JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

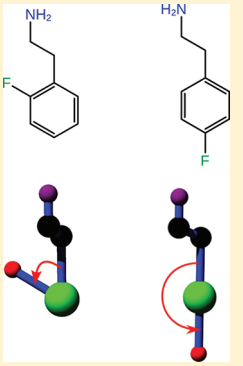
Improving Similarity-Driven Library Design: Customized Matching and Regioselective Feature Trees

J. Robert Fischer,[†] Uta Lessel,[‡] and Matthias Rarey*,[†]

[†]Center for Bioinformatics (ZBH), University of Hamburg, Hamburg, Germany

[‡]Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

ABSTRACT: Reduced graph descriptors, like feature trees, are frequently applied in cases where the relative arrangement of functional groups is more important than exact substructure matches. Due to their ability to deal with fragmented molecules, they are well-suited for fragment space search and library design. We recently presented LoFT, a novel focused library design approach based on feature trees. During evaluation two drawbacks of the reduced graph descriptor were discovered: First, regioisomeric substructures cannot be distinguished in feature tree mappings which results in a large information loss especially when connecting R-groups to cores. Second, the automatic matching procedure might result in undesired alignments, since the knowledge on what is considered as core by the user is not taken into account. In the following, we will present two approaches to overcome those drawbacks. The generation of the feature trees is modified, so that different arene substitution patterns can be recognized and a customized matching is introduced, allowing the user to determine the parts of the query, where the reagents are allowed to match. Subsequently we investigate the improvements on library design by reviewing the design scenarios which were already used for the evaluation of LoFT.



INTRODUCTION

Reduced graph descriptors[1−5] are commonly used for similarity searching and clustering of chemical compounds, representing the molecular graph typically by a tree structure. This preserves the overall topology of functional groups within the compound without the necessity to consider conformational space. Thus reduced graphs complement traditional two- and three-dimensional (2D and 3D, respectively) molecular descriptors.[6] Due to the fact that 3D descriptors have to deal with conformational flexibility, 2D descriptors have been shown to be more successful in ligand-based similarity searching and clustering.[7−9] Nevertheless, the application of 2D descriptors often results in molecules, which are structurally too similar to the query molecule. Recently, reduced graph descriptors have shown their ability for scaffold hopping,[6,10] making it possible to discover novel compounds with a different central core structure using a known active compound as the query molecule.[11]

To calculate the pairwise similarity of two reduced graph descriptors, several methods are available, e.g., pseudo-SMILES,[2] an edit distance approach,[6,12] and the approximate alignment of the two tree structures following a dynamic programming approach.[13−15]

On one hand, finding the best alignment of the tree nodes is computationally more expensive, as for example, the application of a similarity coefficient to 2D vector representations.[16] On the other hand, the approach is more suitable to be used with virtual fragment spaces.[17] These consist of small chemical fragments and a synthetically motivated rule set.[18−20] The fragments can be combined by applying the rules and connecting their so-called link atoms. If a strict hierarchical comparison procedure is used, then the similarity values of the resulting product compounds can be obtained without explicitly building them.[17,21] This way it is possible to traverse efficiently a large chemical subspace. If fragment spaces are based on rule sets of known synthesis protocols, then the chemical feasibility of products out of the space is more or less guaranteed.[22] Accordingly, fragment spaces have been created as collections of combinatorial library synthesis protocols.[10,22] Consequently, the feature trees fragment space approach was extended to multi-objective focused and general library design and implemented in a tool called LoFT.[21] Together with macromolecular physicochemical descriptors, the feature tree descriptor is incorporated into a multi-objective scoring scheme, in order to achieve compounds similar to a given query molecule as well as having desired physicochemical property profiles.

During the evaluation of LoFT, two main problems of reduced graph descriptors were observed. First, in feature trees, regioisomeric structures cannot be distinguished: To generate a node labeled tree from a molecular graph, cyclic dependencies have to be broken. Hence rings are considered as building blocks and condensed into tree nodes. If the nodes are labeled with the steric and physicochemical properties, it results in identical node labels for regioisomeric substructures (Figure 1). Second, the automatic matching procedure might result in alternative alignments, because the replacement for the core structure of the query molecule is too dissimilar. In the following we will describe two approaches to overcome these problems: A penalty function for the distinction of arene substitution patterns and a customized

Received: January 13, 2011
Published: August 18, 2011

matching procedure which allows the user to constrain the possible matchings.

We will further demonstrate the effects of these modifications on three library design test scenarios.

METHODOLOGY

During lead identification the generality of reduced graph approaches is highly desired. Here, an automatic matching procedure is required to select compounds or small sublibraries for testing. But if the task is to optimize a hit which was detected before, then the scientist might have a specific matching of the hit and the corresponding query in mind or wants to include additional information about the binding mode of the hits. Furthermore, in many cases different placements of the core structure are possible depending on its substitution. So, for the design of a focused combinatorial library, it is essential that all products fit in the same orientation. Otherwise there might be combinations of substituents which do not match at all. This is why it is important to be able to direct the matching process. And in contrast to a reactant-based design approach where the core mapping is fixed, the matching procedure can be applied to scenarios, where only specific parts of the query should be matched by certain reagents. For example, using a core with three open valences, only one of the corresponding link atoms should be fixed to a specific mapping. The extension to distinguish between regioisomers allows to add further information, especially if specific substitution patterns are known to be relevant for binding. Instead of pre- or postprocessing the input and output of the library optimization or using additional or different descriptors, controlling the outcome of the similarity comparison itself is advantageous. Without losing information about the detected similarity, the user can shift from a general design to a very specific one using the same approach.

In the following we will describe the approaches for the distinction of substitution patterns and the customization of the feature tree matching algorithm in detail.

Regioselective Feature Trees. A feature tree is generated by identifying the building blocks of a molecule and assigning the properties of the corresponding atoms to the feature nodes. Rings are condensed into single nodes, and special cutting rules are used to create nodes from complex ring systems.[13] Acyclic nonterminal bonds are cut, and terminal atoms are assigned to the same node as the atoms they are bonded to. To cope with arene substitution patterns, the feature tree generation is modified as follows: At rings, hydrogen atoms are considered as part of the ring, while all other atoms are separated. This change allows to distinguish molecules with only small changes in the substitution patterns, like at halogenated rings. Figure 2 shows an example for a histamine H_3 receptor antagonist.[23]

Additionally, the alignment procedure has to be modified in a way, that different substitution patterns can be recognized during the calculation of the similarity value. The feature tree comparison procedure (match−search algorithm) can be roughly described as follows:[17,21]

(1) To obtain the overall similarity of two feature trees, all edges are cut and compared pairwise. By cutting one undirected edge of a feature tree, we achieve two disjunct root subtrees. Due to the fact that the comparison of two subtrees depends only on the comparison of their smaller subtrees,[21] the overall similarity can be calculated by comparing all subtrees to each other and finding the best combination between them.

(2) Starting with the root nodes of two subtrees, the algorithm tries to add subsequent nodes to the matching to increase the similarity.

(3) If the matching cannot be extended, then recursive calls are started for all pairwise combinations of subtrees not included yet.

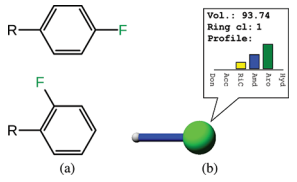


Figure 1. *Para*-fluorophenyl **1** (a, above) and *ortho*-fluorophenyl **2** (a, below) result in the same feature tree with the same node labels (b). Thus, during comparison of the feature trees, these regioisomers cannot be distinguished.

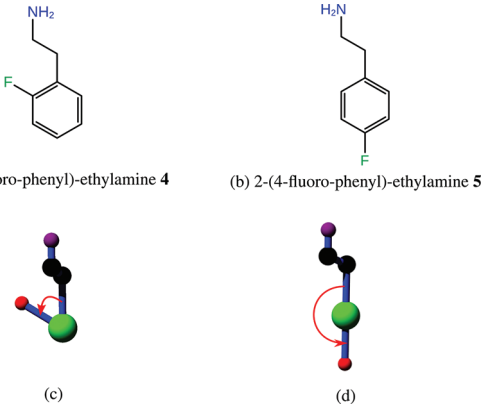


Figure 3. Calculating the angles (red arrows) during the recursive matching procedure. The regiopenalty factor calculated by using eq 1 is multiplied with the match similarity. Thus the two compounds (a and b) can be distinguished during comparison of their corresponding feature trees (c and d).

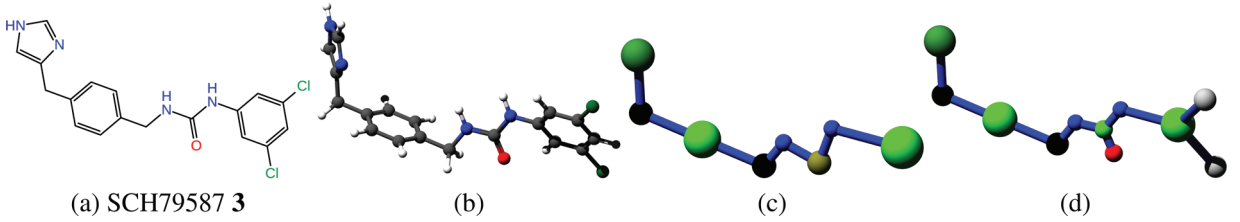


Figure 2. The figure shows a histamine H_3 receptor antagonist (SCH79687)[23] with 2D- and 3D coordinates (a and b, respectively). The corresponding feature tree (c) was generated using the standard rules. Applying the new rules (d), the carbonyl oxygen and the chlorine substituents are assigned to separate nodes.

dx.doi.org/10.1021/ci200014g |J. Chem. Inf. Model. 2011, 51, 2156–2163

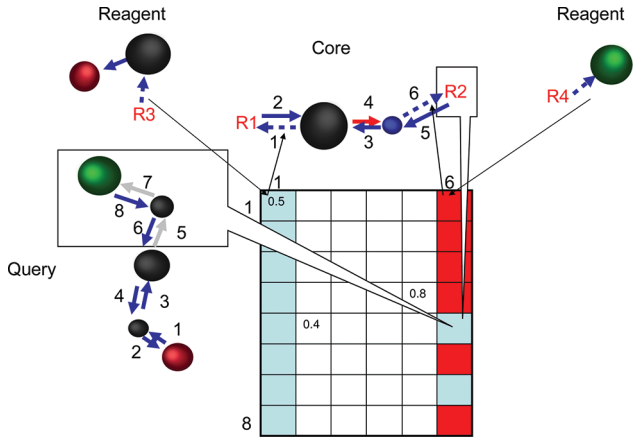
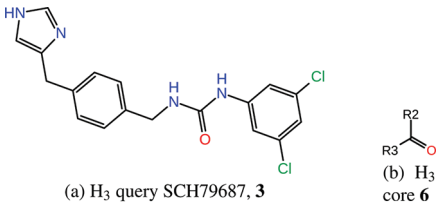


Figure 4. Part of the dynamic programming matrix for the subtree comparison of the core with the directed edges 1−6 and the query with the directed edges 1−8. In this case the reagent link R2 (edge 6) shall be restricted to edges 5 or 7 of the query (gray arrows). In the first column the similarities of the core and the query subtrees are stored if edge 1 is fitted onto edges 1−8 of the query (blue cells). Column 6 contains the prefilled values for the matching of edge 6 and the edges 1−8 of the query. Due to the restriction, only the cells in rows 5 and 7 are prefilled with subtree similarities (blue cells). The other combinations are not allowed. Therefore the corresponding cells are prefilled with high negative values (red cells). This forces the algorithm to select the matches as desired.



(a) H_3 query SCH79687, **3** (b) H_3 core **6**

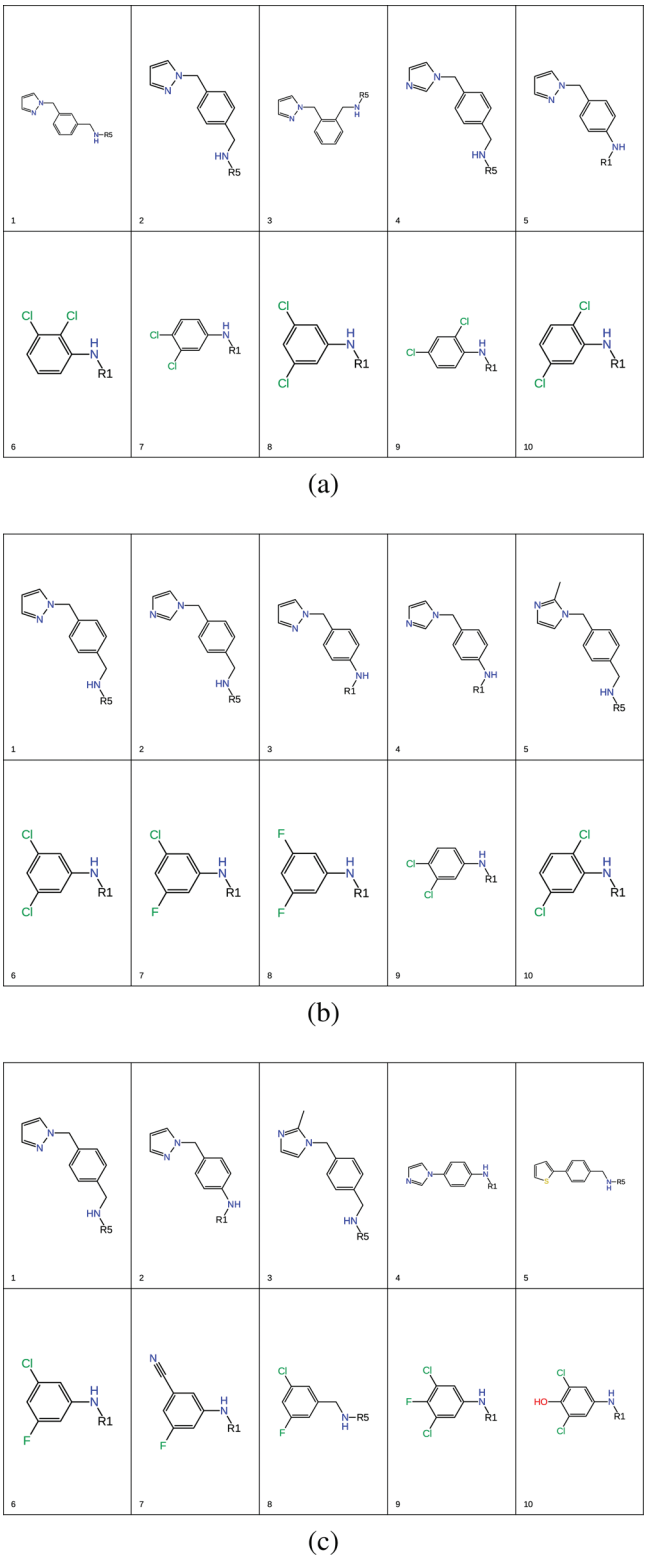
Figure 5. Query and core for the H_3 library.

(4) The similarity value is the highest similarity value of any subtree combination retrieved from the initial splitting.

To distinguish between aromatic rings having different substitution patterns during a recursive call, we have to know the node we are coming from (parent node). Then the angle between the root node of the current call, its parent node, and the considered child node are calculated (see Figure 3). The values computed for the first and the second subtree are inserted in the following equation to compute the regiopenalty factor:

$$1.0 - \left(\text{weight}*\frac{|\text{angle}(a) - \text{angle}(b)|}{\pi}\right) \quad (1)$$

The absolute difference between the angle for nodes a and b is divided by π, which corresponds to the maximum difference of 180°. The resulting penalty factor is then multiplied with the similarity of the corresponding subtrees. Consequently, different substitution patterns are more severely penalized if their underlying subtree is larger. The penalty factor can be adjusted using a weight in the range [0,1]. Moreover no extension of the matching is allowed in this case to prevent the algorithm from matching the substituents only partially. For all substituent combinations, a new recursive call has to be started.



(a)

(b)

(c)

Figure 6. Depiction of the H_3 focused libraries. Reagent linkers are R1 and R5. Reagents 1−5 connect to core linker R3 and reagents 6−10 to core linker R2, respectively. They were retrieved by applying a 200 000 step simulated annealing run taking feature tree similarity to the H3 receptor antagonist SCH79687 as the single objective (a). An additional regiopenalty factor weight of 1.0 was used to generate a focused library (b). To create a more diverse focused library (c), only one reagent from each cluster was allowed.

dx.doi.org/10.1021/ci200014g |J. Chem. Inf. Model. 2011, 51, 2156–2163

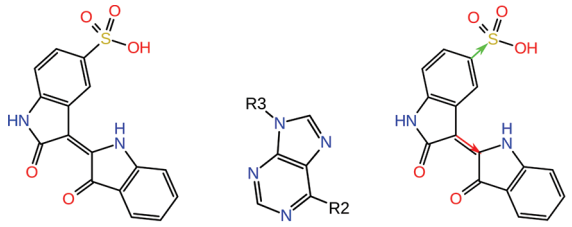


Figure 7. Query 7 (a) and core **8** (b) for the CDK2 library, while (c) depicts the matching customizations of query **7**. The red and green arrows indicate bonds used for customized matching. Only the directional feature tree edge, which is displayed as red arrow on the corresponding bond, is allowed to be matched by the reagents compatible to R2 in the design of the focused library depicted in Figure 8c.

A precondition for this approach is the presence of coordinates. The coordinate of a feature tree node is the centroid of the corresponding atoms' coordinates. If no 2D or 3D coordinates are available for a molecule, then the SDG drawer[24] can be used to generate 2D coordinates. The method can also be applied during fragment-based comparison. Especially if a fragment consists of an aromatic ring or ring system with two or more link atoms as substituents, where the same fragments can be attached, the link discrimination is important to retrieve the correct matching product from the fragment space. Sometimes, connecting two fragments leads to a geometry change, e.g., if the application of a connection rule results in a bond-type change. But in the case of a link atom bonded to an aromatic ring, we can presume that the geometry will not change. Therefore using the link node is suitable for the application of the regiopenalty factor.

In general, the angle calculation can be replaced by, e.g., shortest path counts. Shortest path counts were already introduced for feature trees by Rarey and Stahl.[17] Nevertheless we chose the angle calculation although at least 2D coordinates must be available. By using a continuous measure, the real geometry of the mapped substructures is represented more exactly. This is especially true if mapping five- and six-membered rings. For example, in respect to a query with a 1,4-substituted six-membered ring, a 1,3-substituted five-membered ring is less penalized than a 1,3-substituted six-membered ring. In contrast, the shortest path count between the attachments at both rings is two. Also, substitution patterns of polycyclic substructures can be differentiated without consideration of special cases.

Customized Matching. In many design scenarios a core fragment is already selected to achieve a concrete scaffold replacement. But, especially if the core differs too much from the scaffold of the query molecule in terms of steric or chemical properties, the resulting products might be based on alternative matchings. It is also possible that the positioning of the core depends on its substituents, whereas for a focused library it is important that all products are based on the same matching. For these cases the feature tree comparison algorithm is extended by a constrained matching mode. The user can decide, which edges of the query feature tree can be matched by specific link edges of the reagents.

The feature tree matching algorithm (see Rarey and Dixon[13] for more details) uses a dynamic programming matrix[25] to avoid the recalculation of already computed results. Comparing a query and a product, the comparison matrix of query and core is

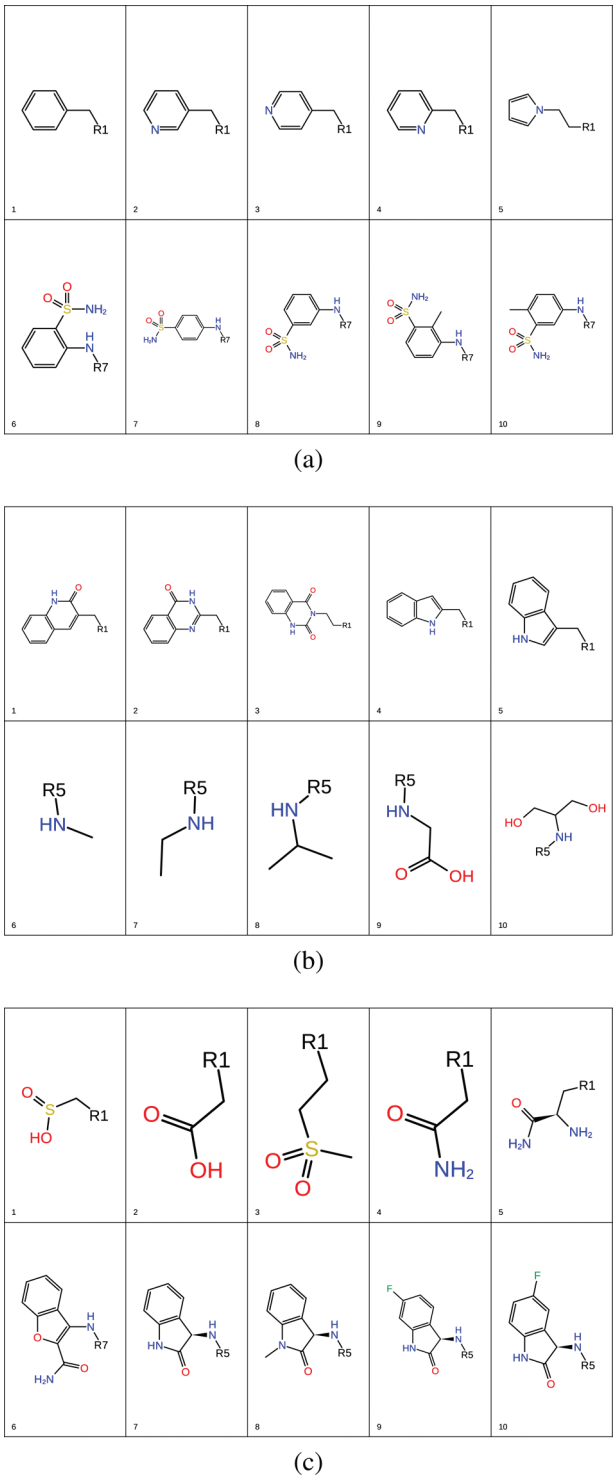


Figure 8. The resulting CDK2 libraries for indirubin-5-sulfonate. Reagents 1−5 connect to the core linker R3, and reagents 6−10 connect to the core linker R2. R1, R5, and R7 are the reagent linkers. The libraries depicted in (a) and (b) were generated by a simulated annealing with 200 000 steps. For library (b) a different starting point was used for the optimization. For focused library (c) the matching was customized (see Figure 7c). The reagents compatible to R2 were only allowed to match the query edge corresponding to the red arrow. Then the core must be placed the other way around by the comparison algorithm (see Figure 9c). Thus all libraries are based on different feature tree alignments (see Figure 9).

dx.doi.org/10.1021/ci200014g |J. Chem. Inf. Model. 2011, 51, 2156–2163

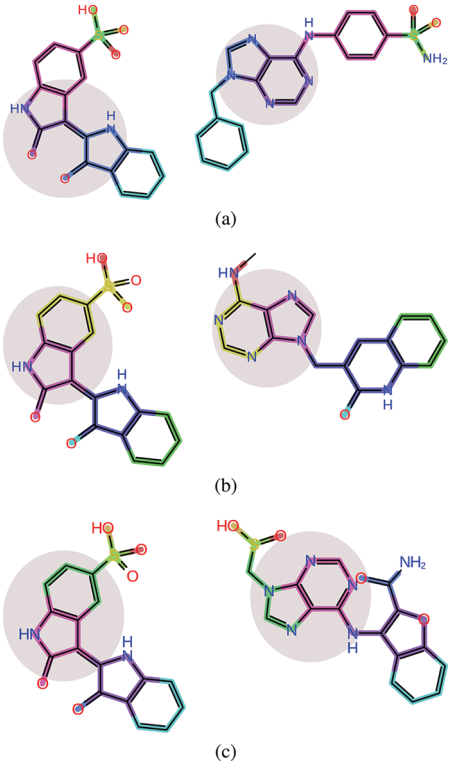


Figure 9. Matching of the query and the products generated from the core and reagents 1 and 6 of the libraries shown in Figure 8, respectively. The matched substructures are highlighted using the same color. Additionally the different core matches are highlighted with gray circles. In (b) reagent 6 is only partially matched.

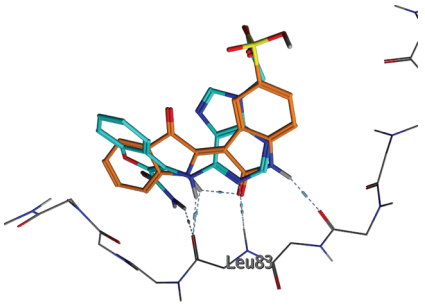


Figure 10. Query 7 and a product from the focused library shown in Figure 8c (core, reagents 2 and 6) were superimposed with ROCS.[35] The aligned product was minimized in the binding site of CDK2 (pdb code 1e9h) with MOE.[36] The kinase hinge pharmacophores of both compounds are matched. In a next design step, the core could be modified in a way that the additional hydrogen bond built up by the query might also be formed by the products.

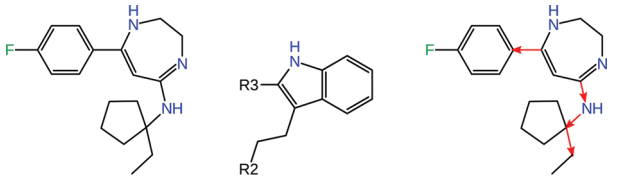
prefilled with the similarity values of the query to reagent comparison (see Figure 4 for an example). Therefore all combinations of reagent link edges and query edges are computed and stored for reuse. The values are inserted into the matrix cells of the corresponding core link edge/query edge combinations. If a combination of a reagent link edge and a query edge is not allowed due to the customized matching, then a high negative similarity value is inserted. Afterward, the computation is started as usual. All core/query edge combinations are used as initial input for the match−search algorithm. To find the best query/product matching, the algorithm recursively tries to add further node pairings (matches) to the matching. During this process, the similarity values of already computed cells are used. Reagent link/query edge combinations which are not compatible to the customized matching are disfavored by their high negative subtree similarities and will usually be overcome by better matching combinations. At the end, it is only important to make sure that all reagent link edges with a matching restriction were properly matched and are not involved in null matches. Therefore, we multiply the computed similarity with the number of properly matched link edges divided by the number of core links. This leads to better performance during a stochastic optimization compared to assigning a score of zero to products with matches which are not allowed.

RESULTS AND DISCUSSION

For analyzing the influence of the regioselective and customized matching extensions described above, we revisit three design scenarios already used for the evaluation of LoFT. To allow for visual inspection, the shown focused libraries have a size of 5 × 5 resulting in 25 products.

Histamine H_3 Receptor. The histamine H_3 receptor is a G protein-coupled receptor modulating the release of histamine[26] as well as other neurotransmitters, such as serotonin and acetylcholine.[27] Therefore H_3 receptor antagonists are being developed to treat a variety of neurological and cognitive disorders.[28] We selected a known H_3 receptor antagonist from the literature with a K_i value of 4 nM on the H3 receptor (SCH79687, see Figure 5a)[23] to illustrate the use of LoFT for the design of a focused urea library, e.g., for hit exploration. Using the core[29] **6** depicted in Figure 5b and 10 314 reagents, we created a urea fragment space. In this case all reagents can connect to both links R2 and R3 of the core.

Performing a simulated annealing run with 200 000 steps and feature tree similarity as the only criterion (regiopenalty 0), for one core link we retrieve the dichloro-substituted phenyls in all variations. More crucial, the reagents connecting to the other core link show different substitution patterns of the phenyl ring, leading to different geometries of the products (Figure 6a). Therefore we set the regiopenalty factor weight to 1.0 to select reagents with more carefully chosen substitution patterns (Figure 6b). Now the reagents selected for the first core link consist of a para-substituted phenyl ring, like the query. The reagent corresponding to the query at the first core link was not part of the urea fragment space. The best reagent for the second core link corresponds now to the substituent of the query. Finally, we apply a diversity criterion allowing only one reagent from each cluster (Figure 6c). The clusters were generated using feature tree similarity considering only the chemical node features and a complete linkage clustering approach with a distance threshold of 0.1.[21] All selected reagents of the first core link have a para-substituted phenyl ring. The 3,5-dichloro-substituted phenyl rings seem to be preferred as reagents for the second core link. Reagent 9 has an additional fluorine and reagent 10 an additional hydroxyl group. Unfortunately, we do not know the activity of the designed molecules, but the selections are plausible according to known structure−activity relationship. Originally, mainly para-substituted [(1H-imidazol-4-yl)methyl] benzamidines and benzylamidines were detected at Schering Plough[30] and led to a 4-benzyl-(1H-imidazol-4-yl) template for H3 receptor antagonists.[23] Furthermore the K_i value of a compound with a



(a) $5HT_{2A}$ query **9** (b) $5HT_{2A}$ core **10** (c) Matching restrictions for query **9**

Figure 11. Query **9** (a) and core **10** (b) for the 5HT2a receptor. Subfigure (c) depicts the matching customizations of query **9**. Only the directional feature tree edges, which are displayed as red arrows on the corresponding bonds, are allowed to be matched by a reagent link edge.

4-benzyl-(1H-imidazol-4-yl) substituent at the first linker and reagent 7 from the library depicted in Figure 6a is 7 nM.[23]

Cyclin-Dependent Kinase 2 (CDK2). CDK2 is a well-known target for cancer treatment, because it plays an important role at two stages of the cell cycle.[31,32] We designed a focused library with core **8** (see Figure 7b)[33] to mimic the known active[34] **7** (indirubin-5-sulfonate, see Figure 7a). The fragment space consists of 11 653 reagents, 10 015 reagents are connectable to R2, and 1638 reagents are connectable to R3.

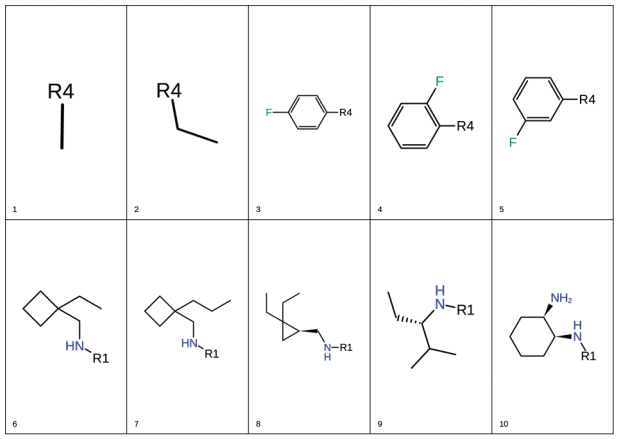
Generating a focused library with similar products to the query molecule, in principle, the core could be placed on the oxindole or the indoxyl substructure of the query. Another possible alignment places the core on both five-membered rings because during feature tree generation, each ring was assigned to its own node, and the edges do not store additional information about the connectivity type. In this context the six-membered ring of the core can be matched with each of the five-membered rings leading to more alternative alignments.

A simulated annealing run with 200 000 steps results in the focused library shown in Figure 8a. In this case the core is mapped on the five-membered rings producing a reasonable alignment from the descriptors point of view (see Figure 9a). Taking a different seed for the optimization, a focused library (Figure 8b) based on a totally different alignment is produced. Here, the core is placed on the oxindole part. The sulfonic acid of the query is only partially matched by reagents 6−8 (see Figure 9b). By restricting the reagents connectable to R2 to match only the edge represented by the green arrow (see Figure 7c), a better reagent selection can be obtained for this core placement (result not shown). Nevertheless, aligning the products with the query, they do not have a kinase hinge pharmacophore at the position where the query binds to the hinge region.

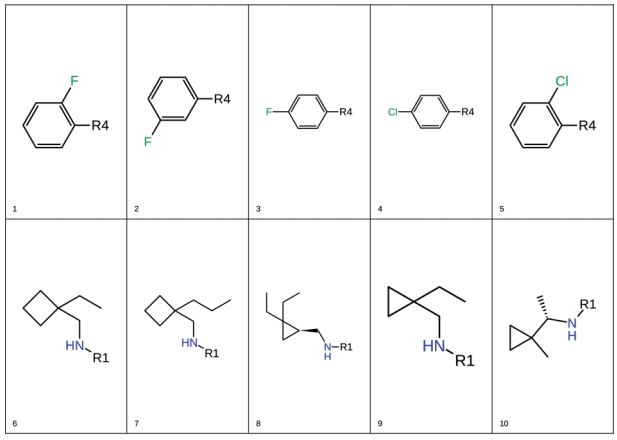
For this purpose the matching procedure has to be customized in a way, that reagents compatible to R2 of the core (see Figure 7b) can be placed only on the feature tree edge corresponding to the bond marked with the red arrow (see Figure 7c). The restriction results in the focused library shown in Figure 8c. Here all reagents connectable to R2 match the specific substructure of the query (see, e.g., Figure 9c), and the restriction leads to focused libraries, with the kinase hinge pharmacophore at the desired position (see Figure 10).

In conclusion, the customized matching extension enables the user to define the orientation of the core and to incorporate his knowledge into the mapping procedure.

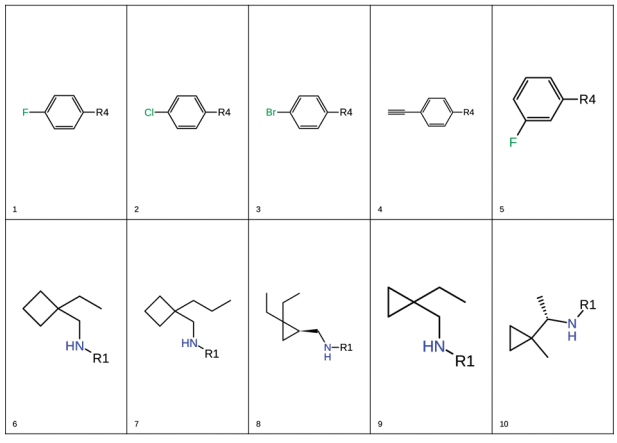
Serotonin $5HT_{2A}$ Receptor. The serotonin $5HT_{2A}$ receptor is a G-protein-coupled receptor, controlling the release of neurotransmitters. The $5HT_{2A}$ receptor antagonists are being developed to treat



(a)



(b)



(c)

Figure 12. The resulting focused libraries for the $5HT_{2A}$ receptor antagonist **9**. Reagents 1−5 with reagent linker R4 connect to the core linker R3 and reagents 6−10 with reagent linker R1 connect to the core linker R2: (a) was generated from a simulated annealing with 200 000 steps, (b) was created applying a customized matching (see Figure 11c), and (c) an additional regiopenalty factor of 1.0 was applied.

e.g., depression, schizophrenia, and sleep disorder.[37] We used the core **10**[38] depicted in Figure 11b and 25 567 reagents (23 486 reagents compatible to link R2; 2081 compatible to link R3) to obtain focused libraries similar to the given query molecule **9** (see Figure 11a).

dx.doi.org/10.1021/ci200014g |*J. Chem. Inf. Model.* 2011, 51, 2156–2163

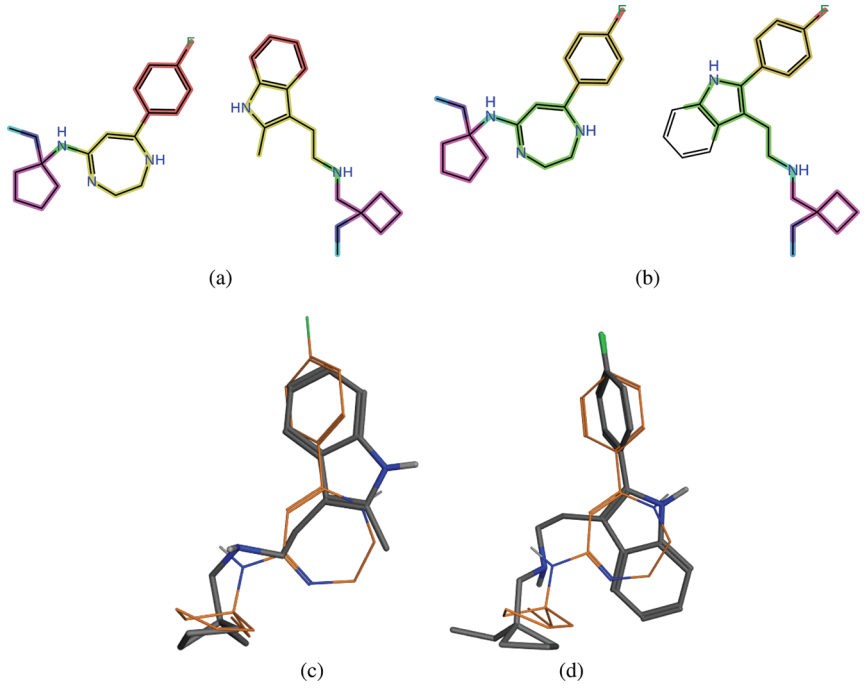


Figure 13. Feature trees matching (a and b) and superpositioning (c and d using FlexS)[39,40] of the query molecule **9** and products of the library shown in Figure 12a generated from the core with reagents 1 and 6 as well as reagents 3 and 6 attached to it, respectively. The basic centers of the products and the query match in both alignment modes.

Separating the terminal nonhydrogen atoms, as described above, leads to a different feature tree for the query, generating one additional node for the fluorine substituent. This leads to two alternative matchings: The first allows the phenyl ring of the indole substructure to match the phenyl ring of the query molecule (see Figure 13a). The second contains a reagent which matches the phenyl ring (see Figure 13b). Both alignments result in high similarity values. Performing a simulated annealing the resulting library contains products showing different alignments. The alignments of the products generated from the core, reagents 1 and 6, as well as from the core, reagents 3 and 6, of the focused library depicted in Figure 12a are illustrated in Figure 13. To avoid this behavior, we constrain the possible matchings of the reagents' link edges. Only the exocyclic bonds of the seven-membered ring are allowed to be matched by the reagent links (see Figure 11c). Thus, the products from this optimization run allow the selected alignment mode (reagents depicted in Figure 12b). The phenyl ring of the reagents at the first core link shows an ortho-, meta-, and para-fluoro- as well as an ortho- and para-chloro-substitution. Here, the feature tree algorithm selects similar substructures without consideration of the substitution patterns. If we use a regiopenalty factor weight of 1.0, the reagents of the first core link consist of para-substituted aryl halides. Additionally a 4-ethinyl- and a 3-fluoro-phenyl ring are selected (see Figure 12c). For the second core link, applying the regiopenalty does not change the selection.

Runtime Considerations. Restricting the allowed matches for the reagent links influences the runtime, because a check of the matches is performed after each similarity comparison. The calculation of the regiopenalty factor does not influence the runtime, but comparing the newly generated feature trees is computationally more demanding, if the corresponding molecule consists of terminal nonhydrogen atoms, which result in additional feature tree nodes. In the case of core **6** (see Figure 5b) and query **3** (see Figure 5a), the runtime is increased from 1:38 to 7:58 min for the simulated annealing with 200 000 steps. The number of core nodes changes from three to four and therefore more possible matchings have to be considered. For core **10** and query **9** (see Figure 11b and a, respectively), the runtime increases from 3:24 to 4:36 min due to the higher number of nodes. On the other hand, using customized matching and regioselectivity reduces the runtime. In general, customizing the matches does not depend on a specific feature tree generation mode. In the example shown, the matching restriction decreases the run time to 4:15 min, because not all query/reagent combinations have to be computed. The additional regioselective search decreases the runtime to 4:05 min, because no match extensions are allowed at aromatic rings. Furthermore, sorting the reagents by their similarity to the best matching query substructure prior to optimization returns a solution within 4 s.

CONCLUSION

We have shown two ways to overcome major drawbacks of reduced graph descriptors during focused library design. Both extensions improve the design process and lead to better results. They work very well together if applied simultaneously. The customization can be used, e.g., to place a core in a user defined way which will not be found by the matching algorithm. Using the restriction to initially sort the reagents by their similarity to specific query substructures, a good sublibrary is obtained very rapidly and can be used for further optimization. The regiopenalty factor for different substitution patterns results in the selection of the most similar reagents tending to the correct substitution patterns according to the query. It is especially valuable in combination with a diversity criterion, e.g., the restriction of the number of reagents from one cluster.

AUTHOR INFORMATION

Corresponding Author
*E-mail: rarey@zbh.uni-hamburg.de.

ACKNOWLEDGMENT

The authors would like to thank Herbert Köppen, Bernd Wellenzohn, and Alexander Weber of Boehringer Ingelheim. We are grateful to Adrian Kolodzik, Tobias Lippert, and Sascha Urbaczek of the ZBH for helpful discussions. Moreover we thank Holger Claussen of BioSolveIT for his support.

REFERENCES

(1) Gillet, V.; Downs, G.; Holliday, J.; Lynch, M.; Dethlefsen, W. Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *J. Chem. Inf. Model.* **1991**, *31*, 260–270.

(2) Gillet, V.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.

(3) Barker, E.; Gardiner, E.; Gillet, V.; Kitts, P.; Morris, J. Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346–356.

(4) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Model.* **1992**, *32*, 639–643.

(5) Fisanick, W.; Lipkus, A.; Rusinko, A. Similarity searching on CAS Registry substances. 2. 2D structural similarity. *J. Chem. Inf. Model.* **1994**, *34*, 130–140.

(6) Harper, G.; Bravi, G.; Pickett, S.; Hussain, J.; Green, D. The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Model.* **2004**, *44*, 2145–2156.

(7) Brown, R.; Martin, Y. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Model.* **1996**, *36*, 572–584.

(8) Brown, R.; Martin, Y. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Model.* **1997**, *37*, 1–9.

(9) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.

(10) Boehm, M.; Wu, T.-Y.; Claussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.* **2008**, *51*, 2468–2480.

(11) Boehm, H.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today: Technol* **2004**, *1*, 217–224.

(12) Birchall, K.; Gillet, V.; Harper, G.; Pickett, S. Training Similarity Measures for Specific Activities. *J. Chem. Inf. Model.* **2006**, *46*, 577–586.

(13) Rarey, M.; Dixon, J. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.

(14) Rarey, M.; Hindle, S.; Maass, P.; Metz, G.; Rummey, C.; Zimmermann, M. Feature Trees: Theory and Applications from Large-Scale Virtual Screening to Data Analysis. In *Pharmacophores and Pharmacophore Search*; Langer, T., Hoffmann, R., Eds.; Wiley-VCH: Weinheim, Germany, 2005; Vol. 32; pp 81−116.

(15) Fischer, J. R.; Rarey, M. SwiFT: an index structure for reduced graph descriptors in virtual screening and clustering. *J. Chem. Inf. Model.* **2007**, *47*, 1341–1353.

(16) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(17) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497–520.

(18) Lewell, X.; Judd, D.; Watson, S.; Hann, M. RECAP−retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(19) Mauser, H.; Stahl, M. Chemical fragment spaces for de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 318–324.

(20) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503–1507.

(21) Fischer, J.; Lessel, U.; Rarey, M. LoFT: Similarity-Driven Multiobjective Focused Library Design. *J. Chem. Inf. Model.* **2010**, *50*, 1–21.

(22) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49*, 270–279.

(23) Aslanian, R.; et al. Identification of a novel, orally bioavailable histamine H(3) receptor antagonist based on the 4-benzyl-(1H-imidazol-4-yl) template. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 937–941.

(24) Fricker, P.; Gastreich, M.; Rarey, M. Automated drawing of structural molecular formulas under constraints. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1065–1078.

(25) Bellman, R. On the Theory of Dynamic Programming. *Proc. Natl. Acad. Sci. U.S.A.* **1952**, *38*, 716–719.

(26) Arrang, J.-M.; Garbarg, M.; Schwartz, J.-C. Auto-inhibition of brain histamine release mediated by a novel class (H3) of histamine receptor. *Nature* **1983**, *302*, 832–837.

(27) Celanire, S.; Wijtmans, M.; Talaga, P.; Leurs, R.; de Esch, I. Keynote review: histamine H3 receptor antagonists reach out for the clinic. *Drug Discovery Today* **2005**, *10*, 1613–1627.

(28) Esbenshade, T.; Fox, G.; Krueger, K.; Miller, T.; Kang, C.; Denny, L.; Witte, D.; Yao, B.; Pan, L.; Wetter, J.; Marsh, K.; Bennani, Y.; Cowart, M.; Sullivan, J.; Hancock, A. Pharmacological properties of ABT-239 [4-(2-{2-[(2R)-2-Methylpyrrolidinyl]ethyl}-benzofuran- 5-yl)-benzonitrile]: I. Potent and selective histamine H3 receptor antagonist with drug-like properties. *J. Pharmacol. Exp. Ther.* **2005**, *313*, 165–175.

(29) Lau, J.; Jeppesen, C.; Rimvall, K.; Hohlweg, R. Ureas with histamine H3-antagonist receptor activity−a new scaffold discovered by lead-hopping from cinnamic acid amides. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5303–5308.

(30) Aslanian, R.; Brown, J.; Shih, N.; wa Mutahi, M.; Green, M.; She, S.; Del Prado, M.; West, R.; Hey, J. 4-[(1H-imidazol-4-yl) methyl] benzamidines and benzylamidines: novel antagonists of the histamine H3 receptor. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2263–2268.

(31) Nigg, E. Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle. *Bioessays* **1995**, *17*, 471–480.

(32) Wadler, S. Perspectives for cancer therapies with cdk2 inhibitors. *Drug Resist. Update* **2001**, *4*, 347–367.

(33) Gray, N.; Wodicka, L.; Thunnissen, A.; Norman, T.; Kwon, S.; Espinoza, F.; Morgan, D.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S.; Lockhart, D.; Schultz, P. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533–538.

(34) Davies, T.; Tunnah, P.; Meijer, L.; Marko, D.; Eisenbrand, G.; Endicott, J.; Noble, M. Inhibitor binding to active and inactive CDK2: the crystal structure of CDK2-cyclin A/indirubin-5-sulphonate. *Structure* **2001**, *9*, 389–397.

(35) *ROCS*, version 3.1.0; OpenEye Scientific Software: Santa Fe, NM, 2010.

(36) *MOE*, version 2010.10; Chemical Computing Group: Quebec, Canada, 2010.

(37) Swain, C.; Teran, A.; Maroto, M.; Cabello, A. Identification and optimization of 5-amino- 7-aryldihydro-1,4-diazepines as 5-HT2A ligands. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 6058–6062.

(38) Smith, A.; Stevenson, G.; Lewis, S.; Patel, S.; Castro, J. Solid-phase synthesis of 2,3- disubstituted indoles: discovery of a novel, high-affinity, selective h5-HT2A antagonist. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2693–2696.

(39) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502–4520.

(40) *FlexS*, version 2.0.0; BioSolveIT: St. Augustin, Germany, 2010.