Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool

Jeremy Kolpak, Peter J. Connolly, Victor S. Lobanov, and Dimitris K. Agrafiotis*

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

Received July 22, 2009

We recently introduced SAR maps, a new interactive method for visualizing structure−activity relationships targeted specifically at medicinal chemists. A SAR map renders an R-group decomposition of a congeneric series as a rectangular matrix of cells, each representing a unique combination of R-groups color-coded by a user-selected property of the corresponding compound. In this paper, we describe an enhanced version that greatly expands the types of visualizations that can be displayed inside the cells. Examples include multidimensional histograms and pie charts that visualize the biological profiles of compounds across an entire panel of assays, forms that display specific fields on user-defined layouts, aligned 3D structure drawings that show the relative orientation of different substituents, dose−response curves, images of crystals or diffraction patterns, and many others. These enhancements, which capitalize on the modular architecture of its host application Third Dimension Explorer (3DX), allow the medicinal chemist to interactively analyze complex scaffolds with multiple substitution sites, correlate substituent structure and biological activity at multiple simultaneous dimensions, identify missing analogs or screening data, and produce information-dense visualizations for presentations and publications. The new tool has an intuitive user interface that makes it appealing to experts and nonexperts alike.

INTRODUCTION

One of the most difficult challenges in data analysis is to be able to represent whatever complexities might be intrinsic to the data in a way that can be easily understood and exploited. The task is compounded when the data spans multiple dimensions and types, a situation all too familiar in drug discovery. Indeed, modern pharmaceutical research involves the simultaneous optimization of chemical compounds against an imposing array of pharmacodynamic, pharmacokinetic, and physicochemical parameters, ranging from binding affinity to the primary and counter-targets, to stability, solubility, absorption, distribution, metabolism, and excretion, to name a few. The task is further complicated by the fact that some of these properties cannot be easily distilled into a single number or at least cannot be easily interpreted in that form. As the complexity of disease biology continues to unravel and our synthetic and screening capacity continue to increase, sifting through large volumes of sparse and heterogeneous data and understanding their relationship to chemical structure is becoming impossible even for the most gifted minds.

In some ways, the practice of medicinal chemistry has remained relatively unchanged throughout the years. The process begins by identifying a lead compound and devising a chemical scaffold and a synthetic strategy that allows the synthesis and evaluation of many related analogs through systematic attachment of different substituents around it. The process uses a divide-and-conquer approach where individual substitution sites are optimized in an iterative manner based largely on trial and error. This process is repeated until the desired potency, selectivity, or pharmacokinetic parameters are achieved or until the potential of that series is exhausted. In the latter case, new chemical scaffolds are designed, new sets of analogs are synthesized, and the cycle continues until a clinical candidate emerges or all hope is abandoned. In the last two decades there have been attempts to industrialize this general process through combinatorial chemistry and parallel synthesis, but the paradigm remains fundamentally the same.

However, there has been a profound change, and that is the speed at which bioassay data are collected and disseminated back to the project teams. Two decades ago, medicinal chemists made compounds that were synthetically accessible, evaluated the biological data a month or so later, and then designed a new set of analogs to explore various SAR hypotheses and test the limits of serendipity. The process was very chemistry driven. Today the data arrive within days after the compound is submitted, and the decision about what molecules to synthesize next necessitates constant review and re-evaluation of the data in real time.

Yet, the medicinal chemist's arsenal for visualizing such structure−activity data is surprisingly limited. Apart from the ubiquitous molecular spreadsheets found in chemical information and molecular modeling packages, most SAR visualization techniques use various forms of clustering to organize the compounds into related families.[1] With few exceptions, this partitioning is usually based on similarity measures that look at the properties of the entire molecule and do not explicitly consider the presence of common structural cores and the synthetic strategies used to embellish them. While clustering methods offer certain advantages, they

* Corresponding author phone: (610)458-6045; fax: (610)458-8249; e-mail:dagrafio@its.jnj.com.

10.1021/ci900264n CCC: $40.75
Published on Web 09/30/2009

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,[2−5] treemaps,[6−8] dendrograms,[9] radial clustergrams,[10] nonlinear maps,[11−14] heatmaps,[15,8] and various forms of conventional statistical plots such as scatter plots, bar charts, pie charts, etc.

Two visualization techniques that take advantage of common substructures are the SAR trees[16] and the scaffold trees.[17,18] A SAR tree represents a collection of compounds as an acyclic graph, where each node represents a common substructure and its children represent the R-groups around it. Each child (R-group) in turn embodies another common substructure that is shared by multiple compounds at that particular attachment site and is recursively split into more refined R-groups, until there are no further variations (a technique that is also employed in ClassPharmer[19]). In scaffold trees, each node represents a unique chemotype at some level of abstraction (e.g., complete molecules, cyclic systems, cyclic system skeletons, reduced cyclic system skeletons, etc.) 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 different levels of structural resolution.

Recently we introduced SAR maps, a new visualization technique that combines the power of R-group analysis with the visual richness of heatmaps.[20] R-group analysis takes as input a set of structures sharing a common scaffold, identifies the variation sites around that scaffold, and tabulates all the substituents at each variation site along with any associated activity values. Traditional SAR tables have become the *lingua franca* of medicinal chemistry; open any random article in a medicinal chemistry journal, and the reader is likely to find at least one such SAR table consisting of a generic structure, accompanied by columns describing different substituents and their biological and/or physicochemical properties. Though automated R-group analysis is available in a number of software packages such as Diva,[21] Accord for Excel,[22] and STN Express,[23] most medicinal chemists still generate and maintain these tables by hand. This process is arduous, time-consuming, and susceptible to transcription errors.

SAR maps automate R-group analysis and simplify the visualization of the resulting data. A SAR map is essentially a heatmap with chemical axes. It renders an R-group decomposition of a congeneric series as a rectangular matrix of cells, each representing a unique combination of R-groups, and thus a unique compound. The cells are color-coded by chemical property or biological activity, which makes activity patterns easy to identify. The tool has built a strong internal user base and has generated significant interest in the broader scientific community. [The paper in ref 20 describing the original SAR maps was the fourth most accessed article published in *J. Med. Chem.* in 2007, and the authors received many inquiries regarding the public availability of the tool.]

After more than a year of extensive internal use, it became apparent that while SAR maps represented a significant step in SAR analysis, their full potential was yet to be realized. The first limitation was that it was only possible to visualize one property (or activity) at a time. While the user interface provided an easy way to rapidly sift through different properties, it was impossible to capture the entire biological profile of the compound set in a single plot. Furthermore, our medicinal chemists quickly identified the need to display more advanced visualizations inside the cells, such as individual activity values, dose−response curves, images, etc. Another desirable feature was the ability to sort the substituents on the chemical axes by any attribute (e.g., size, aromatic character, hydrogen bonding potential, lipophilicity, etc.) or arrange them interactively in a specific, user-defined order.

In this work, we present an enhanced version of the SAR map that makes it possible to display any type of information in any visual form inside the cells. Examples include multidimensional histograms and pie charts, forms, dose−response curves, 3D structure drawings, images, and many others. The remaining sections describe the architectural design and core functionality of the new viewer and provide a few examples from a recently completed medicinal chemistry project that illustrate its potential.

METHODS

Third Dimension Explorer (3DX) and ABCD. The enhanced SAR maps were implemented as a plug-in to Third Dimension Explorer (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,[24] 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 (a feature that is exploited by the new version of the SAR maps), 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 enhanced SAR maps described herein. The program offers a full range of navigation and selection options, 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 new data types and cell renderers, create and (re)arrange viewers, etc. Their functionality and implementation can be very 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 chemically aware data mining tools were introduced in this fashion, including exact structure, substructure, and similarity searching, structure alignment, maximum common substructure detection, chemotype classification, R-group analysis, 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.

Enhanced SAR Map. Just like the original version, the enhanced SAR map renders an R-group decomposition as a rectangular grid of cells. Each cell represents a single compound C_i, defined as the combination of its constituent R-groups $\{R_1(i), R_2(i), ..., R_n(i)\}$, where $R_j(i)$ is the substituent at the j-th variation site in compound i. The default SAR 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 $\{R_{r1,j1=1,...,|R_{r1}|}\} \times \{R_{r2,j2=1,...,|R_{r2}|}\} \times R_{r3,j3} \times ... \times R_{rn,jn}$, where all but 2 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 also includes 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). Since it is extremely rare that a 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 which 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.

The enhanced SAR map extends this core functionality by allowing the user to display any graphical object inside the cells. To understand the user interface, it is important to understand how the SAR map interacts with its host application, 3DX. As mentioned in the preceding sections, the SAR map is a viewer associated with a 3DX table. A 3DX table consists of an array of columns, and each column contains data of a specific type, such as strings, integers, floating point numbers, lists of numbers, dates, etc. Each data type is in turn associated with a number of renderers which are registered for that particular type. For example, a chemical structure type is associated with a 2D molecular renderer, a 3D molecular renderer, and a text renderer which displays the structure as a SMILES string. The user can interactively switch between renderers at run time to display the data in the preferred format. The architecture is illustrated in Figure 1.

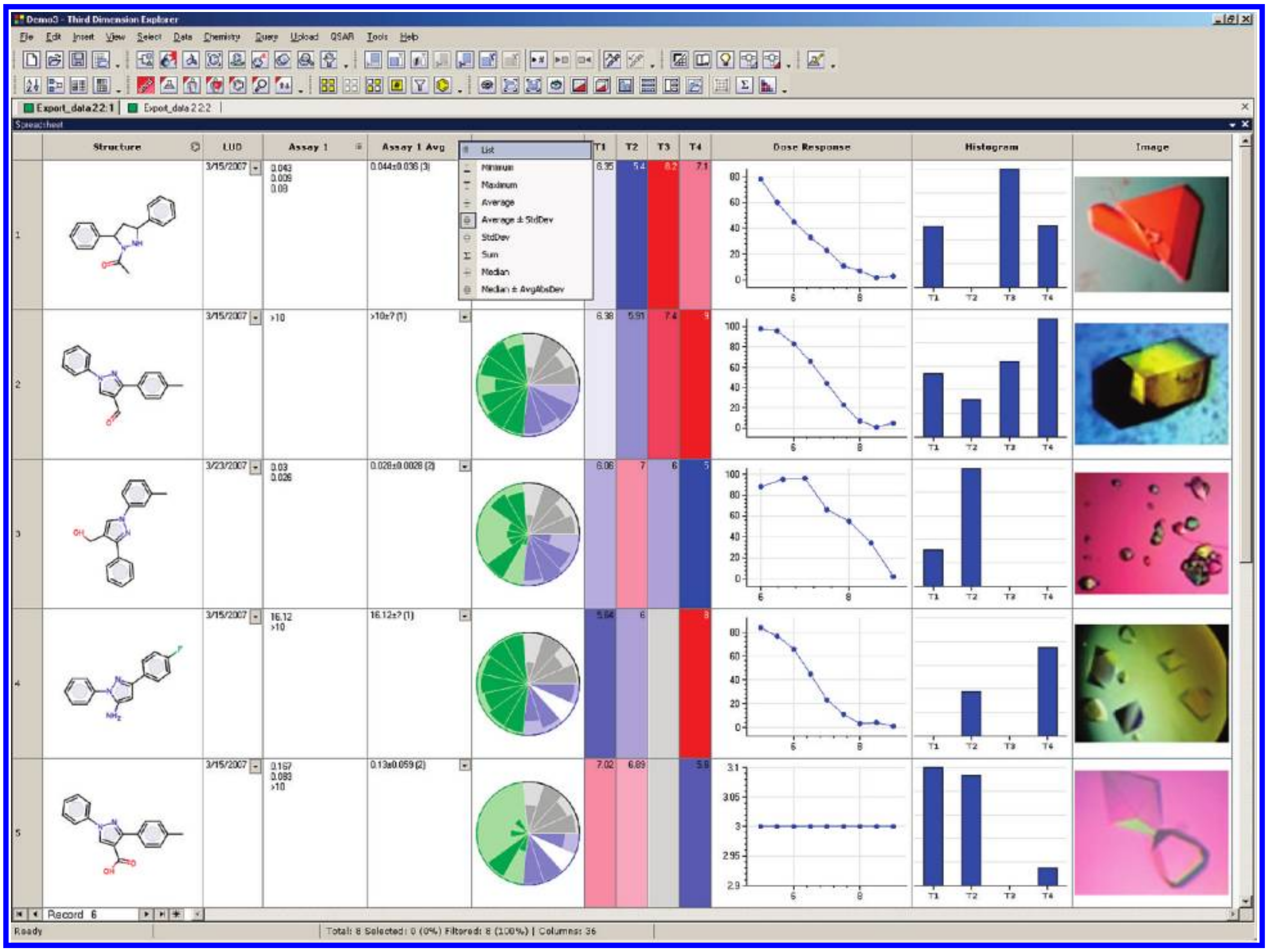
The enhanced SAR map allows the user to select any column from the table and display the contents of that column inside its cells as it is currently shown on the spreadsheet. Since every cell in the SAR map corresponds to a unique combination of R groups and therefore a unique molecule, it is trivial to identify the spreadsheet row that corresponds to that molecule and capture the drawing in the selected column. In essence, the SAR map reuses the renderer that is embedded in the spreadsheet, scales the drawing to fit the current cell size, and renders its contents in its own graphics device context. This architecture is simple, elegant, and highly extensible. Given that both new data types and new renderers can be implemented as plugins, the visualization capabilities of 3DX and the SAR viewer can be extended without modifying the core application or the SAR viewer itself.

The new SAR map offers a number of additional enhancements, such as the ability to sort the substituents at each variation site by any molecular property or by manually dragging them along the chemical axes/sliders, the ability to aggregate the values of multiple rows that contain the same structure, the ability to zoom into a particular region of the map, and the ability to highlight the current and selected records. As a registered viewer, the SAR map has full access to the data in that table and subscribes to all its events. This allows it to automatically refresh itself when the data in the table is changed and respond to any other relevant change that is broadcasted by the table.

The enhanced SAR viewer is implemented as a .Net control and is based on the GDI+ API. The core molecular renderer is based on our implementation of Anti-Grain Geometry (AGG),[25] a high-quality, lightweight, extensible, and platform-independent rendering engine written in standard ANSI C++. AGG provides very fast antialiased graphics with subpixel accuracy, allowing effective visualization of a large number of chemical structures with minimal loss of resolution and clarity.

DISCUSSION

In our original publication, we introduced simple SAR maps to visualize structure−activity relationships of cyclin-dependent kinase (CDK) inhibitors and alpha-1 adrenergic receptor antagonists. These simple SAR maps are limited by design to analysis of one activity parameter at a time. More complex analysis that compares, for example, biological activity at two or more targets requires the user to “walk” the SAR viewer through the individual activity parameters to generate multiple SAR maps. These resulting single-



	LUD	Assay 1	Assay 1 Avg	T1	T2	T3	T4
1	3/15/2007	0.043 0.009 0.08	0.044±0.036 (3)	6.35	5.4	8.2	7.1
2	3/15/2007	>10	>10±? (1)	6.38	5.91	7.4	9
3	3/23/2007	0.03 0.026	0.028±0.0028 (2)	6.06	7	6	5
4	3/15/2007	16.12 >10	16.12±? (1)	5.64	6		8
5	3/15/2007	0.167 0.093 >10	0.13±0.059 (2)	7.02	6.89		5.8

Ready | Total: 8 Selected: 0 (0%) Filtered: 8 (100%) | Columns: 36

Figure 1. Representative custom cell renderers in Third Dimension Explorer (3DX). Several data types are highlighted, including chemical structures, dates, qualified (fuzzy) number lists with on-the-fly aggregation, charts, histograms, pie bar charts, and images.

parameter SAR maps must then be displayed side by side to reveal differences in activities by comparing color/activity patterns. Although the first-generation tool was a significant step forward, a more sophisticated means to analyze multiple activity parameters in a SAR map setting was clearly needed.

The second-generation SAR map tool described herein allows the user to display almost every imaginable type of visualization inside its activity cells. Such flexibility frees the SAR map from the limitations of single-parameter color-coding and greatly improves the ability of the chemist to correlate structure with multiple activity patterns.

To illustrate the utility of the enhanced SAR viewer, we present a SAR map analysis of a recently completed program that was directed toward the identification of kinase inhibitors potentially useful as anticancer agents. The biological targets of this research program were receptor tyrosine kinases (RTKs) in the ErbB family, specifically epidermal growth factor receptor (EGFR, ErbB-1, HER-1) and HER-2 (ErbB-2). Abnormal regulation of ErbB receptor signaling has been identified in solid tumors such as breast, lung, colon, and prostate, and ErbB receptors have been popular targets for drug intervention by monoclonal antibodies targeted to their extracellular domains or small-molecule inhibitors of their intracellular kinase domains.[26]

Approved monoclonal antibody drugs trastuzumab (Herceptin) and cetuximab (Erbitux) selectively target HER-2 and EGFR, respectively, whereas approved kinase domain inhibitors gefitinib (Iressa) and erlotinib (Tarceva) selectively inhibit EGFR and lapatinib (Tykerb) inhibits both EGFR and HER-2. The 4-amino-pyrimidine-5-carbaldehyde oxime scaffold has previously demonstrated its versatility as a platform for building inhibitors of vascular endothelial growth factor-2 (VEGFR-2) kinase[27] and fms-like tyrosine kinase 3 (FLT-3),[28] and we endeavored to extend its utility to the ErbB family. Over the course of the project, nearly 200 4,6-diamino-pyrimidine-5-carbaldehyde oxime analogs were prepared and tested in a battery of target related and off-target *in vitro* assays. Many of these molecules moved forward into ADME protocols and *in vivo* antitumor efficacy models. Papers discussing initial structure−activity relationships in the series[29] and disclosing biological activities of a potential drug development candidate, JNJ-28871063, have been recently published.[30]

The wealth of biological data accumulated for the 4,6-diamino-pyrimidine-5-carbaldehyde oximes in kinase and cell-proliferation assays presents an ideal opportunity to showcase the enhanced functionality of the new SAR viewer. To begin the analysis, we retrieved chemical structures and relevant biological results associated with the chemical series from the ABCD data warehouse using the ABCD query interface available in 3DX. (For this discussion the data set of more than 200 molecules has been limited to approximately 80 representative structures to better illustrate the features and power of the enhanced viewer.) The resulting table of chemical structures and biological data was submitted to R-group analysis to generate a table of R-substituents

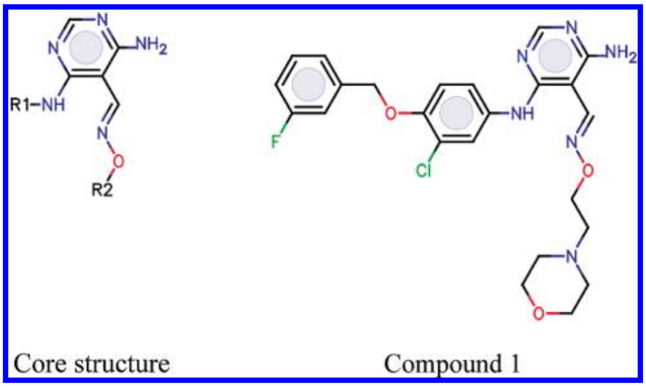


Figure 2. Core structure used for R-group enumeration and generation of SAR maps, plus a representative structure from the data set, Compound **1**.

derived from a user-defined core structure. Figure 2 shows the 4,6-diamino-pyrimidine-5-carbaldehyde oxime core used for R-group generation, with the R1 substituent attached to the 4 amino group and the R2 substituent attached to the oxime oxygen, along with a sample structure from the data table, where R1 is 3-chloro-4-(3-fluorobenzyloxy)phenyl and R2 is 2-(morpholin-4-yl)ethyl (JNJ-28871063).

Because the 4-amino-pyrimidine-5-carbaldehyde oxime scaffold could be derivatized to give molecules that inhibit other kinases, we were interested in monitoring kinase selectivity of our ErbB inhibitors. Indeed, given the ubiquity of protein kinases in the human proteome, the potential for undesired activity or toxicity resulting from nonselective inhibition of nontargeted kinases must always be considered in any small-molecule, ATP-competitive inhibitor program. In our research, we routinely screened ErbB compounds against a small subset of off-target kinases to get a quick read of selectivity. In this analysis, we examined the effect of compounds on the serine-threonine kinase, CDK-1, and the receptor tyrosine kinase, VEGFR-2. The original SAR viewer displayed activity only in the form of color gradient cells. This limited but useful capability has been retained in the new version. Figure 3a shows an original-format SAR map of HER-2 activity. Potent inhibitors are shown in red ($pIC_{50} > 8$), weak inhibitors in blue ($pIC_{50} < 4.5$), 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). Cells with diagonal hash marks indicate that no activity data are available although a molecule having the corresponding R-groups is present in the data set.

A new capability of the second generation SAR viewer is the ability to resort the matrix on the fly using independent parameters for the R1 and R2 axes. In Figure 3 a-d, the R2 (horizontal) axes are sorted by molecular weight (increasing from left to right), and the R1 (vertical) axes are sorted by HER-2 potency (increasing from top to bottom). Figure 3b displays a second original SAR map, still sorted by HER-2 potency, of another targeted kinase, EGFR. It is evident from a cursory comparison of Figure 3a,b that HER-2 and EGFR potency in this set of compounds are closely associated, which is unsurprising given that both kinases are members of the same RTK family. On the other hand, original SAR maps of the two off-target kinases, CDK-1 (Figure 3c) and VEGFR-2 (Figure 3d), show that no analogs with potent HER-2 activity are potent CDK-1 inhibitors, and only a handful of potent HER-2 inhibitors are also potent VEGFR-2 inhibitors.

The single-parameter SAR maps shown in Figure 3 are useful for high-level analysis of activity data, but side-by-side comparison of activities is only possible by taking static screen shots. However, static data display fails to use the SAR map viewer as it was first conceived, i.e. an interactive tool for data analysis. One of the chief enhancements of the second-generation SAR viewer is its ability to display and manipulate any type of custom cell renderer available in 3DX, including chemical structures, identifiers, forms, qualified number lists, charts, histograms, pie bar charts, and images, to mention a few. In addition, the second-generation SAR maps can be manipulated in a highly interactive manner. Zoom regions can be defined by manipulating scroll bars along the R-group axes or by using the mouse to draw a rectangle around the desired matrix cells. Molecules may be selected in the same manner, and selections in the SAR map are completely synchronized with the corresponding rows in the underlying data table and any other viewers that are currently open. Finally, matrix cells may be expanded for easier visualization of their contents by a simple mouse-over action; this is particularly convenient for detail-rich renderers like forms and pie barcharts when the SAR map is zoomed out and matrix cells are small.

Taking advantage of second-generation improvements, the ErbB inhibitor SAR map was enhanced by displaying a simple 4-kinase histogram in the matrix cell in place of a single activity color gradient (Figure 4a-e). Heights of the blue bars in the histogram represent increasing pIC_{50} values for CDK-1, EGFR, HER-2, and VEGFR-2 and provide a clear visual pattern representing kinase potency or lack thereof. Figure 4a shows a mouse-over expansion of the histogram highlighting an EGFR/HER-2 selective compound near the bottom right of the SAR map. Using the vertical scroll bar, this SAR map was zoomed to focus on the lower 6 rows of the data matrix (Figure 4b). Note that most of the molecules in this region inhibit EGFR and HER-2 with about equal potency, and no compounds have significant effect on CDK-1. However, the histograms show quickly and unambiguously that seven molecules inhibit VEGFR-2 and that the molecule in the lower right corner is more potent at VEGFR-2 than either EGFR or HER-2. Because a SAR map can be configured to display other data renderers, it is a simple matter to switch from the histogram to a text-based form. Figure 4c shows the zoomed SAR map with a small form in place of the histogram. Moving to another region of the zoomed-out SAR map, Figure 4d shows a mouse-over expansion highlighting a relatively potent CDK-1 inhibitory compound located in the middle left. Using the vertical scroll bar, this SAR map was zoomed to focus on six rows of the data matrix near the highlighted CDK-1 inhibitor (Figure 4e). Finally, Figure 4f shows the zoomed CDK-1 region SAR map with the small form in place of the histogram. The reader should note that two other compounds exhibiting significant CDK potency are easily identified from the histograms in the SAR map.

Although histograms and text-based forms are very useful for displaying several activity parameters simultaneously, they have their limitations, notably the difficulty differentiating one histogram bar from another or one text row from

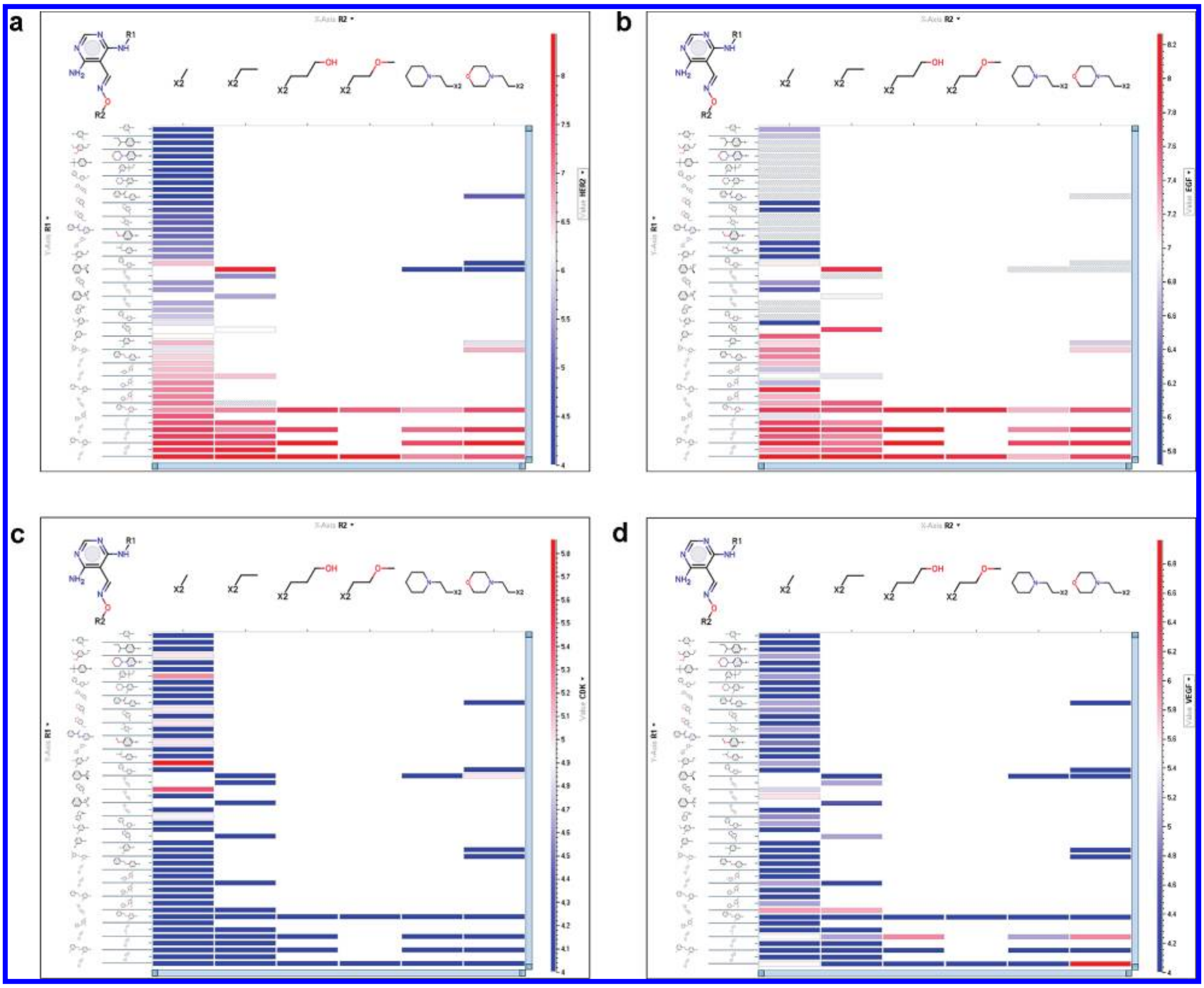


Figure 3. a. Original SAR map of HER-2 activity (pIC_{50}). b. Original SAR map of EGFR activity (pIC_{50}). c. Original SAR map of CDK-1 activity (pIC_{50}). d. Original SAR map of VEGFR-2 activity (pIC_{50}).

another when the number of parameters grows too large. To make the renderings visually acceptable, the histogram or form must be displayed in a larger matrix cell, severely limiting the number of rows and columns that may be displayed practically on the screen. One type of rendering tool implemented in 3DX that is highly suited for the display of large numbers of activity parameters is the VlaaiVis pie barchart.[31] VlaaiVis is a radial plot representing the complex biological profile of a single compound that bears close resemblance to the flower plots introduced more than a decade ago by Martin and co-workers.[32] 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.

Configuration of the pie barchart for the ErbB data set was easily accomplished through the 3DX user interface, illustrated in Figure 5a. The resulting pie barchart is configured to display 20 activity parameters with the color coding shown in Figure 5b: red, EGF (ErbB) family RTKs EGFR, HER-2, HER-2 (phosphorylation assay), and HER-4; blue, non-EGF family RTKs insulin receptor kinase (IRK), RET, and VEGFR-2; green, serine-threonine kinases CDK-1 and aurora-A (AUR-A); gray, cytotoxicity in cultured tumor cell lines SK-BR-3, HeLa, and BT-474; magenta; antiproliferative activity in cultured tumor cell lines A-431, SK-BR-3, BT-474, N87, MCF7, SK-OV-3, HCT 116, and HeLa. All activities are graphed as pIC_{50} values where the lengths of the dark-colored pie wedges (on a lighter-colored background) extending radially from the center indicate increasing potency, from $pIC_{50} < 4$ (no wedge) to $pIC_{50} > 9$ (wedge extends to perimeter of the pie circle). If data are not present for an activity parameter, the background of a wedge region is shown in white. For the ErbB data set, ideal molecules would inhibit ErbB kinases selectively without significantly affecting non-EGF family kinases or serine-threonine kinases (in the pie barchart, filled red pie wedges but minimally filled blue and green wedges). Cytotoxicity or antiproliferative activity (filled gray or magenta wedges) indicates the breadth of potential antitumor activity of a compound, although it is reasonable to expect that ErbB-selective inhibitors act specifically on cell lines that overexpress ErbB receptors like A-431, SK-BR-3, BT-474, and N87 or whose growth is driven by them.

Figure 6 illustrates the ErbB inhibitor SAR map reconfigured to display the information-rich 20-parameter kinase

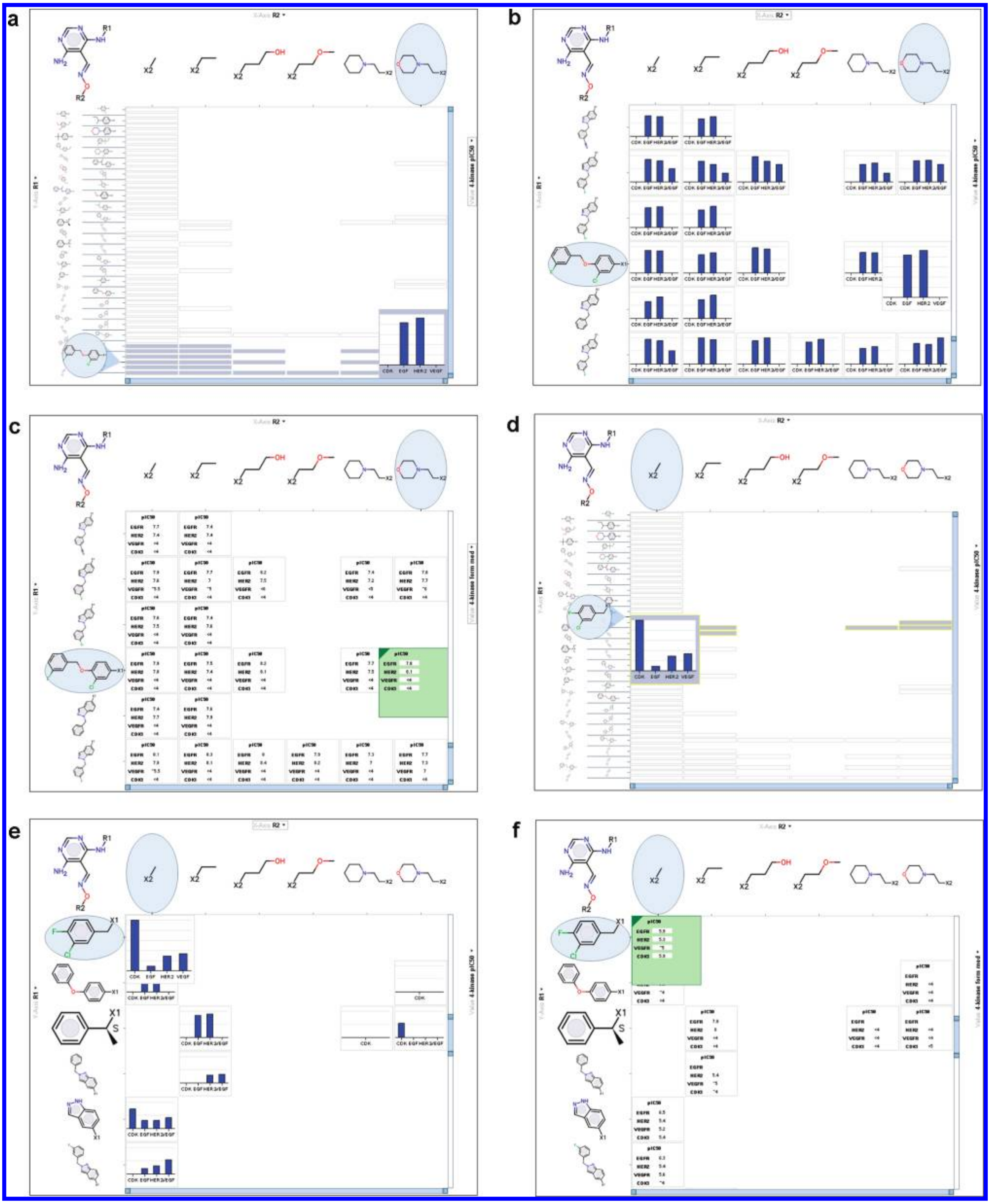
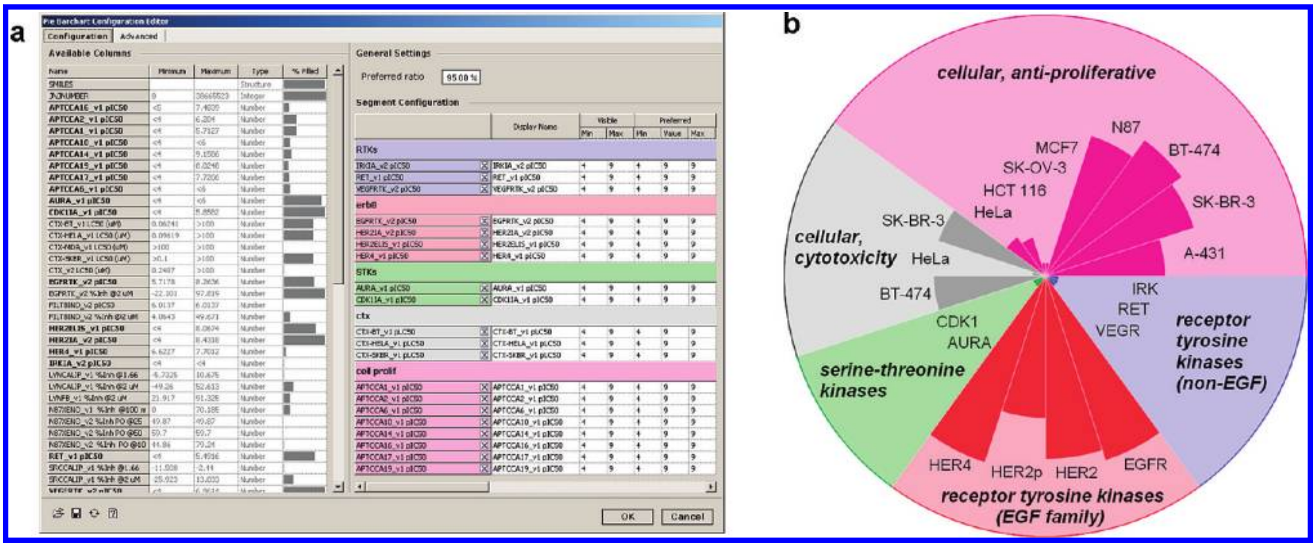


Figure 4. a. Second-generation SAR map showing 4-kinase histogram of an EGFR/HER-2 selective compound. b. Zoomed region showing lower 6 rows of the 4-kinase histogram SAR map. c. Zoomed region showing a text-based form SAR map. d. Enhanced SAR map showing 4-kinase histogram of a CDK-1 inhibitory compound. e. Zoomed region showing 6 middle rows of the 4-kinase histogram SAR map. f. Zoomed region showing a text-based form SAR map.

and cellular activity pie barchart. The two zoomed regions in Figure 4 are shown again in Figure 6a (lower six rows, EGFR/HER-2 selective) and Figure 6b (middle six rows, CDK-1 inhibitory region) as pie barcharts instead of histograms or forms. It is evident from the activity pattern displayed on the pie barcharts in Figure 6a that the highlighted molecule, Compound **1**, having R1 = 3-chloro-4-(3-fluorobenzyloxy)phenyl and R2 = 2-(morpholin-4-yl)ethyl, has a nearly ideal profile. It is a potent and selective inhibitor of ErbB family kinases and has potent antiproliferative activity in all four ErbB overexpressing tumor cell lines. Two other molecules come close to matching its ErbB



Pie Barchart Configuration Editor

Configuration | Advanced

Available Columns

Name	Minimum	Maximum	Type	% Filled
SMILES			Structure	
JNJNUMBER	0	38665523	Integer	
APTCCA16_v1 pIC50	<5	7.4020	Number	
APTCCA2_v1 pIC50	<4	6.204	Number	
APTCCA1_v1 pIC50	<4	5.7127	Number	
APTCCA10_v1 pIC50	<4	<6	Number	
APTCCA14_v1 pIC50	<4	9.1506	Number	
APTCCA19_v1 pIC50	<4	8.0246	Number	
APTCCA17_v1 pIC50	<4	7.7206	Number	
APTCCA6_v1 pIC50	<4	<6	Number	
AURA_v1 pIC50	<4	<6	Number	
CDK1IA_v1 pIC50	<4	5.8582	Number	
CTX-BT_v1 LC50 (uM)	0.06241	>100	Number	
CTX-HELA_v1 LC50 (uM)	0.09619	>100	Number	
CTX-MDA_v1 LC50 (uM)	>100	>100	Number	
CTX-SKBR_v1 LC50 (uM)	>0.1	>100	Number	
CTX_v2 LC50 (uM)	0.2487	>100	Number	
EGFRTK_v2 pIC50	5.7178	8.2636	Number	
EGFRTK_v2 %Inh @2 uM	-22.303	97.819	Number	
FILTBIND_v2 pIC50	6.0137	6.0137	Number	
FILTBIND_v2 %Inh @2 uM	4.0643	49.671	Number	
HER2ELIS_v1 pIC50	<4	8.0674	Number	
HER2IA_v2 pIC50	<4	8.4338	Number	
HER4_v1 pIC50	6.6227	7.7032	Number	
IRKIA_v2 pIC50	<4	<4	Number	
LYNCALIP_v1 %Inh @1.66	-6.7325	10.675	Number	
LYNCALIP_v1 %Inh @2 uM	-49.26	52.613	Number	
LYNFB_v1 %Inh @2 uM	21.917	51.326	Number	
N87XENO_v1 %Inh @100 m	0	70.185	Number	
N87XENO_v2 %Inh PO @25	49.87	49.87	Number	
N87XENO_v2 %Inh PO @50	59.7	59.7	Number	
N87XENO_v2 %Inh PO @10	44.86	70.24	Number	
RET_v1 pIC50	<4	5.4916	Number	
SRCCALIP_v1 %Inh @1.66	-11.508	-2.44	Number	
SRCCALIP_v1 %Inh @2 uM	25.923	13.033	Number	
VEGFRTK_v2 pIC50	<4	6.2614	Number	

General Settings

Preferred ratio: 95.00 %

Segment Configuration

		Display Name	Visible		Preferred		
			Min	Max	Min	Value	Max
RTKs							
IRKIA_v2 pIC50	☒	IRKIA_v2 pIC50	4	9	4	9	9
RET_v1 pIC50	☒	RET_v1 pIC50	4	9	4	9	9
VEGFRTK_v2 pIC50	☒	VEGFRTK_v2 pIC50	4	9	4	9	9
erbB							
EGFRTK_v2 pIC50	☒	EGFRTK_v2 pIC50	4	9	4	9	9
HER2IA_v2 pIC50	☒	HER2IA_v2 pIC50	4	9	4	9	9
HER2ELIS_v1 pIC50	☒	HER2ELIS_v1 pIC50	4	9	4	9	9
HER4_v1 pIC50	☒	HER4_v1 pIC50	4	9	4	9	9
STKs							
AURA_v1 pIC50	☒	AURA_v1 pIC50	4	9	4	9	9
CDK1IA_v1 pIC50	☒	CDK1IA_v1 pIC50	4	9	4	9	9
ctx							
CTX-BT_v1 pLC50	☒	CTX-BT_v1 pLC50	4	9	4	9	9
CTX-HELA_v1 pLC50	☒	CTX-HELA_v1 pLC50	4	9	4	9	9
CTX-SKBR_v1 pLC50	☒	CTX-SKBR_v1 pLC50	4	9	4	9	9
cell prolif							
APTCCA1_v1 pIC50	☒	APTCCA1_v1 pIC50	4	9	4	9	9
APTCCA2_v1 pIC50	☒	APTCCA2_v1 pIC50	4	9	4	9	9
APTCCA6_v1 pIC50	☒	APTCCA6_v1 pIC50	4	9	4	9	9
APTCCA10_v1 pIC50	☒	APTCCA10_v1 pIC50	4	9	4	9	9
APTCCA14_v1 pIC50	☒	APTCCA14_v1 pIC50	4	9	4	9	9
APTCCA16_v1 pIC50	☒	APTCCA16_v1 pIC50	4	9	4	9	9
APTCCA17_v1 pIC50	☒	APTCCA17_v1 pIC50	4	9	4	9	9
APTCCA19_v1 pIC50	☒	APTCCA19_v1 pIC50	4	9	4	9	9

Figure 5. a. Pie barchart configuration editor. b. Pie barchart legend.

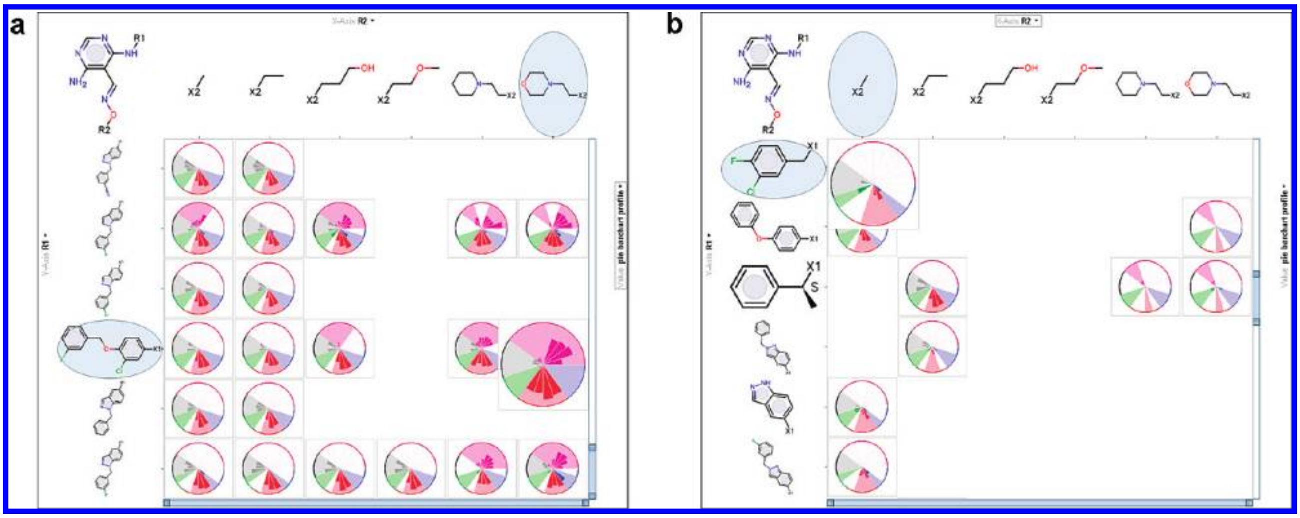


Figure 6. a. Zoomed region showing lower 6 rows of the pie barchart SAR map. b. Zoomed region showing 6 middle rows of the pie barchart SAR map.

potency and selectivity, but Compound **1** has the best overall profile. In Figure 6b, the pie barcharts tell a different story. Importantly, the compounds shown in this region are by definition not potent HER-2 inhibitors and therefore were not tested in cell proliferation and cytotoxicity assays (as evidenced by the white background of the corresponding regions). Also, the highlighted, CDK-1 inhibitory compound (R1 = 3-chloro-4-fluorobenzyl, R2 = methyl), while active in a number of kinases, is not a potent inhibitor of any of them (as evidenced by the short lengths of the gray, green, red, and blue pie wedges). A quick glance is all that is needed to conclude that molecules in this region of the SAR map are simply not interesting ErbB inhibitors.

The second-generation SAR map viewer also provides enhanced capabilities for SAR analysis using more than two R-groups. An example of a multidimensional SAR map based on the ErbB data set is shown in Figure 7a-c. An alternative R-group analysis was first performed using the more specific core structure shown in Figure 7a, wherein the R1, R2, and R3 substituents are attached to the phenylamino group and the R4 substituent is attached to the oxime oxygen. The resulting data are displayed in a four-dimensional (4D) SAR map, with the most diversely populated R1 and R2 groups (each sorted by molecular weight) making up the vertical and horizontal axes, respectively, and the less diverse R3 and R4 groups shown in sliders along the right side of the display. As illustrated in Figure 7b,c, a 4D SAR map facilitates a much more detailed analysis of the effects of particular substituents on biological properties of tested molecules. Activity trends associated with specific groups may be quickly identified by line-by-line or column-by-column comparisons. The pie barcharts in Figure 7b clearly show the dramatic effect of a single change in the R1 substituent from 3-fluorobenzyloxy to morpholin-4-yl (with R2 (Cl), R3 (H), and R4 (methyl) held constant) on HER-2 inhibition (pIC_{50} decreases from 7.8 to <4) and HER-2 phosphorylation (pIC_{50} decreases from 6.4 to <4). More conservative changes in R2, for example from 3-fluorobenzyloxy to benzyloxy or pyridine-3-ylmethoxy, have little effect on HER-2 potency, as revealed by the corresponding pie barcharts. Alternatively, the 4D SAR map view can be instantly changed to focus on alternative R4 groups by a simple click on the R4 slider. Figure 7c illustrates the SAR map where R4 is 2-(morpholin-4-yl)ethyl, once again highlighting Compound **1** and related compounds. This view represents late-stage analogs designed to optimize *in vivo*

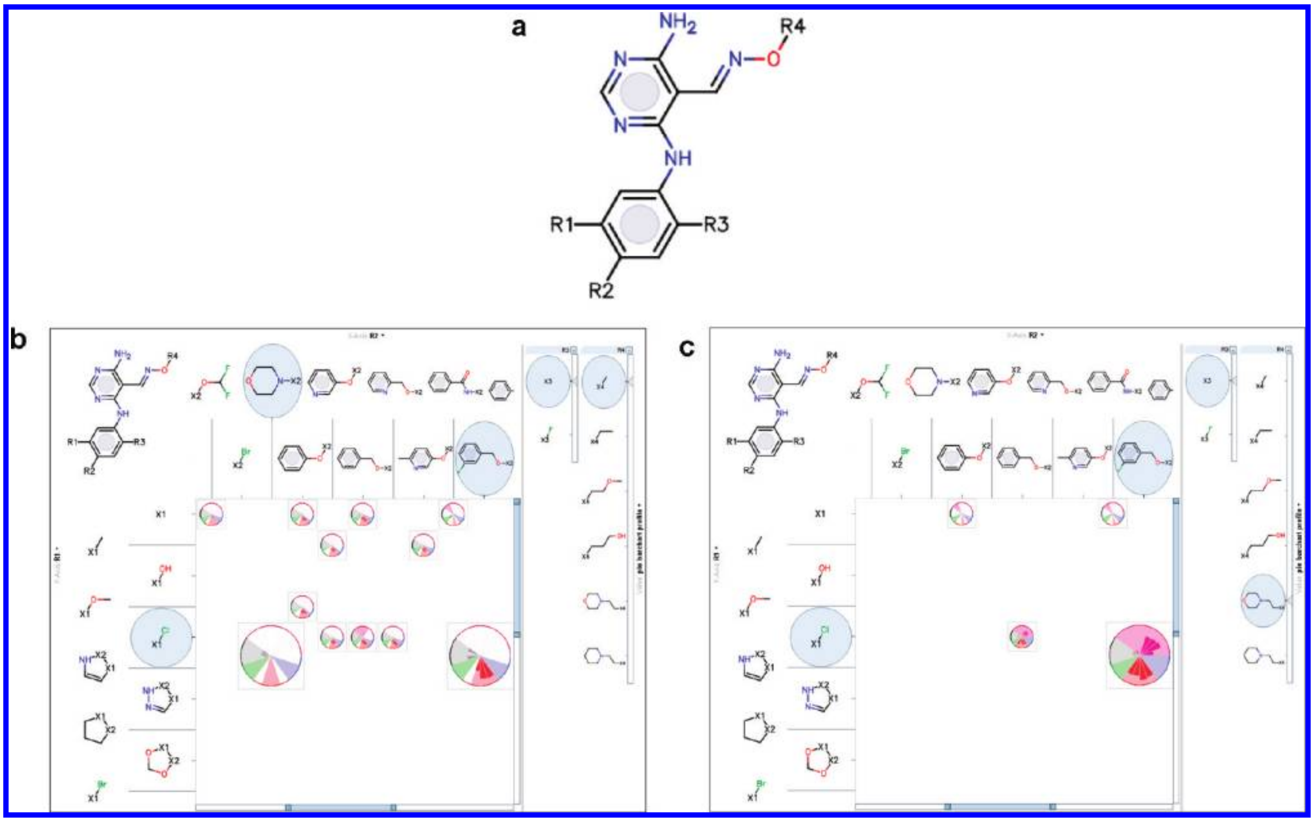


Figure 7. a. Alternative core structure used for R-group enumeration and generation of a 4D SAR map. b. Composite screenshot of the zoomed region of a 4D SAR map, R4 = methyl. c. Zoomed region of a 4D SAR map, R4 = 2-(morpholin-4-yl)ethyl.

properties of the oximes and therefore shows a "leaner" collection of molecules than the view where R4 is methyl (lead-generation stage molecules). On the other hand, many analogs in the late-stage collection were more completely characterized in the *in vitro* kinase and cellular assays, as shown by their completely filled pie barcharts.

The 4,6-diamino-pyrimidine-5-carbaldehyde oximes of the ErbB data set constitute an interesting collection of molecules that allowed us to demonstrate some of the new features of the second-generation, enhanced SAR map viewer. However, screenshots and other static representations of SAR maps do not adequately illustrate its true capabilities as a powerful, highly interactive data analysis tool. Numerous other data renderers can be displayed in the grid cells, and more complicated multidimensional SAR maps of five or more R-group analyses can be easily and clearly configured on the grid. Nevertheless, with this modest ErbB data set, we have shown that the second-generation SAR map viewer offers a wealth of new analytical capabilities to our medicinal chemistry and molecular modeling community.

CONCLUSIONS

The enhanced SAR maps bring an almost infinite gamut of data rendering options to a popular medicinal chemistry tool. By combining the power of R-group analysis with the ability to display any conceivable object inside the cells, they can deliver rich visualizations in the most familiar context to a medicinal chemist. We hope that this publication will help catalyze additional advances in the area of chemically aware data visualization.

Abbreviations. SAR, structure−activity relationships; UI, user interface; CDK1, cyclin-dependent kinase-1; EGFR, epidermal growth factor receptor; VEGFR2, vascular endothelial growth factor receptor-2; KDR, kinase insert domain receptor; RTK, receptor tyrosine kinase, BPH, benign prostatic hyperplasia; LUTS, lower urinary tract symptoms; SOM, self-organizing map; SPE, stochastic proximity embedding; ABCD, advanced biological and chemical discovery; 3DX, third dimension explorer; MCS, maximum common substructure; AGG, antigrain geometry.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(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) Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM T. Graphic.* **1992**, *11* (1), 92–99.

(7) 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.

(8) 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.

(9) 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 07−11, 1995, ACM Press/Addison-Wesley: New York, NY, 1995; pp 401−408.

(10) Agrafiotis, D. K.; Bandyopadhyay, D.; Farnum, M. Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.* **2007**, *47*, 69–75.

(11) Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *C-18* (5), 401–409.

(12) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869–15872.

(13) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215–1221.

(14) Agrafiotis, D. K.; Xu, H. A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 475–484.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) *ClassPharmer Suite,version 3.2−3.5*; Bioreason, Inc.: Santa Fe, NM.

(20) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR maps: A new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **2007**, *50* (24), 5926–5936.

(21) *DIVA 2.1 Software*; Accelrys Inc.: San Diego, CA. http://www.accelrys.com (accessed September 2009).

(22) *Accord for Excel 2.0 Software*; Accelrys Inc.: San Diego, CA. http://www.accelrys.com (accessed September 2009).

(23) *STN Express 8.4 Software*; Chemical Abstracts Service: http://www.cas.org (accessed September 2009).

(24) Agrafiotis, D. K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E. P.; Konstant, P.; Leung, A.; Lobanov, V. S.; Marichal, P.; Martin, D.; Rassokhin, D. N.; Shemanarev, M.; Skalkin, A.; Stong, J.; Tabruyn, T.; Vermeiren, M.; Wan, J.; Xu, X. Y.; Yao, X. Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model.* **2007**, *47* (6), 1999–2014.

(25) Shemanarev, M. The Anti-Grain Geometry Project. http://www.antigrain.com (accessed June 2006).

(26) Seymour, L. Epidermal growth factor receptor inhibitors: An update on their development as cancer therapeutics. *Curr. Opin. Invest. Drugs* **2003**, *4*, 658–666.

(27) Huang, S.; Li, R.; Connolly, P. J.; Xu, G.; Gaul, M. D.; Emanuel, S. L.; LaMontagne, K. R.; Greenberger, L. M. Synthesis and biological study of 4-aminopyrimidine-5-carboxaldehyde oximes as antiproliferative VEGFR-2 inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 6063–6066.

(28) Gaul, M. D.; Xu, G.; Kirkpatrick, J.; Ott, H.; Baumann, C. A. 4-Amino-6-piperazin-1-yl-pyrimidine-5-carbaldehyde oximes as potent FLT-3 inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4861–4865.

(29) Xu, G.; Searle, L. L.; Hughes, T. V.; Beck, A. K.; Connolly, P. J.; Abad, M. C.; Neeper, M. P.; Struble, G. T.; Springer, B. A.; Emanuel, S. L.; Gruninger, R. H.; Pandey, N.; Adams, M.; Pandey, N.; Fuentes-Pesquera, A. R.; Middleton, S. A.; Greenberger, L. M. Discovery of novel 4-amino-6-arylaminopyrimidine-5-carbaldehyde oximes as dual inhibitors of EGFR and ErbB-2 protein tyrosine kinases. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3495–3499.

(30) Emanuel, S. L.; Hughes, T. V.; Adams, M.; Rugg, C. A.; Fuentes-Pesquera, A. R.; Connolly, P. J.; Pandey, N.; Pandey, N.; Butler, J.; Borowski, V.; Middleton, S. A.; Gruninger, R. H.; Story, J. R.; Napier, C.; Hollister, B; Greenberger, L. M. Cellular and in vivo activity of JNJ-28871063: a non-quinazoline pan-ErbB kinase inhibitor that crosses the blood brain barrier and displays efficacy against intracranial tumors. *Mol. Pharm.* **2007**, *73*, 338–348.

(31) 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.

(32) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.

CI900264N