Computational Analysis of Structure-Activity Relationships

—

From Prediction to Visualization Methods

Dissertation

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

vorgelegt von
Anne Mai Wassermann
aus Göttingen

Bonn 2012

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

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

2. Gutachter: Univ.-Prof. Dr. med. Joachim Schultze

Tag der Promotion: 05. Juli 2012

Erscheinungsjahr: 2012

Abstract

Understanding how structural modifications affect the biological activity of small molecules is one of the central themes in medicinal chemistry. By no means is structure-activity relationship (SAR) analysis a priori dependent on computational methods. However, as molecular data sets grow in size, we quickly approach our limits to access and compare structures and associated biological properties so that computational data processing and analysis often become essential.

Here, different types of approaches of varying computational complexity for the analysis of SAR information are presented, which can be applied in the context of screening and chemical optimization projects. The first part of this thesis is dedicated to machine-learning strategies that aim at de novo ligand prediction and the preferential detection of potent hits in virtual screening. High emphasis is put on benchmarking of different strategies and a thorough evaluation of their utility in practical applications. However, an often claimed disadvantage of these prediction methods is their "black box" character because they do not necessarily reveal which structural features are associated with biological activity. Therefore, these methods are complemented by more descriptive SAR analysis approaches showing a higher degree of interpretability. Concepts from information theory are adapted to identify activity-relevant structure-derived descriptors. Furthermore, compound data mining methods exploring prespecified properties of available bioactive compounds on a large scale are designed to systematically relate molecular transformations to activity changes. Finally, these approaches are complemented by graphical methods that primarily help to access and visualize SAR data in congeneric series of compounds and allow the formulation of intuitive SAR rules applicable to the design of new compounds. The compendium of SAR analysis tools introduced in this thesis investigates SARs from different perspectives.

Acknowledgments

I would like to thank

- my PhD supervisor Prof. Dr. Jürgen Bajorath for his continuous inspiration, valuable mentorship, and thoughtful advices on scientific and personal questions

- Prof. Dr. Joachim Schultze for taking the time to review this thesis as a co-referent

- Dr. Martin Vogt who was always approachable for my questions and answered them with immeasurable patience, kindness, and attention to detail

- Dr. Hanna Geppert, Dr. Lisa Peltason, and Dr. Dagmar Stumpfe who introduced me to different aspects of chemoinformatics and from whom I learned in numerous ways

- my dear friend and colleague Dr. Mathias Wawer for his reliable support, his sincere scientific opinion on countless occasions, and the many good times we shared in our office

- Dilyana Dimova, Kathrin Heikamp, and Dr. Britta Nisius for pleasant and productive collaborations that contributed to the completion of this thesis

- Boehringer Ingelheim Pharma for funding

Contents

List of Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
AID	PubChem assay identifier
CAG	combinatorial analogue graph
DAG	directed acyclic graph
DRC	directed R-group combination
DRCG	directed R-group combination graph
DSE	Differential Shannon Entropy
EC_{50}	half maximal effective concentration
ECFP	extended-connectivity fingerprint
GO	Gene Ontology
GPCR	G protein-coupled receptor
H-bond	hydrogen bond
hAR	human adenosine receptor
HTS	high-throughput screening
IC_{50}	half maximal inhibitory concentration
ID	identifier
K_i	inhibitor dissociation constant
LBVS	ligand-based virtual screening

LC	linear combination
MCS	maximum common subgraph
MDDR	MDL Drug Data Report
MI	mutual information
MMP	matched molecular pair
MMPA	matched molecular pair analysis
MOE	Molecular Operating Environment
mtSAR	multi-target structure-activity relationship
pK_i	negative decadic logarithm of K_i
QSAR	quantitative structure-activity relationship
SAK	structure-activity kernel
SAR	structure-activity relationship
SMILES	Simplified Molecular Input Line System
SPP	similarity-property principle
SSR	structure-selectivity relationship
SVM	support vector machine
Tc	Tanimoto coefficient
TGD	typed graph distance
TLK	target-ligand kernel
uPA	urokinase
VS	virtual screening

Chapter 1

Introduction

Chemoinformatics encompasses the development and application of computational methods to solve chemical problems. Although the term chemoinformatics was only introduced about a decade ago, it refers to a research field with a long tradition because individual areas of chemoinformatics, such as chemical structure representation and searching, molecular modeling, and computer-assisted structure elucidation, have their origins back in the 1960s [1]. Since then, the interest in computational methods in chemistry has continuously grown. The rising popularity is largely ascribed to increasing amounts of data that are produced and cannot be dealt with without computational means [2]. In particular, the introduction of combinatorial chemistry and high-throughput screening (HTS) has triggered the need for computational data management and analysis which is facilitated by continuing advances in computational power. These technological innovations have also paved the way for the establishment of chemoinformatics methods in the pharmaceutical industry.

Pharmaceutically-oriented chemoinformatics focuses on small molecules and their interactions with targets, i.e., their biological activity. A prime objective of chemoinformatics methods is the identification of compounds with desired biological activities that might ultimately become drug candidates. To these ends, it is of central importance to analyze and understand the relationship between chemical structure and biological activity of small molecules. Currently available approaches to *structure-activity relationship* (SAR) analysis are multifaceted and of rather different methodological complexity. A general distinction can be made between methodologies that primarily help to access and visualize SAR data obtained from HTS or chemical optimization campaigns and those that ultimately predict biological activities. In this thesis, both kinds of methodologies are employed for the analysis and exploitation of SARs.

Prediction Methods in SAR Analysis

Virtual screening (VS) techniques process large databases of compounds in silico and are often considered as a cost-efficient complement to HTS [3]. Ligand-based VS (LBVS) utilizes information from known bioactive molecules to identify novel structures exhibiting the desired bioactivity. From a conceptual point of view, LBVS is based on the *similarity-property principle* (SPP) [4] that was articulated by Johnson and Maggiora in 1990 and states that overall structurally similar molecules should have similar biological properties. This 'holistic' view of molecular similarity [5] provides the basis for similarity searching [6]. Following this approach, one or multiple known bioactive molecules are used as reference compounds for comparison with database compounds of unknown bioactivity. The degree of 'whole molecule'-similarity between a database compound and the reference set is quantitatively assessed through the representation of entire compounds by sets of structure-derived descriptors and the application of a mathematical function that measures the similarity between descriptor sets. Database compounds are then ranked in order of decreasing similarity to the reference set. According to the SPP, top-ranked molecules are most likely to exhibit the desired bioactivity and constitute prime candidates for biological testing. In compound classification approaches, test molecules are usually compared to active and inactive reference compounds and assigned to the class to which they are more similar.

A more refined view of molecular similarity is used in the analysis of pharmacophores [7] or quantitative structure-activity relationships (QSARs) [8]. These methods focus on 'local' similarities [5] in the study of biological activity determinants. Pharmacophore analysis aims at generating a hypothesis about the spatial arrangement of groups of functionalities in a molecule that render it active by forming interactions with a biological target of interest. These feature arrangements determine the pharmacophore model that is subsequently used in database screening to prioritize molecules with similar geometric features. QSAR correlates biological activities of congeneric compounds with selected structural features and/or properties represented as numerical chemical descriptors. For a set of reference molecules, a numerical relationship between potency and selected descriptors is established to deduce a regression model that takes descriptor values of any compound as input and returns its predicted activity value. It follows that the selection of activity-relevant descriptors is highly critical for the prediction quality of QSAR models. Furthermore, QSAR models are usually built from series of structurally closely related compounds. Therefore, test compounds of a different chemotype than the reference molecules fall outside of the applicability domain of most QSAR models and their activity cannot be reliably predicted [9].

Limitations of Prediction Methods

Over the past few years, the integration of machine-learning and artificial intelligence concepts has led to increasingly sophisticated QSAR and LBVS methods with improved prediction accuracies [10, 11]. Nevertheless, in many cases, these methods still fail to produce reliable predictions for test compounds, even in the presence of highly similar reference compounds [12]. These observations can be explained by the limited validity of the concept underlying these approaches: although the SPP is intuitive and a central paradigm in medicinal chemistry, it is frequently observed that small modifications of chemical structure can significantly alter compound activity [13]. This finding can be rationalized by the specificity of molecular recognition processes that require a high degree of complementarity between interacting surfaces. For example, the precise fit of a molecule into a binding site and the formation of key interactions with a target can easily be hindered by a small structural modification that changes molecular shape or charge distribution. The term *activity cliff* refers to compound pairs showing high structural similarity but large differences in activity [14]. Activity cliffs are exploited by medicinal chemists in hit-to-lead and lead optimization projects where small structural modifications are systematically applied to an initial hit compound in order to achieve improved compound potency [13]. On the other hand, activity cliffs strongly complicate molecular similarity analysis and represent challenges for QSAR and LBVS applications, often making them a hopeless endeavor [12]. In light of the limited applicability of quantitative prediction and compound classification methods and their strong dependency on the nature of the SARs present in the activity class under study (vide infra), it is attractive to extend the spectrum of available SAR analysis methods with descriptive approaches that aim at the extraction of interpretable SAR information from compound data sets. These methods are primarily designed to guide compound design in hit-to-lead and lead optimization campaigns but do not generalize extracted patterns to new data in an automated manner. Furthermore, these approaches circumvent the "black box" character of many prediction methods that do not enable the user to trace back prediction results to molecular structure and do not reveal which structural features are ultimately activity-relevant.

Data Mining and Visualization Techniques in SAR Analysis

Various computational *data mining* and *visualization* methods have been developed to systematically identify interesting SAR features in activity-annotated compound data sets and present them in an interpretable way [15]. Scopes of

these methodologies are highly variable, ranging from the mere structuring of SAR data to the automated extraction of actual SAR information [16]. For example, standard clustering which is often applied in the analysis of large screening data sets makes use of whole-molecule similarity to group structurally similar compounds together. Often, single clusters contain congeneric series of compounds that show only small structural variations. In these cases, a subsequent visual inspection of clustered structures may lead to the recognition of SAR patterns, but the retrieval of interpretable SAR information is left to the medicinal chemist [15,16]. More informative are substructure-centric approaches that provide direct access to bioactivity values associated with structural fragments. A data structure called scaffold tree [17] systematically generates molecular building blocks from sets of bioactive compounds by first pruning all side chains and then iteratively removing rings from molecular structures. The organization of the generated substructures in a hierarchy and their annotation with bioactivity values of the compounds from which they were extracted [18] enables the identification of activity-prevalent molecular frameworks (also termed *scaffolds*) which can subsequently be exploited to design novel active compounds [19]. Such data mining approaches are not limited to the study of compounds sharing the same bioactivity. For example, the systematic comparison of substructures across different targets and target families allows to distinguish fragments that exclusively occur in ligands sharing a specific bioactivity [20,21] from molecular entities that promiscuously bind to many different targets [22,23] and pose a high risk of adverse drug reactions [24]. For later stages of medicinal chemistry efforts that focus on the optimization of series of structurally similar compounds, understanding potency effects of chemical substitutions is of central importance. For this purpose, computational approaches have been designed that either highlight interesting activity patterns for a particular chemotype under study [25] or systematically investigate effects of common chemical substituents on ligand potency across different activity classes [26,27].

Conceptually different from the above mentioned methodologies are so-called *SAR profiling* methods that integrate the analyses of the structural similarity of and potency differences between bioactive compounds and characterize the nature of the SARs underlying a compound set [16]. In principle, three different SAR categories are distinguished [28]. *Continuous* SARs are found in sets of compounds where gradual structural changes result in only small to moderate changes in compound potency, and increasingly diverse structures fall within the same potency range. In compound sets showing a *discontinuous* SAR type, small changes in compound structure lead to large-magnitude changes in potency. Activity cliffs represent the extreme form of SAR discontinuity. However, continuous and discontinuous SARs are not mutually exclusive and often co-occur in different subsets of a compound class. The SAR in these data sets is therefore termed *heterogeneous*. Knowledge about the SAR type inherent to a

class of compounds is of paramount importance and of high practical utility. Whereas discontinous SARs are considered as an indicator for the evolvability of compound sets and can be used for the prioritization of hits for further optimization, continuous SARs are a prerequisite for the successful applicability of QSAR and LBVS methods (vide supra). Numerical functions have been devised to quantitatively describe the nature of SARs [29, 30] and have only recently been complemented by a number of graphical visualization methods that provide an intuitive access to global and local SARs underlying compound data sets and reveal the presence of interpretable SAR rules [31].

Research Topics

In this thesis, various aspects of SAR analysis are investigated and computational methodologies that aid in the elucidation and exploitation of SAR information in different ways are presented. These approaches can be grouped thematically into the four areas of machine-learning, information theory, data mining, and visualization techniques. In the following, a brief overview of all research projects presented in this dissertation is given. A classification of the projects according to key methodological aspects is provided in Table 1.0-1.

Support Vector Machines in Orphan and Potency-Directed Screening

The term *support vector machine* (SVM) [32] refers to a machine-learning technique that has gained wide popularity in LBVS because of its ability to build complex but robust prediction models. Though originally introduced for classification, SVMs have been adapted to generate rankings of test databases and sort compounds by decreasing likelihood of being active [33]. We were interested in the specific tailoring of SVM strategies to two particularly challenging tasks in LBVS: orphan and potency-directed screening.

Orphan screening [11] aims at the identification of ligands for targets for which no known ligands are available. LBVS is usually not applicable in the absence of known active molecules. However, recent studies have introduced LBVS approaches that employ ligand information from similar targets to predict ligands for orphan targets [34–37]. Here, the key question is how to best define target similarity and weight the influence of ligands from different targets in the prediction model. To address this question, we employed different strategies to integrate a variety of target similarity functions and/or differently composed ligand reference sets into SVM learning and systematically compared them in simulated virtual screening trials on two different target protein systems.

Potency-directed screening approaches the question as to how compound potency information can be integrated into LBVS as an additional search pa-

rameter. In contrast to QSAR modeling, LBVS typically does not take compound potency into consideration, although the ability to direct search calculations towards the recognition of potent hits would certainly be attractive for practical applications. Therefore, we designed SVM-based strategies that do not only learn to separate active from inactive molecules but also distinguish between highly, intermediately, or weakly active compounds by incorporating categorized potency labels of reference molecules into model training. To assess their ability to enrich database selection sets with potent hits, the strategies were benchmarked on four HTS data sets.

Selection of Compound Class-Specific Descriptors

Descriptors that capture compound class-specific and biological activity-relevant information are of high interest for the exploration of structure-activity relationships and a major determinant for the success of LBVS and QSAR applications. The identification of compound class-specific information generally requires the comparison of descriptor value ranges and information content in a set of compounds having a desired property and data sets where all or at least the majority of compounds lack this property. This is in principle possible through adaptations of the *Shannon entropy* concept [38] from *information theory*. However, previous adaptations of this concept for descriptor profiling are insufficient to select discriminatory descriptors for data sets that dramatically differ in size. To circumvent these difficulties, we transformed a previously introduced information entropic strategy into mutual information analysis, a related concept, and investigated its utility to identify discriminatory descriptors on more than 160 activity classes.

Systematic Profiling of Activity Cliffs

Activity cliffs are considered to contain high SAR information content and have important implications for drug discovery efforts from more than one point of view (vide supra). Although activity cliffs have been intensely studied in individual compound sets, an open question is how they are globally distributed across available bioactive compounds and protein targets. Furthermore, the frequency of "multi-target cliffs", i.e., pairs of similar compounds showing large potency differences for multiple targets, is currently unknown. Therefore, we designed and carried out a large-scale data mining study that systematically searched for single- and multi-target activity cliffs in public domain compounds with reported activity against human targets.

Potency Effects of Molecular Transformations

In medicinal chemistry, potency effects of structural modifications are often studied on a case-by-case basis for series of structurally related compounds. However, it would also be interesting to know whether certain structural modifications have a higher propensity to retain or considerably change compound potency than others, irrespective of the chemotype or specific bioactivity of the compound they are applied to. To investigate this question, we used the framework of *matched molecular pairs* (MMPs) [39] to systematically extract *molecular transformations* from publicly available compound data annotated with activity values against human targets. An MMP is defined as a pair of two structurally related compounds that differ only at a single localized site and are hence distinguished by a defined substructure. Hence, a characteristic of an MMP is that the compounds forming the pair are related to each other by a well-defined transformation. Identified molecular transformations were then associated with potency changes and defined criteria were applied to identify chemical substitutions that frequently introduce activity cliffs or consistently produce only small to moderate changes across different compound classes and biological targets. Furthermore, potency-retaining structural modifications, herein referred to as *bioisosteric replacements*, were also investigated at the level of individual target families.

SAR Analysis of Analogous Compound Series

In compound optimization during later stages of medicinal chemistry efforts, SAR exploration primarily aims at the analysis and design of analogs of active compounds with further improved properties. A central issue in lead optimization is the improvement of compound potency. SARs for analog series are traditionally studied using R-group tables that contain the core structure common to a series of analogs and rows displaying the substituents of individual compounds. Although user-friendly extensions of R-group tables have been developed, SAR features resulting from combinations of R-groups at multiple substitution sites cannot be analyzed in a straightforward and consistent manner. For this reason, we aimed to develop a visualization method that is specifically tailored towards a systematic exploration and intuitive interpretation of SAR features involving different R-groups and their combinations. To assess the potential of our newly designed data structure to uncover SAR rules, we inspected multiple data sets for the occurrence of predefined information-rich SAR patterns.

Another important theme in lead optimization is the study of molecular specifity of active compounds. Although many currently marketed drugs are known to interact with multiple targets [40], cross-reactivity can also lead to severe side effects and might be a major reason for the failure of many drugs in

Table 1.0-1: Research projects

methodology	projects	key aspects
machine-learning	1. orphan screening 2. potency-directed screening	LBVS, SVM, benchmarking
information theory	1. selection of compound class-specific descriptors	information entropic functions for molecular descriptor profiling
data mining	1. activity cliff profiling 2. potency effects of molecular transformations 3. SAR transfer	large-scale analyses of public domain compound data
visualization	SAR analysis in analog series: 1. substituent potency effects 2. multi-target SAR 3. SAR transfer	graphical methods, intuitive access to information-rich SAR patterns, derivation of SAR rules

clinical trials [41]. Therefore, rendering compounds with multi-target activity target-selective is a major goal of chemical optimization efforts and requires the systematic comparison of SARs for multiple target. However, the computational study of *multi-target SARs* (mtSARs) is still in its infancy. To these ends, we developed a methodological framework for the study of mtSARs that comprises a uniform R-group decomposition of analogs, their comparison on the basis of pharmacophore feature edit distances, and their organization in a previously reported hierachical structure [42] that reflects SAR discontinuity at substitution sites and their combinations. We then tested the approach for its ability to identify substitution sites that are selectivity determinants and to determine preference orders for chemical modifications to improve target selectivity.

In light of the objective of comprehensively improving compound properties in lead optimization before moving a drug candidate to later stages of drug discovery, it frequently happens that a compound series displaying an otherwise promising SAR fails to reach further development, perhaps due to inescapable metabolic or toxicological issues. In such situations, one would ideally like to build upon prior knowledge, utilize available SAR information, and evaluate the possibility of an "SAR transfer", i.e., the exploration of an alternative compound series that displays similar SAR characteristics and potency progression but lacks the liabilities associated with the original chemotype. SAR transfer is also interesting from the point of view that combining the SAR analyses of two chemically differently explored series with transferable SAR might reveal more information than their separate examination. However, a comprehensive literature search did not reveal computational methods available to aid in the identification of such parallel series. This motivated us to design a data mining approach that enables the identification of alternative analog series with

different core structures, corresponding substitution patterns, and comparable potency progression. We applied the methodology to search for alternative chemotypes for selected analog series and to systematically assess SAR transfer potential in a public compound repository.

Thesis Outline

This thesis is structured as follows. *Chapter 2* introduces fundamental concepts of molecular representations in chemoinformatics including the linear notation of molecular graphs, numerical chemical descriptors, and fingerprints. Furthermore, a standard similarity measure for the comparison of small molecule fingerprint representations is briefly described. The second part of the chapter discusses the increasing availability of public domain bioactivity data and gives a short overview about major compound databases that are of paramount importance for computer-aided drug discovery and chemoinformatics.

Chapter 3 reports ligand prediction for orphan targets using support vector machines. First, different approaches to orphan screening are summarized and an introduction into SVM theory is given. Then our study investigating the role of target information in finding ligands for orphan targets is reported in detail and implications of the results for practical applications are discussed.

Chapter 4 describes the adaptation of SVM strategies from orphan screening to potency-oriented LBVS. The design and results of our benchmark study on four HTS data sets are presented and the practical utility of potency-directed SVM strategies for early-stage drug discovery is discussed in light of our findings.

Chapter 5 addresses the selection of descriptors capturing compound class-specific information. Concepts from information theory are presented and their potential to identify discriminatory descriptors is compared on a large array of activity classes. Furthermore, representative examples are chosen to illustrate benefits and shortcomings of individual methods in detail.

Chapter 6 investigates the occurrence of activity cliffs in public domain compound data. Special emphasis is put on activity cliff distributions in different target families and a frequency analysis of multi-target activity cliffs.

Chapter 7 introduces the concept of matched molecular pairs and highlights its utility as a consistent reference framework for the identification of molecular transformations. A public domain compound repository is systematically searched for molecular transformations that are related to resulting potency changes. Sets of chemical replacements with a high potential to introduce activity cliffs or produce compounds with similar potency levels are identified.

Chapter 8 reports a newly designed graph structure representing entire series of analogs in a consistent manner. Graph components are designed to represent well-defined SAR patterns and provide immediate access to activity-

relevant substitution sites and R-group combinations. Exemplary analyses of different analog series illustrate how SAR determinants are identified on the basis of interactive graphical analysis and how such insights can be utilized to design new analogs.

Chapter 9 discusses the adaptation of a previously introduced hierarchical tree-like graph structure for a parallel SAR analysis of multiple targets. Two exemplary applications are reported that illustrate the ability of the approach to derive simple rules for the design of substitutions that are likely to yield target-selective compounds.

Chapter 10 presents a novel computational approach to study the transfer of SAR information from one chemical series to another. The new methodology can be applied to search for alternative analog series if one series is known or, alternatively, to systematically assess SAR transfer potential in compound databases. For both settings, exemplary applications are reported. Possibilities to further rationalize the process of SAR transfer are outlined.

Chapter 11 summarizes major findings and general conclusions of the work presented in this dissertation.

Chapter 2

Molecular Representations and Compound Databases

This chapter discusses basic concepts of molecular representations and similarity analysis, which are recurrent themes throughout this thesis, and reports on public domain repositories of compound structures and activity data as indispensable tools for pharmaceutical research in academic environments.

The analysis of structure-activity relationships requires defined representations of compounds that can be related to their biological activity. Furthermore, in many instances, the extraction of SAR information from compound data sets frequently relies on pairwise structural comparisons of small molecules and, hence, on the application of similarity measures that assess the degree of structural relatedness between compounds. In this chapter, frequently employed representations for small molecules are introduced in section 2.1, with a focus on fingerprint representations and their similarity assessment. Because publicly available bioactivity data provides essential resources for structure-activity data mining and the evaluation of chemoinformatics and drug design methods, section 2.2 is dedicated to the description of different categories of freely accessible compound databases.

2.1 Molecular Representations

Probably the most well-known description of compounds is the molecular graph representation. In a chemical graph, the atoms composing a molecule are represented as nodes and bonded atoms are connected by edges indicating the type of bond, e.g., a single, double, or aromatic bond. A graph does not represent three-dimensional structure information but the topology of a molecule. The graph can also be annotated with stereochemical information, which defines the relative spatial arrangements of selected atoms. To process and store molecular graphs on a computer, they are often encoded as connection tables

that are composed of an atom list and a bond list. In the atom list, the atoms of the molecules are provided in a sequential order, whereas the bond list specifies connections between pairs of atoms.

2.1.1 Linear Notation

A more compact and readable description of molecular graphs are linear notations, such as the popular *Simplified Molecular Input Line System* (SMILES) [43], which capture the structure of a molecule in form of an unambiguous text string using alphanumeric characters. They allow the efficient storage and fast processing of large numbers of molecules. The SMILES language uses the following basic rules for encoding molecules [43]:

1. Atoms are represented by their atomic symbols. Hydrogen atoms saturating free valences are not explicitly denoted.

2. Neighboring atoms stand next to each other and bonds are characterized as being single (-), double (=), triple (#), or aromatic (:). Single and aromatic bonds are usually omitted.

3. Enclosures in parentheses specify branches in the molecular structure.

4. For the linear representation of cyclic structures, a bond is broken in each ring and the connecting ring atoms are followed by the same digit in the textual representation.

5. Atoms in aromatic rings are indicated by lower case letters.

Figure 2.1-1 uses intuitive examples to illustrate the general concepts of the SMILES language. Although SMILES strings are unambiguous in the description of chemical structures, they are not unique because multiple valid SMILES representations exist for the same molecular graph, as also illustrated in Figure 2.1-1. To facilitate the comparison of molecular structures by SMILES representations and enable a fast check for molecular identity, algorithms have been introduced that ensure that the same SMILES string, also termed *canonical SMILES*, is always calculated for a given molecular graph [44]. Canonical SMILES strings are often used to ensure uniqueness of molecules in a database.

2.1.2 Molecular Descriptors

The majority of chemoinformatics methods rely on the representation of molecular structure and properties by numerical descriptors. Such descriptors are suitable as input for statistical and data mining methods. Accordingly, property descriptors are frequently employed in diversity analysis, representative

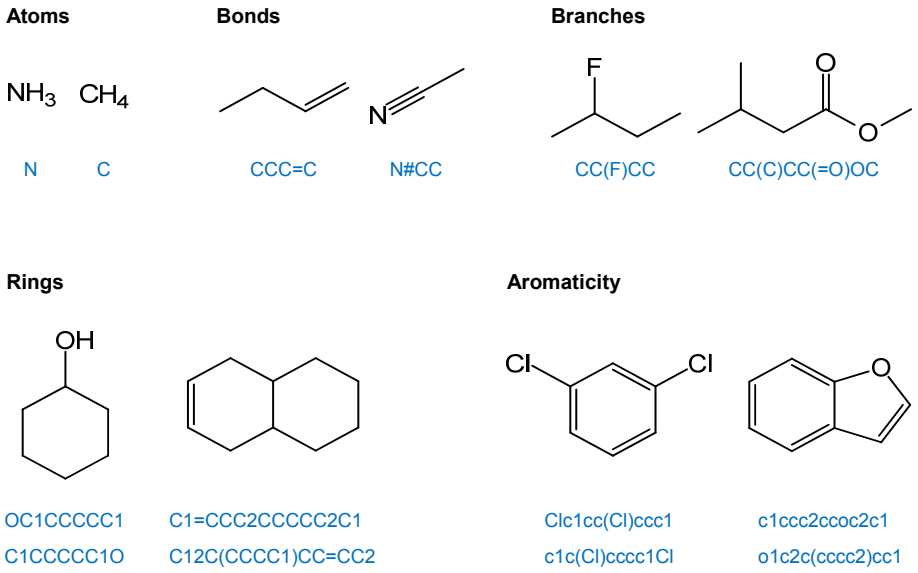


Figure 2.1-1: SMILES concepts Examples for the illustration of basic SMILES syntax rules are provided. Each molecular structure is annotated with one or multiple valid SMILES strings.

compound subset selection, combinatorial library design, and QSAR investigations.

Literally thousands of different molecular descriptors of greatly varying mathematical complexity are available [45]. Molecular descriptors are often classified as one-, two-, or three-dimensional (1D, 2D, or 3D), depending on the molecular representation from which they are calculated [3]. 1D descriptors are derived from the chemical formula. Examples include bulk properties, such as molecular weight, or simple atom counts. 2D descriptors are typically derived from a molecular graph representation and comprise, for example, topological descriptors or computational approximations to experimental measurements, such as solubilities or dipole moments. 3D descriptors are based on molecular conformations and often account for molecular surfaces, volumes, or surface-derived properties.

Typically, combinations of many descriptors are calculated for molecular data sets that then constitute chemical reference spaces where each descriptor adds a dimension to the space. In this reference space, the position of a molecule is determined by its descriptor values that serve as coordinates and similarity between molecules is defined by their spatial proximity that can be calculated by various measures [6]. Many descriptor spaces that are utilized in chemoinformatics are high-dimensional. However, for applications such as compound classification or QSAR, the dimensionality of chemical reference spaces is often reduced in order to focus on features that are most descriptive – and predic-

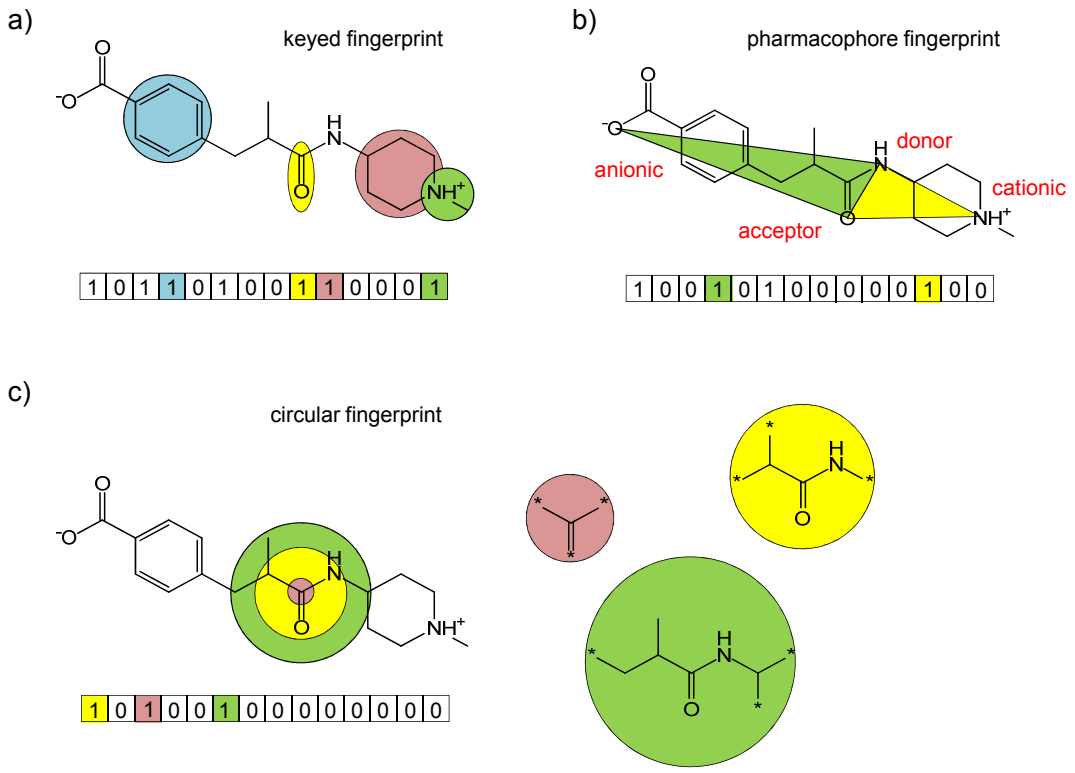


Figure 2.1-2: Fingerprints (a) In a keyed structural fingerprint, each bit accounts for the presence of a defined structural fragment. Here, bit positions encoding an aromatic ring, a carbonyl group, a heteroatom-containing ring, and a nitrogen attached to three carbon atoms are highlighted. (b) In a pharmacophore fingerprint, each bit accounts for the spatial arrangement of a defined atomic feature combination. Here, bits that account for defined geometric arrangements of the combinations "donor - acceptor - anionic" and "donor - acceptor - cationic" are highlighted. (c) In circular fingerprints, local atom environments are mapped to individual bit positions. Around a selected atom, three circular layers up to a diameter of four bonds are drawn, and the extracted substructural fragments are shown on the right. Attachment points constituting non-hydrogen atoms are marked by an asterisk.

tive – for a given data set and to provide a basis for chemical interpretation of the results.

2.1.3 Fingerprints

Fingerprints are a special form of a complex descriptor capturing feature distributions as bit string representations [46]. Apart from a general classification into 2D and 3D fingerprints in analogy to the categorization for molecular descriptors, four broad classes of fingerprints are distinguished: keyed and path-based fingerprints, pharmacophore-based fingerprints, binary circular fingerprints, and circular fingerprints considering counts [47].

Keyed fingerprints consist of a fixed number of bits where each bit accounts for the presence (i.e., the bit is set to one) or absence (the bit is set to zero) of a predefined structural fragment, as shown in Figure 2.1-2a. A popular example is the publicly available set of 166 MDL structural keys, also known as MACCS keys.[1] Path-based fingerprints extract all unique linear fragments up to a prespecified atom number from a molecular graph that are then projected into a fixed-sized bit vector. Despite their conceptual differences, keyed and path-based fingerprints were found to produce similar compound rankings in similarity searching [47].

Figure 2.1-2b illustrates the design principle of a pharmacophore fingerprint where each bit monitors the presence of a predefined geometrical arrangement of atomic features. To encode a molecule in this bit string format, each atom is assigned to one of multiple predefined pharmacophore features (e.g., hydrogen (H-)bond acceptor or donor) and then all possible combinations of (usually two to four) pharmacophore features and their (binned) pairwise atom distances are recorded. In 2D pharmacophore fingerprint design, atom distances are usually calculated as graph distances (i.e., the number of bonds in the shortest path between two atoms) whereas 3D pharmacophore fingerprints calculate atom distances using 3D atomic coordinates. For all spatial feature arrangements present in a molecule, the relevant bits are set to one. The typed graph distance (TGD) fingerprint implemented in the chemical computing software Molecular Operating Environment (MOE) represents a two-point pharmacophore-type fingerprint that is calculated from the 2D molecular graph. It is composed of 420 bits accounting for 15 binned distances between atom pairs in a molecule, with each atom assigned to one of seven possible features (“anion”, “cation”, “donor”, “acceptor”, “hydrophobe”, “polar”, “none of the aforementioned”).

Circular fingerprints are designed to capture local atom environments. Around each atom in a molecule, circular atom environments up to a specified bond distance range are calculated, and each resulting structural fragment is assigned to a fixed position in the fingerprint. Whereas binary circular fingerprints encode only the presence or absence of specific atom environments, circular fingerprints using counts also consider the frequencies of occurrence of these environments. Extended-connectivity fingerprints (ECFPs) [48] implemented in the scientific software Pipeline Pilot are currently the most popular binary circular fingerprints. ECFP4 captures three different substructures with bond diameters zero, two, and four around each atom of a molecule, as schematically depicted in Figure 2.1-2c.

Fingerprint comparisons between two compounds are often used to quantify their molecular similarity. A variety of different similarity coefficients has been

[1] Symyx Software, San Ramon, CA, USA; URL: http://www.symyx.com/
Software (e.g., for the calculation of fingerprints) and databases used in this work are summarized in Appendix A.

introduced for this purpose, with the Tanimoto coefficient (Tc) being the most frequently employed similarity measure for binary fingerprints [46]. For two molecules A and B, the Tc calculates fingerprint overlap by

$$\mathrm{Tc}(A,B) = \frac{c}{a+b-c} \tag{2.1}$$

where a corresponds to the number of bits set on in the fingerprint representation of molecule A, b is the number of bits set on in the fingerprint representation of B, and c reports the number of bits set on in both fingerprints. It follows that the Tc yields values in the range from zero (minimal similarity) to one (maximal similarity). The Tc is frequently employed in similarity searching to assess similarities between reference and test (database) compounds.

2.2 Public Domain Compound Databases

Different from genome sequencing projects that are often carried out by large publicly funded consortia, most drug discovery-relevant compound data has been generated in proprietary pharmaceutical environments (although the amount of data originating from academic settings is currently on the rise). Consequently, such data have only been sparsely distributed in the public domain. Only recently, a few pharmaceutical companies have begun to release rather significant amounts of small molecule activity data [49]. However, limited compound data availability has been, and continues to be, a problematic issue for computer-aided drug discovery and chemoinformatics [50]. Only over the past few years have public domain data repositories evolved to the extent that compound data mining is beginning to yield discovery-relevant insights.

One can roughly distinguish between five categories of public domain compound repositories. For each category, prototypic databases are discussed.

1. First, there are collections of largely non-annotated drug-like molecules. The ZINC database [51] contains millions of compounds offered by chemical vendors and provides modeled 3D conformations of these molecules.

2. Other databases collect biological screening data. As a consequence of the "Molecular Libraries Initiative" of the US National Institutes of Health [52], the repository PubChem BioAssay [53] has become the major source of compound screening data.

3. In addition, there are databases compiling drug and clinical trials data including, first and foremost, the online resources of the US Federal Drug Administration[2] and also databases such as DrugBank [54].

[2] http://www.accessdata.fda.gov/scripts/cder/drugsatfda/

4. A variety of relatively small specialized databases have also been introduced that contain compound information for specific protein families including, for example, ligands of G protein-coupled receptors (GPCRs) [55] or peptide-like and other small molecule inhibitors of proteases [56].

5. Finally, with BindingDB [57, 58] and ChEMBL [59], two large databases have been introduced that contain activity measurements and target annotations for compounds from different stages of medicinal chemistry programs. In general, the information contained in current ligand-target databases is extracted from the medicinal chemistry literature and patent resources. BindingDB and ChEMBL contain different types of affinity data (i.e., inhibitor dissociation constants (K_i), half maximal inhibitory concentrations (IC_{50}), half maximal effective concentrations (EC_{50}), etc.) for a broad spectrum of target classes, with inhibitory/antagonistic potency data for kinases, proteases, and GPCRs being most abundant in both databases. Importantly, these databases are not only a source of drug discovery-relevant information, but also research tools for method development in computational medicinal chemistry and chemoinformatics. The scientific activities of many academic research groups critically depend on the public availability of such data and, as detailed later, compound data sets extracted from BindingDB and ChEMBL have been abundantly used throughout this thesis.

An important caveat that should be considered when using data from public compound databases is that molecular representations for compounds from different sources are most likely not consistent. Therefore, all compound data sets used in projects of this dissertation must be standardized using tools provided in MOE or Pipeline Pilot. The standardization process includes, for example, the removal of salts and the protonation of strong bases and deprotonation of strong acids. Although such modifications might potentially alter activity determinants in molecules, this process is indispensable to ensure uniformness of molecular representations.

Source Information

Sections of the text in this chapter have been taken from [60–62].

Chapter 3

Ligand Prediction for Orphan Targets

Traditionally, drug discovery research has a strong single-target focus and lead optimization efforts predominantly aim at the design of target-selective small molecules for therapeutic intervention [63]. However, evidence is accumulating that many pharmaceutically relevant compounds do not, as originally thought, engage in specific single-target interactions but act on multiple targets to elicit their biological effects [64, 65]. The beneficial effect of multi-target inhibition can be explained by the finding that single-point modifications of biological systems are often compensated by other mechanisms [66], which might make single-target drugs less effective than one might anticipate based on their high potency values measured in vitro. In light of these observations, the analysis of compound activity against biological networks becomes increasingly important and is reflected by the emerging trend in drug discovery to depart from the single-target focus and test compounds against multiple targets. This approach is related to *chemogenomics* [67] that is commonly understood as the systematic exploration of interactions between all therapeutic protein targets and possible small molecule drugs.

As alternatives to brute-force compound screening, computational methods have increasingly been investigated to aid in this process. For example, target fishing aims at the identification of all likely targets for a small molecule and involves profiling of compounds against arrays of target-directed computational models such as Bayesian classifiers [68]. Other studies have employed a more integrated view on ligand-target interactions. For example, Schuffenhauer et al. have introduced "homology-based" similarity searching that investigates how similarity searching can be utilized to identify ligands interacting with the same target and also ligands binding to homologous targets [34]. Such studies take into account that homologous proteins usually have similar binding sites and therefore also structurally related ligands. This concept is particularly attractive

for the identification of ligands for orphan targets, i.e., targets for which no ligand information is available, which was traditionally beyond the scope of LBVS methods. Furthermore, homology-based similarity searching has recently been complemented by the adaptation of machine-learning methods to orphan screening.

In particular, *multi-task learning using support vector machines* (SVMs) has produced promising results in simulated orphan screening trials [35, 36, 69] and yielded better search performance than homology-based similarity searching in exhaustive benchmark calculations [35]. SVMs were originally developed for binary object classification. In a typical chemoinformatics SVM application, training compounds belonging to two different classes (e.g., active and inactive) are projected into chemical feature space and the SVM subsequently derives a hyperplane in this space to separate the two classes. Test compounds are classified based on which side of the hyperplane they fall. Alternatively, if the aim is a ranking of the test database, compounds are sorted by their signed distance to the hyperplane [33, 70]. Furthermore, the use of *kernel functions* in SVM learning enables the classifier to derive a more complex (non-linear) decision boundary and generalize to cases in which the two classes are not linearly separable. Kernel functions can also be thought of as similarity functions that determine how training and test objects are compared in SVM learning and classification. Multi-task learning tries to improve the generalization performance of a classifier by learning multiple related tasks simultaneously while using a shared representation of the tasks. Multi-task learning is especially beneficial in comparison to independent training strategies if high commonalities exist among the learning tasks and training examples for individual tasks are rare. In the context of chemogenomics and orphan screening, the integration of multiple related tasks corresponds to the parallel learning of ligand characteristics for multiple targets with the similarity of targets defining the relatedness of the tasks. Learning in combined target-ligand space can be elegantly achieved by using 'true' and 'false' target-ligand pairs as training examples (instead of active and inactive compounds for a single target) and is facilitated through the design of target-ligand kernel functions that account for pairwise similarities of target-ligand combinations. A target-ligand kernel is frequently calculated as the product of separate kernel functions for pairs of proteins and pairs of ligands. State-of-the art protein kernels used in ligand prediction include, for example, a sequence homology-based classification kernel [35]. Furthermore, among various ligand descriptors that can be used, 2D fingerprints have been found to be efficient small molecule representations for SVMs [70]. Fingerprint similarity can be captured, for example, by calculating the scalar product of bit vectors. However, many other types of kernel functions combining biological target and chemical ligand information can be envisioned, and one might expect that the

biological information content captured by the design of such kernel functions plays a major role for SVM-based ligand prediction.

To investigate the role of kernel functions for the prediction of ligands for orphan targets, we implemented a variety of target-ligand kernels with a particular focus on target kernels capturing different types of target information including sequence, secondary structure, tertiary structure, biophysical properties, ontologies, or structural taxonomy. Using two different SVM strategies that learn from multiple targets, these kernels were tested in ligand predictions for simulated orphan targets in two target protein systems characterized by the presence of different inter-target relationships [71]. For comparison, we also implemented a standard SVM trained on compounds active or inactive against the protein that is most closely related to the orphan target. The methodological background, design and results of our study are reported in this chapter. Section 3.1 gives an introduction to SVM theory. In section 3.2, specific SVM adaptations to learning from multiple targets are discussed and the three different SVM strategies that we used for orphan screening are presented. Section 3.3 reports the design of different kernel functions that are the focus of investigation. Section 3.4 presents the study design including the assembly of appropriate test systems and search calculations that were carried out for testing the various kernel functions in combination with our three different SVM strategies. Results of our study are reported in section 3.5 and conclusions and general implications of the results for practical orphan screening are discussed in section 3.6.

3.1 Support Vector Machine Theory

The term support vector machine refers to a supervised machine learning technique [32, 72]. A computational model is built based on a training set to associate class labels of objects with feature vectors. SVM learning for the purpose of virtual compound screening makes use of training examples $\{\mathbf{x}_i, y_i\}$ $(i = 1, \ldots, n)$ with $\mathbf{x}_i \in R^d$ being the feature vector (fingerprint representation) and $y_i \in \{-1, +1\}$ the class label (positive or negative; active or inactive) of training compound i. An SVM derives the normal vector $\mathbf{w}$ (with Euclidean norm $||\mathbf{w}||$) and the scalar b (called bias) to define a hyperplane H that best separates positive from negative training examples:

$$H : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \tag{3.1}$$

where $\langle \cdot, \cdot \rangle$ defines a scalar product.

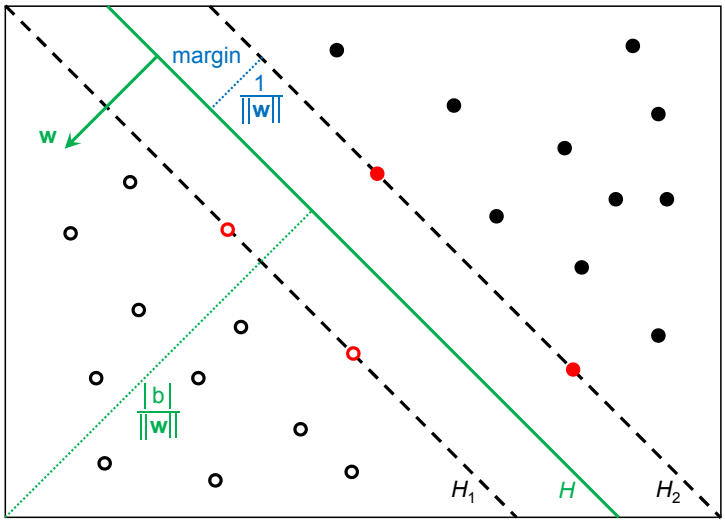


Figure 3.1-1: Maximum margin hyperplane Two classes are separated by the maximum margin hyperplane H that is defined by its normal vector $\mathbf{w}$ and its distance to the origin ($|b| \,/\, ||\mathbf{w}||$). Support vectors are highlighted in red and located on hyperplanes H_1 and H_2 parallel to the decision hyperplane H.

3.1.1 Classification in Original Feature Spaces

For linearly separable training data, an infinite number of hyperplanes exist to correctly classify the data. As shown in Figure 3.1-1, the hyperplane selected by the SVM is the one that maximizes the distance (called *margin*) from the nearest training data points, thereby minimizing the so-termed "structural risk" of overfitting the training data and enhancing the generalization performance of the classifier. Without loss of generality, the following inequality constraints to be met by the training data for correct classification can be formulated:

$$y_i \left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right) \geq +1 \quad \forall i \tag{3.2}$$

The molecules for which equality holds in equation 3.2 are nearest to the hyperplane H and are termed *support vectors*. The distance from H to the support vectors from the positive and the negative training class is $1/\,||\mathbf{w}||$, meaning that maximizing $1/\,||\mathbf{w}||$ or minimizing $||\mathbf{w}||$, given the conditions in equation 3.2, yields the maximum margin hyperplane. If a perfect linear separation of training examples is impossible, no solution for the optimization problem is found. To overcome this problem, violations of the strict constraints specified in equation 3.2 are permitted by the introduction of slack variables ξ_i, which allows training examples to be located within or on the incorrect side of the margin. The value of the slack variable ξ_i correlates with the degree of misclassification of the incorrectly positioned training compound i, as illustrated in Figure 3.1-2. A cost factor C is introduced for penalizing training errors and is adjustable to find a compromise between an optimal fit for the training data and the size of

the margin. The newly introduced parameters ξ_i and C lead to the following reformulation of the minimization problem:

$$\text{minimize:} \quad V(\mathbf{w}, \xi) = ||\mathbf{w}||^2 + C \sum_i \xi_i \tag{3.3}$$

$$\text{subject to:} \quad y_i \left(\langle \mathbf{x}_i, \mathbf{w} \rangle + b \right) \geq 1 - \xi_i \quad \text{with} \quad \xi_i \geq 0 \quad \forall i \tag{3.4}$$

Optimization problems under constraints can be solved by the introduction of Lagrange multipliers [72], which yields a convex quadratic programming problem amenable to standard methods:

$$\text{maximize:} \quad L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.5}$$

$$\text{subject to:} \quad \sum_i \alpha_i y_i = 0 \quad \text{with} \quad 0 \leq \alpha_i \leq C \quad \forall i \tag{3.6}$$

Solving the Lagrangian optimization problem results in the normal vector $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ with α_i being non-negative Lagrange multipliers. Since only support vectors (i.e., those vectors falling on the edge, within, or on the incorrect side of the margin) are assigned factors α_i greater than zero, the position of the hyperplane is exclusively determined by these critical vectors, which facilitates computations in high-dimensional feature spaces. Once the optimization problem formulated in equations 3.5 and 3.6 has been solved and $\mathbf{w}$ and b have been deduced, a test molecule $\mathbf{x}$ is classified on the basis of the decision function

$$f(\mathbf{x}) = \text{sgn}\Big(\sum_{i=1}^{n} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \Big) \tag{3.7}$$

This means that compounds with $f(\mathbf{x}) = +1$ are assigned to the positive class and those with $f(\mathbf{x}) = -1$ to the negative class. Geometrically, the sign indicates on which side of the hyperplane a test molecule falls.

3.1.2 Classification in Transformed Feature Spaces

In many cases, a planar surface might not be capable of separating the data correctly. To solve this problem, the data can be projected into a high-dimensional space $\mathcal{H}$, which might make a linear separation of the training data possible. If one assumes that the projection is accomplished using a mapping $\Phi : \mathbf{R}^d \rightarrow \mathcal{H}$, then the optimization problem specified in equation 3.5 requires the calculation of scalar products in $\mathcal{H}$, which is expressed by $\big\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \big\rangle$. However, the replacement of scalar products in $\mathcal{H}$ by suitable kernel functions, i.e., $K(\mathbf{x}_i, \mathbf{x}_j) = \big\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \big\rangle$, removes the need for an explicit formulation of the high-dimensional feature space $\mathcal{H}$, which is referred to as the kernel-trick [73]

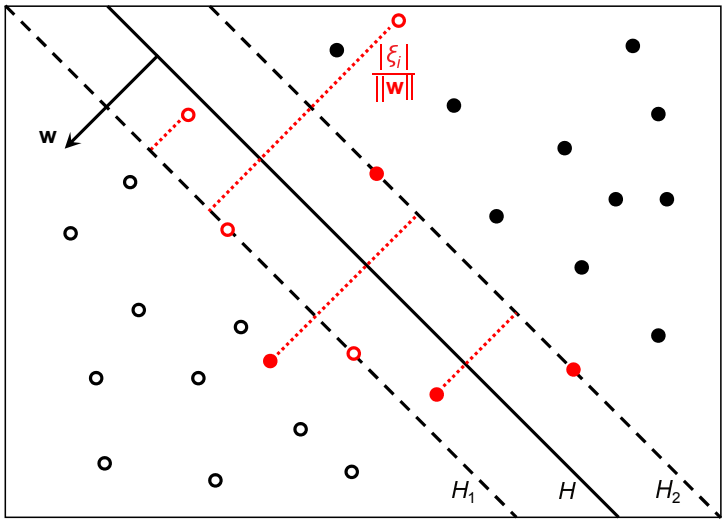


Figure 3.1-2: Hyperplane for imperfect separation Slack variables (ξ_i) permit the misclassification of some training compounds and correlate with their distance to the margin, as indicated by dashed lines in red.

and illustrated in Figure 3.1-3. The embedding function $\Phi(\mathbf{x})$ does not have to be known because the scalar product in the decision function $f(\mathbf{x})$ can also be replaced by the kernel function:

$$f(\mathbf{x}) = \operatorname{sgn}\Big(\sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x}_i, \mathbf{x}\right) + b\Big) \tag{3.8}$$

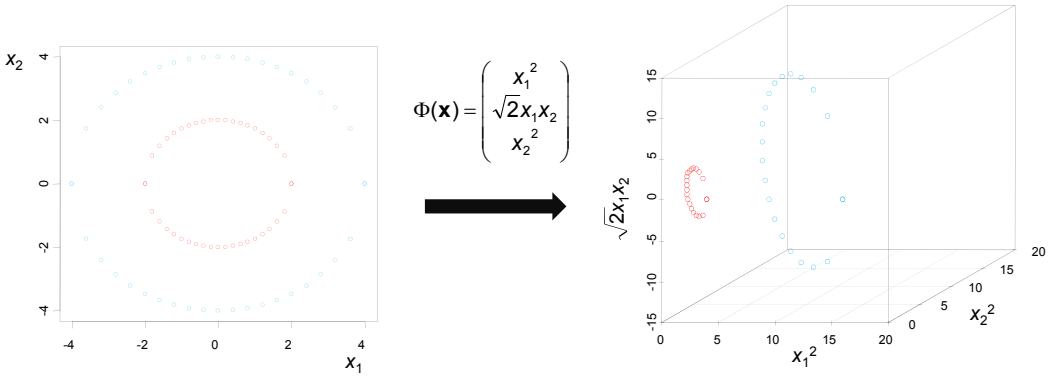
Frequently used kernels for the comparison of fingerprint representations, also termed ligand kernels $K_{\text{ligand}}(\cdot,\cdot)$ in the following, are the linear kernel (i.e., the standard scalar product), the Tanimoto kernel [74] encoding the Tanimoto coefficient as a kernel function, and the Gaussian or radial basis function kernel [75].

$$K_{\text{linear}}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.9}$$

$$K_{\text{Tanimoto}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - \langle \mathbf{x}_i, \mathbf{x}_j \rangle} \tag{3.10}$$

$$K_{\text{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma |\mathbf{x}_i - \mathbf{x}_j|^2\right) \tag{3.11}$$

The Gaussian kernel depends on the parameter γ (also called width) that implicitly regulates the number of support vectors defining the maximum margin hyperplane. The number of support vectors increases with an increasing value of γ and raises the risk of overfitting the training data. On the other hand, selecting a too small value for γ results in a very smooth decision boundary that does not classify the training data with sufficient accuracy [33]. Hence, similarly to the cost factor C, the choice of an appropriate parameter value is critical to SVM classification performance.



$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle = \left\langle \left(x_1^2, \sqrt{2}x_1x_2, x_2^2 \right), \left(z_1^2, \sqrt{2}z_1z_2, z_2^2 \right) \right\rangle = x_1^2 z_1^2 + 2x_1x_2z_1z_2 + x_2^2 z_2^2 = (x_1z_1 + x_2z_2)^2 = \langle \mathbf{x}, \mathbf{z} \rangle^2$$

Figure 3.1-3: Kernel trick Data points are projected into a higher-dimensional space by using the reported embedding function $\Phi(\mathbf{x})$. The transformation makes the two classes linearly separable. For the determination of the decision boundary, scalar products must be calculated in the transformed feature space. However, the explicit transformation of data points and subsequent calculation of scalar products in the higher-dimensional space can be avoided because $\langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle$ corresponds to and can be replaced by the polynomial kernel $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^2$.

3.1.3 From Classification to Ranking

To adapt SVM to virtual compound screening, the transformation of the classification into a ranking function is highly desirable. This can be achieved in a straightforward manner by defining the rank of a test molecule $\mathbf{x}$ according to the distance from its projection $\Phi(\mathbf{x})$ to the separating hyperplane determined in $\mathcal{H}$. Thus, test molecules are ranked from the most distant compound on the positive half-space to the most distant compound on the negative half-space. This ranking methodology corresponds to removing the signum function in equation 3.8 and sorting molecules in decreasing order of

$$g(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) \tag{3.12}$$

As a constant term, the bias b can be removed from the ranking function.

3.2 Chemogenomics-Oriented SVM Strategies

In the following, two different SVM strategies that learn from multiple targets and have previously been applied to orphan screening are briefly explained: *SVM using target-ligand kernel* (SVM TLK) and *SVM linear combination* (SVM LC). The two strategies are conceptually different since SVM TLK is a prime example for multi-task learning (vide supra), whereas SVM LC does

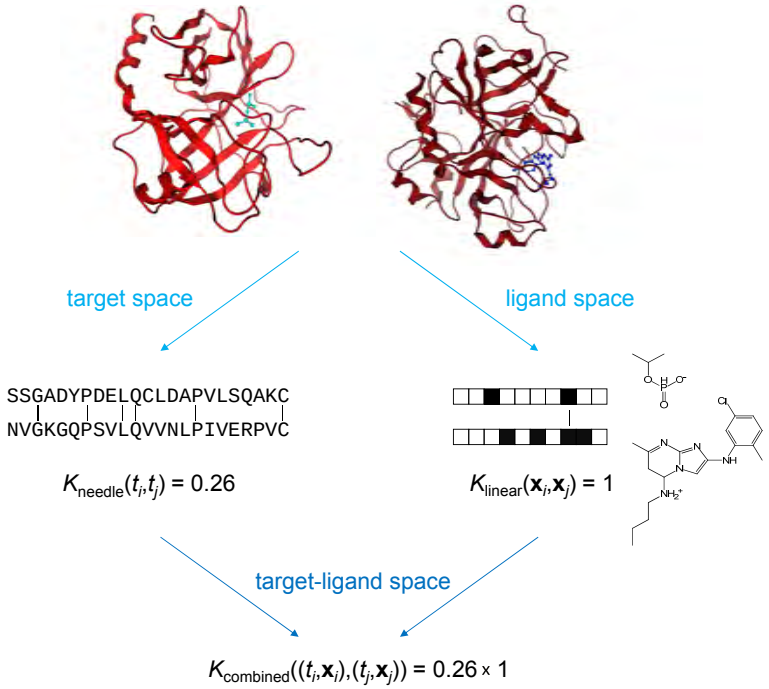


Figure 3.2-1: Target-ligand kernel The comparison of two target-ligand pairs via a target-ligand kernel function is divided into two independent tasks. In this case, the similarity of protein targets is quantified by sequence comparison, while ligand similarity is assessed through comparison of fingerprint representations. The product of the two similarity scores is taken to recombine target and ligand information. The figure is adapted from [71].

not learn from multiple training sets simultaneously but builds separate classification models that are subsequently integrated into a single ranking function. Both methods were found to produce similar results in virtual screening trials that aimed at enriching small database selection sets with ligands for simulated orphan targets [35]. Furthermore, a much simpler approach to orphan screening using a standard binary SVM is introduced.

3.2.1 SVM with Target-Ligand Kernel

In recent chemogenomics-oriented studies [35–37], not only compounds but also true and false protein-compound pairs were used to train SVMs in order to enable learning and classification for multiple targets in parallel. For doing so, scalar products between target-ligand pairs occurring in the objective function of the SVM optimization problem (equation 3.5) and the decision function (equation 3.7) are replaced by suitable kernel functions accounting for the similarity of protein-small molecule pairs in combined target-ligand space. Analogously to the mapping of molecules into high-dimensional descriptor spaces,

a target-ligand pair $(t_i, \mathbf{x}_i)$ is projected into a high-dimensional target-ligand space using an implicit embedding function $\Phi(t_i, \mathbf{x}_i)$ given by the kernel function

$$K_{\text{target-ligand}}\big((t_i, \mathbf{x}_i), (t_j, \mathbf{x}_j)\big) = \big\langle \Phi(t_i, \mathbf{x}_i), \Phi(t_j, \mathbf{x}_j) \big\rangle \tag{3.13}$$

To capture interactions between features of a molecule (represented by feature vector $\Phi_{\text{ligand}}(\mathbf{x}_i)$ in high-dimensional ligand space) and features of its protein target (represented by $\Phi_{\text{target}}(t_i)$ in high-dimensional target space), it was suggested to represent the target-ligand pair by the set of all possible products of features of the target and the ligand, which corresponds to the tensor product of their feature vectors [36]:

$$\Phi(t_i, \mathbf{x}_i) = \Phi_{\text{target}}(t_i) \otimes \Phi_{\text{ligand}}(\mathbf{x}_i) \tag{3.14}$$

Inserting this definition for $\Phi(t, \mathbf{x})$ into equation 3.13 yields

$$K_{\text{target-ligand}}\big((t_i, \mathbf{x}_i), (t_j, \mathbf{x}_j)\big) = \big\langle \Phi_{\text{target}}(t_i) \otimes \Phi_{\text{ligand}}(\mathbf{x}_i), \Phi_{\text{target}}(t_j) \otimes \Phi_{\text{ligand}}(\mathbf{x}_j) \big\rangle \tag{3.15}$$

The application of linear algebra leads to the following factorization of the scalar product between two tensor product vectors:

$$\big\langle \Phi_{\text{target}}(t_i) \otimes \Phi_{\text{ligand}}(\mathbf{x}_i), \Phi_{\text{target}}(t_j) \otimes \Phi_{\text{ligand}}(\mathbf{x}_j) \big\rangle = \big\langle \Phi_{\text{target}}(t_i), \Phi_{\text{target}}(t_j) \big\rangle \times \big\langle \Phi_{\text{ligand}}(\mathbf{x}_i), \Phi_{\text{ligand}}(\mathbf{x}_j) \big\rangle \tag{3.16}$$

Equation 3.16 shows that the comparison of two target-ligand pairs can be reduced to the separate assessment of target and ligand similarities in target and ligand space, respectively, which also avoids the computationally prohibitive calculation of tensor product vectors. The use of kernels functions further reduces the complexity of the problem and leads to equation 3.17, generally defining the target-ligand kernel as the product of two separate kernels for the target pair and the ligand pair:

$$K_{\text{target-ligand}}\big((t_i, \mathbf{x}_i), (t_j, \mathbf{x}_j)\big) = K_{\text{target}}(t_i, t_j) \times K_{\text{ligand}}(\mathbf{x}_i, \mathbf{x}_j) \tag{3.17}$$

The design principle of target-ligand kernels is illustrated in Figure 3.2-1. Independent kernels for protein and ligand representations are used to account for pairwise target and ligand similarities and combined to calculate the similarity of the protein-molecule pairs in target-ligand space.

For orphan screening, a model is built using true and false target-ligand pairs. The false target-ligand pairs are usually derived by combining the same targets that are found in the known target-ligand pairs with randomly selected compounds from the screening database, assuming that most database compounds are inactive. All test compounds are then combined with the orphan

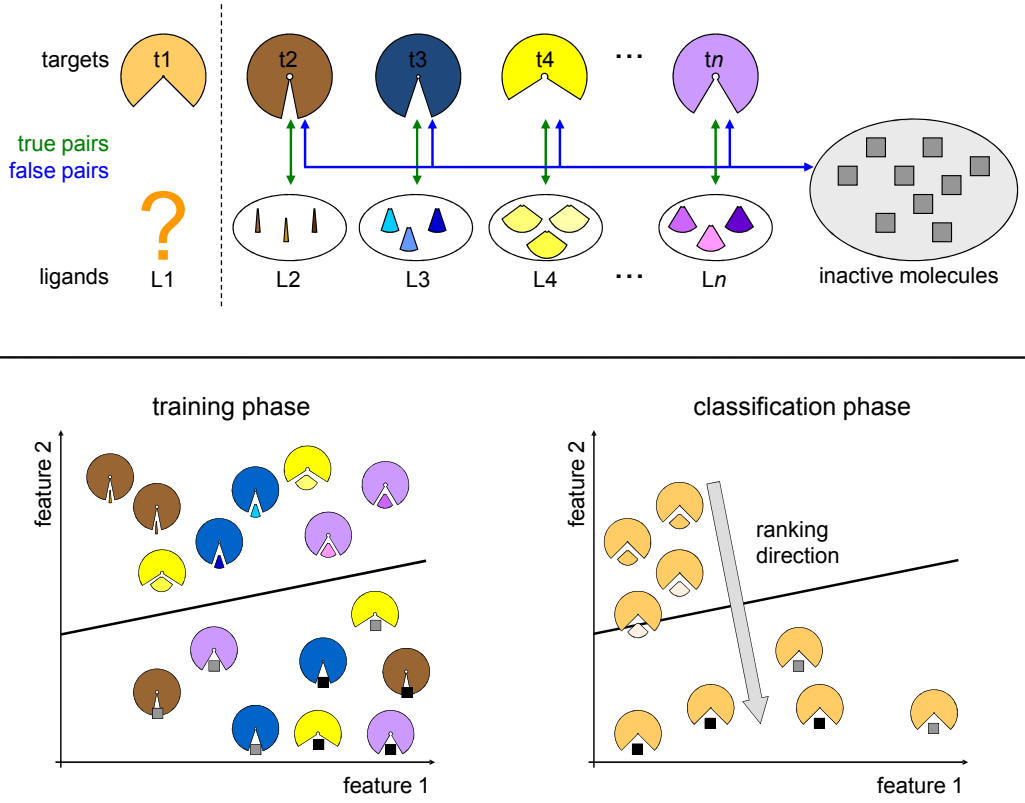


Figure 3.2-2: Orphan screening with SVM TLK Targets in the reference set are paired with their known ligands and presumably inactive compounds taken from the screening database to build positive and negative training examples, respectively, for SVM learning. A decision boundary is derived in combined target-ligand space. Test compounds are subsequently paired with the orphan target, projected into target-ligand space, and sorted by their signed distance to the hyperplane.

target and classified by the derived decision function. Analogously to the transformation applied to conventional small molecule classifiers, a ranking of ligands for an orphan target can be obtained by sorting the orphan target-ligand pairs according to their signed distance from the maximum margin hyperplane (see equation 3.12). Orphan screening using SVM TLK is illustrated in Figure 3.2-2.

3.2.2 SVM Linear Combination

SVM LC was introduced to learn a model for a particular target t_j, which sets it apart from the SVM TLK approach that learns to generally classify target-ligand pairs as true or false. Nevertheless, SVM LC is distinct from conventional small molecule classifiers in that it makes use of ligand sets with different bioactivities for learning. For each target t_i in the training set, an individual normal vector $\mathbf{w}_i$ is calculated by learning a binary SVM classifica-

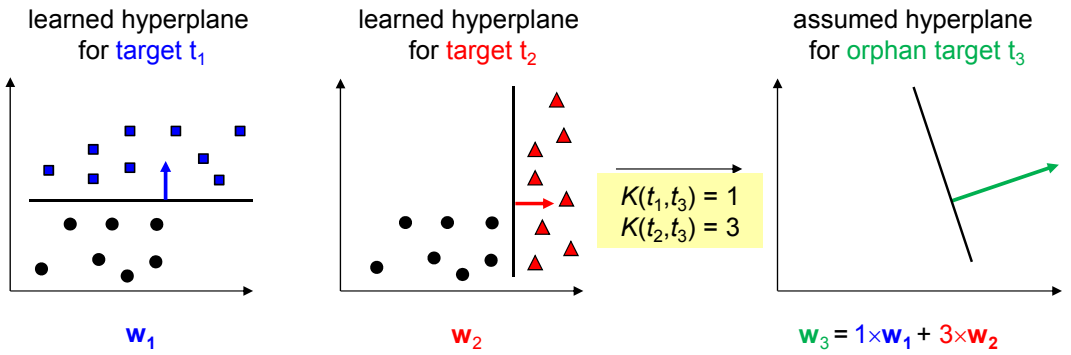


Figure 3.2-3: Orphan screening with SVM LC For each target in the reference set, an individual SVM model is built. The normal vectors of the decision hyperplanes are then linearly combined to yield a final SVM model for sorting the test database and prioritizing molecules for the orphan target. The linear factors for the individual normal vectors are determined by the similarity of the corresponding reference target to the orphan target, as measured by a target kernel function. The figure is adapted from [76].

tion function for its ligand set L_i (that is used as positive class and combined with a set of randomly selected database molecules as negative class). The vector $\mathbf{w}_i$ corresponds to the normal vector of the maximum margin hyperplane $H_i = \left\{\mathbf{x} | \left\langle \mathbf{w}_i, \Phi_{\text{ligand}}(\mathbf{x}) \right\rangle + b = 0\right\}$. For the target of interest t_j, a final normal vector $\mathbf{w}_j^{\text{final}}$ is built by linearly combining the individual $\mathbf{w}_i$:

$$\mathbf{w}_j^{\text{final}} = \sum_i s(t_i, t_j)\mathbf{w}_i \tag{3.18}$$

where a similarity function $s(t_i, t_j)$ is used to determine linear factors for individual normal vectors. Hence, the higher the similarity between targets t_i and t_j, the more contributes the normal vector $\mathbf{w}_i$ to the final vector $\mathbf{w}_j^{\text{final}}$. If ligands for t_j are available, $\mathbf{w}_j$ usually contributes most to the model. If t_j is an orphan target, then $\mathbf{w}_j^{\text{final}}$ does not contain the term $\mathbf{w}_j$, as exemplarily shown in Figure 3.2-3 for an orphan target t_3 and two reference targets t_1 and t_2. The similarity function $s(t_i, t_j)$ can also be thought of as a target kernel function so that

$$\mathbf{w}_j = \sum_i K_{\text{target}}(t_i, t_j)\mathbf{w}_i \tag{3.19}$$

To obtain a compound ranking and prioritize molecules as potential ligands for target t_j, test molecules are sorted in descending order of $g(\mathbf{x}) = \left\langle \mathbf{w}_j^{\text{final}}, \Phi_{\text{ligand}}(\mathbf{x}) \right\rangle$.

3.2.3 Homology-Based SVM

A much less complex search strategy for orphan screening is *homology-based SVM*, which, as revealed by its name, incorporates design principles of homology-

based similarity searching. Among all available targets in the training set, the nearest neighbor, i.e., the target that is most closely related to the orphan target, is determined. Then the ligand set of the nearest neighbor is used as positive training class and an arbitrarily chosen subset of the screening database as negative class for deriving a standard SVM ranking function, as given by equation 3.12.

3.3 Target and Ligand Kernels

We aimed at a systematic investigation of different kernel functions in the three presented SVM strategies for orphan screening, with a particular focus on target kernel functions used by SVM TLK and SVM LC. Using fingerprints as small molecule representations, we considered the linear (equation 3.9) and Gaussian (equation 3.11) kernel as ligand kernels for our study since they had shown good search performance for fingerprints in previous virtual screening trials [33, 70, 77]. The ligand kernels were complemented by 11 different target kernels.

(a) *Uniform kernel* between two targets (t_i, t_j):

$$K_{\text{uniform}}(t_i, t_j) = 1 \tag{3.20}$$

In this case, differences between targets are not considered. For the TLK search strategy, using K_{uniform} corresponds to pooling training molecules for all proteins and deriving a standard SVM model on the pooled compounds.

(b) *Needle kernel* is the percentage sequence identity SI for a protein pair (t_i, t_j) computed using the Needleman-Wunsch algorithm for pairwise global sequence alignment implemented in EMBOSS [78].

$$K_{\text{needle}}(t_i, t_j) = \text{SI}(t_i, t_j) \tag{3.21}$$

(c) *Water kernel.* Each protein pair (t_i, t_j) is also subjected to pairwise local sequence alignment using the Smith-Waterman algorithm implemented in EMBOSS and the alignment scores $S_{\text{SW}}(t_i, t_j)$ are expressed in logarithmic form:

$$K_{\text{water}}(t_i, t_j) = \ln S_{\text{SW}}(t_i, t_j) \tag{3.22}$$

(d) *PROFEAT kernel.* The PROFEAT server [79] computes 1 447 protein descriptors from protein sequence including descriptors developed by Dubchak et al. [80] that account for the composition, transition, and distribution of structural and physicochemical properties, such as hydrophobicity, polarity, charge,

and solvent accessibility. Each descriptor is separately normalized to the value range [0,1] and each target t_i is represented by a vector $\Phi_P(t_i)$ of 1 447 normalized descriptor values. The PROFEAT kernel is then defined as

$$K_{\text{PROFEAT}}(t_i, t_j) = \langle \Phi_P(t_i), \Phi_P(t_j) \rangle \tag{3.23}$$

(e) *Spectrum kernel* is a string kernel introduced by Leslie et al. [81]. It compares sequence strings representing k-mers. Here conventional 3-mers were computed for target sequences. Each protein t_i is represented by a 20^3 dimensional vector $\Phi_S(t_i)$ (for 20 amino acids) where each dimension corresponds to a possible string of three amino acids and reports the count of the number of occurrences of this fragment in the sequence of t_i. To account for different lengths of protein sequences, the kernel is normalized as follows:

$$K_{\text{spectrum}}(t_i, t_j) = \frac{\langle \Phi_S(t_i), \Phi_S(t_j) \rangle}{\sqrt{\langle \Phi_S(t_i), \Phi_S(t_i) \rangle \langle \Phi_S(t_j), \Phi_S(t_j) \rangle}} \tag{3.24}$$

(f) *SSEA kernel.* For each target, the secondary structure is predicted by PSIPRED [82] resulting in a string of residues each represented by one of three letters for the states helix, strand, or coil. Strings for a target pair (t_i, t_j) are then globally aligned using the dynamic programming algorithm implemented in the SSEA web server [83], which yields a score $S_{\text{SSEA}}(t_i, t_j)$ in the range [0,100]. This score is directly used as target kernel:

$$K_{\text{SSEA}}(t_i, t_j) = S_{\text{SSEA}}(t_i, t_j) \tag{3.25}$$

(g) *GO kernel.* Gene Ontology (GO) [84] terms of the Molecular Function category are extracted for all protein targets from the UniProt Knowledgebase [85]. The GO kernel for a target pair (t_i, t_j) counts the number of identical GO terms in the GO term sets of t_i and t_j [86].

(h) *Cleavage kernel.* Peptidases act on specific substrates and their catalytic activity is often restricted to specific sequence recognition sites. For all targets, available cleavage sites of their substrates are extracted from the MEROPS [56] and CutDB [87] databases that collect cleavage sites in natural and synthetic substrates. Cleavage site patterns are reduced to two residues on either side of the scissile bond and for each target t_i, a position-specific frequency matrix is generated. The columns of the matrix are then concatenated to form a 4×20 dimensional feature vector $\Phi_C(t_i)$ and the cleavage kernel is calculated as follows:

$$K_{\text{cleavage}}(t_i, t_j) = \frac{\langle \Phi_C(t_i), \Phi_C(t_j) \rangle}{\sqrt{\langle \Phi_C(t_i), \Phi_C(t_i) \rangle \langle \Phi_C(t_j), \Phi_C(t_j) \rangle}} \tag{3.26}$$

(i) *SCOP kernel.* The SCOP database [88] is hierarchically structured into protein folds, superfamilies, families, and domains and can be represented as a directed acyclic graph (DAG). For a target t_i, $\Phi_{\mathrm{Sc}}(t_i)$ contains as many features as there are nodes in the graph and each feature is set to one if the corresponding node is part of t_i's SCOP hierarchy and zero otherwise. The SCOP kernel is then defined as follows

$$K_{\mathrm{SCOP}}(t_i, t_j) = 2^{\langle \Phi_{\mathrm{Sc}}(t_i), \Phi_{\mathrm{Sc}}(t_j) \rangle} \tag{3.27}$$

(j) *Topmatch kernel.* All protein targets are represented by a 3D substructure comprising all amino acids within an 8 Å radius of the target's catalytic residues. Residues falling within this radius are computed with MOE and then subjected to structure comparison using TopMatch-web [89]. For a target pair (t_i, t_j), TopMatch-web computes a relative similarity score S_T within the range [0,100] that is directly used as the target kernel:

$$K_{\mathrm{Topmatch}}(t_i, t_j) = S_T(t_i, t_j) \tag{3.28}$$

(k) *MEROPS kernel.* The MEROPS database [56] is hierarchically structured into catalytic types, so-called protein clans, families, and subfamilies and can also be visualized as a DAG. Hence, $\Phi_M(t_i)$ can be defined analogously to $\Phi_{\mathrm{Sc}}(t_i)$ and the MEROPS kernel is given by

$$K_{\mathrm{MEROPS}}(t_i, t_j) = 2^{\langle \Phi_M(t_i), \Phi_M(t_j) \rangle} \tag{3.29}$$

A note on the general validity of the kernel functions is found in Appendix B.

3.4 Data and Calculations

We applied the three different SVM strategies using the described kernel functions to search for inhibitors of individual proteases in two different target sets that were regarded as orphan targets and hence not included during SVM learning.

3.4.1 Target and Ligand Systems

Two sets of reference targets were assembled that represented different degrees of inter-target relationships. The first target set included 12 proteases belonging to nine different families (Table 3.4-1) and showing four different catalytic mechanisms: cathepsin D and renin are aspartate proteases; thrombin and trypsin are serine proteases; cathepsin L, calpain 2, and caspase 3 are cysteine proteases;

and matrix metalloproteases 2 and 8, methionyl aminopeptidase 2, glutamate carboxypeptidase 2, and angiotensin-converting enzyme 2 are metalloproteases. Proteases possessing the same catalytic machinery can either be closely or distantly related in sequence, as illustrated in Figure 3.4-1, which organizes targets into clans, families, and subfamilies following the classification scheme of the MEROPS peptidase database. Based on the MEROPS hierarchy, the nearest neighbor target for each protease in our test set was determined. If several nearest neighbor candidates were suggested for a given target based on the MEROPS hierarchy, the protease with highest sequence identity to the target was chosen. For all 12 targets, ligand sets were assembled from the MDL Drug Data Report (MDDR), a commercial database storing structural and activity data for biologically relevant compounds, BindingDB, and original literature sources. In total, 1 359 different protease inhibitors were collected. As reported in Table 3.4-1, each ligand set contained between 14 and 281 compounds having a potency of at least 1 µM (K_i or IC_{50}) against the target. Ligand sets were mutually exclusive in their composition, i.e., a compound reported to inhibit multiple protease targets was only assigned to the target it was most potent against.

The second target set included 11 proteases and was taken from [35]. These targets included the cysteine proteases calpain 1 and 2, caspase 1 and 3, and

Table 3.4-1: Target and ligand data set 1

target	abbr.	MEROPS ID	PDB entry	#ligands	NN target
angiotensin-converting enzyme 2	ace2	M02.006	1r42	28	mmp2
calpain 2	cal2	C02.002	1kfu	49	catL
caspase 3	cas3	C14.003	1cp3	264	catL
cathepsin D	catD	A01.009	1lyb	70	ren
cathepsin L	catL	C01.032	1mhw	78	cal2
glutamate carboxypeptidase 2	mgcp	M28.010	2oot	14	mmp8
methionyl aminopeptidase 2	metap2	M24.002	1b6a	254	mgcp
matrix metalloprotease 2	mmp2	M10.003	1qib	83	mmp8
matrix metalloprotease 8	mmp8	M10.002	1bzs	16	mmp2
renin	ren	A01.007	2ren	164	catD
thrombin	thr	S01.217	1ppb	281	try
trypsin	try	S01.127	1trn	58	thr

For each target protein, a target name abbreviation (abbr.), its MEROPS identifier (ID), a corresponding Protein Data Bank (PDB) [90] entry, the number of ligands (#ligands), and the nearest neighbor (NN) target are reported. The MEROPS identifier is composed of a family-based component (e.g., thrombin and trypsin both belong to the family S01) and an individual target-based component. The PDB entry refers to the 3D protein structure used in calculating the Topmatch kernel and is cross-linked to the SCOP entry used in calculating the SCOP kernel.

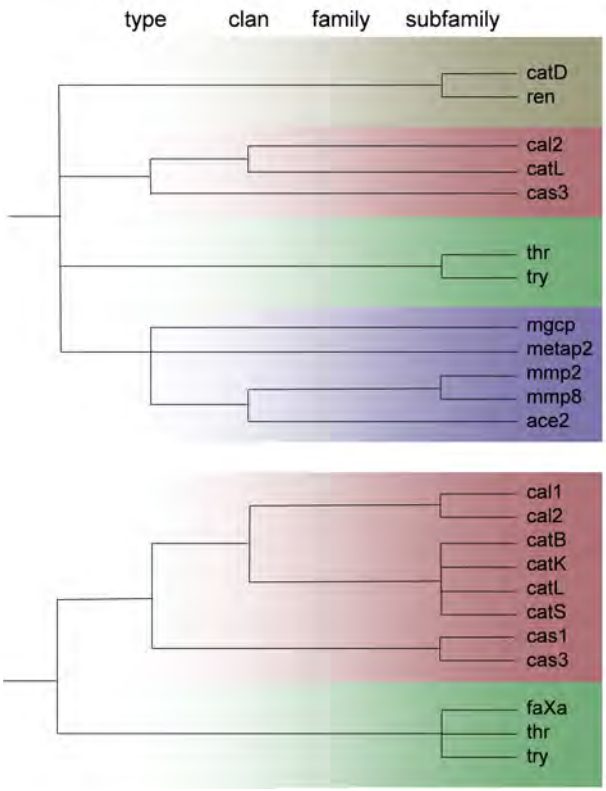


Figure 3.4-1: Target relationships The relationships between the proteases in the target sets 1 (top) and 2 (bottom) are illustrated. The MEROPS classification scheme (i.e., type, clan, family, and subfamily) is applied. From the "subfamily" to the "type" level, target similarity is fading away. The figure is adapted from [71].

cathepsin B, L, K, and S, and the serine proteases factor Xa, thrombin, and trypsin, as summarized in Table 3.4-2. Relationships between these targets are also illustrated in Figure 3.4-1. Ligand sets for these targets were assembled as described above. For proteases shared among both test systems (i.e., calpain 2, caspase 3, cathepsin L, thrombin, and trypsin), the same ligand sets were used. As reported in Figure 3.4-1, the inter-target and nearest neighbor relationships differed between target sets 1 and 2. Whereas each target in set 2 had a nearest neighbor that belonged to the same subfamily, several targets in set 1 had nearest neighbors sharing the same catalytic mechanism, but lacking further evidence of evolutionary relationships. The different inter-target relationships found in data sets 1 and 2 were explored in SVM modeling and ligand-target prediction.

3.4.2 Search Calculations

The performance of alternative kernel functions and SVM ranking strategies was evaluated in systematic search calculations on the two protease systems. All proteases of a system were in turn regarded as orphan targets and hence not included during SVM learning. The model was built on the remaining targets and their ligand sets in the system. As a background database for SVM analysis,

Table 3.4-2: Target and ligand data set 2

target	abbr.	MEROPS ID	PDB entry	#ligands	NN target
calpain 1	cal1	C02.001	1tlo	46	cal2
calpain 2	cal2	C02.002	1kfu	49	cal1
caspase 1	cas1	C14.001	1ice	21	cas3
caspase 3	cas3	C14.003	1gfw	264	cas1
cathepsin B	catB	C01.060	1gmy	17	catS
cathepsin K	catK	C01.036	1yk7	223	catS
cathepsin L	catL	C01.032	1mhw	78	catK
cathepsin S	catS	C01.034	1ms6	221	catK
factor Xa	faXa	S01.216	1mq5	783	thr
thrombin	thr	S01.217	1ppb	281	faXa
trypsin	try	S01.127	1trn	58	faXa

For each target protein, a target name abbreviation (abbr.), its MEROPS identifier (ID), a corresponding Protein Data Bank (PDB) entry, the number of ligands (#ligands), and the nearest neighbor (NN) target are reported.

100 000 compounds were randomly chosen from the ZINC database. MACCS structural keys and the TGD fingerprint were used as ligand descriptors.

For each reported combination of kernel function(s) and SVM strategy, ten different randomly selected training and test sets were studied for each simulated orphan target and the search results were averaged. As negative training examples, 1 000 database compounds were randomly selected in each case. For SVM TLK, the 1 000 compounds were combined with each reference target to build false target-ligand pairs, and for SVM LC, they served as negative training class for each individual model. For simple SVM calculations on each target, five inhibitors of the nearest neighbor target were used as positive training molecules, while for SVM TLK and LC five for each of the remaining targets in the protease set were used. The inhibitor set of the orphanized target was not used during SVM learning but added to the background database as potential database hits during testing. As a measure of performance, recovery rates (number of correctly identified orphan target inhibitors divided by their total number in the test database) were calculated for database selection sets of increasing size and averaged over the ten independent trials per target. All calculations were carried out using SVM^{light}, a freely available SVM implementation [91]. With exception of the parameter γ in the Gaussian kernel that was set to 0.01 after preliminary test calculations, all calculation parameters were SVM^{light} default settings to ensure reproducibility of the calculations. Perl scripts were applied to calculate SVM linear combinations and analyze the results.

Table 3.5-1: Search results for ligand prediction using homology-based SVM (set 1)

	MACCS				TGD			
	linear[a]		Gaussian[a]		linear[a]		Gaussian[a]	
	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]
ace2	0.0	0.9	0.0	0.9	0.9	14.8	0.4	17.4
cal2	19.1	60.2	22.1	63.4	4.8	30.7	5.0	30.2
cas3	2.6	9.5	2.4	10.0	0.6	8.5	0.2	5.4
catD	15.4	51.5	16.0	54.5	40.5	72.2	38.6	77.1
catL	10.4	34.7	10.0	35.1	8.0	23.6	6.2	25.2
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	0.0	0.5	0.0	0.2	0.0	0.0	0.0	0.0
mmp2	49.6	77.1	54.7	78.7	75.6	94.0	77.6	95.3
mmp8	49.1	59.1	50.9	59.1	57.3	67.3	57.3	68.2
ren	30.3	57.9	32.0	58.2	44.3	59.1	41.3	59.9
thr	27.7	70.3	28.0	71.3	28.9	80.5	29.7	81.1
try	36.9	61.4	38.5	61.2	46.9	74.4	46.2	78.7
average	20.1	40.3	21.2	41.0	25.6	43.8	25.2	44.9

[a] Kernel. [b] Set size. Recovery rates (in %) are reported for all targets in data set 1 averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1. The results reported for the Gaussian kernel were obtained with the parameter γ set to 0.01.

3.5 Results

The results of systematic search calculations are described in the following. First, global kernel performance was assessed. Then target- and set-specific differences in prediction rates for alternative target-ligand kernels and the dependence of successful ligand prediction on the nearest neighbor reference target of a simulated orphan target were investigated.

3.5.1 Global Kernel Performance

We first investigated the relative performance of ligand kernels in homology-based SVM calculations searching for active compounds of orphan targets. Table 3.5-1 reports compound recovery rates for activity classes of target set 1, the MACCS and TGD fingerprints, and database selection sets of 100 and 1000 compounds. In these calculations, both ligand kernels produced comparable recovery rates. In some instances, the search calculations failed for any kernel and in others, high recovery rates were consistently observed. Because there was no apparent preference for a ligand kernel in our test calculations, we selected the linear kernel that has lower computational complexity for further

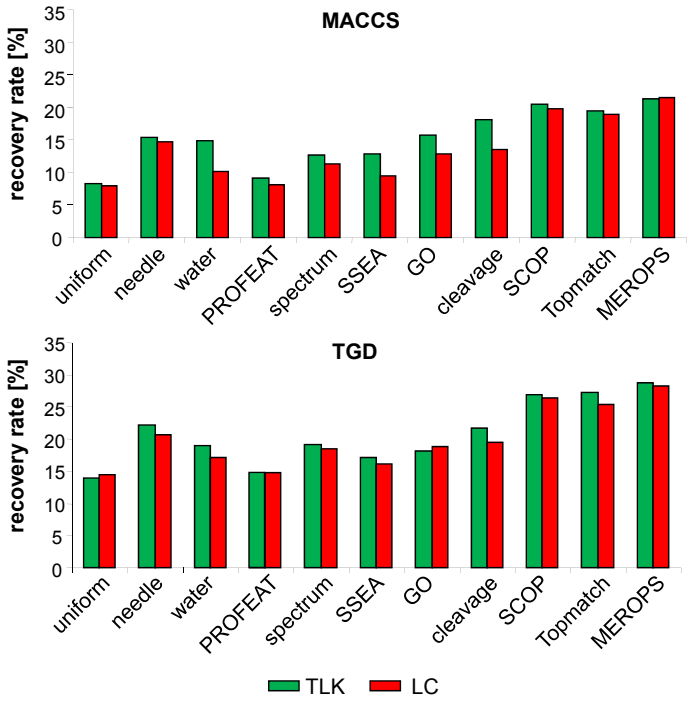


Figure 3.5-1: Global kernel performance (set 1) For the MACCS and TGD fingerprints, recovery rates for SVM TLK and LC strategies are shown for 11 alternative target kernels and selection sets of 100 database compounds. Recovery rates are averaged over all 12 targets and ten independent search trials per target. The figure is adapted from [71].

calculations and combined this kernel with the 11 different target kernels for SVM TLK and LC ligand prediction calculations.

For the 11 different target-ligand kernel combinations, average results for target set 1 and selection sets of 100 database compounds are shown in Figure 3.5-1 and Appendix Tables B-1 and B-2 report recovery rates on a per target basis. The SVM TLK and LC search strategies were found to produce similar compound recovery rates of approximately 8% to 28% for both fingerprints and all target kernels for a database selection set size of 100 compounds. By and large, there was relatively little variation in kernel performance, much less so than anticipated given the significant differences in target kernel complexity and encoded protein information. The PROFEAT kernel, which is based on biophysical descriptors calculated from protein sequence, did not produce higher recovery rates than the uniform kernel that does not take protein similarity into account and hence served as a reference for target kernels. Differences in kernel performance were rather subtle but the overall highest recovery rates were achieved with the MEROPS kernel that encodes a hierarchical protein organization scheme.

Equivalent observations were made for target set 2. Figure 3.5-2 reports average results of the search calculations on set 2 and Appendix Tables B-3

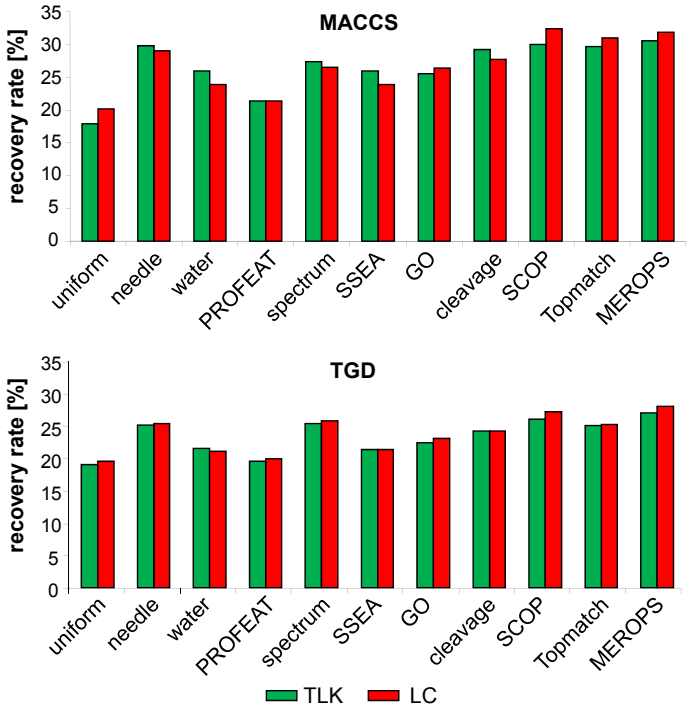


Figure 3.5-2: Global kernel performance (set 2) Search results for target set 2 are shown corresponding to Figure 3.5-1. The figure is adapted from [71].

and B-4 report recovery rates on a per target basis. In this case, the recovery rates were generally higher than for set 1, ranging from approximately 18% to 33%, but differences between SVM search strategies and alternative kernels were even smaller than those observed for target set 1. The uniform kernel produced average recovery rates of close to 20% and several kernels taking protein similarity at different levels into account performed only slightly better. Here, the hierarchical SCOP and MEROPS kernels and the Topmatch kernel that is based on active site structural similarity performed equally well, but only slightly better than the sequence similarity-based needle kernel. Thus, taken together, the results of systematic SVM calculations on our two target sets revealed surprisingly little differences in search performance for target kernels of different design.

3.5.2 Target-Dependent Kernel Performance

As described in section 3.3, the overall best-performing MEROPS kernel differs from other target kernels in that it assigns high weights to closely related targets, due to its exponential formalism (see equation 3.29). In order to explore the contributions of the most closely related targets to ligand recovery, we analyzed the search performance for all individual set 1 targets in SVM TLK

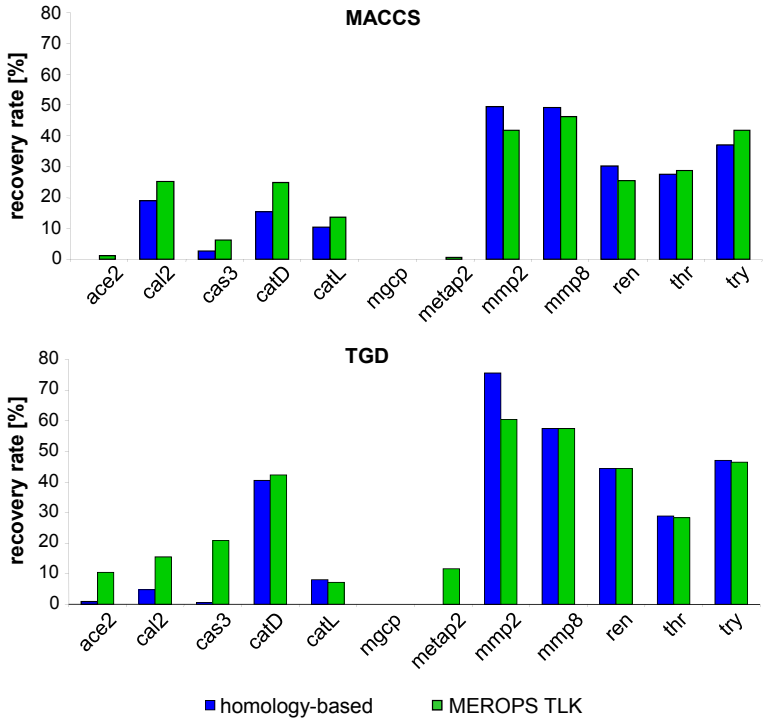


Figure 3.5-3: Target-dependent kernel performance (set 1) For all targets of set 1, recovery rates are shown for homology-based SVM ranking and for the SVM TLK search strategy in combination with the MEROPS target kernel. Recovery rates are compared for selection sets of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1. The figure is adapted from [71].

calculations using the MEROPS kernel and, in addition, homology-based SVM calculations. In the latter case, the SVM was trained on the ligands of the target most closely related to the orphanized target. The results of these SVM TLK and homology-based SVM calculations are shown in Figure 3.5-3. Significant differences in target-dependent search performance were observed. The search performance was found to be highly dependent on the degree of relatedness between the orphan target and its nearest neighbor. For those targets having a closely related nearest neighbor at the subfamily level (i.e., cathepsin D and renin, matrix metalloproteases 2 and 8, thrombin and trypsin; see Figure 3.4-1), highest recovery rates were observed. For these targets, simple SVM calculations using the ligands of the nearest neighbor as positive training examples matched the performance of SVM TLK calculations using the MEROPS kernel. By contrast, homology-based SVM calculations produced only low recovery rates, or failed, for targets that had no closely related neighbor (i.e., all cysteine proteases in set 1, methionyl aminopeptidase 2, glutamate carboxypeptidase 2, and angiotensin-converting enzyme 2; see Figure 3.4-1). The cumulative recall curves shown in Figure 3.5-4a and 3.5-4b illustrate the close correspondence between homology-based SVM and SVM TLK calculations when a closely related

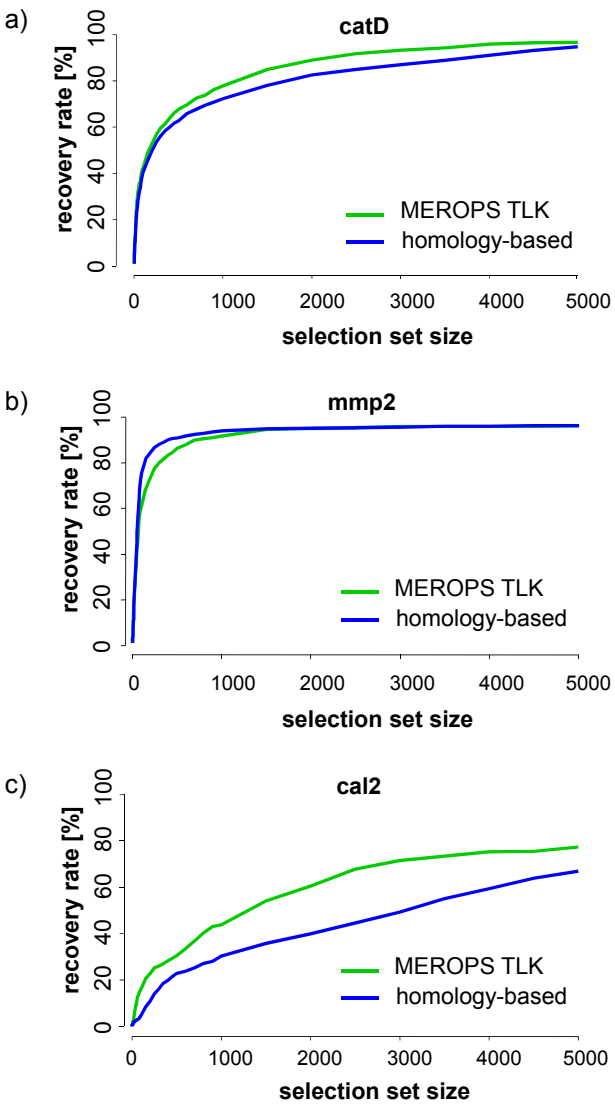


Figure 3.5-4: Cumulative recall curves Representative recall curves for homology-based SVM and SVM TLK in combination with the MEROPS kernel are shown for three targets, (a) cathepsin D, (b) matrix metalloprotease 2, and (c) calpain 2, using TGD as the ligand descriptor. Recovery rates are averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1. The figure is adapted from [71].

nearest neighbor target was available. However, Figure 3.5-4c shows that taking additional target and ligand information into account when no closely related neighbor was available further improved the search performance, a trend that was especially observed for larger database selection sets.

Different from target set 1, each target in set 2 had a nearest neighbor at the subfamily level (Figure 3.4-1). Accordingly, one would expect better target-dependent search performance for targets in set 2 than in set 1. The SVM TLK search calculations with the MEROPS target kernel shown in Figure 3.5-5 confirm this expectation. The majority of targets in set 2 produced recovery rates of at least 20% (with the MACCS fingerprint as ligand representation). In this case, SVM control calculations were also carried out after pooling the ligands of all members of the orphan target's subfamily for training. As illustrated in Figure 3.5-5, the recovery rates observed in these SVM control calculations were

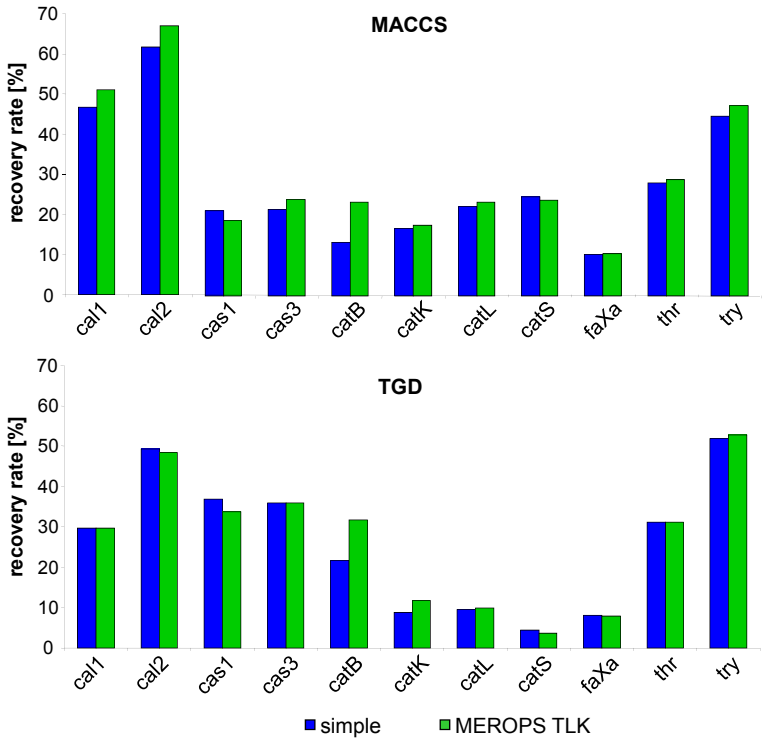


Figure 3.5-5: Target-dependent kernel performance (set 2) For all targets of set 2, recovery rates are shown for simple SVM ranking and for the SVM TLK search strategy in combination with the MEROPS target kernel. For simple SVM ranking, ligands of all members of the orphanized target's subfamily were pooled and used as the positive training class. Recovery rates are shown for a selection set of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-2. The figure is adapted from [71].

almost indistinguishable from those of SVM TLK calculations. Furthermore, in Appendix Table B-5, recovery rates for standard SVM calculations on set 2 targets are reported for selection sets of 100 compounds when either only ligands of the nearest neighbor target were used for training (i.e., homology-based SVM) or, alternatively, ligands of all subfamily members were pooled. The results demonstrate that recovery rates for targets having several closely related subfamily members further improved when ligands from all related targets were taken into account compared to ligands of only the most closely related target.

3.5.3 Nearest Neighbor Effects

The findings discussed above reflect a strong influence of ligand information of nearest neighbor targets on ligand prediction for orphanized targets. In order to evaluate the magnitude of nearest neighbor effects, SVM TLK calculations using the MEROPS kernel were also carried out after removal of the ligands of the nearest neighbor target from SVM learning. The search results for set 1

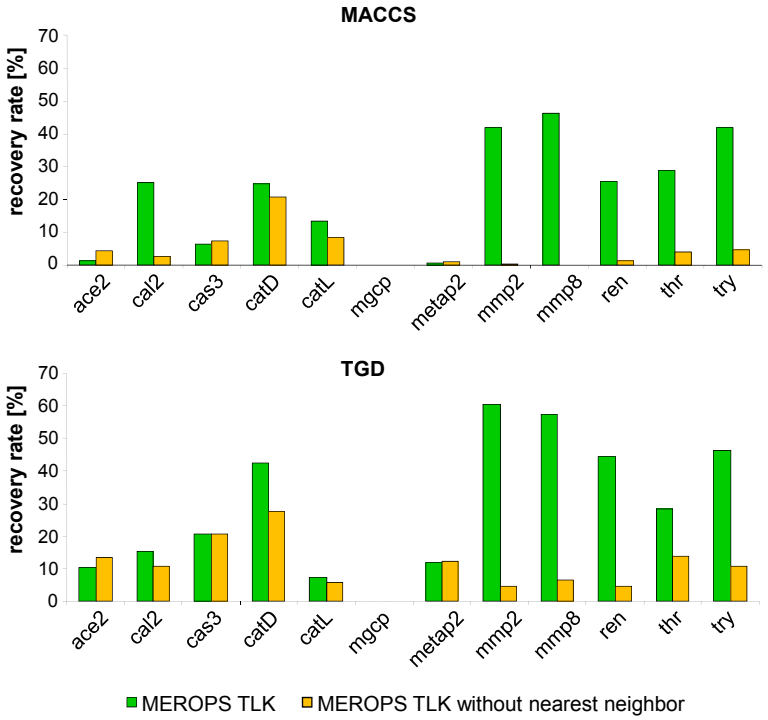


Figure 3.5-6: Dependence of search performance on ligands of nearest neighbor targets For all set 1 targets, search results are reported for the SVM TLK search strategy in combination with the MEROPS kernel. Green bars show recovery rates obtained by learning with ligands of all reference targets, whereas yellow bars show recovery rates obtained when the ligands of the nearest neighbor are excluded from the training set. Recovery rates are shown for a selection set of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1. The figure is adapted from [71].

targets are shown in Figure 3.5-6 and Appendix Table B-6 reports the comparison of SVM TLK and LC calculations. As can be seen in Figure 3.5-6, removal of ligands led to a sharp decline in recovery rates when a nearest neighbor target was available at the subfamily level (effects observed in SVM TLK and LC calculations were similar). By contrast, removal of ligands for targets where no closely related neighbor was available had only little influence on the search performance. Thus, these findings further corroborated the crucial role of nearest neighbor ligand information for orphan target ligand prediction using SVM techniques.

3.6 Conclusions

In this chapter, we have investigated different strategies for SVM-based ligand prediction for simulated orphan targets with special emphasis on the evaluation of alternative target kernel functions that capture protein information at

different levels. Information about orphan targets was not included in SVM model building. Thus, the approaches investigated here aimed at de novo ligand predictions. The way target information was taken into account presented a major variable in these calculations and, accordingly, target kernels of different complexity and information content were designed and evaluated. Surprisingly, these alternative kernel functions influenced the calculations much less than one might anticipate. Rather, nearest neighbor effects were found to be the major determinant of ligand prediction performance. In particular, when ligand information from one or more closely related targets was available, simple SVM calculations utilizing this information met the search performance of SVM TLK and LC calculations. For SVM-based ligand prediction on orphan targets, these findings have significant implications. Rather than focusing on information provided by reference systems capturing protein hierarchies, searching for targets with known ligands that are closely related to orphan targets (e.g., at the subfamily level) should be a primary objective. For this purpose, simple detection of sequence similarity might often be sufficient. In the presence of strong nearest neighbor relationships, SVM-based strategies for ligand prediction can be simplified. In these cases, simple SVM calculations using nearest neighbor ligands for learning are expected to produce promising results. By contrast, if no closely related targets can be identified, SVM learning using target kernels capturing protein hierarchy information is likely to be a preferred approach. Thus, SVM strategies for ligand prediction can be adjusted based on an initial exploration of target relationships.

Source Information

Sections of the text in this chapter have been taken from [71, 77].

Chapter 4

Preferential Detection of Potent Hits in Ligand-Based Virtual Screening

Many ligand similarity-based methods exist that are utilized to mine databases for novel active compounds. However, with the exception of QSAR models, these approaches typically do not consider compound potency as search information [5]. Thus far, it has only rarely been attempted to incorporate potency information into LBVS search algorithms [92], although search strategies tailored to the preferential detection of potent hits would certainly be highly attractive for practical applications.

Potency information has recently been integrated into an SVM-based approach to the prediction of selectivity toward human adenosine receptors (hARs) [93]. Here, a multi-label approach (termed ct-SVM) was used to construct a single model integrating binary classifiers for four different hAR subtypes. Furthermore, three models based on increasingly strict criteria for threshold activity (i.e., K_i threshold values of 500, 250, and 100 nM) were applied sequentially to quantify the biological affinity of test compounds. This analysis demonstrated that SVM-based classification provides an interesting alternative to traditional regression-based QSAR modeling. This study aimed at the annotation of test compounds with predefined potency ranges and represents a non-QSAR SVM-based classification of different biological activity levels.

Considering the learning from multiple classes with different affinity ranges as a typical multi-task problem, we decided to adapt the different SVM strategies presented in Chapter 3 to potency-directed virtual screening. For this purpose, we developed a new structure-activity kernel function and constructed a potency-oriented SVM linear combination that were tested on different public domain screening data sets and compared to conventional SVM ranking [76], as reported in this chapter. Section 4.1 describes the composition of the test data

sets, the applied SVM strategies, and search calculations. In section 4.2, the performance of the three SVM search strategies in the benchmark calculations is reported and conclusions of the study are discussed in section 4.3.

4.1 Data, Search Strategies, and Calculations

To provide a practically relevant search scenario, four different HTS sets were selected as test data sets for the evaluation of our two potency-directed SVM search strategies. Furthermore, a standard binary SVM was included as control in the search calculations to set a baseline performance expectation.

4.1.1 High-Throughput Screening Data Sets

The four HTS data sets were extracted from PubChem BioAssay and included inhibition assays for enzyme targets hydroxyacyl-coenzyme A dehydrogenase type II (assay identifier (AID) 886), 15-human lipoxygenase (AID 887), 15-hydroxyprostaglandin dehydrogenase (AID 894), and aldehyde dehydrogenase 1 (AID 1030). All assays were designated as *confirmatory assays*, indicating that compounds had been tested at different concentrations to generate dose-response curves. Compound potencies were reported as half-maximal inhibitory concentrations (IC_{50} values). Compounds with incomplete or ambiguous activity annotations were removed from these data sets. Then, a 2D unique version of each compound set was generated, i.e., stereoconfigurations of molecules were ignored and of compounds sharing the same 2D graph only the one with highest potency was retained. The composition of the so-prepared compound data sets is summarized in Table 4.1-1. It should be emphasized that active compounds in all four data sets covered wide potency ranges of more than three orders of magnitude. Furthermore, in each data set, there were many more weakly than highly potent compounds and, in addition, many more inactive than active compounds. Thus, for potency-directed LBVS, these data sets provided challenging test cases. For each data set, potency intervals were defined to divide active compounds into four potency categories, termed *C1–C4*, with potency values decreasing from C1 to C4. For each category, the negative decadic logarithm of the potency value of its lower potency threshold was calculated and used as its annotation, $pAct$, as also reported in Table 4.1-1.

4.1.2 Support Vector Machine Search Strategies

We investigated standard SVM calculations as well as two potency-directed SVM techniques including a structure-activity kernel taking reference compound potency differences directly into account and, in addition, a linear com-

Table 4.1-1: Data sets

AID	target	#act	#inact	range	pAct	cat	#mol	#ref
886	hydroxyacyl-coenzyme A dehydrogenase type II	2 409	68 845	10–100 nM	7	C1	20	5
				100 nM–1 µM	6	C2	128	32
				1–10 µM	5	C3	803	200
				10–100 µM	4	C4	1 458	364
887	15-human lipoxygenase	998	70 822	2–200 nM	6.7	C1	16	5
				200 nM–2 µM	5.7	C2	93	29
				2–20 µM	4.7	C3	711	222
				20–200 µM	3.7	C4	178	55
894	15-hydroxy-prostaglandin dehydrogenase	6 318	139 805	1–100 nM	7	C1	12	5
				100 nM–1 µM	6	C2	115	47
				1–10 µM	5	C3	1 452	605
				10–100 µM	4	C4	4 739	1 974
1030	aldehyde dehydrogenase 1	15 817	197 666	10–100 nM	7	C1	138	5
				100 nM–1 µM	6	C2	946	34
				1–10 µM	5	C3	6 091	220
				10–100 µM	4	C4	8 642	313

For four confirmatory high-throughput screening data sets, the numbers of active ("#act") and inactive ("#inact") compounds are reported. For each data set, the potency ranges ("range") of the four categories C1-C4 ("cat") into which active compounds were divided and the potency threshold values pAct are given. Furthermore, the numbers of molecules per potency category ("#mol") and reference compounds ("#ref") taken from each category are reported.

bination of different SVM hyperplanes derived for reference compounds falling into different potency ranges.

4.1.2.1 Standard SVM

For standard SVM calculations, compound training sets of different composition are used. First, active reference compounds from all potency categories are pooled to provide the positive training class, and confirmed inactive molecules are used as the negative training class, as shown in Figure 4.1-1 on the left. Then the maximum margin hyperplane is derived and test molecules with unkown activity status are ranked by their signed distance to the decision boundary (equation 3.12). This SVM strategy is named SVM_{pooled}. Furthermore, for control calculations, positive reference sets exclusively containing highly potent compounds are also utilized, as illustrated in Figure 4.1-1 on the right. A standard SVM using positive reference sets consisting only of compounds taken from the potency range C1 is termed $SVM_{1\,Cat}$. Accordingly, a standard SVM with active training compounds falling into the categories C1 and C2 is referred to as SVM_{2Cat}.

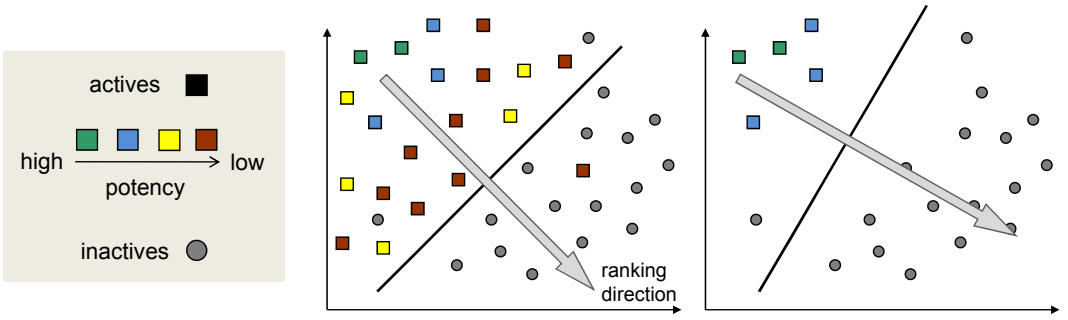


Figure 4.1-1: Standard SVM In $\text{SVM}_{\text{pooled}}$ (left), known active compounds from all potency categories are pooled to form the positive training set, and confirmed inactive compounds constitute the negative training class. Test compounds are ranked according to their signed distance from the hyperplane represented by the arrow. For control calculations (right), only reference compounds from the two highest potency categories are used as positive training examples. The figure is adapted from [76].

4.1.2.2 SVM Linear Combination

SVM LC is adapted for potency-directed SVM searching. Therefore, for each potency category C_i, a hyperplane is constructed using known active ligands of C_i as positive training objects and inactive reference compounds as negative examples. To obtain one overall ranking function, the individual normal vectors $\mathbf{w}_i$ of all hyperplanes are then linearly combined to a single vector $\mathbf{w}_{combined}$ by applying

$$\mathbf{w}_{\text{combined}} = \sum_{i=1}^{n} f_i \mathbf{w}_i \quad \text{with} \quad f_i = a_i - \min_{j=1,\ldots,n}(a_j) + 1 \tag{4.1}$$

where f_i denotes the linear factor, a_i the pAct of potency category C_i, and n the total number of categories. Test compounds are then ranked by $g(\mathbf{x}) = \langle \Phi(\mathbf{x}), \mathbf{w}_{\text{combined}} \rangle$. For this strategy termed LC_{simple}, f_i increases linearly with the pAct of the potency category C_i. To further increase weights on highly active compounds, the $LC_{squared}$ strategy is introduced that utilizes the square product of the linear factors used in $\text{LC}_{\text{simple}}$ as the potency category-specific weight for the linear combination:

$$\mathbf{w}_{\text{combined}} = \sum_{i=1}^{n} f_i \mathbf{w}_i \quad \text{with} \quad f_i = \left(a_i - \min_{j=1,\ldots,n}(a_j) + 1 \right)^2 \tag{4.2}$$

4.1.2.3 SVM with Structure-Activity Kernel

In analogy to the target-ligand kernel discussed in Chapter 3, we designed a structure-activity kernel (SAK). For SVM using the SAK in learning and ranking (SVM SAK), we represent each compound i as a fingerprint-potency category pair $(\mathbf{x}_i, a_i)$. Accordingly, the comparison of two compounds is divided into

a separate assessment of their structural similarity and their activity similarity by two different kernel functions $K_{structure}$ and $K_{activity}$ that are then combined to build the SAK:

$$K\big((\mathbf{x}_i, a_i), (\mathbf{x}_k, a_k)\big) = K_{\text{structure}}(\mathbf{x}_i, \mathbf{x}_k) \times K_{\text{activity}}(a_i, a_k) \tag{4.3}$$

The design principle of the SAK is illustrated in Figure 4.1-2. To make SVM SAK directly comparable to SVM LC (vide infra), the activity kernel is defined as

$$K_{\text{activity}}(a_i, a_k) = \max_{j=1,\ldots,n}(a_j) - \min_{j=1,\ldots,n}(a_j) + 1 - |a_i - a_k| \tag{4.4}$$

for the approach SAK_{simple} or as

$$K_{\text{activity}}(a_i, a_k) = \Big(\max_{j=1,\ldots,n}(a_j) - \min_{j=1,\ldots,n}(a_j) + 1 - |a_i - a_k|\Big)^2 \tag{4.5}$$

for the approach $SAK_{squared}$. As in equations 4.1 and 4.2, a_i denotes the pAct of potency category C_i and n the total number of categories.

For SVM training, positive training objects are obtained by combining the fingerprint combination of each known active compound with its potency category threshold value and negative training examples by combining the fingerprint representation of inactive reference compounds with all possible potency category threshold values. Then a hyperplane is derived to separate true compound fingerprint-potency pairings from false pairing (Figure 4.1-2). For the classification of molecules with unknown activity, test compounds are assigned the threshold value a_{high} of the highest potency category, i.e., $a_{\text{high}} = \max(a_j)$, and a ranking is generated by determining the signed distance from the pairs $(\mathbf{x}, a_{\text{high}})$ to the hyperplane H derived in structure-activity reference space. Test compounds are paired with a_{high} because we aim at the detection of potent hits and want to sort test compounds according to their likelihood of belonging to the highest potency category. It should be noted that, by setting a_k to a_{high}, the kernel function $K_{\text{activity}}(a_i, a_k)$ becomes identical to the factor f_i in the linear combination.

4.1.3 Search Calculations

The performance of the alternative SVM ranking strategies was evaluated in search calculations on the four PubChem HTS data sets. Compounds were encoded as MACCS or ECFP4 bit strings. To compare fingerprint representations, the Tanimoto kernel (see equation 3.10) was utilized. For test calculations, reference compound sets were assembled to reflect the potency distribution in each data set. Accordingly, five compounds belonging to the highest potency category were randomly selected in each case and reference compounds of the

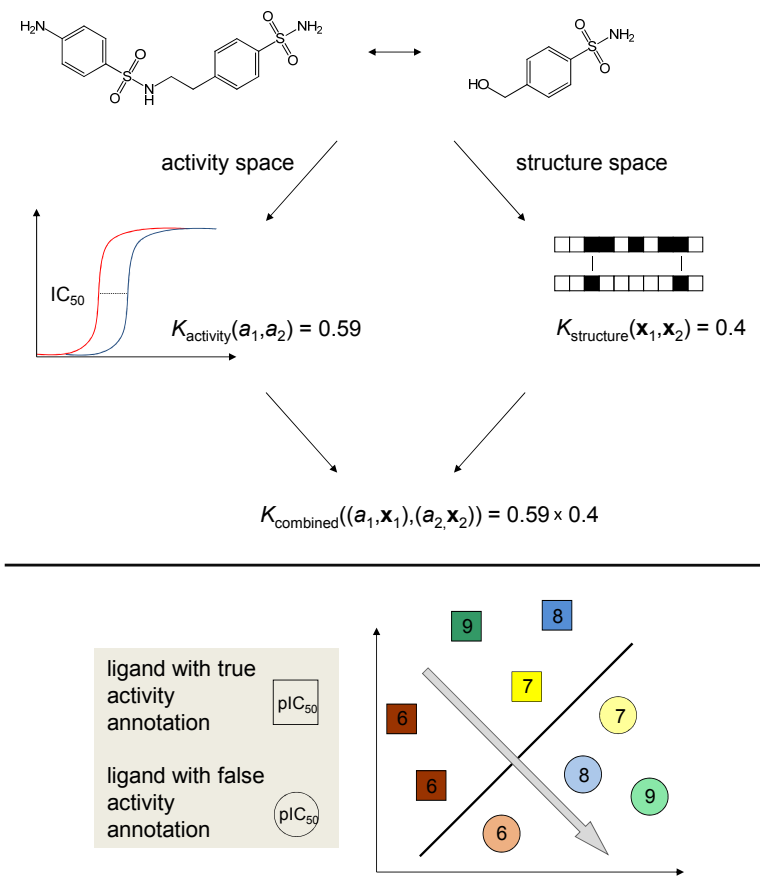


Figure 4.1-2: SVM with structure-activity kernel The comparison of two fingerprint-potency category pairs is divided into two independent tasks such that the structural similarity and activity similarity of two compounds are first separately determined and then combined. A hyperplane is constructed to separate true fingerprint-potency category pairings from false pairings. The figure is adapted from [76].

other categories were chosen such that the reference-to-test molecule ratio was approximately the same for all potency categories, as reported in Table 4.1-1. Positive training classes mirroring the potency distribution of a data set are termed *potency-balanced*. Control calculations were carried out with reference sets containing only highly potent compounds. These biased reference sets were used to evaluate whether potent reference compounds would lead to the preferential detection of potent hits. For all assays and SVM strategies, 1 000 inactive compounds were taken as negative training examples. All remaining molecules from each data set were utilized as screening database. For each combination of a search strategy and fingerprint, ten different trials with randomly assembled reference and test sets were carried out. As a measure of performance, recovery rates were calculated for database selection sets of increasing size and averaged over the ten independent trials per target.

All calculations were carried out using SVM^{light} with default settings for calculation parameters and Perl scripts were applied to calculate SVM linear combinations and organize the results.

4.2 Results

With our study, we aimed to investigate whether compound potency could be incorporated as a search parameter in SVM-based virtual screening to further refine search calculation.

4.2.1 Search Performance

We first compared our three alternative SVM strategies for potency-balanced reference sets. The results for ECFP4 and MACCS representations are reported in Figure 4.2-1 and Appendix Figure C-1, respectively. In these figures, compound recall of all active compounds (regardless of their potency) and of the most potent compounds (categories C1, C2) is separately monitored. Compound recall was generally higher for ECFP4 than for MACCS. Overall, the average recovery rates of all active compounds were comparable for standard SVM (SVM_{pooled}) and advanced SVM strategies. However, in all cases, SVM SAK and LC calculations were found to retrieve a higher percentage of highly potent compounds than standard SVM calculations. Although search results for SVM SAK and LC were very similar, some underlying trends and characteristics of the individual methods were detected. Independent of the fingerprint representation, SVM SAK usually identified more active compounds belonging to the highest potency category C1 than LC. Because the overall compound recall was comparable for all advanced strategies using ECFP4 as fingerprint representation, SVM SAK was considered as the preferred strategy for this fingerprint. However, for the MACCS fingerprint, overall compound recall was consistently higher for LC than for SVM SAK so that there was no clear advantage of one over the other method. Furthermore, the search results for simple and squared weights were also comparable. However, the use of simple weights often led to slightly higher recovery rates for all active compounds, whereas squared weights favored the recovery of highly potent molecules.

In Figure 4.2-2 and Appendix Figure C-2, average recovery rates are reported for a constant selection set size of 1 000 database compounds. Depending on the HTS data set, recovery rates for all active compounds ranged from ca. 3% to 20%. For potent (C1, C2) compounds, higher recovery rates were observed ranging, on average, from approximately 10% to 60%. Here it should be taken into account that many more weakly than highly potent compounds were available in each data set. The comparison of the recall rates of alternative SVM strategies for selection sets of 1 000 database compounds further

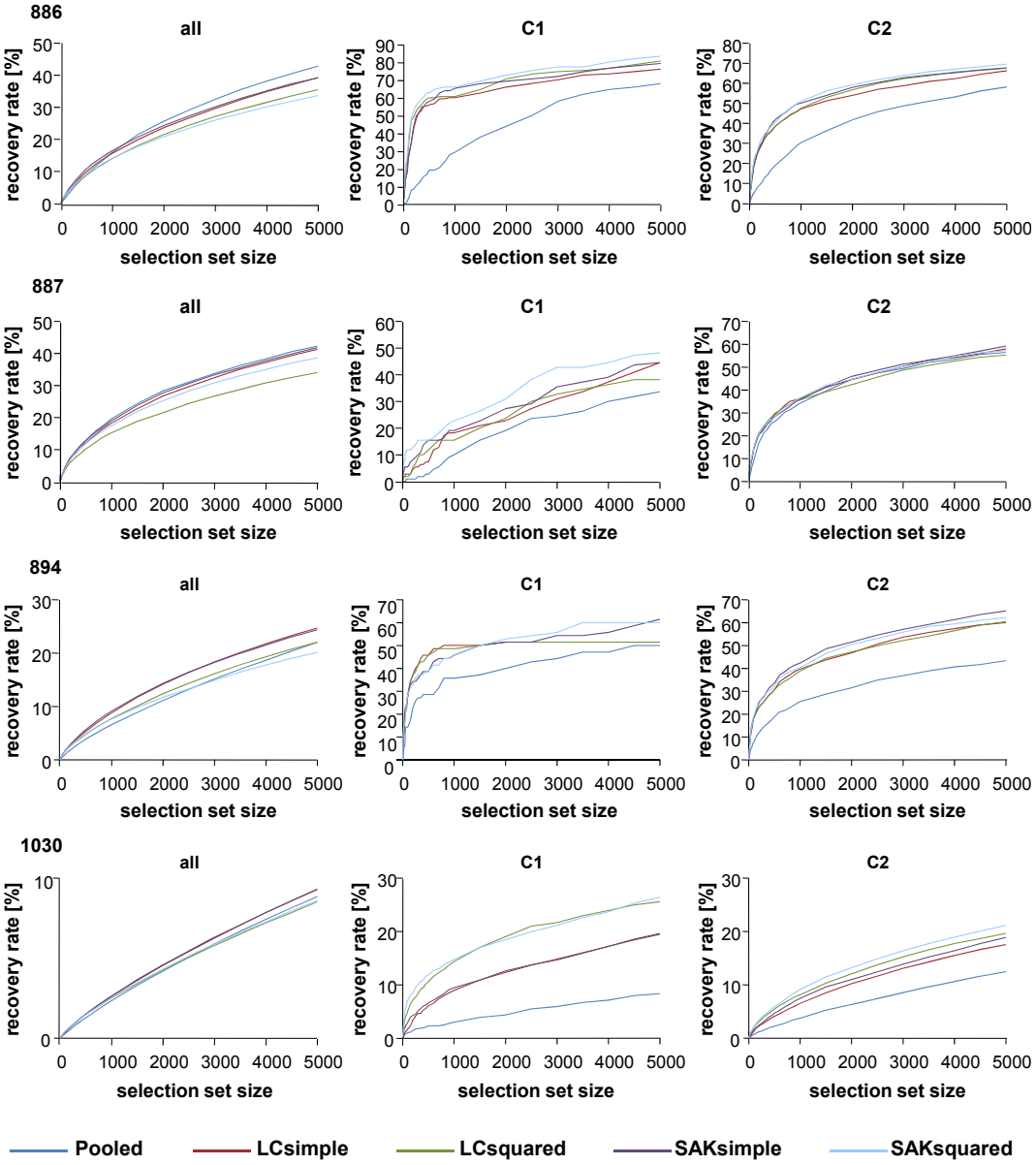


Figure 4.2-1: Cumulative recall curves for potency-balanced reference sets For each bioassay, cumulative recall curves are shown for all active compounds and the highest potency categories (C1 and C2) and different SVM strategies using ECFP4 as fingerprint representation. Recall curves represent the average of ten independent trials using different reference sets. Potency-balanced reference sets consist of compounds spanning the entire potency range in a data set. The figure is adapted from [76].

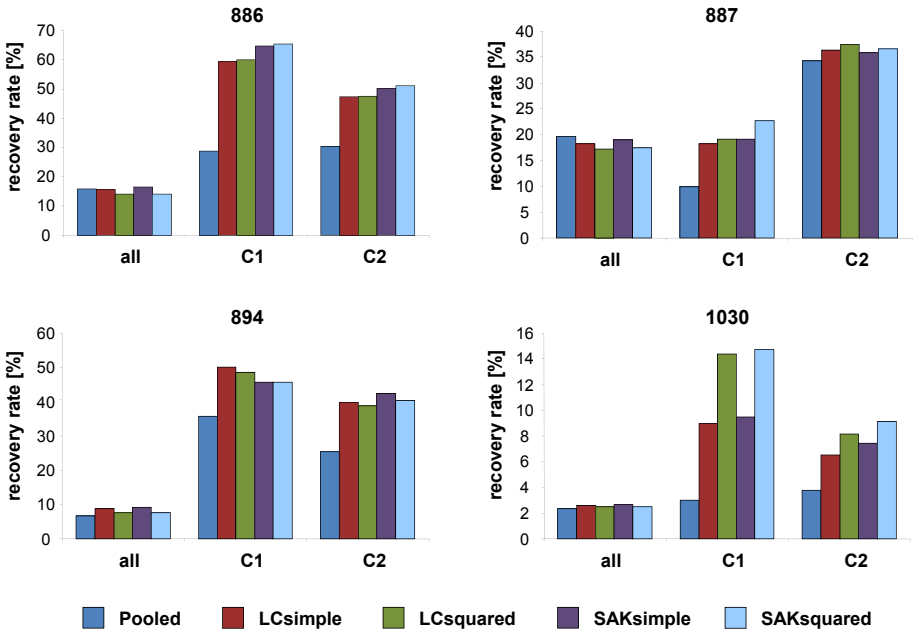


Figure 4.2-2: Support vector machine performance for database selection sets of constant size Recovery rates are shown for the ECFP4 representation, potency-balanced reference sets, and database selection sets of 1 000 compounds. The results are averaged over 10 independent trials per data set. The figure is adapted from [76].

illustrated that SVM SAK and LC calculations consistently detected more potent compounds than standard SVM calculations. Thus, potency-directed SVM searching reached the recall performance of standard SVM classification but led to the desired preferential detection of hits having higher potency.

4.2.2 Control Calculations

We next carried out standard SVM calculations on active reference sets exclusively consisting of potent compounds. The results were then compared to standard SVM and SAK calculations for potency-balanced reference sets. These control calculations were carried out to reveal whether the potency of reference compounds determined the outcome of the search calculations relative to advanced SVM strategies. The results for the ECFP4 and MACCS representations are reported in Figure 4.2-3 and Appendix Figure C-3, respectively. It can be seen that standard SVM using only the five most potent reference compounds as positive training examples (strategy SVM_{1Cat}) produced recovery rates of potent compounds that were significantly lower compared to advanced strategies, which was especially obvious for the recall of potent compounds belonging to category C2. In most cases, even standard SVM using potency-balanced reference sets recognized more C2 compounds. Of course, the C1 reference set was the smallest of all and hence contained the least information about active

molecules. Accordingly, when adding reference compounds falling into potency category C2 to the positive training class (strategy SVM_{2Cat}), recovery rates of potent compounds increased and were found to be overall comparable to advanced strategies (or even slightly better in case of the MACCS fingerprint). Thus, the exclusive use of highly potent reference compounds in standard SVM calculations also led to the preferential detection of potent screening hits. However, there was a price to pay because, in this case, the recovery rates of all active compounds were substantially reduced for standard SVM calculations. Thus, overall much better recall rates of active compounds were obtained for balanced reference sets where potency-directed SVM searching provided a clear enrichment of potent screening hits.

4.3 Conclusions

This chapter introduced SVM-based techniques for potency-directed LBVS, for which alternative methods are currently not available. For many similarity-based search methods, the incorporation of potency as a search parameter is a difficult problem. However, in the context of SVM learning, the use of kernel functions and their combination provides a basis for the design and implementation of a multi-parametric search approach. SVM LC learns separate hyperplanes for training sets of different activity ranges and then combines them by associating a potency-dependent weighting scheme. By contrast, the SAK approach introduced herein compares compound pairs simultaneously in activity and structure space by evaluating structural similarity on the basis of whole-molecule fingerprint descriptors and multiplying it with an assessment of activity similarity for pairs of ligands. Using balanced (unbiased) compound reference sets, both advanced SVM techniques met the active compound recall performance of conventional SVM calculations but achieved a clear enrichment of potent hits. In addition, we demonstrated that reference sets biased towards compounds having high potency also led to an enrichment of potent hits in standard SVM calculations, but only at the cost of overall recall performance. These findings have a number of implications for practical SVM database search applications. We deliberately performed our analysis on HTS data to eliminate the influence of molecular complexity effects [94] on the search results. In typical screening libraries, hits with different potency usually have comparable molecular weight and topological complexity because they are not (yet) chemically optimized with respect to a specific biological activity. This avoids complications that are often associated with benchmark calculations and also practical applications. In typical benchmark settings, highly optimized and potent compounds are usually added to screening databases consisting of lower complexity compounds, which generally yields artificially high recall rates [11] because highly complex reference and active database compounds are relatively easy

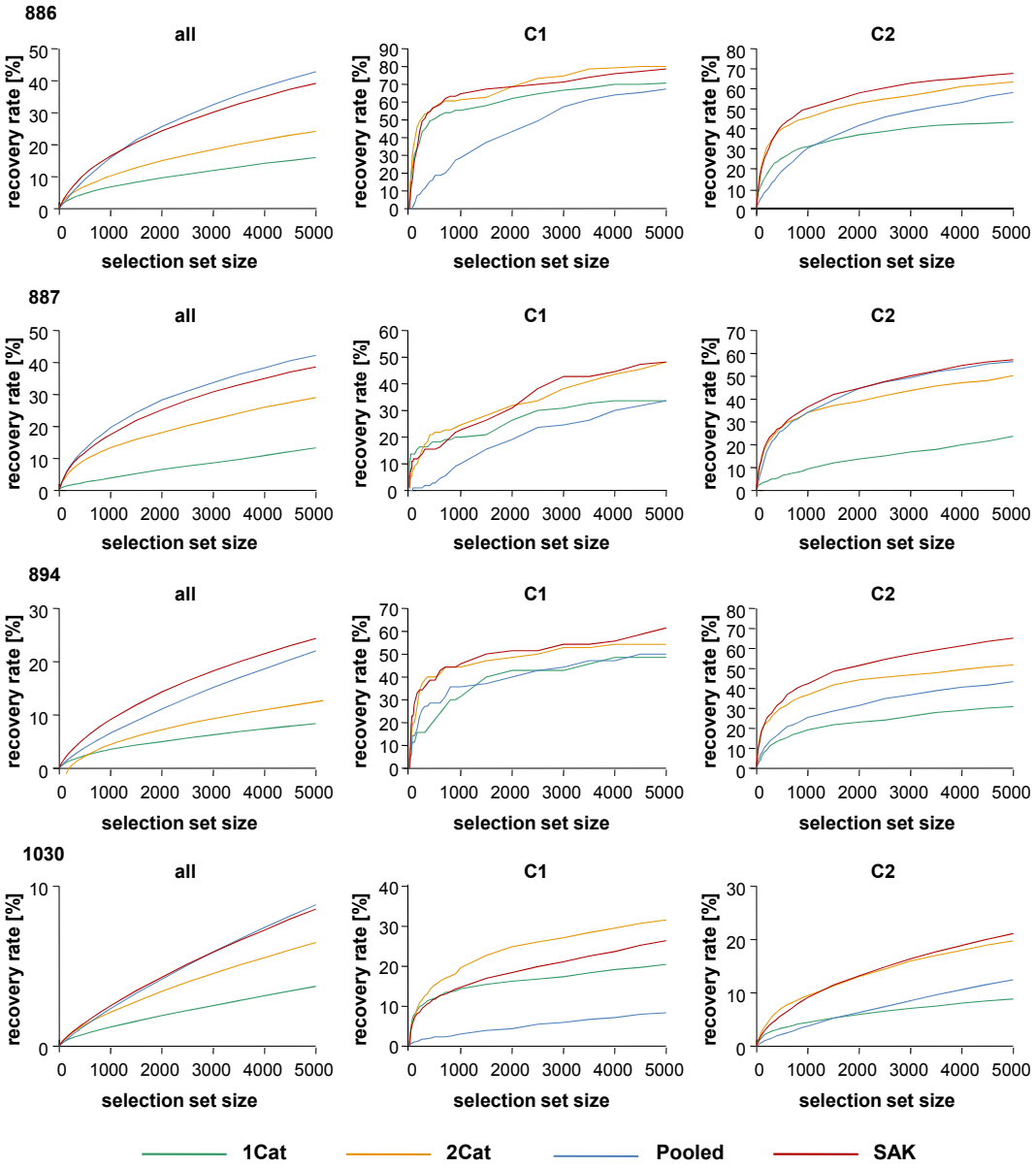


Figure 4.2-3: Control calculations using highly potent reference compounds
For each bioassay, recall curves are shown for all active compounds and the two highest potency categories (C1 and C2). Compound recall is monitored for different SVM strategies using ECFP4 averaged over 10 independent trials. The following strategies are compared: standard SVM with reference compounds from potency category 1 ('1Cat'), 1 and 2 ('2Cat'), and all categories ('Pooled') and SVM structure-activity kernel (SAK). SAK_{simple} is shown for sets 886 and 894, $SAK_{squared}$ for sets 887 and 1030. The figure is adapted from [76].

to distinguish from screening molecules having lower complexity. However, the situation is completely different when highly complex reference compounds are utilized to search for hits having average screening database complexity, which has been shown to provide the by far most difficult practical search scenario for fingerprint-based methods [95]. These considerations would suggest to rather focus on screening hits as reference compounds, even if many of them might only be weakly potent. For SVM learning, we now introduce techniques that take relative compound potency into account and are particularly well suited for this task. Selecting a spectrum of available screening hits for learning, the SVM SAK and LC techniques would be expected to detect many active compounds and direct the search towards potent hits, if available in a screening database. Such calculations should be particularly promising if reference compounds and potential hits would originate from the same screening collection (where many active compounds might have similar chemical properties). For example, this would make the application of these methods attractive in the context of sequential screening [96] where initial screening hits from a fraction of the database are used as reference compounds for search calculations to prioritize another subset of the database (with a putative enrichment of additional hits) for the next round of experimental screening.

Source Information

Sections of the text in this chapter have been taken from [76].

Chapter 5

Selection of Compound Class-Specific Descriptors

Numerical descriptors of chemical structure and properties play a central role in chemoinformatics, and literally thousands of different descriptors are currently available [45]. Descriptors that capture compound class-specific and biological activity-relevant information are of high interest for the exploration of structure-activity relationships. However, the identification of such descriptors is far from being a trivial task. Thus, selection algorithms capable of finding descriptors that contain compound class-specific information are highly desired. A generally applicable way to identify discriminatory descriptors is to compare their data distributions for a given compound activity class and a large database where the vast majority of compounds do not have the desired activity. A descriptor contains compound class-specific information if the value distributions of the descriptor significantly differ for the two compound data sets. By contrast, if value distributions for a descriptor are very similar for the two data sets, i.e., if each descriptor value occurs with roughly the same frequency for the activity class and the database, then the descriptor provides only very little set-specific information. Albeit simple in theory, a systematic descriptor selection is complicated by the fact that different descriptors usually have different units and value ranges so that identified differences in descriptor settings cannot be easily compared. Therefore, descriptor selection approaches that make use of the Shannon entropy (SE) concept [38] from information theory have been developed to quantify the variability of different descriptors independent of their value ranges by representing data distributions as histograms with a defined number of bins [97]. Furthermore, to quantitatively compare the overlap of descriptor value distributions for two different data sets and rank descriptors by their ability to distinguish between compounds of different sources, an extension of the SE approach termed Differential Shannon Entropy (DSE) [98] was introduced. However, as pointed out in this chapter, the DSE approach is intrin-

sically limited in its ability to select set-specific descriptors for compound data sets of very different size. This implies that the DSE approach is not amenable to the exploration of structure-activity relationships in a meaningful way because the identification of descriptors that capture biological activity-relevant information typically requires the comparison of a given activity class containing only a few dozen or hundred molecules and a large database comprising thousands or even millions of compounds. To circumvent these difficulties and reliably assess the class-specific information content of descriptors, we transformed the DSE formalism into mutual information analysis, another concept from information theory, and evaluated our approach by descriptor ranking and correlation analysis on 168 compound activity classes [99]. This chapter describes the evolution from initial Shannon entropy applications to our mutual information-based approach in a stepwise manner. The Shannon entropy concept is introduced in section 5.1. Details of the DSE approach and identified shortcomings are discussed in section 5.2. The transformation of the DSE formalism into mutual information is integral part of section 5.3. A systematic comparison of descriptor rankings produced by DSE and our newly introduced approach is reported in section 5.4. The chapter ends with concluding remarks in section 5.5. All descriptors used in this chapter are available in MOE.

5.1 Shannon Entropy

Introduced in a landmark paper by Claude Shannon in 1948 and originally developed for applications in digital communication, Shannon entropy [38] is a concept from information theory to quantify the average information contained in a "message". In the context of molecular descriptor analysis, the message is simply the value of a descriptor calculated for a compound and the SE is given by the average information content of all values of this descriptor for a compound set. The information content of a certain descriptor value depends on the frequency with which this value occurs in a set of compounds and is calculated as the negative base 2 logarithm of its frequency of occurrence (or probability) p_i (i.e., $-\log_2(p_i)$). Hence, the information content increases with decreasing frequency of occurrence, which is rather intuitive because a rare descriptor value obviously conveys more information about a compound than a frequently occurring value. SE defines the average information contained in a descriptor D and is given by

$$H(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{5.1}$$

where n corresponds to the number of possible values the descriptor adopts. The higher $H(D)$ becomes, the more information is captured by the descriptor D.

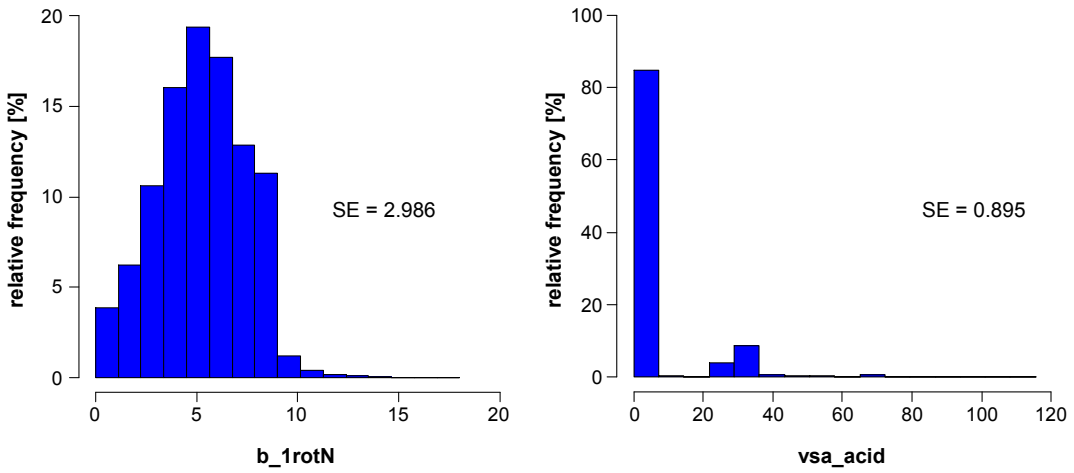


Figure 5.1-1: Descriptor histograms and corresponding SEs Two exemplary descriptor histograms calculated from 100 000 compounds randomly taken from ZINC are shown, and the corresponding SE values are reported. The descriptors "b_1rotN" and "vsa_acid" represent a high- and a low-entropy descriptor, respectively.

The Shannon entropy $H(D)$ is maximal when all descriptor values have the same frequency of occurrence, resulting in an SE equal to $\log_2(n)$. By contrast, $H(D)$ is minimal and adopts a value of zero if only one descriptor value is observed, i.e., if the frequency of a particular descriptor value is one. To facilitate the quantitative comparison of the average information content of different descriptors, a consistent data representation format for their value distributions is desirable. Therefore, all descriptor distributions are represented as histograms where the complete data range of a descriptor is divided into the same number of equally sized data intervals. Exemplary histogram representations of value distributions and the corresponding SE are shown in Figure 5.1-1. For 100 000 compounds randomly taken from the ZINC database, value distributions of the descriptors "b_1rotN" (number of rotatable single bonds) and "vsa_acid" (approximation to the sum of van der Waals surface areas of acidic atoms) are reduced to a discrete set of possible values by partitioning the range between the minimum and maximum value into 16 evenly spaced data intervals. As can be seen, the descriptor "b_1rotN" varies greatly among the database compounds, whereas the values of the descriptor "vsa_acid" mostly fall into a single bin. The differences between these distributions and their information content are reflected by the calculated SE values of 2.986 for "b_1rotN" and 0.895 for "vsa_acid".

In addition to comparing SEs for different descriptors, the information content of a descriptor for two different compound sets **A** and **B** can also be compared. For this purpose, exactly the same bin definitions (i.e., partitions) must be used to represent the value distributions for the two data sets. Therefore, the range of values the descriptor adopts for the union of sets **A** and **B** is deter-

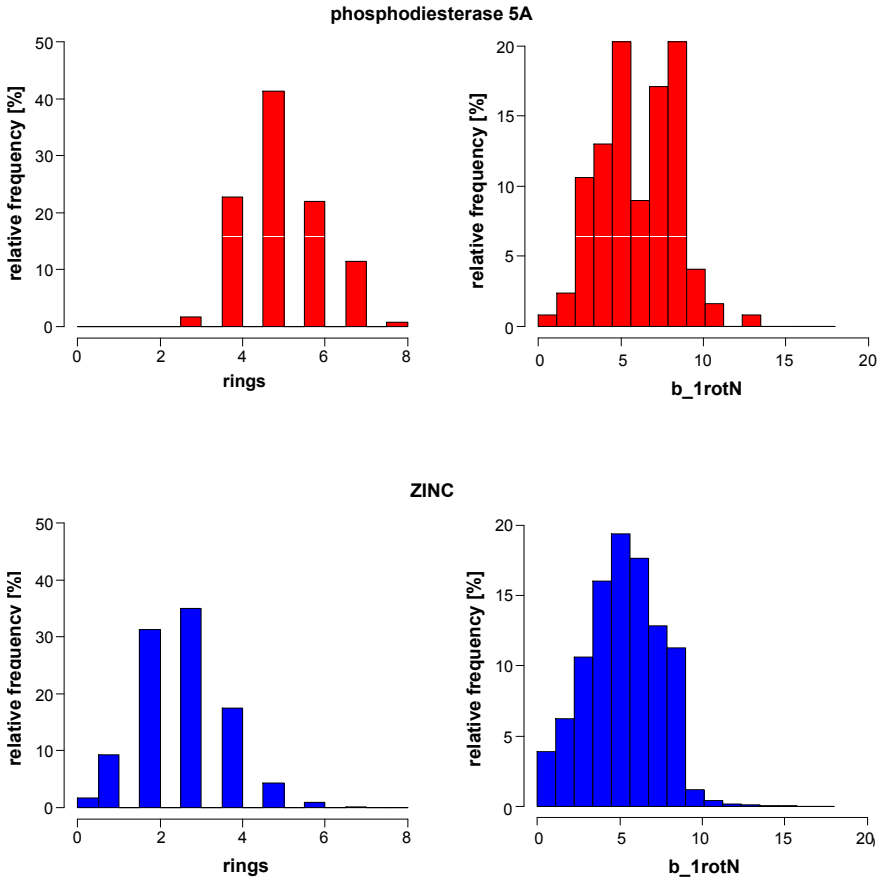


Figure 5.2-1: Discriminatory and non-discriminatory descriptors Exemplary descriptor value distributions are shown for a class of 123 phosphodiesterase 5A inhibitors (red) and 100 000 ZINC compounds (blue). Histograms for the (number of) "rings" descriptor are distinct and cover mostly different value ranges; hence this descriptor contains class-specific information. Histograms for the "b_1rotN" (number of rotatable bonds) descriptor are shown that largely overlap. Therefore, this descriptor is unsuitable to discriminate between the activity class and the reference database.

mined and then divided into a predefined number of equally sized bins. However, the comparison of SEs for two compound sets only accounts for differences in the variability of the corresponding distributions but does not provide information about the distribution overlap. Quantifying the overlap of descriptor value distributions for different data sets is important because descriptors with little overlap can be utilized to distinguish between compounds from different sources.

5.2 Differential Shannon Entropy

In order to compare descriptor value settings for any two classes of compounds, e.g., two different databases or active versus inactive compounds, the SE concepts needs to be extended. Discriminatory and non-discriminatory descriptors for comparison of an exemplary activity class, i.e., a set of phosphodiesterase inhibitors taken from the ChEMBL, and the ZINC subset are shown in Figure 5.2-1. Most active compounds contain more than four rings, whereas ZINC molecules consist mostly of one to four rings. Thus, the descriptor "rings" can be utilized to discriminate between these compound classes. Of course, the discrimination is not perfect because the two histograms overlap. Furthermore, the number of rotatable single bonds is compared for the two data sets. As illustrated in Figure 5.2-1, histograms for the descriptor "b_1rotN" are highly variable for both data sets but cover a similar value range. Hence, this descriptor is clearly non-discriminatory. This example emphasizes an important point, namely that descriptors that are information-rich for single data sets are not necessarily suitable to distinguish between different sets.

The DSE method was introduced to quantify how much information about a given compound class is contained in the value distribution of a descriptor when compared to another [98]. As illustrated in Figure 5.2-2, DSE calculations for a descriptor D and two compound classes **A** and **B** involve the following steps: First, for both sets, the descriptor value distributions are represented as histograms using an equi-distant binning scheme. From these two histograms, the set-specific Shannon entropies $H_{\mathbf{A}}(D)$ and $H_{\mathbf{B}}(D)$ are calculated. Then, a single histogram accounting for the distribution of the entire population of compounds from both sets is generated. For this combined histogram, the frequency for a bin i is calculated according to the following equation:

$$f_{\mathbf{AB}}(i) = \frac{n \times f_{\mathbf{A}}(i) + m \times f_{\mathbf{B}}(i)}{n + m} \tag{5.2}$$

Here, n corresponds to the number of molecules in set **A** and m to the number of molecules in set **B**. In addition, $f_{\mathbf{A}}(i)$ and $f_{\mathbf{B}}(i)$ report bin frequencies for sets **A** and **B**. Based on the combined histogram, $H_{\mathbf{AB}}(D)$ is calculated. Finally, DSE is defined as

$$\mathrm{DSE}(D) = H_{\mathbf{AB}}(D) - \frac{H_{\mathbf{A}}(D) + H_{\mathbf{B}}(D)}{2} \tag{5.3}$$

In Figure 5.2-3, descriptor value distributions binned into 16 data intervals are compared for 10 000 ZINC and 10 000 MDDR compounds and combined histograms are reported. For all ZINC compounds, values for the shape descriptor (topological index) "KierA1" fall into the six lowest bins, with more than 60% of all values accumulating in the third bin, such that the distribution becomes rather narrow. Although this descriptor also preferably adopts low values for

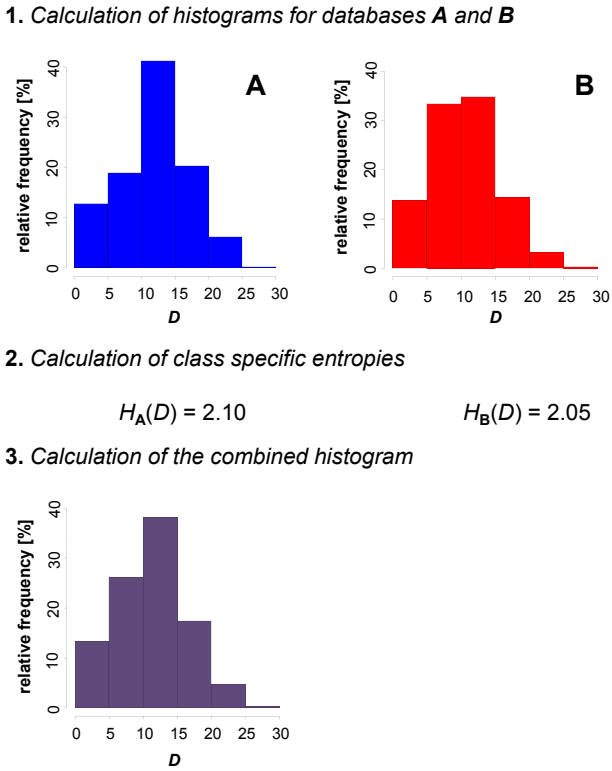


Figure 5.2-2: DSE calculation All steps involved in DSE calculation are illustrated for two hypothetical classes of the same size, classes **A** and **B**. In this example, the value range of descriptor *D* is divided into six bins. The figure is adapted from [62].

MDDR compounds, descriptor values are more evenly spread and the right tail of the MDDR distribution shows that high descriptor values are obtained for a compound subset. Because high descriptor values are exclusively detected for MDDR compounds, the descriptor carries at least some set-specific information. By contrast, for the surface area descriptor "SlogP_VSA5", the distributions for ZINC and MDDR compounds are almost identical.

With its highly populated third bin and right tail, the shape of the combined histogram for the descriptor "KierA1" clearly reflects distinct characteristics of the two underlying distributions. Since the MDDR and ZINC distributions for the descriptor "SlogP_VSA5" were highly similar, it is not surprising that the combined histogram is also hardly distinguishable from the distributions of the individual data sets. As reported in Figure 5.2-3, "KierA1" and "SlogP_VSA5" obtain DSE values of 0.157 and 0.007, respectively. Hence, in this example, DSE successfully quantifies how much set-specific information is captured by the two descriptors.

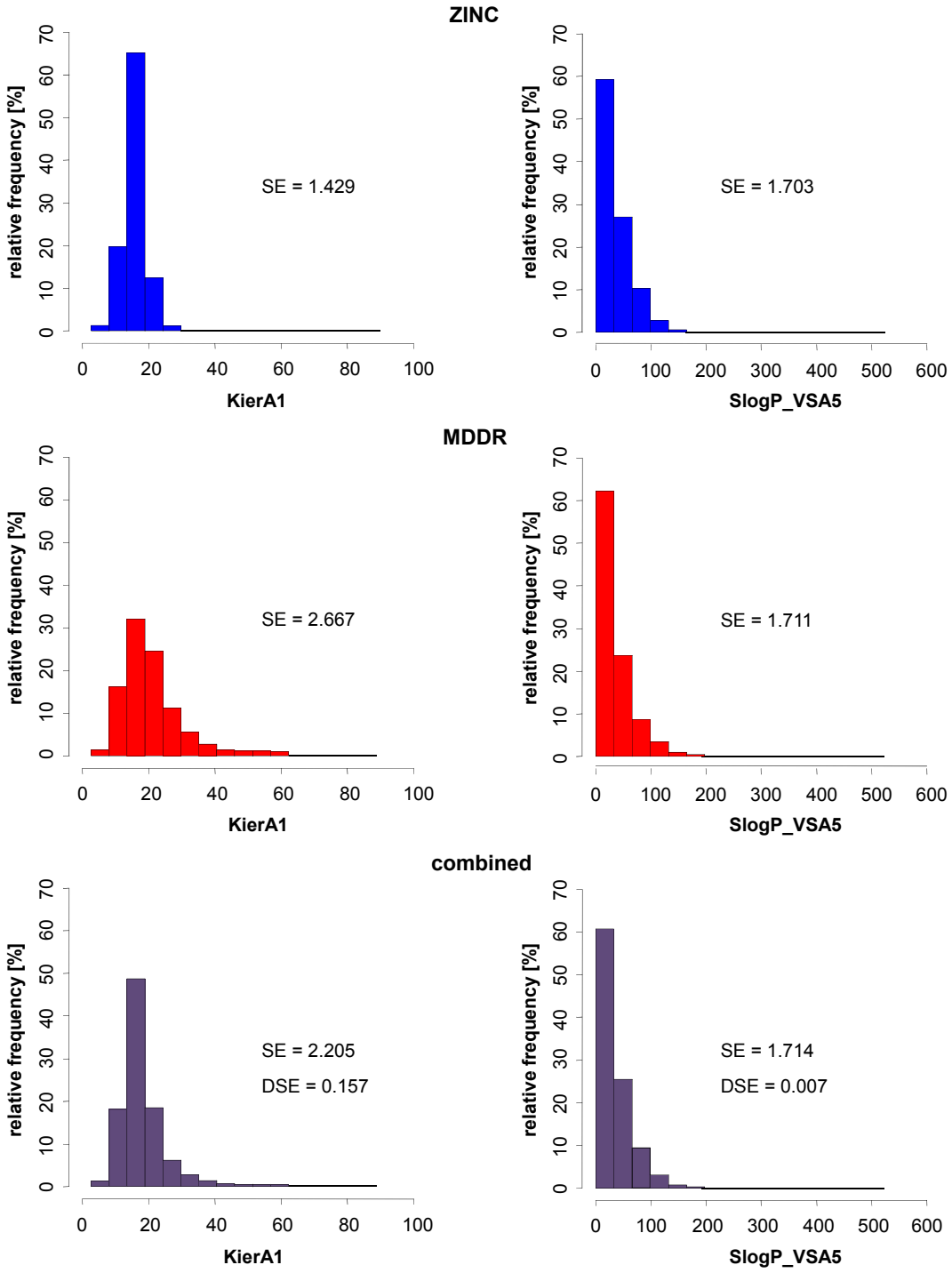


Figure 5.2-3: Assessment of discriminatory power by DSE For the descriptors "KierA1" and "SlogP_VSA5", individual and combined histograms for MDDR and ZINC compounds are shown and corresponding DSE values are reported.

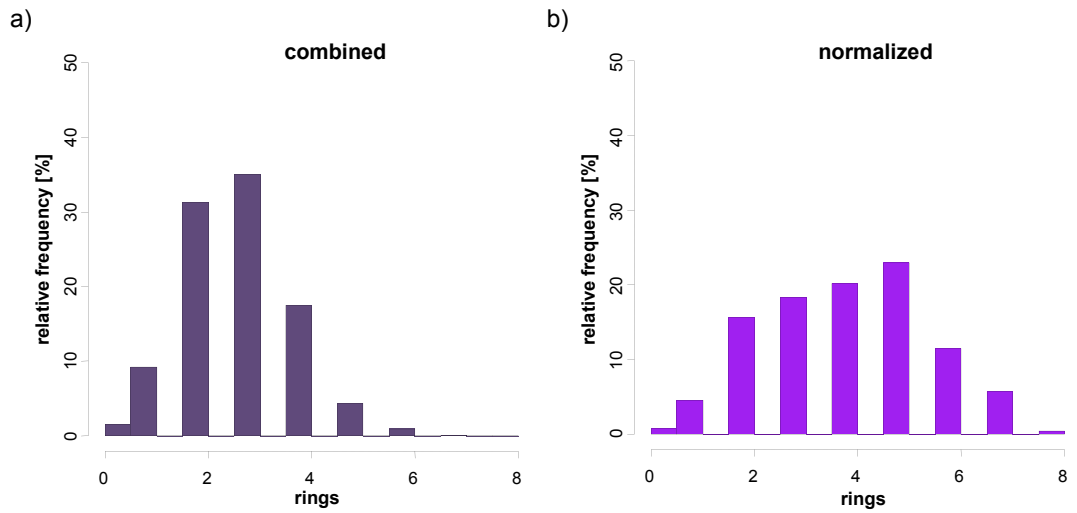


Figure 5.2-4: Combined histograms for DSE and MI-DSE For the descriptor "rings" and the two value distributions shown in Figure 5.2-1, the combined DSE histogram is shown in (a) and the combined MI-DSE histogram in (b).

However, a problem arises if the two compound classes are of significantly different size. In this case, the combined histogram is much influenced by the larger class and its value distribution is biased, as illustrated in Figure 5.2-4a. Although the descriptor "rings" shows distinct value distributions for the exemplary activity class and the ZINC subset (see Figure 5.2-1), the combined histogram reflects the descriptor distribution of the much larger ZINC subset. Hence, $H_{\mathbf{AB}}(D)$ is essentially equal to $H_{\mathbf{B}}(D)$ and equation 5.3 can be reduced to

$$\mathrm{DSE}(D) \approx \frac{H_{\mathbf{B}}(D) - H_{\mathbf{A}}(D)}{2} \tag{5.4}$$

Thus, under these conditions, the magnitude of DSE is mostly determined by descriptors that display high variability in the large compound class but only little variability in the activity class. Then, DSE does no longer quantitatively account for value range dependencies, i.e., the overlap of data distributions, although this is meant to be a key feature of the DSE approach. Therefore, in this case, the method cannot be applied in a meaningful way.

5.3 Mutual Information-DSE

When trying to identify descriptors that contain activity class-specific information, we are always faced with large or very large differences in compound class size. Here, the small class is represented by an activity class and the large class by a database where the vast majority of the compounds do not belong to

the activity class. Therefore, we have developed a descriptor selection approach that is independent of the size of the compound classes under consideration.

In information theory, the concept that quantifies the amount of information about a class of objects (compounds) captured by a descriptor is known as (average) *mutual information* (MI) [100]. MI exactly describes how much information about the class is contained in the value of a descriptor. Formally, MI is defined as the difference between the Shannon entropy of the descriptor for two combined classes and the conditional Shannon entropy of the descriptor given the class:

$$\mathrm{MI}(D, C) = H(D) - H(D|C) \tag{5.5}$$

Here, D is the descriptor and C is the class. $H(D|C)$ quantifies the additional information content of D for class C. For two classes $\mathbf{A}$ and $\mathbf{B}$, $H(D|C)$ is given by

$$H(D|C) = \Pr(C = \mathbf{A}) \times H_{\mathbf{A}}(D) + \Pr(C = \mathbf{B}) \times H_{\mathbf{B}}(D) \tag{5.6}$$

By setting the probabilities $\Pr(C = \mathbf{A}) = \Pr(C = \mathbf{B}) = 0.5$ (which can be seen as an unbiased estimator for the probability that a molecule belongs to either class), the class size dependence of MI is eliminated and the following equation is obtained

$$\mathrm{MI}(D, C) = H(D) - \frac{H_{\mathbf{A}}(D) + H_{\mathbf{B}}(D)}{2} \tag{5.7}$$

Because of the inequality $MI(D, C) \leq H(C) = 1$ (see Appendix D for further information), the calculated MI is normalized to the range [0,1].

We now return to the DSE formalism and the calculation of bin frequencies for combined histograms. Instead of using equation 5.2 where compound classes were weighted according to their size, we calculate the frequencies as follows

$$f_{\mathbf{AB}}(i) = \frac{f_{\mathbf{A}}(i) + f_{\mathbf{B}}(i)}{2} \tag{5.8}$$

On the basis of these frequencies, we can generate the combined histogram of the value distributions of our two compound classes. In the following, we use the term *normalized* for a combined histogram that is calculated based on frequencies calculated according to equation 5.8 instead of equation 5.2. The normalized histogram for the descriptor "rings" is shown in Figure 5.2-4b. In contrast to the original histogram in Figure 5.2-4a, the normalized histogram reflects both the value distribution of the descriptor within the activity class and the screening database. Calculating $H_{\mathrm{norm}}(D)$ from the normalized histogram yields a modified DSE score that exactly corresponds to equation 5.7 and is therefore termed "Mutual Information-DSE" (MI-DSE):

$$\text{MI-DSE}(D) = H_{\mathrm{norm}}(D) - \frac{H_{\mathbf{A}}(D) + H_{\mathbf{B}}(D)}{2} \tag{5.9}$$

This quantity also corresponds to the Jensen-Shannon divergence of descriptor value distributions [101]. The MI-DSE measure has the desired property of yielding normalized scores between zero and one, reflecting the significance of descriptors to capture differential information content. A score of zero indicates that the descriptor distributions for compound classes **A** and **B** are identical and that the descriptor captures no class-specific information, whereas a score of one indicates that the value distributions are fully disjoint and that the descriptor can thus perfectly distinguish between **A** and **B**.

5.4 Applications

MI-DSE was compared to DSE by the evaluation of descriptor ranking and correlation analysis on 168 compound activity classes. Furthermore, top-ranked DSE and MI-DSE descriptors were analyzed for their discriminatory potential.

5.4.1 Descriptor Ranking and Correlation Analysis

To investigate whether MI-DSE and DSE prioritize different descriptors, as would be expected on the basis of our example given in section 5.3, MI-DSE and DSE calculations were carried out for 168 compound activity classes extracted from the ChEMBL database that contained at least 50 inhibitors with minimum potency of 1 µM. Compounds were represented by 171 numerical 1D or 2D descriptors available in MOE, listed in Appendix Table D-1. These descriptors accounted for a number of diverse properties including physicochemical and bulk parameters, atom and bond counts, chemical composition, and surface, topological, or shape properties. Different binning schemes were investigated by dividing the value ranges of all descriptors into 8, 16, 32, or 64 equally-sized bins. For each activity class, all descriptor value distributions were compared to those in a database of 100 000 randomly collected ZINC compounds, MI-DSE and DSE scores were calculated, and the descriptors were ranked in the order of decreasing scores. In order to compare DSE- and MI-DSE-based rankings for an activity class, Spearman correlation coefficients were calculated. The Spearman correlation coefficient provides a measure for the correlation between two data rankings when the values themselves are not of interest, but the relative order they produce. It can be calculated as the Pearson correlation coefficient of corresponding ranking positions. As a control for the choice of binning schemes, DSE-based rankings and MI-DSE-based rankings were first compared among themselves (Appendix Table D-2). The rankings produced with different numbers of bins were highly correlated for the individual methods, reflecting that descriptor ranking was essentially independent of the utilized number of bins. We then compared the DSE and MI-DSE rankings and found that, irrespective of the applied binning scheme, correlations between rankings were in most cases

not detectable or rather low. The average Spearman rank correlation coefficients were 0.151, 0.201, 0.262, and 0.341 for 8, 16, 32, and 64 bins, respectively (hence, with increasing numbers of bins, the rankings became only slightly more similar). This large-scale comparison over many different compound activity classes demonstrated that DSE and MI-DSE calculations produced very different descriptor rankings.

5.4.2 Comparison of Top-Ranked DSE and MI-DSE Descriptors

For each activity class, value distributions for the ten top-ranked DSE and MI-DSE descriptors calculated on 16 equally-sized bins were represented as histograms and compared by visual inspection. The comparison of descriptor value histograms for active and database compounds and resulting DSE and MI-DSE values clearly showed that MI-DSE calculations prioritized descriptors capturing compound class-specific information, much more than DSE calculations. Representative results for two activity classes are shown in Figures 5.4-1 and 5.4-2 (with further information on the descriptor rankings of these activity classes provided in Appendix Table D-3). For antagonists of the muscarinic acetylcholine receptor M2, the most discriminatory descriptor according to MI-DSE is the formal charge ("FCharge") of the molecules. Whereas most active compounds are positively charged, ZINC molecules are predominantly uncharged or negatively charged, as shown in Figure 5.4-1. MI-DSE correctly identifies that the value distributions are mostly disjoint and that the descriptor carries compound class-specific information. As the descriptor has a low SE for the ZINC database and DSE calculations are dominated by the value distribution of the large compound class (vide supra), the descriptor is ranked lowly by the DSE approach. By contrast, the surface area descriptor "SMR_VSA4" is ranked highly on the basis of DSE calculations, but not MI-DSE calculations. Figure 5.4-1 reveals that value distributions overlap for active compounds and the background database. For both sets, descriptor values falling into the first bin are most frequently observed. However, DSE assigns a high score to this descriptor because the SE calculated for the background database is much higher than for the activity class (compare equation 5.4). MI-DSE recognizes that the modes of the two distributions overlap and correctly assigns a considerably lower score than for the formal charge descriptor.

For a set of carbonic anhydrase II inhibitors, the partial charge descriptor "PEOE_VSA-4" is top-ranked by MI-DSE. Corresponding value distributions in Figure 5.4-2 illustrate that much higher descriptor values are obtained for active molecules and that compounds from the two sets mostly populate different bins. Albeit the evident discriminatory nature of the descriptor, DSE calculations yield a negative score because the descriptor values for the carbonic

anhydrase II inhibitors are more widely spread resulting in a higher SE for the activity class than for the large background database. Furthermore, Figure 5.4-2 shows value distributions for the adjacency matrix descriptor "GCUT_-SLOGP_0" that is a top-ranked descriptor for carbonic anhydrase II inhibitors on the basis of both DSE and MI-DSE analysis. This descriptor produces a comparably high SE value for the database and a low SE value for the activity class, resulting in a high DSE value, which is mainly determined by the high SE value for the database. However, the value distributions of this descriptor dis-

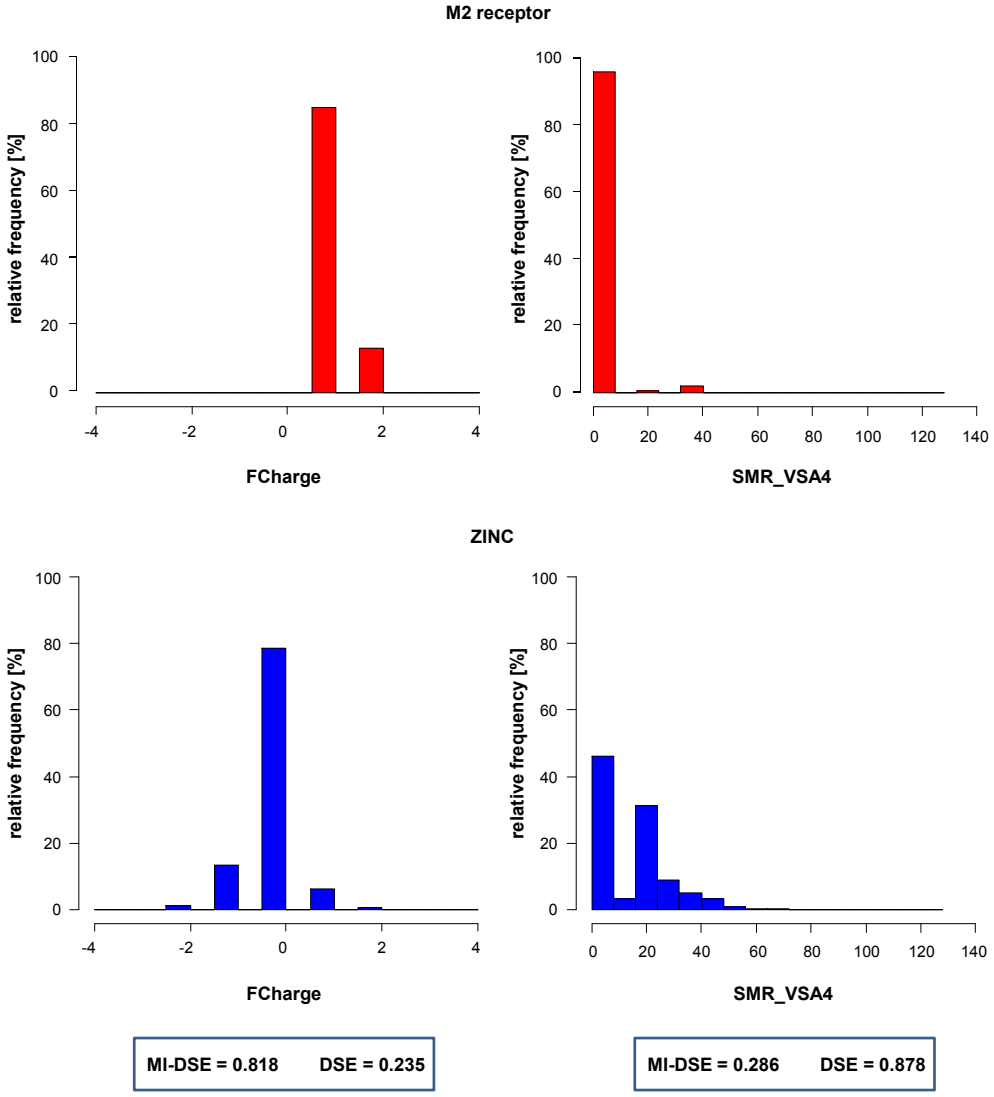


Figure 5.4-1: Descriptor rankings for M2 receptor For the muscarinic acetylcholine receptor M2, value distributions for top-ranked descriptors according to MI-DSE or DSE are shown.

play only limited overlap, which also results in a high MI-DSE value, consistent with the class-specific information captured by the descriptor.

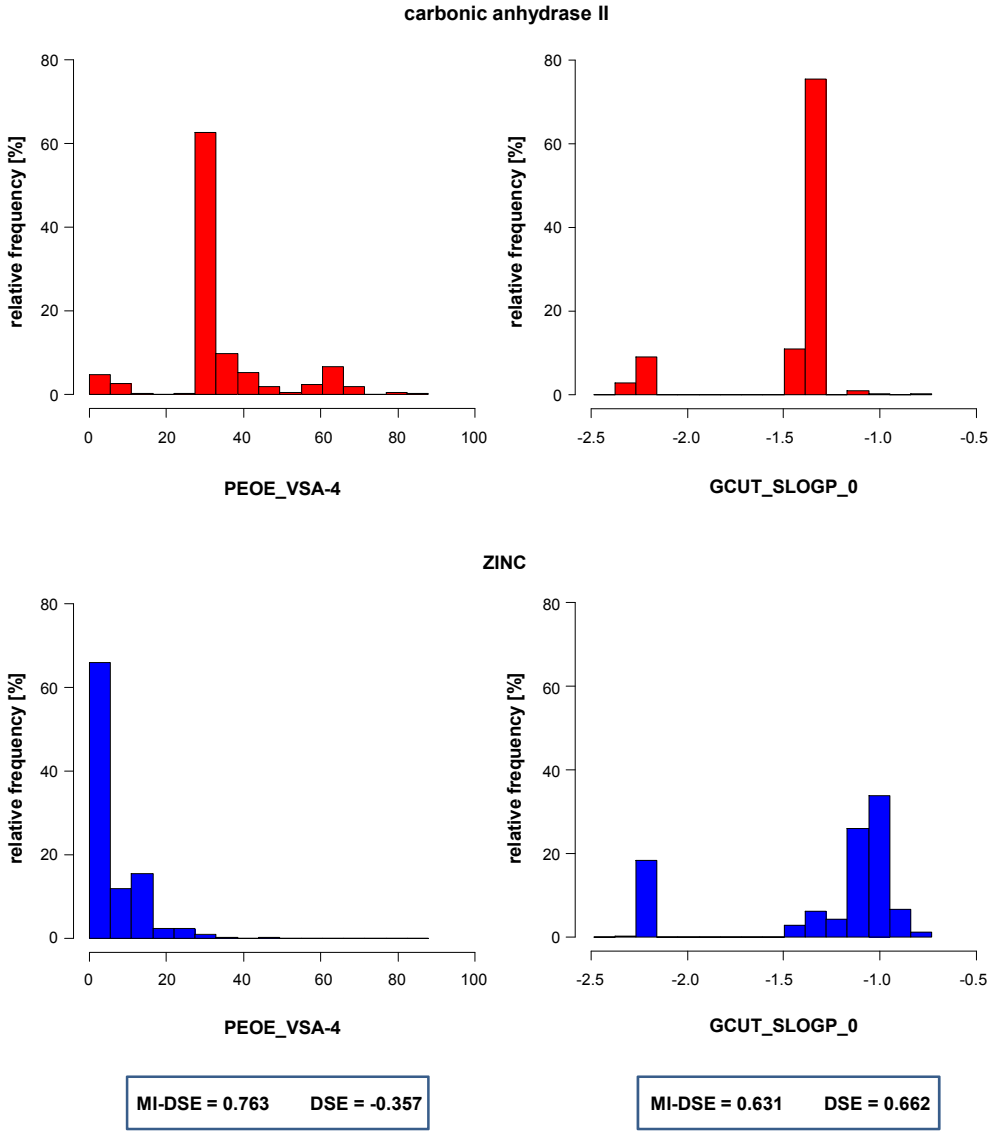


Figure 5.4-2: Descriptor rankings for carbonic anhydrase II For carbonic anhydrase II, value distributions for top-ranked descriptors according to MI-DSE and/or DSE are shown.

5.5 Conclusions

The identification of descriptors that capture compound class-specific information is of high relevance for many chemoinformatics applications. Generally, compound class-specific information must be assessed by comparing sets of compounds having a desired property (such as, for example, biological activity) with data sets where most compounds lack this property. The larger this reference database is, the more reliable the assessment of class-specific information becomes. If only small reference sets are utilized, the analysis is not meaningful. For the study of biological activity, descriptors that systematically differ in their value settings between active and database compounds are highly desired, i.e., descriptors that are activity-relevant. In this chapter, an information-theoretic approach to reliably assess compound class-specific information content of descriptors has been introduced. Importantly, the approach is not biased by intrinsic differences in the size of activity classes and reference databases. This has been accomplished by combining the Differential Shannon Entropy formalism with the mutual information concept. The comparison of value distributions of descriptors and the resulting (DSE and) MI-DSE values and descriptor rankings has confirmed the utility of the MI-DSE approach to identify descriptors containing class-specific information. This newly introduced approach is straightforward and should be useful for large-scale descriptor analysis.

Source Information

Sections of the text in this chapter have been taken from [62, 99].

Chapter 6

Comprehensive Survey of Single- and Multi-Target Activity Cliffs

Molecular similarity-analysis is rooted in the similarity-property principle stating that overall similar molecules should have similar biological activity [4]. Despite the intuitiveness of this principle, it is known that minor structural modifications of an active molecule can dramatically increase or decrease its activity, which corresponds to SAR discontinuity. Structurally similar compounds having large differences in potency form so-termed activity cliffs, as shown in Figure 6.0-1. Therefore, activity cliffs represent the extreme form of SAR discontinuity [28] and their presence in compound sets is often responsible for difficulties in deriving QSAR models for activity prediction [12]. Given their "small structural change – large potency effect" phenotype, activity cliffs are also regarded as the most informative SAR feature in bioactivity-annotated compound data sets. Moreover, the assessment of activity cliffs in compound series plays an important role for chemical optimization efforts.

Activity cliffs have conventionally been analyzed for compounds active against individual targets (single-target cliffs). However, for compounds with activity against multiple targets, multi-target cliffs might also occur [102]. Such cliffs result from differential potency of a compound pair against two or more related targets. These targets might be closely related, i.e., members of the same protein family, or unrelated, if the cliff-forming compounds display polypharmacological behavior [65]. Although activity cliffs are intensely studied, analyses are usually focused on identifying activity cliffs in individual compound sets. We noticed that no systematic assessment of activity cliff distributions had been carried out so far and that it was unknown how activity cliffs are globally distributed across available bioactive compounds and protein targets. Furthermore, it was unclear how frequently multi-target activity cliffs might actually occur in bioactive compounds. To shed light on these questions, we have carried out a systematic analysis of single- and multi-target activity cliffs

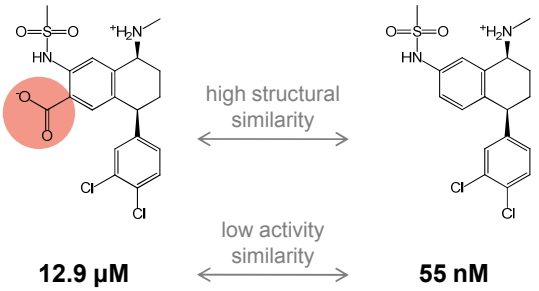


Figure 6.0-1: Activity cliff Shown are two dopamine transporter inhibitors that are structurally very similar but show a large difference in their reported IC_{50} values and therefore constitute an activity cliff.

formed by public domain compounds annotated with activity against human protein targets [103], as reported in this chapter. Employed compound data sets and calculations for similarity-potency comparisons are described in section 6.1. Results of our analysis including separate assessments of global and target family-based activity cliff distributions are presented in section 6.2. The chapter ends with conclusions in section 6.3.

6.1 Data and Calculations

6.1.1 Data Sets

Two major public domain compound repositories, i.e., PubChem BioAssay and BindingDB, were analyzed for the occurrence of activity cliffs. As detailed before, PubChem bioassays contain HTS data, whereas BindingDB predominantly contains compounds taken from the medicinal chemistry literature, mostly originating from chemical optimization efforts. The version of BindingDB used in this analysis has integrated large parts of the ChEMBL compound collection for defined protein targets.

For our analysis, we extracted small compounds consisting of at least five heavy atoms and having a molecular weight of not more than 900 Da. An upper weight threshold was applied because comparisons of large molecules often yield artificially high similarity values [94]. Only K_i or IC_{50} values were considered as activity annotations.

6.1.2 Activity Cliff Calculations

The detection of activity cliffs requires a consistent definition of high structural similarity and activity dissimilarity. For the calculation of pairwise compound similarities, molecules were encoded using ECFP4 representations. As a similarity threshold for activity cliff formation, an ECFP4 Tc value of 0.55 was

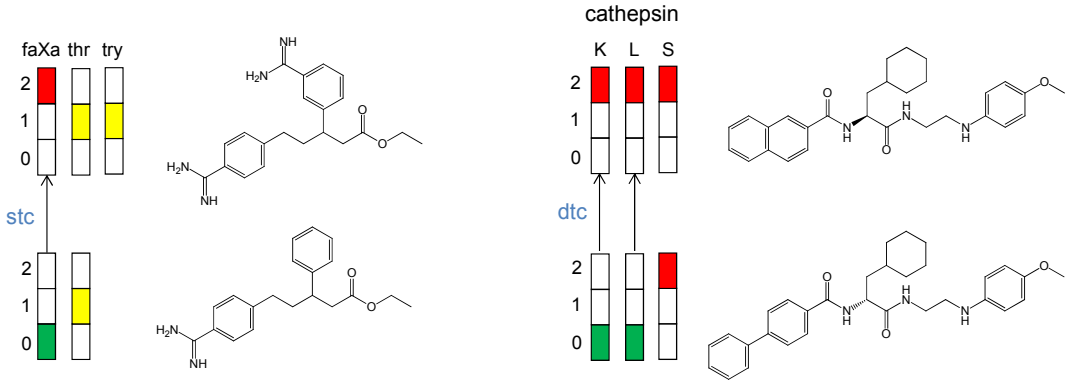


Figure 6.1-1: Single- and dual-target activity cliffs Shown are four compounds and their activity profiles. The two compounds shown on the left are serine protease inhibitors. One of these compounds is active against all three and the other against two of the serine proteases factor Xa (faXa), thrombin (thr), and trypsin (try). Based on their activity profiles, these two compounds form a single-target activity cliff (stc) for factor Xa. The cliff is indicated by an arrow. On the right, two compounds are shown that are active against cathepsins K, L, and S. One compound is highly potent against all three cathepsins, whereas the other is only highly potent against cathepsin S but weakly potent against cathepsins K and L. Hence, these compounds form a dual-target cliff (dtc) for cathepsins K and L, indicated by arrows. The figure is adapted from [103].

applied since it has been shown previously that this ECFP4 Tc threshold value identifies compounds with high structural similarity [104]. For comparison, the calculations were also carried out with another molecular representation, i.e., the MACCS fingerprint. The same number of pairs of similar compounds above the ECFP4 Tc threshold value of 0.55 was obtained for the MACCS Tc calculations when a threshold value of 0.85 was applied.

For all compounds, an activity profile was generated using a constant representation scheme [102]. An activity profile of a compound consisted of binned activity measurements for all annotated targets. Potency values were assigned to three different ranges, i.e., "weakly potent" ($pK_i \leq 5$; bin label "0"), "moderately potent" ($5 < pK_i \leq 7$; bin label "1"), or "highly potent" ($pK_i > 7$; bin label "2"). If multiple measurements were available, a compound was only included in the analysis if all values fell into the same potency bin. For our analysis, we applied the definition that an activity cliff was formed by a pair of compounds that exceeded the defined fingerprint similarity threshold and in which one compound was highly potent against a given target and the other only weakly potent (i.e., representing a "2" vs. "0" potency bin combination against the target). Compounds active against multiple targets can form single- or multi-target activity cliffs. The latter are termed dual-, triple-, quadruple-target cliffs etc., according to the number of targets for which cliffs occur. In Figure 6.1-1, exemplary compound pairs are shown with their activity profiles that form a single- and dual-target cliff, respectively.

To investigate activity cliff frequencies for different target families, the targets in our analysis were grouped together based on the family organization of the protein database UniProt. Targets belonging to the large GPCR 1 family were divided into smaller groups following the protein classification hierarchy available in ChEMBL.

For our analysis, Perl and Java programs were generated.

6.2 Results

A total of 164 165 unique BindingDB compounds were obtained (approximately 85% of which originated from ChEMBL and were subsequently integrated into BindingDB) that were reported to be active against 1 355 non-redundant individual human targets. These compounds yielded 330 526 defined activity annotations (i.e., many compounds were active against multiple targets). From PubChem, 187 confirmatory inhibition assays for human targets were extracted that contained 21 532 active compounds with 30 805 defined annotations against 98 different targets. We systematically searched these compound data sets for single- and multi-target activity cliffs.

6.2.1 Global Activity Cliff Distribution

The activity cliff distribution for BindingDB compounds is reported in Table 6.2-1. When, as an approximation, both K_i and IC_{50} values were considered as potency annotations, 36 063 single-, 1 654 dual-, and 233 triple-target activity cliffs were obtained. The number of multi-target cliffs of higher degrees (target numbers) rapidly declined, although cliffs involving up to seven targets were detected. Table 6.2-1 also reports the corresponding activity cliff distribution when only directly comparable K_i values were considered as measurements. In this case, 10 063 single-, 330 dual-, and 61 triple-target activity cliffs were de-

Table 6.2-1: Activity cliff statistics

activity type	degree							directionality		polypharmacological	all
	1	2	3	4	5	6	7	dir.	undir.		
K_i/IC_{50}	36 063	1 654	233	43	29	21	2	38 019	26	79	38 045
K_i	10 063	330	61	17	2	0	0	10 469	4	4	10 473

Activity cliff statistics are reported for the K_i/IC_{50}- and K_i-based analyses. "degree" denotes the number of targets per activity cliff. "all" gives the sum of single- and multi-target cliffs. Under "directionality", the number of directed ("dir.") and undirected ("undir.") multi-target cliffs is reported. In addition, the number of polypharmacological cliffs ("polypharmacological") is given.

tected (i.e., approximately one fourth of the cliffs found when both K_i and IC_{50} values were considered). Furthermore, 17 cliffs involving four targets and two cliffs involving five targets were identified. Relating the number of compounds forming activity cliffs to the total number of compounds in the K_i/IC_{50}- and K_i-based data sets, we found that 12.5% and 10.9%, respectively, of compounds were involved in the formation of activity cliffs.

We also determined the total number of compound pairs that could potentially form activity cliffs, i.e., pairs that exceeded the fingerprint similarity threshold for activity cliffs. When both K_i and IC_{50} measurements were considered, 1 530 493 qualifying compound pairs yielded a total of 38 045 (single- and multi-target) cliffs. Hence, only 2.5% of all qualifying compound pairs formed activity cliffs and only 5.2% of these compound pairs formed multi-target cliffs. When only K_i values were considered, 574 851 compound pairs were found that yielded a total of 10 473 cliffs, i.e., only 1.8% of these pairs formed activity cliffs and only 3.9% of these were multi-target cliffs. Thus, activity cliffs were only sparsely distributed among pairs of structurally similar compounds. Control calculations using the MACCS fingerprint yielded very similar statistics and are therefore not discussed further.

The active compound pool extracted from PubChem amounted to approximately one seventh of the size of BindingDB. These screening hits had overall lower potency than BindingDB compounds, as expected, and were structurally more diverse. When K_i and IC_{50} values were taken into account, only 13 single-target and no multi-target cliffs were detected. These 13 activity cliffs involved only five different targets. Thus, the occurrence of activity cliffs in PubChem compounds was negligible. Hence, the further analysis of activity cliffs was limited to the BindingDB compound collection.

6.2.2 Target Family Distribution

We then studied the protein target family distribution of all activity cliffs. The results for the K_i/IC_{50}- and K_i-based distributions are provided in Table 6.2-2. For the top ten families of each analysis, ranked according to the total number of activity cliffs, significant differences in cliff numbers were observed. However, on a relative scale (with respect to the total number of pairs of similar compounds), activity cliffs were similarly distributed over these target families. The top ten families included popular therapeutic targets for which many qualifying active compound pairs were available. Most activity cliffs were found for ligands of the short peptide receptor and peptidase S1 families. The family rankings differed for the K_i/IC_{50}- and K_i-based cliff distributions, but in both cases protease, kinase, nuclear hormone receptor, and GPCR families were found among the top ten families. For most highly-ranked families, the percentage of multi-target cliffs among all activity cliffs was small (i.e., 0% to less than 10%), although

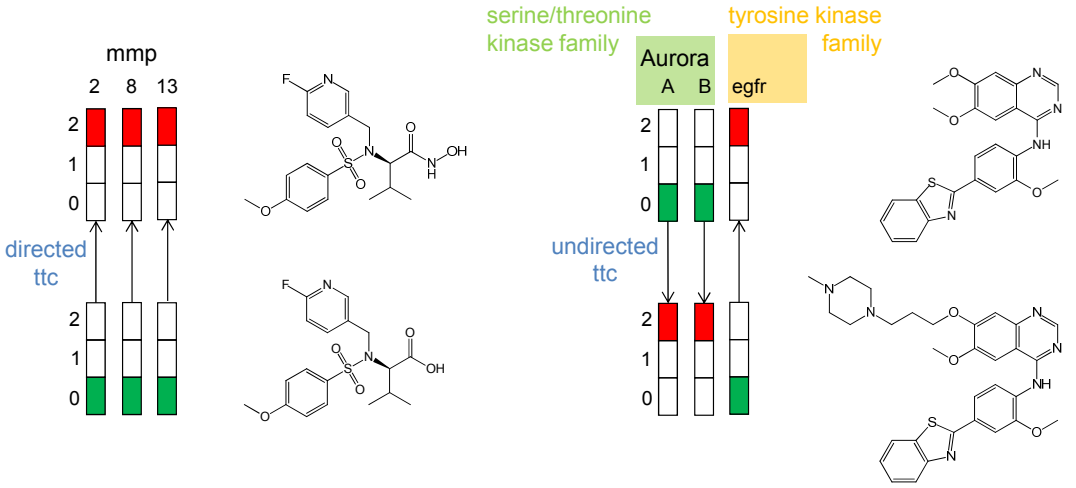


Figure 6.2-1: Directed and undirected activity cliffs On the left, a compound pair is shown that forms a triple-target cliff (ttc) for matrix metalloproteases (mmp) 2, 8, and 13. The potency of one compound is consistently high against all three targets, and the potency of the other is consistently low, i.e., this cliff is directed. On the right, a compound pair is shown that forms a triple-target cliff for Aurora serine-threonine kinases A and B and the tyrosine kinase epidermal growth factor receptor (egfr). In this case, the compounds show differential target selectivity. One compound is highly potent against Aurora kinases A and B and weakly potent against the epidermal growth factor receptor kinase, whereas the other compound displays an inverse activity profile. Accordingly, this triple-target cliff is undirected. Furthermore, as the three kinases belong to two different target families, a polypharmacological cliff is formed by the two compounds. The figure is adapted from [103].

there were exceptions. For example, for the AGC serine/threonine kinase and the peptidase M10A (matrix metalloproteases) families, 24.1% and 17.5% of all activity cliffs were multi-target cliffs, respectively, and for the peptidase C1 family, 13.0% were multi-target cliffs.

6.2.3 Activity Cliff Directionality

Then we analyzed the directionality of multi-target activity cliffs. In a "directed" multi-target cliff pair, the potency of compound A is consistently high for all targets and the potency of compound B is consistently low. By contrast, in an "undirected" multi-target cliff pair, compound A has high potency for at least one target for which compound B is only weakly potent and vice versa. Differences in cliff directionality are illustrated in Figure 6.2-1 where exemplary directed and undirected triple-target cliffs are shown. Importantly, only undirected multi-target activity cliffs contain compounds with different target selectivity. In the K_i/IC_{50}-based distribution, only 26 of 1 982 multi-target cliffs (1.3%) were undirected, and in the K_i-based distribution, only 4 of 410 (1.0%). Thus, nearly all multi-target cliffs were directed and activity cliff-forming compounds with different target selectivity were extremely rare.

Table 6.2-2: Target family distribution of activity cliffs

target family	degree							mt (%)	all	freq. (%)	#targets
	1	2	3	4	5	6	7				
K_i/IC_{50}											
short peptide receptor	6 657	161	17	1	0	0	0	2.6	6 836	2.6	37
peptidase S1	4 006	68	5	3	1	0	0	1.9	4 083	3.4	21
tyrosine kinase	2 656	192	37	0	0	0	0	7.4	3 085	2.9	29
peptidase A1	1 610	13	6	0	0	0	0	1.2	1 629	2.7	69
prostaglandin G/H synthase	1 585	39	0	0	0	0	0	2.4	1 624	7.6	2
AGC serine/threonine kinase	927	162	71	18	22	21	0	24.1	1 221	4.3	16
peptidase M10A	984	152	32	19	4	0	2	17.5	1 193	2.6	9
CMGC serine/threonine kinase	1 052	75	0	0	1	0	0	6.7	1 128	2.5	12
nuclear hormone receptor	899	113	0	0	0	0	0	11.2	1 012	2.2	17
nucleotide-like ligand receptor	955	26	4	0	0	0	0	3.0	985	1.3	6
K_i											
short peptide receptor	3 203	75	9	0	0	0	0	2.6	3 287	2.6	32
peptidase S1	1 732	18	0	3	1	0	0	1.3	1 754	2.4	18
monoamine receptor	677	39	20	2	1	0	0	8.4	739	0.9	24
nucleotide-like ligand receptor	630	30	0	0	0	0	0	4.5	660	0.6	5
carbonic anhydrase	533	15	0	0	0	0	0	2.7	548	3.6	7
peptidase C1	457	58	10	0	0	0	0	13.0	525	9.1	5
G protein-coupled receptor 2	446	0	0	0	0	0	0	0.0	446	3.4	2
lipid-like ligand receptor	406	5	0	0	0	0	0	1.2	411	1.6	12
nuclear hormone receptor	200	8	0	0	0	0	0	3.8	208	1.6	11
TKL serine/threonine kinase	198	0	0	0	0	0	0	0.0	198	28.5	1

The target family distribution of activity cliffs is reported for the K_i/IC_{50}- and K_i-based analyses, respectively. In each case, the top ten target families are ranked according to the sum of single- and multi-target cliffs they cover ("all"). The percentage of multi-target cliffs ("mt (%)") among all activity cliffs is also reported. Furthermore, for each family, the number of compound pairs that form activity cliffs divided by the number of qualifying pairs of similar compounds and the number of targets for which activity cliffs occur are reported in the columns "freq. (%)" and "#targets", respectively.

6.2.4 Polypharmacological Cliffs

In addition, we also searched for what we regard as "polypharmacological cliffs", i.e., multi-target activity cliffs that involved targets belonging to different protein families. Here, only targets with unambiguous family assignments according to UniProt or ChEMBL were considered. For the K_i/IC_{50}-based distribution, we found 79 polypharmacological cliffs that involved a total of 84 compounds. Seventy-one of these cliffs were dual- and eight triple-target cliffs. For the K_i-based distribution, we identified only four (dual-target) polypharmacological cliffs involving seven compounds. Hence, compounds with activity against different target families displayed only limited polypharmacological cliff potential.

6.3 Conclusions

In this chapter, a comprehensive analysis of activity cliffs formed by currently available bioactive compounds was reported. We have searched for cliffs of large magnitude that are in general least affected by measurement inaccuracies and that usually provide focal points of SAR exploration. Furthermore, in our analysis, we have differentiated between single- and multi-target activity cliffs, studied the directionality of multi-target cliffs, and also introduced polypharmacological cliffs that are a special type of a multi-target cliff. In general, single-target cliffs occurred much more frequently than multi-target cliffs and were similarly distributed over different target families. We also found that compounds having different target selectivity only rarely occurred in multi-target cliffs. Although the percentage of qualifying compound pairs that formed activity cliffs was only about 2%, on average more than 10% of compounds active against different target families were involved in the formation of large-magnitude activity cliffs. Thus, for an active compound of interest, a thorough search of its structural neighbourhood is rather likely to reveal activity cliffs from which SAR determinants might be deduced.

Source Information

Sections of the text in this chapter have been taken from [103].

Chapter 7

Relating Molecular Transformations to Potency Effects

In lead optimization, the structure of active compounds is changed by small chemical modifications in the attempt to further improve compound potency and/or in vitro (and in vivo) properties. Traditionally, the design of analogs has largely depended on medicinal chemistry expert knowledge, experience, and intuition [105]. Without doubt, this subjective and expertise-driven approach has to this date been largely responsible for the success of medicinal chemistry programs, despite the increasing use of high-throughput technologies in drug discovery settings. However, attempts have also been made to more systematically address the process of analog design in order to aid medicinal chemists in the decision which compounds to synthesize next in the course of a compound optimization effort [106, 107]. By and large, these attempts have focused, and continue to focus, on analyzing the wealth of available analog and structure-property relationship data for diverse targets in order to identify, and ultimately predict, favorable chemical modifications. For this purpose, compounds are compared in a pairwise manner in order to identify structural changes that are subsequently related to changes in compound properties.

In this chapter, three studies are reported that systematically relate chemical replacements to potency changes [108–110]. Section 7.1 introduces the concept of matched molecular pairs (MMPs) that provides a consistent structural reference frame to generalize chemical modifications. Moreover, a computationally efficient algorithm for the identification of MMPs is described. In section 7.2, the application of the MMP concept to systematically analyze the ability of defined chemical changes to introduce activity cliffs is reported and properties of structural modifications frequently leading to large potency changes are discussed. Sections 7.3 and 7.4, on the other hand, report the exploration of potency-retaining or bioisosteric replacements. Section 7.3 focuses on replacements that act as bioisosteres across different targets and are conservative with

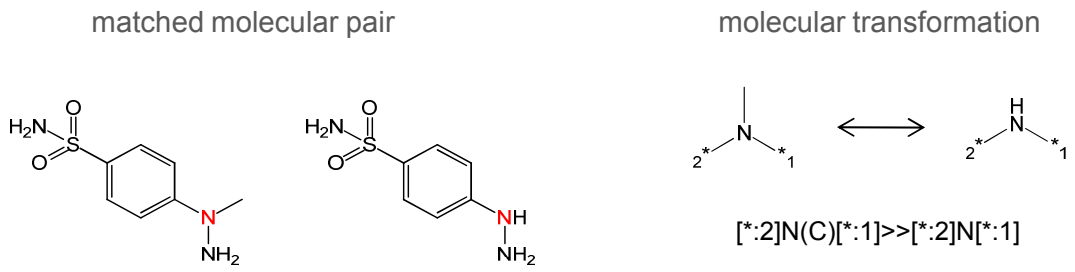


Figure 7.1-1: Matched molecular pair and molecular transformation Shown is a pair of compounds that have similar structures and differ only at a localized site. The fragment exchange relating the compounds to each other is referred to as molecular transformation and encoded as SMIRKS string [111].

respect to different biological activities. In addition, section 7.4 addresses the question whether bioisosteric replacements can be found that preferentially act against a given target family. The chapter closes with a comparison of the chemical replacements identified in the different studies and a summary of key observations in section 7.5.

7.1 Matched Molecular Pairs and Molecular Transformations

An MMP is defined as a pair of compounds that differ only at a single localized site and are distinguished by a defined substituent or molecular fragment [39]. It follows that two compounds forming an MMP are related to each other by a specific molecular transformation, such as the addition of a substituent or the exchange of a ring system. Figure 7.1-1 shows an MMP and the molecular transformation that converts one compound into the other. Matched molecular pair analysis (MMPA) generally aims at identifying all MMPs from a set of compounds and determining associated property changes [112]. MMPA is highly attractive for medicinal chemistry due to the chemical interpretability of the results. However, most search algorithms introduced for comprehensively identifying MMPs in compound sets are computationally expensive and not applicable to large data sets. For example, this is the case for maximum common subgraph-based algorithms that carry out pairwise comparisons to find the largest substructure common to two molecules and then analyze whether differing fragments constitute a single-point change [107,113]. Only recently, Hussain and Rea have introduced an efficient algorithm to systematically extract MMPs from compound data sets [114], thus enabling a large-scale analysis of MMP distributions.

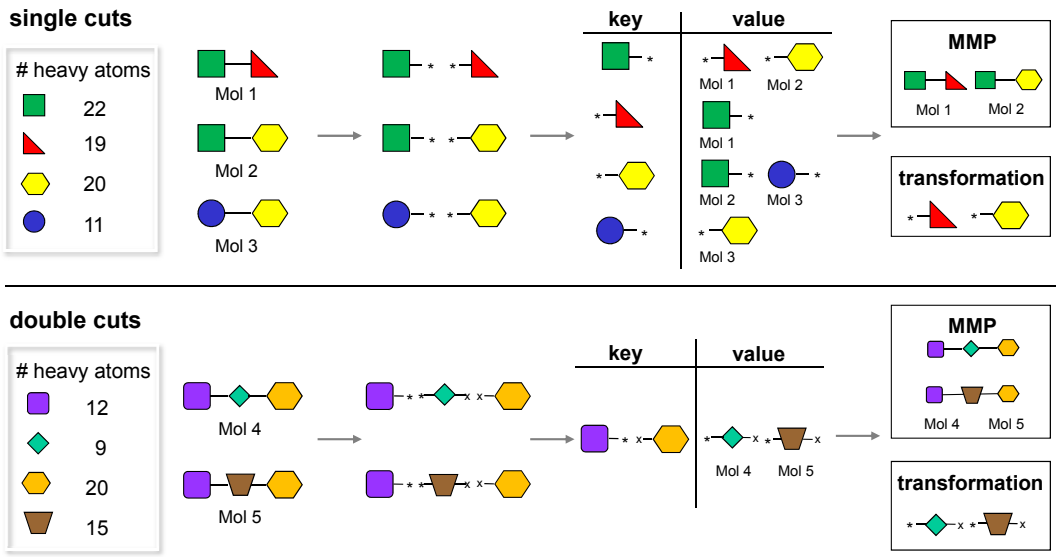


Figure 7.1-2: MMP search algorithm The steps of the Hussain and Rea algorithm leading to the identification of MMPs are illustrated for single and double cuts. Although molecules 2 and 3 in the single cut example share a common key fragment, the pair is not accepted as an MMP because the heavy atom counts of the two value fragments differ by more than eight atoms. The figure is adapted from [108].

7.1.1 Hussain and Rea Algorithm

The first step of the Hussain and Rea algorithm for the identification of MMPs involves fragmentation of all compounds in a data set. This is accomplished by marking all non-ring single bonds between two non-hydrogen atoms in a molecule, followed by systematic deletion of these bonds ("single cut") and their two- and three-bond combinations, called "double cuts" and "triple cuts", respectively. A single cut results in two fragments F1 and F2 that are added to an index list. Two "key-value" pairs are built: fragment F1 is added as a "key" to the index with F2 as the corresponding "value" and vice versa. Double cuts result in a core and two terminal fragments. In this case, the core is considered the value and the two terminal fragments together constitute the key. Of all possible triple cuts, only those are considered that result in a single core and three terminal fragments (i.e., triple cuts that result in two cores and two terminal fragments are not considered). In analogy to double cuts, the core is stored as the value with the three terminal fragments together as the key. For double and triple cuts, connectivity information of the fragments is retained. Furthermore, for all generated key-value pairs, source compound information is stored. Because all compounds sharing a particular key contain the corresponding fragment(s), the pairwise combination of these compounds yields MMPs and the two value fragments define the structural transformation for each MMP. The chemical replacement is denoted using the syntax of the reaction transform lan-

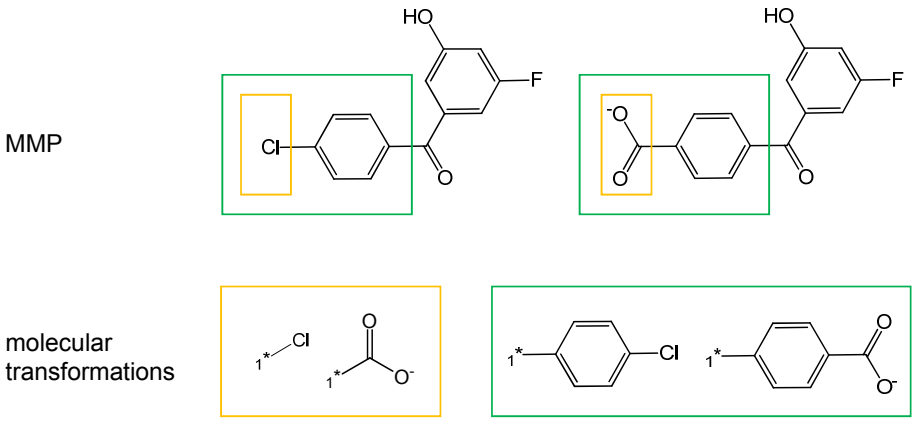


Figure 7.1-3: Structural context of actual substructure exchanges For the MMP shown on the top, multiple value fragment pairs of different sizes are obtained that represent alternative structural transformations.

guage SMIRKS [111], which essentially corresponds to the SMILES strings of the two value fragments connected by ">>" and is illustrated in Figure 7.1-1. A schematic outline of the algorithm is provided in Figure 7.1-2. It should be noted that the algorithm often yields multiple, differently sized fragments that define the transformation of an MMP. For example, the compounds shown in Figure 7.1-3 are inter-convertible by a "carboxylate to chlorine" transformation. However, another valid transformation for this compound pair is a "p-benzoate to p-chlorophenyl" transformation. Hence, for each MMP, transformations describing the structural context of the actual substructure change to different extents can be derived. Another strength of the algorithm is that the search for MMPs is comprehensive and not limited to predefined functional groups or substrucutres. By utilizing single and multiple cuts, transformations involving both R-groups and core structures can be assessed.

In the implementation of the algorithm used to carry out the studies reported in this chapter, a combination of two compounds was only considered an MMP if the heavy (non-hydrogen) atom counts of their distinguishing fragments differed by maximally eight atoms. This criterion was applied to ensure that compounds forming MMPs did not have large differences in size and is illustrated in the algorithm scheme in Figure 7.1-2. Furthermore, transformations relating compounds in a pair were canonicalized such that MMPs could be grouped by transformations. To be included in our analyses, transformations had to occur in at least two compound pairs in which the remaining molecular substructure (that was not involved in the transformation) contained at least as many heavy atoms as the exchanged fragment, thereby avoiding the identification of overly large or specific transformations. Moreover, stereochemistry information of molecules was not considered because this information would

often be lost during molecular fragmentation. The algorithm was implemented using Perl and the Scientific Vector Language[1].

7.2 Activity Cliff-Introducing Chemical Replacements

We have been interested in exploring the potential of defined chemical changes to introduce activity cliffs in compound data sets. In particular, we wished to determine whether defined chemical transformations exist that display a general tendency to introduce activity cliffs across different compound classes and biological targets. If so, this information might be very helpful to evaluate chemical modifications for compound optimization efforts. To address this question, we extracted compound potency data from the ChEMBL database (bioassay repositories such as PubChem BioAssay were not suitable for this analysis, as demonstrated in the previous chapter), systematically identified MMPs and molecular transformations, and finally related structural and potency changes to each other. Defined criteria were applied to identify transformations most frequently introducing activity cliffs.

7.2.1 Compound Data Sets

From the ChEMBL database, compounds with potency measurements against human targets were extracted. Whenever available, K_i values were selected (otherwise IC_{50} values). For compounds with multiple potency values reported against the same target, the arithmetic mean was calculated to yield the final potency. Furthermore, only measurements with a very high confidence for direct compound-target interactions were selected. Very large molecules with more than 45 non-ring single bonds were filtered out prior to MMPA because their fragmentation would have been computationally demanding. Ligand sets were assembled for individual targets when at least five active compounds were available that met our criteria, leading to the selection of a total of 33 497 unique active compounds organized into 523 target-specific sets containing between 5 and 1 528 ligands. Compounds were grouped by targets and MMPs were separately identified for each ligand set because molecular transformations were associated with potency changes that must not be calculated for compounds with different bioactivities.

[1]The Scientific Vector Language is an integral part of MOE.

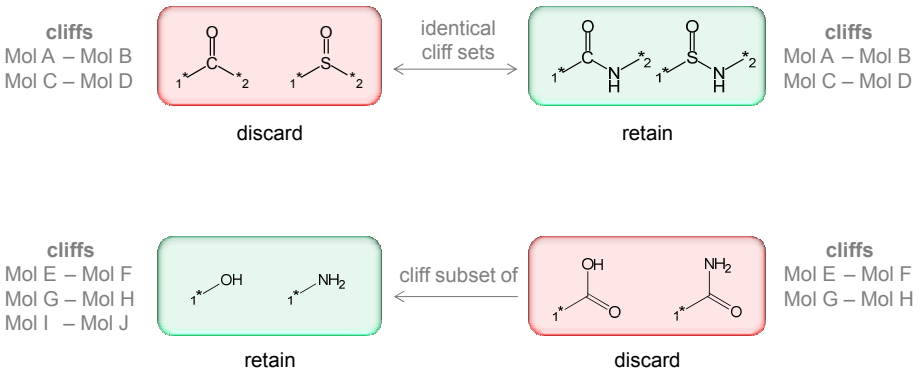


Figure 7.2-1: Redundancy rules In order to unambiguously define frequent cliff-forming transformations, rules are applied that relate the size of exchanged fragments and their number of occurrences in cliff-forming molecule pairs to each other.

7.2.2 Transformation Selection Criteria

For each valid transformation identified by our modified version of the Hussain and Rea algorithm, corresponding MMPs were assembled and logarithmic potency differences between compounds in each pair were recorded. If an MMP was present in more than one ligand set, i.e., if it was annotated with potency values for more than one target, potency differences for all targets were calculated. In the following, the term "potency record" refers to the potency difference calculated for one MMP-target combination. An MMP was considered to form an activity cliff if the potency values of its compounds differed by at least two orders of magnitude. In order to identify chemical replacements with a high propensity to introduce activity cliffs, the following filter criteria were applied:

1) At least 10% of all potency records collected for a transformation represent activity cliffs (this threshold is approximately five-fold higher than the general frequency of occurrence of activity cliffs reported in the previous chapter).

2) The transformation introduces activity cliffs for at least four different targets.

3) The activity cliffs for the transformation are formed by at least four different MMPs.

Because several transformations might generate the same MMP, as illustrated in Figure 7.1-3, we apply the following rules to transformations passing our three filters, thereby clearly defining cliff-forming transformations and avoiding redundancies:

i) If several transformations yield the same set of cliff-forming MMPs, then the largest substructural exchange in terms of heavy atoms is retained. For

example, if the exchange of a carbonyl against a sulfinyl group (transformation 1) always occurrs in the context of an amide versus sulfinamide replacement (transformation 2) in MMPs constituting activity cliffs, only transformation 2 involving the larger number of heavy atoms is retained, as illustrated in Figure 7.2-1 (top). The application of this rule (rule 1) ensures that potential context dependence of structural changes is taken into account.

ii) If a transformation accounts for more cliff-forming compound pairs than a larger substructural exchange and if it includes all cliff-forming compound pairs of the larger fragment pair, then this transformation is selected instead. For example, if both the exchange of a hydroxyl against an amino group and the exchange of a carboxyl against an amide group are identified as activity cliff-introducing transformations, but the more general, i.e., smaller transformation accounts for more activity cliffs, then only this transformation is retained, as depicted in Figure 7.2-1 (bottom). This rule for MMP subset relationships (rule 2) complements rule 1 that is applied for transformations with identical MMP sets. Hence, the two redundancy rules were independently applied for different sets of transformations and not in a sequential manner. Therefore, the structural context dependence of transformations accounted for through the application of rule 1 was not removed by rule 2.

As transformations that occurred in different structural contexts (i.e., compound pairs of various chemotypes) were of high interest, for MMPs representing the same transformation molecular scaffolds were calculated according to Bemis and Murcko [115]. These scaffolds or molecular frameworks correspond to all ring systems and atoms on the direct path connecting two ring systems (linker) in a molecule, as illustrated in Figure 7.2-2. Hence, these scaffolds were obtained from compounds by removal of all side chains from rings and linkers, utilizing a Pipeline Pilot implementation.

7.2.3 Results

The major aim of our analysis has been to determine whether chemical changes exist that generally affect compound potency, i.e., across different compound classes and targets. For the purpose of our analysis, we have applied the MMP concept to explore whether defined chemical changes exist that frequently introduce activity cliffs in different ligand sets. For 513 of our 523 ligand sets, at least one MMP was retrieved. Overall, 323 075 different valid transformations were identified that corresponded to 149 563 different MMPs (the same MMP was represented by multiple transformations) and 29 565 unique compounds (compounds often participated in multiple MMPs).

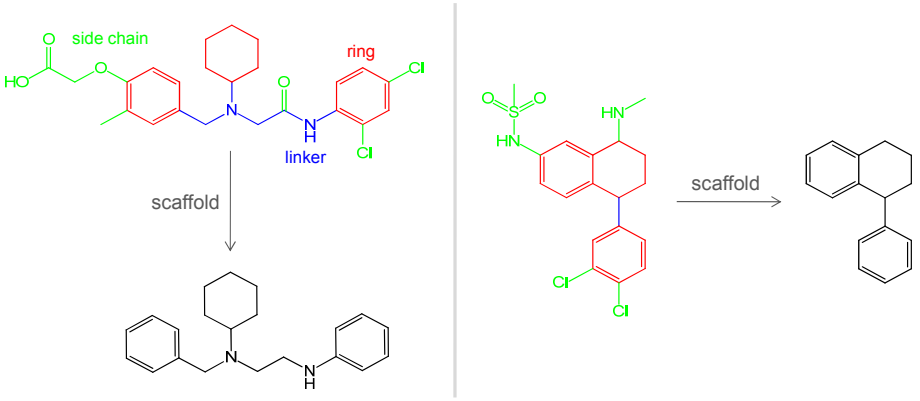


Figure 7.2-2: Molecular scaffolds Two compounds and their scaffolds calculated according to Bemis and Murcko are shown. Scaffolds are obtained by side chain removal and correspond to all rings and linkers found in a molecule.

7.2.3.1 Activity Cliff Statistics for Molecular Transformations

We considered transformations to form an activity cliff if compounds in corresponding MMPs displayed an at least 100-fold difference in potency. Initially, 77 052 cliff-forming transformations were detected. For all of these transformations, the relative frequency with which they introduced cliffs was calculated, i.e., the total number of cliffs they formed divided by their number of potency records. Applying frequency-, target-, and compound pair-oriented filters described in paragraph 7.2.2, the number of transformations was substantially reduced to 487. Finally, the application of the two redundancy rules yielded 146 non-redundant "frequent cliff formers". SMIRKS representations for these transformations, which provided the basis for our further analysis, are given in Appendix Table E-1 and a subset of frequent cliff formers is shown in Figure 7.2-3. The frequency with which they formed activity cliffs ranged from 10% to 83.3%, with a median frequency of 16.3%. Individual transformations occurred in up to 43 cliff-forming MMPs and 26 ligand sets, respectively. As the same MMP might form activity cliffs for multiple target, a single transformation formed a maximum of 47 cliffs. The calculation of corresponding medians yielded five MMPs, five ligand sets, and eight cliffs.

7.2.3.2 Transformation Characteristics

For the set of frequent cliff formers, fragment pairs can roughly be divided into two groups, i.e., (i) non-polar fragments of different size and (ii) fragments that differ in their charge or H-bonding properties. A particular enrichment of fragment pairs was observed where one of the exchanged fragments was negatively charged. Among these, transformations involving a carboxyl group were prevalent. We found that the exchange of a carboxyl group with a variety of other

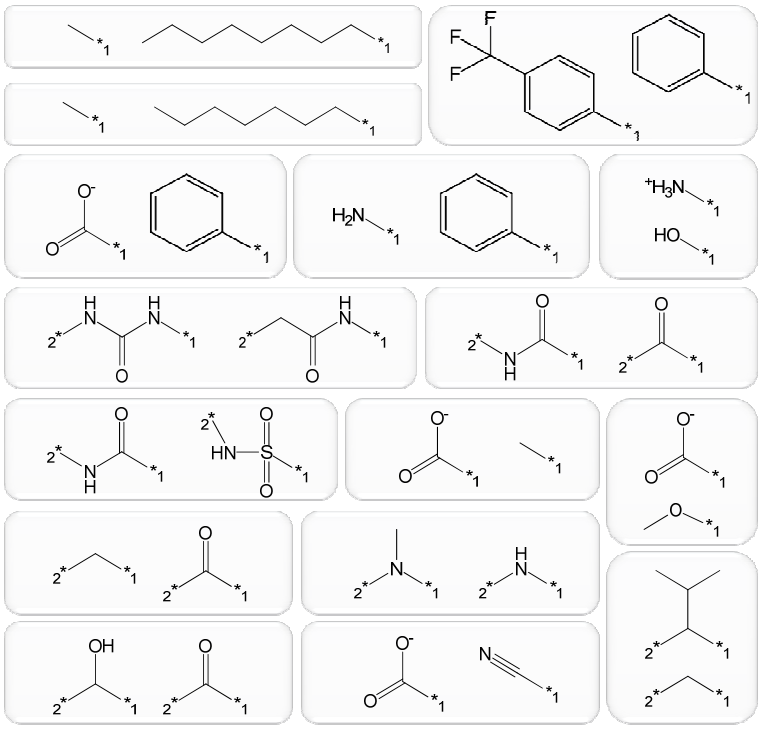


Figure 7.2-3: Frequent cliff formers Shown is a representative subset of structural transformations that frequently introduce activity cliffs.

groups (e.g., halogens, methyl, methoxy, phenyl, or cyano groups) frequently introduced activity cliffs. Similar but weaker tendencies were also detected for transformations involving positively charged nitrogens in different structural environments. Other substructures that were often involved in the formation of activity cliffs included the carbonyl group and amines. By contrast, although changes at ortho, meta, and para substituent positions of phenyl rings were frequently observed, these positional changes rarely introduced activity cliffs. Moreover, a number of very similar transformations were found to display substantially different propensities to introduce activity cliffs, as shown in Figure 7.2-4. For example, introducing a fluorine atom at the para position of a phenyl ring led to an activity cliff in only 5 of 501 cases. However, the relative frequency of cliff formation was found to increase with the size of the halogen substituent, with iodine substitutions introducing cliffs with a nearly 17% frequency. Furthermore, the introduction of a secondary hydroxyl group showed a lower cliff frequency than the introduction of a carbonyl oxygen, another rather unexpected finding. In addition, the exchange of a carbonyl group and a carboxyl ester group displayed a significantly lower tendency to cause large potency changes than the exchange of a carbonyl group and an amide moiety. We also observed that distributions of compound potency differences in MMPs shifted towards high values for transformations that displayed an increasing propensity to introduce activity cliffs, as illustrated in Figure 7.2-5.

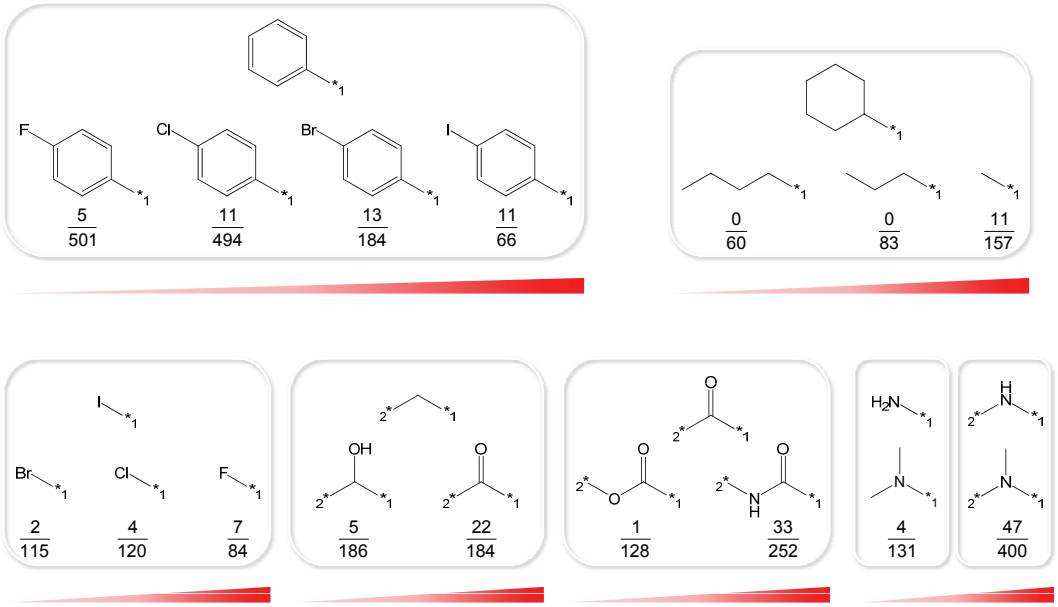


Figure 7.2-4: Transformations with different activity cliff potential Structurally similar transformations are shown together with their relative frequency of cliff formation. If more than two fragments are shown in a box, then the fragment at the top is exchanged with each of the fragments at the bottom. From left to right, transformations are arranged in the order of increasing cliff-forming potential. The figure is adapted from [108].

Thus, transformations that frequently introduced activity cliffs also displayed the tendency to introduce cliffs of large magnitude.

7.2.3.3 Target and Chemotype Distributions of Transformations

In our analysis, transformations that introduced activity cliffs for diverse targets and in different structural contexts were of most interest. Therefore, we analyzed whether the 146 identified transformations preferentially introduced activity cliffs for related targets or different target classes and to what extent frequent cliff formers introduced activity cliffs in different chemotypes. For the target distribution analysis, the protein target classification hierarchy of ChEMBL was adopted and slightly extended to group all targets covered by our ligand sets. For each of the 146 transformations, the number of targets for which it produced activity cliffs was determined and divided by the number of different groups to which these targets belonged. The distribution of all target-to-target group ratios is reported in Figure 7.2-6 (left). For the majority of transformations that produced MMPs active against multiple targets, the targets mostly belonged to different groups, with a median ratio of 1.5 targets per class. Only for a very few transformations accounting for the right tail of the distribution, a target group bias was apparent.

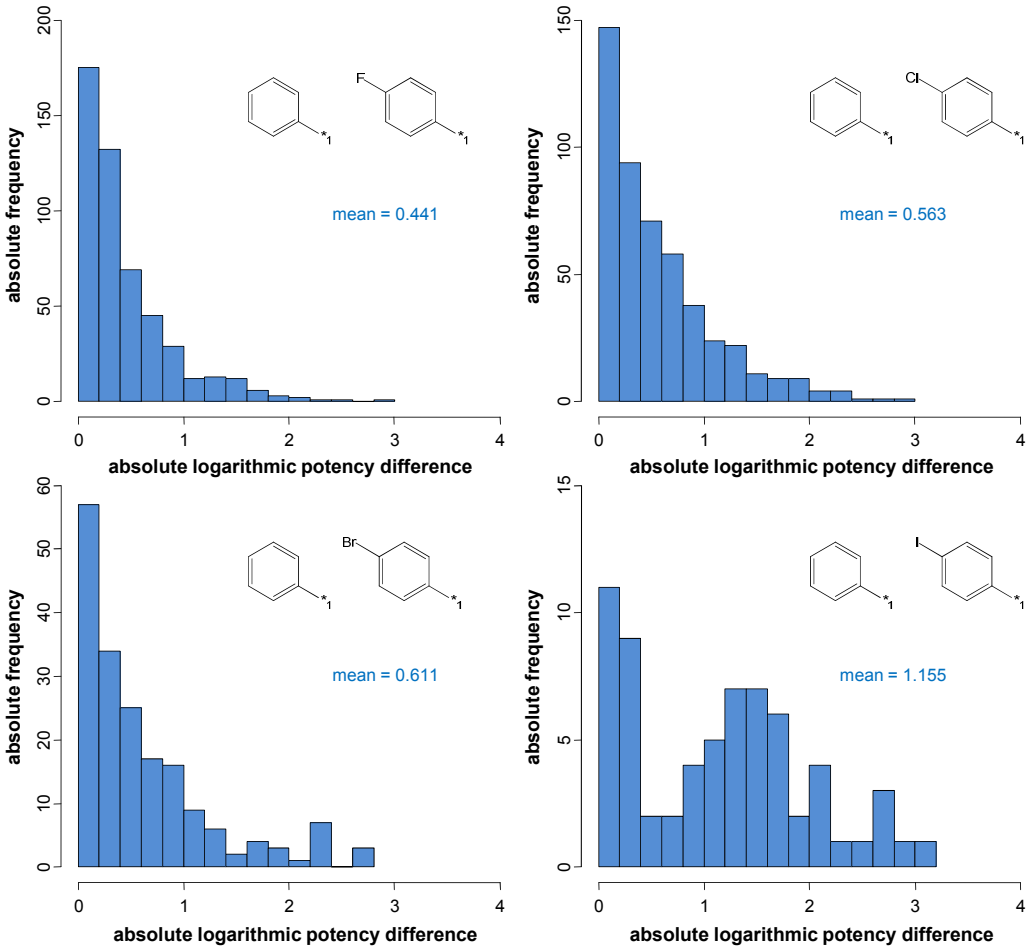


Figure 7.2-5: Potency distributions of selected transformations For four similar transformations, distributions of absolute logarithmic potency differences between compounds in corresponding MMPs are shown, and arithmetic means of the potency differences are reported. The relative frequency of activity cliff formation of these transformations is given in Figure 7.2-4.

To analyze the chemotype distribution, the ratio of the number of cliff-forming MMPs and the number of MMPs representing different scaffold pairs was determined for each frequent cliff former. As shown in Figure 7.2-6 (right), these results were usually close to one, thus indicating that most transformations introduced activity cliffs in variable structural environments. Figure 7.2-7 shows examples of cliff-forming MMPs that were defined by the same transformation and contained rather different scaffolds. In each of these compound pairs, a secondary and tertiary amine was exchanged. These MMPs formed activity cliffs against different targets including melanin-concentrating hormone receptor 1, carbonic anhydrase II, epidermal growth factor receptor, and thrombin. For thrombin, the replacement of the tertiary amine by the secondary amine increased potency by about two orders of magnitude for the compound pair on

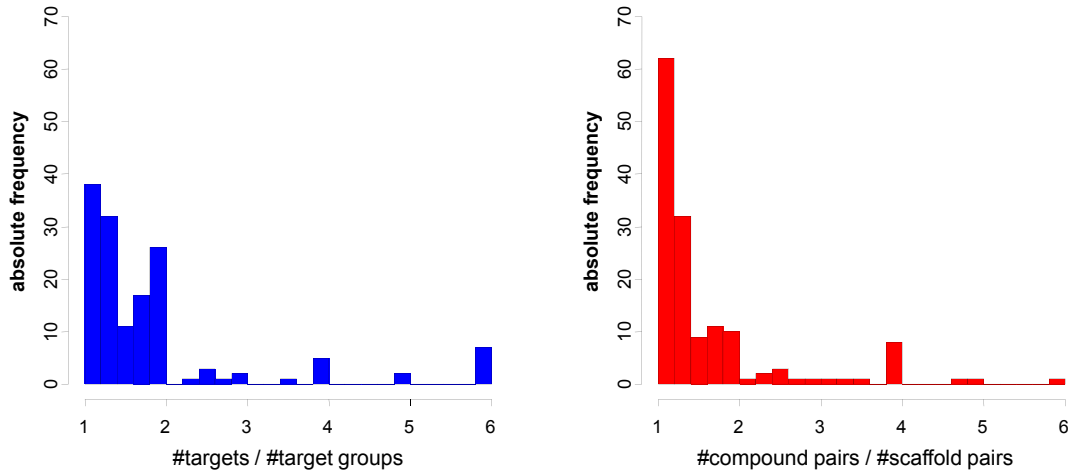


Figure 7.2-6: Target and chemotype distributions of frequent cliff formers For the set of 146 identified transformations, target-to-target group and MMP-to-scaffold pair ratios are reported.

the left in Figure 7.2-7 and decreased potency by more than three orders of magnitude for the compound pair on the right. In light of these findings, we further investigated the direction of potency changes for transformations inducing multiple activity cliffs for at least one target. We found that 70% of these transformations displayed only one potency direction for each target. However, as shown in Appendix Table E-2, there was a clear dependency on the number of different scaffolds in a set of compound pairs. A transformation that occurred in different chemotypes active against the same target had a considerably higher probability to cause large potency changes in different directions.

7.2.4 Summary

We have systematically analyzed chemical transformations in public domain compound data sets that defined matched molecular pairs and determined the potential of these substitutions to introduce activity cliffs. Approximately 150 non-redundant transformations were identified that displayed a strong tendency to introduce cliffs in different chemotypes and across different targets. This general tendency was not necessarily expected. However, clear trends emerged for specific chemical substitutions to globally introduce activity cliffs. By contrast, in other instances, structurally closely related substitutions displayed only little, if any, cliff potential. Hence, concerning the formation of activity cliffs, privileged substitutions exist.

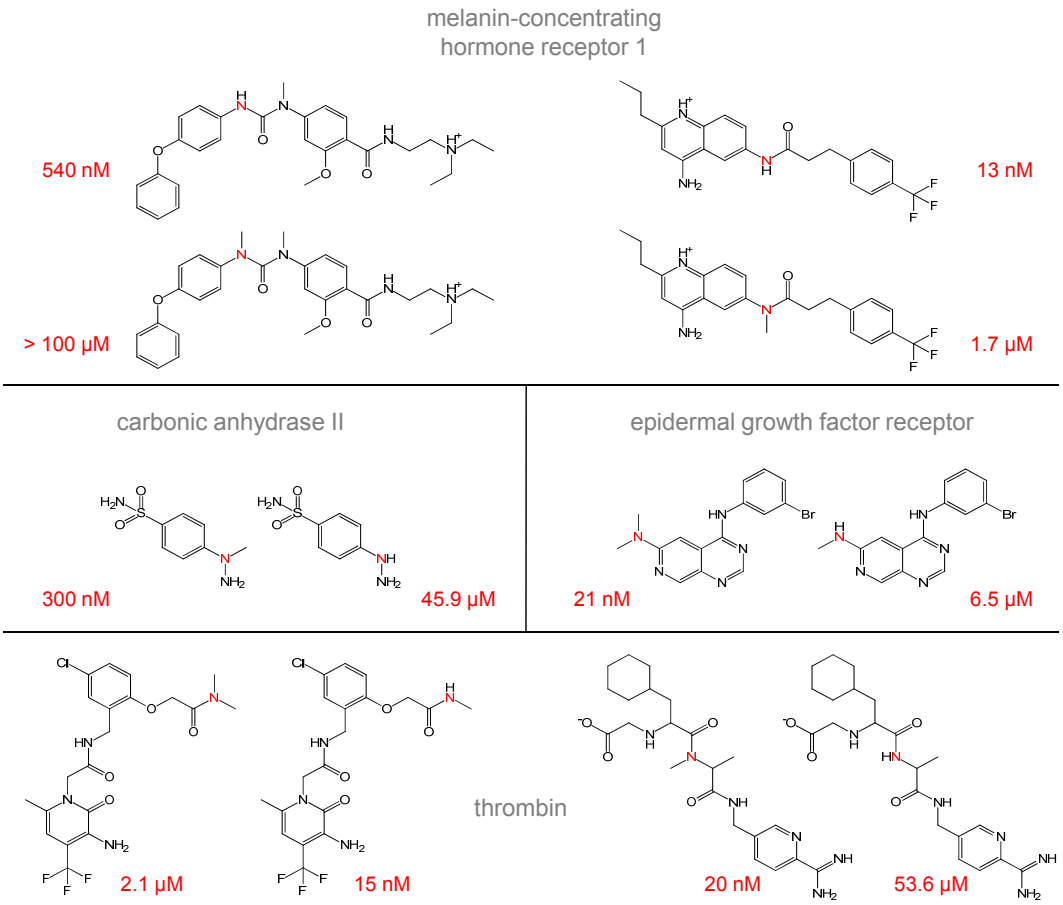


Figure 7.2-7: Structurally diverse MMPs Examples of MMPs are shown that are defined by the same transformation (i.e., the exchange of a secondary against a tertiary amine), contain different scaffolds, and form activity cliffs for different targets. All compounds are annotated with their calculated mean potency values. The figure is adapted from [108].

7.3 Bioisosteric Replacements

In addition to the analysis of identified frequent cliff-forming transformations, we have carried out a systematic search for bioisosteric replacements. In general, bioisosteres are defined as groups of atoms with similar structural and physicochemical properties whose replacement retains the biological activity of small molecules [116]. In medicinal chemistry, bioisosterism is an intensely investigated topic [117, 118] because bioisosteric replacements are used to alter molecular properties in a desired way, for example, by improving solubility and/or metabolic stability or by reducing toxic side effects while not abolishing the biological activity. In our investigation, we have applied stringent criteria for bioisosterism. We not only required that chemical replacements retained

similar biological activity but also comparable compound potency levels across different target families, thereby focusing on truly tolerated substitutions.

7.3.1 Compound Data Sets

Compound sets for human targets were assembled from the ChEMBL database following the criteria applied in the search for activity cliff-forming substitutions, complemented by two additional rules: First, measurements containing threshold values (i.e., reported as "$>$" or "$<$") were not considered. In the search for activity cliff formers, it was essential to include measurements for weakly active compounds (often carrying the modifier value "$>$") because many activity cliffs would have been missed otherwise. However, for the identification of bioisosteric replacements yielding compounds with similar potency levels, potency differences between pairs of compounds must be clearly resolved, making measurements with threshold values unsuitable for this analysis. Second, if multiple measurements were available for the same compound-target pair and individual potency values differed by more than one order of magnitude, all measurements were disregarded. The selection procedure led to the extraction of 30 368 different compounds organized into 472 target-specific sets with at least five ligands each. All targets used in our study were grouped into target families using the sequence-based family annotation of the UniProt database and the classification hierarchy of the ChEMBL database.

7.3.2 Transformation Selection Criteria

MMPs were calculated for all ligand sets using our modified version of the Hussain and Rea algorithm and transformations were associated with potency changes as described in paragraph 7.2.2. To identify transformations that are bioisosteric replacements, we applied the following filter criteria:

1) At least 30 potency records are associated with the transformation.

2) MMPs defined by the transformation are found for targets of at least two different families.

3) If the target family for which most potency records are available is removed from the analysis, at least 15 measurements remain.

4) Not more than 1/15 of all potency records are larger than one, i.e., maximally 6.67% of all potency changes induced by the transformation are larger than one order of magnitude.

5) In the case that more than 50% of all potency records are observed for one target family, the fraction of potency records greater than one magnitude observed for the other target families only is not larger than 2/15.

With these filters, we aimed at identifying transformations that were frequently observed for different target families but rarely introduced large potency changes. Filter 1 was applied to base the analysis on a meaningful statistical ground, whereas filters 1, 3, and 5 prevented that the analysis was biased toward a single target family for which the transformation was frequently observed, as had been the case for a very few identified frequent cliff formers. Filter 4 is the actual "bioisostere filter" in that it identifies those transformations that consistently generate compound pairs with very similar potency levels. In analogy to the rules for structural context dependence and MMP subset relationships applied to cliff-forming transformations, we formulated the following two rules to avoid redundancies:

i) Rule 1: if the MMP sets described by two or more transformations are identical, only the largest transformation is retained as it provides most information about the structural exchange. This rule was applied prior to our selection criteria and was used to reduce the number of transformations subjected to filters 1-5.

ii) Rule 2: all those transformations whose MMP sets are subsets of MMP sets of other bioisosteric replacements are excluded. This rule was applied after filters 1-5 to further reduce the resulting set of bioisosteric replacements.

7.3.3 Results

On the basis of our search strategy, transformations were regarded as potential bioisosteric replacements if they were consistently represented by multiple MMPs with moderate potency differences for targets of more than one family and, in addition, if the corresponding potency records were not significantly biased by a single target family. Potency variations as a consequence of transformations were largely (but not exclusively) limited to an order of magnitude. This range is consistent with the basic idea of bioisosterism because a replacement resulting in a 100- or 1 000-fold reduction in compound potency would hardly be regarded as bioisosteric. On the other hand, a ten-fold increase in potency as a consequence of a replacement would certainly be considered a favorable bioisosteric effect. Furthermore, it should be noted that fragments replaced by transformations meeting our requirements may not always be true bioisosteres because they constitute an unimportant part of the molecule, i.e., one cannot be sure that they are involved in critical interactions with the target.

7.3.3.1 Transformation and Potency Difference Statistics

For 460 of our 472 ligand sets, at least one MMP was identified. The corresponding 460 targets belonged to 141 different target families, with the number of targets per family ranging from 1 to 35. Redundancy rule 1 reduced the

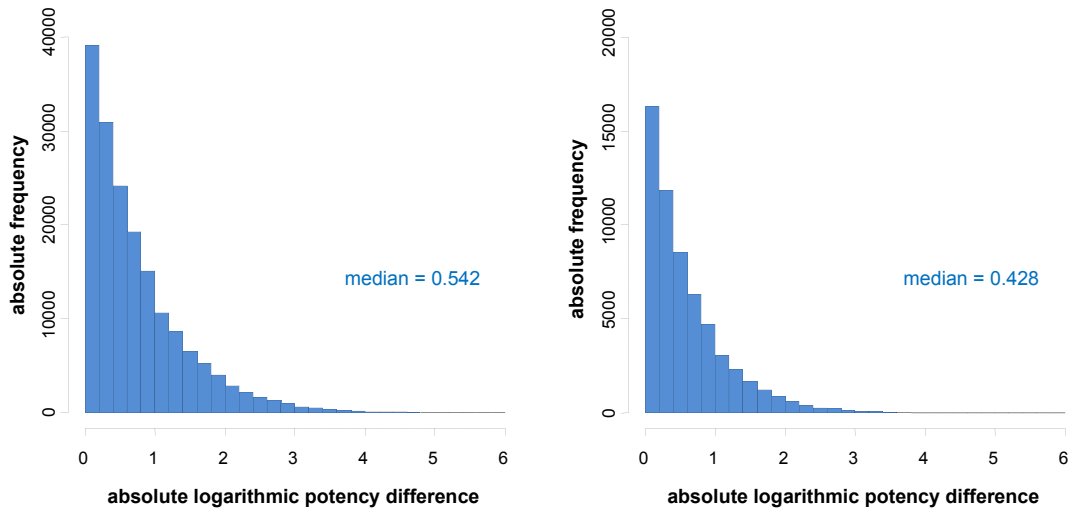


Figure 7.3-1: Potency difference distributions Distributions of absolute logarithmic potency differences are reported as histograms for all matched molecular pairs (left) and matched molecular pairs that were defined by a transformation with at least 30 potency records (right). The figure is adapted from [109].

original set of 276 920 identified different transformations to 55 399 candidate transformations, which provided the pool for the application of our bioisostere selection criteria. For this set of 55 399 transformations, the number of potency records for a single transformation ranged from 2 to 2 360. For compound pairs defined by these transformations, absolute logarithmic potency differences were calculated for all targets and the distribution of these potency records is shown in Figure 7.3-1 (left). The distribution roughly follows an exponential distribution. Thus, compounds of the majority of MMPs showed only small differences in potency and only few MMPs showed large potency differences. The median absolute logarithmic potency difference was 0.542 for compounds forming an MMP.

7.3.3.2 Identified Replacements

In order to identify bioisosteric replacements that were observed for members of multiple target families, filters 1-5 were applied. Of the 55 399 considered transformations, only 1 721 transformations had at least 30 potency records associated with them. Hence, filter 1 already reduced the number of transformations to approximately 3% of the original volume. However, we considered a minimum of 30 potency records required for a statistically meaningful analysis. A total of 1 585 transformations were observed in the ligand sets of different protein families and passed filter 2. However, for 153 transformations, potency records were predominantly reported for one target family with less than 15 records remaining for other families, such that these transformations failed filter

3. For the MMPs described by the remaining 1 432 transformations, a distribution of absolute logarithmic potency differences was calculated and is displayed in Figure 7.3-1 (right). In comparison to the distribution for the complete set of transformations shown on the left in Figure 7.3-1, the overall shape of the distribution is similar, however, the median potency difference is shifted to the left (0.428) indicating a slight tendency to cover smaller potency differences for frequently found transformations. We determined that 19.6% of all calculated potency differences were larger than one order of magnitude. This frequency could be used to estimate the amount of "false positive" bioisosteres identified by our analysis. For a transformation that introduces potency changes of more than one order of magnitude with this mean frequency and for which only 30 measurements are available, the probability that it is identified as a bioisostere given our criteria was smaller than 5%. This demonstrates that a cutoff of 30 measurements enabled a statistically meaningful analysis.

Our bioisostere filter reduced the set of 1 432 transformations to 132 replacements. Hence, in many instances, too many potency records available for a transformation violated the potency criterion. Only 1 of these 132 transformations did not meet the final selection criterion, but the second redundancy rule for subset relationships further reduced the number to 96 transformations that described bioisosteric replacements accepted in the context of our analysis. SMIRKS representations for these bioisosteric replacements are provided in Appendix Table E-3 and are annotated with their number of corresponding potency records (ranging from 30 to a maximum of 439 records). Individual bioisosteric replacements occurred in up to 282 different MMPs and were found in 132 ligand sets belonging to 48 target families. The calculation of corresponding medians yielded 27 MMPs, 25.5 targets, and 13.5 target families.

We have also determined whether the identified bioisosteric replacements occurred in different chemotypes by calculating the ratio of the number of MMPs and the number of MMPs representing different scaffold pairs for each transformation. In most cases, the MMP-to-scaffold pair ratios were close to one, hence indicating that most bioisosteric replacements indeed occurred in different chemotypes.

7.3.3.2.1 Comparison to Known Bioisosteric Replacements Among our set of 96 non-redundant bioisosteric replacements, we found a number of previously described bioisosteres [117, 118], as illustrated in Figure 7.3-2. The halogen substitutions bromine versus chlorine or iodine and the exchange of piperazine and piperidine are examples of "classical" isosteres, as defined by Erlenmeyer [119], i.e., atoms or molecules in which the outer electron shells are considered identical. In addition, previously reported non-classical bioisosteres were also identified, including, for example, the substitution of iodine for a nitro

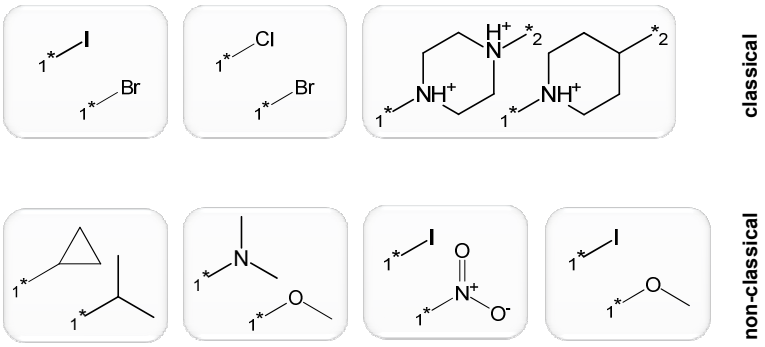


Figure 7.3-2: Traditionally known bioisosteres Classical and non-classical bioisosteres identified in our analysis are shown. The figure is adapted from [109].

or methoxy group, the exchange of a cyclopropyl and isopropyl group, and the substitution of a dimethylamine for a methoxy group.

7.3.3.2.2 Preferred Structural Environments Several other commonly accepted bioisosteric replacements were only found in specific structural environments, as illustrated in Figure 7.3-3. For example, the exchange of a fluorine atom and hydroxyl group, which are for long known as isosteres according to Grimm's Hydride Displacement Law [120], was only accepted as a bioisostere when these substituents were attached to an acyclic carbon atom. Similar structural "restrictions" were also observed for replacements of chlorine by fluorine, and methyl, trifluoromethyl, and hydroxyl groups, the replacement of a nitro or trifluoromethyl group by bromine, and the exchange of a methoxy group and fluorine. Furthermore, the exchange of an –O– and an –S– linker only qualified as a bioisosteric replacement when one of the attached moieties was a 4-substituted phenyl ring.

Because only single bonds between non-hydrogen atoms were deleted to fragment bioactive compounds for MMPA, it was not possible to identify a transformation that replaced a single hydrogen atom. Instead, the substitution of hydrogen atoms by other functional groups was detected as part of transformations involving multiple heavy atoms that provided additional information about the structural context of a substitution. Figure 7.3-3 shows several such transformations where the actual exchange is the replacement of a hydrogen atom by a fluorine or chlorine atom or a methyl or methoxy group.

Furthermore, a subset of our 96 bioisosteric replacements consisted of structural isomers, as shown in Figure 7.3-4. Most of these replacements defined ortho/meta or meta/para regioisomers. By contrast, bioisosteric ortho/para regioisomers were not found.

7.3.3.2.3 Excluded Bioisosteric Replacements A number of commonly accepted bioisosteric replacements were not identified. For example, the substructure pairs shown in Figure 7.3-5 are generally regarded as bioisosteres and

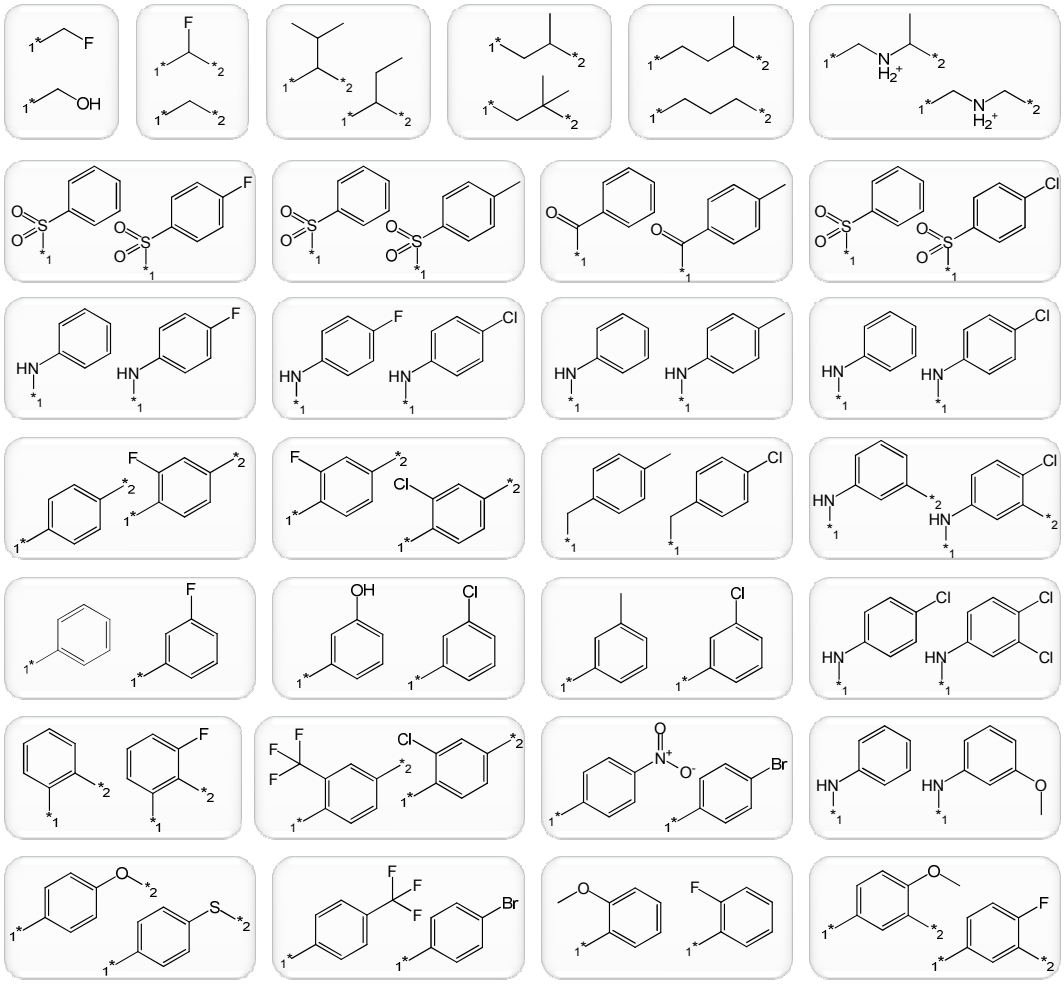


Figure 7.3-3: Structural context dependence Commonly accepted bioisosteric replacements that only met our transformation selection criteria when occuring in specific structural environments are reported. The figure is adapted from [109].

we investigated the reasons for their exclusion on the basis of our selection criteria. The phenyl versus 2-thienyl, 2-furanyl, and 2-pyridyl exchanges frequently occurred across different target families but failed to pass our bioisostere filter because the fraction of potency records within one order of magnitude ranged from 75.7% to 86.3%. Furthermore, we analyzed five known bioisosteric replacements of carboxylate whose substitution with other groups had been found to frequently introduce activity cliffs, as reported in the previous section. For the substitutions of phosphonate and sulfonate for carboxylate, too few potency records were available to meet our selection criteria. However, the limited number of available potency records revealed a substantial difference between these two replacements. For the carboxylate versus phosphonate replacement, only 29.2% of the potency records were within one order of magnitude, whereas

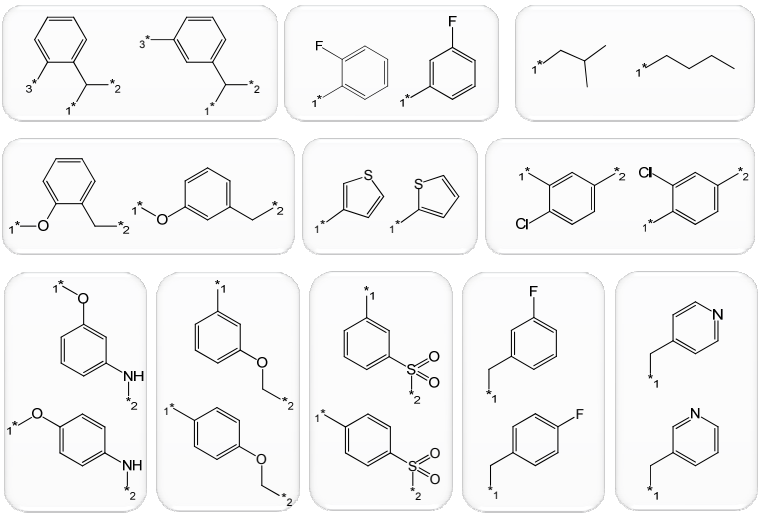


Figure 7.3-4: Structural isomers Molecular transformations that describe constitutional isomers and qualified as bioisosteric replacements in our analysis are shown. The figure is adapted from [109].

93.3% of the potency records for the carboxylate versus sulfonate replacement were within one order of magnitude. The replacements of a primary amide, hydroxamate, and tetrazolate by carboxylate occurred more frequently, but were often accompanied by potency changes larger than one order of magnitude, for example, in 68% of all cases for hydroxamate. In addition, the hydroxamate versus carboxylate exchange was essentially limited to matrix metalloproteases. Overall, exchanges of the divalent linkers –S–, –O–, –NH–, and –C–, which were also frequently observed, retained similar potency levels in about 75% to 85% of all cases. A similar percentage was observed for the replacement of carbonyl by sulfinyl and sulfonyl groups. However, the sulfinyl to carbonyl transformation was only annotated with ten potency records and much less frequent than the sulfonyl versus carbonyl exchange. Furthermore, the three replacements of the amide group shown in Figure 7.3-5 clearly failed to meet our bioisostere criterion, with approximately every fourth potency record being larger than one order of magnitude. Thus, in these cases, previously reported bioisosteres were not selected by our analysis because too few potency measurements were available or the replacements were associated with rather large changes in potency. Of course, for compound optimization efforts, such replacements might well be attractive, but their bioisosteric nature should be considered with some caution.

7.3.3.2.4 Extending the Current Spectrum of Bioisosteres Half of our bioisosteric replacements were previously unobserved. A subset of these replacements is shown in Figure 7.3-6. For example, rather surprisingly, an

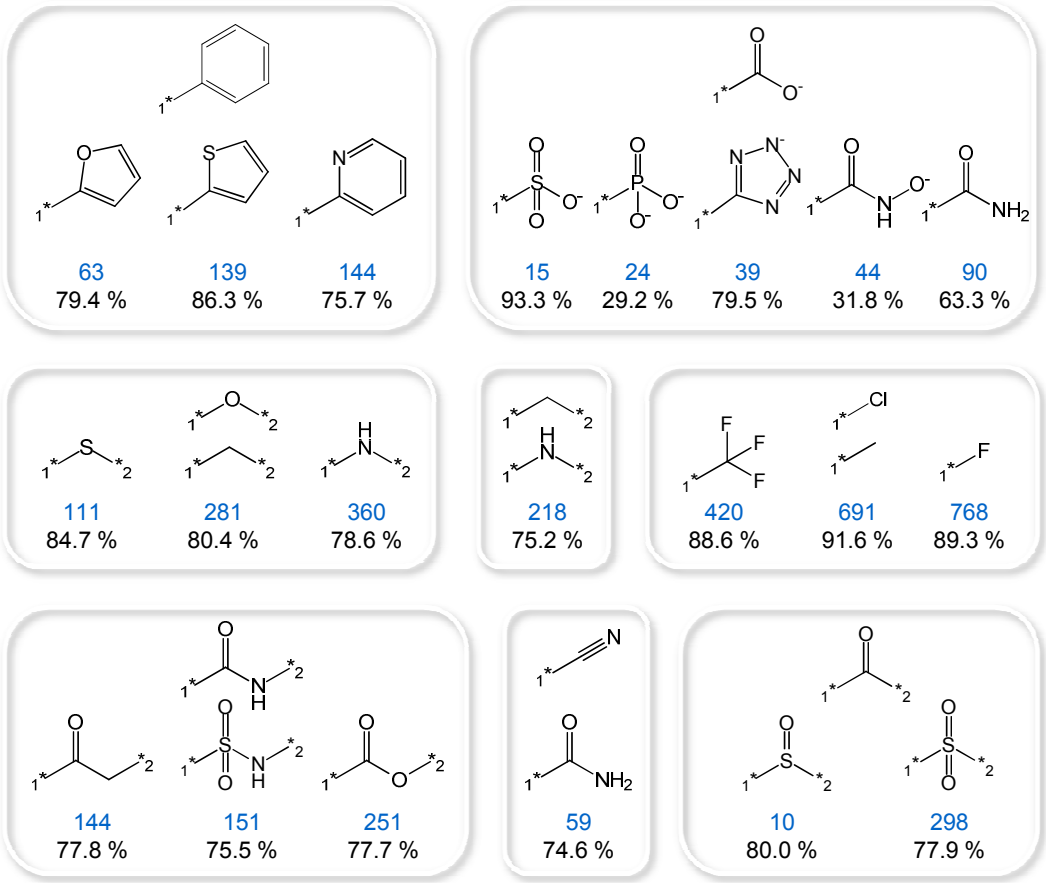


Figure 7.3-5: Non-accepted substitutions Replacements that were excluded on the basis of our selection criteria are shown and annotated with their number of potency records and the percentage of records falling within one order of magnitude. If more than two fragments are shown in a box, the fragment at the top is exchanged with each of the fragments at the bottom. The figure is adapted from [109].

ortho-fluoro-substituted and various meta-substituted phenyl rings, but not the unsubstituted phenyl moiety, were found to be bioisosteres of 2- or 3-thienyl groups. In addition, pyrrolidine, but not the probably more intuitive piperidine, was identified as a bioisostere of morpholine. Furthermore, the 3-methoxyphenyl and benzodioxol rings were also found to be bioisosteric. Moreover, the cyclopentyl ring was not only identified as a bioisostere of the cyclobutyl group but also of isobutyl and propyl groups. Several bioisosteric replacements in Figure 7.3-6 also reveal that the introduction of an ether group was permitted in different structural environments. Thus, about half of the bioisosteres identified in our analysis further extend the spectrum of generally accepted bioisosteric replacements.

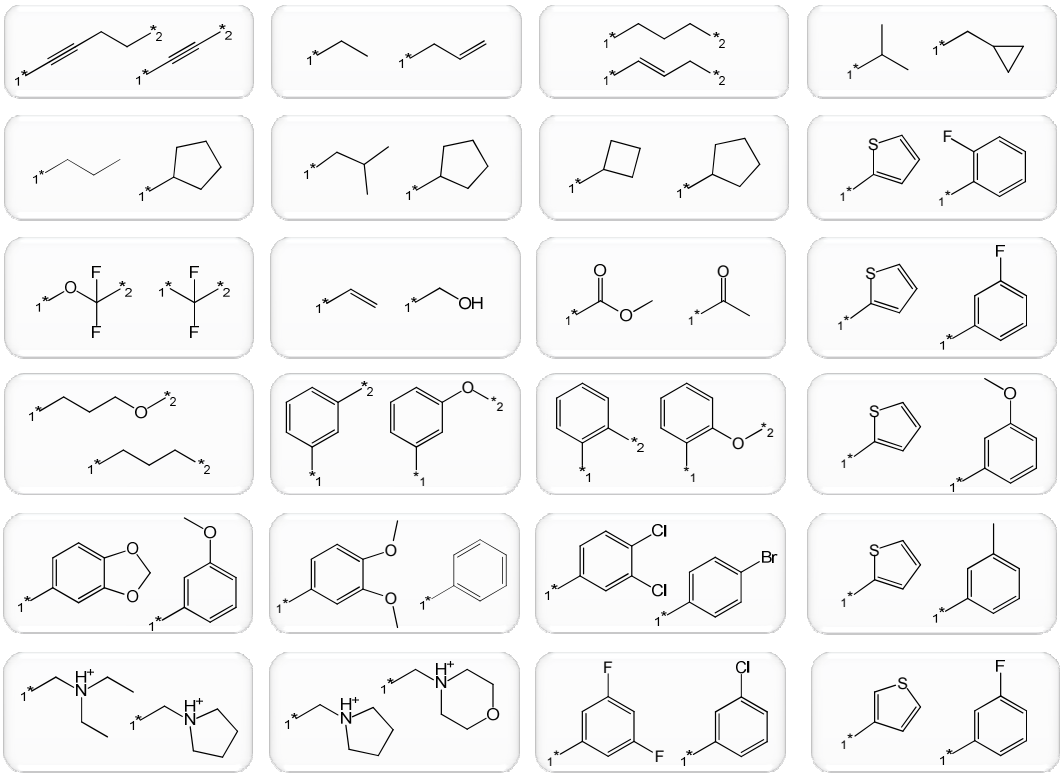


Figure 7.3-6: Extended spectrum of bioisosteric replacements Previously only little-considered or unobserved bioisosteric replacements are shown. The figure is adapted from [109].

7.3.4 Summary

On the basis of systematically generated MMPs, we have identified a set of non-redundant transformations that defined potential bioisosteric replacements, i.e., conservative substitutions of chemical groups that are tolerated by biological targets but have the potential to modulate other compound properties in a desired way. Transformation selection criteria were applied that captured what we considered to be key aspects of bioisosterism. These criteria emphasized the availability of many potency records associated with defined transformations, limited potency differences in MMPs representing a given replacement, and observed activity of candidate replacements across different target families. On the basis of these criteria, a number of previously reported replacements were not considered to be bioisosteric because currently available compound data were too sparse to draw statistically sound conclusions or the replacements were accompanied by large changes in compound potency. Another key observation has been that many commonly accepted bioisosteric replacements depended on a specific structural environment. We identified a set of 96 non-redundant bioisosteric replacements that consistently met our selection criteria

and were well-supported by currently available data. Approximately half of these bioisosteres were thus far only little-considered or unobserved. This revised and further extended set of bioisosteric replacements should be useful for many medicinal chemistry applications.

7.4 Target-Family Directed Bioisosteric Replacements

Bioisosteric replacements discussed thus far have been general, i.e., they have been considered to be bioisosteric across different targets. In compound optimization, one typically considers bioisosteres on the basis of this premise. A question that has not yet been investigated, but that is also of high relevance for medicinal chemistry applications, is whether or not bioisosteric replacements can be found that preferentially act against a given target family. Considering the individual structural constraints that must be met to yield specific target-ligand interactions, it would perhaps not be unlikely that such replacements exist. In order to address this question we have adapted the previously described transformation selection criteria to the search for bioisosteres in individual target families, using the same compound data sets as in the search for general bioisosteric replacements.

7.4.1 Compound Data Sets

We used the sets of target-specific MMPs identified in our search for general bioisosteric replacements. As only target families with MMP sets for at least three targets were considered in our analysis, only 432 of the original 460 MMP sets were retained. The corresponding 432 targets belonged to 40 different target families, with the number of targets per family ranging from 3 to 35. The 432 sets accounted for 22 631 different compounds forming 107 669 MMPs defined by 255 845 (in part redundant) transformations.

7.4.2 Transformation Selection Criteria

Molecular transformations were separately analyzed for each target family. For each transformation found in compound sets of a target family, all corresponding MMPs were assembled and the following search protocol for bioisosteric replacements in target families was applied:

1) At least 20 potency records are associated with the transformation.

2) MMPs defined by the transformation are found for at least two targets and contain at least five different scaffold pairs.

3) If the target or the scaffold pair for which the largest number of potency records are available is removed from the analysis, at least ten potency records have to remain.

4) Not more than 1/15 of all potency records are larger than one, i.e., maximally 6.67% of all potency changes induced by the transformation are larger than one order of magnitude.

5) In the case that more than 50% of all potency records are observed for one target or scaffold pair, the fraction of potency records greater than one magnitude observed for the other targets or scaffold pairs, respectively, is not larger than 2/15.

Filter 2 focuses the search on transformations that occur in different structural contexts. In the search for frequent cliff formers and general bioisosteric replacements, chemotype diversity was retrospectively assessed for the final set of selected transformations but not included in the search criteria. However, on the basis of individual target families, biases of potency record distributions towards single scaffolds are much more likely to occur and were therefore directly prevented by our modified search criteria. Rules for avoiding redundancies in transformations were adopted from paragraph 7.3.2.

7.4.3 Results

We aimed at the identification of bioisosteric replacements at the level of a target family. Transformations met the search requirements if they were consistently represented by multiple MMPs with moderate potency differences for more than one target, if they occurred in different structural environments, and if the corresponding potency records were not significantly biased by a single target or scaffold pair. Importantly, due to the general sparseness of currently available activity annotations, one would not be able to conclude with certainty that replacements meeting these selection criteria would be true bioisosteres for all targets belonging to a family or that they could not act on a target belonging to another family. It is important to note that one can only extract information that currently available compound data provide. Therefore, replacements that ultimately met our criteria for only one target family were classified as target family-directed (rather than family-specific) bioisosteres.

7.4.3.1 Preselected Transformations

A pool of 255 845 transformations provided the starting point for our analysis. After the application of our selection criteria, only 79 non-redundant transformations remained. The selection procedure traced transformations back to 16 target families. However, for a subset of transformations, multiple family

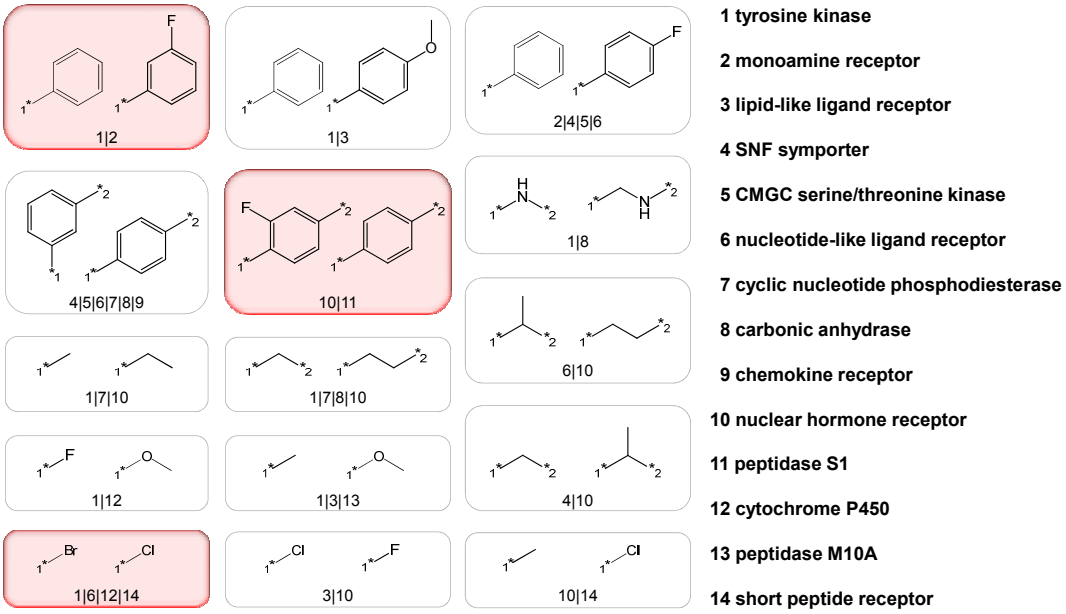


Figure 7.4-1: Bioisosteres directed at multiple target families The 16 bioisosteric replacements that were identified for multiple target families are shown and annotated with their family assignments. Three replacements that were also identified as general bioisosteres across multiple target families are highlighted in red. The figure is adapted from [110].

assignments were obtained. We found more than one family assignment for a total of 15 transformations that are shown in Figure 7.4-1. These transformations involved replacements of small functional groups (e.g., halogen atoms and methyl and methoxy groups), aliphatic linkers of different length, and phenyl rings with different substituents. Most of these transformations were assigned to two target families and, in a few instances, also to three or four. The extreme case has been a rather generic replacement of a meta- versus para-substituted phenyl ring, which was bioisosteric for ligands of six target families. Three of the 15 transformations had also passed the filter criteria in our previous search for transformations acting as bioisosteres across different target families and are highlighted in red boxes in Figure 7.4-1.

7.4.3.2 Identified Replacements

After removal of these 15 transformations, a total of 64 transformations remained that met our single target family constraint and hence qualified as target family-directed bioisosteres. These 64 bioisosteric replacements were directed against 11 target families. Table 7.4-1 reports the distribution of bioisosteric replacements over these families. A total of 22 replacements were found in ligands active against the nucleotide-like ligand receptor family, which represented the largest number, followed by the short peptide receptor, tyrosine

Table 7.4-1: Distribution of target family-directed bioisosteres

target family	abbr.	#bioisosteres	#targets
nucleotide-like ligand receptor	NLR	22	5
short peptide receptor	SPR	17	35
tyrosine kinase	TK	7	30
peptidase M10A	M10A	4	9
peptidase S1	S1	3	17
monoamine receptor	MAR	3	34
carbonic anhydrase	CA	3	9
nuclear hormone receptor	NHR	2	18
AGC serine/threonine kinase	AGC	1	14
peptidase C1	C1	1	6
lipid-like ligand receptor	LLR	1	20

Target families for which family-directed bioisosteres were identified are listed in the column "target family" and are abbreviated ("abbr."). For each family, the number of directed bioisosteres ("#bioisosteres") and the number of targets in the family ("#targets") are reported.

kinase, and peptidase M10A families with 17, 7, and 4 bioisosteric replacements, respectively. Interestingly, while seven bioisosteric replacements were found for inhibitors of tyrosine kinases, only a single bioisostere was detected for inhibitors of AGC serine/threonine kinases. In two more cases including the lipid-like ligand receptor and the peptidase C1 families, only a single qualifying replacement was identified. In Figure 7.4-2, all 64 target family-directed bioisosteric replacements are shown. Four transformations identified for a single target family in this study were also included in the set of 96 replacements acting as bioisosteres across multiple target families and are highlighted in red boxes. This finding underlines the known sparseness of available compound activity data and implies that these transformations would likely have been identified as bioisosteres for additional target families if enough potency records had been available for their ligands. However, as further discussed below, replacements that might, at first glance, look rather generic are indeed target family-directed because of significant potency differences associated with them in different families.

As shown in Figure 7.4-2, chemically different and differently sized replacements were observed for individual protein families. Qualifying replacements involved not only functional groups, but exchanges of linker fragments or substituted ring systems were also frequently found. Thus, for practical compound optimization, some target family-directed bioisosteres involving well-defined functional groups might be of higher interest than others, such as linker fragments, although these fragments also met all formal requirements for bioisosterism. Figure 7.4-3 shows a subset of these preferred bioisosteres that would have also qualified for one to five other target families if potency changes of more than

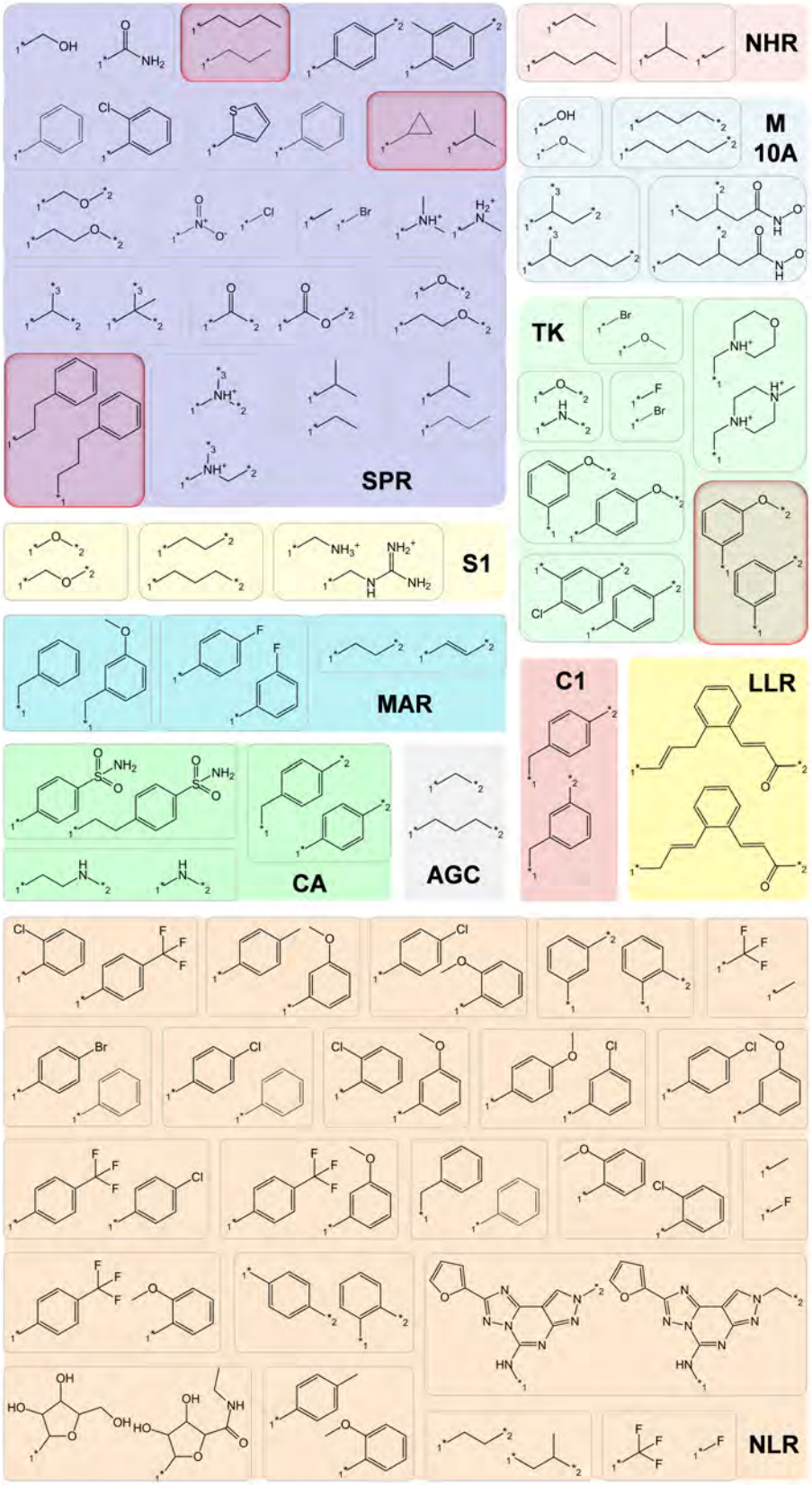


Figure 7.4-2: Target family-directed bioisosteric replacements All 64 single target family-directed bioisosteric replacements are shown and annotated with their family assignment. Abbreviations are used according to Table 7.4-1. Four replacements that were also identified as general bioisosteres across multiple target families are highlighted in red boxes. The figure is adapted from [110].

one order of magnitude would have been permitted in our analysis. As an example, we consider the methyl to trifluromethyl replacement that was found to be directed against the nucleotide-like ligand receptor family. This substitution was also frequently observed for monoamine receptors, tyrosine kinases, short peptide receptors, lipid-like ligand receptors, and sodium neurotransmitter symporters. However, for these families, the replacement often induced large potency differences. For these families, potency changes of more than one order of magnitude were observed with frequencies of 21.1, 18.5, 15.6, 13.3, and 9.7%, respectively. Hence, in these cases, the methyl to trifluromethyl substitution did not qualify as a bioisostere. Nevertheless, although their bioisosterism with respect to the secondary target families might in part be questionable due to large potency alterations, the replacements shown in Figure 7.4-3 can also represent attractive candidates for the optimization of compounds active against these families if potency changes are tolerated or even explicitly desired in the current stage of the optimization process.

7.5 Conclusions

In this chapter, we have adapted the matched molecular pair formalism to comprehensively retrieve molecular transformations from publicly available bioactive compound sets. A strength of the MMP-based approach is that it requires no preconceived chemical notion of groups that might be of interest, but rather systematically detects all substructure exchanges that are defined by single transformations. In addition, structural contexts of actual substructure modifications are taken into consideration. In our analyses, chemical replacements were systematically related to resulting potency changes. Sections 7.2 and 7.3 reported the identification of substructure exchanges with a general propensity to introduce activity cliffs or produce compounds with similar potency levels. Care was taken not to bias the analyses towards individual chemotypes or target families. From vast available chemical transformation space, including both R-group and core substructure changes, compendiums of activity cliff-forming and bioisosteric replacements were assembled that showed clearly distinguishing characteristics. For example, frequent cliff formers often described structural exchanges that led to compounds with different charges or notable differences in size. By contrast, the bioisosteric replacements that we identified mainly consisted of substructures of similar size and charge, in accordance with the general definition that bioisosteres should have similar structural and physicochemical properties. A notable amount of our identified bioisosteric replacements led to pairs of compounds with different numbers of H-bond acceptors, but changes in the number of H-bond donors were only rarely observed. Our set of frequent cliff formers, which is the first and only collection of this kind reported so far, and our revised and further extended set of bioisosteric replacements should be use-

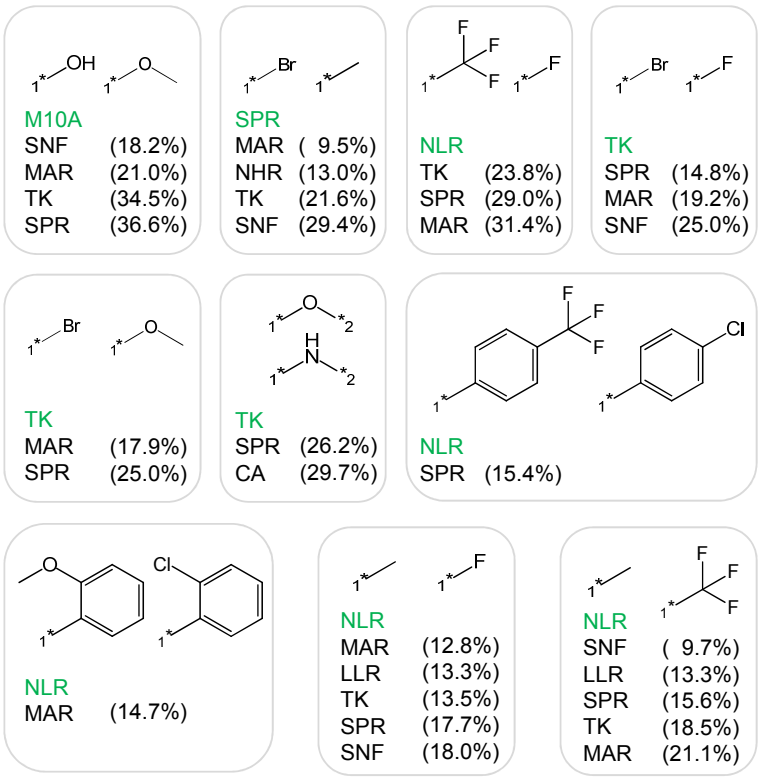


Figure 7.4-3: Comparisons across different target families Target family directed replacements are shown that would have qualified as bioisosteres for more than one target family if potency changes larger than one order of magnitude had been accepted. The qualifying target family is abbreviated in green and other families that did not meet the potency criterion are shown and annotated with the relative frequency with which the transformation introduced potency changes larger than one order of magnitude. Abbreviations are used according to Table 7.4-1. SNF stands for the sodium neurotransmitter symporter family. The figure is adapted from [110].

ful for many medicinal chemistry applications and have important implications for screening library design.

Furthermore, the results presented in section 7.4 suggest that the currently available repertoire of bioisosteric replacements is further extendable through a systematic analysis of bioisosteres at the level of target classes and that a subset of bioisosteres is indeed preferentially directed against individual target families. A differentiation of transformations on the basis of their observed activity profiles is anticipated to be helpful in the establishment of a classification scheme for bioisosteric replacements and of great aid in compound optimization efforts.

In general, systematic compound data analysis approaches, as reported herein, are expected to become increasingly relevant for the analysis of structure-activity relationships in the future as available compound collections further grow in size.

Source Information

Sections of the text in this chapter have been taken from [108–110].

Chapter 8

Structure-Activity Relationship Patterns in Series of Analogs

In addition to the approaches to study structure-activity relationships presented so far, computational visualization methods have been introduced that help to extract SAR information from compound data sets [31]. Different methods have been developed to analyze large and diverse compound data sets including HTS data [15, 121]. The extraction of SAR information from large compound sets represents one of two major tasks in SAR analysis. The other task is compound optimization during later stages of medicinal chemistry efforts. In this case, the focus shifts from larger data sets to individual compound series where SAR exploration primarily aims at the analysis and design of analogs of active compounds with further improved properties. Therefore, computational methods are required that are very sensitive to small structural modifications and analyze SARs with a high resolution at the level of individual substitution sites.

The conventional and still most widely used data structure for the analysis of analog series are R-group tables that contain the core structure common to a series of analogs and rows displaying the substituents of individual compounds and the associated potency measurements. User-friendly extensions of R-group tables have been introduced, such as SAR maps [25] that arrange analogs in rectangular matrices of cells where each cell represents a unique combination of R-groups at two substitution sites. Cells are then color-coded according to a specific molecular property, usually compound potency against a given target. Only a subset of a series is displayed if analogs display variations at more than two substitution sites. Heat maps were also used to display mean potency changes resulting from the exchange of a pair of substituents at a given site [122]. Similarly to SAR maps, multiple views of the same series of analogs are required to display SAR information for more than one substitution site. Another recently introduced data structure of graphical analog analysis is the combinatorial analogue graph (CAG) [42] that systematically organizes substi-

tution sites and their combinations in a tree-like structure and identifies their contributions to SAR discontinuity. Hence, CAGs view analog series differently from R-group tables because they pinpoint substitution sites in the common core structure where R-groups introduce significant changes in potency. However, CAGs do not provide an immediate access to functional groups at these positions (discussed in more detail in Chapter 9).

Other than R-group tables, their extensions, and CAGs, there are currently no graphical SAR analysis methods for analogs available. In particular, SARs between R-group combinations at different sites are difficult to monitor. Therefore, we have developed a graphical data structure that goes beyond the capacities of previously published visualization methods by explicitly using R-group combinations and their (subset) relationships as an organizing principle [123]. The design of this graph structure that emphasizes relationships between different R-group combinations and is termed *directed R-group combination* (DRC) graph (DRCG) is introduced in this chapter. As detailed in section 8.1, R-group combinations are systematically extracted from a given analog series, associated with potency information of all analogs containing a specific combination, and organized according to consistently numbered substitution sites. Interpretable, information-rich SAR patterns emerging from this data structure are described in section 8.2 and exemplary applications to four different analog series are reported in section 8.3. The chapter ends with a summary of key aspects in section 8.4.

8.1 Methodology

The newly designed graph structure represents entire series of analogs in a consistent manner, regardless of their size and complexity of substitution patterns. The approach is specifically tailored towards a systematic exploration and intuitive interpretation of SAR features involving different R-groups and their combinations. Analogs and their potency information are systematically organized on the basis of R-group combinations that are present in a series.

8.1.1 R-group Deconvolution and Signature Formation

Initially, the maximum common subgraph (MCS) of compounds comprising an analog series is determined and all non-hydrogen atoms of the MCS are labeled with numeric identifiers, as illustrated in Figure 8.1-1 (top) for a model series of five analogs. The MCS is then used as the invariant molecular core structure for R-group deconvolution of all analogs. For this purpose, the MCS is mapped onto each analog and the numeric identifiers are transferred to matching atoms. Variable R-groups are identified and unambiguously assigned to corresponding substitution sites for all analogs by extracting groups that are not part of the

alignment and marking them with the numeric identifier(s) of the matching atom(s) to which they are attached. The list of all identified R-groups is used as the signature of the molecule. If multiple mappings of the MCS onto a compound are possible because it contains symmetry elements around rotatable bonds, all mappings and resulting signatures are determined, as illustrated for molecules **C**, **D**, and **E** in Figure 8.1-1 (middle). In this example, the three molecules contain a chlorine atom at the ortho position of the phenyl ring that can be assigned to substitution sites 2 or 4.

In the next step, R-group combinations that are shared by multiple compounds are systematically detected by extracting signature subsets from all analogs. Hence, if an analog contains R-groups at n substitution sites, all possible signature subsets with R-groups for $n-1$ to 1 substitution sites are generated, as illustrated for molecule **E** in Figure 8.1-1 (middle). The original signature and all signature subsets are then added as separate keys to an index table and assigned to the source compound (Figure 8.1-1, bottom). Hence, all analogs belonging to a particular key share the R-group combination defined by the key. If alternative mappings of the MCS onto given analogs of a series are possible, corresponding keys in the index might describe the same R-group pattern for an identical set of compounds with alternatively numbered substitution sites. These keys are identified and combined into a single entry (Figure 8.1-1, bottom).

The R-group decomposition, signature (subset) formation, and index structure generation routines were implemented in Java using the OpenEye chemistry toolkit.

8.1.2 DRCG Design and Visualization

In order to capture subset relationships between keys in the index table (i.e., sets of R-groups at specific substitution sites), a directed acyclic graph is generated, as illustrated in Figure 8.1-2.

8.1.2.1 Graph Structure

Keys correspond to nodes in the DRCG. Each node is associated with the set of molecules that contain the specified R-group combination (and are thus linked to the same key in the index table). Nodes are connected via directed edges to all other nodes that are obtained by removing R-groups from exactly one substitution site of the original set. Thus, nodes connected by directed edges are involved in parent-child relationships and all molecules that are associated with a parent node are also associated with a child node. However, a child node might contain additional analogs.

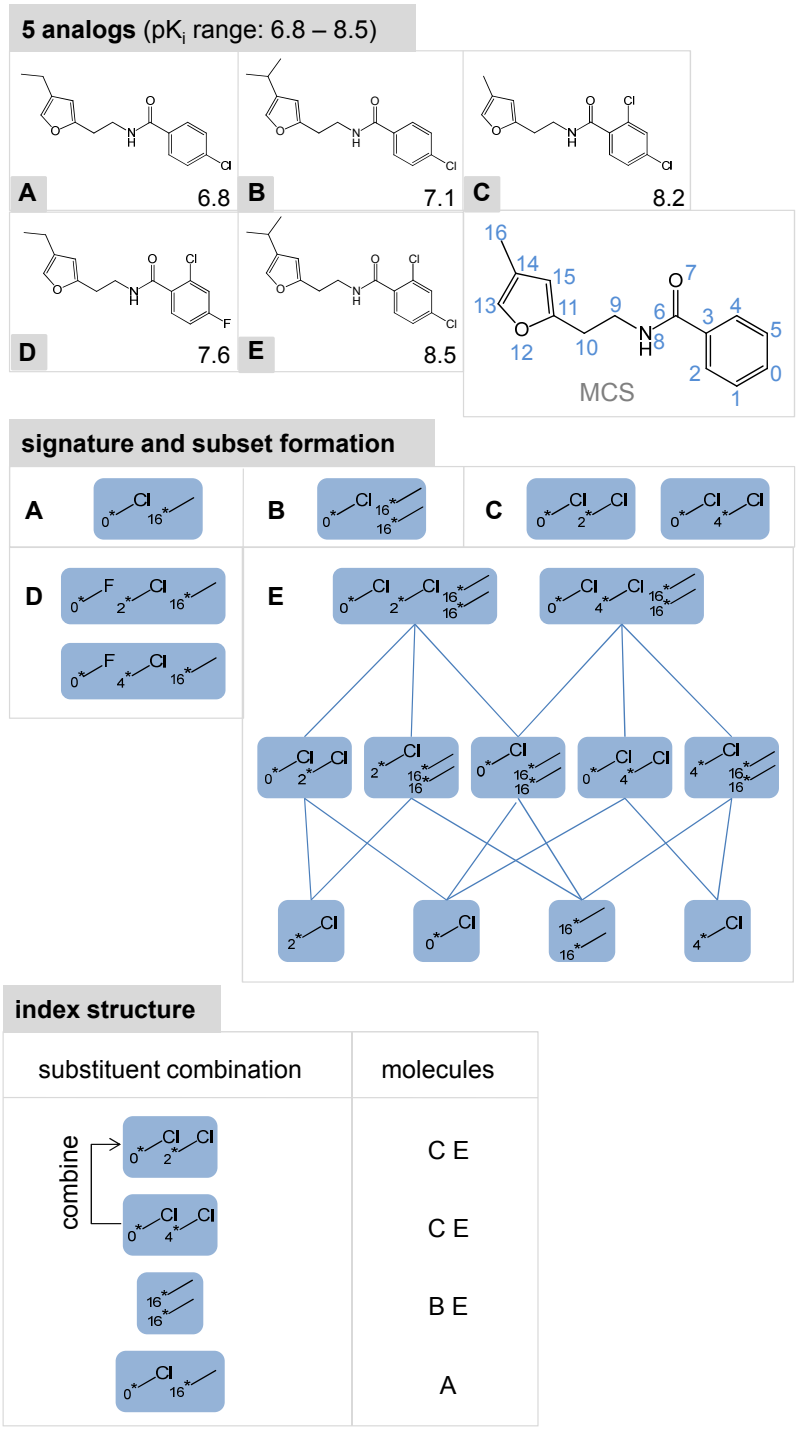


Figure 8.1-1: R-group combinations A model analog series of five compounds is shown (top) together with its MCS. Compounds **A**-**E** are annotated with their logarithmic potency, i.e., pK$_i$ values. Signatures, i.e., sets of R-groups, are extracted from all compounds (middle). For molecule **E**, the generation of signature subsets is illustrated. R-group combinations are then added to an index structure and associated with the analogs in which they occur (bottom). For clarity, only a section of the complete index structure is shown. R-group combinations that are obtained by symmetry-related mappings are combined into a single entry. The figure is adapted from [123].

A child node associated with exactly the same set of molecules as its parent node is removed, which reduces the complexity of the graph by omitting redundant information. This is the case if a smaller R-group combination always occurs in the context of a larger one. If node removal eliminates the only existing pathway between a parent and a grandchild (i.e., another node connected to the child), an edge is inserted that directly connects the parent to its grandchild. The graph structure is iteratively updated after each node removal. The process ends when all redundancies are eliminated. The original unprocessed graph structure for the model analog series in Figure 8.1-1 is shown in Figure 8.1-2a. Because all edges from the top to the bottom of the graph are directed (and follow the same direction), arrows are generally omitted for clarity. Nodes that convey redundant information and are removed from the graph during processing are shown in yellow. The processed graph is depicted in Figure 8.1-2b.

8.1.2.2 Node Types

In the processed graph, two types of nodes are distinguished: nodes that are associated with a single compound are drawn as circles while nodes that represent R-group combinations in multiple analogs are drawn as squares. The size of an analog subset assigned to a square-shaped node is indicated by its frame thickness that increases with the number of compounds.

The different node types are interpreted as follows: the R-group combination represented by a circular node corresponds to the signature (i.e., the complete list of R-groups) of the single molecule that is assigned to the node. Hence, the combination of the signature and MCS defines the molecular structure of the associated compound. However, the signature of a molecule is only associated with a circular node if the corresponding set of R-groups does not occur in any other analog of the series. If the signature of a compound corresponds to a subset of R-groups in other analogs, these analogs are combined and represented by a square-shaped node. Following our terminology, this compound is then "masked" by the square-shaped node. In order to identify a masked compound in the graph, it is symbolized as a rectangle in the lower-right quadrant of the node (Figure 8.1-2b).

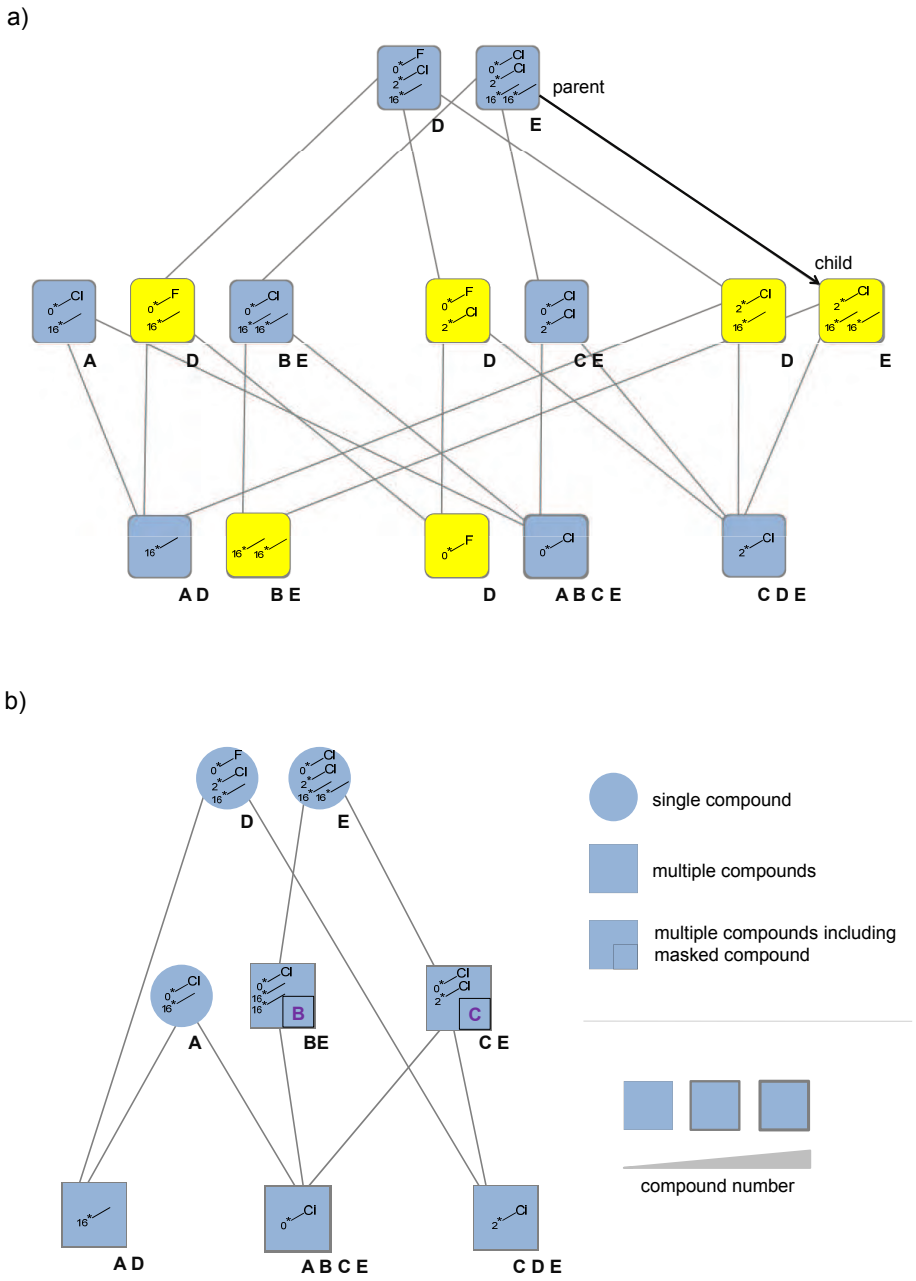


Figure 8.1-2: Graph structure

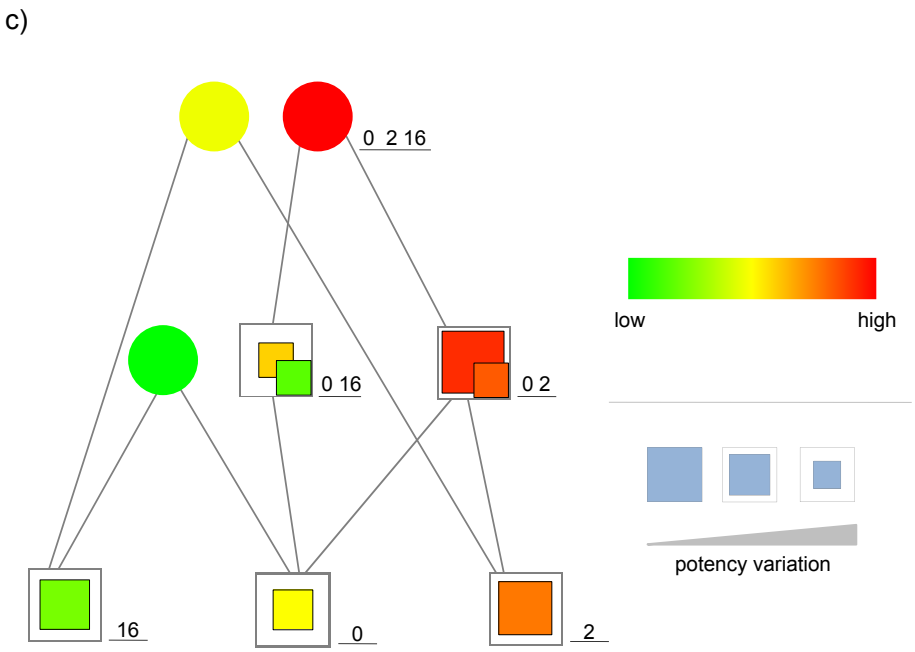


Figure 8.1-2: Graph structure (continued) Schematic illustrations of the graph structure are presented that highlight different design elements. (a) An unprocessed graph is displayed that contains nodes for all substituent combinations found in the model data set shown in Figure 8.1-1. Nodes are associated with all analogs containing the given R-group combination. An exemplary parent-child relationship between two R-group sets and the corresponding directed edge are indicated on the right. Nodes that carry redundant information because they are associated with the same analog subset as a parent node are highlighted in yellow. (b) The processed graph is shown after (i) removal of redundant nodes, (ii) introduction of different node types, and (iii) scaling of node frames according to compound numbers. For two nodes, the masked compounds **B** and **C** are labeled in purple. (c) The graph is displayed with (i) a color-code accounting for (mean) compound potencies and (ii) scaling of color-filled node areas according to potency variations. For clarity, node labels and compound information are not shown. Instead, groups of nodes are labeled with the corresponding substitution site combinations. The figure is taken from [123].

8.1.2.3 Compound Potency Information

All circular nodes are colored according to the potency of the corresponding compounds and the square-shaped nodes are colored according to the mean potency of the associated analogs using a uniform continuous color gradient from green (lowest potency in the data set) to red (highest potency), as illustrated in Figure 8.1-2c. A rectangle symbolizing a masked compound is colored according to its potency (analogous to circular nodes). Square-shaped nodes are often not completely color-filled, for the following reason: if multiple compounds are associated with a node, the area of the node that is colored reflects the standard deviation of potency values. Thus, a node that is completely colored corresponds to a standard deviation of zero, i.e., all associated molecules have the same po-

tency value. For standard deviations larger than zero and smaller than one, the color-filled area continually decreases to half of the original diameter and is then kept constant for standard deviations equal to or larger than one (Figure 8.1-2c). Hence, decreasing color-filled node areas indicate increasing compound potency variations.

As shown in Figure 8.1-2, nodes are arranged in layers that reflect decreasing numbers of substitution sites, i.e., parents are always positioned above their children. Furthermore, within the same layer, nodes representing R-groups at exactly the same substitution sites are grouped together and arranged in order of increasing potency from left to right.

8.1.2.4 Interactive Analysis

R-groups represented by nodes are stored as canonical SMILES strings that serve as node labels. Nodes are associated with tooltips to display R-group structures, report the number of compounds assigned to a node as well as their mean potency, and the potencies of any masked compounds. The graph layout can also be interactively edited. The graph design was implemented using the Java package JUNG.

8.2 SAR Patterns

The DRC graph structure is designed to extract SAR information from R-group patterns in analog series. An important feature of the approach is that any series of analogs can be studied in context, regardless of the number of substitution sites that occur (or the number of compounds). Furthermore, the systematic and hierarchical organization of analogs on the basis of combinations of all R-groups that are available in a series and the analysis of relationships between different sets of R-groups also sets this methodology apart from currently available approaches to study analog series, such as R-group tables and their extensions. In particular, the multiple R-group analysis scheme reveals (i) critical substitution sites, (ii) (un)favorable substituents, (iii) additive and non-additive effects on compound potency as a consequence of multi-site substitutions, (iv) optimization pathways gradually increasing compound potency, and (v) suggestions for analog design. Thus, as shown in the following, the potential of the DRCG approach goes much beyond conventional analysis of analog series.

The DRCG structure contains several well-defined SAR patterns, i.e., subgraphs that reveal immediately interpretable SAR information. These graph components are schematically depicted in Figure 8.2-1 and are rationalized as follows:

- *SAR pattern 1*: R-group combinations that exclusively occur in highly potent compounds are identified by square-shaped nodes filled with red color that, ideally, have a thick frame indicating that the R-group combination has been explored in many different compounds that are consistently highly potent.

- *SAR pattern 2*: critical substitution sites or combinations of sites where structural modifications lead to large differences in potency occur as horizontal node patterns in the graph. In this case, differently colored nodes are grouped together within the same node layer, hence representing different combinations of R-groups at the same substitution sites spanning a wide potency range. It follows that this pattern also provides an immediate access to favorable and unfavorable R-group combinations.

- *SAR pattern 3*: if the nodes forming the horizontal pattern 2 are all connected to the same child node in the subsequent layer, the structural modifications responsible for large potency variations can be traced back to a single substitution site.

- *SAR pattern 4*: a gradual increase in potency resulting from a stepwise addition of R-groups to a starting compound is detected as vertical pattern. Following the path from a node in inverse edge direction (i.e., bottom-up towards its ancestors) leads to increasingly potent analogs.

- *SAR pattern 5*: a parent node that is connected to multiple less potent child nodes indicates that its potency results from the interplay of the different R-group sets associated with the child nodes. The substituent sets of the child nodes are disjoint unless they share a common ancestor. The interplay between different substitution sites and R-groups might result in additive or non-additive effects on compound potency.

- *SAR pattern 6*: under the assumption that favorable R-group effects on compound potency are not compensatory (i.e., that positive effects at two or more sites do not combine in a negative way), compound design suggestions can be easily made on the basis of the DRCG structure. Attractive analogs with presumably high potency can be derived from nodes that represent favorable R-group combinations within the same layer and are connected to a shared less potent child node in the next layer. Thus, starting from the same R-group combination, the introduction of additional R-groups at different substitution sites leads to analogs with increased potency. It follows that new analogs can be immediately suggested that combine the original R-group set with all potency-increasing R-groups introduced at distinct sites.

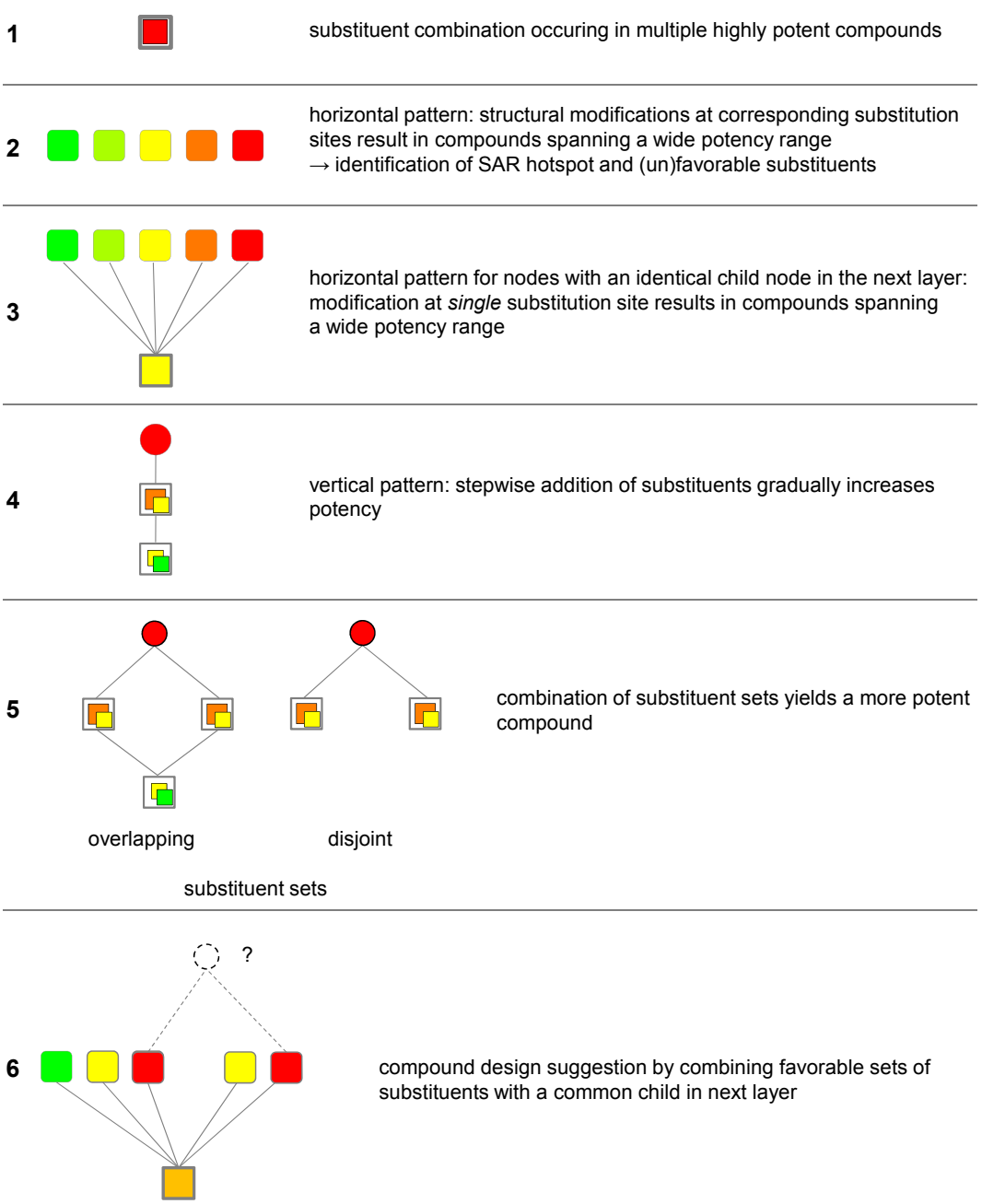


Figure 8.2-1: SAR patterns Subgraph patterns that capture SAR information in a defined manner are shown and explained. The figure is taken from [123].

If SAR information is contained in a series of analogs, it will consistently emerge in the form of the intuitive SAR patterns described above. Therefore, searching a DRCG of any analog series for these characteristic SAR patterns enables the extraction of SAR information, if available in a data set.

8.3 Applications

In the following, four examples of analog series are discussed that contain interpretable SAR information inferrable from DRCG representations.

8.3.1 Compound Data Sets

Four analog series of different composition containing between 31 and 54 compounds were assembled from the ChEMBL database. Compounds active against the human melanocortin receptor 4, norephineprine transporter, and dopamine D1 receptor were extracted. From compounds forming each activity class, Bemis and Murcko scaffolds were extracted and molecules sharing the same scaffold (and activity) were combined into an analog set. Series comprising 30 or more analogs were subjected to DRCG analysis.

8.3.2 Melanocortin Receptor 4 Antagonist Series 1

The first series of melanocortin receptor 4 antagonists consists of 32 analogs with potencies ranging from pK_i 5.2 to 8.9. For this series, a conventional R-group table is provided in Appendix Table F-1 and its DRCG representation is shown in Figure 8.3-1a. Because analogs sharing the same substitution sites are grouped together, the graph reveals that most analogs in this series are characterized by two different substitution site combinations, i.e., sites 0 and 5 (lower horizontal pattern in Figure 8.3-1a) and sites 0, 3, and 5 (upper horizontal pattern). In both cases, R-groups at site 0 vary whereas groups at site 5 or sites 3 and 5 are invariant. This is captured by the graph structure because removal of the substituents at site 0 yields the same child for all nodes of a group, i.e., child nodes annotated with site combinations "3 5" and "5", respectively. The labels of these nodes reveal that the invariant R-groups at positions 3 and 5 are chlorine atoms. Analogs forming both horizontal patterns are arranged in order of increasing potency. In both instances, traversing nodes and associated R-groups from the left to the right reveals that aliphatic amine moieties attached to site 0 via an amide bond or carbamide derivatives are preferred substituents. Exemplary analogs are depicted in Figure 8.3-1b.

Because R-groups at site 0 are highly variable, only three of the analogs in the upper horizontal pattern are derived from others by addition of a chlorine atom to position 3. However, from the vertical pathways involving these analog pairs (highlighted in blue in Figure 8.3-1a) it can be inferred that a chlorine atom at site 3 increases compound potency. The leftmost compound in the upper horizontal pattern is accessible via two edges because an analog that differs from this compound by the absence of the chlorine substituent at site 5 is also available in the data set. The analog without the chlorine at position 5 is also less potent: we can thus conclude that chlorine atoms at sites 3 and also

5 make positive contributions to compound potency, which is also reflected by the framed, red-filled node representing this R-group pair.

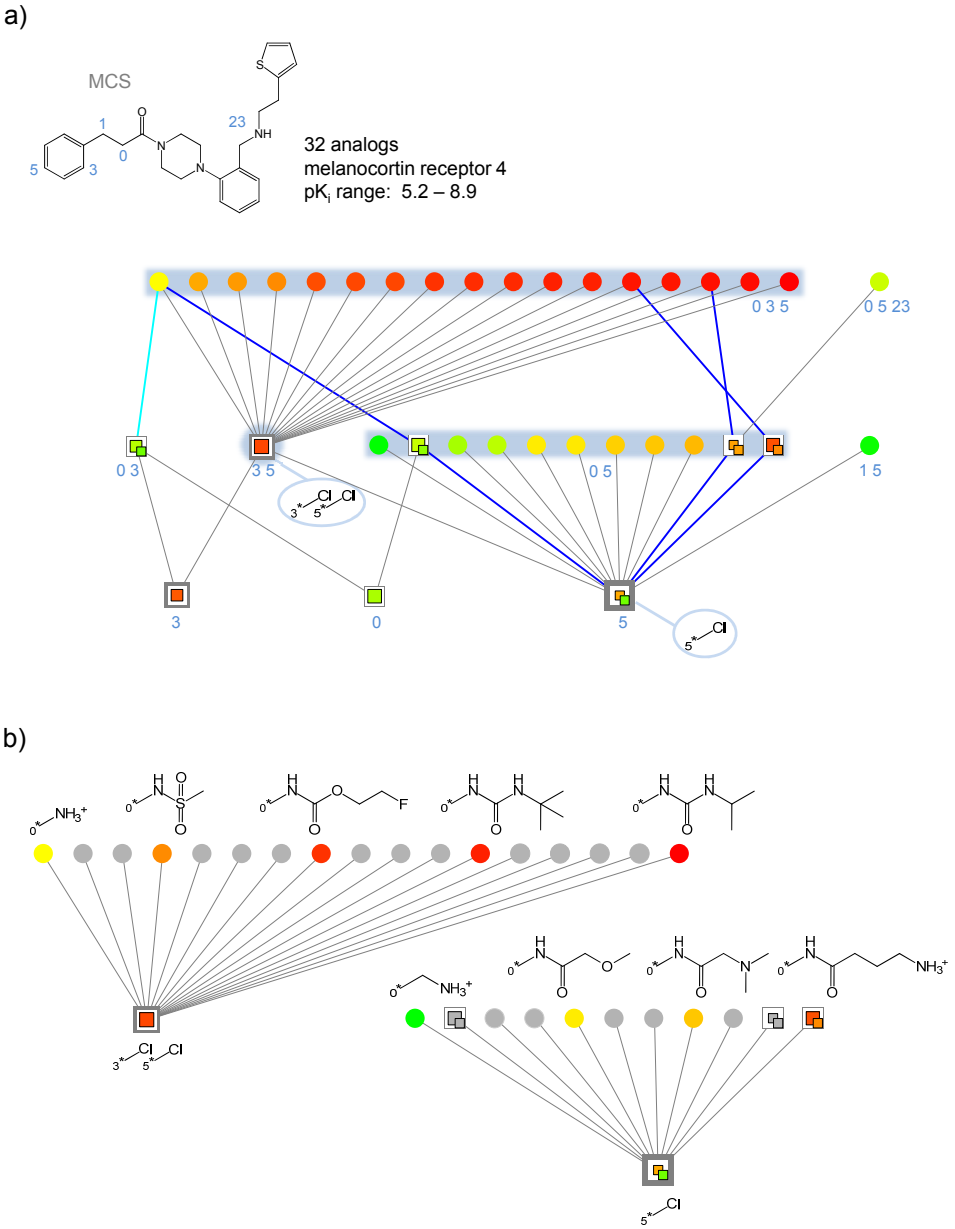


Figure 8.3-1: DRCG for melanocortin receptor 4 antagonists (series 1) (a) The MCS of a series of 32 analogs is shown at the top and substitution sites are labeled with numeric atom identifiers. For two nodes, substituent combinations are provided. Characteristic SAR patterns are numbered according to Figure 8.2-1 and highlighted as follows: patterns 1 and 3, blue node background; pattern 4, blue edges (on the right); pattern 5, combination of blue and turquoise edges (left). (b) Exemplary analogs from horizontal patterns are shown. The remaining nodes are colored gray. The figure is taken from [123].

By comparison with Appendix Table F-1 it is evident that the DRCG structure provides easy access to SAR information contained in this series that would be much harder to extract from an R-group table. In particular, the analysis of multi-site R-group effects would require a comparison of an analog with all others in the R-group table. Furthermore, in the graph structure, a set of R-groups is not only associated with the potency of the analog it defines but also with the mean potency of all compounds in which this R-group combination occurs. For example, the node that represents chlorine at position 5 in Figure 8.3-1a (bottom right) provides the information that this substitution alone yields only a weakly potent compound (masked in this node), whereas its combination with other R-groups at other sites generally produces compounds with increased potency. This type of information conveyed by the DRCG representation helps to identify R-groups that act favorably in combination with others.

8.3.3 Melanocortin Receptor 4 Antagonist Series 2

The DRCG of another structurally distinct series of 54 melanocortin receptor 4 antagonists covering a pK_i range from 5.1 to 8.4 is shown in Figure 8.3-2a. The complete graph reveals that all analogs in this series are substituted at three or more different sites. A region formed by increasingly potent analogs is highlighted. In these analogs, two different combinations of R-groups at substitution sites 2, 16, and 22 are frequently found (corresponding to the two nodes with "2 16 22" annotation at the bottom of the highlighted pattern). Figure 8.3-2b focuses on another combination that consistently produces highly potent analogs. This combination is formed by a 2-methylpropyl-3-(dimethylamino)propanamide group at position 0 and a chlorine atom at site 22. Analog **A** in Figure 8.3-2b that contains only these two substituents is currently untested (as stated above, all known analogs of this series carry R-groups at three or more sites). The hypothetical analog **A** is also associated with the node representing this R-group combination in the displayed graph structure. Edges forming pathways to all compounds that contain this R-group combination are highlighted in blue in Figure 8.3-2b. Furthermore, all analogs that contain the same R-group at site 0 but lack the chlorine substituent at site 22 are reached following the turquoise edges that lead to a cluster of weakly potent compounds highlighted on the right of the graph. However, these compounds are also set apart from the other analogs in the set by the presence of an R-group at site 21, such that it cannot be concluded whether the disruption of the interplay of substituents at sites 0 and 22 or the presence of a substituent at site 21 is detrimental to compound potency. Two analogs **B** and **C** are marked in the graph that contain the 2-methylpropyl-3-(dimethylamino)propanamide group at position 0 and identical R-groups at sites 4 and 16, as indicated by a shared child node. One of these analogs contains an additional chlorine atom at

site 22 and is highly potent whereas the other contains a chlorine atom at site 21 and is only weakly potent. Therefore, it would be suggested to test another hypothetical compound **D** (also shown in Figure 8.3-2b) that does not contain any of the chlorine substituents, which might confirm the potency-decreasing effect of a chlorine substituent at site 21 or the potency-increasing effect of a chlorine atom at site 22.

Overall, most potent compounds are obtained when substitution sites 0, 2, 16, and 22 are simultaneously occupied, as highlighted in Figure 8.3-2a. Many analogs with this site combination contain an isopropyl group at position 16, a chlorine atom at position 22, and a methyl or trifluoromethyl group at position 2, as shown in Figure 8.3-2c. The R-group at site 0 is generally large and has limited structural variability, as illustrated at the top of Figure 8.3-2c, which displays the subgraph associated with these highly potent analogs.

Compounds associated with nodes **B** and **E** in Figure 8.3-2b share the same R-groups at overlapping substitution sites (sites 0, 16, and 22; shared child node) and have consistently high potency. Hence, it would be interesting to combine these favorable R-group sets by generating a new analog **F** that is shown in Figure 8.3-2b.

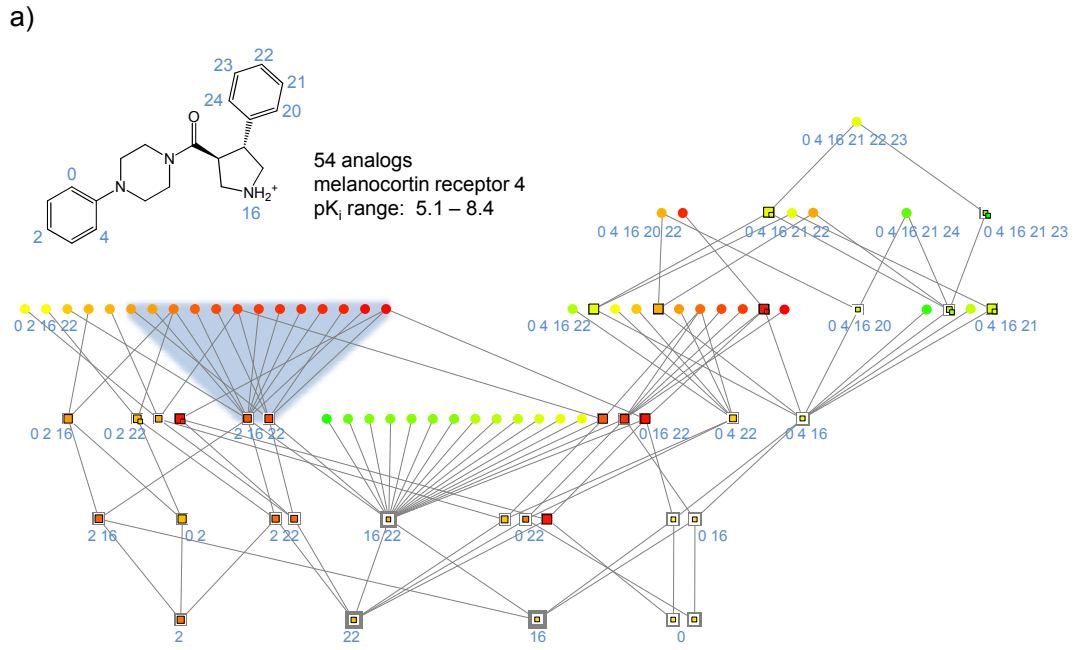


Figure 8.3-2: DRCG for melanocortin receptor 4 antagonists (series 2)

b)

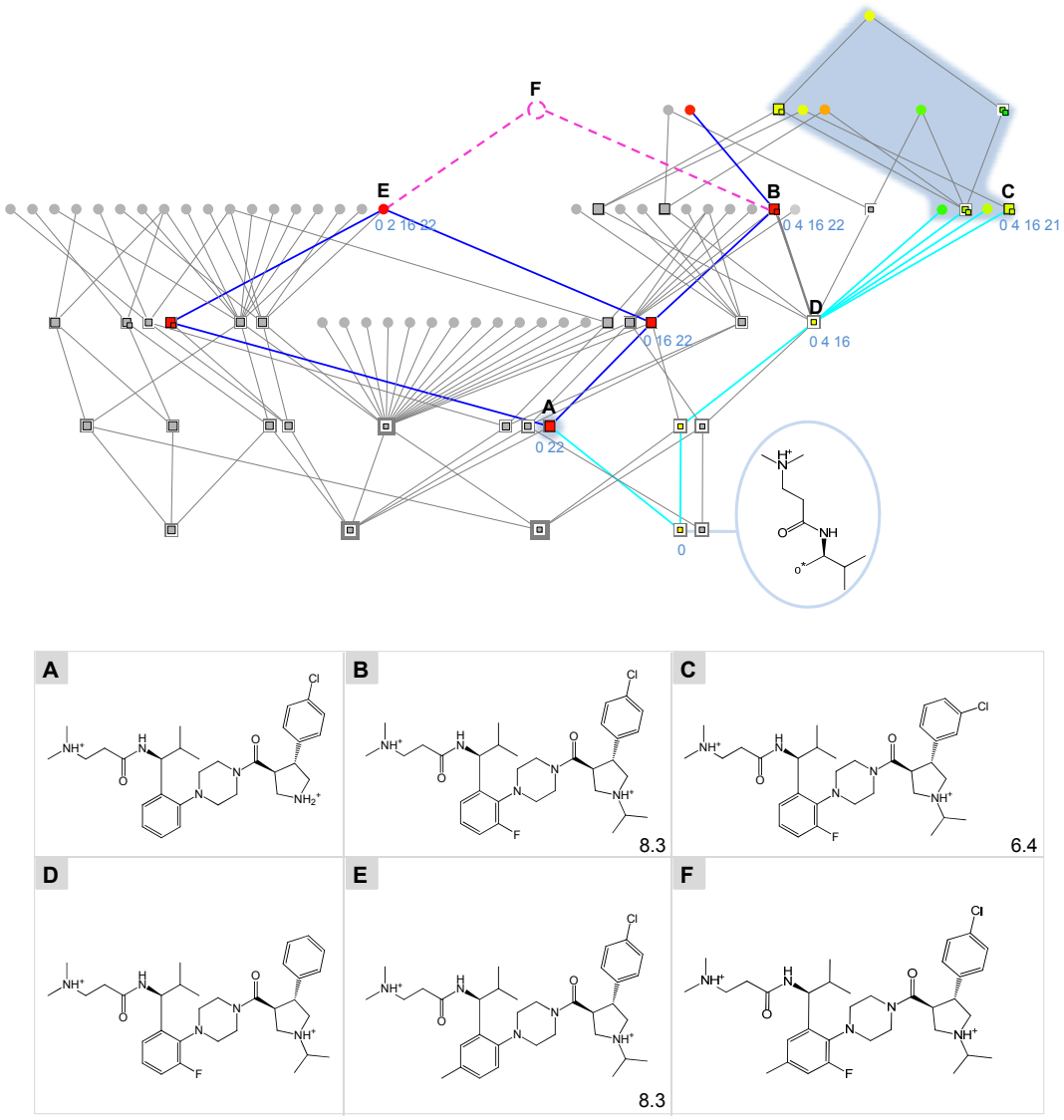


Figure 8.3-2: DRCG for melanocortin receptor 4 antagonists (series 2) (continued)

c)

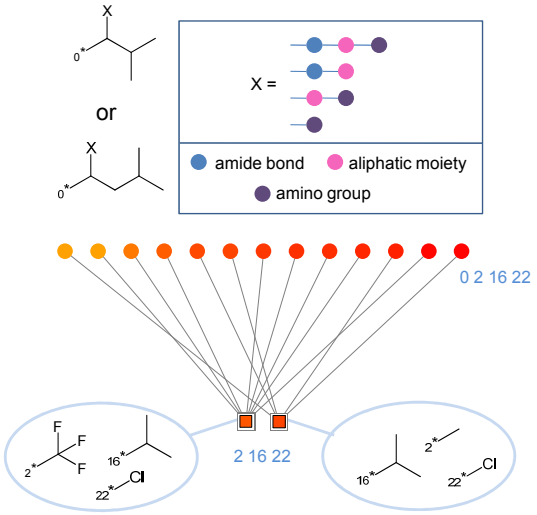


Figure 8.3-2: DRCG for melanocortin receptor 4 antagonists (series 2) (continued) For a series of 54 analogs, the complete graph and SAR information-rich subgraphs are shown. (a) The MCS of all analogs is shown with relevant numeric atom identifiers. A subgraph associated with highly potent compounds and two substituent combinations frequently found in these analogs are highlighted (SAR pattern 1, blue background). (b) Starting from a favorable combination of two R-groups at substitution sites 0 and 22 (SAR pattern 1), edges leading to highly potent compounds containing this combination are shown in blue. A cluster of weakly potent compounds is highlighted (blue background) that contains only one of these two R-groups (displayed substituent at site 0) and additional R-groups at other sites. The path to this cluster is indicated using turquois edges. A compound design suggestion based on SAR pattern 6 is indicated by dashed pink edges. Six nodes are labeled with the identifiers of compounds (**A**-**F**) defined by the corresponding substituent combinations. Structures and, if available, potency information for these compounds are provided at the bottom. (c) The subgraph corresponding to the highlighted SAR pattern in (a) is shown together with R-group information for nodes. The figure is adapted from [123].

8.3.4 Norepinephrine Transporter Inhibitor Series

The DRCG of a series of 41 norepinephrine transporter inhibitors spanning a potency range of approximately three orders of magnitude is shown in Figure 8.3-3. The graph representation reveals that substitution sites 0 and 5 have predominantly been explored in this series. Recurrent R-groups among analogs include a trifluoromethyl group at position 5 and a dimethyl rest at position 0. Interestingly, the introduction of one of these R-groups in isolation only generates a weakly potent compound, but their simultaneous introduction leads to a more than additive increase in potency, yielding one of the most potent analogs in this series (masked compound labeled **A** in Figure 8.3-3). The highlighted horizontal pattern for the combination of substitution sites 0 and 5 shows that the introduction of different R-groups at both sites has large effects on compound potency. Moreover, as revealed by the weakly potent compounds on the

right in Figure 8.3-3, the introduction of additional substituents at the meta positions of the phenyl ring displays a strong tendency to decrease potency. For example, when adding a trifluoromethyl group to the most potent analog of this series (**B** in Figure 8.3-3) at site 8 (yielding analog **C**) or site 6 (analog **D**), potency is reduced by more than two orders of magnitude. In general, the relationships between di- and trisubstituted analogs in the DRCG of this series indicate that the addition of R-groups at sites other than atom positions 0 and 5 does not lead to notable increases in potency.

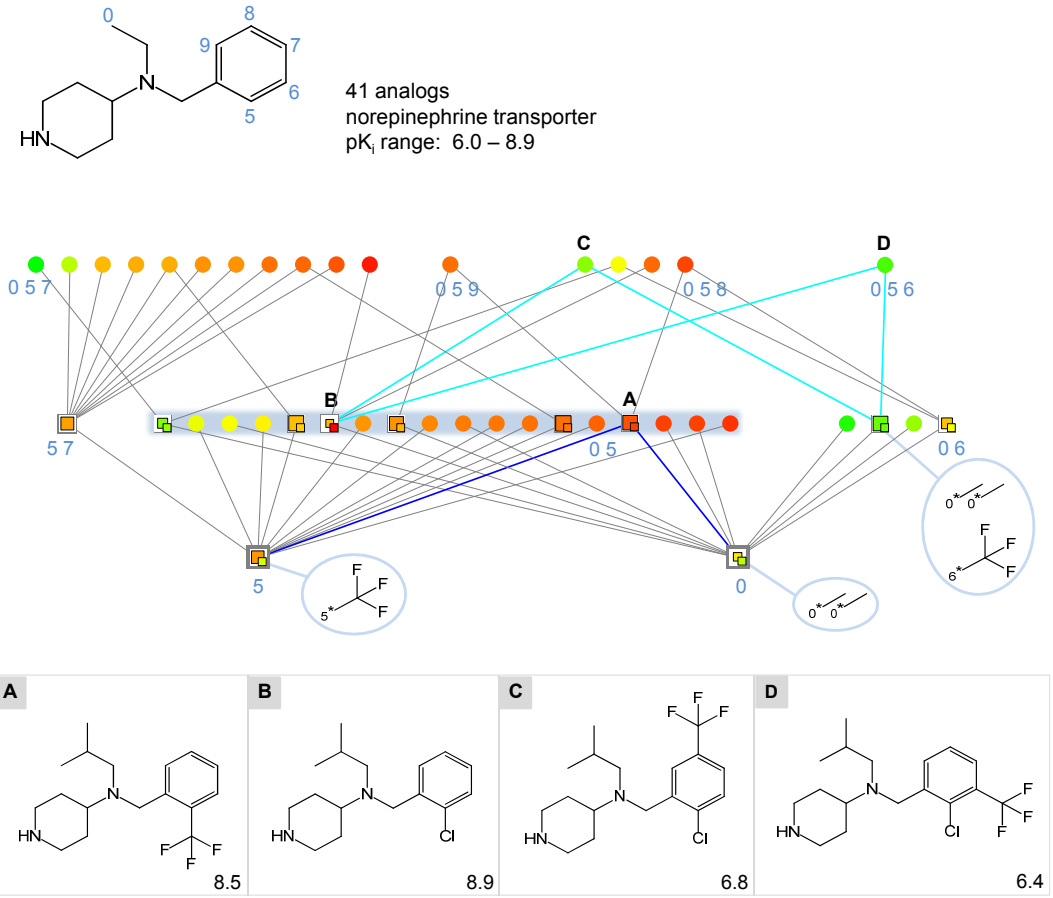


Figure 8.3-3: DRCG for norepinephrine transporter inhibitors The MCS extracted from a series of 41 analogs is shown with numeric atom identifiers. For three nodes, substituent combinations are shown. In addition, four nodes are labeled with the identifiers of compounds (**A**-**D**) defined by the corresponding substituent combinations. Structures and potency information for these compounds are provided at the bottom. SAR patterns are highlighted: pattern 2, blue node background; pattern 5, blue edges. In addition, detrimental effects of R-group combinations on compound potency are indicated using turquoise edges. The figure is adapted from [123].

8.3.5 Dopamine D1 Receptor Antagonist Series

The DRCG of a series of 31 dopamine D1 receptor antagonists that span a comparably narrow pK_i range from 6.8 to 8.8 is shown in Figure 8.3-4a. As revealed by the highlighted horizontal pattern, the introduction of different chemical groups at the meta position of the terminal phenyl ring (designated with numeric identifier 21 in the graph) leads to largest potency fluctuations within this series. At the left and right of this pattern, the trifluoromethyl and methyl group are identified as least and most favorable R-group at this site, respectively. In addition, considerable potency increases are observed for all analogs having an R-group at the ortho position of the phenyl ring (site 20). Two vertical patterns are highlighted in Figure 8.3-4a where the subsequent addition of R-groups leads to stepwise increases in potency. Both pathways begin at the same analog that carries a chlorine substituent at site 11 and is only moderately potent (analog **A**). The addition of a fluorine atom at position 20 then leads to a potency increase of approximately one order of magnitude (analog **B**). Another order of magnitude is gained by adding a trifluoromethyl group at the other ortho position in the ring (site 24, analog **C**). Similar potency changes are detected for the stepwise addition of two methoxy groups at the corresponding positions (analogs **D** and **E**). However, as depicted in more detail in Figure 8.3-4b, potency changes of larger magnitude are observed for compounds **F** and **G** that are also derived from analog **A**. In these cases, both the introduction of a chlorine atom at the ortho position or a methyl group at the meta position of the terminal phenyl ring increase compound potency by two orders of magnitude. Hence, it would be attractive suggesting two additional analogs that combine these favorable substitutions (i.e., hypothetical molecules **H** and **I** in Figure 8.3-4b) in order to further increase compound potency within this series.

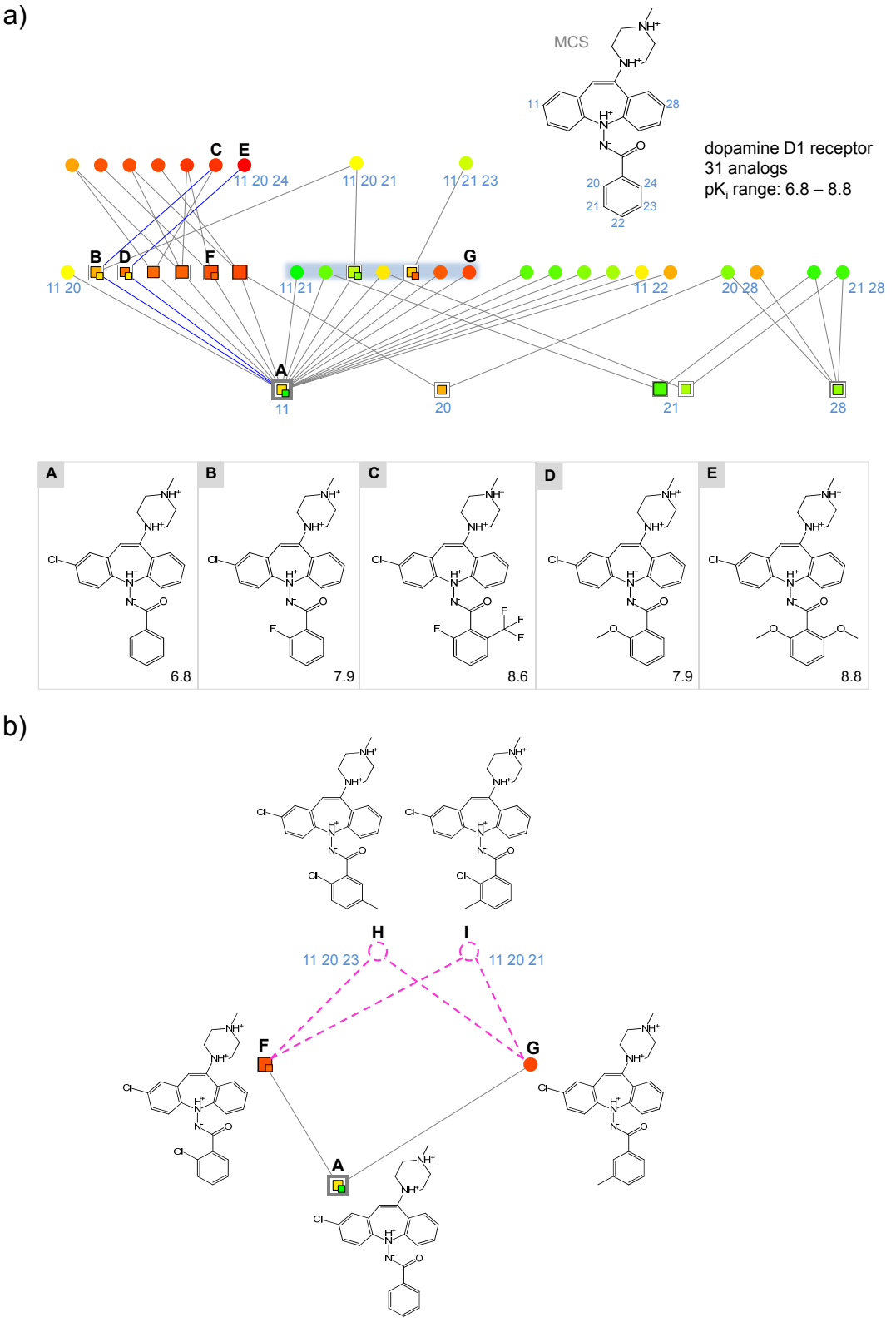


Figure 8.3-4: DRCG for dopamine D1 receptor antagonists For a series of 31 analogs, the complete graph and a subgraph illustrating compound design suggestions are shown. (a) The MCS with relevant numeric atom identifiers is shown. Nodes are labeled with compound identifiers. Structures and pK_i annotations are displayed for compounds **A-E**. Characteristic SAR patterns are highlighted: pattern 3, blue node background; pattern 4, blue edges. (b) Compound design suggestions (**H**, **I**) are made based on SAR pattern 6. The figure is adapted from [123].

8.4 Conclusions

In this chapter, we have introduced a new graphical SAR analysis concept specifically developed for the study of analog series and for compound design. Instead of individual compounds, systematically derived R-group combinations provide the basis for the construction of the DRCG structure, which is a central feature of the approach. The graphical representation contains a number of design elements that emphasize available SAR information. From the hierarchical organization of R-group combinations and corresponding analog sets, characteristic subgraphs emerge that represent well-defined SAR patterns. If analog series are characterized by the presence of multiple substitution sites and R-group combinations, it is usually difficult to rationalize SARs by comparing individual analogs and their potency values in R-group tables or subsets of analogs with modifications at pairs of substitution sites. By contrast, in DRCGs, entire analog series are consistently represented, regardless of the numbers of analogs and substitution sites, and emerging SAR patterns reveal interpretable SAR information. Importantly, subset relationships between R-group combinations emerge from this data structure such that potency changes resulting from the removal (addition) of a substituent from (to) a given R-group combination can be monitored.

Source Information

Sections of the text in this chapter have been taken from [123].

Chapter 9

Selectivity Determinants in Series of Analogs

Lead optimization traditionally has a strong single-target focus. In fact, the premise that specific (and exclusive) interactions of a compound with an individual target are a prerequisite for desired biological efficacy represents a paradigm that has for long been a cornerstone of drug discovery research [63]. However, with increasing evidence of polypharmacological behavior of biologically active compounds, these views are beginning to be revised [65]. It is now better understood that many pharmaceutically relevant compounds act on multiple targets and promiscuous multi-target interactions are known to be favorable in several instances, for example, for the treatment of cancer using protein kinase inhibitors [64].

However, target-specific compounds continue to be of critical importance, considering that adverse drug reactions, which often result from a lack of molecular specifity, are estimated to be the fourth leading cause of death in western countries [124]. Furthermore, high selectivity is required to combat pathogenic infections, for example, when targeting bacterial enzymes for which human orthologs exist. Consequently, if one intends to focus on an individual target, one usually attempts to render compounds with multi-target activity target-selective through chemical optimization efforts. For this purpose, multi-target structure-activity relationships (mtSARs) are studied and compared, which represents a rather complicated task for classical analog design strategies. In order to achieve target selectivity, or specificity, one ultimately needs to identify functional groups that are selectivity determinants, which is often difficult. To study SARs in a consistent manner, we have developed a methodology for mtSAR analysis [125] that combines information about 2D pharmacophore feature similarity and compound potency distributions within compound series with combinatorial analogue graphs (CAGs) that are specifically tailored toward the detection of critical, i.e., activity-determining substitution sites in analog

series [42]. For the study of mtSARs, we revised the CAG data structure to provide better access to types of substitutions responsible for apparent SAR discontinuity and to facilitate side-by-side comparisons of CAGs for multiple targets. The enhanced data structure is capable of extracting available mtSAR information from compound data sets and differentiate it in a target-directed manner, as demonstrated in this chapter. Section 9.1 describes methodological details of our approach and emphasizes modifications that were made to the original CAG implementation it was built on. In section 9.2, it is shown how the introduced analysis scheme can be used to derive preference orders for selectivity-conferring substitution patterns and SAR rules. Section 9.3 reports two exemplary applications of the methodology to publicly available compound data. Concluding remarks are found in section 9.4.

9.1 Methodology

The CAG data structure was designed to analyze SARs of analog series and is generated on the basis of three operations: R-group decomposition, similarity assessment, and SAR discontinuity evaluation [42].

9.1.1 CAG Data Structure

For each analog series, the MCS shared by all analogs is determined using Pipeline Pilot. The MCS is then used as the invariant core structure for R-group decomposition to determine corresponding substitution sites in analogs and assign sets of substituents to these sites. For all possible pairs of compounds, substitution sites that carry different R-groups are determined and compound pairs are grouped into subsets according to the substitution site combination S at which they differ, as shown in Figure 9.1-1. Symmetrically substituted molecules are not firmly associated with only one possible mapping onto the MCS, but a symmetrically substituted compound is compared with all remaining compounds in the analog series, and each compound pair is mapped such that the number of identical substituents shared by the two molecules is maximal. If several mappings are possible for a compound pair based on this selection criterion, all mappings are retained and the compound pair is assigned to multiple subsets. Furthermore, subsets containing pairs of compounds that show differences at more than three substitution sites are discarded. Then, the degree of discontinuity found for compound pairs in each subset is assessed by using the SAR Index discontinuity score [29], a numerical function that monitors the presence of activity cliffs in compound classes and yields high values for sets of molecules that contain structurally similar compounds with large po-

tency differences. For each subset, a raw discontinuity score is calculated for all compound pairs i and j that differ at the given substitution site combination S

$$\text{disc}_{\text{raw}} = \underset{\{(i,j)|i,j \text{ differ at } S\}}{\text{mean}} \Big(|P_i - P_j| \times \text{sim}(i,j) \Big) \tag{9.1}$$

where P_i and P_j denote the negative decadic logarithm of potency values of compounds i and j and sim(i,j) corresponds to a similarity function. Raw discontinuity scores are converted into Z-scores by using the sample mean and standard deviation of scores obtained for analog pairs in target family-specific reference sets. Z-scores are then mapped onto the value range [0,1] by calculating the cumulative probability for each score under the assumption of a normal distribution. Because scores are not specific for a single activity class but normalized with respect to reference compounds from a target family, the magnitude of scores can be directly compared for related targets. A high discontinuity score indicates the presence of activity cliffs in compound data sets. As scores are calculated for subsets of compound pairs that differ only at well-defined substitution sites, they reflect the SAR discontinuity introduced by individual substitutions or combinations of substitutions and highlight atom positions in the MCS where minor structural modifications lead to considerable potency changes.

To provide immediate access to critical substitution sites, the results of R-group decomposition and discontinuity calculations are visualized in a CAG, as shown in Figure 9.1-2. A CAG is a graph that consists of nodes and edges connecting individual nodes. Here, nodes correspond to subsets of compounds varying at specific substitution sites identified by node labels. Edges indicate that compound pairs in connected subsets share modifications at the same substitution sites (e.g., site 1, 1 and 2 (1-2), and 1-2-3). Thus, CAGs hierarchically organize analog series according to substitution sites. Nodes are color-coded according to discontinuity scores using a continuous color spectrum from green (low discontinuity score) to red (high discontinuity score). The discontinuity score for the root node is calculated by taking only those compound pairs into account that differ in no more than three substitution sites. CAGs are calculated and displayed using the statistical computing language R [42].

As shown in Figure 9.1-2, the graph structure allows the straightforward identification of "SAR hotspots" (red nodes), i.e., substitution sites where changes are most likely to introduce SAR discontinuity and produce compounds with large differences in potency. Furthermore, CAGs enable the identification of "SAR holes", i.e., combinations of substitution sites, that have not yet been explored. Hence, useful suggestions which compounds to synthesize next and how to complement a current series can be derived from considering relationships between SAR holes and hotspots. In the analysis of mtSARs, a side-by-side comparison of CAGs is useful to identify commonalities or differences in SAR hotspots for related targets. However, based on the original CAG data

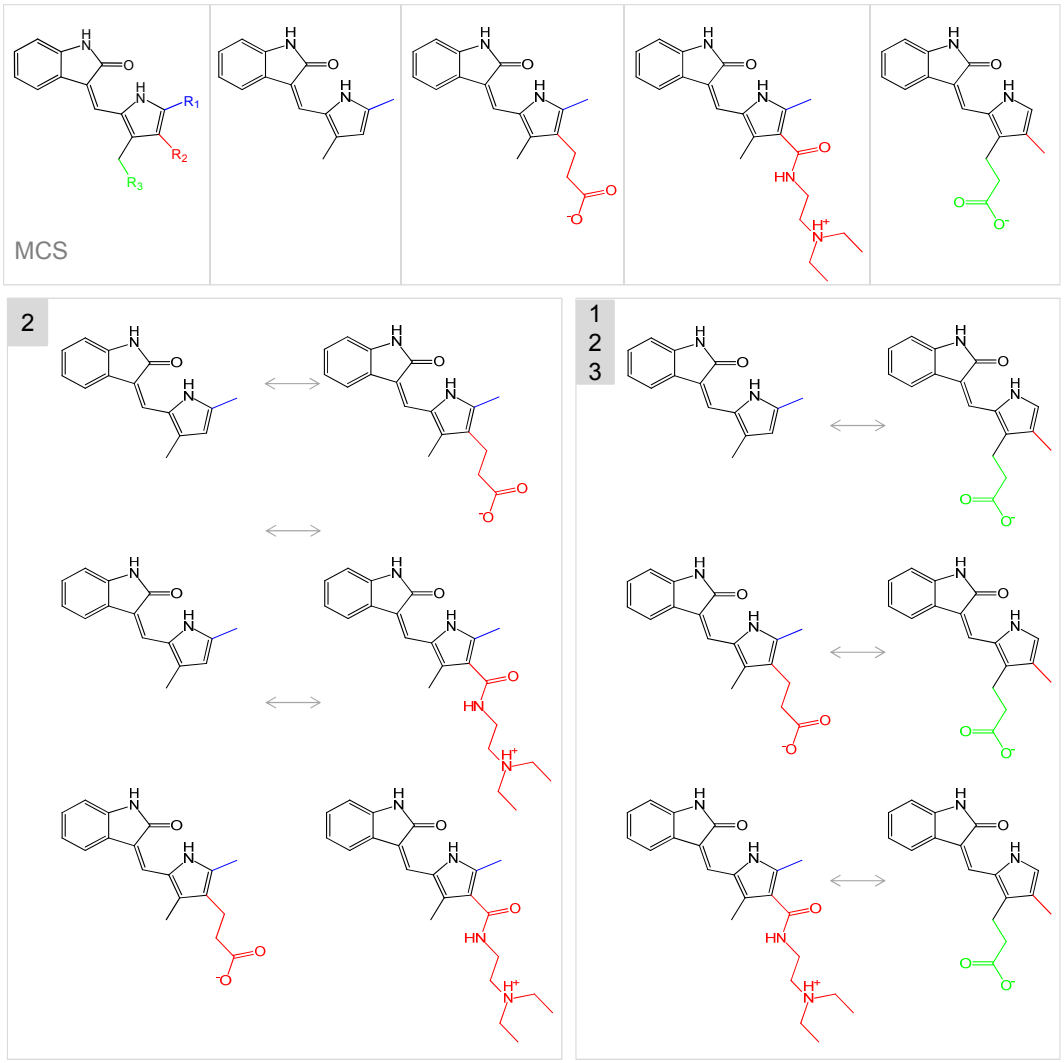


Figure 9.1-1: CAG subset formation Four compounds and their MCS are shown. Pairs of compounds are built that differ either at substitution site 2 or at sites 1, 2, and 3. Compound pairs are grouped accordingly.

structure, it is generally difficult to determine which type of substitutions are responsible for introducing discontinuity and, consequently, SAR rules toward target selectivity are difficult to formulate.

9.1.2 Adaptation to Multi-Target SAR Analysis

The most critical modifications made to the original CAG data structure to facilitate the analysis of mtSARs address the encoding of structural variation and the similarity assessment of molecules. Previously, pairwise compound similarity was calculated as conventional whole-molecule similarity based on MACCS

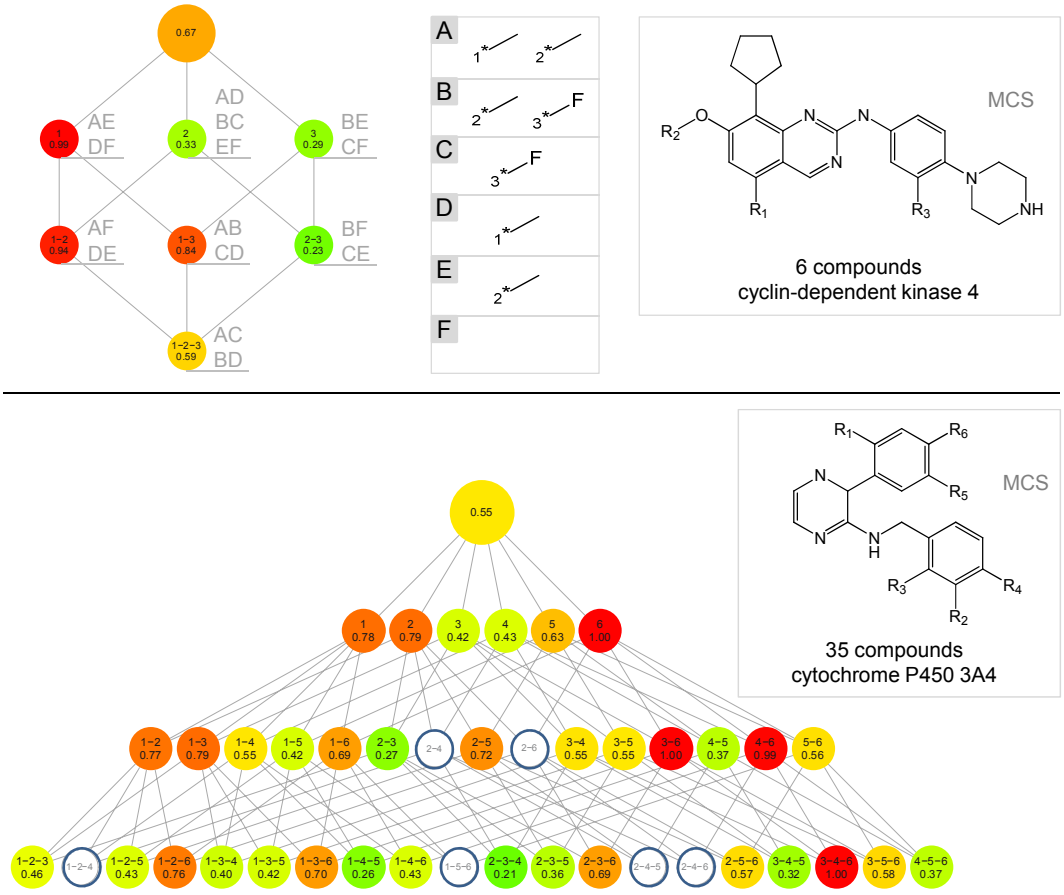


Figure 9.1-2: Combinatorial analogue graphs CAG representations for six and 35 analogs active against the protein targets cyclin-dependent kinase 4 (cdk4) and cytochrome P450 3A4 (cyp P450 3A4), respectively, are shown. For the cdk4 inhibitors, substituents of the individual compounds and the assignment of compound pairs to corresponding nodes are reported. SAR holes in the CAG representation for cyp P450 3A4 are circled in blue. The figure is adapted from [31].

structural keys and the Tanimoto coefficient. In the revised CAG data structure for mtSARs, to directly assess chemical variations of substituents, we compare molecules on the basis of substituent pharmacophore feature similarity. Therefore, substituents are assigned to pharmacophore feature classes by following the scheme of Harper et al. [126]. First, substituents are classified based on whether they are acyclic or contain an aromatic or aliphatic ring structure (three constitutional classes). Then, a substituent is further assigned to one of the six classes "positively charged", "negatively charged", "donor", "acceptor", "donor and acceptor", or "featureless". Combining these six classes with the three constitutional classes yields 18 possible pharmacophore feature categories that are listed in Table 9.1-1. For the purpose of our analysis, the classification scheme was further extended by an additional class: "no substituent". Because a

Table 9.1-1: Substituent pharmacophore feature classes

pharmacophore feature class	abbr.	pharmacophore feature class	abbr.
acyclic positively charged	Ac-P	aliphatic ring positively charged	Al-P
acyclic negatively charged	Ac-N	aliphatic ring negatively charged	Al-N
acyclic donor	Ac-D	aliphatic ring donor	Al-D
acyclic acceptor	Ac-A	aliphatic ring acceptor	Al-A
acyclic donor and acceptor	Ac-DA	aliphatic ring donor and acceptor	Al-DA
acyclic featureless	Ac	aliphatic ring featureless	Al
aromatic ring positively charged	Ar-P	no substituent	(-)
aromatic ring negatively charged	Ar-N		
aromatic ring donor	Ar-D		
aromatic ring acceptor	Ar-A		
aromatic ring donor and acceptor	Ar-DA		
aromatic ring featureless	Ar		

The 19 pharmacophore feature classes used for encoding substituents and their abbreviations ("abbr.") are listed.

given substituent might formally be assigned to more than one feature class, the following order of priority is applied according to Harper et al. [126] to unambiguously define the class membership of each substituent: "positively charged" > "negatively charged" > "donor" or "acceptor". Since pharmacophore features have different degrees of similarity, different edit distances or "costs" are associated with feature replacements, as reported in Table 9.1-2. These costs are defined in analogy to the weights used by Harper et al. for calculating an edit distance for pairs of reduced graphs [126]. For comparison of two analogs, costs are summed for all operations needed to convert one analog into the other, which yields the distance between the compounds. Similarity is then calculated as the complement of the edit distance, as shown in Figure 9.1-3. Hence, for an analog series with n substitution sites, the similarity between two compounds i and j is calculated as follows:

$$\mathrm{sim}(i,j) = 1 - \sum_{s=1}^{n} \mathrm{cost}(R_s^i, R_s^j) \tag{9.2}$$

where R_s^i and R_s^j denote the pharmacophore classes of compounds i and j at substitution site s ($s = 1, \ldots, n$). As costs for a single substitution range from 0 to 0.2, similarity values for compounds that differ only at one substitution site lie between 0.8 and 1, those for compounds that differ at two sites between 0.6 and 1 and those for compounds that differ at three sites between 0.4 and 1. A similarity value of one means that all substituent exchanges always occur within the same pharmacophore class.

Analog series generally display a high degree of whole-molecule similarity. Therefore, using pharmacophore feature edit distances instead of Tanimoto

Table 9.1-2: Pharmacophore feature substitution matrix

Class	(-)	Ac-P	Ac-N	Ac-D	Ac-A	Ac-DA	Ac	Ar-x	Al-x
(-)	0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Ac-P		0	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Ac-N			0	0.2	0.2	0.2	0.2	0.2	0.2
Ac-D				0	0.2	0.1	0.2	0.2	0.2
Ac-A					0	0.1	0.2	0.2	0.2
Ac-DA						0	0.2	0.2	0.2
Ac							0	0.2	0.2
Ar-x								0.1	0.2
Al-x									0.2

Costs are reported for substitutions of pharmacophore features. Feature classes are abbreviated according to Table 9.1-1, except that all aromatic and aliphatic substituents are combined here into two classes Ar-x and Al-x, respectively, because all Ar-x/Al-x substitutions are assigned the same costs. Costs reported apply only to feature class mismatches, i.e., the substitution cost of an aromatic positively charged group to an aromatic negatively charged group is 0.1, but the cost for the replacement of one aromatic positively charged group by another one is zero.

similarity puts more emphasis on small structural modifications that lead to high differences in potency. Furthermore, it is now possible to directly relate apparent SAR discontinuity to underlying pharmacophore feature exchanges, as further explained in the following.

9.2 Derivation of SAR Rules

The introduction of pharmacophore feature transformations for pairs of compounds makes it possible to group compounds according to defined changes in substituents and infer systematic SAR trends for subsitution site combinations that can be compared for the targets under study and exploited in the design of target-selective compounds.

9.2.1 Substituent Preference Orders

Figure 9.2-1 summarizes how the SAR of an analog series against an individual target is analyzed. Analogs in CAG representations only differ at given substitution sites and thus the local discontinuity scores account for SAR contributions of R-groups at these sites. Compounds representing SAR hotspots are most relevant for further analysis, as illustrated for node 2 in Figure 9.2-1. The encoding of substituents as defined pharmacophore features makes it possible to group compounds according to changes in substituents that are required to transform one molecule into the other. For example, the first operation shown in Figure

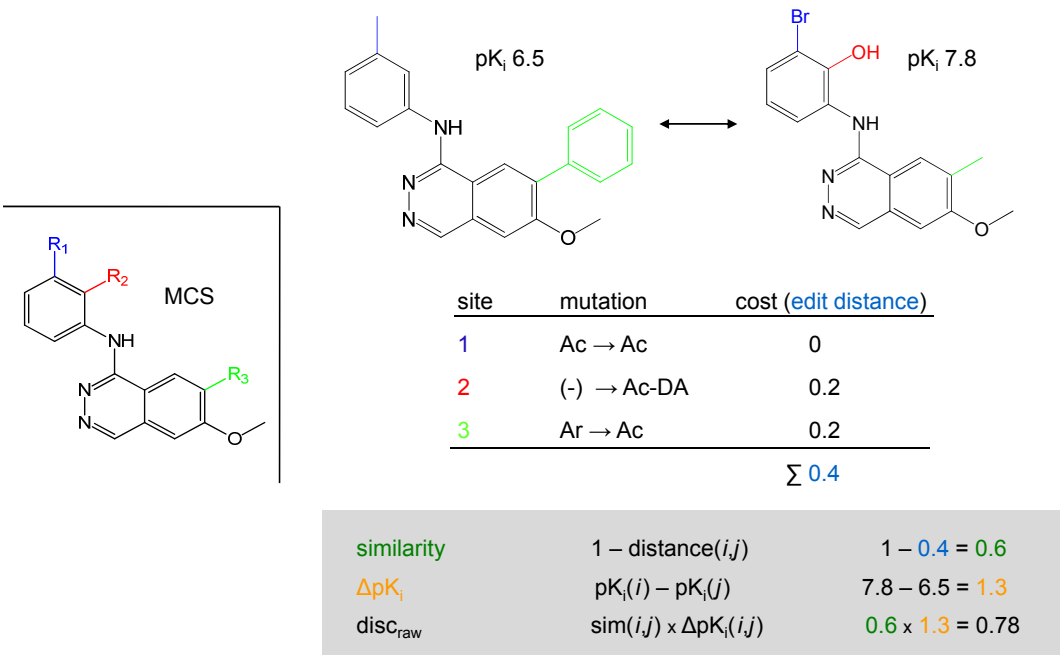


Figure 9.1-3: Similarity calculation Substituents are assigned to classes of pharmacophore features and pairwise compound similarity is calculated based on pharmacophore feature edit distances of different substituents.

9.2-1 is the introduction of a featureless aromatic group ("(-) → Ar") at substitution site 2. One exemplary compound pair falling into this substitution category is shown at the top. The introduction of the featureless aromatic group leads to an increase in potency (the potency difference is given on a logarithmic scale). For all other compound pairs falling into this substitution category, potency changes are recorded and the process is repeated for all other observed substitution categories. The negative sign for the operation listed in the third row in Figure 9.2-1, i.e., substitution of a featureless acyclic group by a featureless aromatic group ("Ac → Ar"), indicates that the substitution decreases potency. Node-specific records are easily interpretable and allow several conclusions to be drawn:

1. Records mirror how thoroughly substitution sites have been explored and which pharmacophore features have been investigated.

2. They identify substitutions causing most significant changes in potency.

3. Given that the same type of pharmacophore transformation is consistently associated with similar potency changes, they enable the derivation of "preference orders" of pharmacophore features for given substitution sites.

Such preference orders are of prime importance to explore and exploit mtSARs because the comparison of preference orders makes it possible to formulate

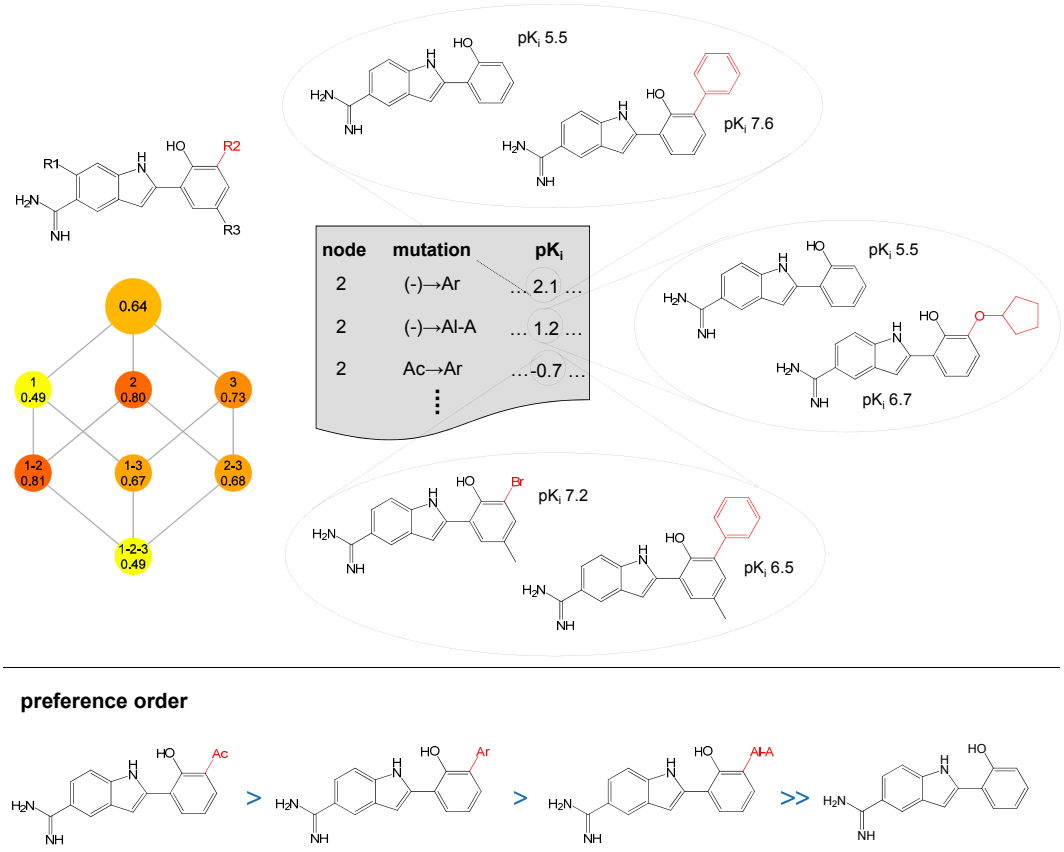


Figure 9.2-1: Analysis of pharmacophore feature-dependent potency changes For each substitution site or combination of sites that is assigned a high discontinuity score, compound pairs are grouped according to the substitutions required to transform one analog into the other and potency changes observed for these substituent exchanges are recorded. Based on the direction (increase or decrease) and magnitude of potency changes, preference orders for pharmacophore features at individual substitution sites or combination of sites can be derived. The figure is adapted from [125].

SAR rules for the design of target-selective compounds. The derivation of a preference order is also illustrated for the compound series in Figure 9.2-2: the introduction of an aliphatic H-bond acceptor and a featureless aromatic group leads to increased potency. The potency change is larger for the aromatic substituent and hence indicates a higher preference for this pharmacophore class. However, it is not the most preferred feature at this site because the comparison of featureless acyclic and aromatic groups reveals that an acyclic group is even more favorable. Taken together, these observations result in the following preference order that prioritizes substitutions at site 2: featureless acyclic group > featureless aromatic group > aliphatic H-bond acceptor > no substituent.

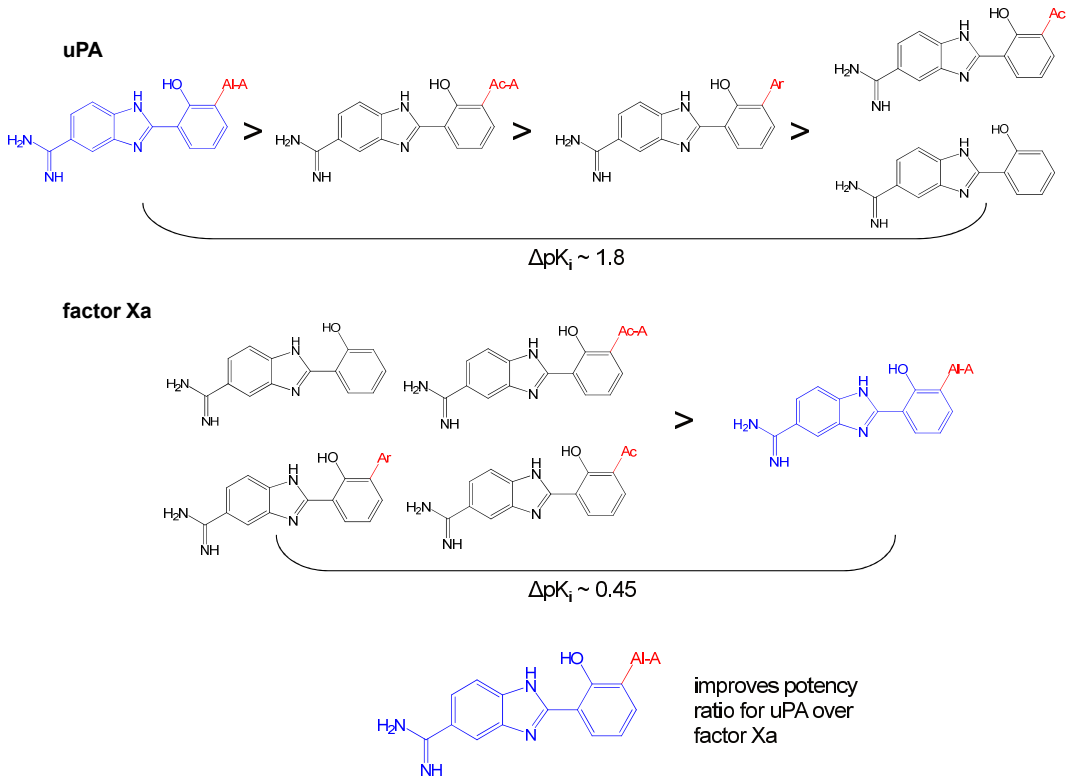


Figure 9.2-2: Identification of selectivity determinants Preference orders for specific chemical modifications at a given substitution site are compared. In this case, an aliphatic H-bond acceptor is the most preferred substituent for inhibitors of enzyme 1 (urokinase) but the least favorable for enzyme 2 (factor Xa). Consequently, the introduction of a substituent corresponding to this pharmacophore feature is likely to increase the potency ratio for enzyme 1 over 2. The figure is adapted from [125].

9.2.2 Design of Target-Selective Compounds

After separate analyses of SARs of an analog series for multiple targets, SAR hotspots can be compared across these targets. If targets have different SAR hotspots or if they share critical substitution sites but differ in their preference order of substituents, rules for the optimization of compound potency and selectivity can be derived. The latter case is illustrated in Figure 9.2-2 for an analog series active against two serine proteases, urokinase (uPA) and factor Xa. Preference orders of pharmacophore features at substitution site 2 are displayed for both enzymes. Pharmacophore features having comparable effects on potency are grouped together and are not separated by ">". In this example, comparing the preference orders reveals that an aliphatic H-bond acceptor is the most preferred substituent at site 2 for uPA, whereas this substituent is least preferred for factor Xa. Hence, the introduction of an aliphatic H-bond acceptor

increases the potency ratio for uPA over factor Xa, which can be exploited to render analogs selective for uPA.

9.3 Applications

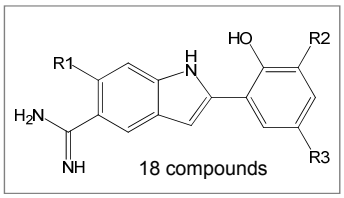
By studying mtSARs, we ultimately address the question of compound selectivity. For compounds active against multiple targets, we aim to understand SAR characteristics for each individual target, compare these SARs, and differentiate between conserved and non-conserved SAR rules, which can be explored to search for selectivity determinants. Thus, the study of mtSARs leads to an analysis of structure-selectivity relationships (SSRs). A major goal of SSR analysis is the derivation of structural/substitution rules for designing target-selective compounds. The utility of CAGs to identify SSR determinants and make compound design suggestions is exemplified in two applications on the serine protease family.

9.3.1 Compound Data Sets

Ten analog series with reported inhibitory activity against the human serine proteases factor Xa, thrombin, trypsin, and uPA were extracted from BindingDB. All analog series were annotated with potency information (K_i or IC_{50}) for at least two related targets. Similarity and discontinuity calculations were carried out for all analog series and the resulting values provided the basis for discontinuity score normalization (see paragraph 9.1.1). In the following, exemplary results are reported and discussed for two analog series of serine protease inhibitors consisting of 18 and 14 compounds, respectively. The two discussed serine protease inhibitor series are active against human uPA, factor Xa, and trypsin. Only compounds with potency values reported against all three proteases were taken from BindingDB. In order to assess predictions of potency-increasing and target selectivity-conferring substitutions, the database was searched for analogs with dual protease annotations that were not included in the analyzed series and hence not utilized to derive target selectivity rules. It was examined whether analogs contained predicted selectivity determinants.

9.3.2 Serine Protease Inhibitor Series 1

Figure 9.3-1 shows CAG representations for serine protease inhibitor series 1 consisting of 18 analogs active against human uPa, factor Xa, and trypsin. Calculation of the MCS yields a core structure with three substitution sites. Comparison of the individual CAGs in Figure 9.3-1 clearly shows that the SAR characteristics of this series against its three targets differ. Especially compounds with variations at site 2 but also site 3 are assigned a high discontinuity



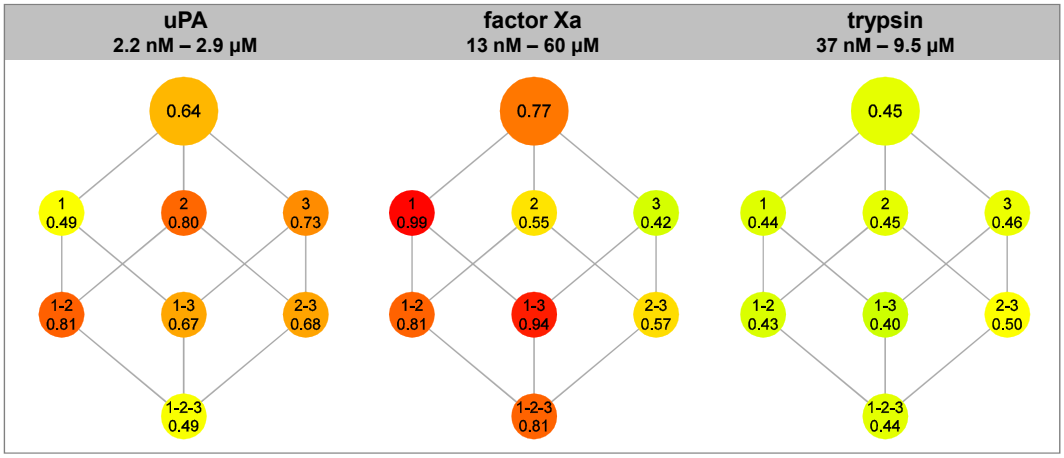
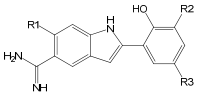
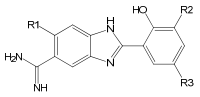


Figure 9.3-1: CAG representations for serine protease inhibitors (series 1) The maximum common subgraph of a series of 18 serine protease inhibitors is shown and the CAG representations of these analog sets for the three enzymes uPA, factor Xa, and trypsin. The CAG topology is conserved because the same analog series is represented. Comparison of these CAGs indicates the presence of different SAR characteristics. The figure is adapted from [125].

score for uPA, whereas site 1 produces discontinuity exclusively for factor Xa. For trypsin, SAR discontinuity and node heterogeneity is much reduced in comparison to the other two enzymes. The overall degree of discontinuity is also reflected by the root node that combines all compound pairs differing at up to three substitution sites and yields a considerably lower score for trypsin than for the other two proteases. For uPA, any substituent at site 2 improves compound potency, in some instances by more than two orders of magnitudes. The priority ordering for this site is featureless acyclic group > featureless aromatic group > aliphatic H-bond acceptor, as reported in Table 9.3-1. Interestingly, similar preferences are also observed for factor Xa and trypsin, but the observed potency increases are smaller. At site 3, the substitution having largest impact on potency for uPA is the introduction of a featureless acyclic group that decreases potency by more than one magnitude. This type of substituent also decreases the potency against trypsin, albeit to a much lesser extent. For factor Xa, substitution site 1 constitutes a prominent SAR hotspot. The introduction of a featureless acyclic group at this site consistently decreases potency by more than one or two magnitudes. Combinations of substitution sites with

Table 9.3-1: Preference orders

analog series	target	site 1	site 2	site 3
	uPA	Ac > (-) > Ac'	Ac > Ar > Al-A >> (-)	Ac-A > Ac-P|Ac-A'|(-) >> Ac|Ac-N
	factor Xa	(-) >> Ac	Ac > Ar > Al-A|(-)	Ac-P > Ac-A > Ac-A'|(-)|Ac|Ac-N
	trypsin	(-) > Ac	Ac|Ar > Al-A > (-)	Ac-A > Ac-P|Ac-A'|(-) > Ac|Ac-N
	uPA	Ac > (-)|Ac'	Al-A > Ac-A|Ar > Ac|(-)	inconclusive
	factor Xa	(-) > Ac	(-)|Ac-A|Ar|Ac > Al-A	inconclusive
	trypsin	(-) > Ac	Ac-A|Al-A|Ar > Ac > (-)	Ac-A|Ac > Ac-N|(-)|Ac-P

The table reports preference orders of pharmacophore features for individual substitution sites in two series of serine protease inhibitors for the three targets uPA, factor Xa, and trypsin. Abbreviations of pharmacophore feature classes are given according to Table 9.1-1. '>>' indicates that the average potency difference observed for substituents belonging to two pharmacophore classes is more than one order of magnitude, whereas '|' indicates that potency values observed as a consequence of feature class substitution are comparable. In cases where notable potency differences for substitutions within the same feature class occur, the class *P* is divided into *P* and *P'* to account for intra-class potency differences and placed twice in the preference order.

high discontinuity scores for uPA or factor Xa were also examined. For site combinations, SAR discontinuity was generally introduced by the same types of substitutions that caused discontinuity at the individual sites.

9.3.2.1 Key substitutions

Taken together, the analysis of protease inhibitor series 1 revealed the presence of differential SAR characteristics for the three target enzymes and indicated that high potency for uPA would be achieved by a combination of a featureless acyclic group at site 2 and an acyclic H-bond acceptor at site 3. However, a compound representing this combination of pharmacophore features was not present in the data set. Moreover, on the basis of preference ordering, an unoccupied substitution site 1 in combination with a featureless acyclic substituent at site 2 and a positively charged group at site 3 would represent the most preferred combination of substituents for achieving high potency against factor Xa. However, a compound having this substitution pattern was also not available

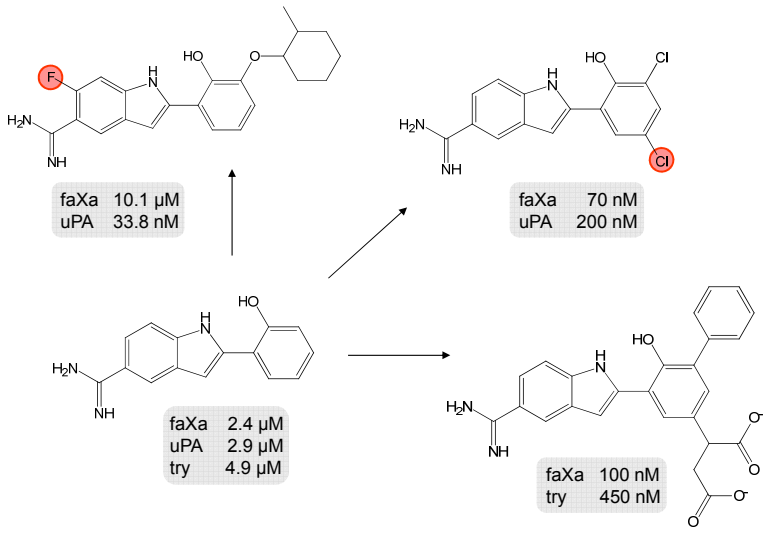


Figure 9.3-2: Verification of selectivity determinants identified for series 1 BindingDB analogs not included in the analyzed compound data set were used to evaluate predicted selectivity determinants. Here, three additional protease inhibitor analogs are displayed and their potency values are compared to those of the least substituted analog in this series. Selectivity-conferring substituents are encircled. faXa stands for factor Xa and try for trypsin. The figure is adapted from [125].

in the data set, indicating that the series might not have been explored to its full potential for factor Xa and uPA.

9.3.2.2 Target Selectivity Rules

Rules for the design of target-selective compounds were also derived. On the basis of our analysis, the following rules were formulated: the most important modification to increase selectivity for uPA or trypsin over factor Xa was the introduction of a featureless acyclic group at site 1. However, for trypsin, this type of modification also decreased potency, albeit to a much lesser extent than for factor Xa, thereby leading to a relative increase in selectivity over factor Xa at the expense of potency. The selectivity for uPA over the other two enzymes was further increased by the introduction of a featureless acyclic or aromatic group at site 2. By contrast, the selectivity for factor Xa or trypsin over uPA was increased by introducing a featureless acyclic group at site 3.

In order to assess these predictions, we searched BindingDB for analogs with matching substitution patterns. Three inhibitors were identified that were not contained in series 1 because they were annotated only with potency information for two of the three targets. These compounds confirmed the predictions and are shown in Figure 9.3-2 with their potency values. The compound at the upper left is selective for uPA, as we would expect considering the halogen substituent (featureless acyclic group) at site 1. The aliphatic H-bond acceptor at site 2 also

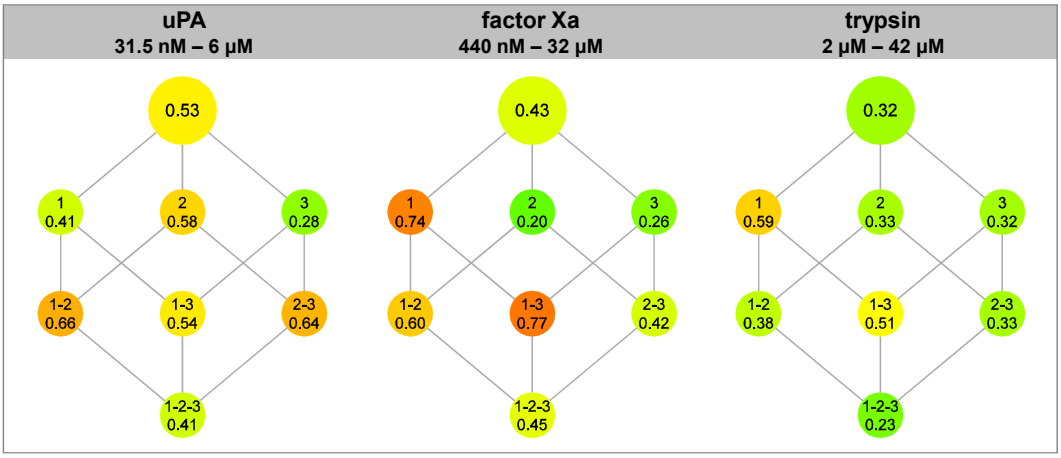


Figure 9.3-3: CAG representations for serine protease inhibitors (series 2) The maximum common subgraph of a series of 14 serine protease inhibitors is shown and the CAG representations of these analog sets for the three enzymes uPA, factor Xa, and trypsin. The figure is adapted from [125].

makes a contribution to selectivity. The compound at the upper right contains a featureless acyclic group at site 3, which favors selectivity for factor Xa over uPA. However, this effect is counter-balanced by the introduction of a featureless acyclic group at site 2 that increases potency for uPA more than for factor Xa. Hence, the net effect of simultaneously introducing these two substituents is only a small increase in selectivity for factor Xa over uPA. The third compound at the lower right contains an aromatic ring at site 2 and a negatively charged substituent at site 3. However, consistent with our prediction that substitutions at these sites would not be critical for distinguishing between factor Xa and trypsin, this compound is approximately ten-fold more potent for both targets than the unsubstituted molecule (lower left) but also non-selective.

9.3.3 Serine Protease Inhibitor Series 2

In Figure 9.3-3, CAG representations for serine protease inhibitor series 2 are shown. The MCS of series 2 is very similar to the one of series 1 in Figure 9.3-1; the indole moiety in series 1 is replaced with a benzimidazole in series 2. However, the CAG representations for these two series differ notably. CAG root nodes for series 2 are assigned much lower discontinuity scores than for series 1,

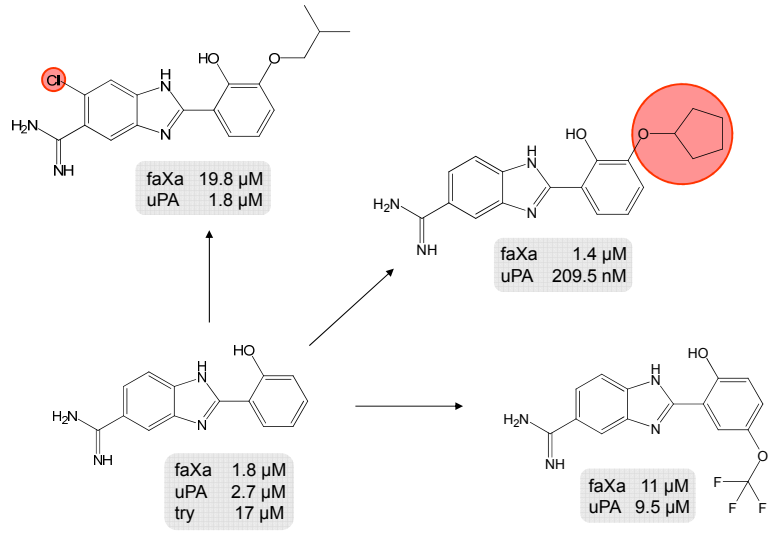


Figure 9.3-4: Verification of selectivity determinants identified for series 2
Three additional protease inhibitor analogs are displayed and their potency values are compared to those of the least substituted analog in this series. Selectivity-conferring substituents are circled. faXa stands for factor Xa and try for trypsin. The figure is adapted from [125].

consistent with the presence of a comparably narrow potency range in series 2. Similar to series 1, substitution site 2 in series 2 is most important for activity against uPA and is involved in all substitution site combinations with high discontinuity scores. However, Table 9.3-1 reporting the preference orders for series 1 and 2 shows that substitution site 2 in series 2 has a preference order that substantially differs from series 1. In this case, the introduction of an aliphatic H-bond acceptor is most preferred. Its introduction is also favorable for trypsin, albeit to a lesser extent, but unfavorable for factor Xa. By contrast, substitution site 1 represents an SAR hotspot in both series 1 and 2 for factor Xa. At this site, potency losses are caused by the introduction of a featureless acyclic group. However, for the benzimidazole derivatives, induced potency changes are much smaller than for the indole derivatives. The introduction of a featureless acyclic group at site 1 is unfavorable for trypsin but does not significantly affect the potency of uPA.

In contrast to series 1, the data for series 2 do not provide insights as to how one would improve the selectivity for trypsin or factor Xa over uPA. Instead, two possibilities emerge to increase selectivity for uPA, i.e., by introducing a) a featureless acyclic group at site 1 and b) an aliphatic H-bond acceptor at site 2. Again, we searched for analogs that were not included in our analysis and found three inhibitors that were only annotated with potency information for human factor Xa and uPA, shown in Figure 9.3-4. As predicted, compounds with a featureless acyclic group at site 1 or an aliphatic H-bond acceptor at site 2 show a higher selectivity for uPA than the unsubstituted molecule. By

contrast, an additional analog with a substituent at site 3 (bottom right) shows no change in selectivity.

9.4 Conclusions

Multi-target SARs are often complex and difficult to analyze. However, decoding of such SARs is critically important for identifying determinants of compound selectivity. In the past, different SARs have in some instances been compared using QSAR techniques to ultimately make predictions of compounds active against a specific target. Such predictions made using different QSAR models could then be compared in attempts to differentiate SAR characteristics [127,128]. Only recently, first attempts have been made to design molecular graph representations that identify selectivity cliffs resulting from potency differences of compounds against pairs of targets [129] or multi-target activity cliffs of different magnitude [102]. Nevertheless, it is fair to say that a more systematic analysis of mtSARs using computational means is currently still in its infancy.

We have introduced a computational approach to study mtSARs that relies on a data structure generated by consistent R-group decomposition, assessment of analog similarity on the basis of pharmacophore feature edit distances, and correlation with potency data. An intuitive and easy graphical access to these data is provided by CAGs that hierarchically organize analog series according to substitution sites. Comparison of these graph representations for multiple targets immediately identifies individual SAR hotspots and reveals target-dependent differences in SAR characteristics of analog series. Then, compound pairs corresponding to SAR hotspots are analyzed and preference orders for pharmacophore feature substitutions are derived, which is a key aspect of the approach presented herein. From these preference orders, simple and intuitive rules for the design of target-selective compounds can often be deduced, as demonstrated in the reported applications. Selectivity predictions made on the basis of mtSAR analysis have been confirmed through searching for relevant analogs not included in our analysis.

Source Information

Sections of the text in this chapter have been taken from [125].

Chapter 10

SAR Transfer

The chemical lead optimization process involves the simultaneous improvement of multiple properties, such as compound potency, selectivity, oral availability, or metabolic stability. There are many possible complications along the path to developing clinical candidates and it is not uncommon that a compound series displaying an otherwise promising SAR hits a roadblock, perhaps due to metabolic liabilities or unwanted side effects, which then prevents its further development. In these situations, it would be highly desirable to reutilize available SAR information and apply learned SAR rules to an alternative chemotype that shows similar SAR characteristics. Accordingly, one would search for alternative molecular core structures (scaffolds) where corresponding chemical substitutions yield comparable SAR trends (consistent with a conserved mechanism of action) but circumvent liabilities associated with the original chemotype. However, an SAR transfer represents a rather challenging task. Importantly, it goes far beyond the identification of alternative scaffolds displaying a specific biological activity [116] because it requires the identification or generation of corresponding analog series with equivalent SAR characteristics. As shown in Chapter 9 by the comparison of substituent preference orders for the two serine protease inhibitor series, even for very similar scaffolds, different SAR trends can be observed. Hence, SAR transfer events are not expected to be abundant and easy to identify.

In addition to replacing one chemical series by another, SAR transfer also has other facets. For example, if parallel compound series with varying degrees of chemical exploration (e.g., different numbers of analogs, different potency levels) would be available, one might learn more about SAR progression than on the basis of a single series. In addition, it might be possible to suggest potent analogs for one series based on another, whose other lead-relevant properties could then be compared.

Figure 10.0-1 illustrates an SAR transfer scenario that emphasizes both the alternative series and learning aspects. Compound series 1 and 2 in Figure 10.0-1 contain distinct core structures and consist of analogs with pairwise

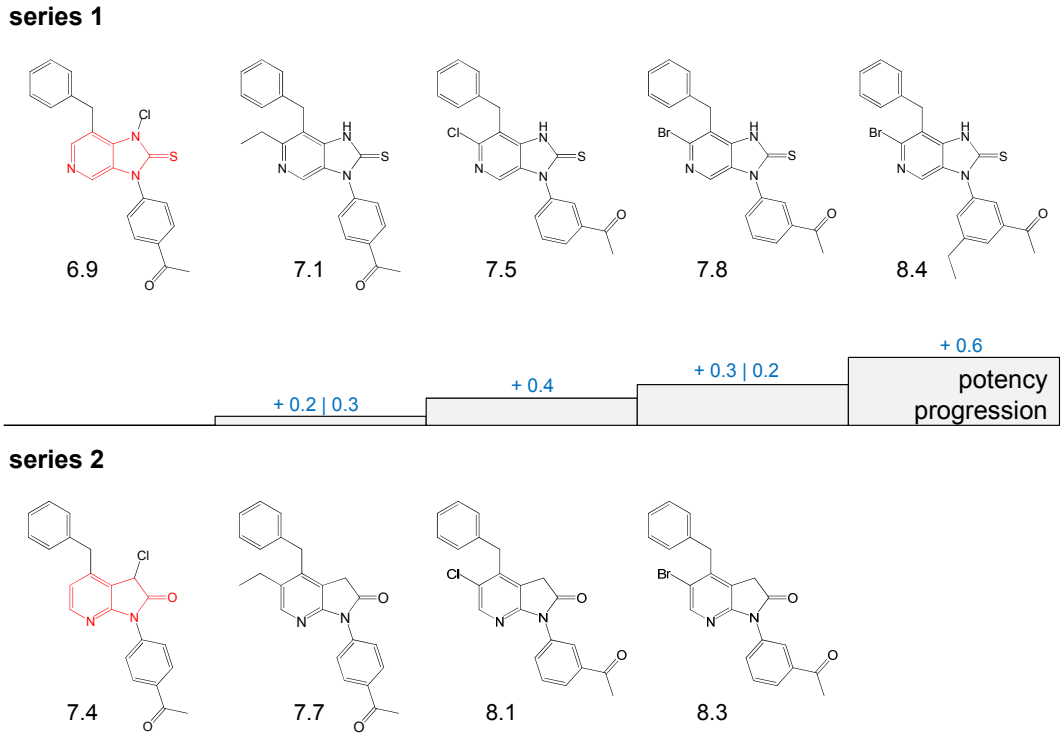


Figure 10.0-1: SAR transfer series The principal idea of SAR transfer is illustrated. Two structurally related model analog series are shown. The ring systems that distinguish the two corresponding scaffolds are highlighted in red in a compound of each series. Compounds in both series that carry the same substituents and only differ in the exchanged rings are aligned vertically, i.e., they form pairs of corresponding analogs. The alignment organizes compounds (labeled with their pK_i values) in the order of increasing potency from left to right. For potency-based ordering, one series serves as a reference. In this case, the two ordered series display a steady potency progression and potency differences (shown in blue) between analogs are comparable for both series, although the absolute potency values of corresponding compounds differ. Hence, the two analog series represent a prototypic SAR transfer.

corresponding R-group patterns that show a comparable increase in potency. Thus, the SARs of series 1 and 2 are essentially interchangeable and the two series represent a prototypic SAR transfer model. Moreover, series 1 contains a potent analog that has no counterpart in series 2 and hence the corresponding analog might be suggested for synthesis. Although chemotype replacement might be considered the primary task of SAR transfer, the comparative learning aspect is also attractive for SAR exploration and exploitation.

Although SAR transfer is of high practical relevance in medicinal chemistry and a frequently discussed topic, comprehensive literature searches did not reveal computational methods available to aid in this process. Therefore, as a first step in this direction, we designed a data mining method to identify chemical series with SAR transfer potential [130]. This chapter presents the methodological concept of our computational approach in section 10.1 and

reports exemplary applications in section 10.2. Applications comprise (i) the identification of chemical series with SAR transfer potential if one analog series is available as a starting point and (ii) the detection of all SAR transfer events that occur in compound databases. A summary of major findings and an outlook for future work is given in section 10.3.

10.1 Methodology

Because the prediction of an "SAR mimic" from first principles would be very difficult, we have approached SAR transfer from a data analysis perspective. We started from a typical, practically relevant SAR transfer situation where a series of analogs with multiple structural modifications and increasing potency was available as a starting point and aimed to identify SAR transfer series through compound database searching. For this purpose, we designed an approach consisting of three different stages that were implemented in Java using the OpenEye chemistry toolkit and are introduced in the following.

10.1.1 Identification of Alternative Scaffolds

Initially, from a known series of active analogs, in the following also termed "template series", the common scaffold is extracted. Single atoms forming exocyclic double bonds to ring atoms (mostly carbonyl oxygens) are considered part of a scaffold and not removed. Then, a database search is carried out to find related but chemically distinct scaffolds (Figure 10.1-1). Candidate scaffolds are permitted to differ from the original scaffold by the replacement of one contiguous ring system (i.e., consisting of one or more rings). There are no restrictions on ring sizes or composition. Hence, depending on the exchanged ring systems, corresponding scaffolds might display different degrees of (dis)similarity. Scaffolds with changes in one ring structure are identified using a variant of the MMP search algorithm of Hussain and Rea presented in Chapter 7. In this adaptation, all contiguous ring systems in a scaffold are separately removed by deleting all connecting bonds between the ring system and the remaining structure. Connectivity information for the resulting fragments is retained by marking the attachment points. For two scaffolds that only differ by a single ring, removal of the exchanged ring structures leads to identical key fragments and this scaffold pair can be identified from an index table, as described in Chapter 7. If an alternative scaffold is identified for the template scaffold, all compounds represented by this scaffold are retrieved (Figure 10.1-1). These analogs constitute a "target series".

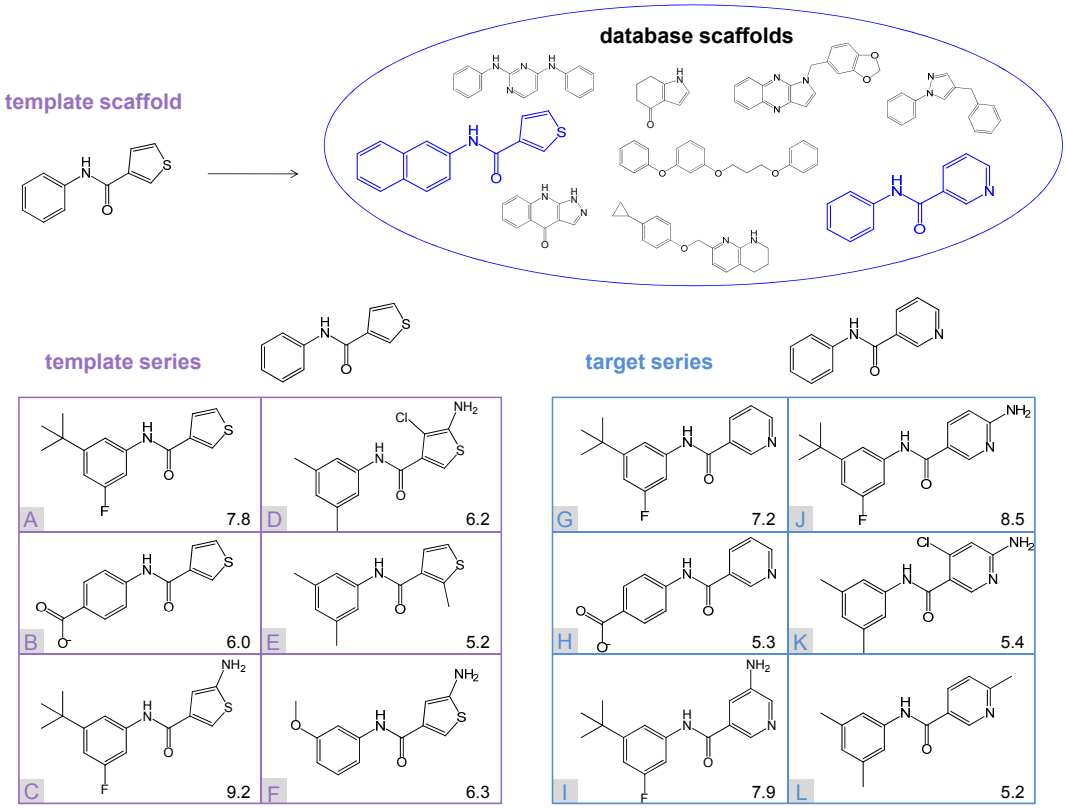


Figure 10.1-1: Database search for alternative scaffolds A scaffold representing a template series is used to search a database for scaffolds that differ by the replacement of exactly one ring system. Two target scaffolds meeting this structural criterion are highlighted in blue. For one of these scaffolds, all analogs are assembled representing a target series. In this example, both the template and the target series contain six analogs. These compounds are consecutively names A-L. The pK_i value of each analog is reported. The figure is adapted from [130].

10.1.2 Identification of Corresponding Analogs

For analogs in the template series, matching compounds in the target series are identified. For this purpose, all compounds represented by the two scaffolds are fragmented by removing the distinguishing ring and attached R-groups. The removed fragment is then further decomposed into the invariant ring structure and the R-groups, which are marked with the numeric identifier of the ring atom to which they were attached. Hence, each molecule of a series can be unambiguously represented by the combination of the residual substructure after ring removal and the set of R-groups. Corresponding compounds of two analog series are required to have identical residual substructures and R-groups. These R-groups are used to define corresponding substitution sites in the exchanged ring systems, as shown in Figure 10.1-2. Especially if R-groups are found for exchanged rings of different size, substitution site correspondences are not clear

and, therefore, all possible pairwise mappings of positions in exchanged rings are explored and alternative mappings are compared, as illustrated in Figure 10.1-2. In order to identify series/mappings showing SAR transfer, all compound pairs complying with the same R-group mapping are assembled into two "matching series".

10.1.3 Potency-Based Compound Ordering and SAR Transfer Score

For all compounds represented by a given scaffold within a matching series, the mean potency is determined and potency values are rescaled with respect to the mean, i.e., zero-centered "relative potency" values for all analogs are calculated by subtracting the mean from actual compound potency values (in pK_i units). In order to meet SAR transfer criteria, template and target series can differ in their absolute potency values but potency differences between ordered analogs in each series should be similar so that corresponding structural modifications entail similar potency effects. The comparison of relative potency values for paired analogs is a straightforward way to account for potency progressions within the two series. If compounds forming a pair are always assigned the same relative potency, substitutions consistently cause the same potency changes, leading to complete SAR transfer for the two matching series.

10.1.3.1 Color-Coded Analog Pair Alignments

To facilitate the comparison of relative potency values for a target and template series and for a visual inspection of SAR transfer potential, color-coded analog pair alignments are utilized, as shown in Figure 10.1-3. From top to bottom, corresponding compound pairs are ranked in the order of decreasing potency of the target series and displayed as colored nodes connected by an edge. A uniform color code is applied to account for the potency difference of a compound from the mean. The color code represents a continuous spectrum from green (over yellow) to red to account for potency differences from the mean within the range from -1.5 (green) to 1.5 (red) pK_i units. Potency differences falling below or above this range are represented in green and red, respectively. Color matches along the ranking indicate whether corresponding replacements have similar effects on potency progression within the analog series. SAR transfer is characterized by paired nodes that are consistently assigned the same color along the alignment. By contrast, if corresponding analogs have differently colored nodes, substitutions have different relative potency effects. If this occurs for only a few pairs of analogs within larger series, SAR transfer is locally incomplete. If no corresponding color patterns are observed, the SARs of the two series are distinct and not transferable.

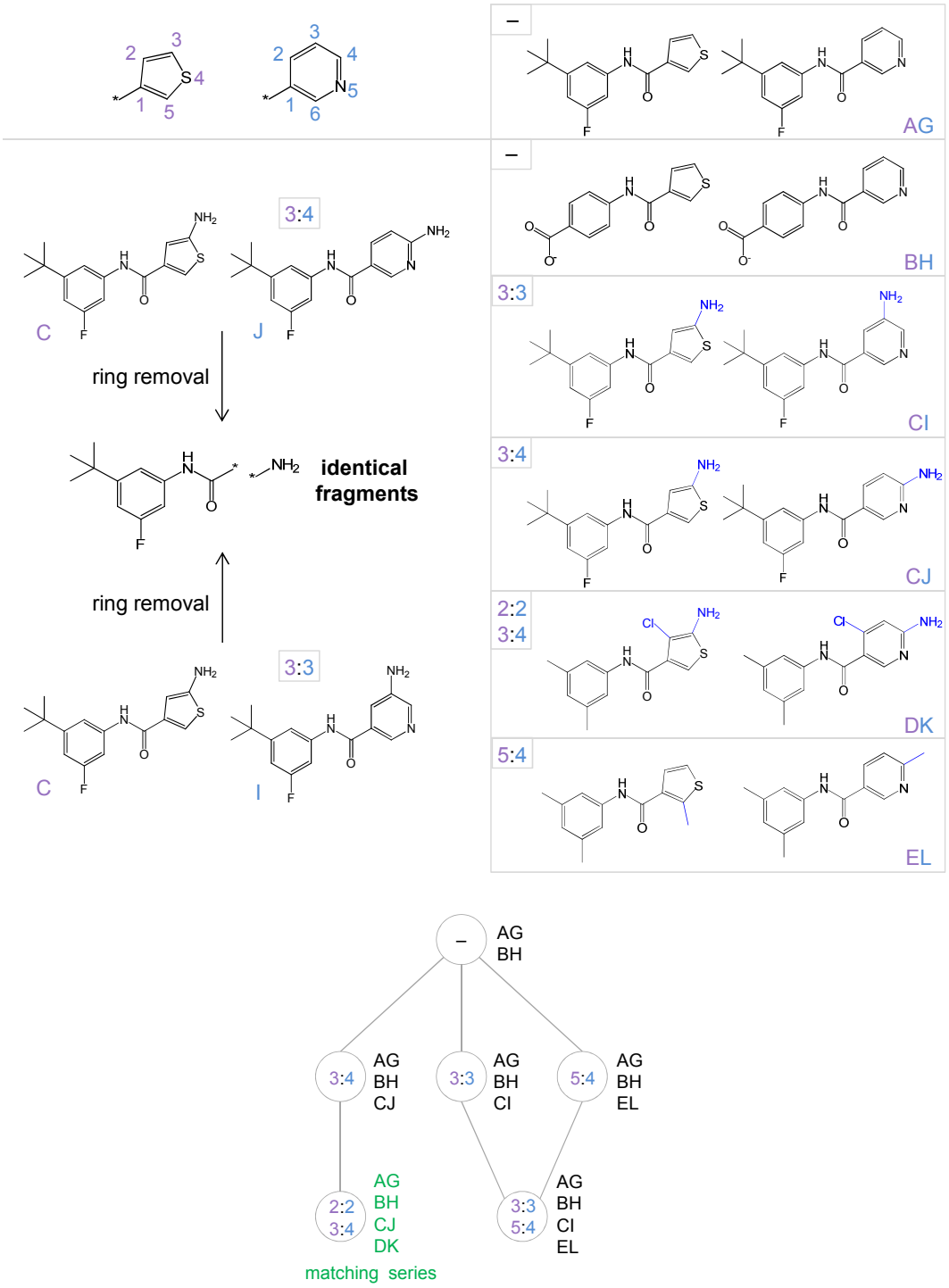


Figure 10.1-2: Pairwise analog assembly Top left: Corresponding compounds in the template and target series are identified by removing the distinguishing ring system. Identical R-groups at the exchanged ring structures determine substitution site correspondences (e.g., "3:3") for each pair of corresponding compounds. Top right: All possible compound pairs retrieved for the template and target series from Figure 10.1-1 are reported and mappings of ring substituent positions are provided. The hyphen (-) indicates the absence of substituents at the exchanged rings. Bottom: All different mappings of ring positions are systematically explored for SAR transfer. Therefore, all compound pairs that comply with a mapping are assembled into so-called matching series (highlighted in green for the mapping "2:2,3:4").

10.1.3.2 Score Calculation

For a quantitative assessment of SAR transfer potential, absolute differences between rescaled potencies of all compound pairs in matching series are calculated, as demonstrated in Figure 10.1-3. The maximal difference between two compounds in a pair is utilized as an SAR transfer score because it describes the largest observed deviation of matching series from an ideal SAR transfer scenario. Accordingly, a score of zero corresponds to perfect SAR transfer with identical potency progressions between ordered pairs of analogs in matching series, whereas high scores indicate a substantial discrepancy in potency progression.

10.2 Applications

We first tested the methodology by searching for alternative analog series using a known series as template and then aimed at a systematic detection of all SAR transfer events in BindingDB.

10.2.1 SAR Transfer Detection for Selected Analog Series

Compounds annotated with K_i values against human targets were extracted from BindingDB. For different activity classes, compound series consisting of multiple analogs were selected as template series and the remaining active compounds were searched for potential target series that contained corresponding compounds with comparable potency progression. In the following, four examples for successful SAR transfer detection for the targets dopamine D3 receptor, thrombin, factor Xa, and carbonic anhydrase I are reported.

10.2.1.1 Dopamine D3 Receptor Antagonists

Figure 10.2-1 reports search results for a template series consisting of eight analogs of dopamine D3 receptor antagonists. Three target series with different SAR transfer potential were detected.

In Figure 10.2-1a, the template (left) and target (right) series are related by the exchange of a benzofuran versus an indole moiety. The target series also consists of eight analogs. In the template and target series, five and six analogs, respectively, are unsubstituted at the exchanged ring and form four corresponding pairs in the alignment, as shown in Figure 10.2-1a. Among these pairs, a clear SAR transfer is observed, indicated by very similar node colors for compounds forming each pair.

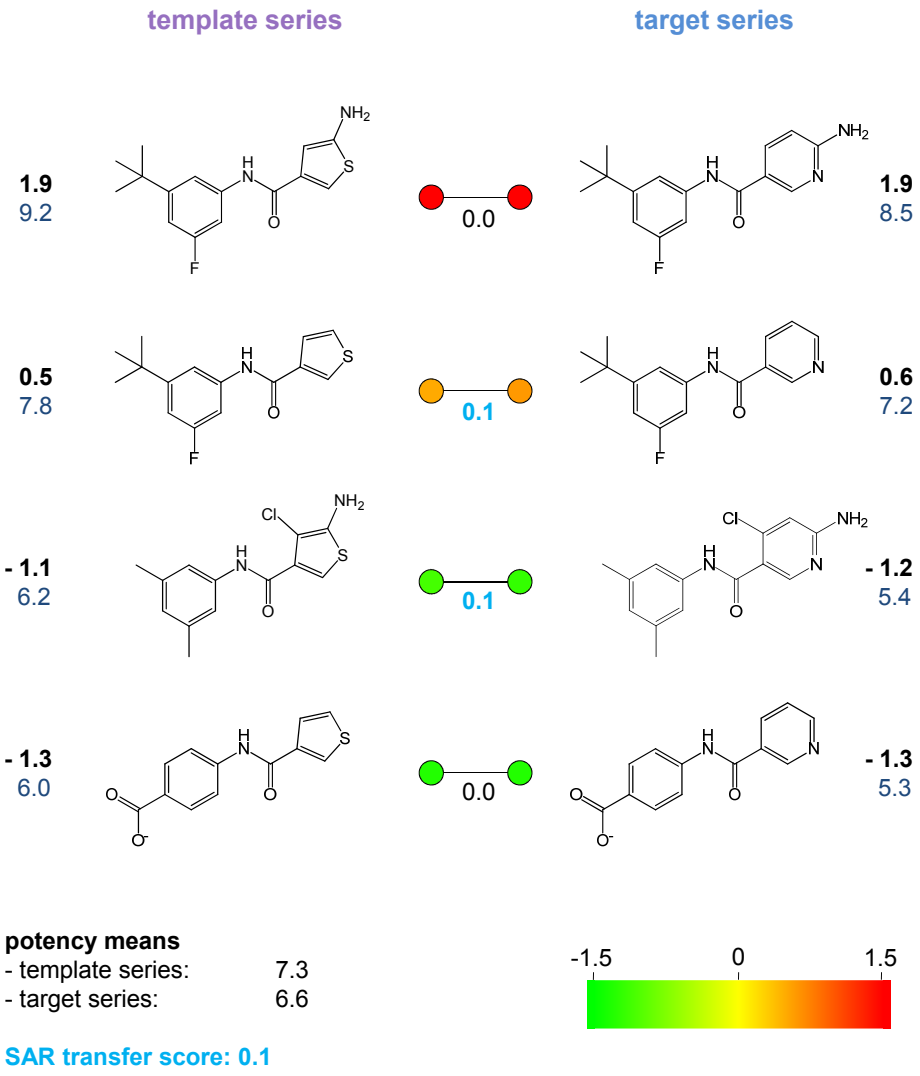


Figure 10.1-3: Analog pair alignment For the mapping "2:2,3:4" (compare Figure 10.1-2), all pairs of analogs are assembled into two matching series. Corresponding compounds of the target (right) and template (left) series are vertically aligned and represented by nodes connected by an edge. From top to bottom, analog pairs are ranked in the order of decreasing potency of the target series. The mean potency for the four aligned compounds from each series is determined, and for each compound, the potency difference from the mean (relative potency) is calculated. The compounds are annotated with their relative (bold) and absolute potency values. Nodes are color-coded by relative potency using a continuous spectrum from green (via yellow) to red. Green indicates lowest and red highest relative potency in a series. Edges are labeled with differences between rescaled potencies of corresponding compounds (connected nodes), and maximal differences representing the SAR transfer score are highlighted in light blue.

In the second target series shown in Figure 10.2-1b, which consists of three analogs, the benzofuran moiety is replaced by a benzothiophene ring. Here, three corresponding pairs are also formed for unsubstituted exchanged rings. In

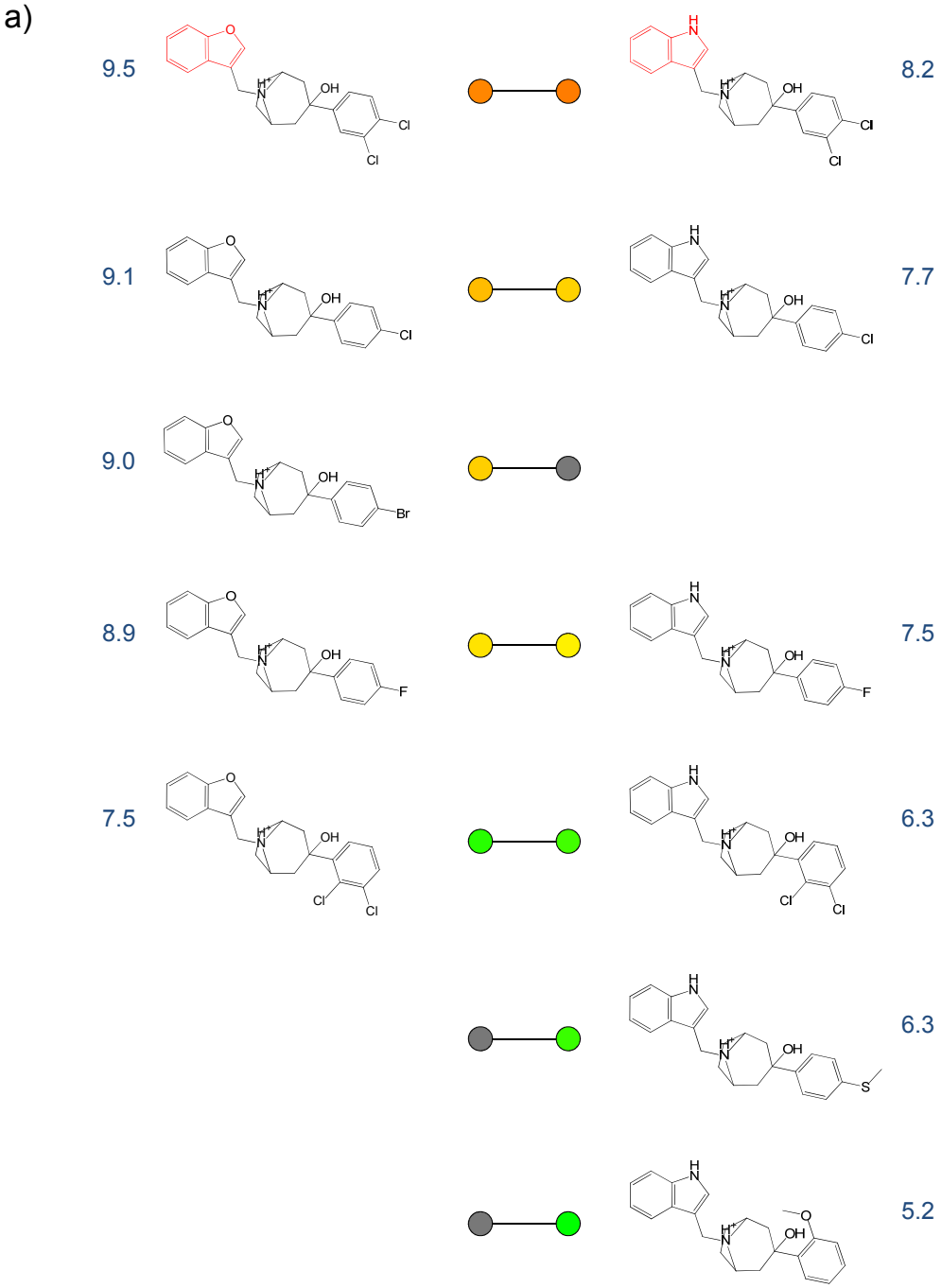


Figure 10.2-1: Template and target series for dopamine D3 receptor antagonists

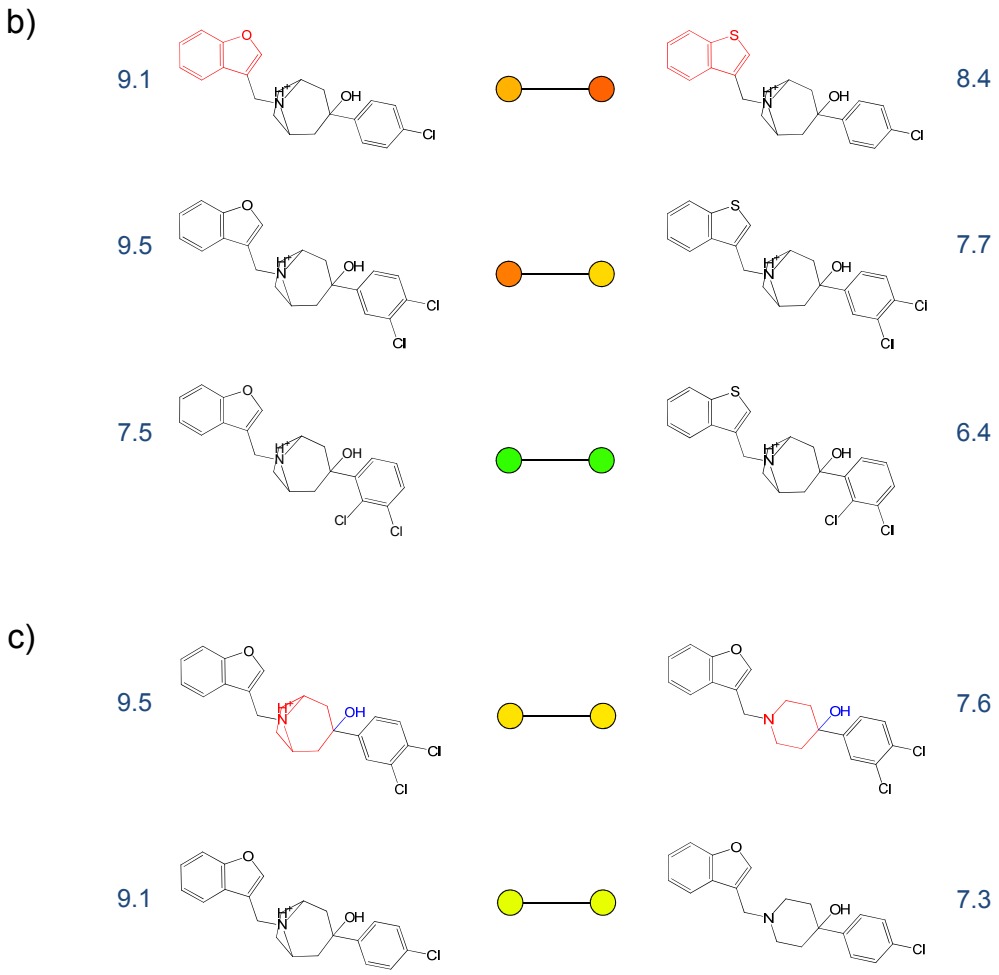


Figure 10.2-1: Template and target series for dopamine D3 receptor antagonists (continued) (a-c) Alignments with three different target series are shown for a template series consisting of dopamine D3 receptor antagonists. In each case, compounds belonging to the template series are shown on the left and the target series is shown on the right. Corresponding compound pairs are ranked in the potency order of the target series. Node colors are determined on the basis of centered potency differences according to Figure 10.1-3. Gray nodes indicate "missing" analogs. For each compound, its pK_i value is reported. In the compound pair at the top of each alignment, the exchanged ring structures are colored red. R-groups at the highlighted exchanged rings are colored blue. The figure is adapted from [130].

contrast to the first target series, no SAR transfer is observed for the second target series, due to inconsistent potency progression. Whereas a 3,4-dichloro-substituted phenyl ring leads to the most potent analog in the template series, a 4-chloro-substituted phenyl ring is preferred in the target series. Hence, although the core structures of the template and the two target series are almost identical and differ only by a single heteroatom, only two of the three series show a similar SAR. This observation is also reflected by calculated SAR trans-

fer scores that amounted to 0.15 and 0.54 for the alignments of the template series with target series 1 and 2, respectively.

Figure 10.2-1c shows a rudimentary yet structurally qualifying match with only two pairs of analogs for the alignment of the template series with the third target series. In this case, the central saturated ring moiety is exchanged and the two rings carry a hydroxyl substituent. Although node colors are similar for the two aligned compound pairs, the alignment conveys comparably little SAR transfer information and it would certainly be inappropriate to deduce SAR transfer based on similar potency effects for only one structural modification. At least, the alignment does not exclude the possibility of an SAR transfer for the two series.

10.2.1.2 Thrombin Inhibitors

Figure 10.2-2 shows a thrombin inhibitor template series consisting of eight analogs and the single target series we identified. The target series contains five analogs, all of which form pairs with template compounds. The template and target series are related by the exchange of a tetrazole versus a triazole ring. The alignment reveals clear SAR transfer character. In addition, the two most potent template compounds have no counterparts in the target series. In light of the observed potency progression, the corresponding analogs would be expected to have higher potency than the currently most potent triazole-containing compound, hence providing a prototypic example for alignment-based compound suggestions and the comparative learning aspect associated with SAR transfer analysis.

10.2.1.3 Factor Xa Inhibitors

We also searched for target series using a large template series of 88 factor Xa inhibitors. In this case, a small target series containing only six analogs was detected. These series were related by the exchange of phenyl and pyridyl rings. Four pairs of corresponding analogs could be aligned (Figure 10.2-3). The alignment shows that the relative potency of corresponding compounds is similar and SAR transfer is observed along all compound pairs. In the most potent paired analogs, the variable core rings contain a fluorine-methyl-fluorine substituent pattern and the two phenyl rings are meta-substituted with carbamimidoyl groups. A comparison to the most weakly active compounds reveals that both series benefit from the exchange of an aminomethyl by a carbamimidoyl group at the meta position of a phenyl ring. Hence, when the substitution pattern of the central ring is kept constant, structural changes at the phenyl ring lead to similar potency progression in both series. Remaining highly potent compounds

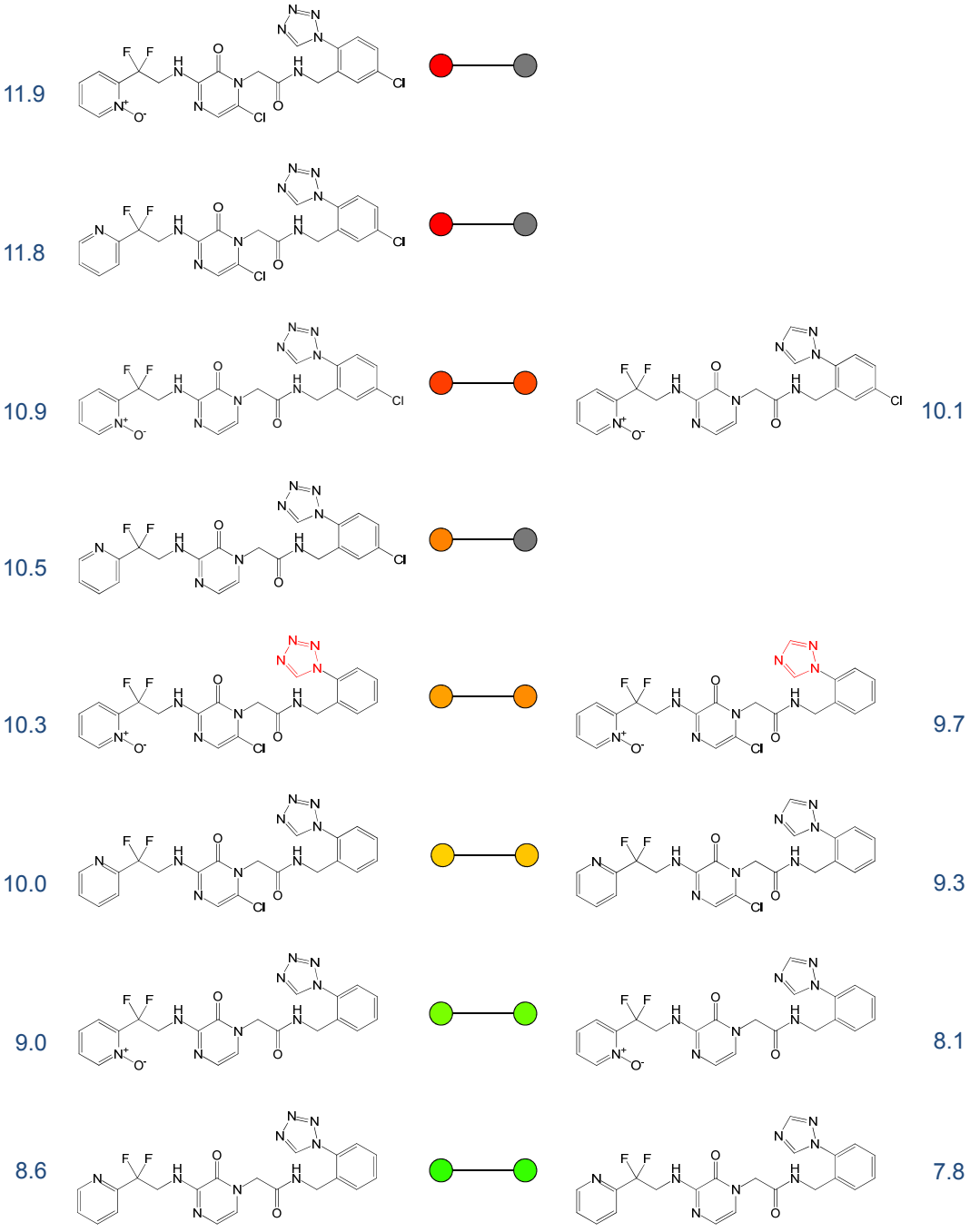


Figure 10.2-2: Template and target series for thrombin inhibitors Shown is a template series consisting of thrombin inhibitor analogs (left) and a single target series (right). The representation is according to Figure 10.2-1. The figure is adapted from [130].

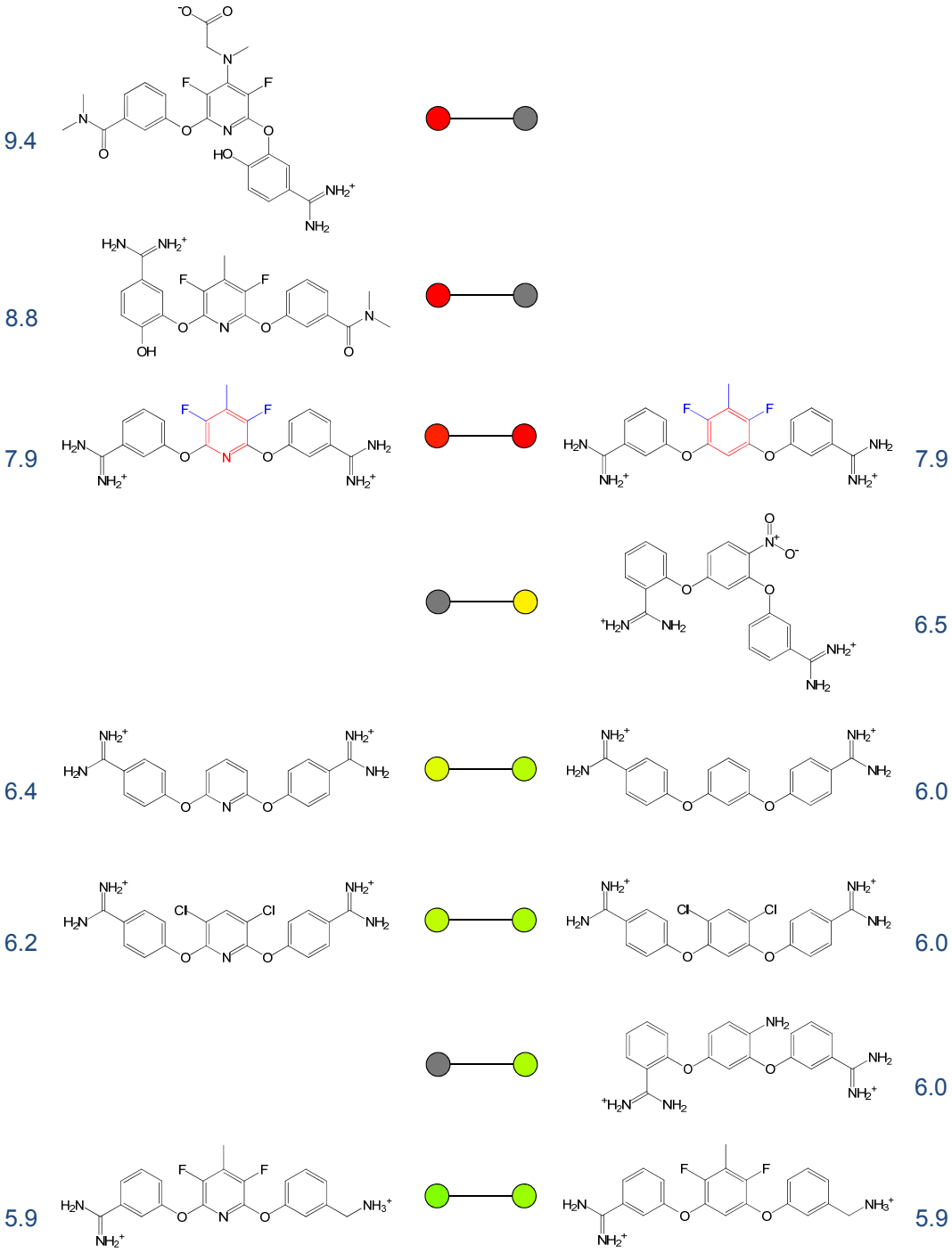


Figure 10.2-3: Template and target series for factor Xa inhibitors Shown is a template series consisting of factor Xa inhibitor analogs (left) and a single target series (right). The representation is according to Figure 10.2-1. The figure is adapted from [130].

of the template series (a representative example is shown in Figure 10.2-3) also contain this fluorine-methyl-fluorine ring substituent pattern. Thus, on the basis of these findings, additional compound design suggestion could be made for the target series.

10.2.1.4 Carbonic Anhydrase I Inhibitors

In Figure 10.2-4, search results are shown for another large template series of sulfonamide-containing carbonic anhydrase I inhibitors (51 analogs). Here, a target series with 39 analogs was identified. In this case, a phenyl and a thiadiazole ring are exchanged that carry the critical sulfonamide substituent. For matching compounds in the target series, this substituent can be transferred to the ortho, meta, or para position of the exchanged phenyl ring and we investigated all three possibilities for SAR transfer. The best alignment consisting of eight analog pairs was obtained for transferring sulfonamide substituents to the meta position of the phenyl ring, as illustrated in Figure 10.2-4. In this alignment, SAR transfer is locally incomplete because the relative potencies of a weakly potent analog pair differ for the two series, as indicated by the color code. For the remaining analog pairs, potency progression is comparable. Thus, the alignment represents an example of partial SAR transfer. Importantly, SAR transfer is observed for the more potent compounds and several potent analogs in both series have no counterparts in the alignment (a few representative examples are shown in Figure 10.2-4), thus providing another opportunity for comparative learning from two series and the design of other potent analogs.

10.2.2 Systematic SAR Transfer Detection

We then systematically searched for possible SAR transfer events in BindingDB to address the question how frequently SAR transfer is observed for chemical series that differ by the replacement of a single ring system.

10.2.2.1 Compound Data Sets

All ring-containing compounds with available K_i values for human targets were extracted from BindingDB. For all molecules with multiple potency measurements against the same target, the arithmetic mean was calculated to yield a final potency value unless reported K_i values spanned a potency range of more than one order of magnitude. In this case, the target activity was excluded from the analysis. A total of 53 760 different qualifying compounds were selected and organized into 708 target-specific compound data sets. In each ligand set, compounds were grouped by scaffolds containing at least two ring systems. Compounds represented by single ring scaffolds were discarded.

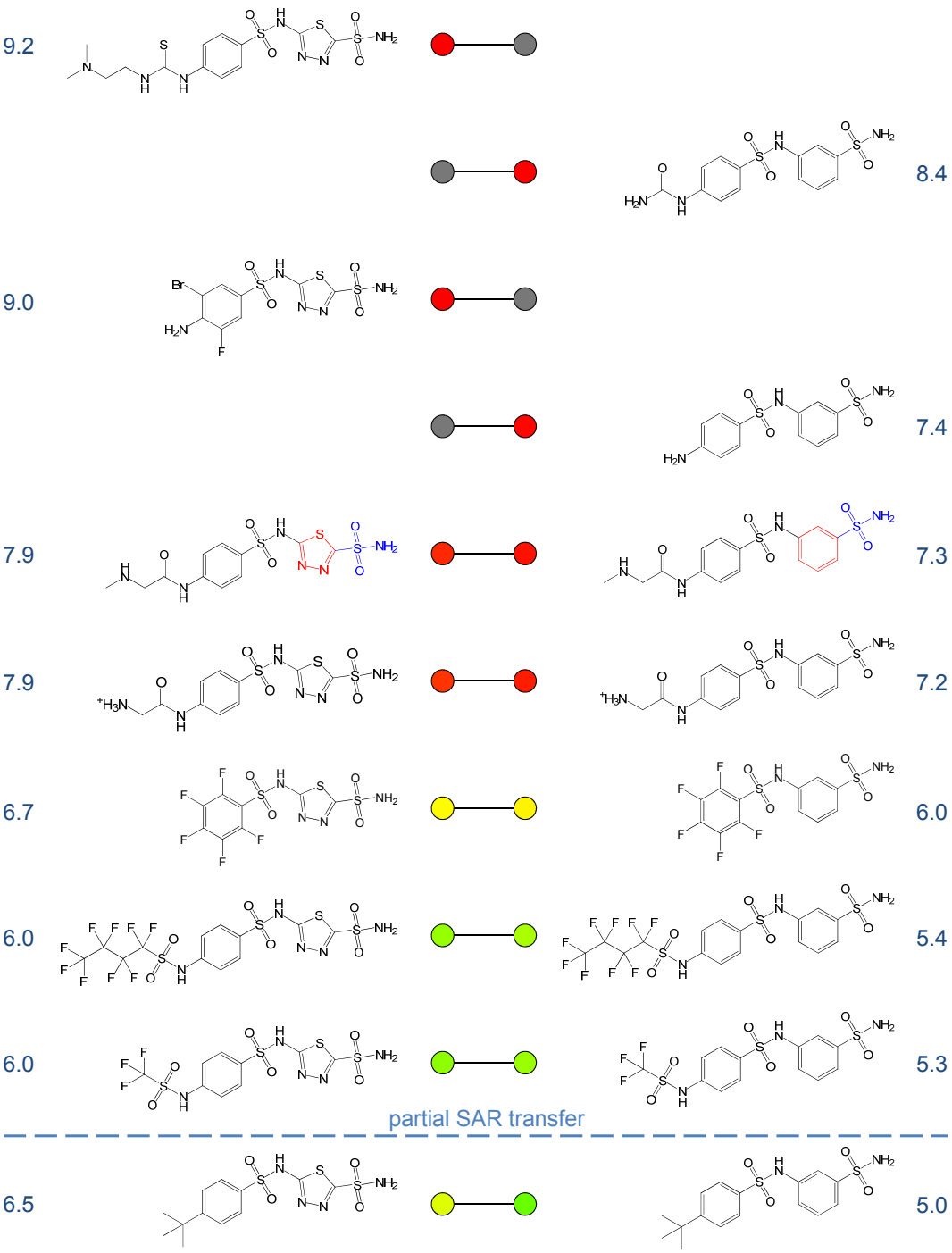


Figure 10.2-4: Template and target series for carbonic anhydrase I inhibitors Shown is a template series consisting of carbonic anhydrase I inhibitor analogs (left) and a single target series (right). For clarity, only a part of the global series alignment is displayed. The representation is according to Figure 10.2-1. Because relative potencies differ for the two paired compounds at the bottom, the alignment represents a partial SAR transfer. The figure is adapted from [130].

10.2.2.2 SAR Transfer Score Distribution

For the general assessment of SAR transfer potential, we searched for all scaffold pairs meeting the single ring exchange criterion in the compound data sets extracted from BindingDB. All possible matching series in a ligand set were identified (in this case, template and target series are not distinguished). Then, all possible mappings of R-group positions in matching series were separately considered for score calculation, i.e., for each possible analog alignment the maximal difference of rescaled potency values for corresponding compounds was calculated. In order to focus the search on series that displayed a comparable potency progression over multiple compounds, we only retained alignments containing at least three compound pairs. Furthermore, analogs in at least one of the matching series must span a potency range of more than one order of magnitude.

On the basis of these criteria, 306 matching scaffold pairs were identified in 93 target sets. Because some scaffold pairs occurred in multiple sets, a total of 405 different scaffold pair-target combinations were obtained. For the general assessment of SAR transfer potential, the matching series yielding the lowest SAR transfer score for a scaffold pair-target combination were selected among alternative mappings.

We found that matching series consisted on average of 4.15 corresponding compound pairs. Figure 10.2-5 shows the score distribution observed for the 405 compound pair alignments. Scores between 0.2 and 1.0 were most frequently obtained, with a mean score of 0.69. However, the right tail of the distribution indicates that very high scores also occurred in some instances, i.e., for some matching series the exchange of a single ring structure led to dramatic SAR discrepancies. In order to investigate whether the number of transferred ring positions had an influence on the SAR transfer potential of a matching scaffold pair, statistics were separately generated for different numbers of mapped R-group sites. Table 10.2-1 reveals that matching series without R-groups at the exchanged rings were most frequently observed and that these series displayed the tendency to yield low scores, consistent with high SAR transfer potential. Furthermore, we observed the trend that with increasing numbers of R-groups at exchanged rings the scores also increased.

10.2.2.3 SAR Transfer Series

Visual inspection of many analog alignments suggested that a score lower than or equal to 0.3 typically represented matching series showing SAR transfer. For example, for the series shown in Figures 10.2-1a, 10.2-2, and 10.2-3 scores of 0.15, 0.12, and 0.24 were obtained, respectively. Hence, we applied a score threshold of 0.30 to search for SAR transfer series. A total of 61 SAR transfer series were identified in BindingDB that occurred in 39 different target sets and

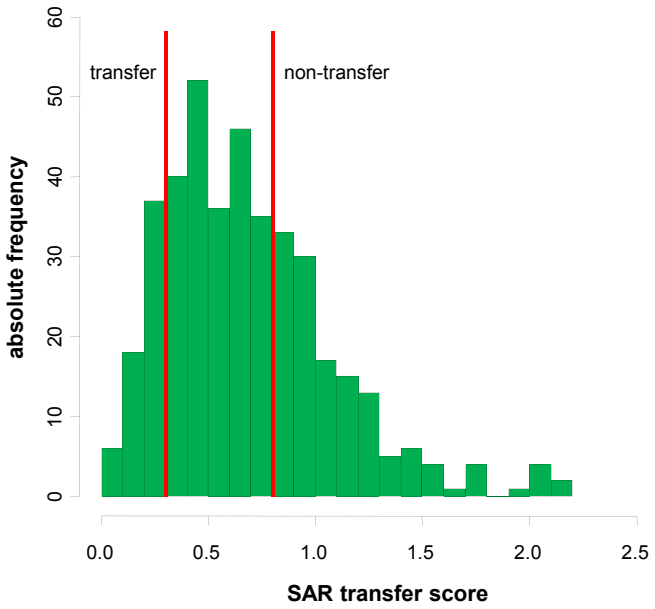


Figure 10.2-5: SAR transfer score distribution The distribution of SAR transfer scores is reported for 405 compound pair alignments extracted from BindingDB. Thresholds for SAR transfer and non-transfer series are displayed. The figure is adapted from [130].

contained 59 different scaffold pairs. As shown in Table 10.2-2, 70% of these SAR transfer series did not carry R-groups at the exchanged ring systems.

10.2.2.4 Similar Chemical Series With Distinct SARs

Finally, we also identified matching series without SAR transfer potential. Therefore, a lower score cutoff of 0.80 was applied. In this case, 135 matching series were identified where the exchange of a ring system resulted in very different potency progression. Hence, matching series with limited or no SAR transfer potential were more frequently found than SAR transfer series. Interestingly, about 65% of matching series with scores larger than or equal to 0.80 comprised compounds with R-groups at the exchanged ring systems (Table 10.2-2). Perhaps surprisingly, the number of matched compound pairs did not correlate with a decreasing SAR transfer potential of two series.

10.3 Conclusions

This chapter introduced a computational approach to search for SAR transfer series, a task of considerable practical relevance in medicinal chemistry, and systematically analyze SAR transfer events in databases. Despite the inherent complexity of the problem, the underlying computational methodology is straightforward and much emphasis was put on the chemically intuitive nature

Table 10.2-1: Global assessment of SAR transfer potential

#R-sites	#occurence	score	#pairs
0	186	0.60	4.04
1	134	0.75	4.11
2	57	0.64	4.37
3	27	1.01	4.63
4	1	1.23	4.00

The 405 compound alignments extracted from BindingDB are grouped by the number of ring positions ("R-sites") mapped for corresponding compounds from different analog series. For each group, its absolute frequency of occurrence ("#occurence"), average SAR transfer score ("score"), and average number of aligned compound pairs ("#pairs") are reported.

Table 10.2-2: SAR transfer and non-transfer series

#R-sites	#occurence	
	SAR transfer series	non-transfer series
0	43	49
1	11	58
2	6	12
3	1	15
4	0	1

The 61 SAR transfer series and 135 non-transfer series are grouped by the number of ring positions ("R-sites") mapped for corresponding compounds from different analog series. For each group, its absolute frequency of occurrence ("#occurence") is reported.

of the approach. Interpretability of the results was ensured by the introduction of analog pair alignments that are simple to analyze and provide a basis for comparative SAR analysis. Furthermore, if SAR transfer is observed and individual compounds exist in one analog series that do not have counterparts with corresponding R-groups in the other, new analogs can be readily suggested, i.e., analog pair alignment information can be translated into compound design. Several representative examples have been discussed and a statistical analysis of SAR transfer events has been presented, which also included the identification of structurally corresponding analog series with differing potency progression. Our systematic analysis revealed that, in many instances, the replacement of a single ring system results in a chemical series with distinct SAR characteristics.

In our approach, partial core structure replacements were considered for the purpose of SAR transfer analysis, with no restrictions on the size, complexity, and composition of exchanged ring systems. However, the reported method can also be easily modified to account for scaffold or substructure replacements other than our preferred ring transformation.

Source Information

Sections of the text in this chapter have been taken from [130].

Chapter 11

Summary and Conclusions

In this thesis, computational methods that address the analysis of SARs from very different perspectives have been discussed. Presented approaches included machine-learning techniques for orphan and potency-directed screening, information-theoretic concepts for descriptor profiling, large-scale data mining of compound databases, and graphical SAR analysis. SARs were elucidated at multiple levels of detail and the major results are summarized in this chapter.

First, support vector machine search strategies for the prediction of ligands for orphan targets have been presented and were retrospectively evaluated in simulated virtual screening trials. In the analysis of search results, implications of our findings for prospective applications were evaluated and discussed. It was demonstrated that ligand prediction for orphan targets using SVMs and various target-ligand kernels was significantly influenced by nearest neighbor effects. Ligand information provided by nearest neighbors of orphan targets dominated SVM performance, much more so than the inclusion of protein information in multi-task learning strategies. As long as ligands of closely related neighbors of orphan targets were available for SVM learning, orphan target ligands could be well predicted, regardless of the type and sophistication of the kernel function that was used. Therefore, simplified strategies for SVM-based ligand prediction for orphan targets were suggested. The identification of targets that are closely related to the orphan target and for which known ligands are available should be a major objective in orphan screening campaigns.

We then extended the current spectrum of SVM approaches for different chemoinformatics applications by the introduction of potency-directed SVM searching. In comparison to conventional SVM ranking, the potency-oriented SVM linear combination and the multi-task learning strategy using the newly designed structure-activity kernel achieved an enrichment of highly potent hits at high ranking positions. One of the attractive features of potency-directed LC and SAK calculations was that high recall rates of active compounds were obtained and searches were not limited to exclusive recognition of highly potent

compounds. Furthermore, benchmarking of the strategies on HTS data sets highlighted their utility for hit identification.

Concepts from information theory were adapted to assess the amount of compound class-specific information captured by numerical descriptors. Different approaches based on the Shannon entropy concept were presented and the utility of the novel MI-DSE approach to identify descriptors containing activity-relevant information was confirmed.

Data mining approaches that explore prespecified properties of available bioactive compounds on a large scale and extract knowledge from the data were presented. First, a large-scale analysis of currently available bioactive compounds was carried out to present a systematic survey of single- and multi-target activity cliffs. It was shown that only approximately 2% of all pairs of structurally similar compounds sharing the same biological activity form activity cliffs. However, on average, approximately one of ten active compounds is involved in the formation of one or two single-target cliffs of large magnitude. Perhaps unexpectedly, activity cliffs were found to be similarly distributed over different protein target families. Moreover, only approximately 5% of all activity cliffs were multi-target cliffs, and only very few of these cliffs were formed by compounds having different target selectivity. After this global frequency analysis of activity cliffs, we asked the question whether large-magnitude potency changes were predominantly induced by specific structural modifications. Indeed, a systematic analysis of activity-cliff inducing chemical replacements on the basis of matched molecular pairs identified 146 replacements, including both R-group and core substructure changes, that displayed a general tendency to form activity cliffs. This means that introduced activity cliffs were formed in the structural context of diverse scaffolds and in compounds active against many different targets.

Bioisosteric replacements with a high propensity to produce compounds with limited potency alterations could also be identified. A compendium of 96 molecular transformations retaining potency across diverse targets and 64 modifications being conservative for single target families was assembled.

Graphical methods for the analysis of congeneric compound series were presented that aimed at the extraction of interpretable SAR rules applicable to the design of new compounds. The DRC graph structure was introduced to extend conventional analysis of analogs using R-group tables and provide more differentiated SAR information. Conventional approaches suffer from the limitation that SARs between R-group combinations at different sites cannot be analyzed in a straightforward and consistent manner. Therefore, subset relationships between different R-group combinations were utilized as organizing principle in the design of the DRCG. It was shown that this organization scheme results in graph components that represent well-defined SAR patterns. Analysis of these patterns provides an immediate access to critical substitution sites, fa-

vorable and unfavorable R-groups, or non-additive potency effects of multi-site substitutions. Furthermore, the data structure makes it possible to design new analogs by combining favorable R-group combinations derived from different compounds.

Then, a methodological framework for the study of mtSARs and identification of selectivity determinants was described. Active analogs were organized in CAGs that were adapted for mtSAR analysis by comparing and grouping compounds on the basis of pharmacophore feature exchanges. Generating this data structure for multiple targets makes it possible to determine preference orders for chemical modifications that improve target selectivity.

Finally, a combined data mining and visualization approach for the detection of SAR transfer from one chemical series to another was presented. The methodology enables the identification of alternative analog series with different core structures, corresponding substitution patterns, and comparable potency progression. Scaffolds can be exchanged between these series and new analogs suggested that incorporate preferred R-groups. The application of the approach to a systematic assessment of SAR transfer potential in publicly available compound data revealed a limited number of SAR transfer events and also confirmed that SARs of chemically related series are often substantially different.

Bibliography

[1] J. Gasteiger. Chemoinformatics: a new field with a long tradition. *Anal. Bioanal. Chem.*, 384: 57 – 64, 2006.

[2] T. Engel. Basic overview of chemoinformatics. *J. Chem. Inf. Model.*, 46: 2267 – 2277, 2006.

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

[4] M. A. Johnson and G. M. Maggiora, editors. *Concepts and Applications of Molecular Similarity.* John Wiley & Sons, New York, 1990.

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

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

[7] R. P. Sheridan, A. Rusinko, R. Nilakantan, and R. Venkataraghavan. Searching for pharmacophores in large coordinate data bases and its use in drug design. *Proc. Natl. Acad. Sci. U.S.A.*, 86: 8165 – 8169, 1989.

[8] E. X. Esposito, A. J. Hopfinger, and J. D. Madura. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.*, 275: 131 – 214, 2004.

[9] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemala, and O. Mekenyan. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.*, 45: 839 – 849, 2005.

[10] L. Michielan and S. Moro. Pharmaceutical perspectives of nonlinear QSAR strategies. *J. Chem. Inf. Model.*, 50: 961 – 978, 2010.

[11] H. Geppert, M. Vogt, and J. Bajorath. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.*, 50: 205 – 216, 2010.

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

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

[14] M. Lajiness. An evaluation of the performance of dissimilarity selection. In C. Silipo and A. Vittoria, editors, *QSAR: Rational Approaches to the Design of Bioactive Compounds*. Elsevier Science, Amsterdam, 1991.

[15] M. Wawer, E. Lounkine, A. M. Wassermann, and J. Bajorath. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov. Today*, 15: 630 – 639, 2010.

[16] M. Wawer and J. Bajorath. Extraction of structure-activity relationship information from high-throughput screening data. *Curr. Med. Chem.*, 16: 4049 – 4057, 2009.

[17] 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.

[18] D. K. Agrafiotis and J. J. M. Wiener. Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *J. Med. Chem.*, 53: 5002 – 5011, 2010.

[19] S. Renner, W. A. L. van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofman, S. Wetzel, A. Schuffenhauer, P. Ertl, T. I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, and H. Waldmann. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.*, 5: 585 – 592, 2009.

[20] Y. Hu, A. M. Wassermann, E. Lounkine, and J. Bajorath. Systematic analysis of public domain compound potency data identifies selective molecular scaffolds across druggable target families. *J. Med. Chem.*, 53: 752 – 758, 2010.

[21] 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.

[22] Y. Hu and J. Bajorath. Polypharmacology directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.*, 50: 2112 – 2118, 2010.

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

[24] A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread, and J. L. Jenkins. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2: 861 – 873, 2007.

[25] D. K. Agrafiotis, M. Shemanarev, P. J. Connolly, M. Farnum, and V. S. Lobanov. SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.*, 50: 5926 – 5937, 2007.

[26] P. J. Hajduk and D. R. Sauer. Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.*, 51: 553 – 564, 2008.

[27] P. Ertl. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.*, 43: 374 – 380, 2003.

[28] L. Peltason and J. Bajorath. Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.*, 14: 489 – 497, 2007.

[29] L. Peltason and J. Bajorath. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.*, 50: 5571 – 5578, 2007.

[30] R. Guha and J. H. van Drie. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, 48: 646 – 658, 2008.

[31] A. M. Wassermann, M. Wawer, and J. Bajorath. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.*, 53: 8209 – 8223, 2010.

[32] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2 edition, 2000.

[33] R. N. Jorissen and M. K. Gilson. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.*, 45: 549 – 561, 2005.

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

[35] H. Geppert, J. Humrich, D. Stumpfe, T. Gärtner, and J. Bajorath. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.*, 49: 767 – 779, 2009.

[36] L. Jacob and J. P. Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24: 2149 – 2156, 2008.

[37] D. Erhan, P. J. L'Heureux, S. Y. Yue, and Y. Bengio. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.*, 46: 626 – 635, 2006.

[38] C. E. Shannon. A mathematical theory of communication. *Bell. Syst. Tech. J.*, 27: 379 – 423, 1948.

[39] P. W. Kenny and J. Sadowski. Structure modification in chemical databases. In T. I. Oprea, R. Mannhold, H. Kubinyi, and H. Timmerman, editors, *Chemoinformatics in Drug Discovery (Methods and Principles in Medicinal Chemistry, volume 23)*, pages 271 – 285. Wiley-VCH, Weinheim, 2005.

[40] S. K. Mencher and L. G. Wang. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.*, 5: 3, 2005.

[41] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today*, 10: 1421 – 1433, 2005.

[42] L. Peltason, N. Weskamp, A. Teckentrup, and J. Bajorath. Exploration of structure-activity relationship determinants in analogue series. *J. Med. Chem.*, 52: 3212 – 3224, 2009.

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

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

[45] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, and H. Timmerman, editors. *Handbook of Molecular Descriptors (Methods and Principles in Medicinal Chemistry, volume 11)*. Wiley-VCH, Weinheim, 2000.

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

[47] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.*, 49: 108 – 119, 2009.

[48] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50: 742 – 754, 2010.

[49] D. A. Fidock. Drug discovery: priming the antimalarial pipeline. *Nature*, 465: 297 – 298, 2010.

[50] J. Mestres, E. Gregori-Puigjané, S. Valverde, and R. V. Solé. Data completeness – the Achilles heel of drug-target networks. *Nat. Biotechnol.*, 26: 983 – 984, 2008.

[51] 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.

[52] C. P. Austin, L. S. Brady, T. R. Insel, and F. S. Collins. NIH Molecular Libraries Initiative. *Science*, 306: 1138 – 1139, 2004.

[53] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte, and S. H. Bryant. PubChem's BioAssay Database. *Nucleic Acids Res.*, 40: D400 – D412, 2012.

[54] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, 39: D1035 – D1041, 2011.

[55] Y. Okuno, A. Tamon, H. Yabuuchi, S. Niijima, Y. Minowa, K. Tonomura, R. Kunimoto, and C. Feng. GLIDA: GPCR-ligand database for chemical genomics drug discovery – database and tools update. *Nucleic Acids Res.*, 36: D907 – D912, 2008.

[56] N. D. Rawlings, F. R. Morton, C. Y. Kok, J. Kong, and A. J. Barrett. MEROPS: the peptidase database. *Nucleic Acids Res.*, 36: D320 – D325, 2008.

[57] 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.

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

[59] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40: D1100 – D1107, 2012.

[60] A. M. Wassermann and J. Bajorath. BindingDB and ChEMBL - online compound databases for drug discovery. *Expert Opin. Drug Discov.*, 6: 683 – 687, 2011.

[61] M. Vogt, A. M. Wassermann, and J. Bajorath. Application of information-theoretic concepts in chemoinformatics. *Information*, 1: 60 – 73, 2010.

[62] A. M. Wassermann, B. Nisius, M. Vogt, and J. Bajorath. Information entropic functions for molecular descriptor profiling. *Methods Mol. Biol.*, 819: 43 – 55, 2012.

[63] F. Sams-Dodd. Target-based drug discovery: is something wrong? *Drug Discov. Today*, 10: 139 – 147, 2005.

[64] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, 26: 127 – 132, 2008.

[65] G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins. Global mapping of pharmacological space. *Nat. Biotechnol.*, 24: 805 – 815, 2006.

[66] P. Csermely, V. Agoston, and S. Pongor. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.*, 26: 178 – 182, 2005.

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

[68] Nidhi, M. Glick, J. W. Davies, and J. L. Jenkins. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.*, 46: 1124 – 1133, 2006.

[69] L. Jacob, B. Hoffmann, V. Stoven, and J. P. Vert. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinformatics*, 9: 363, 2008.

[70] H. Geppert, T. Horváth, T. Gärtner, S. Wrobel, and J. Bajorath. Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J. Chem. Inf. Model.*, 48: 742 – 746, 2008.

[71] A. M. Wassermann, H. Geppert, and J. Bajorath. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model.*, 49: 2155 – 2167, 2009.

[72] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2: 121 – 167, 1998.

[73] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[74] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Netw.*, 18: 1093 – 1110, 2005.

[75] M. J. D. Powell, J. C. Mason, and M. G. Cox. Radial basis functions for multivariable interpolation: A review. In *Algorithms for Approximation*, pages 143 – 167. Clarendon Press, New York, 1987.

[76] A. M. Wassermann, K. Heikamp, and J. Bajorath. Potency-directed similarity searching using support vector machines. *Chem. Biol. Drug Des.*, 77: 30 – 38, 2011.

[77] A. M. Wassermann, H. Geppert, and J. Bajorath. Application of support vector machine-based ranking strategies to search for target-selective compounds. *Methods Mol. Biol.*, 672: 517 – 530, 2011.

[78] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16: 276 – 277, 2000.

[79] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, 34: W32 – W37, 2006.

[80] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 92: 8700 – 8704, 1995.

[81] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, pages 564 – 575, 2002.

[82] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292: 195 – 202, 1999.

[83] P. Fontana, E. Bindewald, S. Toppo, R. Velasco, G. Valle, and S. C. Tosatto. The SSEA server for protein secondary structure alignment. *Bioinformatics*, 21: 393 – 395, 2005.

[84] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11: 1425 – 1433, 2001.

[85] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32: D115 – D119, 2004.

[86] Z. Lei and Y. Dai. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, 7: 491, 2006.

[87] Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J. W. Smith, A. L. Osterman, and A. Godzik. CutDB: a proteolytic event database. *Nucleic Acids Res.*, 35: D546 – 549, 2007.

[88] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247: 536 – 540, 1995.

[89] M. J. Sippl and M. Wiederstein. A note on difficult structure alignment problems. *Bioinformatics*, 24: 426 – 427, 2008.

[90] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28: 235 – 242, 2000.

[91] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, Cambridge, MA, 1999.

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

[93] L. Michielan, F. Stephanie, L. Terfloth, D. Hristozov, B. Cacciari, K. N. Klotz, G. Spalluto, J. Gasteiger, and S. Moro. Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. *J. Chem. Inf. Model.*, 49: 2820 – 2836, 2009.

[94] Y. Wang and J. Bajorath. Advanced fingerprint methods for similarity searching: balancing molecular complexity effects. *Comb. Chem. High Throughput. Screen.*, 13: 220 – 228, 2010.

[95] Y. Wang and J. Bajorath. Balancing the influence of molecular complexity on fingerprint similarity searching. *J. Chem. Inf. Model.*, 48: 75 – 84, 2008.

[96] C. N. Parker, C. E. Shamu, B. Kraybill, C. P. Austin, and J. Bajorath. Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? *Drug Discov. Today*, 11: 863 – 865, 2006.

[97] J. W. Godden, F. L. Stahura, and J. Bajorath. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.*, 40: 796 – 800, 2000.

[98] J. W. Godden and J. Bajorath. Differential Shannon Entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.*, 41: 1060 – 1066, 2001.

[99] A. M. Wassermann, B. Nisius, M. Vogt, and J. Bajorath. Identification of descriptors capturing compound class-specific features by mutual information analysis. *J. Chem. Inf. Model.*, 50: 1935 – 1940, 2010.

[100] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[101] J. Lin. Divergence measures based on Shannon entropy. *IEEE Trans. Inf. Theory*, 37: 145 – 151, 1991.

[102] D. Dimova, M. Wawer, A. M. Wassermann, and J. Bajorath. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.*, 51: 258 – 266, 2011.

[103] A. M. Wassermann, D. Dimova, and J. Bajorath. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem. Biol. Drug Des.*, 78: 224 – 228, 2011.

[104] M. Wawer and J. Bajorath. Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.*, 50: 1395 – 1409, 2010.

[105] J. G. Topliss. Utilization of operational schemes for analog synthesis in drug design. *J. Med. Chem.*, 15: 1006 – 1011, 1972.

[106] D. Y. Haubertin and P. Bruneau. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.*, 47: 1294 – 1302, 2007.

[107] J. W. Raymond, I. A. Watson, and A. Mahoui. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.*, 49: 1952 – 1962, 2009.

[108] A. M. Wassermann and J. Bajorath. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.*, 50: 1248 – 1256, 2010.

[109] A. M. Wassermann and J. Bajorath. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.*, 3: 425 – 436, 2011.

[110] A. M. Wassermann and J. Bajorath. Identification of target family directed bioisosteric replacements. *Med. Chem. Commun.*, 2: 601 – 606, 2011.

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

[112] E. Griffen, A. G. Leach, G. R. Robb, and D. J. Warner. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.*, 54: 7739 – 7750, 2011.

[113] R. P. Sheridan. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.*, 42: 103 – 108, 2002.

[114] J. Hussain and C. Rea. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.*, 50: 339 – 348, 2010.

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

[116] S. R. Langdon, P. Ertl, and N. Brown. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inf.*, 29: 366 – 385, 2010.

[117] G. A. Patani and E. J. LaVoie. Bioisosterism: a rational approach in drug design. *Chem. Rev.*, 96: 3147 – 3176, 1996.

[118] D. C. Young. *Computational Drug Design: A Guide for Computational and Medicinal Chemists*. John Wiley & Sons, Hoboken, NJ, 2009.

[119] H. Erlenmeyer and M. Leo. On pseudoatoms. *Helv. Chim. Acta.*, 15: 1171 – 1186, 1932.

[120] H. G. Grimm. On the systematic arrangement of chemical compounds from the perspective of research on atomic composition; and on some challenges in experimental chemistry. *Naturwissenschaften*, 17: 557 – 564, 1929.

[121] C. Ahlberg. Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today*, 4: 370 – 376, 1999.

[122] D. K. Agrafiotis, J. J. M. Wiener, A. Skalkin, and J. Kolpak. Single R-group polymorphisms (SRPs) and R-cliffs: an intuitive framework for analyzing and visualizing activity cliffs in a single analog series. *J. Chem. Inf. Model.*, 51: 1122 – 1131, 2011.

[123] A. M. Wassermann and J. Bajorath. Directed R-group combination graph: a methodology to uncover structure-activity relationship patterns in series of analogs. *J. Med. Chem.*, doi: 10.1021/jm201362h, in press.

[124] J. Scheiber, B. Chen, M. Milik, S. C. K. Sukuru, A. Bender, D. Mikhailov, S. Whitebread, J. Hamon, K. Azzaoui, L. Urban, M. Glick, J. W. Davies, and J. L. Jenkins. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.*, 49: 308 – 317, 2009.

[125] A. M. Wassermann, L. Peltason, and J. Bajorath. Computational analysis of multi-target structure-activity relationships to derive preference orders for chemical modifications toward target selectivity. *ChemMedChem*, 5: 847 – 858, 2010.

[126] G. Harper, G. S. Bravi, S. D. Pickett, J. Hussain, and D. V. S. Green. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.*, 44: 2145 – 2156, 2004.

[127] M. Böhm, J. Stürzebecher, and G. Klebe. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.*, 42: 458 – 477, 1999.

[128] H. Matter and W. Schwab. Affinity and selectivity of matrix metalloproteinase inhibitors: a chemometrical study from the perspective of ligands and proteins. *J. Med. Chem.*, 42: 4506 – 4523, 1999.

[129] L. Peltason, Y. Hu, and J. Bajorath. From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem*, 4: 1864 – 1873, 2009.

[130] A. M. Wassermann and J. Bajorath. A data mining method to facilitate SAR transfer. *J. Chem. Inf. Model.*, 51: 1857 – 1866, 2011.

[131] H. Saigo, J. P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20: 1682 – 1689, 2004.

Appendix A

Software and Databases

Software and databases used in this thesis are listed in alphabetical order in Tables A-1 and A-2, respectively.

Table A-1: Software

EMBOSS	European Molecular Biology Open Software Suite
description	EMBOSS is a software analysis package for molecular biology and bioinformatics applications including, for example, sequence alignment, rapid database searching with sequence patterns, and protein motif identification.
provider	European Bioinformatics Institute, Hinxton, UK
URL	http://emboss.sourceforge.net/
Java	
description	Java is an object-oriented programming language.
provider	Oracle Corporation, Redwood City, CA, USA
URL	http://www.oracle.com/technetwork/java/
JUNG	Java Universal Network/Graph Framework
description	JUNG is a software library providing a common and extendible language for the modeling, analysis, and visualization of data representable as a graph or network.
provider	Danyel Fisher, Tom Nelson, and Joshua O'Madadhain
URL	http://jung.sourceforge.net/
MOE	Molecular Operating Environment
description	MOE is an interactive computing and molecular modeling tool written in the Scientific Vector Language, a self-contained programming system developed by the Chemical Computing Group. MOE provides applications for the calculation of numerical property descriptors and implementations of numerous fingerprint formats including MACCS structural keys and the TGD fingerprint.
provider	Chemical Computing Group Inc., Montreal, QC, Canada
URL	http://www.chemcomp.com/

Table A-1: Software (continued)

OEChem TK	OpenEye Chemistry Toolkit
description	OEChem TK is a programming library for chemistry and chemoinformatics that is, inter alia, wrapped for Java.
provider	OpenEye Scientific Software Inc., Santa Fe, NM, USA
URL	http://www.eyesopen.com/oechem-tk/
Perl	
description	Perl is a scripting language.
provider	Larry Wall
URL	http://www.perl.org/
Pipeline Pilot	
description	Pipeline Pilot is a scientific informatics platform that provides components for the creation of workflow protocols enabling data analyses and a variety of chemoinformatics applications, e.g., fingerprint and scaffold calculations.
provider	Accelrys Inc., San Diego, CA, USA
URL	http://www.accelrys.com/products/scitegic/
PROFEAT	Protein Feature Server
description	PROFEAT is a web server for computing commonly used features of proteins and peptides from amino acid sequence.
provider	Bioinformatics & Drug Design Group, Computational Science Department, National University of Singapore, Singapore
URL	http://bidd.cz3.nus.edu.sg/cgi-bin/prof/protein/profnew.cgi/
PSIPRED	
description	The PSIPRED server is a web server for protein secondary structure prediction.
provider	Bioinformatics Group, Departments of Computer Science, University College London, UK
URL	http://bioinf.cs.ucl.ac.uk/psipred/
R	The R Project for Statistical Computing
description	R is a programming language and software environment for statistical computing and graphics.
provider	R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria
URL	http://www.r-project.org/
SSEA	Secondary Structure Element Alignment
description	SSEA is a web server for computing either local or global alignments of protein secondary structures.
provider	Biocomputing UP, Department of Biology, University of Padua, Italy
URL	http://protein.bio.unipd.it/ssea/

Table A-1: Software (continued)

SVM^{light}	
description	SVM^{light} is an implementation of support vector machines for classification, regression, and ranking problems.
provider	Thorsten Joachims
URL	http://svmlight.joachims.org/
TopMatch-web	
description	TopMatch-web is a web server for the alignment and superposition of protein structures.
provider	Division of Bioinformatics, Department of Molecular Biology, University of Salzburg, Austria
URL	http://topmatch.services.came.sbg.ac.at/

Table A-2: Databases

BindingDB	
description	BindingDB is a database of approximately 800 000 measured binding affinities, in particular for interactions between proteins considered to be therapeutically relevant and drug-like small molecules (see Chapter 2).
provider	Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA, USA
URL	http://www.bindingdb.org/
ChEMBL	
description	ChEMBL is a database reporting bioactivity data and calculated properties for more than one million drug-like small molecules (see Chapter 2).
provider	European Bioinformatics Institute, Hinxton, UK
URL	https://www.ebi.ac.uk/chembl/
CutDB	
description	The CutDB is a collection of documented proteolytic events for natural proteins in vivo or in vitro, with each entry in the database corresponding to a combination of a protease, a protein substrate, and a cleavage site.
provider	Burnham Institute for Medical Research, La Jolla, CA, USA
URL	http://cutdb.burnham.org/
MDDR	MDL Drug Data Report
description	The MDDR is a database containing more than 150 000 biologically active compounds assembled from patent literature, meetings, congresses, and journals.
provider	Symyx Software, San Ramon, CA, USA
URL	http://www.symyx.com/

Table A-2: Databases (continued)

MEROPS	
description	The MEROPS database provides a structure-based classification of peptidases and is an information resource for proteins and small molecules that inhibit them.
provider	Wellcome Trust Sanger Institute, Cambridge, UK
URL	http://merops.sanger.ac.uk/
PDB	Protein Data Bank
description	The PDB is a central repository for 3D structural data of proteins and nucleic acids obtained by NMR spectroscopy or X-ray crystallography.
provider	European Bioinformatics Institute, Hinxton, UK; Osaka University, Japan; Research Collaboratory for Structural Bioinformatics, USA
URL	http://www.pdb.org/
PubChem BioAssay	
description	The PubChem Bioassay database contains 500 000 descriptions of assay protocols and provides over 130 million bioactivity outcomes. More than 1 600 assays are confirmatory, hence providing quantitative potency measurements.
provider	National Center for Biotechnology Information, Bethesda, MD, USA
URL	http://pubchem.ncbi.nlm.nih.gov/
SCOP	Structural Classification of Proteins
description	The SCOP database provides a (mostly manually curated) classification of protein structural domains based on amino acid and structure similarities.
provider	Laboratory of Molecular Biology, Cambridge, UK
URL	http://scop.mrc-lmb.cam.ac.uk/scop/
UniProt	Universal Protein Resource
description	UniProt is a central repository of protein sequence and functional annotation.
provider	European Bioinformatics Institute, Hinxton, UK; Protein Information Resource, Georgetown University Medical Center, DC, USA; Swiss Institute of Bioinformatics, Lausanne, Switzerland
URL	http://www.uniprot.org/
ZINC	
description	ZINC is a database of over 14 million commercially available compounds provided in 3D formats.
provider	Shoichet Laboratory, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA
URL	http://zinc.docking.org/

Appendix B

Orphan Screening – Additional Information

A comment on the general validity of target kernel functions used in our simulated orphan screening trials reported in Chapter 3 is provided.

Tables B-1 and B-2 report average recovery rates for individual targets of data set 1 using the strategies SVM TLK and SVM LC, respectively. Corresponding results for individual targets of data set 2 are provided in Tables B-3 and B-4.

Average recovery rates for homology-based SVM on individual targets of data set 2 are given in Table B-5.

Table B-6 reports average recovery rates for the SVM strategies SVM TLK and SVM LC that are obtained for individual targets of data set 1 when ligands of the nearest neighbor target are excluded from training.

A Note on the Validity of Kernel Functions
Some of our target kernel functions introduced in the context of orphan screening are not generally valid kernel functions because they are not necessarily positive semi-definite for all protein sets (i.e., for a given set of proteins, an all-versus-all matrix of scores might have some negative eigenvalues). To overcome this problem, one can convert a symmetric into a positive semi-definite matrix by subtracting from the diagonal of the matrix its smallest negative eigenvalue [131]. However, for the analysis presented in this thesis, this conversion was not required because all score matrices were positive semi-definite for our data sets.

Table B-1: Search results for ligand prediction using SVM TLK (set 1)

	uniform	*needle*	*water*	*PROFEAT*	*spectrum*	*SSEA*	*GO*	*cleavage*	*SCOP*	*Topmatch*	*MEROPS*
						MACCS					
ace2	5.7	5.2	4.4	3.9	5.2	3.5	0.0	0.9	0.4	4.4	1.3
cal2	2.3	5.7	6.6	2.3	5.2	2.5	19.1	27.5	26.1	13.6	25.2
cas3	6.7	6.7	8.9	8.1	6.6	8.7	7.4	8.3	8.3	6.3	6.2
catD	21.2	22.2	34.6	27.5	46.0	31.7	37.2	39.5	24.5	24.3	24.8
catL	8.5	9.3	11.0	8.6	8.6	8.8	11.6	17.7	10.8	10.4	13.6
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	1.7	1.2	1.7	1.6	1.4	1.6	1.6	1.5	1.1	1.7	0.7
mmp2	3.3	17.4	16.2	4.0	15.4	8.7	14.9	20.3	34.2	39.6	41.9
mmp8	9.1	30.9	28.2	10.0	20.9	17.3	23.6	27.3	42.7	48.2	46.4
ren	3.0	18.7	9.5	5.0	14.1	9.7	30.5	12.2	26.0	17.9	25.6
thr	17.5	29.2	27.5	18.9	13.4	27.2	13.7	27.5	29.1	29.4	28.9
try	19.6	37.9	30.4	20.4	14.4	34.8	28.3	35.2	41.7	38.5	41.7
average	8.2	15.4	14.9	9.2	12.6	12.9	15.7	18.2	20.4	19.5	21.4
						TGD					
ace2	10.4	16.5	11.3	10.9	14.8	13.0	0.4	7.8	6.1	10.9	10.4
cal2	9.3	6.6	9.3	6.4	7.7	6.1	11.6	12.7	11.1	10.7	15.5
cas3	14.9	14.4	17.4	16.0	20.4	15.0	19.0	21.2	21.5	21.6	20.8
catD	36.0	38.2	41.2	36.3	42.5	39.9	43.7	42.9	43.4	40.5	42.3
catL	6.3	6.2	6.3	6.3	6.3	6.2	7.0	7.5	7.4	6.2	7.3
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	11.5	7.6	11.4	11.2	9.7	11.7	5.1	2.7	10.4	5.5	11.7
mmp2	4.9	33.0	15.4	5.5	16.8	8.7	10.9	18.6	49.6	57.3	60.5
mmp8	10.0	37.3	26.4	12.7	28.2	15.5	22.7	34.6	54.6	57.3	57.3
ren	17.5	43.5	33.8	22.9	39.6	30.6	43.0	40.4	44.7	43.5	44.2
thr	19.5	25.1	22.7	20.3	19.3	23.5	25.9	27.5	28.2	27.0	28.2
try	27.3	38.1	32.9	27.5	24.6	35.2	29.6	43.9	45.6	46.0	46.4
average	14.0	22.2	19.0	14.7	19.2	17.1	18.2	21.7	26.9	27.2	28.7

Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1.

Table B-2: Search results for ligand prediction using SVM LC (set 1)

	uniform	*needle*	*water*	*PROFEAT*	*spectrum*	*SSEA*	*GO*	*cleavage*	*SCOP*	*Topmatch*	*MEROPS*
						MACCS					
ace2	3.5	3.9	3.0	3.5	4.8	3.5	2.2	1.7	0.4	4.4	1.3
cal2	7.3	5.2	7.1	7.5	8.0	7.1	14.1	14.6	22.7	9.8	22.7
cas3	7.1	6.6	7.5	7.0	6.0	7.5	6.3	6.9	7.1	5.6	5.6
catD	28.8	28.2	34.3	29.7	46.6	29.9	36.5	35.7	27.9	37.9	27.5
catL	9.0	8.9	8.9	9.0	8.6	9.2	11.9	14.5	11.8	10.0	12.9
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	1.2	1.2	1.3	1.3	1.2	1.3	1.7	1.4	1.0	1.4	0.6
mmp2	2.6	13.1	5.6	3.0	9.0	3.6	7.1	7.2	27.7	29.6	39.4
mmp8	4.6	25.5	8.2	4.6	12.7	4.6	13.6	11.8	35.5	41.8	44.6
ren	4.1	18.3	7.9	4.5	14.3	7.9	12.3	16.6	29.4	17.6	29.9
thr	11.2	28.2	16.3	12.1	11.4	16.4	28.4	21.4	29.8	28.7	29.8
try	15.6	37.1	21.5	15.4	14.0	22.7	20.2	29.6	43.9	41.0	43.7
average	7.9	14.7	10.1	8.1	11.4	9.5	12.9	13.5	19.8	19.0	21.5
						TGD					
ace2	15.2	16.5	15.2	15.2	16.5	15.7	15.2	10.9	10.9	13.5	13.5
cal2	13.6	8.4	12.7	11.8	10.0	12.3	17.7	17.7	10.9	11.8	16.1
cas3	20.1	16.1	20.5	20.0	21.8	19.5	22.2	23.3	23.1	22.2	21.0
catD	34.6	38.0	39.2	34.5	42.6	38.2	42.8	45.5	44.3	41.4	44.0
catL	6.6	6.6	6.6	6.6	6.7	6.7	6.4	7.3	7.8	6.6	7.4
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	7.5	5.3	7.5	7.8	8.0	7.6	7.5	2.5	9.8	4.1	10.7
mmp2	4.4	23.5	7.1	4.9	11.2	5.5	8.6	6.3	39.1	37.2	53.0
mmp8	11.8	28.2	18.2	11.8	22.7	12.7	16.4	20.0	52.7	54.6	53.6
ren	21.8	45.0	31.8	23.8	39.8	29.6	36.7	40.5	46.4	45.7	45.9
thr	19.5	24.9	21.4	19.8	19.9	21.7	27.0	25.9	27.7	26.6	28.0
try	19.4	35.2	25.6	20.8	22.3	25.2	24.8	35.4	44.8	41.5	45.2
average	14.5	20.6	17.1	14.7	18.5	16.2	18.8	19.6	26.5	25.4	28.2

Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-1.

Table B-3: Search results for ligand prediction using SVM TLK (set 2)

	uniform	*needle*	*water*	*PROFEAT*	*spectrum*	*SSEA*	*GO*	*cleavage*	*SCOP*	*Topmatch*	*MEROPS*
						MACCS					
cal1	38.1	49.5	52.0	50.0	50.5	51.7	49.5	53.2	52.2	53.2	51.0
cal2	32.7	65.7	55.5	44.1	63.0	50.9	52.7	58.9	65.7	63.0	67.1
cas1	6.9	10.0	12.5	10.6	16.3	15.6	19.4	20.0	16.9	20.0	18.8
cas3	6.6	19.5	12.3	7.4	15.6	12.4	9.1	21.4	24.4	21.2	24.0
catB	37.5	35.8	40.0	35.8	39.2	34.2	30.0	39.2	25.0	29.2	23.3
catK	13.7	18.5	16.7	16.3	19.4	16.1	17.8	17.3	16.6	17.3	17.7
catL	22.7	24.4	24.1	23.6	26.0	25.9	23.2	24.3	24.0	24.8	23.4
catS	11.8	24.1	11.0	14.0	23.6	17.6	16.3	14.5	21.9	21.3	23.9
faXa	3.6	9.7	7.7	4.7	8.4	7.9	8.2	9.2	10.4	10.2	10.5
thr	7.3	25.9	19.9	8.8	15.6	19.2	24.7	21.0	28.3	25.9	29.0
try	16.0	44.3	33.0	19.6	23.8	35.1	29.8	41.9	45.1	40.6	47.4
average	17.9	29.8	25.9	21.4	27.4	26.0	25.5	29.2	30.0	29.7	30.5
						TGD					
cal1	21.0	30.7	27.3	22.7	31.0	25.1	28.1	27.3	31.2	30.7	29.8
cal2	14.3	44.3	19.8	16.4	47.7	20.7	20.5	20.9	40.0	26.1	48.4
cas1	16.3	20.0	19.4	16.3	29.4	21.3	24.4	33.8	31.9	35.0	33.8
cas3	18.0	30.1	24.4	19.0	27.3	24.3	22.7	34.9	35.9	33.5	35.9
catB	35.0	35.8	34.2	35.0	33.3	33.3	37.5	34.2	29.2	35.0	31.7
catK	8.3	11.8	9.9	8.8	11.6	9.0	9.4	10.8	11.5	11.3	11.7
catL	9.3	10.3	8.9	9.5	10.8	9.5	9.7	10.3	10.4	10.6	9.9
catS	3.9	3.7	3.8	3.7	3.7	3.4	3.6	3.2	3.6	3.4	3.7
faXa	7.4	7.7	7.6	7.3	7.6	7.7	7.8	7.7	8.1	7.9	7.9
thr	28.0	30.4	30.0	28.6	29.8	29.8	31.6	30.4	31.1	30.7	31.2
try	47.4	50.9	50.8	47.9	47.2	50.6	50.9	53.0	52.8	50.9	52.8
average	19.0	25.1	21.5	19.5	25.4	21.3	22.4	24.2	26.0	25.0	27.0

Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-2.

Table B-4: Search results for ligand prediction using SVM LC (set 2)

	uniform	*needle*	*water*	*PROFEAT*	*spectrum*	*SSEA*	*GO*	*cleavage*	*SCOP*	*Topmatch*	*MEROPS*
						MACCS					
cal1	45.7	48.5	51.2	48.6	49.3	50.0	50.0	50.7	50.0	52.0	50.2
cal2	42.5	67.5	55.7	48.0	63.2	52.5	65.0	65.7	78.6	69.8	74.3
cas1	8.8	6.3	10.0	9.4	12.5	11.3	19.4	20.0	21.3	20.6	21.3
cas3	7.5	16.3	10.2	7.8	14.3	10.3	10.3	17.8	22.7	17.9	22.7
catB	40.0	36.7	40.8	40.8	40.8	39.2	33.3	41.7	35.0	38.3	29.2
catK	17.8	19.2	18.3	18.0	19.3	18.4	19.8	18.7	18.0	18.9	18.1
catL	23.8	25.6	23.8	23.6	26.3	24.5	21.8	21.9	23.0	24.3	24.0
catS	13.2	23.0	12.8	13.9	22.3	15.7	16.9	14.1	21.1	19.4	22.6
faXa	2.6	8.2	4.5	2.9	7.2	4.6	4.6	6.2	10.2	9.6	10.2
thr	7.0	23.2	11.7	7.5	14.2	11.7	27.0	13.8	28.5	26.1	28.5
try	13.8	44.0	23.6	14.9	22.6	24.7	22.1	34.0	48.5	44.0	48.5
average	20.2	29.0	23.9	21.4	26.5	23.9	26.4	27.7	32.4	31.0	31.8
						TGD					
cal1	22.4	31.2	26.8	24.4	31.2	25.9	28.1	28.1	30.5	30.2	30.7
cal2	19.6	46.8	20.9	20.5	50.0	23.2	26.4	21.8	39.8	24.8	50.2
cas1	19.4	21.9	22.5	20.6	30.6	23.1	26.9	36.9	38.8	36.9	38.8
cas3	19.0	29.3	22.8	19.6	27.1	22.3	24.9	33.6	36.0	32.9	36.0
catB	34.2	37.5	32.5	33.3	35.8	33.3	39.2	34.2	35.8	36.7	35.8
catK	10.1	12.3	10.9	10.4	11.8	10.4	11.6	12.6	13.8	12.6	12.2
catL	10.0	11.0	10.0	10.0	11.4	10.0	9.9	10.6	11.2	11.0	10.7
catS	3.0	3.0	3.1	2.9	3.1	2.8	3.4	2.6	3.3	3.0	3.0
faXa	6.6	7.5	7.0	6.7	7.4	7.0	7.4	7.3	7.9	7.7	7.9
thr	28.2	30.5	29.8	28.5	29.9	29.8	31.9	30.3	31.3	31.0	31.3
try	41.7	48.7	46.0	42.5	45.5	46.2	44.9	48.1	50.9	50.0	50.9
average	19.5	25.4	21.1	19.9	25.8	21.3	23.1	24.2	27.2	25.2	28.0

Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 100 compounds averaged over ten independent trials per target. Target abbreviations are used according to Table 3.4-2.

Table B-5: Search results for ligand prediction using homology-based SVM (set 2)

	MACCS				TGD			
	NN target[a]		all[a]		NN target[a]		all[a]	
	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]
cal1	46.8	62.2	46.8	62.2	29.8	40.2	29.8	40.2
cal2	61.8	76.8	61.8	76.8	49.3	65.5	49.3	65.5
cas1	21.3	51.9	21.3	51.9	36.9	69.4	36.9	69.4
cas3	21.6	41.2	21.6	41.2	36.0	54.2	36.0	54.2
catB	10.0	39.2	13.3	63.3	13.3	35.8	21.7	49.2
catK	14.0	42.9	16.7	47.7	8.0	22.4	8.9	33.2
catL	14.5	36.4	22.2	64.3	7.0	26.3	9.5	26.6
catS	14.0	41.5	24.8	60.7	3.5	18.0	4.3	32.6
faXa	8.7	31.7	10.5	39.0	7.6	48.7	8.1	53.0
thr	9.6	33.3	28.1	71.6	23.6	70.9	31.2	82.0
try	12.5	29.1	44.7	67.6	34.0	72.3	51.9	87.7
average	21.3	44.2	28.3	58.8	22.6	47.6	26.1	54.0

[a] Reference target(s). [b] Set size. Recovery rates (in %) are reported for all targets in data set 2 averaged over ten independent trials per target. Searches are either carried out using only the ligands of the nearest neighbor (NN) target as positive training class or using the pooled ligands of all members of the orphanized target's subfamily (column "all") as positive training class. For simulated orphan targets of the calpain and caspase families, reference sets for both settings are identical because only one reference target belonging to the same subfamily exists in the test system. Target abbreviations are used according to Table 3.4-2.

Table B-6: Search results for ligand prediction using SVM TLK and LC in the absence of nearest neighbor information

	MACCS				TGD			
	SVM LC[a]		SVM TLK[a]		SVM LC[a]		SVM TLK[a]	
	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]	100[b]	1000[b]
ace2	4.4	17.8	4.4	15.7	14.4	36.5	13.5	40.0
cal2	3.2	49.6	2.7	41.8	13.0	32.5	10.7	30.0
cas3	7.0	35.8	7.5	36.3	21.3	48.5	20.8	48.1
catD	15.9	54.3	20.6	60.8	28.6	65.5	27.5	64.2
catL	7.5	26.3	8.5	25.8	5.8	12.6	5.8	11.5
mgcp	0.0	1.1	0.0	1.1	0.0	0.0	0.0	0.0
metap2	0.8	7.3	0.9	6.1	11.1	38.8	12.2	40.8
mmp2	0.4	5.6	0.3	4.7	4.6	23.1	4.7	23.9
mmp8	0.0	6.4	0.0	6.4	5.5	24.6	6.4	23.6
ren	1.3	14.2	1.2	16.9	4.9	44.2	4.5	44.5
thr	4.5	29.7	4.1	29.2	14.5	43.3	13.8	41.1
try	5.0	25.4	4.6	24.4	9.8	44.8	10.8	49.2
average	4.2	22.8	4.6	22.4	11.1	34.5	10.9	34.7

[a] Strategy. [b] Set size. Recovery rates (in %) are reported for all targets in data set 1 averaged over ten independent search trials per target. For all orphan targets, the nearest neighbor target and its ligands were excluded from training. All searches were carried out with the MEROPS target kernel. Target abbreviations are used according to Table 3.4-1.

Appendix C

Potency-Directed LBVS – Additional Information

Search results for potency-directed SVM calculations using the MACCS fingerprint are provided.

Figure C-1 reports cumulative recall curves for potency-balanced reference sets and the three different SVM strategies introduced in Chapter 4.

Figure C-2 shows SVM performance for database selection sets of constant size.

Figure C-3 compares control calculations using highly potent reference compounds only to an advanced SVM strategy (SVM SAK) developed for potency-directed LBVS.

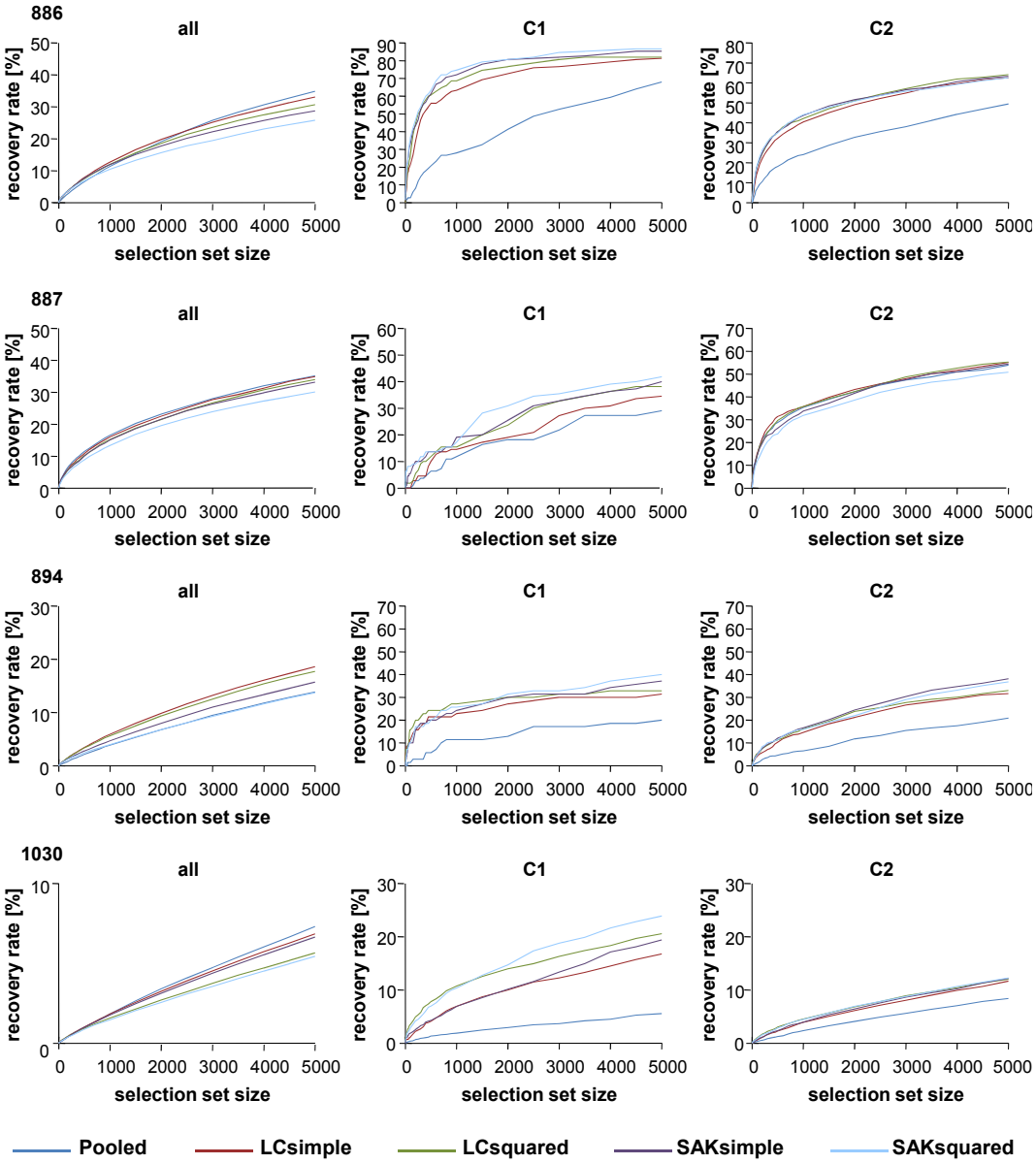


Figure C-1: Cumulative recall curves for potency-balanced reference sets For each bioassay, cumulative recall curves are shown for all active compounds and the highest potency categories (C1 and C2) and different SVM strategies using the MACCS fingerprint. Recall curves represent the average of ten independent trials using different reference sets. Potency-balanced reference sets consist of compounds spanning the entire potency range in a data set. The figure is adapted from [76].

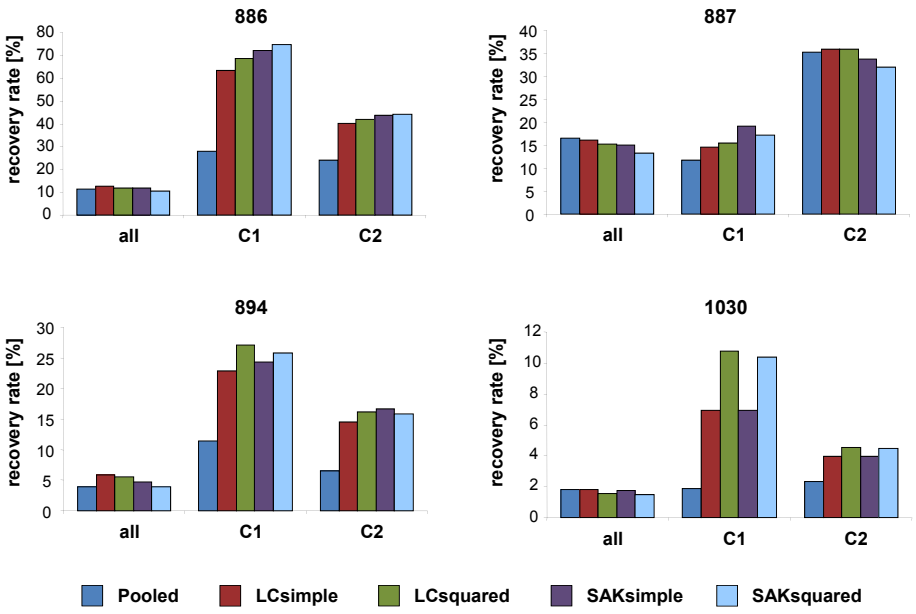


Figure C-2: Support vector machine performance for database selection sets of constant size Recovery rates are shown for the MACCS fingerprint, potency-balanced reference sets, and database selection sets of 1 000 compounds. The results are averaged over ten independent trials per data set. The figure is adapted from [76].

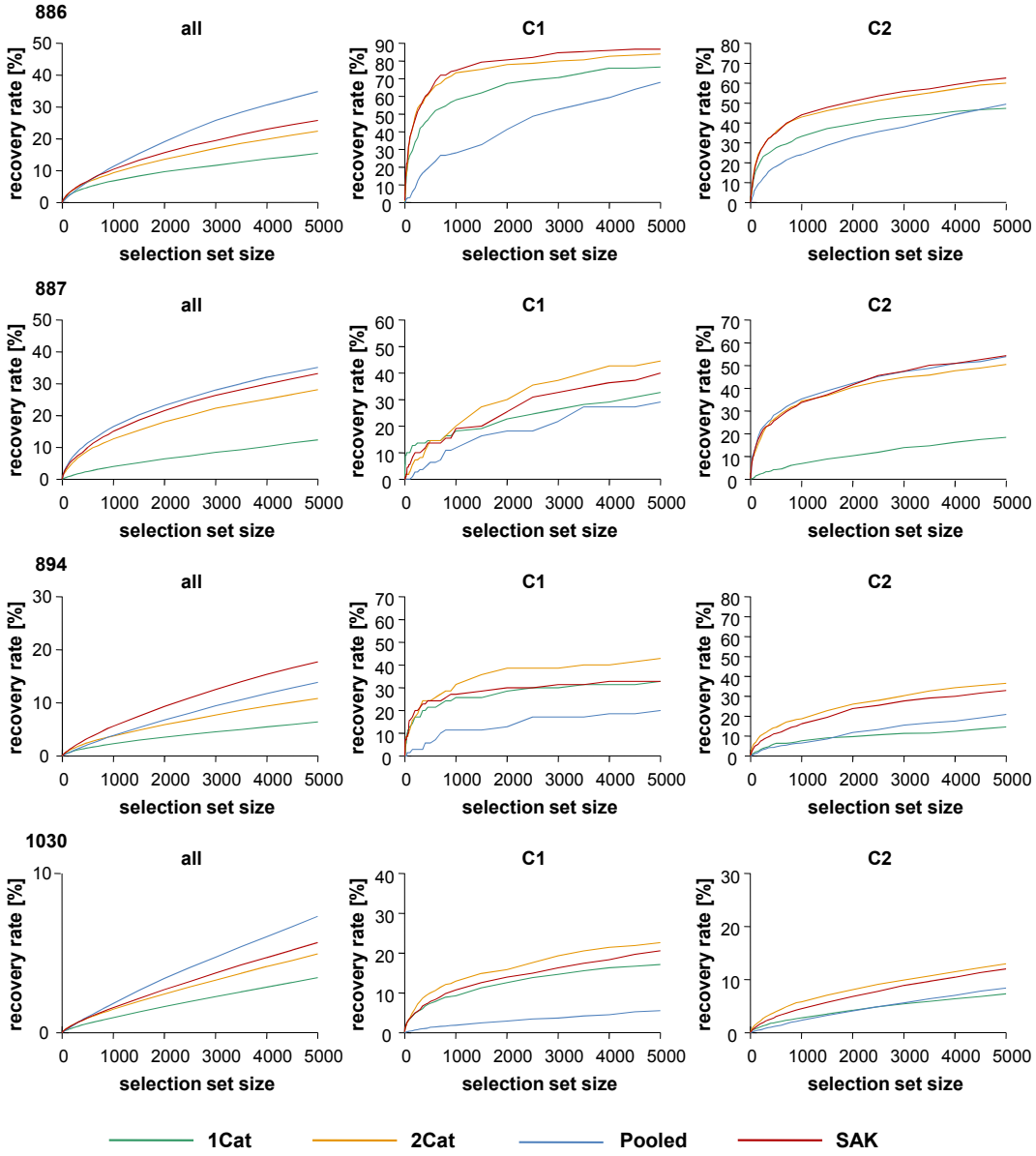


Figure C-3: Control calculations using highly potent reference compounds
For each bioassay, recall curves are shown for all active compounds and the two highest potency categories (C1 and C2). Compound recall is monitored for different SVM strategies using MACCS as fingerprint representation averaged over ten independent trials. The following strategies are compared: standard SVM with reference compounds from potency category 1 ('1Cat'), 1 and 2 ('2Cat'), and all categories ('Pooled') and SVM SAK. SAK_{simple} is shown for sets 887 and 894, $SAK_{squared}$ for sets 886 and 1030. The figure is adapted from [76].

Appendix D

Class-Specific Descriptors – Additional Information

An explanation for the value range [0,1] of the MI-DSE method introduced in Chapter 5 is given.

The 171 descriptors used in our large-scale comparison of DSE- and MI-DSE-based rankings are listed and grouped by descriptor "type" in Table D-1.

Table D-2 analyzes the influence of different binning schemes on DSE- and MI-DSE-based descriptor rankings. For both methods, descriptor rankings using 8, 16, 32, or 64 bins in the histogram calculation are systematically compared on 168 activity classes extracted from the ChEMBL database and average Spearman rank correlation coefficients are reported.

For the targets carbonic anhydrase II and muscarinic acetylcholine receptor M2, the ten top-ranked descriptors based on DSE and MI-DSE are listed in Table D-3.

MI-DSE value range
The mutual information MI(D,C)can be equivalently expressed as

$$\mathrm{MI}(D,C) = \mathrm{H}(D) - \mathrm{H}(D|C) = \mathrm{H}(C) - \mathrm{H}(C|D) \tag{D.1}$$

Since only two classes **A** and **B** are being compared, the maximal entropy that can be obtained for H(C) assuming equal probabilities of $Pr(C = \mathbf{A}) = Pr(C = \mathbf{B}) = 0.5$ is one (i.e., $\mathrm{H}(C) = \log_2(2)$). From equation D.1 it follows that $\mathrm{MI}(D,C) \leq H(C)$, and therefore MI-DSE produces values in the range from zero to one.

Table D-1: MOE descriptors

type	descriptors
physical properties	apol, bpol, density, FCharge, logP(o/w), logS, mr, SlogP, SMR, TPSA, vdw_area, vdw_vol, Weight
subdivided surface areas	SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SMR_VSA0, SMR_VSA1, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7
atom counts and bond counts	a_aro, a_count, a_heavy, a_IC, a_ICM, a_nB, a_nBr, a_nC, a_nCl, a_nF, a_nH, a_nI, a_nN, a_nO, a_nP, a_nS, b_1rotN, b_1rotR, b_ar, b_count, b_double, b_heavy, b_rotN, b_rotR, b_single, b_triple, chiral, chiral_u, lip_acc, lip_don, rings, VAdjEq, VAdjMa
adjacency and distance matrix descriptors	balabanJ, BCUT_PEOE_0, BCUT_PEOE_1, BCUT_PEOE_2, BCUT_PEOE_3, BCUT_SLOGP_0, BCUT_SLOGP_1, BCUT_SLOGP_2, BCUT_SLOGP_3, BCUT_SMR_0, BCUT_SMR_1, BCUT_SMR_2, BCUT_SMR_3, diameter, GCUT_PEOE_0, GCUT_PEOE_1, GCUT_PEOE_2, GCUT_PEOE_3, GCUT_SLOGP_0, GCUT_SLOGP_1, GCUT_SLOGP_2, GCUT_SLOGP_3, GCUT_SMR_0, GCUT_SMR_1, GCUT_SMR_2, GCUT_SMR_3, petitjean, petitjeanSC, radius, VDistEq, VDistMa, wienerPath, wienerPol
pharmacophore feature descriptors	a_acc, a_acid, a_base, a_don, a_hyd, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_hyd, vsa_other, vsa_pol
Kier&Hall connectivity and kappa shape indices	chi0, chi0_C, chi0v, chi0v_C, chi1, chi1_C, chi1v, chi1v_C, Kier1, Kier2, Kier3, KierA1, KierA2, KierA3, KierFlex, zagreb
partial charge descriptors	PEOE_PC+, PEOE_PC-, PEOE_RPC+, PEOE_RPC-, PEOE_VSA+0, PEOE_VSA+1, PEOE_VSA+2, PEOE_VSA+3, PEOE_VSA+4, PEOE_VSA+5, PEOE_VSA+6, PEOE_VSA-0, PEOE_VSA-1, PEOE_VSA-2, PEOE_VSA-3, PEOE_VSA-4, PEOE_VSA-5, PEOE_VSA-6, PEOE_VSA_FHYD, PEOE_VSA_FNEG, PEOE_VSA_FPNEG, PEOE_VSA_FPOL, PEOE_VSA_FPOS, PEOE_VSA_FPPOS, PEOE_VSA_HYD, PEOE_VSA_NEG, PEOE_VSA_PNEG, PEOE_VSA_POL, PEOE_VSA_POS, PEOE_VSA_PPOS, Q_PC+, Q_PC-, Q_RPC+, Q_RPC-, Q_VSA_FHYD, Q_VSA_FNEG, Q_VSA_FPNEG, Q_VSA_FPOL, Q_VSA_FPOS, Q_VSA_FPPOS, Q_VSA_HYD, Q_VSA_NEG, Q_VSA_PNEG, Q_VSA_POL, Q_VSA_POS, Q_VSA_PPOS

For more detailed information on the descriptors, see:
http://www.chemcomp.com/journal/descr.htm.

Table D-2: Spearman rank correlation coefficients

	MI-DSE				DSE			
	8 bins	16 bins	32 bins	64 bins	8 bins	16 bins	32 bins	64 bins
8 bins	1	0.935	0.890	0.841	1	0.919	0.869	0.798
16 bins		1	0.967	0.922		1	0.956	0.895
32 bins			1	0.973			1	0.961
64 bins				1				1

Table D-3: DSE- and MI-DSE-based descriptor rankings

Descriptors (DSE)	Descriptors (MI-DSE)
carbonic anhydrase II	
GCUT_SLOGP_0 (0.66)	PEOE_VSA-4 (0.76)
chiral_u (0.53)	vsa_don (0.73)
PEOE_VSA-1 (0.42)	SMR_VSA4 (0.72)
BCUT_SLOGP_0 (0.35)	SlogP_VSA1 (0.65)
SMR_VSA2 (0.33)	GCUT_SLOGP_0 (0.63)
SlogP_VSA4 (0.29)	vsa_pol (0.53)
a_aro (0.26)	a_nS (0.49)
SMR_VSA3 (0.22)	GCUT_SLOGP_1 (0.49)
chiral (0.20)	PEOE_VSA-3 (0.42)
BCUT_SMR_3 (0.19)	TPSA (0.41)
muscarinic acetylcholine receptor M2	
GCUT_SMR_0 (1.01)	FCharge (0.82)
BCUT_PEOE_3 (0.89)	a_base (0.80)
SMR_VSA4 (0.88)	GCUT_SMR_0 (0.77)
SlogP_VSA1 (0.78)	BCUT_SLOGP_0 (0.73)
BCUT_SMR_0 (0.73)	BCUT_PEOE_3 (0.69)
BCUT_SMR_3 (0.71)	GCUT_PEOE_0 (0.69)
BCUT_SLOGP_0 (0.70)	BCUT_SMR_0 (0.64)
GCUT_PEOE_0 (0.70)	GCUT_PEOE_3 (0.62)
vsa_acc (0.60)	BCUT_SMR_3 (0.56)
PEOE_VSA+5 (0.56)	BCUT_PEOE_0 (0.55)

For the top-ranked descriptors based on DSE ("Descriptors (DSE)") and MI-DSE ("Descriptors (MI-DSE)"), corresponding DSE and MI-DSE values are given in parentheses.

Appendix E

Molecular Transformations – Additional Information

For the 146 molecular transformations that frequently introduce activity cliffs according to the search criteria detailed in Chapter 7, SMIRKS representations are provided in Table E-1.

Table E-2 reports on the potency directionality of cliff-forming transformations.

SMIRKS representations for the identified set of 96 biosisosteric replacements are given in Table E-3.

Table E-1: Activity cliff-introducing chemical replacements

SMIRKS	#cliffs	#targets	#MMPs	#records	freq.
[*:1]C(=O)N[O-]>>[*:1]P(=O)([O-])[O-]	5	4	4	6	83.3
[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1C[NH2+]CC1[*:2])=O >>[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1Cc2c(cccc2)C1[*:2])=O	18	6	5	26	69.2
[*:2]NC1CCCC1OC(=O)c1cc(O)c([*:1])c(O)c1 >>[*:2]NC1c2c(CC1OC(=O)c1cc(O)c([*:1])c(O)c1)cccc2	17	6	4	26	65.4
[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1CC(O)CC1[*:2])=O >>[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1Cc2c(cccc2)C1[*:2])=O	22	6	4	34	64.7
[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1CC(CC1[*:2])C[NH3+])=O >>[*:3]C(=O)c1c(O)cc(cc1[*:1])C(OC1Cc2c(cccc2)C1[*:2])=O	22	6	4	34	64.7
[*:2]NC1CC(CC1OC(=O)c1cc(O)c([*:1])c(O)c1)CO >>[*:2]NC1c2c(CC1OC(=O)c1cc(O)c([*:1])c(O)c1)cccc2	20	6	4	34	58.8
[*:1]CCC([NH3+])CC(P(=O)([O-])[O-])O >>[*:1]CCC([NH3+])CCS(=O)(=O)[O-]	9	4	4	16	56.3
[*:1]CCC([NH3+])CC(=O)[O-] >>[*:1]CCC([NH3+])CC(P(=O)([O-])[O-])O	9	4	4	16	56.3
[*:1]CCC([NH3+])(CO)CO >>[*:1]CCC([NH3+])CC(P(=O)([O-])[O-])O	8	4	4	16	50.0
[*:1][O-]>>[*:1]c1ccccc1	8	4	4	16	50.0
[*:1]C(=O)CCc1c2c([nH]c1)cccc2>>[*:1]C(=O)Cc1ccccc1	8	4	4	17	47.1
[*:2]C(=O)NC(C(=O)NC[*:1])C >>[*:2]C(=O)NC(CCCC)C(=O)NC[*:1]	10	5	5	22	45.5
[*:1]Nc1ccccc1>>[*:1][O-]	4	4	4	9	44.4
[*:2]C(OC1CCC[NH2+]CC1NC(=O)c1ccc([*:1])cc1)=O >>[*:2]C(OC1Cc2c(cccc2)C1NC(=O)c1ccc([*:1])cc1)=O	22	6	6	51	43.1
[*:1]CCCC>>[*:1]C[NH+]1CCOCC1	6	4	6	14	42.9
[*:2]NC1CCCCCC1OC(=O)c1cc(O)c([*:1])c(O)c1 >>[*:2]NC1c2c(CC1OC(=O)c1cc(O)c([*:1])c(O)c1)cccc2	14	5	4	34	41.2
[*:1]c1ccc(cc1)C(C)C>>[*:1]c1ccccc1C	27	4	22	66	40.9
[*:1]C(=O)[O-]>>[*:1]NC(=O)C	10	8	5	25	40.0
[*:2]C(CCP(=O)([O-])[O-])C[*:1] >>[*:2]C(COP(=O)([O-])[O-])C[*:1]	8	5	5	20	40.0
[*:2]C(C(=O)N[*:1])C>>[*:2]C(CC(C)C)C(=O)N[*:1]	10	8	6	26	38.5
[*:1]N=[N+]=[N-]>>[*:1]O	5	4	5	13	38.5
[*:1]Br>>[*:1]C(=O)[O-]	8	8	6	21	38.1
[*:2]C([*:1])CCCC>>[*:2]C[*:1]	6	5	5	16	37.5
[*:1]CCCC[NH3+]>>[*:1]Cc1ccccc1	7	4	7	19	36.8
[*:2]C(=O)NC([*:1])CC(C)C>>[*:2]C(=O)NC[*:1]	7	5	4	19	36.8
[*:1]CC(C)C>>[*:1]CCCCOCc1ccccc1	6	4	4	17	35.3
[*:1]CC(C)C>>[*:1]Cc1c2c(ccc1)cccc2	6	5	4	17	35.3

Table E-1: Activity cliff-introducing chemical replacements (continued)

SMIRKS	#cliffs	#targets	#MMPs	#records	freq.
[*:2]C[*:1]>>[*:2]c1ccccc1[*:1]	27	4	27	77	35.1
[*:2]C(C(C)C)C(=O)N[*:1]>>[*:2]CC(=O)N[*:1]	9	8	6	26	34.6
[*:1]c1cc(OC)c(OC)cc1>>[*:1]c1ccccc1OC	5	4	4	15	33.3
[*:2]C(=O)C([*:1])=O>>[*:2]C([*:1])=O	6	4	4	18	33.3
[*:2]c1ccc([*:1])cc1>>[*:2]c1sc([*:1])nn1	17	4	7	51	33.3
[*:2]C=CC[*:1]>>[*:2]CC[*:1]	12	4	6	36	33.3
[*:3][S+2]([*:2])([*:1])c1ccc(OC)cc1 >>[*:3][S+2]([*:2])([*:1])c1ccc(Oc2ccccc2)cc1	16	6	7	49	32.7
[*:3]C(NC(=O)C([*:2])[*:1])C>>[*:3]C1N(CCC1)C(=O)C([*:2])[*:1]	9	5	8	29	31.0
[*:2]C(=O)C(NC([*:1])=O)Cc1ccc(O)cc1 >>[*:2]C(=O)C(NC([*:1])=O)Cc1nc[nH]c1	19	4	14	62	30.6
[*:2]C(NC(=O)C([*:1])[NH3+])CCCC >>[*:2]C1N(CCC1)C(=O)C([*:1])[NH3+]	7	4	6	23	30.4
[*:1]C>>[*:1]c1cc(Cl)ccc1	7	7	7	24	29.2
[*:1]C(C)C>>[*:1]O	6	4	5	22	27.3
[*:2]C([*:1])O>>[*:2]O[*:1]	7	4	4	26	26.9
[*:1]C(C)C>>[*:1]C[NH3+]	4	4	4	15	26.7
[*:2]C([*:1])Cc1ccccc1>>[*:2]C[*:1]	17	10	15	64	26.6
[*:1]C(C)C>>[*:1]CC[NH3+]	9	6	6	34	26.5
[*:1]C>>[*:1]CCCCCCCC	9	6	7	34	26.5
[*:1]O>>[*:1]OC(=O)NC	6	5	5	23	26.1
[*:1]N>>[*:1]Nc1ccccc1	9	8	7	37	24.3
[*:1]S(=O)(=O)N>>[*:1][N+](=O)[O-]	8	7	4	33	24.2
[*:1]N>>[*:1]NCC[NH+](C)C	4	4	4	17	23.5
[*:1]Br>>[*:1]C(C)(C)C	7	5	4	30	23.3
[*:1]C(=O)[O-]>>[*:1]P(=O)([O-])[O-]	9	7	7	39	23.1
[*:1]c1ccc(Cl)cc1>>[*:1]c1ccc(S(=O)(=O)C)cc1	5	5	4	22	22.7
[*:1]c1ccc(SC)cc1>>[*:1]c1ccccc1	5	4	4	22	22.7
[*:2]C([*:1])C>>[*:2]C1([*:1])CCCC1	6	4	5	27	22.2
[*:2]C[*:1]>>[*:2]Cc1ccccc1[*:1]	5	4	4	23	21.7
[*:1]C>>[*:1]CCC(C)C	8	4	8	37	21.6
[*:1]C(=O)[O-]>>[*:1]C(F)(F)F	6	6	5	28	21.4
[*:1]c1cc[nH+]cc1>>[*:1]c1ncccc1	4	4	4	19	21.1
[*:1]C(=O)[O-]>>[*:1]CC	8	8	4	38	21.1
[*:1]C(=O)[O-]>>[*:1]F	9	9	8	43	20.9
[*:1]C(=O)c1ccccc1>>[*:1]c1ccccc1	7	4	5	36	19.4

Table E-1: Activity cliff-introducing chemical replacements (continued)

SMIRKS	#cliffs	#targets	#MMPs	#records	freq.
[*:1]C(F)(F)F>>[*:1]NC(=O)C	5	4	4	26	19.2
[*:1]C(C)C>>[*:1]c1ccc(Cl)cc1	5	4	5	26	19.2
[*:2]C(=O)N[*:1]>>[*:2][NH2+][*:1]	9	6	6	47	19.1
[*:2]C(=O)C([*:1])C>>[*:2]C(=O)C([*:1])Cc1ccccc1	12	6	10	64	18.8
[*:2]C(=O)CCCCCC[*:1]>>[*:2]C([*:1])=O	5	5	4	27	18.5
[*:1]F>>[*:1]S(=O)(=O)C	10	7	9	55	18.2
[*:1]CC>>[*:1]N	4	4	4	22	18.2
[*:2]O[*:1]>>[*:2]S([*:1])(=O)=O	12	11	9	66	18.2
[*:1]C>>[*:1]CCCCCCC	11	6	9	63	17.5
[*:1]NC>>[*:1]Nc1ccccc1	12	5	11	69	17.4
[*:1]NCc1ccccc1>>[*:1][O-]	4	4	4	23	17.4
[*:1]N>>[*:1]c1ccccc1	5	5	4	29	17.2
[*:1]N>>[*:1]NCc1ccccc1	6	6	6	36	16.7
Ic1ccc([*:1])cc1>>[*:1]c1ccccc1	11	4	10	66	16.7
[*:1]C>>[*:1]C[NH+]1CCOCC1	9	6	7	55	16.4
[*:1]C(=O)[O-]>>[*:1]c1ccccc1	8	7	5	49	16.3
[*:2]C(=O)N([*:1])C>>[*:2]C([*:1])=O	5	5	4	31	16.1
[*:2]C1CC[NH2+]CC1[*:1]>>[*:2]C1C[NH2+]CC1[*:1]	9	6	4	56	16.1
[*:2]CC(=O)N[*:1]>>[*:2]C[*:1]	5	5	4	31	16.1
[*:1]c1ccccc1OC>>[*:1]c1ncccc1	6	6	5	38	15.8
[*:1]C(=O)[O-]>>[*:1]Cl	8	8	7	51	15.7
[*:2]C(=O)C([*:1])C>>[*:2]C(=O)C([*:1])C(C)C	8	4	8	51	15.7
[*:1]OCc1ccccc1>>[*:1]Oc1ccccc1	11	5	8	70	15.7
[*:1]C#N>>[*:1]NC(=O)C	4	4	4	26	15.4
[*:1]C(=O)[O-]>>[*:1]OC	8	8	6	54	14.8
[*:2]CCCCCCCC[*:1]>>[*:2]C[*:1]	5	4	5	35	14.3
[*:1]c1ccc(Oc2ccccc2)cc1>>[*:1]c1ccc(cc1)-c1ccccc1	6	5	5	42	14.3
[*:1]C(F)(F)F>>[*:1]N	6	5	5	42	14.3
[*:2]CCCCCCCCCC[*:1]>>[*:2]CC[*:1]	6	6	6	43	14.0
[*:2]C(=O)NC([*:1])CC(C)C>>[*:2]C(=O)NC([*:1])Cc1ccccc1	6	5	5	43	14.0
[*:1]C(=O)[O-]>>[*:1]O	9	8	8	65	13.8
[*:1]C>>[*:1]CC(F)(F)F	8	6	7	59	13.6
[*:2]C[*:1]>>[*:2]Cc1ccc([*:1])cc1	9	7	9	66	13.6
[*:2]C([*:1])=O>>[*:2]C([*:1])O	13	11	11	96	13.5
[*:1]O>>[*:1]c1ccccc1	6	5	5	45	13.3

Table E-1: Activity cliff-introducing chemical replacements (continued)

SMIRKS	#cliffs	#targets	#MMPs	#records	freq.
[*:2]C[*:1]>>[*:2]c1cc([*:1])ccc1	4	4	4	30	13.3
[*:2]C(=O)CCCCC[*:1]>>[*:2]C([*:1])=O	5	4	5	38	13.2
[*:1]C#N>>[*:1]N	5	5	5	38	13.2
[*:2]C(=O)N[*:1]>>[*:2]C([*:1])=O	33	13	23	252	13.1
[*:1]C>>[*:1][O-]	7	4	6	54	13.0
[*:1]Cc1cc2c(cc1)cccc2>>[*:1]Cc1ccccc1	16	8	11	124	12.9
[*:1]OC>>[*:1]OCc1ccccc1	15	8	11	116	12.9
[*:1]C>>[*:1]CC=C	10	5	7	78	12.8
[*:2]CCCc1ccc([*:1])cc1>>[*:2]CCc1ccc([*:1])cc1	9	4	6	71	12.7
[*:1]N>>[*:1][O-]	18	14	14	142	12.7
[*:1]C#N>>[*:1]C(=O)[O-]	7	7	5	55	12.7
[*:2]c1cc([*:1])c(OC)cc1>>[*:2]c1ccccc1[*:1]	7	6	5	55	12.7
[*:3]C(=O)C([*:2])([*:1])C>>[*:3]C(=O)C([*:2])[*:1]	12	7	12	95	12.6
[*:1]c1ccc(NC(=O)C)cc1>>[*:1]c1ccc(OC)cc1	4	4	4	32	12.5
[*:2]CCCCCCCC[*:1]>>[*:2]CC[*:1]	10	7	9	81	12.3
[*:2]C([*:1])=O>>[*:2]C[*:1]	22	15	22	184	12.0
[*:2]C(=O)C[*:1]>>[*:2]C[*:1]	6	4	6	50	12.0
[*:1]Cc1ccccc1>>[*:1]Cc1nc[nH]c1	5	4	5	42	11.9
[*:1]C>>[*:1]C(=O)[O-]	8	8	6	67	11.9
[*:1]CO>>[*:1]C[NH3+]	13	9	8	109	11.9
[*:2]N([*:1])C>>[*:2]N[*:1]	47	26	43	400	11.8
[*:1]Cc1ccc(cc1)C(F)(F)F>>[*:1]Cc1ccccc1	5	5	5	43	11.6
[*:1]C>>[*:1]CC1CCCCC1	5	4	5	43	11.6
[*:1]C[NH3+]>>[*:1]O	7	6	4	61	11.5
[*:1]Cc1ccccc1>>[*:1]Oc1ccccc1	6	6	5	52	11.5
[*:1]Cc1ccccc1>>[*:1]c1c2c(ccc1)cccc2	5	5	4	44	11.4
[*:2]c1cc(C)c([*:1])cc1>>[*:2]c1ccccc1[*:1]	5	4	4	44	11.4
[*:1]c1cc(Cl)ccc1>>[*:1]c1ccncc1	4	4	4	35	11.4
[*:2]C(=O)N[*:1]>>[*:2]S([*:1])(=O)=O	8	5	7	71	11.3
[*:1]C(=O)C>>[*:1]c1ccccc1	4	4	4	36	11.1
[*:1]C>>[*:1]C(=O)C	9	4	9	81	11.1
[*:1]c1cc(ccc1)C(F)(F)F>>[*:1]c1ccccc1OC	5	5	5	45	11.1
[*:2]C(=O)N[*:1]>>[*:2]N[*:1]	5	4	4	46	10.9
[*:2]C(O[*:1])=O>>[*:2]CC(O[*:1])=O	7	4	7	64	10.9
[*:1]C>>[*:1]CC[NH3+]	8	6	8	74	10.8

Table E-1: Activity cliff-introducing chemical replacements (continued)

SMIRKS	#cliffs	#targets	#MMPs	#records	freq.
[*:2]c1cc(Cl)ccc1[*:1]>>[*:2]c1cc([*:1])ccc1	8	4	8	74	10.8
[*:2]CN[*:1]>>[*:2]N([*:1])C	5	4	5	47	10.6
[*:1]c1ccc(cc1)C(F)(F)F>>[*:1]c1ccccc1C	6	6	6	57	10.5
[*:2]C([*:1])CC>>[*:2]C[*:1]	10	8	9	97	10.3
[*:2]Cc1cc([*:1])ccc1>>[*:2]c1cc([*:1])ccc1	6	6	6	58	10.3
[*:2]C(=O)C([*:1])C>>[*:2]C(=O)C[*:1]	20	5	20	195	10.3
[*:1]N>>[*:1][N+](=O)[O-]	7	5	7	68	10.3
[*:1]F>>[*:1]c1ccccc1	9	8	8	87	10.3
[*:2]CS[*:1]>>[*:2]S[*:1]	6	4	5	59	10.2
[*:2]c1cc(C)c([*:1])cc1>>[*:2]c1cc(Cl)c([*:1])cc1	5	4	5	49	10.2
[*:2]C([*:1])(C)C>>[*:2]O[*:1]	6	6	6	59	10.2
[*:1]N>>[*:1]NCC	5	5	5	49	10.2
[*:2]C[NH2+][*:1]>>[*:2][NH2+][*:1]	19	6	17	187	10.2
[*:2]C([*:1])c1ccccc1>>[*:2]C[*:1]	8	6	7	79	10.1
[*:1]C>>[*:1]C[NH3+]	7	6	7	70	10.0
[*:2]C(=O)Nc1ccc([*:1])cc1>>[*:2]c1ccc([*:1])cc1	4	4	4	40	10.0

SMIRKS representations for all 146 frequent cliff formers are provided. For each transformation, the number of cliffs ("#cliffs") in which it participates, the number of cliff-forming matched molecular pairs defined by this transformation ("#MMPs"), the number of targets ("#targets") for which this transformation forms cliffs, the frequency of occurrence of the transformation over all ligand sets ("#records"), and the relative frequency (percentage) with which activity cliffs are introduced ("freq.") are recorded. Transformations are sorted in descending order of the relative frequency with which they introduce activity cliffs. The number of cliffs reported is often larger than the number of matched molecular pairs because the same compound pair can form activity cliffs for multiple targets.

Table E-2: Directionality of potency changes for individual targets

#scaffold pairs	consensus	no consensus
1	147	13
2	56	18
3	9	5
4	5	1
5	1	1
6	1	0

For those transformations defining multiple cliff-forming MMPs for a single target, the number of different scaffold pairs represented by these compound pairs was extracted. Then, for all compound pairs, the direction of the activity change following the molecular transformation was recorded. If the direction was the same for all compound pairs, it was considered a "consensus" direction for this transformation and ligand set. For different numbers of scaffold pairs present in the cliff-forming ligand sets, it is reported how often a consensus or no consensus was obtained, respectively.

Table E-3: Bioisosteric replacements

SMIRKS	#records	#MMPs	freq.	#targets	#TF
[*:1]Nc1ccc(F)cc1>>[*:1]Nc1ccccc1	56	27	100.0	21	8
[*:1]C(C)C>>[*:1]CC1CC1	39	25	100.0	29	13
[*:1]C1CCCC1>>[*:1]CC(C)C	35	22	100.0	22	12
[*:1]Nc1cc(OC)ccc1>>[*:1]Nc1ccccc1	35	19	100.0	19	9
[*:2]C1CC[NH+]([*:1])CC1>> [*:2][NH+]1CC[NH+]([*:1])CC1	31	25	100.0	11	7
[*:1]c1cc(Cl)ccc1>>[*:1]c1cc(O)ccc1	30	20	100.0	28	16
[*:1]S(=O)(=O)c1ccc(F)cc1 >>[*:1]S(=O)(=O)c1ccccc1	30	17	100.0	20	11
[*:2]CCCO[*:1]>>[*:2]CCC[*:1]	102	94	98.0	17	11
[*:1]NC(=O)CC>>[*:1]NC(=O)CCC	48	28	97.9	15	7
[*:2]S(=O)(=O)c1cc([*:1])ccc1>> [*:2]S(=O)(=O)c1ccc([*:1])cc1	88	57	97.7	43	19
[*:1]Nc1ccc(cc1)C>>[*:1]Nc1ccccc1	38	18	97.4	15	7
[*:2]CC(=O)NCC[*:1]>>[*:2]CC(=O)N[*:1]	37	23	97.3	23	12
[*:1]c1cc(F)ccc1>>[*:1]c1sccc1	36	21	97.2	30	12
[*:2]Oc1ccc([*:1])cc1>>[*:2]Sc1ccc([*:1])cc1	35	25	97.1	25	15
[*:2]C(OCC[*:1])=O>>[*:2]C(O[*:1])=O	35	25	97.1	15	9
[*:1]Nc1ccc(Cl)cc1>>[*:1]Nc1ccc(F)cc1	35	16	97.1	13	5
[*:1]Cc1cc(F)ccc1>>[*:1]Cc1ccc(F)cc1	34	28	97.1	30	16
[*:1]C[NH+](CC)CC>>[*:1]C[NH+]1CCCC1	34	23	97.1	26	10

Table E-3: Bioisosteric replacements (continued)

SMIRKS	#records	#MMPs	freq.	#targets	#TF
[*:1]CC>>[*:1]CC=C	34	24	97.1	21	14
[*:2]C([*:1])F>>[*:2]C[*:1]	34	27	97.1	17	8
[*:1]c1cc(OC)ccc1>>[*:1]c1cc2OCOc2cc1	34	15	97.1	17	7
[*:1]CC(C)C>>[*:1]CCCC	134	109	97.0	51	23
[*:2]COc1cc([*:1])ccc1>>[*:2]COc1ccc([*:1])cc1	33	31	97.0	21	14
[*:2]Nc1cc(O[*:1])ccc1>>[*:2]Nc1ccc([*:1])cc1	33	14	97.0	12	6
[*:1]CF>>[*:1]CO	32	26	96.9	10	7
[*:1]c1ccc(Br)cc1>>[*:1]c1ccc(cc1)C(F)(F)F	62	44	96.8	29	18
[*:1]Cc1ccc(Cl)cc1>>[*:1]Cc1ccc(cc1)C	31	20	96.8	27	15
[*:1]c1cc(ccc1)C>>[*:1]c1sccc1	31	16	96.8	23	10
I[*:1]>>[*:1][N+](=O)[O-]	31	24	96.8	21	13
[*:1]c1cc(OC)ccc1>>[*:1]c1sccc1	30	21	96.7	27	15
[*:1]c1cc(F)ccc1>>[*:1]c1ccccc1	200	148	96.5	89	34
[*:1]CC(C)C>>[*:1]CCC	170	122	96.5	55	23
[*:1]c1cc(Cl)ccc1>>[*:1]c1cc(ccc1)C	83	57	96.4	48	17
[*:1]C1CCCC1>>[*:1]CCC	49	31	95.9	30	14
[*:2]Oc1ccccc1[*:1]>>[*:2]c1ccccc1[*:1]	96	65	95.8	50	19
[*:2]C(C[*:1])(C)C>>[*:2]C(C[*:1])C	48	34	95.8	32	19
[*:1]CCCCC>>[*:1]CCCCCCC	71	58	95.8	35	20
[*:1]Br>>[*:1]Cl	439	282	95.7	132	48
[*:1]CC>>[*:1]OCC	46	34	95.7	35	21
[*:2]c1cc(C(F)(F)F)c([*:1])cc1 >>[*:2]c1cc(Cl)c([*:1])cc1	45	35	95.6	14	5
[*:2]CCCOc1ccc([*:1])cc1 >>[*:2]CCOc1ccc([*:1])cc1	44	31	95.5	20	11
[*:2]C(=O)N1CCC(CC1)C[*:1] >>[*:2]C(=O)N1CCC([*:1])CC1	44	32	95.5	9	5
[*:1]Nc1ccc(Cl)cc1>>[*:1]Nc1ccccc1	42	26	95.2	18	10
[*:1]c1ccsc1>>[*:1]c1sccc1	61	38	95.1	30	17
[*:2]C=CC[*:1]>>[*:2]CCC[*:1]	81	41	95.1	24	13
[*:1]c1cc(OC)c(OC)cc1>>[*:1]c1ccccc1	40	36	95.0	29	15
[*:1]c1ccccc1F>>[*:1]c1sccc1	40	23	95.0	27	14
[*:1]C(C)C>>[*:1]C1CC1	98	76	94.9	45	19
[*:1]S(=O)(=O)c1ccc(Cl)cc1 >>[*:1]S(=O)(=O)c1ccccc1	39	27	94.9	22	14
[*:2]Nc1cc(O[*:1])ccc1>>[*:2]Nc1ccc(O[*:1])cc1	57	35	94.7	26	14

Table E-3: Bioisosteric replacements (continued)

SMIRKS	#records	#MMPs	freq.	#targets	#TF
[*:2]CCCCCCCCC[*:1]>>[*:2]CCCCCC[*:1]	38	26	94.7	17	8
[*:1]CC[NH+](C)C>>[*:1]C[NH+](C)C	56	36	94.6	32	10
I[*:1]>>[*:1]OC	37	24	94.6	30	16
[*:2]Oc1cc([*:1])ccc1>>[*:2]c1cc([*:1])ccc1	91	65	94.5	55	25
[*:1]C(=O)c1ccc(cc1)C>>[*:1]C(=O)c1ccccc1	36	22	94.4	27	14
[*:1]Nc1cc(Cl)c(Cl)cc1>>[*:1]Nc1ccc(Cl)cc1	36	16	94.4	9	3
[*:1]c1cc(F)ccc1>>[*:1]c1ccccc1F	107	80	94.4	65	27
[*:2]c1cc(Cl)c([*:1])cc1>>[*:2]c1cc(F)c([*:1])cc1	89	57	94.4	39	16
[*:2]C(F)(F)O[*:1]>>[*:2]C([*:1])(F)F	89	65	94.4	34	18
[*:1]c1cc(ccc1)C>>[*:1]c1ccccc1OC	35	19	94.3	27	12
[*:1]c1cc(F)ccc1>>[*:1]c1ccsc1	35	23	94.3	24	12
[*:2]C([*:1])C(C)C>>[*:2]C([*:1])CC	35	23	94.3	23	14
[*:2]Nc1cc([*:1])c(Cl)cc1>>[*:2]Nc1cc([*:1])ccc1	35	15	94.3	13	4
I[*:1]>>[*:1]Br	103	51	94.2	53	21
[*:3]C([*:1])c1cc([*:2])ccc1 >>[*:3]C([*:1])c1ccccc1[*:2]	34	17	94.1	17	6
[*:2]C#CCC[*:1]>>[*:2]C#C[*:1]	34	15	94.1	12	5
[*:2]CCCCCCCC[*:1]>>[*:2]CCCCCC[*:1]	101	73	94.1	33	18
[*:1]CCC>>[*:1]CCCC	437	289	94.1	108	45
[*:1]c1ccc(Br)cc1 >>[*:1]c1ccc([N+](=O)[O-])cc1	50	24	94.0	20	9
[*:1]N(C)C>>[*:1]OC	66	54	93.9	42	21
[*:1]c1ccccc1F>>[*:1]c1ccccc1OC	66	43	93.9	41	18
[*:1]Cc1cccnc1>>[*:1]Cc1ccncc1	33	26	93.9	24	14
[*:2]CC(=O)NCC[*:1]>>[*:2]CCC(=O)NCC[*:1]	33	21	93.9	14	6
[*:2]c1c(F)cccc1[*:1]>>[*:2]c1ccccc1[*:1]	49	43	93.9	31	20
[*:1]CCc1ccccc1>>[*:1]Cc1ccc(F)cc1	49	28	93.9	30	16
[*:2]C(CC[*:1])C>>[*:2]CCC[*:1]	49	30	93.9	29	17
[*:1]C[NH+]1CCCC1>>[*:1]C[NH+]1CCOCC1	48	34	93.8	30	12
[*:2]C[NH+]([*:1])CC>>[*:2][NH+]([*:1])C	48	33	93.8	30	11
[*:1]S(=O)(=O)c1ccc(cc1)C >>[*:1]S(=O)(=O)c1ccccc1	32	24	93.8	22	14
[*:2]C([NH2+]C[*:1])C>>[*:2]C[NH2+]C[*:1]	32	25	93.8	11	6
[*:2]CCC[NH+]([*:1])C>>[*:2]CC[NH+]([*:1])C	63	32	93.7	34	12
[*:2]CNC(=O)N[*:1]>>[*:2]NC(=O)N[*:1]	47	41	93.6	14	6
[*:2]CNS([*:1])(=O)=O>>[*:2]NS([*:1])(=O)=O	78	47	93.6	24	8

Table E-3: Bioisosteric replacements (continued)

SMIRKS	#records	#MMPs	freq.	#targets	#TF
[*:1]c1cc(F)ccc1>>[*:1]c1ccccc1OC	31	24	93.6	27	16
[*:2]c1cc(Cl)c([*:1])cc1>>[*:2]c1cc([*:1])c(Cl)cc1	31	25	93.6	22	18
[*:1]C=C>>[*:1]CO	31	17	93.6	21	12
[*:1]c1cc(Cl)ccc1>>[*:1]c1cc(F)cc(F)c1	31	27	93.6	20	12
[*:2]c1cc([*:1])c(F)cc1>>[*:2]c1cc([*:1])c(OC)cc1	31	19	93.6	15	7
[*:2]CCCCCCO[*:1]>>[*:2]CCCO[*:1]	31	20	93.6	11	7
[*:2]c1cc(F)c([*:1])cc1>>[*:2]c1ccc([*:1])cc1	262	186	93.5	76	34
[*:1]CCCc1ccccc1>>[*:1]CCc1ccccc1	92	57	93.5	43	23
[*:1]C1CCC1>>[*:1]C1CCCC1	46	32	93.5	30	15
[*:1]C(=O)C>>[*:1]C(OC)=O	46	38	93.5	29	17
[*:1]c1cc(Cl)c(Cl)cc1>>[*:1]c1ccc(Br)cc1	60	31	93.3	25	13
[*:2]Cc1cc(O[*:1])ccc1>>[*:2]Cc1ccccc1O[*:1]	30	23	93.3	25	13
[*:1]CCC[NH+]1CCOCC1 >>[*:1]CC[NH+]1CCOCC1	30	22	93.3	14	7

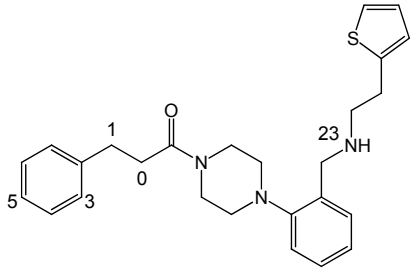
SMIRKS representations for all 96 bioisosteric replacements are provided. For each transformation, the number of its potency records ("#records"), the number of matched molecular pairs ("#MMPs") defined by this transformation, the percentage of potency records smaller than one order of magnitude ("freq."), as well as the number of targets ("#targets") and target families ("#TF") for which this transformation occurs in active compounds are recorded. Transformations are sorted in descending order of the relative frequency with which they produce potency records smaller than one order of magnitude. The number of potency records is often larger than the number of matched molecular pairs because the same compound pair can be found in ligand sets of multiple targets.

Appendix F

R-Group Table

For the analog series of 32 antagonists of the melanocortin receptor 4 discussed in Chapter 8, the common core structure and substitution sites are provided in a conventional R-group table format (Table F-1). For all individual analogs, R-groups and potency values are reported.

Table F-1: SAR table



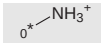
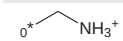
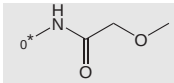
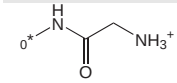
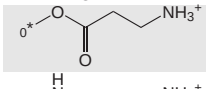
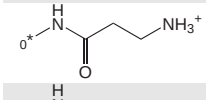
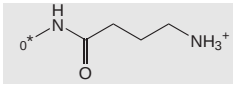
R0	R1	R3	R5	R23	pK_i
					6.0
 (R)					6.2
					5.2
 (R)					7.2
 (R)					7.4
					6.5
 (R)					7.7
 (R)					7.9

Table F-1: SAR table (continued)

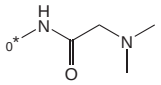
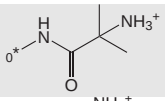
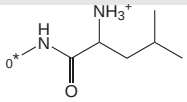
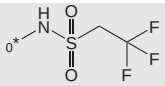
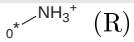
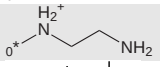
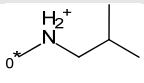
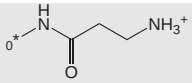
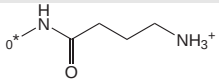
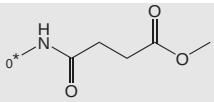
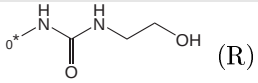
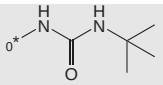
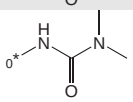
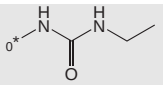
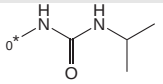
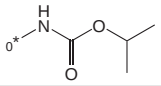
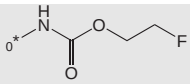
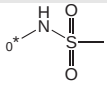
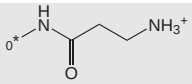
R0	R1	R3	R5	R23	pK_i
					7.4
					7.2
					7.5
					6.4
					7.0
					7.7
					7.6
					8.7
					8.7
					8.6
					8.5
					8.6
					8.7
					8.8
					8.9
					8.6
					8.3
					8.3

Table F-1: SAR table (continued)

R0	R1	R3	R5	R23	pK_i
 (R)					8.3
 (R)					8.5
 (R)					7.9
 (R)					6.7
					5.3
					6.0

For a subset of analogs, the stereocenter at substitution site 0 is in the R-configuration, as indicated in the table (R).