JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

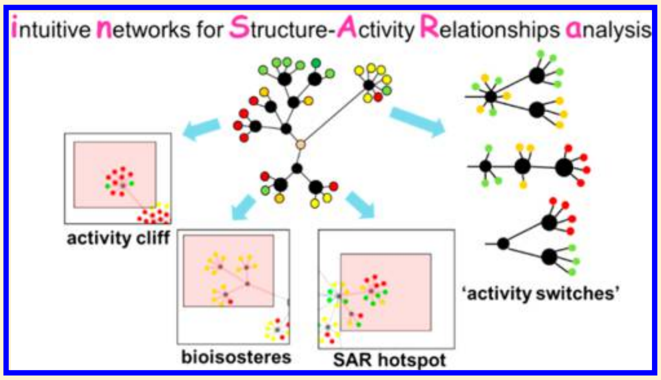
pubs.acs.org/jcim

inSARa: Intuitive and Interactive SAR Interpretation by Reduced Graphs and Hierarchical MCS-Based Network Navigation

Sabrina Wollenhaupt[†] and Knut Baumann*[,†]

[†]Institute of Medicinal and Pharmaceutical Chemistry, University of Technology Braunschweig, Beethovenstrasse 55, 38106 Braunschweig, Germany

Supporting Information


intuitive networks for Structure-Activity Relationships analysis
activity cliff
bioisosteres
SAR hotspot
'activity switches'

ABSTRACT: The analysis of Structure−Activity-Relationships (SAR) of small molecules is a fundamental task in drug discovery. Although a large number of methods are already published, there is still a strong need for novel intuitive approaches. The inSARa (**i**ntuitive **n**etworks for **S**tructure-**A**ctivity **R**elationships **a**nalysis) method introduced herein takes advantage of the synergistic combination of reduced graphs (RG) and the intuitive maximum common substructure (MCS) concept. The main feature of the inSARa concept is a hierarchical network structure of clearly defined substructure relationships based on common pharmacophoric features. Thus, straightforward SAR interpretation is possible by interactive network navigation. When focusing on a set of active molecules at one single target, the resulting inSARa networks are shown to be valuable for various essential tasks in SAR analysis, such as the identification of activity cliffs or "activity switches", bioisosteric exchanges, common pharmacophoric features, or "SAR hotspots".

INTRODUCTION

The analysis of structure−activity relationships (SAR) is ever since an integral component of the drug discovery process. For the medicinal chemist this knowledge is essential at different stages of drug development. It plays an important role not only in the lead or hit to lead optimization process but also in the selection of promising compounds from HTS and denovo-design.

SARs are traditionally manually explored for example by the inspection of so-called R-group tables. However, this way of SAR analysis is not only strongly dependent on the experience of the medicinal chemist, but it is also limited to a series of small size of analogue molecules. SAR maps facilitate this process by enabling interactive exploration of advanced R-group tables.[1] Nevertheless, in the past years a rapid progress in combinatorial chemistry and high-throughput-screening (HTS) was observed leading to a large number of bioactivity data.[2] The organization of these data and the mining of hidden SAR information is one of the major challenges in drug discovery.

The requirements for SAR analysis methods have changed over the past decades. Nowadays, handling a large number of heterogeneous data is a key issue since the analysis of e.g. HTS data is wanted. Moreover, systematically extracting crucial common features and detecting important SAR trends in an automatic manner becomes more and more important. In the past, it was often neglected that medicinal chemists have to understand and accept the way molecules are encoded and compared by a particular method. Therefore, an intuitive, easily interpretable concept with subsequent appealing visualization of the resulting data is a key criterion for broad acceptance of a particular method.

To tackle the mentioned issues, several computational methods aiming at the recognition of SAR patterns and the visualization of SARs and activity landscapes have evolved over the past years.[3] Yet, SAR analysis of small molecules remains a challenging job, and up to now the philosopher's stone has not been found. Therefore, the development of methods with novel characteristics is important.

One key concept in medicinal chemistry is the similar property principle which states that similar molecules tend to have similar (physicochemical) properties including their biological activity.[4] For this assumption to hold, it is of utmost importance to encode molecular similarity in a meaningful way (i.e., the choice of an adequate structural representation and similarity metric is crucial). In the following a brief review of current work in this field is given and contrasted with the novel approach proposed here.

Many of the recent approaches for large-scale SAR analysis and visualization are based on fingerprint-based similarity. This includes, for example, the Network-like Similarity Graphs[5] (NSGs) and extensions for facilitating automated and systematic analysis of large data sets, such as SAR Pathways, Chemical Neighborhood Graphs, and SAR Trees.[6] Further examples are

Received: December 19, 2013
Published: May 21, 2014

the compound-centric Similarity Potency Tree[7] approach and the SALI-Graphs[8] designed for the detection of activity-cliffs. These approaches are often based on a graph or network representation. Recently, this type of representation is increasingly used in drug discovery for the visualization of information because it enables the rapid capture of complex relationships and management of large amounts of data in an intuitive way.[9] SAS maps, another fingerprint-based approach not based on graph representation, characterize and represent 2D activity landscape by relating molecular similarity and activity similarity.[10] As pairwise comparisons are displayed, this approach is less well suited for large-scale analysis.

The advantage of fingerprints is that they are fast to compute and therefore well suited for processing large data sets. Nevertheless, some drawbacks of this widely used concept are the difficulty to figure out the underlying molecular features responsible for the presumed chemical similarity and the high dependence on the fingerprint type used for molecular representation. As the calculated similarity values most often reflect whole-molecule similarity, the resulting relationships are less intuitive. Thus, SAR interpretation may be quite difficult for the medicinal chemist.

More intuitive and easier to grasp approaches use well-defined substructure relationships instead of calculated similarity values.[3] Herein, two promising and related concepts are the maximum common substructure (MCS) and the matched molecular pair (MMP). The MCS is the largest substructure in common between two molecules.[11] A matched molecular pair is defined as a pair of molecules that differ only by small, well-defined, structural changes.[12] That means the two molecules share a large common substructure and the structural change transforming one molecule into the other is rather small, whereas an MCS of two molecules can be of any size. If the MCS size of a pair of molecules exceeds a certain threshold, two molecules can be regarded as MMP.

One recently published graphical approach for SAR analysis based on the MMP concept is the Bipartite Matching Molecular Series Graph (BMMSG).[13] In this approach all MMPs are determined by fragmentation using single, double, and triple cuts, and afterward the MMPs are organized in a network structure. In these networks two node types can be distinguished: Nodes representing the common substructure(s) and molecule nodes connected to them if the substructure is part of the molecule represented by this node. Moreover, hierarchical relationships between different substructures are shown in a separate treelike graph. The BMMSG were shown to be valuable for several tasks in large-scale SAR analysis. However, one drawback of this approach is that it is not able to cope with small variations in the common substructure because substructures are matched exactly without permitting any fuzziness. Therefore, the BMGGS are useful for the analysis of analogue series or parallel series (different scaffold, same substitution pattern).

Another graphical method for SAR analysis, the Combinatorial Analog Graph (CAG), is based on determination of the MCS for molecules with the same scaffold, subsequent R-group decomposition, and discontinuity score calculation for compounds varying in the same substitution site combination.[14] The resulting graphs can be used to identify substitution sites which are crucial for potency modification ("SAR hotspots") or undersampled substitution sites which are not yet explored ("SAR holes"). Yet, the CAG approach is limited to the analysis of one single analogue series, the calculated discontinuity scores lack direct interpretability, and no relationships are established between similar scaffolds.

The inSARa (**i**ntuitive **n**etworks for **S**tructure-**A**ctivity **R**elationships **a**nalysis) approach introduced herein was developed not only to tackle large-scale SAR analysis in a more intuitive way than fingerprint based approaches but also to address some of the problems associated with the methods mentioned above. The method is based on SAR interpretation by network navigation and takes advantage of the synergy resulting from the combination of the reduced graph (RG) and the MCS concept.[15] One key feature of inSARa is that the resulting SAR networks are hierarchically organized and based on the pairwise maximum common substructure in the set of active molecules. The hierarchical structure is analogue to the HierS,[16] Scaffold Tree,[17] layered skeleton−scaffold organization graph,[18] or the maximum common framework[19] approaches which provide a systematic overview about scaffold or framework relationships without considering substitution patterns.

Apart from being intuitive, the MCS concept also shows several drawbacks. Its high computational cost makes the application to larger data sets unfeasible, especially when they consist of analogue molecules with a large number of atoms.[11] If an exact atom match is required, which is commonly the case, very small connected MCSs may result, for example, when comparing compounds with small structural variations (for instance varying linker size) but obviously similar functional units. Chemically less meaningful MCSs can arise if rings are only partially matched.[20] These drawbacks potentially obscure the detection of similarities and SAR interpretation. In order to circumvent the mentioned problems, inSARa uses reduced graphs instead of the original molecule structures for MCS calculation.

Reduced graphs (RGs) reduce groups of atoms of a molecule to single pseudoatoms while retaining the original topology between these functional units. The pseudoatoms represent particular physicochemical properties or pharmacophoric features of the encoded molecule.[21] Hence, RGs provide a larger level of abstraction than e.g. substructure fingerprints. Additionally, RGs only consist of a few pseudoatoms. This enables the tackling of large-scale SAR analysis because the MCS calculations become computationally less demanding.[15] In addition to the gain in performance, the provided insight into similar pharmacophoric features is an important advantage which can be of high value for SAR interpretation, especially for the analysis of more heterogeneous data sets.

Many examples of successful application of reduced graphs in chemoinformatics were reported.[15,21−27] For SAR analysis RGs were successfully used for the automated, SMARTS-based extraction of SAR rules by evolutionary optimization[28] and the identification of bioisosteric exchanges.[29] Moreover, RGs represented as fingerprints were successfully combined with decision trees to build SAR models.[24] Several variants of RG implementations are found in the literature, such as Feature Trees (FT),[26] the Extended reduced Graph (ErG),[25] and different reduced graph approaches which are mainly based on or related to the work carried out at the University of Sheffield.[21,27,29,30]

RGs can be represented in different ways.[22] The inSARa approach is based on graph representation of the RGs.[15] Treating RGs as mathematical graphs has the advantage that graph matching techniques, such as the determination of the maximum common substructure and substructure searching,

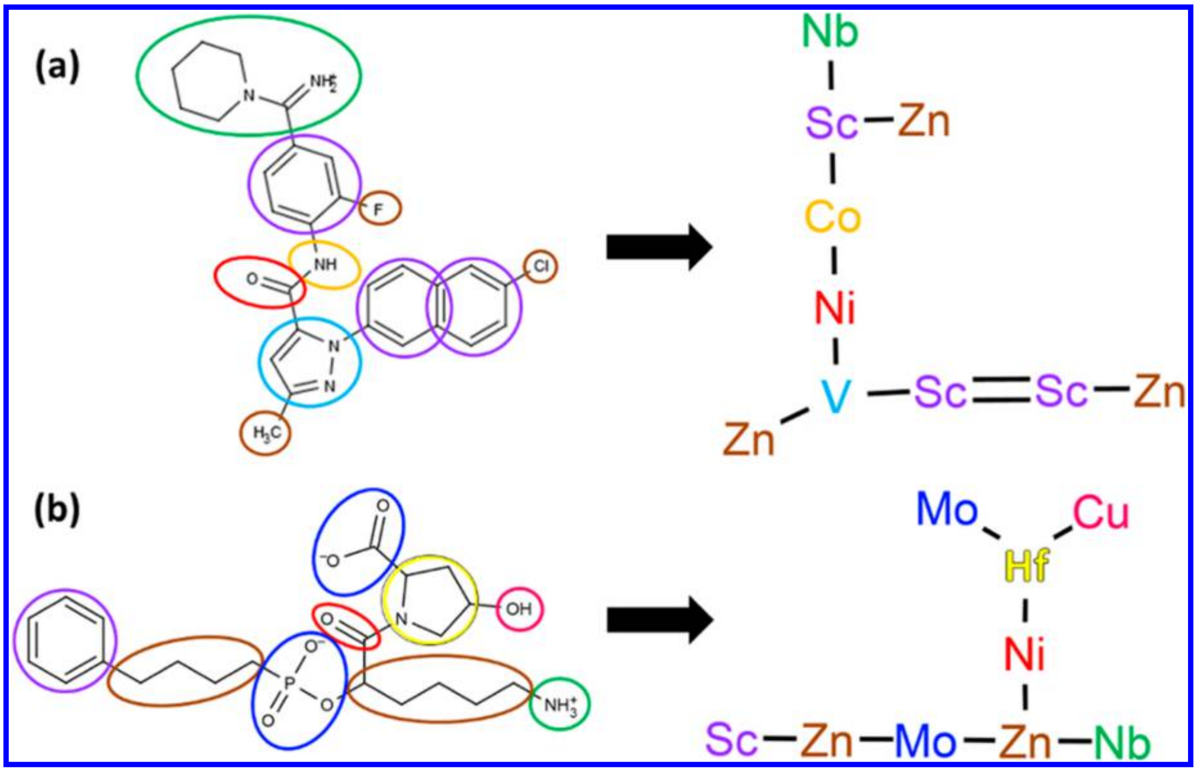


Figure 1. Reduced graph generation illustrated with two different molecules (a) and (b). On the left-hand side the original molecules are given, while on the right-hand side the resulting RGs are shown. The RG node definitions are given in Table 1. The encoded molecular features are encircled and correspond to the pseudoatoms of the same color in the RG.

can be straightforwardly employed. In RGs transition metals are commonly used as pseudoatoms for encoding pharmacophoric features and functional units. This allows the application of standard chemoinformatic software and toolkits for RG generation and handling (e.g., for MCS calculations) or visualization.

In the following section, all necessary steps for the generation of inSARa networks will be described. A summary of the RG generation procedure will be followed by a description of the MCS based network generation and network visualization. Next, parameters for network optimization are discussed. For this, serveral large compound data sets were analyzed. They are assembled from the public domain database BindingDB and reflect targets from different target classes. Additionally, a method to improve the target specificity of inSARa networks is outlined which is based on blacklisting unspecific MCS. Afterward, rules for facilitating interactive SAR analysis and interpretation are proposed. Finally, the application of inSARa to a large Factor Xa data set is described in detail, and a comparison with fingerprint based similarity networks is carried out.

■ MATERIALS AND METHODS

Generation of Reduced Graphs. Different graph reduction schemes were extensively tested. The approach described in the following yielded the best overall results. However, the exact reduction scheme is a tuning parameter and can be changed by the user to fine-tune the method since it will not be possible to find a universally optimal definition. inSARa networks can be generated with all kinds of reduced graphs provided by the user unless a graph representation is used. By adapting the SMARTS[31] definition used for feature recognition, the reduced graph can directly be tuned by the user.

The current RG implementation is based on a modified version of the Ar/F(4) RG definition of Gillet et al.[21] which Barker et al. extended with negatively ionizable (NI) (acid) and positively ionizable (PI) (basic) feature nodes.[30] This reduction scheme is also used by Gardiner et al.[15] In order to avoid having too many or too specific node types PI and NI features are not combined with structure type annotations (acyclic (Ac), aliphatic ring (R), aromatic ring (r)) in this reduction scheme. This means that the whole ring is encoded as NI or PI when an acidic or basic feature is found in the ring system as illustrated in Figure 1a. Hence, it cannot be traced back if this NI or PI node encodes an acyclic or cyclic part of the molecule from the pseudoatom code. The pseudoatom code proposed by Harper et al.[27] (Table 1) is used and adapted for PI and NI features. That means that Nb encodes all PI features and Mo encodes all NI features. Linkers, which are defined as nonterminal acyclic atoms without NI/PI feature or hydrogen bonding capability, are explicitly encoded in the RG as well as terminal nonfeature groups (e.g., terminal alkyl groups or halogen containing groups).[30] The same pseudoatom (Zn) which also represents linker atoms is used. This leads to a higher level of abstraction and reduces the number of RG node types. All allowed combinations of pharmacophoric features can be found in Table 1 which provides a complete list of possible RG nodes together with their codes and abbreviations. The latter will be used below when molecules and reduced graphs are shown together in figures as mnemonic. Pharmacophoric features are assigned by SMARTS matching. The encoded molecular features are listed in Table 2. They are adapted from Gillet et al.,[21] Greene et al.,[32] Taminau et al.,[33] and Zuccotto.[34]

Table 1. RG Pseudoatoms and Encoded Pharmacophoric Features with Their Respective Abbreviations[a]

feature definition	pseudoatom code in RG
– positively ionizable (PI)	Nb
– negatively ionizable (NI)	Mo
(1) **Ring**	
a. aromatic (r)	
– H-bond-acceptor (HBA)	V
– H-bond-donor (HBD)	Ti
– joint H-bond-acceptor and -donor (HBAD)	Cr
– featureless	Sc
b. aliphatic (R)	
– H-bond-acceptor (HBA)	W
– H-bond-donor (HBD)	Ta
– joint H-bond-acceptor and -donor (HBAD)	Re
– featureless	Hf
(2) **Acyclic (Ac)**	
– H-bond-acceptor (HBA)	Ni
– H-bond-donor (HBD)	Co
– joint H-bond-acceptor and -donor (HBAD)	Cu
– featureless (linker or terminal group)	Zn

[a] Adapted from Harper et al.[27]

The RG generation scheme as described in the following is illustrated in Figure 1. First, positively and negatively ionizable features are identified to ensure that for example tetrazoles are recognized as negatively ionizable. Then the cyclic and acyclic parts of the molecule are determined. For ring perception, Figueras's smallest set of smallest rings (SSSR) algorithm[35] as implemented in Open Babel[36] is applied. Cyclic parts are subdivided into aliphatic and aromatic ring systems. For aromaticity perception OEChem's default aromaticity model is used.[37] Following the reduction scheme of Barker et al. fused ring systems are encoded as multiple RG nodes.[30] They are represented in the RG by a double bond between two ring-pseudoatoms. A single bond between two pseudoatoms indicates that the atoms encoded by these two nodes are connected in the original molecule. Only molecules with ring systems consisting of seven ring atoms or less (adopted from Stiefl et al.[25]) are considered in graph reduction. Macrocyclic molecules which represent only a very small portion of molecules in most analyzed target classes are excluded. Hydrogen-bonding capabilities are classified in H-bond-acceptor (HBA), H-bond-donor (HBD), and joint H-bond-acceptor and -donor (HBAD). If more than one pharmacophoric property can be assigned to a node, the following priority order as suggested by Harper et al. is taken into account: PI is preferred to NI, whereas NI is privileged to any hydrogen-bonding feature (HBA, HBD, or HBAD).[27]

Stereochemistry information is lost during RG generation. Therefore, differences in SARs which are based on stereoisomers cannot be distinguished and can lead to activity cliffs. An example will be discussed in detail in the "Results and Discussion" section below.

The whole RG generation process is implemented in Python using routines from the OEChem TK[37] and Open Babel[36] (SSSR).

MCS Based Network Generation. After RG generation, the inSARa networks are generated as described in the following part. No activity information is used for network generation. That is to say that the inSARa networks are unsupervisedly generated.

Table 2. Molecular Features and Functional Groups Encoded by the SMARTS Patterns Used for Pharmacophoric Feature Perception in RG Generation[a]

negatively ionizable center (NI)	positively ionizable center (PI)	hydrogen bond acceptor (HBA)	hydrogen bond donor (HBD)
• carboxylic acid • S-/P acids (sulfonic, sulfinic, phosphonic or phosphinic acid • tetrazole • acid imides • acid sulfonamides • trifluoromethylsulfonamide • other functional groups which are likely to be deprotonated at physiological pH (7.4) (as implemented in MOE's sdwash) • negative charge must not be directly adjacent to positive charge (exclusion of nitro-oxygen)	• aliphatic basic amines • amidine • guanidine • other functional groups which are likely to be protonated at physiological pH (7.4) (as implemented in MOE's sdwash) • positive charge not directly adjacent to negative charge (exclusion of nitro-nitrogen)	• every nitrogen, oxygen, and sulfur atom with at least one nondelocalized lonely electron pair without positive formal charge • exclusion of (sulfon)amide-N atoms due to delocalization • exclusion of amidine-/guanidine-N atoms due to positive charge and delocalization	• every nitrogen, oxygen, and sulfur atom having at least one covalently bound hydrogen atom without negative formal charge

[a] Adapted from Gillet et al.,[21] Greene et al.,[32] Taminau et al.,[33] and Zuccotto.[34]

Table 3. Overview about Data Sets from BindingDB Used for Network Optimization and SAR Analysis

target abbreviation	target	target class	bioactivity annotation type	raw no. of compds	no. of compds after preparation and RG-generation	no. of unique RGs
FXA	Factor Xa	protease (enzyme)	pIC_{50}	1887	1736	912
CDK2	cyclin-dependent kinase 2	kinase (enzyme)	pIC_{50}	2557	1575	979
COX2	cyclooxygenase-2	oxidoreductase (enzyme)	pIC_{50}	4357	2349	1083
CB1	cannabinoid receptor 1	GPCR	pK_i	2712	1957	890
P38	MAP kinase p38 alpha	kinase (enzyme)	pIC_{50}	2937	2446	1409
THR	thrombin	protease (enzyme)	pK_i	3540	2852	1731

RG Duplicate Check. As molecules with different molecular structure can yield the same RG, RG duplicates are filtered before pairwise MCS generation for accelerating the MCS calculations. This is particularly important for larger data sets. In most analyzed data sets, the number of RGs for pairwise comparison could be drastically reduced (around 40−50%) due to the large number of analogue series (see Table 3).

Pairwise Determination of MCSs. In the first step, all pairwise maximum common substructures (MCSs) of the unique Z RGs are determined (Figure 2) and stored in a ZxZ matrix where Z denotes the total number of different RGs. The maximum common edge subgraph (MCES) is used as MCS definition as the maximum common induced subgraph (MCIS) is less intuitive.[11] Only connected MCSs are considered as in this case the generation of a hierarchical interpretable network structure is easier to realize than with disconnected MCSs. By default the exact MCS is determined, but when analyzing larger data sets the performance can optionally be increased by changing to an approximate MCS algorithm. The minimum MCS size for being considered in the subsequent network can be adapted by the user. A minimum of 3 RG pseudoatoms is mandatory. The variation of the minimum MCS size is discussed in detail in the "Network optimization" section. If more than one MCSs of the same size (e.g., disconnected common substructures) is found for a RG pair (molecule X and Y), none of these MCS is prioritized. Instead all MCSs are stored in the MCS matrix at the position [X, Y] and [Y, X].

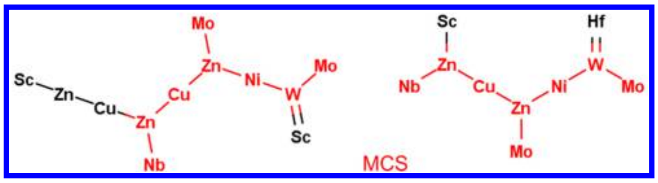


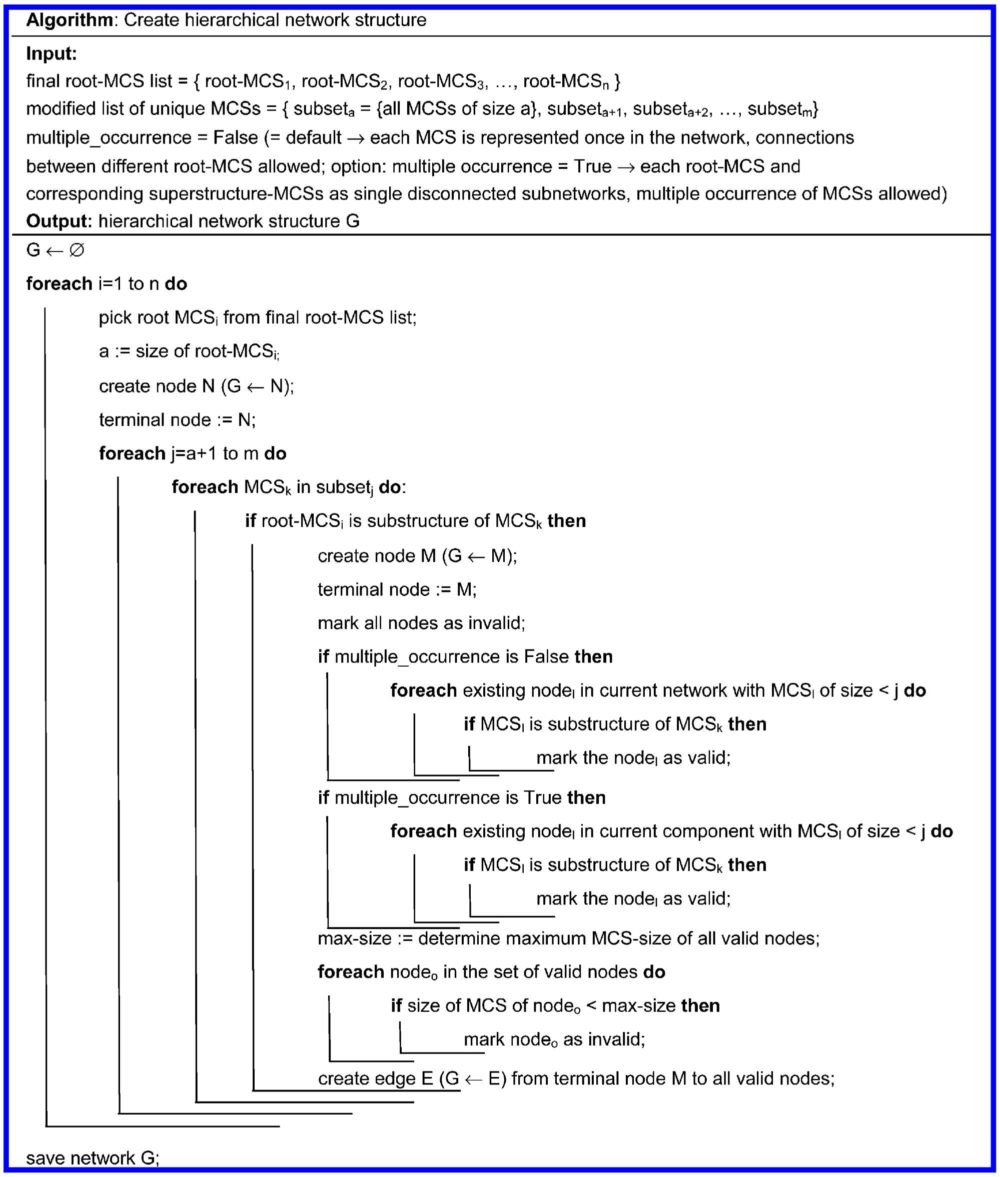
Figure 2. Illustration of the pairwise determination of the maximum common substructure (MCS).

The substructures stored in this matrix are for at least one pair of molecules the MCSs. For other molecules, these MCSs may be only common substructures (CS). Nevertheless, for the sake of simplicity, in the following it will be referred to these CSs as MCSs.

Preselection of Potential Root-MCSs. The next step for starting with the network setup is the determination of possible root-MCSs from the pool of MCSs stored in the MCS matrix. Before doing so, a list of unique MCSs is generated based on filtering the MCS matrix for duplicates using canonical SMILES.[38] Then, based on this list of unique MCSs a list of potential root-MCS is generated. A MCS fulfills the definition of a "root-MCS" if and only if it is only a substructure but in no case a superstructure of any other MCSs in the list of unique MCSs. This root-MCS is referred to as "root node" in the resulting network because in contrast to normal "MCS nodes" it has only successor nodes but no parent nodes.

Selection of Final Root-MCSs. Next, the root-MCS selection process begins. From the list of potential root-MCSs, that MCS is chosen which represents most molecules (i.e., this RG-MCS is substructure of the RGs of these molecules) not yet represented by any other root-MCS already selected. If multiple MCSs represent the same number of unrepresented molecules, the root-MCS representing less redundant information (fewer molecules already represented) is preferred. If there is still a tie which cannot be solved by this rule, the root-MCS is randomly chosen. This procedure of searching for root-MCSs is repeated until a user-defined proportion (e.g., ≤2% unrepresented molecules = default value) of molecules is represented in the network. If the data set is structurally too heterogeneous or if the required minimum MCS size is set too large, there will be a large number of pairs of molecules which will have no MCS of the required size. In these cases, it will happen that no further root-MCSs can be selected which represent any unrepresented molecules although the defined termination criterion is not yet fulfilled. In this case, the root node selection process is automatically stopped. Nevertheless, the selection of final root-MCSs can also be terminated if a user-defined number of root-MCSs has been selected.

Generation of the Hierarchical Network Structure. After the selection of root-MCSs, the hierarchical network structure is generated by iterative super- and substructure searches. The pseudocode for this algorithm is shown in Figure 3. The input for this procedure is the final root-MCS list and a modified list of unique MCSs which is generated by removing all potential root-MCSs from the list of unique MCSs. The first step is to pick a root-MCS from the list of final root-MCSs and to create a root-node. This root-MCS represents the lowest level of complexity in the hierarchy of the network. Complexity is measured by the number of pseudoatoms of corresponding MCS which is also referred to as the size of the MCS. In the next step, by superstructure search in the modified list of unique MCSs, all MCSs which are a superstructure of this root-MCS and at the same time represent the next level of complexity (i.e., MCS size a plus one pseudoatom:= size j) are identified. Next, in the network G, MCS nodes are created representing these MCSs of size j. For each MCS_k of size j, again by substructure searches all MCSs (size < j) in the existing network (default) or only in the current component (optional) are identified which represent a substructure of this MCS. The substructure MCSs of MCS_k with the highest level



```
Algorithm: Create hierarchical network structure
Input:
final root-MCS list = { root-MCS$_1$, root-MCS$_2$, root-MCS$_3$, …, root-MCS$_n$ }
modified list of unique MCSs = { subset$_a$ = {all MCSs of size a}, subset$_{a+1}$, subset$_{a+2}$, …, subset$_m$}
multiple_occurrence = False (= default → each MCS is represented once in the network, connections
between different root-MCS allowed; option: multiple occurrence = True → each root-MCS and
corresponding superstructure-MCSs as single disconnected subnetworks, multiple occurrence of MCSs allowed)
Output: hierarchical network structure G
G ← ∅
foreach i=1 to n do
    pick root MCS$_i$ from final root-MCS list;
    a := size of root-MCS$_i$;
    create node N (G ← N);
    terminal node := N;
    foreach j=a+1 to m do
        foreach MCS$_k$ in subset$_j$ do:
            if root-MCS$_i$ is substructure of MCS$_k$ then
                create node M (G ← M);
                terminal node := M;
                mark all nodes as invalid;
                if multiple_occurrence is False then
                    foreach existing node$_l$ in current network with MCS$_l$ of size < j do
                        if MCS$_l$ is substructure of MCS$_k$ then
                            mark the node$_l$ as valid;
                if multiple_occurrence is True then
                    foreach existing node$_l$ in current component with MCS$_l$ of size < j do
                        if MCS$_l$ is substructure of MCS$_k$ then
                            mark the node$_l$ as valid;
                max-size := determine maximum MCS-size of all valid nodes;
                foreach node$_o$ in the set of valid nodes do
                    if size of MCS of node$_o$ < max-size then
                        mark node$_o$ as invalid;
                create edge E (G ← E) from terminal node M to all valid nodes;
save network G;
```

Figure 3. Pseudocode of the algorithm used for the generation of a MCS-based hierarchical network structure.

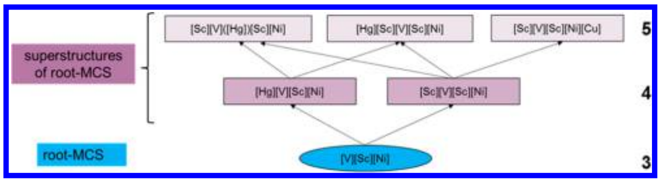


Figure 4. Schematic depiction of the hierarchical relationships between the root-MCS (blue) and its superstructures (purple). The root-MCS has the level of complexity 3 (number of pseudoatoms) and the superstructure-MCSs 4 and 5, respectively. The MCS with the lowest level of complexity is always connected to those superstructure-MCSs which represent the next higher level of complexity in the network hierarchy.

of complexity are connected by an edge to this MCS$_k$. In the following, the hierarchical structure is established by iteratively increasing the level of complexity. That is to say that the size of the MCS considered in superstructure searches is incrementally raised by one pseudoatom. This procedure is repeated for every root-MCS. In the end, each MCS is connected to the next larger MCS(s) representing a superstructure of this smaller MCS and a hierarchical structure is obtained (Figure 4). However, as one MCS node can be the substructure of several larger MCS nodes, the resulting network structure is very complex and difficult to interpret.

In order to reduce network complexity, the network is converted to a tree structure by calculating the minimum spanning tree (MST) using the Kruskal algorithm.[39] The edge weight is set to the inverse number of represented molecules. The MST calculation converts the complex network into a network with tree structures which has the advantage of a clear hierarchical organization and which is easier to interpret.

Mapping of Original Molecules in the Network. By mapping the original data set molecules to the corresponding MCS nodes, inSARa networks are obtained. When visualizing the networks, molecules are only shown on the largest matching RG-MCS node of every tree branch in order to facilitate network interpretation. Molecules can occur multiple times in the resulting network. One drawback of this is the added network complexity and the risk of redundant information. However, this multiple occurrence also helps to get a better understanding of chemical neighborhood and to see one particular molecule in different structural context.

inSARa networks consist of one or more tree structures. However, there are two variants when generating the hierarchical network structure which strongly influence the complexity of the network and the redundancy of information. The default option is that disconnected trees can be joined via common RG-MCSs. This default type of inSARa network avoids redundancy in the networks and facilitates the detection of commonalities between different trees, but it results in larger, more complex networks. The optional variant is that every root node results in a single tree. This has the disadvantage of representing redundant information in different trees, but it usually leads to smaller, less complex subnetworks. All networks shown and discussed in the "Results and Discussion" section are generated using the default option.

For the implementation of the network generation, apart from OEChem TK routines,[37] the open source library NetworkX[40] is used.

Network Layout and Visualization. For visualization, the networks can be imported to the open source network analysis and visualization software Cytoscape.[41] For reasons of network clarity, the force-directed layout algorithm with tuned parameters (for details see Table S 1 in the Supporting Information) is selected. Moreover, it is possible to edit networks manually after layout generation. As the edge length has no chemical meaning, this parameter is used by the algorithm to separate densely connected regions. Nodes representing the RG-MCSs are colored in black. Nodes which represent the original molecules are color-coded depending on the annotated bioactivities (pIC_{50} or pK_i). To be in line with previously published tools, the color spectrum from green (= lowest activity) via yellow (= intermediate activity) to red (= highest activity) is adopted.[42] The thresholds for the different activity classes can be adapted to the analyzed target class. A prototypic inSARa network is shown in Figure 5.

To support interactive SAR exploration, the Cytoscape plug-in chemViz[43] is used to depict the 2D structure of the molecules on the corresponding nodes (see Figure 5b). Depiction of RGs or RG-MCSs is also possible. Facilitating the navigation in the networks, which may be quite complex, the growing RG-MCSs size in the different network branches is visualized by raising the size of the MCS nodes with increasing MCS size.

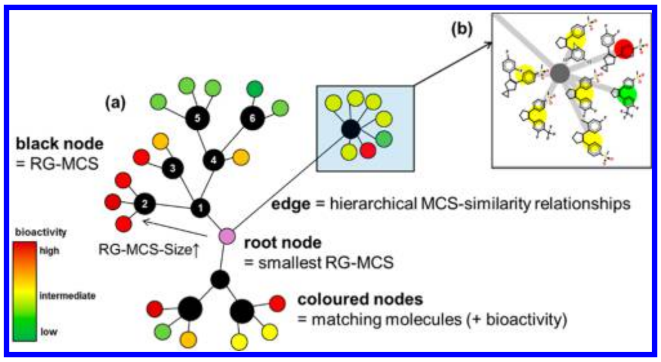


Figure 5. (a) A prototype inSARa network. inSARa networks are hierarchically organized. Edges beween nodes indicate an increase or decrease in RG-MCS by one or more pseudoatoms. Therefore, the following relationships exist: Node 1 is a substructure of nodes 2, 3, 4, 5, and 6 and a superstructure of the root node, whereas node 4 is a superstructure of node 1 and the root node and a substructure of nodes 5 and 6. (b) After import of the network data into Cytoscape, molecules can be depicted on the corresponding nodes using the chemViz plugin.

Compound Data Set Composition. To demonstrate inSARa's capabilities for large-scale SAR analysis, different data sets consisting of more than 1,000 compounds are assembled from BindingDB.[44] Compared to typical QSAR data sets, such as the data sets published by Sutherland et al.[45] or Fontaine et al.,[46] which can also be analyzed with inSARa, the BindingDB data sets are more diverse and much larger. Hence, SAR interpretation is more challenging.

Some restrictions for compound selection were made: Only K_i or IC_{50} values were accepted as bioactivity data.[47] Measurements containing threshold values (i.e., reported as ">" or "<") were not considered. If multiple bioactivity data against the same target were provided, the arithmetic mean was calculated unless the reported potency values differed more than 1 order of magnitude.[48] In this case, the compound was not considered for further analysis in order to avoid errors resulting from large variability. Duplicates were filtered using isomeric canonical SMILES.[38] Moreover, large molecules having a molecular weight higher than 800 Da were excluded from further investigation (analogue to Stiefl et al.[25] and Chen and Reynolds[49]) because they are less drug-like and lead to large RGs. As, for example, the analysis of the distribution of molecular mass during data set compilation showed that the protease inhibitors (e.g., FXa inhibitors, Thrombin inhibitors) tend to have higher molecular weight than usual drug-like molecules, 800 Da was chosen as filtering threshold. Only compounds with a bioactivity value lower than 100 μM were used for data set compilation.

To facilitate performance evaluation of the resulting inSARa networks, targets with well-studied SAR features were preferred. The six analyzed data sets were chosen as they represent different kind of target classes (five enzymes (kinases, proteases, other) and one GPCR) and show differences in size and structural heterogeneity (Table 3). Additional data set characteristics such as potency distribution, fingerprint similarities, and the global SAR Index[50] as calculated with SARANEA[42] can be found in Table S 2 in the Supporting Information.

Data Preparation. To avoid inconsistencies in molecular representation resulting from compound collections from different data sources, a standardization procedure using MOE's sdwash and sdfilter functions[51] and in-house python scripts were applied to every compound data set. This includes the removal of salts, solvents, and other adducts and the adjustment of the protonation state to physiological pH value (7.4). Moreover, only molecules consisting of the standard organic elements (C, H, N, O, S, P, F, Cl, Br, I) were retained. Metalloorganic or hypervalent molecules were rejected by the filter procedure.

Network Optimization: Analysis of the ZINC Database. For further enhancing the benefits from inSARa networks, a large-scale analysis using the latest ZINC database[52] (Version 12) was carried out with the aim of identifying unspecific RG-MCSs which result from the comparison of two randomly chosen molecules. It was investigated how this information can be used for detecting unspecific information during network generation and thus enabling more target-specific networks.

From the ZINC database, the subset "Clean Drug-Like" (~12 million compounds) is chosen because it is most representative for the analyzed data sets from BindingDB. "Clean" means that molecules with undesired functional groups (e.g., toxic or reactive groups) are discarded. Duplicates are filtered using Open Babel's[36] INCHIs.[53] For standardization, MOE's[51] sdwash and sdfilter is applied (see "Data preparation") and reduced graphs are generated as described above. After filtering and RG generation, the database consists of about 11 million compounds. From this pool of compounds two molecules are randomly chosen and the RG-MCS is calculated. The minimal MCS-size is set to 3 pseudoatoms. This procedure is repeated 1 million times for getting a representative statistic about random RG similarities. This statistic contains information about the occurrence of different RG-MCS sizes and RG-MCSs and the relationship between RG-MCS and different fingerprint similarities.

■ RESULTS AND DISCUSSION

Network Optimization. When developing the inSARa method, some parameters influencing the complexity, topology, and specificity of the resulting networks were optimized. Preliminary experiments revealed that the RG definition is one major aspect when searching for parameters for improving the SAR analysis. Furthermore, the root node selection criteria turned out to have a large influence on network specificity and complexity, and the weighting criterion for MST calculation highly influences the network topology.

Apart from these optimized parameters, some default parameters which can optionally be changed by the user were defined. Their influence on network complexity and specificity will be discussed in the following in detail.

Improving Target Specificity by Blacklisting of Unspecific RG-MCSs. To analyze the specificity of the mandatory minimum MCS size and the MCS nodes represented in the inSARa networks, an analysis of random similarity using the ZINC database as described in the 'Methods' section was carried out. The results of this analysis are summarized in Figure 6 and Figure 7.

In Figure 6 the occurrence of different RG-MCS sizes are shown. It can be seen that in about 50% of the random molecule pairs no RG-MCS of size 3 or more atoms can be found. That means that the required minimum MCS size used in the network generation process is already quite specific. As expected, the occurrence of random pair RG-MCSs decreases drastically with increasing MCS size so that a MCS size of 6 atoms occurs only rarely (<1%) in random pairs. From these results it can be concluded that increasing the minimum MCS size to 4 or 5 RG atoms can help to exclude random similarity from the inSARa networks. Thus, it is more likely that the MCSs shown in the networks represent target specific common features.

In Figure 7 the occurrences of the top-20 most occurring RG-MCSs are shown. A complete list of all RG-MCS SMILES with a probability equal to or greater than 0.1% can be found in Table S 3 in the Supporting Information. This information can also be used for increasing the target specificity of the resulting networks. For that purpose, a random probability of 0.1% is chosen as threshold value for lack of specificity. Hence, all RG-MCSs with a probability equal to or greater than 0.1% are regarded as unspecific and discarded from the pool of all unique MCSs and subsequently not considered during network generation. In the following this list will be referred to as the "black list". If this list is used to discard MCSs from the network, this will be shortly denoted 'black list = active'. By default this list is set active; optionally it can be deactivated by the user.

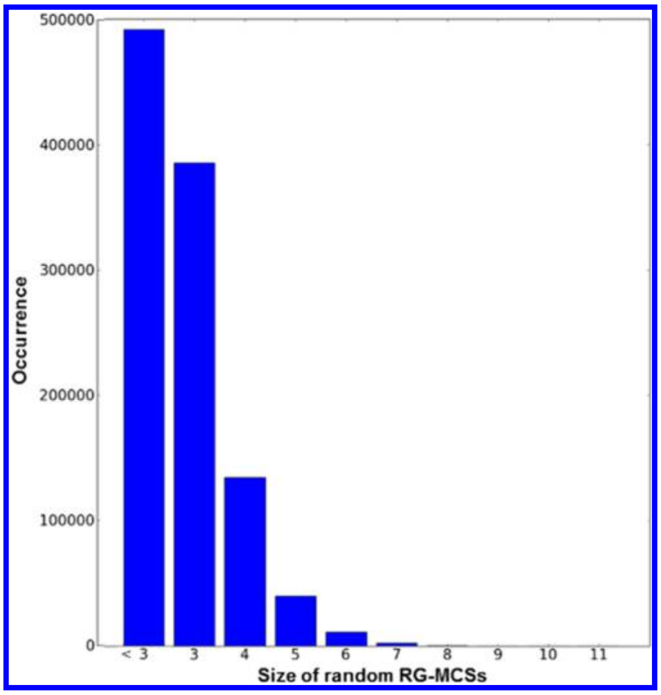


Figure 6. Occurrence of different RG-MCS sizes of unrelated random molecule pairs from the ZINC database. In total 1 million pairs were randomly drawn.

Using this threshold (0.1%), 60 MCSs with a size of 3 or 4 RG atoms are discarded. The most frequent pseudoatoms are Sc (= featureless aromatic ring), Zn (= linker and terminal groups), Ni (= HBA not in ringsystem), and Co (= HBD not in ringsystem). The most frequent SMILES [Co]([Ni])[Zn] represents a linker connected to an HBD which is connected to an HBA. The combination of Co and Ni represents, for example, sulfonamides or amides which are typically found in many peptid-like molecules and is therefore less specific. Sc represents mainly phenyl rings which are one of the most abundant steric features of drug-like molecules. This is just as unspecific as Zn representing any linker or terminal group. In summary, exclusion of unspecific MCSs by using the black list allows for the increase of the target specificity of the inSARa networks. Using the black list has only marginal impact on network complexity for large data sets.

Further Optional Parameters for Optimization. In addition to (in)activation of the black list, the minimum MCS size and the termination criterion for the root node selection are two additional parameters for fine-tuning of the networks which can be set by the user. For analyzing the influence of these optional parameters on network complexity and topology, the data sets from Table 3 were used. After preparing, filtering, and RG-generation, they consist of 1500 to 3000 unique molecules. For optimization, inSARa was applied to each of these data sets with different parameters, and the resulting networks were analyzed with respect to their complexity and topology. The results are summarized in Table 4 and Table 5.

Variation of Minimum MCS Size. In RG generation the minimum MCS size is set to 3 RG atoms. Nevertheless, RG-MCSs consisting of 3 RG pseudoatoms are in most cases of

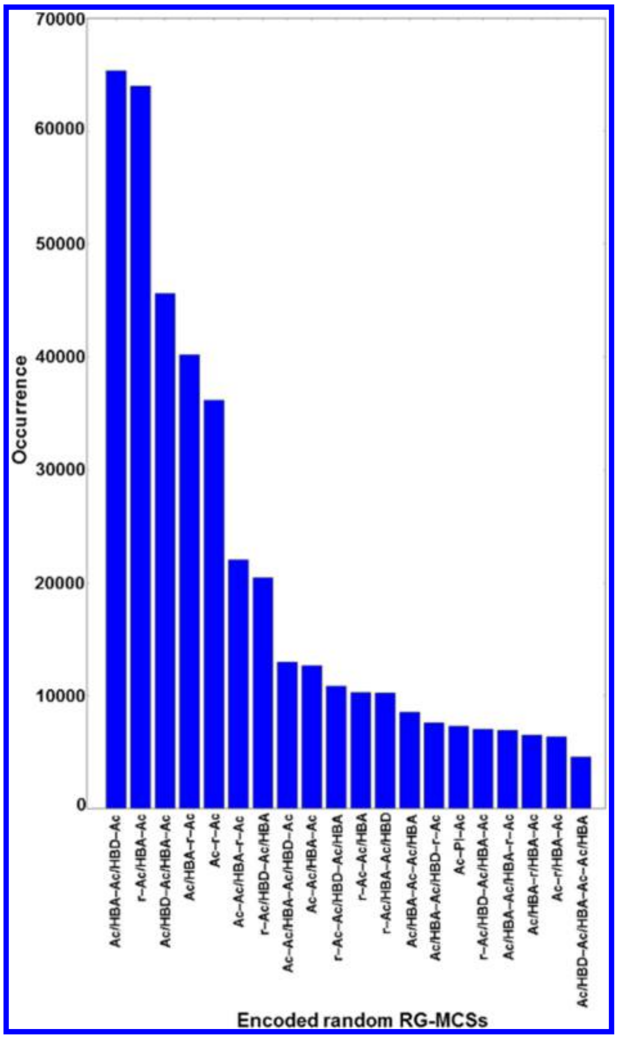


Figure 7. Occurrence of the top-20 most occurring RG-MCSs. One million random molecule pairs from the ZINC database were compared. If different RG-MCSs are determined with the same size for the same molecule pair, all of these MCSs are taken into account in this statistic.

little specificity. This is also emphasized by the analysis of the ZINC database (see above) which indicates that RG-MCSs of a size of 3 atoms occur with a frequency of almost 40% (Figure 6) in unrelated molecules. Increasing the threshold for the minimum RG-MCS size reduces the number of MCSs and subsequently the complexity of the resulting networks (Table 4). In most cases, excluding smaller RG-MCSs results in a marginal loss of information because they often represent broadly occurring molecular features with low information content for SAR analysis. These MCSs often only have linking function for the larger MCSs. Therefore, increasing the minimum MCS size leads to inSARa networks with a higher number of disconnected components (Table 4). However, the important information (usually found at MCS nodes of size 8 or larger) is retained. One major advantage is that these networks are less complex due to the simplified layout, and they are also faster to compute. Expectedly, the number of root nodes increases with a larger minimum MCS size (Table 4). This can be explained by the fact that these larger root nodes are more specific and thus represent less molecules. One problem which might occur when raising the minimum MCS size is that a higher percentage of molecules cannot be represented and the root node selection stops before the termination criterion is fulfilled (e.g., COX2 in Table 4). Hence, what is required is a trade-off between specificity and selectivity of the networks, as an increase in the minimum MCS size leads not only to more specific but also to more selective networks as some molecules will not be represented with larger minimum MCS size. Apart from specificity and selectivity the overall diversity of the data set is important for finding an ideal minimum MCS size since in diverse data sets only small, unspecific RG-MCSs will result. A third factor for a good minimum MCS size is the average RG size of the molecules in the data set (Table 6). As can be seen in Table 4, in data sets with a smaller average RG size (e.g., COX2 or CB1) an increase in the minimum MCS size leads to a higher number of molecules which cannot be represented by MCS nodes of the resulting networks. In summary, a minimum MCS size of 5 RG atoms turns out to be a good trade-off between the potential loss of SAR information and network complexity.

Variation of the Termination Criterion for Root Node Selection. By modifying the termination criterion for root-MCS selection, the network complexity and topology can also be influenced. Similar to the minimum MCS size, this parameter also strongly depends on the analyzed data set. Therefore, it can be optionally modified for network fine-tuning. In Table 5 it can be seen that raising the termination criterion results in an exclusion of root nodes representing only a small number of molecules. Similar to the number of root nodes, also the number of MCS nodes and the number of components decreases. Thus, by stopping root node selection at earlier stages, networks can be simplified.

Another way of excluding root-nodes representing only a small number of molecules would be to select only root nodes which represent a minimum number of molecules. If no further root nodes fulfilling this criterion are found, the root node selection process is stopped.

For the reduction of network complexity postpruning of the network is also possible. By defining a minimum number of molecules which have to be represented by the terminal nodes network complexity could be decreased. The number of MCS nodes shown in Table 4 and Table 5 also include MCS nodes which do not represent any molecule because those molecules that are represented by these nodes are also covered by a larger MCS. These MCS can also be removed from the network if they have no linking function. Further pruning criteria for network simplification can be defined. In summary, depending on the requirements of a particular analysis different ways of network fine-tuning and network simplification are possible.

Rules for Interactive SAR Network Interpretation. *Local SAR Analysis (Focusing on Single MCS Nodes).* inSARa networks will be quite large if large data sets are analyzed (see Table 4). Therefore, simple rules were derived to easily find interesting SARs and SAR spots in the resulting inSARa networks.

First of all, the analysis should focus on terminal nodes because the RG-MCSs are more specific than the MCSs from nonterminal nodes. Thus, the matching molecules are more similar and the SARs are easier to interpret. Empirical analysis revealed that MCS nodes with a corresponding MCS size equal to or greater than 8 or 9 pseudoatoms can be in most cases recognized as chemically "similar" and should be considered for further visual inspection. Thus, interpretable relationships are present and SAR trends can be deduced.

Table 4. Influence of Minimum MCS Size on Network Complexity and Topology[a]

target	minimum MCS size	3	4	5	6	7	8
FXa	unique MCSs in MCS-matrix	1774	1680	1470	1227	996	804
	root nodes in network	12	19	36	57	81	88
	MCS nodes in network	1041	800	660	541	454	366
	unrepresented molecules	33	33	33	34	74	156
	% unrepresented	1.9	1.9	1.9	2.0	4.3	9.0
	no. of components	1	4	16	40	66	74
CDK2	unique MCSs in MCS-matrix	1489	1340	1074	795	581	393
	root nodes in network	26	53	105	129	150	153
	MCS nodes in network	986	813	676	488	341	205
	unrepresented molecules	31	31	31	127	305	527
	% unrepresented	2.0	2.0	2.0	8.1	19.4	33.7
	no. of components	3	30	76	101	126	146
COX2	unique MCSs in MCS-matrix	1750	1615	1322	974	639	366
	root nodes in network	31	61	134	190	207	174
	MCS nodes in network	1336	1117	830	552	315	143
	unrepresented molecules	44	45	76	208	639	1261
	% unrepresented	1.9	1.9	3.2	8.9	27.2	53.7
	no. of components	10	29	80	144	171	164
CB1	unique MCSs in MCS-matrix	1623	1506	1254	938	669	452
	root nodes in network	49	74	96	129	155	158
	MCS nodes in network	1190	968	761	584	387	233
	unrepresented molecules	46	51	88	197	404	643
	% unrepresented	2.4	2.6	4.5	10.1	20.6	32.9
	no. of components	18	41	67	99	127	142
P38	unique MCSs in MCS-matrix	2521	2354	1980	1542	1164	847
	root nodes in network	25	43	119	170	219	232
	MCS nodes in network	1731	1540	1225	992	735	506
	unrepresented molecules	48	48	48	141	314	608
	% unrepresented	2.0	2.0	2.0	5.8	12.8	24.9
	no. of components	1	10	67	102	164	183
THR	unique MCSs in MCS-matrix	3963	3821	3484	2960	2441	2000
	root nodes in network	24	45	105	160	200	206
	MCS nodes in network	2999	2736	2484	2087	1689	1329
	unrepresented molecules	45	56	57	102	217	374
	% unrepresented	1.6	2.0	2.0	3.6	7.6	13.1
	no. of components	1	8	44	102	140	170

[a] Termination criterion = 2% unrepresented molecules, black list = active.

Depending on the motivation for SAR analysis, it should be focused on three different types of nodes which strongly differ in the kind of SAR information they carry.

1) When looking at MCS nodes which are *simultaneously connected to molecule nodes where the majority is red colored and a small number is green colored* or vice versa (i.e., the matching molecules show large differences in potency), *activity cliffs* or *nonbioisosteres* can be identified. An activity cliff is defined as a pair of molecules where small structural differences result in a large potency change.[54] In contrast to fingerprint-based activity cliff detection[55] no thresholds depending on the fingerprint type used have to be defined.

2) If *bioisosteric exchanges* are of interest (e.g., for lead-optimization), MCS nodes only connected to *single-colored molecule nodes* (i.e., all matching molecules have similar bioactivity) should be inspected. Bioisosteres are characterized by retaining the biological activity despite low local structural similarity. They typically have similar steric and pharmacophoric properties.[56]

3) Another interesting spot in inSARa networks are MCS nodes connected to *molecule nodes which have different colors* (i.e., matching molecules show a high variance in bioactivity). In contrast to the first node type no majority of one color type is found but a high variance in color, respectively bioactivity. Focusing on these so-called "*SAR hotspots*"[13,14] might be helpful for the identification of molecular features crucial for bioactivity modification which is a major challenge in lead-optimization.

Subnetwork Analysis (Including Neighboring MCS Nodes). Inspecting entire network paths may reveal what is referred to as "*activity switches*", i.e. a *branch in the network where activity switches from active to inactive* or vice versa.

They are important for the detection of pharmacophoric features which determine activity.

Application and SAR Analysis: Factor Xa. As inSARa networks are generated without considering bioactivity information, they characterize data sets according to common pharmacophoric patterns. Despite the fact that bioactivity has no influence on grouping, it will be shown in the following that molecules cluster reasonably which underpins the significance of the chosen molecular representation.

Factor Xa (FXa) is a well-studied target concerning SARs. The filtered FXa data set (IC_{50}) from BindingDB consists of 1736 molecules. For these reasons, it appears to be a good prototypic target and data set for a detailed study of the

Table 5. Influence of Termination Criterion (Root Node Selection) on Network Complexity and Topology[a]

target (min MCS size)	termination criterion (% unrepresented molecules)	1	2	5	7	10	15	20	25
FXa (5)	root nodes in network	43	36	28	25	21	17	13	11
	MCS nodes in network	666	660	640	629	610	575	532	514
	unrepresented molecules	17	33	81	111	168	244	339	404
	% unrepresented	1.0	1.9	4.7	6.4	9.7	14.1	19.5	23.3
	no. of components	21	16	13	13	13	11	8	7
CDK2 (5)	root nodes in network	105	105	70	59	49	38	31	27
	MCS nodes in network	676	676	660	646	618	570	531	495
	unrepresented molecules	30	31	77	110	155	231	314	384
	% unrepresented	1.9	2.0	4.9	7.0	9.8	14.7	20.0	24.4
	no. of components	74	76	55	47	39	32	26	22
COX2 (3)	root nodes in network	44	31	19	16	14	11	9	7
	MCS nodes in network	1366	1336	1261	1153	1103	1034	985	898
	unrepresented molecules	23	44	110	162	215	316	421	586
	% unrepresented	1.0	1.9	4.7	6.9	9.2	13.5	17.9	24.9
	no. of components	17	10	5	4	4	3	2	1
CB1 (3)	root nodes in network	49	49	28	22	18	14	11	9
	MCS nodes in network	1190	1190	1087	1048	1026	969	855	770
	unrepresented molecules	46	46	94	133	192	274	376	468
	% unrepresented	2.4	2.4	4.8	6.8	9.8	14.0	19.2	23.9
	no. of components	18	18	10	7	5	3	4	4
P38 (5)	root nodes in network	126	119	70	60	49	37	29	24
	MCS nodes in network	1253	1225	1172	1139	1090	115	961	892
	unrepresented molecules	40	48	122	167	244	366	485	608
	% unrepresented	1.6	2.0	5.0	6.8	10.0	15.0	19.8	24.9
	no. of components	67	67	32	29	23	16	12	9
THR (6)	root nodes in network	160	160	122	96	73	31	37	29
	MCS nodes in network	2087	2087	2058	2019	1964	1849	1804	1696
	unrepresented molecules	102	102	142	198	282	423	557	700
	% unrepresented	3.6	3.6	5.0	7.0	9.9	14.8	19.5	24.5
	no. of components	102	102	81	63	46	33	21	16

[a] Black list = active.

Table 6. Distribution of RG Size in Different Data Sets[a]

target	min RG size	max RG size	median RG size	mean RG size
FXa	5	19.0	12.0	11.9
CDK2	3	19	9.0	9.2
COX2	2	20	8.2	8.0
CB1	2	19	9.0	9.0
P38	3	21	10.0	9.8
THR	1	24	12.0	11.5

[a] Histograms are shown in Figure S 1 in the Supporting Information.

application of inSARa to a large data set and subsequent interactive SAR interpretation. Figure 8 and Figure 9 show the resulting inSARa network representing the entire data set for using the recommended minimum MCS size of 5 RG atoms. The inSARa network consists of 16 connected components, 36 root MCS nodes, and a total number of 660 MCS nodes. Due to the default termination criterion of ≤2% unrepresented data set molecules, 33 molecules are not represented in the network.

There are three major connected components which represent most of the molecules (see Figure 8 and Figure 9a/b). In the largest connected component (Figure 8) branches representing molecules with similar bioactivities can often be found. For example, the subnetwork **I** predominantly contains highly active compounds (most terminal leaf nodes colored in red), subnetwork **II** contains intermediately and highly active molecules (most terminal leaf nodes colored in yellow or (dark) red), while part **III** is dominated by lowly or intermediately active molecules (green and yellow molecule nodes); but there are also parts in the network with high variance in bioactivity (see **IV**/**V**). Larger MCS nodes, especially the terminal MCS nodes, are in most cases less heterogeneous with respect to bioactivity than smaller MCS or the root nodes. This is to be expected as the similarity of the matching molecules at nodes representing smaller MCS is lower than the similarity at the terminal or larger nodes.

Seen globally, FXa shows a heterogeneous activity landscape consistent with a SAR Index[50] close to 0.5 calculated for the entire data set. A high proportion of continuous parts[57] where similar molecules exhibit similar bioactivities and where gradual structural changes only lead to modest differences in bioactivity can be found in the network. This is also expressed by a high continuity score (0.78). However, also consistent with the high discontinuity score (0.88), there are a lot of MCS nodes with high variance in bioactivity despite similar molecular features.

Applying the rules proposed above, different characteristic SAR patterns can be identified. This will be illustrated in detail by the following examples (**A**−**G**) extracted from the inSARa network shown in Figure 8 and Figure 9. (Remark: In some figures binding pocket information was manually added for illustration. This is not included in the visualization with Cytoscape. Due to multiple occurrences in the network and MST calculation, some of the molecules shown in the examples might also share larger CSs although terminal nodes are shown. That means other nodes in the network exist where these molecules appear again.)

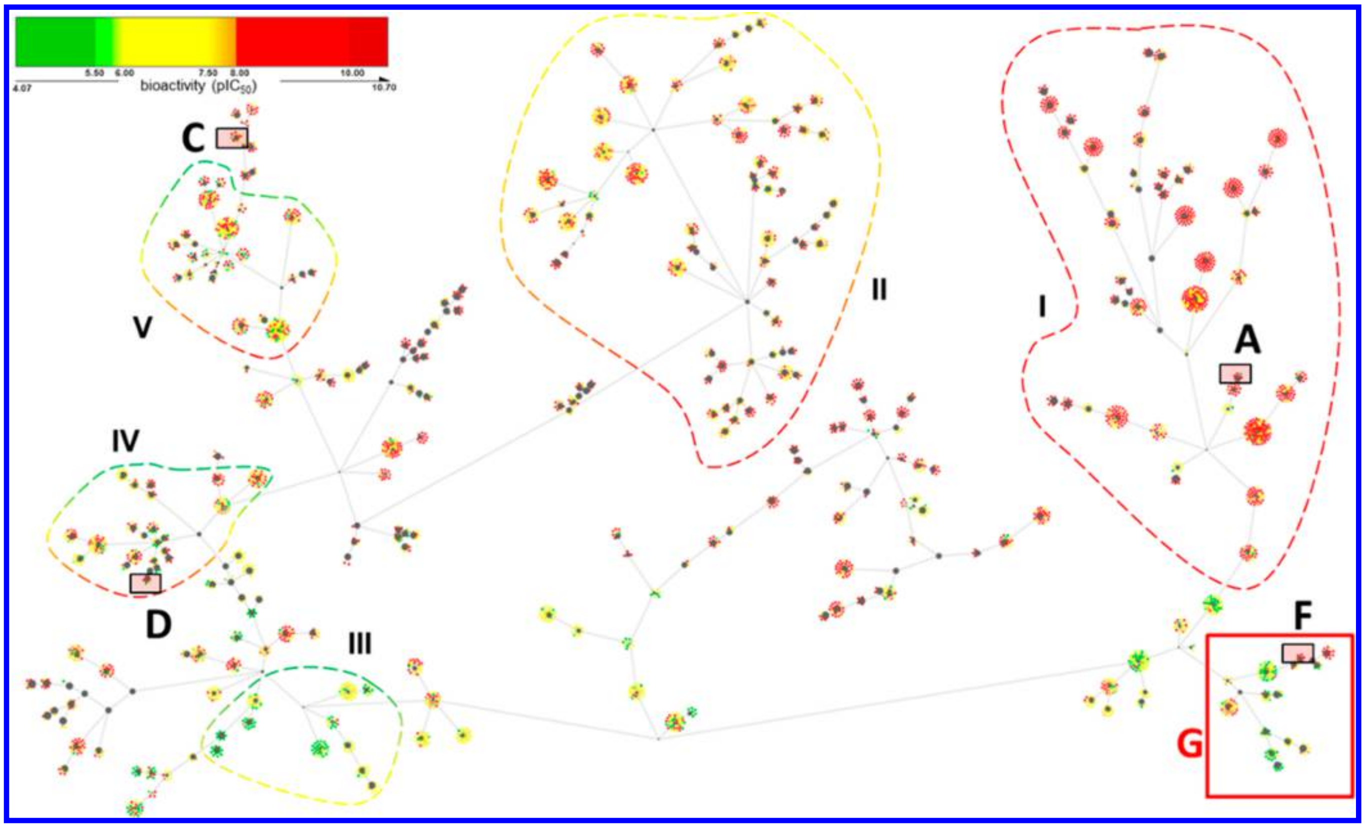


Figure 8. Largest connected component of the FXa inSARa network (parameters: minimum MCS size = 5 RG atoms, black list = active, termination criterion: ≤ 2% unrepresented molecules). The network layout was manually adapted after automated layout generation for reasons of clarity. The encircled subnetworks (**I**–**V**) and labeled nodes (**A**, **C**, **D**, **F**) will be discussed in the text.

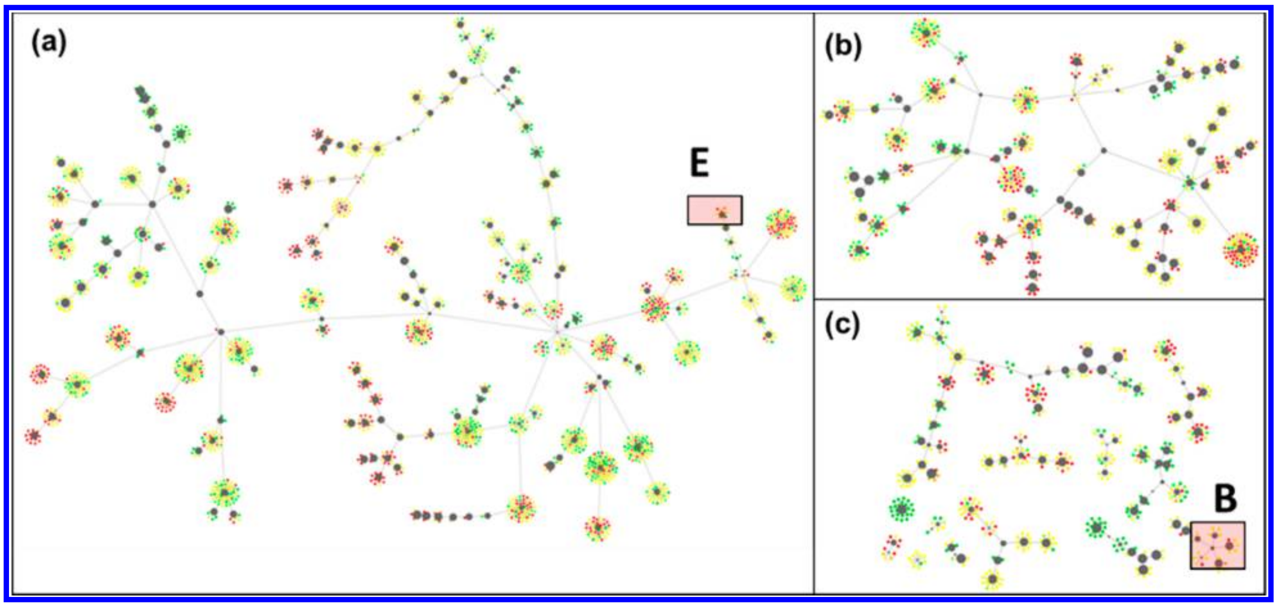


Figure 9. Remaining components of the FXa inSARa network. Shown are the second largest (a), the third largest (b), and the remaining 13 connected components (c). The labeled nodes will be discussed in the text.

A representative example of an *activity cliff* (one green molecule node surrounded by red nodes at one single MCS node) is illustrated in Figure 10. Here, the exchange of the amidine group (P1-element[58]) with an amino-isochinoline leads to a large potency loss although the structural changes are small. Comparing the RGs of molecule 1 and 2 this cliff can be rationalized: The amino-isochinoline (in contrast to molecule 1 the RG of molecule 2 contains no PI-feature) is less basic than the benzamidine and consequently less ionized. Since ionization favors the ionic interaction in the S1-pocket with the Asp$_{189}$, a potency decrease can be expected. In Figure 10 it can also be seen that on the phenyl ring next to the benzamidine different H-bonding acceptors (methyl ether,

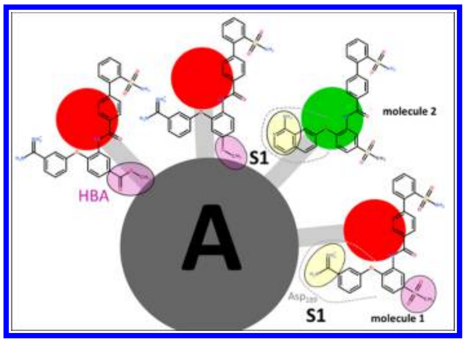


Figure 10. Illustration of the identification of an activity cliff in inSARa networks. The yellow marked molecular features distinguish the molecules 1 and 2 representing the activity cliff. The purple marked functional groups represent H-bonding acceptors which can be exchanged by each other without potency loss.

methyl ester, methyl sulfone) are accepted without potency decrease.

The identification of *bioisosteric groups* is shown in Figure 11. All molecules of this terminal node are of intermediate activity. In (a) bioisosteric positively ionizable groups are marked. They are involved in cation-π interaction in the S4-pocket.[59] In (b) two bioisosteric aromatic rings are identified. The aromatic ring systems are important features for hydrophobic or aromatic interaction in the S1-pocket.[58,60]

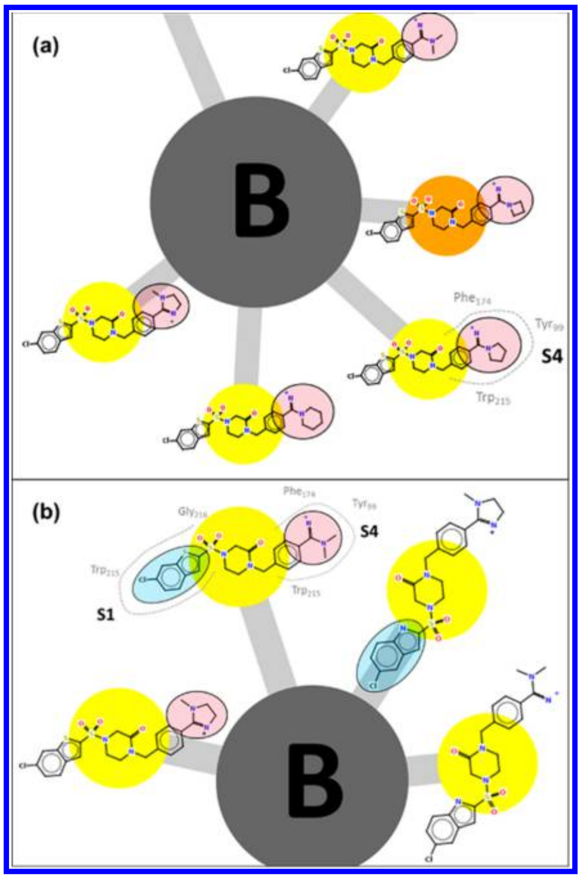


Figure 11. Representative examples for the identification of bioisosteric replacements. (a) Different exchangeable bioisosteric PI groups are identified (marked purple). (b) Two bioisosteric aromatic ringsystems (blue) and positively ionizable groups (purple) are marked. For interactions with subpockets see text.

In Figure 12 and Figure 13 two examples for *SAR hotspots* (occurrence of multiple colors at one single MCS node) are given.

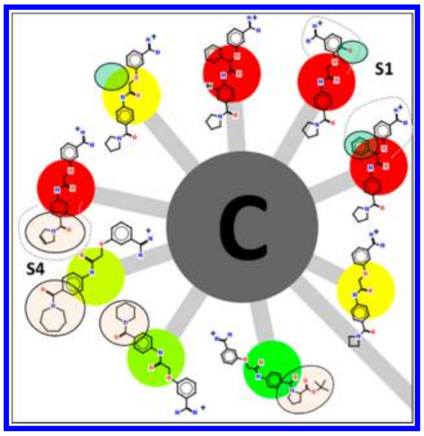


Figure 12. SAR hotspot from the inSARa network. For discussion of groups interacting with subpocket S4 (beige) and S1 (blue) see text.

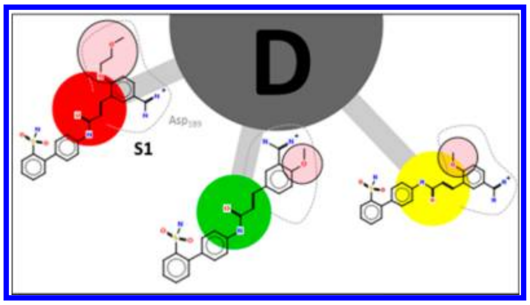


Figure 13. A second example for a SAR hotspot. The substituent circled in purple and its position on the benzamidine-ring strongly influence the biological activity (S1-optimization).

From Figure 12 molecular elements crucial for potency modification can be deduced. For S4-optimization the ring size and ring substitution of the *N,N*-disubstituted amides (beige) is important for optimally fitting in the S4-pocket. It can be seen that bulky rings lead to steric hindrance[61] (strong decrease in potency). For S1-optimization an additional OH-group or phenyl ring is favorable (highly active molecules) as compared to the intermediately active molecules where these functional groups are missing.

Figure 13 shows how the substitution on the benzamidine ring in the S1-pocket strongly influences the biological activity. Substitution on the 4-position is preferred to the 2-position, and the larger 4-methoxyethoxy analogue is more potent than the 4-methoxy compound.[62]

Stereochemistry information is not encoded in the RGs. Therefore, stereoisomers are found at the same MCS nodes. This can be seen in Figure 14 where the influence of stereochemistry and *nonbioisosteric* replacements on biological activity is illustrated. All molecules are encoded by the same RG, nevertheless a high variance in bioactivity is observed. It can be seen that the R-configuration is preferred at the chiral center at the pyrrolidine-ring.[63] Moreover, the chloro-phenyl analogues appear not to be bioisosteric to the chloro- or bromo-thienyl compounds.

In Figure 15 another interesting feature of inSARa networks emphasizing its high value for SAR analysis is exemplified. All the molecules connected to the MCS node **F** are highly potent

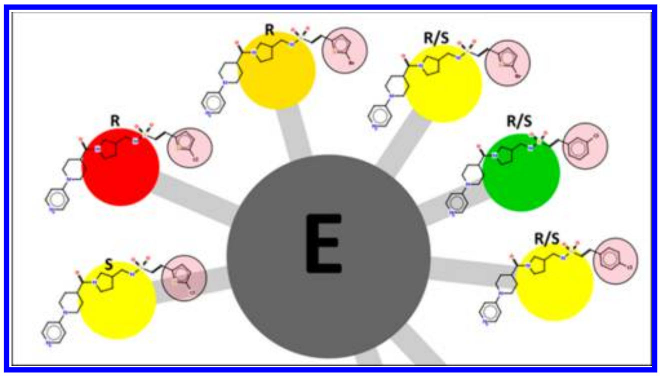


Figure 14. Influence of stereochemistry (not encoded in the RGs) on biological activity. R-configuration at the chiral center is preferred. The marked ring systems (red) are identified as nonbioisosteric.

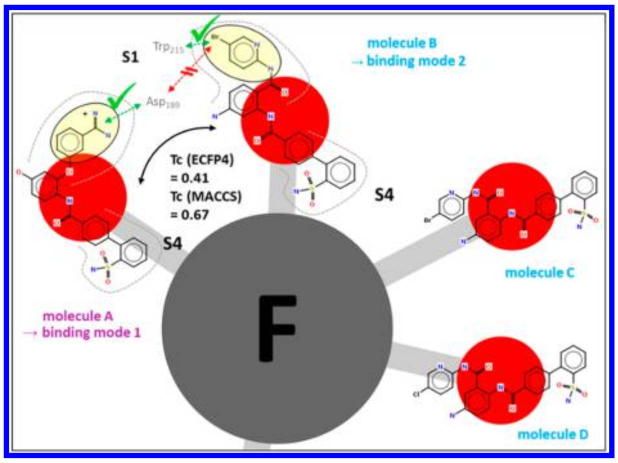


Figure 15. Identification of different binding modes in the S1-pocket of the binding site of FXa using inSARa and comparison with fingerprint-based similarity networks. As the calculated Tanimoto similarities (Tc) for ECFP4 and MACCS are below the thresholds (0.55/0.75) for creating an edge in the fingerprint-based similarity network, no direct relationship will be established. For details see Figure S 2 in the Supporting Information.

inhibitors (all molecule nodes are red-colored). Molecule A has a phenyl-amidine group which is (as seen in Figure 15) one essential molecular feature for interacting with the S1-pocket in the active site of the enzyme. However, in molecules B, C, and D instead there is a chloro- or bromo-pyridinyl group. As these molecular features are unlikely to interact with the Asp_{189} constituting normally a key interaction, evidence for a second binding mode in the S1-pocket is given. This hypothesis can be proven with crystallographic data from the PDB[64] (e.g., 2P3T) and is also well described in the literature.[65] The calculated Tanimoto similarities[66] for molecules A and B using MACCS keys[67] and the ECFP4 fingerprint[68] (MOE implementation[69]) are rather low (see Figure 15). In fingerprint-based similarity networks (NSGs[5] and related networks) each node represents one data set molecule, and edges are only created between two nodes if the Tanimoto similarity exceeds a fingerprint-dependent threshold value. The resulting fingerprint-based similarity network (using ECFP4 fingerprint) for the FXa data set is shown in Figure S 2 in the Supporting Information. When defining 0.55 for ECFP4 and 0.75 for MACCS keys as threshold values, no edges in the fingerprint-based similarity network will be created between the molecules representing different binding modes. Consequently, in contrast to inSARa, no direct relationship between these molecules will be established (as seen in Figure S 2). Hence, inSARa alerts the user to investigate different binding modes for these overall similar molecules that differ mainly in a single strategically important molecular feature.

As previously mentioned, inSARa can also be used for the identification of "activity switches". This is illustrated with the two examples **I)** and **II)** in Figure 17 and Figure 18. Figure 16 shows area **G** from Figure 8 in detail. In **I)** (Figure 17) it can be seen that the two larger MCS nodes on the right-hand side clearly separate highly and lowly active molecules although only one feature is changed as compared to the smaller MCS node

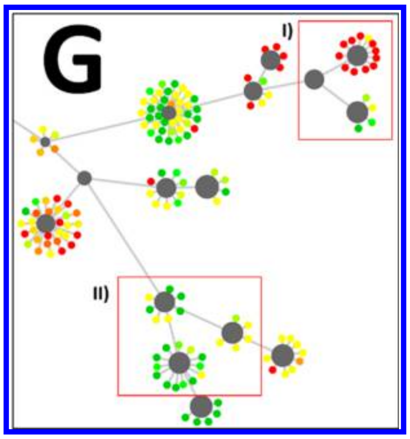


Figure 16. Identification of activity switches using inSARa networks. The network branch in the upper subplot is located in area **G** in Figure 8. Two types of activity switches are illustrated in **I)** and **II)** (detailed plot in Figure 17 and Figure 18). For details see text.

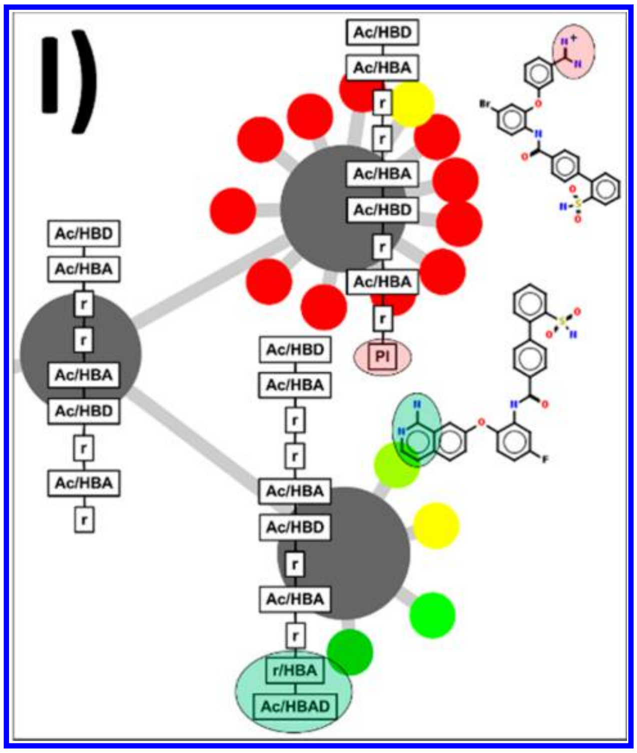


Figure 17. Example **I)** of an activity switch from Figure 16. RG-MCSs translated into encoding features (cf. Table 1). For details see text.

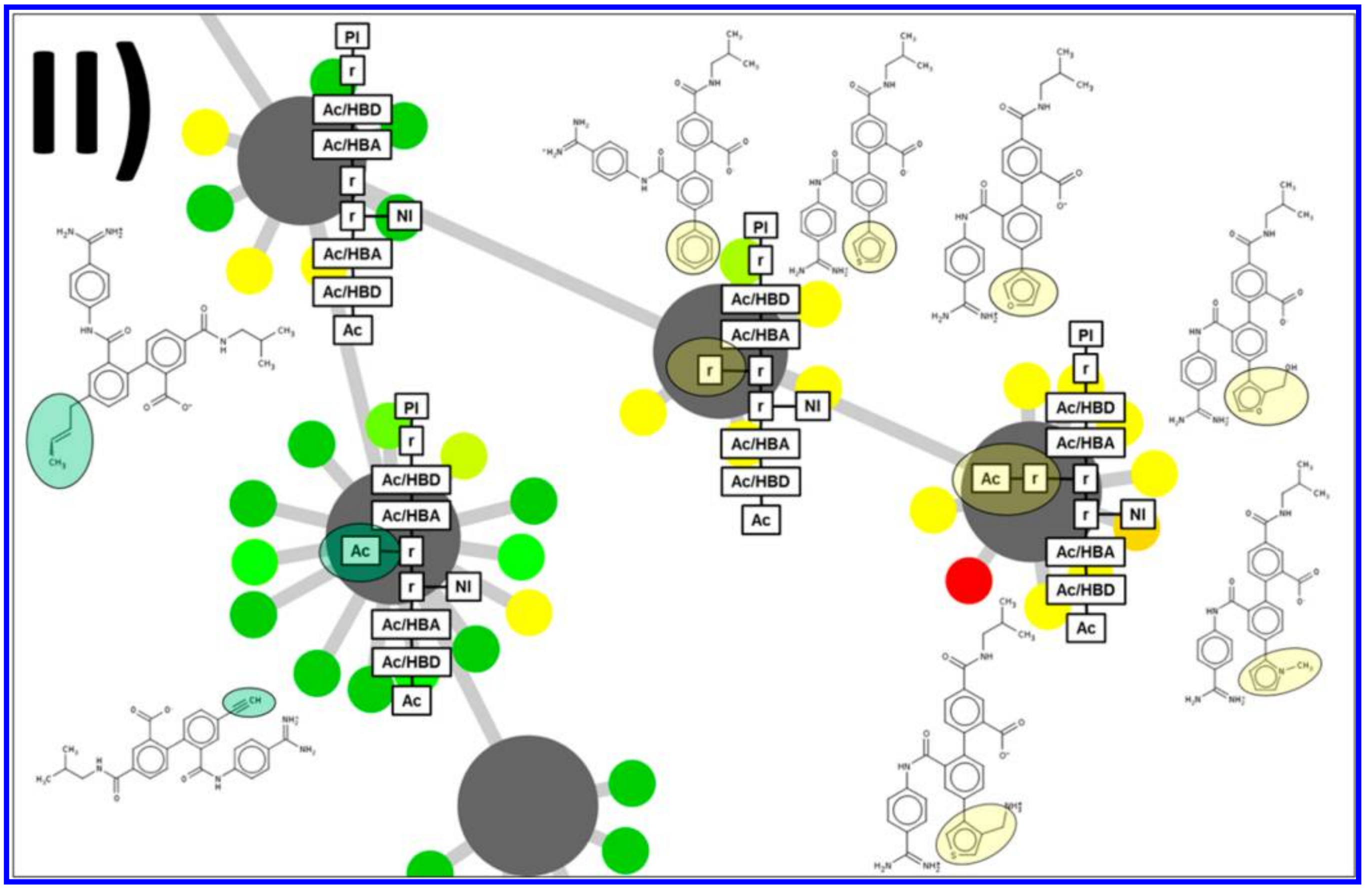


Figure 18. Example **II)** of an activity switch from Figure 16. RG-MCSs translated into encoding features (cf. Table 1). For details see text.

on the left-hand side. By comparing the corresponding RG-MCSs (see also Figure S 3 and Figure S 4 in the Supporting Information for the original chemViz output using RG codes), this SAR trend can easily be interpreted. Adding a PI group (encoded as Nb in the RG, e.g. amidine group in the molecule) to the phenyl ring leads to highly active molecules, whereas adding a non-PI feature results in lowly active molecules. This is already known from the activity cliff example explained in Figure 10. Here, another characteristic of inSARa networks becomes evident. As molecules can appear multiple times in the network, molecules can be seen in a different structural context, and thus SAR trends might be revealed in a different way. The risk of missing important information decreases. Nevertheless, one disadvantage is the higher complexity of the resulting networks. The second example **II)** (Figure 18) demonstrates another type of activity switch. While the features represented by the smallest MCS node at the top are not sufficient to separate lowly and intermediately active molecules (yellow as well as green molecule nodes), the larger neighboring nodes clearly separate both activity classes. As seen in the example below by using the corresponding RG-MCSs, interpretation can be facilitated. The addition of an aromatic group to one of the aromatic rings leads predominantly to intermediately active molecules, whereas adding different acyclic terminal groups, such as alkyl or alkenyl groups, does not result in a potency increase.

When comparing both types of SAR networks (inSARa in Figure 8 and Figure 9, the fingerprint-based similarity network in Figure S 2 (Supporting Information)), several strengths and weaknesses of these distinct approaches can be highlighted. Both techniques target systematic SAR visualization and analysis but capture molecular similarity by a different type of molecular representation and similarity metric. One benefit of inSARa is the hierarchical, treelike network structure which facilitates network navigation and intuitive SAR interpretation. This clear-cut structure is missing in the fingerprint-based network. With the help of the RG-MCS node, the presumed chemical similarity between different molecules connected to one single MCS node can be explained in inSARa networks (as demonstrated above). In fingerprint-based networks, by contrast, the reasons for similarity often remain unclear. This limits SAR analysis. By indicating common pharmacophoric features in inSARa, interpretation is not only facilitated but also abstract relationships not encoded in the usual fingerprint types like ECFP4 and MACCS keys may be revealed. However, due to the level of abstraction sometimes molecules are misleadingly grouped together. Especially on smaller MCS nodes, this probability is high. For effective SAR interpretation of fingerprint-based networks, extensions, such as the SAR pathways,[6] are necessary. One advantage of fingerprint-based networks is the lower complexity of the network (due to the single occurrence of each molecule in the resulting network) and the lower computational cost compared to inSARa where multiple occurrences are allowed and calculations are more demanding based on the complexity inherent in the MCS determination. To sum up, both concepts complement each other as they both have advantages and disadvantages not inherent in the other approach and maximal effect from SAR analysis can be expected by reasonable combination of the information provided by both network types.

CONCLUSION

Herein, a new method for SAR visualization and analysis is introduced. In contrast to fingerprint based approaches, inSARa is based on well-defined substructure relationships using the intuitive concept of the maximum common substructure. Another important key feature is the hierarchical network structure which makes the interpretation straightforward. Due to the encoding of pharmacophoric features by the conversion of molecules to RGs, inSARa complements SAR analysis approaches based for example on MMP analysis (e.g., BMMSGs).

inSARa can be used to analyze data sets of different size (up to some thousands of compounds) and heterogeneity. The degree of network complexity can be adapted by some user-defined parameters. Thus, a trade-off between simplicity (i.e., interpretability) and the potential loss of important SAR information is possible. The analysis of random molecule pairs enables to exclude unspecific similarity information from the networks. Although no kind of bioactivity information is used in network generation, interpretable relationships are found in inSARa networks. This underlines that the molecular representation and comparison of similarity is meaningful. inSARa is able to extract important SAR information from large data sets typically found in lead or hit-to-lead optimization stages. Pharmacophoric patterns, (non)bioisosteric exchanges, "SAR hotspots", and "activity switches" can be identified. Moreover, "activity cliffs" can be easily detected without any similarity threshold which has always to be defined when using fingerprints. By showing molecules in different chemical environments, inSARa can help to reveal relationships (e.g., different binding modes) not directly evident in fingerprint based networks. Due to the hierarchical structure, interactive SAR interpretation is intuitive, and the chemical neighborhood can easily be explored by network navigation. Hence, related scaffolds and substitution patterns can readily be identified. This is potentially helpful for the identification of unexplored chemical space.

Currently, SARs have to be manually explored by navigating inSARa networks. Next steps will be the automated highlighting of the most interesting networks pathways. Alternative methods for the reduction of the network complexity and the introduction of elements of supervised learning in network generation are under investigation.

ASSOCIATED CONTENT

Supporting Information

Table S 1: Layout settings for Cytoscape. Table S 2: Characteristics of compound data sets. Table S 3: Blacklisted RGs with an empirical probability equal to or greater than 0.1%. Figure S 1: RG size distribution for each data set. Figure S 2: NSG of the FXa data set. Figure S 3: Figure 17 with property mnemonics replaced by RG code. Figure S 4: Same as Figure S 3 for Figure 18. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-531-3912751. Fax: +49-531-3912799. E-mail: k.baumann@tu-braunschweig.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully thank OpenEye Scientific Software for providing us with academic licenses of OEChem Toolkit, OEDepict, and Babel. We also gratefully thank ChemAxon for providing us with academic licenses of Marvin software and JChem. We sincerely acknowledge the developers of Open Babel, NetworkX, chemViz, and Cytoscape for providing their software open source. We also sincerely acknowledge the Binding Database and ZINC database for providing the analyzed data.

ABBREVIATIONS

BMMSG, bipartite matching molecular series graph; CAG, chemical analog graph; CB1, cannabinoid receptor 1; CDK2, cyclin-dependent kinase 2; COX2, cyclooxygenase-2; CS, common substructure; ErG, extended reduced graph; FXa, coagulation factor Xa; HBA, H-bond-acceptor; HBD, H-bond-donor; HBAD, joint H-bond-acceptor and -donor; HTS, high-throughput-screening; MCES, maximum common edge subgraph; MCIS, maximum common induced subgraph; MCS, maximum common substructure; MMP, matched molecular pair; MST, minimum spanning tree; NI, negatively ionizable; NSG, network-like similarity graph; P38, map kinase p38 alpha; PI, positively ionizable; RG, reduced graph; SAR, structure−activity relationship; SARI, structure−activity relationship index; SSSR, smallest set of smallest rings; Tc, Tanimoto coefficient; THR, thrombin

REFERENCES

(1) (a) 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. (b) Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K. Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool. *J. Chem. Inf. Model.* **2009**, *49*, 2221−2230.

(2) Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580−588.

(3) (a) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369−378. (b) 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* (15−16), 630−639. (c) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; van Drie, J. H. Navigating Structure−Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698−705. (d) Bajorath, J. Large-Scale SAR Analysis. *Drug Discovery Today: Technol.* **2013**, *10*, e419.

(4) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.

(5) 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.

(6) (a) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure−Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009**, *52*, 1075−1080. (b) Wawer, M.; Bajorath, J. Systematic Extraction of Structure−Activity Relationship Information from Biological Screening Data. *ChemMedChem.* **2009**, *4*, 1431−1438.

(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) Guha, R.; van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(9) Hasan, S.; Bonde, B. K.; Buchan, N. S.; Hall, M. D. Network Analysis has Diverse Roles in Drug Discovery. *Drug Discovery Today* **2012**, *17*, 869−874.

(10) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *Proceedings of the 222nd ACS National Meeting*, Chicago, IL, United States; August 26−30, 2001, Division of Chemical Information, American Chemical Society, Ed.: Washington DC, United States, CINF-032.

(11) Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(12) 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.

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

(14) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure−Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212−3224.

(15) Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007**, *47*, 354−366.

(16) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182−3193.

(17) 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.

(18) 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.

(19) Cho, S.; Sun, Y. Visual Exploration of Structure-Activity Relationship Using Maximum Common Framework. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 571−578.

(20) Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. MultiMCS: A Fast Algorithm for the Maximum Common Substructure Problem on Multiple Molecules. *J. Chem. Inf. Model.* **2011**, *51*, 788−806.

(21) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.

(22) Birchall, K.; Gillet, V. Reduced Graphs and Their Applications in Chemoinformatics. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Methods in Molecular Biology; Humana Press: 2011; Vol. *672*, pp 197−212.

(23) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260−270.

(24) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346−356.

(25) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2005**, *46*, 208−220.

(26) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−490.

(27) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145−2156.

(28) (a) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure−Activity Relationships. 1. Reduced Graph Queries. *J. Chem. Inf. Model.* **2008**, *48*, 1543−1557. (b) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure−Activity Relationship Models. 2. Using Multiobjective Optimization to Derive Multiple Models. *J. Chem. Inf. Model.* **2008**, *48*, 1558−1570.

(29) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of Reduced Graphs to Encode Bioisosterism for Similarity-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 1330−1346.

(30) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503−511.

(31) (a) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems: Los Altos, CA, 2006. (b) Daylight Theory: SMARTS. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed November 25, 2012).

(32) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(33) Taminau, J.; Thijs, G.; De Winter, H. Pharao: Pharmacophore Alignment and Optimization. *J. Mol. Graphics Modell.* **2008**, *27*, 161−169.

(34) Zuccotto, F. Pharmacophore Features Distributions in Different Classes of Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1542−1552.

(35) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.

(36) (a) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33. (b) *Open Babel*, version 2.3.1. http://openbabel.org/ (accessed November 18, 2012).

(37) *OEChemTK,* version 1.9.0; OpenEye Scientific Software Inc.: Santa Fe, NM,http://www.eyesopen.com (accessed February 12, 2013).

(38) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(39) Kruskal, J. B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Am. Math. Soc.* **1956**, *7*, 48−50.

(40) (a) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function Using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA, United States, August 19−24, 2008; Varoquaux, G., Vaught, T., Millman, J., Eds: Pasadena, CA, United States, 2008; pp 11−15. (b) *NetworkX*, version 1.6. http://networkx.lanl.gov/ (accessed November 18, 2012).

(41) (a) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498−2504. (b) Cline, M. S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; Hanspers, K.; Isserlin, R.; Kelley, R.; Killcoyne, S.; Lotia, S.; Maere, S.; Morris, J.; Ono, K.; Pavlovic, V.; Pico, A. R.; Vailaya, A.; Wang, P.-L.; Adler, A.; Conklin, B. R.; Hood, L.; Kuiper, M.; Sander, C.; Schmulevich, I.; Schwikowski, B.; Warner, G. J.; Ideker, T.; Bader, G. D. Integration of Biological Networks and Gene Expression Data Using Cytoscape. *Nat. Protoc.* **2007**, *2*, 2366−2382. (c) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. Cytoscape 2.8: New Features for Data Integration and Network Visualization. *Bioinformatics* **2011**, *27*, 431−432. (d) *Cytoscape*, version 2.8.2. http://www.cytoscape.org/ (accessed November 18, 2012).

(42) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A Freely Available Program to Mine Structure-Activity and Structure-Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 68−78.

(43) (a) Wallace, I. M.; Bader, G. D.; Giaever, G.; Nislow, C. Displaying Chemical Information on a Biological Network Using Cytoscape. *Methods Mol. Biol.* **2011**, *781*, 363−376. (b) UCSF chemViz (chemoinformatics plugin for Cytoscape). http://www.cgl.ucsf.edu/cytoscape/chemViz/ (accessed November 20, 2012).

(44) (a) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002**, *18*, 130−139. (b) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein−Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201. (c) BindingDB. http://www.bindingdb.org/ (accessed November 18, 2012).

(45) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(46) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687−2694.

(47) Stumpfe, D.; Bajorath, J. Assessing the Confidence Level of Public Domain Compound Activity Data and the Impact of Alternative Potency Measurements on SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 3131−3137.

(48) Wassermann, A. M.; Bajorath, J. Identification of Target Family Directed Bioisosteric Replacements. *Med. Chem. Commun.* **2011**, *2*, 601−606.

(49) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.

(50) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.

(51) *Molecular Operating Environment (MOE)*, version 2011.10; Chemical Computing Group: Montreal, Canada. http://www.chemcomp.com/ (accessed November 17, 2012).

(52) (a) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182. (b) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768. (c) ZINC^{12}. http://zinc.docking.org/ (accessed November 18, 2012).

(53) The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/home/publications/e-resources/inchi.html/ (accessed February 28, 2013).

(54) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(55) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(56) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inf.* **2010**, *29*, 366−385.

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

(58) PDB code: 1EZQ; FXa.

(59) PDB code: 2EI6; FXa.

(60) PDB code: 2J34; FXa.

(61) Su, T.; Wu, Y.; Doughan, B.; Kane-Maguire, K.; Marlowe, C. K.; Kanter, J. P.; Woolfrey, J.; Huang, B.; Wong, P.; Sinha, U.; Park, G.; Malinowski, J.; Hollenbach, S.; Scarborough, R. M.; Zhu, B.-Y. Design and Synthesis of Glycolic and Mandelic Acid Derivatives as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2279−2282.

(62) Song, Y.; Clizbe, L.; Bhakta, C.; Teng, W.; Li, W.; Wu, Y.; Jia, Z. J.; Zhang, P.; Wang, L.; Doughan, B.; Su, T.; Kanter, J.; Woolfrey, J.; Wong, P.; Huang, B.; Tran, K.; Sinha, U.; Park, G.; Reed, A.; Malinowski, J.; Hollenbach, S.; Scarborough, R. M.; Zhu, B.-Y. Design, Synthesis, and SAR of Substituted Acrylamides as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1511−1515.

(63) Shi, Y.; Sitkoff, D.; Zhang, J.; Han, W.; Hu, Z.; Stein, P. D.; Wang, Y.; Kennedy, L. J.; O'Connor, S. P.; Ahmad, S.; Liu, E. C.-K.; Seiler, S. M.; Lam, P. Y. S.; Robl, J. A.; Macor, J. E.; Atwal, K. S.; Zahler, R. Amino(methyl) Pyrrolidines as Novel Scaffolds for Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5952−5958.

(64) (a) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242. (b) RCSB Protein Data Bank - RCSB PDB. http://www.rcsb.org/ (accessed November 18, 2012).

(65) Straub, A.; Roehrig, S.; Hillisch, A. Entering the Era of Non-Basic P1 Site Groups: Discovery of Xarelto (Rivaroxaban). *Curr. Top. Med. Chem.* **2010**, *10*, 257−269.

(66) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(67) *MACCS Structural Keys*; Symyx Technologies, Inc.: Sunnyvale, CA. http://www.symyx.com/ (accessed February 28, 2013).

(68) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(69) ECFP4 fingerprints are calculated in MOE using ph4_ExtendedConnectivityFP.svl from SVL Exchange. http://svl.chemcomp.com/ (accessed March 14, 2013).