. Individual structural features are organized in networks that account for their pair-wise co-occurrence in active reference compounds. From these networks, cliques are extracted that represent activity class-characteristic feature combinations. FCoN cliques are ranked based on their occurrence in a large compound database. Furthermore, a search strategy has been designed that assembles compound selection sets based on FCoN clique ranking.

4.1 Structural Fingerprints

Fingerprint representations enable fast computational comparison of reference and database compounds in molecular similarity searching. Here, three molecular fingerprints are compared that use fragment-type descriptors to encode

molecular structure: MDL *Molecular ACCess System* (MACCS),[63] *Extended Connectivity Fingerprints* (ECFP4),[59] and *Activity Class Characteristic Substructures Fingerprints* (ACCS-FP).[48] These fingerprints represent different design strategies that incorporate activity class-specific information to a different extent, as illustrated in Figure 4.1.

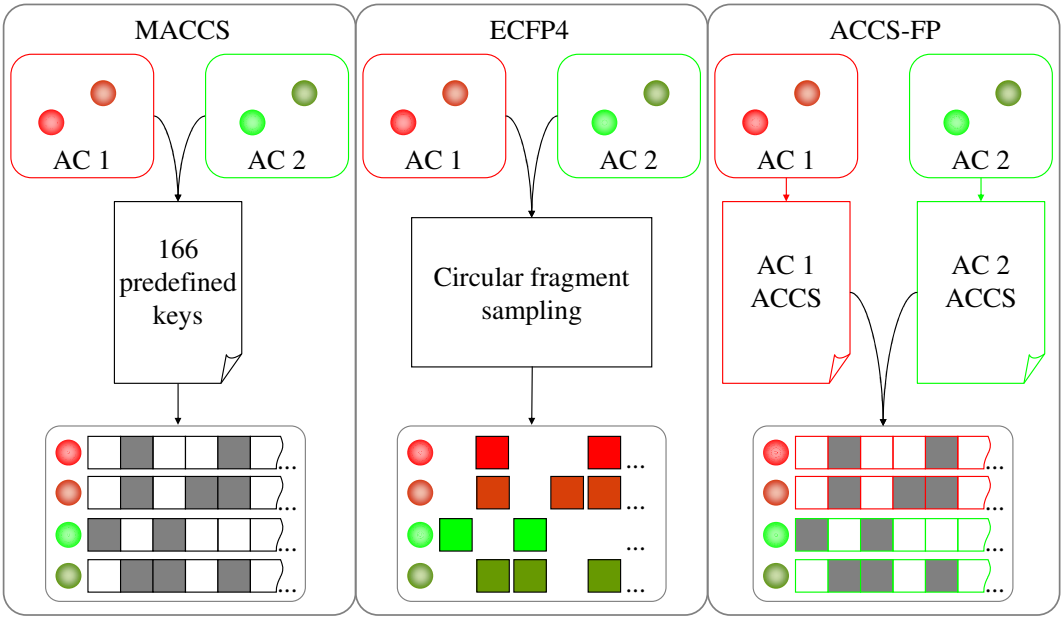


Figure 4.1: Fingerprint design strategies. Three fingerprint design strategies are compared. Circles represent compounds and squares represent fingerprint bit positions or features. MACCS uses a dictionary of 166 predefined structural keys to encode active compounds from different activity classes. In ECFP4, features are generated in a molecule-centric way utilizing layered atom environments. ACCS-FP are derived in an activity class-dependent manner and also represent a structural key-type fingerprint.

MACCS keys are general in their design. Originally, they have been introduced for fast identification of compounds in large databases. They account for the presence or absence of 166 predefined features that were selected to cover a large number of chemical structures. MACCS keys do not take any activity class- or molecule-specific information into account. Instead, they use the same general feature library to encode both reference and database compounds as a fixed-length fingerprint.

By contrast, Extended Connectivity Fingerprints use circular features (see Figure 2.2) to encode molecular structure. In ECFP4, each non-hydrogen atom in a test molecule is combined with bonded atoms at a one- to a four-bond radius. The circular features are converted to integers using a hash function that incorporates information about atom type, charge, and number of bonded atoms. The fingerprint consists of unique integers that have been generated during the hashing procedure. Thus, in ECFP4, fingerprint features are derived

in a molecule-centric manner: individual compounds (i. e. reference as well as database molecules) serve as the source for ECFP4 features. Theoretically, billions of possible ECFP4 features may be produced, but individual compounds typically yield only a small fraction of this hypothetical feature space. Because ECFP4 features are extracted from individual molecules they form fingerprints of variable length.

ACCS-FP utilize ACCS extracted from random fragment populations of reference compounds (see Section 2.2.4). ACCS are organized in hierarchies using the formalism described in Section 3.3.1. This organization is based on fragment co-occurrence patterns within the reference set.[48] Then ACCS are selected that are independent of other ACCS. These de-correlated fragments serve as the basis for fingerprint generation. Thus, ACCS-FP features are generated in an activity class-dependent manner on the basis of multiple reference compounds. ACCS-FP consist of small numbers of features, sometimes only 10.[48] ACCS-FP thus represent an activity-class directed, variable-length fingerprint format.

4.2 Multiple Template Similarity Searching

Fingerprint based similarity searching quantifies the structural resemblance of two compounds using fingerprint overlap as a similarity measure.[13] When only a single reference compound is available, database molecules can be prioritized according to their similarity values. However, when using multiple reference molecules, data fusion techniques are applied in order to exploit information from all fingerprints of the reference set.[30]

4.2.1 Quantification of Fingerprint Overlap

For comparison of binary fingerprints, a variety of different similarity metrics have been proposed.[13] One of the most popular metrics is the Tanimoto coefficient (Tc). It utilizes three sets of features to compare two fingerprints A and B: the number of features present in A (a), the number of features present in B (b), and the number of features shared by A and B (c). The Tc for binary fingerprints is then defined as:[13]

$$Tc = \frac{c}{a + b - c}.$$

For comparison of fingerprints with non-binary values such as feature counts, the general form of the Tc is used. It incorporates dot products of the fingerprint vectors instead of bit numbers:[13]

$$Tc_{general} = \frac{\sum x_{jA} x_{jB}}{\sum x_{jA}^2 + \sum x_{jB}^2 - \sum x_{jA} x_{jB}}.$$

Here, x_j is the j^{th} component of the feature vector $\mathbf{x}$.

4.2.2 Nearest Neighbor Searching

In order to score database compounds on the basis of their similarity to multiple reference compounds, in nearest neighbor searching, individual similarity values between the database molecule and each of the reference compounds are combined. In 1NN searching, the maximal score is retained, i. e. the similarity of the database compound to the reference set is estimated using the most similar reference compound. In the general case of kNN searching, the top k scores are averaged.[30] Figure 4.1 illustrates the different nearest neighbor search strategies.

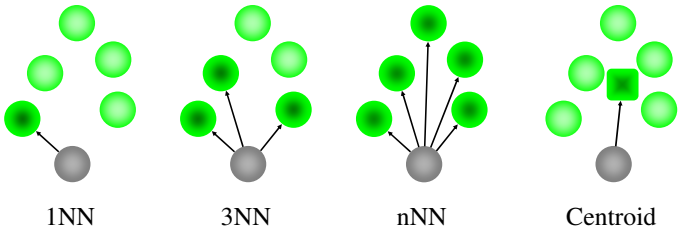


Figure 4.2: Multiple template similarity searching. In kNN searching, a database compound (gray circle) is compared to n reference molecules (green circles). The k highest similarity scores are averaged to yield the final compound score. In the centroid approach, first a centroid (green square) is calculated. The similarity of a database compound to the reference set is estimated through its similarity to the centroid.

4.2.3 Centroid Searching

An alternative to the fusion of pair-wise similarity scores is the fusion of reference compound fingerprints. Therefore, a centroid vector is calculated that reports the relative frequency of each feature within the reference set. Database compound fingerprints are then compared to this centroid rather than to all reference fingerprints individually. Since the centroid is not a binary fingerprint, the general form of the Tc must be used for similarity assessment.[30] Figure 4.1 compares centroid to kNN searching.

Modal fingerprints represent another form of fingerprint fusion. In these fingerprints, a predefined threshold is applied to the centroid and all vector components that meet or exceed this threshold are set to 1, whereas all other positions are set to 0.[43] Thus, modal fingerprints are binary vectors.

4.3 Feature Co-occurrence Networks

In fingerprint overlap calculations, individual features are treated independently of each other and there is no need to determine whether or not features occur in combinations. Consequently, the potential role of feature combinations for fingerprint search performance has so far been only little explored. Fingerprint bits that are strongly conserved in active compounds have been identified[85] as well as characteristic bit patterns[85,86] and, furthermore, consensus fingerprints have been derived for different compound classes.[43] However, whether or not feature combinations might influence fingerprint search performance has not yet been investigated.

In order to determine whether fingerprint features are set in combination and assess the potential impact of such combinations on similarity searching, a generally applicable methodology is presented in this section for the identification of feature combinations of variable size that are preferentially found in active reference compounds. For this purpose, Feature Co-occurrence Networks (FCoN) are introduced that are calculated based on conditional probabilities of feature pair occurrence in active compounds. A clique detection algorithm is then applied to these networks in order to extract feature combinations that are prevalent in different activity classes. The frequency of these feature combinations in a large screening database has been determined, thus enabling compound selection on the basis of selective feature cliques.

4.3.1 FCoN Generation and Clique Detection

Data Sets

For the generation and assessment of Feature Co-occurrence Networks, fourteen activity classes were assembled from the MDDR. Table 4.1 summarizes the biological activities and composition of these classes. The structural homogeneity of activity classes was assessed by pair-wise Tc calculations using the MACCS fingerprint. These 14 classes consisted of six relatively homogeneous (average MACC Tc $\geq$0.5) and eight more heterogeneous (average MACCS Tc 0.4 - 0.5) ones. A randomly selected ZINC subset containing 500,000 molecules was used as a database for virtual screening trials. Each activity class was randomly divided into ten reference and test sets of equal size.

Molecular structure was encoded using three fingerprints described in Section 4.1. ACCS have been generated for each activity class individually by filtering against a randomly selected ZINC subset of 500 compounds (see Figure 2.7). Table 4.1 reports the average number of ACCS generated for individual activity classes.

Activity class	Biological activity	Cmpds.	ACCS	Tc
AA2	Adrenergic alpha 2 antagonists	35	25.3	0.39
BK2	Bradykinin BK2 antagonists	22	31.0	0.55
CAL	Calpain inhibitors	28	23.7	0.48
DD1	Dopamine D1 agonists	30	55.9	0.56
F7I	Factor VIIa inhibitors	23	16.1	0.46
GLG	Glucagon receptor antagonists	33	34.1	0.44
GLY	Glycoprotein IIb-IIIa antagonists	34	34.0	0.57
KRA	Kainate receptor antagonists	22	15.7	0.55
LAC	Lactamase (beta) inhibitors	29	33.2	0.44
SQE	Squalene epoxidase inhibitors	25	7.7	0.40
SQS	Squalene synthetase inhibitors	42	65.4	0.50
THI	Thiol protease inhibitors	34	42.8	0.49
ULD	LDL upregulators	21	16.6	0.43
XAN	Xanthine oxidase inhibitors	35	33.8	0.56

Table 4.1: FCoN data sets. The number of compounds ("Cmpds.") in each activity class is provided together with the average number of activity class characteristic substructures ("ACCS") generated per reference set. "Tc" reports the average intra-class pair-wise similarity of active compounds based on MACCS Tc calculations.

FCoN calculation

In order to identify structural features that predominantly occur in combination in a compound reference set, FCoN are calculated for molecular fingerprints. Individual features encoded by a fingerprint are connected in these networks if they preferably occur in combination. In order to build an FCoN, the frequency of occurrence is calculated first for each feature by dividing the number of compounds containing the feature by the total number of compounds in the reference set, analogously to centroid fingerprints. Pairs of features are assigned a score based on the conditional probabilities of feature pair occurrence:

$$score\,(A,B) = \min\left(\frac{f\,(AB)}{f\,(A)}, \frac{f\,(AB)}{f\,(B)}\right).$$

Here, $f\,(A)$ denotes the frequency of feature A, $f\,(B)$ the frequency of feature B, and $f\,(AB)$ the frequency of the pair AB. Feature pairs that exclusively occur in combination receive the maximal score of 1, whereas features that never occur in combination in any reference compound are assigned the minimal pair score of 0. Scores serve as weights for edges connecting individual features in the FCoN. In order to identify combinations of frequently co-occurring features, a co-occurrence threshold ν is applied and only edges are retained that have a weight equal to or greater than ν. Maximal cliques are

largest completely connected subgraphs that are not contained in any other completely connected subgraph.[11] These cliques are identified in the pruned network and corresponding feature combinations are reported. For clique detection, the Bron-Kerbosch algorithm[87] was implemented in MOE. The clique detection procedure is illustrated in Figure 4.3.

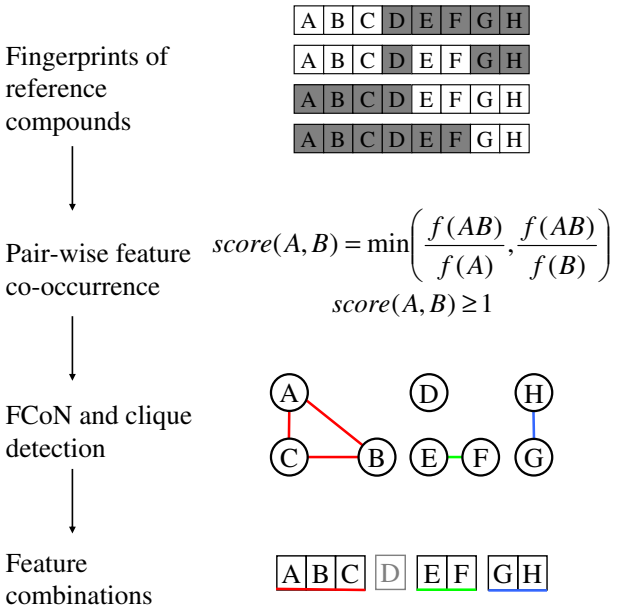


Figure 4.3: FCoN clique detection. FCoN generation is illustrated for a co-occurrence threshold value of one. For four fingerprints of active reference compounds the relative frequency of each feature (bit) is determined. Shaded boxes correspond to fingerprint bit positions that are set on. The corresponding co-occurrence network (FCoN) is calculated by connecting features that occur in combination. Cliques are detected and feature combinations are identified. Although features A, B, and C are only present in two molecules, they form a clique based on a threshold value of one, because neither feature occurs without the other two in any compound. Feature D is not part of a clique and therefore not considered in search calculations.

The co-occurrence threshold ν was systematically varied from 0.5 to 1 in increments of 0.1. Increasing threshold values yield cliques that are highly conserved among reference compounds. In addition to cliques identified for each ν, all cliques that were unique to an activity class have been pooled irrespective of the threshold value.

FCoN were systematically generated for all activity classes and fingerprints. Clique numbers and sizes were determined for each ν and the pooled sets. Statistical analyses were carried out using Perl scripts. Table 4.2 reports the medians of feature clique numbers that were first calculated for individual co-occurrence thresholds for all reference sets and then pooled. Table B.1 in Appendix B provides clique numbers for individual co-occurrence thresholds.

Figure B.1 reports the distribution of clique numbers for pooled sets. ACCS-FP produced the smallest number of feature cliques. Median clique numbers ranged from four to 83 and correlated with the ACCS-FP length, yielding a Pearson correlation coefficient of $R^2 = 0.89$. MACCS and ECFP4 produced on average $\sim$200 and 300 cliques, respectively, and hence significantly more than ACCS-FP.

Activity class	MACCS	ECFP4	ACCS-FP
AA2	5	7	3
BK2	11	9	5
CAL	5	7	4
DD1	5	8	6
F7I	8	6	3
GLG	8	7	3
GLY	6	6	3
KRA	8	11	5
LAC	8	12	4
SQE	7	6	2
SQS	10	13	6
THI	5	5	3
ULD	8	8	4
XAN	5	10	3

Table 4.2: FCoN clique numbers. Median numbers of cliques in pooled sets are reported for each activity class and fingerprint. Medians were calculated from ten independent trials.

As reported in Table 4.3, the median number of features per clique was comparable for MACCS and ECFP4 (five to 13 features), whereas ACCS-FP cliques were smaller (two to six features). However, for all fingerprints, large individual cliques were also detected, in some cases containing more than 40 features, as illustrated in Figure B.2. The median numbers of features per clique for individual co-occurrence threshold values is reported in Table B.2.

Database Distribution of Cliques

The distribution of identified cliques in 500,000 ZINC database compounds was analyzed. As shown in Table 4.4 and Figure 4.4, the three fingerprints significantly differed in the average number of database compounds that contained their feature cliques.

MACCS keys produced cliques that were typically found in large numbers ($\sim$2,000 to 23,000) of database compounds. Thus, the most general of the three fingerprints produced feature combinations that were least specific for active compounds, as one would expect. However, for all three fingerprints, feature combinations were also identified that matched only a few or a single database molecule. Cliques extracted from ECFP4 occurred on average in

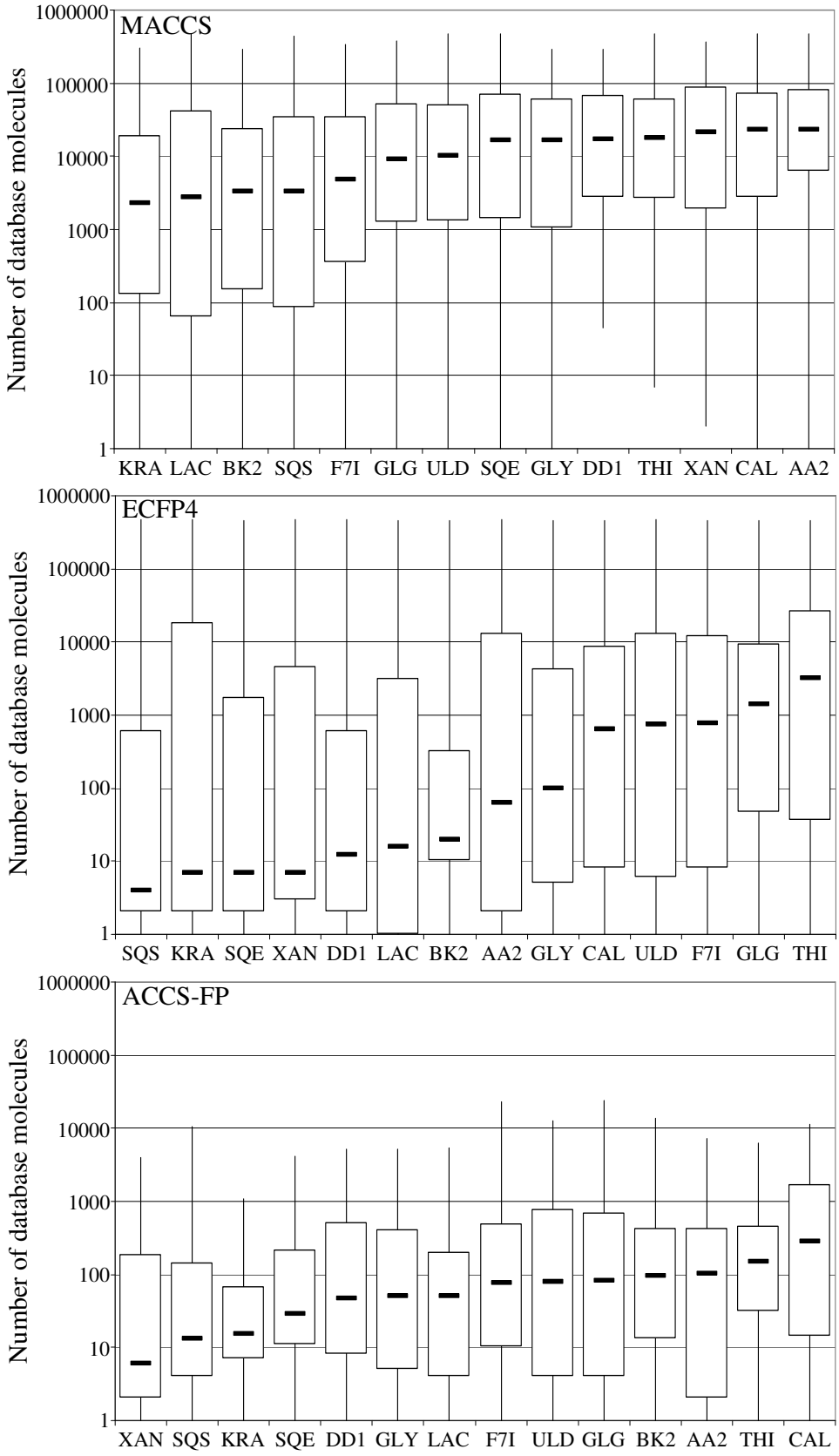


Figure 4.4: Feature clique distribution in database. Box plots show the distribution of cliques among screening database compounds. Thick bars mark median values.

Activity class	MACCS	ECFP4	ACCS-FP
AA2	156	136	21
BK2	184	715	20
CAL	133	201	15
DD1	109	229	46
F7I	248	346	13
GLG	312	325	28
GLY	138	350	24
KRA	141	154	8
LAC	184	275	32
SQE	224	296	4
SQS	239	499	83
THI	243	238	47
ULD	245	287	13
XAN	123	182	23

Table 4.3: FCoN clique size. Median numbers of features in cliques in pooled sets are reported for each activity class and fingerprint.

Activity class	MACCS	ECFP4	ACCS-FP
AA2	23193	62	101
BK2	3273	20	96
CAL	22725	636	282
DD1	17246	12	46
F7I	4886	775	77
GLG	9131	1410	81
GLY	16744	100	51
KRA	2272	7	15
LAC	2713	16	51
SQE	16490	7	29
SQS	3285	4	13
THI	17646	3173	151
ULD	10153	747	80
XAN	21485	7	6

Table 4.4: Database clique distribution. For each activity class and fingerprint, median numbers of database compounds that contained a clique are reported.

considerably smaller numbers of database compounds ($\sim$ 5 to 3,000). Nevertheless, individual cliques were also detected for both MACCS and ECFP4 that occurred in nearly all database molecules. By contrast, ACCS-FP cliques matched only $\sim$5 to 300 database molecules and the most generic ones were found in $\sim$25,000 compounds (i. e. 5% of the database). Thus, compound class-directed ACCS-FP produced the most specific feature combinations.

MACCS structural keys are frequently correlated.[85] For example, a number of individual keys account for combinations of others and, consequently, their bits are typically set on in concert. Furthermore, many ECFP4 features are overlapping and therefore describe similar or identical substructures. By

contrast, ACCS utilized for fingerprint generation are selected to be maximally independent of each other.[48] Thus, correlation effects are expected to play a different role for these three fingerprints. In particular, ACCS-FP feature combinations are not necessarily a consequence of structural fragment correlation effects, as shown in Section 3.3 for ToFI based dependency graphs. Cliques that are formed by overlapping features are expected to frequently occur in database molecules because the substructures they represent tend to be more generic than molecular substructures formed by non-overlapping features. However, the finding that all three fingerprints produced feature combinations that rarely occur in ZINC compounds also shows that MACCS and ECFP4 yield compound class-specific feature cliques that can not be attributed to general correlation effects and that can be identified using FCoN clique ranking. Thus, in addition to generic feature combinations, activity class-specific combinations are formed that can be utilized in clique searching, as described in the next section.

Generally, feature cliques identified at higher thresholds, i. e. highly conserved combinations, were smaller than cliques at lower thresholds and occurred in more database compounds. This finding indicates that fingerprint features that are highly conserved in reference sets are not necessarily a compound class-specific signature because they might also be generic and present in many different, active as well as inactive compounds. Database clique distributions at different co-occurrence threshold levels are reported in Table B.3.

Taken together, the results of FCoN clique detection show that fingerprint features can occur in activity class-characteristic combinations; this corresponds to an activity-dependent extraction of structure correlation patterns in chemical space. For all activity classes and reference sets, multiple cliques were detected. These cliques are often small in size (two to 10 features per clique) and the total number of cliques roughly scales with the length of the original fingerprints. Moreover, the distribution of reference set cliques in a large compound database is an indicator of fingerprint information content. Feature cliques of active molecules extracted from the general MACCS fingerprint are found in many database compounds, the molecule-centric ECFP4 feature combinations occur in fewer database compounds, and many feature combinations produced by the activity class-directed ACCS-FP are only matched by fingerprints of small numbers of database molecules.

4.3.2 Clique-Based Similarity Searching

The distribution of ACCS-FP cliques in active and database molecules suggested the design of a clique-based similarity search strategy that takes activity class specificity of individual cliques into account.

Feature Clique Search Strategy

For search calculations using cliques (rather than fingerprints), a strategy has been designed that takes the degree of compound class-specificity of individual cliques into account. For each clique, the number of database molecules containing all of the clique features is determined. A clique is matched to a database compound if all of its features match. Cliques are ranked in ascending order of their database frequency, i. e. cliques occurring in only small numbers of database compounds are prioritized. Database compounds are selected sequentially according to the ranked clique list. Thus, each feature clique adds compounds to the selection set that have not been retrieved in previous steps by cliques with lower overall database frequency, thereby producing a compound ranking. This selection procedure is illustrated in Figure 4.5.

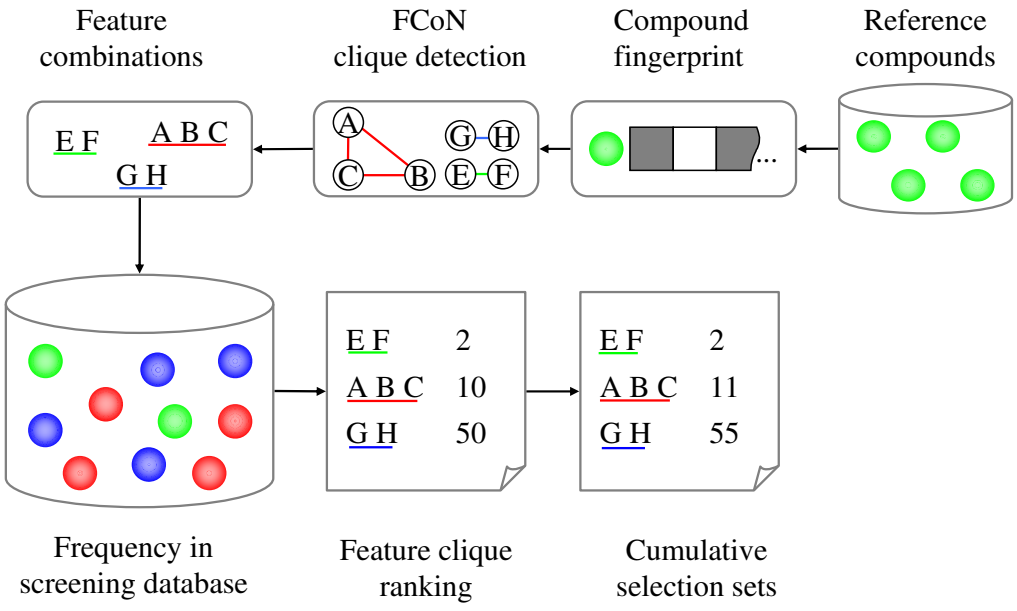


Figure 4.5: Feature clique search strategy. From fingerprints of active compounds, feature cliques are identified. For each clique, the number of database compounds containing all of its features is determined. Cliques are sorted by ascending database frequencies (i. e. rarely occurring cliques are preferred). From this ranking, cumulative compound selection sets are generated. In this example, for clique EF, the selection set contains two compounds. Cliques EF and ABC share one compound, resulting in a selection cumulative selection set of 11 (2+10-1) compounds for clique ABC.

The clique detection strategy extends the concept of conserved bit positions utilized in the centroid approach and modal fingerprints. Calculating relative frequencies of individual features does not provide information about individual feature combinations that might be present in reference set compounds. By contrast, the FCoN strategy explicitly utilizes conditional probabilities of feature co-occurrence, rather than relative frequencies. Thus, it provides an ensemble of feature cliques that can overlap and vary in size.

Computational Complexity

Feature clique searching is computationally more complex than standard fingerprint calculations because cliques need to be identified and mapped. Thus, clique searching has the complexity $O(n^2)$, with n being the number of reference compounds. The time complexity of the centroid approach scales linearly with the number of reference compounds, $O(n)$. Additionally, clique searching has $O(m^2)$ complexity with regard to the total number of unique features m. However, feature combinations are only extracted from small numbers of active reference molecules, the number of unique features in a small compound set is limited, and mapping to database compounds utilizes their fingerprint representations. Thus, additional computational costs are low and practical clique searching requirements are comparable to standard fingerprinting.

Virtual Screening Trials

FCoN selection sets were transformed into ranked compound lists with equal scores assigned to compounds belonging to identical FCoN selection sets. Such non-contiguous score distributions also occur for ranking methods when compounds are assigned the same similarity value. This binning effect is accounted for by calculating the expected number of retrieved compounds for each selection set. For example, given two selection sets with 90 and 110 compounds containing five and seven active compounds, respectively, the number of retrieved active compounds for the selection set size of 100 is calculated as follows. Five active compounds are contained in the set of 90 compounds. The remaining 10 $(100 - 90)$ compounds are randomly selected from 20 $(110 - 90)$ compounds that are additionally present in the set of 110 molecules. The expected number of additional actives in this randomly selected set is given by $(7-5)*10/20 = 1$. Thus, for a selection set of 100 compounds, the method is expected to retrieve six compounds.

Recovery rates for 100 top-ranked database compounds were calculated based on clique sets derived using different co-occurrence thresholds as well as for the pooled sets. Table B.4 reports recovery rates of active compounds at different co-occurrence threshold values and the pooled sets. Generally, pooled clique sets produced results that were better or comparable to those obtained

at individual threshold levels. Therefore, pooled cliques were used for further analysis.

Significance of Feature Combinations for Search Performance

In order to assess the contribution of feature combinations to search performance, clique-based compound selection was compared to 1NN, 3NN, centroid, and modal fingerprint search strategies on the basis of Tanimoto similarity. For modal fingerprints, three thresholds were utilized: 50%, 70%, and 100%. As a measure of search performance, active compound recovery rates for selection sets of 100 top-ranked compounds were calculated for both Tc- and clique frequency-based compound rankings and averaged over ten independent trials. Table 4.5 reports recovery rates for all activity classes and database selection sets of 100 top-ranked compounds. For modal fingerprints, the 50% threshold performed consistently better for all fingerprints than other thresholds and thus only the best results for modal fingerprints are reported.

Depending on the fingerprint, clique-based search performance substantially varied. Figure 4.6 shows recovery rate bar charts for three structurally heterogeneous (CAL, F7I, GLG) and two homogeneous (GLY, KRA) activity classes. For MACCS, feature combinations produced lower recovery rates than standard search strategies for all but one activity class, consistent with the high database frequency of many MACCS-derived cliques, which mirrors the fact that MACCS structural keys have been designed for a broad spectrum of small molecules.[63,88] For the ECFP4 fingerprint that includes features derived from individual molecules, the results of clique searching were overall comparable to the other search strategies. For six out of 14 classes, cliques produced highest recall rates. In particular, this was the case for structurally heterogeneous classes (F7I, GLG, THI) where standard search strategies had difficulties to retrieve active compounds. For ACCS-FP, clique searching produced highest recall for 11 of 14 compound classes. Similar to ECFP4, a notable improvement in recovery rates was observed for heterogeneous classes that represented difficult test cases for nearest neighbor or centroid searching. Table 4.6 summarizes the virtual screening results for the three fingerprints and different search strategies.

The consistently high performance of ACCS-FP clique searching compared to the other search strategies suggests that these cliques preferentially represent activity class-characteristic feature combinations that have high potential to retrieve active compounds, even if they are only partly conserved in reference sets. These findings also indicate that specific feature combinations, rather than individual features, contain most class-specific information in ACCS-FP. By contrast, for structural fingerprints of general design such as MACCS, feature combinations do not determine search performance. In this case, counts of individual features and fingerprint overlap are a more re-

Activity Class	FCoN	Centroid	Modal	1NN	3NN
			MACCS		
AA2	4.23	13.33	10.78	24.47	**28.33**
BK2	18.18	21.82	6.36	**26.36**	17.27
CAL	14.29	22.86	24.25	**32.86**	32.14
DD1	13.81	47.33	56.00	**65.07**	64.67
F7I	6.36	1.67	1.67	**9.17**	7.50
GLG	4.17	5.29	2.35	0.00	**7.65**
GLY	19.60	19.41	15.88	**41.67**	32.94
KRA	61.94	20.91	11.98	**82.73**	63.64
LAC	35.71	24.67	5.87	**60.67**	52.00
SQE	**35.00**	20.77	16.15	3.85	27.69
SQS	36.53	18.57	16.19	46.67	**46.90**
THI	0.97	0.59	1.76	**10.59**	7.06
ULD	3.38	**5.45**	4.74	0.00	3.64
XAN	45.48	43.33	38.33	**58.26**	56.11
			ECFP4		
AA2	32.13	28.61	0.00	46.11	**51.11**
BK2	69.09	46.36	19.09	69.09	**73.64**
CAL	**60.04**	40.71	42.86	55.71	54.29
DD1	75.69	63.33	17.68	84.67	**86.67**
F7I	**67.26**	3.33	1.67	25.00	24.17
GLG	**36.87**	0.59	0.00	6.47	33.53
GLY	**78.82**	34.71	26.47	48.82	67.65
KRA	75.45	50.00	10.00	**89.09**	86.36
LAC	72.26	40.00	2.67	**76.00**	68.00
SQE	**66.36**	31.54	3.08	40.00	63.85
SQS	48.47	25.71	0.00	57.14	**57.62**
THI	**45.86**	4.71	0.00	29.41	33.53
ULD	40.30	13.64	0.00	29.09	**42.73**
XAN	72.94	53.33	15.00	**73.33**	67.22
			ACCS-FP		
AA2	**16.76**	1.24		2.50	9.33
BK2	44.63	46.36	26.28	21.15	**48.18**
CAL	**28.76**	14.84	10.44	7.63	24.53
DD1	65.62	51.01	38.00	54.93	**73.88**
F7I	13.65	5.14	**15.14**	0.62	5.42
GLG	**35.26**	6.47		1.70	12.07
GLY	**37.86**	17.07	2.54	27.98	26.08
KRA	**57.02**	33.94	7.51	26.56	55.12
LAC	**54.99**	23.93	3.57	52.87	42.39
SQE	**32.39**	11.83	13.19	7.53	16.09
SQS	**52.89**	36.19	7.29	12.91	42.86
THI	**10.99**	1.76	6.67	8.18	3.33
ULD	**31.67**	1.82	6.82	0.36	13.78
XAN	**61.21**	42.78	22.78	28.19	55.62

Table 4.5: FCoN virtual screening performance. Average recovery rates in percent are reported. For activity classes AA2 and GLG, no bits were set on in the modal ACCS-FP. Maximal recovery rates are highlighted in bold.

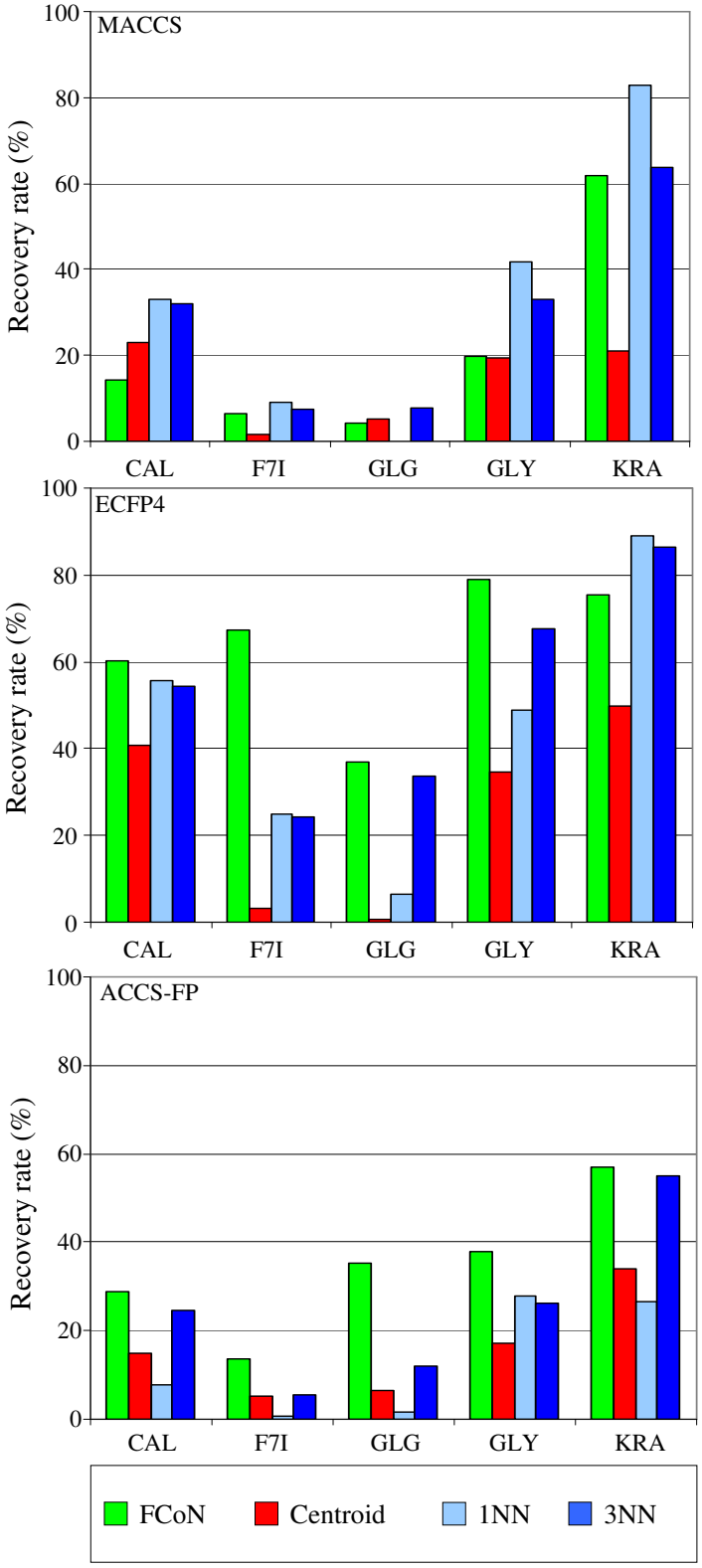


Figure 4.6: FCoN virtual screening trials. Recovery rates are shown for different fingerprints and search strategies. Modal fingerprints behaved similar to the centroid approach and are not shown.

Fingerprint		FCoN	Centroid	Modal	1NN	3NN
MACCS	RR	21.40	19.00	15.17	33.03	31.97
	Best	1	1	0	9	3
ECFP4	RR	60.11	31.18	9.89	52.14	57.88
	Best	6	0	0	3	5
ACCS-FP	RR	38.84	21.03	11.45	18.08	30.62
	Best	11	0	1	0	2

Table 4.6: Fingerprint comparison. Reported are average recovery rates ("RR", in percent) over 14 activity classes. "Best" reports the number of classes with highest recovery rate for each search strategy.

liable measure, which is exploited in the calculation of Tanimoto similarity. For ECFP4, emphasizing individual features or feature combinations produces comparable search results. For ACCS-FP, feature combinations effectively discriminate between active and database compounds and hence clique searching is much superior to established search strategies for multiple reference compounds such as nearest neighbor or centroid calculations.

4.4 Summary

In this chapter, FCoN have been introduced for the systematic extraction of feature cliques from molecular fingerprints. Clique detection in co-occurrence networks can generally be applied to identify feature combinations that are conserved in active compounds. A search strategy has been designed that uses frequency-based ranking of cliques and prioritizes database compounds that contain rarely occurring cliques.

The analysis has revealed three major findings. First, fingerprint features frequently occur in well-defined combinations that can be activity class-specific, even if individual features are of generic origin. Second, feature combinations are highly relevant for the performance of compound class-directed fingerprints, in contrast to generic fingerprints like MACCS. Class-directed cliques rarely occur in the screening database and are capable of significantly enriching selection sets with active compounds. Third, clique-based similarity searching represents a generally preferred strategy for ACCS-FP.

These results are consistent with the identification of activity class-specific pathways in fragment dependency graphs (Chapter 3) and show that combinations of individual structural features often become activity class signatures.

Chapter 5

Fragment Formal Concept Analysis for the Assessment of Complex SARs

In this chapter the activity signature character of fragment combinations is analyzed in more detail. In particular, the focus lies on related activity classes and fragment combinations that they share or are distinguished by.

An adaptation of formal concept analysis (FCA), a data mining and visualization technique originally developed in information science in the 1980s[51] is introduced. First, the general methodological foundation of FCA is described. Then, *Fragment Formal Concept Analysis* (FragFCA) is introduced, which is designed to extract fragment combination signatures of compound sets with non-trivial SAR profiles including multiple activities. Also, fragment combinations specific for defined compound potency ranges can be identified. On the basis of signature fragment combinations extracted using FragFCA, a classification method has been developed that assigns compounds to one of several closely related targets.

5.1 Formal Concept Analysis

Formal concept analysis (FCA) is a method for data analysis, knowledge representation and information management.[51] FCA organizes relationships between a set of *objects* and a set of *attributes* in *concepts* and represents this organization in *concept lattices*.

5.1.1 Concept Lattices

The input to FCA is a *formal context*, which describes binary relationships between objects and attributes of the general form "is" and "is not", e. g. "Aspirin

(object) *is* a cyclooxygenase inhibitor (attribute)". Formal contexts can be reported as a two-dimensional matrix with attributes in columns and objects in rows. Figure 5.1 shows an exemplary formal context describing the distribution of fragment combinations among molecules with different biological activity. Here, the relationships have the form "fragment combination A is/is not contained in molecule B".

Formal concepts are defined as sets of objects that share a set of attributes. Objects and attributes are connected by a so called *Galois connection*. This connection implies that the set of objects can not be extended by additional objects without covering additional attributes and the attribute set can not be extended without losing some of the objects. For any given formal context, the formal concepts are unambiguously defined.

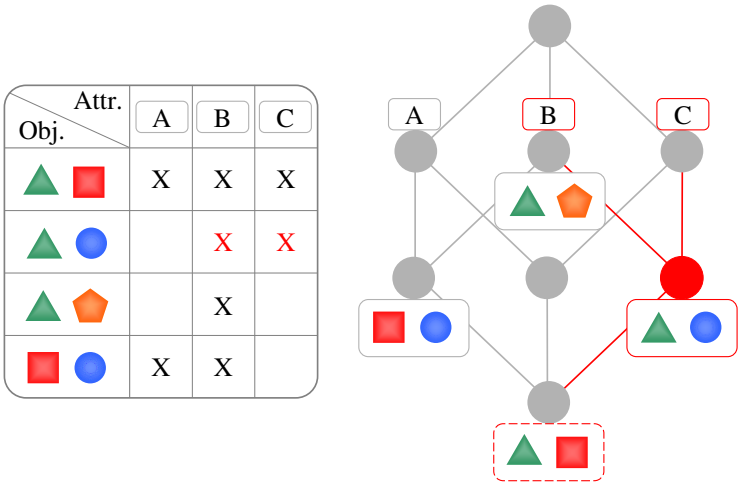


Figure 5.1: Formal concept analysis. A formal context is shown on the left that describes the distribution of formal objects (fragment combinations in FragFCA, colored shapes) among attributes (activity classes in FragFCA) A, B, and C. An "X" indicates occurrence of the combination in the particular activity class. On the right, the corresponding concept lattice is shown. Each node represents a concept. Attributes are written above, objects below nodes. The concept containing fragment combinations characteristic of classes B and C is selected and shown in red. The activity annotations are found by following edges towards the top, the corresponding fragment combinations by following edges towards the bottom of the lattice. The fragment combination inside the dashed box belongs to the concept, but is not selected because it also occurs in activity class A.

Individual formal concepts are not independent of each other, but share objects and/or attributes. These relationships are visualized in concept lattices, as shown in Figure 5.1. Each node in a concept lattice corresponds to one particular concept. Attributes are written above and objects below nodes. The attributes of a concept can be found by tracing all paths towards the top node. Objects that belong to the concept and do not share any other attributes

reported in the lattice are associated directly with the node. In order to extract all objects belonging to the concept, the paths towards the bottom node are followed. Hence, the concept represented by the top node contains all objects, whereas the concept represented by the bottom node contains all attributes.

5.1.2 Scales

For large formal contexts with many attributes, concept lattices quickly become difficult to read and navigate. For example, for three independent attributes eight (2^3) possible concepts exist, whereas for four attributes, 16 (2^4) nodes have to be incorporated into one single lattice in order to cover all possible concepts. This problem is addressed by introducing *scales* that focus on subsets of attributes. Scales correspond to formal contexts that have identical objects but different attributes. Figure 5.2 shows multiple scales. For each scale, a concept lattice is generated. In the following, concept lattices that are based on a specific scale will simply be referred to as "scale". Each scale thus provides information about the distribution of objects among a subset of attributes.

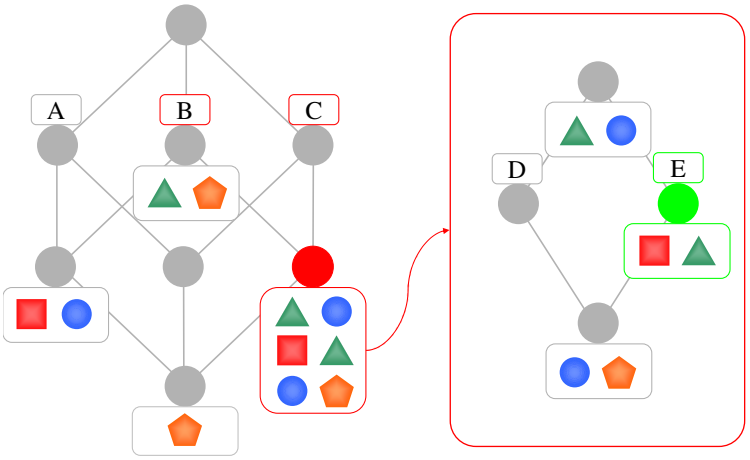


Figure 5.2: FCA scales and scale combination. Two scales are shown focusing on different subsets of activity annotations: "A, B, C" and "D, E". On the first scale, fragment combinations are selected that are characteristic of activity classes B and C, but do not occur in class A (red box). This selection is projected onto the second scale, which reports fragment distribution among classes D and E. On this scale, a subset of the selection is chosen that occurs exclusively in activity class E. Thus, the combination in the green box occurs in B, C, and E, but it is not present in any compound with activity A or D.

Scales are combined in order to extract objects that share attributes reported in different scales. Queries that utilize scale combination play a central role in FCA. During the query procedure, a subset of objects selected on one particular scale is projected onto following scales, allowing the selection of a

(sub-)subset based on different attributes. Figure 5.2 shows how scales are combined in order to define a query.

5.2 Fragment Formal Concept Analysis

This section introduces an FCA adaptation for the analysis of fragment combination distributions among bioactive compounds binding to different but closely related targets. In FragFCA, fragment combinations are formal objects and ligand activity and potency annotations are attributes. Scales are designed to account for a hierarchical organization of desired targets. Furthermore, scales are defined that focus on different potency ranges for specific targets and also report the number of compounds a fragment combination occurs in. The combination of different scales allows the interactive definition of queries for the extraction of signature fragment combinations. It is shown that these signatures are specific for molecules with a defined activity and/or potency profile.

5.2.1 Definition of the Formal Context

Data Set

A previously reported[89] and publicly available set of 267 biogenic amine GPCR antagonists has been utilized in order to define the formal context. Compounds in this set are active against multiple receptors at different potency levels. A compound was assigned to a class if it was active against the target receptor with an IC_{50} value of $10\mu M$ or lower. On the basis of this threshold value, the 267 antagonists received a total of 687 activity assignments. The activity classes and number of activity annotations are reported in Table 5.1.

Activity class	Biological activity	Ann.
D1	Dopamine D1 receptor antagonists	84
D2	Dopamine D2 receptor antagonists	216
D3	Dopamine D3 receptor antagonists	75
D4	Dopamine D4 receptor antagonists	93
5-HT1A	Serotonin 5-HT1A receptor antagonists	95
5-HT2A	Serotonin 5-HT2A receptor antagonists	32
$\alpha 1$	Adrenergic $\alpha 1$ receptor antagonists	92

Table 5.1: FragFCA GPCR dataset. The 267 compounds in this set have multiple activities and represent a total of 687 activity annotations ("Ann.").

Fragment Generation

A hierarchical fragmentation scheme has been applied that divides compounds into rings, linkers, and side chains (see Section 2.2.2). As a refinement of conventional hierarchical fragmentation,[41] condensed rings are not only considered as fragments, but are also further divided into non-fused individual ring components. This is done in order to increase the information content of fragment ensembles. Fragments were generated from all GPCR antagonists and pooled. From the initial set of 701 unique fragments, small fragments with fewer than four atoms and large fragments with more than 20 atoms were removed, resulting in a final set of 427 fragments.

Enumeration of Fragment Combinations

From these 427 GPCR fragments, a structural key-type fingerprint was generated and calculated for each of the 267 antagonists. For each compound, all individual fragments, pairs, triplets, and quadruplets were extracted from its fingerprint representation, yielding a total of 231,464 different combinations, which are formal objects in FragFCA. Because fragment combinations are enumerated from fingerprints, FragFCA can be applied to any fragmentation scheme and structural key-type fingerprint representation.

Activity Annotation of Fragment Combinations

Fragment combinations were annotated with qualitative and quantitative compound activity information, which represent formal attributes in FragFCA. An antagonist was considered active against a GPCR target if its IC_{50} was equal to or below $10\mu M$ and inactive if it was above this value. Furthermore, five different potency ranges were distinguished for active compounds: $\leq 1nM$, $1nM - 10nM$, $10nM - 100nM$, $100nM - 1\mu M$, and $1\mu M - 10\mu M$. If a fragment combination was found in several active compounds with different activity or potency, it was annotated with multiple activities or potency ranges, respectively.

5.2.2 Fragment Formal Concept Analysis

FragFCA Scale Design

Two types of scales have been designed that reflect compound activity distribution among the seven GPCR targets. First, *global* scales were used to qualitatively compare multiple compound activity classes at different levels of detail. Second, *specific* scale types were defined, namely a frequency, activity, and three potency scales for each activity class. The frequency scale determines

the number of active compounds that contain a particular fragment combination. The activity scale distinguishes between fragment combinations that occur only in active, active and inactive, or only in inactive compounds. Potency scales differentiate active compounds according to potency ranges. Scales were designed using the freely available ToscanaJ package.[90]

Figure 5.3 shows the target hierarchy and global scales utilized to distinguish different activity classes and Figure 5.4 shows prototypic specific scales that were calculated for each activity class.

For the generation of global scales, biological activity has been defined in an "all or nothing" manner, i. e. through application of a $10\mu M$ threshold level. However, using scales, other criteria can be readily applied. For example, by combining potency and frequency scales, queries with varying threshold levels for activity and/or fragment occurrence can be designed. Here FragFCA has been applied to 267 active compounds, but compound numbers are not a principal limit of FragFCA.

FragFCA Query Design

The two scale types represent versatile and intuitive tools to build specific and increasingly complex fragment queries. Global scales enable the comparison of biological activity profiles at different levels of detail, while specific scales provide information, for example, about the potency characteristics or frequency of occurrence of individual fragment combinations. The utility of FragFCA goes beyond fragment frequency analysis that has been applied in a number of previous studies[36,44,46] and extends fragment signature identification from individual activity classes to activity and potency profiles.

Combinations of scales define queries that are capable of revealing different types of fragment-based relationships between active compounds, as demonstrated in the following. Due to the presence of overlapping activities and differences in potency, the GPCR antagonists analyzed in this study present complicated SARs. For the analysis of these compound sets, four global GPCR scales were used and, in addition, five specific scales for each of the seven GPCR targets, resulting in a total number of 39 scales.

A major goal of FragFCA is the identification of molecular fragments and fragment combinations that are specific for complex compound activity or potency profiles. These fragment combinations can be used to distinguish ligands of closely related targets from each other.

Nonredundant Fragment Sets

From each fragment set retrieved by a query, redundant fragment information was omitted by removal of fragment combinations that contained selected

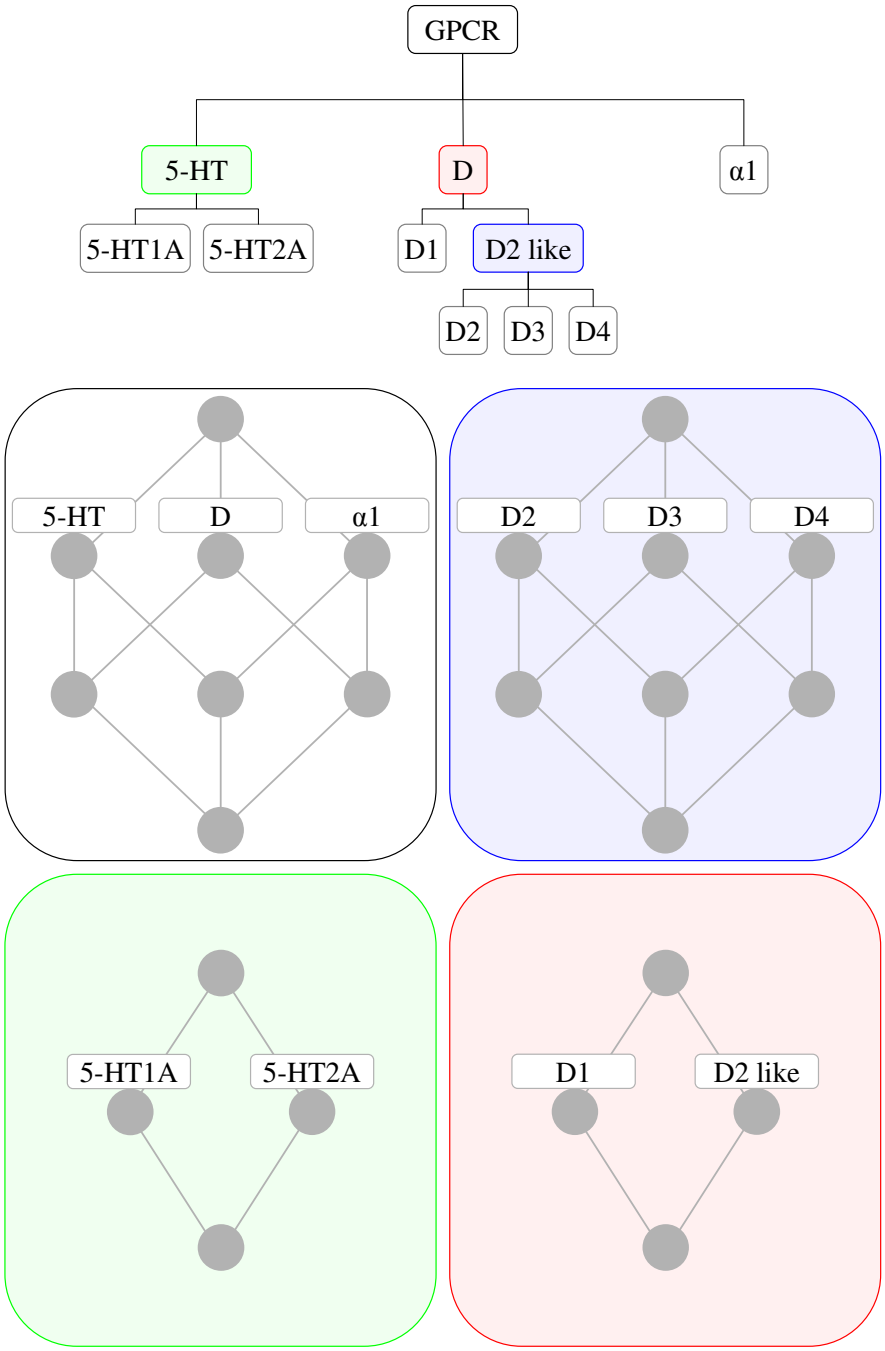


Figure 5.3: General GPCR scales. The tree structure (top) reflects the hierarchical organization of global scales based on target families. Each node and its children correspond to a particular scale (with attributes provided by the children). At the bottom, the different scales are shown. Parents in the tree and corresponding scales are color-coded.

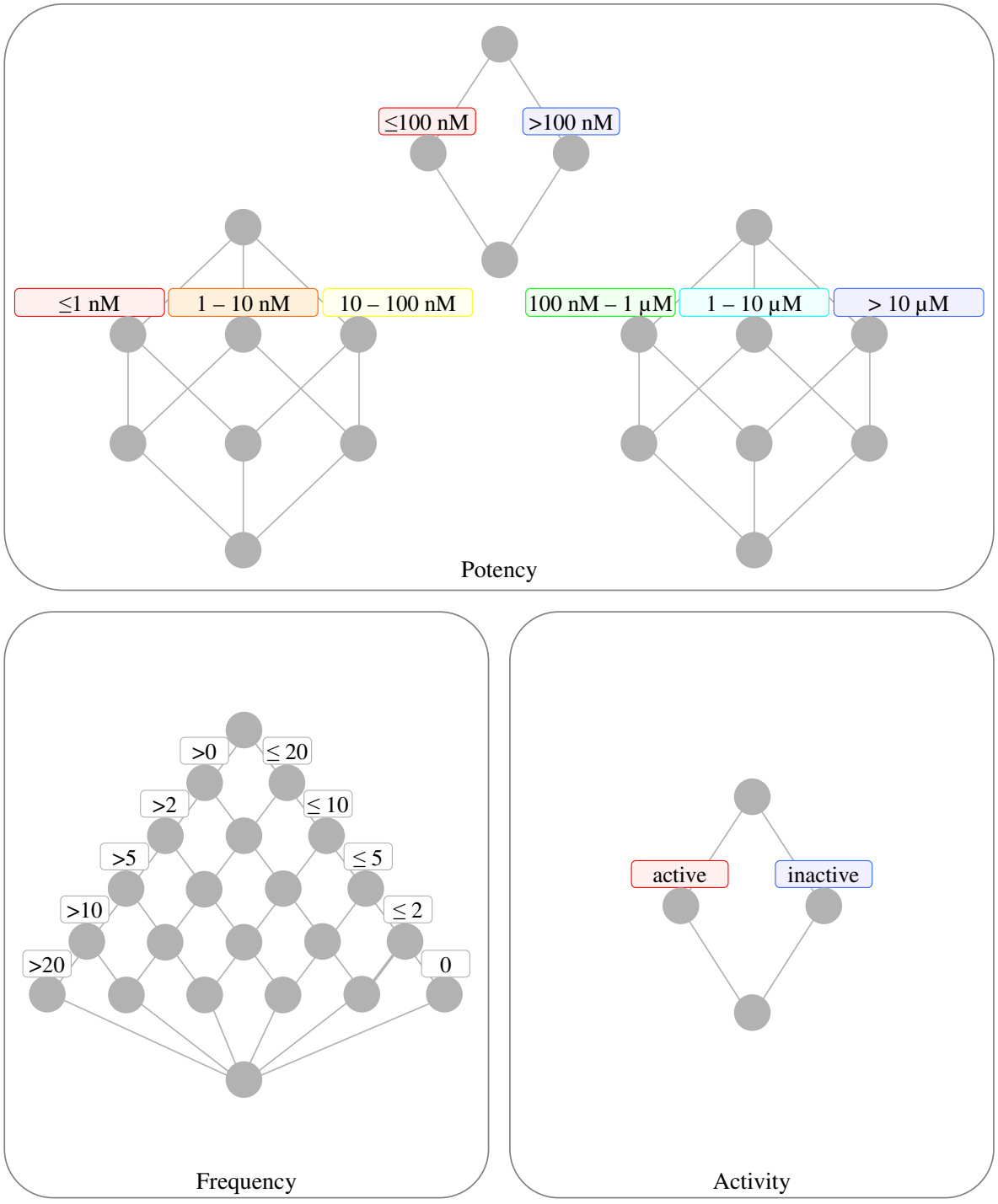


Figure 5.4: Specific GPCR scales. Three types of specific scales are defined for each target. Potency scales account for different potency levels. Frequency scales report the number of molecules a fragment combination occurs in. Activity scales distinguish between fragment combinations that are only present in active compounds and fragment combinations that also occur in inactive molecules.

signature singletons, pairs, or triplets as subsets. Figure 5.5 illustrates the redundancy filtering of fragment combinations that have been identified as an activity profile signature by FragFCA.

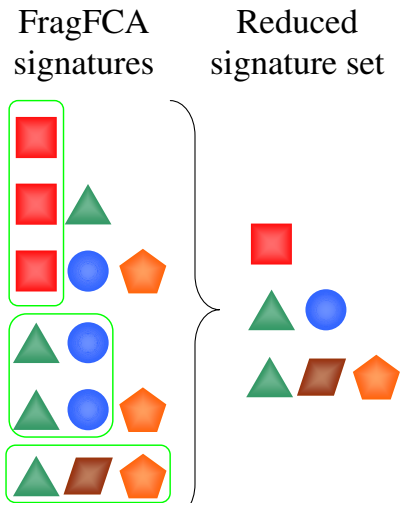


Figure 5.5: Redundancy filtering. Signature fragment combinations identified by FragFCA are filtered to eliminate redundant information. Signature combinations that contain other combinations are removed from the set.

5.2.3 FragFCA Queries

Fragments Characteristic of Dopamine and α_1 Receptor Antagonists

First, fragment distributions in different activity classes have been compared. As an example, fragment combinations characteristic of dopamine antagonists were determined. The D1 activity scale was used to extract 41,049 fragment combinations that occurred in compounds active against D1, but not in inactive compounds. Figure 5.6 shows that 90% (37,098) of these combinations only occurred in dopamine receptor antagonists, 3,367 fragment combinations were shared with serotonin receptor antagonists, but did not occur in α_1 ligands, and 584 combinations were shared by 5-HT, D, and α_1. As also shown in Figure 5.6, 3,367 and 584 shared fragments were unevenly distributed in serotonin receptor antagonists; they mostly occurred in 5-HT1A, rather than 5-HT2A antagonists. In addition, none of the fragment combinations found in α_1 ligands also occurred in 5-HT2A antagonists.

Next, the three fragment subsets were analyzed using the global D2 scale. Most fragment combinations were found to be D2 specific (Figure 5.6). None of the fragments shared with serotonin receptor ligands were specific for either D3 or D4 and the 584 fragment combinations shared among all classes on the global GPCR scale only occurred in D2, but not D3 or D4 antagonists. This

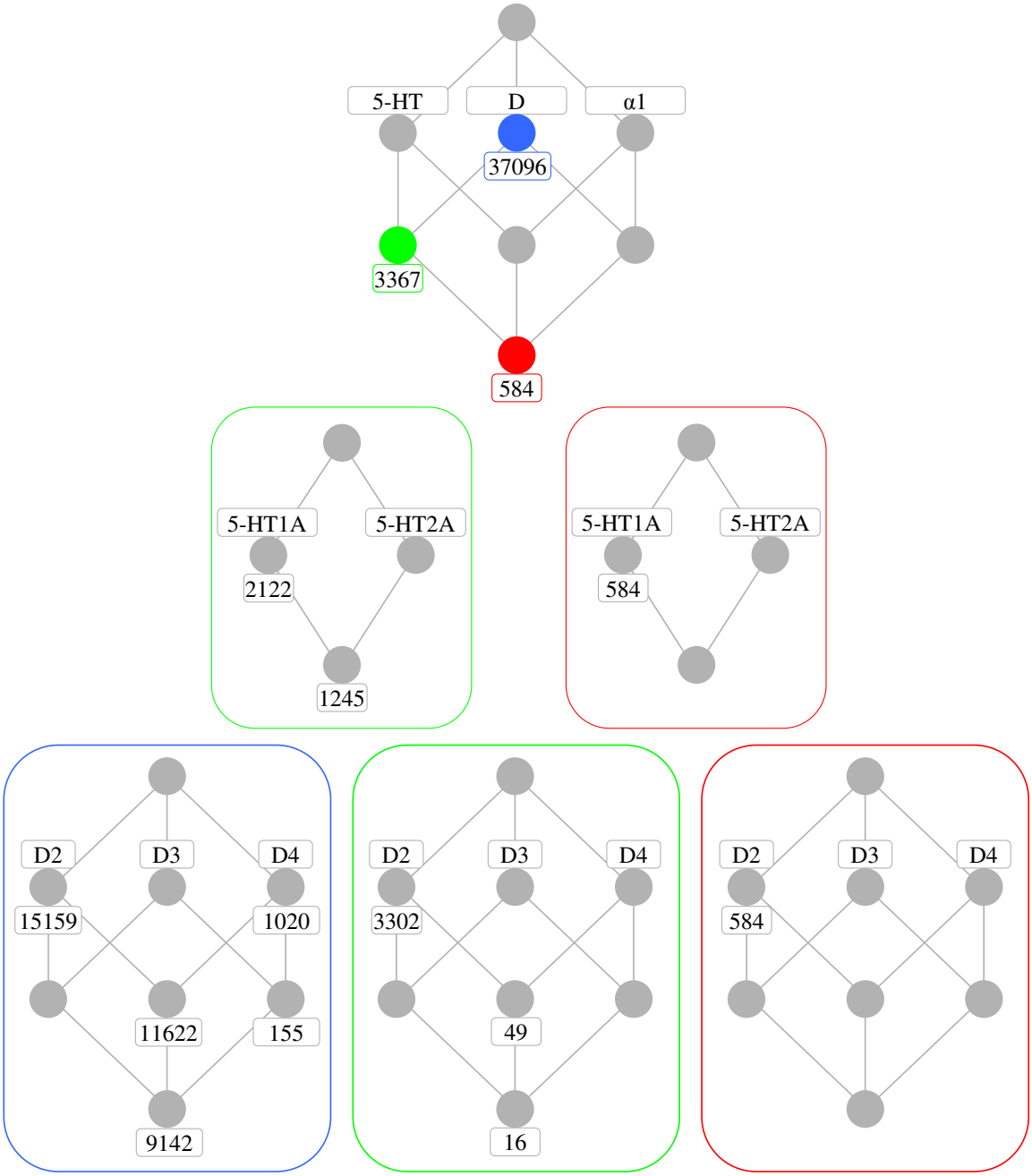


Figure 5.6: D1 signature fragment distribution. At the top, the distribution of 41,049 fragment combinations occurring in D1 antagonists among other targets is shown on the GPCR global scale. Two subsets are selected and projected onto the 5-HT scale and the D2 like scale, respectively. The color-code of the boxes corresponds to each selected subset.

example demonstrates that global scales allow the assessment of complex fragment combination distributions for several activity classes. Using these scales to build queries, signature fragments and combinations can be easily identified and compared. An example is shown in Figure 5.7. The benzimidazol-2-one fragment in the center is found in both D2 and serotonin receptor antagonists. However, in combination with each of the surrounding fragments, it is only present in D2 antagonists.

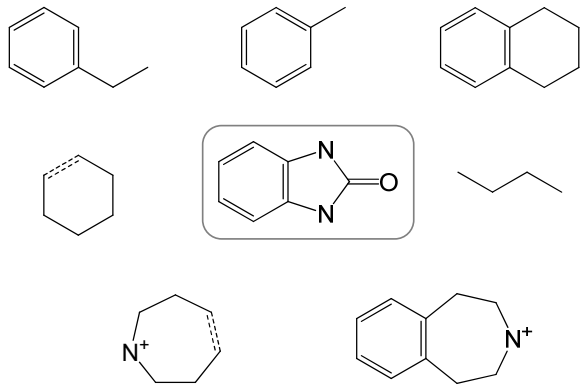


Figure 5.7: D1 signature fragment combinations. The boxed benzimidazol-2-one fragment in the center is found as a singleton in both D2 and serotonin antagonists. By contrast, the combination with each of the surrounding fragments (yielding pairs) only occurs in D2 antagonists.

The same type of FragFCA analysis was carried out for α_1 antagonists. In contrast to dopamine receptor antagonists, no α_1-specific fragment combinations were found, as shown in Figure 5.8. However, fragments with dual receptor specificity existed. For example, a subset of 27,777 fragment combinations was identified that only occurred in α_1 and serotonin receptor antagonists and that could be further reduced to 246 non-redundant combinations. The composition of this non-redundant set is reported in Figure 5.9. As can be seen, the set is dominated by fragment pairs (63%).

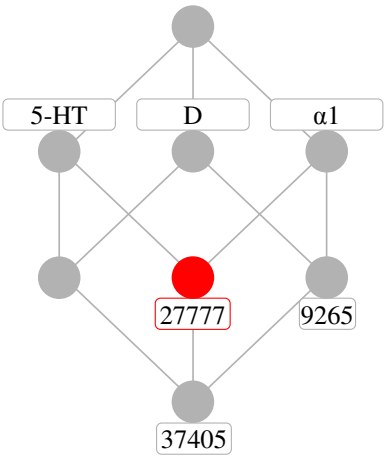


Figure 5.8: $\alpha 1$ fragment distribution. The distribution of fragment combinations occurring in $\alpha 1$ antagonists is shown on the global GPCR scale. A subset of fragment combinations not occurring in dopamine antagonists is selected.

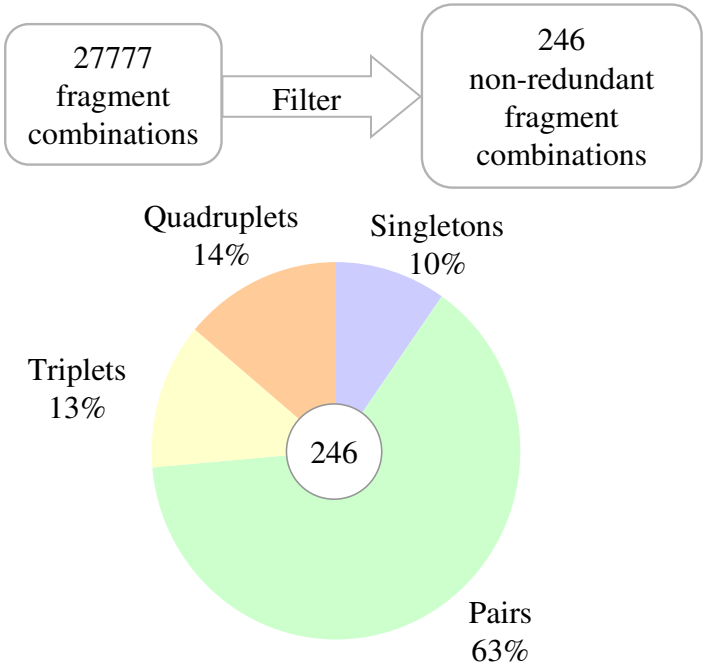


Figure 5.9: Fragment combinations in $\alpha 1$ and serotonin antagonists. The pie chart reports the distribution of 246 non-redundant combinations. 76% of all combinations are pairs or triplets.

Fragments Distinguishing α_1 and D2 from 5-HT Receptor Antagonists

Next, more complex fragment relationships were determined. Fragment combinations were extracted that were shared by α_1 and D2 antagonists, but did not occur in serotonin receptor antagonists. Therefore, the α_1 and D2 activity scales and the global GPCR scale were applied in a sequential manner. A total of 42,265 fragment combinations were extracted shared by α_1 and D2 antagonists and not present in compounds inactive against these two receptors; 9,265 of these combinations did not occur in serotonin receptor ligands. These fragments were reduced to a non-redundant set consisting of 98 fragment combinations. These fragment combinations correctly retrieved all 18 compounds from the GPCR set having the corresponding activity profile (i. e. active against D2 and α_1 but not 5-HT). Moreover, all 18 compounds were described by two fragment pairs covering 12 and 14 compounds, respectively, as shown in Figure 5.10.

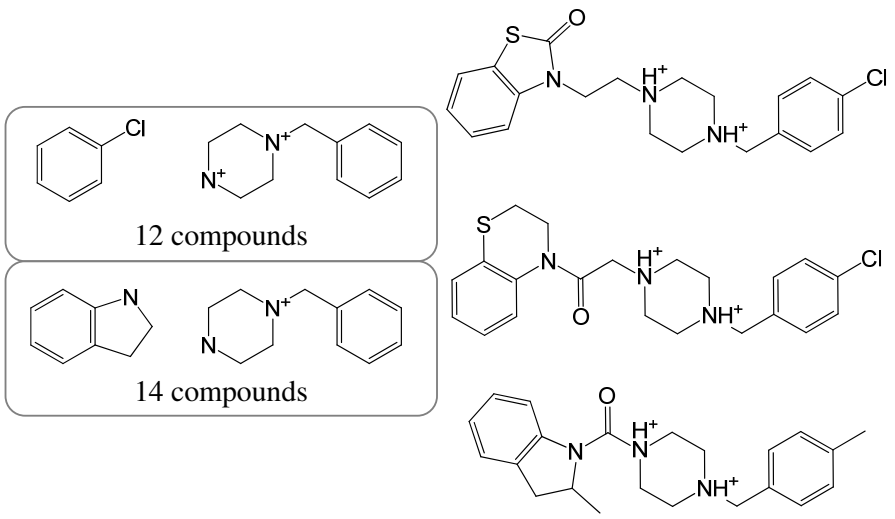


Figure 5.10: Fragment combinations specific for α_1 and D2 against 5-HT antagonists. The minimal set of fragment combinations describing all desired compounds is shown at the left. For each fragment pair, the number of identified molecules is reported. On the right, three representative antagonists are shown that are detected by these fragments.

Fragments Specific for 5-HT1A Receptor Antagonists

A minimal set of fragment combinations that distinguish 5-HT1A from dopamine and α_1 antagonists has been identified. Therefore, the activity scale specific for 5-HT1A was used and fragment combinations were selected that did not occur in 5-HT1A inactive compounds. This query was then further refined using the GPCR global scale. 3,311 combinations from this set were

shared between serotonin and dopamine antagonists, but were not present in α_1 ligands. Most fragment combinations were shared among all three classes or were specific for serotonin receptor antagonists (35% each). A total of 37,641 fragment combinations specific for 5-HT1A were selected and reduced to a non-redundant set of 199 unique fragment combinations. Again, fragment pairs and triplets constituted the major part (80%) of all specific combinations.

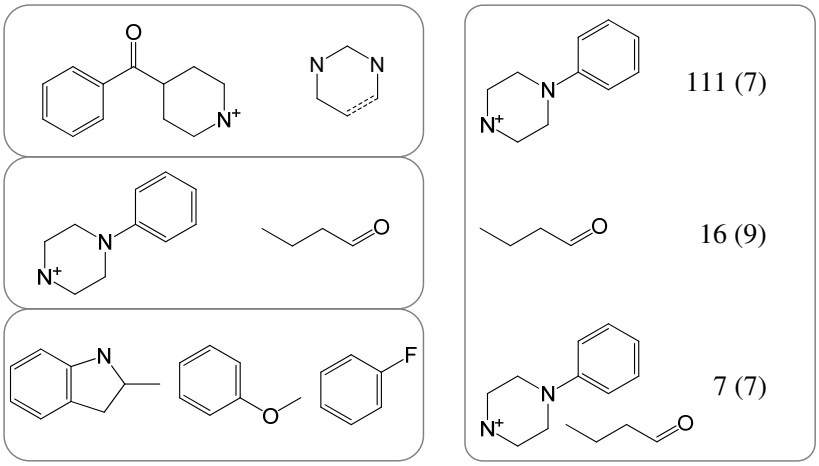


Figure 5.11: Fragment combinations specific for 5-HT1A antagonists. The left panels show three fragment combinations that distinguish all 5-HT1A antagonists from dopamine and α_1 antagonists. The right panel reports the distribution of individual fragments belonging to the second fragment pair (butaldehyde and cationic phenylpiperazine). Reported is the total number of matched molecules and, in parenthesis, the number of matched 5-HT1A antagonists.

Figure 5.11 shows the minimal set of fragment combinations that consisted of two fragment pairs and one fragment triplet. This minimal set identified all nine compounds in the database having the corresponding activity profile (i. e. active against 5-HT1A, but not dopamine or adrenergic receptors). The second fragment pair (butaldehyde and cationic phenylpiperazine) described seven of the nine compounds. This pair was further analyzed with respect to the selectivity of its individual fragments. As also shown in Figure 5.11, the individual fragments were not specific for 5-HT1A antagonists. By contrast, the combination of these two fragments was specific.

Fragments in Potent 5-HT1A and D4 Receptor Antagonists

The 5-HT1A query discussed above was further refined using the 5-HT1A potency scale to initially extract 17,225 fragment combinations occurring in compounds with $\leq 100nM$ potency. The GPCR antagonist set contained six spe-

cific 5-HT1A antagonists at this potency level. Then, the 5-HT1A frequency scale was used to select 3,542 fragment combinations that occurred in all of these compounds. The non-redundant set contained 32 fragment combinations, 27 of which were pairs or triplets. Thus, adding two scales to the pre-defined 5-HT1A query made it possible to identify fragment combinations that were specific for a subset of potent 5-HT1A antagonists.

In order to search for signature fragments of highly potent D4 antagonists, the D4 potency scale was directly applied to extract 35,593 fragment combinations occurring only in D4 antagonists with $\leq 100nM$ potency. The query was further refined using a D4 high potency scale that selected 588 combinations specific for the highest potency range ($\leq 1nM$), as shown in Figure 5.12. The reduced set consisted of nine fragment combinations that identified four of six D4 antagonists with $\leq 1nM$ potency present in the database, and detected no other compounds.

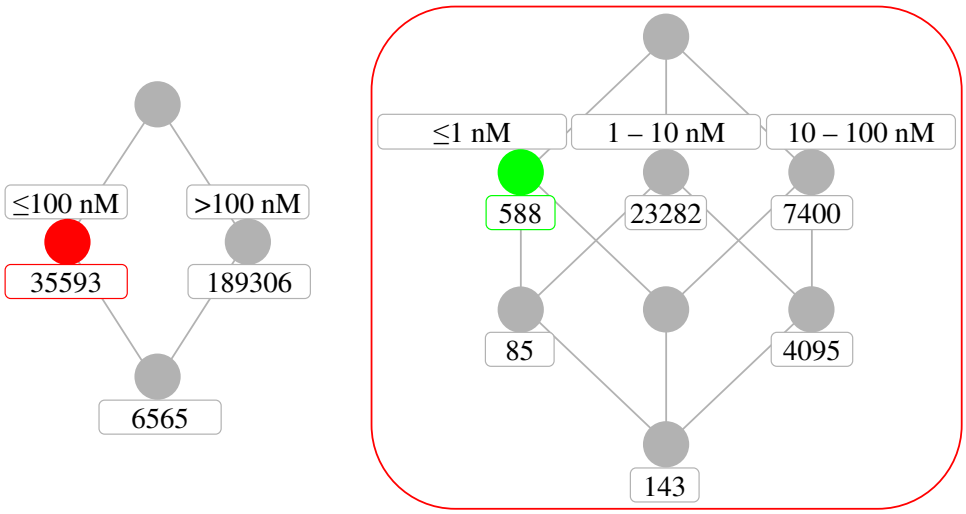


Figure 5.12: Fragment combinations specific for highly potent D4 antagonists. Specific D4 potency scales are combined in order to extract 596 fragment combinations that occur in highly potent D4 antagonists.

5.2.4 Query Summary and Concluding Remarks

The results for all GPCR selectivity queries are summarized in Table 5.2. FragFCA signature sets identified all relevant GPCR ligands in three of four cases, and four of six relevant ligands for the last, most complex, query. No other ligands were mapped by the identified fragment combinations.

Query	Signatures	Cmpds.	Recovered cmpds.
α_1 and D2 vs 5-HT	98	18	18
5-HT1A vs D and α_1	199	9	9
Potent selective 5-HT1A	32	6	6
Potent selective D4	9	6	4

Table 5.2: FragFCA GPCR queries. For each selectivity query, the number of reduced fragment combinations ("Signatures"), the total number of available selective compounds ("Cmpds."), and the number of compounds correctly identified by the query ("Recovered cmpds.") are reported.

The application of FragFCA to the GPCR ligand set has shown that FragFCA is capable of extracting signature fragment combinations that are highly descriptive of complex SARs including potency information. The distribution of fragment combinations among ligands with overlapping activities is visualized in concept lattices that represent defined scales, which focus on individual activity classes or potency ranges. Scales are combined in order to design fragment queries of increasing complexity and provide signature fragment combinations that retrieve compounds with the desired activity and/or potency profile.

Moreover, combinations of two or three fragments are most relevant for distinguishing between different activity profiles or compound potency levels, rather than single fragments or quadruplets. The results demonstrate that FragFCA provides an easy and systematic access to structural signatures of complex profiles in activity space that exploit structural information encoded in feature combinations, rather than individual fragments.

5.3 FragFCA Classifier

The predictive power of FragFCA signatures has also been evaluated. Therefore, eight supersets have been assembled consisting of two to four closely related activity classes from the MDDR. Here, the FragFCA approach is extended by designing queries for selectivity analysis within supersets of closely related targets in a systematic manner. It is demonstrated that fragment signatures resulting from these queries can be successfully applied for compound classification and virtual screening of external data sets.

5.3.1 Data Sets and Fragment Generation

A total of 24 different activity classes were assembled from the MDDR containing between 65 and 1,994 agonists, antagonists, or inhibitors. These activity classes were grouped into eight supersets of compounds active against one of two to four closely related targets. In addition to the MDDR sets, two supersets that shared no compounds with the MDDR sets were assembled from BindingDB. The composition of MDDR and BindingDB supersets is reported in Table B.5.

The hierarchical fragmentation scheme described in Section 5.2.1 has been applied to each MDDR compound superset. Fragments containing between four and 12 heavy atoms were sampled that occurred more than once in a superset fragment population. Table B.5 reports the size of the resulting fragment library for each superset, ranging from 40 to 676 fragments.

5.3.2 Scale and Query Design

Scales and queries have been designed for each superset that correspond to general scales described in Section 5.2.2. For supersets containing two or three targets, single scales were utilized. For supersets containing four targets, three scales were defined. Thus, taken together, a total of 12 scales were used in order to identify signature fragment combinations for 24 activity classes.

Each activity class was 10 times randomly divided into a reference and a test set. Reference sets contained one third of the activity class, but maximally 50 compounds. For all reference compounds, combinations of up to three fragments were enumerated from their fingerprint representations based on the corresponding superset fragment library. Using FragFCA, fragment combinations were isolated for each activity class that only occurred in its reference compounds but no other activity classes belonging to the superset. This systematic scale and query design has been implemented in MOE.

FragFCA queries consistently identified significant numbers of signature fragment combinations for all 24 activity classes in eight supersets. Average numbers of fragment combinations per class ranged from 28 for adenosine A2

agonists to 13,161 combinations for thrombin inhibitors, as reported in Table B.5.

5.3.3 Compound Classification

Sets of activity class-specific fragment combinations were utilized to establish ligand-target relationships by mapping them onto test set compounds. For a successful match, all fragments of a signature combination had to be present in a test compound. Given the fragment combination set of an activity class, the relative frequency of successfully mapped fragment combinations yields a *fragment score* in the interval [0, 1] for each test set compound.

The FragFCA classifier introduced here utilizes signature sets to predict the activity class of database molecules with activity against the target family. A test compound is assigned to the activity class for which it produces the largest fragment score. Protocols for compound classification calculations were implemented in Pipeline Pilot.

The identified signature fragment combinations were applied to classify test compounds taken from each superset. For seven of eight supersets, on average more than 95% of test compounds were successfully matched by signature fragment combinations. Moreover, $\sim$40% of the test compounds only matched fragment combinations of a single class, thus demonstrating the high specificity of fragment combinations identified using FragFCA. Only for one superset, Adenosine, fewer test compounds, $\sim$50%, were matched. Table B.6 reports detailed compound classification results and Figure 5.13 shows the distribution of correctly classified, incorrectly classified, and unclassified test compounds for each superset. On average, more than 75% of matched test compounds were correctly classified for all supersets.

The results for the Adenosine superset illustrate that the presence of only a limited number of mapped combinations did not negatively affect classification accuracy because 80% of the mapped compounds in this superset were correctly classified. Furthermore, the number of correctly classified compounds that matched fragment combinations of only a single activity class was analyzed (Table B.6). More than 90% of single class matches were found to represent correct predictions, which further illustrates the high degree of specificity of the structural information captured by fragment combinations. Taken together, these findings show that FragFCA consistently identified signature fragment combinations that successfully discriminated between compounds with closely related biological activities.

To complement benchmark calculations on the MDDR compound supersets, signature fragment combinations derived for MDDR phosphodiesterase and serine protease inhibitors were also utilized to map non-overlapping corresponding supersets assembled from BindingDB (see Table B.5). For these

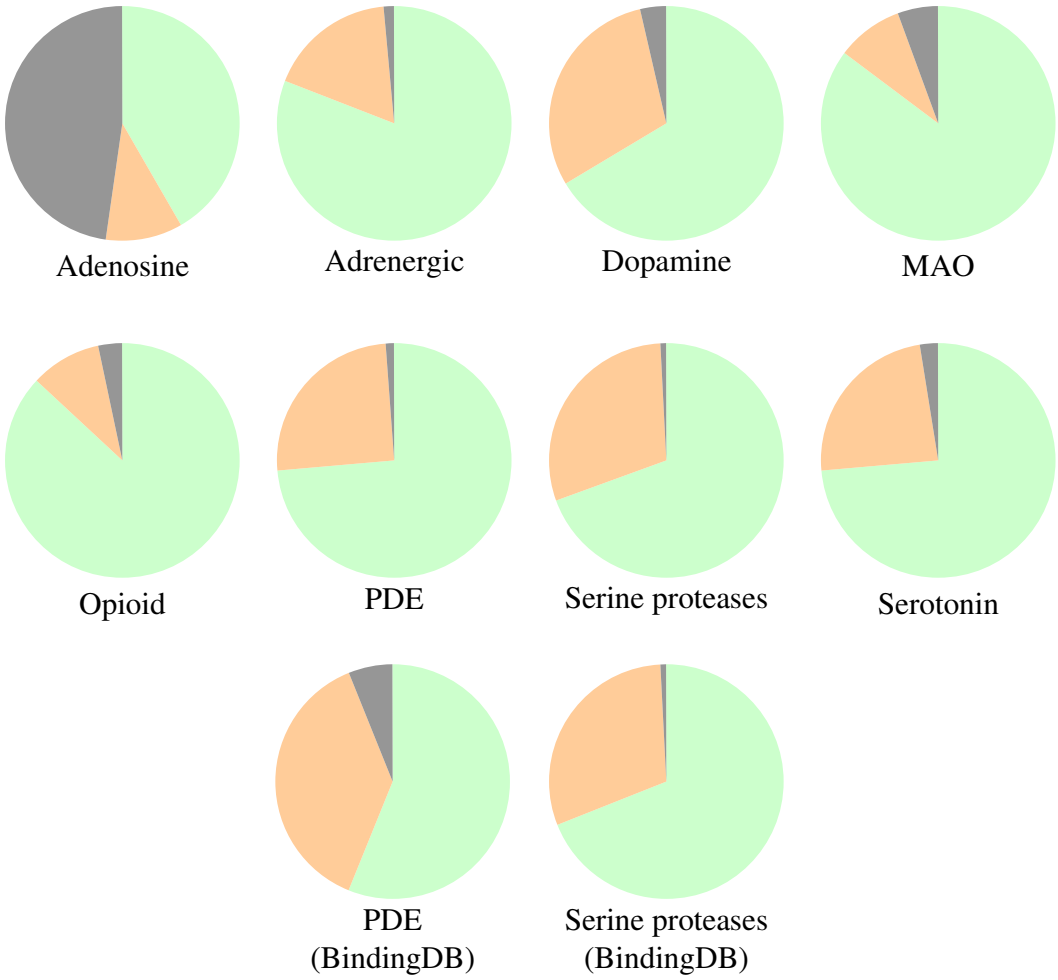


Figure 5.13: FragFCA classification results. Pie charts show the fraction of correctly classified (green), incorrectly classified (red) and not mapped (grey) compounds for each superset. In addition to the MDDR supersets, results for the two BindingDB sets are shown.

external test compounds, no fragments were derived. As shown in Figure 5.13, comparable prediction accuracy was achieved for BindingDB molecules when MDDR-derived fragment combinations were mapped. On average, 56% of the available phosphodiesterase inhibitors and 69% of serine protease inhibitors were correctly classified.

5.3.4 Similarity Searching

In order to evaluate the utility of fragment signature sets for similarity searching, the MDDR test sets were added to a randomly selected ZINC background of 100,000 molecules and fragment mapping of the resulting database was carried out. In a first step, database compounds matching only a single signature set were selected. The selected compounds were then ranked according to their fragment score, i. e. the relative number of fragment combinations that matched the compound. For the ranked database compounds, "Area Under the Receiver Operating Characteristic Curve" (ROCAUC) calculations were carried out to monitor compound recall. A ROC curve is constructed by plotting the true positive rate against the false positive rate for increasing selection set size. ROC curves illustrate the recall performance of a classifier. Figure 5.14 shows three exemplary ROC curves for supersets Opioid, Dopamine, and MAO.

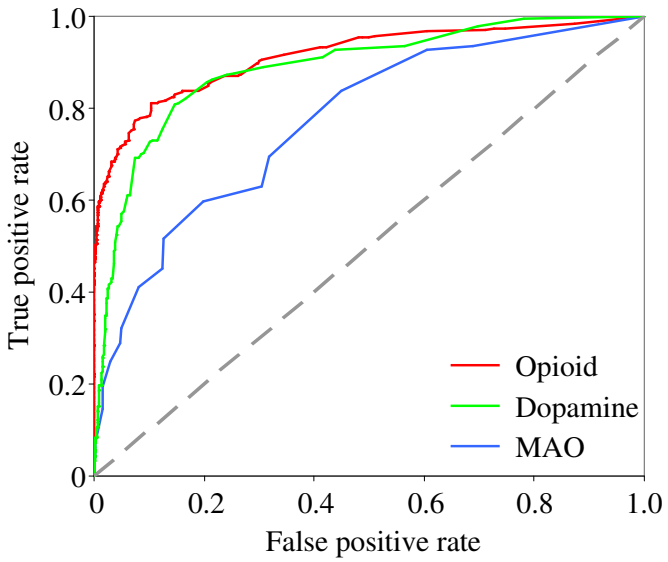


Figure 5.14: FragFCA classification ROC curves. Three representative ROC curves are shown for supersets Opioid, Dopamine, and MAO. The dashed grey line represents random selection of active compounds.

The area under the curve is a convenient parameter that summarizes the curve properties. It can be interpreted as the probability that a randomly chosen active is ranked higher than a randomly chosen inactive compound.[91] A ROCAUC of 1 indicates perfect recall of actives (i. e. no false positives), whereas a score of 0.5 corresponds to random selection. As reported in Table 5.3, up to 30% of the ZINC compounds did not match any signature combination. Compounds matching a single class were ranked according to their fragment score, which produced significant enrichment of active compounds for six of eight su-

persets, with ROCAUC values ¿ 0.8. These results showed that FragFCA signatures in combination with class-directed compound ranking were also effective in database searching for active compounds.

Superset	**ROCAUC**	**Single score**	**Unclassified**
Opioid	0.90	33.87	17.94
Serine Proteases	0.88	20.76	8.12
Dopamine	0.87	23.06	8.59
Adrenergic	0.84	36.01	17.80
Serotonin	0.84	22.53	10.28
PDE	0.82	21.96	6.55
MAO	0.77	58.27	28.54
Adenosine	0.59	48.21	31.54

Table 5.3: Similarity searching using FragFCA classifier. "ROCAUC" reports the Area Under the Receiver Operating Characteristic Curve. "Single score" reports the percentage of the screening database compounds that were only matched by fragments from a single activity class and "Unclassified" the percentage of screening database compounds that were not matched by any signature set.

5.4 Summary

In this chapter, FragFCA has been introduced for the analysis of complex SARs and extraction of signature fragment combinations for defined activity profiles. It has been shown that pairs and triplets of hierarchically generated fragments are capable of distinguishing ligands selective for one of several closely related targets. Using knowledge-based scale design reflecting target hierarchies, FragFCA allows the definition of flexible queries of increasing complexity and identifies discriminative fragment combinations.

Moreover, systematic analysis of the predictive ability of FragFCA signatures reveals that compounds selective for one of several closely related targets can be successfully distinguished using these signatures. Application to similarity searching has shown that FragFCA signatures can also be used to find active and selective compounds in a large database.

FragFCA thus links activity and chemical space by identifying structural feature combinations characteristic of defined activity profiles. FragFCA directly associates molecular fragments and their co-occurrence patterns with biological activity of ligands. The results presented in this chapter confirm that structural patterns captured by fragment combinations can serve as signatures for activity classes and distinguish between complex activity profiles.

Chapter 6

Molecular Formal Concept Analysis for Systematic Exploration of Activity Space

This chapter extends the FCA-based analysis of activity and potency profiles to compound selectivity. First, it is briefly discussed how compound selectivity can be defined and assessed using computational methods. Then, *Molecular Formal Concept Analysis* (MolFCA) is introduced for systematic mining of activity space. Different from FragFCA, MolFCA focuses on complete molecules and extracts diverse compound sets that share a defined selectivity profile. These compound sets can then be further analyzed to explore structure-selectivity relationships (SSRs).

6.1 Compound Selectivity

In chemogenomics, ligand affinity profiles are often generated using biological screens that test ligands across a range of different targets. The availability of such profiles makes it possible to study SSRs that explicitly take into account a ligand's affinity towards multiple targets.[10] Going beyond ligand activity analysis and prediction, recent studies have begun to assess ligand selectivity by computational means.[30,92–97]

In order to define ligand selectivity, potency values measured against multiple targets must be systematically compared and selectivity threshold ratios defined.[10,96] In particular, for closely related targets, ligand binding is often not an "all or nothing" event; rather, active small molecules bind with different potency against related targets and thereby potency differences essentially determine target selectivity. Therefore, potency ratios for each pair of targets are calculated, which provides a measure of compound selectivity.

Computational approaches have recently been developed that address

the issue of target selectivity on the basis of binary potency relationships.[94–97] For example, virtual selectivity searching has been introduced to identify compounds that are selective for one particular target over another.[96] However, computational methods for the exploration of more complex selectivity profiles involving more than two targets have thus far not been available.

6.2 Molecular Formal Concept Analysis

In this section, MolFCA is introduced for the mining of selectivity profiles in biologically annotated compound databases. MolFCA utilizes selectivity annotations as attributes that are derived from compound potency values against targets of interest. Objects in MolFCA are individual compounds. Importantly, MolFCA exclusively uses selectivity profiles as a molecular representation, but no structural descriptors or similarity measures. This makes it possible to identify structurally distinct compounds sharing a desired selectivity profile, without structural bias towards reference molecules.

6.2.1 Compound Selectivity Annotation

Compound selectivity was calculated on a target-pair basis. Given a compound with reported potencies against targets A and B, three selectivity categories were defined: (1) if compound potency for A was $\geq$50-fold higher than for B, the compound was considered selective for A over B (e. g. A: $1nM$, B: $70nM$); (2) if compound potency for B was $\geq$50-fold higher than for A, the compound was defined as selective for B over A (e. g. A: $70nM$, B: $1nM$); (3) if the compound potency ratio for target A and B was $\leq$10, the compound was considered non-selective (e. g. A: $1nM$, B: $7nM$). Different thresholds for selectivity ($\geq$50-fold) and non-selectivity ($\leq$10-fold) were applied in order to avoid boundary effects. Thus, compounds could be reliably classified as selective or non-selective despite possible fluctuations in potency measurements.

6.2.2 MolFCA Scale Design

In MolFCA, two types of scales are utilized. Selectivity scales distinguish between selective, "inverse selective", and non-selective compounds for a target pair. For capturing selectivity relationships, the type of scale depicted in Figure 6.1 was applied throughout the analysis. These scales assess compound selectivity independently of a specific compound potency range. Only compounds with potencies reported against both targets are considered for selectivity annotation. Compounds with no defined selectivity are assigned to the topmost node in MolFCA selectivity scales, which is not used in query definitions.

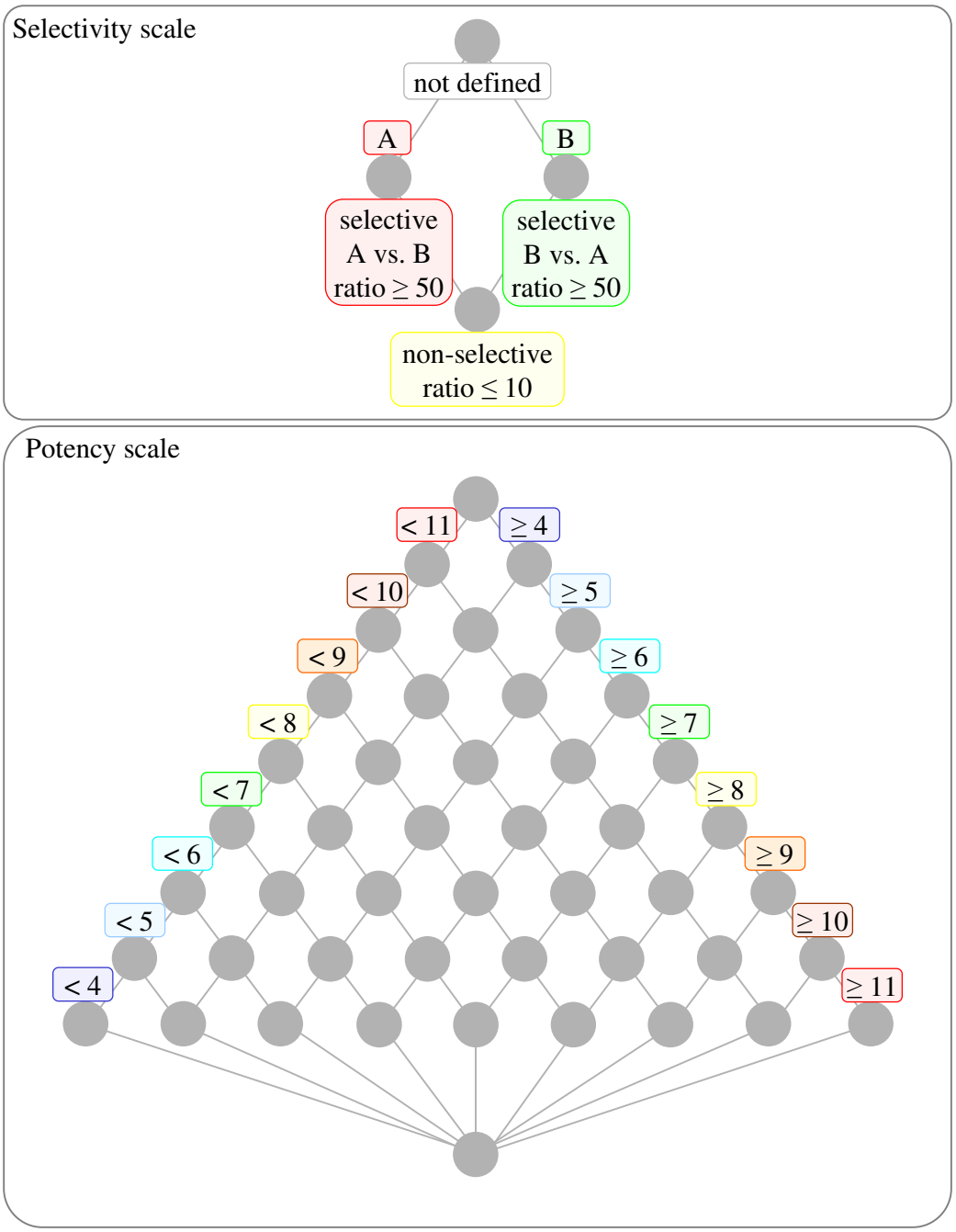


Figure 6.1: MolFCA scales. The top panel shows a prototypic selectivity scale that discriminates between two targets A and B. The lower panel depicts a prototypic potency scale. The numbers correspond to pK_i values.

Potency scales have been designed to report the distribution of compounds among nine potency bins ranging from $100\mu M$ to $10pM$. Each range contains compounds within one log unit (pK_i), e. g. compounds having a potency value falling into the sub-range $10nM$ - $100nM$. A prototypic potency scale is shown in Figure 6.1. Potency scales complement selectivity scales because selectivity is assessed based on potency ratios, rather than specific potency ranges. Hence, two compounds might share the same selectivity profile, but differ in potency. Thus, the design of the two scale types is optimized for selectivity queries that combine different scales.

6.2.3 MolFCA Queries

The central feature of MolFCA is the combination of different selectivity and potency scales for the definition of increasingly complex queries that yield compounds with specific selectivity and potency profiles. The selectivity scale design renders queries and their information content highly variable. For example, selectivity scales can be combined to identify non-selective compounds, as illustrated in Figure 6.2 on page 90. In this example, three scales are combined into a query that yields 22 histone deacetylase (HDAC) inhibitors that have comparable potency against three histone deacetylases (HDAC1, HDAC3, and HDAC6). The total number of compounds reported in the second scale corresponds to the number of compounds selected on the first one and the number of compounds reported in the third scale to the number of compounds selected on the second one. Thus, pre-selected compounds are sequentially transferred from scale to scale, similar to fragment combinations in FragFCA.

MolFCA has been implemented in MOE. The implementation enables the definition of scales, interactive survey of compound databases using these scales, generation of concept lattice representations, and storage of queries. This permits re-querying of updated compound databases without the need to re-assemble individual queries.

Selectivity Profile Mining

MolFCA has been applied to the BindingDB database in order to identify inhibitors with defined selectivity profiles. BindingDB is suitable for this analysis because it contains compound potency annotations (in form of K_i or IC_{50} values) for targets grouped into different target families. In order to obtain single potency values for each compound, multiple potency annotations for human targets, if available, were combined using the geometric mean of provided potency values.

MolFCA queries were generated in order to find inhibitors with defined selectivity profiles directed against one of the following four target families: histone deacetylases, phosphodiesterases, inosine monophosphate dehydrogenases,

and caspases. Table 6.1 reports individual targets for each family.

Target family	Scales	BindingDB targets
Histone deacetylase	3	Histone Deacetylase 1 (HDAC1)
		Histone Deacetylase 3 (HDAC3)
		Histone Deacetylase 5 (HDAC6)
Phosphodiesterase	6	Phosphodiesterase Type 4 (PDE4B)
		Phosphodiesterase Type 4 (PDE4D)
		Phosphodiesterase Type 10 (PDE10A)
		Phosphodiesterase Type 11 (PDE11A)
		Phosphodiesterase Type 1 (PDE1B)
		Phosphodiesterase Type 2 (PDE2A)
		Phosphodiesterase Type 3 (PDE3B)
Dehydrogenase	3	Inosine Monophosphate Dehydrogenase Type 1 (IMPDH1)
		Inosine Monophosphate Dehydrogenase Type 2 (IMPDH2)
Caspase	3	Caspase-3
		Caspase-7
		Caspase-8

Table 6.1: Target families and MolFCA queries. Reported are the four target families studied using five MolFCA queries. "Scales" reports the number of individual scales used to design each query. Names of individual targets reported in BindingDB are given under "BindingDB targets".

Utilizing reference compounds taken from BindingDB or literature sources, MolFCA has been applied to search for BindingDB compounds with corresponding selectivity profiles. Furthermore, specific MolFCA queries were relaxed by omitting or softening individual selectivity constraints in order to identify additional compounds with defined deviations from the reference profile. The reference compounds included: suberoylanilide hydroxamic acid (SAHA, marketed as Vorinostat, brand name Zolinza), a histone deacetylase inhibitor approved for the treatment of cutaneous T-cell lymphoma;[98] SB 207499 (Cilomilast, brand name Ariflo), a selective phosphodiesterase type 4 (PDE4) inhibitor used for the treatment of asthma and chronic obstructive pulmonary disease;[99] mycophenolic acid (MPA, brand name Myfortic), a reversible, non-competitive inosine monophosphate dehydrogenase (IMPDH) inhibitor used as an immunosuppressant to prevent transplant rejection;[100] and IDN 6556, a pan-caspase inhibitor that is used as an apoptosis (i. e. programmed cell death) inhibitor to prevent liver tissue damage during liver transplantation.[101]

The application of MolFCA to search for inhibitors of the four different target families in the BindingDB using reference selectivity profiles of individual reference compounds is described in the following. Table B.7 in Appendix B summarizes all queries.

Selectivity Profile Mining - Vorinostat

The selectivity profile of Vorinostat was derived from BindingDB potency data against histone deacetylases HDAC1 ($93nM$), HDAC3 ($52nM$), and HDAC6 ($43nM$). Thus, according to the MolFCA selectivity classification, this compound was non-selective for each target pair, i. e. HDAC1 vs. HDAC3, HDAC1 vs. HDAC6, and HDAC3 vs. HDAC6. In order to mine the database for compounds matching this selectivity profile, three selectivity scales have been combined (i. e. one scale for each pair). Figure 6.2 illustrates this query.

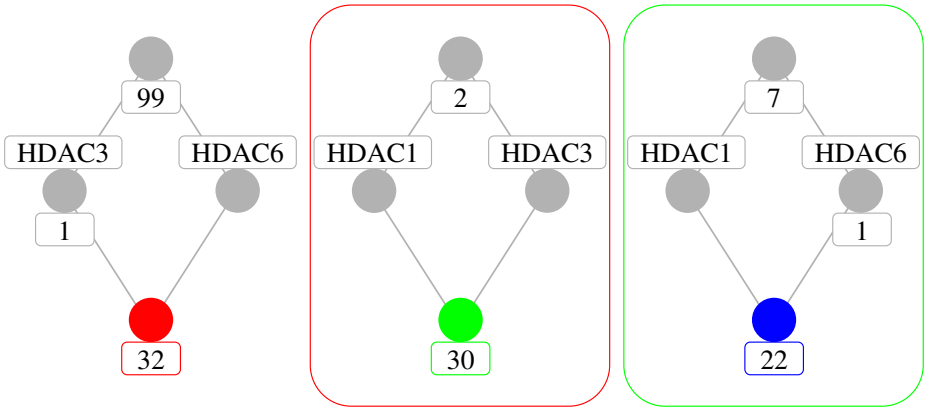


Figure 6.2: Vorinostat selectivity profile query. Three selectivity scales are combined in order to select compounds (colored nodes) active with comparable potency against HDAC1, HDAC3 and HDAC6. Seven compounds with undefined selectivity are removed by the last scale and also one compound selective for HDAC6 over HDAC1.

Only one possible arrangement of scales is shown for clarity. However, in query design, scales can be combined in an arbitrary order. On the first scale, 32 compounds non-selective for targets HDAC3 and HDAC6 were extracted. These compounds were then projected onto the second scale. For two compounds, the selectivity profile for these two targets was not defined. The remaining 30 compounds were transferred to the third scale and the final set of 22 compounds (including Vorinostat) was selected from the bottom node. Thus, 21 additional compounds matching the selectivity profile of Vorinostat were identified. Nine of these compounds belonged to the triazole structural class and eight other compounds were thiophene derivatives. The remaining four compounds were structurally distinct and included LAQ-824, CRA-024781, PXD101, and a 4-phenylimidazole derivative.

Figure 6.3 shows Vorinostat, a representative triazole ligand, a thiophene derivative, and the four additional inhibitors. Strikingly, the thiophene derivatives and the 4-phenylimidazole derivative lack the hydroxamic acid group that is a hallmark of Vorinostat and other HDAC inhibitors. Nevertheless, the

most potent thiophene derivative is comparable to Vorinostat (i. e. HDAC1: $210nM$, HDAC3: $120nM$, and HDAC6: $240nM$). This example illustrates that MolFCA is capable of identifying structurally distinct compounds with corresponding selectivity profiles.

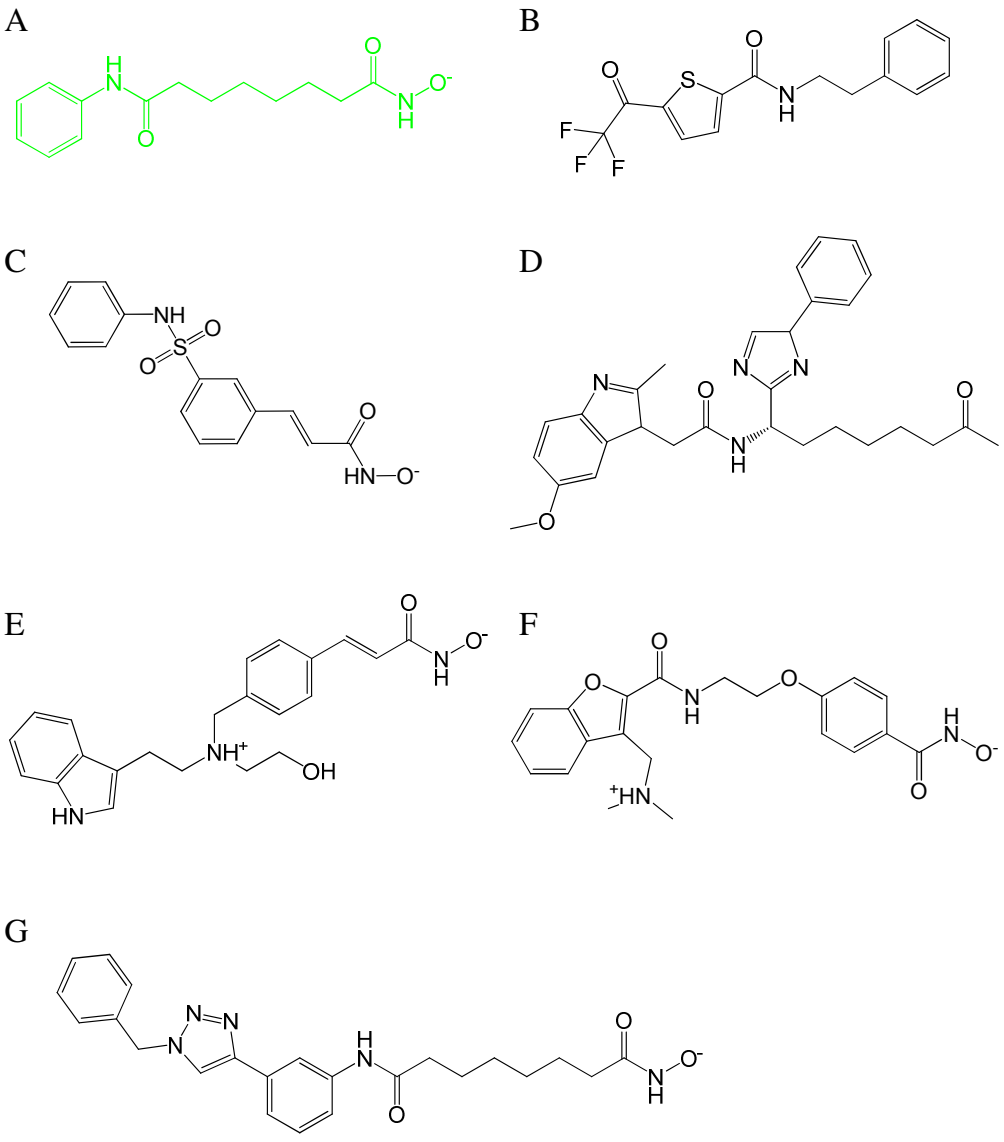


Figure 6.3: Compounds matching the Vorinostat profile. The reference compound Vorinostat (**A**, green) is shown together with two representative thiophene derivatives and two triazole ligands. Four additional compounds not belonging to these structural classes are also shown. **A**: Vorinostat; **B**: thiophene derivative, 3b; **C**: PXD101; **D**: 4-phenylimidazole, 17; **E**: LAQ-824; **F**: CRA-024781; **G**: Triazole Ligand, 10a.

In order to find other compounds that only partly matched the Vorinostat selectivity profile, the HDAC query was relaxed by systematically modifying the selectivity category on each scale or by removing an individual scale, leading to the identification of three additional compounds. First, Apicidin, a cyclic peptide antibiotic acting through protozoal HDAC inhibition and also inhibiting tumor proliferation[102] was found to be non-selective for HDAC1 over HDAC3, but selective for these two targets over HDAC6. Second, an additional triazole ligand was identified as selective for HDAC6 over HDAC1, but non-selective for the target pairs HDAC1/HDAC3 and HDAC6/HDAC3 (HDAC1: $97.8nM$, HDAC3: $13.7nM$, HDAC6: $1.9nM$). Third, another thiophene derivative was also found to be selective for HDAC6 over HDAC1 (HDAC1: $9.7\mu M$, HDAC6: $15nM$) and identified by omitting HDAC3 from the query (potency data for HDAC3 was not available for this compound). Figure 6.4 shows these three compounds and structurally similar molecules that were identified to match the Vorinostat profile. It is evident that only subtle structural deviations between these compounds altered their selectivity profiles. Thus, compounds identified with original and relaxed MolFCA queries were structurally similar but had different selectivity profiles. Because MolFCA does not utilize structural representations as input, any structurally similar or diverse subset of molecules can be identified.

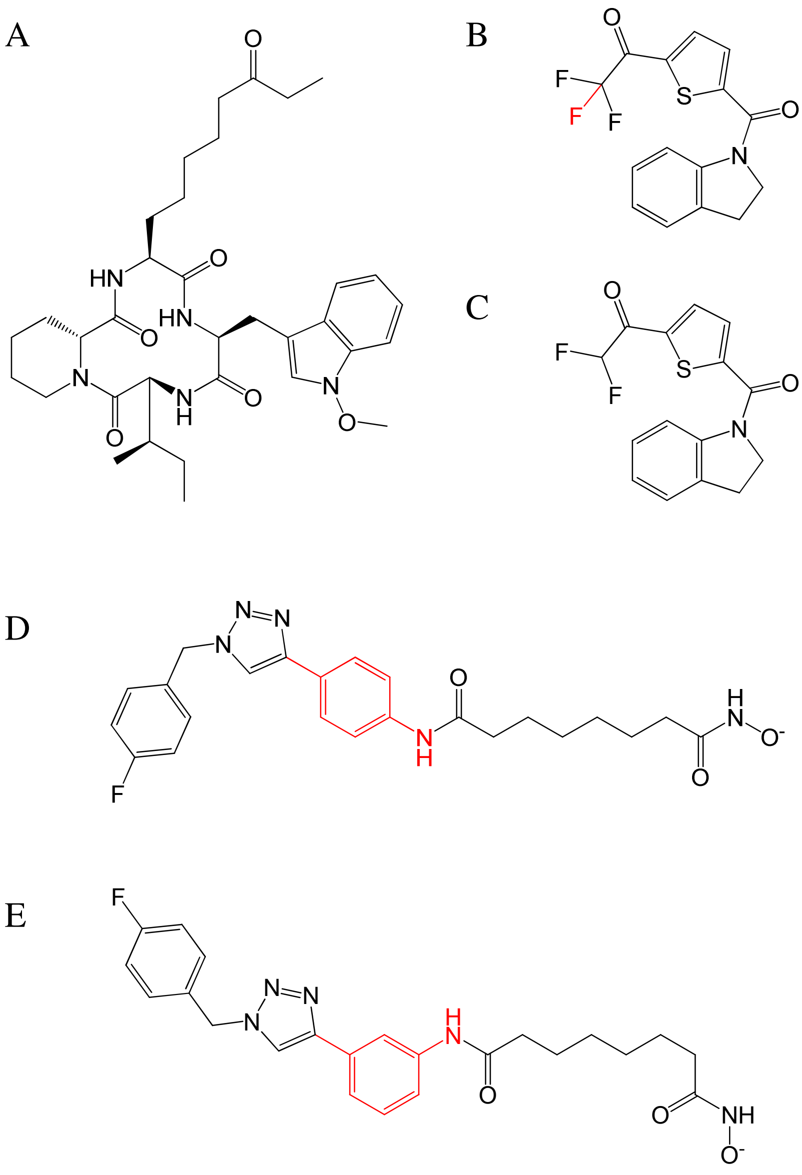


Figure 6.4: Compounds deviating from the Vorinostat profile. Three compounds are shown that deviate from the Vorinostat profile. For the thiophene derivative and the triazole ligand, structurally highly similar compounds that match the Vorinostat profile are also displayed. Structural differences between each pair of similar compounds with distinct selectivity profiles are highlighted in red. **A**: Apicidin (HDAC3 / HDAC6; HDAC1 / HDAC6); **B**: thiophene derivative, 3h (Vorinostat profile); **C**: thiophene derivative, 15h (HDAC6 / HDAC1); **D**: Triazole Ligand, 6b (HDAC6 / HDAC1); **E**: Triazole Ligand, 10b (Vorinostat profile).

Cilomilast

As a second profile mining application, Cilomilast has been chosen as a template, a potent and selective PDE4 inhibitor used for the treatment of respiratory disorders.[99] Based on BindingDB potency data provided for Cilomilast (PDE4B: 25nM, PDE4D: 11nM, PDE1B: 87μM, PDE2A: 160μM, PDE3B: 87μM, PDE10A: 73μM, and PDE11A: 21μM) a selectivity query combining six selectivity scales was assembled. Figure 6.5 depicts the query. Compounds non-selective for PDE4B and PDE4D were selected on the first scale. The subsequent scales assessed the selectivity of these compounds for PDE4B over the other targets. This query yielded one additional compound, Piclamilast, which shared the Cilomilast selectivity profile but was overall more potent (i. e. PDE4B: 41pM, PDE4D: 21pM, PDE1B: 68μM, PDE2A: 54μM, PDE3B: 11μM, PDE10A: 21μM, PDE11A: 1.6μM) and is shown in Figure 6.6.

In order to identify additional compounds that only partly matched the Cilomilast profile, individual scales for PDE4B selectivity over the other related targets were assessed. Three additional compounds were found that that were non-selective for PDE4B and PDE4D, but selective for PDE4B over PDE11A including Zardaverine (PDE4B: 930nM, PDE4D: 390nM, PDE11A: 140μM), Filaminast (PDE4B: 960nM, PDE4D: 1μM, PDE11A: 57μM), and Roflumilast (PDE4B: 840pM, PDE4D: 680pM, PDE11A: 25μM). All of these compounds are currently also evaluated or used as bronchodilatory agents to treat respiratory disorders.[103–105] Furthermore, two other compounds were found that were non-selective for PDE4B and PDE4D, but selective for PDE4B over PDE10A including Mesopram (PDE4B: 420nM, PDE4D: 1.1μM, PDE10A: 63μM) and Rolipram (PDE4B: 915nM, PDE4D: 1.1μM, PDE 10A: 140μM). Different from the inhibitors discussed above, Mesopram is evaluated for the treatment of multiple sclerosis[106] and Rolipram for the treatment of depression,[107] but also has immunosuppressive properties.[108] Figure 6.6 shows these five compounds. Thus, through relaxation of a complex PDE4 selectivity query, compounds with related yet distinct selectivity profiles having in part different therapeutic indications were identified.

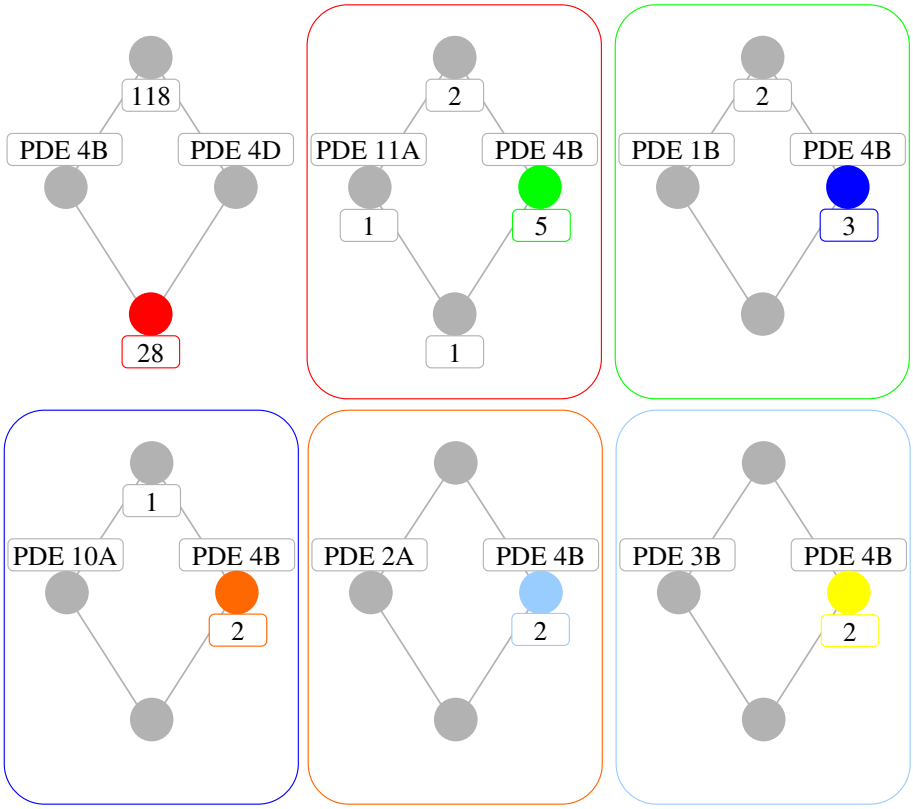


Figure 6.5: Cilomilast query. Six selectivity scales used to define the Cilomilast query are shown. Colored nodes represent subsets selected on each scale. Each subset is projected onto the next scale, indicated by colored boxes.

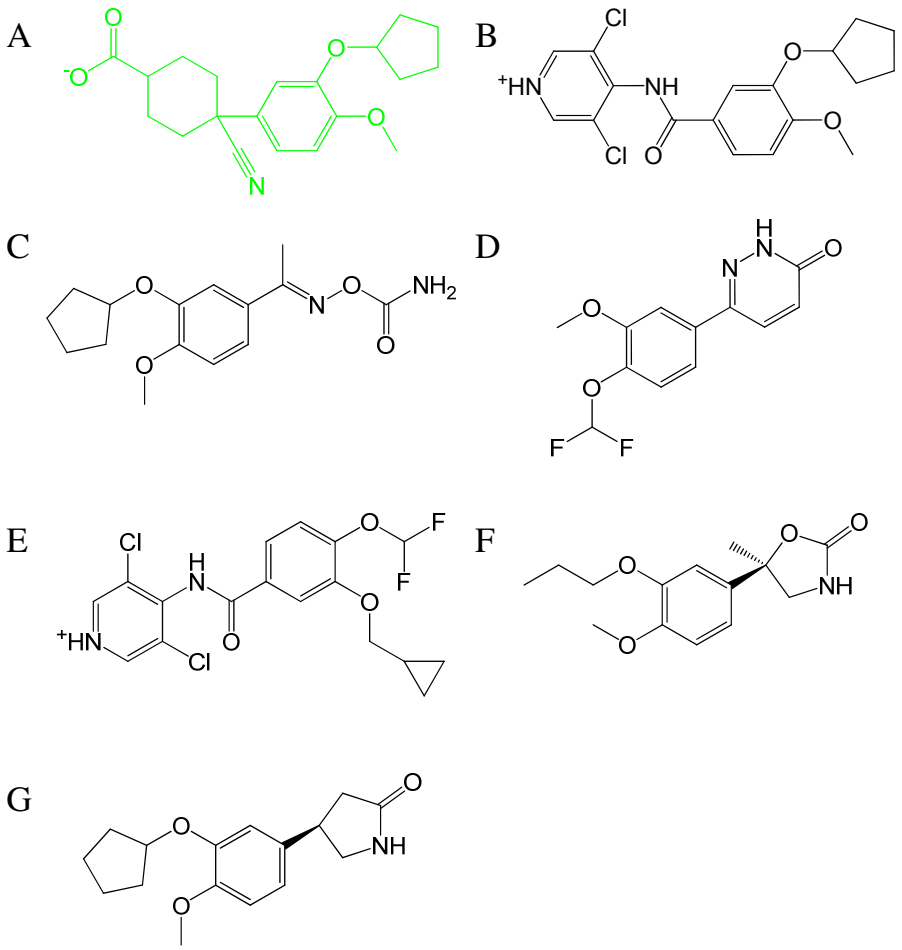


Figure 6.6: Compounds (partly) matching the Cilomilast profile. The reference compound Cilomilast (**A**, green) is shown together with Piclamilast that was identified by MolFCA. Compounds **C-E** partly match the Cilomilast query and are also used to treat respiratory disorders. Compounds **E** and **F** partly match the query and are currently evaluated for the treatment of neurodegenerative disorders. **A**: Cilomilast; **B**: Piclamilast; **C**: Filaminast; **D**: Zardaverine; **E**: Roflumilast; **F**: (R,S)-Mesopram; **G**: Rolipram.

Mining for Highly Potent Inhibitors - Mycophenolic Acid

Mycophenolic acid (MPA) inhibits IMPDH1 and IMPDH2 in a non-selective manner with high potency values of $33nM$ and $11nM$, respectively. In order to identify other potent (i. e. $\leq 100nM$) IMPDH inhibitors, a MolFCA query consisting of one selectivity and two potency scales has been designed. A selectivity scale was applied to select a total of 13 inhibitors that were non-selective against IMPDH1 and IMPDH2. These compounds were then transferred to the IMPDH1 potency scale. Eleven of these 13 compounds (including MPA) were found to fall into the desired IMPDH1 potency range. These eleven inhibitors were then projected onto the IMPDH2 potency scale. Six of these compounds fell into the desired IMPDH2 potency range.

The 12 additional inhibitors belong to two structural classes, nine tiazofurin adenine dinucleotide (TAD) analogues and three mycophenolic adenine dinucleotide (MAD) analogues, shown in Figure 6.7. In addition to MPA, the six highly potent inhibitors included four TAD analogues and one MAD analogue.

De novo Selectivity Profiles - IDN 6556

The caspase inhibitor IDN 6556 was taken from the literature,[101] but was not available in BindingDB. This compound prevents apoptosis in liver transplants and has been indicated to act as a pan-caspase inhibitor by inhibiting both initiator and effector caspases.[101] In order to find potential pan-caspase inhibitors in BindingDB, a selectivity query for pan-caspase inhibitors shown in Figure 6.8 was built. This query was designed to identify caspase inhibitors with comparable potency against caspase 3, 7 and 8. Caspase 8 is an initiator caspase, i. e. it activates downstream caspases by cleavage, whereas the other two caspases are effector caspases, which induce apoptosis by chromatin fragmentation.[109]

Because MolFCA scales utilize two selectivity thresholds, it was possible to design the query, although no exact potency information was provided for the compound. The query identified nine compounds representing five chemotypes, all of which were distinct from IDN 6556 (Figure 6.9). Seven of these nine compounds had potency values in the range of $10nM$ to $500nM$. The two remaining compounds represented a weakly active chemotype with potency values of $40\mu M$ and $4\mu M$. This example illustrates that effective MolFCA queries can also be defined in the absence of suitable reference compounds exclusively on the basis of selectivity criteria.

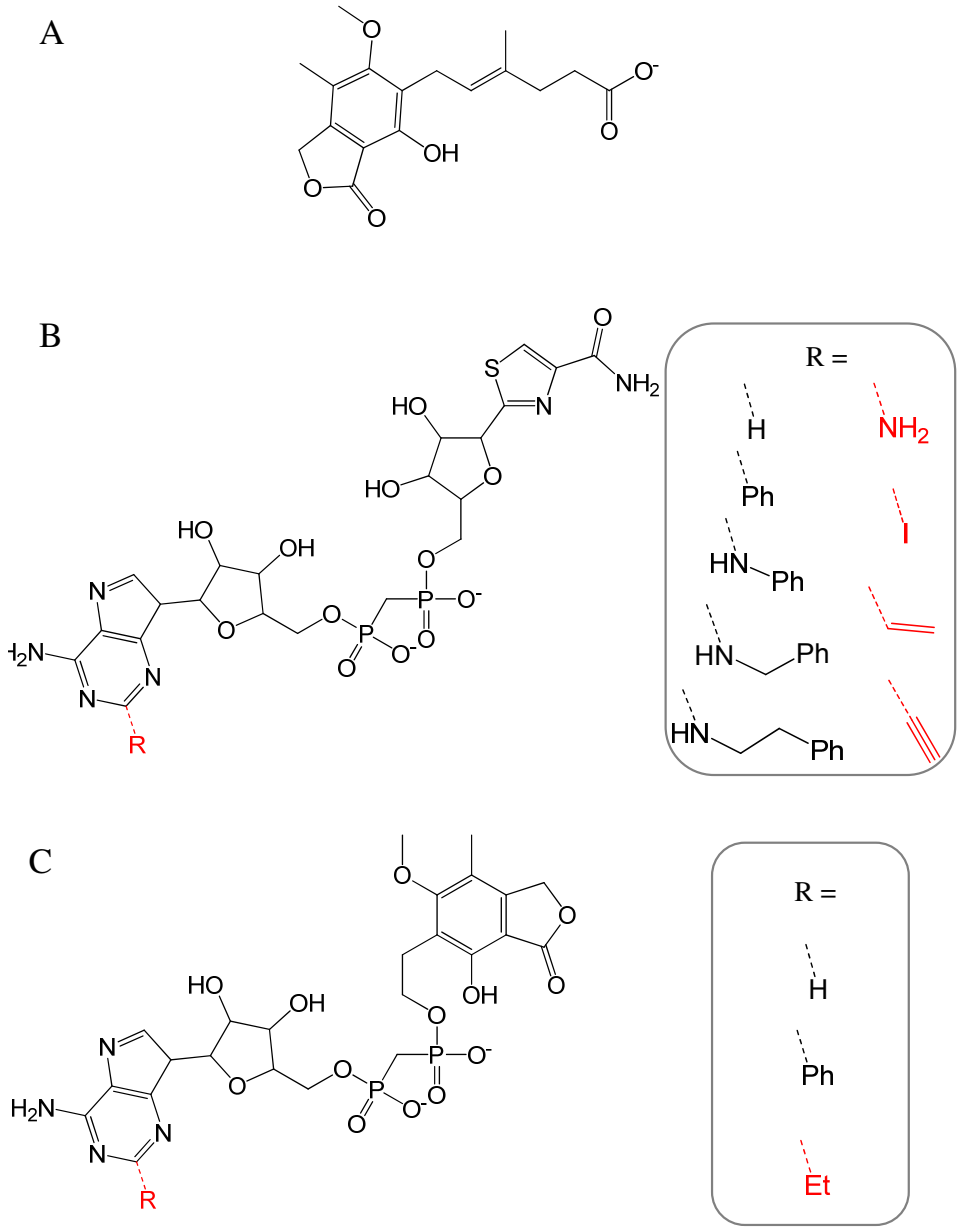


Figure 6.7: Compounds matching the MPA profile. The reference compound MPA is shown together with the core structures of the 12 additional non-selective IMPDH inhibitors. The R-group of each inhibitor is shown in the boxes on the right and R-groups of the five highly potent compounds selected using potency scales are colored red. **A**: MPA; **B**: Tiazofurin Adenine Dinucleotide (TAD) Analogues; **C**: Mycophenolic Adenine Dinucleotide (MAD) Analogues.

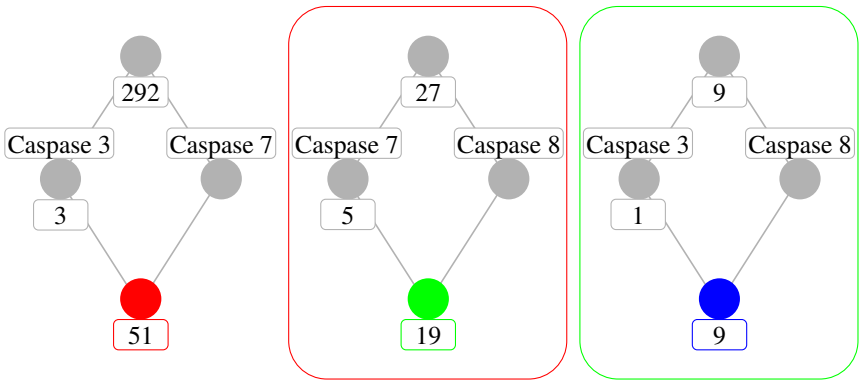


Figure 6.8: De novo MolFCA query design. Three selectivity scales are combined in order to select nine compounds with comparable activity against caspases 3, 7, and 8. Colored nodes represent compound subsets, which are projected onto subsequent scales (colored boxes).

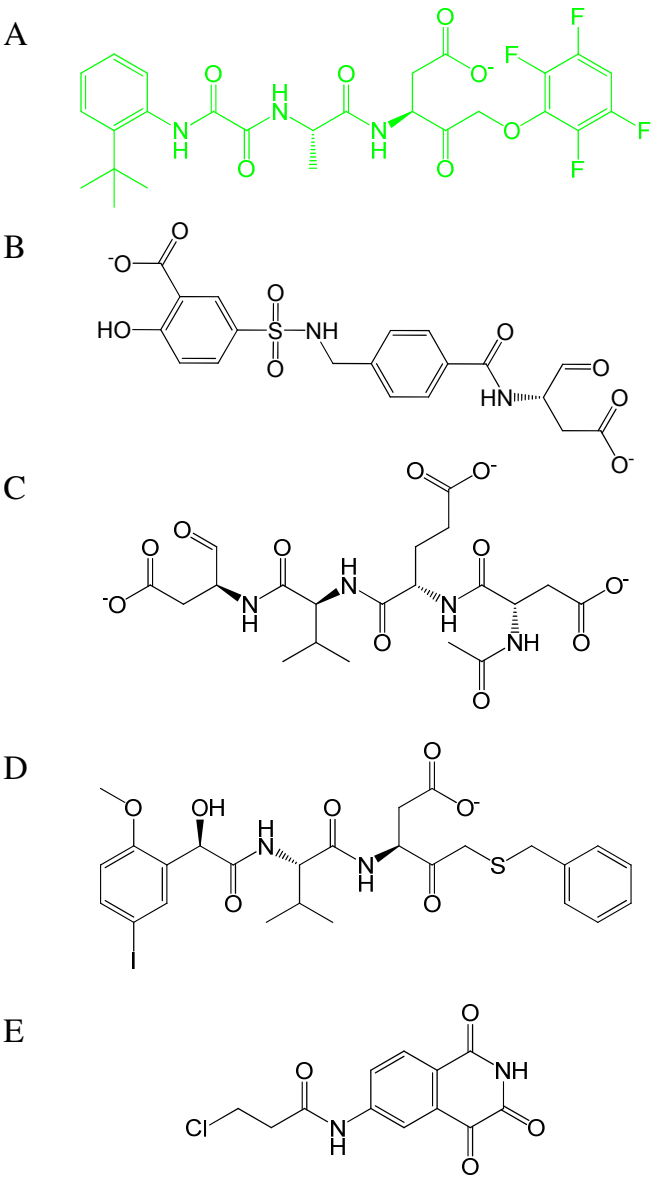


Figure 6.9: Compounds matching the de novo MolFCA query. IDN 6556 (**A**, green), a literature compound with indicated pan-caspase inhibitory activity is shown together with four structurally diverse compounds with high to medium potency that match a MolFCA query for pan-caspase inhibitors. **A**: IDN 6556; **B**: Ac-DEVD_CHO; **C**: Inhibitor 3; **D**: valine aspartyl ketone 35; **E**: Isoquinoline-1,3,4-trione 9k.

6.3 Summary

In this chapter, MolFCA has been introduced for the systematic mining of complex selectivity and potency profiles. Going beyond the analysis of fragment distribution among active compounds, MolFCA focuses on the analysis of target selectivity on a whole-molecule basis. MolFCA queries were designed to identify compounds that shared defined selectivity profiles with reference molecules. The selectivity queries consisted of up to six MolFCA scales and involved up to seven targets.

MolFCA identified structurally diverse compounds matching each selectivity profile. The indentified compounds represented in part very different structure-selectivity relationships. The findings demonstrate that MolFCA is capable of detecting sets of compounds that are active against target families with different selectivity. Diverse compound sets identified by MolFCA can be used to analyze SSRs and relate structural features to molecular selectivity.

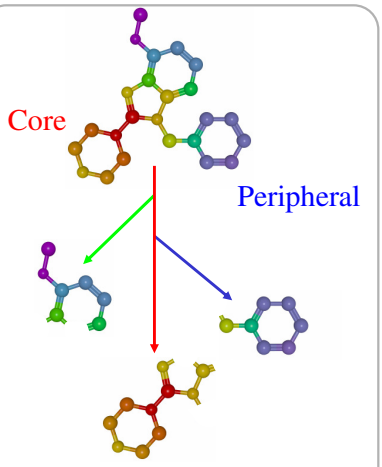
This kind of SSR analysis has not been possible so far, because no generally applicable framework for the definition of selectivity profiles existed. MolFCA provides a basis for the assembly of intuitive and flexible compound queries. A query can be easily extended to an arbitrary number of targets by incorporating the selectivity and potency scales introduced here. Thus, MolFCA represents a flexible, but well-defined framework for the mining of complex selectivity profiles.

Chapter 7

Summary and Conclusions

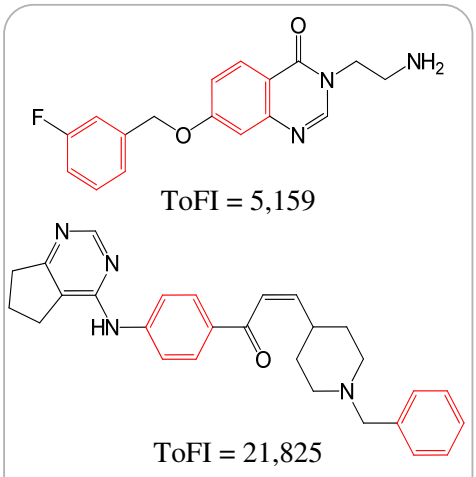
In this thesis, methods have been presented that provide a link between chemical and biological activity space.

Approaches have been introduced to abstract from the chemical structure of individual ligands through the use of fragment-type structural descriptors in an activity class-directed manner. Furthermore, the annotation of fragments with biological activity information was studied. First, an activity class-directed hierarchical fragmentation approach has been presented. The fragments were organized in Core Trees that represent active compounds in form of fragmentation pathways. Core paths correspond to structurally conserved regions within individual activity classes and can be aligned using sequence alignment methods. Multiple core path alignments yield Consensus Fragment Sequences that serve as an activity class signature.

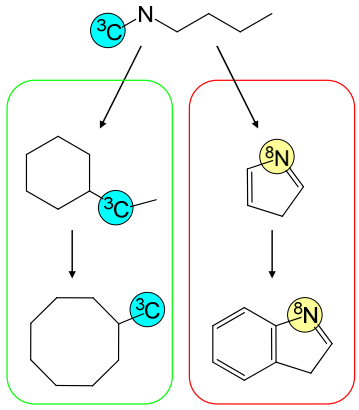


Core Trees distinguish between core and peripheral regions of active compounds.

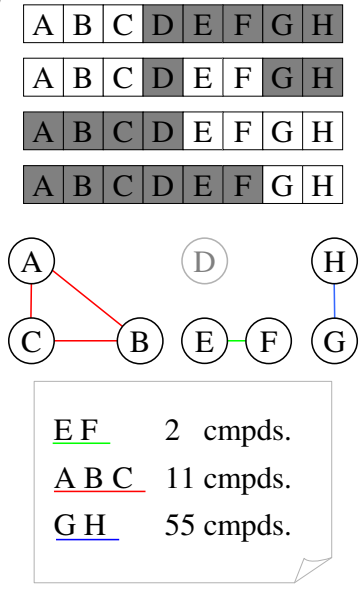
Then, the Topological Fragment Index (ToFI) has been introduced for the assessment of the topological environment of substructures in active compounds. ToFI values distinguish between different topological environments and thus extend fragment counts. RECAP fragments were organized in hierarchies based on ToFI value distributions in compounds with biological activity against multiple, closely related targets. Fragment relationships in these hierarchies allowed the identification of Activity Class Characteristic RECAP Fragments and revealed activity class-specific fragment topology clusters.



ToFI values distinguish between different topological environments of a fragment in two molecules.



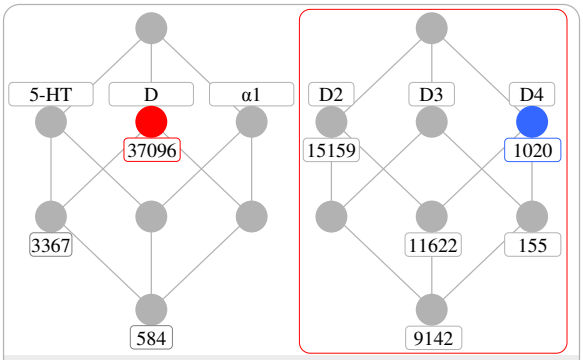
Dependency graphs reflect ToFI distributions in active compounds.



FCoN feature cliques are utilized in virtual screening.

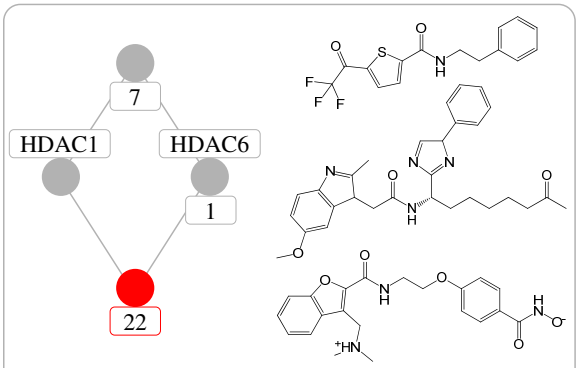
Furthermore, co-occurrence of features extracted from three different molecule fingerprint representations was analyzed. Therefore, Feature Co-occurrence Networks have been introduced that enable the identification of feature cliques characteristic of individual activity classes. A similarity search protocol has been established that utilizes these cliques in order to rank database compounds. It has been found that generic structural features can form combinations that are activity class-specific.

Fragment Formal Concept Analysis has been introduced in order to analyze the distribution of defined fragment combinations in compounds active against closely related targets. Again, fragment combinations, rather than individual fragments best distinguished between compounds with closely related biological activity profiles. Furthermore, it has been shown that signature fragment combinations were predictive in compound classification and virtual screening for active and selective compounds.



FragFCA scales are combined in order to extract signature fragment combinations.

Molecular Formal Concept Analysis (MolFCA) has been introduced for mining of complex selectivity relationships in biologically annotated databases. MolFCA identified structurally diverse compounds that matched selectivity profiles of drug-like reference molecules. Structure-selectivity relationships of these compound sets can be further analyzed. Thus, MolFCA allows the flexible design and exploration of complex selectivity queries for multiple targets.



MolFCA queries yield structurally diverse compounds having a defined selectivity profile.

In summary, fragment-type descriptors of molecular structure are capable of integrating activity and chemical space. Activity class-relevant information is predominantly encoded in fragment combinations, rather than individual fragments. Two adaptations of formal concept analysis have been introduced that allow compound data mining at the level of fragments or molecules in order to elucidate structure-activity and structure-selectivity relationships.

Appendix A

Software and Databases

Software and databases used in this thesis are listed in alphabetical order.

ACCS-FP	Life Science Informatics, University of Bonn (Germany)
Description:	Activity Class Characteristic Substructures Fingerprints are activity class-directed, small key-typed fingeprints.
Reference:	Batista *et al.*[48]

BindingDB	
Description:	BindingDB is a public, web-accessible database of measured binding affinities, mainly focusing on the interactions of proteins considered to be drug targets with small, drug-like molecules.
Reference:	BindingDB[110–113]
WebSite:	*http://www.bindingdb.org/bind/index.jsp*

MACCS	Symyx Software: San Ramon, CA (USA)
Description:	MACCS structural keys represent a 2D fingerprint that consists of 166 structural features.
Reference:	MACCS[63]
WebSite:	*http://www.symyx.com*

MDDR	Symyx Software: San Ramon, CA (USA)
Description:	MDL Drug Data Report (MDDR) is a database containing approx. 160,000 biologically active compounds with target and/or therapeutic annotations.
WebSite:	*http://www.symyx.com/*

MOE	Chemical Computing Group Inc.: Montreal, QC (Canada)
Description:	The Molecular Operating Environment (MOE) provides applications for the calculation of property descriptors and several fingerprint formats including MACCS.
WebSite:	*http://www.chemcomp.com*

Perl	Larry Wall
Description:	Perl is a freely available programming language.
WebSite:	*http://www.activestate.com/activeperl/*

Pipeline Pilot	Accelrys Software Inc.: San Diego, CA (USA)
Description:	SciTegic Pipeline Pilot allows the creation of workflows for chemoinformatics analyses and the calculation of ECFP fingerprints.
WebSite:	*http://www.chemcomp.com*

ToscanaJ	DSTC, the University of Queensland, and the Technical University of Darmstadt
Description:	ToscanaJ is a freely available viewer / browser for concept lattices.
Reference:	ToscanaJ[90]
WebSite:	*http://toscanaj.sourceforge.net/index.html*

Tulip 3.0.1	InfoViz, David Auber
Description:	Tulip is a freely available graph visualization software package.
Reference:	Tulip[84]
WebSite:	*http://www.labri.fr/perso/auber/projects/tulip/news.php*

ZINC	UCSF University of California: San Francisco, CA (USA)
Description:	ZINC (ZINC Is Not Commercial) is a public-domain database of commercially available compounds in predicted 3D conformational states.
Reference:	Irwin *et al.*[114]
WebSite:	*http://blaster.docking.org/zinc*

Appendix B

Additional Data

B.1 Feature Co-occurrence Networks

Figure B.1 shows box plots of clique numbers and Figure B.2 clique size distributions.
Table B.1 reports median clique numbers and Table B.2 clique size distribution for different co-occurrence thresholds.
Table B.3 reports median values of database compounds matching FCoN feature cliques for different co-occurrence thresholds.
Table B.4 provides recovery rates of active test compounds for different co-occurrence thresholds and pooled sets.

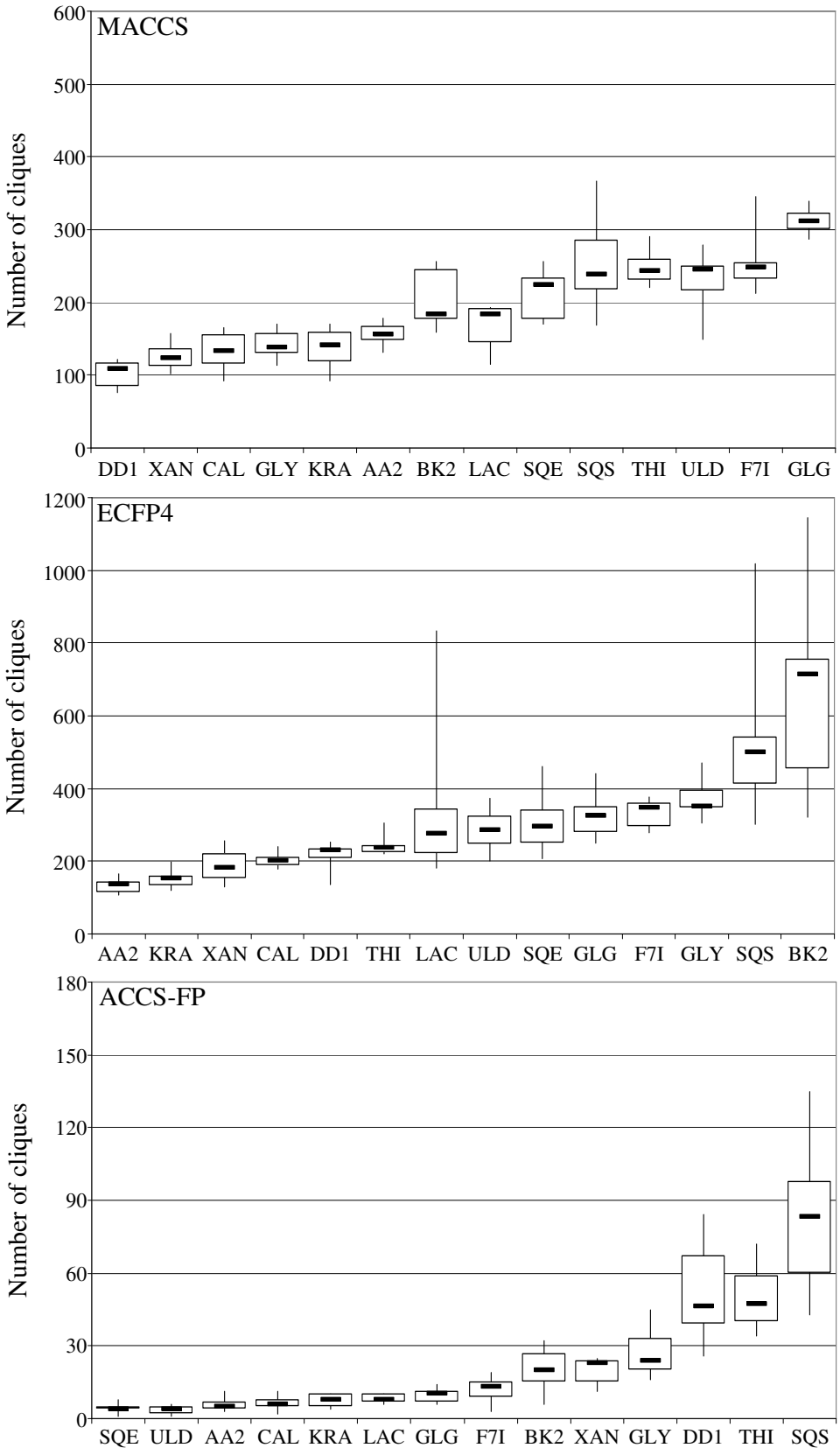


Figure B.1: Clique number distribution. Box plots report the distribution of clique numbers for each activity class calculated from ten independent trials. Thick bars mark median values.

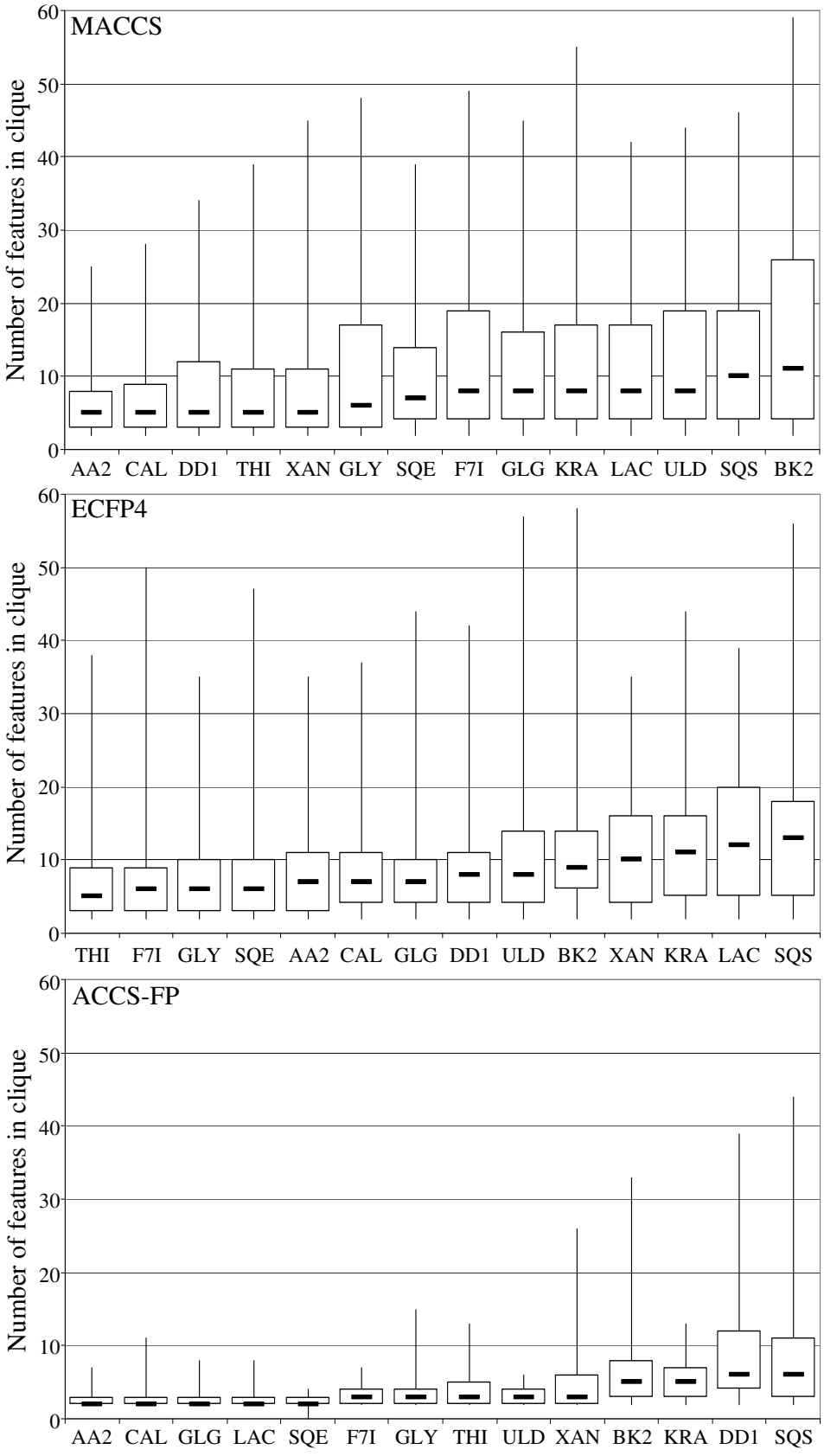


Figure B.2: Clique size distribution. Box plots report the distribution of clique sizes for each activity class. Thick bars mark median values.

Activity Class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1.0$	**Pooled**
			MACCS				
AA2	62	46	36	29	23	20	156
BK2	70	52	45	33	19	18	184
CAL	52	44	33	25	19	20	133
DD1	40	29	25	23	18	16	109
F7I	103	66	50	40	25	22	248
GLG	129	91	58	38	24	19	312
GLY	49	36	30	22	17	13	138
KRA	52	36	28	25	18	19	141
LAC	66	46	32	30	22	18	184
SQE	89	55	44	34	22	20	224
SQS	64	58	58	46	26	12	239
THI	120	75	43	27	15	12	243
ULD	93	65	46	34	20	19	245
XAN	38	33	35	24	21	18	123
			ECFP4				
AA2	57	46	40	36	35	34	136
BK2	182	199	158	88	62	56	715
CAL	58	64	63	55	55	47	201
DD1	81	75	59	46	41	40	229
F7I	178	95	61	49	41	40	346
GLG	144	108	70	65	55	50	325
GLY	158	112	80	75	63	58	350
KRA	57	54	44	36	29	29	154
LAC	105	81	60	46	38	36	275
SQE	143	88	72	77	59	59	296
SQS	132	165	132	88	54	52	499
THI	94	83	66	55	45	43	238
ULD	136	89	60	50	38	38	287
XAN	59	64	60	47	43	41	182
			ACCS				
AA2	9	8	6	7	5	5	21
BK2	7	8	5	6	7	7	20
CAL	6	6	5	6	6	6	15
DD1	12	15	12	10	9	10	46
F7I	7	5	3	3	3	3	13
GLG	14	12	10	10	10	10	28
GLY	14	9	7	7	6	6	24
KRA	3	4	3	4	4	4	8
LAC	11	11	10	10	8	8	32
SQE	3	1	2	2	1	1	4
SQS	15	19	20	20	17	17	83
THI	20	20	13	11	9	9	47
ULD	5	4	4	4	4	4	13
XAN	8	8	7	8	9	9	23

Table B.1: FCoN clique numbers. The median number of cliques calculated from ten independent trials is reported at different co-occurrence thresholds and for pooled sets.

Activity Class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1.0$	**Pooled**
			MACCS				
AA2	7	4	3	3	2	2	5
BK2	20	10	8	4	2	2	11
CAL	8	5	3	3	3	3	5
DD1	7	5	3	3	3	3	5
F7I	15	8	6	3	2	2	8
GLG	13	9	5	3	2	2	8
GLY	7	8	5	3	3	2	6
KRA	12	7	6	4	3	2	8
LAC	14	9	6	4	3	3	8
SQE	11	6	4	3	3	2	7
SQS	20	11	8	5	3	3	10
THI	9	5	3	2	2	2	5
ULD	15	8	5	3	3	3	8
XAN	6	5	4	3	3	2	5
			ECFP4				
AA2	9	7	6	6	6	6	7
BK2	14	9	7	5	3	3	9
CAL	13	8	7	4	4	4	7
DD1	9	8	7	5	6	6	8
F7I	7	5	3	3	3	3	6
GLG	9	6	5	4	4	5	7
GLY	9	5	5	5	4	3	6
KRA	14	10	9	9	9	9	11
LAC	15	16	6	5	4	4	12
SQE	9	5	4	4	3	3	6
SQS	12	7	12	17	5	6	13
THI	7	5	4	4	4	5	5
ULD	11	7	5	4	6	6	8
XAN	13	10	7	6	5	5	10
			ACCS				
AA2	3	3	2	2	2	2	3
BK2	4	4	4	4	3	3	5
CAL	4	4	2	2	2	2	4
DD1	9	7	5	4	3	3	6
F7I	3	3	2	2	2	2	3
GLG	5	3	2	2	2	2	3
GLY	3	3	3	2	2	2	3
KRA	5	4	3	3	3	3	5
LAC	6	4	3	3	2	2	4
SQE	2	3	2	2	2	2	2
SQS	8	9	7	5	2	2	6
THI	5	3	3	2	2	2	3
ULD	5	5	4	3	3	3	4
XAN	4	3	2	2	2	2	3

Table B.2: FCoN clique size. The median number of features per clique is reported at different co-occurrence thresholds and for pooled sets.

Activity Class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1.0$	**Pooled**
			MACCS				
AA2	7730	20885	53094	65435	91088	88006	23193
BK2	158	3541	6425	38528	75803	102747	3273
CAL	7078	24961	54214	59199	54214	64120	22725
DD1	7532	13740	37286	55423	69663	88004	17246
F7I	348	6333	18515	48227	74604	79421	4886
GLG	1136	7842	40932	85941	144425	113971	9131
GLY	3255	11182	31568	50112	75455	92245	16744
KRA	192	2207	7201	19914	43101	44758	2272
LAC	212	1449	5212	26080	67093	71593	2713
SQE	3803	17649	52751	81292	102744	102745	16490
SQS	129	1311	3256	13539	88005	102750	3285
THI	3102	26918	64501	92377	92552	92376	17646
ULD	914	9454	32052	64088	91528	101083	10153
XAN	8175	15494	22415	43582	57061	92247	21485
			ECFP4				
AA2	60	108	103	61	58	13	62
BK2	7	19	58	1354	2336	414	20
CAL	8	38	959	1741	1632	960	636
DD1	3	17	32	46	7	6	12
F7I	89	1354	6971	9026	8065	3751	775
GLG	156	1959	3610	3962	3610	2655	1410
GLY	9	276	573	279	818	384	100
KRA	5	7	77	27	24	24	7
LAC	9	2	223	225	51	25	16
SQE	3	10	32	11	10	10	7
SQS	5	47	5	2	7	6	4
THI	305	3964	8743	11894	2013	759	3173
ULD	127	421	4649	5054	2671	2671	747
XAN	5	7	19	188	188	188	7
			ACCS				
AA2	37	63	219	287	249	249	101
BK2	15	58	99	188	309	309	96
CAL	22	227	344	476	1117	1118	282
DD1	8	20	94	377	511	655	46
F7I	14	95	491	1062	1062	1062	77
GLG	5	175	668	668	702	702	81
GLY	14	82	232	411	434	434	51
KRA	11	12	34	32	32	32	15
LAC	4	34	74	169	432	432	51
SQE	25	26	26	80	218	218	29
SQS	5	6	15	25	289	513	13
THI	75	158	367	517	694	694	151
ULD	8	25	241	662	760	760	80
XAN	2	5	9	8	11	11	6

Table B.3: FCoN database compound retrieval. The median number of database compounds containing a clique is given for each co-occurrence threshold and for pooled sets.

Activity Class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1.0$	**Pooled**
			MACCS				
AA2	**4.8**	0.1	0.5	0.1	0.1	0.1	4.2
BK2	10.8	**18.4**	13.6	8.0	0.2	0.0	18.2
CAL	**14.3**	13.1	4.1	2.3	1.8	1.8	**14.3**
DD1	9.4	**15.1**	9.9	6.4	1.1	0.2	13.8
F7I	5.5	5.7	2.6	1.6	1.6	1.6	**6.4**
GLG	**4.9**	2.3	0.3	0.0	0.0	0.0	4.2
GLY	15.7	13.8	12.1	9.2	3.5	1.0	**19.6**
KRA	61.3	54.5	32.0	26.2	6.9	6.2	**61.9**
LAC	34.2	**36.1**	34.8	34.9	31.2	31.2	35.7
SQE	31.6	33.6	32.0	25.4	10.3	10.3	**35.0**
SQS	33.4	29.8	31.2	10.1	1.0	0.3	**36.5**
THI	**1.0**	0.1	0.1	0.1	0.1	0.1	**1.0**
ULD	3.4	**5.7**	3.9	3.8	3.8	3.8	3.4
XAN	39.1	44.9	21.6	6.1	0.5	0.1	**45.5**
			ECFP4				
AA2	29.0	**34.8**	32.8	34.0	34.3	34.3	32.1
BK2	61.8	66.4	63.6	63.6	62.7	62.7	**69.1**
CAL	50.7	48.7	47.4	52.4	53.6	54.8	**60.0**
DD1	63.6	**76.3**	72.4	67.7	60.8	60.8	75.7
F7I	63.7	58.8	56.8	50.7	53.6	53.6	**67.3**
GLG	36.1	**39.1**	38.2	32.8	34.0	34.0	36.9
GLY	76.2	74.1	74.7	76.1	76.7	76.7	**78.8**
KRA	62.1	64.1	70.0	73.6	**75.5**	**75.5**	**75.5**
LAC	57.9	69.3	69.4	70.0	68.7	68.7	**72.3**
SQE	59.8	62.2	59.4	56.8	57.3	57.3	**66.4**
SQS	39.7	41.2	45.9	47.7	46.0	46.0	**48.5**
THI	40.8	40.7	46.4	**46.5**	46.4	46.4	45.9
ULD	36.3	36.0	35.9	36.9	36.9	36.9	**40.3**
XAN	56.2	67.1	65.3	69.2	69.4	70.6	**72.9**
			ACCS				
AA2	14.6	15.9	15.5	15.0	15.0	15.0	**16.8**
BK2	32.3	43.5	44.5	**44.6**	41.5	41.5	**44.6**
CAL	28.3	**30.5**	16.9	12.6	11.7	11.7	28.8
DD1	56.4	64.1	64.4	59.2	48.4	44.4	**65.6**
F7I	11.7	**13.9**	6.9	5.9	5.9	5.9	13.6
GLG	25.9	27.7	22.0	18.9	11.5	11.5	**35.3**
GLY	37.0	27.8	21.6	17.4	15.2	15.2	**37.9**
KRA	36.0	47.7	47.7	51.3	50.9	50.9	**57.0**
LAC	47.7	49.3	47.2	34.6	22.7	22.7	**55.0**
SQE	28.0	29.4	26.9	19.7	17.9	17.9	**32.4**
SQS	34.0	41.7	50.3	43.4	44.0	43.5	**52.9**
THI	**12.3**	8.3	6.1	5.7	5.9	5.9	11.0
ULD	22.6	27.1	23.4	18.7	18.7	18.7	**31.7**
XAN	36.9	46.3	46.7	51.3	57.1	57.1	**61.2**

Table B.4: FCoN recovery rates. The recovery rates at different co-occurrence thresholds and pooled sets are given in percent. Top recovery rates are highlighted in bold.

B.2 Fragment Formal Concept Analysis

Table B.5 reports the data sets used for FragFCA classifier benchmarking.

Table B.6 reports the FragFCA classification results for all supersets.

Superset	MDDR act. index	Activity class	Cmpds.	Frags.	Sig.
		MDDR			
Adenosine	07707	A1 agonists	146	40	27.8
	07708	A2 agonists	114		28.7
MAO	08410	MAO A inhibitors	67	86	47.2
	08420	MAO B inhibitors	158		75.6
Adrenergic	31251	β1 blockers	65	284	846.5
	31261	α1 blockers	507		1188.9
	31262	α2 blockers	268		1167.8
Opioid	31251	κ agonists	243	201	690.5
	31261	δ agonists	347		464.1
	31262	μ agonists	89		434.8
PDE	78415	PDE I inhibitors	69	589	437.2
	78417	PDE III inhibitors	408		1973.4
	78418	PDE IV inhibitors	1994		2634.3
Serine proteases	37110	Thrombin inhibitors	1037	676	13160.8
	37121	Factor Xa inhibitors	1121		7436.8
	37125	Factor VIIa inhibitors	156		5422.8
Dopamine	07702	D1 antagonists	136	330	332.1
	07701	D2 antagonists	439		2006.8
	07703	D3 antagonists	249		1191.2
	07710	D4 antagonists	647		1009.5
Serotonin	06235	5-HT1A agonists	943	430	3141.1
	06237	5-HT1C agonists	180		579.8
	06246	5-HT1D agonists	528		1702.7
	06251	5-HT1F agonists	110		1116.6
		BindingDB			
PDE		PDE III inhibitors	69		
		PDE IV inhibitors	51		
Serine proteases		Thrombin inhibitors	89		
		Factor Xa inhibitors	597		
		Factor VIIa inhibitors	95		

Table B.5: FragFCA classifier dataset. The composition of the compound data sets is reported. BindingDB sets were used as additional test cases from which no fragment information was derived. "Cmpds.": number of compounds; "Frags.": number of fragments in superset fragment library; "Sig.": average number of signature fragment combinations.

Activity class	CS	CM	IS	IM	NM	E
Adenosine	37.2	4.6	3.4	6.9	47.9	84.3
A1	41.3	5.3	4.4	2.4	46.7	94.5
A2	32.0	3.7	2.3	12.6	49.4	71.7
MAO	78.7	6.7	3.9	5.1	5.6	93.9
MAO A	72.6	12.8	3.5	8.0	3.1	90.1
MAO B	81.3	4.1	4.1	3.9	6.7	95.4
Adrenergic	45.5	35.4	15.2	2.5	1.4	94.8
$\beta 1$	61.3	33.9	4.8	0.0	0.0	100.0
$\alpha 1$	37.7	40.3	18.1	2.6	1.3	93.6
$\alpha 2$	58.8	25.5	11.1	2.8	1.8	95.4
Opioid	55.2	31.9	7.6	2.1	3.3	96.4
κ	52.7	38.6	6.9	1.0	0.7	98.1
δ	59.0	28.6	6.2	1.7	4.7	97.2
μ	44.6	26.6	16.5	7.1	5.1	86.3
PDE	25.4	48.1	23.6	1.9	1.0	93.0
PDE I	49.8	44.9	3.4	1.1	0.9	97.9
PDE III	27.7	46.5	21.8	2.9	1.1	90.4
PDE IV	24.4	48.5	24.4	1.8	1.0	93.3
Serine proteases	12.9	56.5	28.4	1.5	0.8	89.6
Thrombin	13.7	57.1	27.0	1.4	0.8	91.0
Factor Xa	10.1	55.9	31.3	1.8	0.9	85.2
Factor VIIa	32.5	56.5	10.9	0.0	0.1	100.0
Dopamine	30.6	35.9	23.1	6.8	3.6	81.9
D1	54.9	29.8	5.5	5.0	4.8	91.7
D2	22.1	43.7	27.6	4.6	2.0	82.9
D3	38.6	41.8	17.5	1.2	0.9	97.0
D4	29.7	29.8	24.8	10.7	5.4	73.5
Serotonin	27.5	46.0	21.4	2.6	2.5	91.2
5-HT1A	24.1	50.9	21.9	1.9	1.2	92.7
5-HT1C	42.5	37.5	8.7	4.7	6.7	90.1
5-HT1D	28.7	38.9	25.2	3.5	3.8	89.3
5-HT1F	33.2	48.4	14.4	2.8	1.2	92.3

Table B.6: FragFCA compound classification. Compound classification results are reported for each superset and activity class in percent of classified compounds. CS: correct, single class match; CM: correct, multiple classes matched; IM: incorrect, multiple classes matched; IS: incorrect, single class match; E: enrichment, calculated as the ratio of correct matches against only one class over all single class matches (both correct and incorrect) in percent.

B.3 Molecular Formal Concept Analysis

Table B.7 reports reference and identified compounds for each MolFCA query.

Query	Compound name	Note
HDAC	4-phenylimidazole, 17	
	Apicidin	selective HDAC1,HDAC3 / HDAC6
	CRA-024781	
	LAQ-824	
	PXD101	
	SAHA	reference compound
	thiophene derivative, 15h	selective HDAC6/HDAC1
	thiophene derivative, 19c	
	thiophene derivative, 3b	
	thiophene derivative, 3c	
	thiophene derivative, 3d	
	thiophene derivative, 3f	
	thiophene derivative, 3g	
	thiophene derivative, 3h	
	thiophene derivative, 3i	
	Triazole Ligand, 10a	
	Triazole Ligand, 10b	
	Triazole Ligand, 10c	
	Triazole Ligand, 10d	
	Triazole Ligand, 12b	
	Triazole Ligand, 14a	
	Triazole Ligand, 14b	
	Triazole Ligand, 14c	
	Triazole Ligand, 14d	
	Triazole Ligand, 6b	selective HDAC6/HDAC1
PDE	(R,S)-Mesopram	selective PDE2B/PDE10A
	Cilomilast	reference compound
	Filaminast	selective PDE2B/PDE11A
	Piclamilast	
	Roflumilast	selective PDE2B/PDE11A
	Rolipram	selective PDE2B/PDE10A
	Zardaverine	selective PDE2B/PDE11A
IMPDH	C2-Mycophenolic Adenine Dinucleotide (C2-MDA)	
	Mycophenolic Acid (MPA)	reference compound, $\leq 100nM$
	Mycophenolic Adenine Dinucleotide (MAD) Analogue, 37	
	Mycophenolic Adenine Dinucleotide (MAD) Analogue, 38	$\leq 100nM$
	Tiazofurin Adenine Dinucleotide (TAD)	
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 25	$\leq 100nM$
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 26	
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 27	$\leq 100nM$
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 28	
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 29	
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 30	
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 31	$\leq 100nM$
	Tiazofurin Adenine Dinucleotide (TAD) Analogue, 32	$\leq 100nM$
Caspase	Ac-DEVD-CHO	
	Burnham Institute Compound 1	low potency
	Burnham Institute Compound 2	low potency
	Inhibitor 3	
	Inhibitor 66a	
	Isoquinoline-1,3,4-trione 13f	
	Isoquinoline-1,3,4-trione 9k	
	valine aspartyl ketone 14	
	valine aspartyl ketone 35	

Table B.7: Selected BindingDB compounds. For all four queries, the identified BindingDB compounds are reported. "Compound name" corresponds to the BindingDB field "BindingDB Monomer Display Name".

Bibliography

[1] H. Lüllmann, K. Mohr, and M. Wehling. *Pharmakologie und Toxikologie*. Thieme, Stuttgart, 15th edition, 2002.

[2] H. Böhm, G. Klebe, and H. Kubinyi. *Wirkstoffdesign: Der Weg zum Arzneimittel*. Spektrum Akademischer Verlag, 1st edition, 2002.

[3] C.M. Dobson. Chemical space and biology. *Nature*, 432:824–828, 2004.

[4] D. Rognan. Chemogenomic approaches to rational drug design. *Brit. J. Pharm.*, 152:38–52, 2007.

[5] D. Schnur, B.R. Beno, A. Good, and A. Tebben. Approaches to target class combinatorial library design. *Methods Mol. Biol.*, 275:355–378, 2004.

[6] A.P. Russ and S. Lampel. The druggable genome: An update. *Drug Discov. Today*, 10:1607–1610, 2005.

[7] G.V. Paolini, R.H.B. Shapland, W.P.v. Hoorn, J.S. Mason, and A.L. Hopkins. Global mapping of pharmacological space. *Nature Biotech.*, 24:805–815, 2006.

[8] M.A. Yildirim, K. Goh, M.E. Cusick, A. Barabási, and M. Vidal. Drug-target network. *Nature Biotech.*, 25:1119–1126, 2007.

[9] P.R. Caron, M.D. Mullican, R.D. Mashal, K.P. Wilson, M.S. Su, and M.A. Murcko. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.*, 5:464–470, 2001.

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

[11] A.R. Leach and V.J. Gillet. *An Introduction to Chemoinformatics*. Springer, 2007.

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

[13] P. Willett, J.M. Barnard, and G.M. Downs. Chemical similarity searching. *J. Chem. Inf. Comp. Sci.*, 38:983–996, 1998.

[14] L.M. Kauvar, D.L. Higgins, H.O. Villar, J.R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K.E. Bauer, H. Dilley, and D.M. Rocke. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. and Biol.*, 2: 107–118, 1995.

[15] M. Vieth, R.E. Higgs, D.H. Robertson, M. Shapiro, E.A. Gragg, and H. Hemmerle. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. and Biophys. Acta*, 1697:243–257, 2004.

[16] C.M. Krejsa, D. Horvath, S.L. Rogalski, J.E. Penzotti, B. Mao, F. Barbosa, and J.C. Migeon. Predicting ADME properties and side effects: The BioPrint approach. *Curr. Opin. Drug Discov. and Devel.*, 6:470–480, 2003.

[17] J. Singh, Z. Deng, G. Narale, and C. Chuaqui. Structural interaction fingerprints: A new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. and Drug Des.*, 67:5–12, 2006.

[18] M. Campillos, M. Kuhn, A. Gavin, L.J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321:263–266, 2008.

[19] M.A. Fabian, W.H. Biggs, D.K. Treiber, C.E. Atteridge, M.D. Azimioara, M.G. Benedetti, T.A. Carter, P. Ciceri, P.T. Edeen, M. Floyd, J.M. Ford, M. Galvin, J.L. Gerlach, R.M. Grotzfeld, S. Herrgard, D.E. Insko, M.A. Insko, A.G. Lai, J. Lélias, S.A. Mehta, Z.V. Milanov, A.M. Velasco, L.M. Wodicka, H.K. Patel, P.P. Zarrinkar, and D.J. Lockhart. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nature Biotech.*, 23:329–336, 2005.

[20] S. Hiller, V. Golender, A. Rosenblit, L. Rastrigin, and A. Glaz. Cybernetic methods of drug design. I. Statement of the problem–The perceptron approach. *Comp. Biomed. Res.*, 6:411–421, 1973.

[21] M.A. Johnson and G.M. Maggiora. *Concepts and Applications of Molecular Similarity*. Wiley-Interscience, 1st edition, 1990.

[22] H. Eckert and J. Bajorath. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discov. Today*, 12:225–233, 2007.

[23] L. Peltason and J. Bajorath. SAR index: Quantifying the nature of structure-activity relationships. *J. Med. Chem.*, 50:5571–8, 2007.

[24] R. Guha and J.H.v. Drie. Structure–activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, 48:646–58, 2008.

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

[26] M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, and J. Bajorath. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.*, 51:6075–6084, 2008.

[27] M. Wawer, L. Peltason, and J. Bajorath. Elucidation of structure-activity relationship pathways in biological screening data. *J. Med. Chem.*, 52: 1075–1080, 2009.

[28] J.L. Medina-Franco, K. Martínez-Mayorga, A. Bender, R.M. Marín, M.A. Giulianotti, C. Pinilla, and R.A. Houghten. Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *J. Chem. Inf. Model.*, 49:477–491, 2009.

[29] J. Bajorath. Integration of virtual and High-Throughput screening. *Nature Rev. Drug Discov.*, 1:882–894, 2002.

[30] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comp. Sci.*, 43:391–405, 2003.

[31] A. Bender, J.L. Jenkins, M. Glick, Z. Deng, J.H. Nettles, and J.W. Davies. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J. Chem. Inf. Model.*, 46:2445–2456, 2006.

[32] A.E. Cleves and A.N. Jain. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.*, 49:2921–2938, 2006.

[33] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, and B.K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotech.*, 25:197–206, 2007.

[34] J. Hert, M.J. Keiser, J.J. Irwin, T.I. Oprea, and B.K. Shoichet. Quantifying the relationships among drug classes. *J. Chem. Inf. Model.*, 48: 755–765, 2008.

[35] G. Müller. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today*, 8:681–691, 2003.

[36] D.M. Schnur, M.A. Hermsmeier, and A.J. Tebben. Are target-family-privileged substructures truly privileged? *J. Med. Chem.*, 49:2000–2009, 2006.

[37] B.E. Evans, K.E. Rittle, M.G. Bock, R.M. DiPardo, R.M. Freidinger, W.L. Whitter, G.F. Lundell, D.F. Veber, P.S. Anderson, and R.S. Chang. Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.*, 31:2235–2246, 1988.

[38] T. Klabunde and G. Hessler. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem*, 3:928–944, 2002.

[39] A.M. Aronov and G.W. Bemis. A minimalist approach to fragment-based ligand design using common rings and linkers: Application to kinase inhibitors. *Proteins*, 57:36–50, 2004.

[40] A.M. Aronov, B. McClain, C.S. Moody, and M.A. Murcko. Kinase-likeness and kinase-privileged fragments: Toward virtual polypharmacology. *J. Med. Chem.*, 51:1214–1222, 2008.

[41] G.W. Bemis and M.A. Murcko. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, 39:2887–2893, 1996.

[42] G.W. Bemis and M.A. Murcko. Properties of known drugs. 2. Side chains. *J. Med. Chem.*, 42:5095–5099, 1999.

[43] N.E. Shemetulskis, D. Weininger, C.J. Blankley, J.J. Yang, and C. Humblet. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comp. Sci.*, 36:862–871, 1996.

[44] E. Lameijer, J.N. Kok, T. Bäck, and A.P. Ijzerman. Mining a chemical database for fragment co-occurrence: Discovery of "chemical clichés". *J. Chem. Inf. Model.*, 46:553–562, 2006.

[45] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M.A. Koch, and H. Waldmann. The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.*, 47:47–58, 2007.

[46] J.J. Sutherland, R.E. Higgs, I. Watson, and M. Vieth. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.*, 51:2689–2700, 2008.

[47] J. Batista and J. Bajorath. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.*, 47:1405–1413, 2007.

[48] J. Batista and J. Bajorath. Similarity searching using compound class-specific combinations of substructures found in randomly generated molecular fragment populations. *ChemMedChem*, 3:67–73, 2008.

[49] E.v.d. Horst, Y. Okuno, A. Bender, and A.P. Ijzerman. Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. *J. Chem. Inf. Model.*, (49):348–360, 2009.

[50] P.R.N. Wolohan, L.B. Akella, R.J. Dorfman, P.G. Nell, S.M. Mundt, and R.D. Clark. Structural unit analysis identifies lead series and facilitates scaffold hopping in combinatorial chemistry. *J. Chem. Inf. Model.*, 46: 1188–1193, 2006.

[51] U. Priss. Formal concept analysis in information science. *Ann. Rev. Inf. Sci. Technol.*, 40:521–543, 2006.

[52] C. Merlot, D. Domine, C. Cleva, and D.J. Church. Chemical substructures in drug discovery. *Drug Discov. Today*, 8:594–602, 2003.

[53] J. Bajorath. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comp. Sci.*, 41:233–245, 2001.

[54] P. Ertl and S. Jelfs. Designing drugs on the internet? Free web tools and services supporting medicinal chemistry. *Curr. Top. Med. Chem.*, 7: 1491–1501, 2007.

[55] S.W. Benson and J.H. Buss. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *J. Chem. Phys.*, 29:546–572, 1958.

[56] C. Hansch. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, 2:232–239, 1969.

[57] G.W. Adamson, M.F. Lynch, and W.G. Town. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part II. Atom-centred fragments. *J. Chem. Soc.*, C:3702–3706, 1971.

[58] L. Hodes, G.F. Hazard, R.I. Geran, and S. Richman. A statistical-heuristic methods for automated selection of drugs for screening. *J. Med. Chem.*, 20:469–475, 1977.

[59] P. Willett. A screen set generation algorithm. *J. Chem. Inf. Comp. Sci.*, 19:159–162, 1979.

[60] A. Feldman and L. Hodes. An efficient design for chemical structure searching. I. The screens. *J. Chem. Inf. Comp. Sci.*, 15:147–152, 1975.

[61] G.W. Adamson, J. Cowell, M.F. Lynch, A.H.W. McLure, W.G. Town, and A.M. Yapp. Strategic considerations in the design of a screening system for substructure searches of chemical structure files. *J. Chem. Doc.*, 13:153–157, 1973.

[62] G.W. Adamson, J.A. Bush, A.H.W. McLure, and M.F. Lynch. An evaluation of a substructure search screen system based on bond-centered fragments. *J. Chem. Doc.*, 14:44–48, 1974.

[63] M.J. McGregor and P.V. Pallai. Clustering of large databases of compounds: Using the MDL keys as structural descriptors. *J. Chem. Inf. Comp. Sci.*, 37:443–448, 1997.

[64] M. Clark. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.*, 45:30–38, 2005.

[65] H. Matter, K.H. Baringhaus, T. Naumann, T. Klabunde, and B. Pirard. Computational approaches towards the rational design of drug-like compound libraries. *J. Combi. Chem. and High-Throughput Screen.*, 4:453–475, 2001.

[66] T.I. Oprea, A.M. Davis, S.J. Teague, and P.D. Leeson. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comp. Sci.*, 41:1308–1315, 2001.

[67] W.P. Walters and M.A. Murcko. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.*, 54:255–271, 2002.

[68] A. Bender, H.Y. Mussa, R.C. Glen, and S. Reiling. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comp. Sci.*, 44:1708–1718, 2004.

[69] L. Xing and R.C. Glen. Novel methods for the prediction of logP, pK_a, and logD. *J. Chem. Inf. Comp. Sci.*, 42:796–805, 2002.

[70] R.E. Carhart, D.H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comp. Sci.*, 25:64–73, 1985.

[71] S.K. Kearsley, S. Sallamack, E.M. Fluder, J.D. Andose, R.T. Mosley, and R.P. Sheridan. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comp. Sci.*, 36:118–127, 1996.

[72] N. Yu and G.A. Bakken. Efficient exploration of large combinatorial chemistry spaces by monomer-based similarity searching. *J. Chem. Inf. Model.*, 49:745–755, 2009.

[73] A.M. Clark and P. Labute. Detection and assignment of common scaffolds in project databases of lead molecules. *J. Med. Chem.*, 52:469–483, 2009.

[74] X.Q. Lewell, D.B. Judd, S.P. Watson, and M.M. Hann. RECAP–retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comp. Sci.*, 38:511–522, 1998.

[75] D.J. Graham, C. Malarkey, and M.V. Schulmerich. Information content in organic molecules: Quantification and statistical structure via Brownian processing. *J. Chem. Inf. Comp. Sci.*, 44:1601–1611, 2004.

[76] J. Batista, J.W. Godden, and J. Bajorath. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, 46:1937–1944, 2006.

[77] J. Batista and J. Bajorath. Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, 47:59–68, 2007.

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

[79] D. Weininger, A. Weininger, and J.L. Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comp. Sci.*, 29: 97–101, 1989.

[80] E. Lounkine, J. Batista, and J. Bajorath. Mapping of activity-specific fragment pathways isolated from random fragment populations reveals the formation of coherent molecular cores. *J. Chem. Inf. Model.*, 47: 2133–2139, 2007.

[81] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[82] I. Eidhammer, I. Jonassen, and W.R. Taylor. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. Wiley, 1st edition, 2004.

[83] C.A. James and D. Weininger. *Daylight theory manual*. Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA, 2008.

[84] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[85] L. Xue, J.W. Godden, F.L. Stahura, and J. Bajorath. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comp. Sci.*, 43:1218–1225, 2003.

[86] C. Williams. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Divers.*, 10:311–332, 2006.

[87] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16:575–577, 1973.

[88] J.L. Durant, B.A. Leland, D.R. Henry, and J.G. Nourse. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comp. Sci.*, 42: 1273–1280, 2002.

[89] I. Vogt, H.E.A. Ahmed, J. Auer, and J. Bajorath. Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Divers.*, 12:25–40, 2008.

[90] P. Becker, J. Hereth, and G. Stumme. ToscanaJ: An open source tool for qualitative data analysis. In V. Duquenne, B. Ganter, M. Liquiere, E.M. Nguifo, and G. Stumme, editors, *Advances in Formal Concept Analysis for Knowledge Discovery in Databases.*, pages 1–2, Lyon, France, 2002.

[91] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, 2004.

[92] S.V. Frye. Structure-activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. *Chem. and Biol.*, 6:R3–R7, 1999.

[93] X. Xia, E.G. Maliski, P. Gallant, and D. Rogers. Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.*, 47:4463–4470, 2004.

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

[95] I. Vogt, D. Stumpfe, H.E.A. Ahmed, and J. Bajorath. Methods for computer-aided chemical biology. Part 2: Evaluation of compound selectivity using 2D molecular fingerprints. *Chem. Biol. and Drug Des.*, 70: 195–205, 2007.

[96] D. Stumpfe, H. Geppert, and J. Bajorath. Methods for computer-aided chemical biology. Part 3: Analysis of structure-selectivity relationships through single- or dual-step selectivity searching and bayesian classification. *Chem. Biol. and Drug Des.*, 71:518–528, 2008.

[97] H.E.A. Ahmed, H. Geppert, D. Stumpfe, E. Lounkine, and J. Bajorath. Methods for computer-aided chemical biology. Part 4: Selectivity searching for ion channel ligands and mapping of molecular fragments as selectivity markers. *Chem. Biol. and Drug Des.*, 73:273–282, 2009.

[98] M. Haberland, R.L. Montgomery, and E.N. Olson. The many roles of histone deacetylases in development and physiology: Implications for disease and therapy. *Nature Rev. Genetics*, 10:32–42, 2009.

[99] M.A. Giembycz. Cilomilast: A second generation phosphodiesterase 4 inhibitor for asthma and chronic obstructive pulmonary disease. *Exp. Opin. Invest. Drugs*, 10:1361–1379, 2001.

[100] M. Sanford and G.M. Keating. Enteric-coated mycophenolate sodium: A review of its use in the prevention of renal transplant rejection. *Drugs*, 68:2505–2533, 2008.

[101] S. Natori, H. Higuchi, P. Contreras, and G.J. Gores. The caspase inhibitor IDN-6556 prevents caspase activation and apoptosis in sinusoidal endothelial cells during liver preservation injury. *Liver Transpl.*, 9:278–284, 2003.

[102] J.W. Han, S.H. Ahn, S.H. Park, S.Y. Wang, G.U. Bae, D.W. Seo, H.K. Kwon, S. Hong, H.Y. Lee, Y.W. Lee, and H.W. Lee. Apicidin, a histone deacetylase inhibitor, inhibits proliferation of tumor cells via induction of p21WAF1/Cip1 and gelsolin. *Cancer Res.*, 60:6068–6064, 2000.

[103] J.C. Kips, G.F. Joos, R.A. Peleman, and R.A. Pauwels. The effect of zardaverine, an inhibitor of phosphodiesterase isoenzymes III and IV, on endotoxin-induced airway changes in rats. *Clin. and Exp. Allergy*, 23: 518–523, 1993.

[104] D. Spina. PDE4 inhibitors: Current status. *Brit. J. Pharm.*, 155:308–315, 2008.

[105] R.J. Heaslip, L.J. Lombardo, J.M. Golankiewicz, B.A. Ilsemann, D.Y. Evans, B.D. Sickels, J.K. Mudrick, J. Bagli, and B.M. Weichman. Phosphodiesterase-IV inhibition, respiratory muscle relaxation and bronchodilation by WAY-PDA-641. *J. Pharm. Exp. Therap.*, 268:888–896, 1994.

[106] H. Dinter, J. Tse, M. Halks-Miller, D. Asarnow, J. Onuffer, D. Faulds, B. Mitrovic, G. Kirsch, H. Laurent, P. Esperling, D. Seidelmann, E. Ottow, H. Schneider, V.K. Tuohy, H. Wachtel, and H.D. Perez. The Type IV phosphodiesterase specific inhibitor mesopram inhibits experimental autoimmune encephalomyelitis in rodents. *J. Neuroimmun.*, 108:136–146, 2000.

[107] H. Wachtel. Potential antidepressant activity of rolipram and other selective cyclic adenosine 3',5'-monophosphate phosphodiesterase inhibitors. *Neuropharmacology*, 22:267–272, 1983.

[108] N. Sommer, P.A. Löschmann, G.H. Northoff, M. Weller, A. Steinbrecher, J.P. Steinbach, R. Lichtenfels, R. Meyermann, A. Riethmüller, and A. Fontana. The antidepressant rolipram suppresses cytokine production and prevents autoimmune encephalomyelitis. *Nature Medicine*, 1: 244–248, 1995.

[109] B. Alberts, A. Johnson, P. Walter, J. Lewis, M. Raff, and K. Roberts. *Molecular Biology of the Cell.* Taylor and Francis, 5th edition, 2008.

[110] T. Liu, Y. Lin, X. Wen, R.N. Jorrisen, and M.K. Gilson. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35:D198–D201, 2007.

[111] X. Chen, Y. Lin, and M.K. Gilson. The Binding Database: Overview and user's guide. *Biopolymers. Nucleic Acid Sci.*, 61:127–141, 2002a.

[112] X. Chen, Y. Lin, M. Liu, and M.K. Gilson. The Binding Database: Data management and interface design. *Bioinformatics*, 18:130–139, 2002b.

[113] X. Chen, M. Liu, and M.K. Gilson. Binding DB: A web-accessible molecular recognition database. *J. Combi. Chem. High-Throughput Screen.*, 4: 719–725, 2001.

[114] J.J. Irwin and B.K. Shoichet. ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45:177–182, 2005.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Computational Methods for the Integration of Biological Activity and Chemical Space" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

E. Lounkine, J. Batista, J. Bajorath. Random molecular fragment methods in computational medicinal chemistry. *Curr. Med. Chem.*, 15:2108-2121, 2008.

E. Lounkine, J. Bajorath. Core trees and consensus fragment sequences for molecular representation and similarity analysis. *J. Chem. Inf. Model.*, 48:1161–1166, 2008.

E. Lounkine, J. Bajorath. Topological fragment index for the analysis of molecular substructures and their topological environment in active compounds. *J. Chem. Inf. Model.*, 49:162-168, 2009.

E. Lounkine, Y. Hu, J. Batista, J. Bajorath. Relevance of feature combinations for similarity searching using general or activity class-directed molecular fingerprints. *J. Chem. Inf. Model.*, 49:561-570,2009.

E. Lounkine, J. Auer, J. Bajorath. Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. *J. Med. Chem.*, 51:5342-5348, 2008.

F. Krüger, E. Lounkine, J. Bajorath. Fragment formal concept analysis accurately classifies compounds with closely related biological activities. *ChemMedChem*, 2009, in press.

E. Lounkine, D. Stumpfe, J. Bajorath. Molecular formal concept analysis for compound selectivity profiling in biologically annotated databases *J. Chem. Inf. Model.*, 2009, in press.