5002 J. Med. Chem. 2010, 53, 5002–5011
DOI: 10.1021/jm1004495

Journal of Medicinal Chemistry Article

Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes

Dimitris K. Agrafiotis*,† and John J. M. Wiener‡

†Johnson & Johnson Pharmaceutical Research & Development, LLC, Welsh & McKean Roads, Spring House, Pennsylvania 19477, and ‡Johnson & Johnson Pharmaceutical Research & Development, LLC, 3210 Merryfield Road, San Diego, California 92121

Received April 13, 2010

We introduce Scaffold Explorer, an interactive tool that allows medicinal chemists to define hierarchies of chemical scaffolds and use them to explore their project data. Scaffold Explorer allows the user to construct a tree, where each node corresponds to a specific scaffold. Each node can have multiple children, each of which represents a more refined substructure relative to its parent node. Once the tree is defined, it can be mapped onto any collection of compounds and be used as a navigational tool to explore structure−activity relationships (SAR) across different chemotypes. The rich visual analytics of Scaffold Explorer afford the user a "bird's-eye" view of the chemical space spanned by a particular data set, map any physicochemical property or biological activity of interest onto the individual scaffold nodes, serve as an aggregator for the properties of the compounds represented by these nodes, and quickly distinguish promising chemotypes from less interesting or problematic ones. Unlike previous approaches, which focused on automated extraction and classification of scaffolds, the utility of the new tool rests on its interactivity and ability to accommodate the medicinal chemists' intuition by allowing the use of arbitrary substructures containing variable atoms, bonds, and/or substituents such as those employed in substructure search.

Introduction

Chemical Scaffolds. The concept of a chemical scaffold is probably as old as medicinal chemistry itself. Scaffolds are structural cores upon which different types of substituents can be attached. In lead generation, scaffolds are embellished through parallel synthesis and combinatorial chemistry to produce large chemical libraries for high-throughput screening (HTS). In lead optimization, the same strategy is applied on a smaller scale to improve an initial lead by optimizing individual substitution sites in an iterative (and often recursive) manner through successive rounds of synthesis and biological testing. This process is repeated until the desired potency, selectivity, or pharmacokinetic parameters are achieved or until the potential of the series is exhausted. In the latter case, new chemical scaffolds are designed (often as variations of the old ones through a process known as scaffold hopping[1]), new sets of analogues are synthesized, and the cycle continues until a promising clinical candidate emerges or the program is terminated. In practice, medicinal chemistry teams tend to pursue multiple series concurrently in order to maximize utilization of resources and the probability of success.

Computer-based methods can be particularly effective in analyzing and categorizing molecular graphs and have been used extensively for identifying common scaffolds in large collections of molecules. The most common application of automated methods for scaffold classification is in assessing the diversity of large chemical libraries and in analyzing hit lists from high throughput screening experiments. These hits tend to mirror the source libraries from which they emerge, in that they span multiple chemotypes and are highly heterogeneous in terms of structure, physicochemical properties, and synthetic origin. Grouping them into related families not only assists in hit confirmation and follow-up but also allows more predictive local statistical models to be developed and employed in guiding future rounds of analogue design.[2−4]

However, the ways scaffolds are understood by medicinal chemists and computers are not always congruent. To a medicinal chemist, a scaffold is a central core with clearly defined substitution sites. To a computer, it is often an abstract connectivity pattern that is present in a large group of molecules. This is due to the fact that most automated scaffold classification algorithms employ some form of clustering driven by descriptors and similarity measures that look at abstract topological patterns and do not explicitly consider the presence of structural cores with well-defined variation patterns.[5] While clustering methods afford certain advantages, they are often misguided by idiosyncratic patterns in molecular graphs and produce groupings that look "unnatural" to a medicinal chemist. Their most important limitation is that the partitioning depends on the specific data set that is being clustered, the membership rules are not easily interpretable, and the resulting clusters do not represent real equivalence classes with unambiguous rules of membership that are applicable to compounds beyond the original training set. This makes the key determinants of biological activity difficult to identify and even more difficult to exploit in the design of improved analogues.

Scaffold Extraction/Visualization Tools. To address these problems, several groups, most notably Nicolaou et al.[6,7]

*To whom correspondence should be addressed. Phone: (215) 628-6814. Fax: (215) 540-4619. E-mail: dagrafio@its.jnj.com.

 Published on Web 06/04/2010

and Inglese et al.,[8,9] have attempted to generate interpretable structural motifs by identifying the maximum common substructure of the molecules in each cluster in a separate postprocessing step. However, because of the computational cost of MCS,[a] this method is limited to relatively small clusters and is sensitive to "outliers" which happen to fall into a particular cluster because of their overall structure but do not necessarily share a common scaffold with the other molecules in that cluster.

While scaffolds are not required by definition to contain rings, in practice most of them do. By reducing conformational flexibility, rings allow more precise positioning and orientation of key pharmacophoric groups and a more detailed mapping of the active site. The first systematic analysis of ring structures present in drug molecules was reported by Bemis and Murcko,[10] who decomposed molecules into ring systems, linkers, side chains, and frameworks and looked at their relative frequency of occurrence in the CMC database using various shape descriptors. The authors defined ring systems as single or contiguous cycles that share at least one edge, linkers as the paths connecting two ring systems, side chains as the paths that are not part of a ring system or a linker, and frameworks as the networks of ring systems and linkers present in the molecule. Their analysis was carried out at two levels of abstraction, one in which only pure topology was considered (i.e., unlabeled nodes and edges), and one where atom type, hybridization, and bond order were also taken into account. Lewell and coworkers later reported a web-searchable database of rings that appear in drugs and demonstrated its utility in identifying alternative chemical rings for scaffold hopping.[11] Both of these methods produced flat categorizations of ring scaffolds as opposed to navigable hierarchies.

Hierarchical classifications of ring systems have been reported by several authors, including Franco et al.,[12] Schuffenhauer et al.,[13,14] Wetzel et al.,[15] and Wilkens et al.[16] In Franco's approach, which is based on the molecular equivalence work of Xu and Johnson,[17] the levels of the hierarchy represent different degrees of structural abstraction obtained by iterative simplification of the parent molecule, followed by canonicalization of the resulting structures. Four levels were defined: exact molecules which sit at the leaves of the tree, cyclic systems which are derived from exact molecules through removal of side chains, cyclic system skeletons which are obtained from cyclic systems by removing atom and bond type information (i.e., by setting all atoms to carbon and all bonds to single), and reduced cyclic system skeletons which are obtained from cyclic system skeletons by deleting all atoms attached to two other atoms. 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.

In Schuffenhauer's scaffold tree,[13] each node represents a different ring system and the different levels of the hierarchy are traversed through iterative removal of rings until a single root ring is obtained. Just like any other ring-based method, the initial set of rings is obtained by removing all side chains from the target molecule. The resulting ring systems are divided into smaller and smaller scaffolds by removing one ring at a time using a set of prioritization rules, which ensure that the order in which rings are removed is unambiguous and deterministic. In general, central and more complex rings are retained over peripheral simpler ones. The practical implication of this approach is that any given ring scaffold has a unique path leading up to it (i.e., it has only one parent). Although this helps minimize tree size and complexity, it complicates navigation, as the user needs to remember the rules or rely on chemical searching to locate the scaffold of interest. Furthermore, because the activity of a compound can only be assigned to a single parent scaffold even though multiple core scaffolds may be present in the same molecule, the exploration and interpretation of SAR can be more challenging.

This idea is further elaborated in Wetzel's Scaffold Hunter,[15] a hierarchical tree-like representation of a set of compounds constructed through automated fragmentation of the original data set. The program determines which molecules represent relevant, complex scaffolds and then iteratively deconstructs those scaffolds one ring at a time to create more general scaffolds. The resulting tree can be associated with potency data. Notably, in the construction of this tree, intermediate "virtual" scaffolds that are not represented by actual compounds in the original data set but that may hold promise as structurally simpler avenues toward comparable potency can be readily identified.

Wilkens et al[16] did not impose a single parent constraint but truncated the complexity of the tree by removing scaffolds that occurred only once. Their method proceeds by identifying the base ring systems (contiguous cyclic subgraphs obtained when all linkers and side chains are removed), recursively enumerating all their possible combinations, and arranging them in a hierarchical manner. Unlike Schuffenhauer's method, no attempt was made to prune more complex fused ring systems into smaller units.

An alternative approach specifically tailored to combinatorial libraries was described by Katritzky.[18] Here, scaffolds were defined as invariant molecular frameworks that are common in a large portion of the library, and a set of rules was proposed to determine when two scaffolds should be considered equivalent. The distinguishing element of this scheme is that the scaffold is not defined in isolation but only in reference to the other molecules in the library. The same substructure in the same molecule may constitute a scaffold if that molecule is part of library A and an appendage in library B. The assignment depends on what other variations of that structure are present in the library under consideration. The classification scheme takes into account molecular shapes, molecular pharmacophores, substituent orientation, and substituent diversity, and designates scaffolds as equivalent when the summed scores of the transformations that are needed to convert one to another does not exceed a certain threshold.

The recently reported open-source application SARANEA[19] employs a "network-like similarity graph" for visualization and analysis of structure−activity/selectivity relationships in variously sized compound data sets. In this approach, compounds (each represented by a single node in the graph) are graphically connected if their computationally determined similarity score is above a set threshold. Resulting clusters of highly similar nodes are grouped together and separated from other clusters. Potency is displayed using color-coding of nodes and, additionally, the extent to which each compound represents a discontinuity in SAR is shown by the size of that compound's node.

[a]Abbreviations: SAR, structure−activity relationships; CatS, cathepsin S; ABCD, advanced biological and chemical discovery; 3DX, third dimension explorer; MCS, maximum common substructure; UI, user interface.

The problem with the aforementioned techniques is that the scaffolds are automatically extracted from the data set by algorithmic means, typically through iterative pruning, maximum common substructure (MCS), or some variation thereof. While this approach may be suitable for analyzing large chemical libraries of highly heterogeneous compounds and for triaging hit lists from high throughput screening campaigns, it leaves much to be desired in terms of organizing SAR data for the hit-to-lead and lead optimization phases of a discovery project. In most drug discovery projects, scaffolds may be cyclic or acyclic, may contain complete or partial appendages, and may include variable atoms and bonds or larger substructures (e.g., atom X is either a nitrogen or an oxygen; bond Z is double or aromatic, substituent Z is a 5- or 6-membered aromatic ring, etc). Indeed, the most general definition of a scaffold is a substructure, such as those employed for chemical database searching, shared by a collection of molecules.

Scaffold Explorer. In this paper, we describe an interactive tool called Scaffold Explorer that allows the user to construct interactively any hierarchy of scaffolds, where each scaffold represents any arbitrary substructure with variable atoms, bonds, and/or substituents. Although the tree can be populated through automated scaffold extraction algorithms, its true power comes from allowing the users to define the nature and hierarchy of the scaffolds themselves in a way that mirrors their lead optimization strategy. The substructures associated with each scaffold can be recursively elaborated into increasingly refined substructures, representing deeper nodes in the tree. This tree representation was designed specifically to mimic the iterative manner in which medicinal chemists optimize compounds and analyze SAR data. This process involves exploring different core scaffolds, different classes of substituents at each point of variation around these scaffolds, different classes of substituents around those substituents, etc., until the series is thought to be exhausted or at least explored to a reasonable degree. This concept is illustrated by the cathepsin S (CatS) inhibitor program discussed below. For a typical discovery project, the resulting hierarchies tend to comprise a few tens of scaffolds with 3−5 levels of depth at most. Therefore, the manual effort required to build and maintain these hierarchies is very small and well worth the flexibility that it affords.

The Scaffold Explorer offers a rich set of data rendering options that allow the user to obtain a "bird's-eye" view of the entire chemical space spanned by a particular data set, identify the relative population of each scaffold class, map any physicochemical property or biological activity of interest onto the individual scaffold nodes, serve as an aggregator for the properties of the compounds in each of these nodes, and quickly distinguish promising chemotypes from less interesting or problematic ones. The tool can be dynamically connected to any SAR table and serve as an effective navigational tool through linked selections and visualizations (vide infra). Scaffold Explorer is particularly useful in conjunction with the recently described SAR maps,[20,21] which provide more detailed views of the substituent effects around each individual scaffold and can be very effective in driving SAR discussions at project team meetings. In the remaining paragraphs, we describe its core features and demonstrate its practical utility with an internal drug discovery project aimed at designing inhibitors for CatS.

Methods

Third Dimension Explorer (3DX) and ABCD. The Scaffold Explorer was implemented as a component of 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 platform known as ABCD,[22] 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, SAR maps, and the scaffold viewer described herein. The program offers a full gamut 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 as needed. 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, etc. 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, 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.

Scaffold Explorer. The Scaffold Explorer is essentially an editor that allows the user to define a scaffold tree and dynamically "connect" it to an SAR table. A scaffold tree is a singly rooted, acyclic graph, where each node has one parent (except the root) and zero or more children. A node with no children is referred to as a leaf or terminal node. The root node is always present and cannot be deleted and is marked with a red outline (Figure 1). Every node can be associated with a chemical substructure, selected by double-clicking on the node and drawing the pattern in the popup sketcher. If no structure is drawn, the node is referred to as a "pass-through" node and includes all the molecules contained in its parent node. Child nodes represent scaffolds with more refined substructures than their parent nodes, i.e., any specific molecule that contains the substructure of a child node must also contain the substructure of its parent node. Unlike automated scaffold extraction algorithms where each scaffold is typically an exact molecule, the scaffolds in Scaffold Explorer represent fully fledged chemical substructures that can contain generic (query) atoms and bonds like those employed in a typical substructure search (see Discussion).

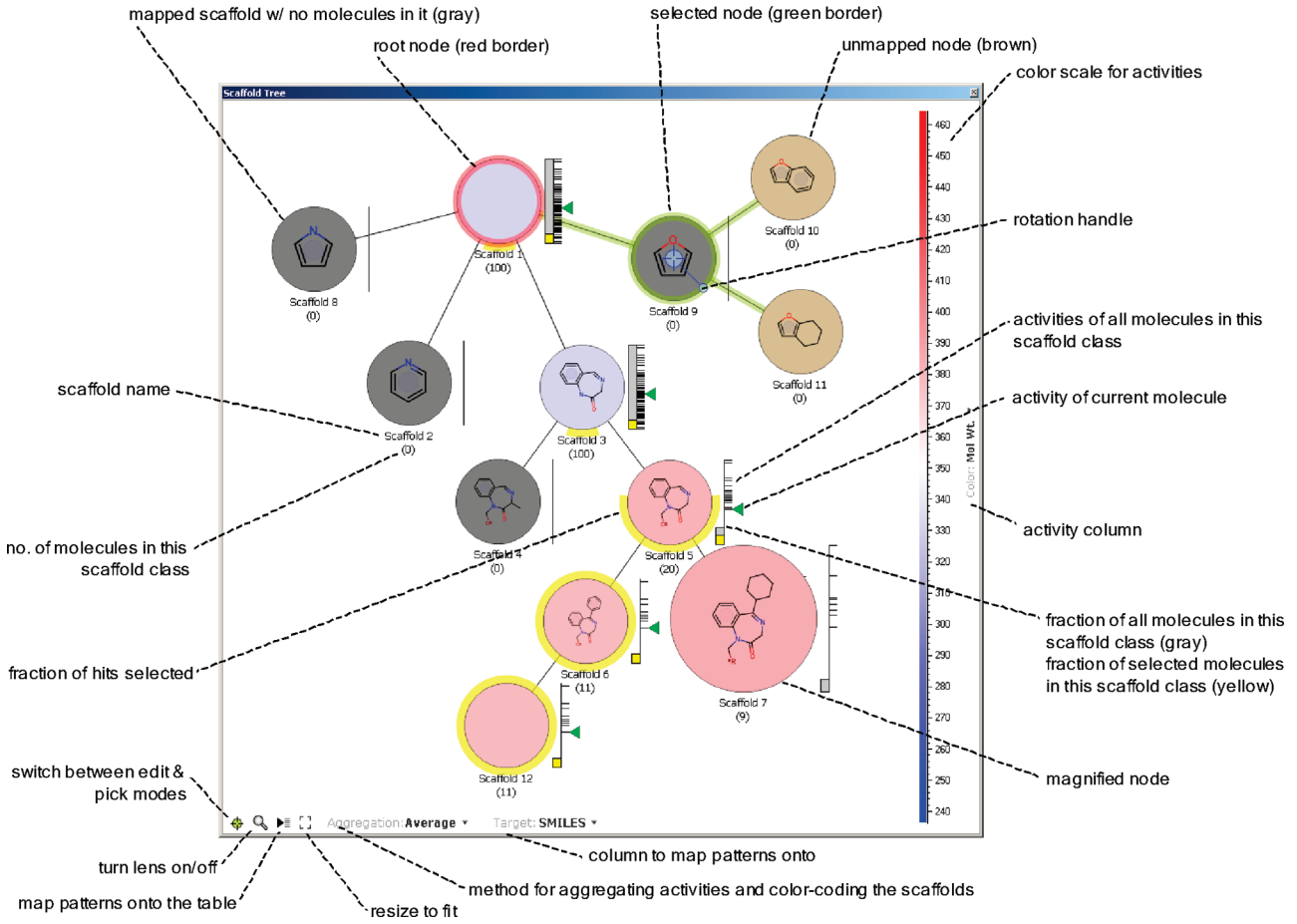

Figure 1. User interface elements of the Scaffold Explorer.

User Interface. The scaffold tree operates in two modes: (1) edit mode, where the tree is constructed and arranged on the drawing window, and (2) select mode, where nodes are used to select/deselect the molecules in the current 3DX table that contain the corresponding substructures. The user can switch between the edit and pick modes by pressing the escape key or clicking on the "mode" button on the bottom left of the viewer. By default, the scaffold tree starts in the edit mode, with a single empty root at the center.

New child nodes are created in edit mode by clicking the left mouse button on an existing node (the parent) while pressing a key modifier and dragging the new child node to its desired position on the canvas. To minimize the drawing effort, the child node automatically inherits the substructure of its parent, which can be edited by double-clicking on the new node, as described above. While in edit mode, nodes can be selected and moved anywhere on the drawing canvas either individually, as an entire branch, or as any arbitrary subset of nodes selected by a lasso. When a nonterminal node is selected, a transient rotation handle appears centered at that node, which allows the user to rotate that node and all of its children around the chosen reference point. A number of additional editing functions are supported, including panning, zooming, and adjustment of the nodes' radii.

As soon as a node is edited using the sketcher, its color (and that of all its descendants) turns to brown to indicate that they are unmapped. Mapping is the process of associating the substructures in the scaffold tree with the molecules in the current 3DX table and identifying which molecules fall into each scaffold class. This mapping is initiated using an explicit command in the toolbar or the context-sensitive menu. When the scaffolds are mapped onto the current table, only the molecules contained in the parent node are used as input for the child nodes, thus significantly truncating the time required for substructure searching. Once the tree is mapped, the nodes are color-coded based on the aggregate activity of the compounds contained within them (vide infra) and a vertical scale appears to the right of each node. This vertical scale serves two purposes. The first is to indicate the relative number of records that belong to each scaffold class, which is indicated by the height of the vertical gray bar on the left of the scale (if any of these records also happen to be selected, they will be indicated by an additional yellow bar overlaid with the gray one). The second is to visualize the activities of the compounds that belong to that scaffold class, which are indicated by individual tick marks on the right of the scale. The activities can be read from any column in the current table, which is selected using the *Color* dropdown box on the right of the color scale. The minimum and maximum of the activity scales of the individual scaffolds are all the same and are controlled by the large color scale to the right of the tree. Adjusting the color scale also adjusts the individual activity scales on each of the tree nodes.

When mapped, each scaffold node contains a subset of molecules from the current table and can thus serve as an aggregator of these molecules. If an activity column is selected, the background color of each node will be determined by the aggregate activity of all the molecules that fall into that scaffold class, using the color mapping of the large color scale on the right-hand side of the viewer (blue to white to red). Four aggregation functions are supported (minimum, maximum, average, and median) and can be interactively changed using the *Aggregation* dropdown box on the lower left corner of the viewer. If no molecules fall under a particular scaffold class, the corresponding node is colored gray.

The activity of the current molecule (the molecule that the user last clicked on in order to inspect its contents) is also indicated with a green triangle to the right of the activity tick

marks on each scaffold node. Note that the same molecule may belong to multiple scaffold classes, and there will be a green arrow for each class where the current molecule is mapped.

Additional features are available through the context-sensitive menu, accessible by right-clicking over the drawing canvas and/or a node. The user interface is illustrated in Figure 1, and the context-sensitive menu commands, mouse controls, and keyboard shortcuts are detailed in Tables 1−3.

Implementation. The Scaffold Explorer was implemented as a .Net control and was written in C# using the GDI+ graphics library available in .Net.

Discussion

In general, a data set with a large number of unique compounds presents the analyst with a significant challenge. These sets are likely to include an array of series and subseries, each with its own biological profile and issues. As discussed, meaningful visualization of such large numbers of compounds can be difficult—a simple tabular representation is far from sufficient given the number of substituents that are variable, not to mention the numbers of substituents on those substituents that are, in turn, variable. We demonstrate the utility of Scaffold Explorer in providing readily viewable and information-rich hierarchical structure analyses of the subclasses within a parent chemical series with a case study using data drawn from a CatS inhibitor program.

CatS Inhibitor Program. CatS is a cysteine protease that mediates cleavage of the major histocompatibility class II (MHC II)-associated invariant chain (Ii), one step in the sequence of events leading to antigen presentation on the cell surface and thus a key constituent of an immune response.[23−27] For this reason, CatS inhibitors have been proposed for treatment of various autoimmune disorders as well as other diseases. Inhibitors of CatS are often covalent-binding active site modifiers, although recently noncovalent inhibitors have been disclosed.[28−35] Crystal structure analyses of human CatS enzyme have revealed several relevant binding pockets (known as S1−S5), and the regions of the molecules that occupy these portions of the enzyme are correspondingly termed P1−P5.[36−39]

Analysis began with a search of the ABCD data warehouse,[22] retrieving P2 pyrazole structures for which human CatS enzymatic binding data (hCats pIC_{50}) had been generated. For the purpose of this discussion, we will focus our attention on a subset of 1294 unique structures that were tested in the assay. These molecules, which were specifically chosen to illustrate the full capabilities of the tool, fall into three main series related to the nature of the substituent on the right-hand side of the pharmacophore (P1 and P3 binding regions): amines, thioethers, and alkynes (Figure 2), with the variable positions defined as described in relevant patent applications.[31−33]

Table 1. Context-Sensitive (Popup) Menu

menu item	function
open	open scaffold tree from a binary file
save as...	save scaffold tree to a binary file
map	map any unmapped nodes onto the current document
fit	rescale tree to fit in the visible window
lens	turn magnification lens on/off
select	select node under the mouse
unselect	unselect node under the mouse
select all	select all nodes in the tree
clear selection	clear selection
rename...	rename the clicked node
delete	delete clicked node and connect its children to its parent
delete branch	delete clicked node along with all its children
delete all	remove all nodes exceept the root (and clear the root)
cut branch	cut the subtree under the clicked node and copy it onto the clipboard
copy branch	copy the subtree under the clicked node onto the clipboard
paste branch	paste the subtree on the clipboard as a child of the clicked node
paste branch as new	paste the subtree on the clipboard as a new tree
copy to clipboard	copy scaffold tree to the clipboard as an image
print...	print the tree; provides page setup, print preview, and print options.

Table 3. Summary of Keyboard Shortcuts

shortcut	function
escape	switch between edit and pick modes
control-A	select all
control-shift-A	unselect all
control-C	copy image to clipboard
control-R	reset (fit to window)
control-D	delete all nodes except the root (and clear the root)
control-M	map scaffolds onto current table
control-L	turn magnifying lens on/off
plus	zoom in
minus	zoom out
shift-plus	zoom in by a larger increment
shift-minus	zoom out by a larger increment
arrow	move selected nodes left/right/up/down
shift-arrow	move selected nodes left/right/up/down by a larger increment
control-arrow	move entire tree left/right/up/down
control-shift-arrow	move entire tree left/right/up/down by a larger increment

Table 2. Summary of Mouse Controls

action	over node	mode	function
right down	yes|no	edit|pick	display context-sensitive menu
control + left down	yes|no	edit|pick	move the entire tree
left down	yes	edit	highlight node
left double click	yes	edit	edit scaffold structure
shift + left down	yes	edit	highlight node and all of its children
left down + move	yes	edit	move highlighted nodes
left down + move	no	edit	draw selection lasso
alt + left down	yes	edit	create new child node
left down	yes	pick	select
shift + left down	yes	pick	unselect

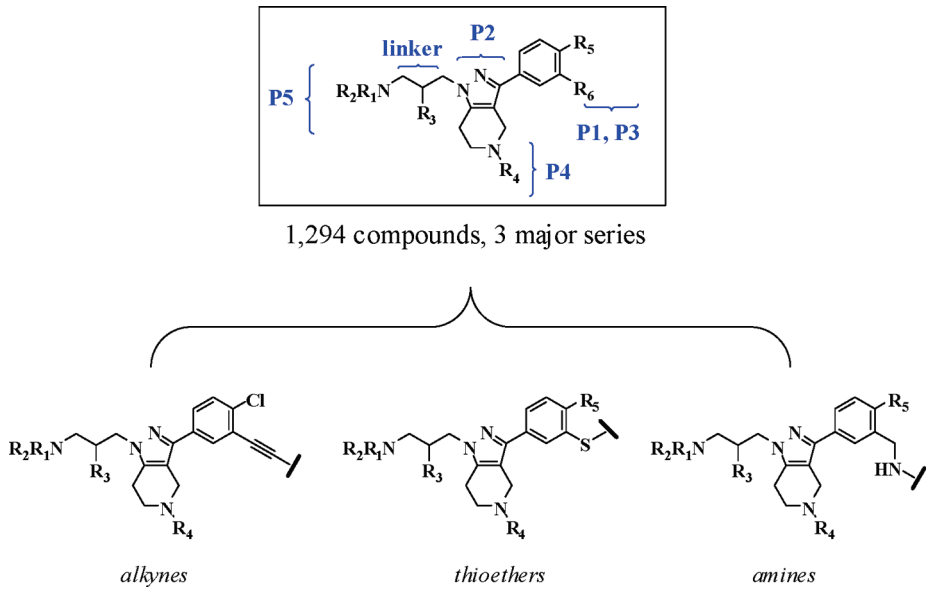

Figure 2. General structure of cathepsin S inhibitors and major series.

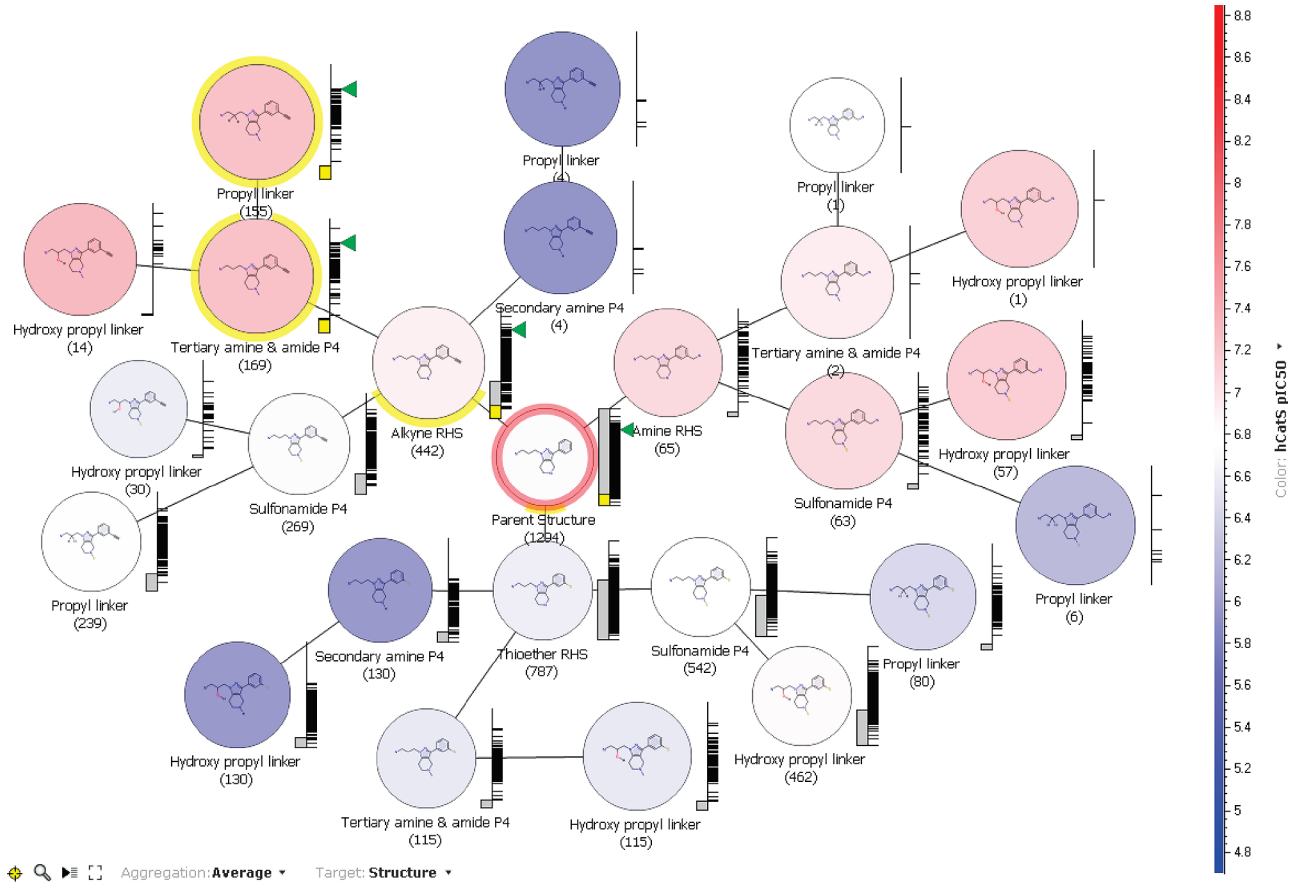

Figure 3. Scaffold tree representation of cathepsin S inhibitor series and subseries, color-coded by average hCatS inhibition (pIC_{50}).

Application of Scaffold Explorer to CatS Inhibitor Program. Using Scaffold Explorer, from the parent structure shown at the center of the diagram in Figure 3 (the root node), emanate three child nodes representing each of these three series as different scaffolds. Mapping the data set onto the tree reveals the population of each scaffold (indicated as a number in parentheses below each node), from which is immediately evident the relative paucity of amines within this data set (65 amine compounds, compared to 442 alkynes and 787 thioethers). Color-coding the nodes according to average hCatS pIC_{50} values allows facile analysis of the aggregate potency of the compounds within these three scaffolds, with the scaffolds containing the most potent compounds shown in red and those with the least potent compounds shown in blue. At this highest level of analysis, it is clear that all three of these scaffolds represent compounds

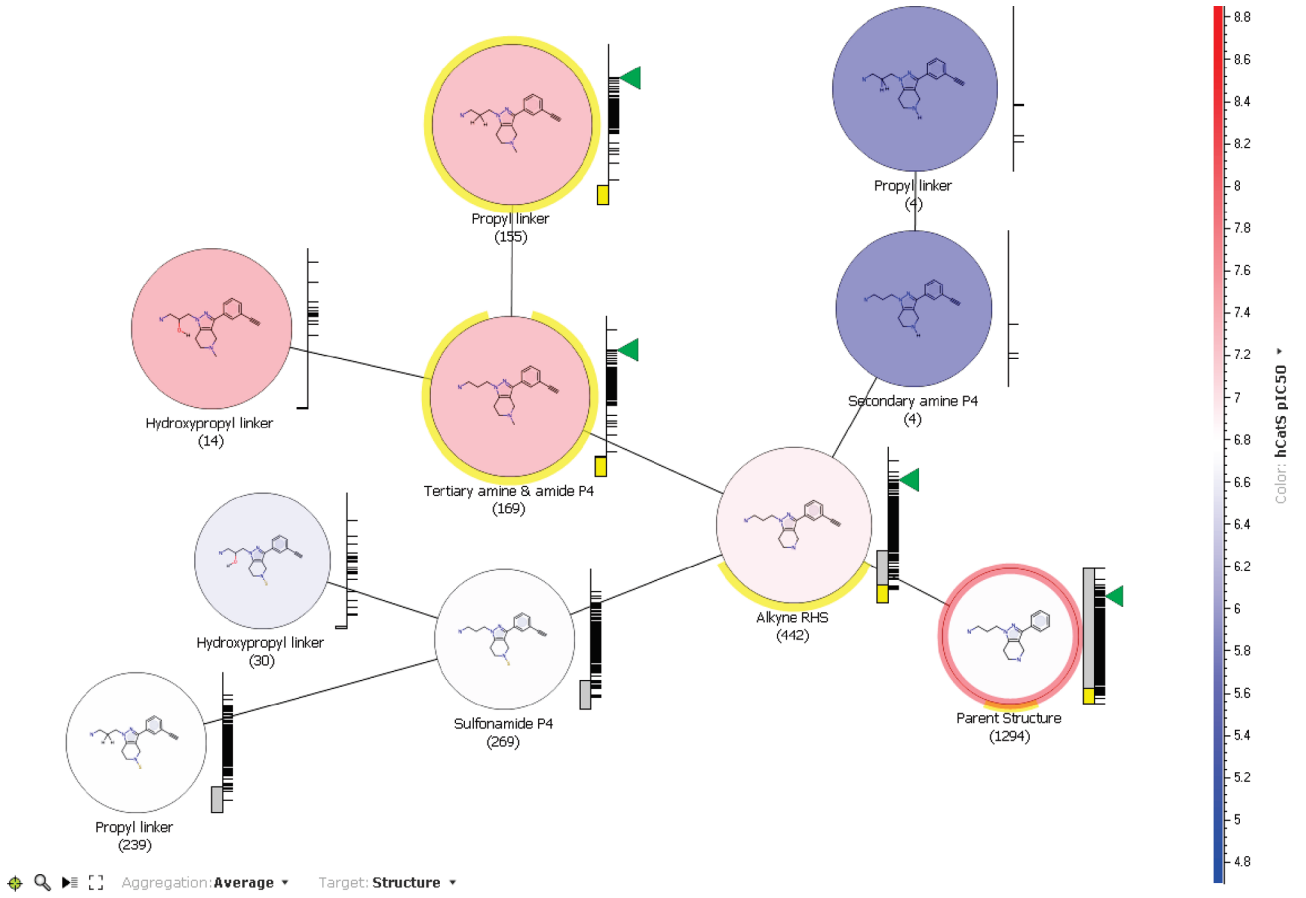

Figure 4. Scaffold tree representation of cathepsin S inhibitor alkyne subseries, color-coded by average hCatS inhibition (pIC_{50}).

with comparable potency, with average pIC_{50}s in the range of 6.8 to 7.2. The vertical scale to the right of each node provides two types of information: the gray bar indicates what fraction of total compounds fall under that particular node, and the black tick-marks indicate the stratification of biological data across the compounds constituting each node (the scale is identical for each node and is the same as the vertical color scale on the right side of the plot). As might be expected for a large data set, each of the alkyne and thioether series contain a large number of molecules with midrange potency and a substantially smaller number of high-potency analogues.

To establish a more refined understanding of SAR, this approach of creating new child nodes representing more specific substructures and mapping data onto them is performed iteratively, thereby creating an increasingly complex structural hierarchy, as shown in Figure 3. Among the relevant scaffolds represented by nodes in this tree are those involving variation of the P4 and P5 substituents and alteration to the linker connecting the P5 substituent to the P2 pyrazole core. Examination of the alkyne series in more detail using a magnified portion of this scaffold tree is informative (Figure 4). Again using differential coloration to explore variation in potency among the compounds within the scaffolds, one terminal node within the alkyne series (the "propyl linker" terminal node highlighted in yellow, deriving from the "tertiary amine & amide P4" node) contains a large number of molecules (155) with promising potency (average hCatS $pIC_{50} \sim 7.2$) and predominant clustering of compounds toward the high end of the potency range. For a more detailed understanding of the SAR of the compounds constituting this terminal node, an SAR map[20,21] is useful (Figure 5). Pairing a scaffold tree with SAR maps can provide elegant and complementary visualization of high-level SAR trends and population densities for many scaffolds as well as specific, detailed SAR within terminal nodes.

The two-dimensional matrix in Figure 5 displays the various P5 substituents along the vertical axis and the P4 substituents along the horizontal axis, with the slider to the right indicating that the P1 substituent is held constant as a *para*-chlorobenzylamine; hCatS pIC_{50} values of compounds are displayed as colored rectangles at the intersection points of these substituents. The P4 and P5 substituents are sorted according to molecular weight, with low-molecular-weight substituents at the left and top of the axes. These data indicate a general trend toward increasing potency with increasing molecular weight of the P5 substituent, an unsurprising feature given the lack of specific interactions proposed to be relevant in the S5 region of the enzyme.[34] More interesting is the compound highlighted in Figure 5 containing a urea P4 substituent and a simple piperidine P5 moiety. These low-molecular-weight fragments contribute to a molecule with surprisingly high potency (hCatS $pIC_{50} = 7.3$). This type of nonadditive SAR is of particular interest in selecting compounds for advanced profiling, based on the desire to identify molecules which might involve a higher degree of specific interactions, allowing them to achieve more optimal physical properties by limiting overall molecular weight and thus enabling further studies with related structures.

Closer examination of the thioether portion of the scaffold tree (Figure 6) reveals a cluster of 115 thioether compounds (the "hydroxyl propyl linker" terminal node highlighted in

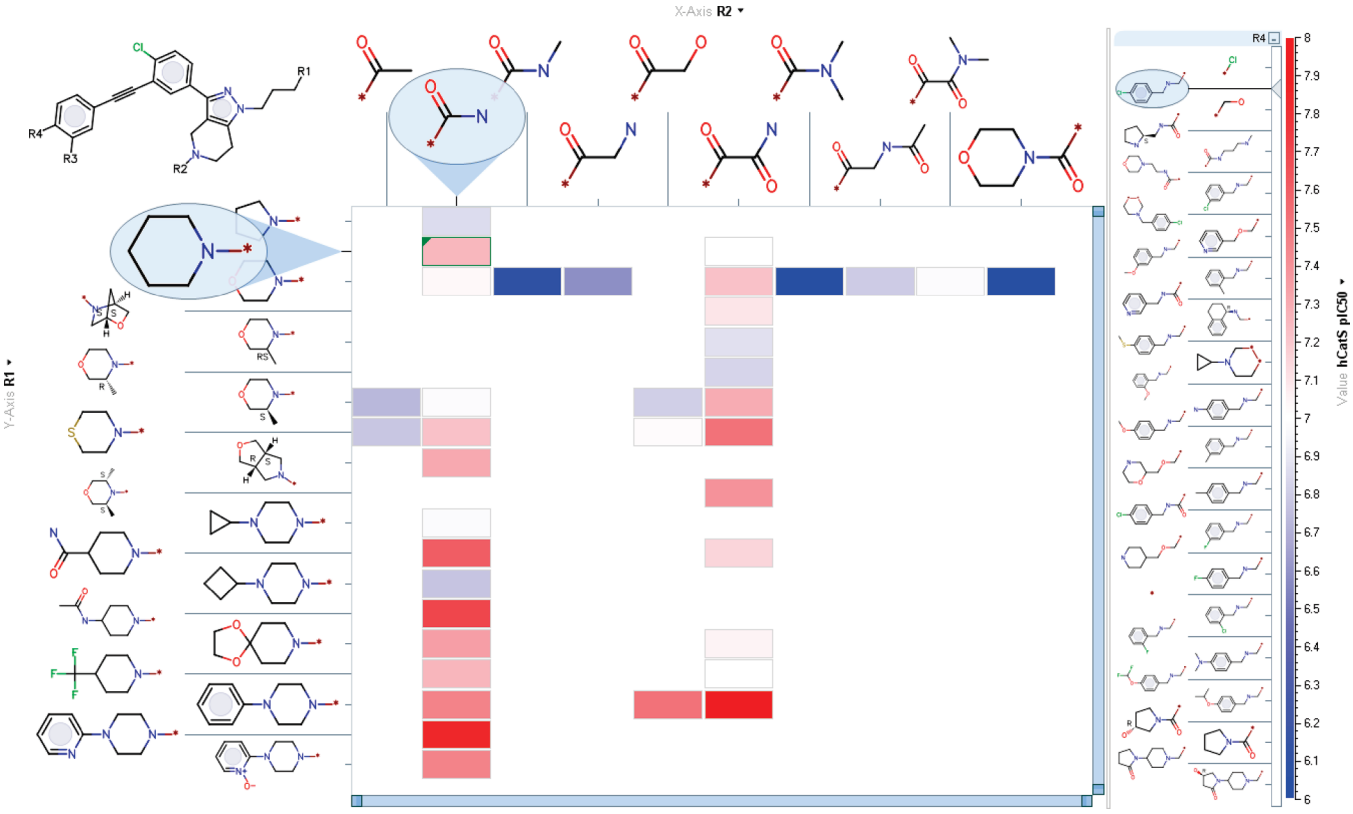

Figure 5. hCatS inhibition (pIC_{50}) of alkynes, sorted by R1 and R2 molecular weight, then by hCatS potency.

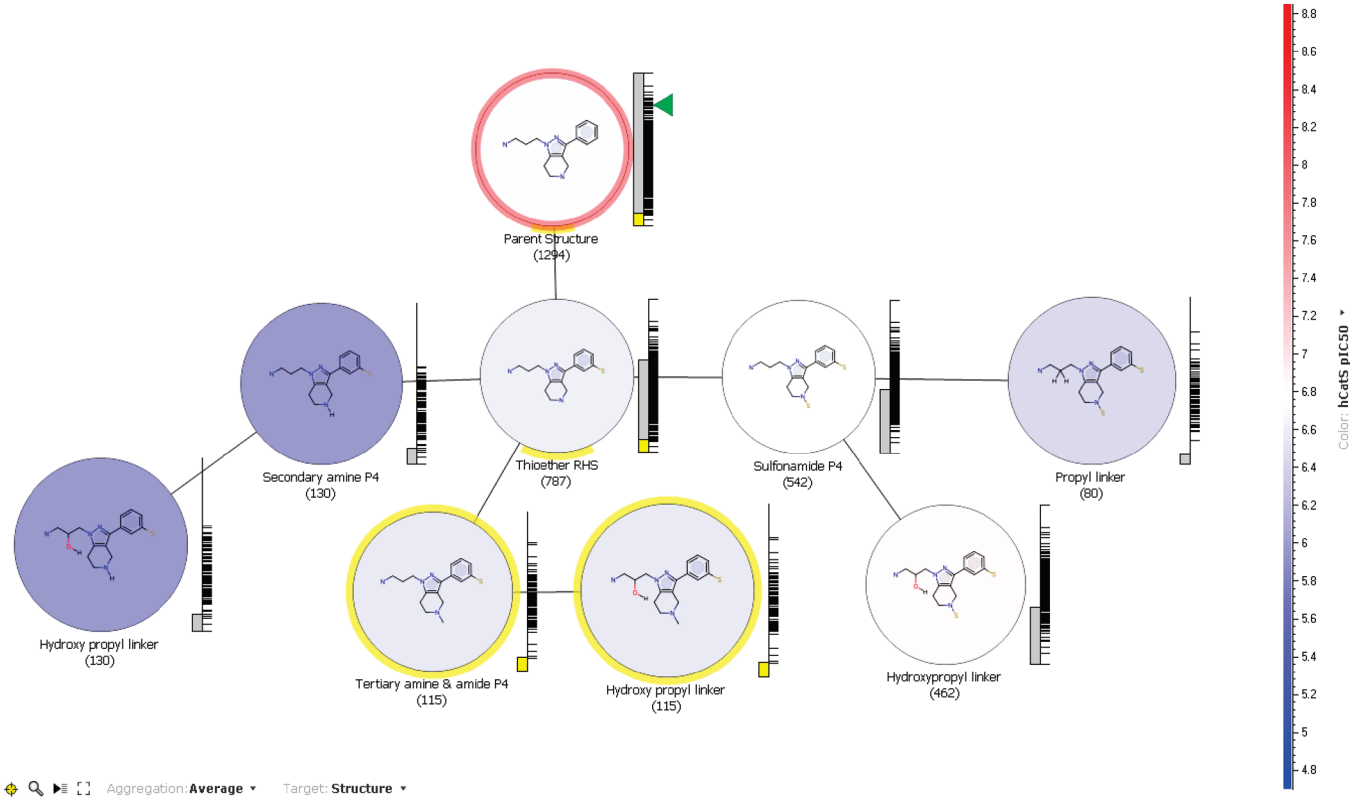

Figure 6. Scaffold tree representation of cathepsin S inhibitor thioether subseries, color-coded by average hCatS inhibition (pIC_{50}).

yellow, derived from the "tertiary amine & amide P4" node) with moderate potency, on average, as indicated by the coloration of the node. Yet, the vertical scale to the right of the node indicates that within this node are a number of high-potency outliers, rendering what might otherwise be an uninteresting terminal node instead one worth exploring in

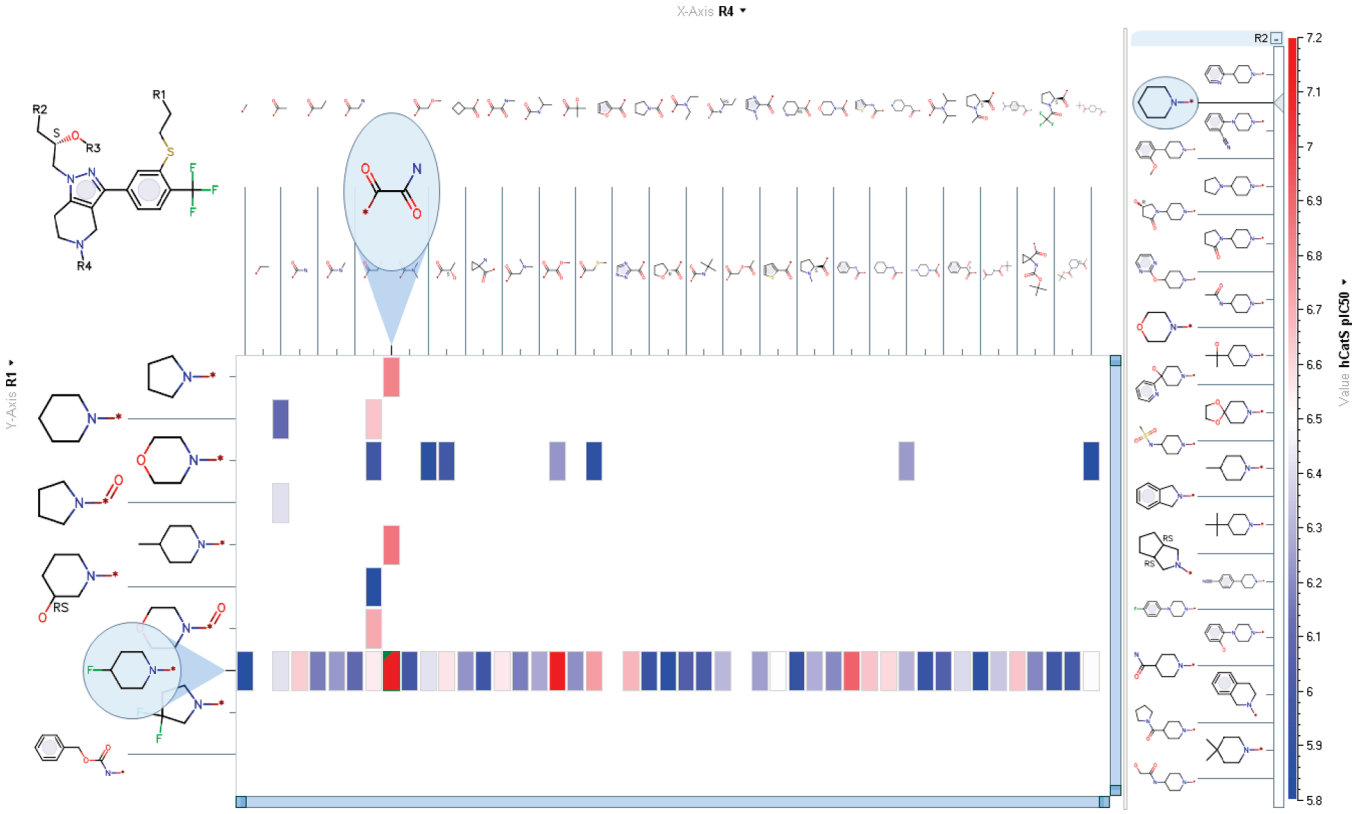

Figure 7. hCatS inhibition (pIC_{50}) of thioethers, sorted by R1 and R4 molecular weight, then by hCatS potency.

more depth. Once again, an SAR map helps to elucidate the SAR of this terminal node (Figure 7). The P4 substituents are arrayed along the horizontal axis and the P3 substituents lie along the vertical axis, both sorted according to molecular weight with the low-molecular-weight substituents at the left and top of the axes. On the basis of the positive findings from the alkyne SAR map relating to the potency of molecules with a piperidine P5 substituent, the slider to the right has been used to fix the P5 substituent here as a piperidine. Attention is quickly drawn to the high potency of a molecule containing the relatively low molecular weight oxamide moiety as a P4 substituent and a 4-fluoropiperidine P3 fragment. Whereas much larger P4 structures can afford reasonable or, in many cases, poor potency, the oxamide allows for remarkably high potency (hCatS $pIC_{50} = 7.2$) with relatively low molecular weight, highlighting the potential of the oxamide to form specific interactions in the S4 pocket, in accord with published crystal structure analyses indicating the presence of hydrogen bond donors in that region of the enzyme.[34]

Conclusions

We described an interactive tool that allows medicinal chemists to define arbitrary hierarchies of chemical scaffolds and use them to explore and visualize their existing project data. Our approach differs from previous automated scaffold classification algorithms in that the scaffolds can be of arbitrary complexity and their precise definition is controlled entirely by the user. The only constraint that the tool enforces in order to improve navigation is that each scaffold must represent a more refined substructure than that of its parent node. Scaffold trees can be dynamically edited when new scaffolds or variations of existing ones are introduced and can be interactively mapped to any collection of compounds to explore structure−activity relationships across multiple chemotypes. This is accomplished by allowing the user to aggregate any biological or physicochemical property of interest at the scaffold level and to display both the aggregate as well as the individual properties of the molecules in each scaffold class. The real utility of this tool comes from its interactivity and its ability to simultaneously explore multiple views of the data through linked visualizations. Our future plans include extending this tool to serve a generalized decision tree that will support a variety of partitioning functions beyond chemical patterns.

Acknowledgment. We thank the numerous users of ABCD and Third Dimension Explorer for providing valuable feedback during the development of this tool, and in particular Dr. Siquan Sun from J&JPRD La Jolla, whose conversations with one of the authors (D.K.A.) during the first release of ABCD provided the inspiration for the work described in this paper. Interestingly, Dr. Sun is a biologist.

References

(1) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today Technol.* **2004**, *1* (3), 217–224.
(2) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 393–404.
(3) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Learning from the data: mining of large high-throughput screening databases. *J. Chem. Inf. Model.* **2006**, *46*, 2381–2395.
(4) Wolohan, P. R. N.; Akella, L. B.; Dorfman, R. J.; Nell, P. G.; Mundt, S. M.; Clark, R. D. Structural unit analysis identifies lead series and facilitates scaffold hopping in combinatorial chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 1188–1193.
(5) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(6) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of large screening datasets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1069–1079.
(7) Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data analysis of high-throughput screening results: application of multi-domain clustering to the NCI anti-HIV data set. *J. Med. Chem.* **2002**, *45*, 3082–3093.
(8) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (31), 11473–11478.
(9) Auld, D. S.; Thome, N.; Nguyen, D.; Inglese, J. A specific mechanism for nonspecific activation in reporter-gene assays. *ACS Chem. Biol.* **2008**, *3* (8), 463–470.
(10) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(11) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46* (15), 3257–3274.
(12) 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.
(13) 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.
(14) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (48), 17272–17277.
(15) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nature Chem. Biol.* **2009**, *5*, 581–583.
(16) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.
(17) Xu, Y. J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.
(18) Katritzky, A. R.; Kiely, J. J.; Herbert, N.; Chassaing, C. Definition of templates within combinatorial libraries. *J. Comb. Chem.* **2000**, *2*, 2–5.
(19) Lounkine, E.; Wawer, M.; Wasserman, A. M.; Bajorath, J. SARANEA—a freely available program to mine structure−activity and structure−selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* **2010**, *50*, 68–78.
(20) Agrafiotis, D. K.; Shemanarev, K.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **2007**, *50* (24), 5926–5937.
(21) Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K. Enhanced SAR maps: expanding the data rendering capabilities of a popular medicinal chemistry tool. *J. Chem. Inf. Model.* **2009**, *49*, 2221–2230.
(22) 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.
(23) Gupta, S.; Singh, R. K.; Dastidar, S.; Ray, A. Cysteine cathepsin S as an immunomodulatory target: present and future trends. *Expert Opin. Ther. Targets* **2008**, *12*, 291–299.
(24) Villadangos, J. A.; Bryant, R. A. R.; Deussing, J.; Driessen, C.; Lennon-Dumenil, A.-M.; Riese, R. J.; Roth, W.; Saftig, P.; Shi, G.-P.; Chapman, H. A.; Peters, C.; Ploegh, H. L. Proteases involved in MHC Class II antigen presentation. *Immunol. Rev.* **1999**, *172*, 109–120.
(25) Villadangos, J. A.; Ploegh, H. L. Proteolysis in MHC Class II antigen presentation: Who's in charge? *Immunity* **2000**, *12*, 233–239.
(26) Chapman, H. A. Endosomal proteolysis and MHC Class II function. *Curr. Opin. Immunol.* **1998**, *10*, 93–102.
(27) Nakagawa, T. Y.; Rudensky, A. Y. The role of lysosomal proteinases in MHC Class II-mediated antigen processing and presentation. *Immunol. Rev.* **1999**, *172*, 121–129.
(28) Link, J. O.; Zipfel, S. Advances in cathepsin S inhibitor design. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 471–482.
(29) Thurmond, R. L.; Sun, S.; Karlsson, L.; Edwards, J. P. Cathepsin S Inhibitors as Novel Immunomodulators. *Curr. Opin. Invest. Drugs* **2005**, *6*, 473–482.
(30) Leroy, V.; Thurairatnam, S. Cathepsin S inhibitors. *Expert Opin. Ther. Patents* **2004**, *14*, 301–311.
(31) Ameriks, M. K.; Arienti, K. L.; Edwards, J. P.; Grice, C. A.; Jones, T. K.; Lee-Dutra, A.; Liu, J.; Mani, N. S.; Neff, D. K.; Wickboldt, A. T.; Wiener, J. J. M. Preparation of tetrahydro-pyrazolo-pyridine thioether modulators of cathepsin S. U.S. Patent US2009-099157-A1, 2009.
(32) Ameriks, M. K.; Axe, F. U.; Edwards, J. P.; Grice, C. A.; Cai, H.; Gleason, E. A.; Meduna, S. P.; Tays, K. L.; Wiener, J. J. M.; Wickboldt, A. T. Preparation of carbon-linked tetrahydro-pyrazolo-pyridines, particularly substituted 1-[3-(monocyclic amino)-2-hydroxypropyl]-3-phenyl-4,5,6,7-tetrahydro-1*H*-pyrazolo[4,3-*c*]-pyridines, as modulators of cathepsin S. U.S. Patent US2008-0200454-A1, 2008.
(33) Allen, D.; Ameriks, M. K.; Axe, F. U.; Burdett, M.; Cai, H.; Choong, I.; Edwards, J. P.; Lew, W.; Meduna, S. P. Monocyclic aminopropyl tetrahydropyrazolopyridines as modulators of cathepsin S and their preparation, pharmaceutical compositions and use in the treatment of CatS-mediated diseases. U.S. Patent US2009-0118274-A1, 2009.
(34) Ameriks, M. K.; Axe, F. U.; Bembenek, S. D.; Edwards, J. P.; Gu, Y.; Karlsson, L.; Randal, M.; Sun, S.; Thurmond, R. L.; Zhu, J. Pyrazole-based cathepsin S inhibitors with arylalkynes as P1 binding elements. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6131–6134.
(35) Ameriks, M. K.; Cai, H.; Edwards, J. P.; Gebauer, D.; Gleason, E.; Gu, Y.; Karlsson, L.; Nguyen, S.; Sun, S.; Thurmond, R. L.; Zhu, J. Pyrazole-based arylalkyne cathepsin S inhibitors. Part II: Optimization of cellular potency. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6135–6139.
(36) McGrath, M. E.; Palmer, J. T.; Bromme, D.; Somoza, J. R. Crystal structure of human cathepsin S. *Protein Sci.* **1998**, *7*, 1294–1302.
(37) Pauly, T. A.; Sulea, T.; Ammirati, M.; Sivaraman, J.; Danley, D. E.; Griffor, M. C.; Kamath, A. V.; Wang, I.-K.; Laird, E. R.; Seddon, A. P.; Menard, R.; Cygler, M.; Rath, V. Specificity determinants of human cathepsin S revealed by crystal structures of complexes. *Biochemistry* **2003**, *42*, 3203–3213.
(38) Patterson, A. W.; Wood, W. J. L.; Hornsby, M.; Lesley, S.; Spraggon, G.; Ellman, J. A. Identification of selective, nonpeptidic nitrile inhibitors of cathepsin S using the substrate activity screening method. *J. Med. Chem.* **2006**, *49*, 6298–6307.
(39) Inagaki, H.; Tsuruoka, H.; Hornsby, M.; Lesley, S. A.; Spraggon, G.; Ellman, J. A. Characterization and Optimization of Selective, Nonpeptidic Inhibitors of Cathepsin S with an Unprecedented Binding Mode. *J. Med. Chem.* **2007**, *50*, 2693–2699.