2D Depiction of Fragment Hierarchies

Alex M. Clark*

Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada H3A 2R7

Received September 16, 2009

Drug discovery projects often involve organizing compounds in the form of a hierarchical tree, where each node is a substructure fragment shared by all of its descendent nodes. A method is described for producing 2D depiction layout coordinates for each of the nodes in such a tree, ensuring that common fragments within molecular structures are drawn in an identical way, and arranged with a consistent orientation. This is achieved by first deriving a common numbering scheme for common fragments, then using this scheme to redepict each of the molecules, one fragment at a time, so that common fragments have common depiction motifs. Once complete, the distinct root branches can be overlaid onto each other, after which all of the fragments and whole molecules have a common layout and orientation. Several methods are described for preparing visual representations of molecular structure hierarchies alongside activity information. Combining high level tree display and structure depiction showing common features readily facilitates insight into structure−activity relationships.

INTRODUCTION

A cornerstone of ligand optimization in drug discovery research is the comparison of activity and property data for a collection of molecules which are related by similar structural cores.[1] In order to rationalize the relationship between structure and activity, it is often beneficial to organize the structures in the form of a hierarchical tree. Structures with a common core fragment are arranged in branches, in which each parent fragment is a smaller, pared-down substructure that is common to each of the children. If the tree is well constructed, considerable insight can be gained regarding which core fragments and which peripheral substituents are responsible for the properties of interest, such as binding affinity against some number of protein targets, toxicity, and relevant physical properties.

Given a collection of arbitrary molecular structures, there is typically no single unambiguous way to arrange them in a tree such that each parent node is a substructure of all its children. If the molecules happened to be synthesized in a particular sequence, such as by introducing a variety of substituents in a stepwise fashion to some number of similar core fragments, it may be sensible to produce a fragmentation tree which is based on the synthetic procedures. Or, if a set of common scaffolds is already known, it may be sensible to start with these scaffolds as the root fragments, and from these, construct the descendency hierarchy. If the collection of molecules has significant structural similarity, but no specific information about common substructures is available, then algorithms exist for estimating which parts of a structure are most *scaffold-like*. These can be used to generate a fragmentation sequence which can then be expressed as a tree.[2]

While the fragmentation techniques have been well developed, one part of the process remains conspicuously undocumented. A fragmentation tree of molecules with similar common scaffolds is significantly less informative if the 2D coordinates of the molecular diagrams are not arranged such that common fragments are drawn with a common layout and uniform orientation. Molecular structures are often sketched by chemists showing a standard orientation when structures are added to a database, but this is not always the case.

For a small group of structures, or for a rare showcase example, it is typically not a Herculean task to manually ensure that all of the fragments are properly drawn and aligned. For larger collections, or if charged with the task of regularly regenerating this data, one would at least begin to feel like Sisyphus. Unfortunately, even if a good 2D depiction algorithm is already available, the procedure of generating 2D layout coordinates that honor common fragment ancestry, with group layout decisions made in the context of the global optimum, is not trivial.

In this work, a method will be described for obtaining 2D coordinates suitable for structure diagrams, which are chosen for both independent aesthetic appeal and for clearly showing the common substructure patterns by means of layout and orientation. Also discussed are higher level presentation methods for making use of the structure−activity information, which is contained in the fragmentation hierarchy.

In the following section, methods will be described for generating a fragment tree, mapping sibling fragments onto each other in an optimal way, then using this mapping to guide the 2D depiction process. The results section will describe methods of presenting this information in the context of examining structure−activity relationships.

METHODS

1. Fragment Generation. For each molecule in the data set, it is necessary to propose a scheme whereby the molecule is peeled away in some number of steps, such that the last remaining fragments are the most *scaffold-like*. The definition

* To whom correspondence should be addressed. Phone: +1 514 393 1055. Fax: +1 514 874 9538. E-mail: aclark@chemcomp.com.

10.1021/ci900350h
Published on Web 12/29/2009

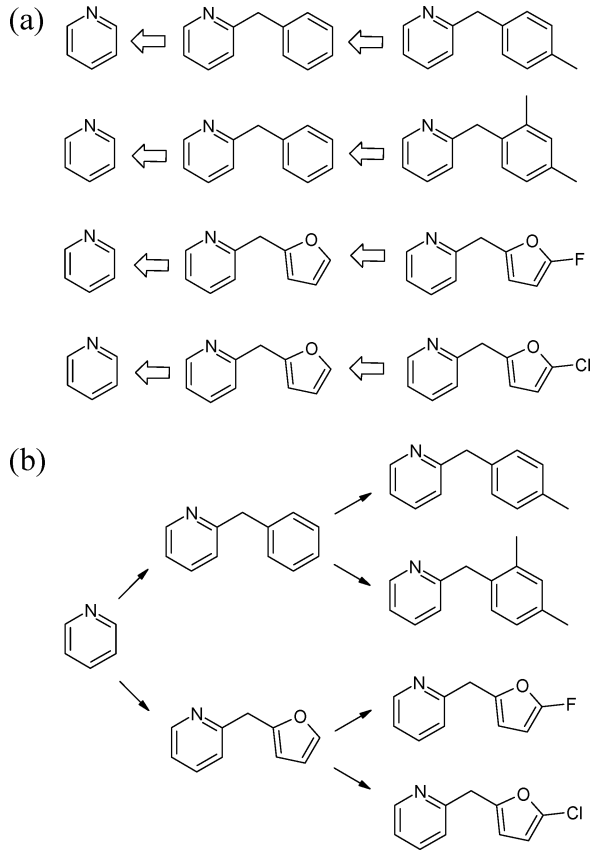
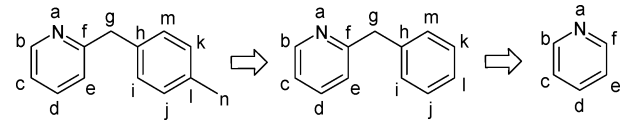


Figure 1. (a) shows the fragmentation sequence for four molecules which share a common root, which are grouped according to common fragments, while (b) shows the corresponding tree representation, which subsumes adjacent, analogous fragments into individual nodes.

of scaffolds, and the fragmentation sequences which relate them to the whole molecules, may be adapted to suit the data. The method published by Schuffenhauer and Ertl et al. has been found to consistently produce agreeable results for pharmacologically relevant molecules.[3] These decomposition rules are such that the last remaining fragments are typically those that have been used as scaffolds and tend to be common substructures for a particular drug discovery campaign. One modification is applied, which is to collapse sequential breaking of fused aromatic ring systems into a single fragmentation step. This fragmentation method is used exclusively in this work, but the layout methods described are general and apply equally well to any scheme which is appropriate to the data.

The fragments generated for each molecule need to be organized in the form of a tree. This can be accomplished by first producing canonical string representations for each fragment (e.g., SMILES).[4] The canonical strings are arranged from smallest to largest for each molecule, and sorted by alphabetical order. In this way, for a given row and column, the fragments above or below are considered to be part of the same tree node if all the fragments up to and including that column are equal, which is illustrated in Figure 1.

In the following section, it will be necessary to know the correspondence between each fragment and its parent fragment. This can be accomplished by assigning an arbitrary label to each atom in the original molecule, and retaining these labels as the fragmentation proceeds, for example



2. Mapping Root Branches. For any node position in the tree which has more than one child molecule, the objective is ultimately to obtain an ideal 2D depiction layout and orientation such that analogous atoms in related molecules are placed in the same location, in order to make visual perception of the common features as easy as possible. Before this can be accomplished, it is first necessary to find an atom-to-atom mapping scheme that relates each of the equivalent substructure fragments to each other.

When all of the fragments are nonsymmetrical and nondegenerate, this process is not especially complicated, since there will only be one correct mapping between any pair of fragments. When the potential mappings are degenerate, it is desirable to consider an ensemble of possible solutions, and find a combination that leads to the most compatible topological overlap of whole molecules. Consider the example shown in Figure 1, where four molecules share a common root fragment, the pyridine core. Pyridine itself has permutational symmetry, which means that when two such fragments are mapped onto each other, it is necessary to choose a mapping from multiple pairwise combinations. Figure 2 illustrates the need to choose wisely. An arbitrary numbering scheme is chosen for the pyridine root fragment of the first molecule. The atom assigned as number 6 has a benzyl substituent for the descendent fragment. For the second molecule, there are two valid numbering schemes. As shown for the first option, the descendent fragment also has a benzyl substituent at atom number 6, for the second option, the benzyl fragment is attached to atom number 2.

For this simple example, it is clear that a suboptimal atom-to-atom mapping scheme would eventually lead to molecules being superposed on top of each other in a manner that misleads rather than elucidates the common structural features. Unfortunately, the number of possible solutions grows combinatorially, which means that it is not viable to rely on an algorithm which examines every combination to find the best topological overlap. While adequate results could be achieved by using a greedy algorithm, that is, assign the possible graph automorphisms for each fragment sequentially and pick the case that best matches the previous assignments, the method described here uses the Unary Quadratic Optimization (UQO) method,[5] which has previously been applied to the problem of common scaffold detection.[6] Each fragment can have one possible assigned permutation, and the quality of the ensemble can be reasonably expressed as the sum of compatibility scores between each pair of assignments. The interaction energy between each pair of states not belonging to the same fragment is expressed in terms of the compatibility between the substituents which are not part of the shared substructure. Figure 3 shows two molecules that share a commonly assigned benzene fragment, with the tentative mapping numbers shown. The similarity of the substitution patterns is obtained by a breadth first search of the molecular graph, starting from each direct substitution point. For the first shell, the molecule in 3a is described as [2-N, 6-C], for a singly bonded nitrogen atom adjacent to position 2 and a singly

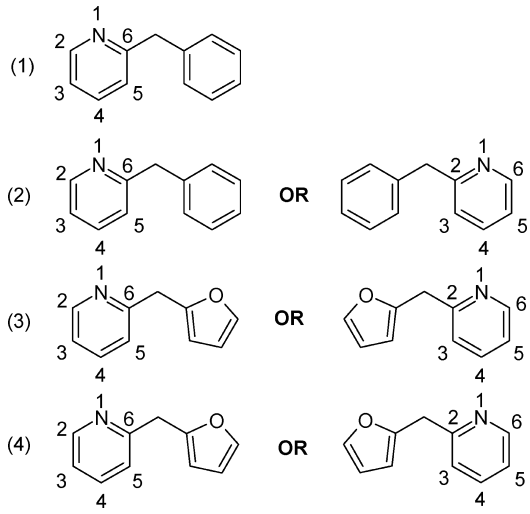


Figure 2. For the pyridine root fragment, there are multiple ways to map the branches to each other, which are not of equal quality.

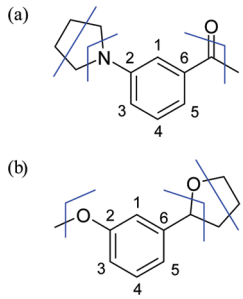


Figure 3. Shell patterns for substituents of disubstituted benzene fragments.

bonded carbon atom at position 6. The molecule in 3b is described as [2-O, 6-C]. For the second shell, the patterns are [2-C, 2-C, 6=O, 6-C] and [2-C, 6-O, 6-C], while the third shell is [2-C, 2-C] and [6-C, 6-C]. Substitution points are assigned as being the topologically closest mapped atom, while the highest bond order is taken if there is a choice of pathways.

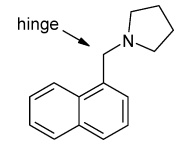
Similarity between two fragments is determined by examining one shell at a time, and matching up pairs of atoms that are assigned to the same mapped atom. The score is computed as follows: 1 point for each pair, 1 point for both being carbon or both being heteroatoms, and 1 point for having the same bond order. When there are multiple mapping permutations, the best matched set of pairs is selected. The score for each shell is divided by N, which is the shell number, which starts at 1 and is incremented with each step in the breadth first search. This method of comparing substituents is simple and fast, and emphasizes the value of sharing any substition at analogous positions, with those having similar topology and heterosubstitution being further favored.

The assignment of a common numbering system is done recursively. For each unique root fragment, its root nodes are considered as a group. For each of these groups, the UQO equation is formulated and solved, which yields a unique solution for each member of the group.

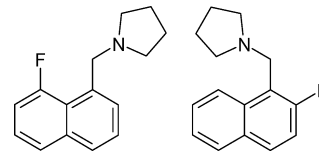
Once the atom equivalency is established for the group of substructures which share a tree node, these assignments are stored in the fragments themselves, and also percolated through down the tree. The descendent nodes therefore share the common mapping within their partial substructures. The process is continued for each child fragment.

Figure 4 illustrates the mapping procedure in a stepwise fashion. In Figure 4a, the assignment has been completed for the pyridine root fragment, which has been mapped for all four molecules. As can be seen, the mapping is such that the substitution is always at the atom labeled as number 6, which is an optimal solution. In Figure 4b and c, the process has been recursively applied to each of the distinct children of the root fragments, and a large proportion of the structures have been mapped. These two steps are carried out separately, so the numbering systems used for the benzyl and the methylfuran substituents are not related. Figure 4d shows the completed mapping, where tree singletons are assigned arbitrary numbers for the remaining atoms.

3. Colligative Depiction. The method for 2D layout of atom coordinates which we use in this work normally operates by searching for a globally optimum aesthetic ideal, which is the appropriate goal for depiction of individual molecules.[7] When examining a set of molecules, such as the fragmentation tree we have described thus far, it is important to be able to represent common fragments with a common depiction motif. While this is often the case for unconstrained depiction, it cannot be relied upon, since arbitrary substituents may influence overall layout decisions in conflicting ways. Consider the following simple fragment, which has a prominent degree of freedom in the form of a methylene linker:



Two different derivatives of this fragment clearly indicate that a free-for-all depiction is not the most suitable way to produce a common layout, since the common substructure fragments which correspond to the overall aesthetic ideal are no longer superimposable:



To solve this problem, we make use of the tree structure and the mapping numbering scheme, to build up the fragment depictions sequentially. While the results must honor the requirement of common layout for common fragments, degrees of freedom within the constituent fragments are dealt with in a way that emphasizes overall layout aesthetics.

As a first step, we submit each of the input molecules to the depiction layout algorithm, without constraints, and store the new coordinates within the molecule datastructure. While many of the structures will have their coordinates further modified by subsequent depiction layout procedures, the coordinates obtained by unconstrained layout are used during the following procedure for clustering purposes.

The tree is analyzed recursively, starting with each root fragment. For each sibling node, the 2D coordinates of the

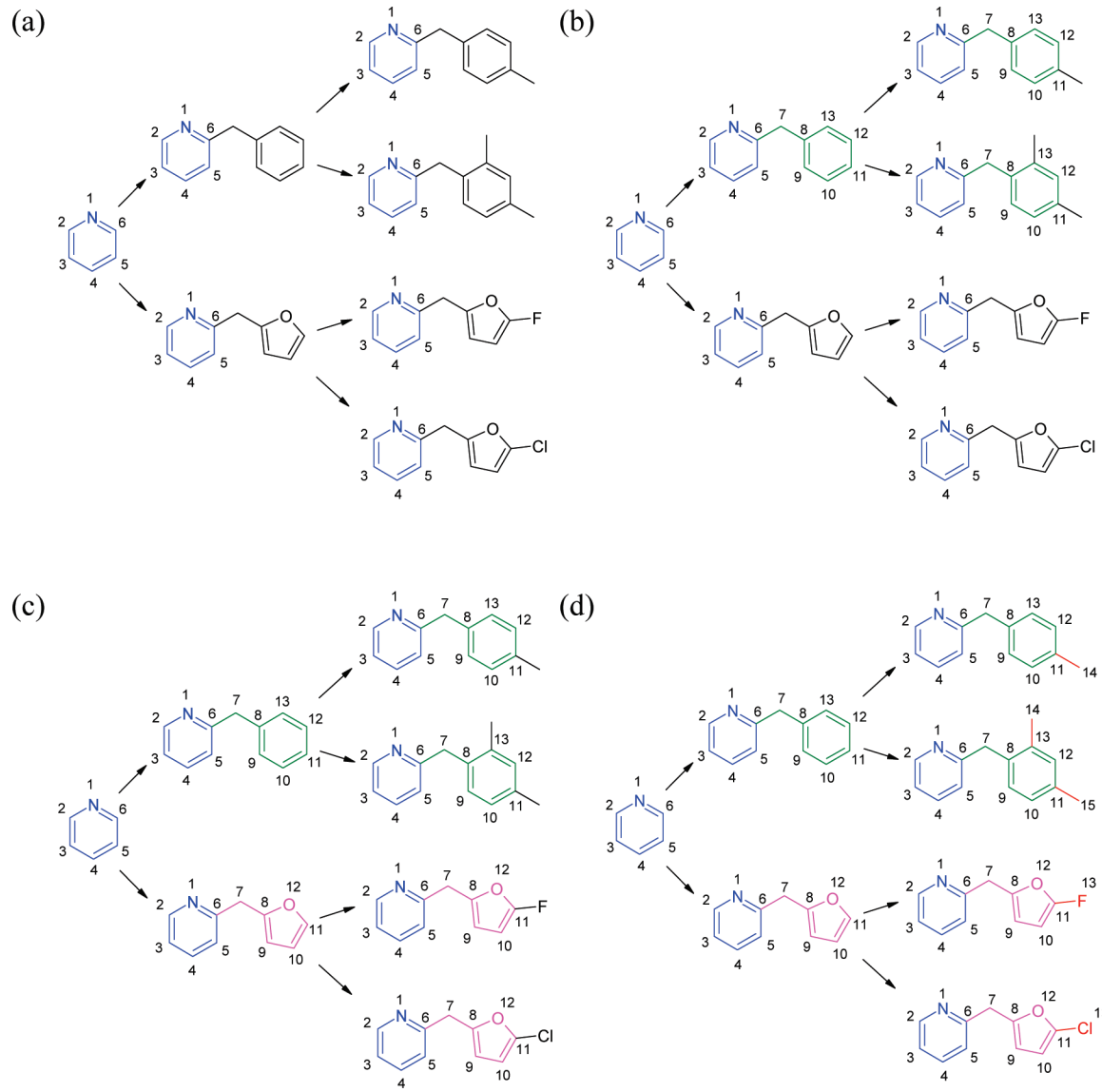


Figure 4. Stepwise assignment of a common numbering system, for a fragmentation tree consisting of 4 molecules.

fragments comprising the node are extracted, and clustered in a greedy fashion. Starting with the first set of coordinates as the first cluster, each subsequent set of coordinates is added to the same cluster as any fragment for which its coordinates can be superposed with an rmsd of less than 0.1 Å or to a new cluster if none. The common atom mapping scheme derived in the previous step is used to enforce pairwise atom matches. If there is more than one cluster, which can occur when fragments have layout degrees of freedom such as rotational symmetry or aliphatic chains, the cluster with the largest number of members is selected. The first set of coordinates in this cluster is used as the definitive reference. For fragments that were found in the same cluster, a translation/rotation/inversion is sufficient to map to the reference coordinates. This transformation is applied simultaneously to all fragments belonging to a particular molecule, not just the fragment under consideration.

Fragments which are not in the reference cluster require more sophisticated treatment. In each case, the whole molecule for the fragment is resubmitted to the depiction layout algorithm, with a special constraint: the coordinates for the fragment are submitted as a *preblock*, which forces the layout algorithm to construct the rest of the 2D molecule around the predefined fragment, whose coordinates are taken from the reference fragment. This process has been previously described for drawing molecules with common scaffolds, and is used to obtain coordinates which are optimal for the collection of molecules, if not necessarily each individual molecule.[6] The coordinates from all of the fragments corresponding to the whole molecule are updated accordingly.

Figure 5 illustrates these distinct steps. The fragmentation patterns for 3 molecules are shown, each of which has been subjected to unconstrained depiction layout. Consider the fragments in column (b), which form a single node in the resulting tree, and are directly descended from the common root fragment shown in (a). The fragments (1,b) and (2,b) can be superposed onto each other, and so they are part of the same cluster of size 2. Fragment (3,b) cannot be superposed, and so it is assigned to its own cluster of size 1. The largest cluster consists of (1,b) and (2,b). The coordinates of (1,b) are used as the reference point. For the fragment (2,b), the coordinates are superposed onto the reference. For fragment (3,b), the coordinates are obtained by redepiction, using the reference coordinates as the preblock. Figure 6

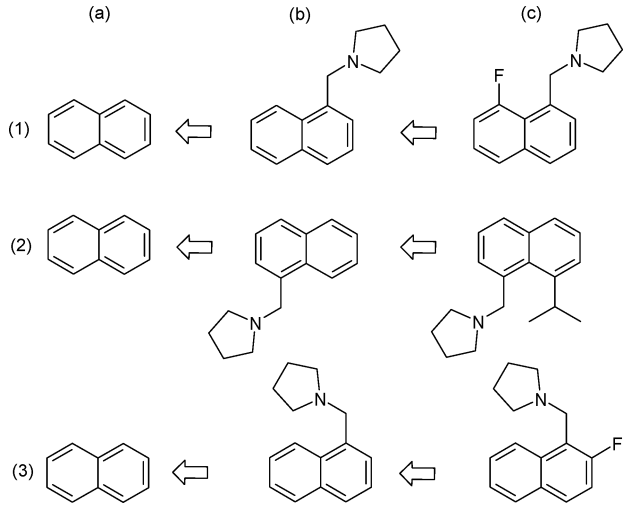


Figure 5. Fragmentation tree where molecule coordinates are obtained by unconstrained depiction layout.

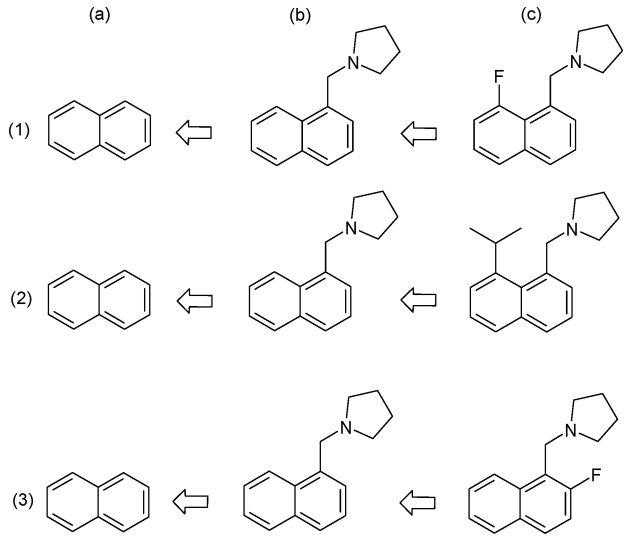


Figure 6. Same fragmentation tree as shown in Figure 5, after the colligative depiction layout procedure has been applied.

shows the same molecules and fragments after this process has been applied.

4. Interbranch Orientation. At the completion of the previous step, each of the root branches has been arranged and oriented to show the common features within the branch. There is as yet no frame of reference for comparing the structures within *different* root branches, because there is no common ancestor and hence not even a partial common mapping system. There is, however, quite a high likelihood that the root branches are structurally related, and so it is useful to devise a scheme to orient them in a common way by means of translation/rotation/inversion.

To do this, we take advantage of the fact that the constituent fragments of the root branches are depicted in a very constrained way. Their 2D shape now encodes a significant amount of information, which is generally not the case for unconstrained depiction layout. Therefore, it is quite viable to search for a single transformation for each whole branch which maximizes the overall shape overlap of the 2D structures.

Since the orientation is a relatively imprecise step, it is sufficient to use a greedy algorithm, rather than a more rigorous clustering method. One begins by first defining the *reference* set to be the root branch with the largest number of constituent molecules. The *subject* set is the root branch with the second highest molecule count. For the subject set, an orientation is selected such that its combined 2D shape is most similar to that of the reference set. The orientation is applied to the subject set, and then it is merged into the reference set. A new subject set is selected, and the algorithm proceeds until all of the root branches have been processed.

To compare the shapes of two sets, each of the molecules in each set is first translated so that the center of the root fragment is at the origin. A grid is defined, which is large enough to capture the bounds of each set as it is rotated around the origin.[8] For each set, grid values are defined by addition of a Gaussian function, for each atom in each molecule:

$$f(r) = \frac{1}{N} e^{-|r|^2}$$

where r is the distance from the grid point to the center of the corresponding atom and N is the number of molecules in the set. The two grids are now directly comparable, and their similarity can be computed:

$$\text{similarity} = \sum_{ij} (|R_{ij} - S_{ij}| + |R_{ij}^{\text{het}} - S_{ij}^{\text{het}}|)$$

where i and j iterate over each of the grid points, R is the grid for the reference set and S is the grid for the subject set. R^{het} and S^{het} are analogous grids where only heteroatoms are plotted. Lower values indicate more similar grids.

To find the most similar orientation, the subject set is varied by rotating about the origin in increments of 5°; inverting along one of the axes; and translating by [dx, dy] $\in \{-2, 1, 0, 1, 2\}$, which makes for a total of 3600 evaluations.

Figure 7 shows the resulting orientations and the corresponding grid pattern for each of 9 root fragments from a database of DHFR inhibitors, which contains 397 structures.[9] As can be seen, at the completion of the orientation step, the layout coordinates of molecules with different root fragments are approximately comparable, and the 2D coordinates now encode significant information about shared structural features.

RESULTS

1. DHFR Inhibitors. Once the fragment tree has been generated and the appropriate steps have been taken to ensure that the layout and orientation of the molecules shows the common fragmentation patterns, there are a number of ways to display this information visually. The most direct method is to display the fragmentation tree as a hierarchical structure, with each root node being a distinct entity.

To make examination of the tree practical from a user-interface perspective, it makes sense to allow nodes to be interactively opened or closed, so that parts of the tree can be displayed as necessary. Because molecules are usually grouped together according to common fragmentation patterns, it is helpful to plot summary information in the vicinity of each node. This elicits information about the activities of

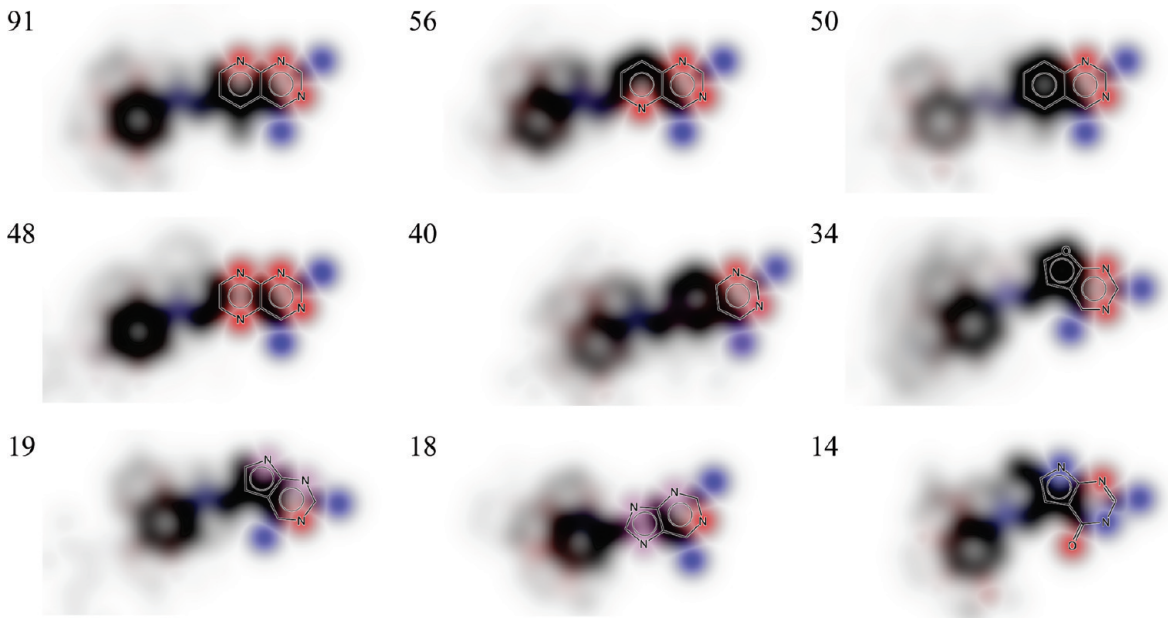


Figure 7. Orientation patterns for 9 root fragments. The numbers indicate how many structures are contained in each group. Hydrogen bond donors are tinted blue; acceptors are tinted red.

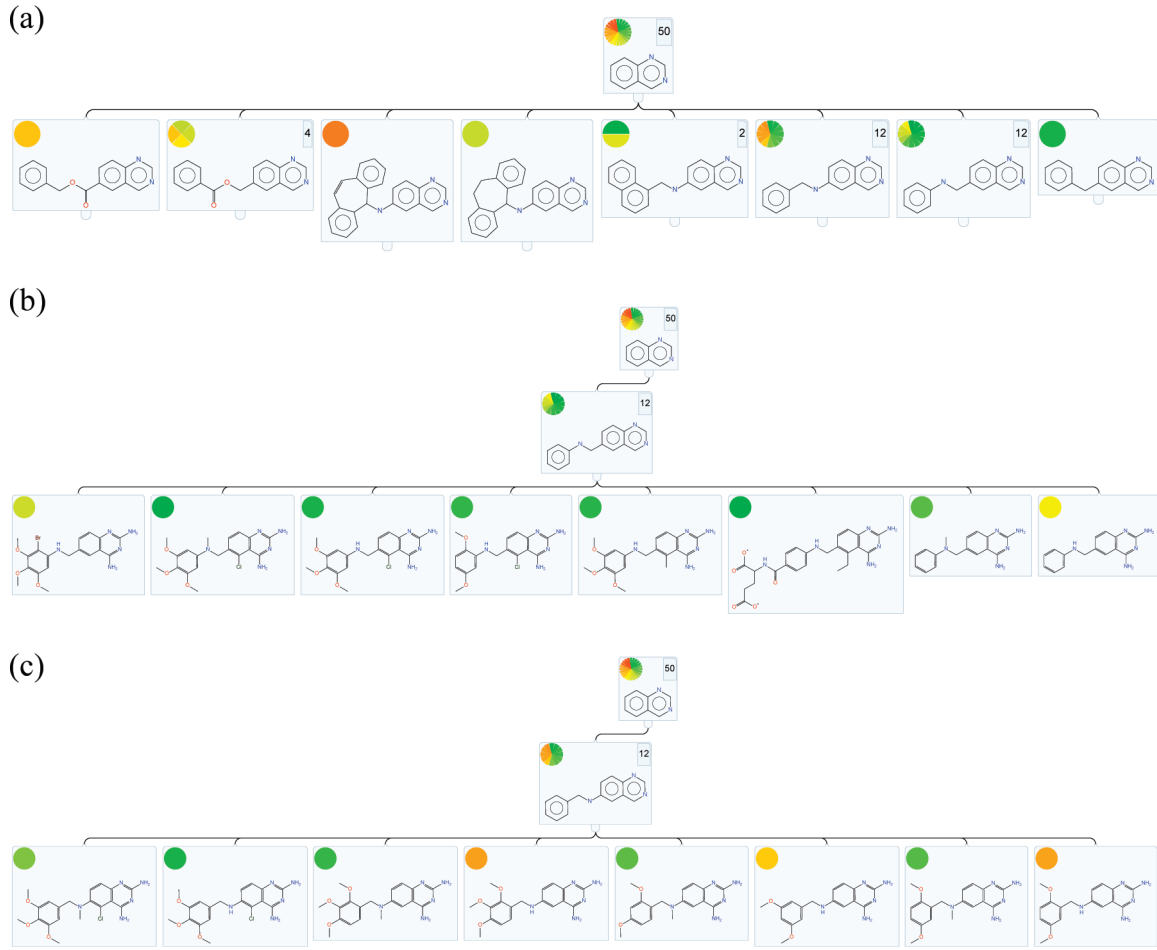


Figure 8. (a) Single root fragment and a selection of its immediate descendents. (b and c) Selected child of the root fragment and a selection of their descendents.

the molecules which are represented by each node, which can be used to identify trends in activity data.

Figure 8 shows several subsets of the fragmentation hierarchy for a data set compiled to show activity against various species of DHFR, using the data set described in the previous section.[9] Figure 8a shows a single root fragment, and several of the immediately descended fragments. Each of the nodes provides an indication of the activity values against *Pneumocystis carinii* DHFR of the constituent molecules by means of a colored *pie chart* display at the

top left, which is divided up into a number of *slices*, one for each constituent compound. The color scheme used is green for highly active (<0.1 μM), red for inactive (>10 μM) and yellow-orange for moderately active ($\sim$1 μM). As can be seen from the color pattern of the root fragment itself, the constituent molecules exhibit a wide spectrum of activity.

By traversing the hierarchy from the root downward, it is possible to reveal groups of structurally related molecules, which show distinctive inhibition patterns. The branch shown in Figure 8b shows a high degree of consistency, which suggests that the inhibition capabilities can be ascribed to the core fragments. The branch shown in Figure 8c, on the other hand, shows a mixture of active and inactive constituents, which suggests that the inhibition properties are determined to a large extent by the substituents, rather than just the scaffolds.

While the hierarchy display is appropriate for detailed analysis of the fragments themselves, it is impractical to examine more than several dozen structures at once with a typical display device. An alternate approach is to display the fragmentation hierarchy of all of the input structures in the form of a dendrogram, where all of the root branches are considered to be descended from a point in the center of a circle, and the branches extend outward from the center.[10] Each of the leaf nodes is located at a position about the rim.

To ascribe significant structural information to this layout, the ordering of the descendency pattern is chosen to maximize the similarity of any two adjacent leaf nodes. In preparation, fingerprints are calculated for each leaf node structure, using the GpiDAPH3 scheme.[11] A *reference* node is selected, which is initially set to the leaf node with the highest average Tanimoto similarity to all other nodes.

The tree is then traversed recursively, according to the following procedure: (1) For each of the immediate child nodes, assemble a list of all of the leaf nodes of which it is an ancestor. (2) Select next-in-sequence child node to be that whose descendent leaf nodes has the highest average Tanimoto similarity to the reference node. (3) If the selected node is a leaf node, set the new reference node to this. Otherwise, recursively process the child node. (4) Goto step 2, until all child nodes have been selected.

Once the ordering of the tree is determined, the actual layout of the dendrogram is straightforward. The final position of each node is obtained by evenly spacing about the rim of a disk, in order of tree traversal. The parent of all root nodes is considered to be the center, and the remaining nodes are arranged in between.

If the background is chosen so that each pixel within the enclosing circle is assigned a color derived from the value of the average activity of the molecules belonging to the nearest fragment node, it is possible to gain considerable insight into structure−activity relationships. This is shown in Figure 9, which plots activity data against *P. carinii*, using the red/yellow/green activity scheme described in the previous section. It is possible to quickly identify which regions of the fragmentation tree have consistently high or low activity, suggesting that the core fragment is responsible for these properties, as well as branches whose activity is quite mixed, suggesting that the substitution patterns are primarily responsible.

Figure 10 shows a cutout section of the dendrogram, where several of the structures of leaf nodes are indicated. The leaf

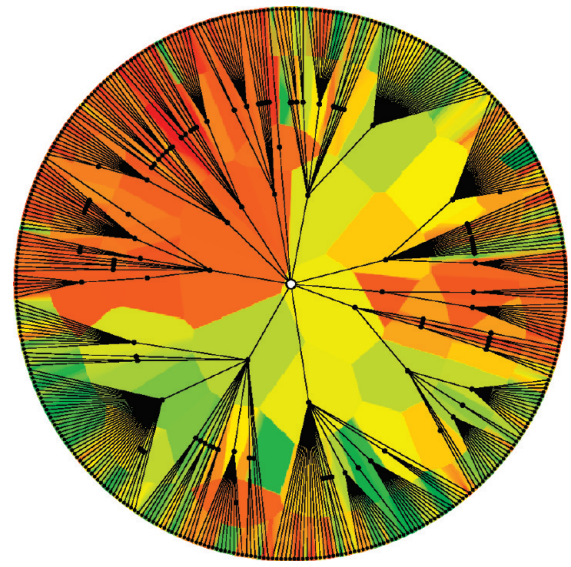


Figure 9. Dendrogram-style fragment tree layout, arranged onto a disk. The background color reflects activities of close-by fragments.

node clustering is evident by comparison of structures (a) and (b), which are descended from a different root fragment, but are almost identical structures, differing only by an additional chlorine atom on (a) and an additional nitrogen into the fused heterocycle scaffold of (b). The structures (c) and (d) have the same core fragmentation sequence, and denote the two ends of a block of very similar compounds clustered together, sharing high activity against *P. carinii* DHFR.

The dendrogram display is very compact and information-rich. Combined with an interactive user interface which allows facile examination of the structures, which are represented by the fragment nodes, a significant amount of structure−activity information can be inferred from the layout and color-coding.

2. COX-2 Inhibitors. A collection of 467 structures with corresponding activity against the COX-2 enzyme[9] was submitted to the fragmentation and depiction sequence described within. Each of the structures was embedded in 3D, and its atom ordering scrambled randomly to ensure that there was no residual bias from the input sketch. Figure 11 shows a selection of substructure fragments, upon which most of the structures are based. In each case, a certain motif is observed: central 5-membered ring, usually a heterocycle, is decorated by two adjacent substituents, which are either 5- or 6-membered rings, a mixture of phenyl, pyridyl, or a 5-membered heterocycle. Some of the fragments show an additional substituent on the opposite side of the central ring, which is a consequence of the fragmentation ordering.

The figure shows each of the fragments with a common orientation, which is chemically intuitive. In all cases, the central heterocycle bears the two adjacent substituents at approximately the 10 o'clock and 12 o'clock positions. The common orientation is a consequence of the layout methods described in this work: each of the branches of the fragmentation tree are depicted in such a way as to trade off aesthetic depiction against showing common substitution patterns in analogous positions, taking degeneracy into account. Because this data set has a high degree of structural similarity within the core fragments and the substitution

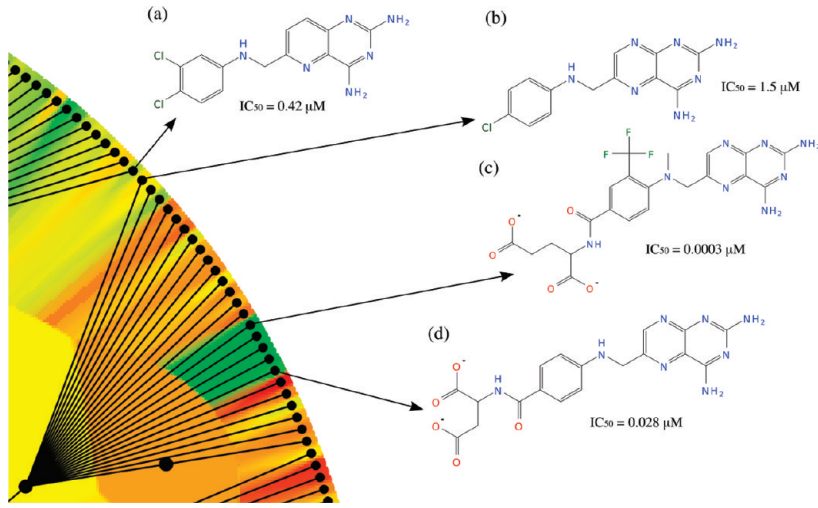


Figure 10. Portion of a dendrogram fragment display, with selected leaf nodes indicated.

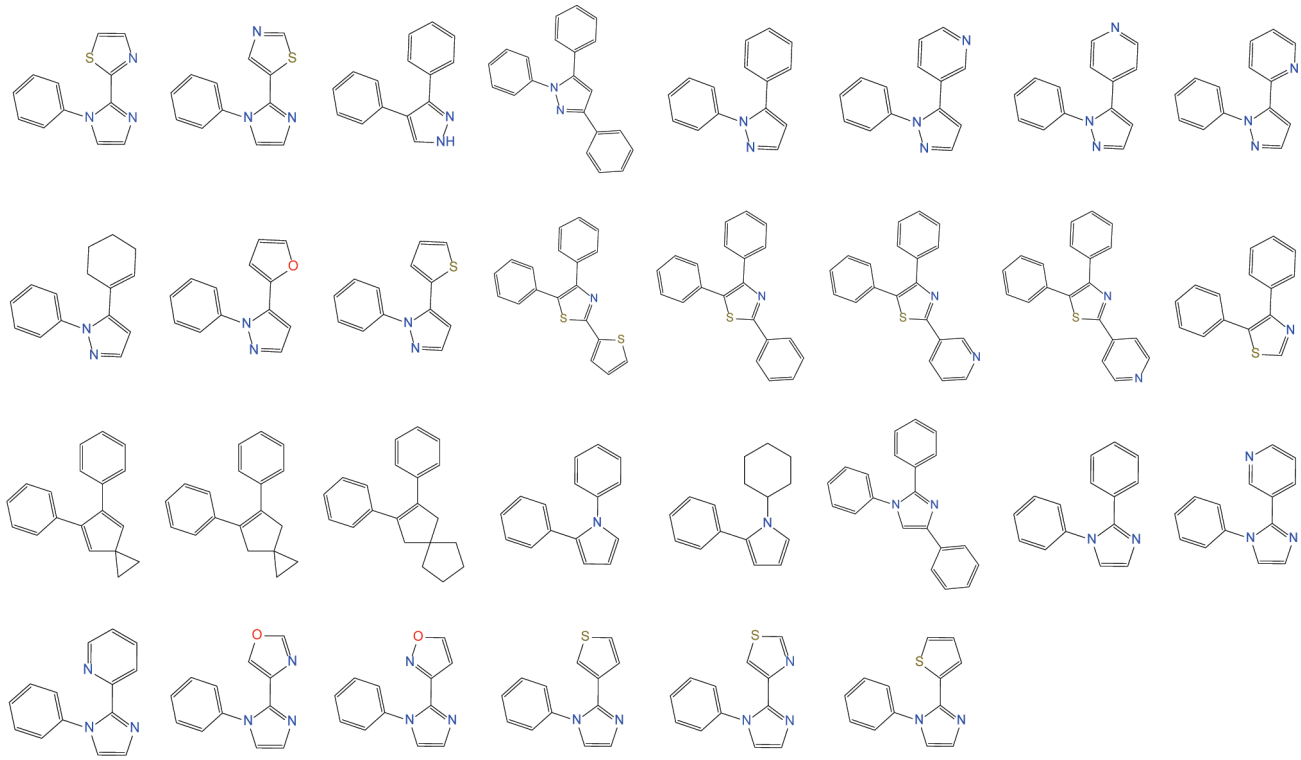


Figure 11. Selected core fragments of COX2 inhibitors.

patterns, the final spatial orientation using 2D structures reveals the alignment shown.

Figure 12 shows a portion of a radial dendrogram, composed from the hierarchy of COX-2 inhibitors, using the same algorithm as described in the previous section. One sub-branch is shown in its entirety, and the structures of six of these compounds are shown. All of these compounds are based on the same 1,2-diphenylimidazole core, with a trifluoromethyl substituent in the 4-position. The activity values vary from 68 to 0.03 μM. Compounds a and b, as well as all of the compounds that are arranged in between them in the linear cluster, feature a sulfonamide substituent in the same position on the phenyl ring at the 1-position, and all of them exhibit high activity, as can be seen by the bright green color coding. The remainder of the compounds in this sub-branch have a methylsulfonyl substituent in place of sulfonamide, and their activity is much more varied and is determined by the substitution pattern on the phenyl ring at the 2-position. For example, compounds c and d, which have a small substituent in the meta position, exhibit strong activity, while compounds e and f, which have larger substituents, exhibit weaker activity.

Each of the compounds shown has a meta-substituted phenyl ring at the 2-position, and in each case, one of the meta substituents is depicted so that it is oriented at approximately the 2 o'clock position. While an unconstrained depiction algorithm will be likely to make these layout choices on the grounds of congestion, the symmetrical phenyl substituents are explicitly mapped to each other on an atom-by-atom basis, in order to maximize the topological similarity of the interstructure mapping. A single depiction is chosen for the precursor fragment, and applied to all descendents, which *ensures* that the most similar substituents are placed at a consistent position.

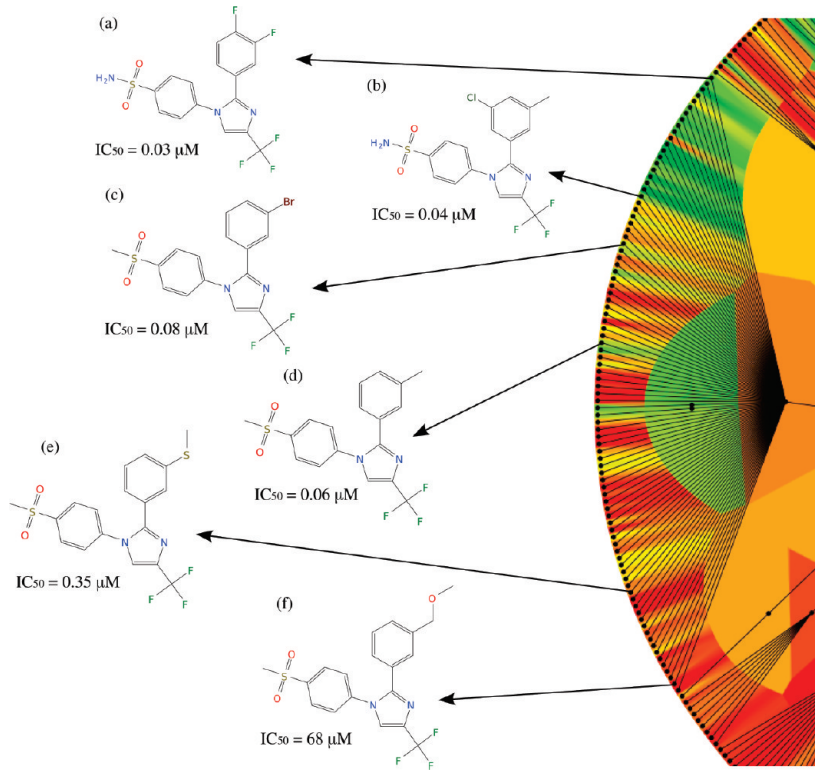


Figure 12. Structurally related COX2 inhibitors.

3. Performance. All algorithms described in this work have been implemented in SVL (Scientific Vector Language), running in MOE (Molecular Operating Environment).[12] The total processing time for the data set of DHFR inhibitors described in part 1 (397 compounds) is approximately 16 s, measured on an Intel 2.0 GHz 32-bit processor. The time taken for each step breaks down roughly as follows: initial depiction (4 s), fragment generation (2 s), deriving common mapping (4 s), colligative redepiction (2 s), and branch orientation (4 s). For the COX-2 inhibitors described in part 2 (467 compounds), the total time taken was approximately 50 s, of which 33 s were devoted to deriving the common mapping scheme. The rate limiting steps scale linearly in proportion to the number of input molecules, with the exception of the mapping step, in the presence of symmetry-containing fragments. An abundance of degenerate partial fragments causes the UQO portion of the common mapping algorithm to dedicate additional computational resources to finding an optimal mapping. This is the principal reason why the COX-2 data set takes longer to analyze than does the DHFR data set, since the DHFR inhibitors are largely based on heterocyclic ring blocks with low symmetry, while the core fragments of the COX-2 inhibitors contain an abundance of substituents, such as phenyl rings, that have rotational symmetry.

CONCLUSION

A method has been demonstrated for treating scaffold-like fragmentation trees so that common ancestor fragments are depicted and oriented in a consistent way that makes common structural features readily evident to the observing chemist. The algorithm operates without supervision and produces aesthetically desirable results using a combination of new and preexisting techniques.

With the molecule layout and orientation method established, software applications that present fragmentation trees become significantly more valuable to medicinal chemists. There is no longer a requirement to sketch the input molecules in any particular way, nor is it necessary to postprocess the structures to elicit insight into the structure−activity information that is encoded in such trees.

The value of presenting structural data in this way has been demonstrated, by using high-level and medium-level hierarchical tree views, which combine constrained clustering with activity color-coding, and the ability to examine individual structures that have a common depiction and orientation. Such views are effective ways to examine structure−activity data, and the algorithms that have been described essentially remove the manual effort required to produce them.

Use of such visualization and grouping tools can allow the user to make intuitive selections of compounds for further development. For example, if a set of compounds which has balanced activity and diverse structure is necessary for producing a QSAR validation set, either the tree or dendrogram presentation methods could be used to identify such a subset, either by visual inspection or automated selection. Alternatively, when considering hypothetical new compounds to add to the collection, it may be informative to depict them under the same conditions as for the existing compounds, and insert them into their appropriate locations within the hierarchy according to the ordering scheme used by the visualization method. Examination of the patterns exhibited by neighboring structures may aid in qualitatively assessing the expected properties of the compound.

Future research will involve developing new ways to present information-rich, easily comprehensible graphics in a concise area, for example, single computer screen, projected

slide or printed page, as well as developing interactive user interfaces for arranging and querying content data. The hierarchical atom-to-atom mapping system developed in this work has been used primarily for depiction purposes, but the assignment can also be used to differentiate scaffolds vs substituents at any level within the hierarchy, which opens up possibilities for fragment-based QSAR and clustering studies.

REFERENCES AND NOTES

(1) (a) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937. (b) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658. (c) 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. (d) Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008**, *51*, 5342–5348. (e) 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. (f) 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.

(2) (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. (b) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193. (c) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442–446. (d) Lounkine, E.; Bajorath, J. Core Trees and Consensus Fragment Sequences for Molecular Representation and Similarity Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1161–1166.

(3) 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.

(4) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(5) Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins: Struct. Funct. Bioinf.* **2009**, *75*, 187–205.

(6) Clark, A. M.; Labute, P. Detection and Assignment of Common Scaffolds in Project Databases of Lead Molecules. *J. Med. Chem.* **2009**, *52*, 469–483.

(7) (a) Clark, A. M.; Labute, P.; Santavy, M. 2D Structure Depiction. *J. Chem. Inf. Model.* **2006**, *46*, 1107–1123. (b) Clark, A. M.; Labute, P. 2D Depiction of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1933–1944.

(8) A suitable grid spacing is 1 Å, assuming that the depiction layout is scaled such that the average bond distance is approximately 1.5 Å.

(9) (a) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554. (b) Broughton, M. C.; Queener, S. F. Pneumocystis carinii Dihydrofolate Reductaste Used To Screen Potential Antipneumocystis Drugs. *Antimicrob. Agents Chemother.* **1991**, *35*, 1348–1355. (c) Rosowsky, A.; Hynes, J. B.; Queener, S. F. Structure−Activity and Structure−Selectivity Studies on Diaminoquinazolines and Other Inhibitors of *Pneumocystis carinii* and *Toxoplasma gondii* Dihydrofolate Reductase. *Antimicrob. Agents Chemother.* **1995**, *39*, 79–86. (d) Gangjee, A.; Elzein, E.; Queener, F.; McGuire, J. J. Synthesis and Biological Activities of Tricyclic Conformationally Restricted Tetrahydropyrido Annulated Furo[2,3-*d*]pyrimidines as Inhibitors of Dihydrofolate Reductases. *J. Med. Chem.* **1998**, *41*, 1409–1416. (e) Nelson, R. G.; Rosowsky, A. Dicyclic and Tricyclic Diaminopyrimidine Derivatives as Potent Inhibitors of *Cryptosporidium parvum* Dihydrofolate Reductase: Structure−Activity and Structure−Selectivity Correlations. *Antimicrob. Agents Chemother.* **2001**, *45*, 3293–3303.

(10) Agrafiotis, D. K.; Bandyopadhyay, D.; Farnum, M. Radial Clustergrams: Visualizing the Aggregate Properties of Hierarchical Clusters. *J. Chem. Inf. Model.* **2007**, *47*, 69–79.

(11) (a) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Diversity* **2006**, *10*, 311–332. (b) Any common molecular fingerprint scheme which distinguishes between pharmacologically relevant features may be used instead.

(12) The source code for the algorithms described within, and those referred to, is packaged as part of MOE and may be examined and used under the terms of the MOE user license, which is available from the Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada. http://www.chemcomp.com (accessed 1 June, 2009).

CI900350H