Computational Methods for Structure-Activity Relationship Analysis and Activity Prediction

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Disha Gupta-Ostermann

aus Kota, Indien

Bonn

May, 2015

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

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

Tag der Promotion: 20 October, 2015
Erscheinungsjahr: 2015

Abstract

Structure-activity relationship (SAR) analysis of small bioactive compounds is a key task in medicinal chemistry. Traditionally, SARs were established on a case-by-case basis. However, with the arrival of high-throughput screening (HTS) and synthesis techniques, a surge in the size and structural heterogeneity of compound data is seen and the use of computational methods to analyse SARs has become imperative and valuable.

In recent years, graphical methods have gained prominence for analysing SARs. The choice of molecular representation and the method of assessing similarities affects the outcome of the SAR analysis. Thus, alternative methods providing distinct points of view of SARs are required. In this thesis, a novel graphical representation utilizing the canonical scaffold-skeleton definition to explore meaningful global and local SAR patterns in compound data is introduced.

Furthermore, efforts have been made to go beyond descriptive SAR analysis offered by the graphical methods. SAR features inferred from descriptive methods are utilized for compound activity predictions. In this context, a data structure called SAR matrix (SARM), which is reminiscent of conventional R-group tables, is utilized. SARMs suggest many virtual compounds that represent as of yet unexplored chemical space. These virtual compounds are candidates for further exploration but are too many to prioritize simply on the basis of visual inspection. Conceptually different approaches to enable systematic compound prediction and prioritization are introduced. Much emphasis is put on evolving the predictive ability for prospective compound design. Going beyond SAR analysis, the SARM method has also been adapted to navigate multi-target spaces primarily for analysing compound promiscuity patterns. Thus, the original SARM methodology has been further developed for a variety of medicinal chemistry and chemogenomics applications.

Acknowledgments

I would like to express deep gratitude to my supervisor Prof. Dr. Jürgen Bajorath for providing me with this excellent opportunity to pursue the doctoral studies and for his constant guidance and support.

I thank Prof. Dr. Michael Gütschow for reviewing my thesis as a co-referent. I also thank Prof. Dr. Thorsten Lang and Prof. Dr. Thomas Schultz for being members of the review committee.

I extend my gratitude to all the colleagues of the LSI group for providing a nice working and learning atmosphere. I further thank Jenny Balfer, Dr. Ye Hu and Dr. Vigneshwaran Namasivayam for the fruitful collaborations. Special thanks to the lunch group for all the fun times spent in the Mensa.

I would like to thank Boehringer Ingelheim for supporting this thesis. Especially I'd like to thank Dr. Peter Haebel and Dr. Nils Weskamp for the helpful discussions and their hospitality.

Further, I would like to thank my family for showering their love on me. Finally, I would like to thank Björn and his family, for being a persistent support during my studies.

Contents

Chapter 1

Introduction

The modern drug discovery process is a complex multistage process that focuses on the development of novel drugs, i.e., chemical entities that elicit a desired response in the biological system by acting on target(s) of interest. The structure of these small molecules plays an important role in their interactions with corresponding biological target(s). Understanding the *structure-activity relationships* (SARs) of bioactive molecules is a key task in medicinal chemistry.[1,2] Since the 1960s, computational approaches have been deployed for SAR exploration.[2] A central principle that underlies SAR analysis is the "similarity property principle" (SPP) which states that similar molecules should have similar properties.[3] The description of molecular structures and the assessment of molecular similarities is critical for conducting relevant SAR studies and obtaining meaningful results.

Molecular Representations and Similarity

The SPP principle is not easy to capture methodologically because the problem lies in defining (dis)similarity in a consistent manner. The assessment of structural similarity of compounds depends on the computational representation of molecules and the similarity metric. Hence, the choice of the representation and similarity metric influences SAR analysis.[4]

The simplest way to represent a molecule is by its empirical formula. This is a one-dimensional (1D) representation. However, one formula represents multiple molecules because it does not contain structural information. Linear notations such as Simplified Molecular Input Line Entry System (SMILES)[5] have been developed, which represent the structural information of molecules in an unambiguous, reproducible and universal manner.

A more intuitive and popular way to represent a molecule is to use its two-dimensional (2D) molecular graph. In a graph, atoms are represented as nodes using atomic symbols and edges correspond to bonds. Hence, the graph represents the topology of the molecule and can be encoded in the form of a connection table. This connection table comprises a sequential list of atoms and a list of bonds connecting these atoms.

Molecules can also be represented in three dimensions (3D) by accounting for the spatial arrangement of their atoms.

Molecular Descriptors

In computational medicinal chemistry, there is no gold standard by which molecules should be represented. One of the most widely used ways is the application of molecular descriptors. Molecular descriptors are mathematical functions that characterize structural and/or physicochemical properties of molecules as numerical values. With the help of these numerical descriptors computational chemical reference spaces in which molecules are projected can be generated.[6] Chemical (dis)similarity is then defined through the intermolecular distance in the space.

A large number of descriptors have been defined that vary in complexity.[7] In general, descriptors can be classified based on the dimensionality of the molecular representations from which they are calculated. For example, 1D descriptors include molecular weight and atom counts, such as the number of carbon or oxygen atoms. These descriptors are calculated from 1D representation of molecules (chemical formula). 2D descriptors are derived on the basis of 2D molecular graphs that characterize, for example, physicochemical or topological properties, such as octanol/water partition coefficient (logP) or various

topological indices. 3D descriptors are generated from 3D conformations, such as pharmacophores and surface area.

Molecular Fingerprints

Apart from numerical representations of molecular structures and properties, bit string representations are also popular. *Molecular fingerprints* (FPs) are bit string representations of chemical structures and properties, which are often encoded in binary formats. The presence and absence of a given feature in the molecule is indicated by setting the corresponding bit to 1 and to 0, respectively.

As with numerical descriptors, FPs can be categorized into 2D or 3D depending on whether the chemical features describing the bit positions are derived from 2D or 3D molecular graph representations. Over the past decades, various FPs have been introduced that vary in their design, composition and length, based on which different FP prototypes can be defined.[8]

FPs in which each feature is assigned to a specific bit position are called keyed FPs. These FPs usually have fixed length, such as Molecular ACCess System (MACCS)[9] that contains 166 predefined structural fragments (substructures).

By contrast, combinatorial FPs capture layered atom environments in molecules up to a predefined bond diameter. Instead of predefined feature sets, molecule-specific features are calculated from individual compounds and thus the corresponding FPs would have a variable length. In addition, each feature is hashed into an integer number that represents the final feature set for a molecule. The most popular combinatorial FPs are the extended connectivity FPs (ECFPs).[10] An important feature of combinatorial FPs is that they capture atom environments in a molecule.

Pharmacophore patterns can be captured by pharmacophore FPs. "Pharmacophores are 3D (or 2D) arrangement of groups (functionalities) in a compound responsible for its bioactivity".[8] In pharmacophore FPs, bit positions are assigned to possible pharmacophore patterns encoded by conformers of a molecule. Pharmacophore patterns are typically defined by triplet or quadruplet feature points and inter-feature distance ranges. These FPs typically con-

tain very large number of bit positions. A comparison of the three different types of FPs is presented in Figure 1.1. For a common molecule three different FP representations are encoded as bit strings.

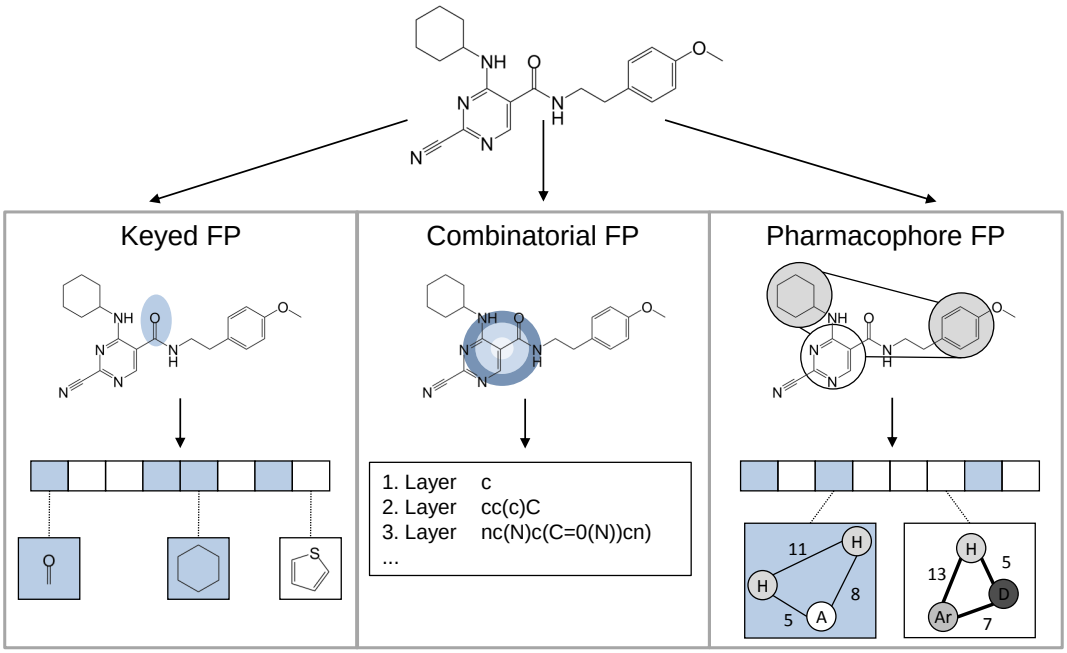


Figure 1.1: Molecular fingerprints. Three different (keyed, combinatorial and pharmacophore) FP designs are shown. Structural information used to obtain the corresponding FP representation is highlighted. Blue- and white-colored bits indicate the presence and absence, respectively, of specific structural features or arrangements in the molecule. Taken from [8].

A number of similarity metrics are available to quantitatively assess similarity between a pair of molecular FPs.[11] The underlying concept is to account for common and distinct structural features. The most widely applied measure is the Tanimoto coefficient (Tc)[11] that counts the number of bits common to two binary FPs with respect to the total number of unique bits that are set on in each FP. Accordingly, the Tc for two binary FP representations A and B is calculated as follows:

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

where c is the number of bits set on in both FPs and a and b refer to the number of bits set on in A and B, respectively. Tc value ranges between 0 and 1, where

0 corresponds to no FP overlap and 1 to identical FPs. However, it should be noted that identical FPs do not necessarily correspond to identical molecules because FPs are only a generalization of the molecular structures.

Depending on the FP one uses, it is very difficult to decide whether a given Tc value indicates the presence of "significant similarity" or not.[4,12] Furthermore, it is difficult to relate specific structural changes in pairs of molecules to quantified similarity values. Thus, the FP-based similarity measure is often difficult for medicinal chemists to use. Substructure-based representations can be chemically more intuitive to relate SARs and guide novel compound designs.

Molecular Scaffolds

The concept of *scaffolds*, which is popular in medicinal chemistry, accounts for a substructure-based representation of molecules. Scaffolds are generally used to describe core structures of molecules that are utilized in drug design or used as building blocks for compound synthesis.[13]

Many different definitions of scaffolds exist. The most widely used definition was introduced by Bemis and Murcko.[14] Bemis and Murcko (BM) scaffolds are generated by removing all side chains from the molecules and retaining ring systems and linkers. This enables the consistent generation of scaffolds and provides a sound basis for molecular framework-based SAR analysis. Following this definition, multiple BM scaffolds with minor differences in their heteroatoms and/or bond orders, are considered structurally distinct. BM scaffolds can be further abstracted to "cyclic skeletons" (CSKs)[15] by changing each heteroatom to carbon and setting all double, triple and aromatic bonds to single bonds. Thus, topologically equivalent BM scaffolds are represented by a common CSK. Figure 1.2 illustrates the compound-scaffold-skeleton hierarchy. Each scaffold represents one or more compounds and each CSK covers one or more scaffolds that share the same topology.

BM scaffolds and CSKs have been used to analyze the diversity of known drugs[13,14] and SAR trends in compound data.[16,17] However, the hierarchical scaffold definition has limitations. For example, the addition of a ring to an existing BM scaffold creates per definition a distinct BM scaffold even though

such modifications are commonly applied during lead optimization.[13] Moreover, the nature and properties of substituents attached to the scaffolds that often influence the SARs are not accounted for. Hence, an alternative representation is required that accounts for well-defined substructural relationships.

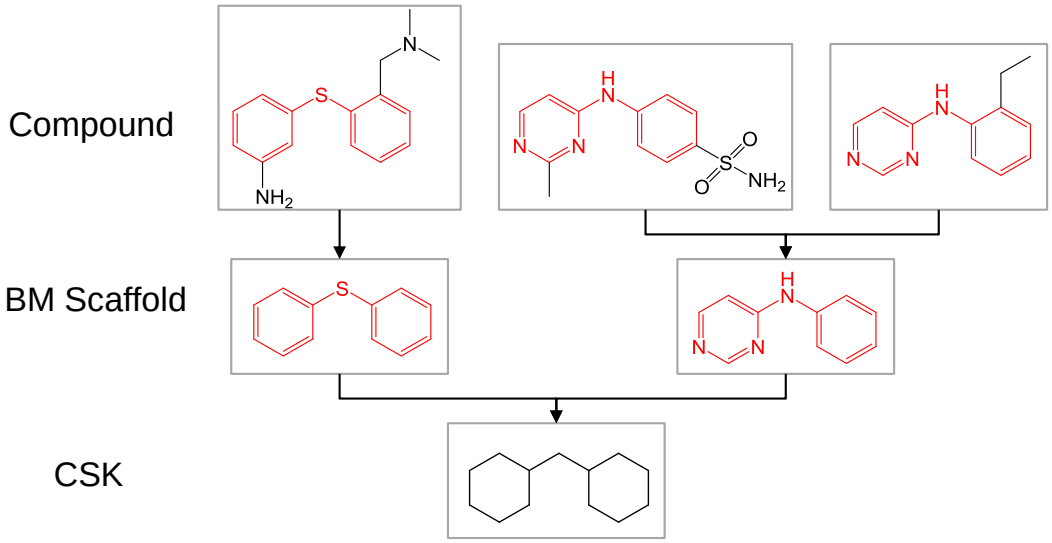


Figure 1.2: Molecular framework A schematic diagram of the hierarchical generation of Bemis and Murcko (BM) scaffolds and cyclic skeleton (CSK) from three compounds is shown. BM scaffolds (red) are generated by removing all side chains and retaining only the rings and linkers of the compounds. BM scaffolds are further converted to CSK by substituting all heteroatoms to carbon and setting bond orders to one.

Matched Molecular Pairs

Substructural relationships between pairs of compounds can be elegantly captured by the concept of *matched molecular pairs* (MMPs).[18] An MMP is a pair of compounds that share a large substructure and differ by a structural change (R-group) at a common site.[18] An exemplary MMP is given in Figure 1.3. The MMP formalism provides a consistent and well-defined framework to assess structural similarity. It helps in correlating structural changes to activity/property changes in a systematic manner as compared to FPs or BM scaffolds. The MMP concept has gained wide recognition in the medicinal chemistry field.[19]

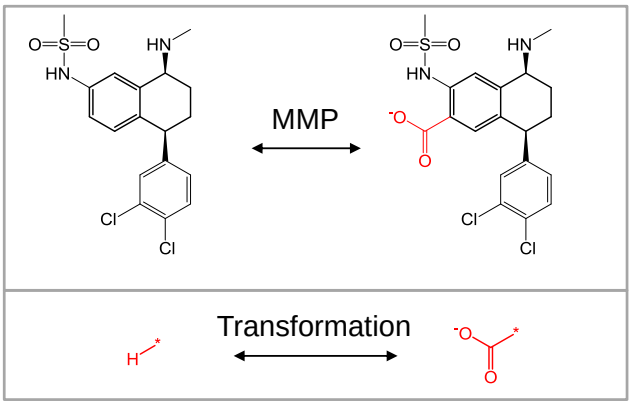


Figure 1.3: Matched molecular pair. A pair of compounds that forms a matched molecular pair (MMP) is shown. The exchanged substituent is highlighted in red and the corresponding transformation is depicted at the bottom.

Different algorithms that systematically extract MMPs from compound data sets are available. Some algorithms utilize direct graph comparison like maximum common substructure (MCS) search between pairs of molecules.[20] The MCS search is an NP-hard problem[21] and requires comparison of molecules in a pairwise manner.[22] Other algorithms involve fragmenting molecules into substructures on the basis of pre-defined rules.[23] The fragmentation step is complemented by subsequent indexing of the identified fragments. The fragmentation is carried out systematically on all single acyclic bonds present between two non-hydrogen atoms in a molecule. The resulting larger fragments are stored as keys and the remaining smaller fragments as values in the index table. If a key fragment already exists, the corresponding value fragment is added to the value list. Thus, the key fragment corresponds to the common substructure shared between the two molecules and the value fragments correspond to the exchange of a pair of substructures, termed chemical transformations,[23] as shown in Figure 1.3. The fragmentation approach is computationally more efficient for large-scale MMP extraction than MCS search. Furthermore, the MMP definition has also been extended to include chemical changes at more than one position by fragmenting molecules at multiple acyclic bonds (typically up to three).[19]

In order to assess compound pairs that are only distinguished by a functional group or a single ring system "transformation size-restricted MMPs" have been introduced.[24] Such MMPs are useful for correlating small structural modifications to activity/property changes.

A recent work has introduced the concept of "fuzzy matched pairs" (FMP)[25] that combines the classical MMP definition with a pharmacophore description. This enables the analysis of compound pairs with transformations that are structurally distinct but share a pharmacophore.

The methods described in this section are different ways to represent molecules and assess their similarity. Each method has its own advantages and limitations. The exploration of SARs is affected by the choice of the representation and the similarity metric. Other factors, such as the origin, composition and size of the compound data set under investigation also affect the analysis of SARs. These factors need to be considered when choosing the method for the analysis of SARs.

SAR Analysis Methods

Current computational approaches to study SARs are multifaceted and of different methodological complexity. In general, the methodologies could be classified as *descriptive* or *predictive*. Descriptive approaches mine the SAR information from the data and then represent it numerically or graphically. The represented SARs can then be analyzed by medicinal chemists. Predictive approaches extract generalized SAR patterns from the reference compounds to predict biological activities of new compounds.[4]

The field referred to as quantitative SAR (QSAR) analysis, was first developed by Hansch et al.[26] and has been invaluable for understanding SARs. In QSAR, a mathematical model is derived that relates structural features and/or molecular properties to bioactivity. QSAR models are built from a set of compounds with known biological activity. These models can be applied to predict

activities of candidate molecules with a structural/chemical composition similar to that of the reference compounds. Candidate compounds that are not reasonably similar to some reference compounds fall outside the applicability domain of the model and their activity cannot be reliably predicted.[27]

Over the years, QSAR modeling has evolved from applications using relatively simple linear regression methods to more complex non-linear machine learning techniques.[28] However, even in the presence of similar compounds these methods fail to reliably predict activities of the candidate compounds in many cases.[29] Outliers result not only from statistical fluctuations or measurement errors but also from the limitation on the part of the SPP principle underlying these approaches. SPP is intuitive and a central paradigm in medicinal chemistry, however, it is frequently observed that small modifications in chemical structures can lead to dramatic changes in compound activity.[29] Pairs of compounds that show high structural similarity and significant difference in activity are called *activity cliffs*[29] and represent exceptions to the SPP principle.

These observations suggest that there are fundamental differences in the nature of SARs. To deconvolute the complex SAR patterns in the data, descriptive approaches have been used. These methods guide compound design in hit-to-lead and lead optimization campaigns by enabling the user to understand on a case-by-case basis the structural features that determine activity. A conventional data structure called R-group table that displays the substituents of individual compounds and their corresponding compound activity is useful to study the effect of small structural changes on compound potency. However, R-group tables are applied to analogs that share the same core structure and are not suitable to analyze large compound sets. Therefore, tools that can be applied on large and structurally heterogeneous compound data sets are indispensable.

Activity Landscapes

The descriptive approaches for SAR analysis include various *data mining* and *visualization* methods to systematically analyze SARs on a large-scale and ex-

tract available SAR information from compound data sets of different sizes and origins.[30] The combination of these methods provides a basis for the exploration of SARs.

The activity landscape concept is an approach that has become popular.[4,30] An activity landscape can be defined as any graphical representation that integrates similarity and potency relationships between compounds having a specific biological activity.[4] It enables the systematic comparison of compound structures and their potencies.

The Nature of SARs

The different natures of SARs can be observed in 3D activity landscapes. A 3D activity landscape is generated by adding activity as the third dimension to a 2D chemical reference space of a set of compounds.[31] In the 2D chemical space, inter-compound distances reflect structural (dis)similarity. Thus, compounds that are close in the 2D space are chemically more similar than compounds that are farther apart. The third dimension, activity, provides information about the distribution of the compounds' potency values. Compounds with large or moderate differences in their potency value can be clearly observed. The activity landscape view resembles geographical landscapes, and contains similar features, e.g. plains, mountains and valleys.

In 3D representations, gently sloped areas, as shown in Figure 1.4a, represent regions of SAR continuity where gradual changes in chemical structure are accompanied by small or moderate changes in potency.[2,4] By contrast, rugged areas, as shown in Figure 1.4b, represent regions of local SAR discontinuity where small modifications in chemical structures lead to large changes in potency.[2,4] In these regions high peaks correspond to activity cliffs. Activity cliffs represent a prominent form of SAR discontinuity and are highly informative. In most cases, a compound data set is represented as a "variable activity landscape"[32] that is a combination of continuous and discontinuous SAR components, as shown in Figure 1.4c. Such variable activity landscapes correspond to the presence of SAR heterogeneity.[4,32]

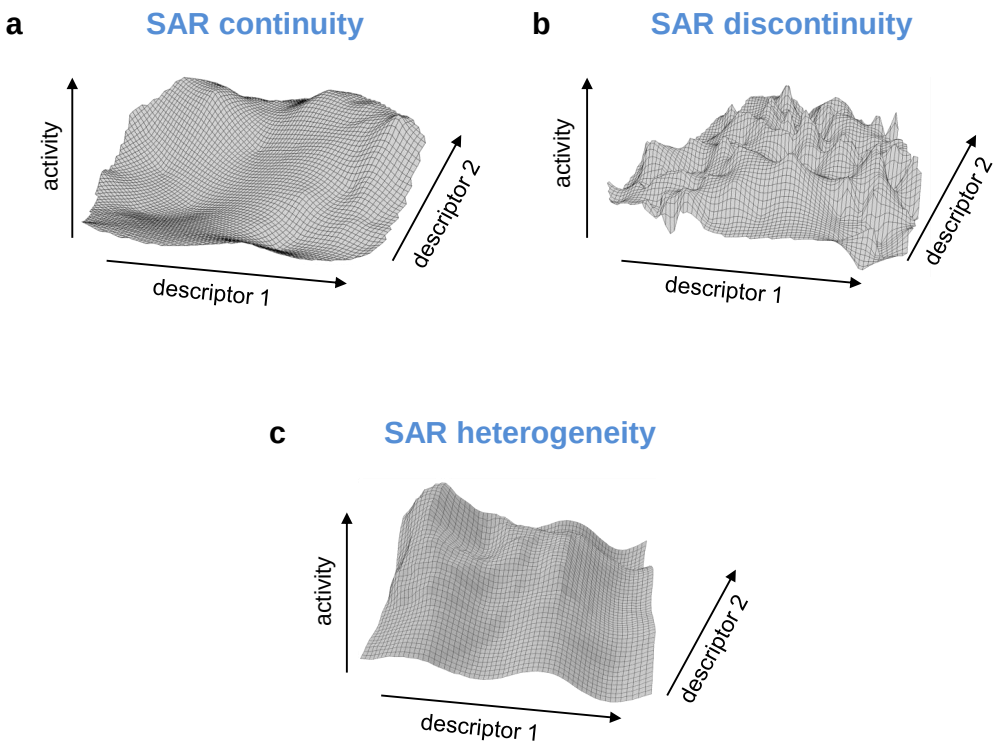


Figure 1.4: SAR characters. Shown are model 3D activity landscapes depicting (**a**. continuous, **b**. discontinuous, **c**. heterogeneous) SAR characters, respectively. For landscape generation, compounds are projected onto a 2D chemical reference space and activity is added as the third dimension. Taken from [4].

The continuous SAR character is a prerequisite for virtual screening or linear QSAR applications. The discontinuous SARs, especially the activity cliffs, are exploited in lead optimization campaigns, in order to improve compound activity.[4,30] Thus, the systematic description of the different SAR characteristics, namely continuous, discontinuous and heterogeneous, helps to choose the relevant application for analysis and/or prediction.

Numerical SAR Analysis

Complementing the activity landscape analysis, numerical functions that quantify different SAR characteristics have also been developed.[33,34] These functions are based on pairwise calculations of structure and activity similarity for data

set compounds. The SAR index (SARI)[33] is a combination of the SAR continuity and SAR discontinuity scores. The SAR continuity and discontinuity scores quantify the continuous and discontinuous characters in compound data sets, respectively, by taking the potency difference and similarity between compound pairs into account. The SARI score is normalized and ranges from 0 to 1. Low, intermediate and high scores correspond to discontinuous, heterogeneous and continuous SAR characters, respectively.

The discontinuity score component of the SARI formalism can be used to interpret the different SAR characteristics at a global level, i.e., for activity classes and at a more local level, i.e., for a cluster of compounds within an activity class.[35] Furthermore, a local discontinuity score can also be calculated to assess individual compound contributions to SAR discontinuity.[35]

Another numerical score reported by Guha et al.[34] called the structure-activity landscape index (SALI) quantifies pairs of compounds based on their differences in activity divided by their distances in chemical space. It emphasizes pairs of structurally similar compounds with large potency differences and is designed to detect activity cliffs in a data set.

Thus, numerical scores can be used to quantify and diagnose the different SAR characters for compound data sets. These functions often complement the landscape based SAR analysis. As graphical representations, the activity landscape models provide intuitive access to the SAR information of compound data sets. However, with steadily growing numbers of active compounds, the activity landscapes become increasingly complex.[36] This requires the design of other novel graphical schemes to effectively extract SAR information. Many different types of graphical schemes have been designed to assist in SAR analysis.

Graphical SAR Analysis

Molecular network representations have become increasingly popular for the visualization of SAR characteristics of compound data sets. The structure-activity similarity (SAS) maps[37] are one of the earliest graph-based activity landscape representations. In SAS maps, pairwise structural and activity similarity is plotted along an xy-plane, such that each data point represents a

pairwise compound comparison. Usually FPs are used as molecular representations and the similarity is accounted for by the Tc metric. Activity similarity is represented as the logarithmic potency difference. Thus, a large difference corresponds to low activity similarity and a small difference to high activity similarity.

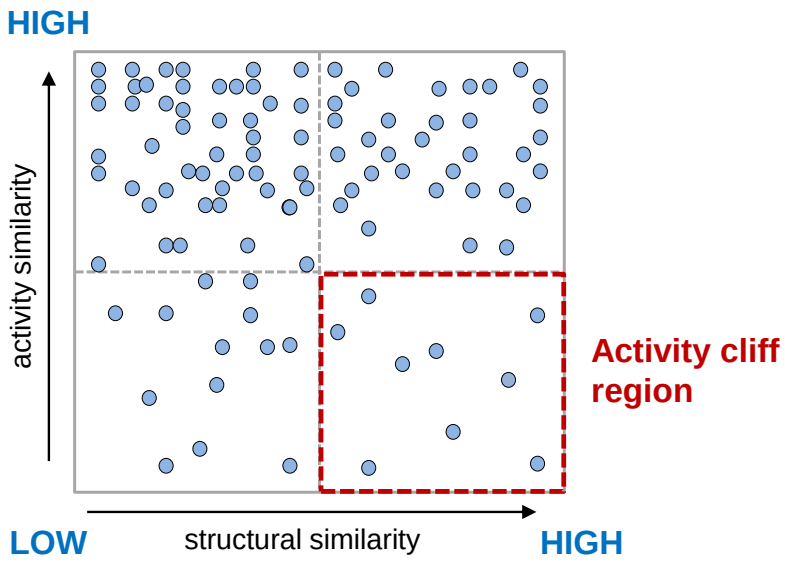


Figure 1.5: Structure-activity similarity maps. A schematic representation of an SAS map is shown that depicts the structural and activity similarity for all compound pairs within a data set in a scatter plot. Each compound pair is mapped to one of the four regions. The activity cliff forming region can be identified at the bottom right section of the SAS map. Adapted from [4].

The SAS map can be subdivided into four sections that capture different SAR characteristics. A schematic illustration of the SAS map is presented in Figure 1.5. The upper-left section contains pairs of compounds with high activity and low structural similarity. This region can aid in the identification of new active scaffolds with similar activity. The upper-right region contains compound pairs with high structural and activity similarity, corresponding to analogs with comparable potency. The lower-left section contains compound pairs with low structural and activity similarity and does not contain any desirable trait for further analysis. By contrast, compound pairs falling into the

lower-right section have high structural and low activity similarity and represent the activity cliff region of an SAS map. These are highly informative for further analysis.

More advanced molecular network representations such as network-like similarity graphs (NSGs)[35] help elucidate local SAR features in relation to the global SAR features of the compound data. Here compounds are represented as individual nodes. Edges are drawn between nodes to account for structural similarity, if the compound pairs exceed a certain predefined Tc threshold. Nodes are color-coded according to compound activity. A continuous color spectrum is applied ranging from green (minimal potency in the data set) over yellow (medium potency) to red (maximal potency). Nodes are also scaled in size by the local per-compound discontinuity scores. Furthermore, cluster discontinuity scores are calculated to characterize the local SARs. A schematic representation of an NSG is shown in Figure 1.6.

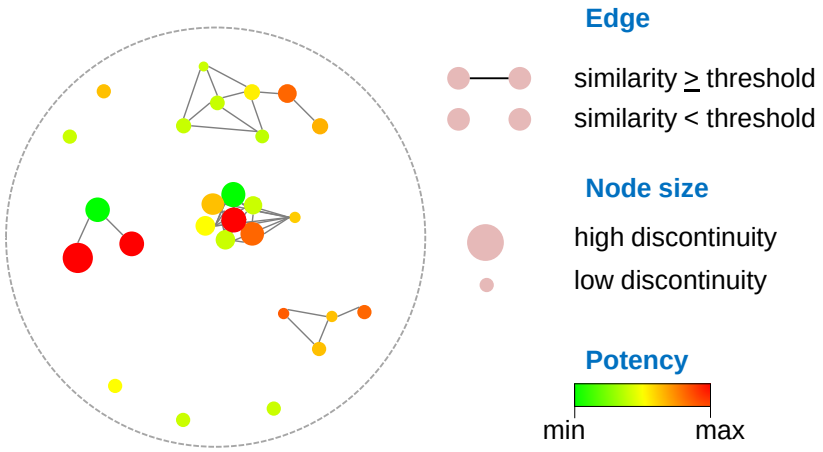


Figure 1.6: Network-like similarity graphs. A schematic illustration of an NSG is shown. Nodes correspond to compounds and edges between nodes represent compound pairs that show structural similarity (Tc) greater than the defined threshold. The node sizes are scaled according to the compound discontinuity scores. The node colors reflect compound potencies as indicated by the color bar at the bottom. Adapted from [4].

The node positions and edge lengths in NSGs are determined by a force-directed graph layout algorithm[38] that separates densely connected clusters from each other. Thus, inter-cluster distances have no chemical meaning. The clusters help to identify the most interesting local SAR regions. For example, clusters of similarly colored and sized nodes highlight regions that are continuous in nature. By contrast, clusters that show different colors and sizes indicate the presence of local SAR discontinuity. Large red and green nodes connected by edges indicate activity cliffs in the compound set.

NSGs have been primarily designed for the analyzes of lead optimization sets, yet they have been proven to be equally intuitive and applicable for the analysis of large screening sets.[39]

NSGs use whole-molecule similarity measures, which leads to additional effort in interpreting structural changes. Representations that capture direct substructure-based relationships overcome this limitation. These approaches directly relate structural fragments with activity information. Substructure-based relationships are captured by molecular scaffolds and MMPs.

A representation called the scaffold tree[40] involves the systematic extraction of molecular building blocks from sets of bioactive compounds. This is achieved by first pruning all side chains and then subsequently removing rings from molecular structures according to predefined chemical rules.[40] This process is carried out until a single ring structure remains. Each generated substructure is organized hierarchically and is annotated with the activity information of the compounds in which it is contained. A schematic illustration of the scaffold tree is provided in Figure 1.7.

Given the rule-based decomposition of ring systems, the scaffold tree hierarchy contains scaffolds that are not contained in the data set compounds. These virtual scaffolds can be utilized in compound design efforts. In a prospective application of the scaffold tree data structure, novel (two- to four-ring) scaffolds for the enzyme 5-lipoxygenase and the nuclear receptor ERα have been designed.[41]

MMP relationships have also been utilized to organize compounds and represent them graphically. One of the first representations that used MMPs was the bipartite matching molecular series graph (BMMSG).[42] BMMSGs not only

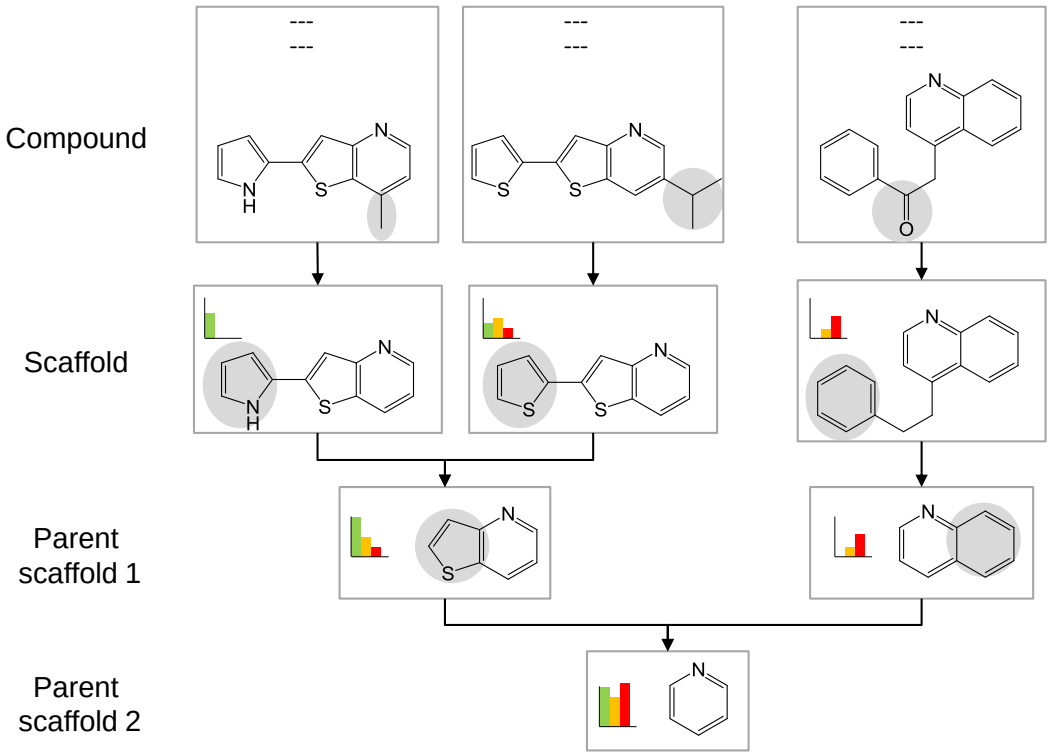


Figure 1.7: Scaffold tree. A scaffold tree representation depicting the hierarchical fragmentation of model compounds is shown. The hypothetical activity distribution of compounds represented by the corresponding scaffold is reported in bar charts. Substructures removed in subsequent steps are highlighted in gray. Adapted from [30].

reveal global SAR trends in compound data set but also show local SAR patterns in matching molecular series (MMS).[42] An MMS constitutes of a set of compounds that share the same core fragment and differ by substitutions at a single site.

The concept of structurally analogous matching molecular series (A_MMS) was formulated on the basis of MMS.[43] A_MMS refers to multiple series with structurally similar cores and overlapping substitution patterns. The SAR matrix[43] data structure organizes compound series as A_MMS, such that the rows represent structurally related cores and columns correspond to different substituents. Each cell represents a compound that is a combination of a core and a substituent. The SARM is designed to systematically elucidate SAR patterns in analog series. Furthermore, core-substituent combinations that do not repre-

sent any data set compounds might arise, thereby extending the chemical space to previously unexplored compounds. These "virtual compounds" are potential candidates for further exploration. Therefore, the SARM data structure provides a link between descriptive SAR analysis and prospective compound design.

Predictive Approaches

Activity landscapes are used to analyze SAR data sets for which activity values have already been obtained from experiments. The numerical and graphical SAR analysis schemes, described so far, characterize SAR patterns in a data set but do not directly help in the activity prediction of novel compounds. The activity landscape concept could be used to predict not just the activity of novel compounds but also their local SAR environment, especially if they are involved in the formation of activity cliffs.

Activity cliffs represent the extreme form of SAR discontinuity and traditional QSAR methods are unlikely to predict very different activities for two structurally similar molecules. So far, the activity cliff analysis has been descriptive in nature. Applications have attempted to mine and analyze activity cliffs from public databases[24,44] or directly identify structural modifications that lead to their formation.[45] Recently, studies have begun to directly predict whether novel molecules would form activity cliffs using the activity landscape paradigm. One study attempted to identify activity cliffs by predicting SALI values for pairs of molecules using random forest[46] models.[47] Predicted SALI value for a novel compound is an indicator of its ability to form an activity cliff when paired with other molecules in the data set. Another study utilized the MMP representation to classify molecule pairs as activity cliff forming and activity cliff non-forming using support vector machine (SVM)[48] approach.[49] The study attempted to identify structural features among compounds sharing a specific activity that are responsible for high and low potency and thus, ultimately, for the formation of activity cliffs. Another study utilized the emerging chemical patterns (ECP)[50] approach to identify distinguishable structural and potency

characteristics from compounds forming activity cliffs.[51] These patterns were used for the prediction of unknown activity cliff forming compounds.

Prediction of a novel compound's local SAR environment, using the ECP method, has also been attempted.[52] Here, per-compound SARI discontinuity scores were calculated and patterns that distinguished compounds with low, intermediate and high discontinuity scores were employed for classifying compounds that mapped to low, intermediate and highly discontinuous SAR regions, respectively.

These methods have attempted to utilize SAR characteristics derived from activity landscape models for prediction purpose. Thereby the role of activity landscape models was extended from descriptive to predictive applications. The predictive approaches complementing descriptive activity landscape methods can help in prospective compound design.

Multi-Target Activity Spaces

Currently it is widely recognized that many pharmaceutically relevant compounds and drugs elicit therapeutic effects by interacting with multiple targets.[53,54] The presence of specific interactions of a compound with multiple targets is referred to as compound promiscuity and provides the basis for polypharmacological effects.[55,56] The analysis of compound promiscuity is useful for chemogenomics applications.

Public compound repositories[57] represent the largest source of publicly available chemogenomics data. The degree of compound promiscuity observed is dependent on the type of activity measurements that are considered.[58] A study analyzing the growth of compound activity data in the ChEMBL[59] repository found out that compound promiscuity rates involving distantly related or unrelated targets increase when assay-dependent (IC_{50}) measurements are utilized.[58]

Systematic data mining efforts and compound and/or target network representations have been deployed to understand compound promiscuity in different contexts.[56] Systematic analyses at the level of molecular scaffolds have identified scaffolds that are selective for closely related targets and scaffolds that are

promiscuously active across multiple target families.[60,61] Network representations were utilized to study scaffold-target family relationships of promiscuous scaffolds. A bipartite network with different types of nodes representing scaffolds and target families was generated. Edges were drawn between the two node types if the scaffold was active against the target family. The network helped in the identification of promiscuity patterns among topologically equivalent scaffolds active against different target families.[61]

In another study, MMP formalism has been utilized to explore structure-promiscuity relationships in a compound profiling data set using 100 sequentially-unrelated proteins.[62,63] 126 compound pairs where small structural modifications led to large-magnitude change in promiscuity, i.e., promiscuity cliffs were detected. A network representation, as shown in Figure 1.8, was utilized to indicate compound pairs that formed promiscuity cliffs. Here, each node represents a compound and is color-coded according to the number of target annotations by applying a continuous color spectrum from black for inactive compounds to white for highly promiscuous compounds. Nodes are connected by edges if the compounds form promiscuity cliffs. Two representative promiscuity cliffs are also shown in Figure 1.8. The presence of promiscuity cliffs suggested that promiscuity is not an inherent feature of molecular scaffolds but can be induced by small chemical substitutions.[63]

Data mining and especially visualization tasks are complicated for chemogenomics data given their multi-target nature and the substantially varying degrees of compound promiscuity.[65] No single data mining effort or network representation is able to capture its complexity. Hence, a combination of novel intuitive mining and graphical methods need to be deployed for the same.[65]

Thesis Outline

The primary objective of this dissertation is to introduce novel methods that facilitate SAR exploration and activity predictions using the activity landscape framework. The studies guide compound design efforts in pharmaceutical research.

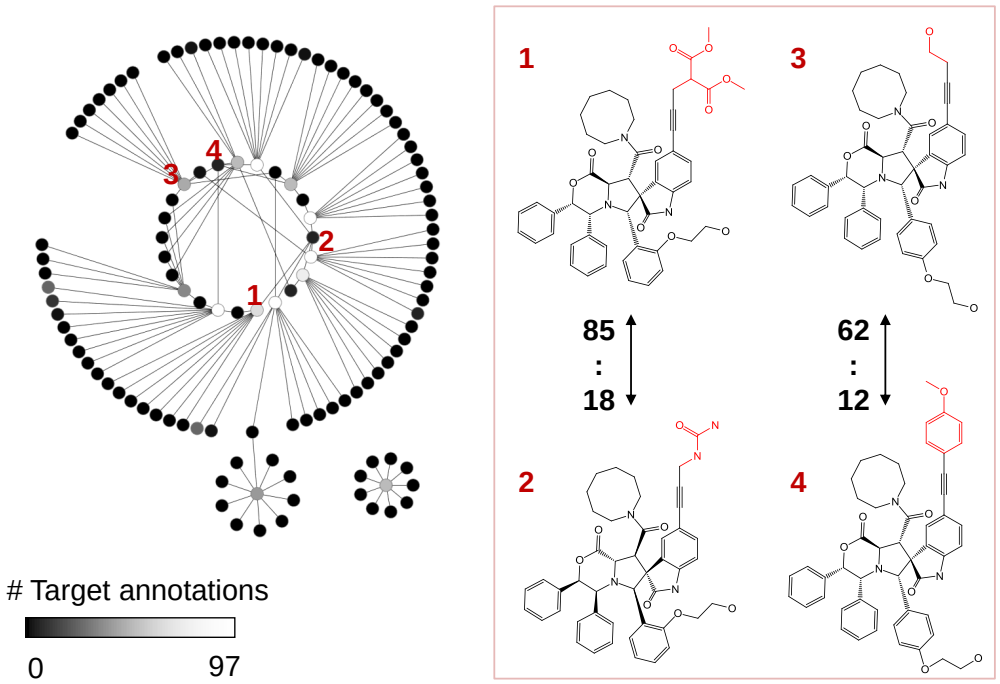


Figure 1.8: Network representation for promiscuity cliffs. 126 promiscuity cliffs are organized in a network representation (left). Nodes represent compounds and edges indicate promiscuity cliffs. Nodes are color-coded according to the number of target annotations using a continuous color spectrum from black (0 targets; inactive) to white (97 targets; most promiscuous). Two representative promiscuity cliffs are shown on the right. The number of targets each compound is active against is reported. Substructural changes are highlighted in red. Taken from [64].

This dissertation consists of six studies that are organized as individual chapters:

- In Chapter 2, a newly designed landscape model that utilizes the canonical scaffold and skeleton representation to organize compound sets in a consistent and hierarchical manner is reported. SAR information can be intuitively extracted from the model. Exemplary analyses of different compound data sets reveal how global and local SAR patterns can be identified.

- The SAR matrix (SARM) methodology based on the MMP formalism is designed to systematically extract structurally related compound series

from compound data sets and organize these in matrices.[4] In Chapter 3, methodological extensions have been introduced for the SARM method. These second generation SARMs are useful for applications in medicinal chemistry and chemogenomics. This study summarizes the methodological advancements and Chapters 4 and 5 report them in detail.

- In Chapter 4, a computational and graphical framework based on the SAR matrix method for the analysis of multi-target activity spaces and compound promiscuity patterns is introduced.

- The virtual compounds emerging in the SAR matrix data structure are potential candidates for further exploration. In Chapter 5, a novel QSAR-based approach utilizing local chemical neighborhood information for virtual compound activity prediction from SAR matrices is reported.

- The prediction method described in Chapter 5 can only be utilized for hit-to-lead or lead optimization sets, where explicit potency values are available. Therefore, a conditional probability-based prediction approach using SAR matrices has been developed and evaluated in Chapter 6. This method is applicable for screening sets and is useful for hit expansion.

- Chapter 7 reports the results of the first prospective application of the SARM-derived probability approach described in Chapter 6.

At the end, the key aspects and results of the work presented in this dissertation are summarized in Chapter 8.

References

[1] Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspectives in Drug Discovery and Design* **1998**, *9–11*, 225–232.

[2] Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Medicinal Chemistry* **2009**, *1*, 451–466.

[3] In *Concepts and Applications of Molecular Similarity*, Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

[4] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2010**, *53*, 8209–8223.

[5] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

[6] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *Journal of Chemical Information and Modeling* **2010**, *50*, 205–216.

[7] Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Combinatorial Chemistry and High Throughput Screening* **2000**, *3*, 363–372.

[8] Heikamp, K.; Bajorath, J. Fingerprint Design and Engineering Strategies: Rationalizing and Improving Similarity Search Performance. *Future Medicinal Chemistry* **2012**, *4*, 1945–1959.

[9] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro,CA 94577.

[10] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[11] Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.

[12] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2014**, *57*, 3186–3204.

[13] Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 1742–1753.

[14] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.

[15] Xu, Y.-j.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 181–185.

[16] Hu, Y.; Bajorath, J. Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs. *Journal of Chemical Information and Modeling* **2010**, *50*, 500–510.

[17] Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 1848–1856.

[18] Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*, Oprea, T. I., Ed.; Wiley-VCH: 2005, pp 271–285.

[19] Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Development Research* **2012**, *73*, 518–527.

[20] Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *Journal of Chemical Information and Modeling* **2009**, *49*, 1952–1962.

[21] Garey, M. R.; Johnson, D. S. In *Computers and Intractability: A Guide to the Theory of NP-Completeness*; W.H. Freeman and Company: New York, U.S.A, 1979.

[22] Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 521–533.

[23] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *Journal of Chemical Information and Modeling* **2010**, *50*, 339–348.

[24] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *Journal of Chemical Information and Modeling* **2012**, *52*, 1138–1145.

[25] Geppert, T.; Beck, B. Fuzzy Matched Pairs: A Means to Determine the Pharmacophore Impact on Molecular Interaction. *Journal of Chemical Information and Modeling* **2014**, *54*, 1093–1102.

[26] Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180.

[27] Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *Journal of Chemical Information and Modeling* **2005**, *45*, 839–849.

[28] Cherkasov, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014**, *57*, 4977–5010.

[29] Maggiora, G. M. On Outliers and Activity Cliffs why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535–1535.

[30] Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.

[31] Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W.; Oprea, T. I. A Practical Strategy for Directed Compound Acquisition. In *Chemoinformatics in Drug Discovery*, Oprea, T. I., Ed.; Wiley-VCH: 2005, pp 317–332.

[32] Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chemistry and Biology* **2007**, *14*, 489–497.

[33] Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *Journal of Medicinal Chemistry* **2007**, *50*, 5571–5578.

[34] Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *Journal of Chemical Information and Modeling* **2008**, *48*, 646–658.

[35] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-Like Similarity Graphs and Local Structure-Activity Relationship Indices. *Journal of Medicinal Chemistry* **2008**, *51*, 6075–6084.

[36] Hu, Y.; Bajorath, J. Learning from 'Big Data': Compounds and Targets. *Drug Discovery Today* **2014**, *19*, 357–360.

[37] Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26–30, 2001*; American Chemical Society: Washington, DC, 2001, abstract no. 77.

[38] Fruchterman, T. M.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Software: Practice and Experience* **1991**, *21*, 1129–1164.

[39] Wawer, M.; Bajorath, J. Extracting SAR Information From a Large Collection of Anti-Malarial Screening Hits by NSG-SPT Analysis. *ACS Medicinal Chemistry Letters* **2011**, *2*, 201–206.

[40] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree–Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *Journal of Chemical Information and Modeling* **2007**, *47*, 47–58.

[41] Renner, S.; van Otterlo, W. A.; Dominguez Seoane, M.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-Guided Mapping and Navigation of Chemical Space. *Nature Chemical Biology* **2009**, *5*, 585–592.

[42] Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *Journal of Medicinal Chemistry* **2011**, *54*, 2944–2951.

[43] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *Journal of Chemical Information and Modeling* **2012**, *52*, 1769–1776.

[44] Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single-and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chemical Biology & Drug Design* **2011**, *78*, 224–228.

[45] Wassermann, A. M.; Bajorath, J. Chemical Substitutions That Introduce Activity Cliffs Across Different Compound. *Journal of Chemical Information and Modeling* **2010**, *50*, 1248–1256.

[46] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

[47] Guha, R. Exploring Uncharted Territories - Predicting Activty Cliffs in Structure-Activity Landscapes. *Journal of Chemical Information and Modeling* **2012**, *52*, 2181–2191.

[48] Vapnik, V. N. In *The Nature of Statistical Learning Theory*; Springer: New York, U.S.A, 2000.

[49] Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *Journal of Chemical Information and Modeling* **2012**, *54*, 1301–1210.

[50] Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *Journal of Chemical Information and Modeling* **2006**, *46*, 2502–2514.

[51] Namasivayam, V.; Iyer, P.; Bajorath, J. Prediction of Individual Compounds Forming Activity Cliffs Using Emerging Chemical Patterns. *Journal of Chemical Information and Modeling* **2013**, *53*, 3131–3139.

[52] Namasivayam, V.; Gupta-Ostermann, D.; Balfer, J.; Heikamp, K.; Bajorath, J. Prediction of Compounds in Different Local Structure-Activity Relationship Environments Using Emerging Chemical Patterns. *Journal of Chemical Information and Modeling* **2014**, *54*, 1301–1310.

[53] Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nature Biotechnology* **2006**, *24*, 805–815.

[54] Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nature Reviews. Cancer* **2010**, *10*, 130–137.

[55] Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *Journal of Medicinal Chemistry* **2014**, *57*, 7874–7887.

[56] Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discovery Today* **2013**, *18*, 644–650.

[57] Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 6987–7002.

[58] Hu, Y.; Bajorath, J. Growth of Ligand? Target Interaction Data in ChEMBL is Associated with Increasing and Activity Measurement-Dependent Compound Promiscuity. *Journal of Chemical Information and Modeling* **2012**, *52*, 2550–2558.

[59] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.

[60] Hu, Y.; Wassermann, A. M.; Lounkine, E.; Bajorath, J. Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds across Druggable Target Families. *Journal of Medicinal Chemistry* **2010**, *53*, 752–758.

[61] Hu, Y.; Bajorath, J. Polypharmacology-Directed Compound Data Mining: Identification of Promiscuous Chemotypes with Different Activity Profiles and Comparison to Approved Drugs. *Journal of Chemical Information and Modeling* **2010**, *50*, 2112–2118.

[62] Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins have Distinct Distributions of Structural Complexity that Correlate with Protein-Binding Profiles. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 18787–18792.

[63] Dimova, D.; Hu, Y.; Bajorath, J. Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity. *Journal of Medicinal Chemistry* **2012**, *55*, 10220–10228.

[64] Hu, Y.; Gupta-Ostermann, D.; Bajorath, J. Exploring Compound Promiscuity Patterns and Multi-Target Activity Spaces. *Computational and Structural Biotechnology Journal* **2014**, *9*, e201401003.

[65] Hu, Y.; Stumpfe, D.; Bajorath, J. Visualization of Activity Landscapes and Chemogenomics Data. *Molecular Informatics* **2013**, *32*, 954–963.

Chapter 2

Introducing the LASSO Graph for Compound Data Set Representation and Structure-Activity Relationship Analysis

Introduction

Many different activity landscape representations that facilitate the understanding of SAR characteristics of bioactive compounds have been developed.[1] Critical parameters in the design of activity landscape models are the choice of a molecular representation and a similarity metric. Graphical representations using different parameters offer opportunities to identify distinct SAR patterns in compound data sets.

In this study, a novel activity landscape representation is introduced in which compounds are organized following a canonical scaffold-skeleton structural hierarchy. Hierarchical and substructural relationships between compounds are captured in a novel graphical design. Compound activity data is

integrated to reveal informative SAR patterns. Exemplary applications demonstrate the prominent SAR features of this novel representation.

Journal of Medicinal Chemistry

Article

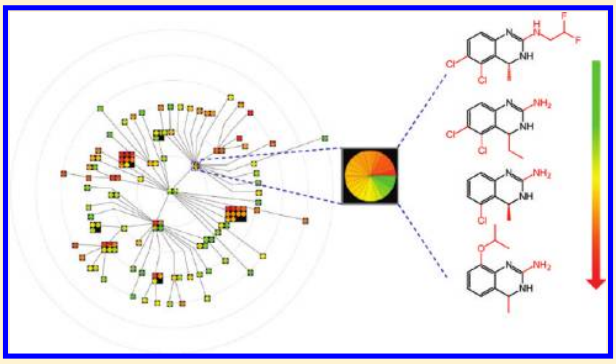
pubs.acs.org/jmc

Introducing the LASSO Graph for Compound Data Set Representation and Structure–Activity Relationship Analysis

Disha Gupta-Ostermann,[†] Ye Hu,[†] and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT: A graphical method is introduced for compound data mining and structure–activity relationship (SAR) data analysis that is based upon a canonical structural organization scheme and captures a compound–scaffold–skeleton hierarchy. The graph representation has a constant layout, integrates compound activity data, and provides direct access to SAR information. Characteristic SAR patterns that emerge from the graph are easily identified. The molecular hierarchy enables "forward–backward" analysis of compound data and reveals both global and local SAR patterns. For example, in heterogeneous data sets, compound series are immediately identified that convey interpretable SAR information in isolation or in the structural context of related series, which often define SAR pathways through data sets.



INTRODUCTION

For the extraction of SAR information from large compound data sets, visualization techniques that view SAR features from different angles have become increasingly popular in recent years.[1,2] For instance, graphical methods have been introduced to globally represent data sets[3−6] or generate compound-centric[7,8] and series-centric views.[9−11] Global graphical analysis approaches include molecular-network-type representations[3,4] or diagrams that compare molecular similarity and activity similarity of compounds in a pairwise manner.[5,6] In these plots, molecular similarity is generally assessed by calculating Tanimoto similarity of test compounds using various descriptors, in particular, fingerprints.[5,6] In SAR networks, similarity relationships (edges) might be established in an analogous manner[3] or by accounting for substructure relationships between active compounds.[4] In addition, local SAR representations might either monitor the structural neighborhood of active compounds[7,8] or concentrate on individual analogue series.[9−11] The latter methods include graphical extensions of conventional R-group tables[9] as well as network-like representations.[10,11] In such networks, analogues might be organized by substituent sites and site combinations[9] or on the basis of systematically determined substructure relationships.[10]

In addition to these global or local compound data set representations, molecular scaffolds originating from active compounds have also been graphically organized in different ways.[12] For example, for SAR monitoring, Scaffold Explorer[13] has been introduced, an interactive editor that links scaffold-like structures to an R-group table. Graphs containing these structures can be interactively built, modified, and annotated with SAR information. The tool is designed to aid medicinal chemists in processing R-group tables containing different core structures. Going beyond interactive analysis, a rule-based organization scheme for scaffolds is provided by the Scaffold Tree data structure.[14] Following this approach, scaffolds are decomposed along pathways by iteratively removing rings from them according to a set of predefined chemical preference rules until single-ring scaffolds remain. Given this rule-based decomposition scheme, scaffolds might be obtained along the tree that are not contained in the original data set compounds, which is a key feature of this approach. These "virtual" scaffolds can then be used for activity prediction, considering the activity of neighboring "real" scaffolds. Compound activity predictions on the basis of virtual scaffolds have been further exploited in an extension of the Scaffold Tree approach termed Scaffold Hunter.[15]

In principal, a scaffold-based representation of a compound data set can be further extended specifically for SAR analysis by following a hierarchical structural organization scheme from active compounds over conventional molecular scaffolds[16] to cyclic skeletons (CSKs),[17] which further abstract from scaffolds by omitting heteroatom and bond order information. This hierarchical organization scheme has previously been applied by us to systematically map target annotations to different compound classes.[12] A key aspect of this approach is that each CSK represents a family of topologically equivalent scaffolds. Hence, scaffolds can be organized according to their topology and further distinguished on the basis of chemical criteria.

In order to utilize this concept for SAR analysis, we have designed a canonical data structure that exploits structural

Received: April 5, 2012
Published: May 9, 2012

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

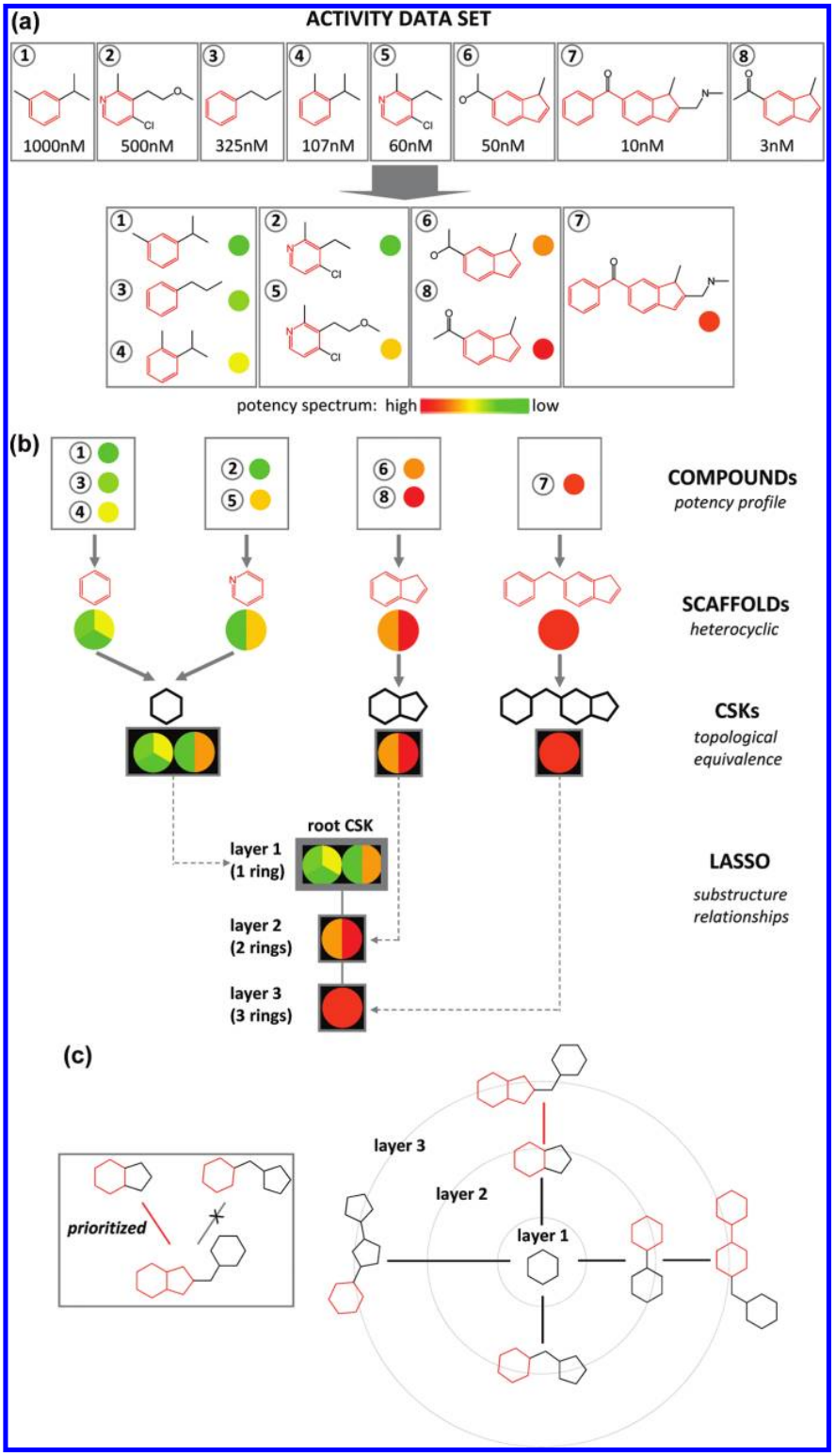


Figure 1. Graph generation. In (a) and (b), the generation of the LASSO graph is illustrated, as described in the text. In the exemplary data set, compounds are labeled with their potency (K_i) values. Scaffolds are colored red. In (c), substructure relationships between CSKs across different graph layers are depicted. In each pair of CSKs connected by an edge, the parental CSK is colored red. In addition, on the left, the prioritized assignment of CSK relationships is illustrated.

compound−scaffold−skeleton hierarchy in a "forward−backward" manner. This is accomplished by first extracting scaffolds and CSKs from active compounds and then organizing the data set in different layers defined by CSKs containing stepwise increasing numbers of rings. These layers capture the associated scaffold and compound information in a graphically intuitive

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

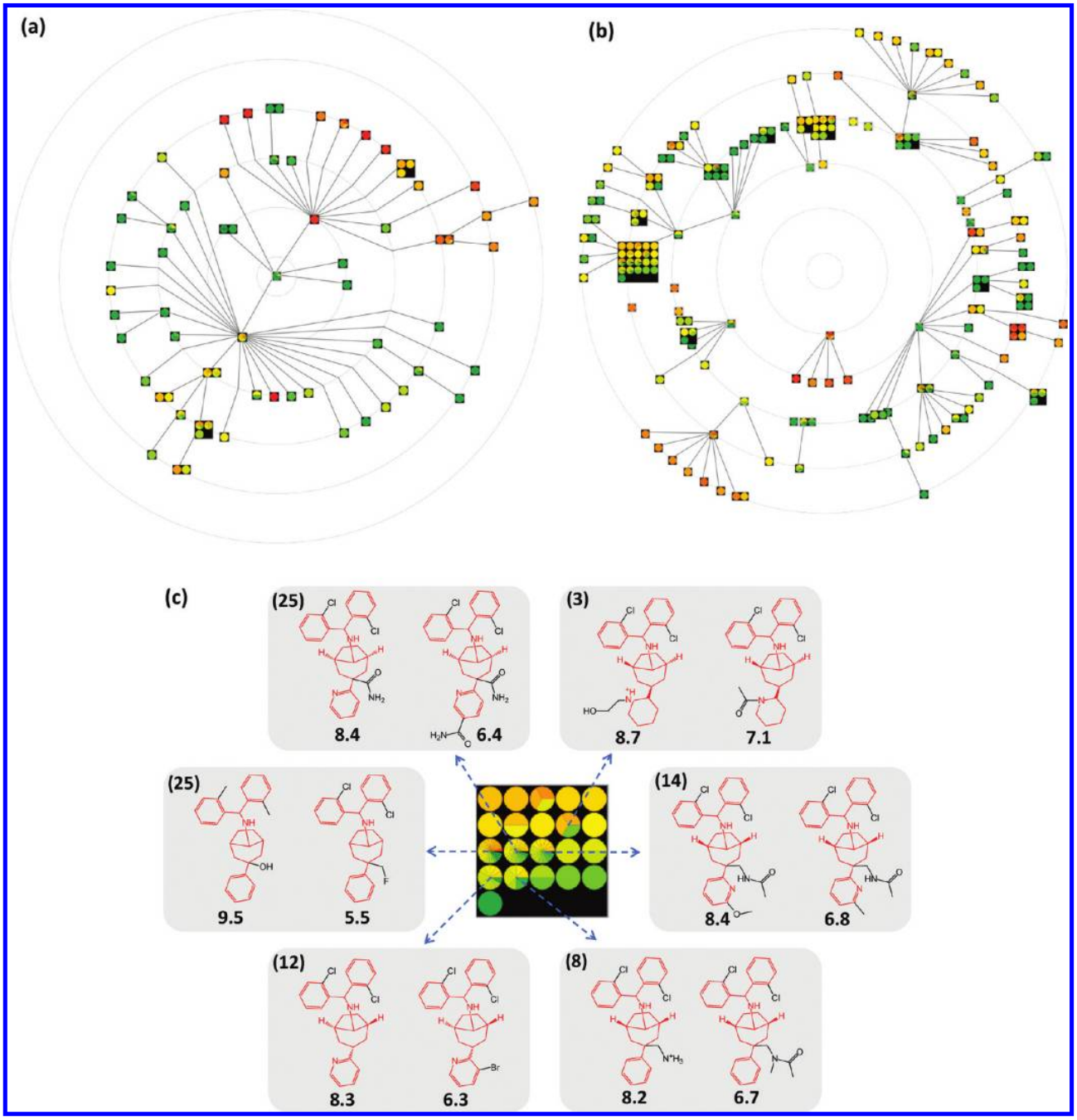


Figure 2. Graph representation. Prototypic LASSO graphs are shown. (a) Melatonin receptor 1A antagonists. (b) Nociceptin receptor antagonists. In (c), the rectangular subgraph on layer 5 on the left is enlarged and for each of six selected scaffolds (red) representing 3−25 compounds (reported in parentheses) the least and most potent analogues are shown. These topologically equivalent scaffolds represent analogue series with different potency progression.

manner. We term this data structure the "layered skeleton−scaffold organization" or LASSO graph (because its layout also reminds us of "roping" SAR information). On the basis of our evaluation, we find the LASSO graph structure to be very well suited for compound data set representation and the exploration of both global and local SAR features. For example, analogue series are immediately identified that convey SAR information in isolation or in the structural context of related series. Furthermore, structural pathways through data sets are obtained that also reveal SAR information. The design of the LASSO graph and exemplary applications are reported herein.

METHODS AND MATERIALS

Scaffold Generation. Scaffolds consisting of ring systems and linkers between them were obtained by removal of all R-groups from compounds following Bemis and Murcko scaffold definition.[16] However, in a departure from this conventional definition, exocyclic double bonds attached to ring atoms were not removed but retained. Hence, substituents with exocyclic

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

double bonds were not considered conventional R-groups. In addition, this modification led to the generation of further diversified scaffold sets. Scaffolds in LASSO graphs also contain stereocenter information. Scaffolds were further transformed into CSKs[17] by changing all heteroatoms to carbons and setting all bond orders to one. Importantly, scaffolds and CSKs are separately accounted for in LASSO graphs as a part of compound−scaffold−skeleton hierarchies.

Graph Design. The organization of the graph is based upon systematically derived substructure relationships between CSKs that are present in a data set. Scaffolds and compounds associated with each CSK are incorporated into the graph representation using different design elements, as discussed in the following. Figure 1a and Figure 1b illustrate the design elements of the graph.

Substructure Relationships. CSKs are organized by the number of rings they contain. Each number of rings (from 1 to *n*) corresponds to a separate layer in the graph. If a CSK contains a condensed ring system, each participating ring is considered a separate entity (and counted separately). Then a parent−child relationship is defined between two CSKs if a CSK at a given layer is completely contained in the structure of another CSK at a higher layer. This might be the next higher layer or a subsequent one, i.e., a parent−child relationship might involve CSKs that differ by more than one ring. Figure 1b illustrates these substructure relationship assignments. If a CSK has multiple possible parents, the relationship with a parent is prioritized that contains fewer linker atoms between rings, as illustrated in Figure 1c.

Layout. The resulting layers are captured in a hierarchical graph structure. A radial graph layout is used for visualization. Each level of the structural hierarchy is represented by a concentric circle onto which CSKs with the corresponding number of rings are placed as nodes. Therefore, the structural complexity of CSKs increases from the inner to outer layers. This layout enables the simultaneous presentation of multiple subgraphs with different roots.

Annotation and Visualization. In Figure 1a, a model compound set is shown and potency-based coloring is illustrated. The compound potency range within a data set is accounted for using a continuous color spectrum from green (lowest) to red (highest potency in the data set). In Figure 1b, CSKs are represented as rectangular nodes. Edges between two CSKs define a parent−child substructure relationship. In addition, individual scaffolds are depicted as circular nodes. Scaffolds represented by a given CSK are embedded within the rectangular CSK node. Thus, by definition, a CSK node must contain at least one scaffold node. Furthermore, scaffolds are color-coded based on the potency of the compounds they represent. If a scaffold represents multiple compounds, it is divided into an equally sliced pie chart where each slice represents an individual compound colored by its potency.

Implementation. All routines required to generate scaffolds and CSKs were implemented in Java using the OpenEye chemistry toolkit.[18] The graph structure was generated using the Java package JUNG.[19]

Data Sets. For graph evaluation and SAR analysis, different compound data sets were extracted from ChEMBL.[20] The data sets can be obtained via the following URL: http://www.limes.uni-bonn.de/forschung/abteilungen/Bajorath/labwebsite (please, see the "downloads" section).

RESULTS AND DISCUSSION

Characteristic Features of the LASSO Method. By design, LASSO is an SAR data mining method. As such, it does not directly provide suggestions for new analogues on the basis of graphical analysis; i.e., it is not a predictive approach. The methodology is devised to identify the most interesting compound subsets or series in large and structurally heterogeneous data sets, which is a particular strength of the underlying hierarchical molecular organization. Once compound series yielding interpretable SAR information have been extracted from such data sets, design of new compounds can be attempted in subsequent steps. Utilizing a hierarchical organization scheme for data mining and analysis has additional implications. For example, molecular hierarchies do not encode synthetic routes to generate compounds. However, they establish substructure and topological relationships between compound series that could not be established on the basis of synthetic criteria or by utilizing R-group tables or related representations. Compared to other hierarchical scaffold organizations such as scaffold trees, the most distinguishing features of the LASSO approach include the addition of topological relationships conveyed in the graphs and the "forward−backward" analysis capacity of scaffold and corresponding compound information, which provides an intuitive access to SAR patterns. An important feature of LASSO is that SAR information is represented in the form of compound−scaffold−skeleton sequences (rather than only using scaffolds), which reflects SAR information at different structural levels and enables a direct comparison of SAR patterns in different compound series.

Prototypic LASSO Graphs. In Figure 2, exemplary LASSO graphs are shown to illustrate general topological characteristics. The graph is arranged in concentric layers according to the presence of increasing numbers of rings in CSKs. Hence, layer 1 always corresponds to one or more CSKs containing a single ring, which might or might not be present in a given data set. If no single-ring CSK is available, layer 1 is empty. From layer to layer, the number of rings contained in CSKs increases exactly by 1. CSKs connected by edges form pathways across different layers, depending on their substructure relationships. The more substructure relationships are present, the more densely connected the graph will be. Depending on these relationships, CSK pathways might not involve each layer. In addition, multiple pathways might originate from the same or different layers (in this case, the JUNG implementation places pathways on layers by balancing radial distances between them). Importantly, any data set compound will appear in the graph, regardless of whether the corresponding CSK is involved in substructure relationships or not. The position of a compound in the graph is determined by the number of rings its CSK contains.

In Figure 2a and Figure 2b, LASSO graphs are shown for sets of antagonists of the melatonin receptor 1A and nociceptin receptor, respectively. Both graphs consist of six layers (i.e., CSKs contain a maximum of six rings), but their topology differs. The melatonin receptor 1A antagonist set in Figure 2a is structurally more homogeneous than the nociceptin receptor antagonist set in Figure 2b that yields a number of singletons and disjoint pathways. Furthermore, the graph of the nociceptin receptor antagonist set has no CSK at the first layer and contains only one CSK with two rings. In Figure 2c, the largest rectangular node of the nociceptin receptor antagonist graph is

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

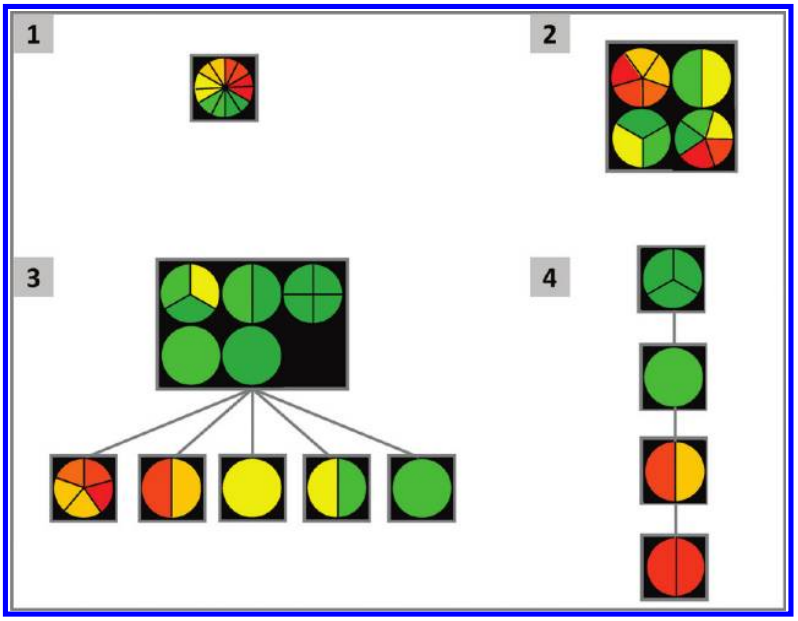


Figure 3. Graph patterns. Shown are four characteristic graph elements or patterns that convey SAR information, as described in the text.

displayed in detail. This subgraph contains analogue series represented by a total of 21 topologically equivalent scaffolds. Exemplary analogues are shown. As can be seen, these related analogue series display rather different potency progression. Hence, this node reveals a high degree of SAR heterogeneity and the compound series it contains are a primary focal point of SAR analysis.

SAR Patterns. As illustrated in Figure 1, compounds are consistently represented in the LASSO graph as a part of the skeleton−scaffold−compound hierarchy. This means that each compound is contained in a rectangular CSK node and a circular scaffold node. The presence of multiple compounds sharing the same scaffold gives rise to a color-coded pie chart representation of the scaffold node. By definition, these compounds form an analogue series. In the LASSO graph, characteristic patterns emerge that contain this basic design element and reflect available SAR information. These characteristic graph patterns are displayed in Figure 3. Pattern 1, the simplest one, represents a series of analogues with steady potency progression. Such a series typically contains interpretable SAR information. Pattern 2 mirrors the presence of topologically equivalent analogue series with varying potency distribution. In this case, SAR features can be compared across different yet related scaffolds and series. Pattern 3 is a characteristic horizontal pattern emerging from a LASSO graph. Here, individual compounds or series share a particular given CSK as the largest common substructure and are only distinguished by the presence and/or position of an individual ring. In this example, potency progression is observed from the right to the left. The potency distribution among such series might be indicative of preferred scaffolds. Furthermore, pattern 4 illustrates a characteristic vertical pattern, resulting from the stepwise addition of a single ring to a CSK. In this case, steady potency progression is also observed along the path. If horizontal or vertical patterns contain compounds or series with different potency distribution, they are arranged in the order of increasing potency, which aids in the identification of SAR-sensitive series and high-priority candidates for further exploration.

Graph Analysis. In the following, two examples are discussed to further illustrate the use of LASSO graphs for SAR exploration.

Serotonin 7 Receptor Antagonists. In Figure 4a, the LASSO graph of a set of 246 antagonists of the serotonin 7 receptor is shown. These compounds yield 119 distinct scaffolds and 76 CSKs. The LASSO graph is characterized by the presence of seven layers and densely connected pathways that originate from the same CSK containing a single ring, hence revealing many substructure relationships. In the graph, three characteristic patterns are labeled (according to the numbering scheme in Figure 3). In Figure 4b, structures forming patterns 1 and 2 are shown in detail. In this and all following illustrations of graph patterns, CSKs, corresponding scaffolds, and (representative) compounds comprising a pattern are shown. Pattern 1 in Figure 4b is formed by 24 analogues spanning a large potency range, which represents a typical SAR hotspot. In addition, pattern 2 is formed by the bipenyl scaffold and three closely related scaffolds. However, similar analogues represented by these scaffolds have significant differences in potency, maximally of more than 3 orders of magnitude. Thus, as revealed by the pattern, a substantial amount of SAR information is available for these biphenyl derivatives. In Figure 4c, the third pattern marked in Figure 4a is shown in detail, which represents a typical horizontal pattern. Five scaffolds and the exemplary compounds they represent are displayed (in several cases, the compounds already represent scaffolds). In this case, the pattern is formed by only a few analogues with steady potency progression of approximately 2 orders of magnitude.

Bradykinin B1 Receptor Antagonists. In Figure 5a, the LASSO graph of 348 antagonists of the bradykinin B1 receptor is shown that yield 132 scaffolds and 68 CSKs. The LASSO graph of this compound set also spans seven layers. Although the numbers of CSKs and scaffolds are comparable to the serotonin receptor antagonist example, in this case, the edge density in the LASSO graph is low and a number of singletons are observed, revealing the presence of only limited CSK substructure relationships. Nevertheless, the graph also contains

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

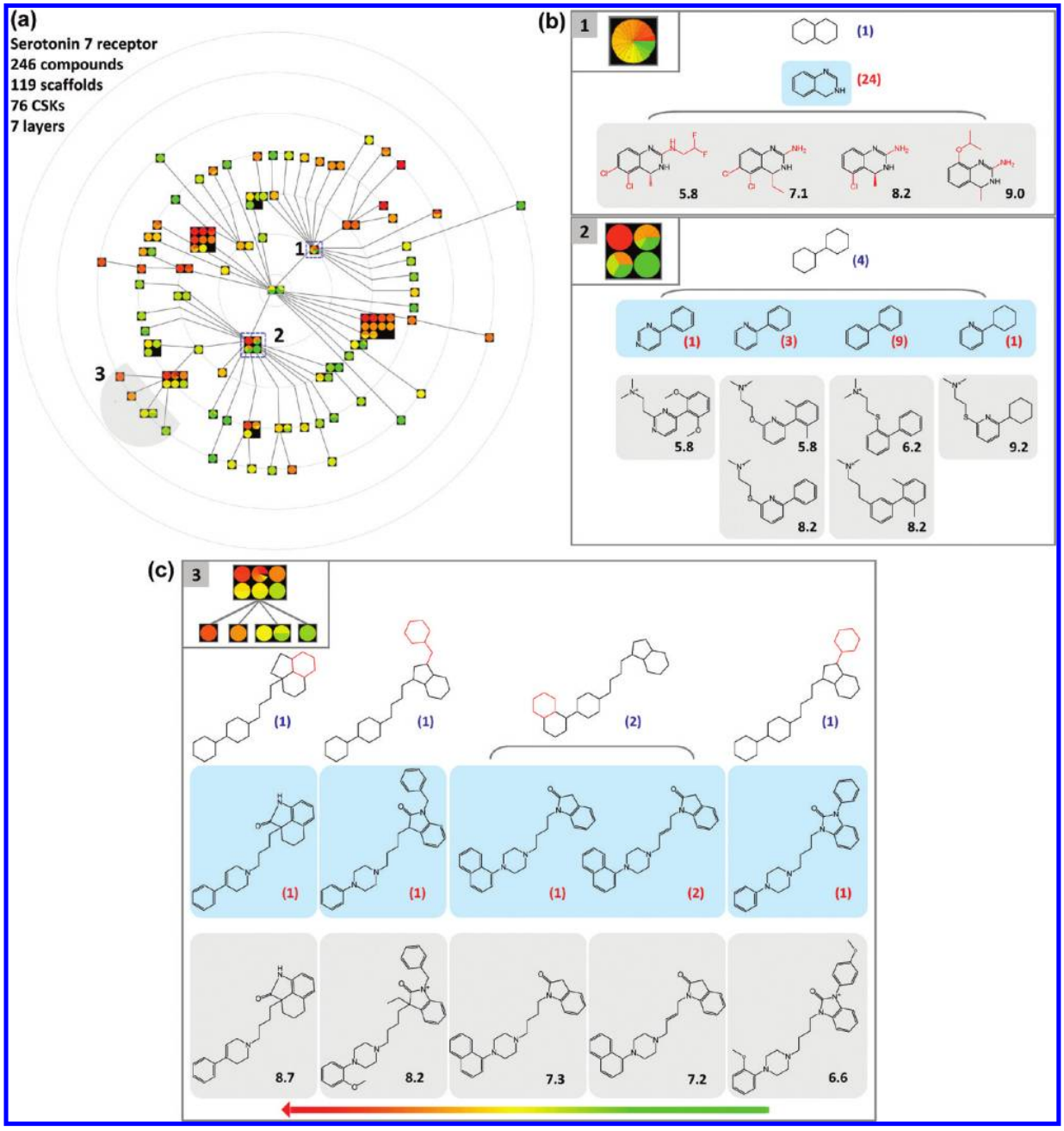


Figure 4. LASSO graph of serotonin 7 receptor antagonists. In (a), the LASSO graph of the compound data set is shown and characteristic patterns are marked and numbered according to Figure 3. In (b) and (c), these patterns and corresponding structures are depicted. For each CSK, the numbers of corresponding scaffolds and compounds are reported in parentheses (in blue and red, respectively). For each pattern, CSKs, scaffolds (on a light blue background), and representative compounds (light gray background) are shown (labeled with their pK_i values). For pattern 1, R-groups in compounds are colored red. For pattern 3, rings that distinguish CSKs are also colored red.

a number of structural pathways and obvious patterns, three of which are marked in Figure 5a (and again numbered according to Figure 3). In Figure 5b, pattern 1 is highlighted, a series of analogues with steady potency progression spanning nearly 4 orders of magnitude, one of the SAR hotspots in this data set. From these analogues, it becomes immediately apparent that the introduction of a nitrile group at the cyclohexane ring significantly increases compound potency. Furthermore, single halogen substituents at the benzene ring are preferred at the ortho and meta positions. However, the largest increase in potency is observed when a bulky trifluoromethyl group is present at the ortho-position. In addition, Figure 5c shows structures that participate in the formation of pattern 2 involving a total of 10 topologically equivalent scaffolds (each containing two pairs of condensed rings and two additional single rings). Among these are chemically very similar scaffolds representing compounds that are only distinguished by one or two ring substitutions and/or stereochemistry at a single stereocenter. However, these modifications cause potency differences of 2−4 orders of magnitude. Hence, this pattern

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

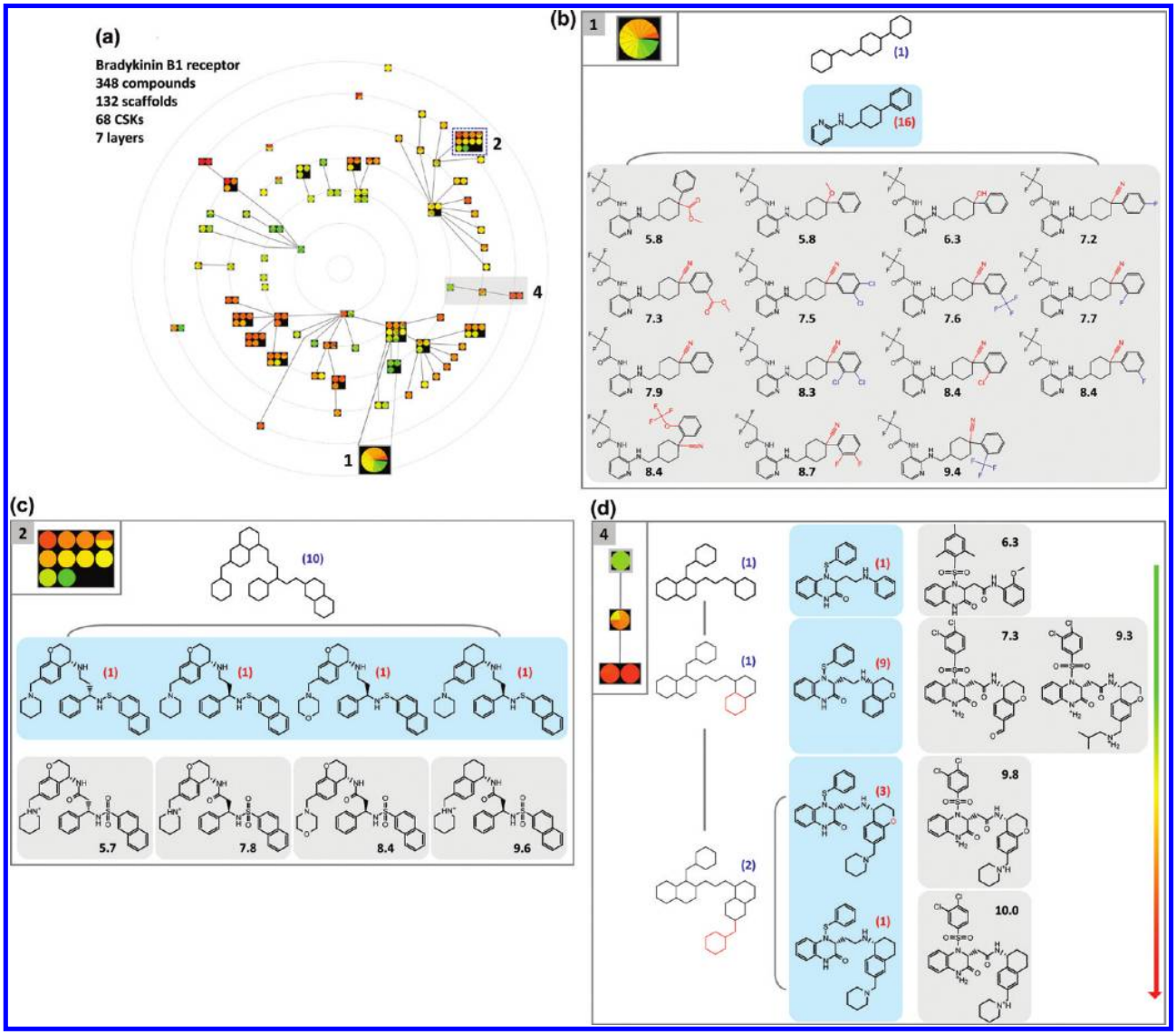


Figure 5. LASSO graph of bradykinin B1 receptor antagonists In (a), the LASSO graph of the compound data set is shown and characteristic patterns are marked and numbered according to Figure 3. In (b), (c), and (d), these patterns and corresponding structures are depicted. The presentation is according to Figure 4. In the analogues corresponding to pattern 1, a conserved substituent at the pyridine moiety is shown in dark gray and R-groups that reveal SAR information are shown in red and blue. For pattern 4, rings added at each layer are colored red.

identifies a highly SAR-informative compound subset. Moreover, in Figure 5d, pattern 4 is analyzed, a characteristic vertical pattern formed by a disjoint structural pathway from layer 4 to layer 6 in the graph. Here, the addition of two single rings in subsequent steps leads to significant progression in potency. The two scaffolds represented by the terminal CSK in layer 6 yield highly potent compounds. A comparison of two representative highly potent compounds (with pK_i values of 9.8 and 10.0, respectively; bottom of Figure 5d) containing a terminal piperidyl ring with a compound represented by the intermediate scaffold in layer 5 (pK_i of 9.3; on the right in Figure 5d) is particularly interesting. The latter compound is much more potent than its closely related analogues and also contains an aliphatic piperidine mimic at the corresponding substitution site. Thus, this comparison clearly implicates the piperidine substituent as an SAR determinant within this series. It also illustrates that vertical graph patterns can reveal detailed SAR information.

■ CONCLUSIONS

Herein we have introduced a new and intuitive graphical data mining method for the structural organization and representation of compound sets and the exploration of SAR information. The LASSO graph globally organizes compound sets according to a well-defined structural hierarchy, integrates compound activity data, and reveals signature patterns that capture SAR information. Conceptually, the LASSO graph is related to the Scaffold Tree data structure. However, different from the Scaffold Tree and its extensions, the LASSO graph is not designed for scaffold decomposition or generation of virtual scaffolds. Moreover, it is focused on a compound data set rather than scaffold representation. LASSO graphs also take an additional structural criterion into account, the topological

dx.doi.org/10.1021/jm3004762 | J. Med. Chem. 2012, 55, 5546−5553

equivalence of scaffolds, which is assessed by considering cyclic skeletons. Characteristic features of the LASSO graph include its constant reference frame for the representation of structurally homogeneous or heterogeneous compound sets and its signature patterns that identify SAR-informative compound subsets. A special feature of LASSO that sets it apart from other scaffold representations is the presence of compound−scaffold−skeleton sequences that capture substructure and topological features of active compounds at different levels and enable "forward−backward" SAR exploration. The data structure emphasizes both global and local structural and SAR features. As such, the LASSO graph further extends the current spectrum of graphical SAR analysis tools. Exemplary applications suggest that ease of interpretation is a particular attractive aspect of LASSO graph analysis.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Author Contributions

†The contributions of these two authors should be considered equal.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS USED

CSK, cyclic skeleton; SAR, structure−activity relationship; LASSO, layered skeleton−scaffold organization

REFERENCES

(1) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure−Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(2) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369−378.

(3) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure−Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075−6084.

(4) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.

(5) Perez-Villanueva, J.; Santos, R.; Hernandez-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. Structure−Activity Relationships of Benzimidazole Derivatives as Anti-Parasitic Agents: Dual-Activity Difference (DAD) Maps. *Med. Chem. Commun.* **2011**, *2*, 44−49.

(6) Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.

(7) Wawer, M.; Bajorath, J. Similarity-Potency Trees: A Method To Search for SAR Information in Compound Data Sets and Derive SAR Rules. *J. Chem. Inf. Model.* **2010**, *50*, 1395−1409.

(8) Wawer, M.; Sun, S.; Bajorath, J. Computational Characterization of SAR Microenvironments in High-Throughput Screening Data. *Int. J. High Throughput Screening* **2010**, *1*, 15−27.

(9) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926−5937.

(10) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational Analysis of Multi-Target Structure−Activity Relationships To Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem* **2010**, *5*, 847−858.

(11) Wassermann, A. M.; Bajorath, J. Directed R-group Combination Graph: A Methodology to Uncover Structure−Activity Relationship Patterns in Series of Analogs. *J. Med. Chem.* **2012**, *55*, 1215−1226.

(12) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742−1753.

(13) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010**, *53*, 5002−5011.

(14) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(15) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581−583.

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

(17) Xu, Y.-J.; Johnson, M. Using Molecular Equivalence Numbers To Visually Explore Structural Features That Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.

(18) *OEChem TK*, version 1.7.4.3; OpenEye Scientific Software Inc.: Santa Fe, NM, 2010.

(19) Java Universal Network/Graph Framework, version 2.0.1. http://jung.sourceforge.net/ (accessed March 7, 2012).

(20) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

Summary

The LASSO graph is a novel representation that provides a bird's eye view of SARs in compound data. It utilizes hierarchical molecular frameworks to organize compound sets for the exploration of SAR information. The compact graphical layout reveals both global and local SAR patterns. Signature patterns that identify SAR-informative compound subsets can be easily identified from the compact graphical layout. Additionally, the graphical topology enables the recognition of structurally homogeneous and heterogeneous compound data sets. My contribution to this study has been the implementation of the LASSO graph and the analysis of different compound data sets using the graph.

The LASSO graph representation is restricted to descriptive SAR analysis. It does not directly provide novel compound design suggestions and activity predictions. The SAR matrix data structure organizes SAR-informative compound series from large data sets and reports several virtual compounds for optimization. This data structure has been further extended to help bridge the gap between data-driven SAR analysis, compound design, and activity predictions and study compound series in multi-target activity spaces. In Chapter 3, the SAR matrix approach and its extensions are presented.

Chapter 3

Second Generation SAR Matrices

Introduction

Activity landscapes are indispensable tools for SAR analysis, however, a steady increase in the size of compound data is making SAR analysis extremely challenging.[2] It is essential to develop computational methods that can handle increasing complexity of data. The tools should provide intuitive approaches for seeking SAR information and compound design.

The SAR matrix (SARM)[3] method provides an opportunity for combining large-scale SAR analysis and prospective compound design. In SARMs, compounds are organized as structurally related series in a format reminiscent of R-group tables. A dual-step MMP fragmentation approach is used to generate these series. SARMs are represented in a matrix format where structurally-related cores are represented in rows and substituents are represented in columns. Each cell represents a compound as a combination of the core and a substituent. A salient feature of the SARMs is the presence of virtual compounds (VCs) that populate chemical space around structurally related series. These VCs represent unexplored combinations of cores and substituents and are potential candidates for further exploration. SARMs are useful for an-

alyzing hit-to-lead and lead optimization sets in order to identify and optimize potential compound series.

In a collaboration with Pfizer, a focused lead optimization set of approximately 10,000 structurally related kinase inhibitors was analyzed using SARMs. The resulting VCs served as potential candidates to synthesize and test. However, prioritization of a few candidates was difficult from a large number of matrices. This triggered the development of a novel approach to predict VC potency using the SARMs. This advancement helps prioritize the VCs and enhances the utility of SARMs for medicinal chemistry applications.

The SARM method is also utilized to study compound series in multi-activity spaces. The purpose here is not SAR analysis but systematic exploration of compound promiscuity patterns in structurally related analogs. These matrices, referred to as compound series matrices (CSMs), offer suggestions for the design of compounds with multi-target activities and are useful for chemogenomics applications.

In this Chapter, the SAR matrix data structure and its novel methodological extensions are discussed.

METHOD ARTICLE

REVISED

The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics [v2; ref status: indexed, http://f1000r.es/3rg]

Disha Gupta-Ostermann, Jürgen Bajorath

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, D-53113, Germany

v2 **First published:** 16 May 2014, **3**:113 (doi: 10.12688/f1000research.4185.1)
Latest published: 23 Jun 2014, **3**:113 (doi: 10.12688/f1000research.4185.2)

Abstract

We describe the 'Structure-Activity Relationship (SAR) Matrix' (SARM) methodology that is based upon a special two-step application of the matched molecular pair (MMP) formalism. The SARM method has originally been designed for the extraction, organization, and visualization of compound series and associated SAR information from compound data sets. It has been further developed and adapted for other applications including compound design, activity prediction, library extension, and the navigation of multi-target activity spaces. The SARM approach and its extensions are presented here in context to introduce different types of applications and provide an example for the evolution of a computational methodology in pharmaceutical research.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 23 Jun 2014		✓ report
version 1 published 16 May 2014	✓ report	✓ report

1 **Herman van Vlijmen**, Janssen Infectious Diseases-Diagnostics BVBA Belgium

2 **Georgia B. McGaughey**, Vertex Pharmaceuticals Inc. USA, **Jonathan Weiss**, Vertex Pharmaceuticals Inc. USA

Discuss this article

Comments (0)

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

How to cite this article: Gupta-Ostermann D and Bajorath J. **The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics [v2; ref status: indexed, http://f1000r.es/3rg]** *F1000Research* 2014, **3**:113 (doi: 10.12688/f1000research.4185.2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 16 May 2014, **3**:113 (doi: 10.12688/f1000research.4185.1)
First indexed: 11 Jun 2014, **3**:113 (doi: 10.12688/f1000research.4185.1)

REVISED **Amendments from Version 1**

We thank all reviewers for their comments. Compound sets used to generate the matrices discussed in the paper are made available as a part of the revision (via ZENODO, please see ref. 16). Readers might also be interested in other publicly available data sets and software tools originating from our laboratory [Hu and Bajorath, F1000Research 2014, 3:69 (doi: 10.12688/f1000research.3979); Hu *et al.* F1000Research 2014, 3:36 (doi: 10.12688/f1000research.3-36.v2]. SARM software is currently not included but the method can be readily implemented on the basis of the information provided.

In the revision, we have expanded the method description (point 3 of the second review) and made corrections (including numbers in Figure 7) and stylistic changes (points 4 and 6). In addition, Figure 1 and Figure 5 have been updated with chemically intuitive compound series. The remaining questions from the second review are answered below (but the answers have not been added to the revision, given the more general nature of these questions):

1. Yes. If such rings are connected via exocyclic bonds to the remaining part of a molecule, the bonds are regularly fragmented and MMPs might be generated.
2. The concept of a Matching Molecular Series was introduced in Wawer and Bajorath, J Med Chem 2011, 54:2944–2951 (doi: 10.1021/jm200026b) and has recently been utilized by O'Boyle *et al.*, J Med Chem 2014, 57:2704–2413 (doi: 10.1021/jm500022q), in the latter case termed Matched Molecular Series.
6. Yes. CSMs can be used to study secondary drug targets and adverse drug reactions. In this context, please also see an MMP application by Hu *et al.* AAPS J 2014, in press (doi: 10.1208/s12248-014-9621-8).

See referee reports

Introduction

Steadily growing numbers of active compounds provide a critically important knowledge base for medicinal chemistry but also challenge Structure-Activity Relationship (SAR) analysis[1]. For important therapeutic targets, compound activity landscapes become increasingly complex[2] and difficult to analyze. Increasing volumes and complexity of compound activity data require the development of computational approaches to effectively extract SAR information from heterogeneous sources[1]. In addition, it is essential to make this information available in an intuitive form that can be appreciated in the practice of medicinal chemistry and utilized in compound design. Therefore, a number of SAR visualization methods and graphical analysis tools have been developed in recent years[2,3] to view SAR characteristics of entire data sets or extract SAR information from compound activity data. Regardless of their algorithmic foundations and design specifics, many (but not all) graphical analysis methods have in common that they provide a bird's eye view of SAR information in compound data sets and depart from the single-series focus that has traditionally governed medicinal chemistry efforts. However, multi-facetted SAR information obtained from heterogeneous compound sources must ultimately again be utilized to advance individual compound series, which is a challenging task.

The Structure-Activity Relationship Matrix (SARM) approach has originally been designed to extract and organize SAR-informative compound series from large data sets[4] and has been further extended to help bridge the gap between data-driven SAR analysis, compound design, and activity predictions[5] and study compound series in multi-target activity spaces[6]. Here, we present the SARM approach and its extensions in context and introduce new features and applications.

Methods

Compound structure analysis and organization

The original design idea underlying the SARM approach was to systematically extract compound series with well-defined structural relationships from data sets and organize them in a matrix format[4]. To convey SAR information, matrix cells representing data set compounds are color-coded according to compound potency. The methodological basis for compound series identification and organization was provided by the matched molecular pair (MMP) concept[7]. An MMP is defined as a pair of compounds that differ only at a single site[7]. Compounds in MMPs can be interconverted by the exchange of a substructure, termed a chemical transformation[8]. In order to generate MMPs on a large scale, compounds must be systematically fragmented. The algorithm by Hussain and Rea[8] (which we re-implemented and further modified in-house) provides an elegant and computationally efficient solution to this task by subjecting compounds to systematic deletion of individual exocyclic single bonds (single-cut) or simultaneous deletion of two (dual-cut) and three (triple-cut) exocyclic single bonds. The resulting fragments are then stored in an index table as keys (core structures) and smaller values (substituents)[8].

The most important aspect of SARM design has been the application of dual fragmentation scheme leading to MMP generation at two levels[4], as outlined in Figure 1. In the first step, MMPs are generated from data set compounds yielding "compound MMPs". In the second step, core fragments from compound MMPs are again subjected to fragmentation leading to the generation of "core MMPs". As a consequence, this hierarchical two-step fragmentation scheme identifies all compound subsets that have structurally analogous cores, i.e., core structures that are only distinguished by a structural modification at a single site. Each subset represents a so-called "structurally analogous matching molecular series" (A_MMS)[4]. Thus, each A_MMS represents a set of compound series with structurally analogous cores. Individual compounds and/or subsets of compounds can belong to multiple A_MMS, hence providing a high-level structural organization of a compound collection that captures all possible (MMP-based) substructure relationships.

SAR matrix design

Each A_MMS is represented in an individual SARM, as illustrated in Figure 1. The SARM is filled with structurally analogous cores resulting from core MMPs (second fragmentation step) and the corresponding substituents obtained from compound MMPs (first fragmentation step). Single-, dual-, and triple-cut matrices are separately generated (*vide supra*). Each cell in a SARM represents a unique compound, i.e., a unique combination of a key and value fragment. Each row contains an individual analog series, i.e., compounds sharing the same core. Each column contains compounds from different series that share the same substituent (single-cut) or substituent combination (dual- or triple-cuts). The series forming a SARM typically contain different sets of substituents, giving rise to "real" compounds (filled cells) and "virtual" compounds (VC;

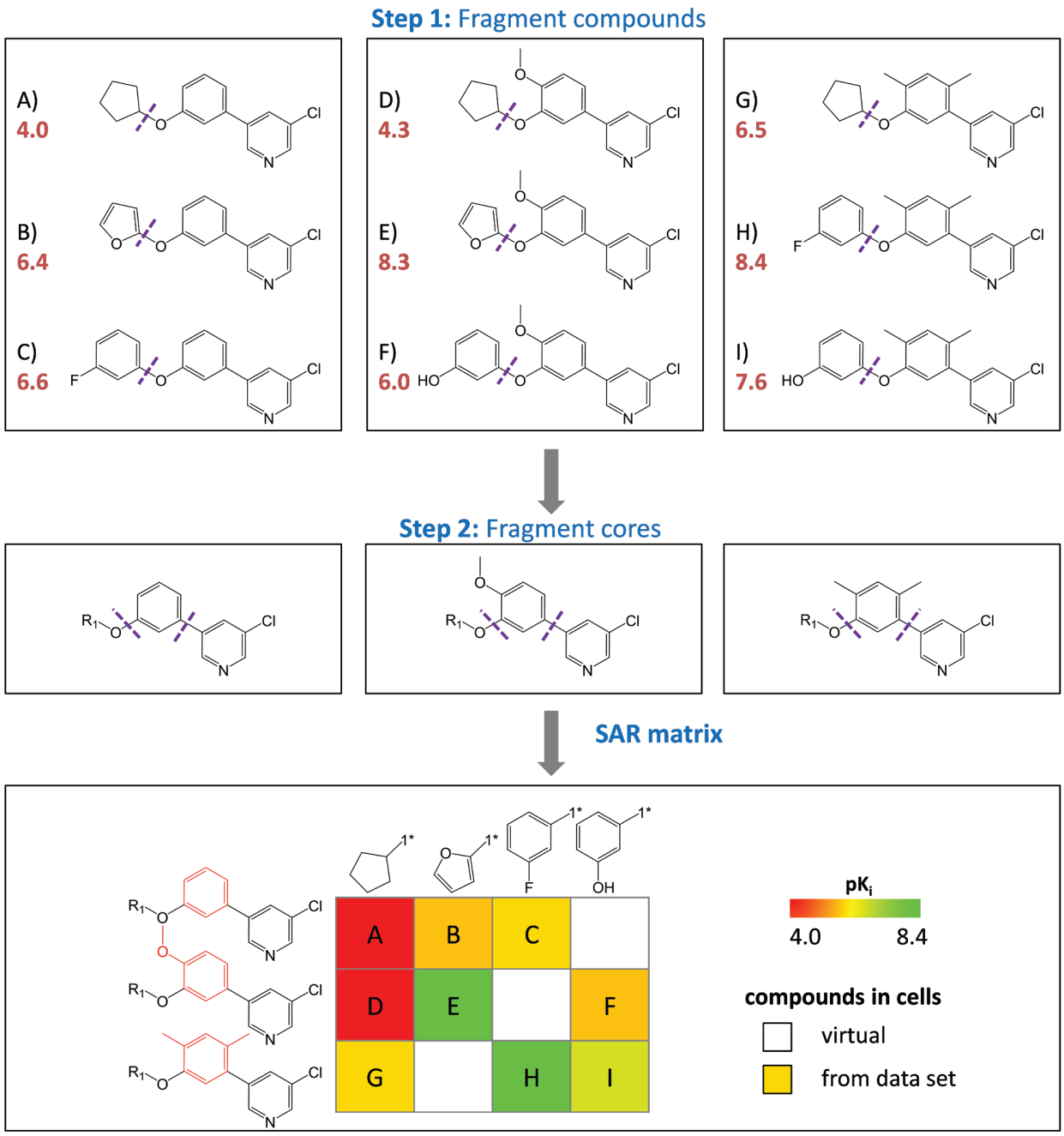


Figure 1. SAR matrix generation. Three model series with three compounds each (**A**–**C**, **D**–**F**, and **G**–**I**) are shown with pK_i values (red). In the first step, all compounds are fragmented at a single bond (purple dotted line) producing compound MMPs that yield a common core (key) and a compound specific substituents (values). In the second step, the cores resulting from the first step are further fragmented to obtain core MMPs. The SARM is then generated by combining series with structurally analogous cores that represent individual rows. In addition, columns represent substituents. In each cell, the combination of a core and a substituent defines a unique compound. Compounds present in the data set are indicated by filled cells that are color-coded according to potency using a continuous spectrum from red (low potency) over yellow to green (high). In addition, empty cells indicate virtual compounds. Substructures distinguishing the core fragments are highlighted in red.

empty cells). As also illustrated in Figure 1, a color spectrum is applied to represent the potency (or ligand efficiency) values of real compounds. Importantly, SARMs resemble standard R-group tables used in medicinal chemistry, although their design and information content is much more complex and comprehensive. Standard R-group tables typically only contain an individual core structure of a single series, all substituents, and associated potency values. However, because SARMs resemble R-group tables, they are readily accessible to medicinal chemists who can inspect individual compounds and their relationships to others.

SAR patterns

In SARMs, different types of SAR patterns become readily apparent. This is illustrated in Figure 2 that shows exemplary SARMs

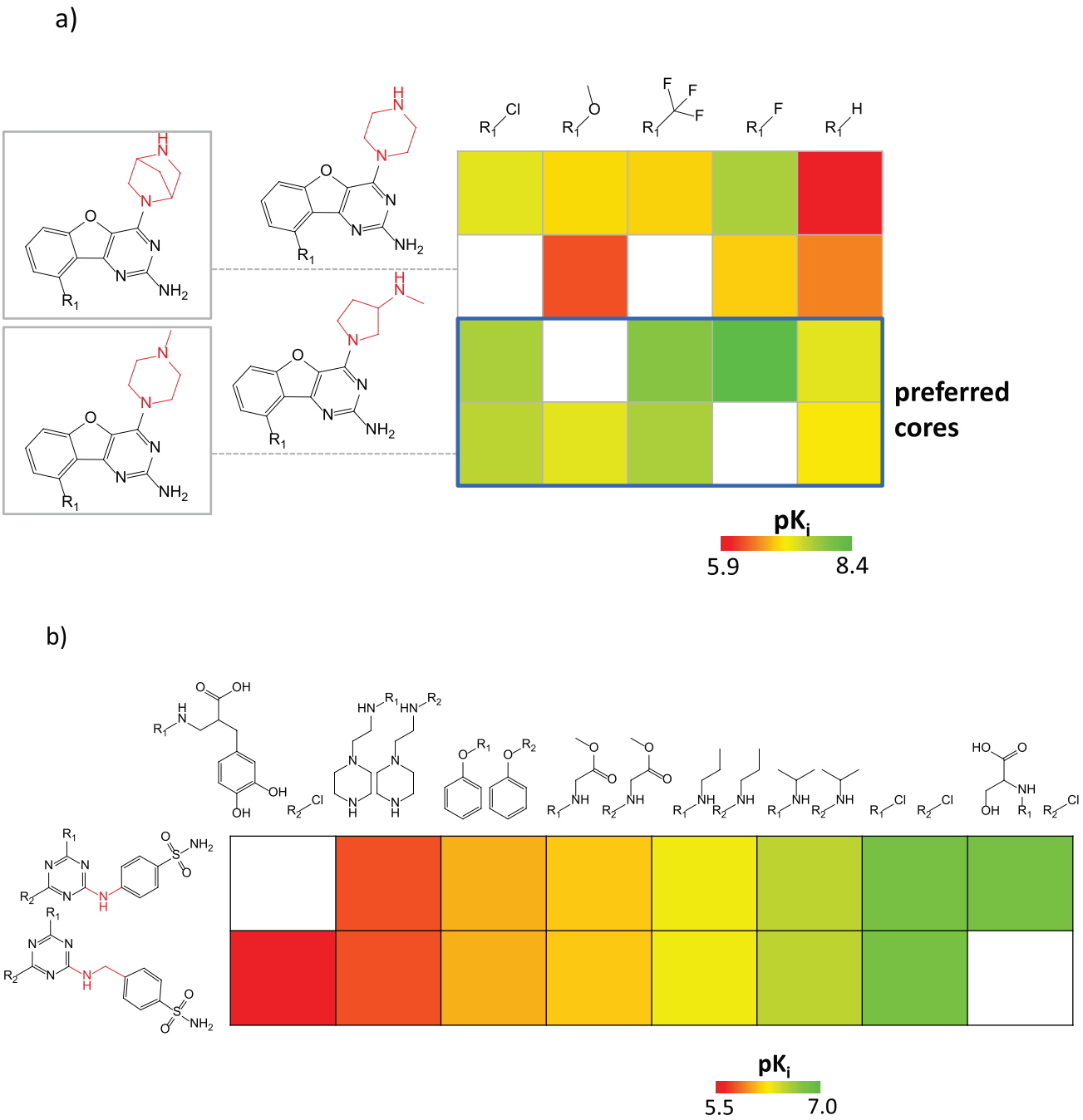


Figure 2. SAR patterns in matrices. In (**a**), a SARM capturing 16 ligands of the histamine H4 receptor is shown. Cells containing analogs with preferred cores yielding potent compounds are framed in blue. Substructures distinguishing the core fragments are highlighted in red. The pK$_i$ value range for the 16 ligands is displayed. In (**b**), a subset of a double-cut SARM is shown that contains series of carbonic anhydrase I inhibitors and an exemplary SAR transfer event.

revealing characteristic patterns (for representation purposes, only small matrices are shown; *vide infra*). For example, the SARM in Figure 2a identifies two preferred core structures that consistently produce potent compounds. Furthermore, the SARM in Figure 2b reveals an SAR transfer event, i.e., the presence of two compound series with related yet distinct core structures that contain pairwise corresponding analogs with similar potency progression. Other SAR patterns that can frequently be detected include, for example, preferred R-groups (or R-group combinations) in related compound series or regions of distinct SAR continuity or discontinuity. Continuous SAR regions are characterized by the presence of compounds with structural modifications that lead to gradual changes in potency, whereas discontinuous SAR regions contain structural analogs with large (and essentially unpredictable) potency variations[2].

Matrix distribution and ranking

Large compound data sets typically yield many SARMs of different size and composition, depending on their degree of structural homogeneity or heterogeneity. Two examples are given to illustrate this point. First, an in-house focused compound library with various substitutions of a small number of core structures comprising 6503 compounds produced a total of 6738 (single-, double- and triple-cut) matrices containing a total of 135,619 VCs. Second, a structurally heterogeneous set of 509 purinergic receptor (P2Y12) ligands generated a total of 181 SARMs containing 17,445 VCs. Again, each SARM contains a unique A_MMS and individual compounds might belong to multiple A_MMS depending on the structural relationships they form. SARMs provide highly resolved views of all of these structural relationships. Depending on the number of compounds forming A_MMS, the size of SARMs can considerably vary. For example, in a survey of 32 different activity classes consisting of 398 to 2497 compounds, SARMs were found to contain between three and 555 compounds, with a median value of 13. Furthermore, we also use a "matrix overlap" measure to account for the overlap between the corresponding substituents (columns) in different A_MMS (rows), which typically varies in SARMs. Matrix overlap is determined as the average over all row overlap values. For individual columns in SARMs, row overlap (RO) is calculated as:

$$\mathrm{RO} = \frac{n\_col - 1}{\#rows - 1}$$

where, *n_col* correspond to the number of data set compounds present in each column. RO yields a numerical score between 0 (no overlap) and 1 (complete overlap). Figure 3 reports the matrix overlap distribution for SARMs from the focused library referred to above, which is a fairly representative distribution for structurally homogeneous data sets. Here 5% of the SARMs have an RO of 0 for each column; hence, the final matrix overlap score is 0 indicating the mutually exclusive nature of the substitution pattern among the A_MMS. By contrast, 30% of the SARMs have an RO of 1 for each column; hence, the final matrix overlap is 1 reflecting the presence of A_MMS with identical substitution patterns. As an alternative measure, "matrix coverage" (C), which accounts for the proportion of cells in a SARM that are populated with real compounds *n_matrix* can be calculated as:

$$\mathrm{C} = \frac{n\_matrix}{\#rows * \#columns}$$

Regardless of the number of SARMs that are obtained from large data sets, there are too many for one-by-one inspection. Hence, ranking schemes should be applied to prioritize and pre-select those SARMs that are most informative for a given application. For instance, SARMs can be easily ranked on the basis of numerical functions that prioritize matrices containing preferred substituent combinations or core structures and SAR transfer events or matrices that capture high degrees of local SAR continuity or discontinuity. For example, Figure 4 shows two SARMs originating from a large data set that are highly ranked on the basis of SAR discontinuity

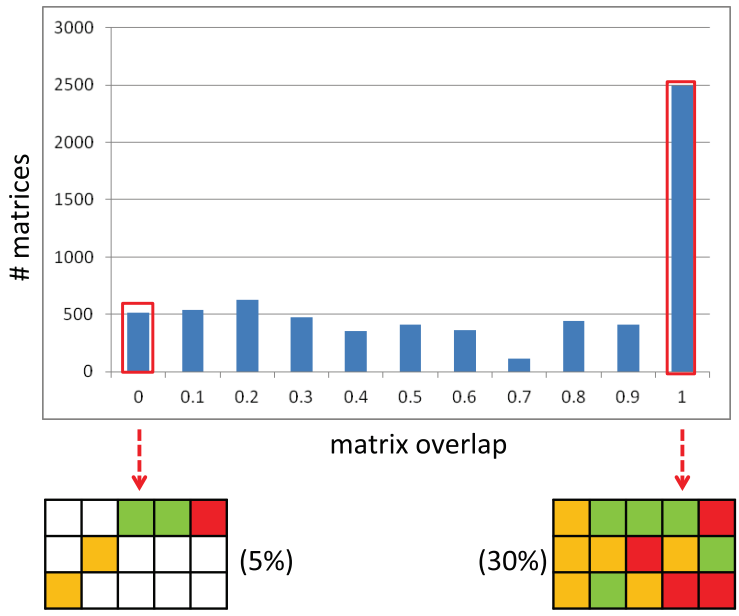


Figure 3. Matrix overlap distribution. Shown is a histogram with the matrix overlap distribution for SARMs from an in-house focused library.

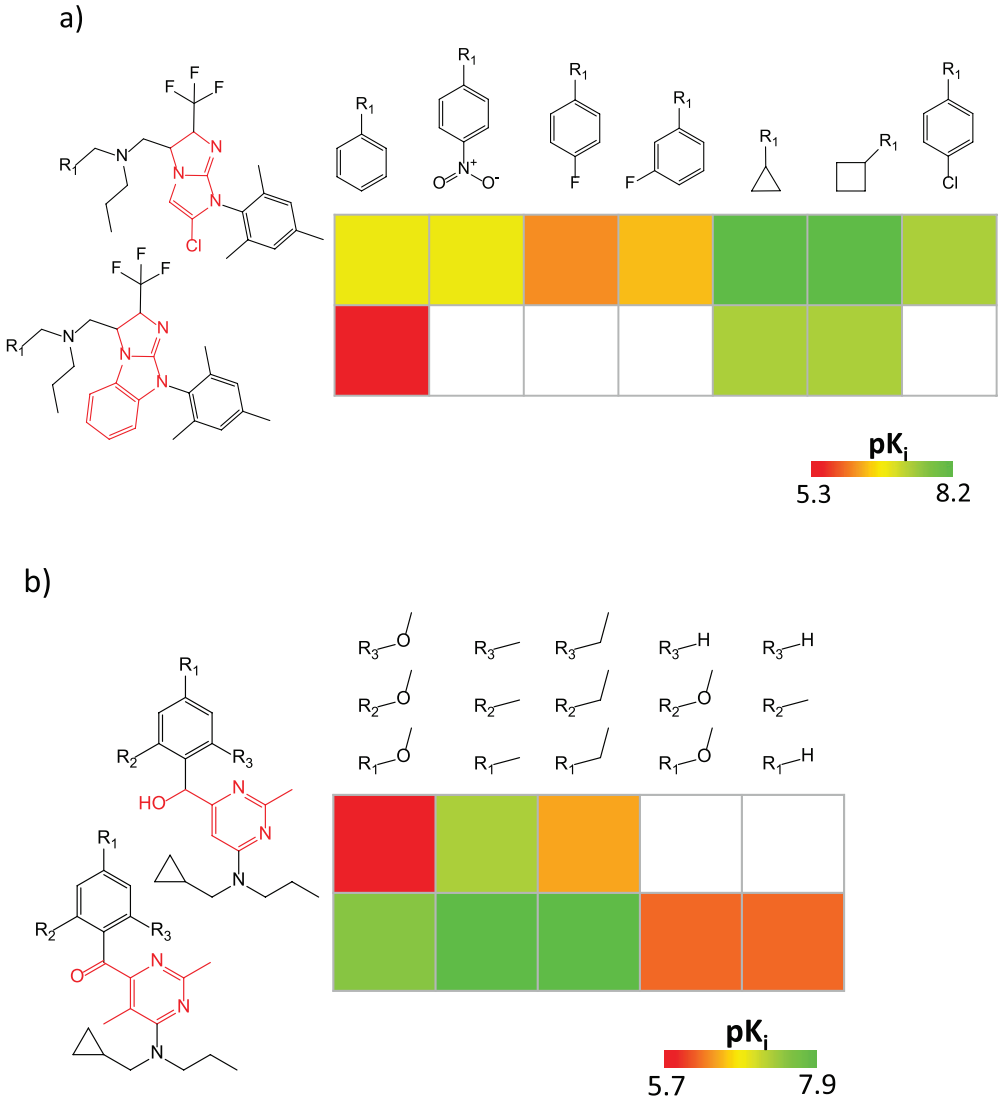


Figure 4. Ranked SAR matrices. In (**a**) and (**b**), two SARMs are shown (resulting from single- and triple-cut fragmentation, respectively) for corticotropin-releasing factor receptor 1 ligands that were highly ranked on the basis of SAR discontinuity scoring.

(as indicated by the presence of multiple analogs with large potency differences). Depending on the applied selection criteria, most informative SARMs can be readily inspected on the basis of a ranked list.

Compound design and activity prediction

VCs contained in SARMs provide immediate suggestions for compound design. Because VCs represent unexplored key-value combinations derived from data set compounds, the union of VCs from all SARMs provides a "chemical space envelope" for a given compound set or library. VCs originating from SAR-informative matrices represent natural focal points for interactive compound design. Moreover, the potency of many virtual compounds can be predicted by applying a compound neighborhood (NBH) principle[5], as illustrated in Figure 5. An NBH of a given VC is defined by three adjacent real compounds that contain the core of the VC (compound G in Figure 5), its substituent (compound E) and the core and substituent of G and E (compound D). The potency of the VC can then be predicted by applying the additivity assumption underlying Free-Wilson analysis[9] using the simple equation shown in Figure 5. The putative potency value of the VC results from the sum of (logarithmic) potencies of the two real compounds sharing the same core and substituent with the VC, respectively, minus the potency of the compound that contains the core structure and substituent of the two other real compounds. Thus, from NBHs, "mini-QSAR" models are derived for activity prediction. For each candidate VC, qualifying NBHs are collected across all SARMs, individual potency predictions are carried out, and their consistency is evaluated, for example, by calculating standard deviations for predictions[5]. In benchmark calculations on six different sets of

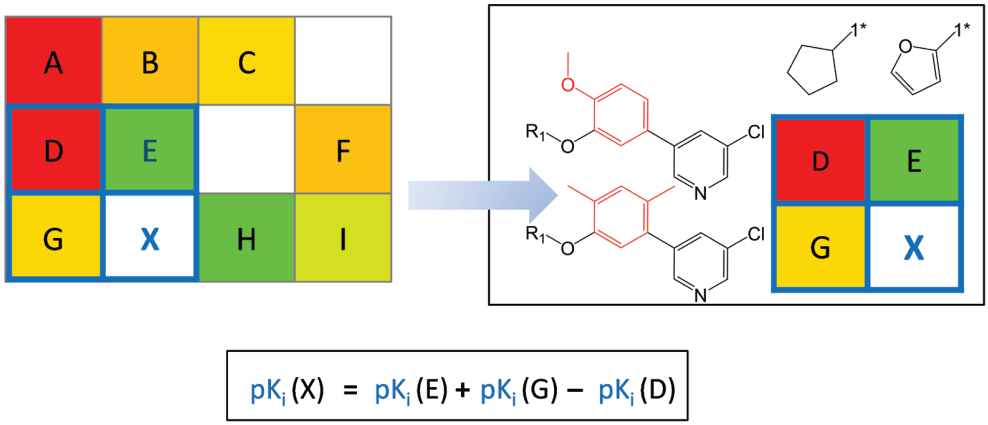


Figure 5. Neighborhood-based potency prediction. An NBH of virtual compound X is marked in blue in a model SARM and displayed in detail. Compounds E and G share the same substituents and core with X, respectively, and the third neighbor D combines the core and substituent of E and G, respectively. At the bottom, the equation to predict the potency of X from the potency values of E, G, and D is shown.

G protein-coupled receptor ligands, potency values of subsets of test compounds falling into continuous local SAR regions were accurately predicted using the NBH-based approach, and prediction accuracy generally increased with the number of qualifying NBHs[5]. This is also relevant for practical applications. For potency prediction, candidate VCs should be prioritized for which multiple NBHs are available. For example, for the set of 509 purinergic receptor ligands (*vide supra*), 5167 of 17,445 VCs were found to have at least three qualifying NBHs. Hence, in these cases, the consistency of potency predictions can be assessed. Such candidate VCs can be explored in a systematic manner. For libraries tested in individual assays, VCs predicted to be consistently active on the basis of multiple NBHs provide preferred candidates for target/assay-dependent library expansion and focusing.

Importantly, the NBH-based mini-QSAR approach is only applicable to candidate compounds falling into SARMs that represent continuous SAR regions, as illustrated in Figure 6a. By contrast, compounds falling into discontinuous SAR regions, as shown in Figure 6b, fall outside the applicability of standard QSAR modeling. Nonetheless, VCs from SARMs representing discontinuous SAR regions are also attractive candidates for compound design. This especially applies to VCs falling into the vicinity of activity cliffs[10] that are formed by pairs of structural analogs with large potency differences, as illustrated in Figure 6b. Activity cliffs represent the pinnacle of SAR discontinuity. VCs in the vicinity of activity cliff can often be expected to display large (positive or negative) potency fluctuations and are hence attractive candidates in the search for potent hits. Although a QSAR formalism cannot be applied to predict the potency of such compounds, they can be easily selected from SARMs containing activity cliffs on the basis of a "guilt-by-association" principle, i.e., VCs are preferentially selected that are neighbors of potent activity cliff partners. For this purpose, SARMs capturing high degrees of local SAR discontinuity are selected on the basis of discontinuity ranking (*vide supra*).

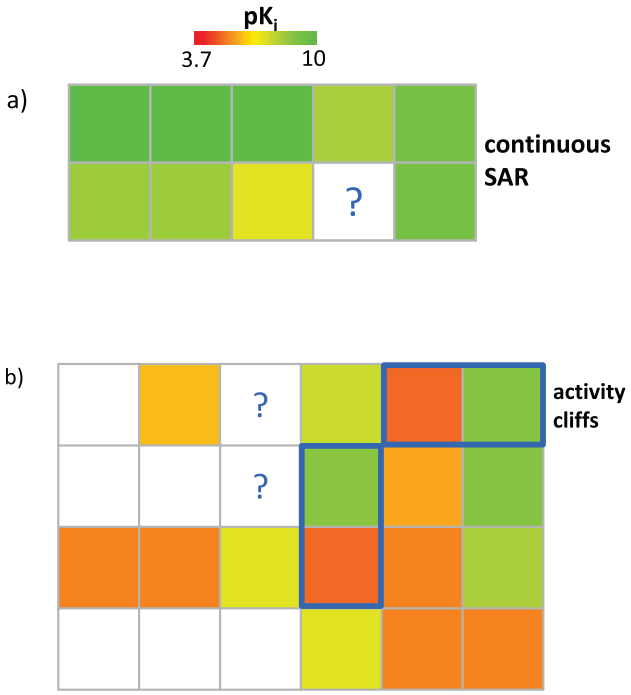


Figure 6. Candidate compound selection and activity prediction. In (**a**), a SARM is shown that represents a highly continuous local SAR environment. In this case, the potency of a virtual compound can be predicted using the NBH-based approach. By contrast, (**b**) shows a SARM representing a discontinuous local SAR. Activity cliff-forming compound pairs are highlighted in blue. Such SAR environments fall outside the applicability domain of NBH-based potency predictions. However, marked VCs represent promising candidates for compound design based on their proximity to activity cliffs. Both SARMs originate from a set of cannabinoid CB1 receptor ligands (compound structures are omitted for clarity).

Multi-target activity spaces

SARMs have also been adapted for the navigation of multi-target activity spaces, which are populated by promiscuous compounds. In this context, promiscuity is defined as the ability of a compound to specifically interact with multiple targets (as opposed to non-specific binding effects)[11]. Here, the primary purpose of the matrix approach is not SAR analysis, but the systematic exploration of compound promiscuity patterns. Therefore, matrices capturing multi-target activities are generated. Such matrices have been designated as Compound Series Matrices (CSMs)[6]. CSMs are of interest for chemogenomics applications in which compound-target interactions are systematically explored[12]. In Figure 7, two exemplary CSMs of different composition and target coverage are shown that reveal different compound promiscuity patterns. In CSMs, data set compounds are color-coded according to the number of targets they are active against (instead of potency-based coloring). In Figure 7a,

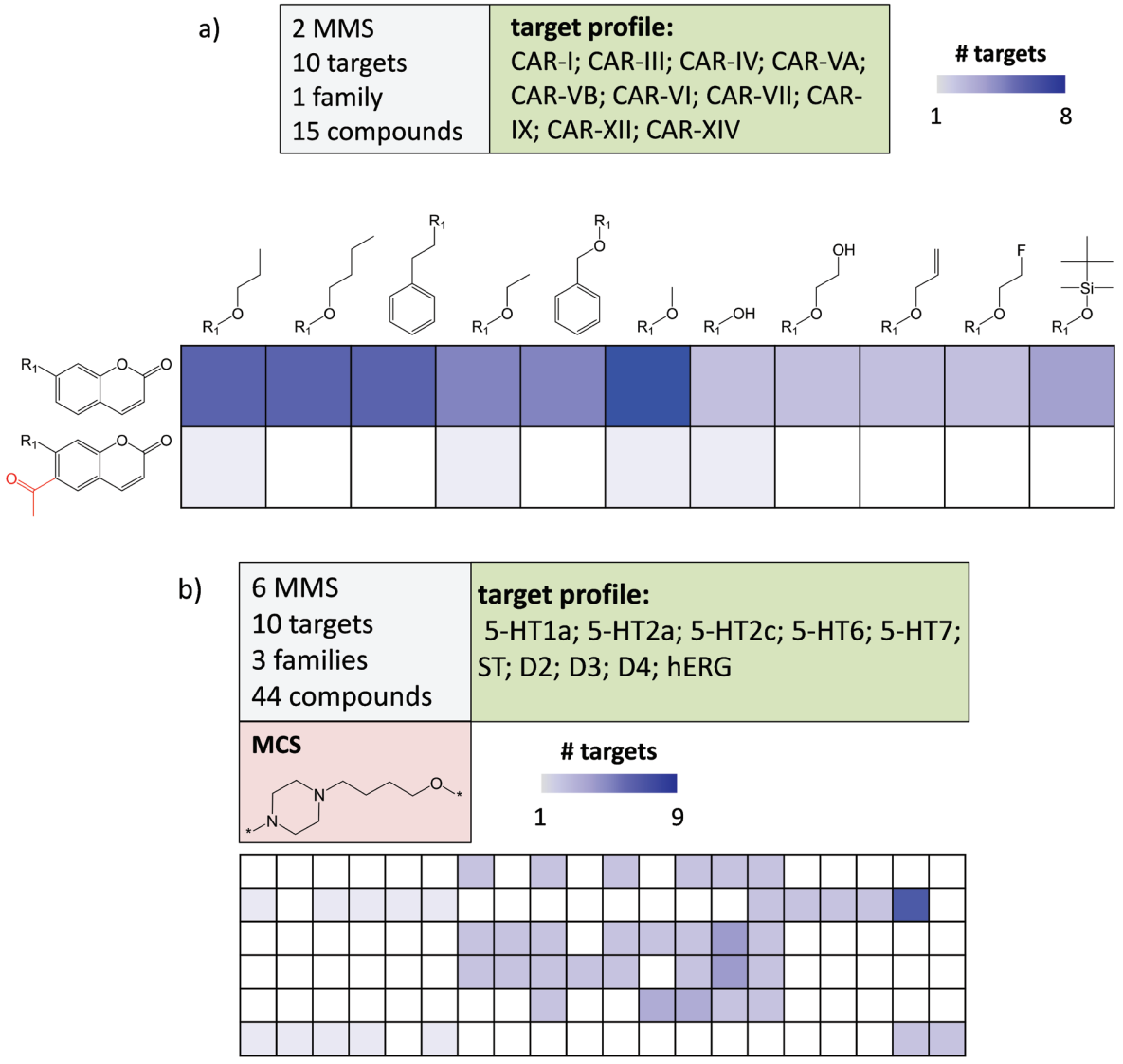


Figure 7. Multi-target compound series matrices. (**a**) shows a CSM containing 15 inhibitors of 10 carbonic anhydrase (CAR) isoforms. Target coverage of analogs is reflected by increasingly dark blue shading of cells. Substructures distinguishing the core fragments are highlighted in red. The matrix composition is summarized (top left) and the target profile reported (top right). (**b**) shows a CSM with 44 analogs active against 10 targets (including the hERG anti-target) belonging to three different families. The maximum common core structure (MCS) of the analog series is displayed. For clarity, compound structures are omitted. Target abbreviations: 5-HT; serotonin receptor, ST; serotonin transporter, D; dopamine receptor, hERG; hERG ion channel.

two structural analogs display very different degrees of promiscuity and in Figure 7b, a center of promiscuity is identified in a sparsely populated matrix. CSMs are designed to mine chemogenomics data sets and also offer immediate suggestions for the design of compounds with different multi-target activities. In addition, it is also readily possible to deconvolute CSMs into individual single-target SARMs, as illustrated in Figure 8. This makes it possible to compare SARMs across different targets and identify compounds that are attractive candidates for testing against additional targets.

Programs and compounds

Java programs were written, in part with the aid of the OpenEye chemistry tool kit[13], to identify A_MMS and generate, rank, and display SARMs. Routines for potency predictions were also implemented in Java. Statistical analyses were carried out using R[14]. All compounds shown herein were obtained from ChEMBL[15].

Concluding remarks

Herein, we have reviewed the design of the SARM methodology and discussed recent extensions and selected applications. In-house implementations of the SARM approach have been continuously developed and further refined to increase the utility of the methodology for medicinal chemistry. Primary reasons for discussing the different aspects and applications of SARMs in context have been to expose this approach to a wider drug development audience and provide an example for the data- and application-driven evolution of a computational medicinal chemistry method. SARMs can essentially be rationalized as local activity landscapes of data sets that are based upon a unique and comprehensive structural organization. SARMs primarily focus on activity information associated with series of closely related compounds but can also be applied to systematically study compound promiscuity patterns. In addition, they can also be easily adapted to explore other structure-property relationships relevant to drug discovery. A special feature of SARMs that sets them apart from many other activity landscape representations is that they closely link descriptive compound data analysis (a primary task of activity landscape modeling) and prospective compound design. Because SARMs are reminiscent of conventional R-group tables, they are readily intuitive to medicinal chemists, thus circumventing the communication barrier that often hinders the effective application of computational approaches in the practice of medicinal chemistry. Future research activities will focus on

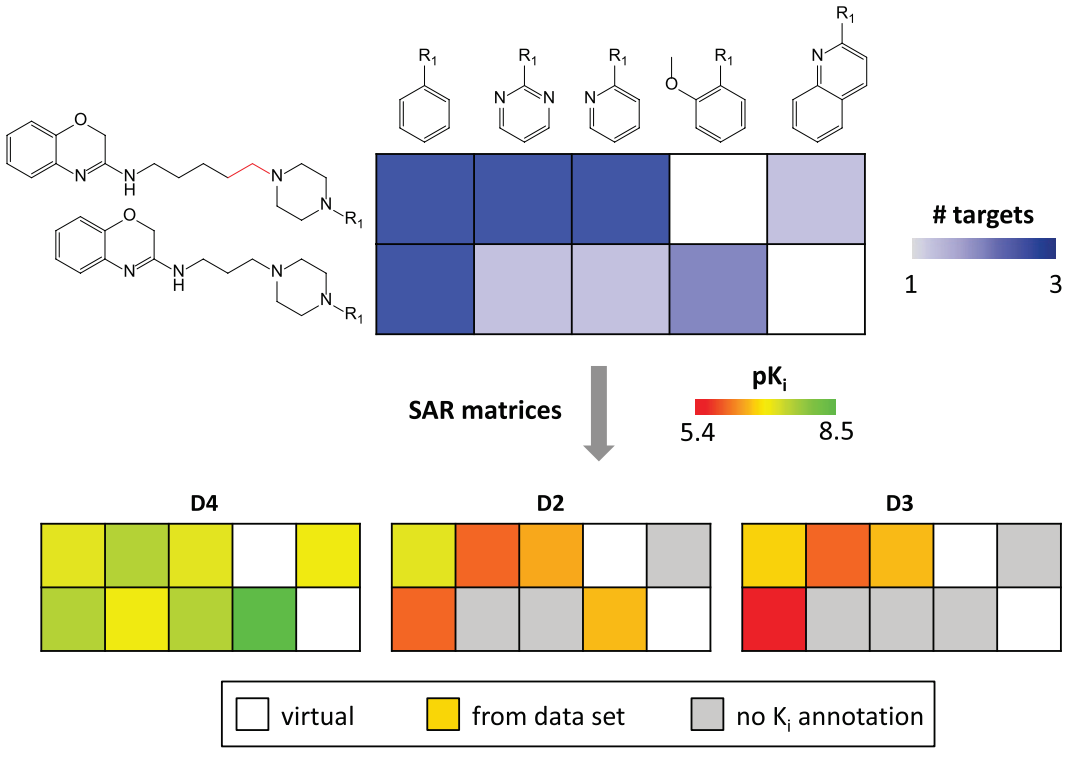


Figure 8. Matrix conversion. The deconvolution of a CSM with eight analogs active against the dopamine D2, D3, and D4 receptor isoforms into three single-target SARMs is illustrated. In all matrices, cells corresponding to VCs are not color-coded. In SARMs, cells of compounds with no available activity annotation for a given target are colored gray.

the design of multi-property SARMs to aid in advanced compound optimization efforts.

Data availability

The compound data sets used to generate the SARMs and the CSMs are available via ZENODO[16].

Author contributions

JB conceived the study, DGO collected the data and generated the representations, JB wrote the manuscript, and both authors examined the manuscript and agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

The authors thank Dr. Anne Mai Wassermann, Dr. Dilyana Dimova, and Dr. Preeti Iyer for key contributions to SARM method development and applications.

References

1. Hu Y, Bajorath J: **Learning from 'big data': compounds and targets.** *Drug Discov Today.* 2014; **19**(4): 357–360.
PubMed Abstract | Publisher Full Text
2. Wassermann AM, Wawer M, Bajorath J: **Activity landscape representations for structure-activity relationship analysis.** *J Med Chem.* 2010; **53**(23): 8209–8223.
PubMed Abstract | Publisher Full Text
3. Stumpfe D, Bajorath J: **Methods for SAR visualization.** *RSC Adv.* 2012; **2**(2): 369–378.
Publisher Full Text
4. Wassermann AM, Haebel P, Weskamp N, *et al.*: **SAR matrices: automated extraction of information-rich SAR tables from large compound data sets.** *J Chem Inf Model.* 2012; **52**(7): 1769–1776.
PubMed Abstract | Publisher Full Text
5. Gupta-Ostermann D, Shanmugasundaram V, Bajorath J: **Neighborhood-based prediction of novel active compounds from SAR matrices.** *J Chem Inf Model.* 2014; **54**(3): 801–809.
PubMed Abstract | Publisher Full Text
6. Gupta-Ostermann D, Hu Y, Bajorath J: **Systematic mining of analog series with related core structures in multi-target activity space.** *J Comput Aided Mol Des.* 2013; **27**(8): 665–674.
PubMed Abstract | Publisher Full Text
7. Kenny PW, Sadowski J: **Structure modification in chemical databases**. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; 271–285.
Publisher Full Text
8. Hussain J, Rea C: **Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets.** *J Chem Inf Model.* 2010; **50**(3): 339–348.
PubMed Abstract | Publisher Full Text
9. Kubinyi H: **Free Wilson analysis. Theory, applications and its relationships to Hansch analysis.** *Quant Struct Act Relat.* 1988; **7**(3): 121–133.
Publisher Full Text
10. Stumpfe D, Bajorath J: **Exploring activity cliffs in medicinal chemistry.** *J Med Chem.* 2012; **55**(7): 2932–2942.
PubMed Abstract | Publisher Full Text
11. Hu Y, Bajorath J: **Compound promiscuity: what can we learn from current data?** *Drug Discov Today.* 2013; **18**(13–14): 644–650.
PubMed Abstract | Publisher Full Text
12. Bajorath J: **Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery.** *Expert Opin Drug Discov.* 2008; **3**(12): 1371–1376.
PubMed Abstract | Publisher Full Text
13. **OEChem, version 1.7.7, OpenEye Scientific Software, Inc.**, Santa Fe, NM, USA, 2012.
Reference Source
14. **R: A Language and environment for statistical computing; R Foundation for statistical computing**, Vienna, Austria, 2008.
Reference Source
15. Gaulton A, Bellis LJ, Bento AP, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1100–D1107.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Gupta-Ostermann D, Bajorath J: **The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics.** 2014.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓

Version 2

Referee Report 02 July 2014

doi:10.5256/f1000research.4876.r5334

Georgia B. McGaughey
Vertex Pharmaceuticals Inc., Cambridge, MA, USA

Thank you for your responses. Revised manuscript looks good and I look forward to reading the paper in AAPS on off target effects.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 11 June 2014

doi:10.5256/f1000research.4481.r4991

Georgia B. McGaughey, **Jonathan Weiss**
Vertex Pharmaceuticals Inc., Cambridge, MA, USA

This is a well written article with a sound description of the Structure-Activity Relationship (SAR) Matrix (SARM) methodology. Only a few questions / suggestions are recommended.

1. In the Methods section, you describe the definition of a MMP. Specifically, you rely on the algorithm by Hussain and Rea. Have you considered MMPs where the only change is in a ring (i.e., an aromatic versus partially saturated ring)?

2. Continuing onwards in the Methods section, you extend the MMP concept to include Matched Molecular Series (MMS). While I believe you were the first to coin this description, there are now others who are also using this formalism (e.g. NextMove (Roger Sayle)) and some reference to these additional methods is advised, particularly to avoid confusion since the names are similar. If there are differences, perhaps you could expand on them.

3. In the Matrix distribution and ranking section, could you expand on the matrix overlap, row overlap and matrix coverage with at least one specific example (ie show the math in the supplementary

material)? Additionally, I recommend using the same variables (i.e. are "*n*" and "*#real compounds*" the same)? If so, then they should be consolidated to one variable. Additionally, it's not obvious how the numbers to the right (5%) and left (30%) of the two matrices are derived.

4. There appears to be some mis-counting of the number of targets in both Figure 7a & b. For Figure 7a, are there 10? And for Figure 7b, we count 10 targets and 4 families. Make sure to check the Figure 7 caption. Consider referring to DOP-D2 as merely D2; likewise DOP-D3 should be merely D3. Check the commas and semicolons as there are inconsistencies.

5. Can these multi-target compound series matrices be tied to Adverse Drug Reactions (ADR)? That seems as a possible extension of this current work. In Figure 7b you consider hERG and hence, it seems as an opportunity to extend beyond merely the primary activity.

6. Consider changing the word "*accessible*" in the concluding remarks to "intuitive" or "interpretable".

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 10 June 2014

doi:10.5256/f1000research.4481.r4803

Herman van Vlijmen
Janssen Infectious Diseases-Diagnostics BVBA, Beerse, Belgium

The manuscript on the SAR matrix method offers a useful approach to extract the relevant information from large datasets with compound activity data, and to present this in an intuitive way to chemists and computational chemists. It is therefore an attractive tool to use in analysis of HTS screens and to find structure activity trends and discontinuities in large groups of structurally related molecules.

The title and abstract cover the content well. The chemogenomics application of the method is related to the use of compound promiscuity instead of the more usual compound activity on a single target. The methods are clearly described and can most likely be reproduced. The datasets used (even the dataset from the public source CHEMBL) are not provided, so the results will be difficult to reproduce. There is no mention in the manuscript on the availability of the tools that were developed. Methods like these could get widespread usage if they would be available to a wider audience. It would also be good if at least the public dataset would be made available so the results can be compared to other approaches.

One significant benefit of this approach is that a large dataset can automatically be processed by the method to create multiple (often very many) SAR matrices. The authors point out that the idea is not that all these (often hundreds or thousands) SAR matrices are inspected visually, but that interesting elements in the matrices can be identified automatically, for instance "virtual compounds" (core-substituent combinations not yet made) that are predicted to have interesting activity. The same virtual compound can appear in different SAR matrices, and therefore multiple predictions can be made for the same compound and the level of consistency could be a good indicator for deciding to make he actual compound. The method also automatically identifies virtual compounds that are close to activity cliffs and are therefore interesting to make and test.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Summary

An overview of the SARM methodology and its extensions has been provided primarily to increase awareness in the drug development audience. The evolution of this method is driven by the available data and its application. In this Chapter, new data was analyzed using the SARMs and CSMs and made public. My contribution to this work herein included the generation and the analysis of SARMs and CSMs.

The detailed study of the SARM-based data mining method, which is used to navigate high-dimensional activity spaces, will be further discussed in Chapter 4.

The predictive approach that helps to prioritize VCs systematically and further enhances the utility of SARMs will be further discussed in Chapter 5.

Chapter 4

Systematic Mining of Analog Series with Related Core Structures in Multi-Target Activity Space

Introduction

Many bioactive compounds have multiple target annotations indicating that compound promiscuity is a general phenomenon.[4] The study of compound-target interactions are crucial to identify the targets against which untested compounds should be tested.

In this study, closely related analog series that are active against multiple targets are identified. The SARM mining method is adapted to extract such series from compound data sets. Here, the method is utilized to identify compound promiscuity patterns among structurally related compounds.

J Comput Aided Mol Des (2013) 27:665–674
DOI 10.1007/s10822-013-9671-5

Systematic mining of analog series with related core structures in multi-target activity space

Disha Gupta-Ostermann · Ye Hu · Jürgen Bajorath

Received: 18 May 2013 / Accepted: 5 August 2013 / Published online: 24 August 2013

Abstract We have aimed to systematically extract analog series with related core structures from multi-target activity space to explore target promiscuity of closely related analogous. Therefore, a previously introduced SAR matrix structure was adapted and further extended for large-scale data mining. These matrices organize analog series with related yet distinct core structures in a consistent manner. High-confidence compound activity data yielded more than 2,300 non-redundant matrices capturing 5,821 analog series that included 4,288 series with multi-target and 735 series with multi-family activities. Many matrices captured more than three analog series with activity against more than five targets. The matrices revealed a variety of promiscuity patterns. Compound series matrices also contain virtual compounds, which provide suggestions for compound design focusing on desired activity profiles.

Keywords Analog series · Core structures · Structural relationships · Compound activity data · Matched molecular pairs · Matching molecular series · Compound series matrix · Biological targets

Introduction

In medicinal chemistry, analog series are usually organized in tables that list R-groups at different sites of a molecular core shared by a given compound series and report corresponding activity values [1]. These tables represent a standard format for SAR analysis. Derivatives of SAR tables have been introduced that organize compound series on the basis of R-group decomposition and display activity data of analogous in a heat map format [2] or in network representations [3]. In addition, approaches that utilize maximum common substructures (MCSs) [4] or scaffold-based compound organization schemes [5–7] are also widely used to represent SAR data. Going beyond the traditional medicinal chemistry focus on single series, methods for the extraction of SAR information from large and heterogeneous compound data sets have been developed in recent years [8]. In this context, SAR matrices have been introduced [9], which capture two or more compound series and display their potency distribution in a colour-coded matrix format. SAR matrices were originally designed to display potency patterns in compound series active against a given target [9]. The SAR matrix data structure utilizes the matched molecular pair formalism [10, 11] for the organization of analog series with related core structures.

There is increasing evidence that many bioactive compounds and drugs specifically interact with multiple targets [12, 13], which extends the traditional single-target focus of medicinal chemistry. Compound promiscuity is intensely studied in pharmaceutical research [14] because it provided the molecular basis of polypharmacology [15]. In recent studies, many compounds with multi-target activities have been identified through data mining [14]. We have been interested in extracting analog series with multi-target activities from currently available active compounds and capturing promiscuity patterns associated with closely related analog series. For systematic mining of such series and the graphical analysis of promiscuity patterns, the SAR matrix data structure [9] has been adapted and further extended.

D. Gupta-Ostermann · Y. Hu · J. Bajorath (✉)
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany
e-mail: bajorath@bit.uni-bonn.de

Materials and methods

Matched molecular pairs

Matched molecular pairs are defined as pairs of compounds that are distinguished by the exchange of a single fragment (substructure) at a specific site [10] *referred to as a chemical transformation* [11]. Distinguishing fragments include R-groups or ring systems and can vary in size. MMPs were systematically calculated using an in-house implementation of the algorithm by Hussain and Rea [11]. All molecules were initially subjected to fragmentation at exocyclic single bonds, so-called single-cuts [11]. The resulting two fragments were stored in an index table as key-value pairs. The larger fragment constituted the key and the corresponding smaller fragment constituted the value. If the two fragments had the same size, each fragment was stored once as a key and the corresponding fragment as a value. If a newly generated key was already contained in the index table, the corresponding value was added to the existing key. For a small compound set shown in Fig. 1a, the generation of the MMP index table is illustrated in Fig. 1b.

Structurally analogous matching molecular series

Matching molecular series (MMS) have previously been introduced as an extension of the MMP concept [16]. *An MMS comprises a series of compounds that share the same core (key) and differ by defined chemical substitutions (values).* In addition, *structurally analogous MMS are defined here as two or more MMS whose core structures (keys) differ at a single site* [16]. For the systematic identification of such structurally related MMS, the keys in the index table were subjected to a second round of fragmentation including all exocyclic single bonds and all combinations of two and three single bonds (so-called dual- and triple-cuts, respectively). Dual-cuts yield three and triple-cuts four fragments. Four-fragment combinations were only retained if they consisted of three fragments with single attachment points and a fragment with three attachment points. Figure 1c illustrates the generation of analogous

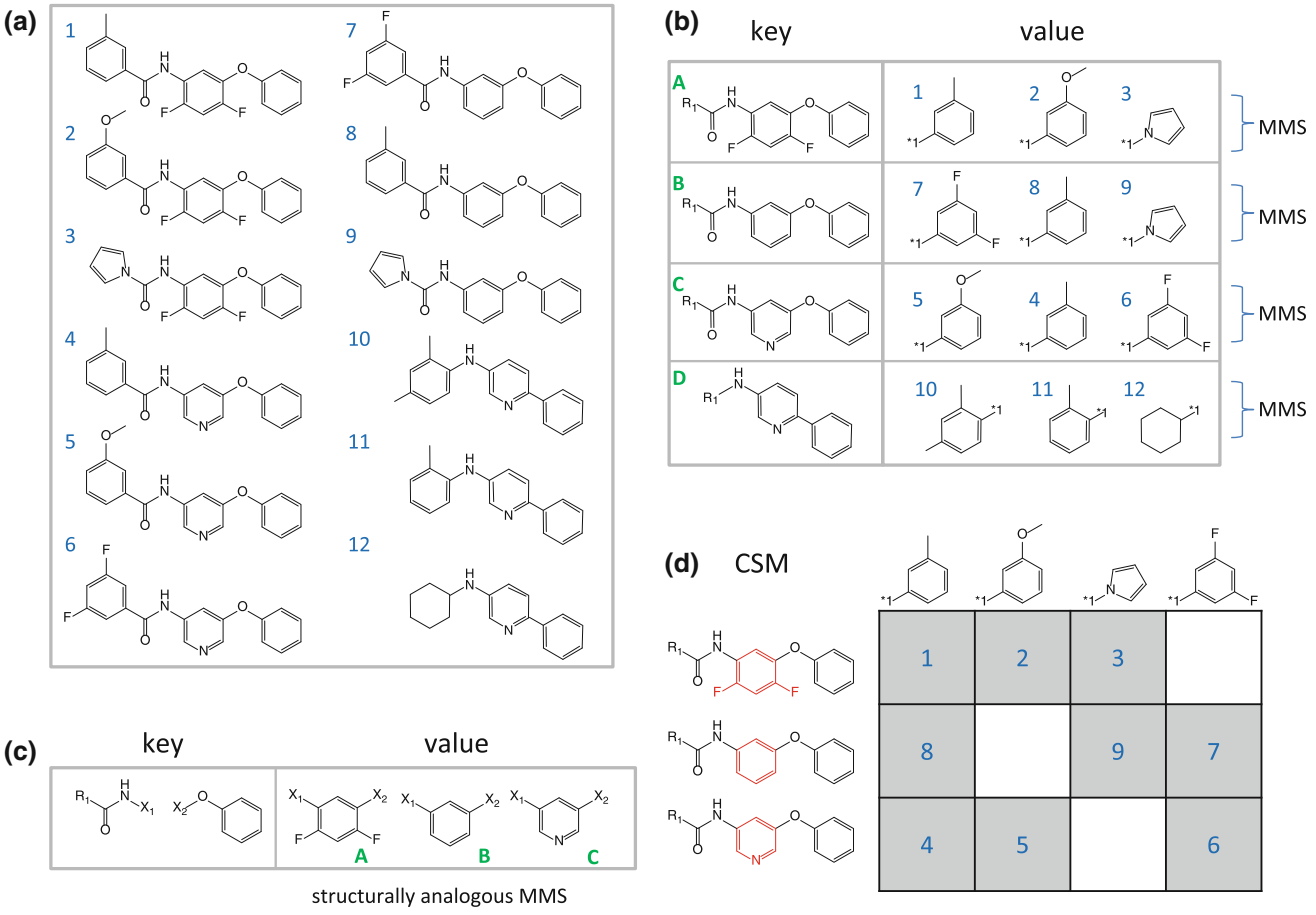


Fig. 1 MMP and matrix generation. **a** A set of 12 test compounds is shown. **b** MMP index table derived from these compounds form four MMS (*A*, *B*, *C* and *D*). **c** Three of these MMS (*A*, *B*, and *C*) that are structurally analogous are revealed by the index table. **d** CSM representing the three structurally analogous MMS. Distinguishing fragments in the structurally related cores are highlighted in *red*

MMS from keys in the original index table (Fig. 1b) through double-cut fragmentation.

Figure 2 shows three exemplary MMS that are further transformed into Bemis–Murcko (BM) scaffolds [17] by removing R-groups from compounds as well as cyclic skeletons (CSKs) [18] by setting all bond orders in BM scaffolds to one and converting all heteroatoms to carbon. Structurally analogous MMS are generally difficult to identify because maximum common substructure methods are not feasible in this case to capture related series. However, they can be systematically identified by applying the MMP-based double-fragmentation procedure described above.

Compound series matrix

MMS that are structurally analogous are combined in a matrix in which rows are formed by keys and columns by corresponding values, as illustrated in Fig. 1d. Each combination of a key and value fragment defines an individual compound. Compounds present in the same row form an individual MMS. The matrix format was adapted from SAR matrices [9] that were designed to analyze SAR information contained in analog series focusing on compound potency, as illustrated in Fig. 3 (left). Here, the matrix structure was utilized for data mining and used to systematically extract all analog series with multi-target activities from currently available active compounds. Accordingly, the matrix structure is termed compound series matrix (CSM). For matrix annotation, all target annotations were collected for available active compounds. Cells in a matrix were color-coded according to the number of targets that compounds were active against, as shown in Fig. 3 (right).

As a consequence of the systematic fragmentation procedure, it is possible that a compound is represented multiple times in matrices as distinct key-value combinations. Therefore, compound redundancy in matrices is minimized as follows:

1. If keys of key-value combinations representing the same compound form substructure relationships, only the larger key fragment is retained.
2. If two MMS sharing the same value fragments yield identical compounds, one of these series is randomly selected and removed.
3. If two different value fragments in an MMS yield the same compounds, the value fragment associated with the smaller number of compounds is removed.

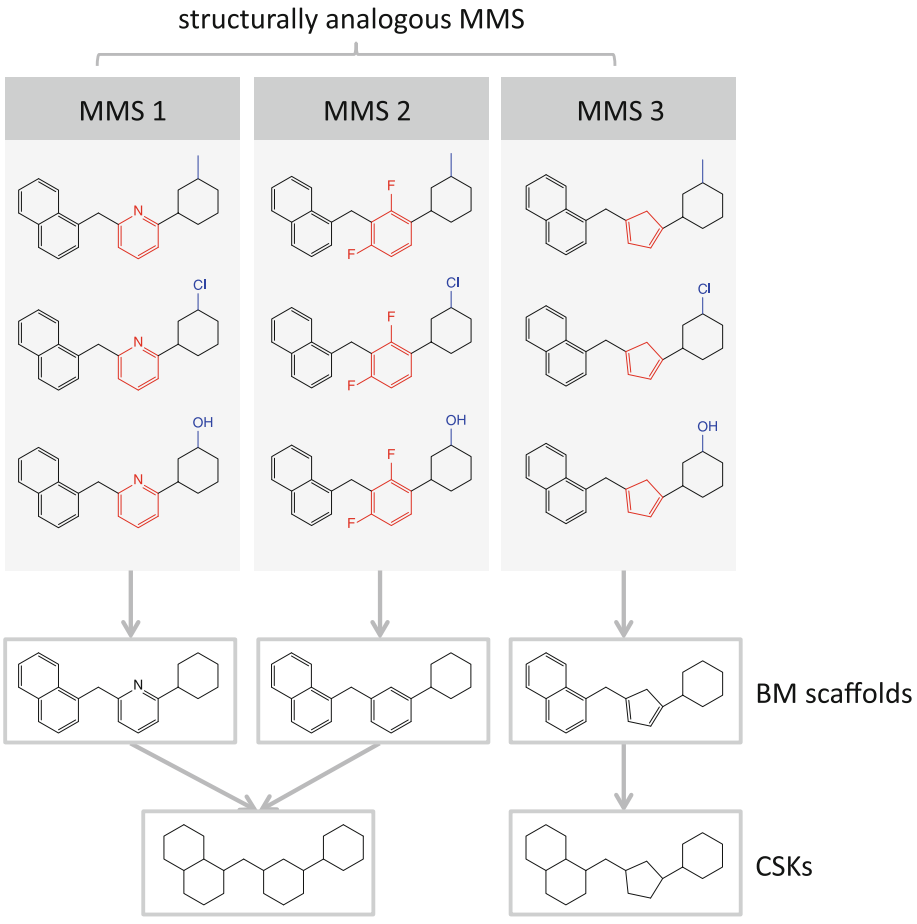


Fig. 2 Structurally analogous matching molecular series. Shown are three exemplary analog series (MMS) with structurally related yet distinct cores. Corresponding substituents are highlighted in *blue* and modifications that distinguish related core structures in *red*. The generation of BM scaffolds and CSKs from analogs is shown at the *bottom*

Matrix coverage

Matrix coverage C defines the proportion of CSM cells that are populated with known active ("real") compounds.

$$C = n/(rows * columns)$$

Here, n gives the number of populated matrix cells and *rows* and *columns* refer to the numbers of rows (keys) and columns (values) forming the matrix. Coverage values range from 0 to 1 and reflect matrix population density for known compounds.

Importantly, by design CSMs consist of known active compounds and virtual compounds that extend/complement structurally related series but are not yet available. As further discussed below, virtual compounds are highly relevant for CSM analysis because they provide compound design suggestions. Accordingly, the matrix coverage parameter provides a measure for virtual compound content of CSMs; the larger C, the larger the population density for real compounds in a given matrix; the smaller C, the sparser the matrix and the larger the population of virtual compounds. Thus, CSMs can be ranked on the basis of matrix coverage. For example, CSMs with larger coverage are prioritized if one searches for extensively explored core structures and/or chemical substitutions. Alternatively, CSMs with smaller coverage are preferred if one searches for opportunities to design new analogs of active compounds or series of interest.

Matrix generation

Keys from original set of MMS comprising of at least three compounds were subjected to a second round of fragmentation to generate CSMs formed by structurally analogous MMS. A CSM was required to contain a minimum of two MMS with at least three compounds each. CSMs comprising compounds that were subsets of other larger matrices were discarded to minimize matrix redundancy. If CSMs consisted of the same set of compounds, the matrix with larger coverage and larger average key size was retained to prioritize analogs with smaller substituent exchanges.

Implementation

Routines required to build the index tables, identify analogous MMS, generate CSMs, and visualize matrices were implemented in Java using OpenEye tool kits [19, 20].

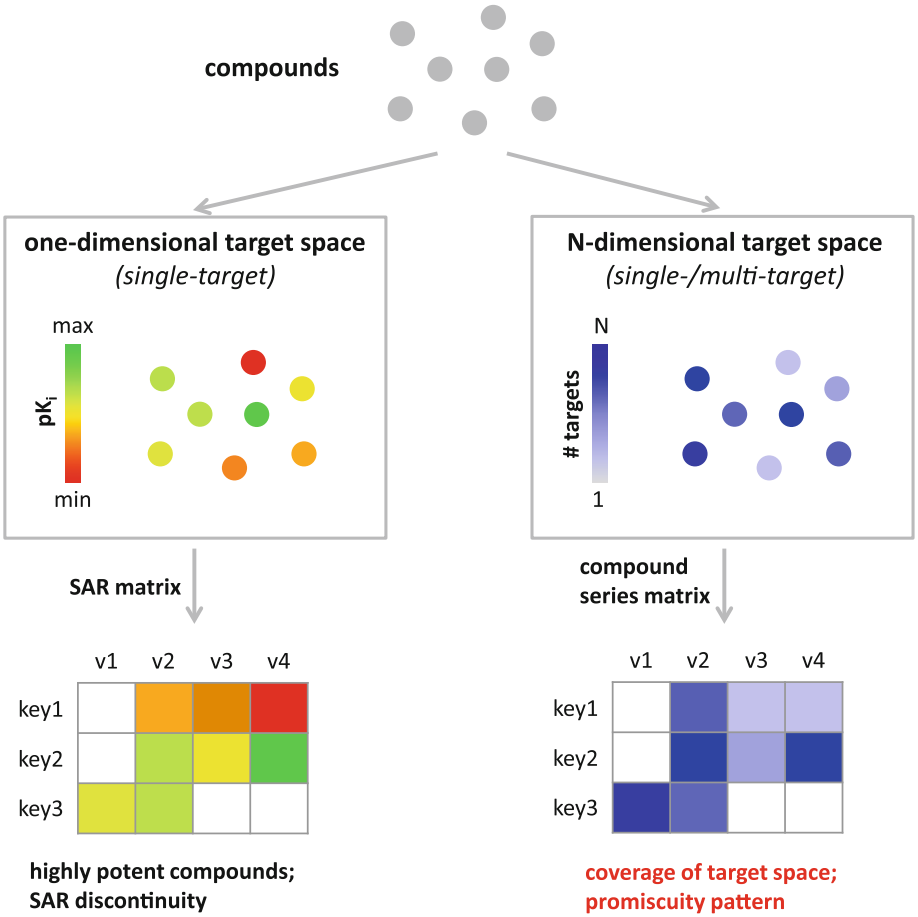


Fig. 3 Matrices of different design concept. A schematic comparison of the SAR matrix method and its CSM extension for mining of multi-target activity space is shown

Data sets

Compounds containing rings and a maximum of 45 acyclic single bonds with activity against human targets were extracted from ChEMBL (release 15) [21]. Compounds with direct target interactions, available K_i values, and a potency of at least 10 μM were selected and their K_i measurement-based target annotations were compiled.

Results and discussion

Study concept

This matrix format was originally introduced as the so-called SAR matrix [9] to monitor potency distributions of analogs active against a given target in series using potency-based color-coding of matrix cells, as illustrated on the left in Fig. 3. SAR matrices were prioritized if they contained many highly potent compounds or displayed SAR discontinuity [9]. Here, we do not utilize the matrix format for SAR analysis but rather adapt the data structure for systematic mining of related compound series in high-dimensional target space, as illustrated on the right in Fig. 3. Transitioning from SAR analysis of single-target compound sets to systematic mining of compound activity data in multi-target space required methodological extensions. In CSMs, an alternative color code was introduced focusing on multi-target activities and efficient compound and matrix redundancy tests were implemented. CSMs were primarily designed to map structurally related analog series in target space, capture multi-target activities associated with closely related compounds, and reveal (visualize) promiscuity patterns. Furthermore, we emphasize another previously unconsidered aspect of matrix mining. Analysis of CSMs makes it possible to bridge between data mining and compound design by focusing on virtual compounds contained in matrices that complement existing series. Depending on their matrix environment and multi-target activity patterns of neighboring compounds, virtual CSM compounds can be prioritized that are likely to display desired activity against selected targets, as further discussed below.

Analog series with related core structures

In Fig. 2, exemplary MMS are shown that are by definition characterized by the presence of closely related yet distinct core structures and corresponding substituents. In this study, we have aimed to systematically identify and characterize such series, which are of particular interest for a comparative SAR analysis and practical medicinal chemistry applications, yet difficult to extract from databases. To these ends, we have adapted an MMP-based dual-fragmentation approach to identify structurally related MMS and organize them in CSMs, as detailed in Methods section.

Mining compound series matrices

On the basis of our selection criteria, 37,850 unique compounds were obtained having a total of 62,784 target annotations. The number of target annotations per compound ranged from one to 35 and the compounds were active against a total of 342 targets.

From the pool of 37,850 compounds, CSMs were systematically generated and a total of 2,337 non-redundant

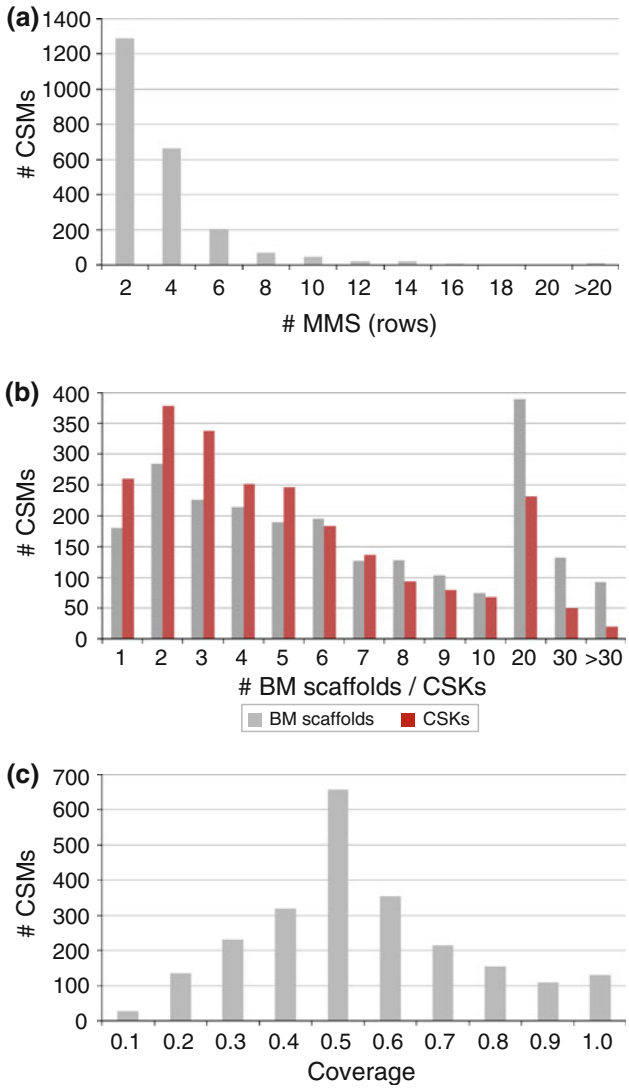


Fig. 4 Compound series matrix content. **a** The number of CSMs containing increasing numbers of MMS (*rows*) is reported in a histogram. **b** The distribution of BM scaffolds (*gray*) and CSKs (*red*) over CSMs is reported. **c** For all CSMs, matrix coverage is reported

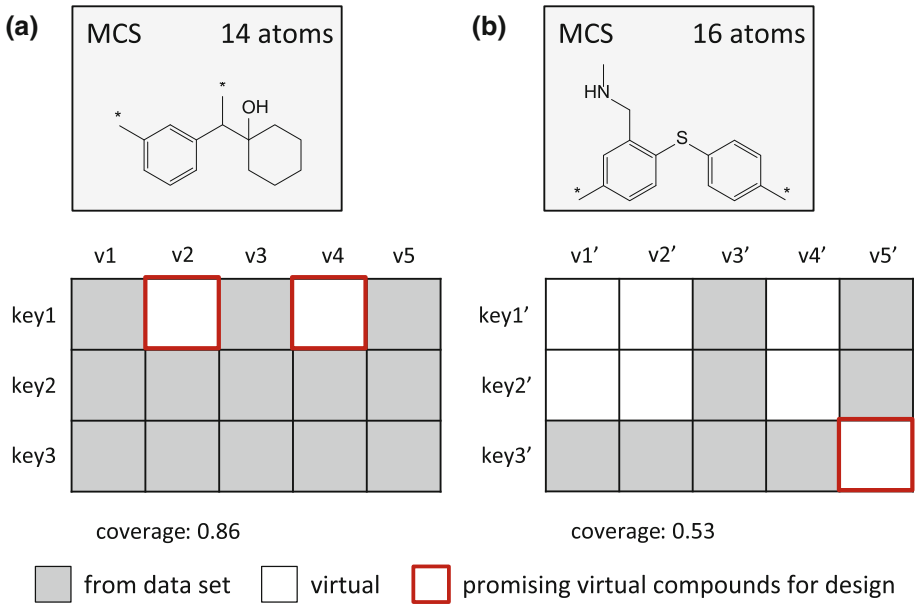


Fig. 5 Compound series matrices with different coverage. Two model CSMs are shown containing the same number of series and substituents but having different matrix coverage. For both CSMs, the maximum common substructure (MCS) is given. Cells are shown in *gray* if they represent data set compounds and *white* if they represent virtual compounds. Cells yielding promising compound design suggestions are highlighted in *red*

matrices were obtained. The structurally related compound series forming each of these matrices are made freely available via the following URL: http://www.lifescience informatics.uni-bonn.de.

Matrix composition

Figure 4a reports the distribution of MMS (rows) over CSMs. The majority of CSMs, i.e., more than 1,300, contained two MMS. In addition, more than 600 and 200 CSMs consisted of three to four and five to six MMS, respectively. Matrices with seven to 14 MMS were also frequently obtained and individual CSMs with more than 20 series were identified. In total, the CSMs represented 5,821 unique MMS. Among these MMS, there were 4,288 series with multi-target activities, 735 of which were active against targets from different families.

Figure 4b reports the distribution of BM scaffolds and CSKs for all CSMs. Among 2,337 CSMs, a total of 180 and 260 CSMs contained compounds that were represented by the same BM scaffolds and CSKs, respectively. The remaining ~89–92 % of CSMs contained multiple core structures yielding different scaffolds and CSKs. Moreover, 689 and 369 CSMs were found to represent more than nine BM scaffolds and CSKs, respectively. On average, CSMs contained compounds with nine scaffolds and six CSKs, respectively.

Figure 4c reports the coverage of CSMs. The distribution displays a peak at a coverage value of 0.5. Hence, in these matrices, 50 % of all theoretically possible analogs were present, providing opportunities for the exploration of additional analogs. CSMs with higher coverage were also frequently observed. Figure 5 shows two model CSMs that

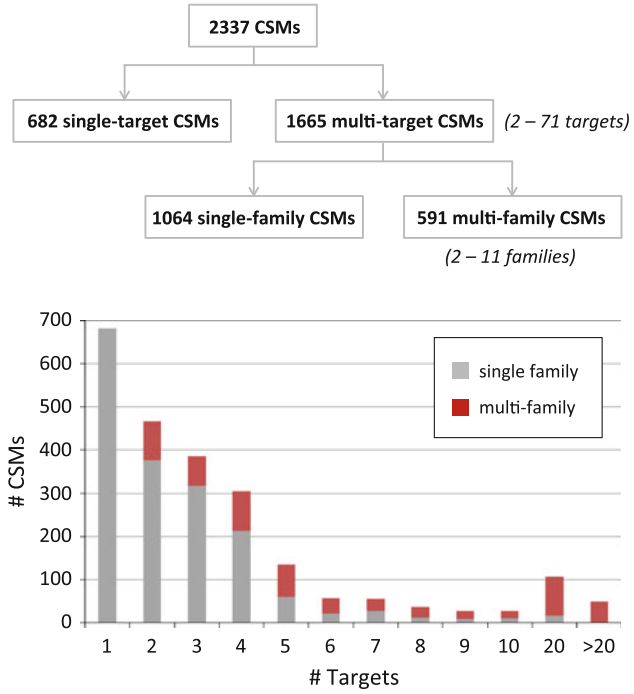


Fig. 6 Mining multi-target activity space. At the top, all generated CSMs are classified according to single- and multi-target activities. At the bottom, the distribution of CSMs annotated with increasing numbers of targets is reported distinguishing between single- (*gray*) and multi-family (*red*) activity

yielded MCSs of comparable size and contained the same number of MMS and value fragments. However, these CSMs were distinguished by different degrees of coverage, i.e., 0.86 versus 0.53. White cells represent virtual compounds combining key and value fragments that occurred in other compounds. Virtual compounds that are adjacent

in the matrix to data set compounds with desired (multi-target) activity (highlighted in Fig. 5) provide promising compound design suggestions.

Target distribution

Major goals of this study have been to mine multi-target activity space using CSMs and identify structurally related analog series with activity against increasing numbers of targets. As reported in Fig. 6, ~29 % of CSMs (682) were composed of compounds with single-target activity. By contrast, a total of 1,064 CSMs identified compounds with activity against multiple targets from the same family. Unexpectedly, 591 CSMs were also found to contain compounds active against targets from two to 11 different families. The target distribution is reported at the bottom of Fig. 6. CSMs containing compounds with reported activity against two to five targets were frequently observed and smaller

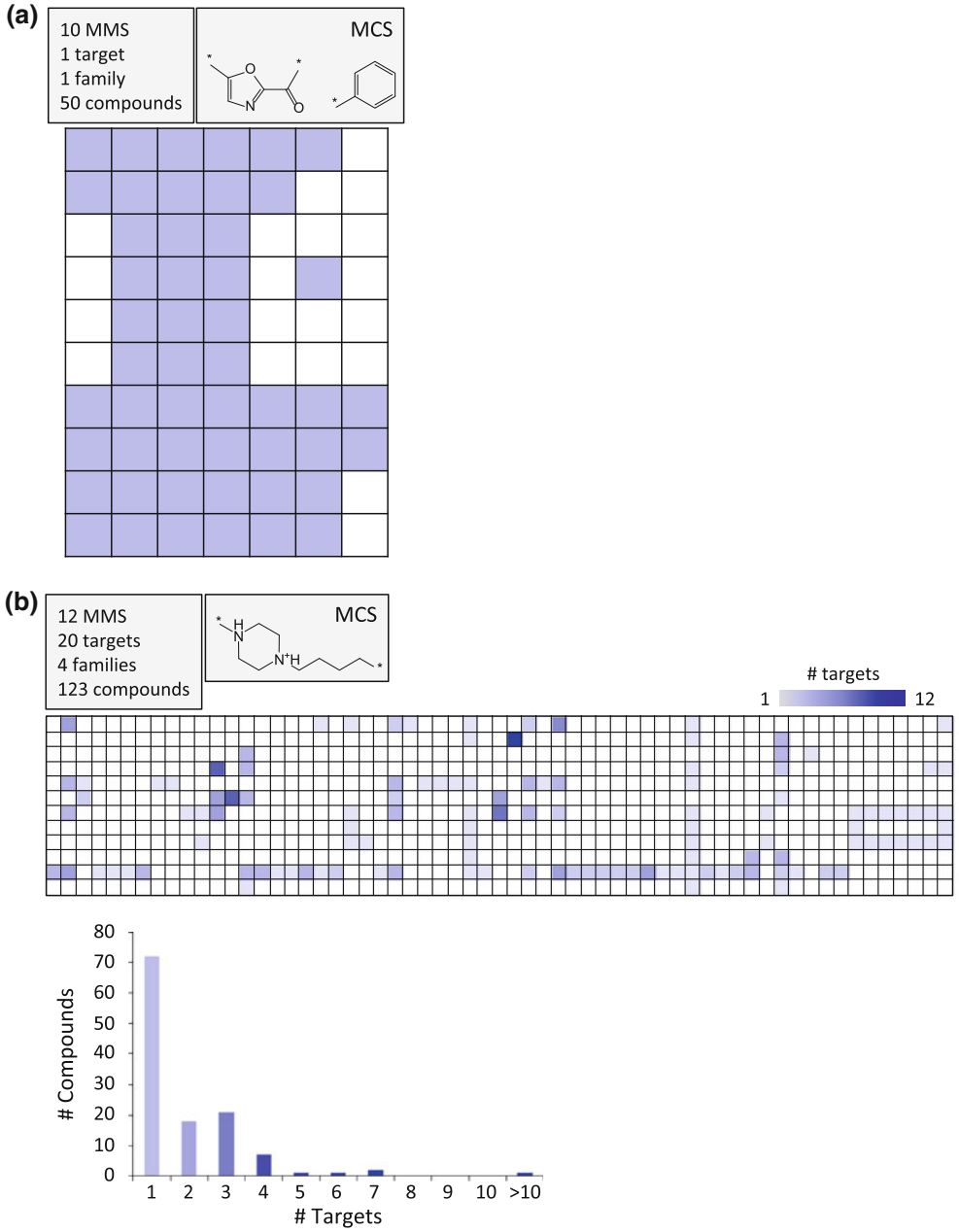


Fig. 7 Single- and multi-target compound series matrices. **a** Single-target CSM with analogs active against anandamide amidohydrolase. The MCS of all series is shown. The CSM label (*top left*) reports the number of MMS, targets the matrix compounds are active against, families these targets belong to, and the total number of analogs. **b** Multi-target CSM with analogs active against 20 targets belonging to four families. Target coverage of analogs is reflected by increasingly *dark blue* shading of cells. The *histogram at the bottom* reports the number of matrix compounds with activity against increasing numbers of targets

numbers of CSMs covered a wide range of up to more than 20 targets. On average, CSMs displayed activity against four targets. In general, the proportion of multi-family CSMs increased with the increasing numbers of targets.

Thus, a large number of multi-target CSMs was identified that captured activity of closely related analog series against targets from different families. The CSM-based identification and structural organization of these series

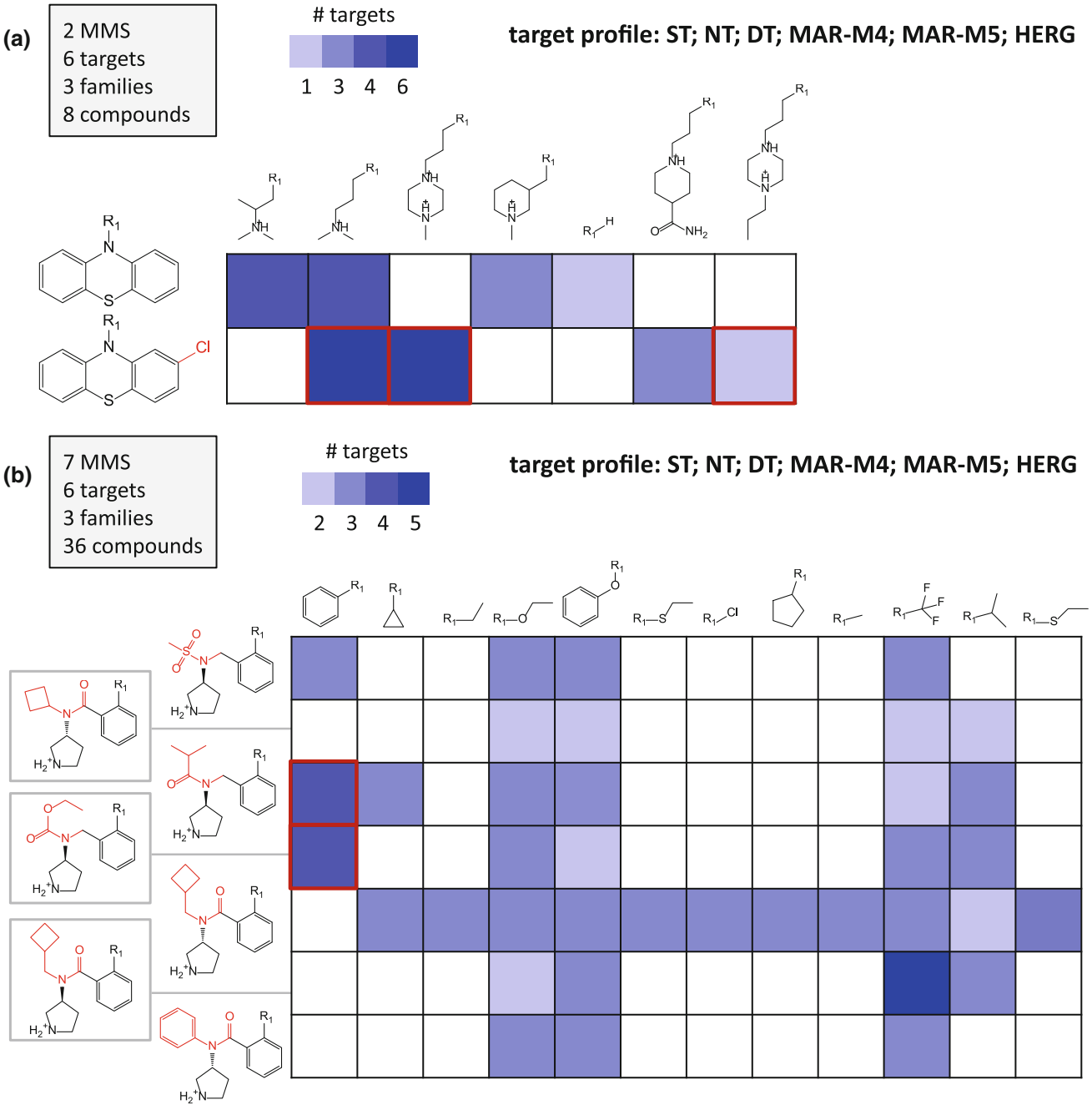


Fig. 8 Compound series matrices representing the same target profile. In **a** and **b**, two exemplary CSMs cover structurally distinct series with activity against the same targets. **a** CSM containing two MMS with eight compounds. **b** CSM containing seven MMS with 36 compounds with activities against 2–5 targets. The representation is according to Fig. 7. Substructures distinguishing the core fragments are highlighted in *red*. The series in these two matrices display different promiscuity patterns. Target abbreviations: *ST* serotonin transporter, *NT* norepinephrine transporter, *DT* dopamine transporter, *MAR-M4* muscarinic acetylcholine receptor M4, *MAR-M5* muscarinic acetylcholine receptor M5, *HERG* HERG ion channel. Cells in the two matrices with reported activity against the HERG anti-target are highlighted in *red*

made it possible to study compound promiscuity patterns in detail.

Promiscuity patterns

In Fig. 7, two exemplary CSMs capturing analog series with single-target activity or target promiscuity are shown. In Fig. 7a, the single-target CSM contained 10 MMS and 50 compounds that were active against anandamide amidohydrolase. By contrast, the larger CSM in Fig. 7b contained 12 series with 123 analogs active against one to 12 targets. The histogram at the bottom of Fig. 7b reports the distribution of target annotations for all analogs, revealing a subset of 51 promiscuous compounds with activity against increasing numbers of targets. In total, the compounds were active against 20 unique targets belonging to four different families. Hence, this example illustrates that CSMs enable the detection of progressive target promiscuity patterns among analogs belonging to closely related series.

Matrices representing the same target profile

As described above, multi-target CSMs were frequently identified. In Fig. 8, two exemplary CSMs with different core structures are shown that contain two and seven MMS, respectively. These series consisted of different numbers of analogs with activity against varying numbers of six targets from three different families. For the majority of these series, closely related analogs were found to be active against overlapping yet distinct targets. These CSMs had very different matrix coverage, i.e., 0.57 (Fig. 8a) versus 0.29 (Fig. 8b). Thus, the CSM in Fig. 8b provided more opportunities to design analogs and "fill" the matrix. However, compounds with HERG anti-target activity were also found in these CSMs (highlighted in Fig. 8). Hence, these compounds point at likely liabilities associated with individual series, which would suggest to carefully investigate potential anti-target activities of closely related analogs captured in these matrices.

Mapping of virtual matrix compounds to drugs

From the 2,337 CSMs reported herein, all virtual compounds were extracted and mapped to 6,081 approved and experimental drugs assembled from DrugBank [22]. A total of 48 drugs were found to match virtual compounds derived from 25 different matrices. Most of these drugs were annotated with targets that overlapped with the target profiles of the corresponding matrices. Hence, virtual matrix compounds are a potential source of interesting drug (-like) molecules.

Conclusions

In this study, we have searched for closely related analog series with multi-target activities by applying the CSM concept and described promiscuity patterns emerging from these series. CSMs represent multiple structurally analogous series and closely related virtual compounds in a well-defined manner. Moreover, the matrix data structure was used here for systematic compound data mining and the exploration of multi-target activity space on the basis of currently available bioactivity data. Virtual analogs adjacent to compounds with desired activities yield design suggestions. We have identified 4,288 series with diverse multi-target and 735 series with multi-family activities. The identification of these series in CSMs and their structural organization makes it possible to analyze promiscuity patterns at the level of structurally related analogs. All CSMs are made freely available to provide a basis for further analysis of analog series with multi-target activities.

References

1. Wermuth CG (ed) (2008) The practice of medicinal chemistry, 3rd edn. Academic Press, San Diego
2. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS (2007) SAR maps: a new SAR visualization technique for medicinal chemists. J Med Chem 50(24):5926–5937
3. Wassermann AM, Bajorath J (2012) Directed R-group combination graph: a methodology to uncover structure–activity relationship patterns in series of analogs. J Med Chem 55(3):1215–1226
4. Cho SJ, Sun Y (2008) Visual exploration of structure–activity relationship using maximum common framework. J Comput Aided Mol Des 22(8):571–578
5. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 47(1):47–58
6. Agrafiotis DK, Wiener JJ (2010) Scaffold explorer: an interactive tool for organizing and mining structure–activity data spanning multiple chemotypes. J Med Chem 53(13):5002–5011
7. Gupta-Ostermann D, Hu Y, Bajorath J (2012) Introducing the LASSO graph for compound data set representation and structure–activity relationship analysis. J Med Chem 55(11):5546–5553
8. Wawer M, Lounkine E, Wassermann AM, Bajorath J (2010) Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov Today 15(15–16):631–639
9. Wassermann AM, Haebel P, Weskamp N, Bajorath J (2012) SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. J Chem Inf Model 52(7):1769–1776
10. Kenny PW (2005) Sadowski J (2005) Structure modification in chemical databases. In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, Weinheim, Germany, pp 271–285
11. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50(3):339–348

12. Knight ZA, Lin H, Shokat KM (2010) Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10(2):130–137
13. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24(7):805–815
14. Hu Y, Bajorath J (2013) Compound promiscuity: what can we learn from current data? Drug Discov Today 18(13–14):644–650
15. Boran AD, Iyengar R (2010) Systems approaches to polypharmacology and drug discovery. Curr Opin Drug Discov Dev 13(3):297–309
16. Wawer M, Bajorath J (2011) Local structural changes, global data views: graphical substructure–activity relationship trailing. J Med Chem 54(8):2944–2951
17. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39(15):2887–2893
18. Xu YJ, Johnson M (2002) Using molecular equivalence numbers to visually explore structure features that distinguish chemical libraries. J Chem Inf Comput Sci 42(4):912–926
19. OEChem TKV (2013) April, Open Eye Scientific Software Inc, Santa Fe, New Mexico
20. OEDepict TKV (2013) April, Open Eye Scientific Software Inc, Santa Fe, New Mexico
21. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
22. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 40:D1035–D1041

Summary

In this study, the SARM mining method was used to identify structurally related compound series active across multiple targets and to analyze promiscuity patterns among them. 2,337 compound series matrices of varying structural and activity composition were identified from high-confidence compound activity data. These compound series matrices identified 4,288 series representing multi-target and 735 series representing multi-family activities. Furthermore, the compound series matrices can also be used to create virtual compounds that provide opportunities to guide the design of novel compounds with desired multi-target activities. My contribution to this study has been the design and the analysis of the compound series matrices.

The following study in Chapter 5 describes a novel method for systematically predicting the activity of virtual compounds from SARMs.

Chapter 5

Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices

Introduction

Virtual compounds resulting from SARMs are difficult to prioritize based on just visual inspection of the matrices. Here, we introduce a methodology to predict and prioritize virtual compounds by systematically utilizing the activity information from their neighborhoods in SARMs. This is done by considering individual contributions of cores and substituents of compounds in qualifying neighborhoods. This approach is well suited for potency prediction during compound optimization considering multiple analog series. Benchmark studies for method evaluation are carried out on six large data sets of G protein coupled receptor antagonists.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

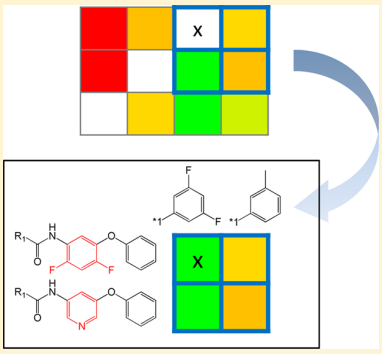
Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices

Disha Gupta-Ostermann,[†] Veerabahu Shanmugasundaram,[‡] and Jürgen Bajorath*,[†]

[†]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, North Rhine-Westphalia, Germany

[‡]Computational Analysis & Design, Center of Chemistry Innovation & Excellence, Worldwide Medicinal Chemistry, Pfizer, Groton, Connecticut 06340, United States

ABSTRACT: The SAR matrix data structure organizes compound data sets according to structurally analogous matching molecular series in a format reminiscent of conventional R-group tables. An intrinsic feature of SAR matrices is that they contain many virtual compounds that represent unexplored combinations of core structures and substituents extracted from compound data sets on the basis of the matched molecular pair formalism. These virtual compounds are candidates for further exploration but are difficult, if not impossible to prioritize on the basis of visual inspection of multiple SAR matrices. Therefore, we introduce herein a compound neighborhood concept as an extension of the SAR matrix data structure that makes it possible to identify preferred virtual compounds for further analysis. On the basis of well-defined compound neighborhoods, the potency of virtual compounds can be predicted by considering individual contributions of core structures and substituents from neighbors. In extensive benchmark studies, virtual compounds have been prioritized in different data sets on the basis of multiple neighborhoods yielding accurate potency predictions.



INTRODUCTION

The conventional approach to the exploration of structure−activity relationships (SARs) in medicinal chemistry is the organization of compound series in R-group tables.[1] Such tables record R-groups at different substitution sites of a core structure common to a compound series and report activity values of analogs. With the aid of R-group tables, new compounds are designed on the basis of medicinal chemistry experience and intuition. Simple comparisons of molecular graphs typically provide a basis for studying similarities and differences between active compounds and for identifying key structural features that correlate with activity.[1,2] In addition to standard R-group tables, other hierarchical structural organization schemes have also been introduced for SAR exploration that are based upon maximum common core structures[3] or molecular scaffolds.[4,5]

Computationally, the design of new compounds on the basis of series information has traditionally been supported by Quantitative SAR (QSAR) analysis methods.[6,7] Following established QSAR principles, mathematical models are derived using descriptors of chemical structure and properties to predict the effects of chemical substitutions on compound potency. In addition, in recent years, graphical methods have been increasingly employed to visually analyze SARs, often going beyond individual compound series.[8,9] However, SAR visualization methods are generally descriptive in nature and do not directly lead to compound predictions.

In addition to scaffold-based approaches and graphical methods, the matched molecular pair (MMP) represents another emerging concept in medicinal chemistry[10,11] that is highly relevant for SAR analysis. An MMP is generally defined as a pair of compounds that differ only by a structural change at a single site, i.e., the exchange of a pair of substructures.[10] Applying the MMP formalism, a chemical change relating two compounds to each other can be directly associated with changes in different types of molecular properties including biological activity.[11,12] Even for large compound data sets, MMPs can be efficiently generated algorithmically,[13] hence providing a comprehensive pairwise substructure-based organization scheme. For SAR analysis, the MMP concept has been further extended by introducing matching molecular series (MMS),[14] which are defined as a series of compounds that only differ by chemical changes at a single site. MMS can be utilized to construct SAR networks and follow substructure changes within and across compound series that lead to SAR progression.[14]

As a data structure that combines intuitive SAR visualization with the MMS concept and further extends this concept, the SAR Matrix (SARM) has been recently introduced.[15] SARM provides an R-group table-like high-resolution view of compound data sets that are automatically organized into structurally related series, as illustrated in Figure 1. It consists of

Received: January 23, 2014
Published: March 4, 2014

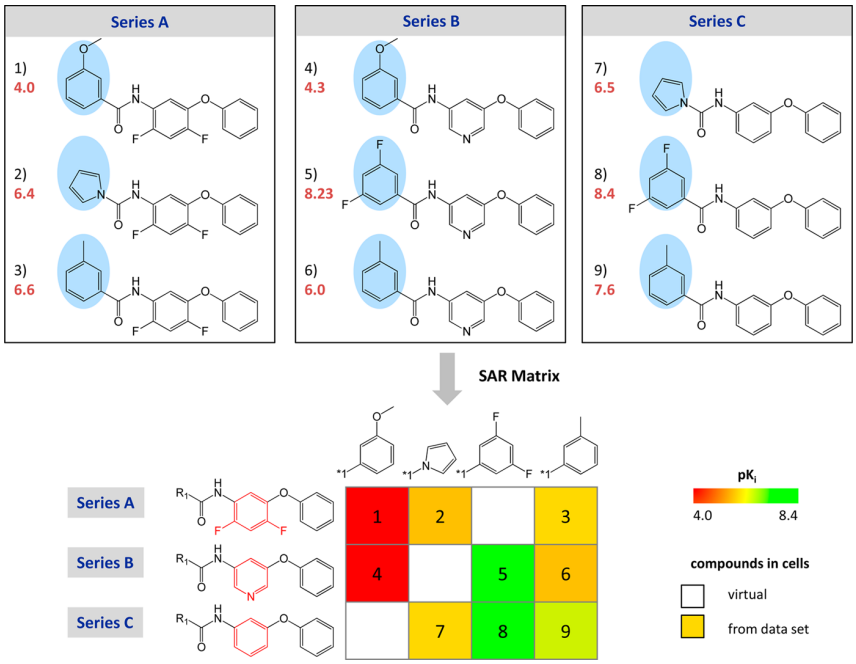


Figure 1. SAR matrix. Three model compound series (A, B, and C) containing three compounds each are shown with their respective pK_i values (red). Compounds in a series share a common core structure and differ by substitutions at a single site (highlighted in blue). The three series contain structurally related cores (bottom left; substructure differences between cores are highlighted in red). The SAR matrix is generated by combining structurally related analog series. Rows and columns represent compounds that share the same core and substituent, respectively. In each cell, the combination of a core and a substituent defines a unique compound. Compounds present in the data set are indicated by filled cells that are color-coded according to potency using a continuous spectrum from red (low potency) over yellow (intermediate) to green (high). In addition, empty cells indicate virtual compounds.

"real" data set and structurally analogous virtual compounds. Herein, we introduce an approach to predict virtual compounds and their actual potency values from SARMs on the basis of compound neighborhood information.

In the following sections, we describe the SARM data structure, introduce the prediction methodology, and report the results of systematic potency predictions of virtual compounds from SARMs of different data sets.

SAR MATRIX DESIGN PRINCIPLES

SARM generation involves a dual compound fragmentation scheme resulting in two-level MMP generation.[15] In the first step, compounds are subjected to MMP fragmentation applying the algorithm by Hussain and Rea,[13] leading to the generation of an index table with large key fragments (core structures) and smaller value fragments (substituents). In the second step, the cores in the index table are subjected to an additional round of fragmentation. The resulting fragments are stored in a new index table with the larger fragment as the key, analogously to the first step. Hence, this dual fragmentation scheme identifies compound series having structurally related cores termed "structurally analogous MMS" (A_MMS),[15] which further extends the MMS concept. In Figure 1, three series A, B, and C are shown, each of which contains a common core structure and differs at a single site (highlighted in blue). The second fragmentation step reveals that the core structures of these three series are related and also only differ at a single site (red substructures at the bottom). Hence, series A, B, and C form an A_MMS, which is represented in a unique SARM. As illustrated in Figure 1, the SARM is filled with structurally related cores resulting from the second MMP generation step and corresponding substituents resulting from the first step. Each row in the matrix contains an individual compound series, and each cell represents an individual compound (a unique combination of a key and value fragment). Hence, by design, SARMs are 2D matrices. The series comprising a SARM typically have overlapping yet distinct sets of substituents, giving rise to combinations of real (filled cells) and virtual compounds (empty cells). As shown in Figure 1, a continuous color spectrum is applied to capture the potency information of real compounds. Alternatively, ligand efficiency values can also be used.

Typically, a large compound data set yields multiple or many SARMs. Depending on the algorithmic fragmentation scheme,[13] single-cut matrices (i.e., one exocyclic bond in a compound is systematically deleted to yield key and value fragments), dual-cut (two exocyclic bonds are simultaneously deleted), and triple-cut matrices (three exocyclic bonds are deleted) are separately generated.[15] The resulting SARMs provide a high-resolution organization of a compound data set that accounts for all possible structural relationships between compound series. In addition, virtual compounds contained in a SARM represent as of yet unexplored key-value combinations and hence provide immediate suggestions for new analogs. As such, virtual compounds systematically captured by SARMs can be rationalized as a "chemical space envelope" that delineates regions surrounding a given data set in chemical space. The appearance of SARMs is akin to R-group tables, which makes them easily accessible to medicinal chemists and permits simultaneous exploration of structurally related compound series. In addition to SAR analysis, the SARM data structure has

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

also been adapted to navigate multitarget activity space and study promiscuous compounds.[16]

The set of SARMs representing a larger data set typically contains many virtual compounds (as further detailed below). This raises the question which of these virtual compounds might be prioritized as design suggestions? SARM in its original conceptualization did not provide selection schemes for virtual compounds. Therefore, we introduce a compound and potency prediction approach to further improve the utility of SARMs for compound prioritization and design, as described in the following.

COMPOUND PREDICTION METHOD

A possible criterion for the selection of virtual compounds from SARMs is their proximity to real data set compounds for which target-specific activity information is available. The underlying idea is that close proximity of a virtual compound to multiple active compounds (i.e., the presence of close structural relationships) increases the probability that this virtual compound might also be active relative to other virtual compounds for which no active neighbors are available. This leads to the notion and assessment of defined compound neighborhoods (NBHs).

Neighborhood Definition. A preferred NBH of a virtual compound is defined as one that consists of three adjacent real compounds, as illustrated in Figure 2a. Because all real compound entries in a given column or row containing a virtual compound are equivalent as potential neighbors, many different three-compound NBHs might exist for a given virtual compound. In general, an NBH of a virtual compound is formed by any combination of three compounds present in the data set of which one shares the core of the virtual compound, the second its substituent, and the third the different core and substituent of these two neighbors. An NBH of this composition is depicted in Figure 2a. If one of these three neighboring compounds is missing, the NBH is incomplete and does not qualify for further analysis. NBHs within the same SARM consist of structurally closely related data set compounds, whereas NBHs from different SARMs might consist of structurally dissimilar compounds.

A virtual compound might already be prioritized based on the number of qualifying NBHs, which establish close structural relationships to known active compounds. However, given the composition of so-defined NBHs, one can go a step further and attempt to predict the potency of a given virtual compound on the basis of potency information of its neighbors.

Neighborhood-Based Potency Prediction. The presence of a three-compound NBH of a virtual compound makes it possible to predict its potency based on a local model utilizing the additivity assumption underlying Free-Wilson analysis,[17,18] as illustrated in Figure 2b. Following this approach, the potency of virtual compound X can be predicted from the sum of logarithmic potencies of compounds A and C, which share the same substituent and core with compound X, respectively, minus the logarithmic potency of compound B, which represents the combination of the core structure and substituent of compounds A and C, respectively:

$$pK_i(X) = pK_i(A) + pK_i(C) - pK_i(B)$$

Hence, subtracting the potency of B from the sum of potencies of A and B corrects for the contributions of the structural fragments of A and C that are not contained in X. Hence, in essence, the local models represent "mini-QSAR"

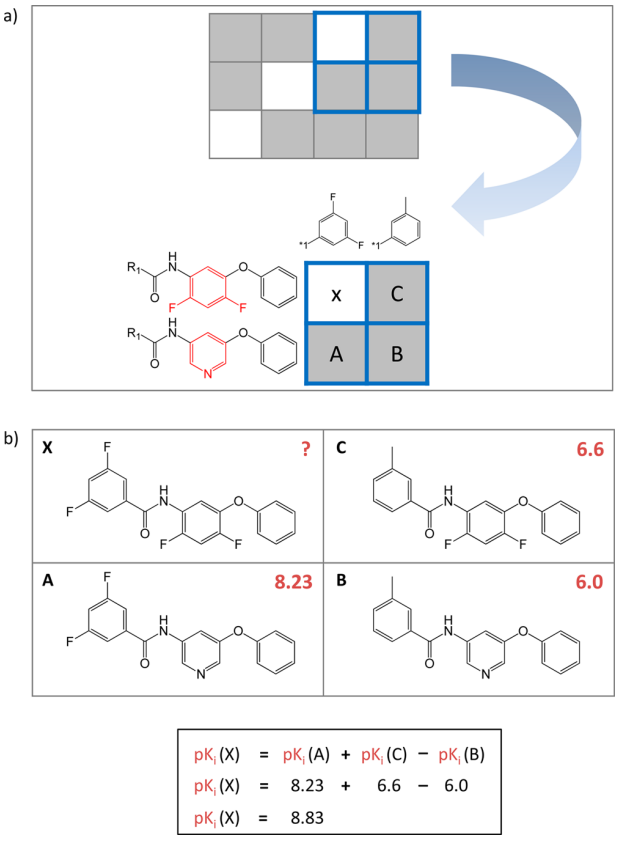


Figure 2. Compound neighborhood and potency prediction. (**a**) The NBH of virtual compound X is marked in blue in the given matrix and shown in more detail at the bottom. Cores and substituents comprising the three compounds forming the NBH are depicted. Compounds A and C share the same substituent and core with X, respectively, and the third neighbor B combines the core and R-group of neighboring compounds A and C, respectively. (**b**) Compound structures and pK_i values (red) of "real" neighbors are given, and calculations are reported to predict the potency of virtual compound X.

models. These Free-Wilson-type predictions are applicable to any qualifying compound NBH. A virtual compound often (but not always) occurs in multiple SARMs obtained for a given data set. Therefore, all qualifying unique NBHs of each virtual compound are systematically identified and subjected to potency predictions, as illustrated in Figure 3. The number of qualifying NBHs might be increased in subsequent iterations by considering compounds with predicted potency for NBH definition (e.g., by generating NBHs consisting of data set and previously predicted compounds). However, prediction accuracy would likely be reduced in such cases.

Comparison with QSAR. The NBH-based prediction concept introduced herein represents a local prediction approach. Local Free-Wilson-type predictions are carried out for all matrices with qualifying NBHs for given virtual compounds. The method does not utilize chemical descriptors such as Fujita-Hansch QSAR or models based on training sets such as Fujita-Hansch and Free-Wilson QSAR. The NBH-based prediction scheme exploits characteristic features of SARMs. By design, all compounds contained in an individual SARM are structurally closely related and hence qualify for QSAR-like predictions of virtual compounds within the same

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

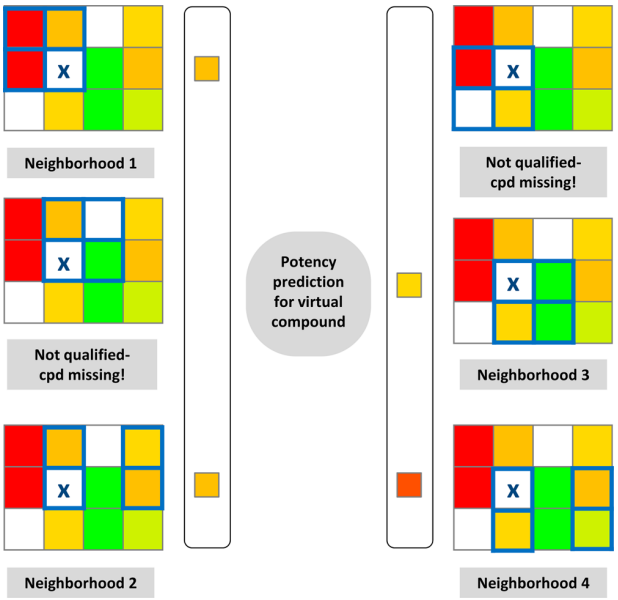


Figure 3. Neighborhood mining. For virtual compound X, the set of all NBHs (outlined in blue) in an SAR matrix is identified and neighborhoods qualifying for potency predictions are determined. In this example, four qualifying NBHs (1 to 4) are identified, and predicted activities are indicated by squares color-coded according to the spectrum in Figure 1.

matrix. In addition, NBH-based predictions do not depend on the population density of data set compounds in a given SARM, only on the presence of local NBHs. This represents an important difference compared to standard QSAR modeling. To predict virtual matrix compounds, QSAR models would need to be built for individual SARMs and would require the presence of sufficient numbers of data set compounds to assemble meaningful training sets.

Applicability Domain. The prediction methodology generally aims to prioritize virtual compounds as candidates for synthesis. In the context of our analysis, preferred virtual compounds would in principle be characterized by the presence of many different NBHs that yield consistent potency predictions. Most interesting would be compounds with consistently predicted high potency. The consistency of potency predictions can be assessed by calculating the standard deviation (SD) of multiple independent predictions; low SD values indicate consistent predictions. It should be noted that low SDs do not necessarily correlate with high prediction accuracy (i.e., predictions can be consistently incorrect). By contrast, high SDs are indicative of inconsistent predictions. Such predictions are likely to occur if virtual compounds have many structurally analogous neighbors with large differences in potency. In this case, virtual compounds map to regions of SAR discontinuity[8] including activity cliffs,[8,19] and local QSAR-type predictions are no longer applicable in a meaningful way.[19,20] Discontinuous compound NBHs yielding predictions with high SDs might still be attractive for the analysis of specific compound environments in SARMs, as further discussed below. However, they fall outside the applicability domain of QSAR modeling and are hence deprioritized in our systematic potency predictions (even if predictions from individual discontinuous NBHs would indicate high compound potency).

IMPLEMENTATION, DATA SETS, AND CALCULATION SETUP

Routines to generate SARMs were implemented with the aid of the OpenEye chemistry toolkit,[21] and potency prediction routines were implemented in Java. Statistical analyses were carried out with R.[22]

For benchmark calculations, six large sets of different G protein coupled receptor antagonists were extracted from ChEMBL (release 15)[23] for which K_i values were available as potency measurements. The targets, sizes, and potency ranges of these data sets are reported in Table 1. These data sets consisted of 1103−1850 compounds that formed SARMs. For all data sets, all possible SARMs were calculated, producing between 655 and 1109 matrices per set, as also reported in Table 1.

Table 1. Data Set Statistics[a]

target name	TID	# SAR matrix	# cpds	pK_i range
dopamine D2 receptor	217	700	1419	3.0 to 10.2
adenosine A1 receptor	226	1104	1825	4.2 to 10.5
adenosine A2a receptor	251	957	1850	4.0 to 11.0
adenosine A3 receptor	256	1109	1547	4.1 to 11.0
melanocortin receptor 4	259	669	1103	3.9 to 9.4
histamine H3 receptor	264	655	1718	4.4 to 10.5

[a]For each compound data set, the ChEMBL target ID (TID), the number (#) of SAR matrices and compounds, and their pK_i range are reported.

Following SARM generation, the following prediction protocol was established:

(i) One third of the compounds were randomly removed from each data set, and all corresponding matrix positions were converted into virtual cells.

(ii) All removed ("pseudovirtual") compounds were recorded as potential targets for predictions (original virtual compounds contained in unmodified SARMs could not be predicted in benchmarking).

(iii) For all "virtualized" cells from potential prediction targets, qualifying compound NBHs were systematically determined.

(iv) Compounds having at least three unique NBHs across SARMs were selected, and potency predictions were carried out on the basis of each NBH. Then, the consistency of predictions was assessed.

RESULTS OF SYSTEMATIC PREDICTIONS

Following the protocol reported above, pseudovirtual compounds having qualifying NBHs were identified across SARMs generated for different data sets and systematic potency predictions were carried out.

Prioritization of Pseudovirtual Compounds. Between 76 and 179 pseudovirtual compounds with at least three qualifying NBHs were identified in SARMs obtained for the different data sets, as reported in Table 2. For all of these compounds, potency predictions were carried out for individual NBHs, and SDs of the predictions were calculated. SD values falling into the first quartile of the distributions of all SDs within a data set were classified as low SDs, and the corresponding pseudovirtual compounds were designated low SD (L_SD) compounds. Figure 4 reports the distributions of SDs of predictions for all L_SD compounds. Median SD values were close to 0.5 pK_i units for all data sets. Hence, alternative

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

Table 2. Pseudovirtual Compounds and Potency Predictions[a]

TID	# pseudovirtual cpds	# prioritized cpds	# L Sd cpds
217	473	91	23
226	608	179	45
251	616	126	32
256	515	146	37
259	367	76	19
264	572	134	34

[a]For each data set (indicated by TID according to Table 1), the total number (#) of pseudovirtual compounds (i.e., data set compounds removed from matrices), number of prioritized virtual compounds with at least three qualifying compound neighborhoods, and number of prioritized virtual compounds yielding potency predictions with low standard deviations (L_SD) are reported.

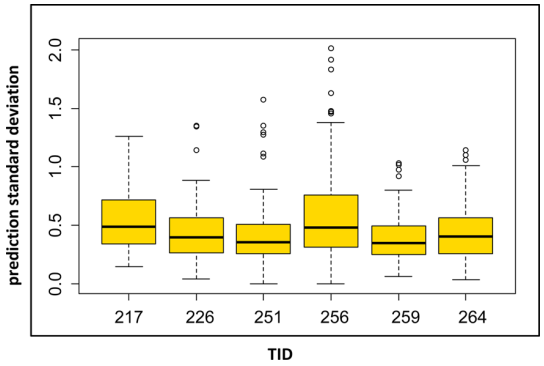


Figure 4. Standard deviations of predictions. Box plots report the distributions of standard deviations (*y*-axis) of potency predictions for single pseudovirtual compounds having at least three qualifying NBHs. Data sets are indicated by TIDs according to Table 1 (*x*-axis).

potency predictions for these compounds generally fell well within 1 order of magnitude. Between 19 and 45 L_SD compounds were obtained from ∼400 to ∼600 potential candidates for the different data sets, as also reported in Table 2. By increasing the SD threshold beyond the first quartile of the global distribution, additional compounds can be obtained. However, for our proof-of-principle investigation, the number of L_SD compounds in Table 2 was readily sufficient. On the basis of our above considerations, L_SD compounds fell within the applicability domain of local prediction models, and their potency predictions were further analyzed.

Prediction Performance. In Figure 5, distributions of prediction errors are reported for L_SD compounds from all data sets. With very few exceptions, prediction errors associated with potencies averaged over NBHs were well within 1 order of magnitude with a median of only ∼0.25 order of magnitude. This is further illustrated in Figure 6 by a comparison of observed vs predicted potency values across the different potency ranges captured by the data sets. Regardless of the potency ranges, observed and predicted potency values were consistently very similar. Hence, for prioritized compounds, accurate potency predictions were obtained.

As control calculations, we also generated Free-Wilson QSAR models using R[22] for individual SARMs to predict L_SD compounds they contained (as discussed above). In these cases, data set compounds contained in a SARM were used as a training set for deriving a matrix-based Free-Wilson model. We then compared prediction errors of these Free-

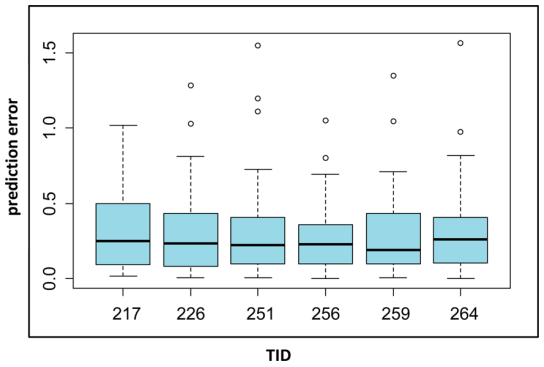


Figure 5. Prediction errors. Box plots report the distributions of prediction errors for individual pseudovirtual compounds yielding potency predictions with low standard deviations (L_SD compounds). Prediction errors were calculated on the basis of averaged predicted pK_i values and are reported on the *y*-axis as Δ pK_i units relative to observed compound potencies. On the *x*-axis, data sets are given.

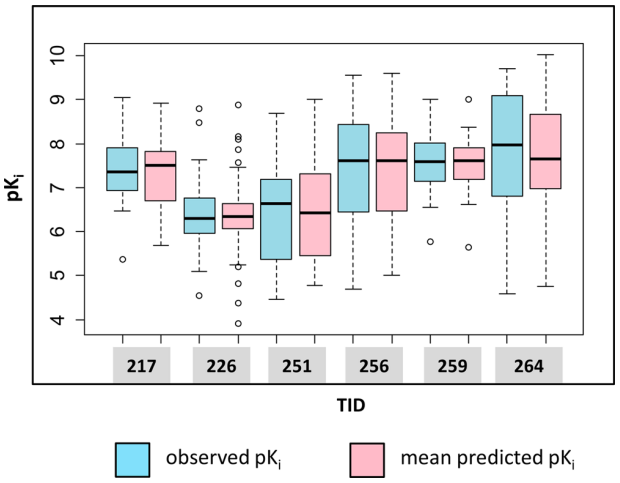


Figure 6. Observed vs predicted potency values. Box plots report the distribution of observed (blue) and mean predicted pK_i values (pink) for L_SD compounds for each data set.

Wilson models for L_SD compounds with NBH-based predictions originating from the same SARM. In test calculations on individual SARMs, we observed very similar (low) prediction errors for matrix-based Free-Wilson and NBH-based predictions, indicating that NBH information was sufficient to achieve accurate predictions for L_SD compounds.

Neighborhood Frequency. L_SD compounds can also be ranked according to the number of NBHs qualifying for prediction. Compounds frequently predicted to have high activity are generally the most interesting candidates for further exploration. In Figure 7, potency predictions and the frequency of NBHs are compared for individual L_SD compounds. Compound data points are scaled in size according to prediction errors. Depending on the data sets, individual compounds were found to have a maximum of ∼10 to ∼30 NBHs across SARMs. If prediction errors exceeding 1 order of magnitude were detected, they were exclusively observed for virtual compounds having only three or four NBHs. With further increasing numbers of NBHs, predictions became increasingly accurate (and prediction errors typically very small). These findings reflect a general relationship between

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

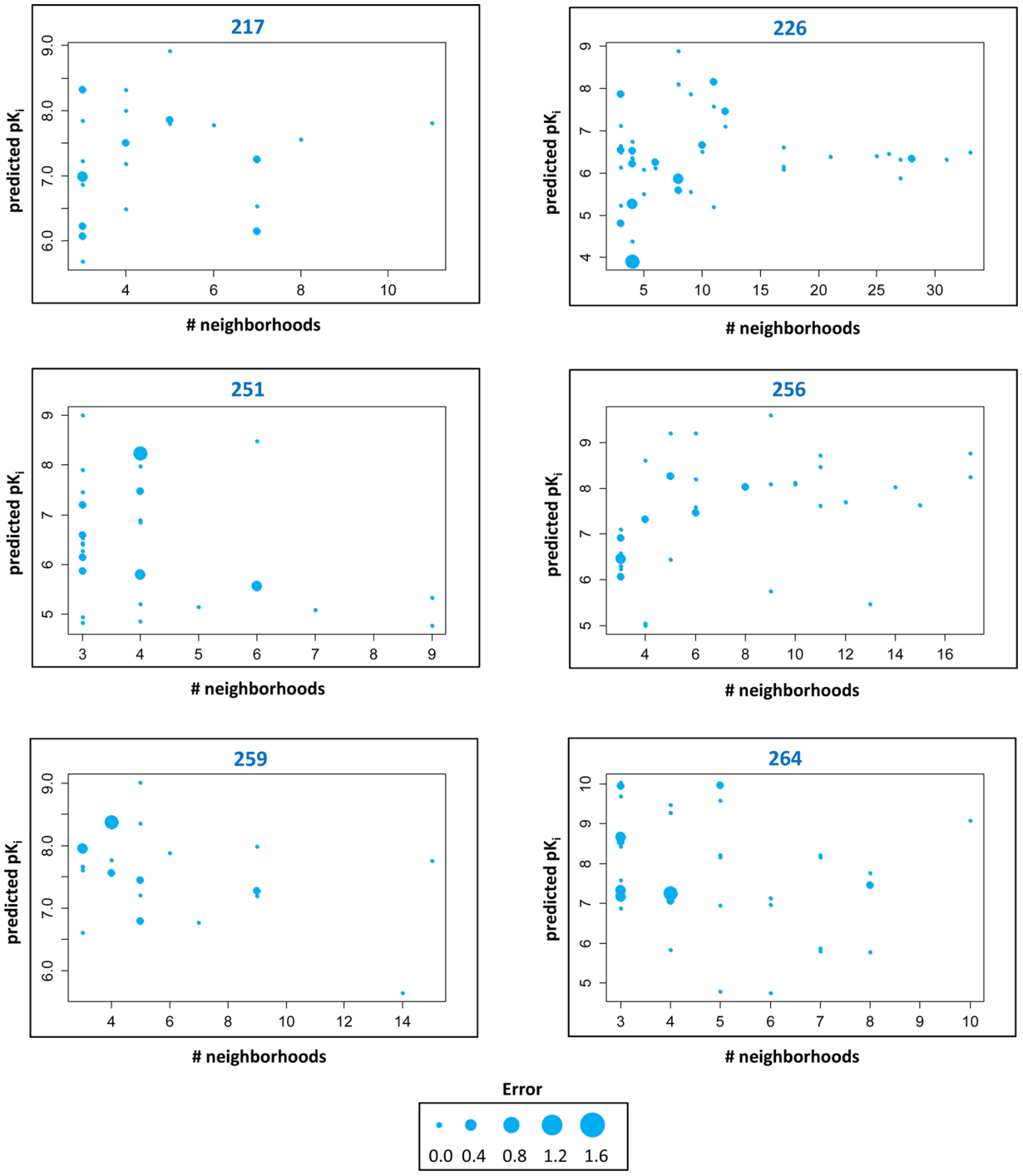


Figure 7. Neighborhood counts vs potency predictions. For L_SD compounds from each data set (TID at the top), the number of NBHs qualifying for predictions and the averaged predicted pK_i value is reported. Compounds are represented as dots that are scaled in size according to prediction errors (Δ pK_i units), as indicated at the bottom.

increasing numbers of qualifying NBHs and prediction accuracy.

Exemplary Compounds and Predictions. In Figure 8, potency predictions are reported for exemplary L_SD compounds together with their NBH information. In Figure 8a, predictions are reported for a histamine receptor antagonist for which 10 different NBHs were available. Despite different potency distributions of data set compounds across the NBHs, the nanomolar potency (pK_i 9.0) of this pseudovirtual compound was accurately predicted. Similarly, in Figure 8b, an adenosine A3 receptor antagonist for which six qualifying NBHs were available yielded an accurate potency prediction. By contrast, in Figure 8c, another adenosine A3 receptor antagonist is shown with only four available NBHs. Although a very low SD value (0.15 pK_i units) was also observed for individual predictions in this case, the potency of the

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

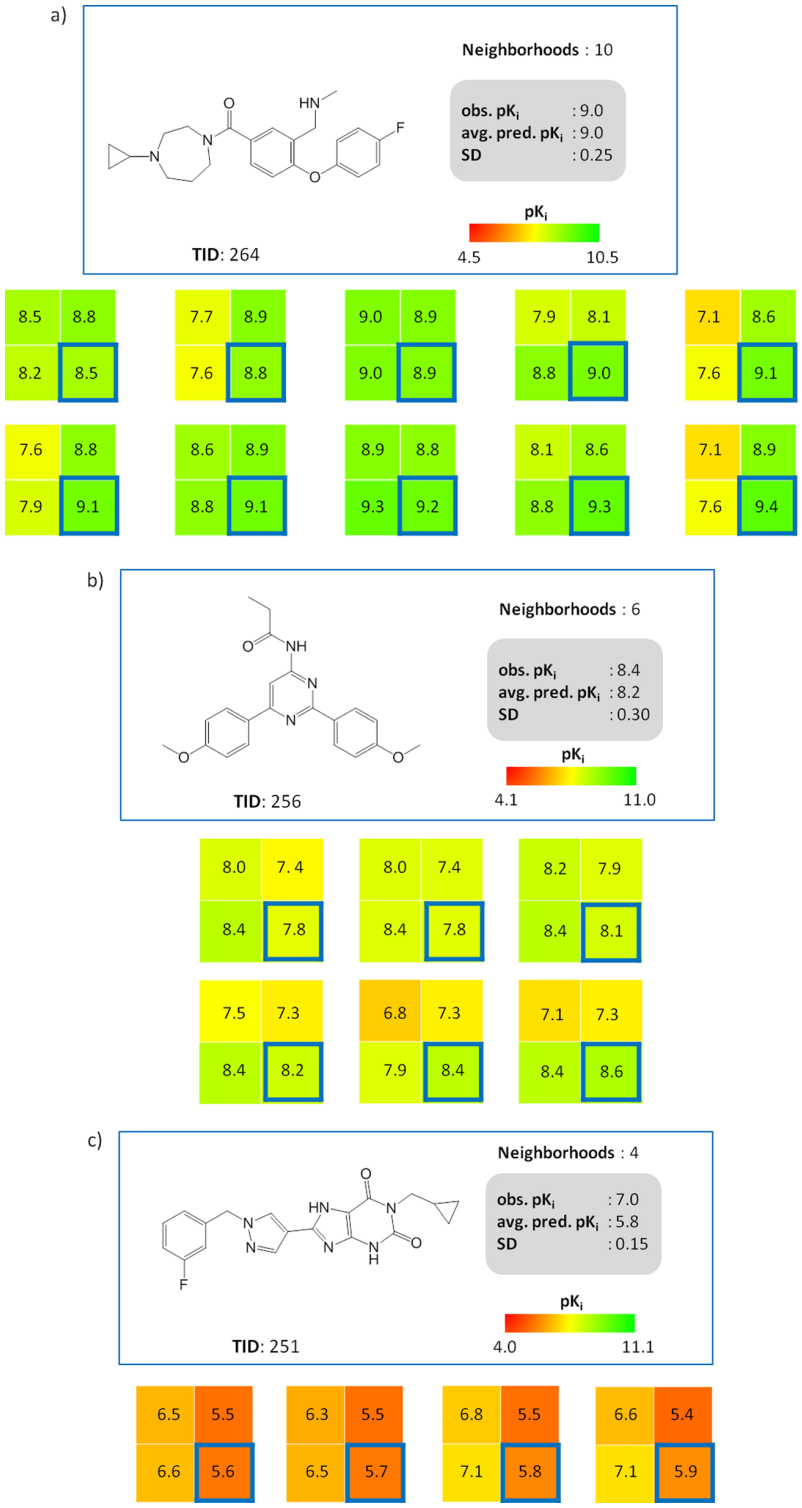


Figure 8. Exemplary predictions. In (**a**) and (**b**), the structures and NBHs of a histamine and adenosine A3 antagonist are shown, respectively, for which accurate potency prediction were obtained. In (**c**), adenosine A2 antagonist is shown for which a prediction error of more than 1 order of magnitude was observed. All three examples are L_SD compounds. Cells in NBHs are annotated with pK_i values of corresponding compounds. Cells framed in blue represent pseudovirtual compounds with predicted potency values.

compound was underpredicted in all four NBHs, yielding a final prediction error of 1.2 orders of magnitude, one of the largest errors observed; a rare case in our predictions of prioritized compounds. For increasing numbers of NBHs qualifying for prediction, only small errors were observed, as reported above.

For comparison, Figure 9 shows a compound yielding high SD values over multiple predictions (hence falling outside the

dx.doi.org/10.1021/ci5000483 | J. Chem. Inf. Model. 2014, 54, 801−809

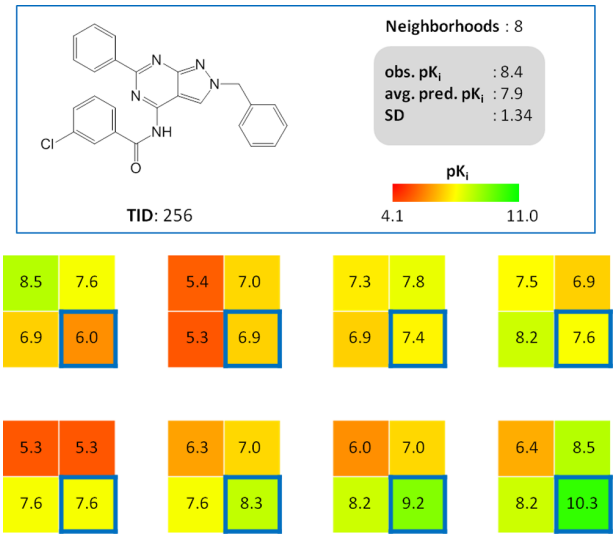


Figure 9. Neighborhoods in discontinuous SAR regions. Shown is an adenosine A3 antagonist for which NBHs map to discontinuous SAR regions and thus yield a wide range of potency predictions (falling outside the applicability domain of the NBH-based prediction concept). The representation is similar to Figure 8.

applicability domain of our approach). These predictions are characterized by the presence of NBHs falling into discontinuous local SAR regions. Although potency values cannot be accurately predicted for such NBHs, virtual compounds mapping to such regions might still be attractive prediction targets because large potency fluctuations are expected for such compounds (and one might hope to hit a potency "home run"). Therefore, high SD values can also be used as a diagnostic to systematically identify virtual compounds with NBHs in discontinuous SAR regions.

DISCUSSION AND CONCLUSIONS

The SARM data structure was originally designed to organize compound sets on the basis of core structures and substituents and all possible structural relationships between cores. This was accomplished through the systematic generation of A_MMS and their organization in a matrix format. A characteristic feature of SARMs is that they contain large numbers of virtual compounds that represent as of yet unexplored core structure and substituent combinations and hence offer suggestions for compound design. However, the SARM data structure does not enable a direct prioritization of such virtual compounds. Rather, visual analysis of SARMs is required to study virtual compounds. Therefore, we have developed a methodology to predict novel active compounds from SARMs. The central idea underlying this approach is the compound neighborhood concept. For each virtual compound in SARMs, NBHs exclusively consisting of known active compounds are systematically assessed. Virtual compounds can be ranked according to numbers of such NBHs. Prioritization on this basis is akin to a "guilt by association" approach, which assumes that the likelihood of a virtual compound to be active increases with the number of neighboring structural analogs. However, as shown herein, one can go a step further and utilize the NBH concept for potency predictions. These predictions are facilitated by applying a Free-Wilson-like additivity principle to individual neighborhoods. This leads to the prediction of the potency of a virtual compound on the basis of differential core and substituent contributions from active neighbors. Of course, the approach is distinct from classical Free-Wilson analysis that derives a mathematical model for the activity of a series of analogs by additively accounting for contributions from all R-groups. However, adapting the additivity principle essentially limits NBH-based potency predictions to the applicability domain of QSAR approaches, thus requiring the presence of SAR continuity and the absence of activity cliffs and cooperative SAR effects. A distinguishing feature of our NBH-based prediction approach is that predictions over multiple NBHs are prioritized. Then, one can assign confidence to consistent predictions resulting in low SD values. As demonstrated herein, accurate potency predictions were obtained in such cases across different data sets, with prediction accuracy further increasing with the number of qualifying NBHs. Depending on the composition of NBHs, virtual compounds with higher potency than known active neighbors can be predicted, and these predictions can also be easily prioritized. Moreover, potency predictions over multiple NBHs can be used as a diagnostic for local SAR environments. For example, predictions yielding high SD values are indicative of discontinuous SAR regions surrounding virtual compounds in which structurally analogous neighbors might have very different potencies. Although these regions usually fall outside the applicability domain of potency predictions employing an additivity principle, they are nonetheless interesting for compound design. This is the case because one might hope to hit a potency "home run" in discontinuous regions that are rich in NBHs, which are easily identified using the approach introduced herein.

In conclusion, neighborhood-based matrix analysis and potency predictions enable the prioritization of virtual compound from SARMs and are thus anticipated to further increase the attractiveness and utility of the SARM data structure for medicinal chemistry applications.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Dilyana Dimova for help with data sets and Liying Zhang for helpful discussions. D.G.-O. is supported by Boehringer Ingelheim.

REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, USA, London, UK, 2008.

(2) Kubinyi, H. Similarity and dissimilarity. A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 225−252.

(3) Cho, S. J.; Sun, Y. Visual exploration of structure-activity relationship using maximum common framework. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 571−578.

(4) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree — visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(5) Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO graph for compound data set representation and structure-activity relationship analysis. *J. Med. Chem.* **2012**, *55*, 5546−5553.
(6) Martin, Y. C. A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J. Med. Chem.* **1981**, *24*, 229−237.
(7) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.
(8) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(9) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discovery Today* **2010**, *15*, 631−639.
(10) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.
(11) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.
(12) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched molecular pair analysis in drug discovery. *Drug Discovery Today* **2013**, *18*, 724−731.
(13) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(14) Wawer, M.; Bajorath, M. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.
(15) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769−1776.
(16) Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Systematic mining of analog series with related core structures in multi-target activity space. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 665−674.
(17) Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(18) Kubinyi, H. Free Wilson analysis. Theory, applications and its relationships to Hansch analysis. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121−133.
(19) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.
(20) Maggiora, G. M. On outliers and activity cliffs − why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(21) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2012.
(22) R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2008.
(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

Summary

The study aimed to utilize compound neighborhoods present in SARMs in a systematic manner to predict the potency of unknown virtual compounds. The predictions are carried out by applying the Free-Wilson additivity principle to individual neighborhoods. Each neighborhood represents a "mini-QSAR" model. Predictions from multiple neighborhoods can be used as a diagnostic for local SAR environments of virtual compounds. Virtual compounds with consistent predictions map to continuous SAR regions and those with inconsistent predictions map to the discontinuous SAR regions. Compounds that map to continuous SAR region are predicted with high accuracy, whereas those that map to the discontinuous SAR region fall outside the applicability domain of this method. The neighborhood-based potency predictions enable the prioritization of virtual compounds from SARMs. My contribution to this study has been the design, implementation and the analysis of the prediction methodology.

The method reported herein is useful for hit-to-lead or lead optimization data sets where explicit potency values are available. However, the methodology is not optimal for data sets where approximate potency values are available, for example, in screening sets. In the next study in Chapter 6, a novel conditional probability-based prediction methodology is introduced to help prioritize active virtual compounds from screening sets.

Chapter 6

Hit Expansion from Screening Data Based upon Conditional Probabilities of Activity Derived from SAR Matrices

Introduction

Advances in high-throughput screening (HTS) technology and combinatorial chemistry have led to the fast accumulation of large amounts of activity data for chemical compounds.[5] With the screening results of a subset of a library at hand, different chemoinformatics approaches can be utilized to expand the number of compounds and chemotypes around the primary hits. Approaches such as nearest-neighbor search[6] or machine-learning (ML) methods[7] have gained popularity. However, successful hit expansion approaches must ultimately find a balance between mathematical complexity and chemical interpretability.

A novel methodology for hit expansion using SAR matrices is introduced in this study. The method is especially suited for raw screening data and is conceptually different from the QSAR-like approach discussed in the previous Chapter. It derives conditional probabilities of activity of individual cores and substituents from the activity information of compounds in SARMs. The per-

formance of the newly designed SARM-based method is compared with state-of-the-art machine learning methods.

Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201400164

Hit Expansion from Screening Data Based upon Conditional Probabilities of Activity Derived from SAR Matrices

Disha Gupta-Ostermann,[a] Jenny Balfer,[a] and Jürgen Bajorath*[a]

Abstract: A new methodology for activity prediction of compounds from SAR matrices is introduced that is based upon conditional probabilities of activity. The approach has low computational complexity, is primarily designed for hit expansion from biological screening data, and accurately predicts both active and inactive compounds. Its performance is comparable to state-of-the-art machine learning methods such as support vector machines or Bayesian classification. Matrix-based activity prediction of virtual compounds further extends the spectrum of computational methods for compound design.

Keywords: Matched molecular pairs and series · SAR matrices · Virtual compounds · Conditional probabilities · Activity prediction · Hit expansion

1 Introduction

Organizing sets of active compounds according to structural criteria is a pre-requisite for the exploration of structure-activity relationships (SARs). Popular approaches attempt to decompose compound series into core structures and substituents by applying the scaffold concept[1] or by considering reaction information such as for the generation of standard R-group tables. While R-group tables are limited to the representation of individual series, scaffold-based organization schemes can represent structurally diverse sets of active compounds and are often supported by graphical representation techniques.[1–3] Another powerful approach for systematically organizing large compound data sets and exploring SARs is the matched molecular pair (MMP) concept.[4,5] An MMP is defined as a pair of compounds that differ only by a structural change at a single site, which corresponds to the exchange of a pair of substructures.[4,5] Accordingly, changes in the activity of structurally related compounds can be directly associated with a well-defined structural modification.[5] MMPs can be algorithmically generated in a computationally efficient manner.[6] For SAR analysis, matching molecular series (MMS)[7] have been introduced as an extension of the MMP concept. An MMS is defined as a series of compounds that only differ by chemical changes at a single site. Hence, MMS generated for a compound data set include available analog series.[7]

The structure-activity relationship matrix (SARM) has been introduced as a data structure to systematically organize large compound data sets on the basis of MMS and identify SAR-informative compound series.[8] SARMs extract all possible structural relationships between compound series from data sets of any source and degree of heterogeneity. In addition to data set compounds forming MMS, SARMs contain an abundance of new (virtual) compounds that represent as of yet unexplored core-substituent combinations and often new chemistry (because systematically derived fragments from different compounds are combined). These virtual compounds can be rationalized to represent a chemical space envelope around a given data set and provide immediate suggestions for compound design. Therefore, one would like to prioritize analogs, preferably on the basis of activity predictions taking known activity information from existing data set compounds into account.

The computational prediction of new compounds on the basis of series information is typically attempted by the application of quantitative SAR (QSAR) approaches.[9,10] QSAR derives mathematical models using molecular descriptors that relate compound structure and properties to biological activity and predict the effects of chemical substitutions on activity. On this basis, new analogs can be predicted to complement existing series and further improve activity.[10]

Because SARMs contain both data set compounds with known activity and virtual compounds, matrix neighborhoods of virtual compounds that exclusively consist of known active compounds can be systematically identified and neighborhood information can be used for activity prediction.[11] In SARMs, compounds are represented as core-substituent combinations. Therefore, the activity of virtual

[a] *D. Gupta-Ostermann, J. Balfer, J. Bajorath*
Department of Life Science Informatics; Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn, Germany
tel: +49-228-2699-306; fax: +49-228-2699-341
**e-mail: bajorath@bit.uni-bonn.de*

compounds can be predicted from core and substituent contributions of known actives forming their neighborhoods on the basis of Free-Wilson additivity principles,[12,13] which represents one of the classical QSAR approaches. Accordingly, a 'mini Free-Wilson model' was built for each qualifying SARM neighborhood of a given virtual compound to predict its activity.[11] Virtual compounds often have multiple qualifying neighborhoods in SARMs, which makes it possible to assess the consistency of activity predictions.

Local Free-Wilson models introduced for SARM predictions require numerical activity values as input such as explicit IC_{50} or K_i values. Hence, they are best applied to hit-to-lead or lead optimization data sets. However, a major limitation of QSAR models including SARM-based Free–Wilson models is that they are only applicable to compound series representing continuous SARs, i.e., when structural modifications lead to gradual (and additive) changes in activity.[14] By contrast, the presence of SAR discontinuity (including activity cliffs) falls outside the applicability domain of standard QSAR modeling.[14,15] The presence of inconsistent activity predictions over multiple neighborhoods of a given virtual compound in SARMs is an indicator of SAR discontinuity in compound neighborhoods and such predictions are thus eliminated from further consideration.[11] Despite these general QSAR limitations, a strength of the SARM approach is that it provides many virtual candidates to further expand any given data set, which is a consequence of the underlying MMS-based structural organization. Therefore, it should be attractive to explore alternative activity prediction schemes.

We have aimed to develop a conceptually different, non-QSAR-like prediction approach that should be applicable to all SAR environments. Such an approach would enable matrix-based activity predictions for virtual compounds originating from data sets in which SARs have not been characterized in detail or in which only limited SAR information is available such as biological screening data. Herein we introduce a novel probability-based approach for qualitative activity predictions that is generally applicable to virtual candidate compounds, irrespective of their SAR features. We show that the methodology is suitable for the prediction of active compounds from screening data, a task generally referred to as hit expansion.

2 Concepts and Methods

2.1 SAR Matrix Generation

SARMs are 2D matrices generated by subjecting compounds to dual-step MMP fragmentation,[8] for which we use an in-house implementation of the efficient MMP algorithm by Hussain and Rea.[6] In the first step, data set compounds are systematically fragmented at one, two, or three exocyclic single bonds. The resulting larger fragment (core) is stored in an index table as a key and the smaller fragment(s) (substituents) as the corresponding value(s). In the second step, the cores in the index table are subjected to an additional round of fragmentation. The resulting fragments are stored in a new index table with the larger fragment representing the key, as before. Hence, this dual fragmentation scheme further extends the MMS concept[7] by identifying all compound series having structurally related cores termed 'structurally analogous MMS' (A_MMS).[8] This process is illustrated in Figure 1. Three model series are shown, each of which contains a common core structure and differs at a single site. When the cores of these series are subjected to a second round of fragmentation, core MMP relationships are identified because these core structures only differ at a single site. Hence, series 1, 2, and 3 form A_MMS, which are represented in a unique SARM. As illustrated in Figure 1, the SARM consists of compound series that are structurally analogous. Each row in the matrix contains an individual series and each cell represents an individual compound (i.e., a unique combination of a key and value fragment). The series comprising a SARM typically have overlapping yet distinct sets of substituents, giving rise to combinations of filled cells representing 'real' data set compounds and empty cells representing virtual compounds (i.e., key-value combinations that have not yet been explored). Cells corresponding to data set compounds are annotated with their activity information using a color code. In Figure 1, cells are colored red if they are inactive in a given assay and green if they are active. Alternatively, a continuous color spectrum can be used to represent relative activities.[11]

Typically, large compound data sets (including screening data) yield many MMPs, MMS, and SARMs. In our experience, the number of MMPs obtained from large data sets, including structurally heterogeneous sets, is not a limiting factor for the SARM approach. SARMs capture all possible core structure relationships contained in a given set and represent all A_MMS. A given real or virtual compound might participate in different A_MMS and hence occur in different SARMs. It should also be noted that the majority of SARMs (but not all SARMs) are incomplete from the point of view that they consist of both data set and virtual compounds, which provides the basis for compound design and activity prediction.

2.2 Matrix-Based Probability of Activity

The assignment of SARM-based probabilities of activity to virtual compounds is a central aspect of the new prediction methodology. In SARMs according to Figure 1, all data set compounds either belong to the *active* or the *inactive* class. Hence, we base our considerations upon a binary classification of assay activity.

Given the distribution of individual cores c and values v in active and inactive compounds, we can derive respective *conditional probabilities* for the classes y. Class contributions of cores and values are assumed to be *dependent* on each

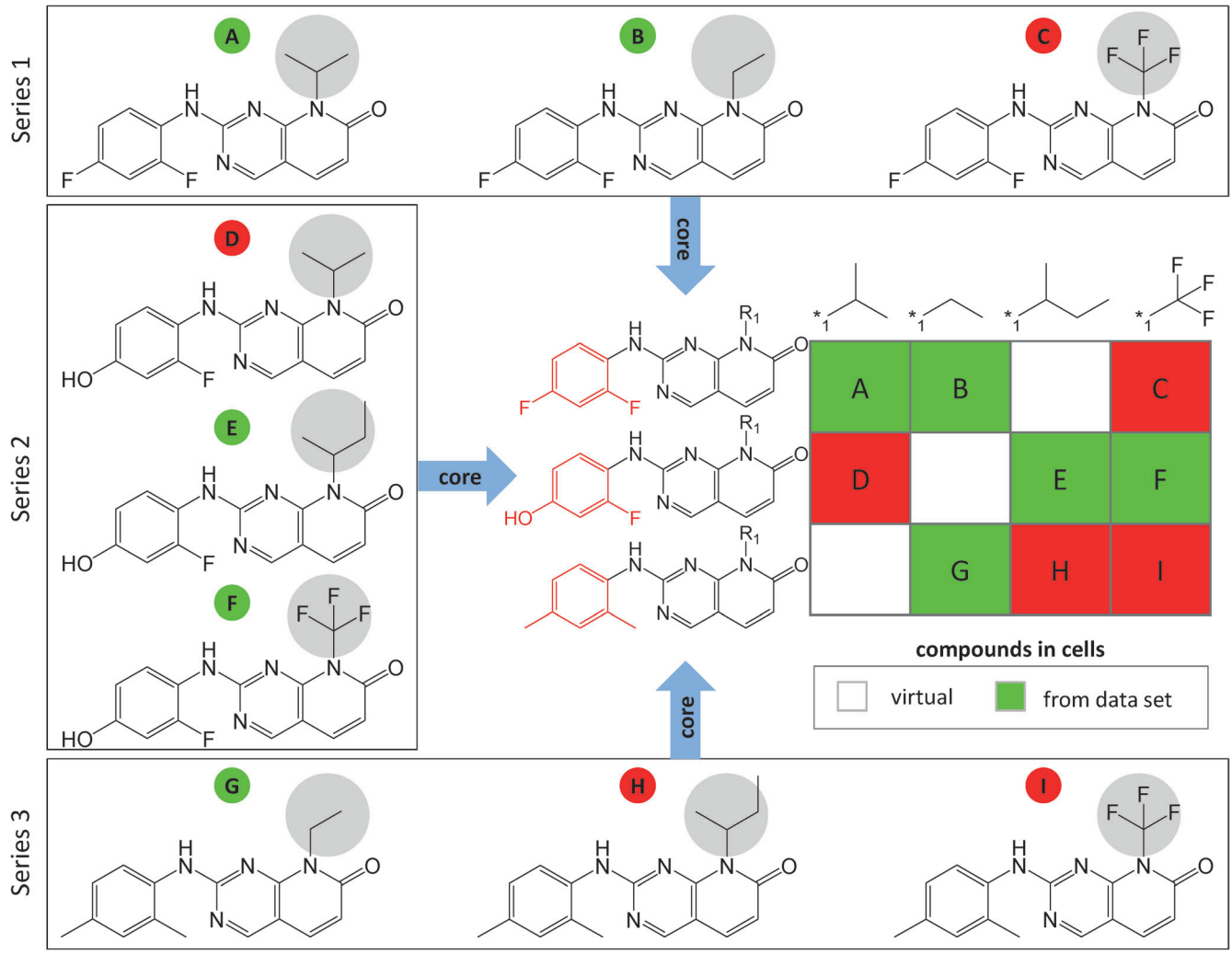


Figure 1. SAR matrix. Three model compound series (1, 2, and 3) containing three compounds each (A–C, D–F, and G–I) are shown with their activity annotations (inactive, red; active, green). Compounds in a series share a common core structure and differ by substitutions at a single site (highlighted in gray). The three series contain structurally related cores. Substructure differences between cores are highlighted in red. The SAR matrix is generated by combining all analog series with structurally related cores. Rows and columns represent compounds that share the same core and substituent, respectively. In each cell, the combination of a core and a substituent defines a unique compound. Compounds present in the data set are indicated by filled cells that are color-coded according to activity. In addition, empty cells represent virtual compounds (core-substituent combinations).

other because compounds in a matrix are represented as a *combination* of individual cores and values. To account for this dependence, *weights* based upon initially determined probabilities are assigned to compounds in order to calculate *class contributions* for each core and value. The *class probabilities* of a virtual compound are then derived by combining the contributions of its core and value, and the compound is predicted to be active or inactive on the basis of these probabilities. The calculation steps for the activity assignment and classification method are further described below and illustrated for a model SARM in Figure 2.

2.2.1 Core and Value Class Probabilities

In the first step, the *class probability* for each core and value in a given matrix is calculated. Accordingly, $P(y|c)$ and $P(y|v)$ are the conditional probabilities that describe how likely it is to observe a given specific class $y \in \{\text{active, inactive}\}$ for a core c and a value v, respectively. If $c^{(x)}$, $v^{(x)}$, $y^{(x)}$ are the core, value, and class of a given compound x, we can express the conditional class probabilities as the fraction of compounds with a core c or value v and class y over all compounds containing this core or value:

$$P(y|c) = \frac{\sum_x \mathbf{1}(c^{(x)} = c)\mathbf{1}(y^{(x)} = y)}{\sum_x \mathbf{1}(c^{(x)} = c)} \quad (1)$$

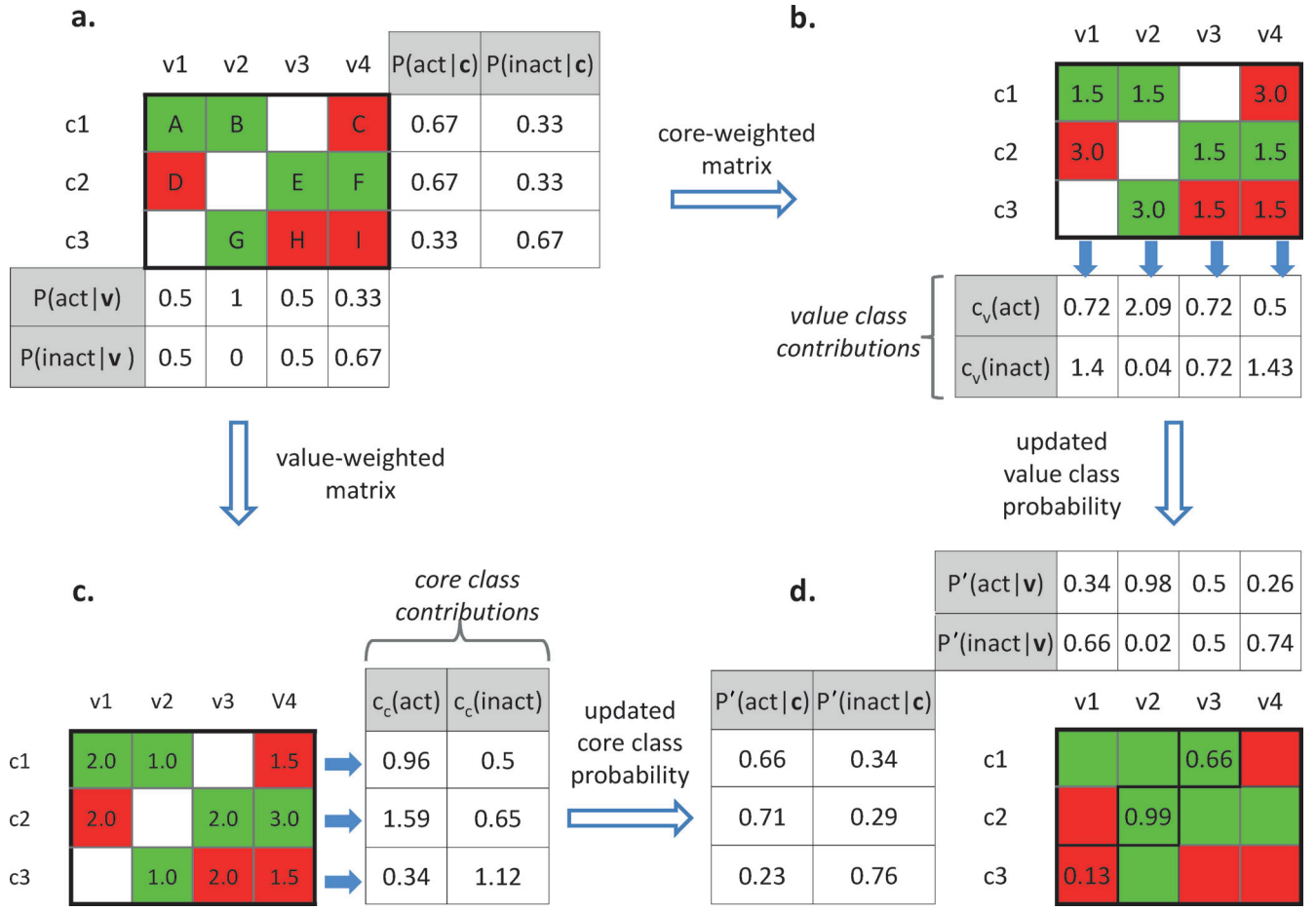


Figure 2. SARM classification. For the model SARM with three cores (c1–c3) and four values (v1–v4), represented according to Figure 1, step-wise class probability calculations are illustrated. (a) Initial class probabilities for the cores and values are reported. (b) shows the *core-weighted matrix*. Compounds are assigned weights depending on the inverse core class probabilities. Value class contributions calculated from the weights of each value are shown. (c) shows the *value-weighted matrix*. Compounds are assigned weights depending on the inverse value class probabilities. Core class contributions calculated from the weights of each core are reported. (d) Updated value and core class probabilities are derived from the matrices in (b) and (c), respectively. Activity probability (p_x) values are reported for virtual compounds and the corresponding cells are color-coded red (predicted inactivity) or green (predicted activity) according to the classification scheme detailed in the text.

$$P(y|v) = \frac{\sum_x \mathbf{1}(v^{(x)} = v)\mathbf{1}(y^{(x)} = y)}{\sum_x \mathbf{1}(v^{(x)} = v)} \quad (2)$$

Here, $\mathbf{1}(a=b)$ is a function that returns 1 if $a=b$ and 0 otherwise. Core and value class probabilities are given in Figure 2a for the three core (*c1–c3*) and four value fragments (*v1–v4*) of the SARM in Figure 1. For the core *c1*, the class probability of activity (0.67) is higher than of inactivity (0.33), whereas for value *v3*, the class probabilities are equal.

2.2.2 Core- and Value-Weighted Matrices

Inverse class probabilities are used as *core weights* and *value weights* of data set compounds:

$$w_c(x) = P(y^{(x)}|c^{(x)})^{-1} = \frac{\sum_x \mathbf{1}(c^{(x)} = c)}{\sum_x \mathbf{1}(c^{(x)} = c)\mathbf{1}(y^{(x)} = y)} \quad (3)$$

$$w_v(x) = P(y^{(x)}|v^{(x)})^{-1} = \frac{\sum_x \mathbf{1}(v^{(x)} = v)}{\sum_x \mathbf{1}(v^{(x)} = v)\mathbf{1}(y^{(x)} = y)} \quad (4)$$

Two weighted matrices are derived from the inverse class probabilities of the cores and values and are referred to as the *core-weighted* and *value-weighted matrix*, respectively. In case of the core-weighted matrix, the inverse class probabilities of the cores are mapped to the compounds that represent the corresponding core and class. In Figure 2b, a core-weighted matrix is shown in which the inverse probability for the inactive and active class for core *c1* is 3.0 and 1.5, respectively. Hence, the inactive compound C (Figure 2a) is assigned a weight of 3.0, and the

active compounds A and B are each assigned a weight of 1.5. Accordingly, the less frequently observed class is assigned a higher weight, as further rationalized below. For virtual compounds (empty cells), the class is unknown and no weight is assigned. Similarly, inverse class probabilities of values are mapped to compounds that contain the corresponding values and belong to the corresponding class, thus producing a value-weighted matrix as shown in Figure 2c.

2.2.3 Core and Value Class Contributions

From the weighted matrices, *class contributions* are calculated. The core-weighted matrix is used for deriving the value class contributions and the value-weighted matrix for deriving the core class contributions. This rationalizes the calculation of weights from the previous step: the less frequently observed class for a core is assigned a higher weight, which leads to a stronger class contribution of the corresponding value of a compound. In the case of core *c2* in Figure 2b, a weight of 3.0 for compound D (Figure 2a) given by value *v1* indicates that the compound is predominantly inactive due to the influence of value *v1*, more so than due to the influence of core *c2*. Thus, for calculating the class contributions for value *v1*, individual core weights of compounds represented by value *v1* are considered:

$$c_v(y) = \frac{\alpha + \sum_x w_c(x)\mathbf{1}\left(v^{(x)} = v\right)\mathbf{1}\left(y^{(x)} = y\right)}{2\alpha + \sum_x \mathbf{1}\left(v^{(x)} = v\right)} \quad (5)$$

Similarly, individual core contributions can be calculated by using the weights from the value-weighted matrix:

$$c_c(y) = \frac{\alpha + \sum_x w_v(x)\mathbf{1}\left(c^{(x)} = c\right)\mathbf{1}\left(y^{(x)} = y\right)}{2\alpha + \sum_x \mathbf{1}\left(c^{(x)} = c\right)} \quad (6)$$

The smoothing parameter α is used to prevent zero probabilities in cases where there is no compound with a certain core or value for a class. For example, this is the case for value *v2* in Figure 2 that does not represent inactive compounds. Without application of this smoothing factor, the value class contribution for *v2* and the inactive class would be unlikely for a limited sample size:

$$c_{v2}(inactive) = \frac{\sum_x w_c(x)\mathbf{1}\left(v^{(x)} = v2\right)\mathbf{1}\left(y^{(x)} = inactive\right)}{\sum_x \mathbf{1}\left(v^{(x)} = v2\right)} = \frac{0}{2}$$

Therefore, Laplacian smoothing is applied to account for the possibility that there might be a small number of inactive instances with value *v2* that are not used for training.

From the class contributions, the corresponding *updated core class* and *value class probabilities* $P'(y|c)$ and $P'(y|v)$ are obtained through normalization:

$$P'(y|c) = \frac{c_c(y)}{\sum_{\hat{y}} c_c(\hat{y})} \quad (7)$$

$$P'(y|v) = \frac{c_v(y)}{\sum_{\hat{y}} c_v(\hat{y})} \quad (8)$$

Individual *value* and *core class contributions* are given in Figure 2b and c, respectively, and the corresponding *updated class probabilities* are reported in Figure 2d.

2.2.4 Combined Class Probabilities

Finally, to predict the *activity probability* p_x of a virtual compound *x*, the corresponding updated core and value class probabilities are combined:

$$p_x = \frac{P'\left(y|c^{(x)}\right)P'\left(y|v^{(x)}\right)}{\sum_{\hat{y}} P'(\hat{y}|c^{(x)})P'(\hat{y}|v^{(x)})} \quad (9)$$

The resulting value ranges from 0 to 1 and an increasing value reflects the increasing probability for a compound to be active. Calculated values for predicting the activity state of virtual SARM compounds are given in Figure 2d. Two of three virtual compounds with values of 0.66 and 0.99 are predicted to be active, whereas one (0.13) is predicted to be inactive.

2.2.5 Exemplary Calculations

As an example for the calculations summarized in Figure 2, let us consider the virtual compound consisting of core *c3* and value *v1*. Its core is shared by compounds G, H, and I and its value by compounds A and D. Consequently, the contributions of these compounds are taken into account for activity prediction.

First, activity probabilities are calculated for all participating cores and values, i.e., not only for *c3* and *v1*, but also *c1*, *c2*, *v2*, *v3*, and *v4*, because they are present in compounds containing *c3* or *v1*. These frequency-based probability calculations are summarized in Figure 2a. The resulting probabilities suggest that our virtual compound is two times more likely to be inactive than active due to the presence of core *c3* and that value *v1* does not convey activity information. These initial estimates are then further refined (updated) by taking information of all compounds in the SARM into account.

To arrive at an updated probability of core activity, the value-weighted matrix is derived. Here, the value weights of compounds G, H, and I, which share core *c3* with the virtual compound, are taken into consideration. These value weights are calculated according to Equation 4 and give the contribution of each value to a class. For example, compound G has a value weight of 1, because only active compounds contain this value. Furthermore, compound I has a value weight of 1.5 because it belongs to the majority class (inactive) of compounds with value *v4*. Finally, com-

pound H has the highest value weight of 2.0 because its class (inactive) is observed equally frequent as the active class. Hence, the value weight of a given compound is increasing with the number of compounds belonging to the other class. From compounds G-I, we can calculate the core class contributions for *c3* according to Equation 6 with a smoothing factor of $\alpha = 0.1$

$$c_{c3}(active) = \frac{0.1 + 1.0}{0.2 + 3} = 0.34$$

$$c_{c3}(inactive) = \frac{0.1 + 2.0 + 1.5}{0.2 + 3} = 1.12$$

Through normalization using Equation 7, we arrive at the updated core class probabilities of 0.23 and 0.76, respectively. By weighting compounds H and I higher than G, the core contribution to inactivity was increased.

Analogously, to calculate the class contributions for value *v1*, the core weights of compounds A and D are used. A is active and belongs to the majority class of *c1* and hence has a lower weight than D, which is inactive and belongs to the minority class of *c2*. Consequently, the inactive compound D influences the class probability of *v1* more than the active compound A, which leads to an increased probability of inactivity. The underlying idea is that compound D is unlikely to be inactive because of its core; hence, its value should be responsible for inactivity.

2.3 Activity-Based Classification

2.3.1 Concept

SARMs were generated for compound data sets from individual screening assays taken from PubChem[16] (see below). For each classification trial, 20% of the SARM compounds were randomly selected as test compounds. These test compounds were converted into virtual compounds for predictions. The training compounds (80%) were used to calculate core and value class probabilities of individual matrices. A test compound can appear in different matrices. In each matrix, it represents a unique combination of core and value fragment. Thus, a test compound can be assigned multiple matrix-dependent p_x values. Therefore, a *mean activity probability value* $\hat{p}_x$ was assigned to a test compound contained in multiple SARMs. Test compounds for which no training compounds with corresponding core and value fragments were available (due to the random removal of test instances from SARMs) were omitted. Each qualifying test compound was assigned to one of three different categories (as further discussed below) based on the calculated mean activity probability values: inactive, $\hat{p}_x < 0.5$; inconclusive, $\hat{p}_x = 0.5$; active, $\hat{p}_x > 0.5$.

2.3.2 Calculations

For each data set, 10 different trials with randomly assembled training and test sets were carried out. Routines to generate SARMs were implemented with the aid of the OpenEye chemistry toolkit[17] and classification routines were implemented in Java. For Laplacian smoothing, we consistently used a factor of $\alpha = 0.1$.

2.4 SARM Selection

For probability-based activity predictions, SARMs are preferred that have a high compound density and overlap between corresponding values (columns) in different A_MMS (rows). Therefore, a *matrix overlap* measure is applied. For a given SARM, matrix overlap is determined as the average of all *row overlap* (*RO*) values. For individual columns in SARMs, *RO* is calculated as

$$RO = \frac{n_c - 1}{\#rows - 1} \quad (10)$$

Here n_c corresponds to the number of data set compounds present in each column. *RO* yields a numerical score between 0 (no overlap) and 1 (complete overlap). Thus, an *RO* of 0 for each column of a matrix will result in the matrix overlap score of 0, reflecting a mutually exclusive nature of the substitution patterns among the A_MMS comprising the SARM, whereas an *RO* of 1 for each column will result in the matrix overlap of 1, reflecting the presence of A_MMS with identical substitution patterns.

Figure 3 reports the distribution of SARMs with a matrix overlap score of < 0.5 (dark blue) and ≥ 0.5 (light blue) for different data sets. For our classifications, only matrices that had an overlap score ≥ 0.5 were preselected.

Furthermore, qualifying SARMs were assigned to three different categories depending on the *class composition* (*CC*) of the compounds: *exclusively inactive*, SARMs only containing inactive compounds; mixed, SARMs containing both active and inactive compounds (with varying ratio); *exclusively active*, SARMs containing only active compounds.

Figure 4 reports the distribution of SARMs with a *CC* for the different assay data sets over these three states.

The set of qualifying SARMs (matrix overlap score ≥ 0.5) is in the following referred to as *a-SARMs*. In addition, a subset of qualifying SARMs was generated for model building and predictions in which *exclusively active* and *exclusively inactive* SARMs were omitted. This subset was generated to avoid potential bias of class probability calculations by these exclusive CC categories and is referred to as *b-SARMs*. The SARM subset selection is illustrated in Figure 5.

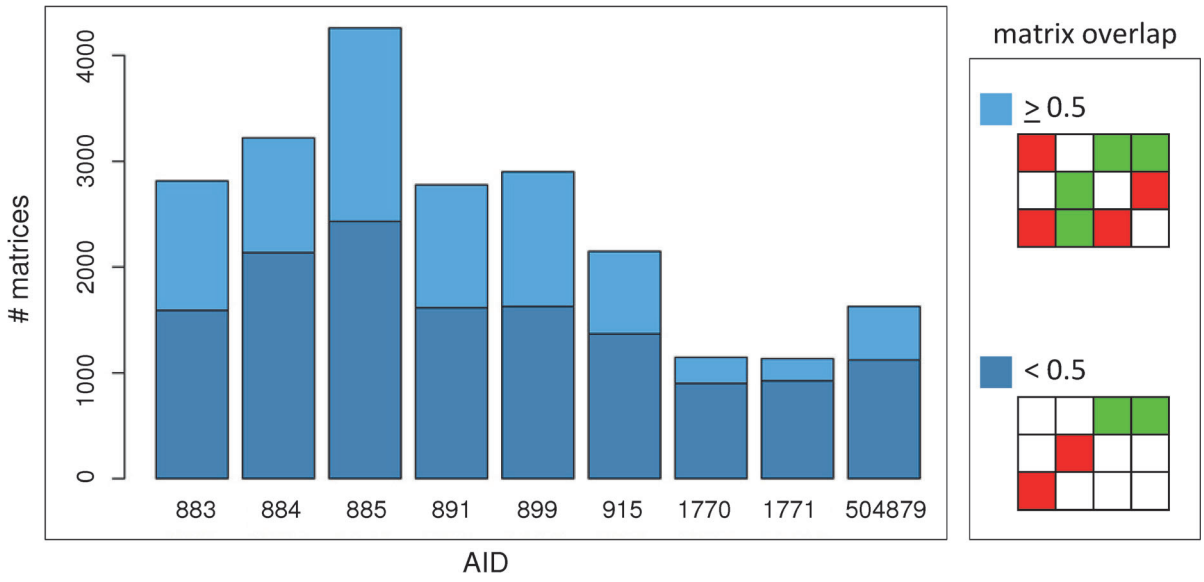


Figure 3. Matrix overlap distribution. Bar plots represent the distribution of matrix overlap scores below 0.5 (dark blue) and equal to or greater than 0.5 (light blue) over all assays. For these scoring ranges, exemplary SARMs are shown on the right.

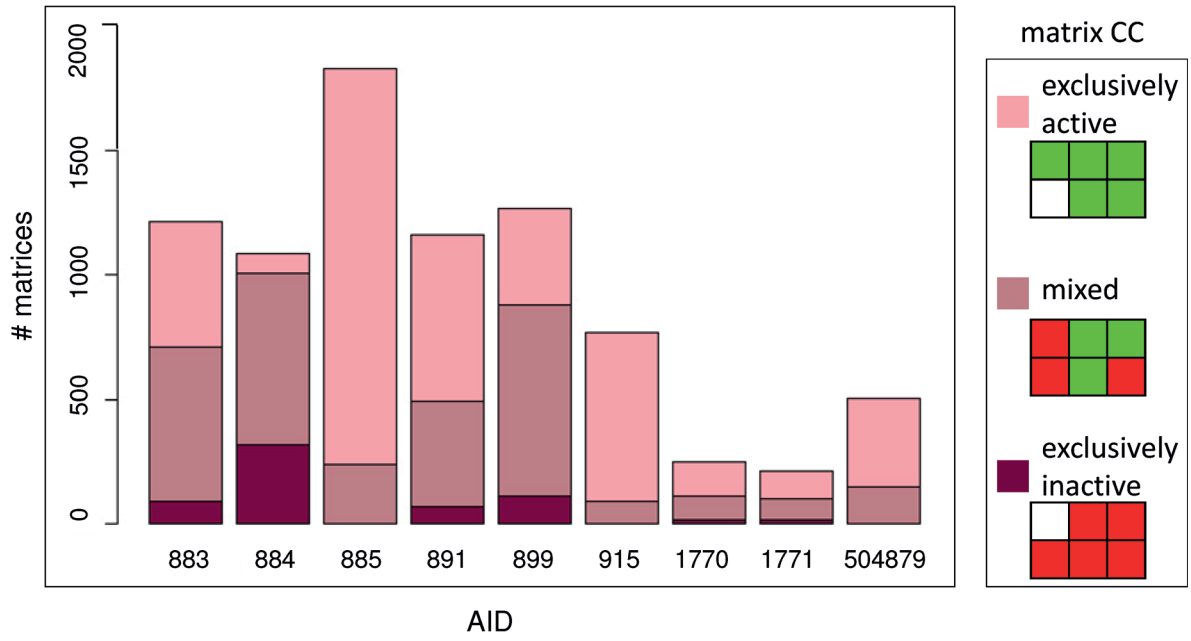


Figure 4. Class composition distribution. Bar plots represent the distribution of matrix compound class composition for 'exclusively inactive' (maroon), 'mixed' (dark pink), and 'exclusively active' (light pink) types over all assays. For these composition types, exemplary SARMs are shown on the right.

2.5 Assays

Nine different confirmatory bioassays were assembled from PubChem. From each screening compound set, SARMs were systematically generated and analyzed. Table 1 reports the assay data sets, their composition, the total number of SARMs, and the compounds comprising these SARMs. Table 2 summarizes the number of a-SARMs and their compound content, and Table 3 b-SARMs and their composition. SARM-based activity predictions and control calculations (see below) were carried out for both a-SARMs and b-SARMs. It should be noted that only assays that resulted in at least 50 SARMs comprising at least 100 compounds were considered for SARM-based predictions. The nine selected assays satisfied these criteria for both a-SARMs and b-SARMs.

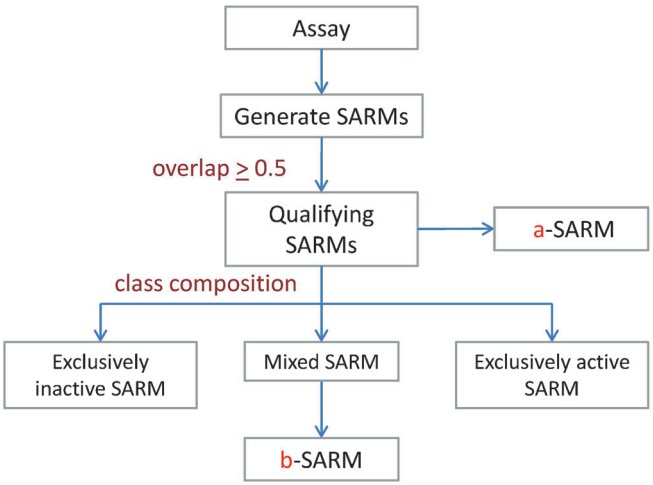


Figure 5. SARM selection. Selection criteria for a-SARM and b-SARM (compound) subsets are summarized.

These assays were pre-selected on the basis of MMS and matrix generation to yield a significant number of overlapping A_MMS and informative matrices that ensure statistically sound predictions. In prospective practical applications (for which no statistical validation is required), sparsely populated matrices can also be used.

2.6 Control Calculations

To put the results of SARM-based activity prediction into perspective, control calculations were carried out using three state-of-the-art machine learning methods including naïve Bayesian classification (NB),[18] random forests (RF),[19] and support vector machines (SVM).[20] NB and RF models also produce probability scores while SVMs yield discriminative models.

For all calculations, extended connectivity fingerprints with bond diameter 4 (ECFP4)[21] were used as compound descriptors. For NB, the Bernoulli formulation was applied to account for the binary nature of the fingerprint descriptors. In addition, SVM models were generated using the Tanimoto kernel.[22] Furthermore, to account for the imbalance of active and inactive training compounds in the different assay data sets (i.e., more inactive compounds were available), we applied sample weights inversely proportional to the number of actives and inactives in the training set, respectively. For each control method, 10 individual trials were performed on the same training and test sets used for SARM-based predictions. Implementations of the freely available Python library scikit-learn[23] were used.

Table 2. Statistics for a-SARMs. For each assay (indicated by AID according to Table 1), the number of selected a-SARMs and the number of compounds contained in these a-SARMs (including inactives and actives) is reported.

AID	# a-SARM	# cpds	# inactive	# active
883	1215	2889	2275	614
884	1089	2749	1101	1648
885	1828	3930	3843	87
891	1164	2567	1938	629
899	1265	2876	1944	932
915	771	2214	2183	31
1770	251	910	677	233
1771	215	865	610	255
504865	504	1454	1402	52

Table 3. Statistics for b-SARMs. For each assay, the number of selected b-SARMs and the number of compounds contained in these b-SARMs (including inactives and actives) is reported.

AID	# b-SARM	# cpds	# inactive	# active
883	620	2550	1980	570
884	687	2519	1013	1506
885	238	1954	1867	87
891	426	2315	1697	618
899	766	2650	1748	902
915	89	584	557	27
1770	97	680	485	195
1771	81	634	428	206
504865	151	953	901	52

Table 1. PubChem assay data and SAR matrices. For each assay, the PubChem Assay ID (AID), the target name, the number (#) of total compounds (cpds) in the assay, the number of SARMs, and the number of compounds covered by these SARMs (including inactives and actives) is reported.

AID	Target	# total cpds	# SARMs	# cpds	# inactive	# active
883	Cytochrome P450/2C9	7461	2808	3988	3210	778
884	Cytochrome P450/3A4	9685	3226	4952	2675	2277
885	Cytochrome P450/3A4	11982	4253	6475	6372	103
891	Cytochrome P450/2D6	7213	2782	3795	2920	875
899	Cytochrome P450/2C19	7547	2896	4064	2847	1217
915	Hypoxia-inducible factor 1	7647	2142	3749	3629	120
1770	CDC-like kinase 4	1126	1153	1126	794	332
1771	CDC-like kinase 4	1103	1138	1103	747	356
504865	USP1 protein	5764	1633	2551	2465	86

As additional controls, we also carried out standard k-nearest neighbor (*k*-NN) similarity search calculations (1-NN, 5-NN, and 10-NN) using fingerprints on a subset of assays. The prediction accuracy of these k-NN search calculations was consistently lower than of all machine learning and SARM-based classification calculations. Therefore, the results of *k*-NN control calculations were not included in the detailed comparison presented below.

2.7 Performance Measures

Average statistics were calculated over all 10 trials and used for performance evaluation. The following performance measures were applied:

$$Balanced\ accuracy = BAC = \frac{0.5*TP}{actives} + \frac{0.5*TN}{inactives}$$

$$F1\ score = \frac{2*TP}{2*TP + FN + FP + IA}$$

TP, TN, FP, and FP define the number of true positives, true negatives, false positives, and false negatives, respectively, and IA denotes the number of compounds that were active but predicted to be inconclusive. In addition, area under the receiver operating characteristic (ROC) curve (ROC AUC) [24] values were also calculated as performance measures.

The control calculations produce predictions for all compounds including those that are inconclusive using the SARM approach. Hence, the performance measures were adjusted to classify inconclusive SARM predictions as false (yielding a conservative estimate for the performance of the SARM approach).

3 Results and Discussion

3.1 Study Design

SARMs provide a systematic MMS-based organization of a compound data set and capture all possible structural relationships between compound series. Because SARMs also reveal all possible key-value combinations arising from systematic MMP fragmentation, they also provide many virtual compounds for further exploration that expand the chemical space around a given compound set. Therefore, activity predictions are highly desirable to prioritize virtual candidate compounds for synthesis. QSAR-like predictions are feasible for virtual candidates with qualifying neighborhoods on the basis of local Free-Wilson models[12] and are particularly suited for hit-to-lead or other chemical optimization campaigns. However, these local QSAR models typically require the presence of advanced compound data sets with significant SAR information content and are confined to continuous SAR environments in SARMs and thus not generally applicable.[11] Therefore, we have set out to develop a conceptually distinct activity prediction method that is generally applicable to data sets of any source, with a particular focus on hit expansion from screening data, for which QSAR-type approaches are not well applicable. Hit expansion focuses on narrowly defined chemical space around "activity islands" which can be well mapped using SARMs. These considerations have led to the design of the conditional probability-based SARM prediction approach presented herein.

3.2 SARM Statistics

The number of SARMs generated for the nine assay data sets (comprising 1103 to 6475 compounds) ranged from 1138 to 4253. Three assays (AID: 885, 915, and 504865) had a strongly unbalanced class composition with significantly fewer active than inactive compounds, as one would expect from confirmatory screening data (Table 1). Some assays were directed against the same target; for example, AID 884 and 885 addressed cytochrome P450 3A4. These two assays shared 9468 compounds but only 6067 compounds had the same activity state (active/inactive) in both the assays. This was also reflected by the activity distribution of SARM compounds from these two assays. SARMs of AID 884 retained many more actives (2277) compared to AID 885 (103). Furthermore, AID 1770 and 1771 reported CDC-like kinase 4 inhibitors. In this case, different from the P450 assays, there was significant overlap of active/inactive compounds (1051).

The number of SARMs selected for model building and prediction on the basis of matrix overlap scores ranged from 215 to 1828 (Table 2). On average, pre-selected SARMs retained 67% of the assayed compounds. Hence, although screening sets are often diverse, SARMs detected many MMS and structural relationships between data sets compounds.

3.3 Prediction Accuracy

Activity predictions were systematically carried out on the basis of pre-selected a-SARMs and their b-SARM subsets, as detailed in the Methods section. We separately determined the prediction accuracy for these sets.

3.3.1 a-SARMs

Average ROC AUC values obtained for a-SARM predictions are reported in Table 4. Despite the highly variable class composition of a-SARMs, our probability-based approach accurately predicted active and inactive compounds with ROC AUC values for SARM consistently above 90%.

3.3.2 b-SARMs

Compared to a-SARMs, the proportion of retained inactive compounds was significantly reduced in b-SARMs, due to

Table 4. ROC AUC values for a-SARMs. For each assay, average ROC AUC values (and standard deviations) from 10 trials for SARM, NB, RF, and SVM predictions of compounds contained in a-SARMs are reported.

AID	SARM	NB	RF	SVM
883	0.94 (0.01)	0.90 (0.01)	0.93 (0.01)	0.94 (0.02)
884	0.95 (0.01)	0.88 (0.01)	0.95 (0.01)	0.96 (0.01)
885	0.97 (0.02)	0.94 (0.02)	0.93 (0.03)	0.98 (0.02)
891	0.94 (0.01)	0.91 (0.01)	0.93 (0.01)	0.95 (0.01)
899	0.92 (0.01)	0.87 (0.02)	0.92 (0.01)	0.93 (0.01)
915	0.92 (0.05)	0.93 (0.02)	0.83 (0.07)	0.96 (0.02)
1770	0.97 (0.02)	0.92 (0.02)	0.97 (0.02)	0.97 (0.02)
1771	0.98 (0.01)	0.94 (0.01)	0.98 (0.01)	0.99 (0.01)
504865	0.92 (0.04)	0.93 (0.06)	0.88 (0.08)	0.95 (0.03)

Table 5. ROC AUC values for b-SARMs. For each assay, average ROC AUC values (and standard deviations) from 10 trials for SARM, NB, RF, and SVM predictions of compounds contained in b-SARMs are reported.

AID	SARM	NB	RF	SVM
883	0.92 (0.02)	0.87 (0.02)	0.91 (0.02)	0.93 (0.01)
884	0.95 (0.01)	0.88 (0.02)	0.94 (0.01)	0.96 (0.01)
885	0.96 (0.02)	0.92 (0.03)	0.94 (0.04)	0.97 (0.02)
891	0.94 (0.01)	0.90 (0.01)	0.93 (0.01)	0.94 (0.01)
899	0.92 (0.01)	0.86 (0.02)	0.92 (0.01)	0.93 (0.01)
915	0.83 (0.07)	0.79 (0.13)	0.81 (0.08)	0.85 (0.09)
1770	0.95 (0.02)	0.91 (0.02)	0.96 (0.03)	0.96 (0.01)
1771	0.95 (0.03)	0.93 (0.03)	0.96 (0.02)	0.98 (0.02)
504865	0.87 (0.05)	0.89 (0.08)	0.86 (0.08)	0.91 (0.06)

Table 6. Inconclusive SARM predictions. For each assay, the average percentage of active and inactive compounds yielding inconclusive SARM predictions is reported for the a-SARM and b-SARM subsets.

AID	a-SARM		b-SARM	
	% active	% inactive	% active	% inactive
883	0.0	0.08	0.37	0.18
884	0.10	0.13	0.10	0.0
885	0.0	0.0	0.0	1.17
891	0.08	0.52	0.26	0.51
899	0.22	0.11	0.3	0.06
915	0.0	7.0	4.0	10.6
1770	0.25	1.5	0.27	1.79
1771	0.0	0.90	1.5	2.18
504865	0.0	2.0	0.0	0.0

the removal of exclusively inactive matrices, whereas the proportion of active compounds essentially remained constant, despite the removal of exclusively active matrices. Thus, b-SARMs further balanced the composition of learning and test sets. Table 5 reports the ROC AUC values obtained for b-SARMs that were similar to those obtained for a-SARMs (Table 4). Hence, the exclusion of exclusively active/inactive matrices did not notably compromise average prediction accuracy across all assay data sets.

Table 6 reports the average percentage of active and inactive compounds that predicted to be inconclusive on the basis of a-SARMs and b-SARMs, which was consistently very small and negligible.

3.4 Method Comparison

Given the perhaps surprisingly high performance of the probability-based activity prediction approach, we compared SARM-based calculations with state-of-the-art machine learning methods including NB, RF, and SVM using alternative performance measures. In Tables 4 and 5, average ROC AUC values of the control calculations are compared to SARM-based predictions using the same training and test sets. The comparison reveals that global prediction performance of the SARM-based approach reached or exceeded performance levels of the machine learning methods. It should be noted that the SARM-approach uses MMP-based compound representations, while the control calculations were carried out using fingerprints. Different compound representations compromise direct comparisons.

Since only limited numbers of active compounds were available as test instances in unbalanced data sets, BAC and F1 scores were also calculated as performance measures. Figure 6 shows that prediction performance in part significantly differed for different assay data sets. Balanced accuracy equally weights true positives and true negatives, regardless of the composition of data sets, and accounts for the fraction of individual test compounds that were correctly predicted. In Figure 6 (left), average balanced accuracy for individual assays was highest for SVM when a-SARMs were considered. Balanced accuracy for SARM-based predictions was comparably high ($>80\%$) for data sets with balanced class composition but lower ($\sim 60\%$) for unbalanced data sets. Both NB and SVM classifiers reached higher performance than RF and SARM-based predictions for unbalanced sets. F1 scores were also compared, which account for the fraction of true positive, but not true negative predictions. In this case, the performance was comparable for all methods. Similar trends were observed for average BAC and F1 scores across balanced and unbalanced a-SARMs and b-SARMs. For example, average BAC and F1 scores for AID 884 and 885 varied for a-SARM and b-SARM sets. For AID 884, for which much larger numbers of active training compounds were available than for AID 885, higher F1 and BAC scores were observed for all methods (except BAC values for SVM and NB). By contrast, BAC and F1 scores for AID 1770 and 1771 were comparable, consistent with their large compound overlap, as discussed above.

3.5 Prediction Visualization

In Figure 7, an exemplary SARM-based prediction is shown for the USP1 assay (AID: 504865). The SARM contains seven

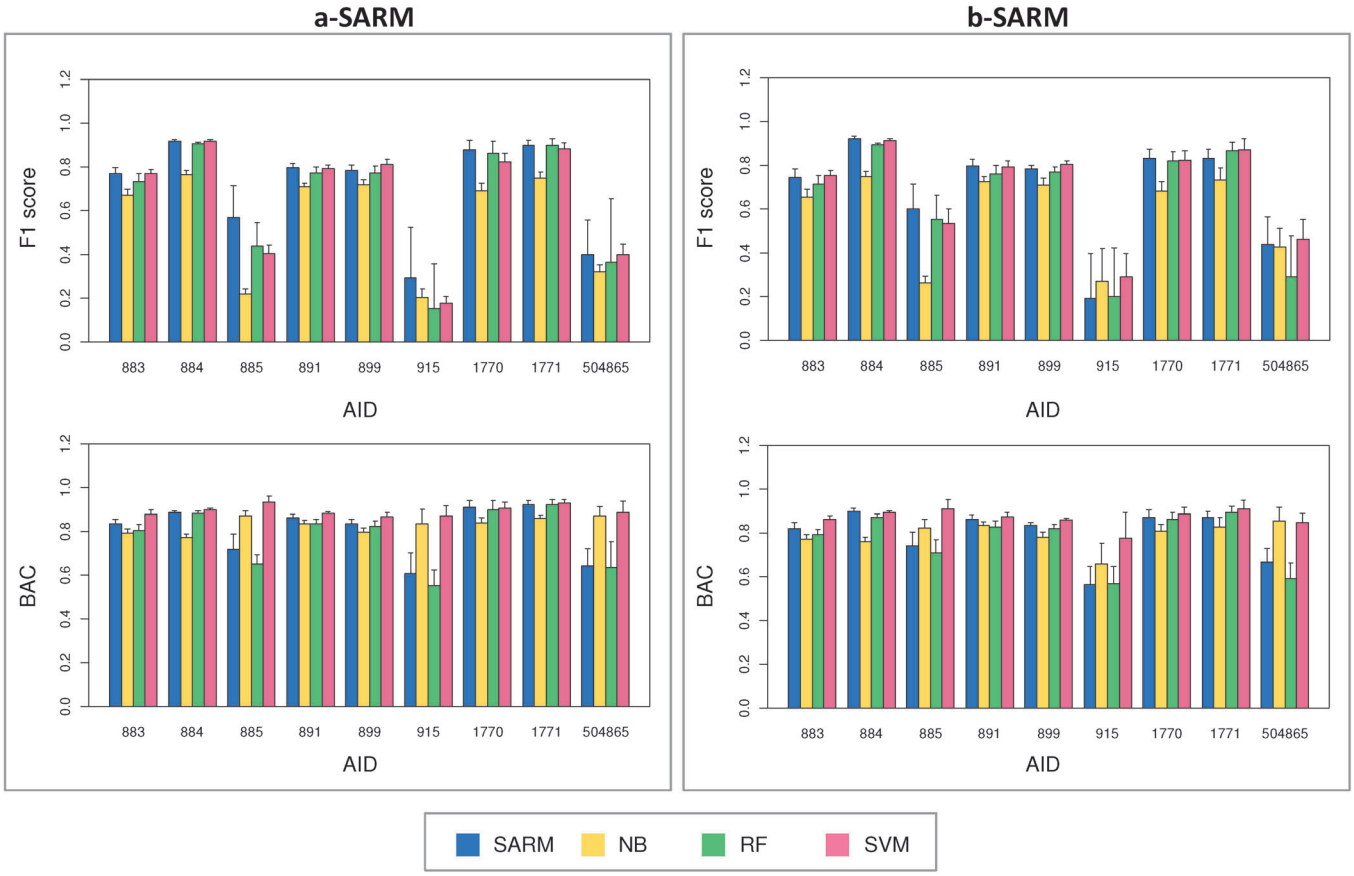


Figure 6. Prediction performance for SARM subsets. Average F1 scores (top) and balanced accuracy (BAC, bottom) are reported for SARM, NB, RF, and SVM predictions on all assays for **(a)** a-SARM and **(b)** b-SARM compound subsets. Vertical lines give standard deviations over multiple predictions. Assay IDs are provided according to Table 1.

structurally related cores and 14 value fragments, which form 60 analogs (filled cells), 12 of which were randomly selected as test instances. In addition, the SARM contains 38 virtual compounds (that would be prediction targets in practical applications). Individual calculated p_x values are given for each test compound. All inactive test compounds were correctly predicted with p_x values ranging from 0.0 to 0.07. In addition, one of two active test compounds was correctly predicted with a p_x of 0.72, whereas the other was assigned a p_x of 0.03, which yielded a false negative prediction. The corresponding updated core and value class contributions for the two active test compounds are also given. The example illustrates the potential of the newly introduced probability-based prediction approach for hit expansion on the basis of SARMs.

4 Conclusions

Herein we have introduced a new methodology for compound activity prediction that derives conditional probabilities of activity from compounds in SARMs and is readily applicable to predict the activity of virtual candidate compounds. The development of the conditional probability-based approach generalizes activity predictions for virtual compounds and is particularly suited for hit expansion. As such, it complements QSAR-type predictions at later stages of compound optimization efforts. The methodological concept of the conditional probability approach has been described in detail. Furthermore, its predictive performance on different screening data sets has been assessed and found to be generally high and comparable to state-of-the-art machine learning approaches (which are mostly used for other compound classification applications). The SARM-based activity prediction method is interpretable and much more intuitive than, for example, support vector machines, yet predicts active compounds with comparable accuracy. Its major goal is the prioritization of virtual candidate compounds in SARMs for hit expansion. The reliable prediction of inactive and active virtual compounds from SARMs, regardless of the presence of SAR continuity or discontinuity, is considered a substantial advance for practical applications.

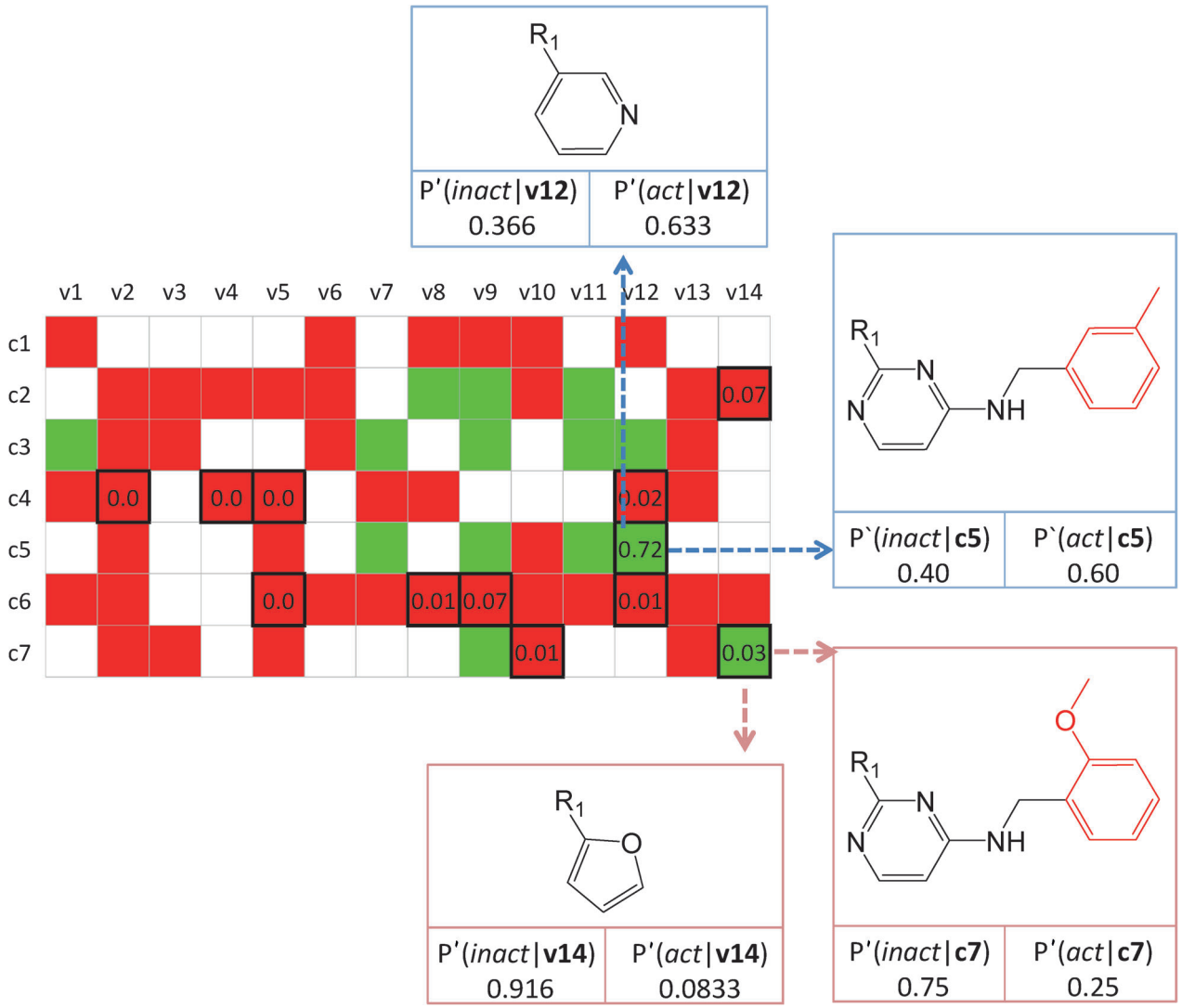


Figure 7. Exemplary SARM prediction. A SARM with seven structurally related cores (c1–c7) and 14 substituents (v1–v14) capturing 60 compounds tested against USP1 is shown. Test compounds are framed in black and the calculated activity probabilities p_x are reported. The corresponding core and value fragments for two active test compounds are shown (indicated by arrows) and their corresponding class likelihoods are reported. Substructures distinguishing the cores are highlighted in red.

Conflict of Interests

No conflict of interests declared.

Acknowledgement

The authors would like to thank *Ye Hu* for help with data sets. DG-O is supported by *Boehringer Ingelheim*.

References

[1] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
[2] D. Gupta-Ostermann, Y. Hu, J. Bajorath, *J. Med. Chem.* **2012**, *55*, 5546–5553.
[3] M. Wawer, E. Lounkine, A. M. Wassermann, J. Bajorath, *Drug Discovery Today* **2010**, *15*, 631–639.
[4] P. W. Kenny, J. Sadowski, in *Chemoinformatics in Drug Discovery* (Ed: T. I. Oprea), Wiley-VCH, Weinheim, Germany, **2005**; pp 271–285.
[5] E. Griffen, A. G. Leach, G. R. Robb, D. J. Warner, *J. Med. Chem.* **2011**, *54*, 7739–7750.
[6] J. Hussain, C. Rea, *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
[7] M. Wawer, J. Bajorath, *J. Med. Chem.* **2011**, *54*, 2944–2951.
[8] A. M. Wassermann, P. Haebel, N. Weskamp, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 1769–1776.
[9] Y. C. Martin, *J. Med. Chem.* **1981**, *24*, 229–237.
[10] E. X. Esposito, A. J. Hopfinger, J. D. Madura, *Meth. Mol. Biol.* **2004**, *275*, 131–214.
[11] D. Gupta-Ostermann, V. Shanmugasundaram, J. Bajorath, *J. Chem. Inf. Model.* **2014**, *54*, 801–809.
[12] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.
[13] H. Kubinyi, *Quant. Struct.-Act. Relat.* **1988**, *7*, 121–133.

[14] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535 – 1535.
[15] D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2012**, *55*, 2932 – 2942.
[16] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte, S. H. Bryant, *Nucleic Acids Res.* **2012**, 40, D400 – D412.
[17] O*EChem*, Version 1.7.7, OpenEye Scientific Software, Inc. Santa Fe, NM, USA, **2012**.
[18] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification,* 2nd ed., Wiley-Interscience, New York, **2000**.
[19] L. Breiman, *Machine Learn.* **2001**, *45*, 5 – 32.
[20] V. N. Vapnik, *The Nature of Statistical Learning Theory,* 2nd ed., Springer, New York, **2000**.
[21] D. Rogers, M. Hahn, *J. Chem. Inf. Model* **2010**, *50*, 742 – 754.
[22] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, *Neural Netw.* **2005**, *18*, 1093 – 1110.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825 – 2830.
[24] A. P. Bradley, *Pattern Recog.* **1997**, *30*, 1145 – 1159.

Received: November 4, 2014
Accepted: December 2, 2014
Published online: January 30, 2015

Summary

The methodology introduced here further expands the utility of the SAR matrix data structure in prioritizing and predicting virtual compounds from screening sets. The methodological concept of the conditional probability utilizes the assumption that either the core, substituent, or their combination might be responsible for the (in)activity. The method performs at par with machine learning methods and has low computational complexity. My contribution to this study has been the design, implementation and analysis of the SARM-based prediction approach.

In this study, benchmark calculations of the SARM-based probability method on publicly available assays are reported. The first prospective application of this approach on a raw screening set is reported in the next Chapter 7.

Chapter 7

Prospective Compound Design using the SAR Matrix-Derived Conditional Probabilities of Activity

Introduction

A collaboration study with PRISM Biolab Corporation is presented in this Chapter. A library of approximately 10,000 compounds is analyzed using the SARMs to identify novel compounds. The compounds represent alpha helical turn mimetics and comprise of well-defined scaffold-substituent patterns. These compounds were screened once at a single concentration to search for new inhibitors of the Wnt/β-catenin pathway.[8]

Prediction methods reported so far in the dissertation provided opportunities to prioritize the resulting virtual compounds. However, in this case, the NBH-based prediction approach (Chapter 5) is difficult to utilize for hit expansion because of the approximate nature of activity annotations obtained from raw screening data. Therefore, the conditional probability-based approach (Chapter 6), which involves the binary classification of compounds into actives and inactives, is utilized. This study is the first prospective application of the

SARM-derived probabilities for prediction and might be of interest for medicinal chemistry applications.

METHOD ARTICLE

REVISED Follow-up: Prospective compound design using the 'SAR Matrix' method and matrix-derived conditional probabilities of activity [v2; ref status: indexed, http://f1000r.es/59v]

Disha Gupta-Ostermann[1], Yoichiro Hirose[2], Takenao Odagami[2], Hiroyuki Kouji[2], Jürgen Bajorath[1]

[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, D-53113, Germany
[2]PRISM BioLab Corporation, Kanagawa, 226-8510, Japan

v2 **First published:** 23 Mar 2015, **4**:75 (doi: 10.12688/f1000research.6271.1)
Latest published: 15 Apr 2015, **4**:75 (doi: 10.12688/f1000research.6271.2)

Abstract

In a previous Method Article, we have presented the 'Structure-Activity Relationship (SAR) Matrix' (SARM) approach. The SARM methodology is designed to systematically extract structurally related compound series from screening or chemical optimization data and organize these series and associated SAR information in matrices reminiscent of R-group tables. SARM calculations also yield many virtual candidate compounds that form a "chemical space envelope" around related series. To further extend the SARM approach, different methods are developed to predict the activity of virtual compounds. In this follow-up contribution, we describe an activity prediction method that derives conditional probabilities of activity from SARMs and report representative results of first prospective applications of this approach.

Open Peer Review

Referee Status: ✓ ✓ ✓ ?

	Invited Referees 1	2	3	4
REVISED **version 2** published 15 Apr 2015				
version 1 published 23 Mar 2015	✓ report	✓ report	✓ report	? report

1 **Hans Matter**, Sanofi-Aventis Deutschland GmbH Germany

2 **Georgia B. McGaughey**, Vertex Pharmaceuticals Inc. USA

3 **Stefan Laufer**, University of Tübingen Germany

4 **Dragos Horvath**, CNRS-Université de Strasbourg France

Discuss this article

Comments (0)

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

How to cite this article: Gupta-Ostermann D, Hirose Y, Odagami T *et al.* **Follow-up: Prospective compound design using the 'SAR Matrix' method and matrix-derived conditional probabilities of activity [v2; ref status: indexed, http://f1000r.es/59v]** *F1000Research* 2015, **4**:75 (doi: 10.12688/f1000research.6271.2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 23 Mar 2015, **4**:75 (doi: 10.12688/f1000research.6271.1)
First indexed: 31 Mar 2015, **4**:75 (doi: 10.12688/f1000research.6271.1)

REVISED **Amendments from Version 1**

We thank all four reviewers for their comments. In our revision, the following points have been addressed.

Georgia B. McGaughey:

An exemplary calculation protocol and SARMs generated from library compounds have been made available for test calculations via a separate open access data deposition. In addition, further methodological explanations have been added to the revision and the similarity of library and virtual candidate compounds has been briefly discussed.

Hans Matter:

Results of the suggested QSAR modeling exercise are summarized in a comment to the review (rather than in the revision) and the similarity of library and predicted compounds has been briefly discussed.

Dragos Horvath:

The description of the conditional probability methodology has been further detailed and formulas have been explained. Furthermore, differences between naïve Bayes modeling and the SARM-based probability approach have been explained in a comment to the review.

Stefan Laufer:

A comment to the review has been added.

See referee reports

Introduction

In recent years, graphical methods have substantially expanded the medicinal chemistry repertoire for analyzing Structure-Activity Relationships (SARs)[1,2]. The development of computational techniques to visualize SAR patterns and identify key compounds has in part been catalyzed by increasing volumes and complexity of activity data in medicinal chemistry. Going beyond a purely descriptive nature of graphical SAR exploration, as exemplified by activity landscape representations[1], the SAR Matrix (SARM) approach[3] was conceptualized to combine large-scale graphical SAR analysis and compound design. SARM calculations generate many virtual compounds (VCs) that populate chemical space around structurally related series. In order to prioritize virtual candidate compounds from SARMs in a target/assay-specific manner, activity prediction methods have been developed including local Quantitative SAR (QSAR) models utilizing compound neighborhood information in SARMs[4] and an approach that derives conditional probabilities of activity from SARMs[5].

In a previous Method Article[6], the SARM methodology and extensions have been described including matrix-based QSAR[4] and navigation of multi-target activity spaces[7]. In this follow-up contribution, we focus on a conditional probability-based approach to activity prediction, which is distinct from QSAR analysis, and report results of first prospective applications. While we are currently unable to disclose the structures of active compounds (due to patent issues of PRISM Biolab Corporation), the prediction statistics and exemplary results we report for an actual drug discovery project should be helpful to put SARM-based predictions into perspective, beyond computational benchmarking, and might spark the interest of practitioners in this field.

Methods

Since details of the SARM methodology and matrix-based QSAR modeling have been presented in the accompanying article[6], we initially provide only brief summaries of these methods, followed by a detailed description of the conditional probability approach.

SAR matrices

To generate SARMs compounds are subjected to a systematic two-step fragmentation procedure yielding matched molecular pairs (MMPs)[8]. An MMP is defined as a pair of compounds that only differ at a single site. In the first step, compounds are fragmented into core structures and substituents. In the second step, resulting core structures are subjected to fragmentation. This two-step fragmentation protocol identifies series of compounds with related core structures (forming "core MMPs"). Series of compounds with cores forming MMP relationships are organized in individual SARMs, as illustrated in Figure 1. Each matrix cell defines a unique combination of a core and substituent (reminiscent of yet distinct from R-group tables). Following MMP terminology, the core is called key fragment and the substituent value fragment[8]. Each filled cell represents an actual compound color-coded by activity or potency and each empty cell a VC representing a previously unexplored core-substituent (key-value) combination. Accordingly, VCs are thought to generate a "chemical space envelope" around structurally related compound series. Depending on the structural relationships that are present within a given compound set, varying numbers of SARMs are obtained that systematically organize available analog series and provide many VCs for further consideration. The more similar data set compounds are to each other, the more SARMs are typically obtained.

Matrix-based local QSAR models

A compound neighborhood (NBH) approach was developed for potency prediction of VCs based on known potencies of structural analogs[4], as illustrated in Figure 2. A qualifying NBH consists of two known active compounds that contain the key and value fragment of a given VC, respectively (D and G in Figure 2a), and a third active compound (E) that consists of the key of D and value of G. The potency of a VC can then be predicted from its neighbors by applying the additivity assumption underlying Free-Wilson analysis[9] using the equation shown in Figure 2a. For a given VC, all qualifying NBHs are identified across all SARMs, as illustrated in Figure 2b, and for each NBH, an individual potency prediction is carried out using a local "mini-QSAR" model. The average potency over all NBHs is then calculated to yield the final prediction.

The NBH approach is based upon numerical values and thus well suited for potency prediction during compound optimization considering multiple analog series. Principal limitations of QSAR modeling also apply to the NBH methodology, given its Free-Wilson foundation. Hence, meaningful potency predictions can only be expected in the presence of SAR continuity (when small structural changes are accompanied by gradual changes in potency). By contrast, SARMs capturing discontinuous SARs or activity cliffs[10] fall outside the QSAR applicability domain. Because

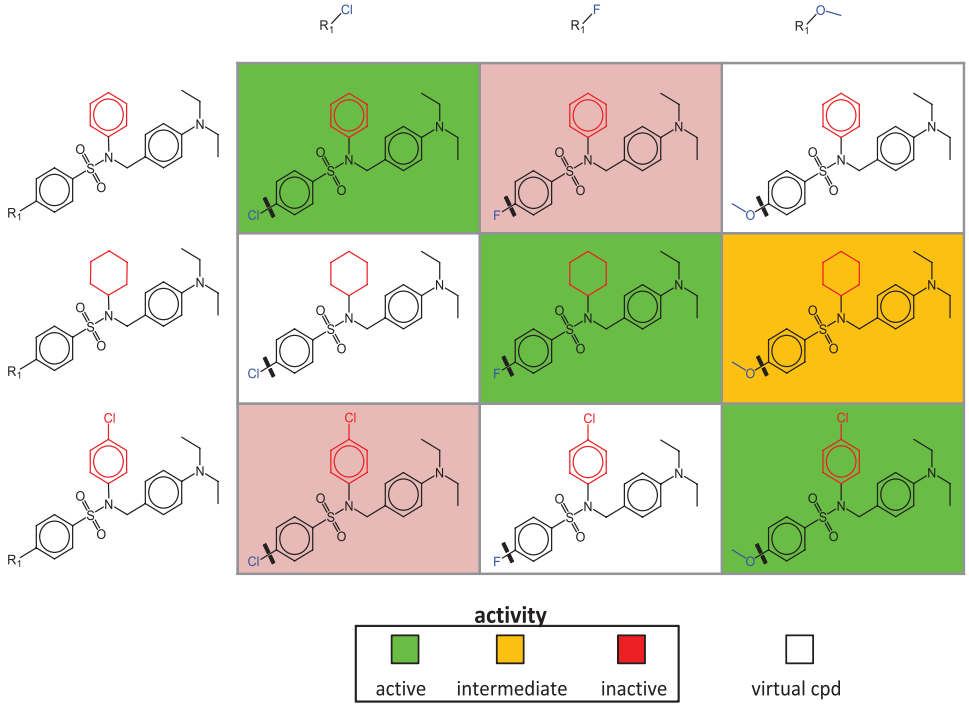


Figure 1. SAR matrix. A schematic representation of a SARM is shown. Compound fragmentation (indicated by thick lines in matrix cells) yields three analog series with structurally related cores (keys). Each series consists of analogs that share a core and differ by a single substituent (value, blue). Structural differences between the cores of the three series are highlighted in red. Each SARM combines all analog series with structurally related cores available in a compound set. Rows and columns represent compounds sharing the same core and substituent, respectively. In each cell, the combination of a core and a substituent defines a unique molecular structure. Compounds present in the data set are represented by filled cells that are color-coded according to activity. In addition, empty cells represent virtual compounds (i.e., previously unexplored key-value combinations resulting from MMP fragmentation).

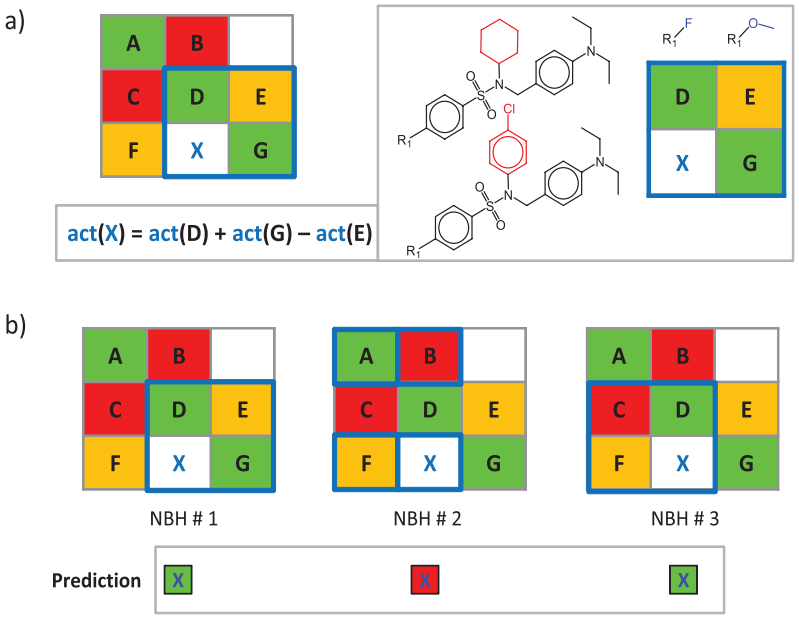


Figure 2. Neighborhood-based activity prediction. (**a**) A NBH of virtual compound X is marked in blue in a model SARM and compounds forming this NBH are displayed. Compounds D and G share the same substituents and core with X, respectively, and the third neighbor E consists of the core of D and substituent of G. At the lower left, the equation to predict the potency of X from the values of D, E, and G is shown. (**b**) The process of NBH mining is illustrated. For X, the set of all qualifying NBHs (marked in blue) in a given SARM are identified and potency values are predicted for individual NBHs (indicated by color-coded squares). "act" stands for activity (in this case, numerical potency values are used).

potency predictions are carried out over multiple NBHs in different SARMs, standard deviations of predictions provide a simple yet effective indicator of prediction reliability. High and low standard deviations indicate the presence of SAR discontinuity and continuity, respectively, for compound subsets involved in the predictions. When standard deviations are low, accurate SARM-based potency predictions can be expected[4].

Predictions based on conditional probabilities of activity

A conceptually different approach was developed for hit expansion from screening data based upon conditional probabilities of activity derived from SARMs, as outlined in Figure 3. In contrast to NBH-based prediction of numerical potency values, the conditional probability method can utilize approximate potency measurements (e.g., primary screening data) leading to a binary classification of

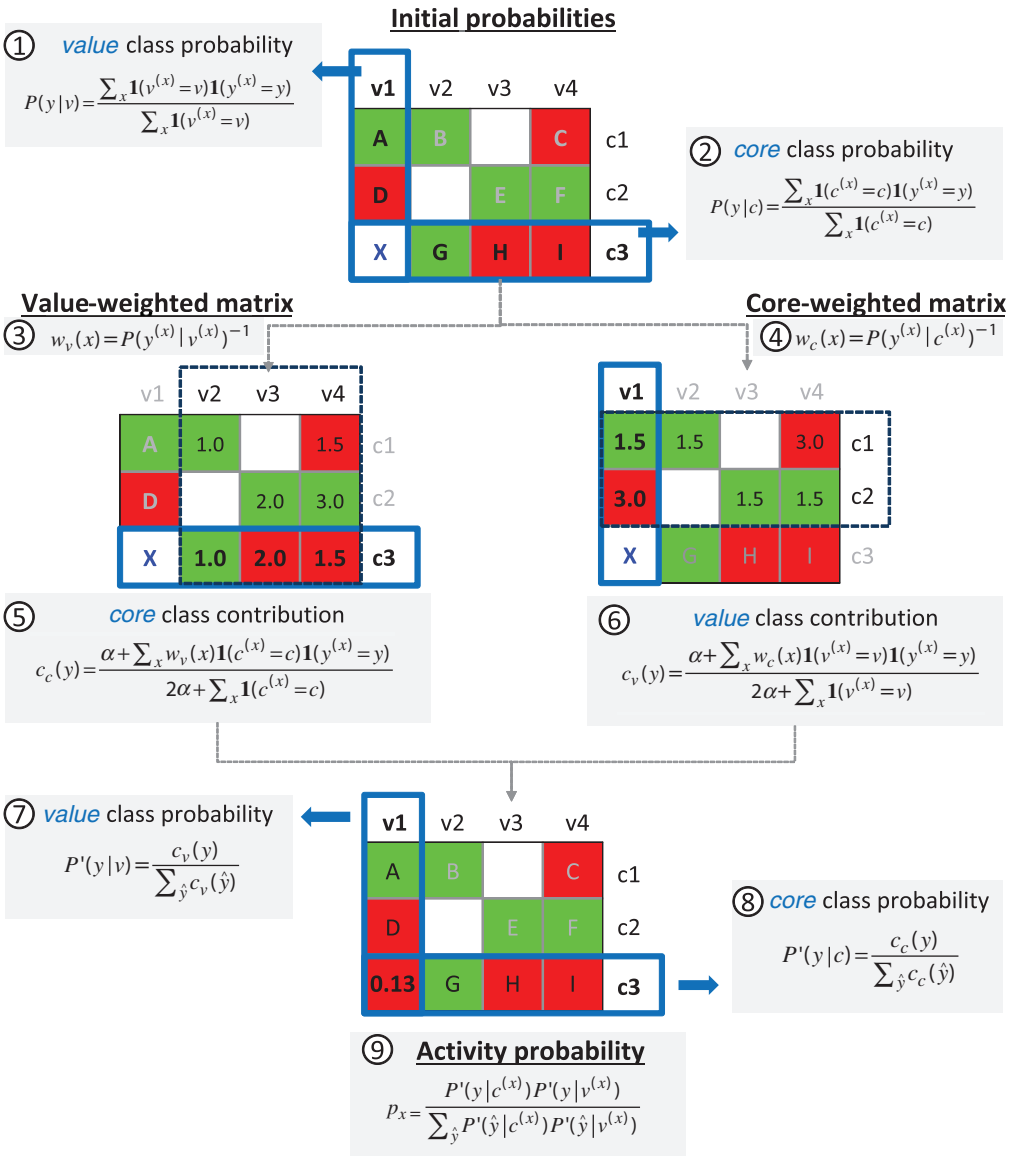


Figure 3. Predictions based on conditional probabilities of activity. Steps and equations required to derive probabilities of activity from SARMs for prediction of virtual compound X are shown using a model SARM with nine compounds (A–I) that contain three cores (*c1–c3*) and four values (*v1–v4*). Matrix cells are color-coded according to compound activity (red, inactive; green, active). In the first step, initial class probabilities are calculated for all cores and values using equation 2 and 1, respectively. Value- and core-weighted matrices are then derived via equations 4 and 3. The class contribution of core *c3* is obtained from the value-weighted matrix using equation 5. Analogously, the class contribution of value *v1* is obtained from the core-weighted matrix using equation 6. The value and core class contributions are then normalized using equations 7 and 8. Finally, the activity probability p_x of 0.13 is obtained for X by combining the normalized core *c3* and value *v1* class probabilities using equation 9.

inactive vs. active data set compounds and ensuing prediction of a probability of activity for VCs.

The conceptual basis of the approach is provided by the following ideas: based on the observed frequency of occurrence of given core and value fragments in active versus inactive compounds (in the following referred to as the active versus inactive class), probabilities of activity and inactivity can be derived for cores and values. Importantly, the contributions of cores and values are thought to be influenced by each other because compounds are represented in SARMs as combinations of individual core and value fragments. Considering the conditional nature of core and value contributions to activity, initial probabilities are weighted to derive class probabilities for any core and value. For a given VC, probabilities of its core and value are then combined to yield a final probability of activity.

Key steps of the methodology are summarized in Figure 3 (and for each step, the respective equation is provided). To illustrate the approach in an intuitive manner, we will go through an exemplary probability calculation for a given VC, guided by Figure 3.

Core and value class probabilities

The SARM in Figure 3 contains nine compounds (A–I) that comprise three cores (c1–c3) and four values (v1–v4). The probability of activity will be predicted for virtual compound X that shares core *c3* with compounds G, H, and I and value *v1* with compounds A and D.

Given the distribution of individual values v and cores c in active and inactive compounds, probabilities of activity and inactivity are calculated using equation 1 and equation 2. Here, $P(y|v)$ and $P(y|c)$ are the conditional probabilities that describe how likely it is to observe a given specific class $y \in \{\text{active, inactive}\}$ for a value v and a core c, respectively. If $c^{(x)}$, $v^{(x)}$, $y^{(x)}$ is the core, value, and class of a given compound x, we can express the conditional probabilities as the fraction of compounds with a core c or value v and class y relative to all compounds containing this core or value. In case of value *v1*, both class probabilities are equal (i.e., 1/2) because *v1* is contained in one active and one inactive compound. By contrast, the probability of inactivity is two times higher for core *c3* than its probability of activity (2/3 vs. 1/3).

Core- and value-weighted matrices

These initial estimates are further refined by taking information from all SARM compounds into account. For this, the inverse of value and core class probabilities is used to derive the *value-weighted matrix* and *core-weighted matrix*, respectively. In case of the value-weighted matrix, the inverse class probabilities of the values are mapped to the compounds that represent the corresponding value and class. Analogously, the core-weighted matrix is derived by mapping the inverse class probabilities of the cores to the compounds that represent the corresponding core and class. The value-weighted matrix results from the assignment of a weight to each compound using equation 3 and the core-weighted matrix is obtained using equation 4.

Refinement of core and value class contributions

In this step, core probabilities using value-weighted matrices and value probabilities using core-weighted matrices are derived. The underlying idea is to statistically assess if a core or value contributes more to activity or inactivity. This rationalizes the calculation of weights from the previous step: the less frequently observed class for a core or value is assigned a higher weight, which leads to a larger class contribution of the corresponding value or core of a compound, respectively. For example, the class probability of core *c3* is updated by considering information from values *v2*, *v3*, and *v4* in compound G, H, and I, respectively. All compounds containing value *v2* are active (2/2); hence, the core class probabilities of compounds B and G are assigned a weight of 1.0 (through value-weighting). For value *v3*, the compounds show equal class frequency of (in)activity (1/2); thus, both active and inactive compounds are assigned the same weight of 2.0. Finally, two of three compounds containing value *v4* are inactive. Accordingly, inactive compound I receives a lower weight of 1.5 indicating that its inactivity is more likely due to *v4*. It follows that with increasing frequency of inactivity for a given value, core weights of inactive compounds decrease (and *vice versa*), indicating that the value is likely to be responsible for inactivity. Analogous considerations apply to assess probabilities of activity.

From the value-weighted matrix, core class contributions are calculated with equation 5. For core *c3*, contributions of 0.34 and 1.12 to activity and inactivity are obtained, respectively, using a smoothing factor of α=0.1 (this factor is applied to prevent zero probabilities when no compound is available to represent a possible core or value class):

$$C_{c3}(act) = \frac{0.1 + 1.0}{0.2 + 3} = 0.34$$

$$C_{c3}(inact) = \frac{0.1 + 2.0 + 1.5}{0.2 + 3} = 1.12$$

Through normalization using equation 7 core class probabilities between 0 and 1 are obtained; for *c3* values of 0.23 (activity) and 0.76 (inactivity).

Analogously, value class probabilities are refined using the core-weighted matrix (generated using equation 4). For example, class probabilities of value *v1* are adjusted by considering information from cores *c1* and *c2* in compounds A and D that contain *v1*. Compound A is active and belongs to the majority class of *c1* and is thus assigned a lower weight than D, which is inactive and belongs to the minority class of *c2*. The higher weight assigned to compound D means that its inactivity is statistically more likely to result from value *v1* than core *c2*. Weighted value class contributions calculated using equation 6 give activity and inactivity contributions of 0.72 and 1.40, respectively, for value *v1* (applying a smoothing factor of α=0.1):

$$C_{v1}(act) = \frac{0.1 + 1.5}{0.2 + 2} = 0.72$$

$$C_{v1}(inact) = \frac{0.1 + 3.0}{0.2 + 2} = 1.40$$

Normalization using equation 8 then yields updated *v1* class probabilities of 0.34 (activity) and 0.66 (inactivity).

Combined activity probability

Finally, the normalized core and value probabilities are combined via equation 9 yielding an activity probability p_x (ranging from 0 to 1) for any core-value combination representing a VC. Increasing p_x values indicate an increasing probability of activity. For classification, a threshold value of activity must be set (e.g., 0.5). In our example, the normalized core and value class probabilities for *c3* and *v1* result in an activity probability p_x of 0.13 for virtual compound X representing this core-value combination. Thus, given the low probability of activity, this VC is predicted to be inactive. In benchmark calculations on sets of known active and inactive compounds, conditional probability calculations yielded reasonably accurate predictions of activity, at least comparable to current state-of the-art machine learning approaches[5].

Because the conditional probability approach is statistically grounded, prediction accuracy is expected to increase with sample sizes and matrix density[6]. Therefore, it makes sense to exclude SARMs from the calculations that contain only a small number of data set compounds or have limited row overlap (accounting for shared substitution patterns among structurally related series)[6]. Accordingly, SARMs with more than 50% row overlap are typically considered informative and prioritized for probability calculations.

Different from the NBH approach, the conditional probability method is generally applicable and not confined to compound subsets representing continuous SARs. Thus, QSAR applicability domain restrictions do not apply in this case.

Application

The conditional probability method has been used for activity predictions (hit expansion) starting from the results of a screen of the PRISM library of alpha helical turn mimetics[11,12] carried out in search of new inhibitors of the Wnt/β-catenin protein-protein interaction and pathway[13,14]. The Wnt pathway is implicated in a variety of disease states including several forms of cancer. Consequently, inhibitors of the Wnt/β-catenin interaction are thought to have high therapeutic potential[13,14]. PRISM's current helix mimetics library contains more than 10,000 small molecules with closely related scaffolds[11,12] suitable for SARM analysis. These compounds are analogs containing closely related scaffolds with three substitution sites each. The library screen was carried out using a luciferase reporter gene assay of the Wnt pathway[15,16] and the stably transfected cell line Hek-293, STF1.1[11]. Figure 4 summarizes SARM analysis of the library and activity predictions. The library contained a total of 10,540 compounds that yielded 11,033 stereochemistry-sensitive SARMs (i.e., matrices explicitly accounting for all stereoisomers) with a total of 231,143 VCs. This matrix distribution was solely determined by structural relationships between library compounds.

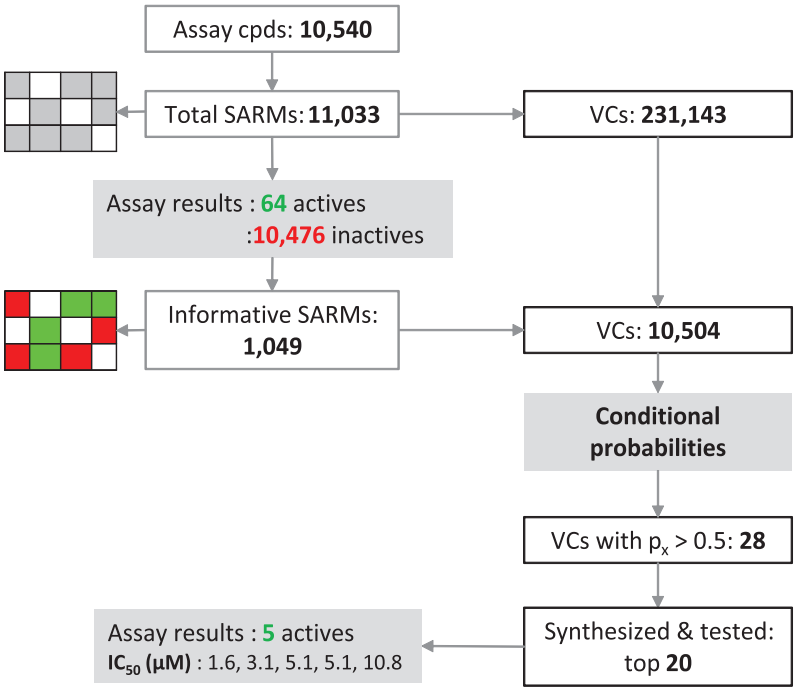


Figure 4. SAR matrix and prediction statistics. SAR matrix statistics for a library of alpha helical turn mimetics are provided and activity predictions for virtual compounds are summarized. For these predictions, conditional probabilities of activity were derived from library screening data.

Screening of the library in the reporter gene assay yielded 64 active and 10,476 inactive compounds (applying a threshold of less than 50% residual luciferase activity). Hence, only a limited number of compounds were classified as active applying this threshold. Active and inactive compounds were then mapped to SARMs and a subset of 1,049 informative matrices (with at least 50% row overlap) was selected that contained 10,504 VCs. Probability calculations predicted 28 VCs to be active. Twenty candidates were synthesized, re-screened, and tested in confirmatory assays, leading to the identification of five novel hits with activities in the low-micromolar range. These five novel actives were, by design, analogs of library compounds having previously unconsidered substitution patterns involving two different sites.

Data availability

In a deposition on the open access ZENODO platform[17], the following data have been made available. Detailed probability calculations for the matrix in Figure 3 are provided in an excel sheet. Furthermore, SARMs generated from the PRISM library on which the calculations were based are made available without compound structures (compounds are represented by unique identification). On the basis of these SARMs, the predictions can be fully reproduced.

Concluding remarks

In this contribution, we have discussed methodological advances for activity prediction on the basis of SARMs, which systematically account for structural/analog relationships in compound sets of any source, organize structurally related compound series, and yield virtual candidate compounds. In combination with the SAR matrix method, compound neighborhood analysis based upon Free-Wilson principles and derivation of conditional probabilities of activity are applicable to predict novel active compounds at different stages of chemical optimization efforts. The conditional probability approach detailed herein is particularly suitable for hit expansion and can be applied to raw screening data. Going beyond benchmark calculations, first prospective applications have yielded promising results. For example, screening data of the PRISM library of helix mimetics made it possible to prioritize a small number of candidate compounds for synthesis from a pool of ~10,000 preselected VCs on the basis of only 64 preliminary screening hits. These predictions ultimately resulted in the identification of five new active compounds by considering only 20 candidates. These compounds provide new starting points for chemical optimization efforts. Of course, further prospective validation studies will need to be performed to better understand the performance of SARM-based activity predictions for different compound classes, targets, and screening assays. However, considering the well-defined scaffold-substituent patterns of compounds representing alpha helical turn mimetics and the systematic design of the library, which plays into the strength of the SARM approach, successful activity predictions are also anticipated for library screens using assay systems and targets engaged in other therapeutically relevant protein-protein interactions.

Author contributions

JB conceived the study and DGO carried out SARM analyses and activity predictions. YO and TO synthesized candidate compounds and collected assays data. YO, TO, and HK analyzed the experimental data. JB wrote the manuscript, JB and DGO designed and generated display items, and all authors examined the manuscript and agreed to its final content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

The authors thank Dr. Anne Mai Wassermann, Dr. Dilyana Dimova, Dr. Preeti Iyer, and Jenny Balfer for valuable contributions to the development of the SARM approach and activity prediction methods. DGO gratefully acknowledges support of doctoral studies from Boehringer Ingelheim.

References

1. Wassermann AM, Wawer M, Bajorath J: **Activity landscape representations for structure-activity relationship analysis.** *J Med Chem.* 2010; **53**(23): 8209–8223.
PubMed Abstract | Publisher Full Text
2. Stumpfe D, Bajorath J: **Methods for SAR visualization.** *RSC Adv.* 2012; **2**(2): 369–378.
Publisher Full Text
3. Wassermann AM, Haebel P, Weskamp N, *et al.*: **SAR matrices: automated extraction of information-rich SAR tables from large compound data sets.** *J Chem Inf Model.* 2012; **52**(7): 1769–1776.
PubMed Abstract | Publisher Full Text
4. Gupta-Ostermann D, Shanmugasundaram V, Bajorath J: **Neighborhood-based prediction of novel active compounds from SAR matrices.** *J Chem Inf Model.* 2014; **54**(3): 801–809.
PubMed Abstract | Publisher Full Text
5. Gupta-Ostermann D, Balfer J, Bajorath J: **Hit expansion from screening data based upon conditional probabilities of activity derived from SAR matrices.** *Mol Inf.* 2015; **34**(2–3): 134–146.
Publisher Full Text
6. Gupta-Ostermann D, Bajorath J: **The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics [v2; ref status: indexed, http://f1000r.es/3rg].** *F1000Res.* 2014; **3**: 113.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Gupta-Ostermann D, Hu Y, Bajorath J: **Systematic mining of analog series with related core structures in multi-target activity space.** *J Comput Aided Mol Des.* 2013; **27**(8): 665–674.
PubMed Abstract | Publisher Full Text
8. Hussain J, Rea C: **Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets.** *J Chem Inf Model.* 2010; **50**(3): 339–348.
PubMed Abstract | Publisher Full Text
9. Kubinyi H: **Free Wilson analysis. Theory, applications and its relationships to**

Hansch analysis. *Quant Struct-Act Relat.* 1988; **7**(3): 121–133.
Publisher Full Text

10. Stumpfe D, Bajorath J: **Exploring activity cliffs in medicinal chemistry.** *J Med Chem.* 2012; **55**(7): 2932–2942.
PubMed Abstract | Publisher Full Text

11. Kouji H, Kogami Y, Odagami T: **Alpha helix mimetic compositions for treating cancer and other CBP/catenin-mediated diseases and conditions**. US 8691819 B2, 2014.
Reference Source

12. Odagami T, Kogami Y, Kouji H: **Alpha helix mimetics and methods thereto**. WO 2010128685 A1, 2010; US 20120088770 A1, 2012.
Reference Source

13. Moon RT, Kohn AD, De Ferrari GV, *et al.*: **WNT and beta-catenin signalling: diseases and therapies.** *Nat Rev Genet.* 2004; **5**(9): 691–701.
PubMed Abstract | Publisher Full Text

14. Klaus A, Birchmeier W: **Wnt signalling and its impact on development and cancer.** *Nat Rev Cancer.* 2008; **8**(5): 387–398.
PubMed Abstract | Publisher Full Text

15. Molenaar M, van de Wetering M, Oosterwegel M, *et al.*: **XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos.** *Cell.* 1996; **86**(3): 391–399.
PubMed Abstract | Publisher Full Text

16. Veeman MT, Slusarski DC, Kaykas A, *et al.*: **Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements.** *Curr Biol.* 2003; **13**(8): 680–685.
PubMed Abstract | Publisher Full Text

17. Gupta-Ostermann D, Hirose Y, Odagami T, *et al.*: **Follow-up: Prospective compound design using the 'SAR Matrix' method and matrix-derived conditional probabilities of activity.** *Zenodo.* 2015.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓ ?

Version 1

Referee Report 08 April 2015

doi:10.5256/f1000research.6727.r8067

Dragos Horvath
Laboratoire de Chémoinformatique and Laboratoire d'Infochimie, UMR 7140 CNRS (LCS), CNRS-Université de Strasbourg, Strasbourg, France

This is an interesting upgrade of the SARM methodology, now endeavored with a probability-driven activity prediction tool described in this paper. Unfortunately, I cannot recommend indexation as is, because the methodology is not comprehensively described: some formulae embedded in a Figure are never rigorously explained, except by means of some hand-waiving example. Therefore, I (hope I) got the principle of the method - looks very much like naive Bayes to me. If so - does it do better than standard naive Bayes, with some fragment count descriptors? Honestly, I could not write a piece of code implementing it on the basis of what is said in the paper. Don't understand cryptic annotations like 1(v(x)=v) - suppose it's some Kronecker delta symbol 'sum only over lines with v(x)=v'... but, by the way, what is 'x'? Molecules? Matrix columns? Rows? This is the best-kept secret of the publication...

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 09 Apr 2015

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

The SARM-based probability approach and Naïve Bayesian (NB) classification are both probabilistic in nature and attempt to estimate the posterior probability of a given compound *x* to belong to a class *y,* i.e. $P(y \mid x)$. However, following Bayes' theorem, NB modeling derives the posterior from the prior and class likelihood. The class likelihood, $P(x \mid y)$, is estimated from the data. By contrast, the SARM-based approach assigns weights to the cores (keys) based on substituents (values) and vice versa. This is done under the assumption that either cores, values, or their combination might be responsible for the activity. Thus, the approach estimates the posterior from the data and then applies a re-weighting (refinement) scheme by calculating core and value class contributions.
All methodological details of the SARM-based probability of activity approach are provided in reference 5 of the paper.

Competing Interests: None

Referee Report 31 March 2015

doi:10.5256/f1000research.6727.r8160

Stefan Laufer
University of Tübingen, Tübingen, Germany

This Method article mostly describes an extension of the SAR Matrix approach to predict active compounds from many virtual candidates that are contained in matrices derived from compound libraries.

The new activity prediction method is generally applicable to screening data to facilitate hit expand and can make use of approximate activity measurements such as % inhibition. This would be attractive in practice.

The conditional probability method, which was first published in an informatics journal, is not trivial and probably not easy to understand for many medicinal chemists.

Therefore, the authors were obviously motivated to make this prediction methodology accessible to wider audience in screening and medicinal chemistry. They have done so by going step by step through exemplary calculations that illustrate ideas behind this approach and show how active compounds are predicted.

In addition, they report first practical applications that should make this method in combination with the SAR matrix structure attractive to many.

Although the application on a library of helix mimetics is essentially proprietary (structures of active compounds cannot be shown), the statistics of the predictions are interesting. Of thousands of virtual compounds the SAR matrix approach generates for this library, only 28 were predicted to be active using the new method when processing reporter gene assay data probing the Wnt pathway. Twenty of these compounds were synthesized and tested and 5 new hits were identified with low-micromolar potency. Clearly, if the combined SAR matrix / activity prediction approach produces similar results in additional applications of this library or other screening libraries, it will be
rather useful for hit expansion.

Taken together, the authors have attempted to make a relatively complex computational approach easier to appreciate by a screening or chemistry audience by providing easy to follow examples and practical applications. This could hardly be accomplished in a specialized computational journal.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 09 Apr 2015

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

We thank the reviewer for pointing at a primary motivation for this publication and for emphasizing accessibility to a wider than an expert computational audience.

Competing Interests: NoneNone

Referee Report 31 March 2015

doi:10.5256/f1000research.6727.r8069

Georgia B. McGaughey
Vertex Pharmaceuticals Inc., Cambridge, MA, USA

Gupta-Ostermann's "follow-up" manuscript is well written and clearly laid out. I only have a few (minor) recommendations, which I believe would help readers more easily replicate their work.

The added value of this manuscript lies in figure 3 where "conditional probabilities of activity" are explained. The authors have explained conditional probabilities with figures, text and associated mathematical equations and have even gone so far as to carry out the math for the weighted core class contributions. For interested readers who want to implement the conditional probabilities concept in their own research, I highly suggest that real (or toy) data be included, in the very least, as supplemental material with all the data completely worked out, not just the weighted core class contributions. This would allow one to implement the concept, carry out the math and compare the results to the published results more easily. Additionally, although text is included to explain conditional probabilities, I found myself having to read this section a few times to fully understand the clear impact this method could have. I think this section needs to be expanded with more text.

Finally, although it is understandable that the work carried out herein with PRISM BioLab Corporation, is proprietary, it is unfortunate that more information regarding the "twenty synthesized candidates" can not be elaborated upon. Any information regarding the similarity of these compounds to the actives (or even the similarity range of the actives themselves) would be informative.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 31 March 2015

doi:10.5256/f1000research.6727.r8159

Hans Matter
Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

This interesting contribution by Bajorath *et al.* nicely extends the idea of graphical methods for SAR analysis in computational medicinal chemistry. The SARM method was shown earlier to capture SAR information from larger collections by matched molecular pairs (MMPs) and to present it in an intuitive way. Furthermore the combination of large-scale SAR analysis with virtual compounds allows guiding synthesis to explore straightforward ideas as direct outcome of SAR interpretation. Therefore this approach is attractive to rapidly identify activity trends and cliffs.

The paper reports a conditional probability-based approach to activity prediction from SAR knowledge. Such a conditional probability measures the probability of activity for one compound given that a structurally related compound was active. Individual probabilities are extracted from rows and columns in the underlying SARMs. While such a probabilistic approach only works for SARMs, which are sufficiently populated and have shared substitution pattern, the approach is not restricted to compound subsets representing continuous SAR only.
The prospective application of this interesting concept suffers from the lack of chemical structures, so that the degree of similarity between actives and follow-up design cannot be assessed. Furthermore the description of the HTS assay, substructure alerts, additional filtering, assay validation and retesting rates, compound QCs for actives is missing. This makes it difficult to evaluate the true HTS outcome using potentially noisy data for such a challenging PPI target.

To illustrate the value of the novel activity estimation approach from matrices, it might be useful constructing a standard 2D-QSAR model and check is for predictivity of the synthesized top-20 design proposals in comparison to the matrix-derived conditional probability. It might be of interest to see, how robust both approaches work with noisy primary screening data.

The manuscript title and abstract cover the content well. The chemoinformatics approach is clearly described and can most likely be reproduced. As this is not the case for the HTS actives and the assays for this study, the results will be difficult to reproduce. The authors might also want to mention, whether software tools from their study are available to the public.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 09 Apr 2015

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

A conventional QSAR model has been difficult to derive in this case because of the rather approximate nature of activity annotations obtained from raw screeing data. Instead, a cross-validated binary QSAR model has been generated from the screening data using the Molecular Operating Environment (version 2013.08; Chemical Computing Group Inc., Montreal, Canada) and applied to predict the activity state of the 20 test compounds, producing an accuracy of 0.65 for active and inactive compounds.

Competing Interests: None

Summary

This study reports the first prospective application of the SAR matrix-derived conditional probabilities for hit expansion. From the PRISM library of helix mimetics with approximately 10,000 compounds, out of which only 64 were active, a pool of approximately 10,000 VCs was obtained using the SARM method. Predictions on these resulted in the prioritization of 20 VCs, which were synthesized and tested. From these ultimately five new actives were identified. This study demonstrates the successful application of the method for data sets comprising of well-defined scaffold-substituent patterns. Further studies would be required to better understand the performance of SARM-based probability method on data sets representing different compound classes, targets and screening assays. My contribution to this study was to carry out the activity predictions and to analyze the data.

Chapter 8

Conclusions

The major objectives of this dissertation have been the development of computational methods for SAR analysis and activity predictions to aid in prospective compound design. A number of representative studies have been presented.

In the first study, a newly designed activity landscape model, LASSO graph, was introduced that utilizes molecular frameworks to organize compounds hierarchically into sets of scaffolds and cyclic skeletons (CSKs). The design scheme facilitates the "forward-backward" exploration of SARs and reveals signature SAR patterns. The graph topology is compact and shows global and local SAR trends in compound data (Chapter 2).

The remainder of the dissertation was dedicated to develop methodological advancements of the SAR matrix method. Activity landscapes are descriptive in nature. They reveal SAR trends in compound data but do not guide compound design directly. SARMs represent a crucial data structure that expand the chemical space envelope of a compound data, giving rise to various unexplored compounds. These virtual compounds are novel design suggestions and can be prioritized for synthesis and testing. Thus, SARMs provide a close link between descriptive SAR analysis and prospective compound design. New methodologies were incorporated in the SARM method to enhance its applicability in the fields of chemogenomics and medicinal chemistry (Chapter 3).

The aim of the original SARM methodology was large-scale SAR analysis of structurally-related compound series active against a given target. Depart-

ing from SAR analysis, the SARM-based structural organization scheme was adapted for chemogenomics applications, in which compound-target interactions are systematically explored (Chapter 4). These matrices, called the compound series matrices, identified closely related analog series with multi-target activities in the public domain. Compound series matrices are useful in exploring compound promiscuity patterns, thereby aiding in the identification of compounds that are attractive for testing against additional targets. Virtual compounds resulting in these matrices can be useful to design novel compounds with desired activity profiles.

Utilizing matched molecular pair relationships in SARMs, an approach was developed to predict compound activities of virtual compounds (Chapter 5). Here, neighborhoods of virtual compounds were systematically utilized as "mini-QSAR" models for activity prediction. Multiple neighborhoods act as a diagnostic for the local SAR environments of the virtual compounds. The approach resulted in accurate activity predictions for compounds mapping to continuous SAR regions. Compounds mapping to discontinuous SAR regions fall outside the applicability domain of the methodology. This approach is not applicable to screening sets where explicit activity values are not available. Therefore, a conceptually different approach was developed for hit expansion from screening data based upon conditional probabilities of activity derived from SARMs (Chapter 6). The method utilizes a binary classification of inactive vs. active data set compounds to predict probability of activity for virtual compounds. The method performs comparable to state-of-the-art machine learning methods and has low computational complexity. This method expands the utility of the SARMs from hit-to-lead and lead optimization data to screening libraries.

Finally, a prospective application of the conditional probability-based prediction approach on the SARM method is introduced (Chapter 7). The study was carried out on the PRISM library of alpha helical turn mimetics, where well-defined scaffold-substituent patterns existed. Out of approximately 10,000 original compounds with 64 actives, approximately 10,000 virtual compounds were generated and pre-selected. 20 of these were predicted to be active. After synthesis of these 20, five novel actives with IC_{50} values in the micromolar

range were found. This study provides the first prospective application of this method beyond benchmarking.

In conclusion, this dissertation reports novel computational methods for SAR analysis and activity prediction. Major methodological advancements were developed on the SAR matrix method, thereby rendering it highly attractive for practical applications.

Additional References

[1] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2010**, *53*, 8209–8223.

[2] Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.

[3] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *Journal of Chemical Information and Modeling* **2012**, *52*, 1769–1776.

[4] Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discovery Today* **2013**, *18*, 644–650.

[5] Wawer, M.; Bajorath, J. Extraction of Structure-Activity Relationship Information from High-Throughput Screening Data. *Current Medicinal Chemistry* **2009**, *16*, 4049–4057.

[6] Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-Directed Nearest-Neighbor Searching. *Journal of Medicinal Chemistry* **2005**, *48*, 240–248.

[7] Glick, M; Jenkins, J.; Nettles, J. H; Hitchings, H; Davies, J. W Enrichment of High-Throughput Screening Data with Increasing Levels of Noise using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers. *Journal of Chemical Information and Modeling* **2006**, *46*, 193–200.

[8] Moon, R. T.; Kohn, A. D.; De Ferrari, G. V.; Kaykas, A. WNT and β-Catenin Signalling: Diseases and Therapies. *Nature Reviews Genetics* **2004**, *5*, 691–701.

Additional Publications

[4] Hu, Y.; Gupta-Ostermann, D.; Bajorath, J. Exploring Compound Promiscuity Patterns and Multi-Target Activity Spaces. *Computational and Structural Biotechnology Journal* **2014**, *9*, e201401003.

[3] Namasivayam, V.; Gupta-Ostermann, D.; Balfer, J.; Heikamp, K.; Bajorath, J. Prediction of Compounds in Different Local Structure-Activity Relationship Environments Using Emerging Chemical Patterns. *Journal of Chemical Information and Modeling* **2014**, *54*, 1301–1310.

[2] Gupta-Ostermann, D.; Bajorath, J. Identification of Multitarget Activity Ridges in High-Dimensional Bioactivity Spaces. *Journal of Chemical Information and Modeling* **2012**, *52*, 2579–2586.

[1] Gupta-Ostermann, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Graph Mining for SAR Transfer Series. *Journal of Chemical Information and Modeling* **2012**, *52*, 935–942.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Computational Methods for Structure-Activity Relationship Analysis and Activity Prediction" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch an keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nächstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

[1] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO Graph for Compound Data Set Representation and Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2012**, *55*, 5546–5553.

[2] Gupta-Ostermann, D.; Bajorath, J. The 'SAR Matrix' Method and its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, 113.

[3] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Systematic Mining of Analog Series with Related Core Structures in Multi-Target Activity Spaces. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 665–674.

[4] Gupta-Ostermann, D.; Shanmugasundaram, V.; Bajorath, J. Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices. *Journal of Chemical Information and Modeling* **2014**, *54*, 801–809.

[5] Gupta-Ostermann, D.; Balfer, J.; Bajorath, J. Hit Expansion from Screening Data Based upon Conditional Probabilities of Activity Derived from SAR Matrices. *Molecular Informatics* **2015**, *34*, 134–146.

[6] Gupta-Ostermann, D.; Hirose, Y.; Odagami, T.; Kouji, H.; Bajorath, J. Follow-Up: Prospective Compound Design Using the 'SAR Matrix' Method and Matrix-Derived Conditional Probabilities of Activity. *F1000Research* **2015**, *4*, 75.

Disha Gupta-Ostermann

Bonn, 2015