Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation

James Law,[†] Zsolt Zsoldos,[†] Aniko Simon,*[,†] Darryl Reid,[†] Yang Liu,[†] Sing Yoong Khew,[†] A. Peter Johnson,[‡] Sarah Major,[‡] Robert A. Wade,[§] and Howard Y. Ando[§]

SimBioSys, 135 Queen's Plate Drive Unit 520, Toronto, ON, M9W 6V1, Canada, School of Chemistry, University of Leeds, Leeds LS2 9JT, United Kingdom, and Pfizer Global R&D, Groton, CT

Received July 8, 2008

Route Designer, version 1.0, is a new retrosynthetic analysis package that generates complete synthetic routes for target molecules starting from readily available starting materials. Rules describing retrosynthetic transformations are automatically generated from reaction databases, which ensure that the rules can be easily updated to reflect the latest reaction literature. These rules are used to carry out an exhaustive retrosynthetic analysis of the target molecule, in which heuristics are used to mitigate the combinatorial explosion. Proposed routes are prioritized by an empirical rating algorithm to present a diverse profile of the most promising solutions. The program runs on a server with a web-based user interface. An overview of the system is presented together with examples that illustrate Route Designer's utility.

1. INTRODUCTION

The field of computer-assisted synthesis design (CASD) has experienced both successes and failures. Early pioneering work provided ground-breaking formalization of the strategies and tactics required to solve problems of organic synthesis.[1] Despite, by modern standards, an entirely resource-starved computer infrastructure, these early attempts met with surprising success and laid down certain fundamental principles, which have provided the basis for subsequent work.[2,3] Synthetic planning software packages were developed by a number of computational chemistry groups.[4−8] The basic goal of these programs was to accept an input target molecule and produce a selection of synthetic paths optimized against certain criteria.

However, despite the considerable initial successes in the development of the methodology, these systems did not enjoy widespread acceptance by synthetic chemists. A possible reason was that the knowledge bases on which these systems depended contained the descriptions of chemical reactions that were created manually and, because of the considerable expertise and effort involved, never covered more than a small fraction of the expert synthetic chemist's knowledge base. In contrast, reaction databases and the software tools to search them gained ready acceptance by synthetic chemists. While the rules of chemistry might be difficult to capture, the electronic transcription of specific instances of known chemical reactions was relatively straightforward. The task, while tedious, amounted to careful translation (and ultimately computer-assisted transcription) of reactions from chemical literature.[9] This task was undertaken by several entrepreneurial companies and was so successful that electronic databases have almost universally replaced the traditional literature searches of previous generations.[10,11] Modern reaction databases provide facilities for sophisticated substructure matching, so that the user can easily determine the historical precedents of any *single* step in a proposed synthetic path.[12] However, the universal acceptance of reaction databases and the software to search them has been accompanied by a lull in the development of systems that tackle the much more difficult problem of proposing complete routes to synthetic targets (CASD systems).

In the past there have been many software tools designed to assist with route generation.[13−15] In Figure 1, various strategies for organic synthesis design are shown with representative software packages.

Some tools provide extensive reaction assessment calculations,[16−18] while others have the objective of suggesting good starting materials based on detailed perception of the target structure.[19,20] Conceptually, these are regarded as synthetic approaches.

Alternatively, a retrosynthetic approach begins from the target molecule and progresses toward simpler component reactants by applying retrosynthetic structural transformations. Some retrosynthetic software tools base their decisions on encoded generalized reaction rules,[4,5,8,21] while others work with formalized reaction constraints.[15,16,22]

While some of these systems do provide the capability of noninteractive execution, the combinatorial complexity of the synthesis problem has often restricted their application to relatively straightforward cases. Where systems have been designed specifically to function without user intervention,[7,23−28] heuristics are employed to direct the retrosynthetic analysis (i.e., limit the combinatorial nature of the problem) to "optimal" paths.

For rule-based systems, retrosynthetic analysis also depends on the available chemical transforms. Unlike transformations found in a reaction database, these transforms are keyed by substructures (RETRONs)[29] that encode the minimum substructure, which is invariably present in the

* To whom correspondence should be addressed. E-mail: aniko@simbiosys.ca.
[†] SimBioSys.
[‡] University of Leeds.
[§] Pfizer Global R&D.

10.1021/ci800228y CCC: $40.75
Published on Web 02/06/2009

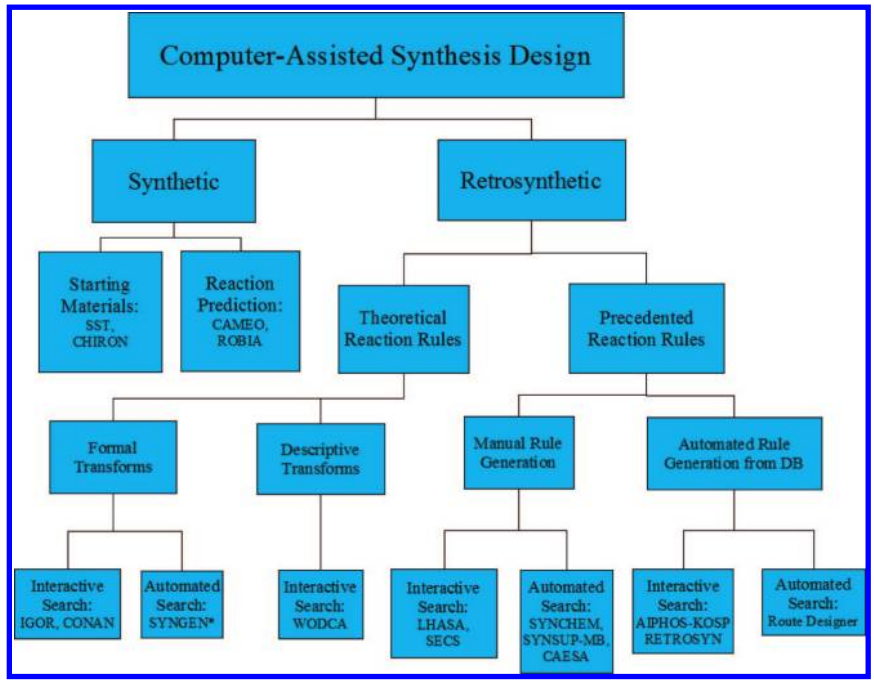


Figure 1. Classification of various CASD route generation tools.

product of a particular reaction. Generally, these transforms are manually created by expert chemists and provide the necessary and sufficient conditions for each transform. These transform rule sets can be of high quality, but the effort needed to create them means that there is often sparse coverage of many important areas of synthetic organic chemistry. To develop better coverage of the entire reaction space there have been several attempts to automate this rule generation process.[30−32] The resulting transform quality has not been anywhere near the level of the manually generated sets because of the difficulty of replicating in a computer program an expert chemist's knowledge of the detailed chemical requirements and limitations of a given reaction.

The Route Designer program was designed to address many of these issues by using transform rules that are automatically derived from modern reaction databases and heuristic controls to handle the combinatorial problem arising from large reaction rule sets. The Route Designer algorithm can create rules from reaction database training sets of varying sizes and has been tested successfully with the Beilstein Crossfire database containing over 4 million example reactions. The program contains a rule extraction mechanism that automatically extracts the essence of each reaction in the database and then creates groups of similar reactions. Each group is generalized to form a representative reaction rule. To address the lack of published failed reactions, the extent of generalizations has been intentionally conservative. Thus, the rules that describe the constraints at any given core atom have been chosen to specifically reflect at least one published example. In this manner, the system is able to capture all of the chemistry contained in a reaction database in an extensive transform rule set usable for retrosynthetic analysis. The system then uses this large rule set in a retrosynthetic search designed to heuristically control the combinatorial complexity of the resulting synthetic trees. The presentation of the results is designed to highlight diverse pathways, while still allowing the chemist to efficiently explore the entire solution space of possible synthetic pathways.

The relationship of Route Designer to various CASD route generation tools is also displayed in Figure 1 above. Note that SYNGEN, while using rules that are generated independently of any reaction precedent, provides a mechanism to search external databases for precedents to any given transformation. In addition, note that CAESA performs a heuristically driven retrosynthetic analysis against a precomputed virtual synthetic library of starting materials (generated by synthetic transformations of real available starting materials) during synthetic accessibility calculations.

2. RULE EXTRACTION

A key component of any computer-aided retrosynthetic analysis program is the set of rules that represent the chemistry used during the search. In the past, many CASD tools, using manually created knowledge bases of chemical reactions, were capable of generating excellent routes within limited knowledge domains. However, these systems required a more systematic rule generation technique if they were to handle a wider variety of synthetic targets.

Systems using machine generated chemistry rules started appearing in the early 1990s.[32] One such example is SYNCHEM,[30] which was adapted to machine learning to increase the program's knowledge base. The KOSP[31] program (knowledge base-oriented system for synthesis planning) attempts to extract rules from reaction databases by clustering reactions based on characteristics of atoms within three bonds of a disconnection site. The focus of KOSP is only disconnective transformations, and the program uses an interactive search process to generate one transform at a time. Similarly, RETROSYN also provided an interactive search based on finding single disconnections by similarity with precedent reactions.[32] From a substructure containing the chemical environment as described by Wilcox and

Levinson,[33] the degree of similarity was assessed by counting overlaps of every subgraph with example reactions from the relatively tiny ORGSYN database. The user was expected to manually determine compatibility and apply any prerequisite functional group modifications.

In the Route Designer rule extraction method, the focus is on extending the reaction cores to capture all of the necessary chemical environments, regardless of distance from the reaction or disconnection site. These entire extended reaction cores are grouped by similarity, then generalized to create a usable set of reaction rules. The Route Designer algorithm is scaleable and has been applied to very large reaction databases such as the Beilstein Crossfire reaction database.

Five steps are used in the Route Designer rule extraction process:

1. Reaction cores are identified. A reaction core is the set of atoms where connections or bonds have changed by going from reactant to product.
2. Cores are extended to encapsulate the *relevant* neighboring atoms, that is, functional groups that influence the reaction.
3. Extended reaction cores are clustered into common groups.
4. Each cluster group produces a generalized rule template that can represent all the specific examples from the given group.
5. Generalized rule templates are converted into completed reaction rules, where each generalized element has been refined to represent either the full range or the statistical mode (for leaving groups) within the associated examples from the cluster group.

2.1. Identifying Reaction Cores. A reaction core is identified using the atom-to-atom mapping for reactants to products included in the reaction databases. If mapped atoms do not have the same attributes in the reactants and products, they are included as part of the core. Such attributes include the number of neighbors, bond types, neighboring atom types, charge (if any) and presence of radical designation. Some atoms found in reaction databases are unmapped. These unmapped atoms are found either only in the reactants or only in the products. When they are found in the reactants, they are considered part of the leaving group and assumed to end up as part of the reaction environment rather than as a product of interest. When they are found only in the products, they are typically small oxygen/nitrogen-based structures originating from reagents rather than reactants. The extracted reaction core alone does not contain all the necessary chemical information to accurately represent the reaction and therefore must be extended to include any neighboring atoms and groups that influence the reaction.

The reaction shown in Figure 2 can be used to illustrate the rule extraction process in Route Designer. Here a Michael Reaction is shown where a "β-ketoester" moiety reacts with an "α,β-unsaturated ketone" to form a new carbon−carbon bond. The extracted reaction core shown in Figure 2b contains only the atoms that have changed properties, that is, carbons 5, 12, and 13, where the number of attachments or bond order have changed. However, it is clear that the extracted core lacks information about activating groups, which, although unchanged, are essential for the reaction to occur and therefore need to be included in any retrosynthetic rule for the reaction.

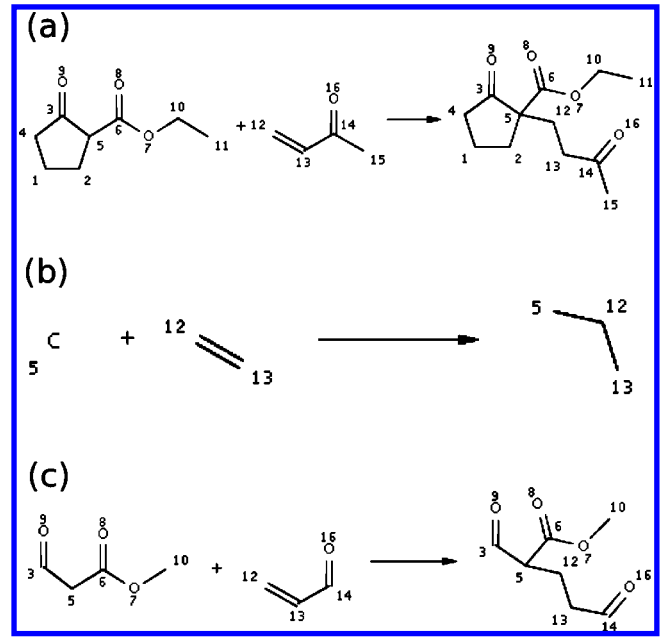


Figure 2. (a) Sample reaction to illustrate core extraction and extension. (b) Extracted core. (c) Extended reaction core includes the chemically relevant environment.

2.2. Reaction Core Extension. The identified reaction cores tend to be underspecified with respect to the relevant chemical environment and therefore must be extended. The guiding principle behind extending the core is to select enough of the neighboring functionality in both reactant and product to specify the minimum substructures on both sides that encompass the essence of the reaction. This may include leaving groups that appear only on the reactant side of the reaction and atoms in the product that are not present in the reactants, but rather come from reagents. In earlier seminal work on retrosynthetic analysis, this minimum substructure on the product side was defined as the RETRON for the reaction.[29]

The core extension algorithm in Route Designer examines all of the border bonds attached to the set of extracted core atoms. The handling of each bond depends on the nature of the noncore atom. If this atom is mapped in the product then it is considered to be part of the nonreacting (but potentially still necessary) neighborhood; otherwise it is considered to be part of a leaving group.

The goal of reaction core extension is depicted in Figure 3. Consider a hypothetical reaction where a bond is being formed between core atom A and core atom X. The bond between core atom A and noncore atom B will be referred to as a primary bond. The bond between noncore atom B and noncore atom C will be referred to as a secondary bond.

In the first instance, it is desirable to extend the core to include all the atoms of any influencing functionality. Functional group completion involves extending the core until a carbon atom is reached which bears an external nonaromatic carbon−carbon single bond. This completion is performed in addition to the following extension rules based on specific structural features.

For cases where there is no mesomeric withdrawing group across the primary bond, the core is also extended across any secondary double or triple bonds, otherwise, these

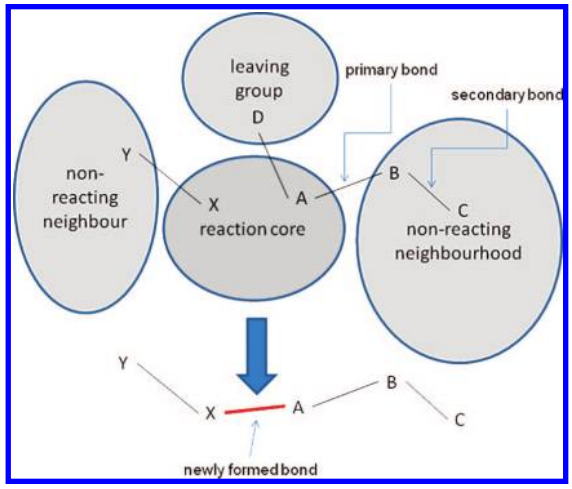


Figure 3. Schematic of reaction core and extended core identification.

carbon−carbon bonds are ignored. For aromatic primary bonds, the core is extended to include the entire aromatic system, while aromatic secondary bonds are only extended whenever atom B is heteroaromatic.

For leaving groups (i.e., unmapped groups which are present in the reactant but not the product), the extension includes the entire set of unmapped atoms and bonds. The nature of the leaving group is algorithmically determined to be nucleofuge, electrofuge, or unspecified (usually carbon based). The determination is based on primary electronegativity difference except that halogens are always nucleofuges and sulfones are electrofuges.

Returning to the example in Figure 2, this process creates the extended core for the reaction (Figure 2c). This extended core contains the two ketones and the ester functional groups that influence the reaction. The extracted core is still part of the extended core, but now sufficient chemical environment has been included to accurately define the requirements for the reaction to occur. It is important to note that carbon 2 is not included in the extended reaction core. This carbon is not essential to the reaction and is correctly excluded. However, it would be (incorrectly) included if core extension were to be performed based only on inclusion of shells of all atoms a particular bond distance from the core. This core extension to *relevant* neighboring functionality is the key feature that distinguishes the Route Designer method from the core extension found in CLASSIFY[34] and KOSP, where extension is based entirely on bond distance from the reaction core. There are some specific exceptions (such as nucleophilic substitution at a carbonyl group, where any directly attached aromatic rings have little effect) where neighboring functionality is not included.

2.3. Clustering Reaction Cores. For each reaction, a canonical name (64-bit Morgan-type number) is calculated for the substructures describing the extended reaction core. These canonical reaction numbers are generated by concatenation of the sum of 32-bit Morgan-type numbers for the reactants and the products. This reaction number, along with the similar number for the initial extracted core reaction, will determine the initial groupings of the reactions. These groups are internally clustered by graph matching on all identified properties (basic atom/bond properties such as hybridization, ring inclusion, and aromaticity).

Clustering of the reaction cores includes translating the mapping between the example reaction and the reaction core, so that the reaction cores of two or more examples can be compared. These clusters represent a partitioning of the example database into similar reactions. The size of each cluster provides a crude relative measure of the reliability of the associated reaction rule. Clusters with many examples have been successful in many situations, while clusters with fewer examples may represent reactions of more limited utility. Finding a cluster containing only a single example is quite common. These singleton clusters are generated, but the associated rules are not currently used by the Route Designer search engine.

Figure 4 shows three esterification reactions with their reaction cores identified. These three example reactions are clustered into the same group by Route Designer. It can be seen that all the reactions are esterifications with the same core and similar extended cores which differ only in the given leaving group. In such cases, when the retrosynthetic analysis is carried out, a representative leaving group is chosen and displayed to the user (Figure 4f).

2.4. Generalizing Reaction Cores. After the reaction cores are clustered, the range of permitted values of certain properties of atoms and bonds in the extended core is determined by examining the original example reactions.

Atom hybridization is defined (e.g., sp^3, sp^2, sp,...) as a single value if all examples agree; but, as with other properties, alternative values are also allowed. Halogen types are grouped (e.g., HL = F, Cl, Br, I) to reflect similar properties. Neighbor counts (carbons, hydrogens, halogens, heteroatoms, and the count for the number of lone pairs) account for the variation across all example reactions. Ring sizes are restricted to the precise sizes found among examples. Peripheral carbons not found as part of the extracted core have relaxed property constraints, so neighbor counts and ring sizes are ignored for carbon atoms which are strictly part of the extended core.

2.5. Complete Reaction Rules. The range for each generalized property across the original example reactions is represented in the completed reaction rule. Where there are alternative leaving groups, the final step in creating a usable reaction rule is the selection of the leaving group to be shown in the reaction display. This is found by examining the leaving group variability found in the rule cluster and selecting the most commonly occurring leaving group as the representative. Halogen types are now refined (HL = F, Cl, Br, I), (CX = Cl, Br, I), and (BX = Br, I) to reflect differences in capability when acting as leaving groups.

2.6. Results of Ruleset Generation. The Route Designer rule generation algorithm has been tested with the MOS reaction database from Accelrys and the Beilstein Crossfire reaction database from Elsevier. The MOS database was processed in 2 h (on a single Pentium 4 class CPU) and generated 4028 unique rules from 42 333 example reactions. A sample of the Beilstein database containing 61 3074 example reactions generated 50 702 unique rules after 18 h of processing. In both cases, each unique rule was associated with a clustering bucket containing at least two example reactions.

The full Beilstein reaction database has also been processed through the Route Designer algorithm. After 200 h of processing, the source database of 4.5 million example

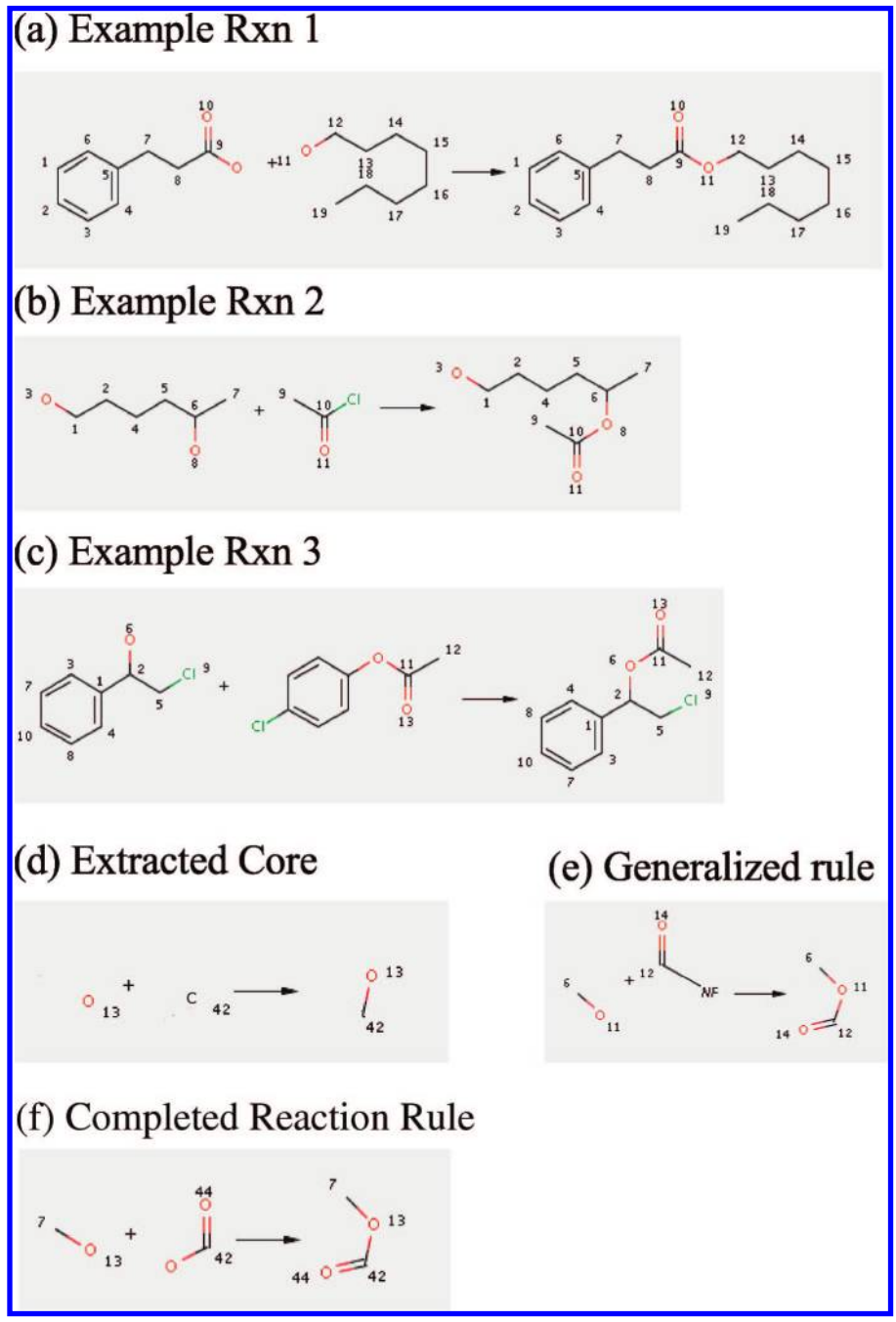


Figure 4. (a, b, c) Three examples of esterification reactions with a common extracted core. (d) The common extracted core for these reactions. (e) The generalized reaction rule for this reaction cluster, note NF represents a nucleofuge. (f) The completed reaction rule, where the generic leaving group is replaced with specific atom types.

reactions generated 285k cluster groups. Technically, each cluster group could be used to generate a unique rule, but rules generated from more examples invariably provide fuller descriptions of the acceptable variability of the reaction neighborhood. So, the ruleset generated from the full Beilstein was chosen to require at least five examples per rule. This produced a set of 92 781 unique rules.

2.7. Fundamental Transforms. In addition to the automatically extracted transforms, Route Designer uses a small group of fundamental transforms (FT) in the initial stages of the retrosynthetic search to improve both the speed of the search and the quality of the results. These fundamental transforms are mainly disconnective in nature, and they allow Route Designer to "break-up" the target molecule into smaller fragments, which are then subjected to the full retrosynthetic analysis. They are intended to reflect the routine disconnections that a synthetic chemist would automatically apply on inspection of a synthesis target.

These FTs currently include rules for approximately 30 common organic reactions (e.g., esterification or amide formation) that have been manually created and added to the reaction database. FTs are applied at the start of each search.

3. SEARCH

The search phase of Route Designer uses the extracted rules from the rule extraction phase to retrosynthetically disconnect a target molecule. When using very large reaction rule databases, mitigating the combinatorial nature of ret-

Table 1. Reaction Classifications Used in Route Designer along with the Associated Transform Type and Frequency of Occurrence in the Rule Set Derived from MOS

type	name	category	no. of MOS rules
1	regular disconnective	RD	2860
2	functional group removal	FGR	92
3	functional group interconversion	FGI	492
3	bond-oriented FGI	BFGI	236
3	nondisconnective rearrangement	NR	3
4	functional group addition	FGA	69
0	deprotection reaction	DP	180
0	bare carbonyl reaction	BCR	12
0	simple alkane product	SAP	16
0	unactivated educt reaction	UER	7

rosynthesis is a real challenge for the Route Designer program. A number of heuristics and user definable limits are used to achieve this goal. These include (a) maximum search depth, which provides one termination condition for the search, (b) choice of rule sets, including the minimum number of examples required for a rule to be applied in the retrosynthetic analysis, (c) choice of available starting materials databases (matching an available starting material provides another termination condition for the search), (d) ability to specify one or more required starting materials, and (e) ability to specify unbreakable bonds or required disconnections, which must appear in every solution.

Given a target molecule the Route Designer search proceeds as follows:

1. Perceive certain features (aromaticity etc) of the target molecule.
2. Add the perceived molecule to the search queue and mark it as unsearched.
3. Select the first unsearched molecule from the search queue which needs to be explored.
4. Check the reaction database for possible/valid retrosynthetic transforms.
5. Apply all transformations, checking for valid pattern match, and identify resulting reactants.
6. Add to the queue any reactant molecules which do not match any molecules that are currently in the queue.
7. Repeat steps 3−7 until all molecules in the search queue are marked as searched or found as starting materials, therefore not requiring further searching.

3.1. Transforms. Transforms in Route Designer are retrosynthetic rules that are applied to a target molecule and are classified by Route Designer into various categories during rule generation. Within each category, the rules are applied in the same manner during the search. The reaction categories are listed in Table 1, along with the number of rules found from processing the 42k examples in the MOS database. Some categories carry the designations used in LHASA,[35] while others were created to address the specific needs of the Route Designer search algorithm. Each category is assigned a type, which determines the default behavior of rules from that category. In Route Designer, the chemist can manually modify the classification and behavior for any specific reaction rule.

Type 1 transformations are normal transformations, which disconnect the carbon skeleton of the target. Type 1 reactions are permitted to occur (i.e., to be applied to the target molecule during search) at all times because they simplify the overall carbon framework of the target. All fundamental transforms are type 1.

Type 2 transformations are known as restricted usage transforms. These transforms are simplifying but are never applied immediately following opportunistic transforms (type 3 or 4). Functional group removal (FGR[35]), where there is a retrosynthetic removal of functional atoms without any modification to the carbon skeleton, falls into this category.

Type 3 transformations are opportunistic transforms. These transforms are used to modify functionality to permit matching to a normal transform or an available starting material. After the application of a type 3 transform, only those type 1 transforms that could not be matched in the parent node are allowed to be applied in the next step. Type 3 transforms include functional group interconversion (FGI[35]) reactions, where there is a retrosynthetic modification of functional atoms without any modification to the carbon structure.

Type 4 transformations are also nonsimplifying and, like type 3, are applied to assist the matching to normal transforms. Unlike type 3, these transforms are only applied if no other transforms were found to match the structure under consideration. Functional group additions (FGA[35]) are type 4 transforms. FGAs are transforms where there is a retrosynthetic addition of atoms but no bonds between heavy atoms are broken. In the synthetic direction this corresponds to removal of a functional group, for example conversion of a bromide to a saturated alkane via hydrogenation. Type 4 transformations actually increase complexity and as such are applied only when no simplifying transformations are applicable.

In addition to these categories, Route Designer has defined a set of transforms that are currently not applied in search but may have a role in the future. These type 0 transforms, unused transforms, include deprotection reactions (DP), simple alkane product (SAP) reactions (the product pattern is a small, unfunctionalized alkane and the retron is likely to match too many targets), and unactivated educt reaction (UER) (educt/reactant patterns are all small, unfunctionalized alkanes and the reaction is likely to require extreme reaction conditions and is not likely to be useful in laboratory synthesis).

3.2. Prioritizing Transforms. The search queue in Route Designer stores all the molecules found during the retrosynthetic analysis of the target molecule. Molecules in the queue are in one of three states: saved, queued for search, or searched. Saved molecules are those found during search but for which no additional analysis is to be performed (i.e., either starting materials or reactants from transformations with low priority, these molecules are termination nodes of their branch) and need be saved to the database. Queued molecules are those that were found during the search, but require further analysis (i.e., intermediates of promising branches/routes). Nodes marked as searched are molecules that have undergone retrosynthetic analysis (i.e., intermediates with computed reactants).

In addition to the state of each molecule, the search depth and a list of all retrosynthetic transformations that were found to match the molecule is also stored. Each retrosynthetic solution is a list of search nodes beginning with the target molecule node and proceeding down the tree until a terminal node is reached. Ideally, this terminal node should represent an available starting material.

The choice of whether a saved molecule is scheduled to undergo further retrosynthetic analysis is based on several criteria:

1. If a node represents a molecule that exactly matches a starting material it is saved, but never searched, since the goal of retrosynthetic analysis is to break down the target into starting materials.
2. Child nodes of fundamental transforms are always queued for search.
3. For all other transforms the decision to queue a molecule for search is based on the relative importance of the transform and the specific transformation when compared with other transformations found for the same parent node. The comparison depends on the following attributes of the transformations:

(a) Prefer maximal starting material coverage. If a transformation produces reactants that are all available starting materials then it is ranked higher than transformations thst do not produce non-starting-material precursors.

(b) Prefer disconnective transformations. If a transformation is disconnective and another is not, then the disconnective transform is ranked higher. The extension of this is that the disconnective transform with the greater number of reactants is ranked higher.

(c) Prefer less total wastage in reactions, that is, prefer efficient reactions. Wastage calculates a weighted sum of unmapped heavy atoms in the reactants that are not present in the product.

(d) Prefer more balanced splits. Given identical wastage, preference is given to the transformation with the larger "second largest mapped reactant".

(e) Prefer thoroughly explored chemistry. Select the transform with the most examples from the reaction database.

(f) Prefer the smaller primary reactant. Given two transformations, comparing the largest reactant of each transformation, select the transformation where this largest reactant is smaller.

After these transformations are ordered, the number of transformations selected for further expansion is calculated (the number of reactions whose reactants will be marked for further processing). In addition, the reactants from FGIs are always explored because the role of the FGI is to create the functionality for simplifying transforms.

3.3. Selecting Molecules from the Queue to be Searched. The molecules that have been marked to be searched are checked for search depth against the maximum depth of the current search. The entire queue is processed until no unsearched molecules are found that are both marked for search and below the requested search depth. Once the queue has been traversed without finding a new molecule to search, the search is considered completed.

3.4. Checking a Transformation for Validity against the Current Molecule. When applying transforms to the selected molecule, a validity check occurs to ensure the applicability of that transform to the specified molecule. This validity check occurs in two stages. The first stage is a quick bit-wise check to ensure that various global property counts are sufficient to match the requirements of the transform. This includes properties such as the number of aromatic atoms, the number of charged atoms, the number of halogens, the number of fused bonds, etc.

After passing these checks, the rule pattern is matched against the target molecule by subgraph (connectivity) matching. This procedure considers atomic number, atomic charge, hybridization, aromaticity, carbon neighbors, bond type, ring size, etc. The atomic number and bond type are always matched while the others are matched as available.

Once the transform has been validated it is applied to the target molecule. Each generated reactant then undergoes a three-step postprocessing check to ensure the validity of the reaction. First a check for unstable structures is performed. Highly reactive structures are allowed to survive only when the immediate subsequent reaction removes the unstable moiety. Subsequent reactions that do not are disallowed. Unstable structures include organometallics, acyl halides, acyl azides, acyl cyanides, isocyanate/isothiocyanates, diazo/diazonium compounds, sulfonyl halides, and alkoxy/aminomethylhalides.

Second, a check is made to ensure that only plausible chemical structures are produced. Here primarily valence/charge balance is checked along with the inexplicable creation or loss of aromaticity or the creation of a triple bond within a small (<8-membered) ring.

Finally, the user-specified settings regarding required disconnection and unbreakable bond settings are checked. If the user requires specific bonds to be broken then each applied transform must break at least one of those bonds until each specified bond has been broken. If a transform attempts to break a bond marked unbreakable then that transform is rejected.

Once all reactants have been generated and validated, each reactant is compared against all previous molecules in the search queue. If a match is found then the transformation is included as an alternate route leading to the same intermediate. Any reactants which do not match already found molecules are added to the search queue. New molecules are added with the save attribute and are then prioritized as described above.

3.5. Starting Materials. The starting material databases provide the successful termination points of the search. These vendor-provided databases are preloaded into the system and analyzed for similarity (duplicates are stored as multiple records referencing the same molecular structure). The starting material database is fully extendable by providing molecules stored in SDF format.

3.6. Assembling Search Results. After all molecules in the search queue have been searched, match starting materials, or the user-specified search depth has been exceeded, Route Designer proceeds to construct synthesis trees for the target molecule. The goal is to provide the chemist with a small number of varied synthetic paths consisting of strong candidate solutions. Variety is achieved by selecting solutions from different branches where possible.

Each synthetic path is rated based on an empirical weighting function, which is a cascading weighted sum of reactions based on wastage, example counts, and balanced disconnections. Yield was not selected as a criterion because our analysis of electronic reaction databases found that yield was often omitted or irregularly specified, so it could not be used reliably; however, yield is still presented for every reaction step when available.

The top-rated solutions are stored in the database and the search results are presented. No more than 50 top rated solutions are presented to the chemist, and the typical solution set consists of less than 20 solutions. The theoretical number

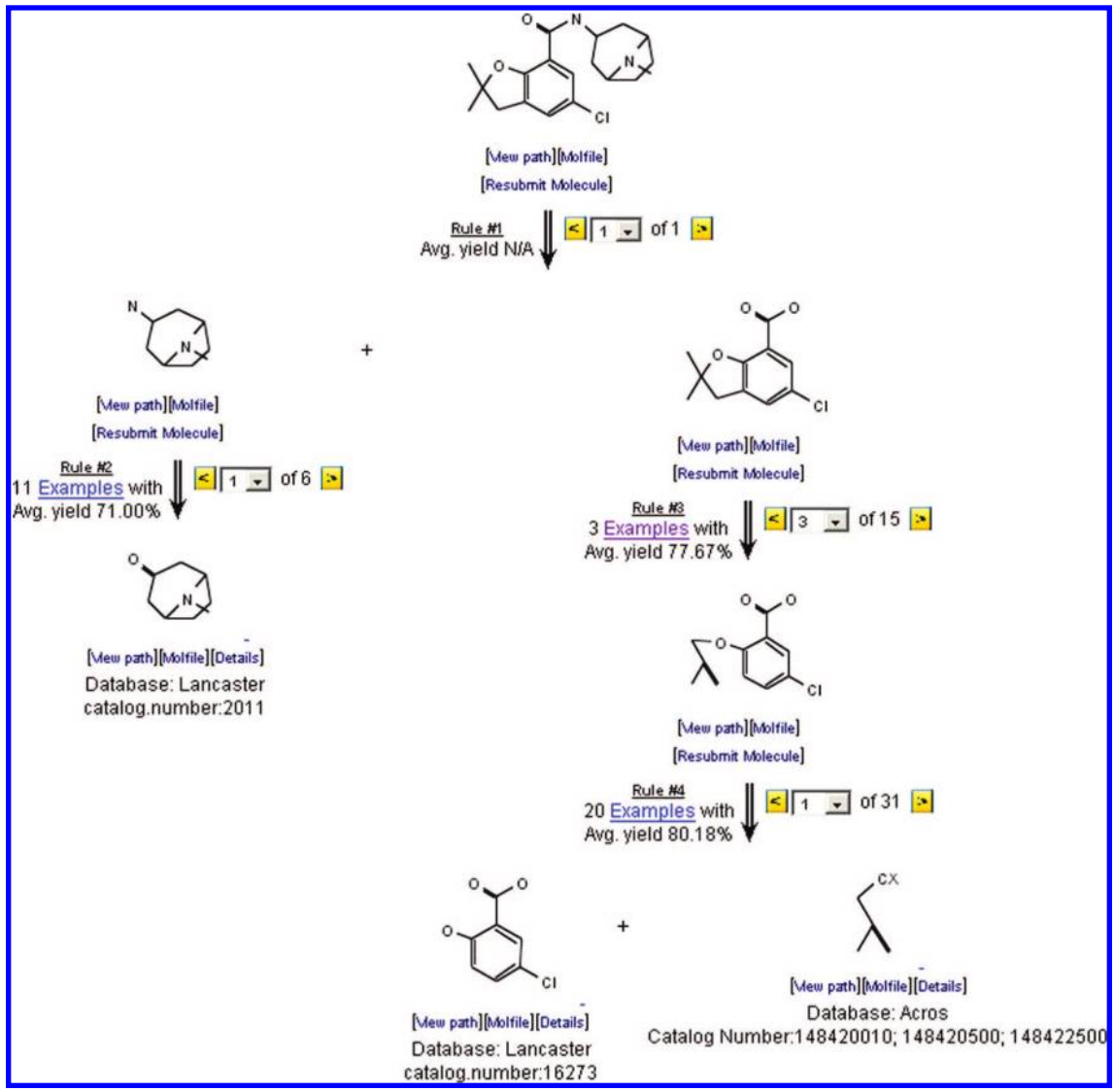


Figure 5. Sample synthetic route found by Route Designer for Zatosetron as the target molecule.

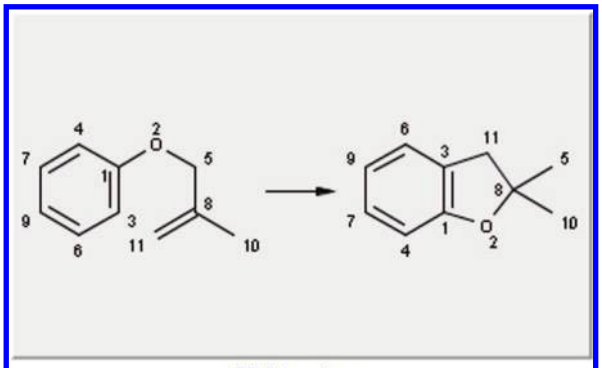
of algebraically different synthetic routes is often in the millions, so this computer-assisted distillation of results is required. The chemist is still able to view all retrosynthetic intermediates generated from a specific node, so the entire solution space is still accessible.

4. EXAMPLE REACTION ROUTES

The following example was found by Route Designer using rules extracted from the MOS database. The MOS-derived rule set was specified to require at least three examples per rule. This produced a total rule set of 1575 rules. The maximum depth for the search was set to 3. The starting material database consisted of 120k compounds from the Lancaster, Acros, and Aldrich catalogues. The complete search result set was generated in less than 5 min. The target molecule was Zatosetron,[36] a potent, selective and long acting 5HT receptor antagonist from Lilly used in the treatment of nausea and emesis associated with oncolytic drugs, including cisplatin. The retrosynthesis shown in Figure 5 was the top-rated solution found through Route Designer.The route shown is virtually identical to the published route for Zatosetron.[36] The first step in the retrosynthesis is an amide formation from a primary amine which is one of our FT transformations. This transformation divided the target into two smaller targets and was essential to directing the retrosynthetic analysis. The aminotropene was produced by reductive amination (rule 2) of tropinone similar to the published route.

Examination of Figure 5, which is a Route Designer screenshot, shows various tools available to the chemist. To the left of each retrosynthetic transformation (downward arrow), a link to literature examples and the average yield of all examples are presented. A selection tool is available on the right to allow the chemist to individually view alternative transformations at every step. The output of the "examples" link for rule 3 presents the rule derivation from several examples of heterocyclization reactions following rearrangement found in the MOS database, as shown in Figure 6.

Below each molecule are tools to display the synthesis path from that molecule and to download the MOL file for use in other applications. In addition, commercially available molecules are displayed with vendor and catalog number along with a tool for displaying details of alternate vendors. Starting material matches for the 5-chloro-2-hydroxybenzoic acid shown in Figure 5 are displayed in Figure 7. Intermedi-



Click to enlarge

<< < 1 of 3 > >>

Reaction details: MOS	
Bibliographic	Bernard A M, Cocco M T, Onnis V, Piras P P, Synthesis, (1), p. 41-43, 1997
Yield	75
Reagent	No data available
Catalyst	No data available
Solvent	No data available
Time	No data available
Temperature	110
Conditions	Mo(CO)6
Conditions	MOS0034762
Conditions	55 h.
Conditions	No data available
Conditions	No data available
Conditions	No data available
Number of Stages	1 Step
External ID	MOS0034762
Reaction Name	C-Alkylation. Catalysis. Claisen rearrangement. Heterocyclisation. Rearrangement. Ring formation
Literature Abstract	Mo(CO)6 catalyzed one-pot conversion of allyl aryl ethers to dihydrobenzofurans.
Comments	The heterocyclic product is favoured over the phenol by 75 to 25%. Fourteen examples given; yields are 15-91%.
Product	No data available
Reactant	No data available

Figure 6. Example details for rule 3 from Figure 5.

ate molecules that are not found to be starting materials have the additional option of resubmission for further retrosynthetic analysis, where the chemist can change any search parameter (e.g., other rule sets, starting materials, search depth).

5. CONCLUSIONS

The described Route Designer application contains a very powerful automatic rule extraction engine, which can capture and generalize chemical knowledge implicitly contained in large reaction databases. The generated rules are used in a search approach that attempts to balance the tradeoffs between exhaustiveness and computational time by prioritizing reaction transformations and setting (user-defined) limits

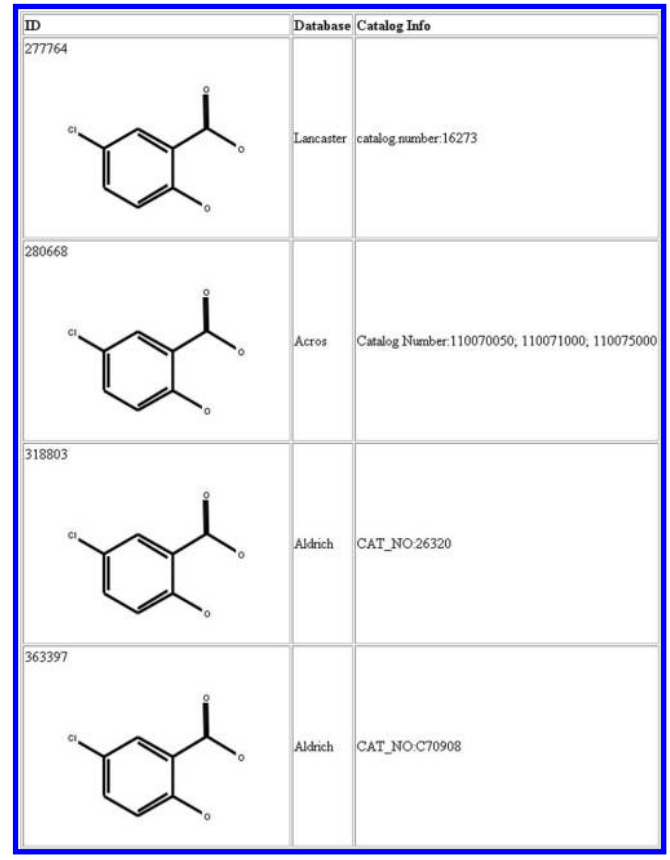


Figure 7. Starting material details for 5-chloro-2-hydroxybenzoic acid from Figure 5.

on search depth. This approach routinely finds interesting synthetic routes to a wide variety of target molecules. A key feature of the system is that it provides access to an exhaustive set of synthetic routes, together with examples of suggested reactions and the associated literature references.

Ongoing improvements to Route Designer include enhancing the rule extraction engine to improve treatment of regioselectivity, stereoselectivity, and interfering functional groups. Incremental improvements in rule extraction and the ordering of the search results to better reflect chemists' preferences will increase the efficiency and usability of the system.

ACKNOWLEDGMENT

The authors are grateful to chemists at Pfizer Inc. for validating the system and providing useful feedback throughout the project. The help and support of Accelrys Inc., in particular of Eric Jamois (now with In-SiliChem, http://www.insilichem.com/ accessed July 2008) and Robert Brown for providing reaction files as input to the system is acknowledged. We would also like to thank Elsevier, in particular Juergen Swienty-Busch, for providing the Beilstein reaction files.

REFERENCES AND NOTES

(1) Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **1967**, *14*, 19–37.
(2) Bersohn, M.; Esack, A. Computers and organic synthesis. *Chem. Rev.* **1976**, *76*, 269–282.
(3) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166*, 178–192.

(4) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General methods of synthetic analysis—Strategic bond disconnections for bridged polycyclic structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116–6124.
(5) Corey, E. J.; Jorgensen, W. L. Computer-assisted synthetic analysis—Synthetic strategies based on appendages and use of reconnective transforms. *J. Am. Chem. Soc.* **1976**, *98*, 189–203.
(6) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-assisted analysis in organic-synthesis. *Science* **1985**, *228*, 408–418.
(7) Gelernter, H. L.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzer, G. A.; Searleman, J. E. Empirical explorations of SYNCHEM. *Science* **1977**, *197*, 1041–1049.
(8) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artif. Intell.* **1978**, *11*, 173–193.
(9) Buntrock, R. E.; Valicenti, A. K. End-users and chemical information. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 203–207.
(10) Methods in Organic Synthesis (MOS). www.accelrys.com/products/chem_databases/databases/methods_organic_synthesis.html (accessed Dec 16, 2008).
(11) Blake, J. E.; Dana, R. C. CASREACT: more than a million reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394–399.
(12) Parkar, F. A.; Parkin, D. Comparison of Beilstein CrossFirePlus Reactions and the Selective Reaction Databases under ISIS. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 281–288.
(13) Ihlenfeldt, W. D.; Gasteiger, J. Computer-assisted planning of organic syntheses: The second generation of programs. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2613–2633.
(14) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.
(15) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Computer-assisted solution of chemical problems—The historical development and the present state-of-the-art of a new discipline of chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201–227.
(16) Gasteiger, J.; Jochum, C.; Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A logic-based program for synthesis design. *Top. Curr. Chem.* **1978**, *74*, 93–126.
(17) Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic-reactions 0.1. Overview. *J. Org. Chem.* **1980**, *45*, 2043–2051.
(18) Socorro, I. M.; Goodman, J. M. The ROBIA program for predicting organic reactivity. *J. Chem. Inf. Model.* **2006**, *46*, 606–14.
(19) Hanessian, S.; Franco, J.; Larouche, B. The psychobiological basis of heuristic synthesis planning man, machine and the chiron approach. *Pure Appl. Chem.* **1990**, *62*, 1887–1910.
(20) Wipke, W. T.; Rogers, D. Artificial-intelligence in organic-synthesis—SST—Starting material selection-strategies—An application of superstructure search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.
(21) Barone, R.; Chanon, M. Search for strategies by computer: The CONAN approach—Application to steroid and taxane framework. *Tetrahedron* **2005**, *61*, 8916–8923.
(22) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A logic-based program for synthesis design. *J. Am. Chem. Soc.* **1985**, *107*, 5228–5238.
(23) Agarwal, K. K.; Larsen, D. L.; Gelernter, H. L. Application of chemical transforms in SYNCHEM2, a computer-program for organic synthesis route discovery. *Comput. Chem.* **1978**, *2*, 75–84.
(24) Baber, J. C. CAESA: Computer-aided estimation of synthetic accessibility—Improved algorithms for the identification of starting materials. PhD thesis, University of Leeds, Leeds, U. K., 1998.
(25) Funatsu, K.; Sasaki, S. I. Computer-assisted organic synthesis design and reaction prediction system Aiphos. *Tetrahedron Comput. Method* **1988**, *1*, 27.
(26) Gillet, V.; Myatt, G.; Zsoldos, Z.; Johnson, A. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34–50.
(27) Myatt, G. J. Computer aided estimation of synthetic accessibility. PhD thesis, University of Leeds, Leeds, U.K., 1994.
(28) Takahashi, M.; Dogane, I.; Yoshida, M.; Yamachika, H.; Takabatake, T.; Bersohn, M. The performance of a noninteractive synthesis program. *J. Chem. Inf. Model.* **1990**, *30*, 436–441.
(29) Corey, E. J. *The Logic of Chemical Synthesis*; John Wiley & Sons Inc.: New York, 1995.
(30) Gelernter, H.; Rose, J. R.; Chen, C. H. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492–504.
(31) Satoh, K.; Funatsu, K. A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 316–325.
(32) Blurock, E. S. Computer-aided synthesis design at RISC-Linz: Automatic extraction and use of reaction classes. *J. Chem. Inf. Model.* **1990**, *30*, 505–510.
(33) Wilcox, C. S.; Levinson, R. A. A self-organized knowledge base for recall, design, and discovery in organic chemistry. *ACS Symp. Ser.* **1986**, *306*, 209–230.
(34) CLASSIFY: The Infochem Reaction Classification Program, version 2.9. http://infochem.de/content/downloads/classify.pdf (accessed Dec 16, 2008).
(35) Corey, E. J. Computer-assisted Analysis of Complex Synthetic Problems. *Q. Rev. Chem. Soc* **1971**, *25*, 455–482.
(36) Robertson, D. W.; Lacefield, W. B.; Bloomquist, W.; Pfeifer, W.; Simon, R. L.; Cohen, M. L. Zatosetron, a potent, selective, and long-acting $5HT_3$ receptor antagonist: Synthesis and structure—activity relationships. *J. Med. Chem.* **1992**, *35*, 310–319.

CI800228Y