JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

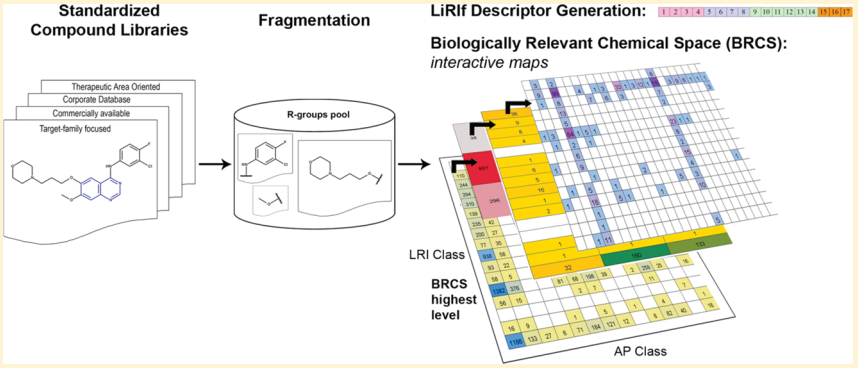
pubs.acs.org/jcim

Using Novel Descriptor Accounting for Ligand−Receptor Interactions To Define and Visually Explore Biologically Relevant Chemical Space

Obdulia Rabal and Julen Oyarzabal*

Small Molecule Discovery Platform, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pio XII 55, E-31008 Pamplona, Spain

Supporting Information



ABSTRACT: The definition and pragmatic implementation of biologically relevant chemical space is critical in addressing navigation strategies in the overlapping regions where chemistry and therapeutically relevant targets reside and, therefore, also key to performing an efficient drug discovery project. Here, we describe the development and implementation of a simple and robust method for representing biologically relevant chemical space as a general reference according to current knowledge, independently of any reference space, and analyzing chemical structures accordingly. Underlying our method is the generation of a novel descriptor (LiRIf) that converts structural information into a one-dimensional string accounting for the plausible ligand−receptor interactions as well as for topological information. Capitalizing on ligand−receptor interactions as a descriptor enables the clustering, profiling, and comparison of libraries of compounds from a chemical biology and medicinal chemistry perspective. In addition, as a case study, R-groups analysis is performed to identify the most populated ligand−receptor interactions according to different target families (GPCR, kinases, etc.), as well as to evaluate the coverage of biologically relevant chemical space by structures annotated in different databases (ChEMBL, Glida, etc.).

INTRODUCTION

Chemical biology, the discovery of tool compounds to advance the molecular understanding of biology (i.e., target validation), and medicinal chemistry, evolving from initial hit molecules to clinical candidates, play a major role in the drug discovery process. Thus, it is clear that to perform an efficient drug discovery project we have to direct our navigation strategies to the overlapping region where chemistry and therapeutically relevant targets, those related to ADME/Tox processes and primary activities as well as network pharmacology (off-targets), reside: "the biologically relevant chemical space".[1] These navigation strategies involve compound acquisition, library design and synthesis, library comparison, the selection of molecules for screening campaigns, and so forth. Therefore, taking into account that "much of chemical space contains nothing of biological interest",[2] a misleading navigation through infertile areas of the chemical universe, may involve an improper use of resources and time. In fact, recent reports suggest that "screening libraries may be improved by increasing the bias toward biogenic molecules"[3] and that "small molecules synthesized according to the biology-oriented synthesis approach are enriched in bioactivity".[4] This approach employs a hierarchical classification of bioactive compounds to select scaffolds as starting points for the synthesis of compound collections.

Thus, the definition and pragmatic implementation of the "biologically relevant chemical space" may be the critical point required to focus our navigation strategies on the chemical space most likely to interact with biological molecules and consequently to improve the drug discovery process. In this scenario, given that molecular representation is a key aspect of molecular diversity analysis, as it dictates the metrics and the techniques that can be used,[5] chemoinformatics plays an important role in defining, developing, and implementing a descriptor that captures protein recognition ligand−receptor interactions (LRI). Such a new descriptor may also lead to a representation of the "biologically relevant chemical space" (BRCS).

Received: December 29, 2011
Published: April 8, 2012

The chemical space definition is a relative concept that depends on the descriptors used,[6] which are typically structural and/or physicochemical. Computational chemical space representations do not aim to accurately reflect the chemical universe, but rather to provide reference frames for the projection or design of well-defined compound data sets.[7] In general, structural descriptors and two-dimensional (2D) fingerprints do not capture biological activity information.[8] There are many precedents where metrics for chemical diversity are questioned as diversity criteria to cover biological activity[9] simply because we move in a space where chemical diversity does not involve diversity in LRI; in fact, explored chemical space does not always overlap with the biological structure spaces formed by the target molecules.[10] Therefore, there is no well-defined relationship between the calculated similarity and the observed biological activity similarity.[8] Nevertheless, these 2D fingerprints have a history of successful applications in virtual screening for novel active compounds;[8,11] and this chemical diversity is still the underlying assumption of current drug design efforts.[12]

The design and navigation of chemical space representations, as well as the understanding of their topology, are very important topics in current chemoinformatics research. The first approach to navigating the chemical space, ChemGPS, was proposed one decade ago.[13] However, there is once again an "increasing interest in a systematic analysis of chemical space representations that are based on property descriptors or fingerprints".[7] Recent reports, for example, are focused on (i) how well compound libraries cover drug relevant chemical space[14] and (ii) how predictive and relevant reference spaces are using different types of descriptors,[12] as well as (iii) the development of reference tools, contours associated with each delimited reference chemical subspace (DRCS), to perform visual comparison of chemical libraries.[15] As very recently described, no generally applicable chemical space representations in chemoinformatics have been reported; in fact, key aspects of chemical space design have not been debated much, although the majority of chemoinformatics applications depends on these reference spaces.[7] Comparing the chemical space of compound collections is therefore not a trivial task because it is highly dependent on the method used and the structural representations of the compounds.[6,16]

Thus, our efforts are initially focused on these two important points: not only on the representation of the BRCS, but also on a generally applicable representation for which no reference space is required. Bearing in mind that techniques used to visualize chemical spaces are based on the underlying descriptors,[17] the initial step in our strategy involves the design and development of a descriptor that is relevant to and responds to different biological activities, accounting for LRI. Designing this descriptor, the ligand−receptor interaction fingerprint (LiRIf), leads to a definition of the BRCS that is fully independent of the analyzed chemical structures and therefore ready for any comparison analyses. The BRCS is predefined, based on established LRI types, and remains constant regardless of the specific structural features shown by the studied compounds. The BRCS is therefore populated according to the LiRIf generated for each analyzed structural motif, and any chemical structure can be represented in it. Structures from different compound libraries may be reported in the same representation of the BRCS, or each analyzed library may be used to independently describe the coverage of the BRCS, performing visual comparisons among compound collections. The representation of the BRCS has been implemented on an interactive visual platform with which the user can easily navigate from the highest level, where comparisons are performed, to the lowest, very detailed, level.

In this work, in addition to developing a new descriptor (LiRIf) and representing the BRCS, four reported compound data sets associated with four different target families (GPCRs, kinases, ion channels, and nuclear receptors) are analyzed with this novel descriptor and visually compared using their representations in the BRCS.

Methods for analyzing the structural features of molecules, known as chemical structure mining, can be broadly divided into two categories: methods that focus on structural parts, or fragments, and methods that consider the complete set of possible substructures of a molecule.[18] Analysis of fragment frequencies has proven useful in the description and comparison of molecular databases and for the identification of "chemical clichés".[19,20] Molecular fragmentation is therefore performed in this work. The reported analysis is focused on those fragments branching off from central scaffolds (R-groups); a scaffold analysis involves additional considerations and further methodological development, and will be reported in due time. Instead of clustering R-groups by linearly ordering them with respect to various classes of structural features,[21−23] fragments from each analyzed data sets are described by a LiRIf. Consequently, these R-groups are classified according to their LRI and represented in the BRCS.

An unbiased comparison between the reference independent space representations, one per analyzed target family, reveals which LRI plays a critical role (from the BRCS highest level), as well as which specific chemical substructures, or R-groups, are more frequent at each scenario (from the BRCS lower levels). The identification of the driving LRI for each target family, as well as an analysis of the co-occurrence of fragments, may yield valuable information,[18] not only for the design of new ligands, but also for similarity searching and library comparisons or acquisitions.

In summary, we report a simple and robust method for representing the biologically relevant chemical space (BRCS) and analyzing fragments accordingly from a drug discovery perspective. Underlying our method is the generation of a descriptor (LiRIf) that converts structural information from analyzed fragments into a one-dimensional string that accounts for their plausible ligand−receptor interactions, as well as for topological information. Capitalizing on ligand−receptor interactions as a descriptor enables one to cluster, profile, and compare libraries of compounds from a chemical biology and medicinal chemistry point of view. Thus, the proposed method should be able to facilitate the following tasks: (i) definition of the biologically relevant chemical space to focus our navigation tasks during the drug discovery process, (ii) data organization to organize and cluster the structures in a meaningful way, (iii) data visualization, (iv) data analysis, and (v) data mining.

METHODS

The overall workflow proposed in Figure 1 shows that the proposed strategy requires a preliminary data preparation process, which involves a customary database standardization followed by decomposition of compounds into fragments using a specific schema (discussed in the Case study section). Next, fragments are characterized on the basis of the potential of their substitution patterns to interact with a protein receptor, that is, on the basis of their "interacting structural moieties". In this

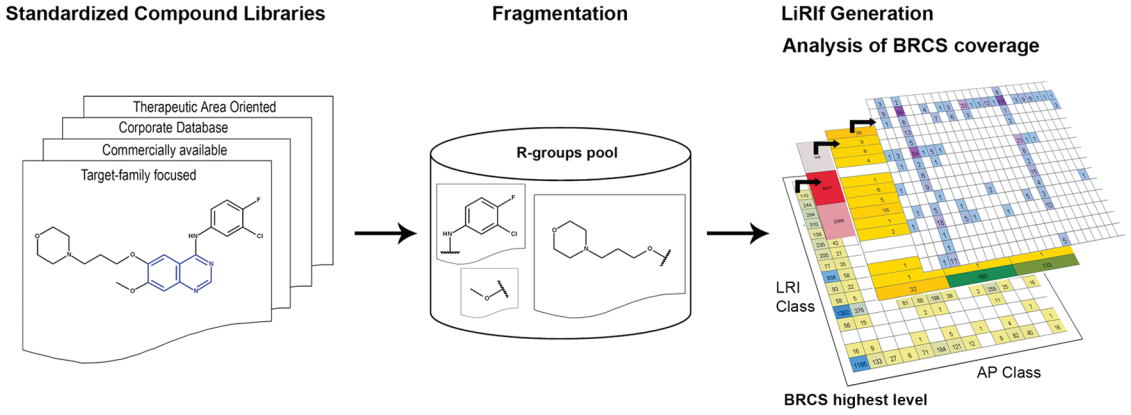


Figure 1. General workflow proposed for analyzing coverage of ligand−receptor interactions by R-groups extracted from compound libraries.

case, the fragment analysis is focused on R-groups. These moieties are defined as continuous sets of atoms in a given chemical environment that may account for certain type(s) of ligand−receptor interactions such as hydrophobic, polar hydrophobic, H-bond acceptor, H-bond, positively charged, negatively charged, and so forth. In most cases, the concept of "an atom in a given chemical environment" essentially agrees with the definition of organic functional groups (amides, ethers...) although in some cases, for chemical biology and medicinal chemistry purposes, we use broader definitions that do not fully match a single substructural search for a particular functional group, e.g., aliphatic heterocycles with functional groups or fused ring systems. Finally, the BRCS sampled by these fragments is assessed in terms of coverage of the ligand−receptor interaction classes (LRI classes) following the classification schema outlined in this article.

Initially, the discussion is focused on the procedure we have adopted to generate the fingerprint (LiRIf) that characterizes fragments, followed by a report of the data sets and fragmentation routine used in this paper.

Fragment Description. The process for describing and clustering fragments has four main sequential stages: (1) atom type assignation, (2) analysis and characterization of the direct chemical environment of the cleavage site or attachment point, (3) exploration of the remainder of the fragment after excision, and (4) generation of a fingerprint (LiRIf) encoding all extracted features that account for ligand−receptor interactions. These stages are described below and illustrated in Figure 2. On the basis of this new descriptor, fragments are clustered and represented in the BRCS, which is independently predefined by the LRI classes described below and consequently ready for visual analysis, comparison, and navigation.

(1). Atom Type Assignation. In order to identify and describe atom types that account for known ligand−receptor interaction, our analysis was based on (i) explicit experimental data annotated in the Protein Data Bank (PDB)[24] extracted from crystal structures of protein−ligand complexes and (ii) implicit information obtained from frequently occurring functional groups in known drugs, which covers those ligand−receptor interactions not exemplified yet, or poorly populated, in the PDB (e.g., compounds interacting with membrane proteins such as GPCRs, ion channels, and so forth). Thus, the list of atom types was inspired by the following two analyses: (i) the set of ligand atom types established in knowledge-based scoring schemes to account for the contribution of the interaction energies of protein−ligand

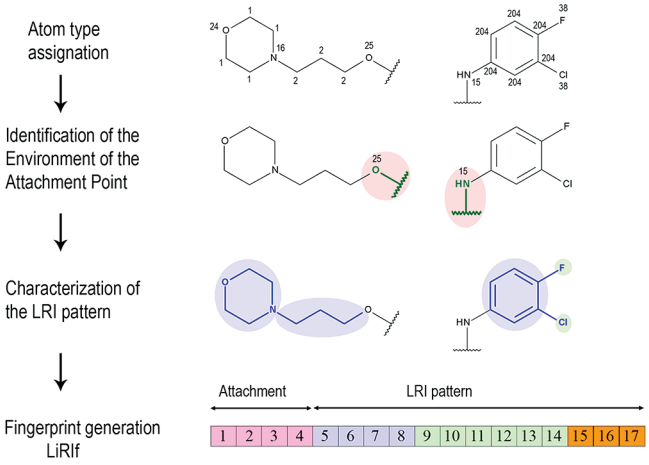


Figure 2. General workflow for fragment description. (1) Atoms are assigned an integer identifier. (2) The direct chemical environment of the connection point is perceived and analyzed. (3) The remainder of the fragment (LRI pattern) is analyzed based on its borne interacting structural moieties. (4) A 17-length array fingerprint (LiRIf) is generated containing features found for each fragment.

atom pairs calculated as potentials of mean force (PMF)[25,26] and M-Score,[27] and (ii) Ghose's analysis of functional groups in known drugs based on the Comprehensive Medicinal Chemistry database (CMC).[28] While we have mainly followed the updated Muegge's ligand atom type list,[26] some changes were required to include a detailed atom characterization for certain elements. For example, the CW polar sp2 carbon in PMF04 is differentiated here by the heteroatom bonded to the carbon; as well as by consideration of the tertiary amine in an aliphatic environment that is not explicitly described in PMF04 (this is part of NP class), but plays a critical role according to Ghose's analysis (monoaminergic GPCRs). Some other atom types considered in PMF04 (metals) are not included in this first version. By combining the LRI types detected in these two different compound-sourced studies, we believe that the atom type assignation is wide enough to allow for a proper classification of the fragments.

To identify the interacting structural moieties and to avoid manually drawing substructures, atoms are labeled with an integer value that corresponds to an atom type by mapping SMARTS[29] queries to each incoming fragment. The substructure mapper distinguishes 83 atom types (refer to Table S1 of the Supporting Information for a complete list including

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

their corresponding SMARTS). These are composed of various properties of the atom itself (element, ring membership, aromatic state, number of implicit hydrogens), as well as of its direct neighbors (atomic number, bond type, ring membership). For brevity, a subset of 35 atom types (13 of them also differing for cyclic and acyclic atoms) covering a broad spectrum of organic functional groups is listed in Table 1. The remaining atom types are designed to match a set of less common and more specific functional groups (Table S2 of the Supporting Information). Explicit hydrogen atoms are ignored; however, they are implicitly considered (Table 1).

Table 1. Abbreviated List of Broad-Spectrum Atom Types Used in This Work[a]

Nonpolar carbon sp3 aliphatic[b]
Polar carbon sp3/sp2 bonded to fluorine[b,c]
Polar carbon sp3/sp2 bonded to Cl,Br,I[b−d]
Nonpolar carbon sp2 aliphatic[b]
Polar carbon sp2 not aromatic bonded to oxygen(carbonyl)[b]
Polar carbon sp2 not aromatic bonded to sulfur (thiocarbonyl)[b]
Polar carbon sp2 not aromatic bonded to nitrogen (iminium)[e]
Polar carbon sp2 not aromatic bonded to nitrogen (amidinium)[e]
sp carbon
Aromatic carbon
Nitrogen with 3 connected atoms as an H-bond donor (NRH2)
Nitrogen with 3 connected atoms as an H-bond donor (NRRH)[b]
Nitrogen with 3 connected atoms as an H-bond acceptor (NRRR)[b]
Nitrogen with two connected atoms as an H-bond donor (=NH)
Nitrogen with two connected atoms as an H-bond acceptor (=NR)[b]
Nitrogen sp3 positively charged
Nitrogen (other than sp3) nonaromatic positively charged (iminium, amidinium)
Aromatic nitrogen as an H-bond acceptor
Aromatic nitrogen as an H-bond donor
Aromatic nitrogen substituted (pyrrole)
Aromatic nitrogen positively charged
Oxygen bonded to hydrogen (hydroxyl)
Oxygen bonded to atoms other than hydrogen[b,f]
Negatively charged oxygen
Aromatic oxygen
Sulfur bonded to hydrogen (thiol)
Sulfur with two connected atoms other than hydrogen[b]
Negatively charged sulfur
Positively charged sulfur (sulfonium)
Aromatic sulfur
Sulfonyl sulfur[b]
Sulfinyl sulfur[b]
Halogen bonded to aromatic carbon
Halogen bonded to nitrogen
Phosphorus

[a]For brevity, only the all-purpose atom types (not assignable to a unique organic functional group) are included. Atom types matching special features (rare organic functional groups) are listed in detail in Tables S1 and S2 of the Supporting Information. [b]Ring and acyclic atoms are labeled separately. [c]Independent of the number of halogen atoms on this carbon. [d]Carbons bearing combinations of fluorine and other halogens are assigned to this atom type. [e]Assigned atom identifier varies depending on the nature of the bonded nitrogen (NH, NR, and N+ between iminium cations and amidinium cations). [f]The specific neighbor (carbon or heteroatoms) is later recognized by the a DFS function (see text). For terminal H-bond acceptor oxygen (carbonyl) or sulfur (thiocarbonyl), no identifier is assigned for ease of implementation. The algorithm detects it on the basis of the nature of the corresponding neighbor carbon.

Although the central theme underlying this approach is closely related to pharmacophore models, it should be noted that the proposed annotation schema goes into more detail than standard pharmacophoric annotation points (e.g., carbonyl oxygen and ether oxygen, regardless of whether they are cyclic or acyclic, are typically labeled as hydrogen bond acceptors without further distinction; this also occurs for ether oxygen, regardless of whether it is in an aliphatic or aromatic environment). In addition to a finely tuned description of ligand−receptor interactions, we want to retain information on reactivity ("attachment point" to the central core), thus obtaining a detailed and chemically meaningful classification that will provide added value to medicinal chemists.

(2). Attachment Point (AP). Apart from synthetic accessibility considerations, the attachment point to the main scaffold of a series may play an important role on explicit ligand−receptor interactions, as in the case of kinase hinge binders, a fact that is sometimes not sufficiently highlighted.[30] In addition, depending on the nature of the attachment point, the electrostatics around the central core changes, which may be related to primary activity and/or off-target selectivity.[31] Thereby, its detailed analysis may lead to new areas of the BRCS for library comparison and space exploration, which may even have an impact on patentability.

With regard to the attachment point analysis, given a fragment with one attachment point, the direct chemical environment of the cleavage site is first perceived and analyzed. As atoms are individually annotated, a depth-first search (DFS) function was implemented to detect single continuous chemical functionalities starting from the connection point. This works by recursively grouping all bonded nonaromatic neighbors that are different from nonpolar aliphatic sp3/sp2 or sp carbon (nonaromatic carbons bonded to atoms other than hydrogen or carbon). By doing so, the possibility of different functional groups sharing atoms is excluded. These atom identifiers are collected into an initial variable-size string according to their order of appearance from the attachment point and attaching bonds. Each string unequivocally translates into a functional group conventional name or a well-defined chemical name for the underlying graph (i.e., cyclic systems) conversion that is done at the visualization stage to ease interpretation and is carried out using a comprehensive internal dictionary that has a total of 1414 unique string combinations. Figure 3 exemplifies several types of chemical environments (in green) for a set of representative fragments with their variable-size string shown in parentheses. For example, for fragment XVII, the string "25−7" corresponds to an O-linked ester (as opposed to "7−25", which corresponds to a carbonyl-linked ester, fragment III).

In the second round, functional atoms (with a pharmacophoric role) within an aliphatic ring are scanned to determine whether the ring assembly is composed of a noncontinuous combination of them (e.g., morpholine, piperazine) or not (e.g., piperidine). This feature is conveniently tagged by incorporating an extra flag into the variable-size string. An example of this is the morpholine of fragment X in Figure 3, which is captured as the concatenation of its constituent heteroatoms (16 for cyclic 3-connected nitrogen and 24 for cyclic oxygen) plus the "150" flag. For medicinal chemistry reasons, the −OCF3 group is considered as a single interacting moiety.

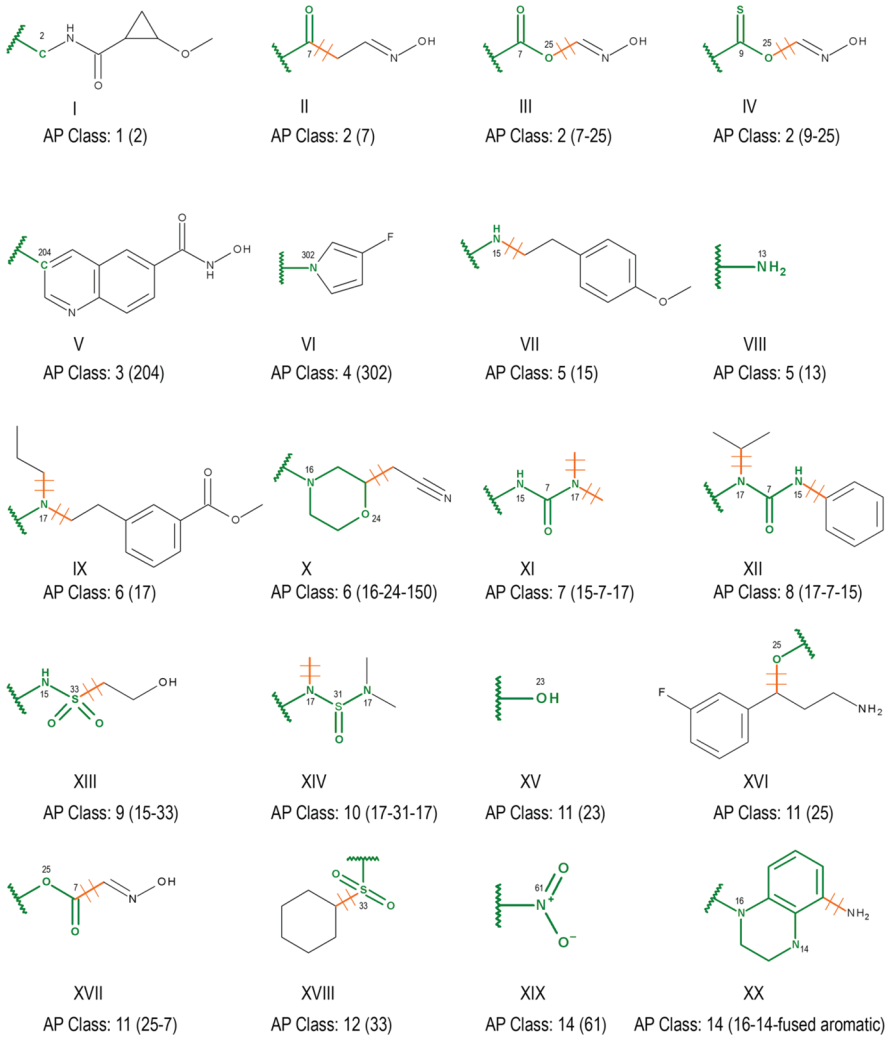


Figure 3. Examples of the detection of the environment of the attachment point (green). The final assignation to the attachment point class is given, and the variable size string representing the functionality of the attachment point is given in parentheses.

For simplicity, these chemical environments are clipped away after examination, and the atoms at the clip points are marked as new points of attachment, which are not considered in the following steps for atom type assignation where the rest of the R-group structure is analyzed. To clip certain linkers, some rules are applied to decide which group is the leaving group, and these are schematically indicated in Figure 3 as orange broken bonds.

- If the atom next to the anchor point is aromatic (fragments V, VI) or is an aliphatic carbon (different from nonaromatic Csp2 double bonded to O, S, N) (fragment I), the fragment is left uncleaved.
- If the functional group contains noncyclic bisubstituted nitrogens, as in tertiary amines (fragment IX), amides, sulpho(i)amides (fragment XIV),..., or ureas (fragments XI and XII), bonds for each growing vector are cut. From this point forward, each side chain is individually examined, and their associated information is later assembled for full description of the original fragment.
- If the atom bonded to the connection point is a cyclic nonaromatic nitrogen, the whole ring assembly is cut off, and all substituents on the ring assembly are separately analyzed (e.g., the morpholine of fragment X). Exocyclic double bonds attached to the removed ring atoms are considered part of the assembly and are removed as well.

At this point, small terminal fragments without further diversity such as halogen atoms, hydroxyl (fragment XV), nitro (fragment XIX), carboxylate,..., primary amine (fragment VIII), cyano, and single hydrogen are ready for classification and do not advance further.

For clustering purposes, each chemical environment will be assigned (see later in the paper) to any of the 14 different attachment point environment classes (AP classes) that we propose for this model (Table 2). These were defined considering not only the kind of the interaction pattern they provide (NH H-bond donor versus NR), but also from the perspective of the most common groups of substitution patterns covered in patent claims for the general Markush structure. It may be argued that this classification is not suitable for all purpose users, as some AP classes seem to be quite general (e.g., o-linked attachment points). However, as shown later in the paper, these classes are navigable according to different environments (e.g., o-linked esters vs o-linked ethers). Moreover, these subclasses could be easily upgraded to form a new AP class at the highest level of the BRCS map if necessary; thus, a fully customizable representation of BRCS is possible.

Table 2. List of Attachment Point Environment Classes[a]

AP Class	14 Attachment Point Environment Classes
1	A0-CX4 (C aliphatic)
2	A0-CO-X (A0-CS-X)
3	A0-c (C aromatic)
4	A0-n (N aromatic)
5	A0-NH-X
6	A0-NR-X
7	A0-NH−CO-X
8	A0-NR-CO-X
9	A0-NH-SO2-X
10	A0-NR-SO2-X
11	A0-O-X (A0-S-X)
12	A0-SO2-X (A0-SO-X)
13	Small Special Groups (A0-X, A0-CN)
14	MaxMix: phosphorus, rare functional groups and complex environments

[a]A0 refers to the attachment point to the central core from which R-groups are fragmented. For brevity, hydrogen is placed within the AP class 1.

Table 3. List of Ligand−Receptor Interaction (LRI) Classes[a]

LRI Class	17 Ligand−Receptor Interaction Classes
1	Hydrogen
2	Aliphatic (alkyls and fluorinated alkyls)
3	Aryl (planar hydrophobic)
4	Aromatic Heterocycles (Hydrophobic−polar, H-bond acceptor)
5	Primary/Secondary Amines (H-bond donor)
6	Tertiary Amines (Positively ionizable: H-bond donor)
7	Hydroxyl/Thiol (H-bond donor/acceptor)
8	(Thio)Ethers (H-bond acceptor)
9	(Thio)Ketones/Sulfones (H-bond acceptor)
10	Amides/Carbamates/Ureide/Sulfonamide (H-bond donor/acceptor)
11	Esters (H-bond acceptor)
12	Acids (H-bond acceptor/donor)
13	Aliphatic Heterocycles (H-bond acceptor/donor, positively ionizable, electronically deficient hydrogens)
14	Charged Groups; e.g. Guanidine/Amidine (positively ionizable: H-bond donor)
15	Small Special Groups (A0-X, A0-CN)
16	Rare functional groups and phosphorus
17	MaxMix: ≥2 branched functionalities

[a]Names assigned to each class are very generic in order to indicate the type of functionality they symbolize. However, most of them encompass less frequent functionalities related to them (i.e., hydroxylamines are grouped with hydroxyl groups). These functionalities initially grouped together, at the highest level, can be explored in detail by navigating from higher- to lower-level clusters in the hierarchy (see text). A0 refers to the direct attachment point to the central core.

The AP class number 14 groups together phosphorus, functional groups rarely found in MedChem libraries (azide, isocyanate,..., halides) as well as complex attachment point environments (combinations of contiguous functionalities). For example, attachment point environments consisting of cyclic tertiary nitrogen fused to aromatic rings are allocated to this class (fragment XX in Figure 3). The "small special groups" (AP class 13) allows for nitrile and halides to directly bind to the main scaffold in the original molecule, i.e., small commonly occurring fragments in structure−activity relationships (SAR) explorations.

(3). LRI Pattern. In the second step, after stripping off the functionality of the attachment point, the algorithm examines what is left of R-group (Figure 2). As mentioned, this subfragment will be referred to as the LRI pattern of the fragment, as we assume that in most MedChem programs the substitution pattern of the linker to the scaffold is commonly where the diversity of a R-group exploration for potency and ADME optimization resides. As was done for the attachment point environment, the next step is the classification of the LRI pattern. According to the two analyses described above,[25−28] we distinguish 17 different LRI classes for grouping together all LRI patterns (Table 3) and, on the basis of assigned atom types, the LRI patterns are clustered accordingly.

As previously mentioned, our aim when classifying fragments is to have a plausible and interpretable description of the covered BRCS. This means that for a fragment to be assigned to a given LRI class, its LRI pattern must be unequivocally attributable to any of the first 16 classes listed in Table 3. If this is not the case, it is assigned to the MaxMix class. Obviously, the number of fragments that contain several LRI classes is much greater than the number of perfectly "clean fragments" bearing just one LRI class: (i) most polar chains that explore solvent-exposed areas of the receptor normally contain combinations of donor/acceptor nitrogens and donor/acceptor oxygens (for instance, the gefitinib compound shown in Figure 1) and (ii) aromatic rings (benzene)/heteroaromatic rings (pyridine) are typically substituted to circumvent P450-mediated metabolism. If all these fragments were sent to the MaxMix class, the clustering schema we propose would be totally inadequate. Therefore, the MaxMix class for the total LRI pattern is restricted to those fragments that contain two (or more) branched LRI structural moieties from the attachment point (Figure 3, fragment XVI), as they may potentially interact with different pockets of the receptor in a "bidirectional branched" way, and a SAR interpretation is not unequivocally attributable to the presence/absence of any of these two (or more) LRI moieties. Monodirectional LRI patterns exhibiting a first nearest LRI class from the attachment point that are further substituted (Figure 3, fragments I and V) are assigned to the corresponding LRI class of this nearest interacting unit (from 1 to 16 in Table 3) and information on further substitutions is later indexed (as described in the next paragraph).

The aryl and heteroaryl classes deserve special attention as these LRI patterns comprise, apart from aromatic monocycles, fused ring structures that would usually be regarded from a MedChem perspective as scaffolds rather than scaffold decoration. Under the same assumption as above, interpretability of SAR features, the maximum size of so-called "aromatic" fused ring assemblies (contiguous ring structures containing at least one aromatic ring), is limited to two rings, allowing a maximum of two aromatic ring assemblies per fragment. Fragments that do not satisfy these requirements are sent directly to the MaxMix LRI class without further inspection.

The "rare functional groups" class stands for phosphorus and all those unwanted chemical motifs that are typically discarded in library design or that are seldom found in druglike libraries: alkyl halides (nonfluorine), diazene, nitro, azoxy, azides, acyl halides, isonitrile, nitrosamines, cyanoimines, and so on. Fluorinated alkyl groups (either cyclic or acyclic) are classified as part of the aliphatic class. The "small special groups" LRI class is equivalent to the "AP class" and was included to enable the displaying of these fragments at the LRI class level (see the section on BRCS description).

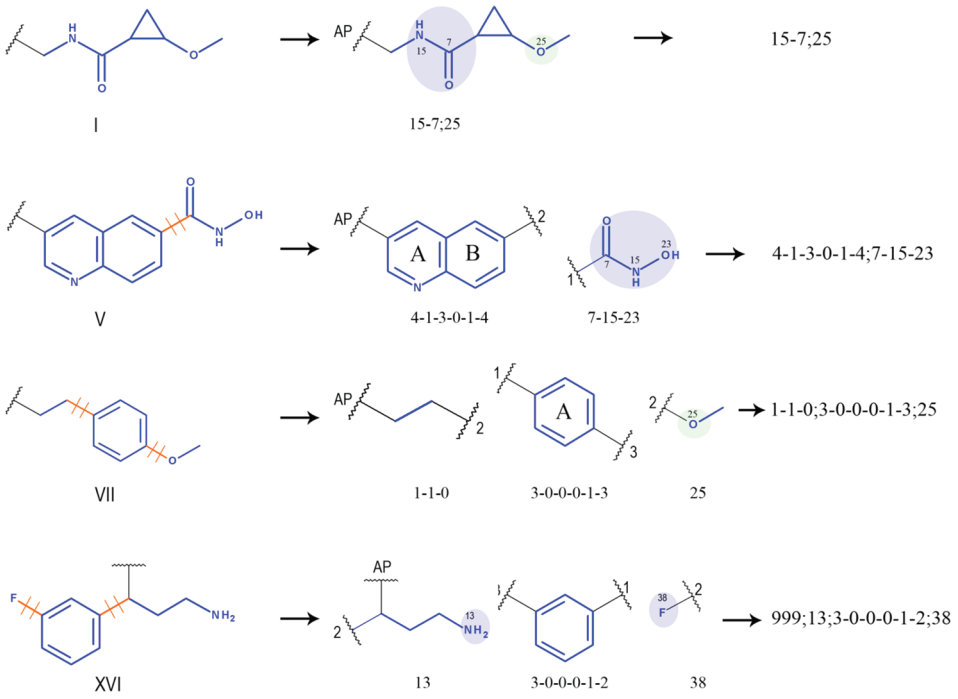


Figure 4. Examination of the LRI pattern for some representative fragments.

The analysis of the LRI pattern is implemented as follows. The LRI pattern is separated into "aromatic" fused ring assemblies and nonaromatic side chains: acyclic substructures and ring assemblies consisting only of nonaromatic heterocycles or aliphatic carbocycles. Thus, only acyclic linker bonds attached to "aromatic" assemblies are broken. Exocyclic double bonds between terminal heteroatoms (O,N,S) and the "aromatic" assembly are kept. An extra label is added to atoms that form part of a bond that is cut in order to retain knowledge of the connectivity between these generated subfragments. The process is illustrated in Figure 4 for four representative fragments. Hereafter, the two types of subfragments are characterized in a different manner depending on their aromaticity.

(a) "Aromatic" fused ring assemblies with a maximum of two rings (where at least one of them is aromatic) are indexed using a six-digit code (X) which provides information on the following:
- X1: ring-type (aryl, heteroaryl, aliphatic carbocycle or nonaromatic heterocycle) for the closest ring to the attachment point in the original fragment (*ringA*, e.g., pyridine ring of fragment V in Figure 4).
- X2: If *ringA* is a heteroaryl, this second digit takes into account atom-type distinctions such as number and nature (atomic number and H-bond donor or H-bond acceptor character) of their constituent heteroatoms (number of aromatic acceptor nitrogens, donor nitrogens, aromatic oxygens, and aromatic sulfurs).
- X3: equivalent to X1, calculated for the second ring (*ringB*, e.g., benzene ring of fragment V in Figure 4).
- X4: equivalent to X2, calculated for the second ring (*ringB*). Atoms that are shared between adjacent rings are counted just one time for the first ring (*ringA*).
- X5: degree of substitution of the complete assembly such as unsubstituted (clean assembly), monosubstituted, or polysubstituted (>1 substitution).
- X6: *ortho*, *meta*, *para*, or *other* (different from the previous three) substitution patterns for the monosubstituted assemblies.

Note that the words used here (aryl, heteroaryl, monosubstituted, ortho, etc.) and other chemical terms across the paper are convenient labels for this discussion. For computation, integer numbers (detailed in Scheme S1 of the Supporting Information) are used (in Figure 4 examples are provided with numerical values).

(b) Subfragments without aromatic ring moieties. The above commented depth-first search iterator is applied to detect single chemical functionalities, using the atom marked with an extra label as initial node to guide the exploration of the directed graph. It should be emphasized that atom type annotation is carried out before shredding the original fragment, and that atoms are therefore correctly recognized within their context in the original fragment. The result is a variable-size string constructed by sequentially concatenating each substring matching a functional group together (suitably separated by special characters), sorted according to its topological position when traversing the graph from the atom. That is, the first functionality corresponds to the one closest to the clip point, the second one is a further substitution of the first one and so on (see fragment I in Figure 4). Subfragments with branched functionalities are outputted with a dummy string indicator of the MaxMix class. Aliphatic heterocycles bearing more than one functional group (e.g., two nonconnected heteroatoms such as in the morpholine ring) are tagged with a flag indicating that these two functionalities belong, from the viewpoint of classification, to a single interacting moiety (as for the case of the AP class exemplified in Figure 3).

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

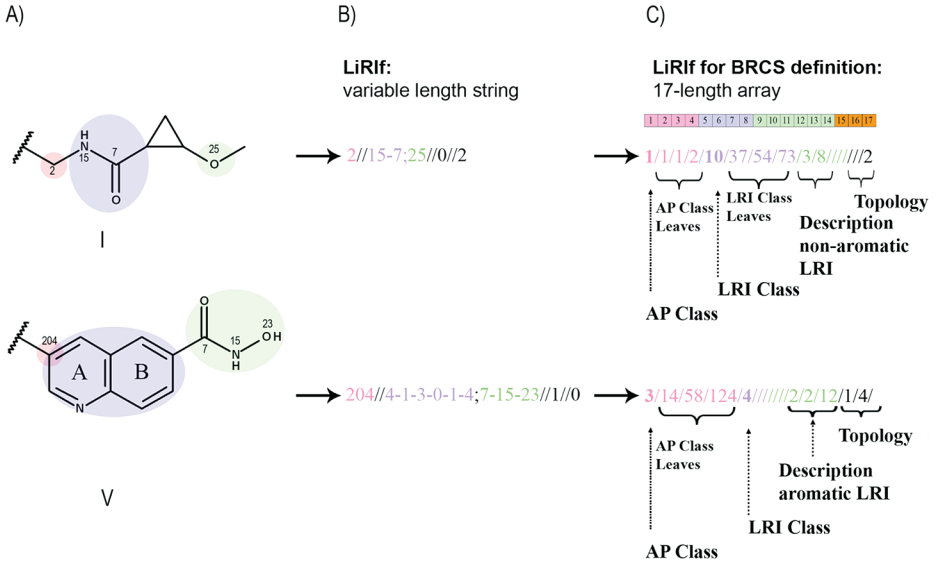


Figure 5. Fingerprint, or LiRIf, representation. LiRIf is a variable length string (b) whose dimensionality is reduced in the last step (c) to a 17-length array for BRCS representation for navigation and visualization purposes.

Finally, aliphatic chains consisting only of nonpolar carbon atoms are accordingly labeled as alkane (1-1-X), alkene (1-2-X) or alkyne class (1-3-X), where X = 0 or 1 for acyclic and cyclic chains, respectively (ethyl chain of fragment VII in Figure 4).

LRI fragments obtained by clipping away the direct chemical environment of the attachment point (compare with corresponding initial fragments in Figure 3) are further disconnected by cutting all acyclic bonds attached to "aromatic" ring assemblies (Fragments V and VII). Note that acyclic bonds attached to nonaromatic rings are kept, e.g., the cyclopropyl ring of fragment I. The breeding subfragments retain knowledge of their original connectivity by adding an extra mark to the newly generated clip points according to their position relative to the original attachment point (labeled as AP in Figure 4). For instance, in the case of fragment VII, the ethyl subfragment (first fragment) is bonded, as indicated in the figure, to subfragment "2" (benzene). In turn, benzene is connected to fragments "1" (ethyl) and "3" (methoxy). In the second round, fragments are classified according to their type. Aromatic moieties are assigned a six-digit code (see text). The quinoline of fragment V is coded as: 4 (heteroaryl *ringA*) − 1 (1 acceptor nitrogen) − 3 (aryl *ringB*) − 0 (featureless) − 1 (monosubstituted) − 4 (*other* position of the substitution). The benzene ring of fragment VII is: 3 (aryl) − 0 (featureless) − 0 (monocycle) − 0 (featureless) − 1 (monosubstituted) − 3 (at *para* position). In the final step, all codes are reconnected in the same configuration as the fragments in the original fragment. MaxMix combinations are found and tagged, as for example for fragment XVI, with a final string including the label "999" indicative of branched arrangements, i.e., comparing the position of the amine group with respect to benzene ring and the original attachment point.

(4). Generation of a Fingerprint, LiRIf, Encoding All Extracted Features. In the final step, all subfragment codes are merged into a unique string, describing from the attachment point to the LRI pattern, in the same configuration as the fragments in the original fragment. Again, fragments composed of combinations of branched functionalities or combinations of branched aromatic assembly and functionality are marked as MaxMix members (e.g., fragment XVI in Figure 4).

Additionally, topological distances between the attachment point of the original fragment and (a) the closest aromatic atom (if any) or (b) the closest atom in a functional moiety of the LRI pattern (if any) are also computed and annotated in the string describing each analyzed R-group (Figure 5b).

Definition and Representation of the Biologically Relevant Chemical Space (BRCS). From the above analysis, each analyzed chemical structure is characterized by a variable-length string containing information on the chemical environment of its attachment point, its LRI pattern, as well as the cited topological distances. Each fragment is therefore fully described by its corresponding LiRIf (Figure 5).

The BRCS is defined by the potential ligand−receptor interactions any analyzed chemical structure may establish. For ease of navigation in the BRCS, the proposed definition and representation of the BRCS is based on a two-dimensional (2D) map constituted, on the one hand, by the nature of the attachment point and, on the other, by the LRI pattern type. A hierarchical classification schema of this data, based on a tree-like set of rules that outputs a 17-length array encoding the fragment assignation to clusters, has been established (Figure 5), which leads to a navigation system from the highest level, where a 14 × 17 matrix (corresponding to nature of attachment points and LRI pattern types) is represented, to more detailed levels providing specific information for each particular cluster (Figure 6). Clustered data is presented in the form of heat maps. In an earlier work at Pfizer,[32] heat maps were introduced for SAR visualization, and their applicability is particularly suited here as they do not convey the spatial relationship between cluster members.

From now on, this 17-length array will be referred to as a "fingerprint", although we would like to point out that this fingerprint has been initially designed for clustering and visualization purposes, i.e., qualitative analysis, and it is not suited to similarity measures using customary metrics; further work to achieve a quantitative use is ongoing. Thus, although out of the scope of this application, the variable-length string could be formatted to create a binary bit-string encoding the

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

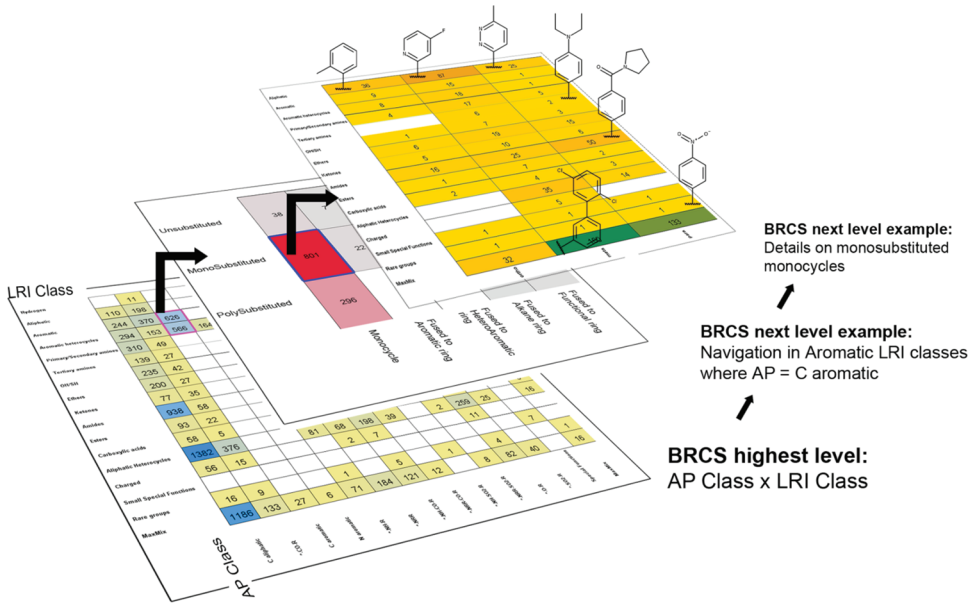


Figure 6. Visual representation of the BRCS: An interactive navigable 4D map. From the initial highest-level representation (AP class × LRI class), one can navigate to the BRCS next level(s) within selected cluster(s) by simply double-clicking on the cluster(s) of interest, thereby moving in the third dimension. In addition, this map can be color-coded by cluster population, by biological activity (best value within each cluster), or by any other property annotated together with the analyzed fragments, adding a fourth dimension.

presence/absence of certain fragment features in a similar manner to other standard substructurally based fingerprints such as MDL Keys.[33] Concerning the current qualitative format, this fingerprint contains an adequate amount of information to achieve our current goals, as details regarding up to the second site of variation of the LRI pattern are reflected. Nevertheless, it could also be easily modified to incorporate new categories that would enable higher-level detail nodes in the cluster tree (i.e., electrodonor or electrowithdrawing character of substitution patterns of aromatic rings, exhaustive characterization of LRI classes for more than two structural interacting moieties embedded in the LRI pattern, etc.).

The above-mentioned different categories of the 17-length array (Figure 5c) include the following:

(a) A hierarchical representation of the nature of the attachment point (dimensions 1−4). The assignation of the chemical environment of the attachment point to any of the 14 AP classes in Table 2 is saved in the first dimension of the fingerprint. For simplicity, the AP class names shown in Table 2 are very general prototypes of the chemical moieties contained within. Thus, the oxygen representation includes its equivalent role-based sulfur atom from an LRI viewpoint (carbonyl vs thiocarbonyl, ether vs thioether, etc.). Each of these 14 AP classes is further split into subgroups in a tree-like manner (if available for a particular AP class).
- The second dimension of the fingerprint adds further detail on the nature of the R substitution pattern (as annotated in Table 2) within the general AP class. For example, given the carbonyl AP class (A0-CO-R), this dimension differs between amides (R = NR′R″), esters (R = OR′), imides (R = NR′COR″) and the corresponding thiocarbonyl derivatives. Details are provided in a comprehensive list that is reported in Scheme S2 of the Supporting Information.
- For those functional groups having potential nonaromatic H-bond donor nitrogens, the third dimension differentiates between fragments carrying at least one NH donor and fragments without it. For example, for ureas under AP class 7 (A0-NHCO-R), the value assigned to dimension 3 indicates whether R = −NHR′ or R = −NR′R″.
- Finally, the fourth dimension corresponds to the leaves of the small tree for the AP class: well-defined chemical entities, that is, explicit specification for a functional group and particular substitution pattern for NR′R″ groups.

(b) A hierarchical representation of the LRI pattern (dimensions 5−17). Dimension 5 of the fingerprint stores the value of the LRI class as shown in Table 3. Depending on its value, two different sets of subsequent bit positions are filled up.

(b.1) If the LRI class matches a nonaromatic class (different from hydrogen or MaxMix classes), that is, class numbers 2 or 5 to 16 in Table 3, then bit positions from 6 to 11 give detailed information on the following:
- As for the case of the AP class (dimensions 2 to 4), dimensions 6−8 particularize the functional group contained within the higher/upper LRI class. For example, the general *aliphatic class* (LRI class 2 in Table 3) is divided into alkane, alkene, alkyne, and fluorinated alkyl groups, as defined by the number saved at the sixth position of the fingerprint. LRI class number 9 encompasses (thio)aldehydes,(thio)ketones, imines, sulfones, sulfoxides, and nitrile groups that are also discerned at this sixth level. Details are provided in a list that is reported in Scheme S3 of the Supporting Information.
- Dimension 9 specifies whether the LRI class is further substituted (second substitution

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

site) as exemplified in Scheme S4 of the Supporting Information.

- Providing that there is a second substitution site, dimension 10 specifies its precise LRI class using the same classification schema (Table 3) as for the main LRI pattern.
- If the second substitution site is of the (hetero)aromatic type, dimension 11 specifies whether it is an unsubstituted aromatic monocycle (offering a clean LRI interpretation) or not.

(b.2) If the LRI class corresponds to an aromatic class (classes 3 or 4 in Table 3), dimensions 12−14 account for the following:

- Dimension 12: Classification of the first ring (*ringA* in Figure 4) according to its type of assembly: (i) monocycle, (ii) fused to a second aromatic ring, (iii) fused to a second heteroaryl, (iv) fused to a carbocycle, and (v) fused to a functional nonaromatic heterocycle.
- Dimension 13: Degree of substitution of the assembly: (i) unsubstituted (clean assembly), (ii) monosubstituted, or (iii) polysubstituted (>1 substitution).
- Dimension 14: For monosubstituted assemblies, this dimension specifies the LRI class of its substitution motif. Here, for a derivatization pattern to match any of these classes (Table 3), the substitution pattern of the side chain must be exclusively composed of the named functionality (with either aliphatic linkers or aromatic clean monocycles). Otherwise, it is categorized as MaxMix class. Finally, topological distances and position of growing vector for aromatic cycles are organized as follows:
- Dimension 15: Topological distance between the attachment point of the original fragment and the closest aromatic atom (if any).
- Dimension 16: Position of the substitution in monosubstituted aromatic cycles (if applicable): (i) *ortho*, (ii) *meta*, (iii) *para*, and (iv) *other*.
- Dimension 17: Topological distance between the attachment point of the original fragment and the closest atom in a functional moiety of the LRI pattern (if applicable).

The fragment description and classification procedure have been implemented in-house, within Pipeline Pilot environment, by using the Perl API for Molecular toolkit.[34]

On the basis of the two analyses previously reported in the manuscript,[25−28] the atom type definition together with its corresponding identification process has been implemented to describe and then cluster R-groups according to their attachment points and LRI patterns. Thus, analyzed fragments can be represented in the BRCS, which is defined by a 2D-map based on known LRI (the 14 × 17 matrix). However, once new data is generated from a chemical biology and/or medicinal chemistry perspective, the current definition of the BRCS may change as a result of new types of LRI that may take place on the basis of novel chemistry and/or unexplored biological targets. The main limitation of the BRCS is therefore that it is based on a known LRI, which means that it should be updated whenever new LRI are described. Very recently, in the course of this year, four new papers have been published that report an updated analysis of the Protein Data Bank (PDB),[24] explicit LRI,[35−37] and the Drug Bank database.[38] The Drug Bank database combines detailed data about drugs and drug candidates with comprehensive drug-target information and in this case, proteins belong to many different classes, including pharmaceutically useful ones[38] and complementary target families poorly exemplified in the PDB. In order to validate the proposed definition of atom types, the corresponding clustering of fragments in the proposed AP and LRI classes, as well as the representation of the BRCS, data and results reported in those four articles are checked according to our classification method to assess our approach—are we really covering all known LRI?

In the case of the DSX scoring function,[35] we compared the overlap between the full set of 158 *fconv* atom types[39] and our proposed BRCS representation. Among these 158 *fconv* atom types, 25 correspond to water, halide ions, and metal ions, which are out of our current scope for describing the most common fragments describing LRI. The main difference arises from the distinction made by *fconv* for the different phosphorus and oxygen bonded to phosphorus atom types (13 atom types), which we group into a single atom type for phosphorus; thus, taking into account these two points, 121 *fconv* atom types are explicitly considered. Then, we can conclude that all these 121 atom types are perfectly represented on the BRCS. Very recently, two atomic distance-dependent statistical scoring functions were developed to model and rank protein−ligand interactions.[37] In this case, 26 atom types were used; our BRCS representation perfectly covers all of them. Additionally, our proposed model is able to project (Figure S2 of the Supporting Information) onto the BRCS all of the 315 unique fragments detected in the analysis of fragment-residue interaction profiles of ligands found in the PDB.[36] A total of 44 complex fragments, composed of phosphorus and aromatic ring assemblies with more than two rings are projected into the rare and MaxMix classes, respectively. As far as the analysis of the Drug Bank database is concerned,[38] authors found a total of 363 different bit positions of the PubChem Fingerprint (excluding atom count keys) corresponding to correlated chemical substructures, all of which are projectable in our BRCS proposal (100%).

Case Study. Data Mining and Data Sets Comparison. Once this new descriptor accounting for LRI is defined, the corresponding atom type assignation leads to the implementation of a proper classification of the fragments and a visual representation of the BRCS. This approach is then utilized in a real case study to mine information contained in chemical data sets of ligands for different target families and, taking advantage of the reference-independent nature of the BRCS, to compare mined chemical data among biological families. Through this analysis, our goal is to perform an unbiased and exhaustive mining of information from chemical data sets of pharmaceutically relevant target families to identify patterns of frequently occurring ligand−receptor interactions, as well as particular substructures that are discriminative for each target family ligands—general key driving forces for affinity. A comparison within a BRCS covered by fragments borne by ligands reported for each biological family pinpoints, from a generalist

Table 4. Data Set Composition[a]

	# Molecules	# Unique molecules after standarization	# Acyclic molecules (not fragmented)	# Molecules without R-groups (only rings without decoration)	# Unique R-groups generated	# R-groups filtered out (other elements not in FP analysis)	# R-groups filtered out (MW > cutoff of 350)	# R-groups for analysis
ChEMBL _Ion Channels All	32017	31837	273 (0.8)	168 (0.5)	**14989**	23 (0.15)	1837 (12.3)	**13129**
ChEMBL _Ion Channels BioActive	19570	19446	180 (0.9)	113 (0.6)	**9276**	12 (0.13)	1151 (12.4)	**8113**
KinaseSARFari All	51090	49650	127 (0.3)	191 (0.4)	**24380**	51 (0.21)	1836 (7.53)	**22493**
KinaseSARFari Bioactive	26714	26655	8 (0.03)	102 (0.4)	**13316**	29 (0.22)	896 (6.73)	**12391**
GPCR SARFari All	118834	110321	553 (0.5)	104 (0.1)	**57805**	88 (0.15)	20830 (36.0)	**36887**
GPCR SARFari Bioactive	80727	80164	271 (0.3)	62 (0.1)	**43960**	63 (0.14)	15511 (35.3)	**28386**
ChEMBL _NuclearReceptor All	23830	23699	100 (0.4)	15 (0.06)	**13190**	110 (0.83)	1200 (9.10)	**11880**
ChEMBL _NuclearReceptor BioActive	10586	10561	15 (0.1)	3 (0.03)	**5807**	1 (0.02)	531 (9.14)	**5275**
KKB Kinases All	201430	199185	373 (0.2)	232 (0.1)	**90081**	94 (0.10)	11600 (12.9)	**78387**
KKB Kinases Actives	108899	107979	46 (0.04)	104 (0.1)	**51383**	39 (0.08)	6328 (12.3)	**45016**
KKB Kinases All Patents	163903	162950	32 (0.02)	84 (0.1)	**78922**	56 (0.07)	10205 (12.9)	**68661**
KKB Kinases All Articles	30751	30316	329 (1.1)	104 (0.3)	**14659**	36 (0.25)	1348 (9.20)	**13275**
Glida	21192	21187	84 (0.4)	24 (0.1)	**13801**	11 (0.08)	3571 (25.9)	**10219**

[a]The initial number of molecules (# molecules), a summary of the number (with the percentage given in parentheses) of compounds/fragments rejected at each filtering step, and the final number of generated fragments for analysis (#R-groups for analysis).

perspective, those LRI that may lead to the achievement of selectivity. This case study is also used as an additional validation test to check if analyzed fragments can be properly represented in the defined BRCS (the 14 × 17 matrix).

Data Sets. Target-family related ligand databases were assembled from (a) the ChEMBL database[40] for Ion-channel, GPCR (GPCR SARFari), kinase (Kinase SARFari), and nuclear receptor inhibitors; (b) the GPCR-ligand GLIDA database;[41] and (c) the Kinase Knowledgebase database (KKB).[42] All databases, with the exception of the KKB, are publicly available online resources. GPCR SARFari and Kinase SARFari are a compilation of curated SAR data extracted from literature (Journal of Medicinal Chemistry and Bioorganic Medicinal Chemistry Letters) from 1980 onward. Particularly, GPCR SARFari is focused on Class A GPCRs. The KKB (Q1 2011 release) contains biological activity data for kinase inhibitors mined from scientific journals and patents. Finally, the GLIDA set was collected from various public resources, including PubChem[43] and the K_i Database.[44] The composition of these data sets is summarized in Table 4.

Compounds were standardized using a custom Pipeline Pilot protocol: salt and ion removal, consistent protonation state, mesomer representation, and duplicate elimination. For comparison purposes, compounds whose activity values (IC_{50}, K_i, K_d) were lower than 10 μM in enzymatic assays (reported as "Biochemical data" in ChEMBL data sets) were labeled "bioactive"; thus, according to this criteria, the active pool of compounds from all the data sets was separately extracted. The compounds in the KKB collection were also divided according to their data source: compounds retrieved from either articles or patents. Unique structures matching both categories were rejected.

Next, compounds were fragmented using an internally developed routine that detects, by using a very simple set of hierarchical rules, a single cyclic scaffold per molecule and then prunes all their decorating substituent positions (R-groups). Scaffolds are filtered out and not considered for further analysis. Although molecular fragmentation is not part of this study, because of the fact that the fragmentation scheme plays an important role and that the results depend on it, a brief description of the applied procedure is reported. This set of rules was designed to bias the search in favor of traditional medicinal chemistry criteria for scaffold detection and to abstract its role as a central skeleton adorned with R-groups (as done in classical SAR maps for congeneric series[45]). Thus, ring assemblies are ranked with regard to their number of aromatic fused rings, where the ranking is inversely proportional to the number of fused rings. Next, aromatic monocycles take preference over nonaromatic cycles. At each step, if scaffolds are equally scored, the ring assembly with the highest number of growing vectors or diversity points is chosen (with the central core taking precedence over peripheral rings). If it is still not possible to differentiate between ring assemblies, priority is given to those ring assemblies with more heteroatoms (N,O,S).

We are aware that more sophisticated methodologies have been formulated for scaffold detection,[46−48] but our main concern here was to develop a fast strategy for perceiving a single scaffold for the molecules rather than finding shared chemotypes in screening libraries. Our in-house-developed approximation for breaking down compounds in an onion-like manner[31] was not chosen for this study as it proceeds by successively identifying as scaffolds all ring systems contained within a molecule. By breaking down the list of R-groups at each position for each scaffold, the generated fragment space would be overrepresented by certain substructures, which would bias the analysis carried out here. On the other hand, the popular Bemis-Murcko molecular framework definition[49] was not contemplated here as it does not cleave linker chains between ring systems. However, from a MedChem perspective,

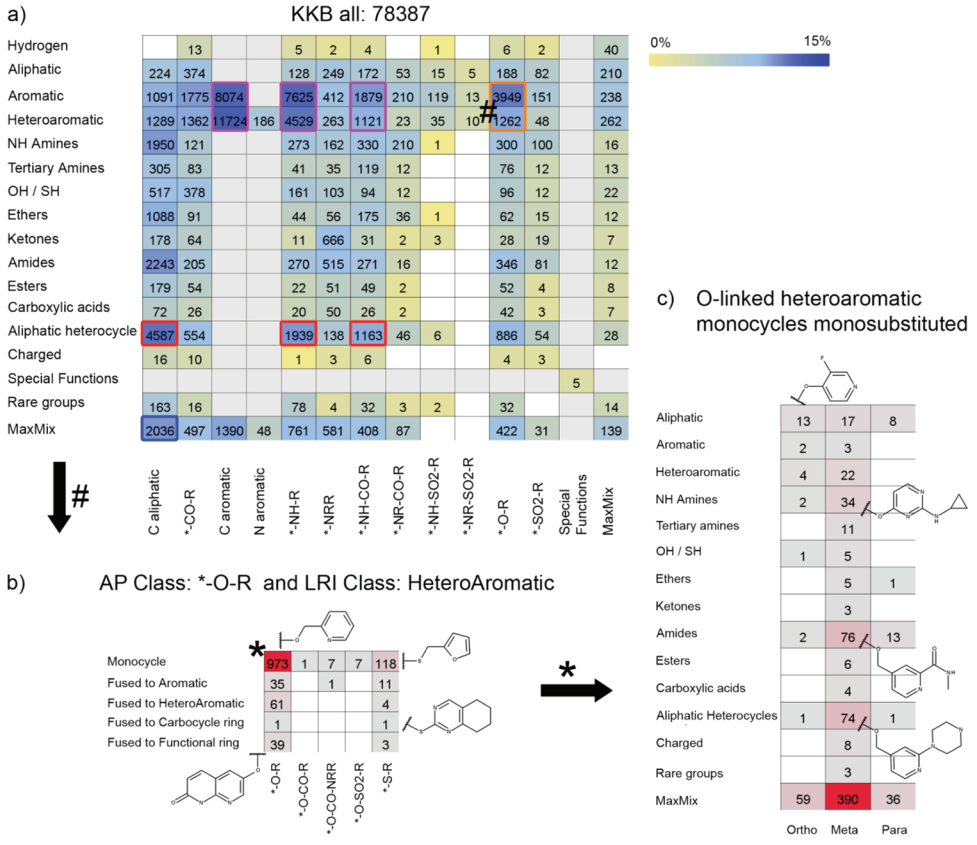


Figure 7. Heat maps showing the distribution of fragments extracted from (a) the KKB into the 238 clusters defined by the 14 AP classes (*x*-axis) and the 17 LRI classes (*y*-axis)—the BRCS highest level. Heat maps are color coded by cluster coverage (in percentage, logarithmic scale) and labeled with the total number of fragments. Light gray classes correspond to the 58 nonexisting pairwise combinations, i.e., the Special Functions AP class only matches, by definition, the Special Functions LRI class. (b) BRCS next level. The heat map of a focused subset of 1262 fragments, obtained after double-clicking on the corresponding cluster, with AP Class = 11 and LRI Class = 4 (marked with a black hash in Figure 7a). Classes are distinguished by the type of ring assembly (*y*-axis) and the specific functionality of the oxygen linker (*x*-axis). (c) BRCS next level. The distribution of the monosubstituted fragments, obtained after navigating through the selected cluster in Figure 7b, marked with a black asterisk, according to the LRI class of the substitution (*y*-axis) and its ring position (*x*-axis). Examples of substructures populating some of those clusters are also reported, which are also obtained from the interactive representation of the BRCS.

individual pendant rings are more frequently regarded as decorating R-groups than as part of the main scaffold.

Acyclic molecules (unfragmented), compounds consisting only of the central selected scaffold (without further decoration), fragments containing elements not regarded in this first version of the atom typing schema (metals, boron, silicon; see Table 4), and fragments with molecular weight >350 Da were discarded from this analysis. The number of resulting R-groups and the statistics of each filtering step for each collection are also shown in Table 4. Elimination of nonincluded atom types does not significantly impact results, as they represent on average a 0.18% of generated fragments. Computation times (standardization and fragmentation) ranges from ∼2 min for the smallest data set (ChEMBL NuclearReceptor Bioactive; 10586 compounds) to ∼30 min for the largest KKB set (201430 initial molecules).

■ RESULTS AND DISCUSSION

To illustrate the usefulness of our approach in inspecting and comparing libraries, we apply it to a set of target-focused libraries in order to examine the distribution of their resulting fragments into the different clusters defined by the fingerprint schema. At the highest hierarchical level of representation, the AP class assignation is depicted along the *x*-axis together with the LRI class value along the *y*-axis, yielding a total of 238 regions (14 AP classes ×17 LRI classes). Of the 238 regions, 58 regions are not truly accessible, as the corresponding pairwise combinations are unrealistic by definition. For example, all R-groups belonging to the Special Function AP class are assigned to the Special Function LRI class and vice versa; there are no additional options. Similarly, when "C aromatic" is AP there are only three options for LRI: "aromatic", "heteroaromatic" or "MaxMix", which contains R-groups with additional functions to aromatic or/and heteroaromatic. This heat map is the default initial representation for examining libraries, although any pairwise combination of fingerprint dimensions can be interactively selected and plotted at any time to gain more detailed knowledge of cluster composition at deeper hierarchy levels. The cluster coverage, expressed as the percentage of fragments within a cluster, is used for color-coding the heat map, although any numerical attribute (e.g., biological activity) is accepted. Fingerprint description and heat map visualization are integrated within a single protocol that takes between 3 and 40 min for the smallest and the largest data set (the R-group description is stored to speed up future visualizations).

Because kinase and GPRC inhibitor data sets are prime targets of interest, both for academia and the pharmaceutical industry, our discussion is focused on them. Details obtained

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

for the two remaining libraries, directed at ion-channel and nuclear receptors, are given in the Supporting Information in the form of heat maps (Figures S8, S9, S14, and S15), as well as tabulated data (SI_Distribution.xls). At this point, we want to emphasize that addressing the problem of detecting precise so-called "privileged substructures"[18,50−54] goes beyond what is relevant to our purposes of presenting the concept of clustering fragments holding similar LRI patterns and illustrating its use. Thus, the discussion will not be focused on individualized fragment frequencies, but on general LRI patterns sampled by different chemical fragments: the identification of privileged interaction types. However, as illustrated in Figure 7b,c and Figure 9c, by navigating through clusters of interest, those substructures that populate them are identified.

Kinase Focused Libraries. Analysis of KKB (Figure 7a) and Kinase SARFari (Figure 8) fragments reveals similar

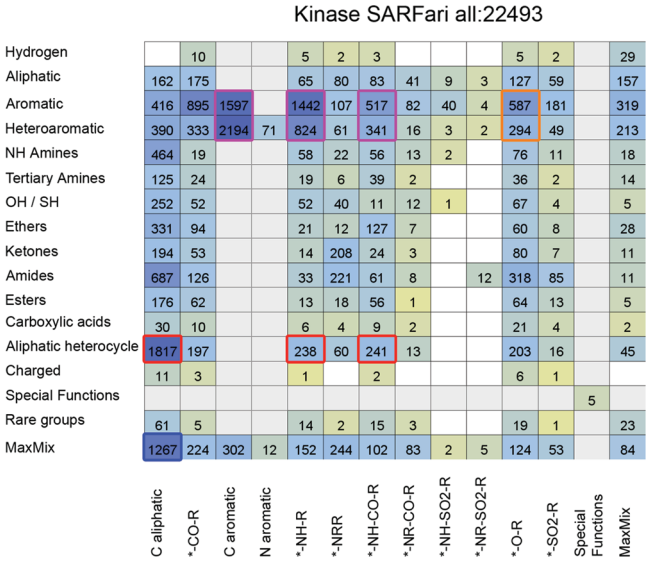


Figure 8. Heat map showing the distribution of fragments extracted from Kinase SARFari into the 238 clusters defined by the 14 AP classes (*x*-axis) and the 17 LRI classes (*y*-axis)—BRCS highest level. Heat maps are color coded by cluster coverage (in percentage, logarithmic scale) and labeled with the total number of fragments. Light gray classes correspond to the 58 nonexisting pairwise combinations.

profiles in cluster occupancy, despite differences in absolute percentage of cluster coverage. In part, this can be explained by the degree of chemical overlap between the two original sets: 24769 inhibitors are shared between the two libraries, which accounts for 49% and 12% of the Kinase SARFari and KKB collections, respectively. However, the KKB collection shows higher percentage of cluster coverage in those clusters bordered in pink and orange. Is this highlighting then a focused exploration on those clusters?

Heteroaromatic and aryl ring assemblies directly bonded to the main scaffold by an aromatic carbon or by an −NH- linker are the most populated clusters (bordered in pink in Figures 7a-8), representing 45% (KKB) and 31% (Kinase SARFari) of the total number of fragments. This is not surprising because these are recurrent fragments in kinase inhibitor design.[50,51,53] Heteroaromatic planar fragments mimic parts of the adenine moiety, where hydrophobicity together with a donor/acceptor role with the "hinge" residues of the protein are common LRI patterns for many kinase inhibitors, and decorating R-groups establish hydrophobic contacts with the back pocket of the ATP-binding site which is accessible through a "gatekeeper".[30] Note that scaffolds, primary defined as fused assemblies with a higher number of diversity points and heteroatoms, are not considered in this analysis. Enrichment in -NH- containing linkers (either as primary amines *-NH-R or bonded to a carbonyl *-NH-CO-R such as amides) is partly justified by its putative role as a hinge binder. This is true for most cases, but not all, e.g., the aniline R-group of gefitinib (Figure 1) does not interact via H-bond to the hinge region of EGFR.[55] Alternatively, H-bond donor amide and ureas (*-NH-CO-R linker) derivatized with aromatic groups are highly conserved features of type II kinase inhibitors targeting the specificity pocket: the DFG-"out" form, the closed conformation of the activation loop that exposes an additional hydrophobic site. Indeed, *-NH-R linkers can be found approximately 3−5 times more frequently than to *-NRR linkers. Concerning carbonyl linkers, *-NH-CO-R linkers are 6−8 times more likely to occur than *-NR-CO-R linkers. Oxygen is a common replacement for nitrogen-linked fragments, especially for cases where the H-bond donor is not required to interact with the hinge motif. Thus, the two clusters bordered in orange in Figures 7a and 8 comprise 6.6% and 3.9% of total fragments for KKB and Kinase SARFari, respectively.

Aside from the aromatic LRI patterns, the class of aliphatic heterocycles is also predominant (bordered in red in Figure 7a and Figure 8). These fragments are typically incorporated as solubilizing groups, although some of them are also well-established hinge binders (e.g., morpholino groups). Finally, the high number of compounds assigned to the MaxMix LRI class (dark blue in Figure 7a and Figure 8) is explained by the presence of tertiary nitrogens (tertiary amines or amides) that are disubstituted with two LRI classes different from aliphatic. Nevertheless, this percentage is lower in the KKB collection. Thus, again, is this pinpointing more focused and "clean" (reducing the MaxMix LRI class as much as possible) explorations in the KKB collection?

After examining the distribution of fragments at the top hierarchical level, we navigate through the interactive BRCS to lower levels, thus exemplifying the use of the fingerprint in focusing the analysis on particular clusters of interest. In Figure 7b, the 1262-membered cluster defined by AP class *-O-R and heteroaryl LRI Classes from the KKB library (marked with a black hash in Figure 7a) is further divided according to the specific functional group of the general AP Class (dimension 2 of the fingerprint, *x*-axis) and the type of ring assembly (dimension 12, *y*-axis). By doing so, 77.1% (973) of the fragments are distinguished as monocycles bonded to the scaffold via an ether moiety. Among them, 804 are monosubstituted (dimension 13) and hold the pattern of derivatization depicted in Figure 7c by plotting dimension 16 (*x*-axis) versus dimension 14 (*y*-axis) of the fingerprint. Dimension 14 is deliberately very restrictive with respect to the number of substitutions (only a single substitution that is not further derivatized is enabled to assign the LRI class at this stage), thereby increasing the occupancy of the MaxMix class.

When focusing on the subset of fragments extracted from known active compounds, the same general trend is observed as for the collections as a whole (Figures S3 and S4 of the Supporting Information). The most notable feature is an increase in the representation of the aromatic classes highlighted in pink in Figure 7a and Figure 8, especially for the Kinase SARFari collection (from 31% to 44%). The corresponding increase for the KKB is much

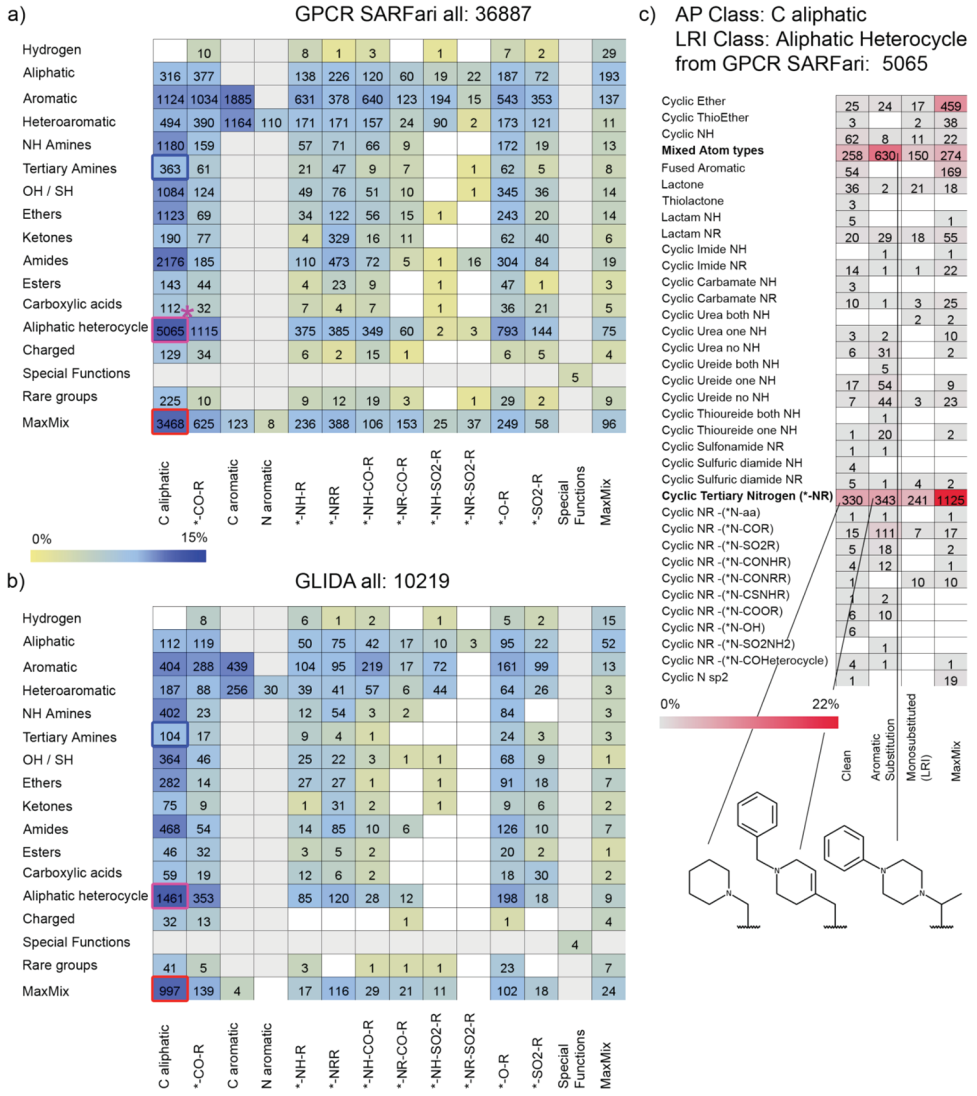


Figure 9. Heat maps showing the distribution of fragments extracted from (a) GPCR SARFari and (b) GLIDA into the 238 clusters defined by the 14 AP classes (*x*-axis) and the 17 LRI classes (*y*-axis)—the BRCS highest level. Heat maps are color coded by cluster coverage (in percentage, logarithmic scale) and labeled with the total number of fragments. Light gray classes correspond to the 58 nonexisting pairwise combinations. (c) BRCS next level. Heat map of a focused subset of 5065 fragments, obtained after double-clicking the corresponding cluster, with AP Class = 1 and LRI Class = 13 (marked with a pink asterisk in Figure 9a). Classes are distinguished by the specific functionality of the aliphatic heterocycle (*y*-axis; dimension 7 of the fingerprint) and the substitution pattern of this heterocycle (*x*-axis; dimension 9).

more modest (from 45% to 47%), probably because of the high rate of patent-described compounds lacking annotated biological data that are in most cases putative active compounds but which are not reflected in this subset. In fact, for fragments generated from compounds in patents, these classes have a rate of appearance 1.3-fold higher than the corresponding classes for fragments extracted from scientific journals (Figures S5 and S6 of the Supporting Information).

GPCR Focused Libraries. Relationships between the GLIDA and GPCR SARFari data sets are readily apparent from the heat map in Figure 9a (GPCR SARFari) and Figure 9b (GLIDA), although the initial sets of molecules partly overlap. In fact, the chemical space overlap between these two sets is not as high as one would expect: there are only 1909 common compounds (9% and 1.6% of the GLIDA and GPCR SARFari sets, respectively). In terms of biological space, GLIDA and GPCR SARFari ligands target a total of 377 and 446 receptors, respectively, and the target overlap is 275 GPCRs. This translates into a total of 20972 (GLIDA, 98.9%) and 99604 (GPCR SARFari, 83.8%) compounds targeting the same biological space. Therefore, although the two sets can be regarded as complementary in terms of chemical diversity, the represented biological space is pretty much the same, explaining the overall profiles obtained for these collections. There are 24 clusters covered by GPCR SARFari (13.3% of the total number of 180 possible clusters to be explored) and not by GLIDA. Although these fragments only account for 188 fragments (0.5%), GPCR SARFari apparently provides better sampling of the BRCS.

The most significant LRI class corresponds to aliphatic heterocycles, particularly linked with aliphatic carbon chains, grouping together 14.3% and 13.7% of the fragments from GLIDA and GPCR SARFari, respectively (bordered in pink in Figure 9a,b). The detailed analysis for this cluster for the GPCR SARFari set (5065 fragments; marked with a pink asterisk in Figure 9a) is shown in Figure 9c. The cyclic tertiary nitrogen (piperidine-like rings, labeled in bold as *Cyclic tertiary nitrogen* *NR in Figure 9c) and other combinations of planar tertiary

dx.doi.org/10.1021/ci200627v | J. Chem. Inf. Model. 2012, 52, 1086−1102

nitrogens with heteroatoms (piperazine, morpholine...labeled as *mixed atom types* in Figure 9c) are the most relevant groups: 2039 (40.3%) and 1312 (25.9%), respectively. This complies with the fact that many reported ligands interact with monoaminergic GPCR subtypes (adrenergic, dopamine, serotonin, histamine...).[56] In this sense, the low coverage of the tertiary amines LRI class (~1% for aliphatic carbon linkers, bordered in blue in Figure 9a,b) is due to the fact that these amines, as for the kinase, are grouped into the MaxMix class (9.4−9.7% for aliphatic carbon linkers, bordered in red in Figure 9a,b) on the basis of the bivalent LRI classes (both of them different from aliphatic carbons) attached to the nitrogen. These amines can be conveniently retrieved by navigating deeper into this class.

Concerning the active pool extracted from GPCR SARFari (Figure S7 of the Supporting Information), similar trends to that observed for the entire set are observed.

Through this case study we are able to identify the main capabilities achieved by the proposed method on the basis of the novel descriptor LiRIf. First, the BRCS is automatically represented as an interactive heat map and populated according to analyzed chemical structures (Figures 7, 8, 9, as well as those included in the Supporting Information). This representation does not require a reference space to compare analyzed data sets, and we can therefore directly evaluate and compare the coverage of the BRCS by each database, e.g., GPCR SARFari vs GLIDA (Figure 9a vs Figure 9b). Second, this cluster-based representation allows one to use visual exploration to identify the most frequent LRI according to analyzed target families and their corresponding bioactive molecules (e.g., the active pool of the GPCR SARFari set, Figure S7 of the Supporting Information). In addition, through this interactive representation, we can also identify those chemical substructures that populate each LRI cluster of interest (e.g., Figure 7b,c and Figure 9c).

CONCLUSION

Nowadays, the ability to define, represent, and visually explore the biologically relevant chemical space is a critical issue in performing an efficient drug discovery project. Misleading navigation through infertile areas of the chemical universe that contain nothing of biological interest is a problem that needs to be addressed. We therefore propose a simple and robust method for representing the BRCS. By means of this interactive representation of the BRCS, a four-dimensional heat map, we are able to focus our navigation tasks (library comparisons, compound design, etc.) during the drug discovery process.

The deterministic nature of the strategy should be noted: chemical structures will fall into the same class regardless of the context. This is in contrast to distance-based clustering schemas, where clusters are defined with respect to the particular pool of compounds to be grouped. Another advantage is that the chemist can immediately recognize the precise criterion defining the class, either by visualizing fragments contained within, or by reading the chemically driven names associated to each class (an intuitive guideline for medicinal chemists). This is in contrast to distance-based methods, which generate high-dimensional spaces difficult to visualize and where cluster commonality relies on the combination of the employed descriptors, metrics, and cluster method.

Underlying our method is the generation of a novel descriptor (LiRIf) that converts structural information from analyzed fragments into a one-dimensional string accounting for their plausible ligand−receptor interactions as well as for topological information. Capitalizing on ligand−receptor interactions as a descriptor enables one to cluster, profile, and compare libraries of compounds from a chemical biology and medicinal chemistry point of view. Thus, the proposed method facilitates the following key tasks: (i) definition of the BRCS, (ii) data organization from a LRI perspective, (iii) data visualization leading to an easy interpretation of trends in structure−activity relationships (SAR) by medicinal chemists, (iv) data analysis to enable the comparison in a reference independent space as well as the profiling and identification of relevant chemical features, and (v) data mining to help search for structures that contain key interactions or specific features.

The definition of the BRCS is based on current knowledge of reported ligand−receptor interactions, and the representation of the BRCS will therefore need to be updated when new LRI are identified. Currently, on the basis of this novel descriptor, we are working on visualization and decision-making tools for drug discovery projects, patents analysis, reagents selection, and libraries acquisition, the results of which will be reported in due course.

ASSOCIATED CONTENT

Supporting Information

Tables S1 and S2 describing the complete list of atom types used for labeling atoms as well as matching particular functional groups. Scheme S1 provides details for the characterization of aromatic fused ring assemblies, Scheme S2 describes specific functionalities under each AP, and Scheme S3 enumerates those functionalities under each LRI class. Scheme S4 reports detailed specifications of the 17-length fingerprint (for dimensions 9, 10, and 11). Figures S3 to S15 are also included; they represent the corresponding top-level hierarchy heat maps for each data set reported in the manuscript. Finally, a tabulated data set, as a xls file, is also provided. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*Phone: +34 948 194700. E-mail: julenoyarzabal@unav.es.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the Foundation for Applied Medical Research (FIMA) of the University of Navarra for financial support and Dr. Heidi Rohwer for manuscript editing. In addition, we thank Dr. Hiroaki Yabuuchi and Dr. Yasushi Okuno for providing access to the GLIDA data set and the reviewers for their constructive feedback. We dedicate this article to Professor Francisco Palacios on the occasion of his 60th anniversary. This work was partially supported by MINECO and FSE (Inncorpora-Torres Quevedo grant), PTQ-11-04781.

REFERENCES

(1) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432* (7019), 824−828.

(2) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432* (7019), 823.

(3) Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **2009**, *5*, 479−483.

(4) Wetzel, S.; Bon, R. S.; Kumar, K.; Waldmann, H. Biology-oriented synthesis. *Angew. Chem., Int. Ed.* **2011**, *50*, 10800−10826.

(5) Gorse, A. D. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* **2006**, *6* (1), 3−18.
(6) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48* (1), 240−248.
(7) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50* (2), 205−216.
(8) Heikamp, K.; Bajorath, J. How do 2D fingerprints detects structurally diverse active compounds? Revealing compound subset-specific fingerprint features through systematic selection. *J. Chem. Inf. Model.* **2011**, *51* (9), 2254−2265.
(9) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350−4358.
(10) Wess, G. How to escape the bottleneck of medicinal chemistry. *Drug Discovery Today* **2002**, *7* (10), 533−535.
(11) Oyarzabal, J.; Zarich, N.; Albarran, M. I.; Palacios, I.; Urbano-Cuadrado, M.; Mateos, G.; Reymundo, I.; Rabal, O.; Salgado, A.; Corrionero, A.; Fominaya, J.; Pastor, J.; Bischoff, J. R. Discovery of mitogen-activated protein kinase-interacting kinase 1 inhibitors by a comprehensive fragment-oriented virtual screening approach. *J. Med. Chem.* **2010**, *53* (18), 6618−6628.
(12) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108−119.
(13) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical Space. *J. Comb. Chem.* **2001**, *3* (2), 157−166.
(14) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* **2009**, *49* (4), 1010−1024.
(15) Le Guilloux, V.; Colliandre, L.; Bourg, S.; Guénegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual characterization and diversity quantification of chemical libraries: 1. Creation of delimited reference chemical subspaces. *J. Chem. Inf. Model.* **2011**, *51* (8), 1762−1774.
(16) Medina-Franco, J. L.; Martínez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the chemical space in drug discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (4), 322−333.
(17) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 325−330.
(18) Van der Horst, E.; Okuno, Y.; Bender, A.; Ijzerman, A. P. Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. *J. Chem. Inf. Model.* **2009**, *49* (2), 348−360.
(19) Lameijer, E. W.; Kok, J. N.; Bäck, T.; Ijzerman, A. P. Mining a chemical database for fragment co-occurrence: Discovery of "chemical clichés". *J. Chem. Inf. Model.* **2006**, *46* (2), 553−562.
(20) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F.; Schenck, R. J.; Trippe, A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73* (12), 4443−4451.
(21) 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* (4), 912−926.
(22) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1302−1314.
(23) Pastor, J.; Oyarzabal, J.; Martinez, S. Navigating MedChem Space: Key for Multi-Factorial Optimization in Drug Discovery. Presented at the Computer-Aided Drug Design Gordon Research Conference [Online], Tilton, NH, July 31−August 5, 2005. http://www.grc.org/programs.aspx?year=2005&program=cadd (accessed December 22, 2011).
(24) 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* (1), 235−242.
(25) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791−804.
(26) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49* (20), 5895−5902.
(27) Yang, C. Y.; Wang, R.; Wang, S. M-Score: A knowledge-based potential scoring function accounting for protein atom mobility. *J. Med. Chem.* **2006**, *49* (20), 5903−5911.
(28) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1* (1), 55−68.
(29) (a) Weininger, D. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36. (b) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.
(30) Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "gatekeeper door": Exploiting the active kinase conformation. *J. Med. Chem.* **2010**, *53* (7), 2681−2694.
(31) Oyarzabal, J.; Howe, T.; Alcazar, J.; Andrés, J. I.; Alvarez, R. M.; Dautzenberg, F.; Iturrino, L.; Martínez, S.; Van der Linden, I. Novel approach for chemotype hopping based on annotated databases of chemically feasible fragments and a prospective case study: New melanin concentrating hormone antagonists. *J. Med. Chem.* **2009**, *52* (7), 2076−2089.
(32) Kibbey, C.; Calvet, A. Molecular property eXplorer: A novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* **2005**, *45* (2), 523−532.
(33) *MACCS-II Database System*, version 1; Molecular Design Limited: San Leandro, CA, 1984.
(34) *Pipeline Pilot*, version 8.5; Accelrys: San Diego, CA, 2011.
(35) Neudert, G.; Klebe, G. DSX: A knowledge-based scoring function for the assessment of protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51* (10), 2731−2745.
(36) Wang, L.; Xie, Z.; Wipf, P.; Xie, X. Q. Residue preference mapping of ligand fragments in the Protein Data Bank. *J. Chem. Inf. Model.* **2011**, *51* (4), 807−815.
(37) Fan, H.; Schneidman-Duhovny, D.; Irwin, J. J.; Dong, G.; Shoichet, B. K.; Sali, A. Statistical potential for modeling and ranking of protein−ligand interactions. *J. Chem. Inf. Model.* **2011**, *51* (12), 3078−3092.
(38) Yamanishi, Y.; Pauwels, E.; Saigo, H.; Stoven, V. Extracting sets of chemical substructures and protein domains governing drug-target interactions. *J. Chem. Inf. Model.* **2011**, *51* (5), 1183−1194.
(39) Neudert, G.; Klebe, G. Fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27*, 1021−1022.
(40) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2010. http://www.ebi.ac.uk/chembl (accessed December 22, 2011).
(41) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. *Nucleic Acids Res.* **2008**, *36* (suppl_1), D907−912.
(42) *Kinase Knowledgebase (KKB)*; Eidogen-Sertanty, Inc: San Diego, CA, 2011.
(43) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2008**, *36* (Database issue), D13−D21.

(44) Roth, B. L.; Lopez, E.; Beischel, S.; Westkaemper, R. B.; Evans, J. M. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.* **2004**, *102* (2), 99−110.
(45) 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* (24), 5926−5937.
(46) 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* (1), 47−58.
(47) Clark, A. M.; Labute, P. Detection and assignment of common scaffolds in project databases of lead molecules. *J. Med. Chem.* **2009**, *52* (2), 469−483.
(48) 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* (2), 503−511.
(49) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(50) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and kinase-privileged fragments: Toward virtual polypharmacology. *J. Med. Chem.* **2008**, *51* (5), 1214−1222.
(51) Vieth, M.; Erickson, J.; Wang, J.; Webster, Y.; Mader, M.; Higgs, R.; Watson, I. Kinase inhibitor data modeling and *de Novo* inhibitor design with fragment approaches. *J. Med. Chem.* **2009**, *52* (20), 6456−6466.
(52) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **2006**, *49* (6), 2000−2009.
(53) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51* (9), 2689−2700.
(54) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 1037−1050.
(55) Yun, C. H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **2007**, *11* (3), 217−227.
(56) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* **2002**, *3* (10), 928−944.