

Systematic Computational Analysis of Structure–Activity Relationships

Dissertation

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

vorgelegt von
LISA BERTHA PELTASON
aus Ulm/Donau

Bonn, 2009

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

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

2. Referent: Univ.-Prof. Dr. rer. nat. Christa E. Müller

Tag der Promotion: 09.04.2010

Erscheinungsjahr: 2010

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn unter
http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Abstract

The exploration of structure–activity relationships (SARs) of small bioactive molecules is a central task in medicinal chemistry. Typically, SARs are analyzed on a case-by-case basis for series of closely related molecules. Classical methods that explore SARs include quantitative SAR (QSAR) modeling and molecular similarity analysis. These methods conceptually rely on the similarity–property principle which states that similar molecules should also have similar biological activity. Although this principle is intuitive and supported by a wealth of observations, it is well-recognized that SARs can have fundamentally different character. Small chemical modifications of active molecules often dramatically alter biological responses, giving rise to “activity cliffs” and “discontinuous” SARs. By contrast, structurally diverse molecules can have similar activity, a situation that is indicative of “continuous” SARs. The combination of continuous and discontinuous components characterizes “heterogeneous” SARs, a phenotype that is frequently encountered in medicinal chemistry.

This thesis focuses on the systematic computational analysis of SARs present in sets of active molecules. Approaches to quantitatively describe, classify, and compare SARs at multiple levels of detail are introduced. Initially, a comparative study of crystallographic enzyme–inhibitor complexes is presented that relates two-dimensional and three-dimensional inhibitor similarity and potency to each other. The analysis reveals the presence of systematic and in part unexpected relationships between molecular similarity and potency and explains why apparently inconsistent SARs can coexist in compound activity classes. For the systematic characterization of complex SARs, a numerical function termed SAR Index (SARI) is developed that quantitatively describes continuous and discontinuous SAR components present in sets of active molecules. On the basis of two-dimensional molecular similarity and potency, SARI distinguishes between the three basic SAR categories described above. Heterogeneous SARs are further divided into two previously unobserved subtypes that are distinguished by the way they combine different SAR features. SARI profiling of various enzyme inhibitor classes demonstrates the prevalence of heterogeneous SARs for many classes. Furthermore, control calculations are conducted in order to assess the influence of molecular representation and data set size on SARI scoring. It is shown that SARI scores remain largely stable in response to variation of these critical parameters.

Based on the SARI formalism, a methodology is developed to study multiple global and local SAR components of compound activity classes. The approach combines graphical analysis of Network-like Similarity Graphs (NSGs) and SARI score calculations at multiple levels of detail. Compound classes of different global SAR character are found to produce distinct network topologies. Local SAR features are studied in subsets of similar compounds and

systematically related to global SAR character. Furthermore, key compounds are identified that are major determinants of local and global SAR characteristics. The approach is also applied to study structure–selectivity relationships (SSRs). Compound selectivity often results from potency differences for multiple targets and presents a critical factor in lead optimization projects. Here, SSRs are explored for sets of compounds that are active against pairs of related targets. For this purpose, the molecular network approach is adapted to the evaluation of SSRs. Results show that SSRs can be quantitatively described and categorized in analogy to single-target SARs. In addition, local SSR environments are identified and compared to SAR features. Within these environments, key compounds are identified that determine characteristic features of single-target SARs and dual-target SSRs. Comparison of similar compounds that have significantly different selectivity reveals chemical modifications that render compounds target-selective.

Furthermore, a methodology is introduced to study SAR contributions from functional groups and substitution sites in series of analogous molecules. Analog series are systematically organized according to substitution sites in a hierarchical data structure termed Combinatorial Analog Graph (CAG), and the SARI scoring scheme is applied to evaluate SAR contributions of variable functional groups at specific substitution sites. Combinations of sites that determine SARs within analog series and make large contributions to SAR discontinuity are identified. These sites are prime targets for further chemical modification. In addition to determining key substitution patterns, CAG analysis also identifies substitution sites that have not been thoroughly explored.

Für meine Familie.

Acknowledgments

I would like to take the opportunity and thank the persons who accompanied me during the work on this dissertation project and contributed to its completion in many different ways.

I have been fortunate to participate in an excellent working group, with a dedicated supervisor and great colleagues. To Prof. Dr. Jürgen Bajorath, I would like to express my honest gratitude for his invaluable guidance and his continuous scientific and personal support. Discussions with him have always motivated and inspired me and provided the fundamental basis for the success of this thesis. Sincere thanks go to Prof. Dr. Christa Müller for taking the time to act as co-referee. I would also like to thank our project partners from Boehringer–Ingelheim, Dr. Andreas Teckentrup and Dr. Nils Weskamp, for the successful collaboration. Many insightful suggestions and enjoyable meetings in Bonn and Biberach have substantially contributed to the progress of this work.

This thesis has also greatly benefited from the work with my colleagues. Special thanks are due to Mathias Wawer for his valuable scientific and creative support, patient advice and proof-reading on numerous occasions, and for his sense of humor. Pleasant collaborations with Ye Hu and Mihiret Tekeste Sisay have also advanced my scientific work. Finally, I would like to express my gratitude to my colleague and friend Dr. Hanna Geppert for her continuous encouragement and understanding, and to all my colleagues at the Life Science Informatics group for motivation, advice, and the good times we shared.

Contents

1	Introduction	1
2	Qualitative SAR Characterization	11
2.1	SARs and Target–Ligand Interactions	12
2.2	Molecular Similarity Assessment	12
2.2.1	2D Similarity Calculation	13
2.2.2	3D Similarity Calculation	14
2.3	Relationships between Similarity and Potency	17
2.3.1	Data and Calculations	17
2.3.2	Results	18
2.4	Summary and Conclusions	23
3	Quantitative SAR Description	27
3.1	SARI Methodology	28
3.1.1	Continuity Score	28
3.1.2	Discontinuity Score	29
3.1.3	Normalization	29
3.1.4	SARI Score	31
3.2	SAR Profiling	31
3.2.1	Data and Calculations	31
3.2.2	Results	34
3.2.3	Discussion	38
3.3	Control Calculations	39
3.3.1	Data Sets	39
3.3.2	Fingerprint Dependence	40
3.3.3	Influence of Compound Set Size	41
3.3.4	Discussion	42
3.4	Related Methods	43
3.5	Conclusions	44
4	Global and Local SAR Analysis	45
4.1	Methodology	46

4.1.1	Compound Clustering and Cluster Scoring	46
4.1.2	Compound Discontinuity Scores	46
4.1.3	Score Normalization	47
4.1.4	Network-like Similarity Graphs	47
4.2	Analysis of Network-like Similarity Graphs	48
4.2.1	Network Topology	49
4.2.2	SARs in Compound Clusters	54
4.2.3	Cluster SARs versus Global SARs	55
4.2.4	Compound Discontinuity and Key Compounds	56
4.2.5	Summary	58
4.3	Application to Screening Data Sets	59
4.4	Conclusions	61
5	Structure–Selectivity Relationship Analysis	63
5.1	Selectivity Data Sets	64
5.2	Potency and Selectivity NSGs	64
5.3	Selectivity NSG Analysis	66
5.3.1	Global SAR and SSR Features	66
5.3.2	Comparison of SAR and SSR Elements	67
5.3.3	Local SSR Environments	71
5.3.4	SAR and SSR Key Compounds	73
5.3.5	Selectivity Determinants	74
5.4	Conclusions	77
6	SAR Determinants in Analog Series	79
6.1	Methodology	80
6.1.1	Data Sets and Analog Series Identification	80
6.1.2	R-Group Decomposition	82
6.1.3	SAR Contributions from R-Groups	82
6.1.4	Combinatorial Analog Graphs	83
6.2	SAR Analysis in Analog Series	84
6.2.1	Interpretation of CAGs	84
6.2.2	SAR Hotspots	86
6.2.3	SAR Holes	88
6.3	SAR Determinants for Multiple Targets	92
6.4	Conclusions	95
7	Summary and Conclusions	97
	Bibliography	101
	A Software and Databases	107

B Enzyme–Inhibitor Complexes	111
C SAR Tables	113

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
AID	PubChem Assay Identifier
CAG	Combinatorial Analog Graph
cat	cathepsin
CID	PubChem Compound Identifier
HTS	High-Throughput Screening
IC ₅₀	half maximal Inhibitory Concentration
K _i	Inhibition Constant
MCS	Maximum Common Substructure
MDDR	MDL Drug Data Report
MOE	Molecular Operating Environment
NSG	Network-like Similarity Graph
PDB	Protein Data Bank
pIC ₅₀	negative decadic logarithm of IC ₅₀
pK _i	negative decadic logarithm of K _i
QSAR	Quantitative Structure–Activity Relationship
SAR	Structure–Activity Relationship
SARI	Structure–Activity Relationship Index
SSR	Structure–Selectivity Relationship
Tc	Tanimoto coefficient

Chapter 1

Introduction

It is a central paradigm in medicinal chemistry that molecules having similar structure should also share similar biological activity. This viewpoint has been articulated in 1990 by the “similarity–property principle” (Johnson and Maggiora, 1990) and continues to be widely accepted in the medicinal chemistry community. Specifically, this concept provides the basis for numerous established computational methods supporting the drug discovery process, including molecular similarity searching, compound library design, and quantitative structure–activity relationship (QSAR) modeling (Bajorath, 2001). Although this concept is intuitive and supported by a wealth of observations, medicinal chemists also know that small chemical modifications can render active molecules completely or nearly inactive or, by contrast, increase their potency dramatically (Kubinyi, 1998). Moreover, it has been shown that compounds that are similar to known active molecules are themselves far less frequently active than one might expect (Martin et al., 2002). This apparent inconsistency suggests that there must be fundamental differences in the nature of structure–activity relationships (SARs) characterizing different classes of active molecules (Eckert and Bajorath, 2007).

Understanding the relationship between chemical structure and biological activity of small molecules is a key challenge in medicinal and pharmaceutical research. The identification of novel active molecules and their systematic chemical optimization require the thorough exploration of the underlying SARs. Traditionally, SARs are studied on a case-by-case basis for series of closely related molecules. However, with the advance of high-throughput screening (HTS) technologies that generate ever growing amounts of biological data, computational approaches to SAR analysis gain increasing importance. Medicinal chemists are challenged to prioritize active molecules that are most promising for further exploration in hit-to-lead projects and have a high potential for chemical optimization. Systematic evaluation of the SAR features present in sets of active molecules could guide this process in a directed manner. These

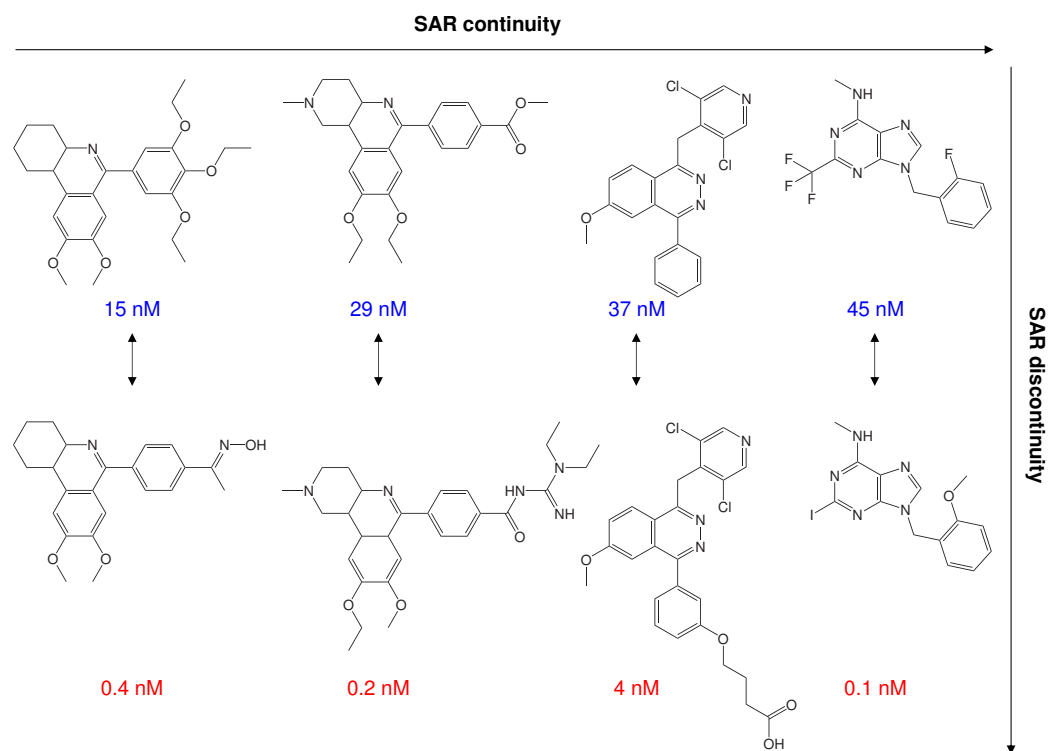


Figure 1.1: Heterogeneous SAR Inhibitors of phosphodiesterase IV are shown that combine continuous and discontinuous SAR features. At the top, nanomolar inhibitors of increasing structural diversity are displayed. The inhibitors belong to different chemotypes but display only gradual potency differences, thus presenting a prime example of a continuous SAR. At the bottom, close analogs to each of these compounds are shown that display a notable increase in potency and hence cause considerable SAR discontinuity. Combination of continuous and discontinuous SAR features within a data set characterizes the heterogeneous SAR phenotype.

considerations have motivated the development of methods to systematically classify and compare SARs, which are presented in this dissertation.

The Nature of Structure–Activity Relationships

SARs are essentially distinguished by the way active compounds respond to chemical alterations. Depending on the types of molecules under investigation, the magnitude of biological responses can vary considerably. Structural modifications of active molecules can be accompanied by only small or moderate changes in potency. In such cases, the underlying SAR is “continuous” in nature. In the presence of continuous SARs, similar molecules display comparable activity. Furthermore, structural departures from a known active compound may result in gradual potency changes, giving rise to a spectrum of increasingly diverse structures having similar activity and often a relatively narrow potency

distribution. Accordingly, a hallmark of continuous SARs is the presence of different chemotypes sharing the same biological activity. This SAR type is consistent with the similarity–property principle and presents a prerequisite for the successful application of whole-molecule similarity methods that aim at the identification of novel structural motifs having a specific biological activity (often referred to as “scaffold hopping”; Schneider et al., 2006). By contrast, large-magnitude biological responses to minor chemical changes are characteristic of “discontinuous” SARs. In the presence of this SAR type, a small chemical modification can dramatically alter the activity of a molecule. SAR discontinuity is thought to result from the presence or absence of structural patterns that are required for biological activity. Accordingly, the primary indicator of discontinuous SARs is the occurrence of “activity cliffs” marked by similar molecules having large differences in potency (Maggiora, 2006). In medicinal chemistry, this situation is exploited in lead optimization efforts where active compounds are modified in a systematic manner to achieve an increase in potency (Kubinyi, 1998). However, discontinuous SARs fall outside the scope of the similarity–property principle and greatly complicate molecular similarity analysis. In particular, in the vicinity of an activity cliff, structurally similar compounds might have distinctly different potency, which presents a major obstacle for any similarity method.

Importantly, continuous and discontinuous SAR types are not mutually exclusive because we frequently also observe that different structural classes share the same biological activity, but that close analogs within each class might have large differences in potency (Eckert and Bajorath, 2007). The corresponding SAR phenotype is termed “heterogeneous” because it combines continuous and discontinuous components. Figure 1.1 shows an exemplary compound set that illustrates the presence of different SAR phenotypes. Molecules of increasing structural diversity belonging to different chemical series are shown that retain nanomolar potency, which represents an exemplary continuous SAR. For each of these compounds, a close analog is found that provides a notable increase in potency and thus reflects SAR discontinuity. The combination of such continuous and discontinuous SAR elements characterizes heterogeneous SARs, which are of practical importance for medicinal chemistry because they provide the opportunity to identify diverse active molecules (in continuous SAR regions) and subsequently optimize them (by exploring activity cliffs). In essence, the continuous, discontinuous, and heterogeneous SAR categories define the spectrum of small-molecule SARs one encounters in medicinal chemistry.

Activity Landscapes

In order to understand these SAR phenotypes and rationalize SAR information, it is generally required to relate compound similarity and potency to each other.

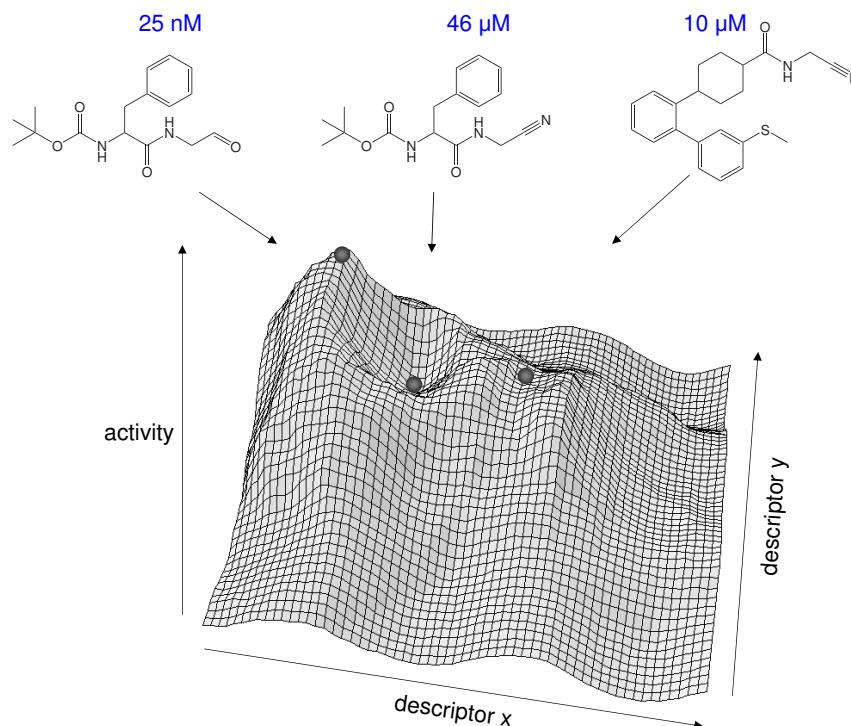


Figure 1.2: Hypothetical activity landscape Activity landscapes visualize the potency distribution of a set of active molecules projected into a two-dimensional chemical reference space. Shown are three exemplary cathepsin S inhibitors on a schematic representation of a hypothetical activity landscape that contains rugged and smooth regions. The two structures on the left and in the middle represent the same chemotype but have potency values that differ by several orders of magnitude, thus forming an activity cliff. By contrast, the structure on the right represents a distinct chemotype but has similar potency as the second structure, indicating a continuous SAR region.

Similarity and potency information can be combined in an activity landscape to conceptualize SAR characteristics, as illustrated in Figure 1.2. Models of activity landscapes can be envisioned as topological maps that project chemical compounds into a two-dimensional plane spanned by molecular descriptors and add compound potency as a third dimension (Maggiora, 2006). Hence, the xy -plane represents a projection of chemical space where data points represent active compounds and the distance between them is proportional to chemical dissimilarity. Thus, the further two compounds are apart in the chemical reference space, the more dissimilar they are. Potency is reported along the z -axis, producing a surface where elevated regions correspond to high potency levels. For different sets of active molecules, activity landscapes display specific topologies reminiscent of geographical landscapes that can directly be associated with the different SAR categories discussed above. For example, a gently

sloped activity landscape is produced by structurally diverse compounds having only small or moderate differences in potency, which is the characteristic feature of continuous SARs. By contrast, rugged landscapes are indicative of SAR discontinuity and are produced by compounds with significant potency differences. In this topology, activity cliffs are the most prominent feature, where small moves within the xy -plane are accompanied by a large change in z -direction. Finally, activity landscapes corresponding to heterogeneous SARs are characterized by gently sloped regions that are interspersed with activity cliffs. For the exploration of SARs, a major challenge is posed by the need to account for these variable regions within an activity landscape.

Traditional Computational SAR Analysis

In medicinal chemistry, SARs are traditionally explored on a case-by-case basis, evaluating individual series of related compounds to infer rules of how to modify a given chemotype and optimize its potency. This exercise typically involves iterative steps to select, modify, and test compounds and relies to a large extent on a medicinal chemist’s experience and intuition. The standard tool to support this process are SAR tables that report core structures, substituents, and biological activities of the studied compounds in a spreadsheet-like manner. SAR tables present a common concept in medicinal chemistry and also serve as a basis for combinatorial QSAR analysis. Recently, attempts have been made to enhance their design, for example by incorporating interactive functionality or combining them with additional representation types (Agrafiotis et al., 2007b). In addition, various computational tools have been developed for the graphical representation of property distributions in large compound data sets (Wawer and Bajorath, 2009). The spectrum of visualization techniques comprises basic types of diagrams such as histograms, scatter plots, or heat maps, as well as displays tailored toward the analysis of multifactorial data, like tree maps or radial clustergrams (Agrafiotis et al., 2007a; Kibbey and Calvet, 2005).

Given their mostly graphical nature, the methodologies described thus far are designed to structure and visualize SAR data, but they do not reveal any SAR information by themselves. Rather, they support the subjective derivation of SAR hypotheses by providing an intuitive access to the analysis of SAR features (Wawer and Bajorath, 2009). A step toward automation of this task has recently been taken by Birchall et al. (2006) who have attempted to extract chemically intuitive SAR rules from screening data through the development of reduced chemical graph queries using an evolutionary algorithm. This is one of the rare examples where the application of machine learning techniques yields interpretable SAR information, in contrast to the usual “black box” character of such methods.

For the aim of deriving and modeling quantitative SAR information, the

QSAR paradigm has become a cornerstone of computational medicinal chemistry (Esposito et al., 2004). QSAR analysis attempts to establish mathematical models that relate chemical structure (or deduced properties) to compound potency in a quantitative manner. The underlying hypothesis is that if such a numerical relationship can be established for sets of known active molecules, then the model can be applied to predict the potency of newly designed compounds. QSAR models can also be utilized as a guidance for compound modification and analog design. Originating from classical linear 2D QSAR, a variety of QSAR methodologies have been developed over the years, including 3D approaches (Kubinyi, 1997) and nonlinear extensions (Kubinyi, 1977; Manallack et al., 1994). Regardless of the conceptual design of different methods, QSAR models are essentially restricted to series of congeneric molecules. Hence, their ability to extrapolate from learning sets to test compounds that represent different chemotypes is generally limited. Furthermore, in order to successfully model an SAR, a continuous activity landscape is required, i.e. successive structural alterations of analogs should be accompanied by gradual changes in potency. The presence of activity cliffs, which characterizes many activity landscapes, cannot (or only inaccurately) be accounted for in QSAR models (Johnson, 2008). In addition, compounds representing activity cliffs are often considered statistical outliers and removed from the analysis, although actually they represent the most interesting compounds for lead optimization (Maggiore, 2006).

Besides quantitative approaches, a number of methods within the medicinal chemistry spectrum explores the relationship between molecular structure and biological activity in qualitative terms. Methods that focus on molecular similarity make use of SAR information that is implicitly encoded in molecular structure rather than trying to deduce explicit SAR rules. For example, in chemical similarity searching, known active molecules are taken as templates and compound databases are screened for similar compounds that are supposed to have similar biological activity, according to the similarity–property principle (Willett et al., 1998). Hence, in similarity analysis, it is of fundamental importance that chosen molecular representations be related to biological activity; in other words, that they display “neighborhood behavior” (Patterson et al., 1996). Different from whole-molecular similarity analysis, pharmacophore modeling investigates local similarity (Sheridan et al., 1989). Preliminary SAR information extracted from known active molecules is utilized to derive pharmacophore patterns that are likely to be responsible for biological activity. As discussed above, molecular similarity methods generally require the presence of continuous SARs and smooth activity landscapes; in rugged regions of an activity landscape, they are likely to fail.

All of the approaches discussed thus far have in common that they are designed to explore SARs on the basis of series of analogous or at least highly similar compounds. In addition, methods like QSAR modeling or similarity

searching rely on the presence of continuous SARs. Hence, the presented methods are capable of reflecting only a limited region of an activity landscape. This distinguishes them from the approaches introduced in this dissertation, which aim at the systematic assessment of SARs present in compound classes on a global scale.

Research Topics

The primary goal of this dissertation has been to develop approaches for the systematic assessment and comparison of structure–activity relationships within sets of active molecules. Established methods for the analysis of SARs traditionally focus on individual compound series and investigate SARs on a case-by-case basis; comparative studies that depart from this paradigm have until recently not been reported. However, qualitative evidence pointing at fundamental differences in the nature of small-molecule SARs is accumulating, emphasizing the need for approaches that are capable of detecting and unambiguously evaluating distinct SAR features.

In light of these considerations, an initial study has been designed to gain qualitative insights concerning the nature of SARs. Accounting for the fact that SARs are essentially the result of specific target–ligand interactions, the analysis focuses on crystallographic complex structures for well-established target enzymes. Systematic comparison of inhibitor similarity, binding modes, and potency reveals previously unobserved relationships and demonstrates the highly variable character of small-molecule SARs. These findings directly lead to the first central goal of this dissertation.

Goal 1: Design of a conceptual framework to systematically characterize and classify SARs present in sets of active molecules.

Following the qualitative characterization of SARs, the next step toward this goal attempts to put the assessment of different SARs on a formal basis. Therefore, a numerical function is developed that captures the elementary SAR features within sets of active molecules in a quantitative manner. Based on molecular similarity and potency data, this function implements a scoring scheme that distinguishes between three basic SAR categories and provides a framework for the classification and comparison of SARs within different compound activity classes. This study relies on a two-dimensional molecular representation, thereby departing from the target-centric view adopted in the initial analysis. This makes it possible to extend the analysis to a wide spectrum of activity classes for which no, or only few, relevant crystal structures are available.

Application of the SAR analysis function to various compound classes shows that different SAR elements can coexist within classes of specifically active

compounds. Thus, a second major goal of this dissertation is to study local SAR features associated with individual compounds or compound series within an activity class.

Goal 2: Development of a methodology to explore SARs at multiple levels of detail that enables the investigation of local SAR features and relationships between global and local SARs.

In order to extend the quantitative SAR analysis to the level of compound series, we divide activity classes into subsets of similar molecules that provide the basis for the analysis of local SAR features. The previously developed scoring scheme is used to quantify local SAR character within these compound subsets. Furthermore, a modified SAR analysis function is introduced that assesses how individual molecules contribute to local and global SAR character of a compound class. In order to relate these different SAR elements to each other, a graphical representation is developed that visualizes similarity and potency distributions of an activity class and makes it possible to investigate local environments of different SAR character. SAR contributions made by individual compounds are also visualized, which permits the identification of key compounds that strongly influence local and global SARs.

Having established a methodology to assess the role that individual molecules play for SARs within a compound class, we are also interested in investigating SAR contributions at the sub-molecular level. Hence, the final goal of this thesis is to systematically quantify SAR contributions made by functional groups in a molecule.

Goal 3: Quantitative evaluation of SAR contributions from functional groups and identification of sub-molecular SAR determinants.

For the assessment of SAR contributions from well-defined parts of a molecule, we focus on series of analogous compounds sharing a common molecular scaffold. Within these analog series, comparison of molecules that differ only at specific substitution sites makes it possible to directly assign observed SAR behavior to variations of functional groups at these sites. The SAR analysis function introduced herein is applied to quantify SAR contributions from substitution sites and combinations of sites. A graphical organization scheme visualizes these SAR contributions, enabling an intuitive analysis of SAR characteristics within series of analogous molecules. Thus, key substitution patterns are identified that largely determine the SAR character within series of analogous molecules.

Outline of the Thesis

This thesis is organized as follows. *Chapter 2* presents the initial study that provides qualitative insights into the nature of small-molecule SARs including target information. Fundamental considerations concerning SARs as a result of target–ligand interactions and the assessment of molecular similarity as a basic tool for computational SAR analysis are discussed. Methodological details of the applied similarity measures are also provided. Then, a comparative study of two-dimensional and three-dimensional compound similarity and potency is presented. Instructive results and their significance for the exploration of SARs are discussed.

Chapter 3 addresses the first goal presented above. Initially, the conceptual design of a quantitative SAR analysis function is presented. Then, the methodology is applied to study SARs within 16 compound activity classes, and exemplary classes are discussed in detail. The second part of this chapter reports the results of control calculations that have been conducted to assess the stability of the scoring scheme. Finally, methods that are related to our approach are summarized.

Chapter 4 is concerned with the second goal of this dissertation. An approach for multi-level SAR analysis is introduced and the methodology is described in detail. The method is applied to six representative compound classes and the results are discussed with regard to key aspects of global and local SAR analysis. In addition, an exemplary high-throughput screening (HTS) data set illustrates the utility of the approach for the analysis of complex SARs present in such data sets.

In *Chapter 5*, the multi-level approach introduced in Chapter 4 is extended to the analysis of structure–selectivity relationships (SSRs). First, the utilized selectivity data sets and the methodological details of the SSR analysis approach are summarized. Then, two representative compound sets with activity against pairs of related targets are studied in detail, including the comparison of local SAR and SSR features and the identification of molecular and sub-molecular selectivity determinants.

In *Chapter 6*, the third major goal of this thesis is addressed. A methodology for the quantification of SAR determinants in analog series is introduced. Key aspects of the approach are discussed using representative compound series. Furthermore, the method is also applied to the analysis of SARs within series of analogs active against multiple related targets.

Finally, *Chapter 7* summarizes the major results and presents general conclusions of this dissertation.

Chapter 2

Qualitative Characterization of Structure–Activity Relationships

In medicinal chemistry, it is widely recognized that biological responses to structural modifications of active molecules are often highly variable and that the underlying structure–activity relationships can have fundamentally different nature (Eckert and Bajorath, 2007). Taking into account that the biological activity of small molecules results from specific interactions with a macromolecular target, many SAR features can directly be related to binding characteristics at the molecular level of detail. However, general analyses comparing protein–ligand interactions and SAR features have rarely been reported. Therefore, we systematically explored information about two-dimensional ligand structure, three-dimensional binding geometry and compound potency (Peltason and Bajorath, 2007a). 2D similarity between ligands was assessed to account for chemical modifications, and a 3D similarity measure captured changes in binding modes. Similarity relationships were systematically compared and related to potency differences to better understand SARs. This chapter presents the study of experimentally determined inhibitor structures for four classical enzyme targets. In Section 2.1, SARs are discussed in the context of target–ligand interaction. General aspects of molecular similarity assessment as a basic tool for computational SAR analysis are addressed in Section 2.2, and the 2D and 3D similarity measures utilized in this study are described. Section 2.3 presents the data sets and results for each enzyme inhibitor set. Conclusions and general implications of the results are discussed in Section 2.4.

2.1 SARs and Target–Ligand Interactions

For a small molecule, efficient binding to a target, most often an enzyme or receptor protein, requires a high degree of geometrical and chemical complementarity. Geometrical complementarity involves the precise fit of the ligand into the target’s binding site, as originally postulated by the lock-and-key analogy (Fischer, 1894) or the induced-fit model of ligand binding (Koshland, 1958), which is often more appropriate. Chemical complementarity implies the ability to form highly specific chemical interactions including hydrogen bonds, electrostatic or ionic interactions, and van der Waals interactions. In addition, hydrophobic or other solvation effects often also contribute to a specific binding event.

Given these well-defined binding requirements, the SAR behavior of active molecules can often be rationalized. Accordingly, the frequently encountered occurrence of “activity cliffs” (Maggiola, 2006) can be assigned to the presence of key features that are crucial for target–ligand binding. Hence, a minute structural modification that prevents a specific key interaction might render an inhibitor completely inactive. In contrast to such “all-or-nothing” binding events, many targets permit at least some degree of ligand variability. Binding sites can often adapt to different chemotypes, giving rise to an “activity radius” that is populated by active molecules of increasing structural diversity (Eckert and Bajorath, 2007). This situation is indicative of continuous SARs and can also be interpreted from a target-centric point of view. Distinct molecular structures that adopt similar spatial conformations and arrange their interaction-relevant features in a preferred way might interact with the target in a similar manner. In conclusion, small molecule SARs are to a large extent determined by the degree of plasticity of the binding site and the presence of more or less stringent binding constraints. Systematic analyses that go beyond the study of individual cases aim at obtaining a more general view on SAR features that are prevalent for specific inhibitors and how they might be related.

2.2 Molecular Similarity Assessment

Structure–activity relationships are characterized by the way chemical modifications of small molecules affect their biological activity. Consequently, the analysis of SARs requires the systematic evaluation of these modifications through pairwise comparison of molecular structures. For this purpose, whole-molecule similarity assessment (Johnson and Maggiola, 1990) presents a well-established technique that has become an integral part of many chemoinformatics applications including virtual screening of compound databases, compound clustering, and the design of targeted or diverse structural libraries (Bajorath, 2001, 2002).

Molecular similarity assessment conceptually involves two independent aspects: the computational representation of molecular structure and a metric to numerically compare these representations. For the representation of chemical structures, a wealth of different descriptors has been designed that capture structural features, physicochemical properties, surface or shape attributes of a molecule (Todeschini et al., 2000). The descriptors that are used to represent a set of molecules span a chemical reference space of which each descriptor defines one dimension. Molecules are located in a reference space according to their descriptor values; molecular “coordinates” in the reference space correspond to the values that descriptors adopt for individual compounds. Similarity or dissimilarity between molecules is defined through their proximity or distance in reference space. Depending on the type of descriptors used, several measures to calculate the similarity or distance between them are available (Willett et al., 1998). For numerical descriptors, popular distance metrics are, for example, the Euclidean distance or the Hamming distance. Common similarity coefficients include the Cosine, the Tversky or the Tanimoto coefficient, which is the most widely used similarity measure in conjunction with binary fingerprints (Willett, 2006).

It is important to note that there is no generally applicable chemical reference space and for different applications and compound classes, different descriptor sets might prove useful (Sheridan and Kearsley, 2002). However, the choice of molecular representations and the definition of molecular similarity strongly influence the shape of an activity landscape. Hence, for the analysis of SARs, similarity assessment is a critical parameter. A major challenge is posed by the need to identify molecular descriptors that are related to compound activity and capable to consistently model an activity landscape. At the same time, similarity assessment must always be chemically meaningful, i.e. evident structural similarity should be numerically reflected by similarity calculations. For example, a similarity measure that (artificially) discriminates between closely related structures in order to account for potency differences might lead to a biased representation of an activity landscape.

2.2.1 2D Similarity Calculation

For the comparison of molecules based on their chemical graph representation, the use of binary molecular fingerprints has become widely accepted (Willett, 2006). Fingerprints are composite numerical descriptors that are represented by arrays of bits accounting for specific structural patterns. Available fingerprints incorporate a number of different chemical features and differ in part substantially in their design and complexity. Simple structural key type fingerprints monitor the presence of a collection of predefined molecular substructures and often consist only of a few hundred bit positions. Hybrid fingerprints

have also been introduced that combine structural keys and property descriptors (Eckert and Bajorath, 2007). Other common 2D fingerprint types are based on topological pharmacophore patterns, atom environments or extended atom connectivity (Bajorath, 2002).

For many applications, similarity assessment using structural keys is intuitive and leads to chemically meaningful and easily interpretable results. For this reason, we selected the widely used MACCS structural keys for the representation of molecular structures in our studies. The publicly available set of MACCS keys¹ consists of 166 bits that indicate the presence of predefined structural features in the molecular graph. Each of these structural features is represented by a position in the fingerprint bit string. If a specific substructure is found in a molecule, the corresponding bit is set to 1 (“on”); otherwise, it is set to 0 (“off”). The similarity between two molecules is then determined by comparison of their fingerprint representations. In the present work, the Tanimoto coefficient (Tc) was utilized to calculate MACCS fingerprint similarity. The Tc presents a measure of bit string overlap and is defined as follows for two binary fingerprints A and B :

$$\text{Tc}(A, B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (2.1)$$

Here, N_{AB} is the number of bits that are set on in both fingerprints, and N_A and N_B refer to the number of bits that are set on in A and B , respectively. Given this formulation, identical fingerprints obtain a maximal Tc value of 1, whereas non-overlapping fingerprints are assigned a Tc value of 0. Fingerprint representations were calculated using the Molecular Operating Environment (MOE).

2.2.2 3D Similarity Calculation

A variety of different methods have been developed for the purpose of three-dimensional molecular comparison (Willett et al., 1998). Irrespective of the specific method, 3D molecular similarity is calculated either on experimentally determined or on modeled molecular conformations. Some representations are also capable of accounting for molecular flexibility by using multiple conformations (Senese et al., 2004; von Korff et al., 2008). Many 3D similarity methods rely on descriptors that are calculated from molecular conformations, taking into account molecular surface, volume, or three-dimensional charge distributions (Todeschini et al., 2000), or fingerprints accounting for 3D pharmacophore patterns (Mason et al., 2001) or molecular shape (Good et al., 1995; Haigh et al.,

¹Fingerprint methods, software and databases used in this work are summarized in Appendix A.

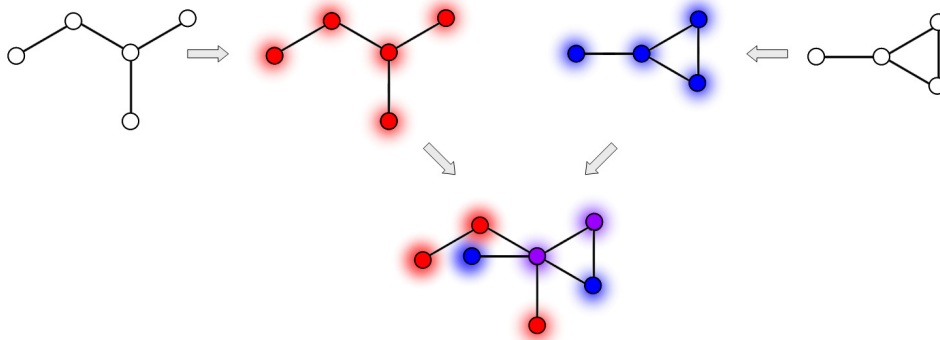


Figure 2.1: 3D similarity calculation The calculation of atomic property density overlap for two molecules is illustrated schematically. The atoms of each molecule are represented by spherically symmetric density functions, indicated by fading spheres (top). The similarity between two overlapping molecular conformations is then calculated as the intersection of their density functions (bottom).

2005). Another class of 3D similarity methods transforms molecular structures into histograms or “spectra” and then calculates the overlap between these histograms (Ankerst et al., 1999; Schuur et al., 1996). By contrast, superposition-based similarity methods directly try to map the compared molecules onto each other by optimizing the overlap of atoms or “fields” calculated around atoms, e.g. electrostatic, steric or atom property derived fields (Lemmen and Lengauer, 2000). Although computationally more demanding, an advantage of superposition methods is that they establish direct equivalences between corresponding parts of molecules.

For the spatial comparison of target-bound enzyme inhibitors, we utilized a modified superposition approach based on the overlap of atomic property density functions (Labute et al., 2001), as illustrated schematically in Figure 2.1. The aim was to compare experimentally determined binding conformations of the inhibitors while taking into account their absolute orientation and position within the binding site. Consequently, we first established a common reference frame by superposing the protein α carbon atoms of all corresponding enzyme–inhibitor complex structures using the protein superposition function in MOE. As a result, the actual binding geometries of the bound inhibitors became directly comparable. Then, a property density function for the coordinates of each ligand was defined and calculated as follows. For each atom i , the following four properties were calculated using a pharmacophore atom typing scheme implemented in MOE (Bush and Sheridan, 1993). A corresponding property weight w_i^P was assigned accordingly, obtaining the value 1 if atom i had the property P and the value 0 otherwise.

Aromatic The aromatic property was assigned to an atom i if it was sp^2 -hybridized and belonged to a ring that obeyed the Hückel rule. In that

case, the corresponding property weight was set to 1, i.e. $w_i^{aro} = 1$.

Donor The H-bond donor property was assigned by setting $w_i^{don} = 1$ if atom i was classified as “donor” or “basic” under the pharmacophore atom typing scheme.

Acceptor The H-bond acceptor property was assigned by setting $w_i^{acc} = 1$ if atom i was classified as “acceptor” or “acidic” under the atom typing scheme.

Hydrophobic The hydrophobic property was assigned by setting $w_i^{hyd} = 1$ if atom i was of type “hydrophobe” under the atom typing scheme.

For a given atom property P , each atom i was represented by a spherically symmetric Gaussian density function f_i^P centered at the position x_i of the atom nucleus; the width of the Gaussian was determined by the van der Waals atom radius r_i :

$$f_i^P(x) = w_i^P \left(\frac{a^2}{2\pi r_i^2} \right)^{3/2} \exp \left\{ -\frac{a^2}{2r_i^2} |x - x_i|^2 \right\} \quad (2.2)$$

Here, the parameter a was used to scale the atom radii simultaneously and was set to 2 in our calculations. The property density f^P for a molecule was then defined as the mean of the property density functions of its n atoms:

$$f^P(x) = \sum_{i=1}^n \frac{w_i^P}{n} \left(\frac{a^2}{2\pi r_i^2} \right)^{3/2} \exp \left\{ -\frac{a^2}{2r_i^2} |x - x_i|^2 \right\} \quad (2.3)$$

For the comparison of two molecules or conformations X and Y , the overlap of their property densities was calculated, obtaining again a sum-of-Gaussians density. Let x_1, \dots, x_n and y_1, \dots, y_m denote the spatial positions of the atoms in conformations X and Y , and r_1, \dots, r_n and r'_1, \dots, r'_m be their van der Waals radii. Let further be w_1^P, \dots, w_n^P and $w'_1{}^P, \dots, w'_m{}^P$ the property weights of the atoms in X and Y , respectively. Then, the density overlap of X and Y for property P was defined as follows:

$$F^P(X, Y) = \sum_{i=1}^n \sum_{j=1}^m \frac{w_i^P w'_j{}^P}{nm} \left(\frac{a^2}{2\pi(r_i^2 + r_j'^2)} \right)^{3/2} \exp \left\{ -\frac{a^2}{2} \frac{|x_i - y_j|^2}{r_i^2 + r_j'^2} \right\} \quad (2.4)$$

This formulation generalizes to more than one property through summation of the overlap equations for the individual properties. For the four atom properties

listed above, the density overlap for two molecules X and Y was defined to be

$$F(X, Y) = \sum_{i=1}^n \sum_{j=1}^m \frac{w_i^{aro} w_j^{aro} + w_i^{don} w_j^{don} + w_i^{acc} w_j^{acc} + w_i^{hyd} w_j^{hyd}}{nm} \times \left(\frac{a^2}{2\pi(r_i^2 + r_j'^2)} \right)^{3/2} \exp \left\{ -\frac{a^2}{2} \frac{|x_i - y_j|^2}{r_i^2 + r_j'^2} \right\} \quad (2.5)$$

A final normalization was carried out in order to obtain 3D similarity values between 0 (distinct spatial arrangement with no common atom positions) and 1 (identical conformation and position). The final 3D similarity values were obtained by dividing the overlap of the molecular property density functions by the mean self-overlap of the respective conformations:

$$F_{norm}(X, Y) = \frac{F(X, Y)}{\frac{1}{2} [F(X, X) + F(Y, Y)]} \quad (2.6)$$

2.3 Relationships between 2D and 3D Similarity and Potency

In order to evaluate SAR characteristics in a detailed manner, we systematically analyzed crystallographic enzyme-inhibitor complex structures. Pairwise 2D and 3D similarity relationships of the inhibitors were quantitatively assessed, compared and related to differences in compound potency.

2.3.1 Data and Calculations

As a data basis for the analysis, sets of inhibitors were assembled for which experimentally determined complex structures and potency measurements for a given target were available. On the basis of a survey of the PDBbind database (Wang et al., 2004, 2005), we selected four representative target enzymes for which sufficient inhibitor data were available: elastase and coagulation factor Xa, two serine proteases; the metalloenzyme carbonic anhydrase II; and the RNA-cleaving enzyme ribonuclease A. These enzymes are long-established targets in pharmaceutical research and represent active sites of distinct chemical and spatial architecture. Table 2.1 summarizes the structural data used for the analysis, and their PDB codes are provided in Appendix B.

For each set of inhibitors, pairwise 2D and 3D similarity coefficients were calculated as described above. In order to support the systematic assessment of similarity relationships, scatter plots were created that correlated 2D and 3D similarity values of every compound pair. Figure 2.2 shows the 2D-3D similarity plots for individual inhibitor sets. These plots facilitated the detection of

correlations or discrepancies between 2D and 3D similarity relationships. Furthermore, in order to relate molecular similarity to compound potency, each data point was colored according to the potency difference of the corresponding compound pair using a continuous spectrum from black for smallest to red for largest potency differences in a data set. For this purpose, absolute differences between pK_i or pIC_{50} values were used. In addition, Pearson correlation coefficients between 2D and 3D similarity were calculated for each inhibitor set and are reported in Table 2.1. Results for individual enzyme inhibitor sets are discussed in the following section.

2.3.2 Results

Ribonuclease A A characteristic feature of the active site of ribonuclease A is the presence of a positively charged binding pocket that inhibitors need to fill in order to bind efficiently. This binding constraint is reflected by the structure of the studied inhibitors. The nine selected compounds are nucleotide derivatives containing adenine or uracil and one or more phosphate groups that are accommodated in the binding pocket and compensate the positively

Table 2.1: Summary of inhibitor data sets

	carbonic anhy- drase II	elastase	factor Xa	ribonuclease A
no. structures	27	14	16	9
2D similarity				
minimum	0.07	0.34	0.24	0.76
maximum	1.00	0.92	1.00	0.98
average	0.59	0.52	0.50	0.87
3D similarity				
minimum	0.00	0.09	0.28	0.13
maximum	0.99	0.96	0.96	0.87
average	0.60	0.36	0.58	0.44
cor. 2D/3D	0.79	0.31	0.46	0.58
potency				
minimum	0.03 nM	0.46 nM	0.007 nM	27 nM
maximum	125 μ M	890 μ M	131 nM	82 μ M

Potency and similarity distributions are given for the four enzyme inhibitor sets discussed in the text. 'no. structures' reports the number of inhibitor structures and 'cor. 2D/3D' denotes the correlation coefficient between pairwise 2D and 3D similarity calculated as described in the text.

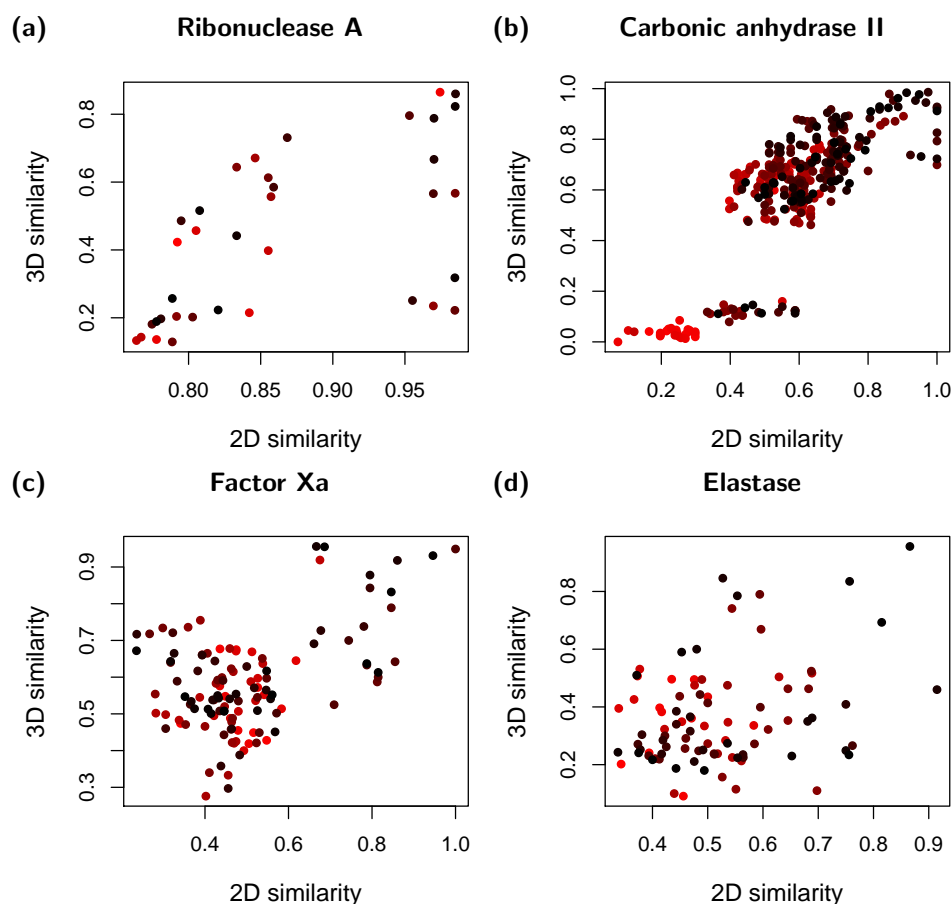
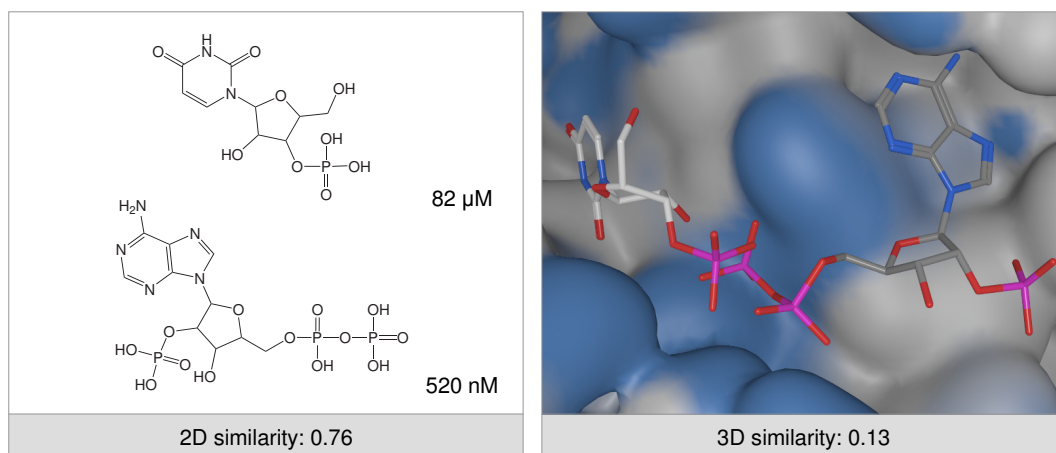


Figure 2.2: Comparison of 2D and 3D similarity Each dot represents 2D and 3D similarity values of a pairwise comparison of two inhibitors. Data points are color-coded according to potency differences by using a continuous spectrum from black for smallest to red for largest potency difference within each compound set. Scatter plots were created using R.

charged residues. Accordingly, the inhibitors have very similar structures and obtain pairwise MACCS Tc similarity values greater than 0.75 (Figure 2.2(a)). However, despite their distinct structural similarity, significant 3D variations are observed among ribonuclease inhibitors. Figure 2.2(a) shows that pairwise 3D similarity values essentially cover the entire range from a minimum of 0.13 to a maximum of 0.87. These varying levels of 3D similarity are due to the fact that inhibitors containing different nucleobases adopt distinct binding modes. As illustrated in Figure 2.3(a), overall similar structures can bind very differently as long as the phosphate group constraint is satisfied. Furthermore, there is little correlation between structural similarity and potency. Similar structures have different potency levels irrespective of whether their binding modes are similar or not. In fact, inhibitors with nearly identical binding conformations are found to differ by up to three orders of magnitude in potency, solely due to

(a)



(b)

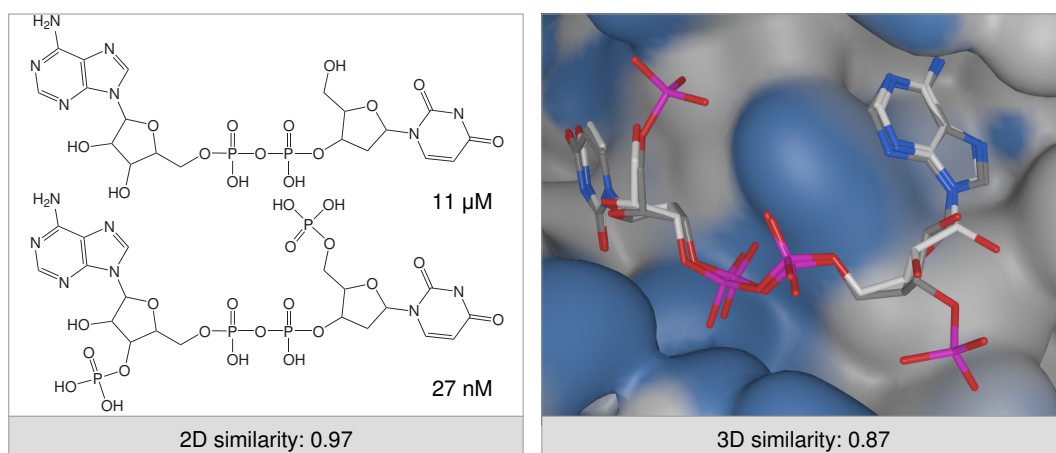


Figure 2.3: Ribonuclease A inhibitors On the left side, the 2D structures of selected inhibitors are shown. On the right, the same inhibitors are shown in their binding conformations within the active site of ribonuclease A. In the 3D representation, the carbon atoms of the inhibitor whose 2D structure is shown at the top of each subfigure are colored in light gray, and the carbons of the inhibitor whose structure is shown at the bottom are colored in dark gray. Blue areas on the protein surface indicate regions of positive partial charge and highlight the phosphate binding pocket that presents a major binding constraint. (a) A pair of inhibitors that adopt different binding modes in the ribonuclease active site. (b) Two closely related analogs with very similar binding modes but dramatic differences in potency.

two additional phosphate groups in the highly potent analogue (Figure 2.3(b)). Hence, in this case, binding characteristics are determined by local structural features and are only weakly related to whole-molecule similarity. The binding constraint posed by the positively charged phosphate-binding pocket makes

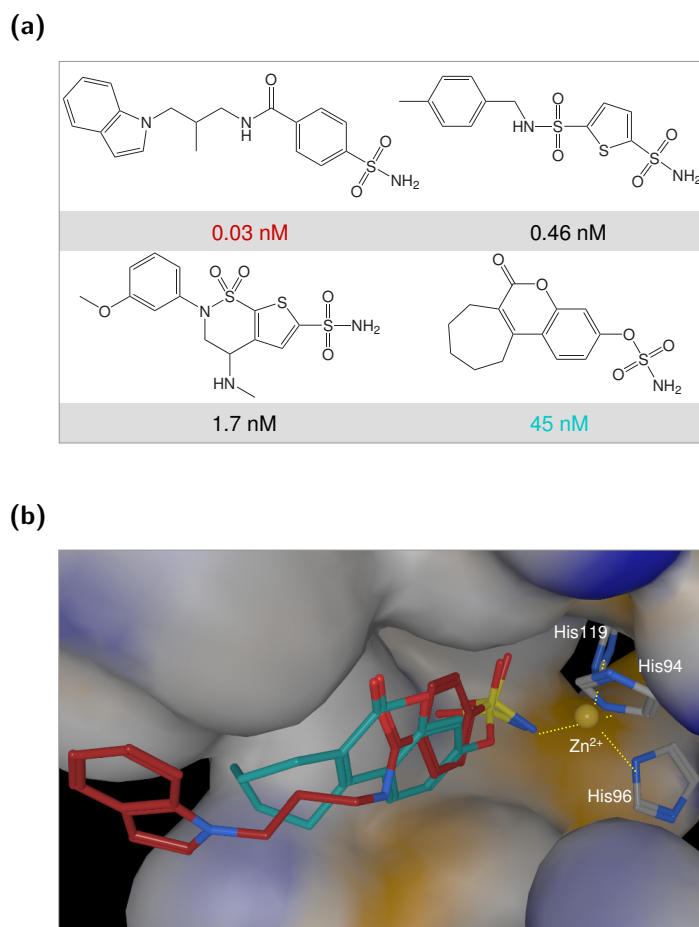


Figure 2.4: Carbonic anhydrase II inhibitors (a) Highly potent inhibitors representing different chemotypes that share a sulfonamide group. Potency values of the most and the least potent molecules are colored according to ligand colors in the 3D representation shown in part (b). (b) Compounds from (a) with highest and lowest potency are displayed in the carbonic anhydrase binding pocket. The compounds have distinct chemical scaffolds and display differences in binding geometry. The zinc cation within the active site, represented as yellow sphere, is coordinated by three histidine residues and the sulfonamide group of the ligand.

ribonuclease inhibitors a prime example for discontinuous SARs that manifest themselves on both the 2D and 3D level.

Carbonic Anhydrase II Carbonic anhydrase presents another example of an enzyme whose binding site architecture poses a severe structural constraint on ligand binding. In this metalloenzyme, the need to coordinate a catalytically important zinc cation within the active site presents the major prerequisite for ligand–target interaction. Known inhibitors meet the constraint by means of a sulfonamide group, which is a hallmark of carbonic anhydrase inhibitors. On

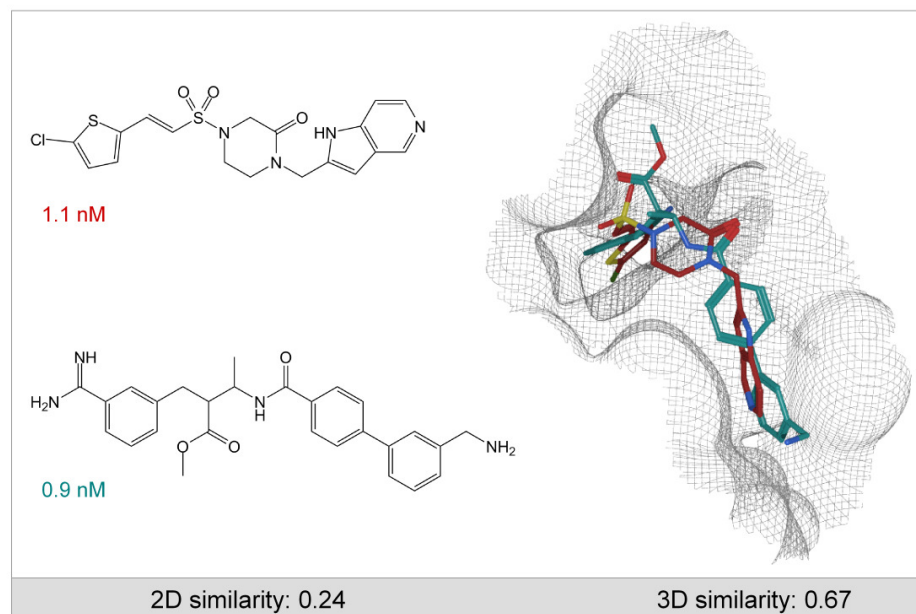


Figure 2.5: Factor Xa inhibitors Inhibitor pair with low 2D but notable 3D similarity and potency in the low nanomolar range.

the basis of these characteristics, one might expect to observe discontinuous SARs, similar to ribonuclease A. However, analysis of the inhibitors reveals a different situation. The studied carbonic anhydrase inhibitors cover a wide spectrum of 2D similarity and are structurally much more diverse than the ribonuclease inhibitors discussed above (Figure 2.2(b)). Although most inhibitors share the sulfonamide group, other structural moieties show in part substantial variations. Different chemotypes are found among highly potent compounds, as shown in Figure 2.4(a). Moreover, 3D binding similarity strongly correlates with structural similarity ($r = 0.79$), i.e. 2D similar molecules bind in a similar manner and with comparable potency, whereas molecules with limited similarity also show differences in their binding geometry (Figure 2.4(b)). Moreover, largest potency differences are observed among dissimilar compounds, indicated by the red data points in Figure 2.2(b). Thus, in this case, continuous SARs exist proximal to an activity cliff formed by the sulfonamide constraint.

Factor Xa The coagulation factor Xa is found to have less stringent requirements for inhibitor binding than the enzymes discussed so far. Accordingly, the majority of inhibitors are related by a continuous SAR. There is a significant degree of structural diversity and most similar 2D structures bind very similarly and with comparable potency. By contrast, compounds with limited 2D and 3D similarity display the largest differences in potency (Figure 2.2(c)). A per-

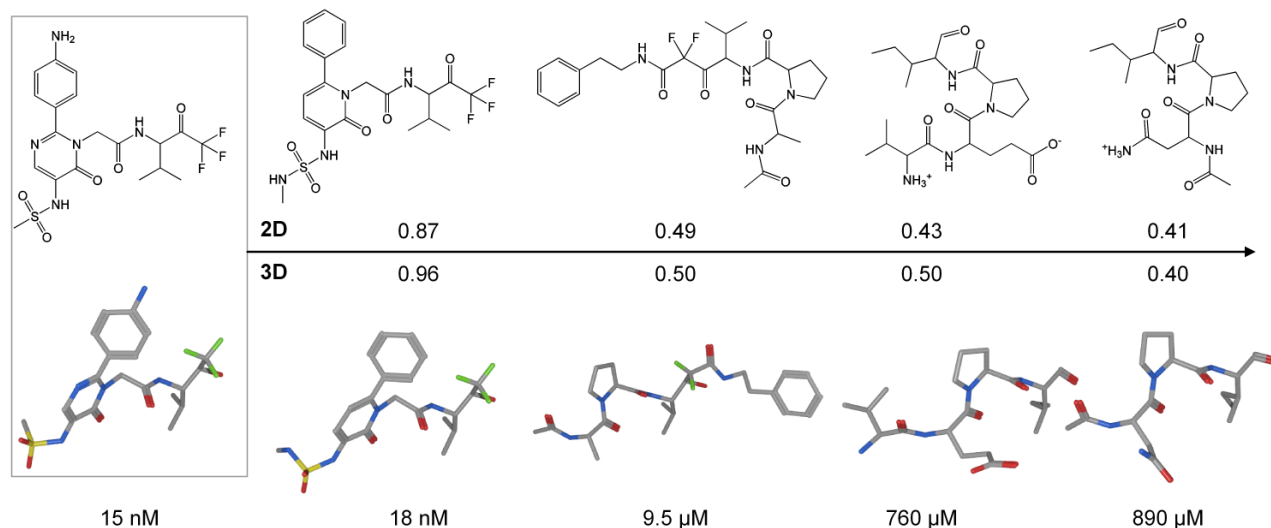
haps unexpected characteristic of diverse factor Xa inhibitors is their tendency to bind in similar conformations. Structures with distinct chemical scaffolds adopt comparable binding modes that match the shape of the binding pocket (Figure 2.5). This indicates that in this case, binding to the target protein is largely governed by shape complementarity, which provides the basis for SAR continuity among factor Xa inhibitors. The active site of the enzyme tolerates structural variations provided a high degree of spatial complementarity is maintained and a few key interactions are formed.

Elastase In the case of elastase, another serine protease with a comparably permissive binding site, structurally related compounds display only minor potency differences and, in addition, potent inhibitors represent diverse structural features, similar to factor Xa. However, analysis of 3D structures reveals a more complex picture than observed for factor Xa. There is no significant correlation between 2D and 3D similarity. In fact, different subsets of elastase inhibitors are identified for which 2D and 3D similarity is either strongly or inversely correlated. For the inhibitor series shown in Figure 2.6(a), strong correlation between structural and binding similarity is observed ($r = 0.82$), and compound potency is found to decrease with decreasing similarity. More precisely, if we consider the most potent compound in Figure 2.6(a) as a reference point, structural departure from a preferred inhibitor is accompanied by a gradual loss in potency, which is prototypic for a continuous SAR. On the other hand, for a series of trifluoro-acetyl-dipeptide anilides with overall comparable potency, 2D and 3D similarity show an inverse correlation ($r = -0.52$; Figure 2.6(b)). This means that within this series of inhibitors, decreasingly similar compounds adopt increasingly similar binding modes, which represents a different type of a continuous SAR. How can these observations be rationalized? As shown in Figure 2.7, elastase accepts multiple binding modes which can be adopted by structurally diverse inhibitors that present their functional groups in spatially corresponding positions. Interestingly, binding modes appear to have no significant influence on compound potency. Thus, in the case of elastase, different continuous SARs can be distinguished, characterized either by a potency gradient accompanying changes in 2D/3D structure or by the presence of 2D and 3D diverse inhibitors with comparable potency.

2.4 Summary and Conclusions

In summary, the analysis of inhibitor structures of four well-studied target enzymes reveals complex similarity–potency relationships. The enzymes ribonuclease A and carbonic anhydrase II impose significant constraints on inhibitor binding due to the architecture of their binding sites, but with different effects

(a)



(b)

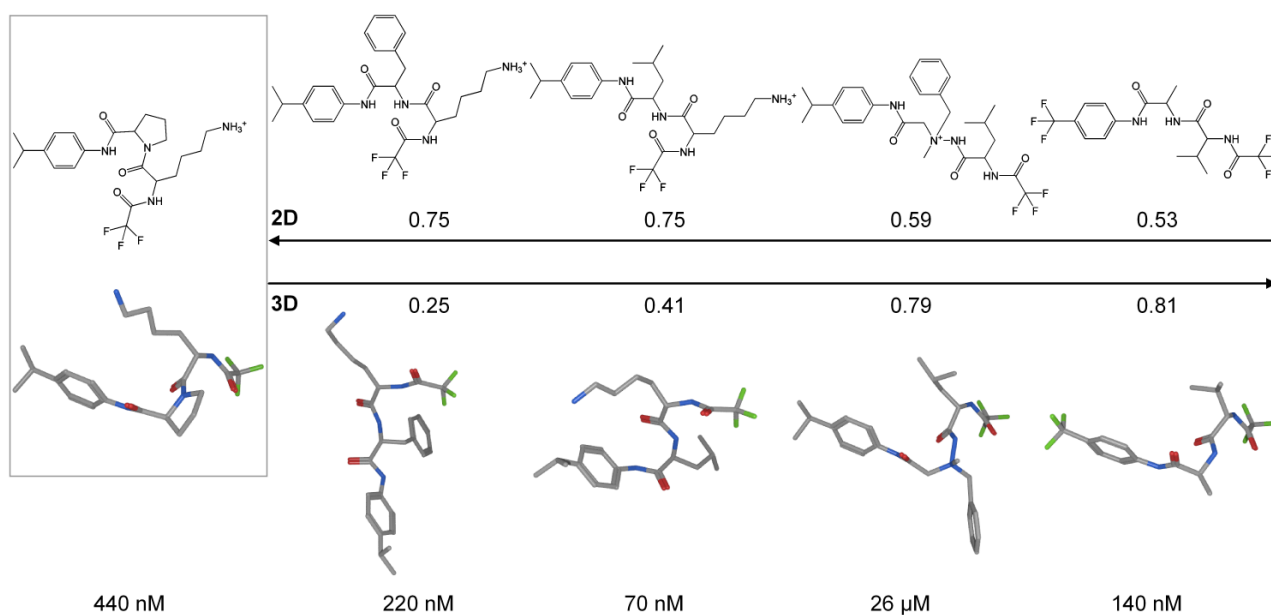


Figure 2.6: Elastase inhibitors For inhibitor series with direct or inverse correlation between 2D and 3D similarity, 2D (top) and 3D structures (bottom) are shown. The compound at the left is used as reference compound, and 2D and 3D similarity values to the reference compound are reported for each inhibitor in a series. Potency values are reported below each 3D structure. (a) Direct correlation between 2D and 3D similarity in a subset of elastase inhibitors. Structural departure from a preferred inhibitor is accompanied by steady potency decrease. (b) Inverse correlation between 2D and 3D similarity in another subset of elastase inhibitors with overall comparable potency.

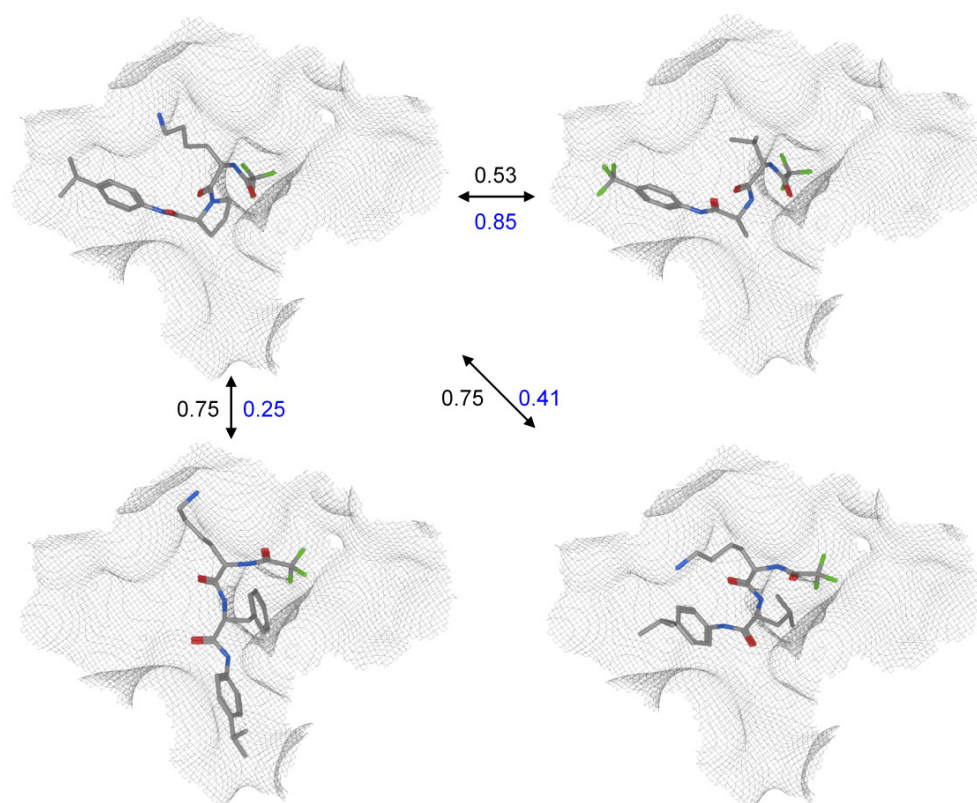


Figure 2.7: Alternative binding modes in elastase Binding conformations of four selected elastase inhibitors are shown. 2D and 3D similarity values to the reference molecule shown in the upper left corner are reported in black and blue, respectively. The two molecules at the top have only limited structural similarity but share the same binding mode. The molecules at the bottom are structurally more similar to the reference compound but adopt different binding modes.

on ligand SARs. The studied set of ribonuclease inhibitors is marked by the lack of structural diversity and displays discontinuous SARs, albeit with a remarkable degree of 3D variability. By contrast, carbonic anhydrase inhibitors are related by continuous SARs within the boundaries determined by a structural binding constraint. Furthermore, a prime example of continuous SARs is presented by factor Xa inhibitors that include a high degree of structural diversity. For factor Xa and carbonic anhydrase, 2D and 3D similarity correspond well to each other, which gives rise to SAR continuity also at the 3D level. In the case of elastase, by contrast, different continuous SARs characterize individual inhibitor series. In one series, 2D similarity correlates with 3D similarity and is consistent with observed potency differences. Another series of elastase inhibitors is characterized by different binding modes that are adopted by similar structures having comparable potency levels.

Taken together, the comparison of 2D and 3D similarity with compound potency reveals that relationships between similarity and potency are variable

and often highly complex. Even in the presence of severe binding constraints, targets permit a sometimes surprising variability of ligand structure and binding modes. These findings revise previous views that similar ligands generally bind in a similar way to a target (Boström et al., 2006). Moreover, the analysis shows that different continuous SARs can coexist in an enzyme, as well as continuous and discontinuous SARs, depending on the structural features of ligands. Thus, the results provide evidence for the presence of multiple and heterogeneous SARs within target-specific activity landscapes. These observations imply that the nature of SARs is not uniquely “dictated” by target features, but is also influenced to a comparable extent by the chemical nature of ligands. The picture that emerges from our analysis is that different SAR features are not mutually exclusive and that SARs are generally more heterogeneous in nature than often thought. These findings suggest that continuous and discontinuous regions coexist in many activity landscapes, which has also practical implications for drug design. In particular, focusing on continuous regions should in principle enable the identification of small molecules with diverse structures but similar activity for many different protein targets.

Chapter 3

Quantitative Description of Structure–Activity Relationships

The qualitative analysis of similarity–potency relationships presented in the previous chapter has elucidated the often highly complex nature of SARs. The coexistence of distinct SAR features in many compound activity classes yields heterogeneous SARs that are characterized by variable activity landscapes where activity cliffs are separated by gently sloped or even flat regions. It is evident that the variable nature of small molecule SARs to a great extent complicates their systematic study or classification. Typically, SARs are investigated on a case-by-case basis for classes of closely related molecules. Methods to systematically explore SARs on a large scale have only recently been introduced (Peltason and Bajorath, 2009). This chapter presents a numerical function, termed SAR Index (SARI), that attempts to put the characterization and comparison of SARs on a quantitative basis (Peltason and Bajorath, 2007b). The approach is based on systematic assessment of structural similarity and potency relationships and thus departs from the target-centric view of the 3D similarity-oriented studies presented in Chapter 2. Limiting similarity assessment to 2D molecular representations makes it possible to extend quantitative SAR analysis to targets for which no, or only few, relevant X-ray structures are available. The SARI formalism provides a consistent framework for the evaluation of activity landscape topology and classifies SARs in compound activity classes into three different categories. Section 3.1 presents these categories and the conceptual basis of SARI. The results of SAR profiling for 16 enzyme inhibitor sets are reported in Section 3.2. Furthermore, the influence of fingerprint representations and data set size is investigated for another set of activity classes in Section 3.3. Finally, we present approaches that are related to SARI in Section 3.4 and discuss general conclusions in Section 3.5.

3.1 SARI Methodology

SARI presents a scoring scheme designed to quantitatively capture the nature of SARs for a given set of compounds active against a specific target. The SARI score is calculated from two individual components, the continuity and discontinuity score, that quantify the composition of smooth and rugged parts of an activity landscape, respectively. To these ends, pairwise 2D similarity relationships and differences in compound potency are assessed and related to each other. Structural similarity between molecules is calculated as the Tc for comparison of MACCS fingerprints and potency is represented by pK_i or pIC_{50} values.

3.1.1 Continuity Score

The continuity score estimates the continuous character of SARs, corresponding to smooth regions in the activity landscape. Continuous SARs are characterized by gradual biological responses to chemical changes and ultimately delineate an activity radius that is populated by increasingly diverse structures with similar potency. Therefore, the continuity score measures the potency-weighted structural diversity within a class of active compounds. For this purpose, the similarity between each pair of compounds is assessed and a weighted mean of the reciprocal pairwise compound similarity is calculated for all compound pairs. The weights combine the potency values of both compounds in a pair and the difference in potency between them. The “raw” (i.e. non-normalized) continuity score for a compound class A is defined as follows:¹

$$\text{cont}_{\text{raw}}(A) = \frac{\sum_{\{(i,j) \in A | i \neq j\}} w_{ij} \frac{1}{1 + \text{sim}(i, j)}}{\sum_{\{(i,j) \in A | i \neq j\}} w_{ij}} \quad (3.1)$$

$$w_{ij} = \frac{P_i \cdot P_j}{1 + |P_i - P_j|} \quad (3.2)$$

Here, $\text{sim}(i, j)$ stands for the similarity between compounds i and j , w_{ij} denotes the weight for the compound pair and P_i and P_j denote their potency values, respectively. Hence, the continuity score measures the global diversity in a set of active compounds, assigning high weights to compound pairs with high potency but low potency differences. This weighting scheme takes into account that SAR continuity is primarily characterized by the presence of comparably potent inhibitors of increasing structural diversity. Compound pairs with overall low

¹Addition of 1 to the similarity in the score definition and to the potency difference in the weight definition prevents division by 0.

potency and/or high differences in potency convey only a limited amount of information concerning the continuous nature of an SAR.

3.1.2 Discontinuity Score

In discontinuous SARs, on the other hand, the most prominent characteristic is the presence of activity cliffs formed by similar compounds having markedly different potency. Accordingly, the discontinuity score accounts for potency differences among similar compounds. It is defined as the mean of potency differences between pairs of similar molecules multiplied by pairwise similarity. Here only compound pairs are considered that exceed a predefined similarity threshold. We set this similarity threshold to a MACCS Tc of 0.65, which is a relatively "soft" threshold, in order to be able to detect also activity cliffs between remotely similar compounds. However, multiplication with similarity puts more emphasis on potency differences between highly similar molecules. Furthermore, we apply a cutoff for the pairwise potency difference. For a compound pair to be considered for discontinuity score calculation, we require a potency difference of more than one order of magnitude in order to focus the analysis on significant activity cliffs. Hence, the discontinuity score for a compound class A is calculated as follows:

$$\text{disc}_{\text{raw}}(A) = \frac{\text{mean}_{\left\{ \begin{array}{l} (i,j) \in A \\ \text{sim}(i,j) > 0.65, \\ |P_i - P_j| > 1 \end{array} \right\}} (|P_i - P_j| \cdot \text{sim}(i, j))}{1} \quad (3.3)$$

3.1.3 Normalization

For ease of comparison, the raw continuity and discontinuity scores are standardized and normalized to the value range [0,1]. For this purpose, a panel of activity classes is taken as a basis, and the raw scores of each class are normalized with respect to the score distribution within this reference panel, as described in the following. Initially, the sample mean ($\overline{\text{cont}}_{\text{raw}}$, $\overline{\text{disc}}_{\text{raw}}$) and sample standard deviation (s_{cont} , s_{disc}) of the scores within the set of reference classes are calculated. These reference values are then used to calculate standardized or Z-scores from the raw scores of each activity class A :

$$\text{cont}_{\text{zscore}}(A) = \frac{\text{cont}_{\text{raw}}(A) - \overline{\text{cont}}_{\text{raw}}}{s_{\text{cont}}} \quad (3.4)$$

$$\text{disc}_{\text{zscore}}(A) = \frac{\text{disc}_{\text{raw}}(A) - \overline{\text{disc}}_{\text{raw}}}{s_{\text{disc}}} \quad (3.5)$$

Z-scores report how many standard deviations a score value is above or below the mean. Thus, the continuity and discontinuity scores are expressed in

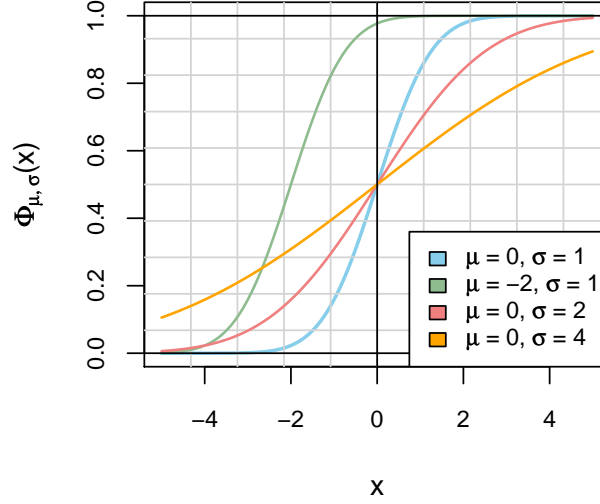


Figure 3.1: Cumulative distribution function for the normal distribution For different values for the mean (μ) and standard deviation (σ) of a normal distribution, the cumulative distribution function is plotted. The standard normal distribution with a mean of 0 and a standard deviation of 1 is indicated by the blue line.

units of standard deviations and can be directly compared. Finally, the scores are mapped onto the value range $[0,1]$ by calculating the value of the cumulative distribution function for each Z-score under the assumption of a standard normal distribution:

$$\text{cont}_{\text{norm}}(A) = \Phi(\text{cont}_{\text{zscore}}(A)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{cont}_{\text{zscore}}(A)} \exp\left(-\frac{1}{2}x^2\right) dx \quad (3.6)$$

$$\text{disc}_{\text{norm}}(A) = \Phi(\text{disc}_{\text{zscore}}(A)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{disc}_{\text{zscore}}(A)} \exp\left(-\frac{1}{2}x^2\right) dx \quad (3.7)$$

This function indicates for a given Z-score value the probability of the event that the standardized score of a randomly chosen activity class is less than or equal to this value. Hence, a Z-score of 0 obtains a value of 0.5 because it corresponds to the mean of the entire raw score distribution, and other Z-scores have a probability of 0.5 to fall into the range below or above the mean, respectively. As illustrated in Figure 3.1, increasing Z-scores obtain values closer to 1, and decreasing Z-scores approach a value of 0. Hence, normalized continuity and discontinuity score values near 0 correspond to a low degree of SAR continuity and discontinuity, respectively, whereas values near 1 indicate the opposite situation.

3.1.4 SARI Score

The final SARI combines the normalized continuity and discontinuity scores:

$$\text{SARI}(A) = \frac{\text{cont}_{\text{norm}}(A) + (1 - \text{disc}_{\text{norm}}(A))}{2} \quad (3.8)$$

Given this formulation, the SARI also falls into the value range between 0 and 1. Since continuity and discontinuity account for contrary SAR character, the discontinuity score is transformed to its complementary value by subtraction from 1. Accordingly, high SARI values result from high continuity and low discontinuity scores and correspond to continuous SARs, whereas low SARI values are produced by the inverse score combination and indicate SAR discontinuity. Intermediate values of SARI around 0.5 correspond to heterogeneous SARs that combine continuous and discontinuous elements. As will be discussed in detail below, intermediate SARI values can arise from two different situations: either high continuity and discontinuity scores, or low values for both score components. The former situation indicates the coexistence of continuous and discontinuous SAR elements and can be envisioned as a smooth activity landscape that contains diverse active compounds and is interspersed with activity cliffs. We call this phenotype “heterogeneous-relaxed” as opposed to the other heterogeneous subtype that is marked by a low degree of continuity and discontinuity. Here, the lack of structural diversity within a data set is accompanied by the absence of significant activity cliffs, a situation which is often indicative of continuous SARs within the boundaries of an activity cliff imposed by a structural binding constraint. Accordingly, this subtype of heterogeneous SARs is termed “heterogeneous-constrained”.

3.2 SAR Profiling

In order to evaluate the design and utility of the SAR index, we calculated SARI scores for a panel of 16 inhibitor classes directed at a variety of target enzymes. The compound sets were classified according to their SARI profile and the results were compared to qualitative observations on the basis of representative molecular structures.

3.2.1 Data and Calculations

For our analysis, we searched for compound activity classes that covered a wide range of targets and had varying degrees of structural diversity and significantly different potency distributions. Given these criteria, we selected 16 sets of enzyme inhibitors consisting of between 9 and 33 molecules taken from the PDB-bind and MDDR databases, as summarized in Table 3.1. The selected activity

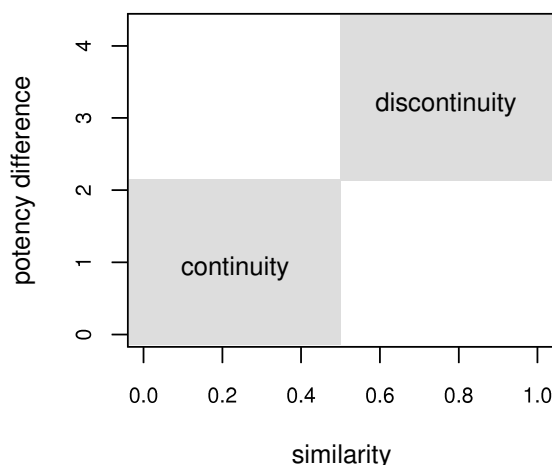


Figure 3.2: Different regions in a similarity–potency plot Four major regions can be identified in a plot of potency difference versus similarity, as indicated by the gray and white fields. The lower left region contains compounds that have diverse structures but similar potency and represent SAR continuity. By contrast, the upper right region contains similar compounds with high differences in potency that participate in activity cliffs and give rise to SAR discontinuity.

Table 3.1: Enzyme inhibitor classes and their similarity and potency distributions

target	cpds	MACCS Tc			potency [nM]	
		min	max	avg	min	max
poly(ADP-ribose) polymerase	23	0.29	0.82	0.47	5	35 000
coagulation factor Xa	16	0.24	1	0.50	0.007	131
cyclooxygenase 2	21	0.21	0.96	0.48	0.09	3 380
cyclin-dependent kinase 2	27	0.25	0.99	0.49	3	38 000
protein-tyrosine phosphatase 1b	22	0.13	0.91	0.49	1.8	63 000
carbonic anhydrase II	27	0.07	1	0.59	0.03	125 000
thromboxane synthase	23	0.21	1	0.48	0.8	33 000
acetylcholine esterase	19	0.22	1	0.46	0.13	8 900
elastase	14	0.34	0.91	0.52	0.46	890 000
trypsin	33	0.14	1	0.49	5.2	32 500 000
dihydrofolate reductase	23	0.34	0.84	0.54	0.1	19 500
peptidylprolyl isomerase	14	0.09	0.99	0.54	0.2	500 000
thrombin	28	0.23	1	0.49	0.0014	85 000
thymidylate synthase	18	0.29	0.98	0.66	2	36 000
ribonuclease A	9	0.76	0.99	0.87	27	82 000
adenosine deaminase	19	0.34	1	0.67	0.0001	9 000

For 16 activity classes, the number of compounds ('cpds'), minimum ('min'), maximum ('max') and average ('avg') MACCS Tc similarity values and minimum and maximum potency values are reported.

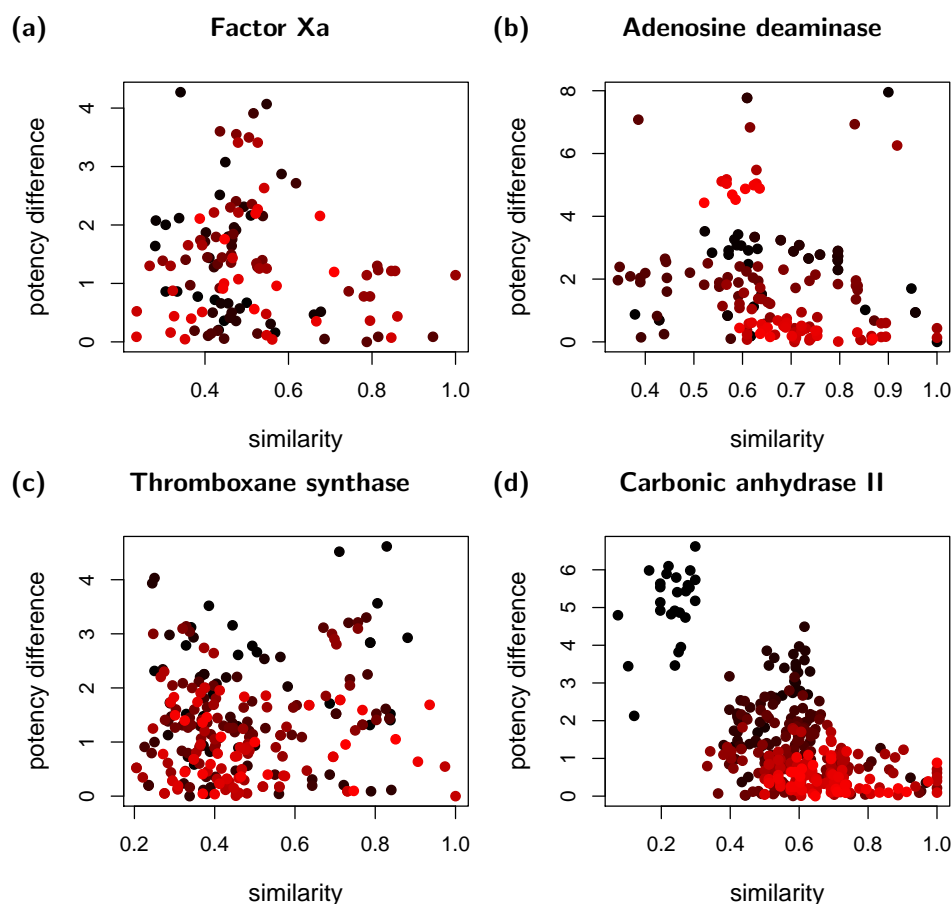


Figure 3.3: Potency differences versus 2D similarity Each data point represents a pairwise comparison of inhibitors within an activity class. Compound similarity is assessed using MACCS Tc values, and potency differences are calculated between pK_i (or pIC_{50}) values. Hence, one unit on the y-axis corresponds to a difference between K_i (IC_{50}) values of one order of magnitude. Data points are color-coded according to the sum of their logarithmic potency values using a continuous spectrum from black (lowest) to red (highest potency). Shown are distributions for four exemplary inhibitor sets discussed in the text.

classes covered a large intra-class diversity spread and differed significantly in their potency distribution. To illustrate the distinct similarity and potency distributions within individual compound sets, we plotted potency differences against similarity values in a pairwise manner. In order to highlight pairs of highly potent compounds, a color code was applied that colored each data point according to the sum of pK_i or pIC_{50} values of the corresponding compounds in a pair, using a continuous spectrum from black for lowest to red for highest potency sums in a class. Hence, red data points indicate compound pairs consisting of two highly potent molecules. In these plots, four major regions representing different SAR information can be distinguished, as illustrated in Figure 3.2. The lower left quadrant contains data points for compound pairs

with limited structural similarity having low potency differences. Thus, these compounds reflect SAR continuity. By contrast, the upper right quadrant is populated by compound pairs that have similar structures but large differences in potency. This plot region accounts for activity cliffs and SAR discontinuity. The other two regions either contain compound pairs with low similarity and high potency differences (upper left) or compound pairs with similar structure and potency (lower right). From these parts of the plot, only limited SAR information can be derived. Figure 3.3 shows similarity–potency scatter plots for four representative classes that will be discussed in the following.

3.2.2 Results

We calculated SARI scores for the selected enzyme inhibitor sets and found that they covered a wide spectrum of SAR characteristics. For normalization of the continuity and discontinuity scores, we utilized the same 16 activity classes as reference, i.e. the scores are scaled to the score distribution within this set of activity classes. The calculated continuity, discontinuity, and SARI scores for the inhibitor sets are reported in Table 3.2 and differ substantially. Continuity scores ranged from less than 0.01 to 0.84, discontinuity scores from 0.05 to

Table 3.2: SAR indices for different classes of enzyme inhibitors

target	SARI scores			SAR category
	cont	disc	SARI	
poly(ADP-ribose) polymerase	0.82	0.05	0.89	continuous
coagulation factor Xa	0.71	0.08	0.82	continuous
cyclooxygenase 2	0.81	0.45	0.68	continuous
cyclin-dependent kinase 2	0.74	0.41	0.66	continuous
protein-tyrosine phosphatase 1b	0.77	0.47	0.65	continuous
carbonic anhydrase II	0.27	0.06	0.61	heterogeneous-constrained
thromboxane synthase	0.84	0.75	0.54	heterogeneous-relaxed
acetylcholine esterase	0.84	0.78	0.53	heterogeneous-relaxed
elastase	0.61	0.54	0.53	heterogeneous-relaxed
trypsin	0.38	0.39	0.50	heterogeneous-constrained
dihydrofolate reductase	0.55	0.65	0.45	heterogeneous-relaxed
peptidylprolyl isomerase	0.16	0.39	0.39	heterogeneous-constrained
thrombin	0.69	0.90	0.39	heterogeneous-relaxed
thymidylate synthase	0.14	0.60	0.27	discontinuous
ribonuclease A	0.01	0.52	0.24	discontinuous
adenosine deaminase	0.13	0.98	0.07	discontinuous

Shown are continuity ('cont'), discontinuity ('disc'), and SARI scores. In addition, the SAR category is reported for each activity class.

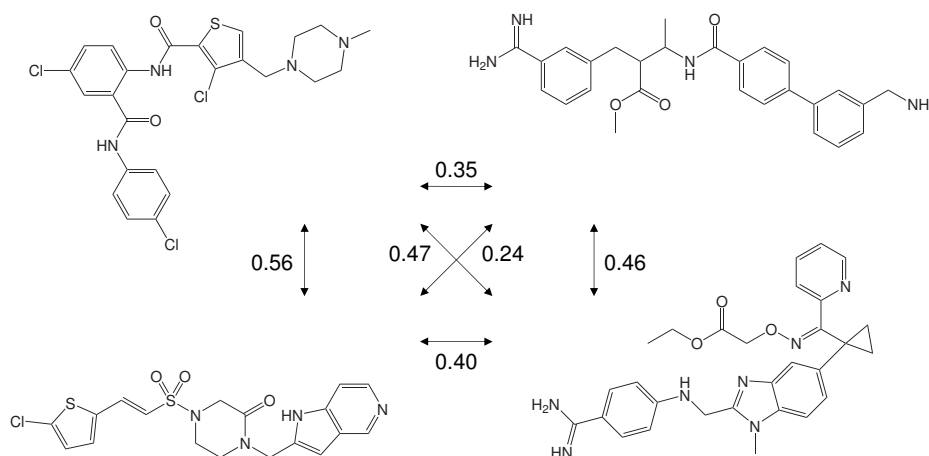


Figure 3.4: Diverse inhibitors of factor Xa Shown are potent factor Xa inhibitors that represent different chemotypes. The numbers report pairwise MACCS Tc similarity values.

0.98, and the resulting SARI scores from 0.07 to 0.89. Three sets of inhibitors produced SARI scores smaller than 0.3 and two other classes had SARI scores greater than 0.8. A total of 8 of our 16 classes fell into an intermediate scoring range between 0.39 and 0.61. In the following, we discuss the results of our analysis for four exemplary inhibitor classes that are representative of the four SAR categories described above.

Factor Xa The distribution of pairwise similarity and potency differences shown in Figure 3.3(a) reveals that studied factor Xa inhibitors cover a wide similarity range of Tc values from 0.2 to 1. As indicated by red data points, pairs of highly potent compounds essentially cover the entire similarity range and are also found in the lower left part of the plot where structurally diverse compound pairs are located. Furthermore, potency differences between similar compounds are only minor, as reflected by the absence of data points in the upper right portion of the distribution plot. These findings are well in accord with the high continuity (0.71), low discontinuity (0.08) and consequently high SARI (0.82) scores that clearly indicate a continuous SAR. Accordingly, this class of factor Xa inhibitors is characterized by high structural diversity among highly potent molecules, as illustrated in Figure 3.4 that shows a spectrum of diverse inhibitors all of which have nanomolar potency. Moreover, structurally similar compounds also have similar potency. This set of factor Xa inhibitors corresponds to the factor Xa data set discussed in Chapter 2, and that the qualitative characterization provided there is also reflected by SARI score calculations.

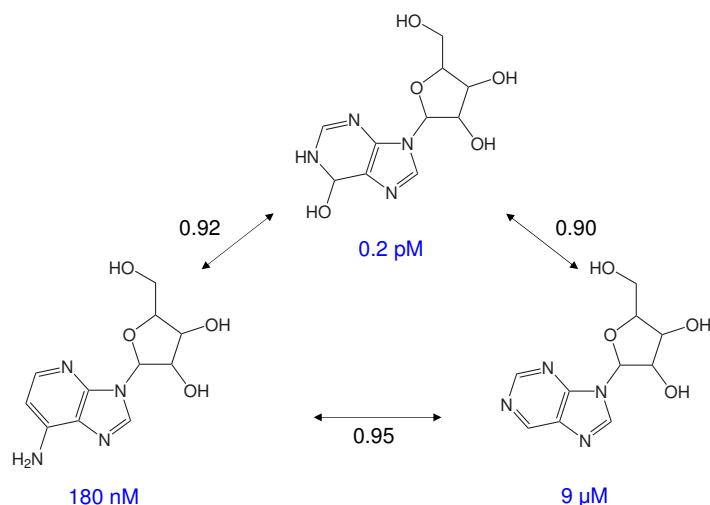


Figure 3.5: Adenosine deaminase inhibitors Shown are three analogous inhibitors that have potency differences of several orders of magnitude. Pairwise MACCS Tc similarity values and potency values are reported (black and blue numbers, respectively).

Adenosine Deaminase In marked contrast to factor Xa, the adenosine deaminase inhibitor set produced a very low continuity score of 0.13, a discontinuity score close to 1 (0.98) and a SARI score of 0.07, revealing a strongly discontinuous SAR. Within this class, many potent inhibitors are structurally similar, but compound pairs with low similarity ($T_c < 0.5$) have only moderate potency in the micromolar range. In Figure 3.3(b), this is mirrored by the unbalanced distribution of red data points that correspond to highly potent compound pairs. This observation is reflected by the low continuity score. The SAR is dominated by an activity cliff that results from the requirement to coordinate a zinc cation in the active site of the enzyme. Potent inhibitors fulfill this requirement by means of a hydroxyl group which increases potency dramatically, as illustrated in Figure 3.5. Potency differences between similar compounds can amount to several orders of magnitude. This discontinuous SAR is recognized by SARI calculations only on the basis of 2D structural information and potency values of inhibitors.

Thromboxane Synthase In contrast to the inhibitor classes discussed above, the set of thromboxane synthase inhibitors is characterized by an intermediate SARI score of 0.54, which results from high continuity (0.84) and discontinuity (0.75) scores. These values suggest that continuous and discontinuous elements coexist within the activity landscape of this inhibitor class. The presence of SAR continuity can be appreciated in Figure 3.3(c) where highly potent inhibitor pairs (red points) accumulate at lower similarity levels. However, there is also a significant number of similar structures having large differences in

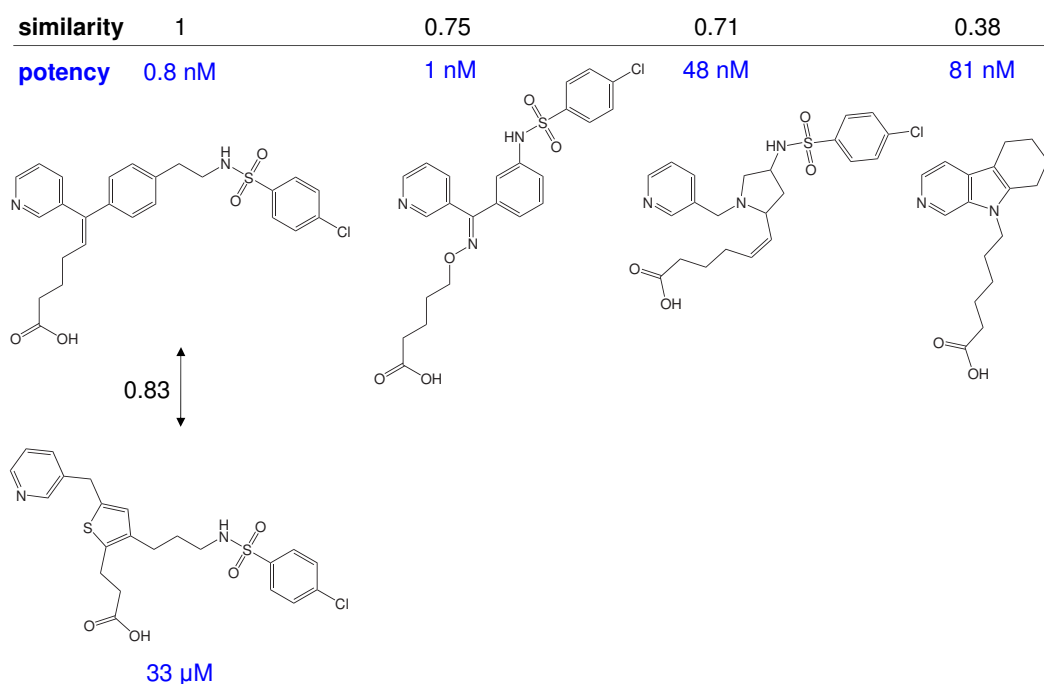


Figure 3.6: Heterogeneous SAR for thromboxane synthase inhibitors The top row presents inhibitors with decreasing similarity and potency values, taking the first compound as a reference. The molecule at the bottom is structurally similar to the reference compound but has significantly lower potency.

potency, located in the upper right part of Figure 3.3(c). Accordingly, the SAR also displays a detectable degree of discontinuity. The individual SAR elements are illustrated in Figure 3.6. As an example of continuous SARs, the upper part of the figure presents a series of thromboxane synthase inhibitors for which similarity and potency are gradually decreasing, considering the leftmost molecule as a reference. By contrast, at the bottom, a molecule is depicted that is closely related to the reference molecule but has significantly lower potency, which is characteristic of a discontinuous SAR. Thus, in the case of thromboxane synthase, different continuous and discontinuous SAR components coexist. This enzyme tolerates different types of small molecule SARs and presents a prototypic example for heterogeneous-relaxed SARs.

Carbonic Anhydrase The set of carbonic anhydrase inhibitors, which corresponds to the data set discussed in Chapter 2, also represents a heterogeneous SAR. With a SARI score of 0.61, it tends toward the continuous value range. However, the intermediate SARI score of carbonic anhydrase inhibitors is the result of low continuity (0.27) and low discontinuity (0.06) scores. Figure 3.3(d) shows that inhibitor pairs with high potency have relatively high similarity val-

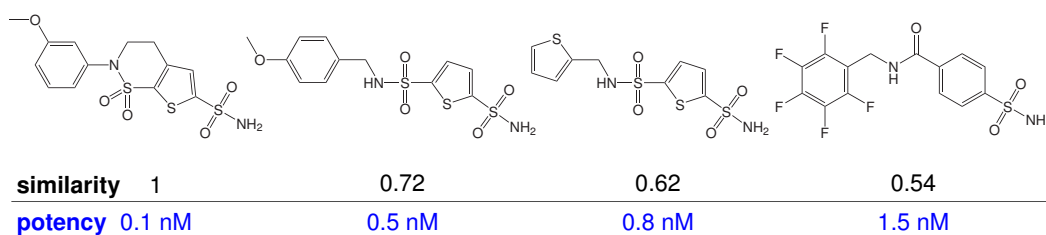


Figure 3.7: Continuous SAR for sulfonamide-inhibitors of carbonic anhydrase Taking the leftmost compound as a reference, MACCS Tc similarity and potency gradually decrease from left to right.

ues within the range of 0.5 to 1. Moreover, the lower left and upper right portions of the plot, which are responsible for SAR continuity and discontinuity, are essentially empty. Although perhaps puzzling at first glance, this SAR phenotype can be well rationalized: it is characterized by SAR continuity within the boundaries of a structural constraint. As discussed in Chapter 2, the major determinant for carbonic anhydrase inhibition is the presence of a sulfonamide group that complexes a zinc cation in the enzyme's active site, similar to adenosine deaminase inhibition discussed above. However, in contrast to adenosine deaminase inhibitors, the studied carbonic anhydrase inhibitors display significant scaffold diversity. Moreover, similar compounds always have comparable potency, resulting in the absence of activity cliffs. This is illustrated in Figure 3.7, which reveals a continuous SAR for diverse sulfonamide-containing inhibitors with potency in the low (sub-)nanomolar range.

The heterogeneous SAR exemplified by the carbonic anhydrase inhibitor set is distinct from the heterogeneous SAR phenotype presented by thromboxane synthase inhibitors. In the latter case, different continuous and discontinuous SARs coexist, whereas in the case of carbonic anhydrase, a continuous SAR is observed within the limits of a structural constraint. In contrast to heterogeneous-relaxed SARs, the characteristic features of heterogeneous-constrained SARs are low continuity and low discontinuity scores. Both subtypes of heterogeneous SARs are clearly distinguished on the basis of SARI analysis.

3.2.3 Discussion

SAR profiling of 16 inhibitor classes directed at diverse target enzymes has shown that the SARI formalism is capable of numerically characterizing SAR elements prevalent in sets of active compounds. Relying solely on 2D similarity and potency values, SARI calculations can quantitatively distinguish between different SAR categories and provide a framework for the compari-

son of compound activity classes on the basis of their SAR character. Thus, these calculations are useful to classify SARs on a large scale. The majority of compound classes investigated in this initial analysis produced intermediate SARI scores that are indicative of heterogeneous SARs. These findings are consistent with earlier proposals that many small molecule SARs should be heterogeneous in nature (Eckert and Bajorath, 2007) and confirm the conclusions that were drawn from the qualitative SAR study presented in Chapter 2. Moreover, SARI analysis makes it possible to further divide heterogeneous SARs into two previously unobserved categories, heterogeneous-relaxed and heterogeneous-constrained, that are distinguished by the magnitude of continuity and discontinuity scores and reflect different activity landscape topology. Using the SARI scoring scheme, different SAR characteristics are identified that are consistent with qualitative observations.

3.3 Control Calculations

Encouraged by our initial findings, we applied the SARI scoring scheme on another set of activity classes. The aim of these calculations was to provide a sound data basis in order to further establish the SARI formalism also on larger data sets of different composition. In addition, it is evident that the SARI formulation depends on parameters that are generally critical for the study of small molecule SARs, in particular, the representation of molecular structure and the composition of the compound sets under investigation. To evaluate the influence of these parameters on SARI scoring, a number of control calculations were carried out using alternative fingerprint representations and compound data sets of varying size.

3.3.1 Data Sets

For the analysis, we assembled compound classes from the MDDR that were of larger size and (as a consequence) of more inhomogeneous composition than the data sets analyzed in the initial study. As summarized in Table 3.3, selected compound sets are active against a variety of targets and include between 71 and 252 molecules. SARI calculations using these 13 activity classes as reference for normalization show that these classes cover a broad spectrum ranging from discontinuous to heterogeneous and continuous SARs (Table 3.4). Similar to the compound sets discussed above, a multitude of activity classes fall into the intermediate value range and accordingly have heterogeneous SARs.

3.3.2 Fingerprint Dependence

Considering the fact that SAR descriptions generally depend on the nature of the chosen molecular representation, we compared MACCS-based results with two different molecular fingerprints, Molprint2D (Bender et al., 2004) and TGT. Molprint2D generates molecular representations based on layered atom environments, whereas TGT is a topological three-point pharmacophore fingerprint implemented in MOE. Pairwise compound similarity was calculated using the Tc on each of the alternative fingerprint representations, and similarity thresholds for SARI discontinuity scores were adjusted to the similarity distribution for the individual fingerprints. SARI scores and individual score components for the different fingerprints are presented in Table 3.4. We observe that for many activity classes, SARI scores calculated on the basis of different fingerprints yield comparable results and are overall well correlated (MACCS–TGT: $r = 0.85$, MACCS–Molprint2D: $r = 0.63$, Molprint2D–TGT: $r = 0.77$). The majority of inhibitor sets are assigned to the same SAR category according to SARI scores for at least two fingerprints. For classes ACH, FAR, LIP, and THR, for example, the SAR type remains invariant for all three alternative

Table 3.3: Enzyme inhibitor classes used for control calculations

class	target	cpds	MACCS Tc			Potency [nM]	
			min	max	avg	min	max
5HT	5-HT transporter	129	0.12	1	0.46	0.01	2 700
ACA	ACAT	195	0.11	1	0.45	0.26	120 000
ACH	acetylcholinesterase	112	0.15	1	0.46	0.02	85 000
COX	cyclooxygenase 2	149	0.05	1	0.45	0.09	50 000
ELA	elastase	92	0.12	1	0.48	0.007	6 000
FAR	farnesyl transferase	146	0.01	1	0.45	0.036	304 000
FXA	factor Xa	152	0.11	1	0.50	0.007	30 000
HIV	HIV-1 protease	179	0.14	1	0.53	0.000014	43 000
LIP	lipoxygenase	252	0.02	1	0.36	1	100 000
PH4	phosphodiesterase IV	209	0.11	1	0.45	0.0025	348 000
PH5	phosphodiesterase V	71	0.26	0.99	0.56	0.006	1 000
SQA	squalene synthase	71	0.08	1	0.44	0.071	500 000
THR	thrombin	172	0.14	1	0.55	0.0019	30 000

A set of 13 enzyme inhibitor classes is summarized. Column ‘class’ provides an identifier code and ‘cpds’ reports the number of compounds for each activity class. The distribution of MACCS Tc similarity and potency values is given in the following columns (‘min’ stands for minimum, ‘max’ for maximum and ‘avg’ for average).

fingerprints. In other cases, however, the scores differ significantly, e.g. in the case of PH5 that is classified into three different categories according to SARI scores based on MACCS, Molprint2D and TGT. These in part substantial differences can be attributed to the different design and resolution of the utilized fingerprints.

3.3.3 Influence of Compound Set Size

A second set of control calculations was carried out in order to assess how composition and size of the data sets might influence SARI scores. For this purpose, we randomly extracted compound subsets of increasing size from the 13 activity classes described above and calculated SARI scores for these random samples. From each of the activity classes, we successively selected subsets of 10, 20 and 50 compounds. For classes consisting of more than 100 (or 200) compounds, also subsets of 100 (and 200) molecules were sampled. Beginning with a compound subset of size 10, the subsets were incrementally extended by randomly adding compounds. The subset selection process was independently repeated 10 times for each activity class. SARI scores were then calculated for each compound subset and averaged over the 10 subsets of a given size. Fig-

Table 3.4: SARI scores for different fingerprint representations

class	MACCS			Molprint2D			TGT		
	cont	disc	SARI	cont	disc	SARI	cont	disc	SARI
5HT	0.55	0.22	0.67	0.77	0.40	0.68	0.91	0.27	0.82
ACA	0.57	0.28	0.65	0.18	0.26	0.46	0.20	0.21	0.49
ACH	0.62	0.72	0.45	0.53	0.64	0.44	0.72	0.61	0.56
COX	0.74	0.21	0.76	0.51	0.24	0.63	0.56	0.23	0.66
ELA	0.36	0.59	0.39	0.77	0.48	0.65	0.64	0.54	0.55
FAR	0.58	0.71	0.44	0.58	0.64	0.47	0.76	0.81	0.47
FXA	0.30	0.27	0.52	0.62	0.34	0.64	0.21	0.39	0.41
HIV	0.12	0.53	0.30	0.01	0.48	0.26	0.07	0.48	0.30
LIP	0.99	0.04	0.97	0.95	0.16	0.89	0.91	0.06	0.92
PH4	0.66	0.59	0.54	0.62	0.35	0.63	0.52	0.40	0.56
PH5	0.08	0.53	0.27	0.76	0.33	0.71	0.36	0.38	0.49
SQA	0.79	0.99	0.40	0.33	1.00	0.17	0.79	0.99	0.40
THR	0.08	0.67	0.21	0.16	0.57	0.30	0.03	0.77	0.13

For three different fingerprints, the continuity ('cont'), discontinuity ('disc'), and SARI scores are reported for 13 activity classes. Class identifiers are according to Table 3.3.

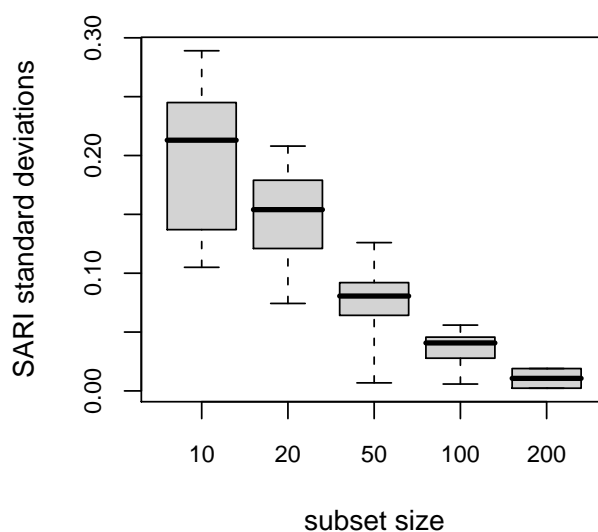


Figure 3.8: Standard deviations of SARI scores for compound subsets of increasing size
The distribution of standard deviations of the SARI score calculated on random subsets taken from 13 activity classes is shown as box plot. From each of the 13 activity classes, 10 random subsets of a given size were sampled. Each box represents the distribution of standard deviations for the scores calculated on compound subsets of a given size for each class. The bottom and top of each box report the lower and upper quartile, and the dashed lines mark the most extreme data points. The median of the distribution is represented as thick horizontal line.

Figure 3.8 reports the distribution of SARI standard deviations for the individual subset sizes in the 13 activity classes. It should be noted that for subset size 100, classes ELA, PH5 and SQA did not contain sufficient compounds and for subset size 200, only classes LIP and PH4 were considered. The figure shows that standard deviations of SARI scores are consistently lower than 0.3 already for small sets of 10–20 compounds and rapidly decrease when compound subsets are enlarged. Although large score variations might be observed in individual cases, these findings suggest that SARI scores calculated for subsets of varying size remain essentially stable.

3.3.4 Discussion

Application of the SARI scoring scheme to a second set of inhibitor classes further established the methodology for quantitative assessment of SAR characteristics and demonstrated its applicability also to large and diverse data sets. Consistent with our initial study, the heterogeneous SAR type was found to be prevalent among the analyzed data sets. Control calculations suggested that SARI scores are remarkably robust with respect to variations of molecular representation and data set size, which are critical parameters for the assess-

ment of SARs. Scores for three different fingerprint representations were compared and found to produce similar results in many cases. However, alternative fingerprints generally capture distinct molecular features and yield different levels of resolution, which might substantially affect SARI calculations in individual cases. Accordingly, molecular representations should be chosen that consistently describe the activity landscape of a given compound set and enable chemically meaningful similarity assessment. Furthermore, the influence of data set size on SARI scoring was investigated. The analysis showed that SARI scores calculated on compound subsets of increasing size remained to a large extent stable. Hence, SARI calculations yield meaningful results also for very small compound data sets. It should be noted, however, that SARI scores reflect the composition and SAR features of a given data set and cannot necessarily extrapolate or predict different sets of compounds sharing the same biological activity.

3.4 Related Methods

Quantitative assessment of SARs in sets of active compounds is still in its infancy, and only few methods have been reported thus far. Earlier approaches include Structure–Activity Similarity (SAS) maps that compare biologically active compounds in a pairwise manner and relate similarity in potency and structural similarity to each other, such as the similarity–potency plots described above (Shanmugasundaram and Maggiora, 2001). An information-theoretic measure was used to compare similarity–potency relationships within an SAS map to idealized activity landscapes in order to estimate the “smoothness” or “roughness” of the activity landscape under investigation. A more recent approach has been presented by Houghten and coworkers that also utilized similarity–potency plots for the detection of “consensus activity cliffs” on the basis of different 2D and 3D similarity methods (Medina-Franco et al., 2009).

With the Structure–Activity Landscape Index (SALI), another scoring function for the quantification of SAR features has been introduced (Guha and van Drie, 2008). In order to account for activity cliffs of varying magnitude, SALI characterizes pairs of compounds by means of their pairwise potency difference divided by compound similarity. SALI scores can be visualized in a so-called SALI graph that represents compounds as nodes that are connected by edges if their SALI score exceeds a user-defined threshold value. Conceptually, SALI resembles the discontinuity score of the SARI framework insofar as both methods aim at the identification of activity cliffs through assessment of potency differences of similar compounds. Regardless of differences in their design, all of these methods can be traced back to the analysis of “neighborhood behavior”, which describes how changes in descriptor settings or molecular representations

relate to changes in the biological activity of test compounds (Papadatos et al., 2009; Patterson et al., 1996).

3.5 Conclusions

With SARI and its individual score components, we have introduced an approach to quantitatively describe the nature of SARs. SARI provides a framework to classify the SAR character with compound activity classes and also makes it possible to compare SARs between different compound classes. Adopting a global view on SARs in compound activity classes, the scoring scheme departs from the traditional case-by-case study of SARs and enables their analysis on a large scale. Three principal SAR types that have long been recognized based on qualitative evidence are for the first time described in numerical terms: continuous, discontinuous, and heterogeneous SARs. Moreover, SARI calculations distinguish between two previously unobserved subtypes of heterogeneous SARs that reflect different composition of continuous and discontinuous elements. Profiling of various activity classes has shown that many small-molecule SARs are heterogeneous in nature, which is consistent with earlier observations and has practical relevance for medicinal chemistry. The heterogeneous-relaxed SAR phenotype is considered particularly attractive for compound screening and chemical optimization efforts because it is likely that structurally diverse active compounds can be identified (in continuous SAR regions) and also optimized (if they map to the vicinity of activity cliffs). Taken together, our findings suggest that SARI presents a simple and robust method for the numerical assessment, classification and comparison of structure–activity relationships within sets of biologically active molecules.

Chapter 4

Analysis of Global and Local Structure–Activity Relationships

The qualitative and quantitative characterization of structure–activity relationships has demonstrated that many activity landscapes are heterogeneous in nature and often contain regions of fundamentally different SAR character. The SARI scoring scheme presented in the previous chapter permits global assessment of SARs in compound activity classes and enables their comparison between different classes. However, this method cannot be applied to study multiple SAR features contained within a set of active compounds at the level of compound subsets or individual molecules. Open questions include, for example: Can we systematically identify subsets of compounds that display different SAR behavior? How are local and global SAR elements related to each other? How do individual compounds influence global SARs?

In order to dissect activity landscapes and analyze multiple SAR components of compound classes with different SAR character, we have developed a SARI score variant that is capable of accounting for SAR contributions from individual compounds. In addition, this chapter introduces Network-like Similarity Graphs (NSG) that provide a detailed graphical representation of potency and similarity relationships within sets of active compounds (Wawer et al., 2008). In computational medicinal chemistry, molecular network representations have previously been used to represent target–ligand relationships (Mestres et al., 2006; Paolini et al., 2006) or relationships between different classes of drug molecules (Hert et al., 2008), among other applications. In this chapter, we utilize NSG representations and SARI scoring on the basis of individual compounds and compound subsets to describe different SAR features that coexist in compound activity classes. This approach makes it possible to better understand how local SAR characteristics are related to each

other and identify individual molecules that are SAR determinants. The design of NSGs and compound SARI scores is described in detail in Section 4.1. The methodology is applied in Section 4.2 to thoroughly analyze SARs in six representative compound sets and utilized also for the characterization of more complex SARs, as discussed in Section 4.3 for an exemplary screening data set.

4.1 Methodology

In order to characterize and compare global and local SAR elements, SARI scoring was applied at three different levels of detail. SARI scores were calculated for entire compound activity classes, for compound subsets identified through similarity-based clustering, and on the basis of individual molecules. NSG representations were designed to visualize relationships between SARs at these different levels.

4.1.1 Compound Clustering and Cluster Scoring

For the identification and characterization of multiple local SARs, activity classes were divided into subsets of similar molecules. For this purpose, the molecules of an activity class were subjected to hierarchical clustering using their pairwise MACCS Tc similarity values and Ward’s minimum variance linkage method (Ward, 1963), which yielded intuitive cluster distributions for our data sets. The resulting cluster dendrograms were pruned at heights between 1 and 2 for different classes to obtain clusters of reasonable size and constitution. For each compound cluster, SARI discontinuity scores were calculated as described in Section 3.1 in order to estimate subset-dependent SAR features. High discontinuity score values indicated subsets with a high degree of local SAR discontinuity including similar molecules with significant potency differences. For our analysis on the level of compound subsets, the continuity score was not considered because it was designed to capture structural diversity, which is primarily a feature of global SARs.

4.1.2 Compound Discontinuity Scores

In order to estimate the contributions that individual compounds make to global SAR discontinuity, we developed a variant of the SARI discontinuity score calculated on a per compound basis. The aim was to focus on compounds responsible for introducing activity cliffs in an activity landscape; hence, local continuity score calculations were not required for our analysis. The compound discontinuity score was designed to account for potency differences between a given active molecule and all molecules that are similar to it, again applying

a MACCS Tc similarity threshold of 0.65. In contrast to global SARI calculations for compound classes or clusters, no potency difference cutoff is required here because for the assessment of discontinuity contributions from individual compounds, all potency differences among similar compounds must be taken into account. For a given molecule i in the activity class A , the compound discontinuity score is defined as

$$\text{disc}_{\text{raw}}(i) = \text{mean}_{\{j \in A | j \neq i, \text{sim}(i,j) > 0.65\}} (|P_i - P_j| \cdot \text{sim}(i, j)) \quad (4.1)$$

This function assigns high scores to molecules that have significantly different potency from their neighbors and are involved in the formation of activity cliffs.

4.1.3 Score Normalization

Global and local discontinuity scores were normalized to adopt values between 0 and 1 by calculation of Z-scores and cumulative distribution functions, as described in Section 3.1. However, we utilized different reference values for standardization of both score variants. Global SARI scores calculated on entire activity classes were normalized with respect to the score distribution within the set of 13 MDDR activity classes presented in Section 3.3. The same reference set was also used for normalization of discontinuity scores calculated for compound clusters. This common normalization reference made it possible to directly compare local cluster discontinuity scores to global scores for an entire activity class and also across different classes. By contrast, discontinuity scores calculated for individual compounds were standardized relative to all compound scores within the same activity class. Hence, key compounds making largest SAR contributions in a given activity class could be readily identified; however, this design does not permit the comparison of compound scores across different classes.

4.1.4 Network-like Similarity Graphs

Similarity and potency relationships within an activity class were visualized using NSGs. In these graphs, compounds are represented by nodes (circles), and edges (lines) between them display similarity relationships. Figure 4.1 shows a schematic representation of an NSG and the information it conveys. Five different levels of information can be distinguished. Firstly, similarity relationships between molecules are reflected by edges that connect two nodes in an NSG if the corresponding molecules exceed a MACCS Tc similarity threshold of 0.65. Secondly, the potency distribution is represented by node colors. Nodes are color-coded according to the pK_i or pIC_{50} values of the corresponding compounds using a color gradient from green via yellow to red, with green

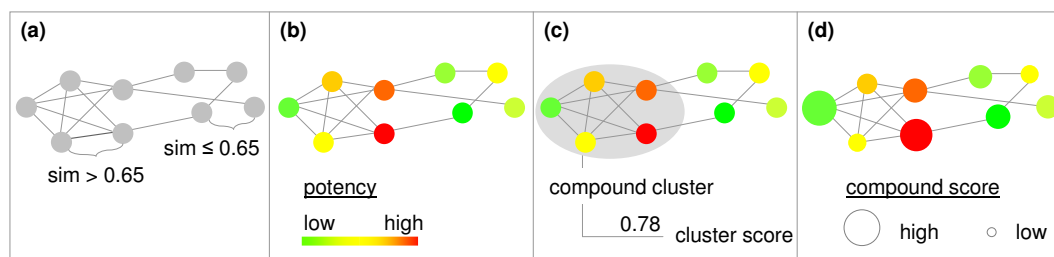


Figure 4.1: Schematic representation of NSG information levels (a) Nodes represent compounds and are connected by edges if their MACCS Tc similarity exceeds a predefined threshold value. (b) Nodes are color-coded according to potency. (c) Compounds are clustered based on pairwise similarity values. For compound clusters, SARI discontinuity scores are calculated. (d) Compound discontinuity scores are calculated and nodes are scaled in size according to the magnitude of the scores.

indicating lowest and red highest potency within a class. A third level of information is presented by compound clusters that indicate subsets of similar molecules. It should be noted that the applied clustering algorithm might assign compounds to the same cluster even if they are not connected by an edge (because their similarity value is below the threshold), and compounds that are connected by an edge can be assigned to different clusters. Hence, compound clusters complement the binary similarity information provided by edges and signify compound subsets that are associated by remote similarity relationships. The fourth level of information is provided by local discontinuity scores in compound clusters. In NSGs, clusters are annotated with their discontinuity score, which makes it possible to highlight regions of different SAR character within a compound set. Finally, presenting the fifth level of information, compound discontinuity scores reveal contributions to overall SAR discontinuity made by individual compounds. Nodes in an NSG are scaled in size according to compound discontinuity scores, with the largest nodes corresponding to compounds that make most significant contributions to global SAR discontinuity in a class.

NSGs were calculated and displayed using the R *igraph* package (Csardi and Nepusz, 2006). The layout of NSGs was calculated on the basis of node connectivity using the Fruchterman–Reingold algorithm (Fruchterman and Reingold, 1991). Accordingly, distances between nodes are not scaled by similarity values but rather indicate how densely nodes within regions of a network are connected by edges.

4.2 Analysis of Network-like Similarity Graphs

Our analysis focused on how to identify SAR features that coexist in activity classes and explore potential relationships between them at the level of com-

Table 4.1: Global and local SAR character for different enzyme inhibitor classes

class	cpds	clusters	global scores			cluster scores		
			cont	disc	SARI	min	max	avg
LIP	252	11	0.99	0.04	0.97	0.02	0.15	0.06
COX	149	7	0.74	0.21	0.76	0.00	0.70	0.26
FXA	152	5	0.30	0.27	0.52	0.02	0.45	0.26
FAR	146	8	0.58	0.71	0.44	0.01	1.00	0.59
SQA	71	8	0.79	0.99	0.40	0.00	1.00	0.48
THR	172	6	0.08	0.67	0.21	0.01	0.80	0.53

For six sets of enzyme inhibitors, the global continuity ('cont'), discontinuity ('disc'), and SARI scores are reported. In addition, cluster discontinuity score distributions are given ('min', minimum; 'max', maximum; 'avg', average scores). Columns 'cpds' and 'clusters' report the number of compounds and clusters in each class, respectively. Class identifiers are according to Table 3.3.

pound subsets and individual molecules. Therefore, we calculated global and local SARI scores and generated NSG representations for the set of 13 activity classes presented in Section 3.3. The size and heterogeneity of these classes made them instructive test cases for the application of the NSG-SARI methodology. In the following, we discuss the results for six activity classes representing different SAR categories: LIP and COX (continuous), FXA (heterogeneous-constrained), FAR and SQA (heterogeneous-relaxed), and THR (discontinuous). Table 4.1 reports their global SARI scores and the distribution of cluster discontinuity scores.

4.2.1 Network Topology

The topology of NSGs is determined by pairwise similarity relationships and their distribution within an activity class. Figure 4.2 shows that the six exemplary classes studied here produce NSGs of different topology and that distinct topologies are also observed for compound sets belonging to the same global SAR category. For example, the LIP and COX compound sets display globally continuous SARs, as indicated by their high SARI scores. With an average pairwise MACCS Tc of 0.36, LIP is characterized by a high degree of intra-class structural diversity, which is clearly reflected in the topology of the corresponding NSG. The network consists of several distinct subgraphs, and many nodes are only sparsely connected. However, a number of densely connected clusters with very low discontinuity scores are also observed (Figure 4.2(a)). Thus, the high degree of structural diversity within this activity class partly results from

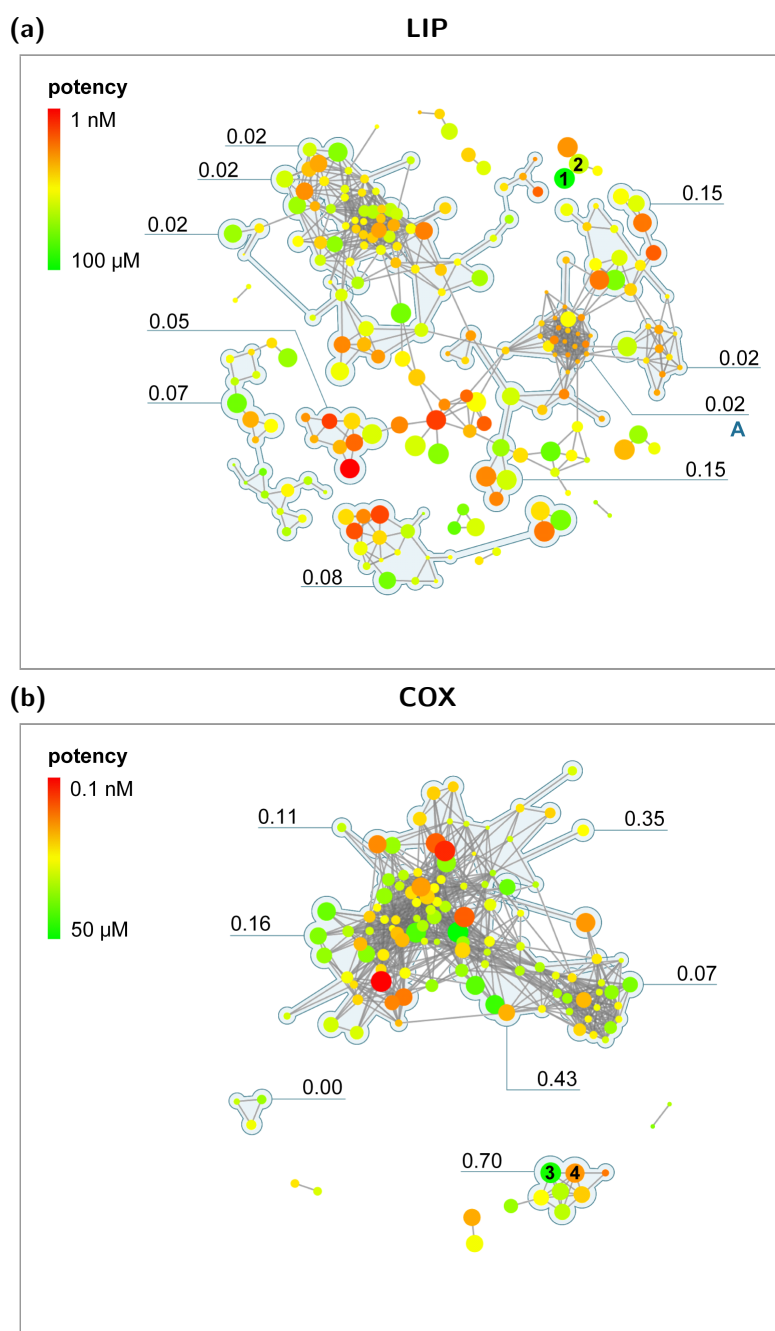


Figure 4.2: NSG representations for six classes of enzyme inhibitors

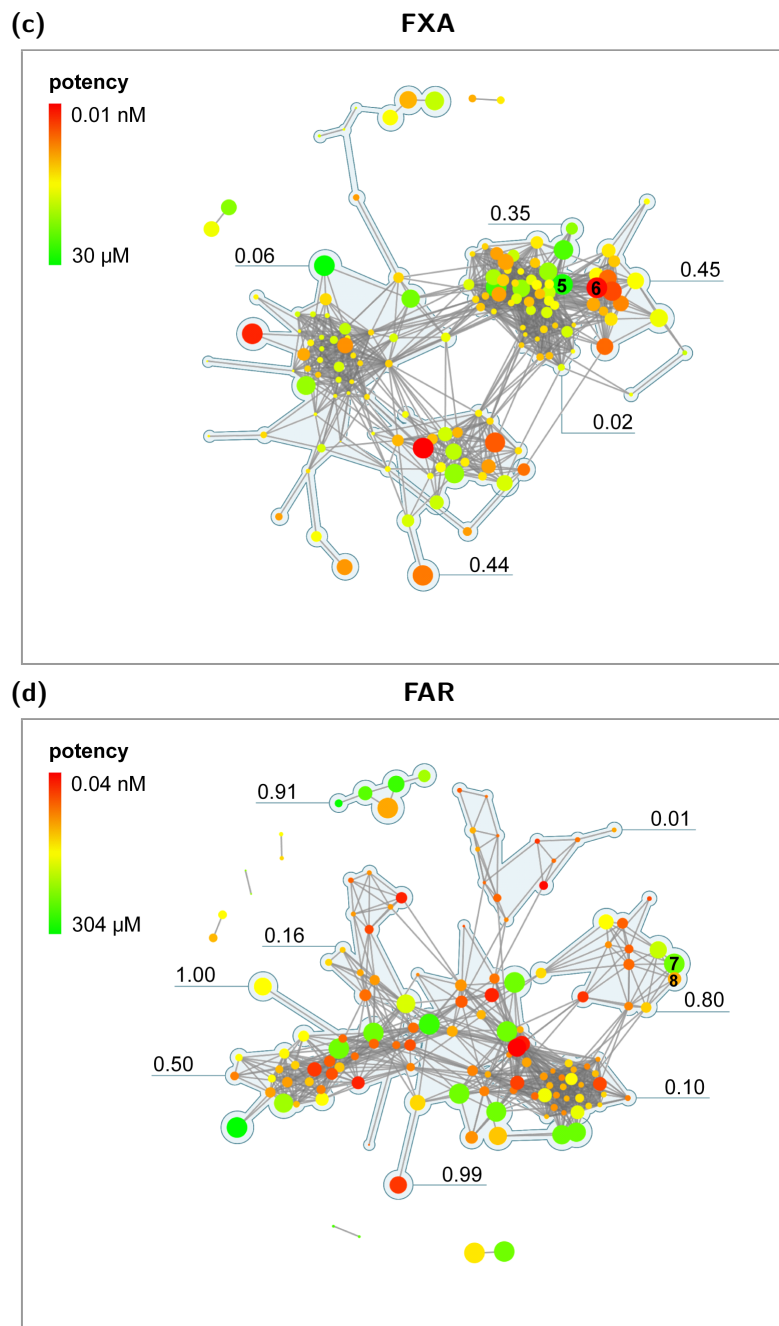


Figure 4.2: NSG representations for six classes of enzyme inhibitors (continued)

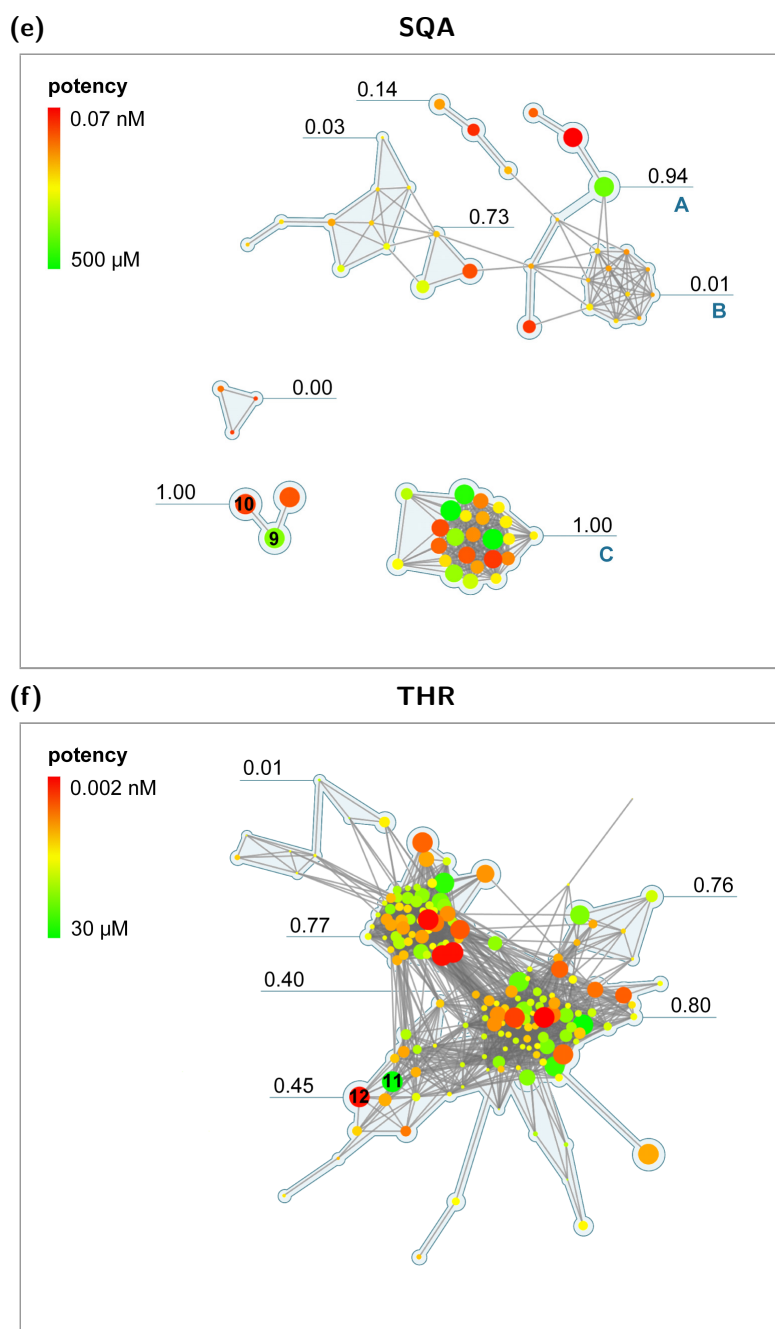


Figure 4.2: NSG representations for six classes of enzyme inhibitors (continued) Nodes represent compounds and edges are drawn between them if pairwise MACCS Tc values exceed 0.65. Nodes are color-coded according to potency using a continuous spectrum from green to red and scaled in size according to their compound discontinuity scores. Compound clusters are displayed on a light blue background and annotated with their discontinuity scores. Clusters and key compounds discussed in the text are labeled with capital letters and numbers, respectively. Unconnected nodes are omitted for clarity.

the presence of chemically different series of active compounds. Compared to LIP, COX is less structurally diverse (average MACCS $T_c = 0.45$) and its NSG is overall more densely connected (Figure 4.2(b)). Most compounds are organized in a large major network component, and only a few peripheral clusters can be found.

In heterogeneous-constrained SAR types, by contrast, the degree of structural diversity is typically limited. Structural variations of active compounds occur within the boundaries of activity cliffs and are frequently accompanied by only minor potency changes, giving rise to continuous SARs within the limits of a discontinuous one. Often, functional groups involved in key interactions are conserved, while other molecular regions are more variable. The FXA inhibitor set studied here belongs to this category, as indicated by the low values for global continuity and discontinuity scores and an intermediate SARI score. As illustrated in Figure 4.2(c), its NSG consists of a few densely connected components, reflecting the relatively low degree of structural diversity in this class (average MACCS $T_c = 0.50$). By contrast, heterogeneous-relaxed SARs are marked by the coexistence of continuous and discontinuous SAR components and compound classes belonging to this category are often structurally diverse, similar to continuous compound sets. This SAR type also produces an intermediate SARI score but is distinguished from its heterogeneous-constrained counterpart by high global continuity and discontinuity score values. Class FAR provides a representative example. The structural diversity within this class is comparable to COX and so is its NSG topology (Figure 4.2(d)). SQA also belongs to the heterogeneous-relaxed SAR category and has intra-class structural diversity comparable to FAR but produces an NSG of different topology (Figure 4.2(e)). The graph consists of several distinct components that are clearly separated from each other. These subgraphs are well-defined and correspond to structurally distinct subsets of compounds that display different SAR characteristics, as indicated by their cluster discontinuity scores.

Finally, the set of THR inhibitors analyzed in this study represents the structurally most homogeneous class with an average MACCS T_c of 0.55 and displays the most discontinuous SAR character. Its NSG contains a single and densely connected major network component, which reflects a high degree of intra-class similarity. Consistent with the global SAR type, many individual compounds are found to make large local contributions to SAR discontinuity (indicated by the large size of the corresponding nodes).

In all NSGs studied here, edges are found that connect individual compounds belonging to distinct clusters with in part very different discontinuity scores, as in the case of FAR or SQA. Thus, structurally similar compounds can be identified that are involved in different local SARs and can be seen as “chemical bridges” between local SAR environments. Furthermore, all NSGs show a clear correspondence between compound clusters and graph or subgraph

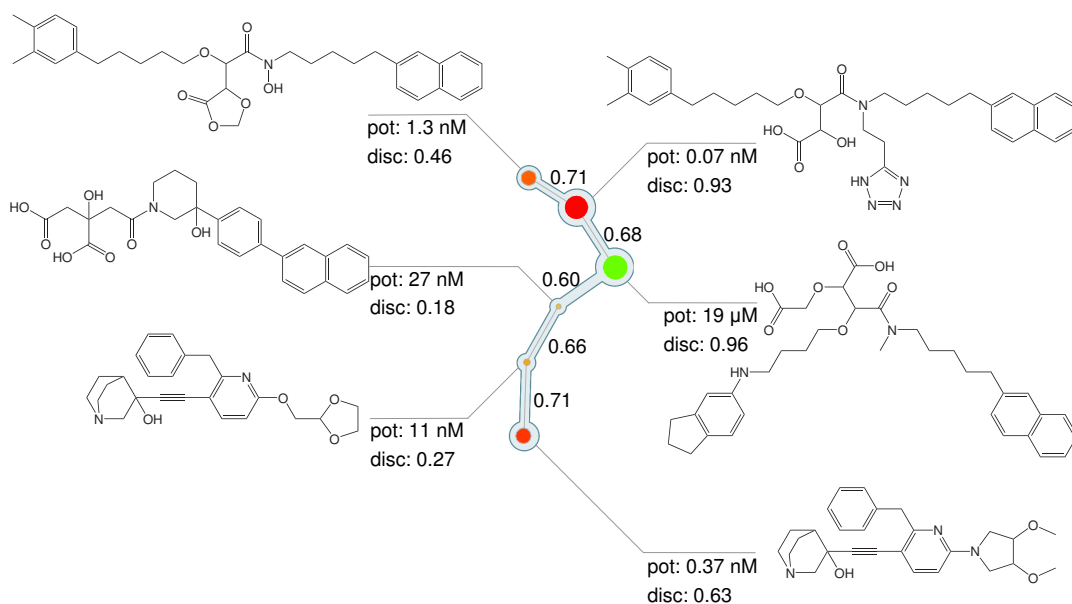


Figure 4.3: Exemplary compound cluster for class SQA Shown are the compounds belonging to cluster A in Figure 4.2(e) with their potency values ('pot') and compound discontinuity scores ('disc'). Pairwise MACCS Tc values are reported along the edges in the cluster.

communities, which are formed by topologically distinct collections of densely connected nodes. Thus, compound clusters can serve as an adequate basis for studying local SAR characteristics.

4.2.2 SARs in Compound Clusters

SARI discontinuity scores were calculated for compound clusters in order to investigate local SAR features present in subsets of similar compounds. Individual clusters can be isolated from NSGs and analyzed separately. This makes it possible to select compound subsets on the basis of their SAR characteristics and study their composition in detail. Figure 4.3 shows an exemplary cluster for class SQA. According to its high discontinuity score (0.94), it includes structurally related compounds with distinct potency differences. For a detailed analysis of this cluster, similarity relationships and the individual compound discontinuity scores are reported. Structurally related representatives of two compound series can be identified. Moreover, two compounds are found that make large local discontinuity contributions (corresponding to the large green and red nodes) and are largely responsible for the overall discontinuous SAR character within this cluster. These compounds can be considered to mark the beginning (green) and end (red) of a compound optimization pathway. Thus, detailed analysis of individual clusters makes it possible to elucidate compound

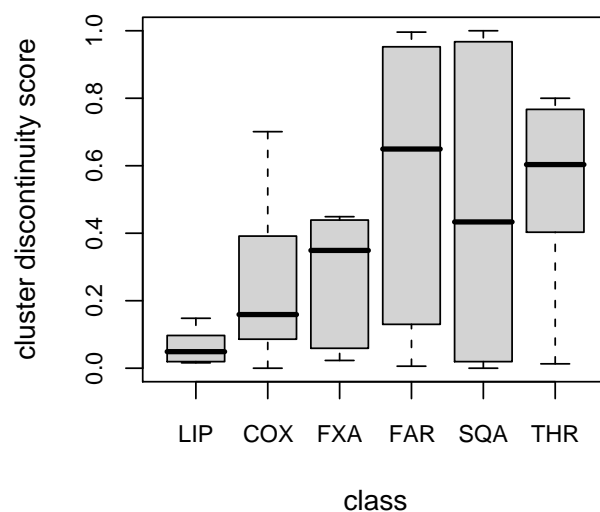


Figure 4.4: Distribution of cluster discontinuity scores For each activity class, the distribution of discontinuity scores for all clusters is presented as a box plot.

and cluster SARs and identify compounds that largely determine local and global SAR character.

4.2.3 Cluster SARs versus Global SARs

In order to relate discontinuity within individual compound subsets to global SAR features, we have calculated the distribution of cluster discontinuity scores for each activity class, as shown in Figure 4.4. Classes representing different SAR types were found to display characteristic score distributions, which becomes also apparent by comparison of the cluster scores given in Figure 4.2.

Compound classes having globally continuous or heterogeneous-constrained SARs lack steep activity cliffs. Accordingly, the cluster discontinuity scores for these classes range from low to intermediate values. Both continuous classes LIP and COX contain clusters with overall low discontinuity scores, and scores in the heterogeneous-constrained class FXA are on average only slightly higher. COX contains only one cluster of noteworthy discontinuous character; LIP none. In LIP, most compound clusters are assigned discontinuity scores at the lower end of the spectrum. For example, cluster A in Figure 4.2(a) obtains a score of 0.02. The low degree of SAR discontinuity within this cluster is a result of the presence of similarly potent compounds (indicated by small orange nodes), as illustrated in Figure 4.5. In globally discontinuous classes like THR, clusters show significantly higher discontinuity scores, as expected. THR contains a few clusters that obtain high discontinuity scores and include strong activity cliffs

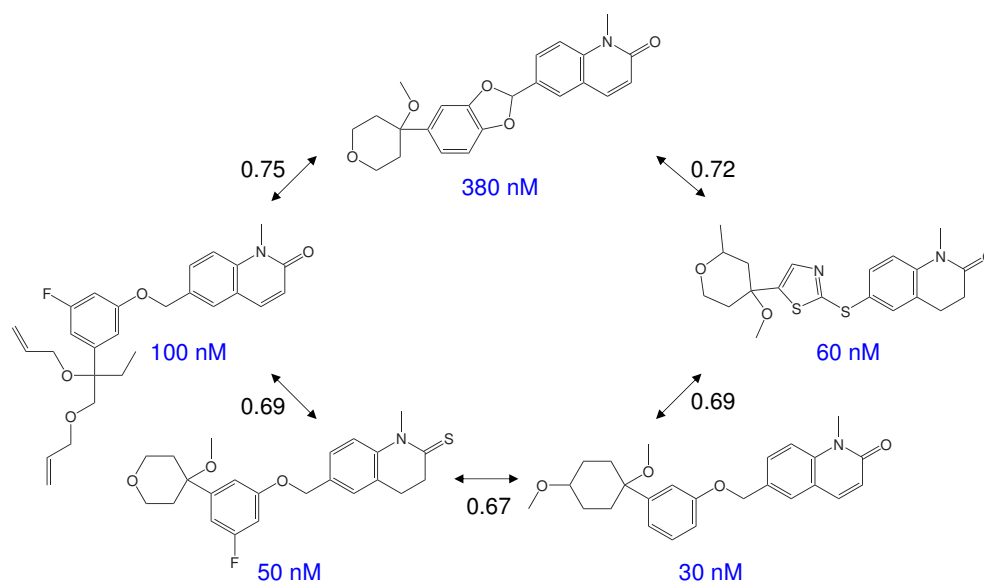


Figure 4.5: Representative compounds from class LIP forming a continuous local SAR Shown are selected compounds from cluster A in Figure 4.2(a) that represent a continuous local SAR over a relatively narrow potency range. Pairwise MACCS Tc values are reported along the arrows.

that dominate global SAR features. However, cluster scores in heterogeneous-relaxed compound classes show the largest variations, due to the coexistence of continuous and discontinuous local SAR components. In classes FAR and SQA, several clusters can be identified whose discontinuity scores range from 0 to 1. The heterogeneous SAR character can be easily discerned in the NSG of class SQA (Figure 4.2(e)) which contains clusters that correspond to continuous or even flat SARs (e.g. cluster B) or, by contrast, represent prototypic instances of a rugged activity landscape that contains similar compounds with large potency spread, corresponding to steep activity cliffs (cluster C).

4.2.4 Compound Discontinuity and Key Compounds

In order to focus on individual activity cliffs, discontinuity scores were calculated on a per compound basis. For all activity classes, highly and weakly potent compounds making large contributions to SAR discontinuity were identified, irrespective of the global SAR phenotype. Selected compound pairs are labeled in Figure 4.2 and displayed in Figure 4.6. These key compound pairs have similar structures but significant differences in potency and are thus activity cliff markers that are easily identified in NSGs as pairs of large red and green nodes. Because compound discontinuity scores are standardized relative to each individual compound class, key compounds from different classes typ-

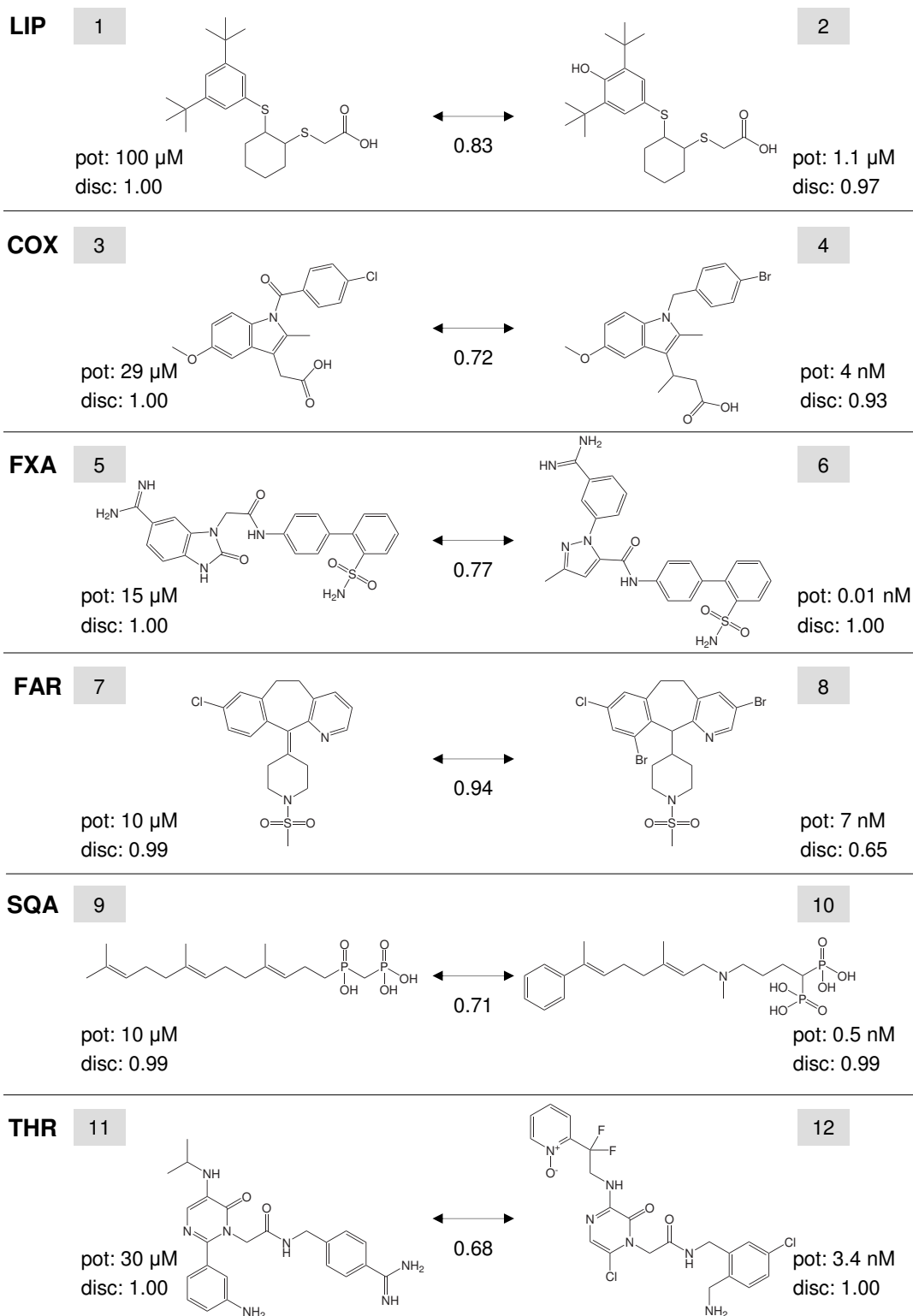


Figure 4.6: Selected key compounds Shown are compounds with high compound discontinuity score together with another high-scoring neighbor that form activity cliffs of varying magnitude. Potency values ('pot') and discontinuity scores ('disc') are reported for each molecule. The arrow annotations provide pairwise MACCS Tc similarity values for compound comparison. Compound numbers correspond to node labels in Figure 4.2.

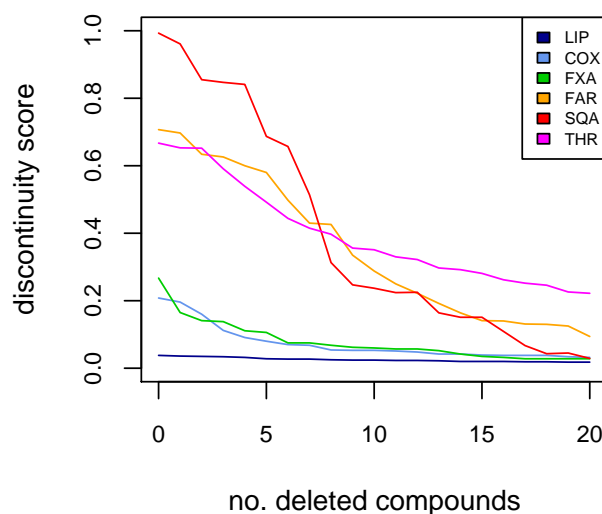


Figure 4.7: Influence of key compounds on global SAR discontinuity For each activity class, compounds with highest compound discontinuity scores were iteratively removed from the compound sets and global discontinuity scores were recalculated following each iteration.

ically represent different levels of discontinuity, depending on the overall score distribution within the given class. The ability to identify activity cliff markers for all compound classes studied here emphasizes the value of discontinuity scores on the basis of individual compounds.

In order to estimate the influence of key compounds on global SARs, molecules with highest local scores were selected and iteratively removed from activity classes. Global SARI scores and compound discontinuity scores were recalculated after each step. The results presented in Figure 4.7 show that for discontinuous and heterogeneous-relaxed SAR types the global discontinuity score constantly decreases and thus shifts global SARs more toward the continuous range. Thus, in these cases, key compounds having high discontinuity scores also strongly influence global SAR characteristics. By contrast, for continuous or heterogeneous-constrained classes, activity cliffs identified on the basis of compound scores are of moderate magnitude and accordingly have only limited influence on global SARs.

4.2.5 Summary

The quantitative assessment of global and local SARs in conjunction with graphical representation of similarity and potency distributions has made it possible to relate different SAR components to each other. The global SAR character of compound activity classes was found to be closely linked to the topology of molecular similarity graphs. Continuous or heterogeneous-relaxed

SAR phenotypes are characterized by a notable degree of intra-class structural diversity, which was reflected by the formation of distinct subgraphs in the corresponding NSG. By contrast, classes of more homogeneous composition such as globally discontinuous or heterogeneous-constrained classes produced a more densely connected network topology. Furthermore, analysis of SARs within individual compound clusters has shown a clear correspondence between cluster discontinuity score distributions and global SAR type. As expected, continuous classes contained clusters with discontinuity scores at the lower end of the spectrum, whereas discontinuous classes obtained overall highest cluster scores. Heterogeneous-relaxed classes covered the widest range of cluster scores, as a result of the coexistence of continuous and discontinuous local SARs. Discontinuity scores on a per compound basis were utilized to identify key compounds forming activity cliffs and contributing strongly to overall SAR discontinuity. In all activity classes studied here, key compounds were identified that corresponded to activity cliffs of varying magnitude. In a number of instances, these compounds were capable of shaping the activity landscape of compound sets.

4.3 Application to Screening Data Sets

The dissection of SAR phenotypes at different levels of detail is particularly valuable for large collections of active molecules that cannot be easily organized manually. Therefore, the NSG methodology was thought to be useful for the analysis of hit sets identified by high-throughput screening (HTS), which is of considerable practical relevance in drug discovery (Wawer and Bajorath, 2009). In order to investigate the utility of the NSG-SARI approach for this type of data, we applied the analysis to screening data sets extracted from PubChem BioAssay. Screening data sets principally differ from compound activity classes extracted from biologically annotated databases or lead optimization sets because they are much larger in size and contain many weakly active hits including false-positives. Taking also into account their high degree of structural diversity, one would expect that global SAR features of screening data should be highly continuous in nature. Consistent with this assumption, we observed that SARI scores calculated for screening data sets were generally high (Wawer et al., 2009). In order to explore local SAR features within the globally continuous activity landscape, NSG analysis was applied. For this purpose, the similarity threshold for discontinuity score calculation and edges was set to 0.75. Figure 4.8 shows the NSG for an exemplary cytochrome P450 inhibitor set (PubChem AID 884) containing 3439 hits. For clarity, cluster annotations were omitted in this example. With a global SARI score of 0.98, this data set displays strongly continuous SAR character. However, the network reveals a notable degree of local SAR heterogeneity. Many compounds form

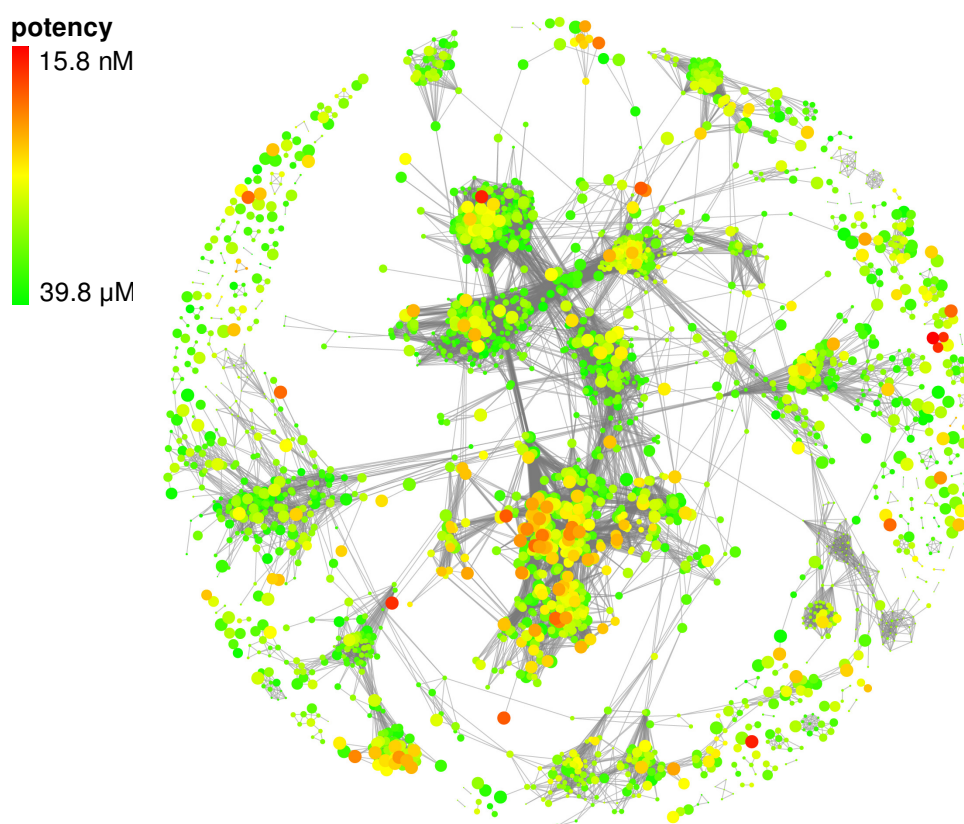


Figure 4.8: NSG representation of a cytochrome P450 inhibitor set A set of 3439 screening hits from a P450 3a4 inhibition assay are displayed in an NSG which reveals different local SAR environments.

communities that correspond to local environments with distinct SAR character. A number of potent compounds cluster in one large network component and form activity cliffs with similar compounds of lower potency. However, potent hits and activity cliffs are also found in other regions of the network. It should be noted that activity cliffs in screening data sets are mostly of smaller magnitude than those observed in optimized compound series, due to the typically narrow potency range. We also clustered the molecules according to their MACCS Tc similarity and calculated discontinuity scores for the resulting compound subsets. While the vast majority of clusters obtained scores close to 0, individual clusters were identified that displayed a considerable degree of discontinuity (with scores of 0.84 and 0.50 for the two most discontinuous clusters). Hence, despite the highly continuous global SAR characteristic of HTS data, regions of relative SAR discontinuity and even (moderate) activity cliffs were observed. These initial findings have been further extended in a systematic

manner and exploited for the automatic extraction of SAR information from compound pathways in screening data NSGs (Wawer et al., 2009).

4.4 Conclusions

A comparative analysis of potency and similarity distributions among different classes of active compounds has been carried out in order to study SARs at varying levels of detail. SARI scores have been calculated on the basis of compound activity classes, compound subsets and individual molecules. Complemented by visualization in network-like similarity graphs, this multi-level approach makes it possible to dissect SAR phenotypes and relate local and global SAR features to each other. Compound discontinuity scores have been introduced to identify key compounds that are activity cliff markers and make strong contributions to global SAR discontinuity. These compounds can be considered SAR determinants and often present the beginning and end points of compound optimization pathways. The methodology has also been applied to analyze SAR components present in screening data. The NSG-SARI approach provides an opportunity to identify distinct local SAR environments within molecular networks and can in practice assist in the prioritization of compounds for further study on the basis of their SAR character.

Chapter 5

Analysis of Structure–Selectivity Relationships

Approaches for the analysis of structure–activity relationships traditionally focus on target-specific compound potency and ultimately aim at predicting highly potent molecules. However, potency is only one of several critical factors considered in lead optimization. Importantly, a promising drug candidate must also have a desired selectivity profile against a number of targets and anti-targets. Contrary to accepted views, target selectivity of active compounds is often not the result of exclusive target binding events but rather emerges from differential potency profiles against multiple targets (Hopkins, 2008). In particular, this implies that collections of active molecules might form multi-target SARs for closely related members of protein families that determine different degrees of compound selectivity.

The analysis of such multi-target SARs has only recently been facilitated through the design of selectivity benchmark systems for chemical biology applications and the adaption of fingerprint similarity searching to identify and distinguish target-selective molecules (Stumpfe et al., 2007, 2008; Vogt et al., 2007). However, for the systematic evaluation of structure–selectivity relationships (SSRs), no computational methods have as yet been established. In this chapter, we report a first step in this direction by extending the numerical and graphical functions developed for the analysis of target-specific SARs to the study of SSRs (Peltason et al., 2009a). Taking into account that SSRs are often a consequence of SARs against multiple targets, we present a comparative study of single-target SARs and dual-target SSRs using the NSG–SARI methodology described in the previous chapter.

5.1 Selectivity Data Sets

For the study of SARs and SSRs, we have analyzed inhibitor sets for cathepsin (cat) B, K, and L for which potency measurements for at least two related cysteine proteases were available. These cathepsin inhibitor data were taken from previously reported compound collections assembled from original literature (Stumpfe et al., 2008). A pool of 287 inhibitors was subdivided into two partly overlapping sets of compounds with potency against two target pairs, respectively: cat L versus B (LB), and cat K versus L (KL). The LB set contained 159 inhibitors, and KL 234 inhibitors. Both sets shared 106 inhibitors for which potency measurements were available for all three cathepsins. The two inhibitor sets are summarized in Table 5.1.

For a given target pair, compound selectivity was determined on the basis of differential potency against the corresponding targets. Selectivity values for target A over target B were calculated as the difference between pK_i or pIC_{50} values of each compound:

$$S_i = P_i(A) - P_i(B) \quad (5.1)$$

Here, S_i stands for the selectivity value of compound i for target A over target B, and $P_i(A)$ and $P_i(B)$ denote its potency values for targets A and B, respectively. In order to distinguish between selective and non-selective compounds, we adopted the selectivity criterion introduced by Stumpfe et al. (2008) that considered compounds to be selective for target A if they had a logarithmic selectivity value greater than 1.7 and selective for target B if their selectivity value was below -1.7 . Compounds falling within this range were considered non-selective. This threshold corresponds to a 50-fold difference in potency for the two targets.

5.2 Potency and Selectivity NSGs

For each data set, NSG–SARI analysis was carried out on the basis of potency values for individual targets and target-pair selectivity values. Three different graph representations were generated using potency values against targets A and B and selectivity values for target A over B, respectively, and the corresponding potency- and selectivity-based SARI scores were calculated. First, we separately calculated global SARI scores, cluster and compound discontinuity scores and NSGs utilizing the potency information for individual targets, thus producing two “potency NSGs”, NSG_A and NSG_B . These potency NSGs provided the basis for characterization of single-target SARs. In order to enable a direct comparison between potency NSGs for related targets, compound discontinuity scores were normalized relative to all compound scores in a given data

Table 5.1: Summary of selectivity data sets

	LB	KL
cpds	159 / 26 / 4	234 / 76 / 47
potency range		
A	3.82 – 10.40	4.00 – 11.05
B	3.00 – 8.07	3.82 – 10.40
selectivity range		
A/B	–2.21 – 3.17	–5.08 – 4.96
SARI scores (A)		
cont	0.41	0.12
disc	0.41	0.84
SARI	0.50	0.14
SARI scores (B)		
cont	0.44	0.09
disc	0.26	0.66
SARI	0.59	0.22
SARI scores (A/B)		
cont	0.44	0.09
disc	0.33	0.79
SARI	0.55	0.15

The composition of compound data sets and global SARI scores are reported. The row ‘cpds’ reports the number of compounds in each set; the first number reports the total number of compounds, the second the number of compounds selective for target A (the first target in a pair), and the third the number of compounds selective for target B (second target). In addition, SARI scores are reported for calculations using potency values for target A or B and selectivity values for target A over B (‘A/B’). ‘cont’ and ‘disc’ stand for continuity and discontinuity score, respectively.

set calculated for both activities. Due to this common normalization scheme, compound scores and node sizes can be directly compared in both potency NSGs. Furthermore, the same color spectrum was used to color-code nodes in NSG_A and NSG_B according to their potency values, ranging from the lowest to the highest potency value observed for one or the other target. Hence, corresponding colors in both potency NSGs denote the same potency values.

In addition, a “selectivity NSG” (NSG_{AB}) was generated for each data set on the basis of selectivity values for target A over target B. Nodes in selectivity NSGs were colored according to selectivity values using a gradient from red for the highest observed selectivity for one target to green for the correspond-

ing inverse selectivity value. Non-selective compounds with similar potency values for both targets and a resulting selectivity value close to 0 were represented by yellow nodes. Global SARI scores and discontinuity scores for compound clusters and individual compounds were calculated as described in Section 4.1 using selectivity values instead of potency values. This definition of selectivity-based SARI scores was appropriate because selectivity values represent potency differences and can accordingly be employed in the same manner. Thus, selectivity-based discontinuity scores served to identify regions of SSR discontinuity and “selectivity cliffs” formed by structurally similar molecules having markedly different selectivity. Selectivity-based compound scores were normalized with respect to all selectivity-based scores in the data set to reflect its unique selectivity distribution and thus cannot be directly compared to scores in potency NSGs or selectivity NSGs for other target pairs.

In order to relate global SSR character to the established SAR categories, selectivity-based SARI scores for entire data sets and for individual clusters were normalized to the same reference as potency-based SARI scores. For this purpose, the set of 13 activity classes from the MDDR described in Section 3.3 was utilized as reference. This common normalization scheme could be applied because selectivity-based scores were of the same dimension as their potency-based counterparts.

5.3 Selectivity NSG Analysis

The aim of our analysis was to characterize different global and local SSRs and investigate their relationships to single-target SARs. Specifically, we explored NSG environments that represented distinct local SSR features. Focusing on such environments made it possible to distinguish individual compounds making large contributions to single-target SARs and dual-target SSRs and identify structural modifications that were selectivity determinants.

5.3.1 Global SAR and SSR Features

For the analysis of global SAR and SSR character, global SARI scores were calculated for the selectivity data sets on the basis of compound potency and selectivity. Selectivity-based SARI scores fall into the value range between 0 and 1 and permit the classification of SSRs in analogy to SARs. High scores close to 1 are indicative of continuous SSRs where gradual changes in molecular structure are accompanied by moderate changes in compound selectivity. By contrast, low scores close to 0 reflect discontinuous SSRs that are characterized by the presence of similar molecules having different selectivity and thus forming selectivity cliffs. Similar to their potency-based counterparts, intermediate selectivity SARI values around 0.5 indicate heterogeneous SSRs that combine

different SSR elements. The global SAR and SSR characteristics of the two selectivity data sets are reported in Table 5.1. For the LB compound set, both target SARs are heterogeneous as well as the SSR for cat L over B. By contrast, for the KL set, both single-target SARs are globally discontinuous and the SSR for cat K over L is also characterized by strong discontinuity. Thus, in the case of the two cathepsin data sets studied here, SAR and SSR phenotypes are well in accord.

Figure 5.1 shows both potency NSGs and the selectivity NSG for the two compound sets. The topology of the NSGs is determined by pairwise similarity relationships between compounds and is thus identical in potency and selectivity NSGs for a data set. Comparison of the LB and KL network topology reveals distinct features that are characteristic of heterogeneous and discontinuous compound classes, respectively. Representing the heterogeneous category, the LB topology is marked by the presence of several distinct subgraphs or communities of varying potency and selectivity composition, as indicated by different node color distributions within local NSG regions. In NSG_{LB} , highly selective compounds (red and green nodes) are distributed over different network regions. This phenotype is a hallmark of SAR and SSR heterogeneity. In contrast, graph representations for the KL data set are more densely connected, which reflects a higher degree of structural homogeneity of the KL than the LB set. One large central network component is found that includes many large nodes representing key compounds, which is characteristic of SAR and SSR discontinuity. However, in NSG_{KL} , highly selective compounds are also found in different network environments, similar to NSG_{LB} .

5.3.2 Comparison of SAR and SSR Elements

In order to compare the composition of local SAR and SSR elements, we calculated the distribution of cluster discontinuity scores in both potency and selectivity NSGs for the LB and KL compound sets. Figure 5.2 shows that potency-based and selectivity-based scores essentially cover the entire value range between 0 and 1, which reflects a high degree of local SAR and SSR variability. For both sets, the cluster discontinuity for one target is significantly higher (L for LB and K for KL) and largely determines the overall SSR heterogeneity. Although the LB set displays globally heterogeneous SAR and SSR character and KL SARs and SSR are more discontinuous, cluster score distributions are overall comparable for both sets. Thus, differences in global SAR and SSR categories for the two sets can be largely assigned to different levels of chemical diversity.

In order to investigate how SAR and SSR characteristics are related to each other and how they are determined by specific compound subsets, corresponding network regions in potency and selectivity NSGs were compared. For

(a)

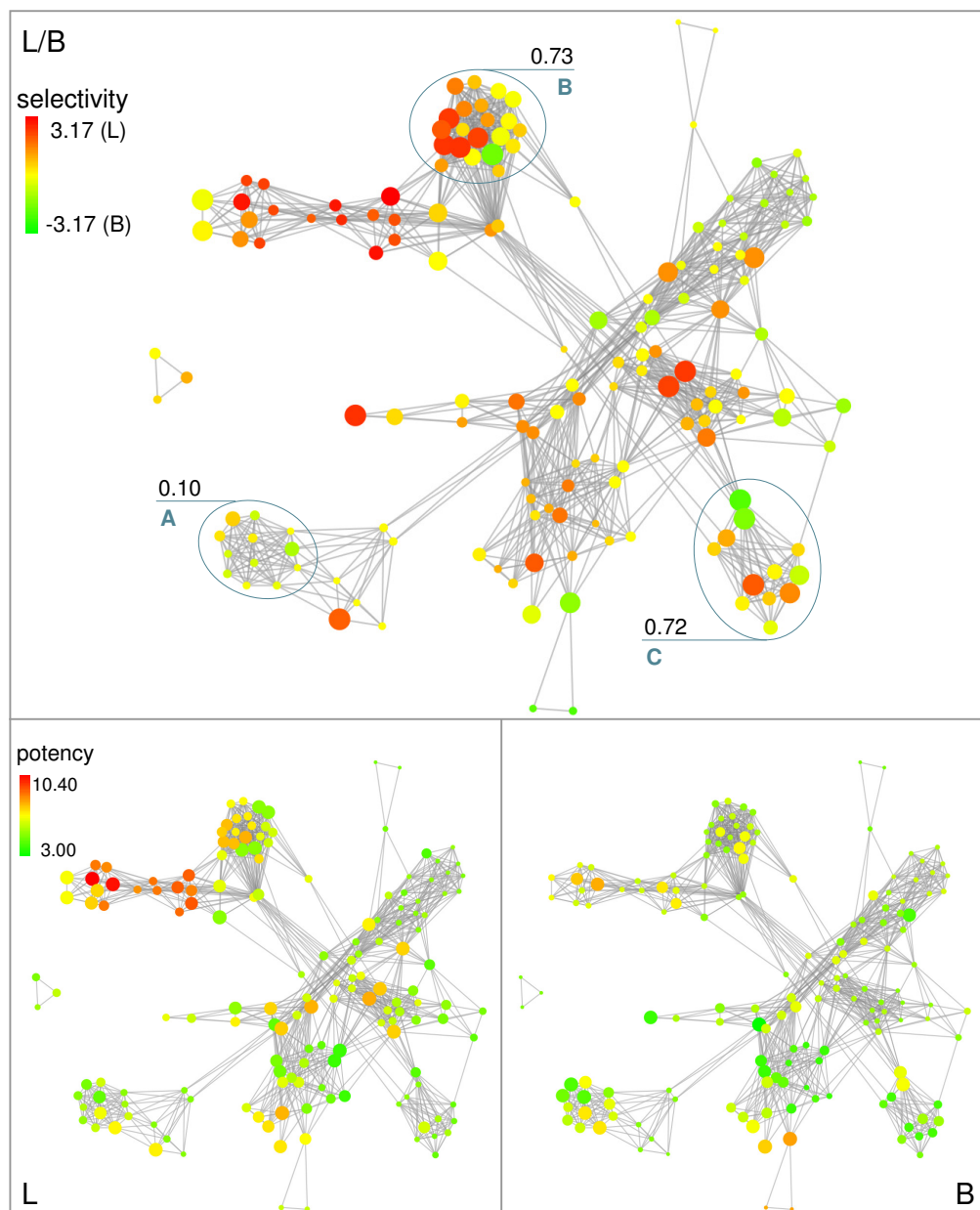


Figure 5.1: NSG representations for selected target pairs

(b)

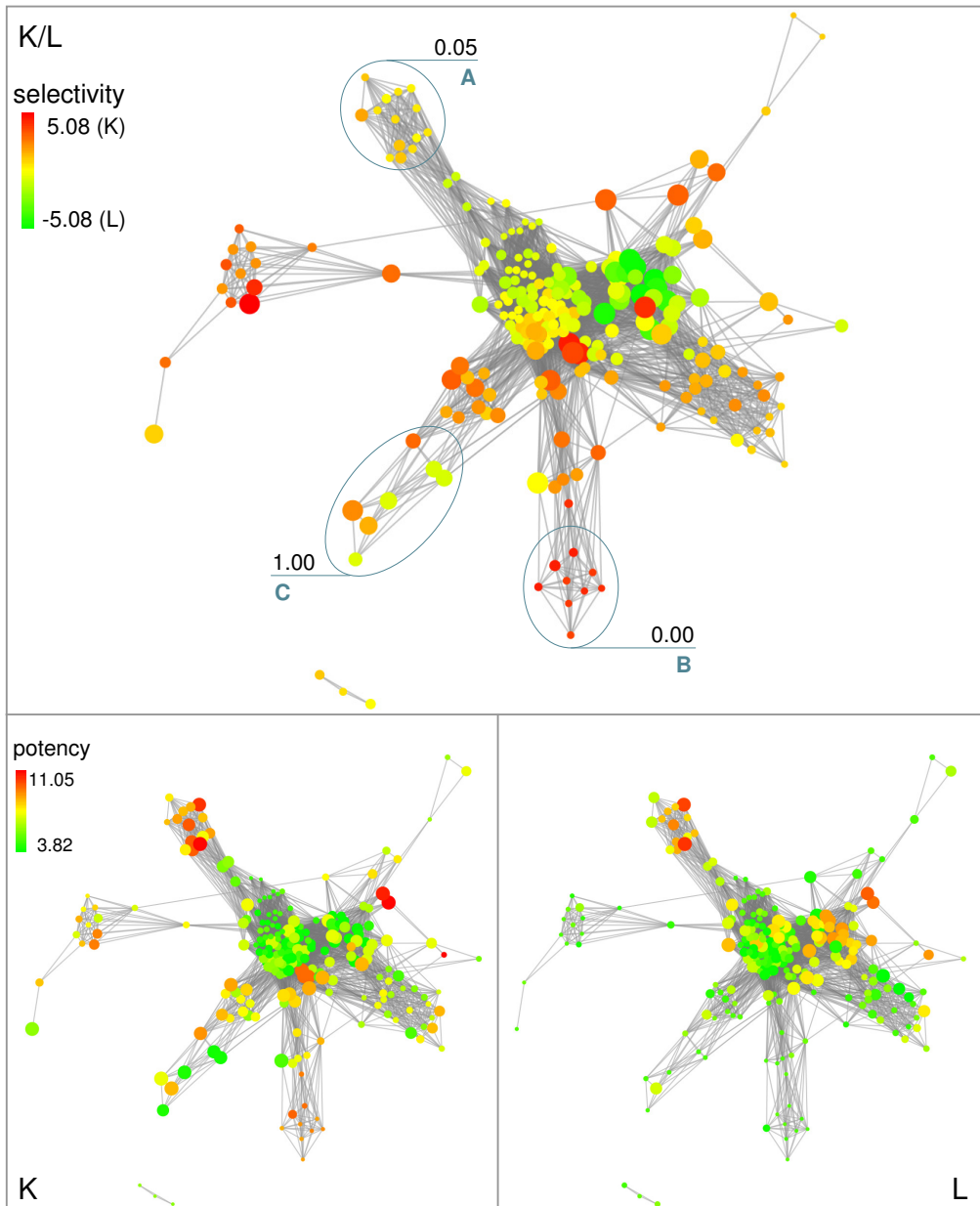


Figure 5.1: NSG representations for selected target pairs (continued) The graph at the top of each subfigure shows the selectivity NSG for a target pair. At the bottom, the corresponding potency NSGs for individual targets are shown. (a) Potency and selectivity NSGs for cat L versus B. (b) Potency and selectivity NSGs for cat K versus L.

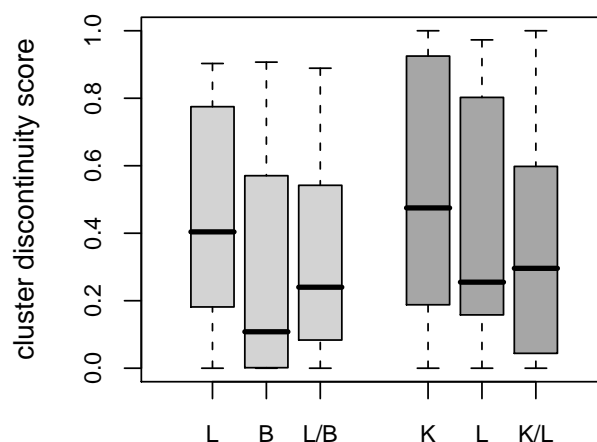


Figure 5.2: Distribution of cluster discontinuity scores Cluster discontinuity score distributions are reported as box plots for the LB and KL selectivity sets. For each set, three boxes are shown that represent the cluster scores for the two potency NSGs and the corresponding selectivity NSG, respectively.

example, compounds in the upper left regions of NSG_L in Figure 5.1(a) make significant contributions to local SAR discontinuity, as reflected by the large size of the nodes. By contrast, in NSG_B , these compounds have only minor potency differences and accordingly form a continuous local SAR, indicated by the smaller size of the corresponding nodes. The compounds represented in this network region respond differently to cat L and B, as illustrated by the different coloring of the nodes in both potency NSGs. Accordingly, the corresponding region in the selectivity graph NSG_{LB} shows that many of these inhibitors are highly selective, whereas also non-selective compounds are found in the same cluster. Hence, this network region is marked by strong local SSR discontinuity and significantly contributes to global SSR heterogeneity.

By contrast, compounds in the network region of NSG_K and NSG_L containing cluster A in Figure 5.1(b) include highly and weakly active compounds that form activity cliffs for both targets and make similarly strong contributions to local and global SAR discontinuity. However, these compounds respond to both targets in a similar way and thus are non-selective or only weakly selective for cat K. Consistent with the comparable selectivity behavior of similar compounds, this network region in NSG_{KL} is characterized by a highly continuous local SSR, as illustrated by the presence of small yellow and orange nodes. Thus, comparison of local SAR and SSR elements reveals the complementary nature of SAR and SSR information captured in NSG representations. In selectivity NSGs, node colors represent potency information for two related targets and

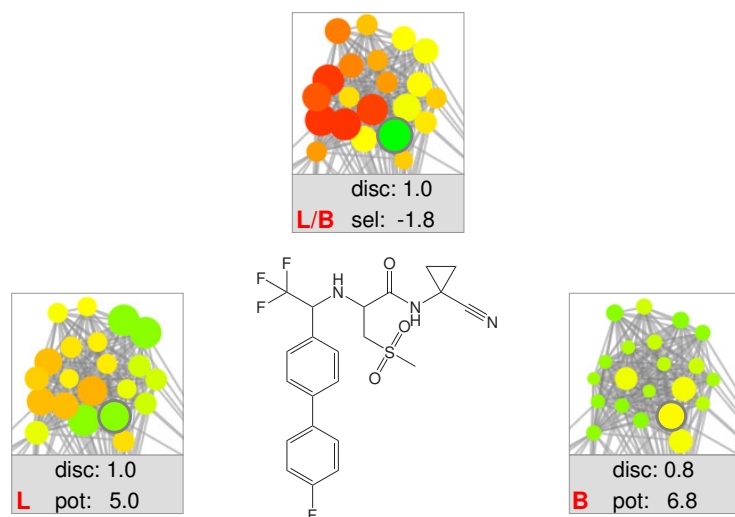
node sizes reflect the selectivity differences of individual compounds compared to their neighbors. Local SSR character is often highly variable, depending on potency distributions and the way similar molecules respond to related targets.

5.3.3 Local SSR Environments

After analyzing relationships between local SAR and SSR features, we focus on local network environments of distinct SSR character. The analysis aims at distinguishing between continuous and discontinuous local SSRs and identifying key compounds that correspond to selectivity determinants. In Figure 5.1, selected clusters in the selectivity graphs are annotated with their cluster discontinuity scores. Despite their different topology, the selectivity NSGs for both data sets, LB and KL, contain distinct local SSR environments, represented by compound clusters with very low or high discontinuity scores. Continuous SSR regions are formed by groups of similar molecules with comparable selectivity. Such continuous environments can either be composed of non-selective compounds or compounds that are selective for the same target. For example, cluster A in NSG_{LB} or cluster A in NSG_{KL} mostly consist of non-selective inhibitors that make essentially no contributions to SSR discontinuity and are represented as small yellow, pale green, or orange nodes. Such environments of local SSR continuity frequently occur in selectivity NSGs and identify compound subsets that provide only little information for the exploration of selectivity at the molecular level. By contrast, cluster B at the bottom of NSG_{KL} is formed exclusively by K-selective compounds (nodes colored in bright red) that show only little differences in their selectivity values. Due to its homogeneous selectivity composition, this cluster represents a continuous SSR region, as reflected by the small size of the corresponding nodes and a cluster discontinuity score of 0.

Other SSR environments are strongly discontinuous in nature. For example, cluster C in NSG_{KL} consists of selective (orange) and non-selective (pale green) compounds that make large contributions to SSR discontinuity. Furthermore, clusters B and C in NSG_{LB} or clusters in the central network component of NSG_{KL} contain inhibitors with high (red) and low (yellow) or even inverse (green) selectivity. These clusters obtain high discontinuity scores. Hence, cluster discontinuity scores detect environments in NSGs that are characterized by a high degree of SSR discontinuity and thus represent the most interesting regions for the selection of compounds to explore selectivity determinants. Within these local environments, compound discontinuity scores provide a measure for the identification of molecules that play a key role for target-pair SSRs.

(a)



(b)

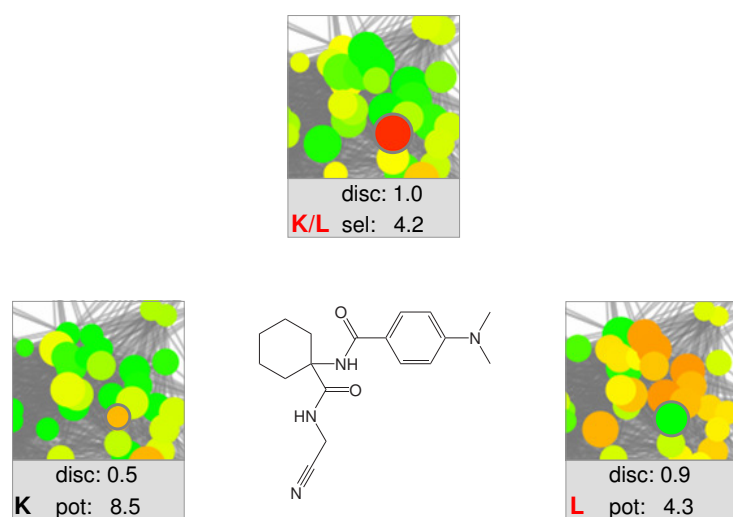


Figure 5.3: Key compounds

(c)

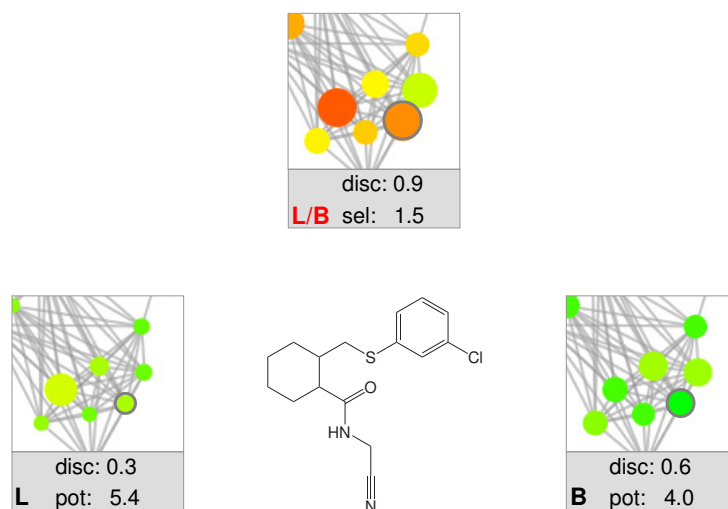


Figure 5.3: Key compounds (continued) Key compounds are shown together with their NSG environments. Network details of the two potency NSGs (left/right) and the corresponding selectivity NSG (top) of a target pair show the key compound (encircled node) and its closest neighbors. Graph labels are colored red if the selected compound is an activity or selectivity cliff marker. For each key compound, logarithmic potency ('pot') and selectivity ('sel') values and compound discontinuity scores ('disc') are reported.

5.3.4 SAR and SSR Key Compounds

Network regions of strong local SAR and SSR discontinuity contain key compounds that have high compound discontinuity scores and are thus involved in the formation of activity or selectivity cliffs. Most prominent selectivity cliffs are formed by pairs of structural analogs where one molecule is selective for target A and the other for target B (i.e. a pair of large red and green nodes in the selectivity NSG). In the selectivity NSGs shown in Figure 5.1, several selectivity cliffs are apparent in discontinuous local environments. On the basis of compound discontinuity scores, we have selected compounds that are activity and/or selectivity cliff markers and hence major determinants of SAR and/or SSR features. Figure 5.3 shows compounds that are selectivity cliff markers but contribute to single-target SARs in different ways. In Figure 5.3(a), an inhibitor belonging to cluster B in the LB selectivity graph is shown that is selective for cat B and strongly contributes to the formation of selectivity cliffs, as reflected by a maximal compound discontinuity score of 1.0. In NSG_{LB} , this inhibitor is the only B-selective compound (green node) within a region containing structurally similar non-selective (yellow) or L-selective compounds (red nodes). The selected compound is B-selective due to its low potency for L ($pK_i = 5.0$) and intermediate potency for B ($pK_i = 6.8$). With this potency

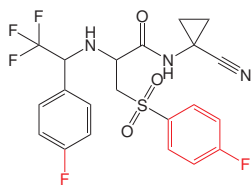
profile, this compound has the lowest potency for L and the highest potency for B compared to its neighbors in the network environment. Accordingly, it also strongly contributes to local SAR discontinuity in NSG_L and NSG_B with discontinuity scores of 1.0 and 0.8, respectively. However, selectivity cliff markers are not always also activity cliff markers. Figure 5.3(b) shows a compound from the central NSG component of the KL data set that is a prominent selectivity cliff marker because it is highly selective for K (red node), whereas its neighbors are mostly selective for L. The selectivity of this compound is largely determined by its low potency against cat L, as illustrated by the complementarity of the node colors in its NSG_{KL} and NSG_L environments. In NSG_L , this compound contributes significantly to local SAR discontinuity because it has considerably lower potency than its neighbors. In NSG_K , however, there is no activity cliff in the corresponding region, because the selected compound is similarly potent against cat K as its neighbors. Hence, this compound represents a selectivity cliff marker that is also an activity cliff marker for one target, but not for the other. Moreover, compounds that do not play a key role for individual SARs might also become key compounds in selectivity NSGs. Figure 5.3(c) shows an example from the LB set belonging to cluster C in NSG_{LB} . This inhibitor is only weakly potent against cat B and L, similar to the other compounds within its network environment. Hence, the compound contributes only little to single-target SAR discontinuity. However, due to its higher potency for L than B, this inhibitor is L-selective, whereas its neighbors are mostly non-selective or selective for cat B. Thus, the selected compound induces a significant degree of local SSR discontinuity in NSG_{LB} . Hence, key compounds that are involved in the formation of selectivity cliffs can have different influence on single-target SARs. The NSG neighborhood of such selectivity cliff markers is populated by similar compounds that have markedly different selectivity and thus present interesting starting points for the analysis of molecular selectivity determinants.

5.3.5 Selectivity Determinants

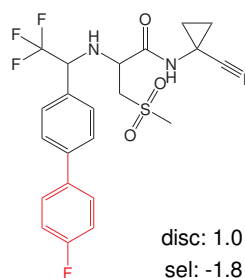
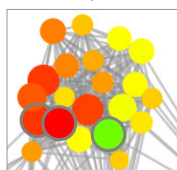
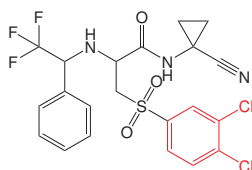
In addition to identifying key compounds that are responsible for SAR and SSR discontinuity, another major goal of selectivity NSG analysis is the exploration of structural features that determine compound selectivity. This can be accomplished by focusing on collections of similar compounds with varying selectivity levels, represented by sets of large connected nodes with different colors in the network. Inspecting the environment of key compounds for similar molecules having different selectivity has led to the identification of analogs that are distinguished by selectivity-determining substitutions.

Four exemplary sets of analogs and their network environments are presented in Figure 5.4. Figure 5.4(a) presents the B-selective key compound from the LB data set discussed above (Figure 5.3(a)) together with two of its clos-

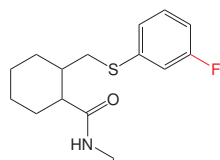
(a)

disc: 1.0
sel: 2.5

L/B

disc: 1.0
sel: -1.8disc: 1.0
sel: 2.5

(b)

disc: 0.6
sel: 0.1

L/B

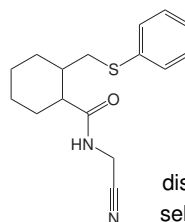
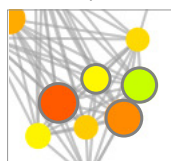
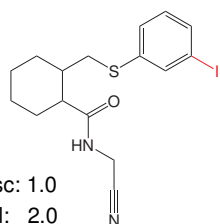
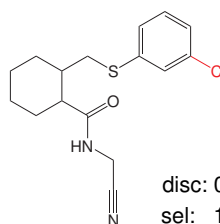
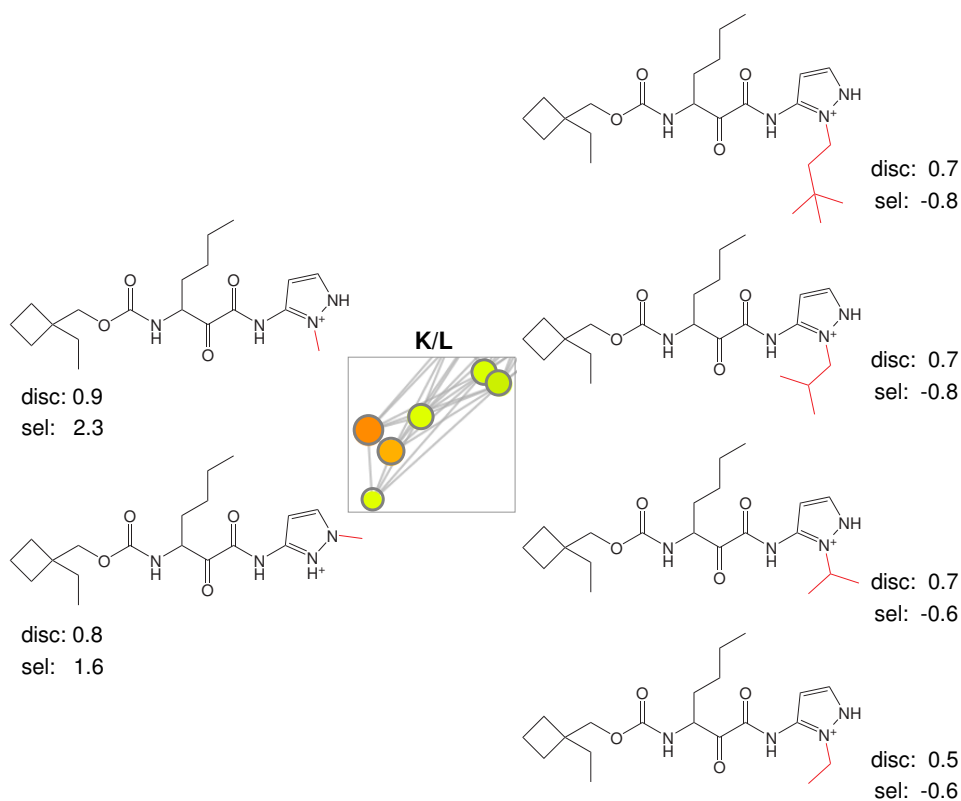
disc: 0.9
sel: -0.7disc: 1.0
sel: 2.0disc: 0.9
sel: 1.5

Figure 5.4: Selectivity determinants

(c)



(d)

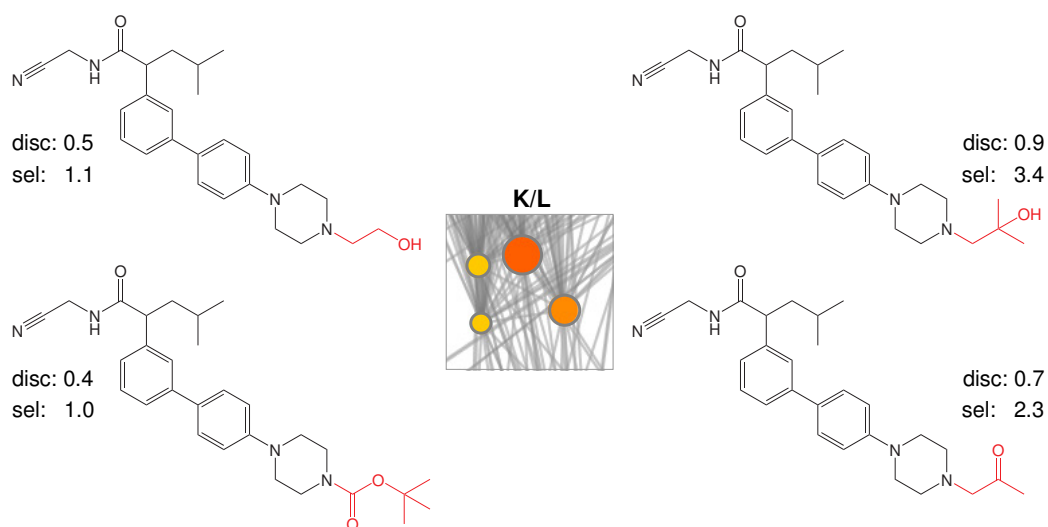


Figure 5.4: Selectivity determinants (continued) In (a) – (d), examples of structurally analogous compounds from the environment of key compounds are shown that form selectivity cliffs of different magnitude. The network environments of these compounds in the selectivity NSG and their discontinuity scores and selectivity values are also displayed. Substituents that distinguish between compounds having different selectivity are colored red.

est neighbors that are highly selective for cat L. Hence, these compounds are prominent selectivity cliff markers. The presence of nitrile groups (or other strong nucleophiles) generally represents a hallmark of non-selective cathepsin inhibition. Thus, target selectivity must be determined by other functional groups. Comparison of the three inhibitors indicates that halogenated phenyl substituents at the sulfonyl group render analogs L-selective (indicated by red coloring), whereas the halogenated biphenyl derivative is B-selective (green). The second key compound from the LB set discussed in the previous section is shown in Figure 5.4(b) together with three other analogs that differ only in a substituent at the phenyl ring. The compound containing the iodine substitution is highly selective for cat L (red node). Selectivity decreases for the more electronegative substituents chlorine (orange) and fluorine (yellow node) and is shifted toward the inverse for a non-substituted phenyl derivative (green node). In addition, Figure 5.4(c) shows a series of analogs from cluster C in NSG_{KL} that are either selective for K or non-selective with a tendency toward cat L. The six compounds only differ in a hydrophobic substituent at the pyrazole moiety. With increasing bulkiness of this substituent, compound selectivity is gradually shifted from cat K (orange) to L (green nodes). Furthermore, in the series of analogs shown in Figure 5.4(d), various oxygen-containing N-substituents at the piperazine ring are observed that determine whether a compound is selective for L (orange nodes on the right) or non-selective (yellow nodes on the left).

These examples illustrate crucial aspects of potency and selectivity NSG analysis. Discontinuous local SSR environments can be readily identified containing key compounds that are responsible for SAR and/or SSR discontinuity and often play different roles for SARs and SSRs of target pairs. Furthermore, series of analogs found in network neighborhoods of key compounds can be utilized to explore SSRs at the level of individual compounds and identify selectivity-determining substitution sites and patterns.

5.4 Conclusions

Target selectivity of compounds active against multiple targets is a critical aspect in medicinal chemistry. In order to quantitatively assess relationships between molecular structure and selectivity, we have applied the NSG-SARI analysis to sets of inhibitors active against two related cathepsins. The potency-centric analysis has been complemented by score calculations and graphical analysis based on selectivity values that were calculated from potency differences. Following this approach, structure-selectivity relationships could be classified into the same categories that were originally established for SARs. Similar to SARs, global SSR types were found to correspond to characteristic graphical

features and were composed of different local SSR elements. Comparison of such elements with local SAR features revealed the variability of SSRs and their dependence of potency distributions in the corresponding network environments. Furthermore, we detected regions of local SSR discontinuity containing selectivity cliffs of different magnitude. Key compounds involved in the formation of selectivity cliffs were identified that influenced SSRs and SARs in similar or different ways. In the network environments of such key compounds, analogous molecules having different selectivity were found. These molecules were distinguished by well-defined substitutions that determined their selectivity.

For medicinal chemistry, the comparative study of SARs and SSRs has considerable practical utility. Specifically, NSG–SARI analysis can aid in the selection of compounds that have a desired potency and selectivity profile and present promising starting points for further optimization. In order to facilitate the systematic optimization of compound selectivity, discontinuous regions in selectivity NSGs can be explored and selectivity determinants at the structural level can be identified. This makes the NSG–SARI methodology a useful tool for the exploration of relationships between molecular structure, compound potency and selectivity.

Chapter 6

Structure–Activity Relationship Determinants in Analog Series

Methods for the systematic analysis of structure–activity relationships as discussed thus far have aimed at classifying global and local SARs present in sets of active molecules. Such compound collections are typically composed of several compound series representing different chemotypes. The introduced methods are particularly useful for the identification of local SAR features and the prioritization of compounds based on their SAR character. Hence, they are designed to aid in hit selection. The requirements change when selected compounds are subjected to hit-to-lead projects. In hit-to-lead or lead optimization efforts, one primarily focuses on individual chemotypes and systematically explores chemical modifications to optimize their potency and other desired properties. This process of designing analogs and evolving leads is largely guided by SAR information that is already available and investigates one chemical modification at a time in order to plan the next step. Hence a central question is, which compounds should be tested in order to obtain as much additional SAR information as possible? Specifically, it is often unclear which parts of a molecule are relevant for a given SAR and, accordingly, at which positions modifications should be made. Often, the analysis is complicated by the variable nature of SARs and the presence of multi-layered SAR information in analog series.

For this reason, we have adopted the SARI formalism and developed Combinatorial Analog Graphs (CAGs) that provide ways and means to organize existing SAR information in analog series with a focus on contributions from individual functional groups and combinations of groups (Peltason et al., 2009b). These graph representations hierarchically organize compounds according to substitution patterns and are annotated with SARI discontinuity scores in order to account for SAR discontinuity at the level of functional groups. The approach makes it possible to identify undersampled regions and highlight key substitution patterns that determine the SAR of a compound series. The methodology

Table 6.1: Source data sets

target	source	no. cpds	no. series	potency range
hsd17b4	PubChem AID 893	1366	134	251 nM – 40 μ M
thrombin	PubChem AID 1215	51	6	1 nM – 50 μ M
cyt P450 3a4	PubChem AID 884	1251	134	25 nM – 40 μ M
hadh2	PubChem AID 886	400	42	32 nM – 40 μ M
cathepsin K	(Stumpfe et al., 2008)	264	37	0.01 nM – 1 mM
cathepsin L	(Stumpfe et al., 2008)	290	43	0.04 nM – 150 μ M
cathepsin S	(Stumpfe et al., 2008)	296	42	0.13 nM – 1 mM

Data sets containing a number of analog series were collected from PubChem or from compound selectivity sets and served as reference for score normalization. ‘no. cpds’ reports the number of compounds and ‘no. series’ the number of analog series with distinct molecular scaffolds present in a data set. ‘hsd17b4’ stands for hydroxysteroid-17 β -dehydrogenase 4, ‘cyt’ for cytochrome, and ‘hadh2’ for hydroxyacyl-CoA dehydrogenase II.

is presented in Section 6.1, and key aspects of the analysis are discussed in Section 6.2 for four exemplary analog series directed at different targets. Furthermore, as demonstrated in Section 6.3, the analysis is also applied to series of analogous cathepsin inhibitors in order to compare SAR determinants for multiple related targets.

6.1 Methodology

In order to analyze SARs of analog series at the level of individual substitution sites, compound series were extracted from various data sources and divided into subsets of molecules that differed only at specific substitution sites or site combinations. Substitution sites were identified through R-group decomposition. For the resulting compound subsets, SARI discontinuity scores were calculated that directly reflected SAR contributions of functional groups at variable sites. Compound subsets distinguished by modifications at well-defined substitution sites and the corresponding discontinuity scores were then organized in a hierarchical graph structure. Figure 6.1 illustrates the subsequent steps.

6.1.1 Data Sets and Analog Series Identification

Analog series were extracted from screening data sets available in PubChem BioAssay including inhibitors of hydroxysteroid-17 β -dehydrogenase 4 (hsd17b4, AID 893), thrombin (AID 1215), cytochrome P450 3a4 (AID 884), and hydroxyacyl-CoA dehydrogenase II (hadh2, AID 886). Compounds considered to be

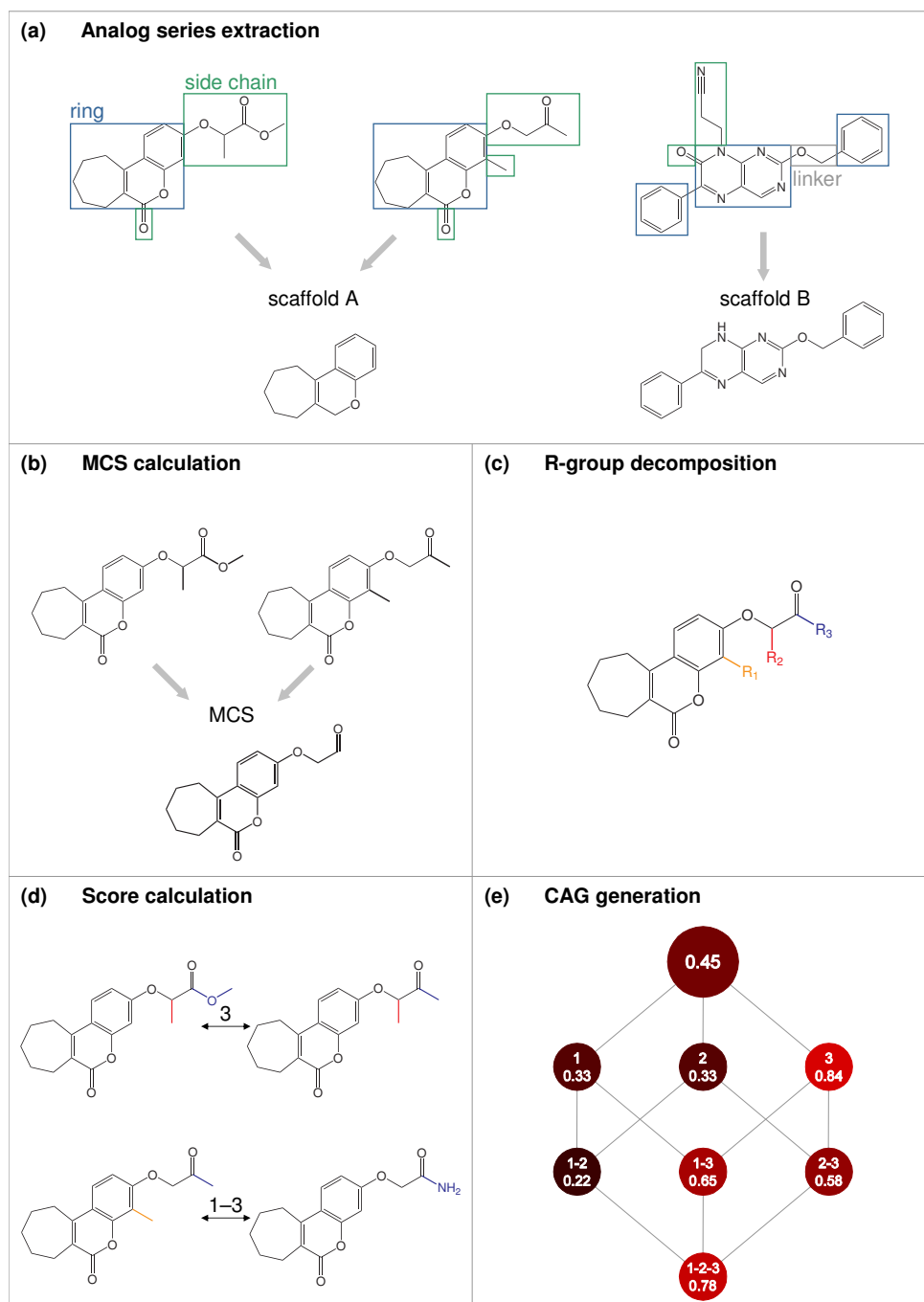


Figure 6.1: Combinatorial analog graph calculation (a) For all molecules in a compound data set, molecular scaffolds are calculated by deleting all side chains. Molecules with identical scaffolds are grouped into the same analog series. (b) For a series of analogous molecules sharing a common scaffold, the maximum common substructure (MCS) is calculated. (c) Variable functional groups are consistently numbered and assigned to corresponding substitution sites through R-group decomposition. (d) SARI discontinuity scores are calculated for subsets of molecules that differ at well-defined substitution sites. Shown are two pairs of molecules that differ at site 3 (top) or at sites 1 and 3 (bottom). (e) Discontinuity scores reflect SAR contributions from individual substitution sites and are organized in a combinatorial analog graph (CAG).

inactive under screening conditions were assigned a potency value equal to the chosen activity threshold. In addition to analog series collected from screening data, inhibitors of cathepsin K, L, and S were taken from previously reported compound sets that included optimized and highly selective compounds (Stumpfe et al., 2008). From these source data sets, series of analogous structures were automatically extracted through analysis of molecular scaffolds following the definition of Bemis and Murcko (1996). Accordingly, scaffolds were derived by deleting all side chains (R-groups) from a molecule, and rings and linkers were retained together with atom element, hybridization, and bond order information. Molecules with identical scaffolds were then grouped into analog series. Table 6.1 summarizes the data sets used in this study.

6.1.2 R-Group Decomposition

Compounds in analog series were divided into constant and variable regions through R-group decomposition. Typically, invariant regions included the molecular scaffold and possibly R-groups that were conserved in all compounds of a series. Initially, invariant molecular regions were determined by calculating the maximum common substructure (MCS) shared by all analogs in a series. The MCS was then used as core structure for R-group decomposition, which defined the substitution sites and functional groups for each molecule. For this purpose, the MCS was mapped onto each molecule in a series and the substituents were assigned to corresponding R-groups and consistently numbered. MCS identification and R-group decomposition were automatically carried out with Pipeline Pilot. SAR tables that report core structures, substitution sites, and R-groups for all series discussed in this chapter are found in Appendix C.

6.1.3 SAR Contributions from R-Groups

In order to assess SAR contributions of functional groups, we organized series of analogs into subsets of molecules that differed only at specified substituent positions. Thus, to quantify contributions of a specific substitution site, all compounds were selected from a series that had different R-groups attached to this site, but were otherwise identical. For the resulting compound subsets, the SARI discontinuity score was calculated as described in Section 3.1. As in previous studies, pairwise compound similarity was calculated using MACCS Tc values. For this study, however, the application of a similarity threshold value was not required because all analogs have highly similar structures. Furthermore, no potency difference threshold was applied to enable the detection of varying levels of SAR discontinuity. Accordingly, for a set S of analogous

compounds, the modified discontinuity score was then defined as follows:

$$\text{disc}_{\text{raw}}(S) = \text{mean}_{\{(i,j) \in S | i \neq j\}} (|P_i - P_j| \cdot \text{sim}(i, j)) \quad (6.1)$$

Because compounds in a subset were only distinguished at well-defined substitution sites, observed SAR discontinuity could be directly attributed to R-group variation at these sites. Furthermore, SAR contributions from combinations of substitution sites were calculated for compounds that had different R-groups attached at site pairs or triplets but identical substituents at the remaining sites. Combinations of up to three different substitution sites were considered. For a given substitution site or combination of sites, several subsets might exist that consist of compounds that differ only at the given sites but are distinguished at another site (see Subsection 6.2.1 for an example). Discontinuity scores for these subsets were calculated independently and averaged to yield the final score for the substitution site combination under consideration. In addition, in order to estimate the SAR character within a given analog series, the SARI discontinuity score as defined in equation 6.1 was calculated also for the entire series, irrespective of individual substitution patterns of compounds.

The “raw” discontinuity scores for an analog series and corresponding compound subsets were normalized by Z-score calculation and mapped to the value range [0,1] by calculating the cumulative distribution function as described in Chapter 3. As summarized in Table 6.1, all analog series used in this study were taken from source data sets consisting of several analog series. The score distribution of all compound subsets from all analog series within a source data set served as the reference for score normalization of its analog series. Accordingly, the scores reflect the target-specific score distribution in the entire data set, which makes it possible to differentiate relatively narrow potency distributions. Using this scoring scheme, scores for different analog series originating from the same source set can be directly compared, thus allowing to discriminate between compound series having different degrees of SAR discontinuity. However, for analog series taken from different data sets, the magnitude of scores cannot be compared.

6.1.4 Combinatorial Analog Graphs

SAR features of analog structures were visualized in a hierarchical graph representation. In a CAG, nodes correspond to compound subsets and edges indicate that compounds in connected subsets have modifications at the same substitution sites (see below). The root node represents the entire analog series and non-root nodes represent subsets of compounds that only differ at individual substitution sites or unique site combinations. Node labels identify these substitution sites and report discontinuity scores for the corresponding compound

subsets. Furthermore, nodes in a CAG are color-coded according to discontinuity scores using a color gradient from black (score 0) to red (score 1) and hierarchically arranged in layers according to the number of substitution sites that are considered. Substitution site combinations for which no compounds are available are shown as small white nodes and represent “SAR holes” (i.e. unexplored sites or combinations). Edges are drawn from a node to all other nodes in the next layer whose substitution site combinations include all of the sites represented by the originating node (e.g. node 2 is connected to nodes 1–2 and 2–3, but not to 1–3). However, it should be noted that in CAGs, only the location of substitutions is considered and not their chemical character. Hence, connected nodes might contain compound sets with distinct substituents at corresponding sites.

6.2 SAR Analysis in Analog Series

A primary goal of our analysis has been to systematically evaluate the SAR contributions of combinatorial R-group patterns in analog series and identify substitution sites that are SAR determinants (“SAR hotspots”) and preferred targets for further chemical exploration. The CAG–SARI approach combines the hierarchical organization of analog series according to substitution site combinations with a quantitative SAR analysis function to assess site-dependent contributions to SAR discontinuity. In the following, key aspects of the approach are discussed on the basis of four representative compound series extracted from screening data sets for different targets.

6.2.1 Interpretation of CAGs

To illustrate the compound organization scheme, Figure 6.2 shows a prototypic CAG representation generated for five exemplary hydroxysteroid-17 β -dehydrogenase 4 inhibitors with three substitution sites. The compounds and their common core structure are also shown. The root node at the top represents the entire compound set and reports its discontinuity score. With a score value of 0.45, this small series is characterized by intermediate SAR discontinuity. Each subsequent node corresponds to a unique combination of substitution sites and reports the degree of SAR discontinuity induced by modifications at these sites. In Figure 6.2, nodes are annotated with compound subsets (in this case, pairs of compounds) that differ only at the corresponding substitution sites and provide the basis for score calculations at the individual nodes. The figure illustrates that analogs usually participate in different subsets, given the distribution of substituents, and multiple compound subsets might exist for individual nodes. For example, the compounds forming the pairs AB and CD only differ at substitution site 3 and are thus assigned to the corresponding node.

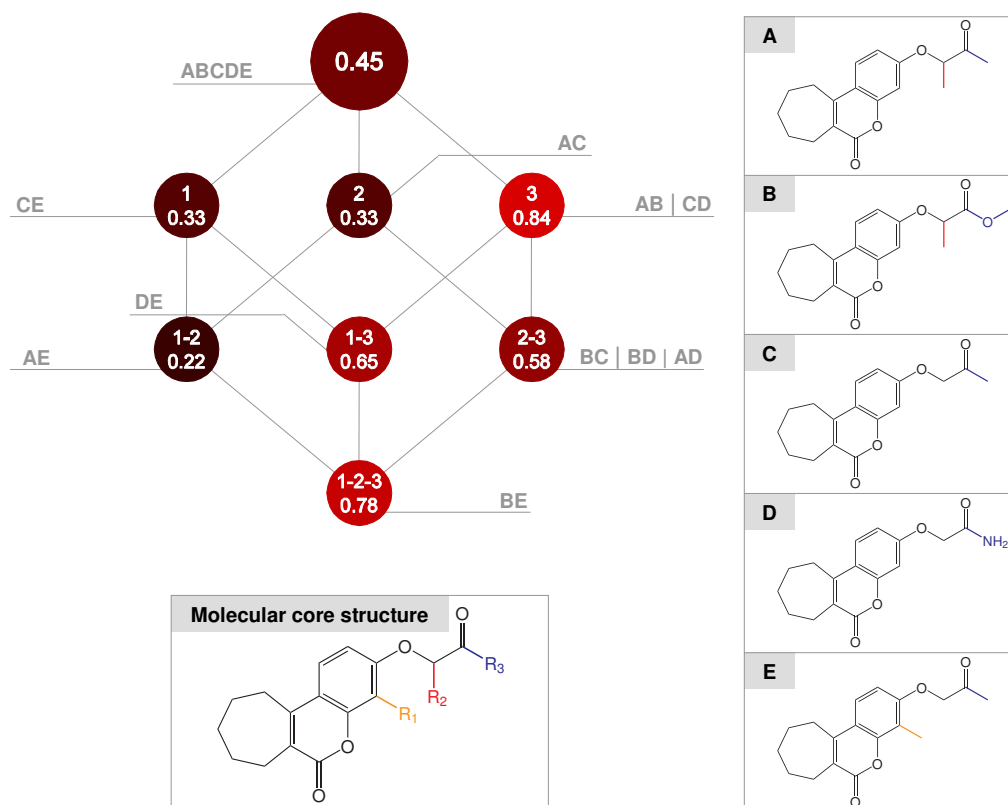


Figure 6.2: CAG for inhibitors of hydroxysteroid-17 β -dehydrogenase 4 An exemplary CAG representation for five analogous inhibitors of hydroxysteroid-17 β -dehydrogenase 4 is shown. Nodes in the CAG correspond to compound subsets: the root node represents the entire analog series and non-root nodes correspond to subsets of compounds that differ only at predefined substitution sites. Node labels identify variable substitution sites and report SARI discontinuity scores calculated for the corresponding compound subsets. Nodes are color-coded according to discontinuity scores (black: 0, red: 1) and annotated with inhibitor labels that provide the basis for score calculations at the individual nodes.

However, these two compound subsets are distinguished from each other at site 2. Thus, for each pair, the discontinuity score is separately calculated and both scores are averaged to yield the final score that reflects the overall discontinuity introduced by R-group variation at site 3. For the exemplary compound set shown in Figure 6.2, simultaneous modifications at all three substitution sites are detected and hence, all possible nodes are populated.

Although CAG–SARI analysis of small data sets is meaningful, larger compound series provide more SAR information for CAG representations. Due to the combinatorial nature of the representation scheme, the complexity of a CAG increases with the number of substitution sites present in a compound series. For example, for three substitution sites, one bottom node with a three-site

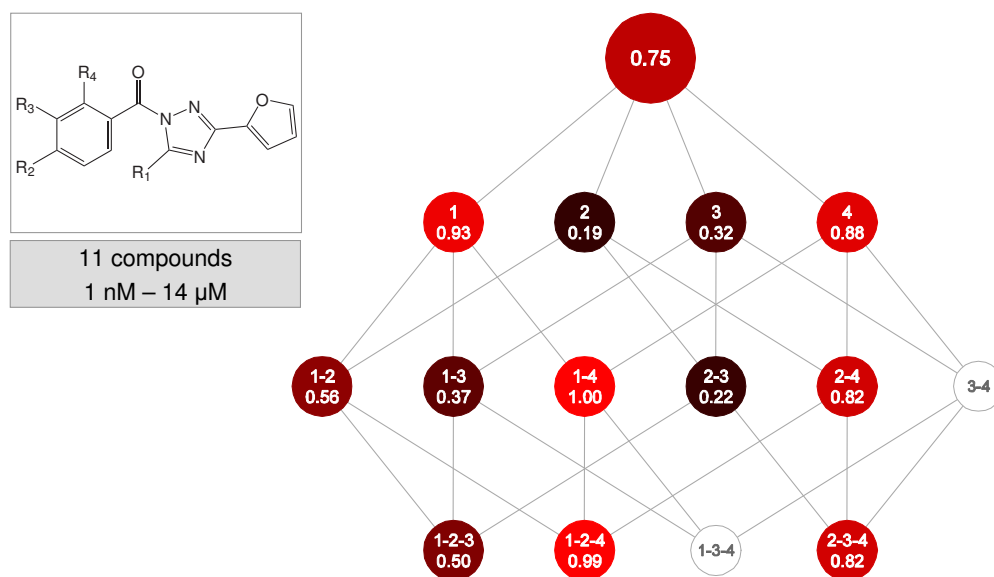


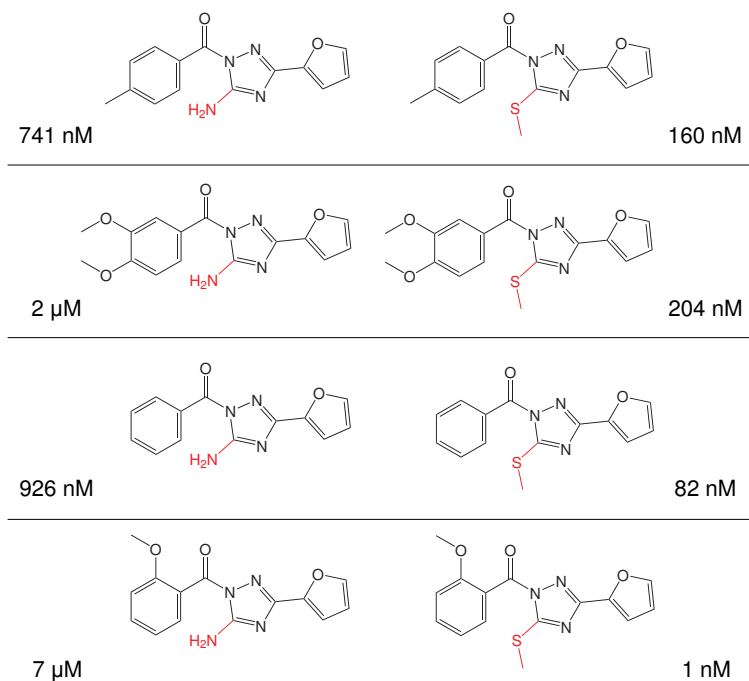
Figure 6.3: CAG for thrombin inhibitors Shown is the CAG representation and the common core structure for a series of 11 analogous thrombin inhibitors with variations at four substitution sites and potency in the low nanomolar to micromolar range.

combination is obtained (Figure 6.2) but for four sites, there are four bottom nodes, as illustrated in Figure 6.3.

6.2.2 SAR Hotspots

Figure 6.3 shows the CAG representation for a series of 11 thrombin inhibitors that cover a wide potency range (1 nM – 14 μ M). With a discontinuity score of 0.75, the entire series shows a considerable degree of discontinuity. This SAR character can be assigned to well-defined substitution patterns, represented by CAG nodes that obtain high discontinuity scores. These nodes are associated with compound subsets that include variations at site 1 (nodes 1, 1–4, 1–2–4) and at site 4 (nodes 4, 2–4, 2–3–4). Figure 6.4(a) presents compound pairs with variations at site 1 that form activity cliffs of increasing significance, with potency differences of up to four orders of magnitude. Modifications at sites 1 and 4 are consistently responsible for SAR discontinuity, whereas modifications at other sites have only limited effects. This is illustrated in Figure 6.4(b) that shows three compounds that are distinguished at sites 2, 4, or both. Removal of the chlorine substituent at site 4 increases potency by more than one order of magnitude, regardless of the simultaneous addition of a methoxy group at site 2. Adding this group at site 2 without simultaneously changing site 4 only

(a)



(b)

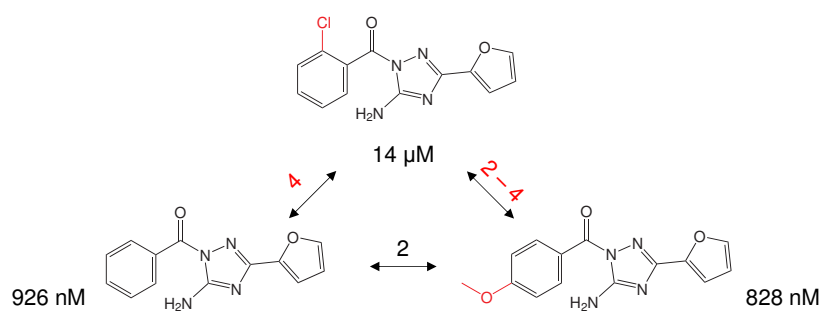


Figure 6.4: Thrombin inhibitors (a) Pairs of compounds with modifications at site 1 that form activity cliffs of increasing magnitude. Variable functional groups are colored red. (b) Compounds with variations at sites 2 and 4. Numbers along the arrows identify variable substitution sites for pairwise compound comparison that make strong (red) or weak (black) discontinuity contributions. Individual modification of site 4 or simultaneous modification of sites 2 and 4 of the molecule shown at the top increases potency by more than one order of magnitude. Variation of site 2 alone does not affect the potency of the two compounds at the bottom. Thus, site 4 presents an SAR hotspot for this inhibitor series.

yields a minor increase in potency. Hence, in this series, the SAR is determined by two SAR hotspots at substitution sites 1 and 4.

For data sets spanning a more limited potency range, the presence of multiple steep activity cliffs is unlikely. However, CAGs highlight the most significant discontinuity contributions within a given data set and thus reveal SAR features that are characteristic for the data set. Figure 6.5(a) presents the CAG for 35 cytochrome P450 3a4 inhibitors with six different substitution sites that span a relatively narrow potency range. Nevertheless, at each layer in the graph, a number of different sites or site combinations are found that produce significant SAR discontinuity, for example, nodes 2, 2-4, 2-6, 2-4-6, or nodes 1-2-5 and 1-2-6. Different from the thrombin inhibitor series, the observed SAR discontinuity is not associated with well-defined node patterns in the CAG. For example, not all nodes including substitution site 2 obtain comparably high scores. However, similar to the thrombin series, individual SAR hotspots can be detected that are responsible for overall SAR discontinuity. As illustrated in Figure 6.6, simultaneous modification of sites 2, 4, and 6 induces a high degree of discontinuity, similar to individual modification of site 2. The corresponding modifications introduced separately at sites 4 and 6 do not have any measurable effect. Hence, site 2 is an SAR hotspot that largely determines the overall discontinuity within this compound series.

Furthermore, to demonstrate the significance of SAR hotspots in CAGs for analog selection, the analysis was repeated after removal of the most potent compounds with potency lower than 200 nM from the series. Figure 6.5(b) shows the CAG recalculated for the remaining 23 active compounds. As expected, the overall discontinuity decreases due to the more limited potency range. Comparison of both graphs in Figure 6.5 shows that SAR hotspots at nodes 1-2-4, 1-2-5 and 1-2-6 are retained, although in the second graph the most potent compounds were not taken into account. However, in this graph, nodes corresponding to the most potent compounds are now empty. These nodes capture variation of sites 2, 2-4 or 2-6. If we utilize the CAG representation in Figure 6.5(b) to predict which substitution sites should be further explored, combinations involving site 2 would have high priority because this site consistently contributes to high-scoring nodes and has not been thoroughly explored. Thus, we focus on site combinations capturing the most potent analogs in Figure 6.5(a). It follows that the information provided by CAGs can be utilized to identify molecular regions where changes are most likely to introduce SAR discontinuity and yield highly potent analogs.

6.2.3 SAR Holes

In addition to revealing SAR hotspots, CAG analysis readily identifies SAR holes, i.e. missing substituent combinations within analog series. This is il-

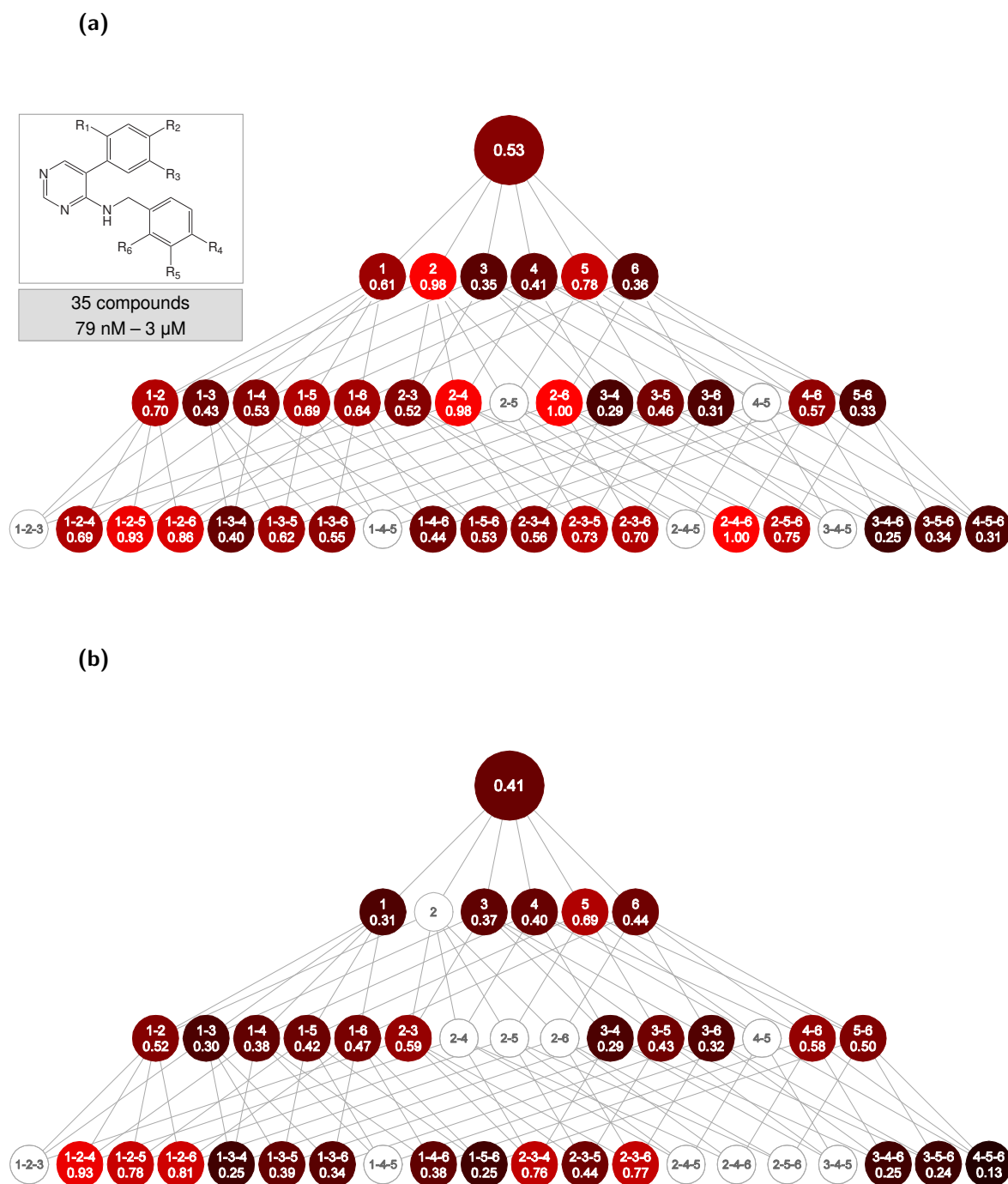


Figure 6.5: CAG for cytochrome P450 inhibitors (a) For a series of 35 analogs with six substitution sites, the CAG representation is shown together with the molecular scaffold shared by all compounds in this series. (b) CAG representation for the same analog series after removal of 12 inhibitors with potency lower than 200 nM. Nodes 2, 2–4, and 2–6 that present SAR hotspots in (a) are now empty.

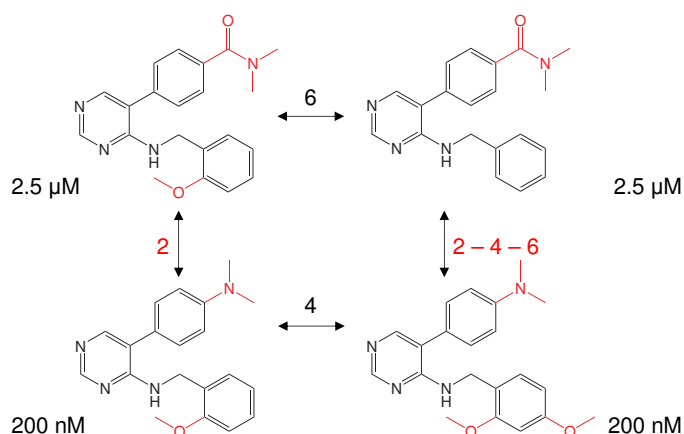


Figure 6.6: Inhibitors of cytochrome P450 3a4 Shown are four analogs that differ in their substituents at sites 2, 4, and 6, colored in red. Individual modifications at sites 4 and 6 have no effect on potency (compound pairs at the top and bottom, respectively), whereas simultaneous variations at sites 2, 4, and 6 lead to significant potency changes, similar to individual modification of site 2.

illustrated in Figure 6.7, which describes a set of analogous hydroxyacyl-CoA dehydrogenase II inhibitors with variations at up to six substitution sites. Similar to the P450 inhibitor series discussed above, this series also shows a notable degree of SAR heterogeneity (discontinuity score: 0.52), which is reflected by score variations between individual nodes. Substitution patterns at specific site combinations produce considerable SAR discontinuity, for example, sites 1–3, 3–6, 1–3–6, and 4–6. However, at the level of individual sites, no significant discontinuity contributions are observed. For site 3, which is involved in combinations that obtain highest discontinuity scores, individual variations have not been tested. Hence, this node remains empty and presents an SAR hole that needs to be explored in order to complete available SAR information. For this purpose, substitution patterns found at combinations of site 3 with other sites present promising starting points. Figure 6.8 shows two compound pairs with modifications at site 1 or at sites 1 and 3, corresponding to nodes 1 and 1–3. The compound pair at the top is distinguished by the presence of an ethyl acetate side chain at site 1 and a hydroxyl group at site 3. These two co-occurring modifications cause a considerable potency difference, consistent with the high score for node 1–3. By contrast, individual addition of the ethyl acetate group to site 1 in another analog has only a minor effect, as illustrated in Figure 6.8 (bottom). In order to elucidate the effect of the site 3 OH group, individual modifications of site 3 and of site 1 in the presence or absence of this group might be explored. In addition, the two compound pairs shown in Figure 6.8 have different R-group configurations at other sites, which might also influence their SAR behavior and should be tested individually.

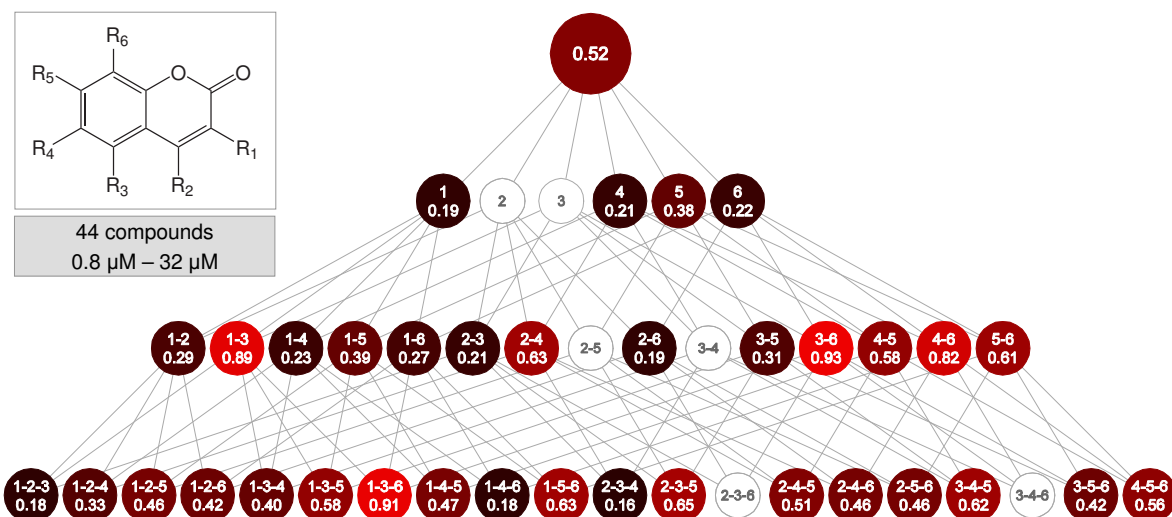


Figure 6.7: CAG for hydroxyacyl-CoA dehydrogenase II inhibitors Shown is the CAG representation for a series of 44 inhibitors of hydroxyacyl-CoA dehydrogenase II and their common core structure that contains six substitution sites.

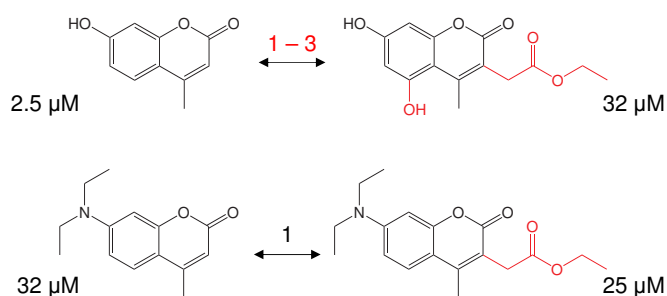


Figure 6.8: Hydroxyacyl-CoA dehydrogenase II inhibitors The two analogous inhibitors shown at the top differ at substitution sites 1 and 3 and have a potency difference of one order of magnitude. The two compounds at the bottom have corresponding substitutions at site 1 but have comparable potency.

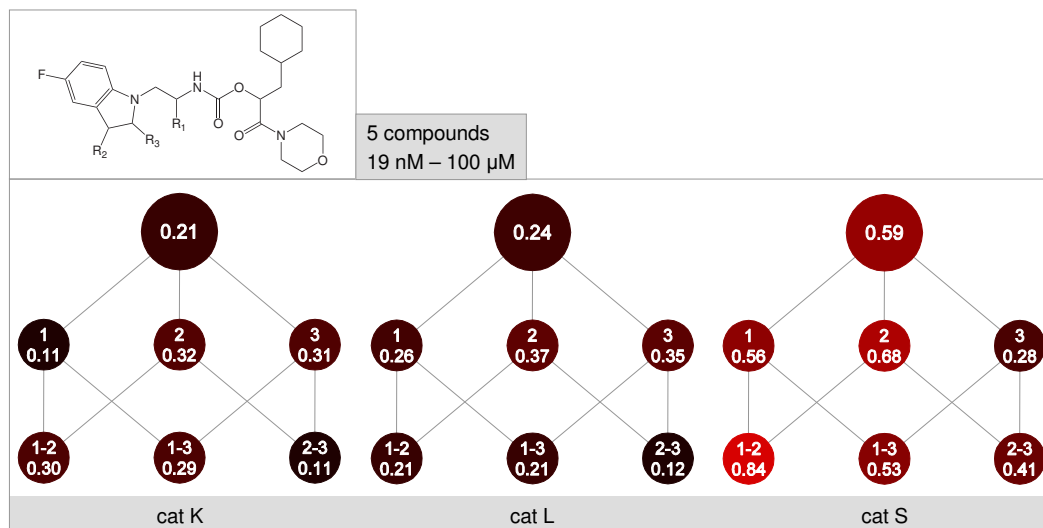
Furthermore, inspection of the compounds that correspond to CAG nodes reveals regions of the analog space that have not been thoroughly sampled. For example, nodes for site combinations 3–6, 1–3–6, and 4–6 indicate SAR hotspots, but the corresponding single-site nodes 1,4 and 6 display similarly low SAR discontinuity. Accordingly, it is not apparent from the CAG if the SAR discontinuity at these nodes can be assigned to individual sites or if it results from modifications at site combinations that act in concert. Thus, in order to estimate the influence of individual sites, the substituents found in nodes 3–6, 1–3–6, and 4–6 at these sites should be systematically varied without modifying other sites. In this manner, CAG representations can provide guidance for further analog design to complement existing SAR information.

6.3 SAR Determinants for Multiple Targets

CAG–SARI analysis was also applied to study multi-target SARs in series of cathepsin (cat) inhibitors with potency measurements against cat K, L, and S. Figure 6.9 shows the CAG representations for three different analog series that were found to inhibit the three related cysteine proteases at significantly different levels. Comparison of the graphs for related targets reveals variable SARs and substitution patterns that influence target-specific SARs in different ways.

The analog series in Figure 6.9(a) has very similar SAR characteristics for cat K and L. Scores for the entire series and all subsets are low and of comparable magnitude for both targets. This phenotype is indicative of flat SARs that often present difficult cases in medicinal chemistry because it remains unclear whether or not compounds can be further optimized. By contrast, this series behaves differently against cat S. Here, the overall discontinuity is intermediate and there is clear SAR heterogeneity among the substitution sites and their combinations, with node 1–2 presenting an apparent activity cliff. Accordingly, this series shows highest changes in potency for cat S and includes compounds that are highly selective for this target. Thus, substitution patterns observed in this series would be expected to offer greater potential for compound optimization against cat S than cat K or L. Furthermore, the inhibitor series in Figure 6.9(b) displays similar overall SAR discontinuity and score variability among substitution site combinations against cat L and S but differs in the behavior against cat K. In the latter case, SAR discontinuity is much reduced compared to the other two enzymes, with only node 2–3 displaying a considerable degree of discontinuity. This node also points at SAR hotspots in cat L and S. For these targets, however, SAR discontinuity is also observed for other substitution sites, e.g. sites 2 (L) or 3 (S). Variations at these sites are likely to determine compound selectivity for these targets. Hence, substitution sites

(a)



(b)

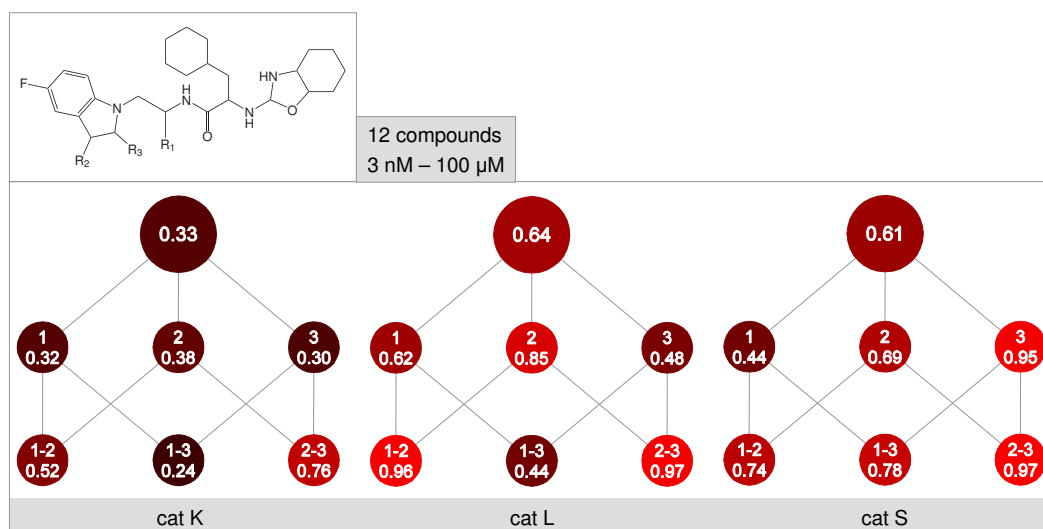


Figure 6.9: CAG for cathepsin inhibitors

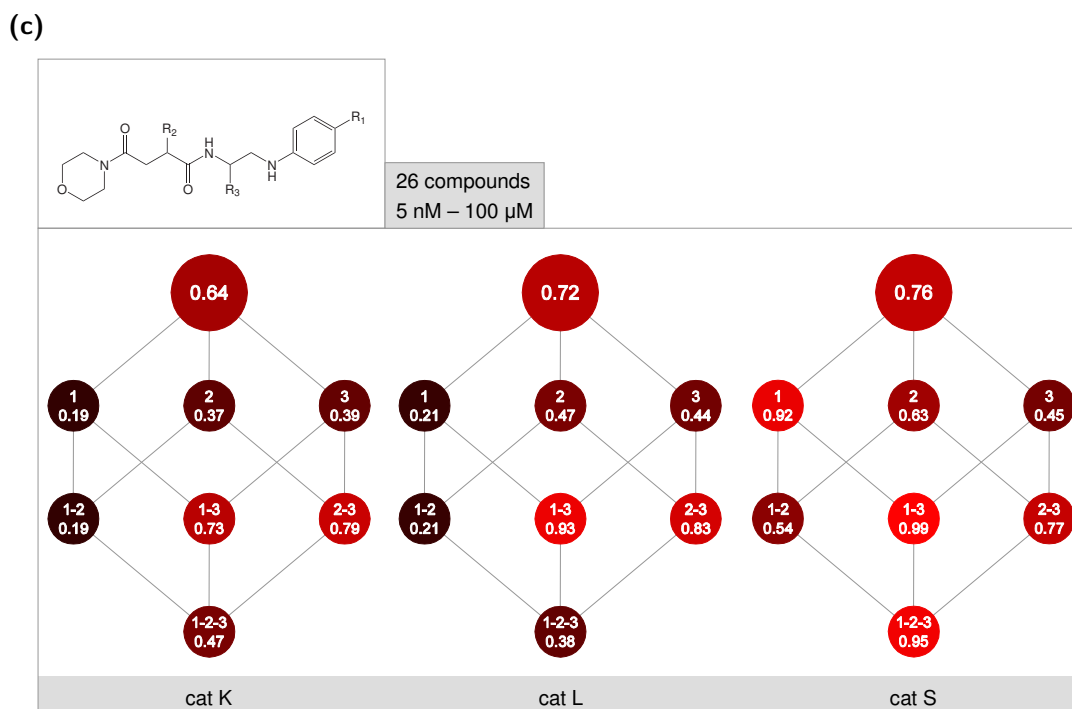


Figure 6.9: CAG for cathepsin inhibitors (continued) For three different series of cathepsin inhibitors, CAG representations are shown utilizing potency values for cat K, L, and S. Parts (a), (b), and (c) represent graphs and molecular scaffolds for individual series.

that indicate SAR hotspots of similar or different magnitude for related targets can be used to explore R-group substituents that determine target selectivity.

Differences in the distribution of SAR discontinuity are also observed for the analog series described in Figure 6.9(c). This series is characterized by a significant degree of discontinuity against all three enzymes and includes several SAR hotspots. For example, combinations 1–3 and 2–3 contribute strongly to SAR discontinuity for cat K, L, and S, whereas site 1 analogs have low discontinuity for cat K and L but high discontinuity for cat S. Individual sites 2 and 3 produce only low to moderate discontinuity levels. In this case, discontinuity in combinations of sites 2 and 3 results from a synergistic effect of two substituents at these sites, as illustrated in Figure 6.10. For the compound in the upper left part of the figure, separately exchanging the trifluoromethyl benzene at site 2 to a methyl cyclohexane or adding an isopropyl substituent at site 3 does not have a measurable effect on potency against cat K and only a weak effect for L. However, combination of these two variations induces a potency leap for both enzymes, leading to the molecule shown at the lower right part of Figure 6.10. For cat S, by contrast, the same modifications of sites 2 and 3 individually lead to a notable potency increase and achieve an additive effect when they are combined. Thus, different discontinuity levels in CAG nodes corresponding

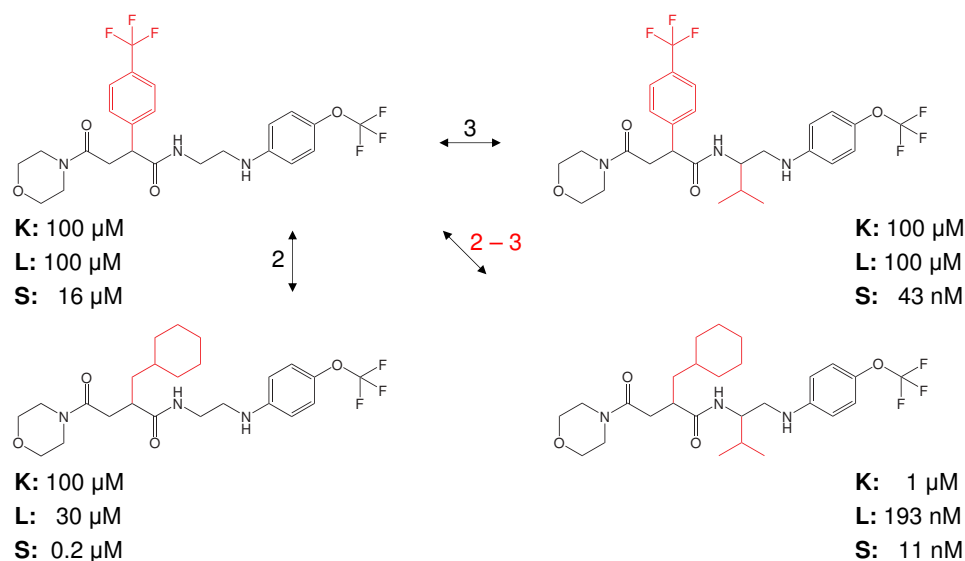


Figure 6.10: Cathepsin inhibitors Four analogous cathepsin inhibitors from the series in Figure 6.9(c) are shown that differ at sites 2 and 3. For the compound displayed at the top left, individual modifications at sites 2 and 3 have no significant effect on potency against cat K and L, but lead to an improved potency for cat S. Combination of the same variations at these sites yield the compound at the bottom right, which has significantly increased potency for all three related targets.

to single substitution sites and their combinations reveal the effects of specific substitution patterns and their mutual dependence.

6.4 Conclusions

By organizing analog series in hierarchical graph structures and applying a simple and robust scoring scheme, SAR contributions of substitution sites and their combinations have been quantitatively analyzed in a systematic manner. Contributions of molecular sites to the overall SAR character have been quantified using the SARI discontinuity score that yielded meaningful results also for relatively narrow potency ranges, as typically observed in screening data. Pursuing a whole-molecule approach, assessment of SAR discontinuity at the level of functional groups was enabled by using a compound organization scheme that divides analog series into subsets of compounds that are distinguished at well-defined substitution sites. This approach permits the exploration of SAR characteristics within large series of analogous compounds. The graph representations introduced herein make it possible to analyze the distribution of substitution site combinations in a straightforward and intuitive manner. Attention is immediately focused on SAR hotspots, i.e. site combinations that make largest

contributions to SAR discontinuity and include SAR determinants, which are prime targets for chemical optimization efforts. In addition, SAR holes and missing substitution combinations are readily discovered. Moreover, it is possible to compare multi-target SARs for series including highly optimized and selective compounds and describe differential characteristics in detail.

Our approach is distinguished from related methods such as SAR tables, analysis of matched molecular pairs (Leach and Law, 2006), and Free–Wilson analysis (Free and Wilson, 1964), by the systematic exploration of substitution sites and site combinations without directly considering the chemical nature of substituents. This makes it possible to prioritize sites in a molecule that are susceptible to chemical modifications that affect potency and thus provide promising starting points for chemical optimization. Focusing on combinations of substitution sites often reveals key substitutions for individual sites or, alternatively, the mutual dependence of individual sites and modifications that act in concert, thus departing from classical additive QSAR or Free–Wilson approaches. Taken together, our findings suggest that the CAG–SARI method has the potential to significantly aid in extracting SAR information from different compound series. By highlighting key substitution patterns, undersampled regions and differential SAR characteristics for related targets, the approach can guide analog design to complement existing SAR information and optimize compound potency and selectivity.

Chapter 7

Summary and Conclusions

This thesis focuses on the systematic computational analysis of structure–activity relationships (SARs) of small molecules. Guided by three central goals stated in the introductory chapter, several novel approaches have been introduced to characterize, quantify, and compare SARs in a systematic manner. The major results of this dissertation are summarized in this chapter. Figure 7.1 illustrates key aspects of the presented methods.

Goal 1: Design of a conceptual framework to systematically characterize and classify SARs present in sets of active molecules.

A comparative study of crystallographic enzyme–inhibitor complexes presented an initial step toward this goal. Comparison of 2D and 3D inhibitor similarity and potency revealed systematic and in part unexpected trends. A notable degree of variability in ligand structures and binding modes was observed even in the presence of severe structural constraints posed by the architecture of an enzyme’s active site. Furthermore, it was shown that different SAR features are not mutually exclusive but often coexist within classes of active compounds. The results revealed that relationships between similarity and potency are often complex and provide evidence of the heterogeneous nature of many SARs.

In order to put the evaluation of variable SARs on a formal and quantitative basis, a numerical scoring scheme was developed. Relying solely on 2D inhibitor similarity and potency data, the SAR Index (SARI) combines two individual scores that quantitatively describe continuous and discontinuous components of an activity landscape. The SARI function was designed to distinguish between three elementary SAR categories; continuous, discontinuous, and heterogeneous. These well-recognized SAR phenotypes were for the first time described in numerical terms. In addition, two previously unobserved subtypes of heterogeneous SARs were established that combine continuous and discontinuous elements in different ways. SARI calculations were applied to

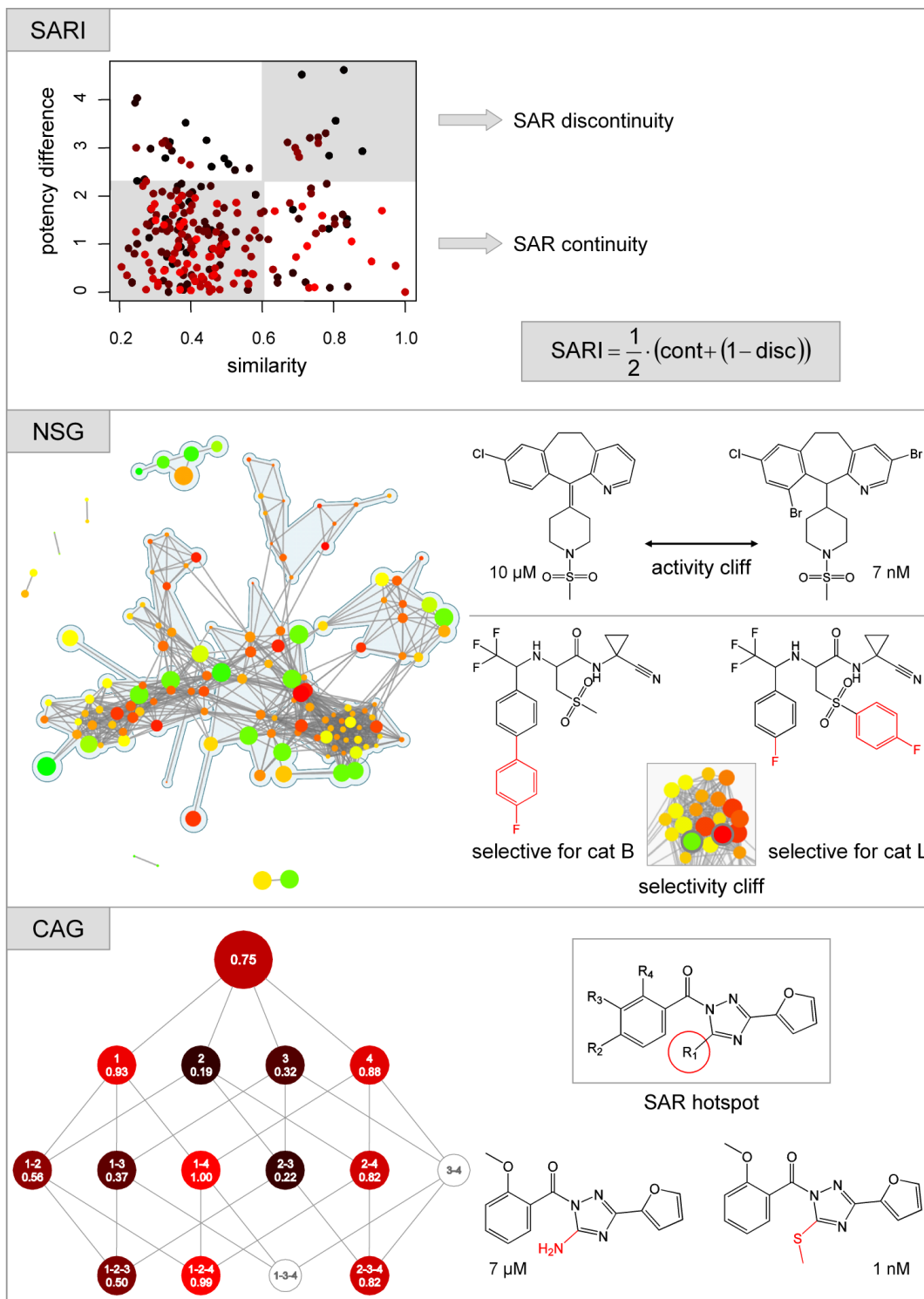


Figure 7.1: Graphical SAR analysis methods

profile various enzyme inhibitor classes. The results showed that heterogeneous SARs are prevalent among many classes, consistent with previous findings. Additional control calculations demonstrated the robustness of the SARI scoring scheme against variation of molecular representations and data set size. Hence, with the SARI framework, we have introduced a methodology that enables for the first time the quantitative classification and comparison of SARs on a large scale.

Goal 2: Development of a methodology to explore SARs at multiple levels of detail that enables the investigation of local SAR features and relationships between global and local SARs.

For the evaluation of different local SAR components that might coexist in compound classes, a methodology was developed to explore SARs at the level of entire compound classes, series of similar compounds, and individual molecules. Network-like Similarity Graphs (NSGs) were designed to visualize potency distributions and similarity relationships within compound classes. In these graphs, subsets of similar molecules were identified and provided the basis for local SAR characterization using SARI scores. Furthermore, a modified SARI score was introduced to assess SAR contributions from individual molecules. This made it possible to identify key compounds that were activity cliff markers and strongly influenced the SAR character of a collection of active molecules. The NSG approach combined with SARI scoring at multiple levels provides ways and means to dissect SAR phenotypes and relate local and global SAR features to each other. Hence, it is readily possible to elucidate multiple SAR components present in large data sets and prioritize compound subsets for further analysis and chemical optimization. The NSG–SARI approach was also applied to study structure–selectivity relationships (SSRs) within sets of compounds active against multiple related targets. Accounting for the fact that target selectivity often results from differences in compound potency against multiple targets, a comparative analysis of single-target SARs and target-pair SSRs was conducted. The quantitative SAR analysis approach was successfully adapted to evaluate SSRs, demonstrating that SSR phenotypes can be categorized in analogy to SARs. Different local SSRs were detected and compared to corresponding SAR features. In addition, key compounds involved in the formation of selectivity cliffs were identified, which made it possible to identify structural patterns that determined compound selectivity.

Goal 3: Quantitative evaluation of SAR contributions from functional groups and identification of sub-molecular SAR determinants.

In addition to the identification of individual molecules that were SAR and SSR determinants, we also investigated SAR contributions made by molecular

substructures. For this purpose, we focused on series of analogous molecules that shared a common molecular scaffold and were distinguished at well-defined substitution sites. Applying the SARI scoring scheme to subsets of molecules that differed only at specific substitution sites made it possible to relate observed SAR characteristics to different functional groups present at these sites. A hierarchical organization scheme termed Combinatorial Analog Graph (CAG) was devised to visualize the levels of SAR discontinuity that resulted from variations at individual substitution sites or combinations of sites. Hence, key substitution patterns that were responsible for SAR discontinuity and thus presented the most promising starting points for chemical optimization could immediately be identified. In addition, CAG representations also highlight substitution sites that have not been thoroughly explored. Thus, the CAG-SARI approach enables the intuitive analysis of SAR contributions from functional groups and can be used to guide analog design in a directed manner.

In summary, the approaches introduced in this dissertation provide the opportunity to systematically explore different aspects of small-molecule SARs in a quantitative manner. Departing from conventional case-by-case analysis, these methods complement and extend existing approaches. Key aspects are their ability to quantify SARs on a large scale and characterize SARs at different levels of detail. Graphical representation of SAR features plays a central role for the intuitive application of these methods and the interpretation of the results. While the systematic analysis of SARs is still a relatively new area of research, a paradigm shift in the SAR analysis field can be anticipated in the coming years. Future challenges include the integration of SAR analysis, compound selection, and prospective compound design, as well as the incorporation of other parameters such as bioavailability criteria or chemical accessibility into the currently potency-centric SAR analysis methodologies.

Bibliography

- Agrafiotis, D. K., Bandyopadhyay, D., and Farnum, M. (2007a). Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.*, *47*, 69–75.
- Agrafiotis, D. K., Shemanarev, M., Connolly, P. J., Farnum, M., and Lobanov, V. S. (2007b). SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.*, *50*, 5926–2937.
- Ankerst, M., Kastenmüller, G., Kriegel, H., and Seidl, T. (1999). 3D shape histograms for similarity search and classification in spatial databases. In R. H. Gutting, D. Papadias, and F. Lochovsky (Eds.), *Advances in spatial databases* (pp. 207–226). Berlin/Heidelberg: Springer.
- Bajorath, J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, *41*, 233–245.
- Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, *1*, 882–894.
- Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, *39*, 2887–2893.
- Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004). Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.*, *44*, 170–178.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucl. Acids Res.*, *28*, 235–242.
- Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2006). Evolving interpretable structure–activity relationships. 1. Reduced graph queries. *J. Chem. Inf. Model.*, *46*, 577–586.

- Boström, J., Hogner, A., and Schmitt, S. (2006). Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.*, *49*, 6716–6725.
- Bush, B. L., and Sheridan, R. P. (1993). PATTY: a programmable atom type and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.*, *33*, 756–762.
- Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.
- Eckert, H., and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, *12*, 225–233.
- Esposito, E. X., Hopfinger, A. J., and Madura, J. D. (2004). Methods for applying the quantitative structure–activity relationship paradigm. *Methods Mol. Biol.*, *275*, 131–214.
- Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, *27*, 2985–2993.
- Free, S. M., and Wilson, J. W. (1964). A mathematical contribution to structure–activity studies. *J. Med. Chem.*, *7*, 395–399.
- Fruchterman, T. M. J., and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exper.*, *21*, 1129–1164.
- Good, A. C., Ewing, T. J. A., Gschwend, D. A., and Kuntz, I. D. (1995). New molecular shape descriptors: application in database screening. *J. Comput. Aided Mol. Des.*, *9*, 1–12.
- Guha, R., and van Drie, J. H. (2008). Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, *48*, 646–658.
- Haigh, J. A., Pickup, B. T., Grant, J. A., and Nicholls, A. (2005). Small molecule shape-fingerprints. *J. Chem. Inf. Model.*, *45*, 673–684.
- Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I., and Shoichet, B. K. (2008). Quantifying the relationships among drug classes. *J. Chem. Inf. Model.*, *48*, 755–765.
- Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, *4*, 682–690.
- Johnson, M. A., and Maggiora, G. M. (Eds.). (1990). *Concepts and applications of molecular similarity*. New York: John Wiley & Sons.

- Johnson, S. R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.*, *48*, 25–26.
- Kibbey, C., and Calvet, A. (2005). Molecular property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.*, *45*, 523–532.
- Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, *44*, 98–104.
- Kubinyi, H. (1977). Quantitative structure–activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J. Med. Chem.*, *20*, 625–629.
- Kubinyi, H. (1997). QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discovery Today*, *2*, 457–467.
- Kubinyi, H. (1998). Similarity and dissimilarity: a medicinal chemist’s view. *Persp. Drug Discov. Des.*, *9-11*, 225–252.
- Labute, P., Williams, C., Feher, M., Sourial, E., and Schmidt, J. M. (2001). Flexible alignment of small molecules. *J. Med. Chem.*, *44*, 1483–1490.
- Leach, A. G., and Law, B. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.*, *49*, 6672–6682.
- Lemmen, C., and Lengauer, T. (2000). Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.*, *14*, 215–232.
- Maggiora, G. M. (2006). On outliers and activity cliffs—why QSAR often disappoints. *J. Chem. Inf. Model.*, *46*, 1535.
- Manallack, D. T., Ellis, D. D., and Livingstone, D. J. (1994). Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.*, *37*, 3758–3767.
- Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, *45*, 4350–4358.
- Mason, J. S., Good, A. C., and Martin, E. J. (2001). 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.*, *7*, 567–597.
- Medina-Franco, J. L., Martínez-Mayorga, K., Bender, A., Marín, R. M., Giulianotti, M. A., Pinilla, C., and Houghten, R. A. (2009). Characterization of activity landscapes using 2D and 3D similarity methods: consensus ac-

- tivity cliffs. *J. Chem. Inf. Model.*, *49*, 477–491.
- Mestres, J., Martin-Couce, L., Gregori-Puigjane, E., Cases, M., and Boyer, S. (2006). Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J. Chem. Inf. Model.*, *46*, 2725–2736.
- Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nat. Biotech.*, *24*, 805–815.
- Papadatos, G., Cooper, A. W. J., Kadiramanathan, V., Macdonald, S. J. F., McLay, I. M., Pickett, S. D., Pritchard, J. M., Willett, P., and Gillet, V. J. (2009). Analysis of neighborhood behavior in lead optimization and array design. *J. Chem. Inf. Model.*, *49*, 195–208.
- Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. (1996). Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.*, *39*, 3049–3059.
- Peltason, L., and Bajorath, J. (2007a). Molecular similarity analysis uncovers heterogeneous structure–activity relationships and variable activity landscapes. *Chem. Biol.*, *14*, 489–497.
- Peltason, L., and Bajorath, J. (2007b). SAR index: quantifying the nature of structure–activity relationships. *J. Med. Chem.*, *50*, 5571–5578.
- Peltason, L., and Bajorath, J. (2009). Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. *Future Med. Chem.*, *1*, 451–466.
- Peltason, L., Hu, Y., and Bajorath, J. (2009a). From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem*, *in press*.
- Peltason, L., Weskamp, N., Teckentrup, A., and Bajorath, J. (2009b). Exploration of structure–activity relationship determinants in analogue series. *J. Med. Chem.*, *52*, 3212–3224.
- Schneider, G., Schneider, P., and Renner, S. (2006). Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.*, *25*, 1162–1171.
- Schuur, J. H., Selzer, P., and Gasteiger, J. (1996). The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.*, *36*, 334–344.

- Senese, C. L., Duca, J., Pan, D., Hopfinger, A. J., and Tseng, Y. J. (2004). 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.*, *44*, 1526–1539.
- Shanmugasundaram, V., and Maggiora, G. M. (2001). *Characterizing property and activity landscapes using an information-theoretic approach*. (Presented at the 222nd American Chemical Society National Meeting, Division of Chemical Information, 2001; Abstract No. 77.)
- Sheridan, R. P., and Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today*, *7*, 903–911.
- Sheridan, R. P., Rusinko, A., Ramaswamy, N., and Venkataraghavan, R. (1989). Searching for pharmacophores in large coordinate data bases and its use in drug design. *Proc. Natl. Acad. Sci. U.S.A.*, *86*, 8165–8169.
- Stumpfe, D., Ahmed, H. E. A., Vogt, I., and Bajorath, J. (2007). Methods for computer-aided chemical biology. Part 1: Design of a benchmark system for the evaluation of compound selectivity. *Chem. Biol. Drug Des.*, *70*, 182–194.
- Stumpfe, D., Geppert, H., and Bajorath, J. (2008). Methods for computer-aided chemical biology. Part 3: Analysis of structure–selectivity relationships through single- or dual-step selectivity searching and bayesian classification. *Chem. Biol. Drug Des.*, *71*, 518–528.
- Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., and Timmerman, H. (2000). *Handbook of molecular descriptors* (Vol. 11). Weinheim: Wiley-VCH.
- Vogt, I., Stumpfe, D., Ahmed, H. E. A., and Bajorath, J. (2007). Methods for computer-aided chemical biology. Part 2: Evaluation of compound selectivity using 2D molecular fingerprints. *Chem. Biol. Drug Des.*, *70*, 195–205.
- von Korff, M., Freyss, J., and Sander, T. (2008). Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *J. Chem. Inf. Model.*, *48*, 797–810.
- Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.*, *47*, 2977–2980.
- Wang, R., Fang, X., Lu, Y., Yang, C., and Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem.*, *48*, 4111–4119.

- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.
- Wawer, M., and Bajorath, J. (2009). Extraction of structure–activity relationship information from high-throughput screening data. *Curr. Med. Chem.*, *in press*.
- Wawer, M., Peltason, L., and Bajorath, J. (2009). Elucidation of structure–activity relationship pathways in biological screening data. *J. Med. Chem.*, *52*, 1075–1080.
- Wawer, M., Peltason, L., Weskamp, N., Teckentrup, A., and Bajorath, J. (2008). Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices. *J. Med. Chem.*, *51*, 6075–6084.
- Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, *11*, 1046–1053.
- Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, *38*, 983–996.

Appendix A

Software and Databases

Fingerprint methods, databases, and software used in this dissertation are listed in alphabetical order.

Table A.1: Fingerprints

MACCS	MACCS structural keys
Description	MACCS structural keys are a binary molecular fingerprint. In this thesis, the publicly available set of 166 bits coding for 166 structural fragments was utilized for representation of molecular structures.
Provider	Symyx Software, San Ramon, CA (USA)
URL	http://www.symyx.com/
<hr/>	
Molprint2D	
Description	Molprint2D is a topological molecular fingerprint that is based on layered atom environments (Bender et al., 2004).
Provider	Unilever Centre for Molecular Science Informatics, University of Cambridge, Cambridge (UK)
URL	http://www.molprint.com/
<hr/>	
TGT	Typed Graph Triangles
Description	TGT is a topological 3-point pharmacophore fingerprint implemented in MOE that encodes graph distances between triplets of typed pharmacophore points in a molecule.
Provider	Chemical Computing Group Inc., Montreal, QC (Canada)
URL	http://www.chemcomp.com/

Table A.2: Databases

MDDR	MDL Drug Data Report
Description	The MDDR is a commercial database containing over 150 000 biologically active compounds from patent literature, journals, meetings, and congresses.
Provider	Symyx Software, San Ramon, CA (USA)
URL	http://www.symyx.com/
<hr/>	
PDBbind	
Description	The PDBbind database is a comprehensive collection of experimentally measured binding affinity data for protein–ligand complexes deposited in the Protein Data Bank (Berman et al., 2000; Wang et al., 2004).
Provider	Shaomeng Wang Laboratory, University of Michigan, Ann Arbor, MI (USA)
URL	http://www.pdbbind.org/
<hr/>	
PubChem BioAssay	
Description	The PubChem BioAssay database contains results from more than 1700 bioactivity screens of chemical substances, including over 700 confirmatory assays that provide quantitative potency measurements.
Provider	National Center for Biotechnology Information (NCBI), Bethesda, MD (USA)
URL	http://pubchem.ncbi.nlm.nih.gov/

Table A.3: Software

MOE	Molecular Operating Environment
------------	---------------------------------

Description	MOE is a chemical computing and molecular modeling tool that provides a variety of chemoinformatics applications, including an implementation of the 166 publicly available MACCS keys.
Provider	Chemical Computing Group Inc., Montreal, QC (Canada)
URL	http://www.chemcomp.com/

Pipeline Pilot	
-----------------------	--

Description	Scitegic Pipeline Pilot is a graphical software for creating workflow protocols and provides components for data analysis and various scientific applications.
Provider	Accelrys Inc., San Diego, CA (USA)
URL	http://www.accelrys.com/products/scitegic/

R	The R Project for Statistical Computing
----------	---

Description	R is a language and free 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/

Appendix B

Enzyme–Inhibitor Complexes

Crystallographic structures of the enzyme–inhibitor complexes analyzed in Chapter 2 are summarized in Table B.1. For each enzyme, the PDB codes of the studied complex structures are given. In addition, the corresponding inhibitors are identified by their unique PDB ligand ID (in parentheses). Peptide inhibitors do not obtain a unique ligand identifier and are instead signified ‘ n -mer’, where n denotes the number of residues.

Table B.1: Crystallographic enzyme–inhibitor structures

carbonic anhydrase II	elastase	factor Xa	ribonuclease A
1a42 (BZO)	1bma (4-mer)	1ezq (RPR)	1afk (PAP)
1avn (HSM)	1eas (TFK)	1f0r (815)	1afl (ATR)
1bcd (FMS)	1eat (TFI)	1f0s (PR2)	1jn4 (139)
1bn1 (AL5)	1ela (4-mer)	1fjs (Z34)	1o0f (A3P)
1bn3 (AL6)	1elb (4-mer)	1g2l (T87)	1o0h (ADP)
1bn4 (AL9)	1elc (4-mer)	1ksn (FXV)	1o0m (U2P)
1bnm (AL1)	1eld (4-mer)	1lpg (IMA)	1o0n (U3P)
1bnq (AL4)	1ele (4-mer)	1lpk (CBB)	1o0o (A2P)
1bnt (AL2)	1gvk (4-mer)	1lpz (CMB)	1qhc (PUA)
1bnu (AL3)	1h9l (4-mer)	1mq5 (XLC)	
1bnv (AL7)	1inc (ICL)	1mq6 (XLD)	
1bnw (TPS)	1qr3 (8-mer)	1nfu (RRP)	
1cil (ETS)	4est (5-mer)	1nfw (RRR)	
1cim (PTS)	5est (3-mer)	1nfx (RDR)	
1cin (MTS)		1nfy (RTR)	
1cnw (EG1)		1xka (4PP)	
1cnx (EG2)			
1cny (EG3)			
1g1d (FSB)			
1g52 (F2B)			
1g53 (F6B)			
1g54 (FFB)			
1if7 (SBR)			
1if8 (SBS)			
1okl (MNS)			
1okn (STB)			
1ttm (667)			

Appendix C

SAR Tables

The following SAR tables report substituents ('R1', 'R2', ...) and potency values ('pot') for all compounds in the seven analog series discussed in Chapter 6. Compounds from PubChem BioAssay data are identified by their unique PubChem CID. Compounds from selectivity data sets are identified by an arbitrarily assigned index. Attachment points are marked with 'Z'.

Table C.1: Hydroxysteroid-17 β -dehydrogenase 4 inhibitors

CID	pot [μ M]	R1	R2	R3
890639	13	Z		Z
890163	25			Z-NH ₂
2938438	25		Z	Z
662549	32			Z
2938604	32		Z	Z-O

Table C.2: Thrombin inhibitors

CID	pot [nM]	R1	R2	R3	R4
977140	1	Z-S			Z-O
1088427	82	Z-S			
1088428	159	Z-S	Z		
976363	204	Z-S	Z-O	Z-O	
828590	741	Z-NH ₂	Z		
1084416	828	Z-NH ₂	Z-O		
828588	926	Z-NH ₂			
828591	1462	Z-NH ₂	Z-Cl		
969825	2227	Z-NH ₂	Z-O	Z-O	
969710	6933	Z-NH ₂			Z-O
828593	13951	Z-NH ₂			Z-Cl

Table C.3: Cytochrome P450 3a4 inhibitors

CID	pot [nM]	R1	R2	R3	R4	R5	R6
3235235	79						
3235476	79						
3234995	100						
3234666	126						
3235489	126						
3234829	158						
3232982	199						
3233999	199						
3234568	199						
3234784	199						
3234813	199						
3235150	199						
3232886	251						
3233050	251						
3235328	251						
3232698	316						
3233287	316						
3233374	316						
3234593	316						
3235521	316						
3233147	398						
3233799	398						
3233983	398						
3234079	398						
3235193	398						

Table C.3: Cytochrome P450 3a4 inhibitors (continued)

3233488	501				
3234501	501				
3235200	501				
3232915	631				
3233258	631				
3235026	631				
3234812	1259				
3234434	1995				
3232748	2512				
3234651	2512				

Table C.4: Hydroxyacyl-CoA dehydrogenase II inhibitors

CID	pot [μ M]	R1	R2	R3	R4	R5	R6
5273569	0.8						
716094	1.3						
5421461	2.0						
5280567	2.5						
5310805	2.5						
5740001	2.5						
92249	3.2						
93864	3.2						
5384392	3.2						
935137	3.2						
4223974	3.2						
5553318	3.2						
5739939	3.2						
5739885	3.2						
5351036	4.0						
731730	4.0						
889783	4.0						
3237311	4.0						
390799	4.0						
179503	10.0						
1212755	12.6						
5281416	12.6						
778746	15.8						
933257	20.0						
1751698	20.0						

Table C.4: Hydroxyacyl-CoA dehydrogenase II inhibitors (continued)

148769	25.1		Z/		
646866	25.1		Z/		Z/
875166	25.1		Z/		Z/
647115	25.1		Z/		
1799746	25.1		Z/		
659294	25.1				
879545	25.1		Z/		
2955775	25.1		Z/		
889425	25.1		Z/		
890072	25.1		Z/		
906996	25.1		Z/		
843236	25.1	Z/	Z/		
662287	25.1				Z/
7050	31.6		Z/		
890738	31.6		Z/		
975169	31.6		Z/		
666304	31.6	Z/	Z/		Z/
663048	31.6				Z/
5381321	31.6		Z/		

Table C.5: Cathepsin inhibitors

(a)

compound	pot cat K [nM]	pot cat L [nM]	pot cat S [nM]	R1	R2	R3
1	100000	100000	19	Z/	Z<	
2	100000	100000	80	Z/		Z<
3	30000	100000	143	Z-CH ₂ OH		
4	30000	30000	226	Z/		
5	30000	30000	2950			

Table C.5: Cathepsin inhibitors (continued)

(b)



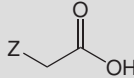



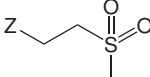

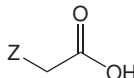
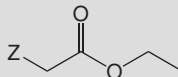

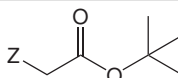

compound	pot cat K [nM]	pot cat L [nM]	pot cat S [nM]	R1	R2	R3
1	3710	123	3			
2	20000	10000	12			
3	100000	4670	15			
4	14700	849	15			
5	4870	369	21			
6	10000	2830	26			
7	100000	70000	27			
8	100000	50000	71			
9	30000	9670	151			
10	30000	10000	222			
11	100000	100000	730			
12	100000	30000	12300			Z=O

Table C.5: Cathepsin inhibitors (continued)

(c)

compound	R1	R2	R3	pot cat K [nM]	pot cat L [nM]	pot cat S [nM]
1				260	98	5
2				995	193	11
3				217	2678	12
4				557	285	13
5				907	267	16
6				100000	100000	21
7				11530	11530	24
8				11530	11530	25
9				84	372	27
10				24290	18900	28
11				3706	30000	31
12				1140	8990	38
13				100000	100000	43
14				291	588	60
15				5410	1520	69
16				100000	100000	70
17				30000	65000	105
18				30000	100000	134
19				100000	30000	194

Table C.5: Cathepsin inhibitors (continued)

20				2331	100000	408
21				30000	30000	1368
22				100000	100000	1590
23				100000	100000	15600
24				100000	30000	18530
25				100000	100000	100000
26				100000	100000	100000

Eidesstattliche Erklärung

An Eides statt versichere ich, dass ich die Dissertation

“Systematic Computational Analysis of Structure–Activity Relationships”

selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist.

Peltason, L., and Bajorath, J. (2007). Molecular similarity analysis uncovers heterogeneous structure–activity relationships and variable activity landscapes. *Chem. Biol.*, *14*, 489–497.

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

Peltason, L., and Bajorath, J. (2008). Molecular similarity analysis in virtual screening. In A. Varnek and A. Tropsha (Eds.), *Cheminformatics: an approach to virtual screening* (pp. 120–149), Cambridge:RSC Publishing.

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

Wawer, M., Peltason, L., and Bajorath, J. (2009). Elucidation of structure–activity relationship pathways in biological screening data. *J. Med. Chem.*, *52*, 1075–1080.

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

Peltason, L., and Bajorath, J. (2009). Systematic computational analysis of structure–activity relationships: concepts, challenges, and recent advances. *Future Med. Chem.*, *1*, 451–466.

Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., and van Drie, J. (2009). Navigating structure–activity landscapes. *Drug Discovery Today*, *14*, 698–705.

Peltason, L.* , Hu, Y.* , and Bajorath, J. (2009). From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem*, *in press*.

Sisay, M. T.* , Peltason, L.* , and Bajorath, J. (2009). Structural interpretation of activity cliffs revealed by systematic analysis of structure–activity relationships in analog series. *J. Chem. Inf. Model.*, *in press*.

* shared first authorship

Bonn, den 05. Oktober 2009

Lisa Peltason