

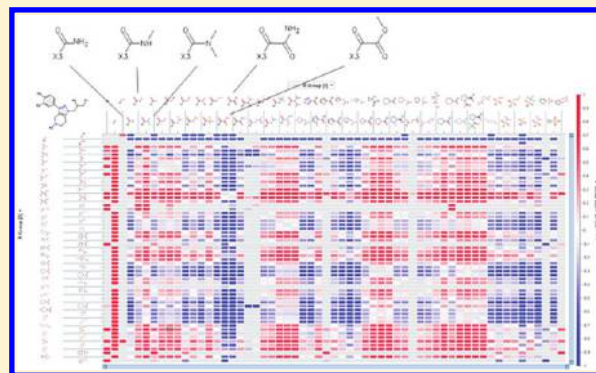
# Single R-Group Polymorphisms (SRPs) and R-Cliffs: An Intuitive Framework for Analyzing and Visualizing Activity Cliffs in a Single Analog Series

Dimitris K. Agrafiotis,<sup>\*,†</sup> John J. M. Wiener,<sup>‡</sup> Andrew Skalkin,<sup>†</sup> and Jeremy Kolpak<sup>†</sup>

<sup>†</sup>Informatics, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., Welsh & McKean Roads, Spring House, Pennsylvania 19477, United States

<sup>‡</sup>Medicinal Chemistry, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 3210 Merryfield Road, San Diego, California 92121, United States

**ABSTRACT:** We introduce Single R-Group Polymorphisms (SRPs, pronounced ‘sharps’), an intuitive framework for analyzing substituent effects and activity cliffs in a single congeneric series. A SRP is a pair of compounds that differ only in a single R-group position. Because the same substituent pair may occur in multiple SRPs in the series (i.e., with different combinations of substituents at the other R-group positions), SRP analysis makes it easy to identify systematic substituent effects and activity cliffs at each point of variation (R-cliffs). SRPs can be visualized as a symmetric heatmap where each cell represents a particular pair of substituents color-coded by the average difference in activity between the compounds that contain that particular SRP. SRP maps offer several advantages over existing techniques for visualizing activity cliffs: 1) the chemical structures of all the substituents are displayed simultaneously on a single map, thus directly engaging the pattern recognition abilities of the medicinal chemist; 2) it is based on R-group decomposition, a natural paradigm for generating and rationalizing SAR; 3) it uses a heatmap representation that makes it easy to identify systematic trends in the data; 4) it generalizes the concept of activity cliffs beyond similarity by allowing the analyst to sort the substituents according to any property of interest or place them manually in any desired order.



## INTRODUCTION

While the armamentarium of medicinal chemists and our understanding of biology and disease etiology have grown tremendously in recent years, the basic process for optimizing drug candidates has remained largely unchanged. 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. These substituents are optimized in an iterative manner 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, new sets of analogs are synthesized, and the cycle continues until a clinical candidate emerges or the project is terminated.

At the heart of this process lies the hypothesis (or, more accurately, historical observation) that similar molecules tend to exhibit generally similar biological properties. While this principle has led to the discovery of many of the breakthrough medications of the modern era, they are not without many exceptions. In fact, if one looks critically at any drug discovery program, for every successful modification of a lead compound, one will find

numerous other attempts that did not yield the expected outcome. This also explains why some efforts to capture structure–activity relationships in a quantitative model have had questionable prospective value.<sup>1,2</sup>

The challenges associated with predictive QSAR have been attributed to the rugged nature of structure–activity landscapes and the presence of activity cliffs, that is, cases where seemingly small changes in chemical structure result in significant changes in activity.<sup>3</sup> Of course, such cliffs are to be expected given that the formation of a single additional hydrogen bond can result in more than a 10-fold increase in potency. What complicates SAR analysis is that it is often difficult to separate experimental variability (noise or measurement errors in the biological response, unknown compound purity, etc.) from observations that simply contradict the physical assumptions of the model. Regardless of these complexities, one thing remains certain: activity cliffs are of inherent interest because they point to potentially invalid assumptions about the underlying SAR and can provide new ideas for compound optimization.

**Received:** February 4, 2011

**Published:** April 19, 2011

The study of activity cliffs has attracted significant attention in recent years, as manifested by a number of original publications and reviews<sup>4,5</sup> and a dedicated symposium at the Fall 2010 ACS national meeting.<sup>6</sup> Activity cliffs are usually detected by comparing a set of molecules to each other in a pairwise fashion and identifying those pairs where the change in activity relative to the change in structure exceeds a certain threshold. Two specific numerical indices that have been proposed to guide this process are the SALI<sup>7,8</sup> and SARI<sup>9</sup> scores, which characterize individual pairs and entire SAR landscapes, respectively. While both of these indices can in theory identify activity cliffs, they are virtually impossible to digest in their raw numerical form, particularly if the data set exceeds a few tens of compounds. Hence, they are often visualized in the form of a graph, where nodes represent individual compounds and edges similarity relationships between them that meet certain criteria. There are several variations of these graphs, which differ in the method used to assign the edges, the way in which the nodes are laid out on the canvas, and the visual annotation of the various graph elements.<sup>10,11</sup> (Alternative nongraph-based visual representations have also been proposed, such as the SAS maps<sup>12</sup> where each point represents a pair of compounds plotted against structural similarity on one axis and activity similarity on the other, a concept reminiscent of the neighborhood plots of Patterson et al.<sup>13</sup>).

In a more recent publication that is closely related to the present work, Bajorath's group presented a complementary approach for analyzing substitution patterns in a single analog series.<sup>14,15</sup> More specifically, the authors use the methodology described by Bemis and Murcko<sup>16</sup> to automatically divide a large data set into groups of compounds that share a common core molecular framework. Within each framework, variable regions are identified using R-group analysis, and SARI discontinuity scores are computed for each pair of compounds that differ in one, two, or three substituents. The resulting information is presented in a so-called combinatorial analog graph (CAG), where each node corresponds to a subset of compounds that differ in one or more combinations of substituents. Nodes are arranged in layers according to the number of substitution sites that are considered and gray-scaled according to their average discontinuity score. Edges are drawn from a node to all other nodes in the next layer whose substitution site combination includes all of the sites represented in the originating node.

While the aforementioned graph-based representations offer a glimpse into the complexity of the underlying SAR space, in our opinion they suffer from three major drawbacks. The first is the density of the graphs, which becomes unmanageable as the size of the data set increases beyond a certain level (CAGs being the only exception). Unless the cutoff value is sufficiently stringent, the graphs can become cluttered with edges, complicating the layout and obscuring important detail (a problem common to many techniques for visualizing large graphs, often referred to as "too much ink"). The second is their dependence on the precise definition of molecular similarity. Maggiora and others have pointed out that chemical spaces are not invariant with respect to representation, and the concept of chemical neighborhood depends critically on the choice of descriptors and similarity functions.<sup>3</sup> Thus, what is a rugged SAR landscape in one representation may be a smooth terrain in another, further complicating interpretation.<sup>17,18</sup> Although alternative similarity functions can and have been used,<sup>19</sup> the need to select an appropriate representation for the problem at hand is still present. But perhaps the most important problem is that the chemical structures of the

compounds (or the individual substituents in the case of CAGs) are completely absent from these visualizations, which limits the ability of the medicinal chemists to place the SAR in the proper chemical context.

Although our group has spent considerable effort on quantitative methods for structure–activity analysis,<sup>20–24</sup> our most recent work has focused on intuitive visualization techniques that can provide insights into general trends and can be used by experts and nonexperts alike. Two tools that have proven particularly useful in this regard are Scaffold Explorer and SAR maps.

Scaffold Explorer allows users to organize their compounds into a hierarchy of scaffolds, where each scaffold represents any arbitrary substructure with variable atoms, bonds, and/or substituents.<sup>25</sup> The substructures associated with each scaffold can be recursively elaborated into increasingly refined substructures, representing deeper nodes in the tree. 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.

Scaffold Explorer is particularly useful in conjunction with SAR maps,<sup>26,27</sup> which provide more detailed views of the substituent effects around each individual scaffold. 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 any chemical property or biological activity, which makes patterns easy to identify. The tool was recently extended<sup>27</sup> to allow many different types of visualizations to be rendered inside the cells, such as multidimensional histograms and pie charts that visualize the biological profiles of compounds across an entire panel of assays, dose–response curves, aligned 3D structure drawings, and many others. These enhancements allow the medicinal chemist to interactively analyze complex scaffolds with multiple substitution sites, identify missing analogs or screening data, and correlate substituent structure and biological activity at multiple simultaneous dimensions.

Although SAR maps are extremely intuitive, their navigation becomes more challenging when the series contains more than two variable sites. In such situations, the effects of a particular substituent at a given position across the breadth of the entire series, encompassing changes at all of the other positions, are not immediately obvious, nor do the tools referenced above offer significant assistance without extensive interactive manipulation. For example, in a data set encompassing a core structure having five variable positions  $R_1$ – $R_5$ , some of the key questions a scientist would ask in dissecting the SAR and making decisions for future efforts are as follows: (1) which substituents make a consistent difference across all other R combinations, e.g., do any  $R_1$ s consistently outperform all other  $R_1$ s for all combinations of  $R_2$  through  $R_5$ ? (2) what type of functionality is needed at each particular part of the molecule (variation site), and which changes in structure represent activity cliffs with respect to a given molecular property such as size, lipophilicity, basicity, hydrogen bonding potential, etc.? Stated differently, which substituent changes are unexpected and may refute the current working hypotheses?

| Core_id  | Core | R1_id   | R1 | R2_id | R2 | R3_id  | R3 | R4_id | R4 | R5_id    | R5 |
|----------|------|---------|----|-------|----|--------|----|-------|----|----------|----|
| C19H14N2 |      | C4H8NO  |    | H     |    | CH3O2S |    | Cl    |    | C16H13ON |    |
| C19H14N2 |      | C4H8NO  |    | H     |    | CH3O2S |    | Cl    |    | C16H12O2 |    |
| C19H14N2 |      | C4H8NO  |    | HO    |    | CH3O2S |    | Cl    |    | C16H13ON |    |
| C19H14N2 |      | C5H10NO |    | H     |    | CH2NO2 |    | Cl    |    | C16H13ON |    |
| C19H14N2 |      | C5H10NO |    | H     |    | CH2NO2 |    | Cl    |    | C16H13ON |    |

Figure 1. Typical SAR table augmented with R-groups.

To answer these questions, we introduce the concept of Single R-Group Polymorphisms (SRPs, pronounced ‘sharps’), a convenient framework for analyzing substituent effects and activity cliffs in a single congeneric series spanning multiple variation sites. A SRP is a pair of compounds that have the same substituents in all but one variation site around the core. The method involves four basic steps: 1) for each variation site and each pair of substituents at that site, identify all SRPs that involve those two substituents, that is, all pairs of molecules ( $i, j$ ) that differ only in those two substituents and have the same substituents at every other position; 2) for each such pair of molecules, compute their difference in activity,  $\Delta A_{ij}$ ; 3) aggregate  $\Delta A_{ij}$  over all pairs of molecules that exhibit that particular SRP using a suitable aggregation function (mean, median, count, etc); and 4) visualize these aggregate differences in activity in a chemical heatmap.

Because the same substituent pair may occur in multiple SRPs in the series (i.e., with different combinations of substituents at the other sites), SRP analysis makes it easy to identify systematic substituent effects and activity cliffs at each point of variation. We term these systematic effects R-cliffs to distinguish them from conventional activity cliffs which look at differences between individual molecules. SRPs can be visualized as a symmetric heatmap where each cell represents a particular pair of substituents color-coded by the average difference in activity between the compounds that contain that particular SRP. SRP maps offer several advantages over existing techniques for visualizing activity cliffs: 1) they are based on R-group decomposition, a natural paradigm for generating and rationalizing SAR; 2) they use a heatmap representation that compresses large volumes of data in a compact visualization and makes it easy to identify systematic trends; 3) the chemical structures of all the substituents at a given site are displayed simultaneously on a single map, thus directly engaging the pattern recognition abilities of the medicinal chemist; 4) they generalize the concept of activity cliffs beyond

abstract similarity by allowing the analyst to sort the substituents according to any property of interest.

While the underlying concept also lends itself to quantitative analysis (and such analysis could indeed be very useful), this paper is focused exclusively on visualization because of the authors’ experience with how medicinal chemists prefer to analyze SAR. Heatmaps are a familiar paradigm with a long history in data visualization, particularly in the life sciences, and seemed quite appropriate for the task at hand.

In the remaining sections, we provide a detailed description of the SRP analysis algorithm and user interface and demonstrate the utility of the method using a case study drawn from a cathepsin S (CatS) inhibitor program.

## METHODS

**Third Dimension Explorer (3DX) and ABCD.** The SRP analysis and visualization components described in the sequel were implemented as plugins 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 platform known as ABCD,<sup>28</sup> 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. 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 the previously described SAR maps, which form the basis of the work described in this paper. The program offers a full gamut of navigation and selection options, augmented through linked



visualizations and interactive filtering and querying. The versatility of 3DX as a general purpose visualization tool has been demonstrated in a variety of other domains beyond discovery research.<sup>29</sup>

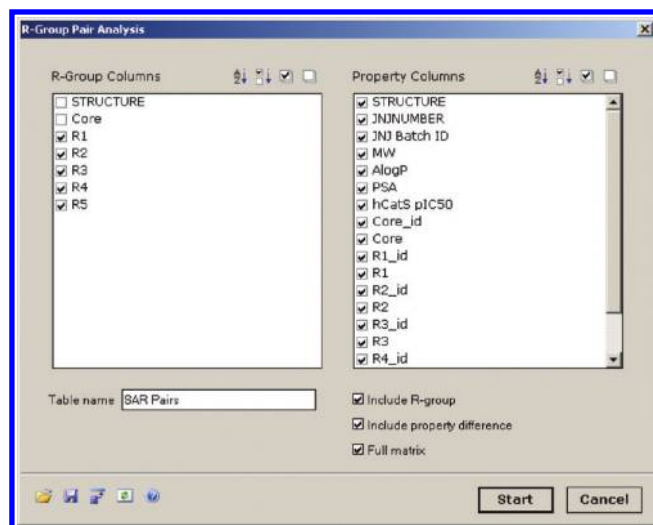
**R-Group Analysis.** The first part of SRP analysis is to decompose a set of molecules sharing a common core into a list of substituents at each point of variation around that core. This process, which is commonly referred to as R-group analysis, takes as input a conventional SAR table (i.e., a molecular spreadsheet) and augments it with additional columns containing the structures and identifiers of the core and the substituents at each variation site (Figure 1).

The common core, which can be any arbitrary substructure and may contain both cyclic and acyclic components, is specified either through manual sketching or through an automated search for the maximum common substructure (MCS). Since the core substructure can be mapped onto a target molecule in multiple ways, our R-group analysis algorithm<sup>26</sup> is designed to identify the mapping that minimizes the number of attachment points on the core. The resulting R-groups are extracted into separate columns labeled R<sub>1</sub>, R<sub>2</sub>, etc., and their attachment points are replaced with dummy atoms labeled X. Similarly, the attachment points on the core are replaced with dummy atoms labeled R<sub>1</sub>, R<sub>2</sub>, etc.

**SRP Analysis.** The SRP analysis plugin computes property differences between compounds that share a common core and differ in only one R-group position. It takes as input an R-group table produced by the R-group analysis plugin described above and produces one or more new SRP tables containing either individual or aggregate SRPs (*vide infra*). The user interface is shown in Figure 2. The dialogue prompts the user to specify the columns containing the R groups (large selection box on the top left labeled “R Group Columns”) and the columns containing the properties whose difference is to be computed (large selection box on the top right labeled “Property Columns”). Because each column in a 3DX table is associated with a particular data type (number, text, image, structure, etc.), these selection boxes are respectively prefilled with those columns in the current 3DX table that are of type Structure (R Group Columns) or any numerical type (Property Columns). In addition, structure columns whose names are prefixed with “R” (suggestive that they contain R-groups as opposed to full structures — a convention employed by the R-Group analysis plugin) are autoselected, further reducing the amount of input required by the user.

Three checkboxes allow the user to specify the information included in the output table(s). “Include R-group” specifies whether the structures of the substituents of the two compounds exhibiting the SRP are included in the output table (checked by default). “Include property difference” specifies whether the output table should include the difference in the selected properties, in addition to the property values of the two molecules (checked by default). Finally, the “Full matrix” checkbox specifies whether the output should include a single table with all SRPs listed as separate rows along with an identifier of the R-group position where they occur, or as multiple tables, one for each position of variation.

By default, the SRP analysis plugin produces a single SRP table for all R-group positions, where each SRP is listed on a separate row (Figure 3). The columns of this table include the R-group position where the SRP occurs, the structures of the substituents of the two molecules exhibiting the SRP (if “Include R-group” is checked), and a set of three columns for each selected property, containing the property values of the two compounds, and



**Figure 2.** User interface of SRP analysis plugin. Because each column in a 3DX table is associated with a particular data type (number, text, image, structure, etc.), the two large column selection boxes on the dialogue labeled “R Group Columns” and “Property Columns” are prefilled with those columns in the current 3DX table that are of type Structure and numerical columns, respectively. These boxes allow the user to specify the columns containing the structures of the R-groups and the property columns that should be included in the output table. The “Include property difference” and “Full matrix” checkboxes allow the user to specify whether property differences should be included in the output table, and whether the output should include a single table with all SRPs listed as separate rows along with an identifier of the R-group position where they occur, or as multiple tables, one for each R-group position.

(optionally) the difference between them. As shown in Figure 3, we use the convention “*property* [1]”, “*property* [2]”, and “*property* [2] – [1]” to label the columns containing the property values of the first and second molecule and their difference. These tables lend themselves to computing and visualizing general SRP statistics, such as the relative prevalence of SRPs at each R-group position, the distribution of property differences, etc., and can be further augmented with various computed properties of the R-groups, and sorted in a way that makes activity cliffs easy to identify.

If the “Full matrix” checkbox is checked, the SRP analysis plugin produces a separate SRP table for each variation site. Unlike the previous option, these individual R-group tables contain all possible pairs of substituents, including the ones that are not observed in the data set. More specifically, if there are  $n$  distinct substituents at the R<sub>1</sub> position, there will be  $n \times (n + 1)/2$  rows in the resulting SRP table for R<sub>1</sub>. If no SRP is observed for a particular pair of substituents, the corresponding property entries in the SRP table will be empty. Similarly, if the same pair of substituents is involved in multiple SRPs (with different combinations of substituents at the other R-group positions), the individual SRPs will appear as different entries in multivalued cells (Figure 4). These cells can be aggregated on-the-fly by selecting an appropriate aggregation function in the column header, as shown in Figure 4. These tables are ideally suited to visualization in the form of a heatmap as described in the next section.

The five buttons at the bottom-left of the SRP analysis dialogue are shared among all 3DX UI-based plugins and allow the user to read the dialogue settings from a file, save the current

|   | R Group # | R Group [1] | R Group [2] | hCats pIC <sub>50</sub> [1] | hCats pIC <sub>50</sub> [2] | hCats pIC <sub>50</sub> [2]-[1] | MW [1] | MW [2] | MW [2]-[1] | AlogP [1] | AlogP [2] | AlogP [2]-[1] | PSA [1] | PSA [2] | P = |
|---|-----------|-------------|-------------|-----------------------------|-----------------------------|---------------------------------|--------|--------|------------|-----------|-----------|---------------|---------|---------|-----|
| 1 | R5        |             |             | 6.07                        | 7.41                        | 1.35                            | 670.84 | 682.01 | 11.97      | 2.69      | 0.98      | -1.71         | 103.17  | 120.24  |     |
| 2 | R5        |             |             | 6.07                        | 7.04                        | 0.98                            | 670.84 | 684.82 | 13.99      | 2.69      | 2.07      | -0.62         | 103.17  | 120.24  |     |
| 3 | R1        |             |             | 6.07                        | 6.1                         | 0.032                           | 670.84 | 676.84 | 6          | 2.69      | 1.95      | -0.84         | 103.17  | 136.87  |     |
| 4 | R5        |             |             | 6.07                        | 6.22                        | 0.15                            | 670.84 | 757.94 | 87.1       | 2.69      | 1.68      | -1.01         | 103.17  | 114.69  |     |
| 5 | R5        |             |             | 6.07                        | 6.47                        | 0.4                             | 670.84 | 744.94 | 74.1       | 2.69      | 2.48      | -0.21         | 103.17  | 114.61  |     |
| 6 | R5        |             |             | 6.07                        | 7.15                        | 1.09                            | 670.84 | 736.07 | 60.03      | 2.69      | 1.26      | -1.41         | 103.17  | 120.68  |     |
| 7 | R5        |             |             | 6.07                        | 6.96                        | 0.89                            | 670.84 | 714.92 | 44.09      | 2.69      | 3.38      | -0.69         | 103.17  | 94.38   |     |

**Figure 3.** SRP table with all individual SRPs listed in separate rows (form 1). The core structure is not included but is the same as the one shown in Figure 1.

settings on a file or store them as an item on the application menu (providing accelerated invocation for frequently used plugins), clear/reset the form, and obtain help through the online system.

**SRP Maps.** The full SRP matrices for each R-group can be visualized using the previously described SAR map technique.<sup>26</sup> While the appearance is identical, a SRP map differs from a SAR map in two important respects. First, in a SAR map each axis displays a different R-group, whereas in an SRP map both axes display the same R-group with the substituents sorted in the same order (see Figures 6–8). Second, in a SAR map each cell represents a single molecule, whereas in a SRP map it represents a set of molecule pairs. If  $q$  is the variation site being displayed on the map, and  $R_q(i)$  and  $R_q(j)$  are the substituents at the  $i$ th row and  $j$ th column, respectively, then the cell  $(i, j)$  represents all pairs of molecules where the first molecule contains substituent  $R_q(i)$  and the second molecule substituent  $R_q(j)$  at position  $q$ , and where all other substituents on that pair of molecules are the same (i.e.,  $R_k(i) \equiv R_k(j) \forall k \neq q$ ). A third difference that emerges from this definition is that SRP maps do not require the additional chemical sliders that are needed to display more than two R-groups in a regular SAR map.

The graphical interface is completed with a color-scale and an additional dropdown box that allows the user to interactively select which property to use for color-coding the cells. Because each cell represents a collection of SRPs, the color represents an aggregate function of these SRPs, such as the average or median difference of a particular property. The aggregation function is specified interactively by the user using the dropdown box on the corresponding column header in the SRP table (Figure 4). As with SAR maps, cells associated with nonobserved SRPs are colored in gray.

Because the substituents on the X and Y axes are sorted in the same order, a SRP map appears symmetric, with the exception

that the colors of the  $(i, j)$  and  $(j, i)$  cells are inverted (since the sign of the property difference is the opposite). Because they use the same underlying component, SRP maps incorporate all the features introduced in the enhanced SAR maps,<sup>27</sup> such as the ability to sort the substituents by any molecular property or by manually dragging them along the chemical axes, the ability to zoom into a particular region of the map, and the ability to highlight the current and selected records.

## DISCUSSION

**CatS Inhibitor Program.** To demonstrate the value of the new method, we use a case study from a CatS inhibitor program. The cysteine protease CatS mediates cleavage of the major histocompatibility class II (MHC II)-associated invariant chain (Ii). This cleavage is a key step in the events leading to antigen presentation on the cell surface and, as such, is a crucial portion of an immune response.<sup>30–34</sup> CatS inhibitors have thus been proposed as therapeutics for a variety of autoimmune disorders and other diseases. Both covalent-binding active site-modifiers and non-covalent inhibitors of CatS have been disclosed.<sup>35–44</sup> Several binding pockets of the CatS enzyme (known as S1–S5) have been identified, with the regions of the molecules occupying these portions of the enzyme named correspondingly P1–P5.<sup>45–48</sup>

Analysis began with a search of the ABCD data warehouse, retrieving P2 pyrazole structures for which human CatS enzymatic binding data (hCats pIC<sub>50</sub>) had been generated under the same assay protocol (experimental conditions). For the purpose of this discussion, we will focus our attention on a subset of 1317 unique structures that were tested in the assay. These molecules, which were specifically chosen to illustrate the full capabilities of the tool, represent a core P2 pyrazole structure adorned with













|     | R Group [1]   | R Group [2]   | Count | Row [1]  | Row [2]   | hCatS pIC <sub>50</sub> [2]-[1] | hCatS pIC <sub>50</sub> [1]   | hCatS pIC <sub>50</sub> [2]   |   |
|-----|---|---|-------|--|---|---------------------------------|---|---|---|
| 127 |  |  | 6     | 648<br>678<br>682<br>721<br>725<br>728                             | 723<br>725<br>724<br>727<br>941<br>1257                             | 0.2 (6)                         | 6.8399<br>7.1003<br>7.0489<br>7.1549<br>6.8495<br>7.0775  | 7.7614<br>6.9381<br>8<br>6.3753<br>7.3979<br>6.8239   | <div> List </div> <div> Minimum </div> <div> Maximum </div> <div> Average </div> <div> Average ± StdDev </div> <div> StdDev </div> <div> Sum </div> <div> Median </div> <div> Median ± AvgAbsDev </div> |
| 128 |  |  | 12    | 577<br>680<br>681<br>735<br>764<br>781<br>783<br>800<br>804<br>834 | 796<br>935<br>803<br>930<br>806<br>985<br>983<br>987<br>986<br>1123 | 0.5 (12)                        | 5.8746<br>6.3768<br>6.1831<br>6.2924<br>6.0211<br>7.228<br>6.5062<br>6.5431<br>6.3835<br>5.1297 | 5.6616<br>6.7967<br>6.5442<br>7.699<br>6.1088<br>6.7696<br>6.7101<br>6.7012<br>6.9648<br>6.7596 |   |
| 129 |  |  | 2     | 126<br>144   | 416<br>374  | 1.17 (2)                        | 5.9684<br>6.2444  | 6.9583<br>7.639   |   |
| 130 |  |  |       |  |   |                                 |   |   |   |
| 131 |  |  |       |  |   |                                 |   |   |   |
| 132 |  |  |       |  |   |                                 |   |   |   |

Figure 4. SRP table for aggregate SRPs in a single R-group position (full matrix).

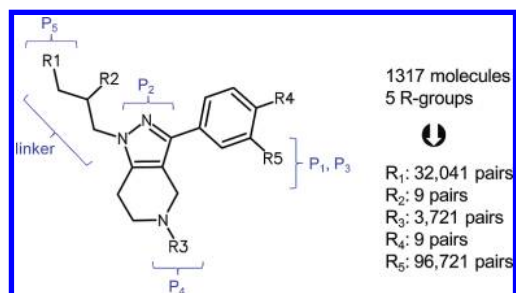


Figure 5. General structure of cathepsin S inhibitor series.

substituents at five positions (Figure 5), with the variable positions defined as described in relevant patent applications.<sup>40–42</sup>

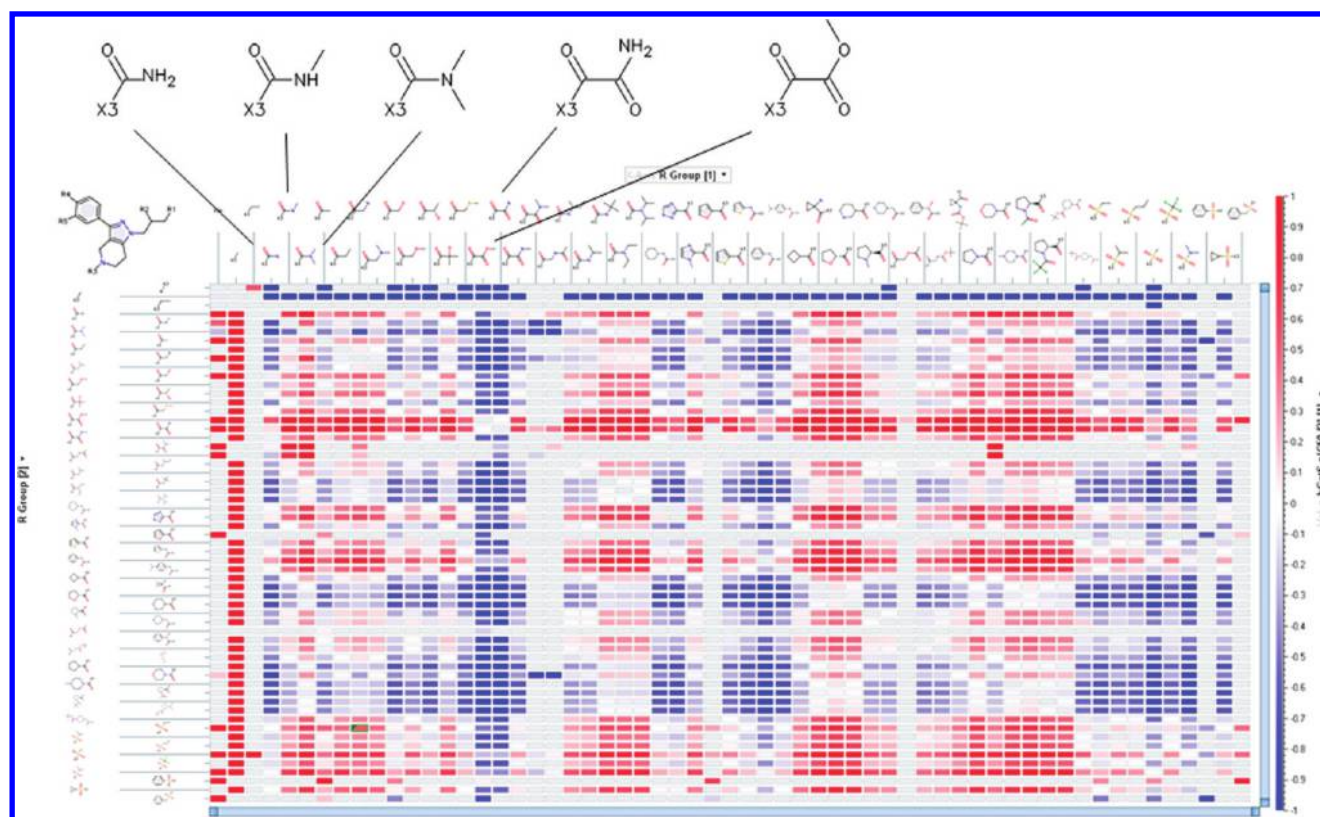
**Application of SRP Maps to CatS Inhibitor Program.** Having defined the core structure shown in Figure 5 (the scaffold used to create all the analogs), the R-group analysis plugin in 3DX was used to identify all R-groups at positions R<sub>1</sub>–R<sub>5</sub>. Next, the SRP analysis plugin was used to generate all possible pairs of SRPs across the entire data set, listing also the R-group number and the difference in CatS enzymatic activity, expressed as pIC<sub>50</sub>s, for each pair of molecules. The SRP analysis plugin also generated individual R-group SRP matrices; for each R<sub>n</sub>, one SRP matrix table was created within which all possible SRP pairs were listed and the pIC<sub>50</sub>-difference data aggregated. From a given SRP matrix, an SRP heatmap of the substituents at that R<sub>n</sub> was

then created, plotting all of the particular R<sub>n</sub> groups along the X and Y axes in the same order, with the colored cells representing pIC<sub>50</sub> differences. Notably, the ordering of substituents on the axes can be adjusted to reflect any desired property, from computed abstract structural similarity of substituents, to physical properties such as molecular weight or logP.

As is evident upon inspection of the SRP map in Figure 6, red rectangles at the intersection of two substituents indicate that, on average, the group on the X axis is less potent than the group on the Y axis across all of the compounds containing that pairing of substituents, regardless of whatever other substituents are present at other variable positions. Conversely, blue rectangles highlight X axis substituents which are, on average, more potent than their Y axis counterparts.

Of course, for aggregate activity of these SRPs to diagnose meaningfully the presence of an R-cliff, the underlying number of compounds should be as large as possible. Given the relative sparseness of many data sets (the number of prepared compound pairs vs the total possible number of pairs), such considerations are all the more relevant. One benefit of these SRP maps is their ability to readily display such information, simply by toggling the color-coding from displaying pIC<sub>50</sub> difference to displaying count (Figure 7). As with many data sets, it can be seen here that the preponderance of SRPs are each drawn from a small number of total compounds (white rectangles), with far fewer SRPs representing a large set of analogs (red rectangles).





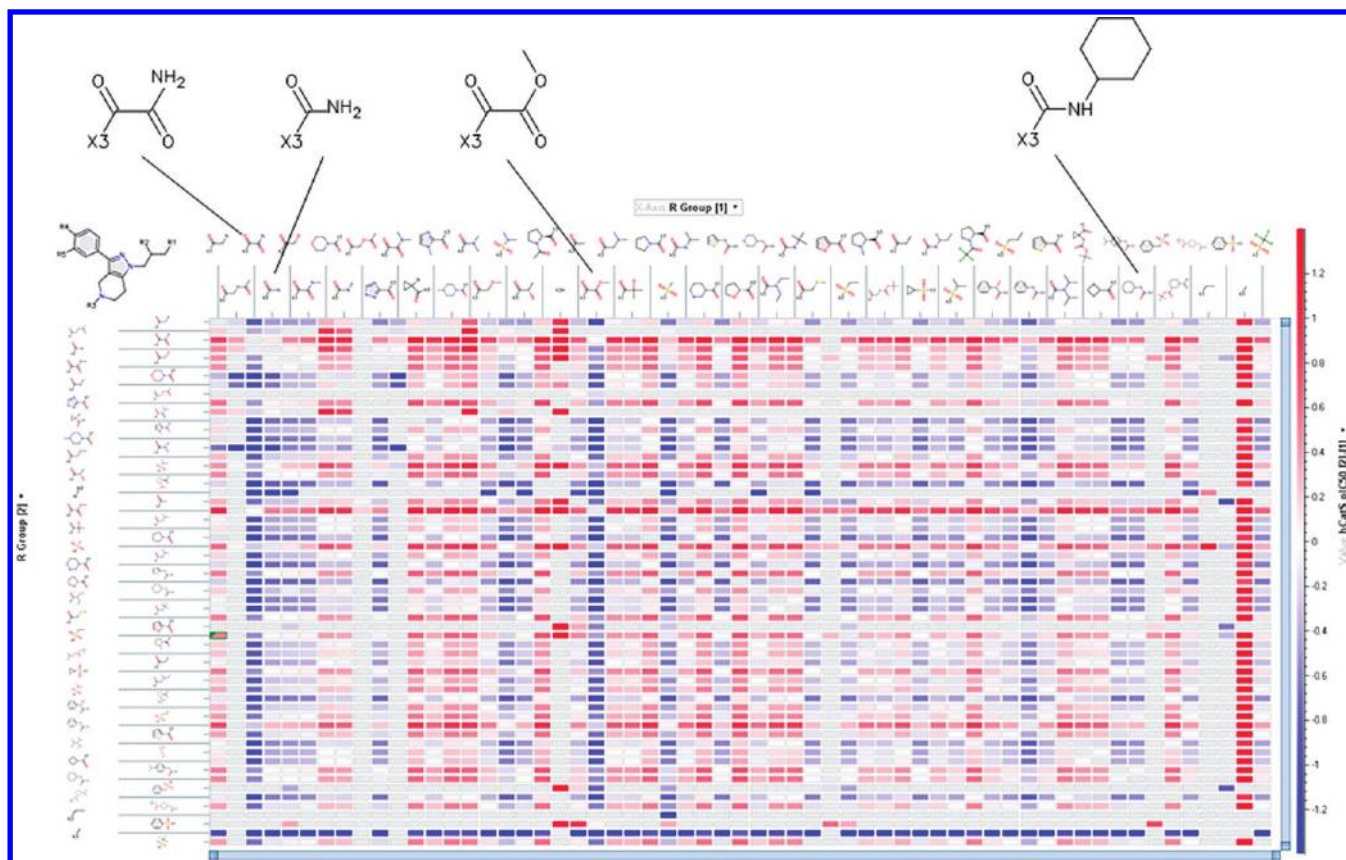
**Figure 6.** SRP map representation of cathepsin S inhibitor series R<sub>3</sub> substituents, color-coded by average hCatS inhibition difference (pIC<sub>50</sub>), and axes sorted by abstract similarity.



**Figure 7.** SRP map representation of cathepsin S inhibitor series R<sub>3</sub> substituents, color-coded by count, and axes sorted by abstract similarity.

In particular, in the SRP map shown in Figure 6 analyzing the R<sub>3</sub> substituents, ordered by structural similarity (derived by computing 117 topological descriptors for each substituent, and normalizing and embedding them into a 1-dimensional space

using stochastic proximity embedding<sup>49–51</sup>), several conclusions are immediately apparent from simple visual inspection. The unsubstituted urea substituent is typically *more* potent than other R<sub>3</sub>s across various R<sub>1</sub>, R<sub>2</sub>, R<sub>4</sub>, R<sub>5</sub>, as indicated by the entire



**Figure 8.** SRP map representation of cathepsin S inhibitor series  $R_3$  substituents, color-coded by average hCatS inhibition difference ( $pIC_{50}$ ), and axes sorted by AlogP.

column below that substituent along the X axis containing only blue rectangles. The two columns immediately to the right of this blue column are significantly more red; as the axes are sorted by similarity, the analyst immediately knows that closely related substituents impart this contrasting aggregate activity profile. In fact, the less active substituents are also ureas, differing merely by substitution with one and two methyl groups. Here, then, simple structural changes make, on average, consistent and large activity differences, thus representing an R-cliff. Using spreadsheets in an attempt to compare all pairs of substituents across all other variable positions would be an incredibly onerous and complex multivariate exercise, whereas SRP maps allow instant visual insight. Other R-cliffs exist in this data set and are also obvious through examination of the SRP map. For example, the oxoacetate and unsubstituted oxamide substituents are distinctly more potent than other related moieties across a range of other substitutions.

An alternative approach to analyzing R-cliffs through use of SRP maps can involve ordering the substituents on the axes, not by similarity, but by a relevant physical property such as lipophilicity. In Figure 8, the  $R_3$  substituents are reordered by ascending AlogP as a measure of lipophilicity. In the case of these large CatS inhibitors, replacement of lipophilic groups with groups making specific interactions is a desirable goal from the standpoint of druggability. Examination of this SRP map quickly highlights the unsubstituted oxamide and urea as substituents with low AlogP that are consistently active, in contrast to neighboring substituents with similar lipophilicity. Indeed, as discussed

previously, this oxamide and the essentially equipotent oxoacetate are structurally similar; in this alternate view, though, the value of the oxamide as a less lipophilic substituent is clear. As well, the unsubstituted urea is typically more potent than the cyclohexyl substituted urea, a much more lipophilic substituent. Taken together, these results indicate the potential of carefully chosen substituents to effect specific interactions in the P4 pocket of the enzyme, as has been reported previously.<sup>25,35</sup>

## CONCLUSIONS

The impetus behind this work was the need to understand the effects of different substituents in analog series containing more than two variable sites. While the previously reported SAR maps provide an ideal interface for visualizing SAR in fewer dimensions, more complex substitution patterns are more difficult to grasp without extensive interactive manipulation. SRP analysis addresses this critical need by compressing all the relevant information into a single plot and revealing patterns that are not immediately evident through other means.

Some caution, however, is required in constructing and interpreting these maps. First, the R groups in a SRP pair can be significantly different. In such cases, the observed R-cliffs are not “real” activity cliffs in a classical sense, but apparent ones; the structural changes may be too large to allow the analyst to draw meaningful conclusions. Second, as shown in the examples above, the interpretation of SRP maps also depends on the molecular representation and precise measure of similarity that is



used to sort the substituent axes (be it structure or property similarity), just like other methods for analyzing activity cliffs. R-cliffs are easier to identify when the relevant substituents are adjacent to each other on the map, which obviously depends on the sorting method. What our tool provides is convenience in interactively switching from one property to another and in manually placing the substituents in any desired order. Finally, while the method can be used with any arbitrary structure as a core, it is probably most useful with a classical definition of a scaffold, i.e., a core structure that is decorated using the same synthetic strategy in a linear or combinatorial way. Despite these caveats, we hope that the examples described above will convince the reader that this method represents a novel and useful addition to the arsenal of the medicinal chemist for understanding and rationalizing SAR.

## AUTHOR INFORMATION

### Corresponding Author

\*Phone: (215)628-6814. Fax: (215)540-4619. E-mail: dagrafio@its.jnj.com.

## ACKNOWLEDGMENT

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

## REFERENCES

- (1) Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug. Discovery* **2003**, *2*, 665–668.
- (2) Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25–26.
- (3) Maggiora, G. M. On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
- (4) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating structure-activity landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.
- (5) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations in structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (6) The emerging concepts of activity landscapes and activity cliffs and their role in drug research. ACS National Meeting, August 22–26, 2010, Boston, MA. Bajorath, J.; Maggiora, G.; Lajiness, M., organizers.
- (7) Guha, R.; Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 645–658.
- (8) Guha, R.; Van Drie, J. H. Assessing how well a modeling protocol captures a structure-activity landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716–1728.
- (9) Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
- (10) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.
- (11) Lunkine, E.; Wawer, M.; Wassermann, 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.
- (12) Shanmugasundaram, V.; Maggiora, G. M. *Characterizing property and activity landscapes using an information-theoretic approach*. 222nd American Chemical Society National Meeting, Chicago, IL, United States, Cinf-032; American Chemical Society: Washington, DC, 2001.
- (13) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
- (14) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of structure-activity relationship determinants in analog series. *J. Med. Chem.* **2009**, *52*, 3212–3224.
- (15) Sisay, M. T.; Peltason, L.; Bajorath, J. Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series. *J. Chem. Inf. Model.* **2009**, *49*, 2179–2189.
- (16) Bemis, G. W.; Murcko, M. A. The properties of known drugs. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (17) Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. Towards a systematic characterization of the antiprotazoal activity landscape of benzimidazole derivatives. *Bioorg. Med. Chem.* **2010**, *18*, 7380–7391.
- (18) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.
- (19) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational analysis of multi-target structure-activity relationships to derive preference orders for chemical modifications toward target selectivity. *ChemMedChem* **2010**, *5* (6), 847–858.
- (20) Izrailev, S.; Agrafiotis, D. K. Variable selection for QSAR by artificial ant colony systems. *SAR QSAR Environ. Res.* **2002**, *13*, 417–423.
- (21) Agrafiotis, D. K.; Cedeño, W. Feature selection for structure-activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45*, 1098–1107.
- (22) Agrafiotis, D. K.; Cedeño, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903–911.
- (23) Cedeño, W.; Agrafiotis, D. K. Using particle swarms for the development of QSAR models based on k-nearest neighbor and kernel regression. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 255–263.
- (24) Seierstad, M.; Agrafiotis, D. K. A QSAR model of hERG binding using a large, diverse and internally consistent training set. *Chem. Biol. Drug. Des.* **2006**, *67* (4), 284–296.
- (25) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold explorer: an interactive tool for organizing and mining SAR data spanning multiple chemotypes. *J. Med. Chem.* **2010**, *53* (13), S002–S011.
- (26) 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.
- (27) 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.
- (28) 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.
- (29) Cepeda, M. S.; Lobanov, V. S.; Farnum, M.; Weinstein, R.; Gates, P.; Agrafiotis, D. K.; Stang, P.; Berlin, J. A. Broadening access to electronic health care databases. *Nat. Rev. Drug Discovery* **2010**, *9*, 84.
- (30) Gupta, S.; Singh, R. K.; Dastidar, S.; Ray, A. Cysteine Cathepsin S as an immunomodulatory target: Present and future trends. *Exp. Opin. Ther. Targets* **2008**, *12*, 291–299.
- (31) 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.
- (32) Villadangos, J. A.; Ploegh, H. L. Proteolysis in MHC Class II antigen presentation: Who's in charge?. *Immunity* **2000**, *12*, 233–239.
- (33) Chapman, H. A. Endosomal proteolysis and MHC Class II function. *Curr. Opin. Immunol.* **1998**, *10*, 93–102.

- (34) 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.
- (35) Wiener, J. J. M.; Sun, S.; Thurmond, R. L. Recent Advances in the Design of Cathepsin S Inhibitors. *Curr. Top. Med. Chem.* **2010**, *10*, 717–732.
- (36) Wiener, J. J. M.; Wickboldt, A. T.; Wiener, D. K.; Lee-Dutra, A.; Edwards, J. P.; Karlsson, L.; Nguyen, S.; Sun, S.; Jones, T. K.; Grice, C. A. Discovery and SAR of novel pyrazole-based thioethers as cathepsin S inhibitors. Part 2: Modification of P3, P4, and P5 regions. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2375–2378.
- (37) Link, J. O.; Zipfel, S. Advances in Cathepsin S inhibitor design. *Curr. Opin. Drug Discovery* **2006**, *9*, 471–482.
- (38) Thurmond, R. L.; Sun, S.; Karlsson, L.; Edwards, J. P. Cathepsin S inhibitors as novel immunomodulators. *Curr. Opin. Invest. Drugs* **2005**, *6*, 473–482.
- (39) Leroy, V.; Thuraiatnam, S. Cathepsin S inhibitors. *Expert Opin. Ther. Patents* **2004**, *14*, 301–311.
- (40) 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. US-2009-099157-A1, 2009.
- (41) 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-1H-pyrazolo[4,3-c]pyridines, as modulators of cathepsin S. US-2008-0200454-A1, 2008.
- (42) 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. US-2009-0118274-A1, 2009.
- (43) 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.
- (44) 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.
- (45) McGrath, M. E.; Palmer, J. T.; Bromme, D.; Somoza, J. R. Crystal structure of human Cathepsin S. *Protein Sci.* **1998**, *7*, 1294–1302.
- (46) 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.
- (47) 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.
- (48) 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.
- (49) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869–15872.
- (50) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215–1221.
- (51) Agrafiotis, D. K.; Xu, H. A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 475–484.