

SAR Maps: A New SAR Visualization Technique for Medicinal Chemists

Dimitris K. Agrafiotis,* Maxim Shemanarev, Peter J. Connolly, Michael Farnum, and Victor S. Lobanov

Johnson & Johnson Pharmaceutical Research and Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341

Received July 11, 2007

We present structure–activity relationship (SAR) maps, a new, intuitive method for visualizing SARs targeted specifically at medicinal chemists. The method renders an R-group decomposition of a chemical series as a rectangular matrix of cells, each representing a unique combination of R-groups and thus a unique compound. Color-coding the cells by chemical property or biological activity allows patterns to be easily identified and exploited. SAR maps allow the medicinal chemist to interactively analyze complicated datasets with multiple R-group dimensions, rapidly correlate substituent structure and biological activity, assess additivity of substituent effects, identify missing analogs and screening data, and create compelling graphical representations for presentation and publication. We believe that this method fills a long-standing gap in the medicinal chemist's toolset for understanding and rationalizing SAR.

Introduction

Research programs involving the identification of structural leads or optimization of their activities against biomolecular targets inevitably require the medicinal chemist to rely on a traditional paradigm for analysis and presentation of structure–activity relationships (SARs^o). This common paradigm may be found by examining a random article in any recent issue of a medicinal chemistry publication, which will likely contain at least one SAR table consisting of a generic structure, accompanied by table columns associating chemical substituents with biological and chemical properties. Such SAR tables are the lingua franca of the field and are widely understood and accepted. Nevertheless, the generation of these tables is often a time-consuming process that involves manual dissection of chemical structures into their component parts and correlation of scaffolds and substituents with activity or potency at a biological target. Usually, SAR tables rely on static textual and numerical correlations at the expense of flexibility and clarity.

Interestingly, the chemoinformatics community has paid little attention to this intuitive way of presenting structure–activity information. Apart from the ubiquitous molecular spreadsheets found in chemical information and modeling packages, most SAR visualization techniques rely on various forms of clustering as a means of organizing the compounds into related subgroups.¹ With few exceptions, this partitioning is usually driven by similarity measures that look at the properties of the entire molecule and do not explicitly consider the presence of distinct scaffolds or substituent groups, which form the basis of most medicinal chemistry projects. A chemical scaffold is more than a common subgraph shared by a family of molecules; it often embodies a specific synthetic strategy that allows systematic exploration of SAR space using a divide-and-conquer approach. While clustering methods offer certain advantages, they are often misguided by idiosyncratic patterns in molecular graphs and produce groupings that look “unnatural” to a medicinal chemist.

This makes the key determinants of biological activity difficult to pinpoint and even more difficult to exploit in the design of improved analogs. Examples of such visualization techniques include self-organizing maps (SOMs), treemaps, dendrograms, radial clustergrams, nonlinear maps, heatmaps, and various forms of conventional statistical plots, such as scatter plots, bar charts, pie charts, and so on.

SOMs or Kohonen networks² map a set of objects onto a two-dimensional (2D) lattice in a way that preserves the topology of the underlying data. Similar objects (which are represented as points in a multidimensional vector space) map onto the same or proximal cells, whereas dissimilar objects map onto distant cells. In essence, SOMs partition the objects into a set of clusters whose relative position on the lattice reflects their degree of relatedness. Gasteiger has used SOMs in conjunction with 2D and three-dimensional (3D) autocorrelation descriptors to successfully separate dopamine from benzodiazepine receptor agonists embedded in a diverse set of commercially available compounds,³ to model the activity of steroids and polyhalogenated aromatics against the corticosteroid binding globulin and Ah receptor, respectively,⁴ and to visualize the diversity of combinatorial libraries.⁵

Dendrograms have historically been the method of choice for visualizing clusters derived by hierarchical algorithms. This layout reveals both the proximity of data items in the clusters as well as the number of levels in the cluster hierarchy. As in other tree layout methods, the difficulty in displaying a dendrogram increases exponentially with the number of nodes, a problem that has been partially alleviated in other domains through the use of hyperbolic geometry.⁶


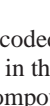
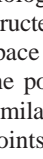

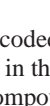
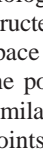

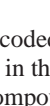
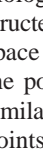

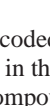
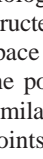

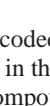
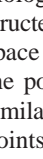
Treemaps⁷ visualize large cluster hierarchies in a space-filling manner. They recursively subdivide the screen space available using horizontal and vertical rectangles at alternating levels of the tree, each with a thickness proportional to the size of the node that it represents. Treemaps have been used to visualize hierarchical clusters of chemical libraries⁸ and SAR data sets,⁹ as well as other drug discovery data such as gene expression profiles¹⁰ and gene ontologies.¹¹

Radial clustergrams¹² represent an alternative space-filling technique that arranges the clusters into a series of layers, each representing a different level of the tree. Starting with the root of the tree at the center, child nodes project outward within the arc subtended by their parents, with a sweep angle that is proportional to the number of points they contain. Each segment

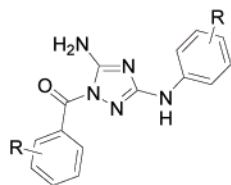
* To whom all correspondence should be addressed. Tel.: (610) 458-6045. Fax: (610) 458-8249. E-mail: dagrafio@prdu.s.jnj.com.

^o Abbreviations: SAR, structure–activity relationships; CDK1, cyclin-dependent kinase-1; VEGFR2, vascular endothelial growth factor receptor-2; KDR, kinase insert domain receptor; SOM, self-organizing map; ABCD, advanced biological and chemical discovery; 3DX, Third Dimension Explorer; MCS, maximum common substructure; AGG, Anti-Grain Geometry.

(a)

	Core_id	Core	R1_id	R1	R2_id	R2
1	C2H3N5		C7H3F2O		C6H6NO2S	
2	C2H3N5		C7H7O5		C6H6NO2S	
3	C2H3N5		C7H3F2O		C7H8NO2S	
4	C2H3N5		C8H5F2O		C7H8NO2S	
5	C2H3N5		C8H5F2O		C11H15N2	

(b) 3,5-diamino-1,2,4-1H-triazole CDK1 inhibitors



core structure for R-group analysis (Figures 2a,b,c)

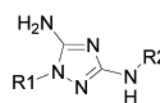


Figure 1. (a) Sample table of core, R₁, and R₂ substituents from R-group analysis of 1-acyl-3,5-diamino-1,2,4-1H-triazoles; and (b) generic triazole structure and core structure for SAR maps.

is color-coded by a user-defined aggregate property of the elements in that node, such as the average or maximum activity of the compounds in a particular biological assay. Compared to classical dendrograms and hyperbolic trees, radial clustergrams make much more efficient use of screen real estate; compared to treemaps, they are more effective in conveying hierarchical structure and displaying properties of nodes at all levels of the tree.

Nonlinear maps¹³ have been used to visualize individual molecules in a way that conveys both molecular similarity and biological activity in a single plot. Nonlinear maps are constructed by embedding a set of molecules into a low-dimensional space (typically 2D or 3D) in such a way that the distances of the points on the map match as closely as possible the (dis)similarities of the corresponding molecules. Color-coding of the points on the map allows one to display additional properties of the compounds, such as binding affinity, selectivity, and so on. While nonlinear mapping has historically been limited to relatively small data sets because of the (minimally) quadratic complexity of distance embedding algorithms, a newer, linearly scaling technique known as stochastic proximity embedding^{14–16} allows the approach to be applied to much larger collections.

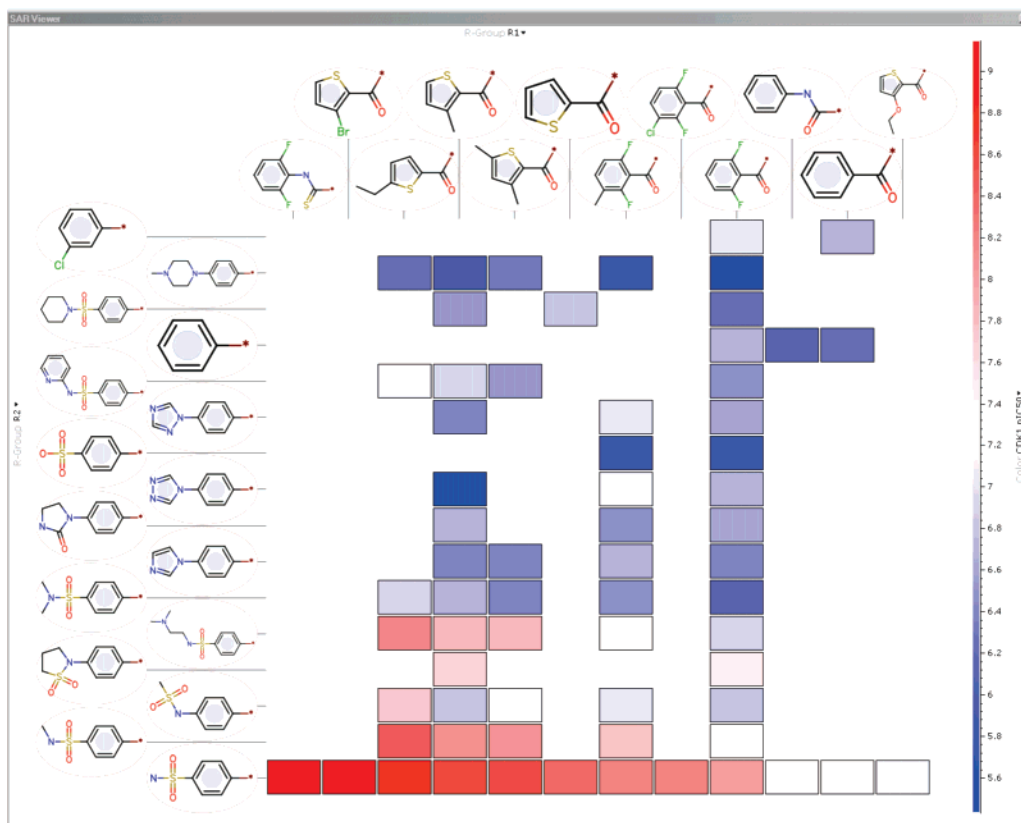
Heatmaps are typically applied to visualize data points with multiple attributes measured on the same scale. A heatmap is a rectangular array of cells, each representing a particular attribute of a particular object. Objects are typically arranged vertically along the y-axis, while attributes are arranged horizontally along the x-axis. Cells are colored according to the values of their corresponding attributes, obtained through mapping onto a gradient color scale. Heatmaps can be combined with dendro-

grams to group together closely related rows and columns and reveal patterns in the data. Heatmaps have been widely used in the analysis of microarray data¹⁷ and were recently adopted for visualizing SARs.¹⁸

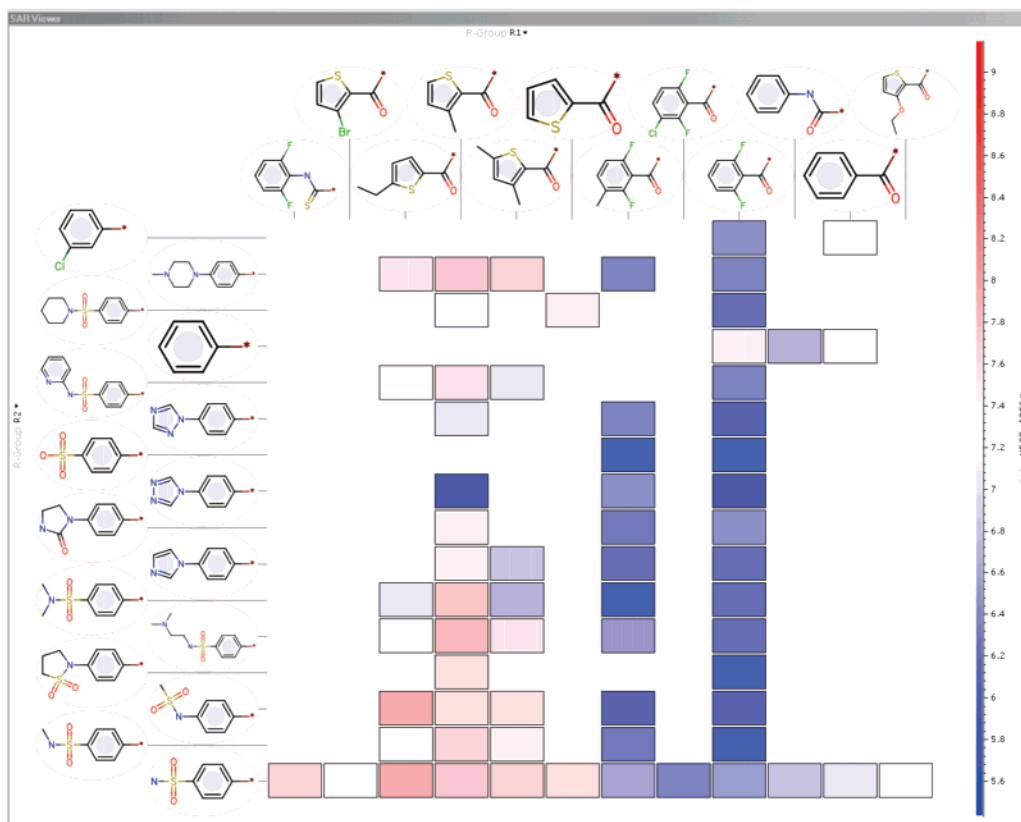
A particularly informative visualization of multifactorial structure–activity data is a special form of a pie chart called VlaiiVis.¹⁹ VlaiiVis is a radial plot representing the property profile of a single compound. Each slice of the “pie” represents a normalized response to a particular assay or property. The circumference of the pie represents the target values of each property, and the length of each slice indicates the deviation of that property from its target value. The method is ideal for visualizing not only how closely a compound meets a complex property profile, but also in determining the number of tests that have been performed on a particular molecule. A large number of compounds may be visualized simultaneously, either vertically on a scrollable spreadsheet or side-by-side on a rectangular grid.

More closely related to the present work is a substructure analysis and visualization technique known as SAR trees.²⁰ An SAR tree is a graph composed of core, subcore, or leaf nodes, which represent distinct chemical substructures, and attachment nodes, which indicate how these substructures are connected to one another. The core node represents a substructure that is common to all of the compounds in the collection and is connected in a radial fashion to a set of attachment nodes, which represent different variation sites (R-groups) around it. Each attachment node is, in turn, connected to a set of subcore or leaf nodes. A subcore node represents a chemical substructure that is shared by multiple compounds at that particular attachment site and contains further variations around it. Leaf nodes

(a)



(b)



represent terminal substructures that are distinct from any other chemical fragments at a given variation site. Thus, individual

molecules are represented by a set of structural variation paths emanating from the core and terminating in a leaf node, and

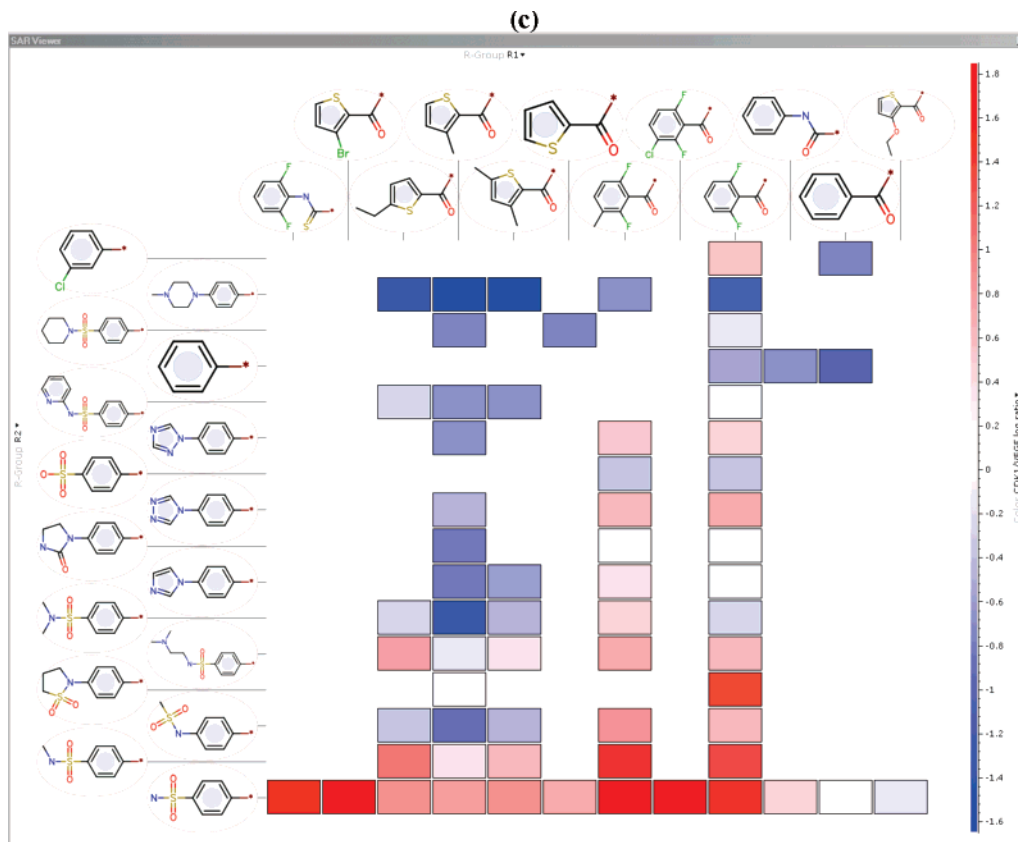


Figure 2. (a) CDK1 inhibition (pIC₅₀) of triazoles, sorted by R₂ AlogP, then CDK1 potency; (b) VEGFR2 inhibition (pIC₅₀) of triazoles, sorted by R₂ AlogP, then CDK1 potency; and (c) Log of (VEGFR2 IC₅₀)/(CDK1 IC₅₀), sorted by R₂ AlogP, then CDK1 potency.

the union of these paths represents the entire library. This type of recursive definition of substructural patterns also forms the basis of ClassPharmer,²¹ a technique that utilizes maximum common substructures (MCSs) and a phylogenetic-like tree algorithm to partition the compounds into a set of clusters, which can then be correlated with biological activity.

Scaffold trees are an alternative way of visualizing substructure hierarchies in large, heterogeneous data sets.^{22,23} Each node in the tree represents a unique chemotype at some level of abstraction (in ref 22, for example, these can be complete molecules, cyclic systems, cyclic system skeletons, and reduced cyclic system skeletons). The hierarchies are obtained through iterative removal of side chains and rings from the parent molecule, followed by canonicalization of the resulting structures. By mapping compounds onto the tree and examining the relative occupancy of actives and inactives at each node, one can assess the degree of enrichment at several levels of structural resolution.

R-group analysis is at the root of many scaffold-based methods and is supported by several software packages such as Diva,²⁴ Accord for Excel,²⁵ and STN Express.²⁶ Diva is probably the first application to offer the capability to find and label R-group substituents around a specified core. These substituents are displayed in a molecular spreadsheet, with additional columns providing associated activity values.

Here, we present SAR maps, a new visualization technique that combines the power of R-group analysis with the visual richness of heatmaps. SAR maps allow the medicinal chemist to interactively analyze complicated datasets with multiple R-group dimensions, rapidly correlate substituent structure and biological activity, assess additivity of substituent effects, identify missing analogs and screening data, and create

compelling graphical representations for presentation and publication. SAR maps have been fully integrated into Third Dimension Explorer (3DX), a state-of-the-art data analysis tool developed as part of J&JPRD's discovery informatics platform known as ABCD.²⁷ By bringing together pictorial representations of chemistry and graphical visualization of biological and property data, the SAR map establishes a new paradigm for SAR analysis that can greatly facilitate the drug discovery process.

Methods

3DX and ABCD. SAR maps were implemented as a component of 3DX, a .Net application designed to address a broad range of data analysis and visualization needs in drug discovery. 3DX is part of a broader initiative known as ABCD,²⁷ which aims to connect disparate pieces of chemical and pharmacological data into a unifying whole and provide discovery scientists with tools that allow them to make informed, data-driven decisions.

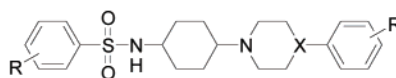
3DX is a table-oriented application, similar in concept to the ubiquitous Microsoft Excel. A 3DX document contains a collection of tables, each of which contains a collection of columns and rows. Each column contains data of the same type, such as strings, integers, floating point numbers, "fuzzy" or qualified numbers (floating point numbers with range or uncertainty qualifiers), number lists, dates, time intervals, chemical structures and substructures, images, graphs, and many others. Much of 3DX's analytical power comes from its ability to handle very large data sets through its embedded database technology, to associate custom cell renderers with each data type in the spreadsheet, and to visualize the entire data set using a variety of custom viewers, such as 2D and 3D scatter plots, histograms, heatmaps, correlation maps, and the SAR maps described herein. The program offers a full gamut of navigation and selection options,

(a)

	Core_id	Core	R1_id	R1	R2_id	R2
1	trans-C10H18N2		C6H4Cl2NO2S _2		C9H11O	
2	cis-C11H19N		C6H4ClFNO2S _2		C9H9O	
3	trans-C11H19N		C6H4ClFNO2S _2		C9H9O	
4	cis-C10H18N2		C6H4ClFNO2S _2		C9H11O	
5	trans-C11H19N		C6H4ClFNO2S _2		C8H6F3O	
6	trans-C10H18N2		C6H4ClFNO2S _2		C9H10FO	
7	cis-C10H18N2		C6H4ClFNO2S _2		C9H10FO	

(b)

N-[4-(piperazin-1-yl)cyclohexyl]- and *N*-[4-(piperidin-4-yl)cyclohexyl]-sulfonamides (X = N, CH)



core structures for R-group analysis (Figures 4a-d, 5a-d)

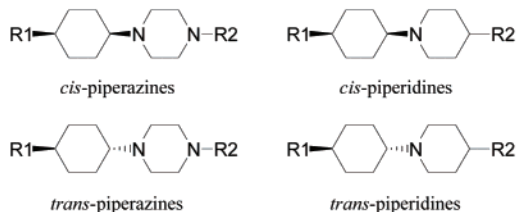


Figure 3. (a) Sample table of core, R₁, and R₂ substituents from R-group analysis of *N*-[4-(piperazin-1-yl)cyclohexyl]- and *N*-[4-(piperidin-4-yl)cyclohexyl]-sulfonamides; and (b) generic piperazine/piperidine structure and core structures for SAR maps.

augmented through linked visualizations and interactive filtering and querying.

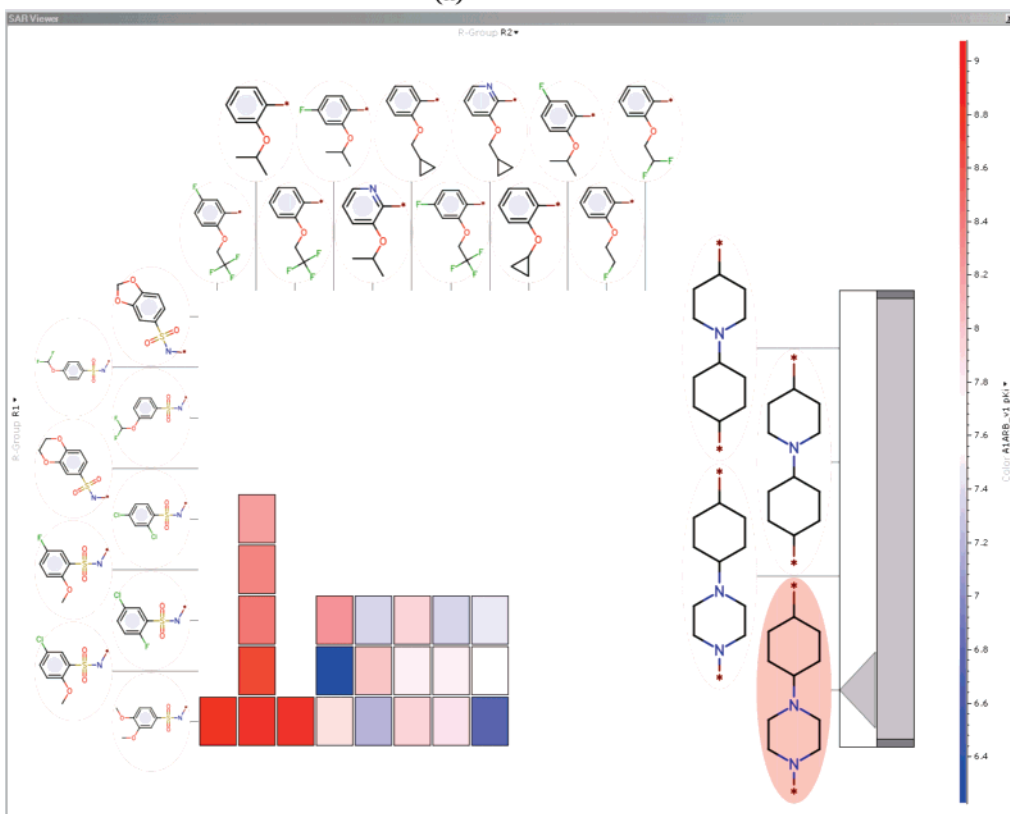
3DX uses a plug-in architecture that allows new functionality to be developed independently of the main application and delivered to the user either automatically or on a per-need basis. Plug-ins can be UI or non-UI driven and have full programmatic access to the 3DX core and the data, allowing them to create and remove tables, insert and remove columns, edit data, create and (re)arrange viewers, and so on. Their functionality and implementation can be extremely diverse, bringing a wealth of data retrieval, processing, analysis, visualization, and reporting capabilities to the end users, without requiring them to leave the application. An array of powerful, chemically aware data mining tools were introduced in this fashion, including exact structure, substructure, and similarity searching, structure alignment, MCS detection, chemotype classification, R-group analysis (vide infra), physicochemical property calculation, combinatorial library generation, diversity analysis, and many others. The plug-in architecture is also used to provide seamless integration with the ABCD warehouse through the ABCD wizard, a graphical query builder that allows users to mine the

ABCD database without requiring knowledge of SQL or its relational schema and to retrieve the results in a variety of tabular formats.

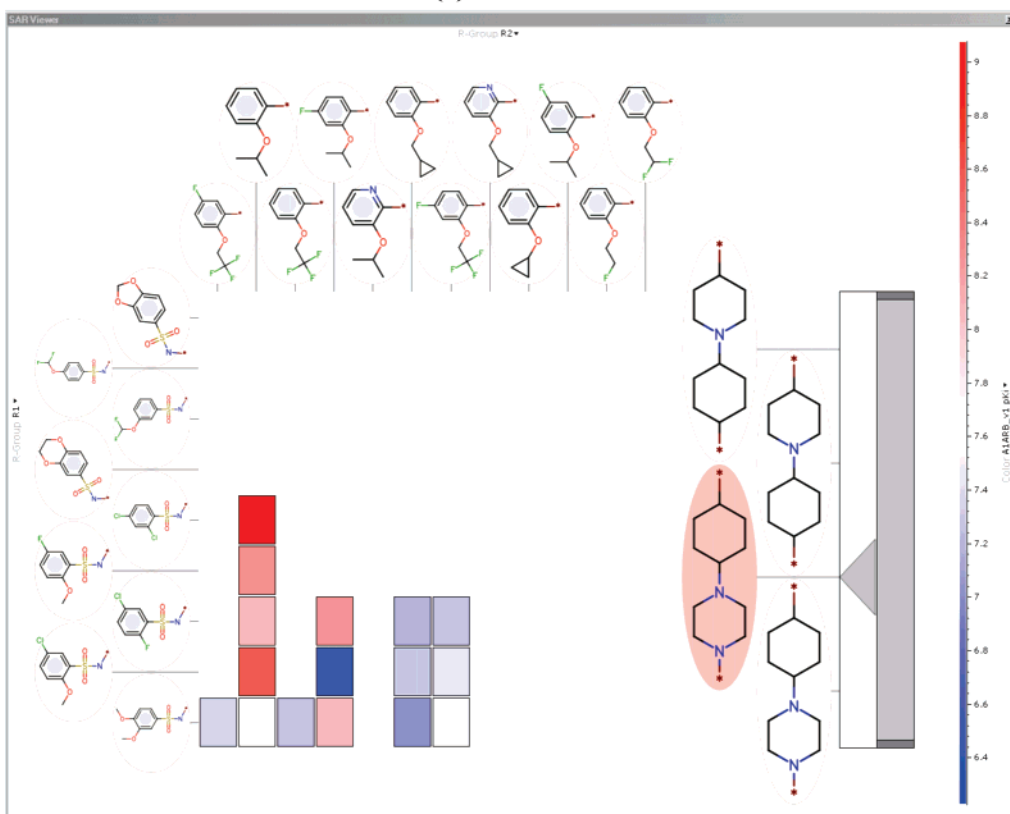
R-Group Analysis. R-group analysis takes as input a list of chemical structures bearing a common substructure (typically a ring scaffold) and decomposes them into a list of substituents or R-groups around that substructure, along with their attachment points to it. The common core is specified either through manual sketching or through an automated search for the MCS. Both the MCS search and the R-group analysis algorithms are implemented as a 3DX plug-in.

A comprehensive search for an MCS among a list of chemical graphs is a complex procedure whose execution time grows exponentially with the number of compounds considered. In addition, identifying the MCS between a pair of molecules is by itself a nontrivial task that requires an exponentially longer time for larger molecules. Hence, manual sketching of the common core is often the fastest option, especially if the chemist is analyzing a known series. When the common core is unknown, an MCS search provides a much better alternative to visual inspection. A naïve

(a)



(b)



MCS search algorithm would involve traversing the list of molecules and iteratively computing the MCS of the i th molecule

to the MCS of the previous $i-1$ molecules (i.e., computing the recursive relation $mcs(i) = MCS(mcs(i-1), mol(i))$, where $mcs(1)$

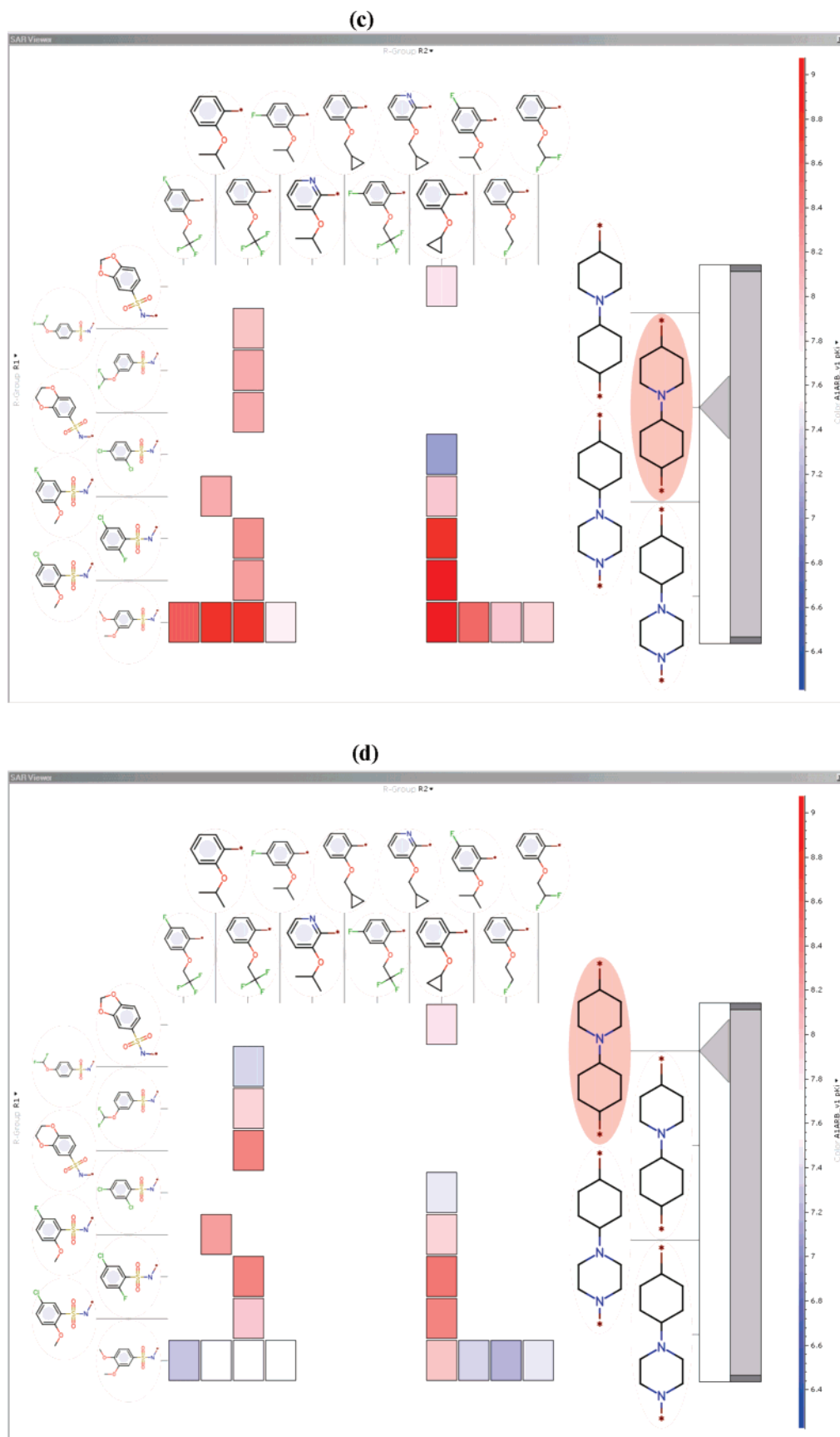


Figure 4. α_{1a} binding affinities (pK_i) of (a) *cis*-piperazines; (b) *trans*-piperazines; (c) *cis*-piperidines; and (d) of *trans*-piperidines.

$\equiv \text{mol}(1)$). Because the MCS is order-dependent, in theory all $N!$ permutations of the N input molecules need to be considered (each involving $N-1$ pairwise MCS comparisons). To achieve acceptable

performance and ensure a positive user experience, we use a fast, approximate algorithm that identifies the correct MCS in most cases but does not necessarily guarantee it (i.e., sometimes the resulting

common substructure may not be the maximum). The algorithm starts by sorting the molecules in descending order of bond count, designating the largest molecule as the initial MCS, and iteratively reducing the MCS by comparing it with the next molecule in the list. Our experience suggests that this $O(N)$ algorithm performs well in most cases and is extremely fast.

Once the common core is identified, it is converted to a substructure pattern that is mapped onto all the molecules in the list. Generally, a pattern can be mapped onto a molecule in a number of ways, for example, a benzene ring can be mapped onto aniline in six different ways, each placing the attachment point for the amino group on a different atom in the benzene ring. Because the goal of R-group analysis is to identify structural variations at the same position on the common core, we need to minimize the number of attachment points so that, using the example above, both aniline and toluene would be considered as having an amino or a methyl R-group, respectively, at the same position on the benzene core. Hence, after all molecules have been mapped, our R-group analysis algorithm identifies a mapping that minimizes the number of attachment points on the core scaffold. In a second pass, the preferred mapping of the common core is applied to each of the molecules to identify the fragments connected to each attachment point. The resulting R-groups are extracted into separate columns labeled R_1 , R_2 , and so on, and their attachment points to the core are replaced with dummy atoms labeled X. Similarly, the attachment points on the core are replaced with dummy atoms labeled R_1 , R_2 , and so on.

SAR Map. An SAR map renders an R-group decomposition as a rectangular grid of colored cells. Each cell represents a single compound C_i , defined as the combination of its constituent R-groups $\{R_1(i), R_2(i), \dots, R_n(i)\}$, where $R_j(i)$ is the substituent at the j th variation site in compound i . The map has the appearance of a heatmap, with the exception that the usual horizontal and vertical text labels are replaced by the chemical structures of the substituents at the two variation sites displayed on the X and Y axes.

When the scaffold contains only two variation sites ($n = 2$), all compounds in the data set are visible on the map, with R_1 and R_2 placed along the X and Y axis, respectively (or vice versa). When $n > 2$, the remaining dimensions are displayed on the side using a set of chemical sliders that allow the user to view all the substituents available at each variation site but limit the selection to a single member of each list. In this case, the SAR map displays the submatrix of compounds formed by the Cartesian product $\{R_{1,j=1,\dots,[R_1]}\} \times \{R_{2,j=1,\dots,[R_2]}\} \times R_{3,j=3} \times \dots \times R_{m,j=n}$, where all but two dimensions are fixed to a single R-group (i.e., the maps display a hyperplane in the n -dimensional combinatorial substituent space).

Two dropdown boxes allow the user to select which variation sites to display along the X and Y axes; the remaining R sites are displayed in sliders arranged by their R numbers. The graphical interface is completed with a color-scale and additional dropdown box that allows the user to interactively select which property and scale to use for color-coding the cells. The color scale handles both numerical and categorical variables (using smooth gradients for numerical variables and discrete colors for categorical ones). Because it is extremely rare that an SAR data set will contain all possible combinations of all the substituents, cells associated with missing compounds are not drawn at all, whereas cells associated with compounds that are present but whose property values are null (e.g., those whose biological activity has not been measured) are colored in gray. This provides a very effective way of assessing the coverage of the combinatorial space and the degree of completeness of the biological characterization.

Given the potentially large number of substituents that need to be displayed, we based our implementation on Anti-Grain Geometry (AGG),²⁸ a high-quality, lightweight, extensible, and platform-independent rendering engine written in standard ANSI C++. AGG provides very fast anti-aliased graphics with sub-pixel accuracy, allowing effective visualization of a large number of chemical structures with minimal loss of resolution and clarity. The SAR viewer itself is implemented as a .Net control and is built upon a

.Net version of the AGG library written in C++/CLI (available with .Net 2.0).

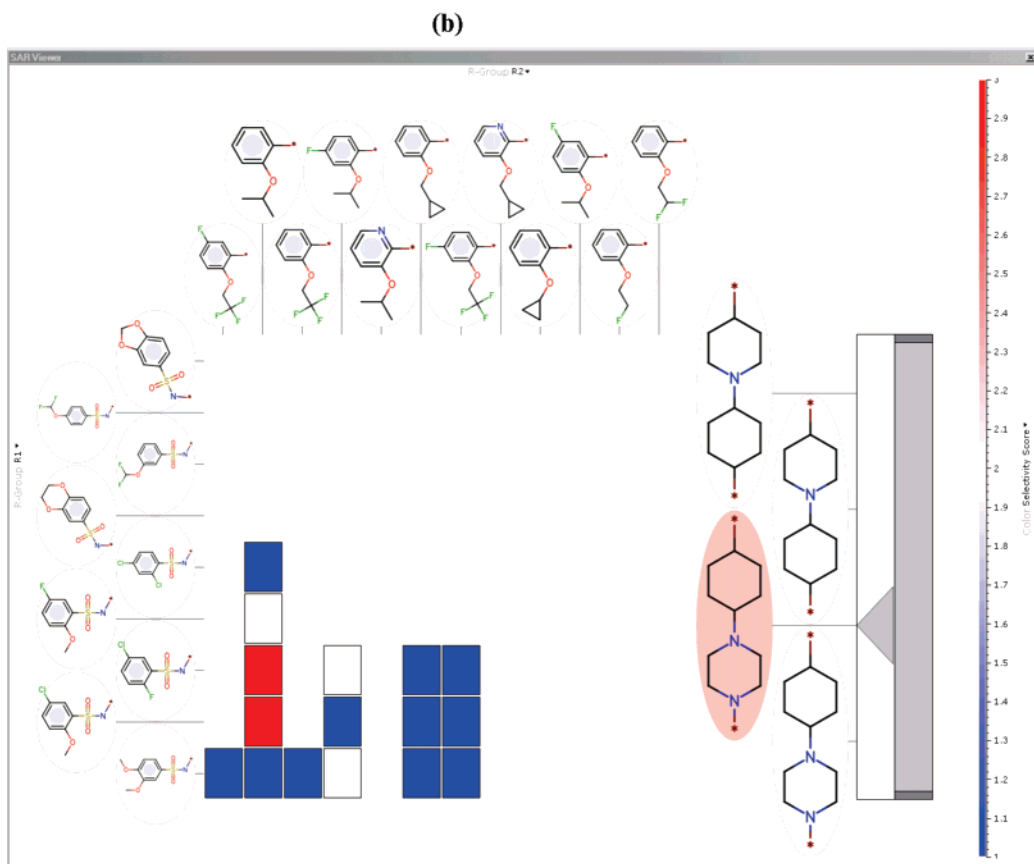
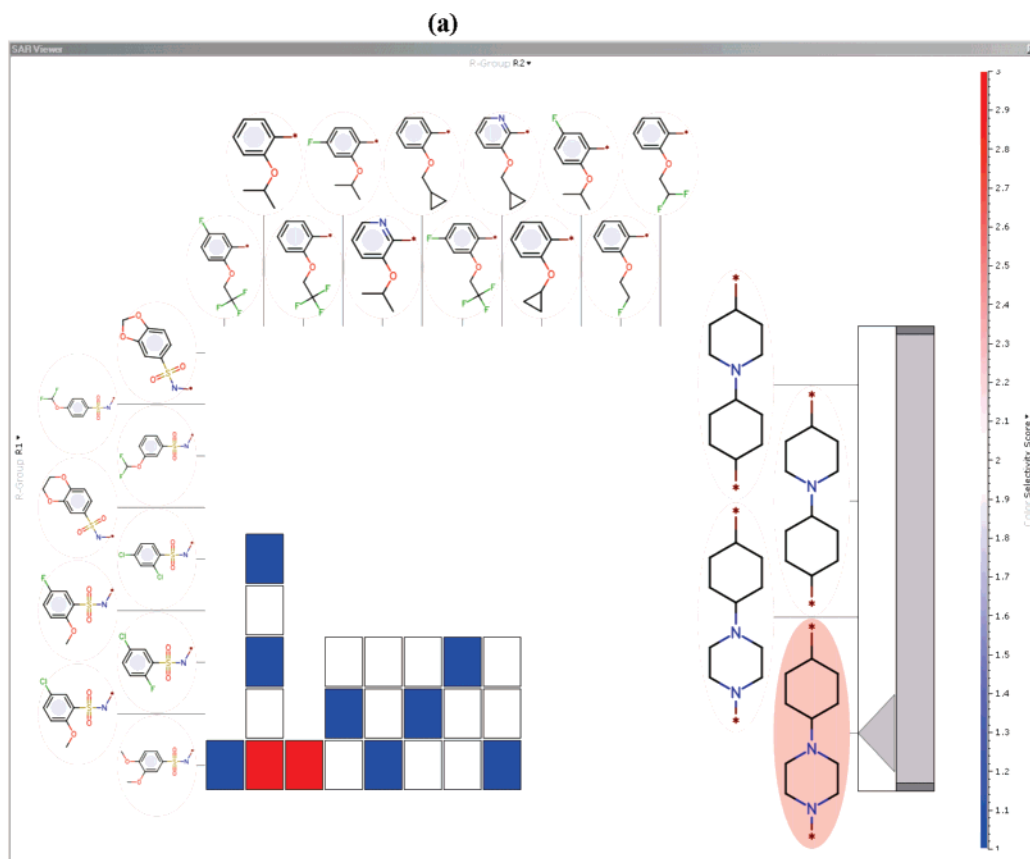
Discussion

We illustrate the utility of SAR maps (and the ABCD data integration, in general) using a recently completed program directed toward the identification of inhibitors of cyclin-dependent kinase-1 (CDK1), which are useful as anticancer agents.²⁹ Analysis began with a substructure search of the ABCD data warehouse for compounds belonging to the chemical series of interest (1-acyl-3,5-diamino-1,2,4-1*H*-triazoles) followed by extraction of relevant biological results. The resulting table of chemical structures and biological data was submitted to R-group analysis to generate a table of R-substituents derived from an algorithmically determined MCS or user-defined core structure. Figure 1 shows the core and R_1 and R_2 substituents for a representative sample of the 74 analogs described in the table.

The results of the R-group analysis may be displayed using chemistry-rich 2D matrices generated by the SAR map tool. In Figure 2a, for example, structures for the R_1 groups (the substituents attached to N-1 of the 3,5-diamino-1,2,4-1*H*-triazole core) are displayed horizontally across the top of the matrix, and structures for the R_2 groups (the substituents attached to the 3-amino group of the 3,5-diamino-1,2,4-1*H*-triazole core) are displayed vertically along the left side. The colored rectangles represent individual compounds having particular R_1 and R_2 substituents, and the color represents biological activity at a given target; in this case, it is the pIC_{50} against CDK1. Potent inhibitors are shown in red ($pIC_{50} > 9$), weak inhibitors in blue ($pIC_{50} < 5.5$), and inhibitors of intermediate potency by a linear gradient from red to blue through white (indicated by the scale on the right side of the matrix). In Figure 2a, the R_2 groups are sorted according to their calculated octanol/water partition coefficients (AlogP), with substituents having higher AlogP values at the top and those with lower AlogP values at the bottom.

The clustering of red at the bottom of the matrix in Figure 2a clearly shows that R_2 AlogP is loosely associated with CDK1 inhibition. In particular, highly potent CDK1 inhibitors (shown by the red rectangles) contain N-substituted 4-sulfamoylphenyl R_2 groups, and the most potent inhibitors contain an unsubstituted 4-sulfamoylphenyl group (AlogP = 0.54). Analogs with lipophilic R_2 groups like 3-chlorophenyl (AlogP = 2.49) or 4-(4-methylpiperazin-1-yl)phenyl (AlogP = 1.95) are among the least potent CDK1 inhibitors in the dataset. Furthermore, exceptions to the correlation hypothesis are simple to identify using the SAR map. In this particular example, the break in color trend, where $R_2 = 4$ -methanesulfonylaminophenyl, is visually striking and suggests that factors other than lipophilicity, such as hydrogen bonding or steric interactions, are at play.

Once the R-group analysis is set up, the SAR map allows the analyst to easily test the same hypothesis (R_2 AlogP correlates with activity) against other biological targets. For the triazole CDK1 inhibitors, inhibition data for other kinase targets are available and can be trivially retrieved from ABCD. It is informative to examine activity at enzymes from different kinase families to obtain a measure of the overall selectivity of these molecules. One such enzyme is VEGFR2 kinase (also known as KDR), a member of the receptor tyrosine kinase family, evolutionarily distinct from the CDK family of serine-threonine kinases. Figure 2b shows the SAR map of the CDK1 inhibitor compound set color-coded by VEGFR2 kinase activity. Arrangement of R_1 and R_2 groups in the matrix and color scaling



(kinase potency in pIC_{50}) are identical to those used in Figure 2a, allowing a direct side-by-side visual comparison of VEGFR2

inhibition with CDK1 inhibition. The evident lack of bright red color indicates that the triazoles are less potent inhibitors of

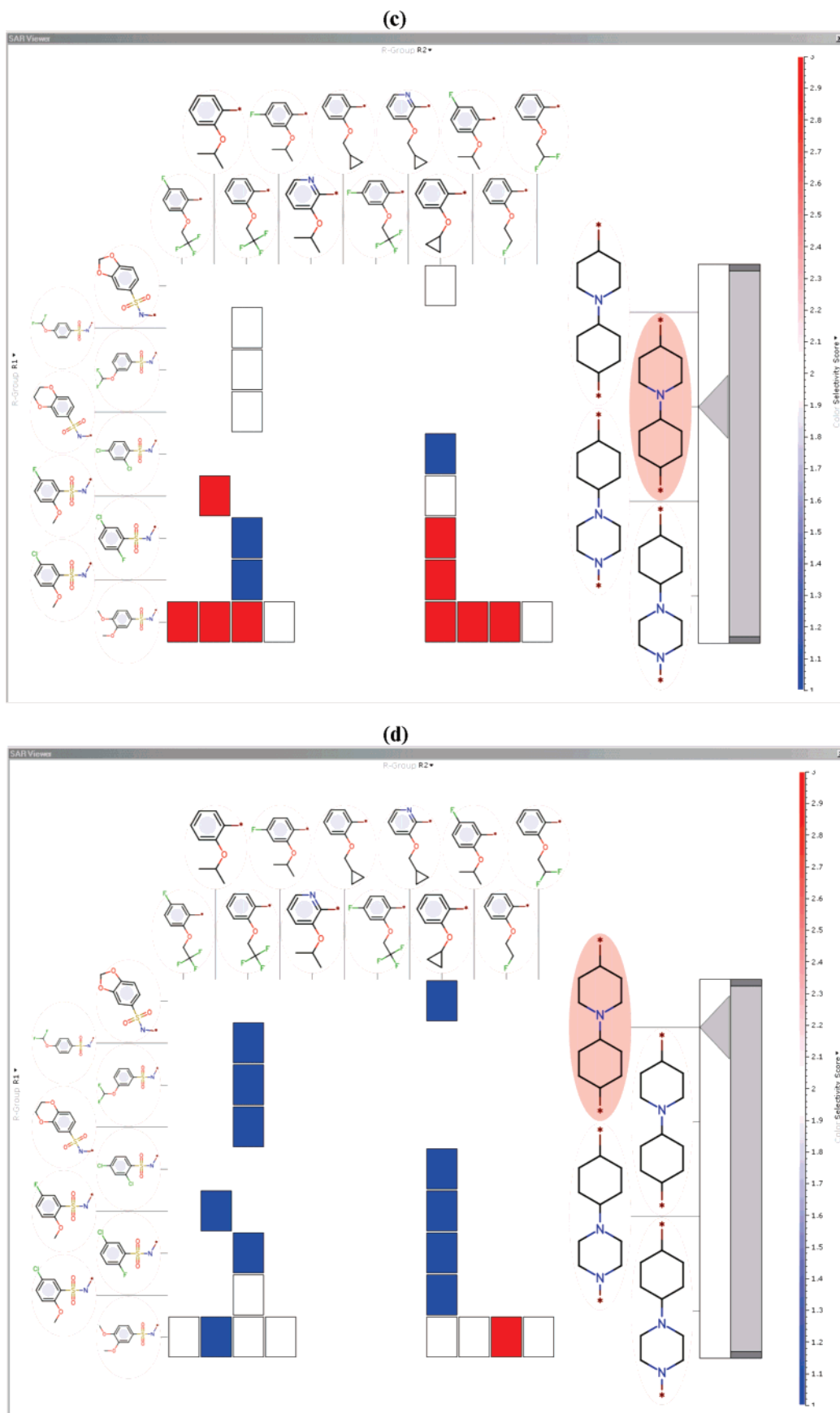


Figure 5. (a) Selectivity scores of *cis*-piperazines; (b) selectivity scores of *trans*-piperazines; (c) selectivity scores of *cis*-piperidines; and (d) selectivity scores of *trans*-piperidines. Legend: score 3 = α_{1d}/α_{1a} K_i ratio 0.33–3 and α_{1b}/α_{1a} K_i ratio >50; score 2 = α_{1d}/α_{1a} K_i ratio 0.2–0.33 or 3–5 and α_{1b}/α_{1a} K_i ratio >50; score 1 = α_{1d}/α_{1a} K_i ratio <0.2 or >5 or α_{1b}/α_{1a} K_i ratio <50.

VEGFR2 kinase (maximum pIC_{50} = 7.9) than CDK1 (maximum pIC_{50} = 9.2). Also, the VEGFR2 inhibition pattern is markedly

different from the CDK1 pattern. Instead of a cluster of potent inhibition around R_2 groups with low AlogP, VEGFR2 inhibi-

tion correlates with specific R₁ substituents, namely, the thiophen-2-carbonyl groups (pale red). Such a correlation implies that the inhibitor binding site in VEGFR2 kinase might bind the heterocyclic thiophene ring with higher affinity than the benzene ring present in the less potent (blue) inhibitors. Note that the analogs containing the 4-sulfamoylphenyl group (along the bottom row of the SAR map), which are the most potent inhibitors of CDK1 (Figure 2a), are not consistently potent inhibitors of VEGFR2 kinase (Figure 2b).

To more clearly visualize CDK1 versus VEGFR2 selectivity, the log of the ratio of VEGFR2 IC₅₀ to CDK1 IC₅₀ may be used as the color-coding parameter (Figure 2c). Some aspects of the color pattern in Figure 2c are similar to the pattern observed for CDK1 potency (Figure 2a), clearly showing that the most potent CDK1 inhibitors (R₂ = 4-sulfamoylphenyl) are for the most part also the most selective. Another group of selective analogs are those where R₁ is 2,6-difluorophenyl or 2,6-difluoro-3-methylphenyl, as shown by the seventh and ninth columns of the SAR map.

The SAR map is also a powerful tool for visualizing SARs in datasets with more complex R-group patterns. As an example, we analyzed a set of substituted *N*-[4-(piperazin-1-yl)cyclohexyl]- and *N*-[4-(piperidin-4-yl)cyclohexyl]-sulfonamide derivatives that were discovered to be high affinity ligands for subtypes of the α_1 -adrenergic receptor.^{30,31} Analogs possessing a desired selectivity profile, namely, high affinity for the α_{1a} and α_{1d} subtypes and lower affinity for the α_{1b} subtype, are desirable as potential agents for the treatment of benign prostatic hyperplasia and lower urinary tract symptoms. A dataset of 76 compounds was extracted from the ABCD database and was submitted to R-group analysis, as described above, producing another table of R-substituents derived from four core structures, *cis*-piperazines, *trans*-piperazines, *cis*-piperidines, and *trans*-piperidines, representing the four distinct chemical subseries in the dataset (Figure 3). SAR maps allow the data for all four core structures and the various R₁ and R₂ substituents to be displayed conveniently in chemistry-rich multidimensional matrices. Figure 4a–d illustrates binding affinities of compounds in each subseries, respectively, for the α_{1a} -adrenergic receptor subtype. The SAR map tool allows the user to interactively switch between different core structures to see the relative populations, diversity, and activity pattern of each subseries. For example, it is apparent that different sets of R₂ substituents were incorporated in the *cis/trans*-piperazine cores compared to the *cis/trans*-piperidine cores. Piperazine SAR is highly concentrated in the first eight R₂ substituent columns, while piperidine SAR skips columns five through eight. Also, it is evident that mono-, di-, and trifluoroethoxy R₂ substituents are much more broadly represented in the piperidines. SAR maps can easily identify gaps like these so that holes in the SAR matrix may be filled, if desired.

Examination of SAR maps for the piperazines (Figure 4a,b) and the piperidines (Figure 4c,d) reveals that, for analogs that share common R₁ and R₂ substituents (shown in the lower left corners of the figures), binding affinities are comparable. On the other hand, direct comparison of *cis*- and *trans*-isomer pairs (alignment of Figure 4a with 4b and Figure 4c with 4d) shows that in most cases the *cis*-isomers have higher α_{1a} binding affinities than the corresponding *trans*-isomers. This difference is readily apparent for the piperazine subseries, but is more striking in the piperidine subseries.

Although high α_{1a} binding affinity is a critical component of the activity profile, α_1 -adrenergic receptor subtype selectivity is equally important. The target selectivity profile for this

research program was approximately equal binding affinity for α_{1a} and α_{1d} subtypes and high selectivity versus the α_{1b} subtype. To enable rapid classification of analogs, a scoring parameter was computed for each compound according to the following criteria. Compounds having α_{1d}/α_{1a} K_i ratios between 0.33 and 3 and α_{1b}/α_{1a} K_i ratios >50 were assigned scores of 3 (highest selectivity). Compounds having α_{1d}/α_{1a} K_i ratios between 0.2 and 0.33 or between 3 and 5 (i.e., intermediate α_{1d}/α_{1a} selectivity) and α_{1b}/α_{1a} K_i ratios >50 were assigned scores of 2. Compounds having α_{1d}/α_{1a} K_i ratios <0.2 or >5 or α_{1b}/α_{1a} K_i ratios <50 were assigned scores of 1 (i.e., either low α_{1d}/α_{1a} selectivity, low α_{1b}/α_{1a} selectivity, or both). When the selectivity score is used as the activity parameter, SAR maps may be used to rapidly identify compounds of high interest. Figure 5a–d show the selectivity scores of each analog in the four subseries in a clear color-coded pattern, where red, white, and blue represent highly selective (score = 3), moderately selective (score = 2), and “nonselective” (score = 1) compounds, respectively. This analysis demonstrates that high binding affinity does not necessarily correlate with high selectivity. While many piperazines bind strongly to the α_{1a} subtype (pK_i values >8, Figure 4a,b), only four analogs (two *cis*-piperazines and two *trans*-piperazines) meet the strictest selectivity criteria (score = 3, Figure 5a,b). In contrast, 10 analogs in the piperidine subseries (nine *cis*-piperidines and one *trans*-piperidine) meet the strictest selectivity criteria (score = 3, Figure 5c,d), and in most cases the high selectivity goes hand in hand with high α_{1a} affinity (Figure 4c,d). Therefore, in terms of overall potential, the *cis*-piperidines appear to offer the best combination of potency and selectivity.

Conclusion

By combining the power of R-group analysis with the visual capacity of heatmaps, SAR maps deliver dense information visualizations rich in chemical context. The method can be combined with other visualization techniques such as the activity-normalized VlaaiVis pie charts,¹⁹ to display additional dimensions of chemical and biological data. SAR maps mirror the way in which therapeutic agents are being discovered and optimized and allow structure–activity patterns to be easily identified and exploited in analog design. We believe that this component, when properly implemented and integrated into an interactive data analysis application with dynamically linked displays, provides an effective solution to a significant unmet need in SAR analysis and visualization.

We wish to thank the numerous users of ABCD and Third Dimension Explorer for providing valuable feedback during the development of this tool.

References

- (1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
- (2) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Heidelberg, 1996.
- (3) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213.
- (4) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (5) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem., Int. Ed. Engl.* **1996**, *34* (23–24), 2674–2677.

- (6) Lamping, J.; Rao, R.; Pirolli, P. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Denver, Colorado, May 7–11, 1995, ACM Press/Addison-Wesley: New York, 1995; pp 401–408.
- (7) Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graphics* **1992**, *11* (1), 92–99.
- (8) Yamashita, F.; Itoh, T.; Hara, H.; Hashida, M. Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J. Chem. Inf. Model.* **2006**, *46* (3), 1054–1059.
- (9) Kibbey, C.; Calvet, A. Molecular property explorer: A novel approach to visualizing SAR using treemaps and heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523–532.
- (10) McConnell, P.; Johnson, K.; Lin, S. Applications of treemaps to hierarchical biological data. Applications of tree-maps to hierarchical biological data. *Bioinformatics* **2002**, *18* (9), 1278–1279.
- (11) Babaria, K. Using treemaps to visualize gene ontologies. 2001, preliminary report available at <http://www.cs.umd.edu/hcil/treemap/GeneOntologyTreemap.pdf> (accessed September 18, 2006).
- (12) Agrafiotis, D. K.; Bandyopadhyay, D.; Farnum, M. Radial clustergrams: Visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.* **2007**, *47*, 69–75.
- (13) Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *C-18* (5), 401–409.
- (14) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869–15872.
- (15) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215–1221.
- (16) Agrafiotis, D. K.; Xu, H. A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 475–484.
- (17) 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.* **1998**, *95*, 14863–14868.
- (18) Kibbey, C.; Calvet, A. Molecular property explorer: A novel approach to visualizing SAR using treemaps and heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523–532.
- (19) Howe, T. J.; Mahieu, G.; Marichal, P.; Tabruyn, T.; Vugts, P. Data reduction and representation in drug discovery. *Drug Discovery Today* **2007**, *1–2*, 45–53.
- (20) Patel, A.; Chin, D. N.; Singh, J.; Denny, R. A. Methods for describing a group of chemical structures. Int. Patent. Appl. WO 2006/023574, 2006.
- (21) ClassPharmer is marketed by Simulations Plus, Inc., <http://www.simulations-plus.com/>.
- (22) Medina-Franco, J. L.; Petit, J.; Maggiora, G. M. Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem. Biol. Drug Des.* **2006**, *67* (6), 395–408.
- (23) 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.
- (24) Diva is marketed by Accelrys, www.accelrys.com.
- (25) Accord for Excel is marketed by Accelrys, www.accelrys.com.
- (26) STN Express is marketed by Chemical Abstracts Service, <http://www.cas.org/>.
- (27) Agrafiotis, D. K.; et al. Advanced biological and chemical discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Info. Model.* **2007**, in press.
- (28) Shemanarev, M. The anti-grain geometry project; <http://www.anti-grain.com> (accessed June 1, 2006).
- (29) Lin, R.; Connolly, P. J.; Huang, S.; Wetter, S. K.; Lu, Y.; Murray, W. V.; Emanuel, S. L.; Gruninger, R. H.; Fuentes-Pesquera, A. R.; Rugg, C. A.; Middleton, S. A.; Jolliffe, L. K. 1-Acyl-1*H*-[1,2,4]-triazole-3,5-diamine analogs as novel and potent anti-cancer cyclin-dependent kinase inhibitors: Synthesis and evaluation of biological activities. *J. Med. Chem.* **2005**, *48*, 4208–4211.
- (30) Chiu, G.; Li, S.; Connolly, P. J.; Pulito, V.; Liu, J.; Middleton, S. A. (Arylpiperazinyl)cyclohexylsulfonamides: Discovery of $\alpha_{1a/1d}$ -selective adrenergic receptor antagonists for the treatment of benign prostatic hyperplasia/lower urinary tract symptoms (BPH/LUTS). *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3292–3297.
- (31) Chiu, G.; Li, S.; Connolly, P. J.; Pulito, V.; Liu, J.; Middleton, S. A. (Phenylpiperidinyl)cyclohexylsulfonamides: Development of $\alpha_{1a/1d}$ -selective adrenergic receptor antagonists for the treatment of benign prostatic hyperplasia/lower urinary tract symptoms (BPH/LUTS). *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3930–3934.

JM070845M