Journal of Medicinal Chemistry

Article

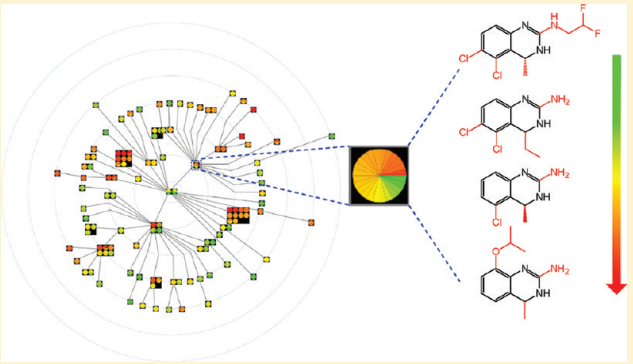
pubs.acs.org/jmc

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

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

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

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



INTRODUCTION

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

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

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

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

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

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

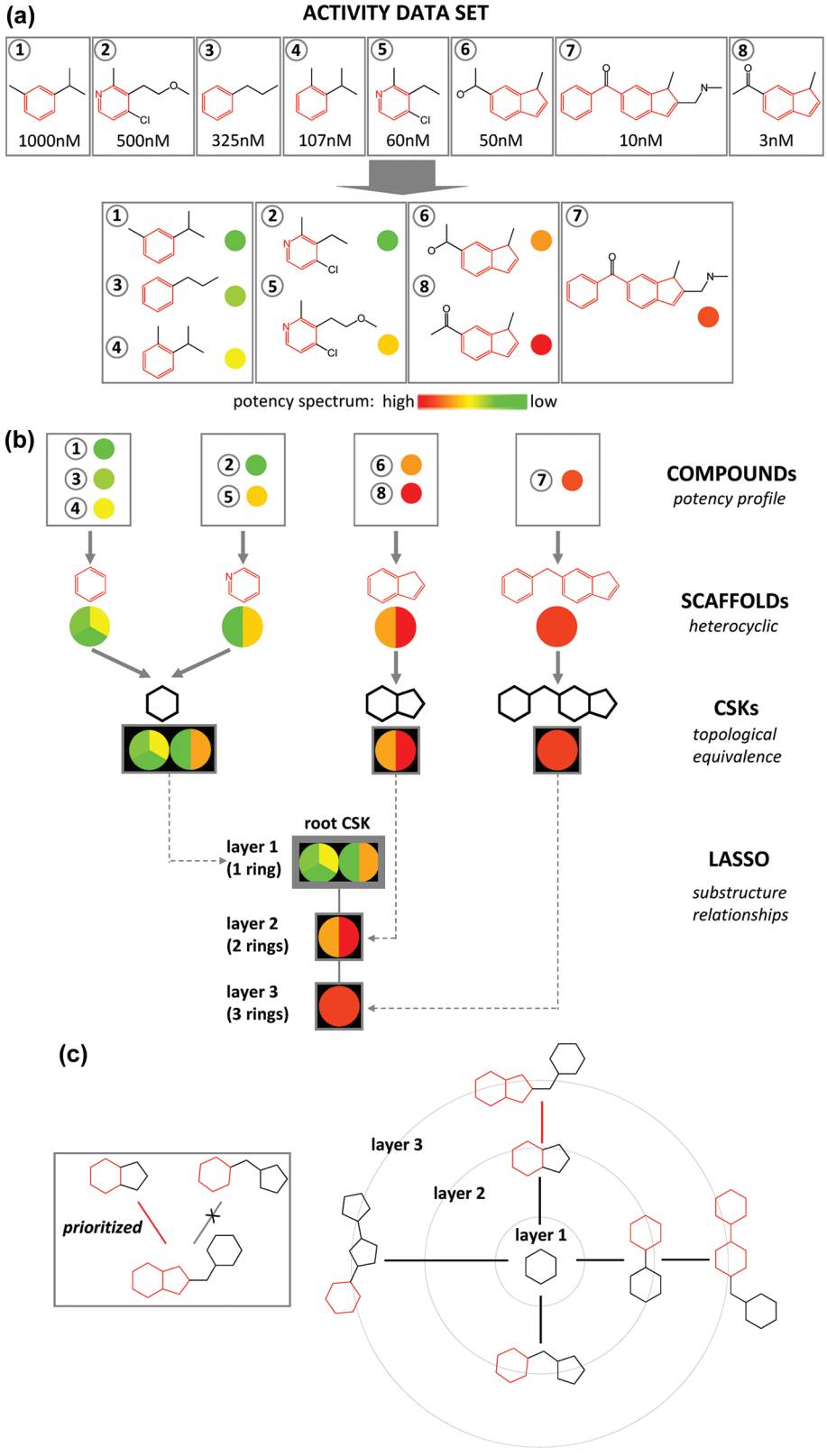


Figure 1. Graph generation. In (a) and (b), the generation of the LASSO graph is illustrated, as described in the text. In the exemplary data set, compounds are labeled with their potency (K_i) values. Scaffolds are colored red. In (c), substructure relationships between CSKs across different graph layers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

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

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

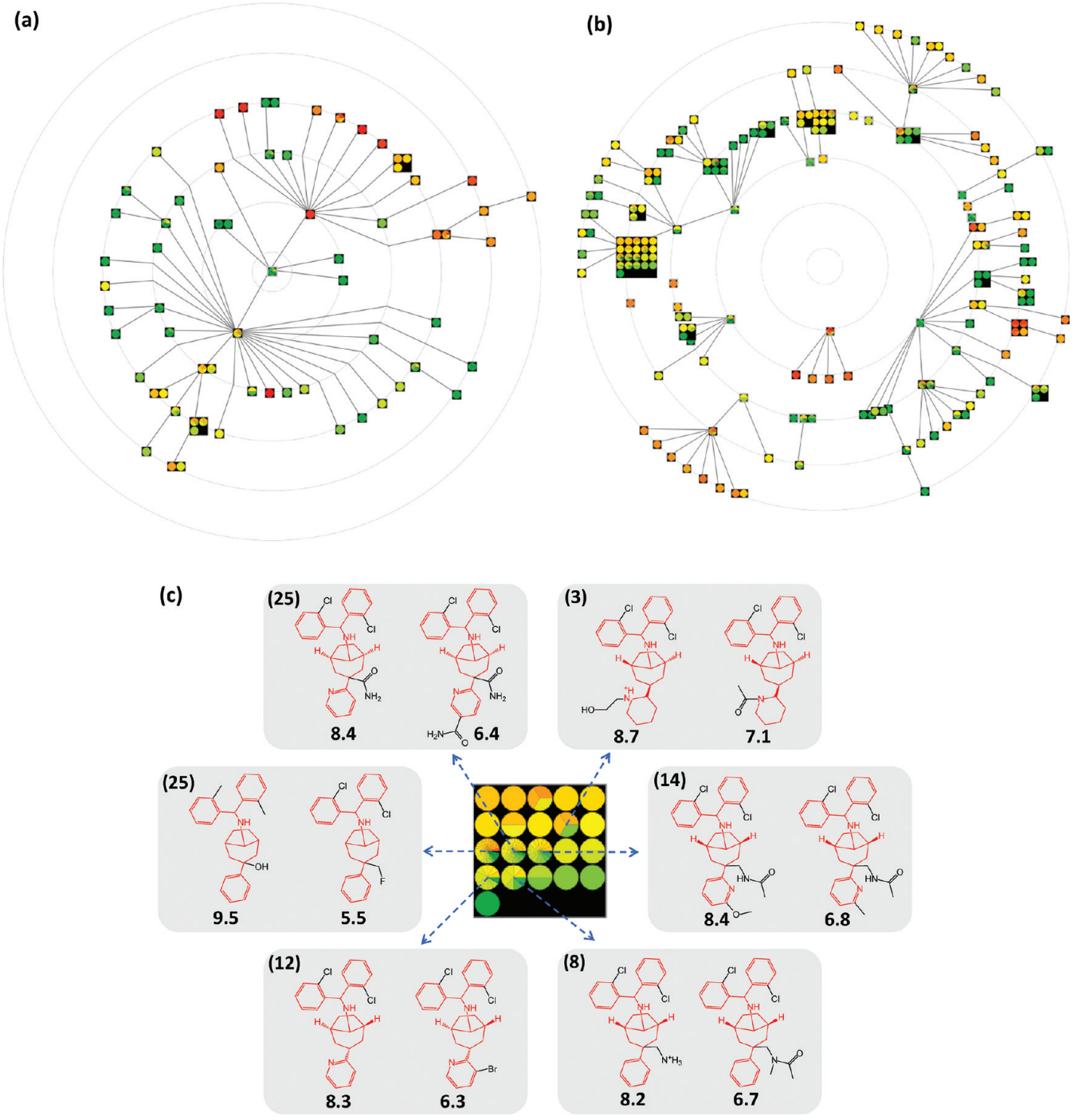


Figure 2. Graph representation. Prototypic LASSO graphs are shown. (a) Melatonin receptor 1A antagonists. (b) Nociceptin receptor antagonists. In (c), the rectangular subgraph on layer 5 on the left is enlarged and for each of six selected scaffolds (red) representing 3−25 compounds (reported in parentheses)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

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

METHODS AND MATERIALS

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

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

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

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

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

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

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

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

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

RESULTS AND DISCUSSION

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

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

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

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

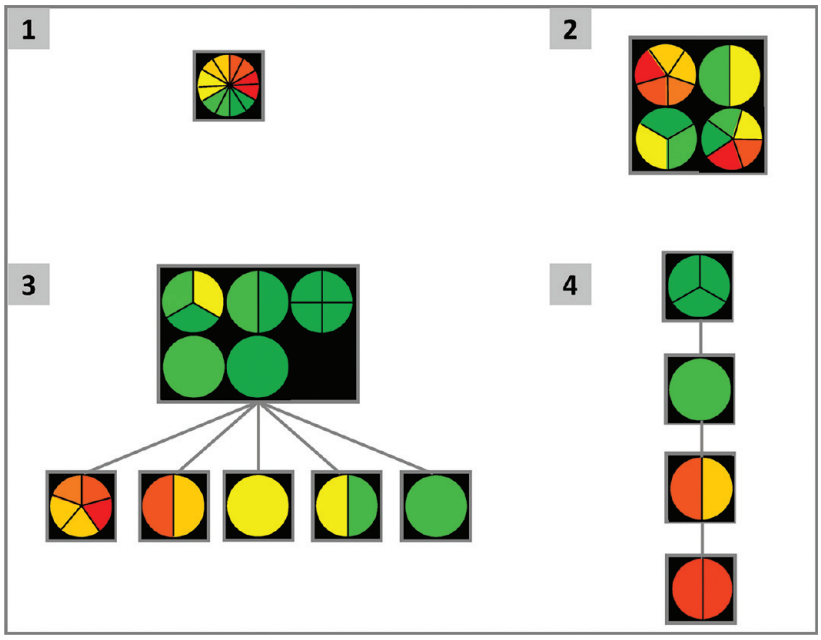


Figure 3. Graph patterns. Shown are four characteristic graph elements or patterns that convey SAR information, as described in the text.

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

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

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

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

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

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

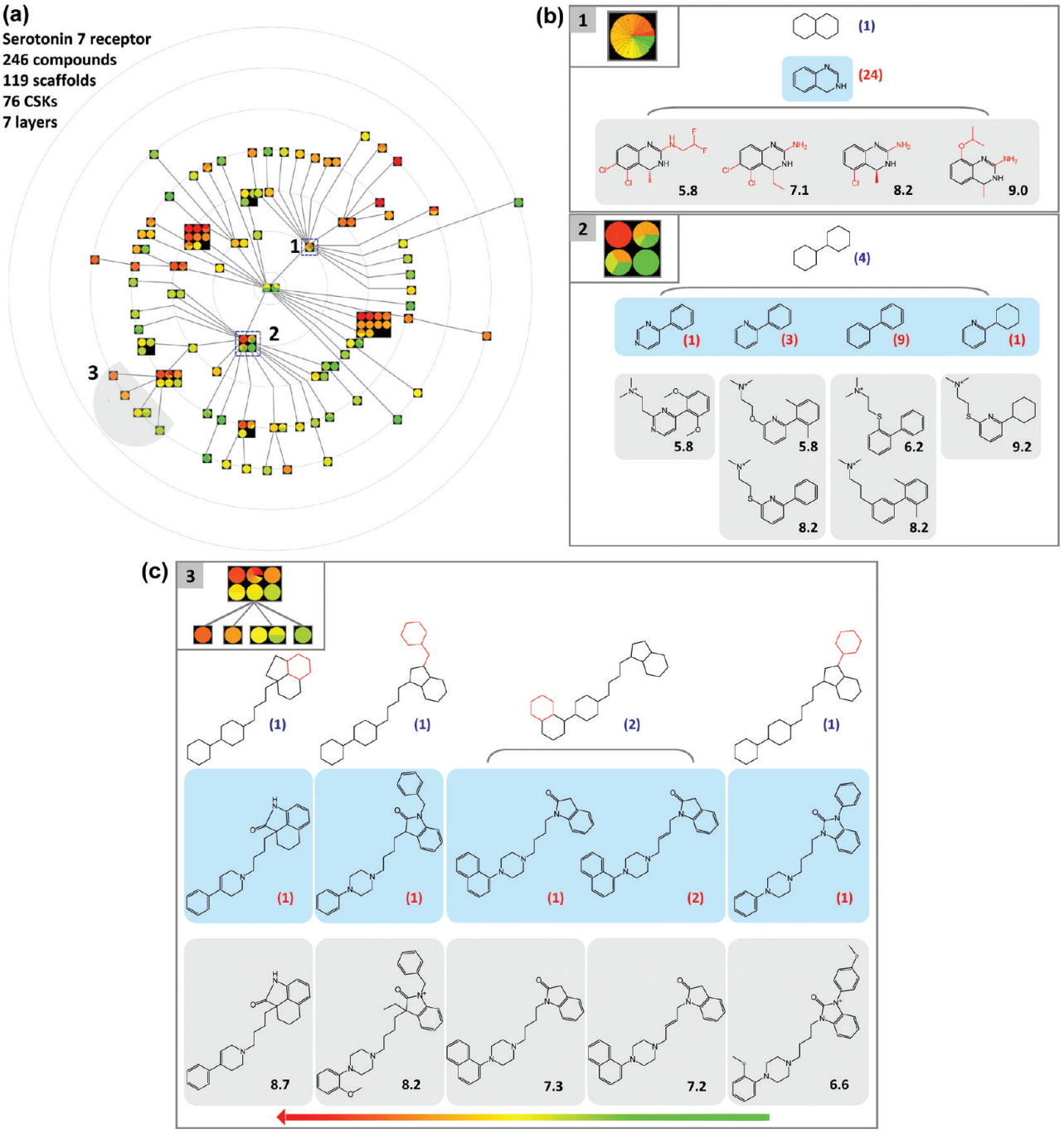


Figure 4. LASSO graph of serotonin 7 receptor antagonists. In (a), the LASSO graph of the compound data set is shown and characteristic patterns are marked and numbered according to Figure 3. In (b) and (c), these patterns and corresponding structures are depicted. For each CSK, the numbers of corresponding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

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

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

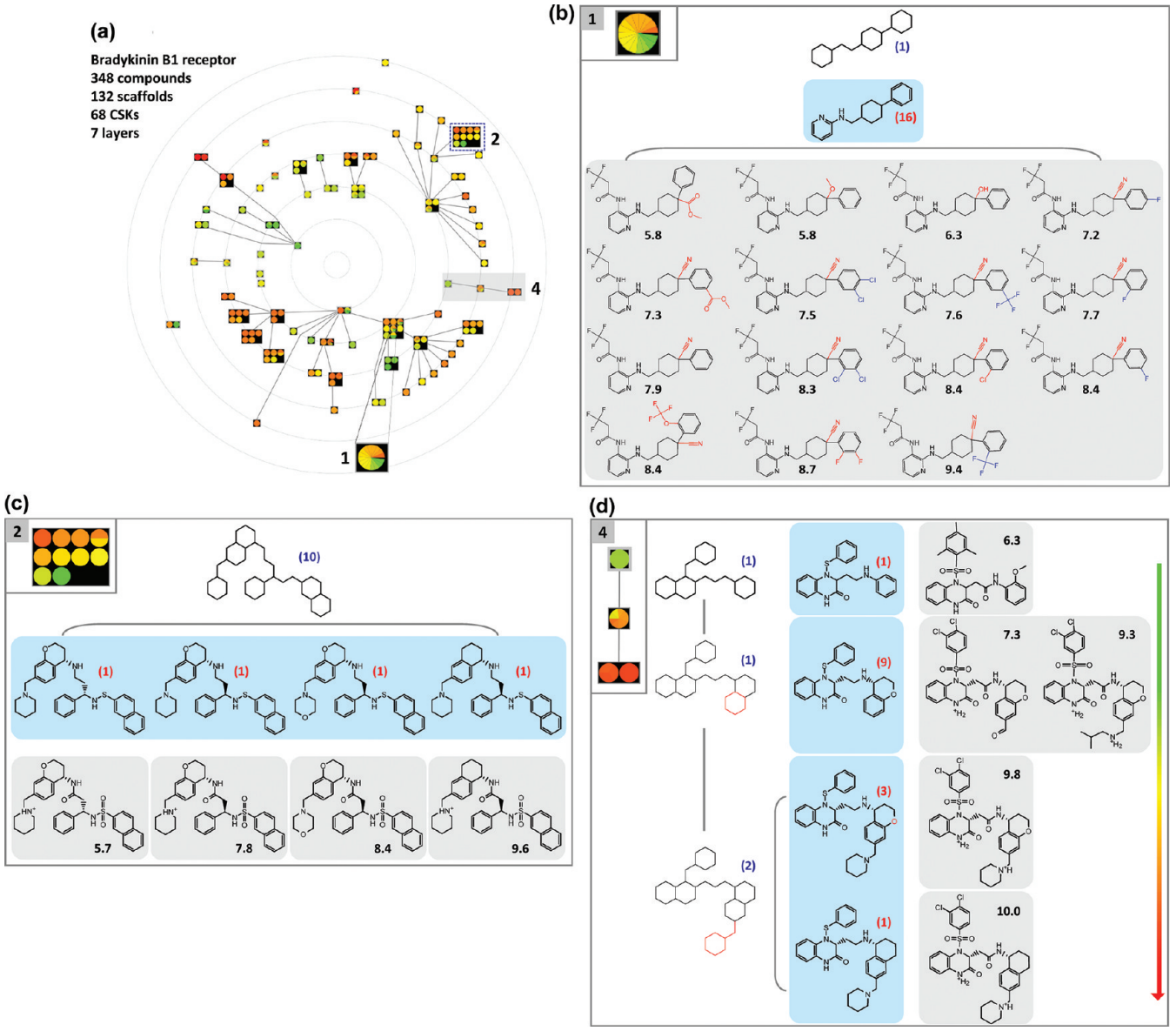


Figure 5. LASSO graph of bradykinin B1 receptor antagonists In (a), the LASSO graph of the compound data set is shown and characteristic patterns are marked and numbered according to Figure 3. In (b), (c), and (d), these patterns and corresponding structures are depicted. The presentation is according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

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

CONCLUSIONS

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

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

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

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

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

Notes

The authors declare no competing financial interest.

ABBREVIATIONS USED

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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