JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets

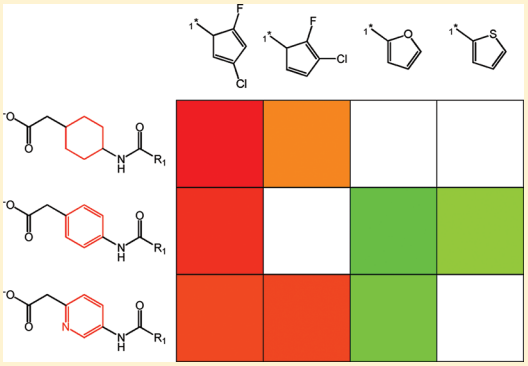
Anne Mai Wassermann,[†] Peter Haebel,[‡] Nils Weskamp,[‡] and Jürgen Bajorath*,[†]

[†]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

[‡]Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, D-88397 Biberach/Riss, Germany

Supporting Information

ABSTRACT: We introduce the *SAR matrix* data structure that is designed to elucidate SAR patterns produced by groups of structurally related active compounds, which are extracted from large data sets. SAR matrices are systematically generated and sorted on the basis of SAR information content. Matrix generation is computationally efficient and enables processing of large compound sets. The matrix format is reminiscent of SAR tables, and SAR patterns revealed by different categories of matrices are easily interpretable. The structural organization underlying matrix formation is more flexible than standard R-group decomposition schemes. Hence, the resulting matrices capture SAR information in a comprehensive manner.



INTRODUCTION

The evaluation of structure−activity relationships (SARs) of small molecules is a critically important task in high-throughput screening (HTS) data analysis and medicinal chemistry.[1,2] For example, in HTS data analysis one often selects the most potent hits and their structural neighbors from a primary screening campaign for further exploration. However, it is equally, if not more, important to identify series of compounds that contain interpretable SAR information as an indicator of sustainable and evolvable SARs. The presence of such information often renders compound series promising starting points for further chemical exploration and hit-to-lead projects. To these ends, standard clustering using whole-molecule similarity measures is often applied to group structurally similar compounds together[3] and subject individual clusters to statistical[4] or graphical analyses[5,6] of corresponding activity data. However, whole-molecule similarity measures might often not be sufficient to capture SAR information in chemical interpretable ways, due to the predominantly local molecular nature of SAR determinants. This potential caveat can be circumvented, for example, by employing substructure-based approaches that organize compound sets on the basis of shared molecular building blocks.[7] For example, a data structure termed 'scaffold tree'[8] has been introduced that derives molecular frameworks from compounds by pruning side chains and uses a set of predefined chemical rules to further decompose scaffolds until only single rings remain. The hierarchical organization of generated substructures and their annotation with activity information of the compounds from which they originate makes it then possible to identify frameworks that are associated with specific biological activities.[8,9] In addition, bioactive frameworks can also be identified on the basis of R-group decomposition schemes.[10] However, when adhering to conventional definitions of molecular frameworks and hierarchies,[7] subtle differences in heteroatom composition of ring systems or in the size of rings and linker fragments yield building blocks that are considered distinct but are in fact often very similar from a chemical point of view. Such frameworks might thus better be considered to belong to the same series, rather than as building blocks of different series, which complicates SAR exploration. Furthermore, additional analysis steps are generally required to extract detailed SAR information from a set of compounds that are represented by a framework associated with a specific activity. For the visual analysis of SARs in medicinal chemistry, methods focusing on maximum common substructures that define the core of congeneric compound series (analogs) are also widely applied.[10,11] In this case, analog series are often represented in R-group tables that contain the common core structure of a series and list chemical groups that are present at individual substitution sites.[10] However, for the analysis of large compound data sets, the determination of maximum common substructures for compound subsets is a computationally expensive task. In addition, R-group tables become difficult to navigate with increasing compound numbers, and the extraction of SAR information from them is in such cases far from being straightforward.

We have been interested in the development of a methodology to extract compound subsets from large data sets that are rich in SAR information, without adhering to predefined definitions of molecular cores or rules of chemical similarity. Building upon the recently introduced concept of 'matching molecular series'

Received: April 27, 2012

   dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

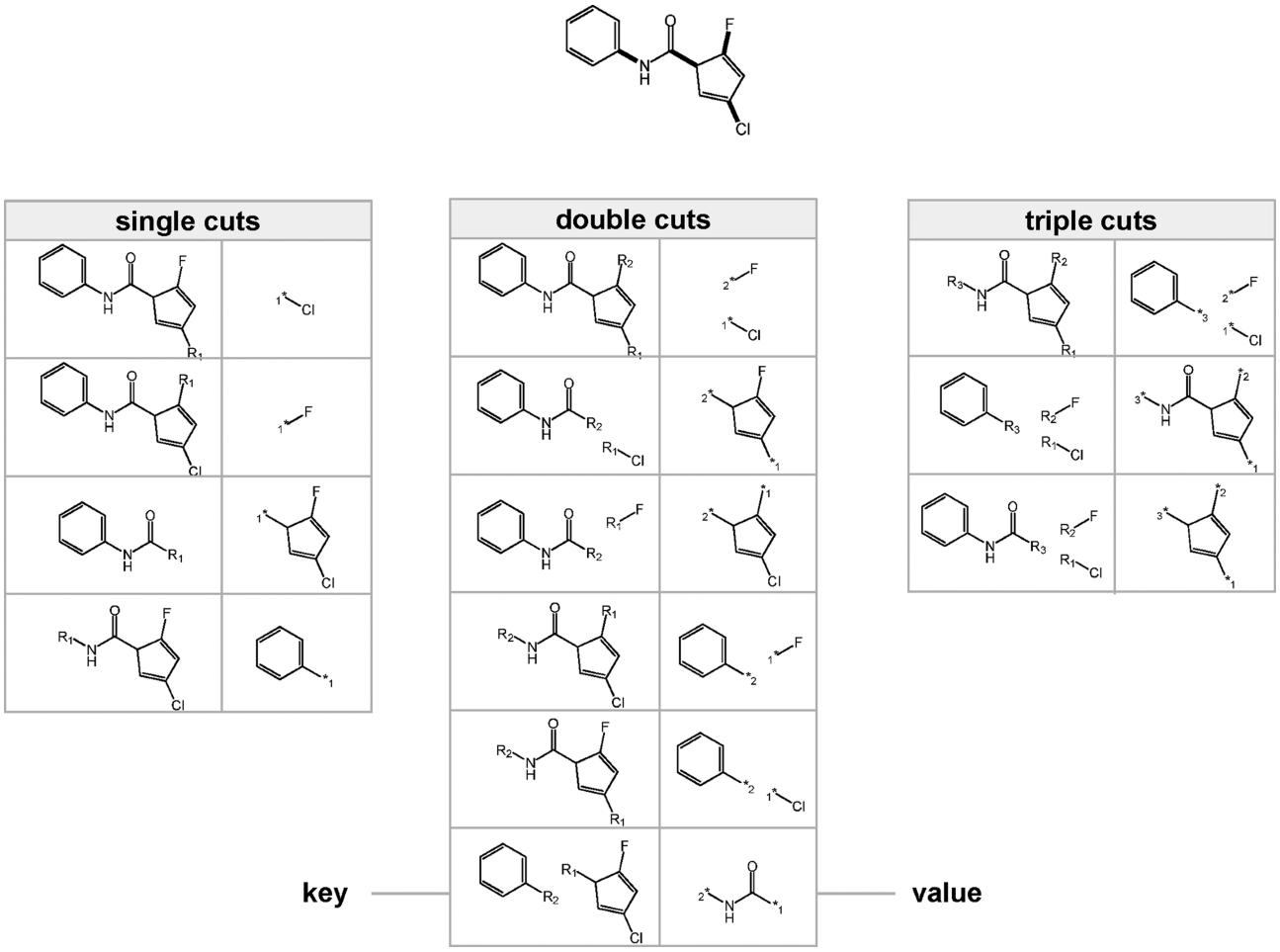


Figure 1. Molecule fragmentation. An exemplary compound is exhaustively fragmented through systematic deletion of all combinations of one to three exocyclic bonds (drawn in bold). If more than two fragments are generated, fragments with a single break point are combined. Fragments are then added to an index table where the fragment (or set of fragments) having the larger number of heavy atoms constitutes the key and the remaining substructure(s) the corresponding value.

(MMS),[12] we have designed and implemented a computationally efficient approach to identify groups of structurally related compounds and use emerging structural relationships to organize core structures and substituents in an SAR table-like format termed '*SAR matrix*'. This display format is chemically intuitive and readily interpretable. In addition, different scoring schemes were devised to prioritize SAR matrices that represent different types of information-rich local SAR environments and capture SAR information in different ways.

In the following, we first introduce the SAR matrix approach and data structure and then report exemplary applications to screening and chemical optimization data sets.

METHODS AND MATERIALS

Index Table. Initially, all molecules in a data set are fragmented by systematically deleting all exocyclic single bonds and their two- and three-bond combinations. Deletion of a single bond generates two fragments that are added to an index table ('single cut table'), with the larger fragment constituting the key and the smaller fragment the corresponding value. If the two fragments contain the same number of heavy atoms, they are added to the index table twice because each fragment is considered once as the key. The simultaneous deletion of two bonds results in the formation of one fragment with two 'break points' and two fragments having a single break point. These single-point fragments are grouped together. Again, the fragment or set of fragments that contains the larger number of heavy atoms is selected as the key for the index table ('double cut table'). Of the possible fragment combinations resulting from the simultaneous deletion of three bonds, only those containing a fragment with three break points are retained and processed in analogy to double cut fragments. This means that triple cuts that produce two fragments with two break points each and two fragments with one break point are ignored. In Figure 1, the systematic fragmentation of a molecule and the generation of index tables are illustrated. This fragmentation and indexing procedures represent a variant of the Hussain and Rea algorithm for the identification of matched molecular pairs (MMPs).[13] In our implementation, cuts are limited to exocyclic single bonds, and the definition of keys and values is more flexible because values are not limited to single fragments.

Matching Molecular Series. All molecules that are associated with the same key in an index table form an MMS,[12] i.e., a series of compounds that share a common substructure or core (key) and differ by defined chemical replacements (values). From standard index tables, molecules that contain a given key fragment as a substructure but carry a hydrogen atom at a break point cannot be identified. Therefore, in order to add these molecules to an MMS, one to three break points of keys consisting of a single fragment are systematically replaced by hydrogen atoms.

dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

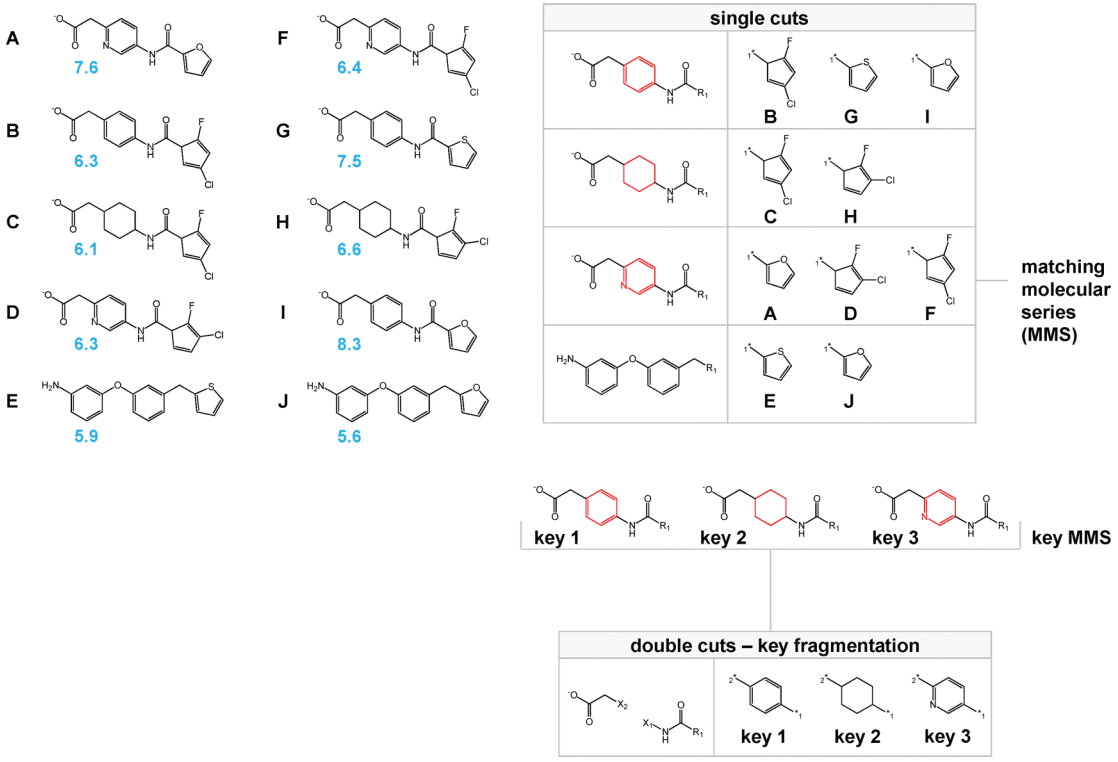


Figure 2. Identification of structurally similar MMS. A model data set consisting of 10 compounds A-J with hypothetical pK_i values is shown on the left. A section of the 'single cut table' of this compound set is displayed on the right. Compounds associated with the same key share a common substructure and form an MMS. Structurally related MMS are then identified by fragmenting keys and generating a separate index table (bottom right). In this index, keys that differ at a given single site (highlighted in red) form a 'key MMS'.

The resulting structure is then used to identify compounds in input data sets that also qualify for an MMS. If one or two break points remain, the modified structure is compared to keys in the single and double cut tables, respectively, and compounds belonging to a matching key are added to the MMS. These additional steps ensure that congeneric compound series involving substitutions of hydrogen atoms are correctly identified, following Hussain and Rea.[13]

Structurally Similar MMS. For the identification of 'structurally similar MMS', i.e., series of compounds having similar cores, keys are fragmented and indexed in separate tables essentially following the procedures applied to fragment original compounds, with two exceptions: (i) one of maximally three bonds that are cut is permitted to connect a break point and a heavy atom that is not a ring atom and (ii) values must now consist of a single fragment. In these newly generated separate index tables, keys from the single, double, and triple cut tables that only differ at a given site are grouped together, as illustrated for a model data set in Figure 2, and form a set of structurally similar MMS, also termed 'key MMS'.

Matrix Generation. For key MMS, all value fragments are pooled, and a matrix is generated in which rows correspond to keys and columns to values. Each combination of a key and a value defines a possible compound, and the corresponding cell in the matrix is colored if this molecule is contained in the data set, as illustrated in Figure 3. The color code reflects the compound potency distribution and follows a traffic light spectrum from red (low potency) over yellow to green (high potency).

Due to the systematic deletion of single bonds during fragmentation, it is principally possible that a compound occurs multiple times within the same matrix in different cells representing distinct key-value pairings. To remove compound redundancies

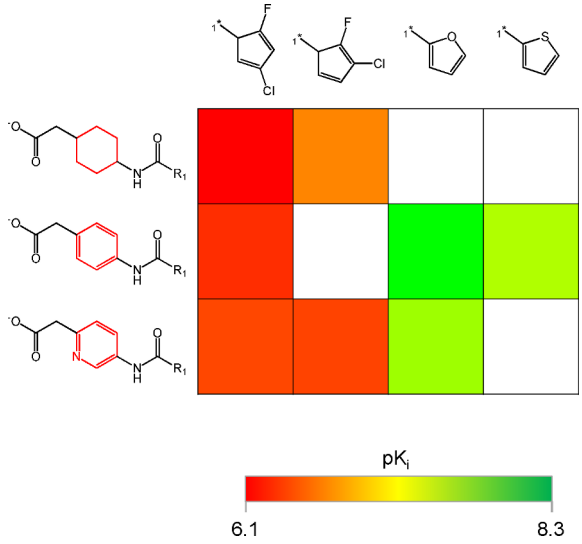


Figure 3. SAR matrix generation. For the key MMS identified in Figure 2, a matrix is generated in which the three keys correspond to rows and their associated value fragments to columns. The combination of a key and a value defines a possible compound and the corresponding cell is colored if this molecule is present in the data set. The color code reflects compound potency and ranges from red (lowest potency in the matrix) via yellow to green (highest potency).

from SAR matrices, the following rules are sequentially applied to prioritize fragment combinations: (i) If among multiple key-value pairings representing the same compound a key is a substructure of another key, then higher priority is assigned to the larger key paired with a smaller value fragment. (ii) If identical compounds are produced by the same value fragments for two

dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

matching series, one of the two series is randomly removed. (iii) If two value fragments generate identical compounds for the same matching series, then one of the two value fragments is randomly discarded. This rule is important for value fragments that are identical except for the numerical identifiers of their attachment points. If for a given value fragment only a subset of the molecules associated with another value fragment exists, then this value fragment is removed, because it is found in fewer molecules. The application of these rules does not guarantee that an individual molecule occurs only once in a generated matrix, as illustrated in **Figure S1** of the Supporting Information. In our implementation, a warning message is displayed if a compound occurs more than once in a matrix.

Optionally, hierarchical clustering can be performed on both horizontal and vertical axes to bring rows (keys) and columns (values) together that yield similar potency patterns, in analogy to gene expression heatmaps.[14]

Matrix Ranking. Due to their systematic generation, many different key MMS matrices are typically obtained for a compound data set. If the same subset of molecules in a data set yields multiple matrices with different keys and values, the matrix with the lowest cut level is selected. This is done because matrices in which molecules yield small numbers of fragments are generally easiest to analyze. Furthermore, four different scoring schemes are introduced to select key MMS among multiple matrices having the same cut level and rank matrices representing different compound subsets. Specifically, matrices are prioritized that contain the following:

(1) *Highly Potent Compounds*. For this purpose, the Kolmogorov−Smirnov (KS) statistic[15] is applied, similar to Varin et al.[16] Accordingly, potency distributions of compounds in a matrix and in the complete data set are determined, and a *p*-value is calculated for a one-sided KS test[15] to detect a shift toward more potent molecules in the matrix compared to the data set distribution (i.e., the smaller the *p*-value, the more significant the activity shift). Matrices are then ranked in the order of increasing *p*-values.

(2) *SAR Discontinuity*. Potency differences are calculated for all compounds in the same row or column of a matrix, and matrices are ranked in the order of decreasing average potency differences. This scoring scheme favors the presence of SAR discontinuity,[17] i.e., matrices with structurally similar compounds having large differences in potency.

(3) *Most Potent Compounds in a Single Column.* Thereby, preferred value fragments (substituents) are identified for similar keys. For each column containing at least three molecules, the potency of the least potent compound is assigned to the corresponding value fragment. Then, the highest assigned potency of all values is selected as a threshold. The percentage of compounds that are less potent than this threshold is calculated as a matrix score, and matrices are ranked in the order of decreasing scores.

(4) *SAR Transfer Series.* In the case of SAR transfer, the same structural modifications in two MMS lead to comparable potency changes between pairs of corresponding compounds.[18] For two MMS in a matrix sharing at least three value fragments, a transfer score[19] is calculated

$$\text{score}(\text{series}_{i,j}) = \frac{\max_{s=i,j}(\text{range}^s)}{\text{sd}(d) + 10^{-8}} \cdot \log(n)$$

with $\text{range}^s = \max_{k=1,\ldots,n}(\text{pot}_k^s) - \min_{k=1,\ldots,n}(\text{pot}_k^s)$ and $d_k = \text{pot}_k^i - \text{pot}_k^j$.

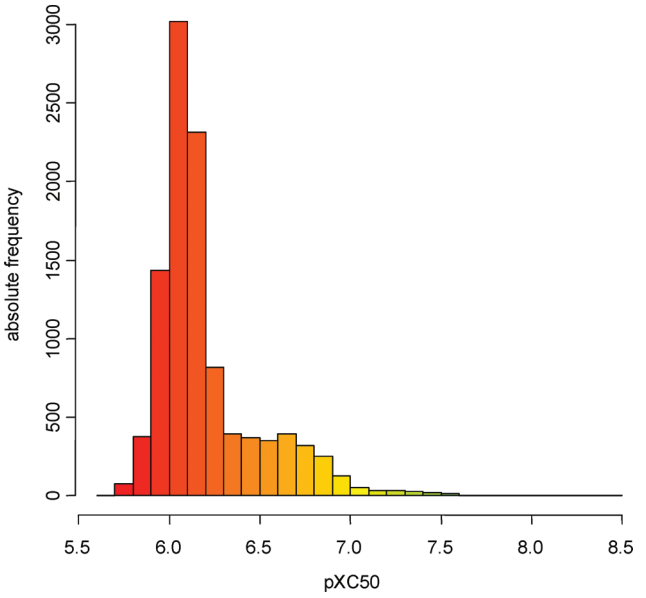


Figure 4. Antimalarial screening data. The potency (pXC50) distribution of 10,437 test compounds is reported as a color-coded histogram (from lowest (red) to highest (green) data set potency).

Here, n corresponds to the number of value fragments shared by the two MMS, pot_k^s gives the potency of the compound in series s that contains the k^{th} value fragment, range^s denotes the potency range spanned by the n compounds in series s, and $\text{sd}(d)$ is the standard deviation of potency differences between compounds containing the same value fragment. This scoring function yields high scores for two MMS having a large number of corresponding compounds with identical substituents (values), constant potency differences between matching compounds, and large potency ranges (a minimal potency range of 1 order of magnitude is set as a threshold). The highest SAR transfer score obtained for two MMS in a matrix is used as a score to rank generated SAR tables in the order of decreasing SAR transfer potential.

Implementation. All calculations required to build index tables, identify structurally related MMS, and generate, rank, and display SAR matrices were carried out using in-house written Java programs. SAR matrix representations shown here are automatically generated. Routines to generate MMPs and draw chemical structures were implemented using the OpenEye chemistry and depict tool kits.[20,21]

Data Sets. In 2010, GlaxoSmithKline released an antimalarial screening data set[22] containing a total of 13,533 compounds displaying at least 80% inhibitory activity in parasite growth assays at two μM concentration. A total of 10,437 of these compounds were found to represent previously unknown inhibitory chemotypes.[23] This compound subset has been subjected to fragmentation and SAR matrix generation in our current study. As activity annotation, the negative decadic logarithm of estimated compound concentrations yielding 50% inhibition of *Plasmodium falciparum* growth (pXC50) was used. In addition, we also assembled 1892 inhibitors of human carbonic anhydrase I from BindingDB[24] for which pK_i values were available. It has previously been shown that SAR transfer series frequently occur in carbonic anhydrase inhibitor sets.[19]

dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

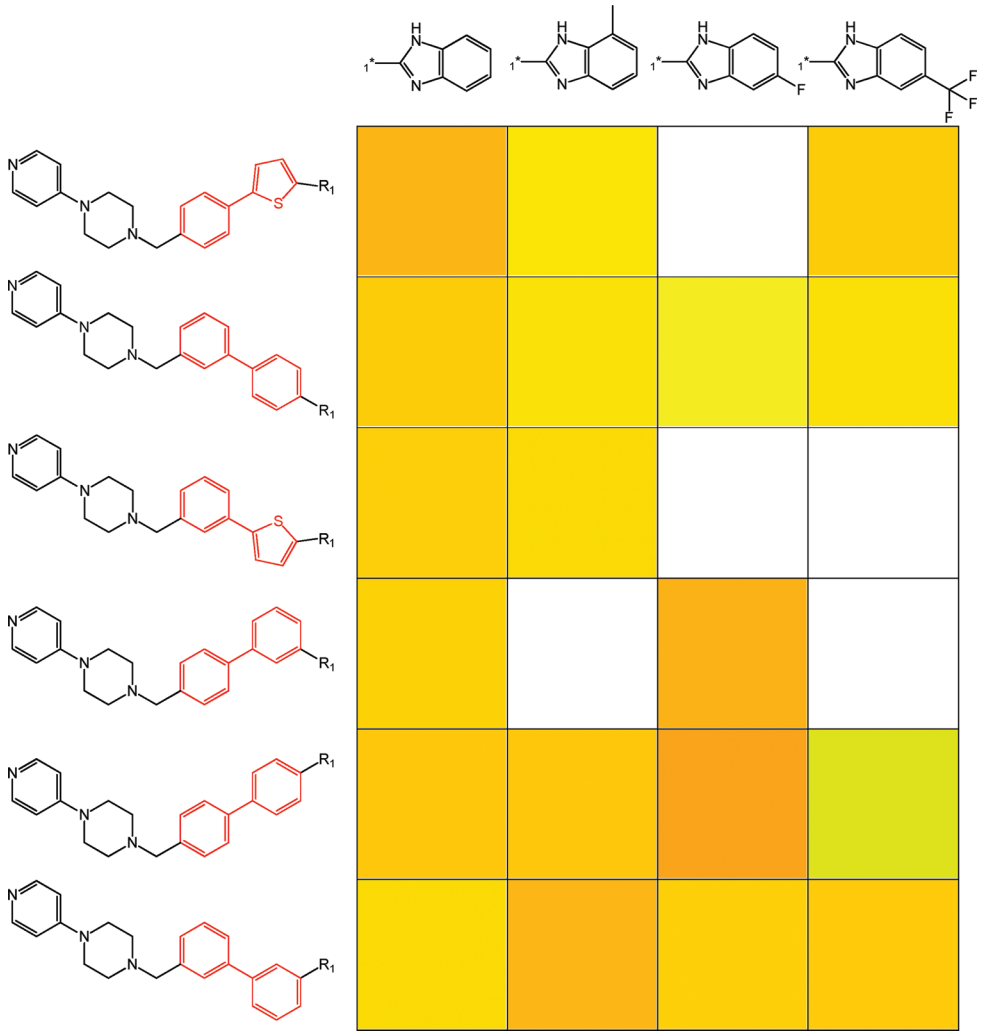


Figure 5. Matrix with potent compounds. The top-ranked antimalarial compound matrix according to KS statistic-based scoring is displayed. For the 19 molecules contained in the matrix, pXC50 values range from 6.6 to 7.2. The cell color code is according to Figure 4. Substructures that distinguish different keys are colored red.

RESULTS AND DISCUSSION

The two data sets that we analyzed were of rather different chemical composition and size. The antimalarial screening set was selected as a structurally heterogeneous compound set that contained many weakly potent hits. The potency distribution of the 10,437 screening set compounds is shown in Figure 4. Many hits displayed inhibitory activities in the micromolar range. The minimum, maximum, and average pXC50 values were 5.7, 8.5, and 6.2, respectively. In addition, given its phenotypic screening readout, potential targets of active compounds were not defined and activity annotations were only approximate in nature. From such phenotypic screening data, it is generally difficult to extract SAR information.[22,23] Hence, this large data set presented a challenging test case for SAR exploration. On the other hand, the carbonic anhydrase I inhibitor set contained many optimized congeneric series of potent inhibitors active in the low nanomolar range. In this case, equilibrium constants were available as potency measurements. Hence, this set represented a typical lead optimization set, with higher and more accessible SAR information. To evaluate the SAR matrix approach, we attempted to generate SAR matrices from these two very different data sets. In the following, representative examples are presented.

Analysis of Antimalarial Screening Data. We searched this data set for SAR matrices that contained a key MMS with at least five individual series and at least two compounds per series. Applying these criteria, we obtained 941 single cut, 1791 double cut, and 348 triple cut matrices from the GSK data.

Potent Screening Hits. The matrices were first ranked according to the KS-based scoring scheme, hence attempting to select matrices enriched with potent screening hits. The top-ranked single cut matrix is shown in Figure 5. All keys in this matrix contained a 1-benzyl-4-(pyridin-4-yl)piperazine moiety. The fourth ring in each key was either another phenyl or a thiophene ring, and the value fragments were differently substituted benzimidazoles. The matrix consisted of 19 different compounds with pXC50 values between 6.6 and 7.2. The average pXC50 value of the matrix compounds was 6.85. Hence, this subset was considerably more potent than an average screening hit (see above). Importantly, on the basis of a standard framework definition, hits contained in this matrix would not have been identified as a related series because all six MMS forming the matrix represented different Bemis and Murcko frameworks[25] and cyclic skeletons.[26] **Figure S2a** of the Supporting Information shows the top-ranked triple cut matrix for KS-based scoring. Low-scoring SAR matrices are abundant, do not contain interpretable SAR

dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

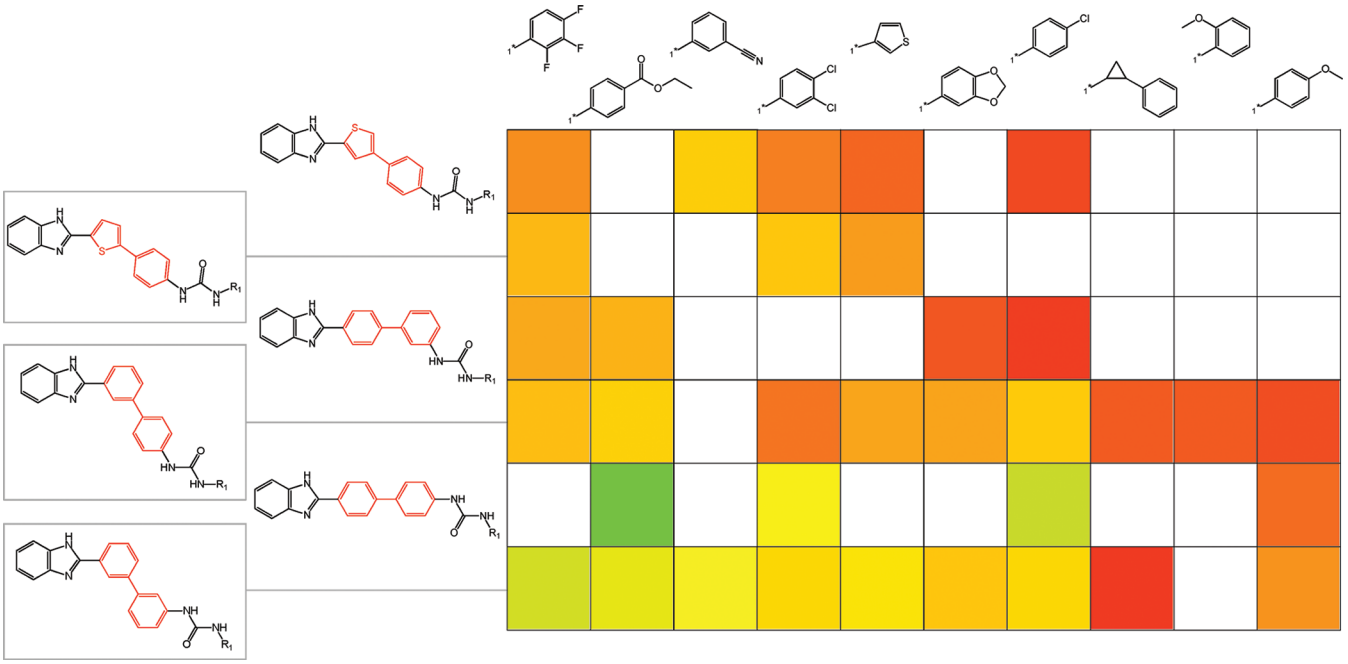


Figure 6. Matrix capturing SAR discontinuity. An antimalarial compound matrix containing 34 molecules with varying pXC50 values (6.0 to 7.7) is shown. On the basis of discontinuity-based scoring, this matrix is assigned rank 20 of all single cut matrices. The cell color code is according to Figure 4. Substructures that distinguish keys are colored red.

information, and do not need to be considered during SAR analysis.

SAR Discontinuity. We then ranked the screening set matrices on the basis of discontinuity scoring. A representative highly ranked single cut matrix is shown in Figure 6 (the top-ranked matrix is provided in **Figure S2b** of the Supporting Information). The 34 compounds in this SAR matrix spanned a pXC50 value range from 6.0 to 7.7. Compounds that contained different keys but shared the same substituent and also compounds with the same key structure but different substituents showed large potency differences within the matrix. For example, the attachment of a 4-chloro-phenyl ring to five different key structures yielded compounds with variable potency (pXC50 values from 6.1 to 7.3). For the fifth MMS series in this matrix that was defined by a 1-(4′-(1*H*-benzo[*d*]imidazol-2-yl)-biphenyl-4-yl)urea key structure, a rather subtle structural modification, i.e., the replacement of a ethyl-4-benzoate by a 4-methoxy-phenyl, led to a notable potency decrease (from a pXC50 value of 7.7 to 6.3). The presence of SAR discontinuity is often considered a valuable indicator for the potential to further evolve a chemical series.[5] Hence, the hit set represented by this matrix should be an interesting starting point for further compound optimization efforts. A number of different matrices with these characteristics were extracted from the screening data set.

Analysis of Carbonic Anhydrase I Inhibitors. For the set of carbonic anhydrase I inhibitors, we obtained 49 single, 97 double, and 56 triple cut matrices containing at least three similar MMS each represented by at least four compounds.

Preferred Substituents. We next scored all matrices for the presence of potent compounds within a single column, thus aiming at the identification of a preferred substituent (value) for similar keys. Figure 7 shows a highly ranked single cut matrix consisting of 24 inhibitors that were active within a pK_i range from 3.5 to 8.4. For six structurally related keys, a 5-sulfamoyl-1,3,4-thiadiazol-2-yl-aminosulfonyl-4-phenyl group was identified as a preferred substituent yielding inhibitors with pK_i values from 8.1 to 8.4 that were much more potent than other matrix compounds. These compounds included analogs with a 4-aminosulfonyl-phenyl substituent and different aliphatic linkers that were only weakly potent in the high micromolar range. Hence, this matrix immediately revealed a critical role of the terminal thiadiazol-sulfonamide group. By contrast, the modifications made to the central pyridyl ring of the key structures were only of minor importance because potency differences between the six MMS were only subtle. The larger top-ranked matrix is shown in **Figure S2c** of the Supporting Information.

SAR Transfer. Finally, we applied the SAR transfer scoring scheme to matrices of the carbonic anhydrase I inhibitor set. Clear SAR transfer events over significant potency ranges were not detected in the GSK screening data, as one might expect (by contrast, multiple instances of the other three categories of SAR matrices at different cut levels were detected in both data sets). Because SAR transfer scoring focused on pairs of MMS, we lowered our threshold for the minimum number of MMS per matrix to two, thereby avoiding the loss of potentially interesting SAR transfer events. This modification further increased the total number of generated matrices from 202 to 524. In Figure 8, the seventh-ranked triple cut matrix is shown that consisted of two MMS with 13 compounds each. This matrix displayed nearly ideal stepwise increases in potency for different substituents. The top-ranked matrix is shown in **Figure S1** of the Supporting Information. This matrix also meets formal SAR transfer criteria because identical substituents induce similar potency changes for different MMS. The two MMS in Figure 8 differed by the exchange of a methyl group and a phenyl ring at the substituted pyridine moiety of their keys. Interestingly, for these MMS, only pairs of analogs with exactly the same R-group pattern at three corresponding

dx.doi.org/10.1021/ci300206e | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

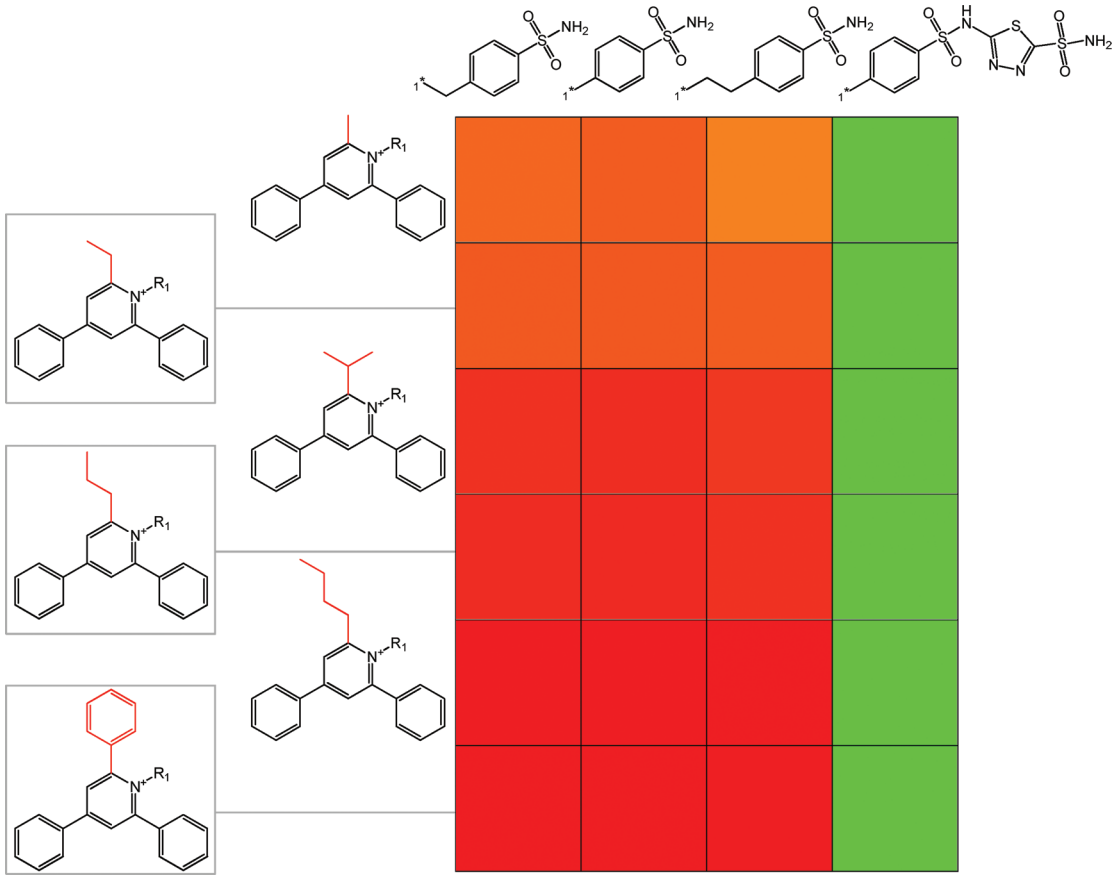


Figure 7. Preferred substituents. A matrix containing 24 human carbonic anhydrase I inhibitors covering a pK_i range from 3.5 to 8.4 is displayed. The matrix obtained ranks 9 among all single cut matrices on the basis of substituent-directed scoring. The cell color code is adjusted to reflect the potency range of the depicted carbonic anhydrase I inhibitor subset. Substructures that distinguish keys are colored red.

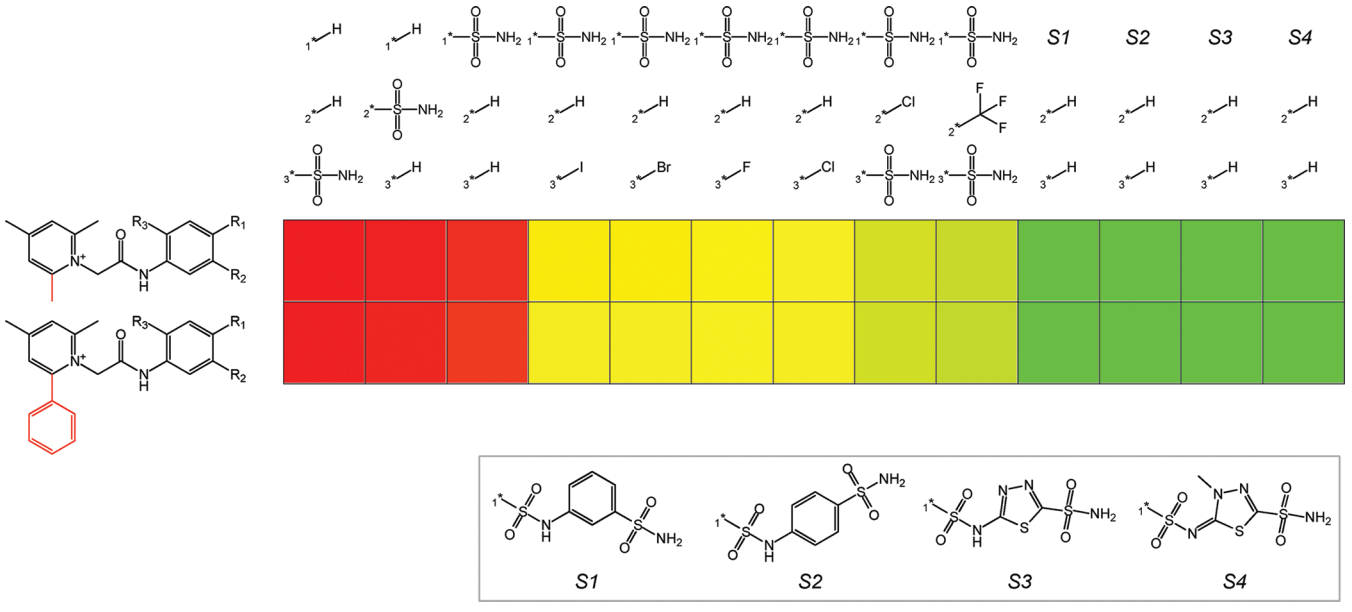


Figure 8. SAR transfer. A matrix with two MMS is shown. These MMS contain 13 carbonic anhydrase I inhibitors each and reveal nearly perfect SAR transfer within a pK_i range from 4.5 to 8.0. Value fragments are sorted in the order of increasing compound potency in the top row, and the cell color code is adjusted to the potency range covered by matrix compounds. Substructures that distinguish keys are colored red. This matrix ranks seventh among all triple cut matrices on the basis of SAR transfer scoring.

substitutions sites were retrieved from the data set, a rather unusual case. As can be seen, potency values for pairs of analogs with identical substituents were very similar. The same structural modifications led to comparable potency changes in the two series, representing a textbook-like SAR transfer event.

dx.doi.org/10.1021/ci300206e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

CONCLUDING REMARKS

Herein we have introduced the SAR matrix data structure for the elucidation of SAR information extracted from large compound data sets. The matrix format is intuitive and straightforward to interpret, and the structural decomposition scheme upon which the matrices are based is flexible and makes it possible to capture SAR information in different ways. We have shown that SAR matrices can be ranked according to alternative criteria, thus emphasizing the presence of different SAR patterns. Ease of interpretation is another characteristic of the SAR matrix structure. In order to evaluate the methodology, we have systematically generated SAR matrices for different compound sets including raw and approximate screening and detailed chemical optimization data. Representative application examples have been discussed, and it has been shown that compound subsets are obtained through matrix generation and ranking that contain potent compounds, display SAR discontinuity, identify preferred substituents for similar core structures, or represent SAR transfer events. On the basis of our findings, we conclude that the SAR matrix data structure is versatile and easily adaptable for different applications.

ASSOCIATED CONTENT

Supporting Information

***Supplementary* Figure S1** illustrates possible compound redundancies in SAR matrices. ***Supplementary* Figure S2** shows top-ranked matrices for the four different ranking schemes. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.M.W. is supported by Boehringer Ingelheim.

REFERENCES

(1) Kubinyi, H. Similarity and dissimilarity. A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 225−252.
(2) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11*, 694−699.
(3) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(4) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical practice in high-throughput data analysis. *Nat. Biotechnol.* **2006**, *24*, 167−175.
(5) Wawer, M; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discovery Today* **2010**, *15*, 631−639.
(6) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(7) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons learned from molecular scaffold analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742−1753.
(8) 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.
(9) Agrafiotis, D. K.; Wiener, J. J. Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *J. Med. Chem.* **2010**, *53*, 5002−5011.
(10) 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.
(11) Cho, S. J.; Sun, Y. Visual exploration of structure-activity relationship using maximum common framework. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 571−578.
(12) Wawer, M.; Bajorath, M. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.
(13) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(14) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *25*, 14863−14868.
(15) Birnbaum, Z. W.; Tingey, F. H. One-sided confidence contours for probability distribution functions. *Ann. Math. Stat.* **1951**, *22*, 592−596.
(16) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *J. Chem. Inf. Model.* **2011**, *51*, 1528−1538.
(17) Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.
(18) Wassermann, A. M.; Bajorath, J. A data mining method to facilitate SAR transfer. *J. Chem. Inf. Model.* **2011**, *51*, 1857−1866.
(19) Gupta-Ostermann, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Graph mining for SAR transfer series. *J. Chem. Inf. Model.* **2012**, doi: 10.1021/ci300071y
(20) *OEChem TK* v2012.Feb; OpenEye Scientific Software Inc.: Santa Fe, NM, 2012.
(21) *OEDepict TK* v2012.Feb; OpenEye Scientific Software Inc.: Santa Fe, NM, 2012.
(22) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanserwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. S.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305−312.
(23) Wawer, M.; Bajorath, J. Extracting SAR information from a large collection of anti-malarial screening hits by NSG-SPT analysis. *ACS Med. Chem. Lett.* **2011**, *2*, 201−206.
(24) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(25) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(26) Xu, Y.-J.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181−185.