JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

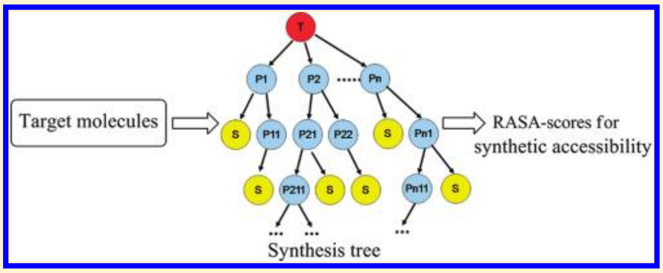
pubs.acs.org/jcim

RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules

Qi Huang,[†] Lin-Li Li,[†] and Sheng-Yong Yang[†]

[†]State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Sichuan 610041, China

Supporting Information



ABSTRACT: In this account, a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules, called RASA (**R**etrosynthesis-based **A**ssessment of **S**ynthetic **A**ccessibility) is devised. RASA first constructs a synthesis tree for the target molecule based on retrosynthetic analysis; in this process a series of strategies are suggested for limiting combinatorial explosion of the synthesis tree. A scoring function (RASA-score) for the assessment of synthetic accessibility is then proposed based on the optional effective synthetic routes, the complexity of reaction, and the difficulty of separation/purification associated with the most favorable synthetic route. The contributions of individual components are calibrated by linear regression analysis based on the synthetic accessibility estimates of a training set (100 compounds) given by a group of medicinal chemists (G1). Two external test sets (TS1 and TS2), whose synthetic accessibility estimates were given by the group G1 medicinal chemists and another group (G2) of medicinal chemists (from literature), respectively, were adopted for the evaluation of RASA. The correlation coefficient between the calculated RASA-score values and the estimated scores by medicinal chemists for TS1 is 0.807 and that for TS2 is 0.792, which demonstrate the validity and reliability of RASA. The validity and reliability as well as the high speed of RASA and its capability of suggesting synthetic routes enable it a useful tool in drug discovery.

INTRODUCTION

Discovery of novel lead compounds has always been the biggest bottleneck in the drug R&D process. Recent development of computer aided drug discovery (CADD) techniques likely brings some hopes for getting rid of this bottleneck. The most widely used CADD technique is the so-called virtual high throughput screening (vHTS), which involves a rapid *in silico* assessment of large libraries of chemical structures in order to identify those structures most likely to bind to a drug target.[1] A big success has been achieved by vHTS, and lots of successful stories of vHTS application have been reported in recent years.[2−5] However the disadvantages of using this method are becoming increasingly obvious. The most pronounced shortcoming is that the compounds obtained through vHTS are pre-existing, which implies a limited IP (intellectual property) position for these compounds. The second most obvious shortcoming could be that retrieving a new lead compound from the existing chemical library is becoming more and more difficult due to the current very limited chemical structural space, particularly that with biological relevance. An alternative technique is the *de novo* design.[6] The *de novo* design approach can produce novel molecular structures with desired pharmacological properties; these compounds have a preferential IP position. In addition, it makes more chemicals available in the whole chemical space that is estimated to be in the order of 10^{60} to 10^{100}.[7] However, due to that the *de novo* designed molecules usually do not exist, whether the chemicals can be easily synthesized, i.e. their synthetic accessibility, becomes the biggest challenging problem.[8] Actually, not only in the *de novo* design process, synthetic accessibility is also a critical problem in vHTS since not all of the compounds retrieved via vHTS can be purchased from off-the-shelf catalogues.[9] Those compounds not available from the market are still needed for synthesis although their synthetic methods may exist somewhere and may not be difficult. From another point of view, the ability to accurately predict bioactivity of a chemical compound is still very limited.[8] Therefore, in order to increase the chance of success, a considerable number of compounds are generally chosen to undergo experimental *in vitro* testing from the compound list suggested by either vHTS or *de novo* design. The synthetic accessibility is a key factor that should be considered in determining which compounds should be selected in order to reduce the late risk associated with their synthetic accessibility.

Although manually predicting synthetic accessibility of a small set of compounds by medicinal chemists is not a difficult task, it is not practical in the case of a large set of structures, which is a general situation when using *de novo* design method or vHTS. Recently some automated tools have been suggested to deal with this type of problem.[10−13] These tools can be roughly classified into three categories: complexity-based, starting material-based, and retrosynthesis-based classes.

Received: May 31, 2010
Published: September 21, 2011

Complexity-based analysis is probably the most widely used technique in the estimation of synthetic accessibility due to that it is easy to actualize and computationally inexpensive.[14−16] In principle, the complexity-based methods are based on graph and information theories. They try to locate and count all the features that are difficult to synthesize, such as spiro-rings, nonstandard ring fusions, stereocenters et al., in a target structure, which are then associated with the synthetic accessibility. Although a much higher speed of calculation, it does not incorporate any information of starting materials, which might result in an inaccurate or completely wrong estimation of the synthetic accessibility.

Starting material-based methods carry out the estimation of synthetic accessibility through assessing how much of a target compound is covered by available starting materials. Two approaches are generally used for the identification of possible starting materials: methods based on the exact matching of substructures and those based on similarity − either of substructures or the target structure as a whole.[11,12] Very recently, various transform-based similarity search methods have also been introduced, in which similarity criterions are based on some generalized reactions or common structural features.[17] The starting material-based methods overcome the shortcomings of complexity-based ones in a way. However, some small differences in structures and functional groups may lead to significantly different synthetic accessibility.[8] The reason is that chemistry is not sufficiently considered in the starting material-based methods.

Comparing with complexity-based and starting material-based methods, retrosynthesis-based methods incorporate more knowledge of chemistry. It has been thought to be the most acceptant to experimental medicinal chemists since it attempts to mimic the thought process employed by them. In retrosynthesis-based method, a target molecule is first disconnected to starting materials by retrosynthetic analysis that is carried out using methods similar to those used in Computer Assisted Organic Synthesis (CAOS) system,[18] based on which the synthetic accessibility is then estimated.

Among the three categories of automated tools, the complexity-based and starting material-based methods have been more widely applied to the estimation of synthetic accessibility in the practical *de novo* drug design or vHTS due to their simplicity and hence a high running speed. Although the retrosynthesis-based strategy likely outperforms the other two methods in principle, its practical application in drug discovery is not spread yet due to the following challenging problems involved in the current retrosynthetic analysis systems. At present, most of the retrosynthetic analysis systems, such as WODCA,[17] LHASA,[19,20] and SECS,[21] perform interactively a comprehensive retrosynthetic analysis. These highly interactive systems can be easily used to estimate the synthetic accessibility. However, the use of these systems needs considerable knowledge of organic chemistry. Further, they can only be used to analyze individual compounds within a reasonable time scale. Recently some automated retrosynthetic analysis systems have been suggested;[22] however, they face the problem of combinatorial explosion of the synthesis tree and are also very time-consuming, which restrict their use in rapid assessment of synthetic accessibility to a large number of chemicals in drug discovery.

In this account, we shall propose a new rapid method for the assessment of synthetic accessibility, named RASA (**R**etrosynthesis-based **A**ssessment of **S**ynthetic **A**ccessibility), which is a retrosynthesis-based method. In RASA, some new strategies are suggested to reduce the synthesis tree, hence limiting the combinatorial explosion. A new scoring function for the assessment of synthetic accessibility is then developed based on the miscellaneous information from the established synthesis tree.

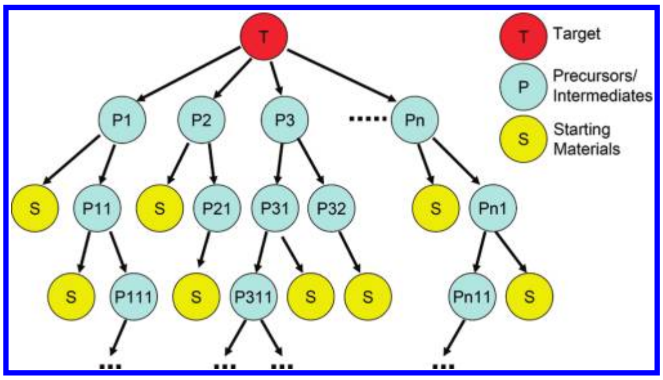


Figure 1. Schematic diagram of a synthesis tree.

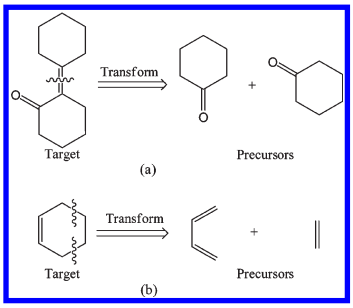


Figure 2. Two examples of transform: aldol condensation (a) and Diels−Alder reaction (b).

METHODOLOGY

Construction of Synthesis Tree Based on Retrosynthetic Analysis. The concept of retrosynthetic analysis was formalized first by Corey in 1960s.[23] Retrosynthetic analysis is achieved by disconnecting a target molecule into simpler precursor structures without assumptions regarding starting materials according to a set of transforms that represent the reversal of chemical reactions. Each precursor is examined using the same method. This procedure is repeated until simple or commercially available structures are reached. This process finally results in a comprehensive synthesis tree, which is schematically shown in Figure 1. In the synthesis tree, the root corresponds to the target compound, leaves represent starting materials, and other nodes are the precursors or intermediates. A path from the root to a leaf represents a possible synthetic route of the target molecule. However, an automated retrosynthetic analysis strategy faces some challenges including the combinatorial explosion of the synthesis tree and the time-consuming problem, which limits its use in rapid assessment of synthetic accessibility to a large number of chemicals in drug discovery.

RASA implements a very similar philosophy of retrosynthetic analysis as formulated by Corey.[24] In RASA, there are 143 retrosynthetic transforms which were derived from the classical and frequently used chemical reactions collected by us. Each transform represents a subtype of reactions. These transforms are encoded in a modified Transform Description Language.[25] As examples, Figure 2 illustrates two transforms: aldol condensation

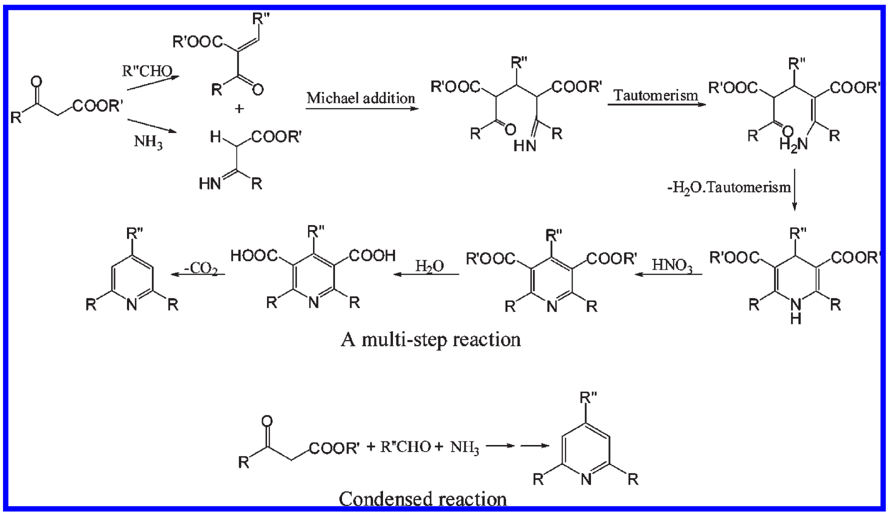


Figure 3. An example of condensing a classical multistep reaction to a one-step reaction.

and Diels–Alder reaction. Ten more transforms (as examples) presented in the modified Transform Description language code are given in the Supporting Information. Based on the established database of retrosynthetic transforms, the synthesis tree is constructed by using a breadth first algorithm. In order to overcome the combinatorial explosion of the synthesis tree and the time-consuming problem, in RASA, a series of strategies, which are derived from the known knowledge of chemistry, are suggested to prune the synthesis tree. Details for these strategies are given as follows.

(1). Limiting the Search Depth of the Synthesis Tree. In RASA, two termination conditions are used to limit the search depth of the synthesis tree. One is based on the starting materials. A database (in SD format) of available starting materials was established in advance, which contains 76,648 compounds currently. These chemicals are all on the off-the-shelf catalogues of providers (Users are allowed to modify the library according to the updated market information.). Following each retrosynthetic transform, the created precursor structures are used as a query to search the starting materials database. If all the precursors in a certain node can find their matched structures in the starting materials database, the corresponding path will be terminated at this node. The second termination condition is based on the maximum search depth. If the search depth of a reaction path exceeds a specified maximum depth of the synthesis tree, the disconnecting process will be terminated. Roles played by each strategy on reducing the search space depend on the structures of target molecules.

(2). Condensing Classical Multistep Reactions to One-Step Reactions. In order to shorten the search depth and hence reduce the synthesis tree, some well-known classical multistep reactions are condensed to one-step reactions. All the intermediates in these multistep reactions are ignored. These one-step reactions condensed are encoded as transforms. For an example, pyridine derivatives can be synthesized by the Hantzsch reaction,[26] which is a seven-step reaction from starting materials (see Figure 3). In RASA, just the pyridine derivative and its starting materials occur in the synthesis tree with the intermediates in the route being ignored. Currently, in RASA, a total of 23 multistep reactions have been compressed to one-step reactions.

(3). Ignoring the Reaction Routes That Suffer Too Much Unfavorable Electronic and/or Steric Properties. According to the knowledge of organic chemistry, electronic properties of substituents of reactants, such as electron-withdrawing and electron-donating properties, considerably influence the activity of the reactive center (RC). Unfavorable electronic properties often lead to a higher reaction barrier, indicating a rigorous reaction condition. In RASA, we expanded the Transform Description Language so that it can deal with the electronic and steric effects in transforms. A database containing electron-withdrawing groups and electron-donating groups was developed in advance. For a given transform, the electronic properties of substitutes for each reaction site are determined by mapping them to the database. Figure 4 shows one typical example that is a Diels–Alder reaction.[27] An electron-donating substituent in the reaction center (RC) atom or its neighbor, such as CH_3 (see Figure 4a) in C-2 of buta-1,3-diene benefits the addition reaction, and an electron-withdrawing substituent, such as nitrile (see Figure 4b), hinders this reaction. In RASA, a basal score is assigned in advance for each transform with the reactants (precursors) having no substituent except hydrogen. In a specific transform, an awarding score will be offered if the electronic properties of substituents on precursors benefit the reaction. Otherwise, a penalty score will be given.

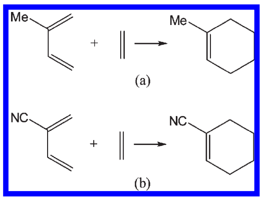


Figure 4. An example of the Diels–Alder reactions showing that the electron-donating substituent CH_3 in C-2 of buta-1,3-diene benefits the addition reaction (a) and the electron-withdrawing substituent nitrile hinders this reaction (b).

Similar to the electronic effect, the steric effect is also a very important factor that influences reaction activity. For example, in S_N2 reactions, the steric effect is generally one of the major factors that determine the reaction rate. In most cases, bulky substituents often hinder the reactants to come close to each other and hence not benefiting the reaction. In RASA, the size of a bulky substituent group is determined by counting the number of its non-hydrogen atoms. When applying to a specific

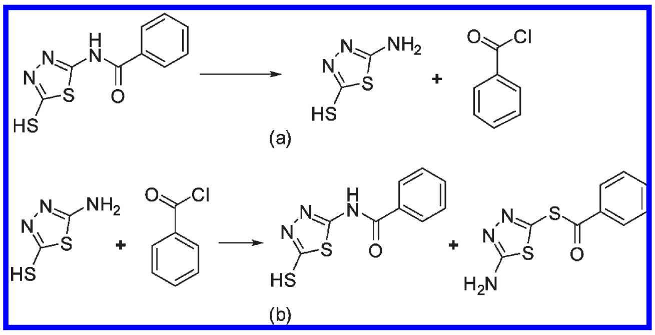


Figure 5. An example of reactions with one of the reactants having many reaction points.

transform, RASA will check the RC atom and its closest neighbors whether there are bulky substituents. If a bulky substituent is found, an appropriate penalty score that is based on the size of the bulky group will be assigned to this transform.

If the accumulated total score associated with the electronic and steric properties is lower than a specified threshold value, the corresponding route will be ignored in the synthesis tree. Finally it is necessary to mention that how much the introduction of a reactivity parameter actually reduces the search space depends on the structure of a target molecule. In general, this strategy would play a relatively more significant role in reducing the search space for a target molecule containing more heteroatoms and bulky substitution groups.

(4). Omitting the Reaction Routes with Which Many Concomitant Products May Be Produced. A target molecule may be disconnected to some precursors who have many reaction points, implying that when these precursors are used in synthesis the product might be a complicated mixture. Figure 5 shows one example. The target molecule N-(5-mercapto-1,3,4-thiadiazol-2-yl) benzamide is supposed to disconnect to two precursors, 5-amino-1,3,4-thiadiazole-2-thiol and benzoyl chloride (see Figure 5(a)). Actually, the reaction of 5-amino-1,3,4-thiadiazole-2-thiol and benzoyl chloride can produce two products, namely N-(5-mercapto-1,3,4-thiadiazol-2-yl)benzamide and S-5-amino-1,3,4-thiadiazol-2-yl benzothioate (see Figure 5(b)), since 5-amino-1,3,4-thiadiazole-2-thiol has two reaction points. In these cases, the separation/purification of the target product is relatively difficult, and the yield of the target product is also reduced. In RASA, a very simple method was used to search possible alternative reaction points. First, we find out the characteristic functional group of the reaction point by comparing the precursor and its target, followed by searching the characteristic functional group in the precursor. If the same characteristic functional group is found somewhere different from the original reaction point, it is thought as a reaction with multiple reaction points. This method is very simple and easily implemented although it may not be the best one. In RASA, a penalty score is assigned to the target product (a node in the synthesis tree) based on the number of reaction points of the precursors (a child node in the synthesis tree). If the accumulated penalty score associated with the whole reaction route is larger than a specified threshold value, the corresponding route will be omitted from the synthesis tree. Again, it is necessary to mention that the effect that this strategy has on reducing synthesis tree depends on the structure of a target molecule.

(5). Ignoring the Reaction Routes That Lead to Chemically Not-Allowed or Unstable Intermediates. In some cases, a transform may produce chemically not-allowed or unstable intermediates. For example, when disconnecting some heteroaromatic compounds with substituent N, O, or S atoms linked to the reaction center, it may happen that the produced intermediates are chemically not-allowed or unstable. The syntheses involving chemically not-allowed or unstable intermediates usually need rigorous reaction conditions, such as a very low temperature or anhydrous condition. Following suggestions of several experienced organic chemists, we define some cases which are usually thought as chemically not-allowed or unstable, for examples, two or more heteroatoms bond to the same carbon atom, such as O—C—O. Figure 6 depicts some of the not-allowed or unstable situations defined in RASA. If a transform results in any one of the not-allowed situations, it will be rejected. By the way, we have to mention that ideally the encoded retrosynthetic transforms would not allow these intermediates to be generated. However it is impractical to make transforms that are both general enough and never wrong, which means that this step is also useful.

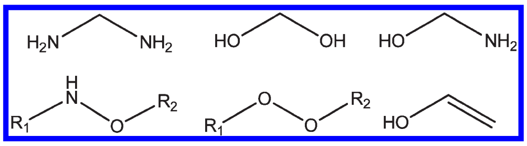


Figure 6. Some of the not-allowed or extremely unstable structures defined in RASA.

(6). Ignoring Transforms That Lead to Substantially More Complex Precursors. In order to reduce the synthesis tree, transforms that lead to substantially complex precursors will be ignored unless the complex precursors are available starting materials. In RASA, the complexity of a molecule is calculated by a formula (see eq 1) proposed by us recently;[10] this is actually a modified version of Barone's molecular complexity,[15] in which the contribution of chirality to the synthetic accessibility was also involved in addition to the contributions of rings, interatomic connections, and atom types. If the difficulty (F value, see eq 1) of any precursor of a target compound is substantially larger than that of the target compound, the transform will be ignored

$$F = F_{ring} + F_{connect} + F_{type} + F_{chirality} \tag{1}$$

where F_{ring} represents the contribution of rings, which is calculated by the following formula

$$F_{ring} = \sum_{i=1}^{i=nring} cons \times size(i) \tag{2}$$

where *nring* is the number of rings in the molecule, *size(i)* is the number of atoms in the ring *i*, and *cons* is a constant coefficient (which is set to 6 here).

$F_{connect}$ in eq 3 refers to the contribution of interatomic connections. The calculation formula is

$$F_{connect} = \sum_{i=1}^{natom} f(atom_i) \tag{3}$$

$$f(atom_i) = \begin{cases} 24 & if \quad con_{d}egree = 4 \\ 12 & if \quad con_{d}egree = 3 \\ 6 & if \quad con_{d}egree = 2 \\ 3 & if \quad con_{d}egree = 1 \end{cases} \tag{4}$$

where *natom* is the number of atoms in the molecule, and *con_degree* is the connection degree of atom, i.e. the number of atoms bonded to the specified atom.

F_{type} in eq 1 refers to the contribution of atom type. If it is a carbon atom, the value is set to 3. Otherwise, the value is 6

$$F_{type} = \sum_{i=1}^{natom} f(at_i) \tag{5}$$

$$f(at_i) = \begin{cases} 3 & \text{if} \quad atomic_t ype = C \\ 6 & \text{if} \quad atomic_t ype \neq C \end{cases} \tag{6}$$

The last term in eq 1, namely $F_{chirality}$, represents the contribution of chiral centers. This calculation formula is

$$F_{chirality} = diff \times nchiral \tag{7}$$

where *diff* refers to a "difficulty" coefficient for each chiral center, and it is set to 20 here. *nchiral* is the number of chiral centers in the molecule.

The Scoring Function Based on the Synthesis Tree. In RASA, a new scoring function, namely RASA-score, based on the synthesis tree for the estimation of synthetic accessibility is proposed. The scoring function involves contributions from the optional effective synthetic routes, the complexity of reaction, and the difficulty of separation/purification associated with the most favorable synthetic route. These individual components are combined to an overall score of synthetic accessibility by an additive scheme (see eq 8)

$$\text{RASA-score} = aC_{\text{esr}} + bC_{mfr} + cC_{sp} \tag{8}$$

Here C_{esr} represents the contribution of optional effective synthetic routes, C_{mfr} is the complexity of the most favorable reaction, C_{sp} stands for the difficulty of the separation and purification associated with the most favorable reaction, and *a*, *b*, and *c* are the weights of the individual components, which were calculated by linear regression analysis based on estimated synthetic accessibility scores of a selected training set compounds given by medicinal chemists. This strategy of calculating synthetic accessibility, namely by summing the contributions from individual components with weights of individual components obtained by linear regression analysis, has previously been used by Boda et al. in a study of structure and reaction based evaluation of synthetic accessibility.[12] In their investigation, this strategy has been shown to be a very good and effective approach to the estimation of synthetic accessibility.

(1). The Contribution of Optional Effective Synthetic Routes. As indicated before, a path from the root to any leaf on a synthesis tree corresponds to a possible reaction route. There are two types of reaction routes: "effective" and "ineffective" paths. A reaction route is called "effective" route if the leaf corresponds to starting materials, otherwise it is an "ineffective" route. In general, a compound that can be made in multiple ways would be logically considered easier to synthesize than one with a single route. In other words, a synthesis tree with more effective routes means that the chemical synthesis of the "root" compound is relatively easier. Thus, in our proposed scoring function, we define a term C_{esr} to reflect this fact

$$C_{\text{esr}} = 1/N_{effective} \tag{9}$$

where $N_{effective}$ is the number of the 'effective' routes.

(2). The Complexity of the Most Favorable Reaction. We first assigned a synthetic accessibility score for each transform in the transform database; these scores are actually the average estimates provided by five invited experienced medicinal chemists. The medicinal chemists were asked to give their scores within the same scale: the maximum value 10 meaning most difficult to synthesize and the minimum value 1 easiest to synthesize. When deciding the score for each transform, the mainly considered factors by the medicinal chemists include the experimental conditions and the yield rate of reaction. The experimental conditions include the reaction temperature, pressure, and catalysts used. A transform with rigorous conditions, for example, an extremely high/low temperature or pressure, expensive catalysts, will be assigned a higher score, otherwise a lower score. A transform with a higher yield rate is assigned a lower score or else a higher score. When applying the transforms to a target compound, the contributions of electronic and steric properties as well as reaction points of the precursors are also incorporated into the reaction complexity; the contributions of electronic and steric properties as well as reaction points are represented by the awarding or penalty scores, which have been calculated and saved in the process of constructing synthesis tree (see above). The complexities of transforms along a reaction path are summed up to form the total complexity of the reaction path. The reaction path with the lowest complexity is the most favorable reaction path.

(3). The Difficulty of Separation/Purification Associated with the Most Favorable Synthetic Route. Separation and purification of the products are critically important for the synthesis of drug molecules. In general, the difference of polarity between reactants and products is larger, the separation and purification are easier. Here, the difference of logP (the logarithm of n-octanol/water partition coefficient) of chemicals is adopted to represent the difference of polarity since the logP of chemicals can be easily and accurately calculated based on their chemical structures.[28] The reciprocal of the absolute value of ΔlogP (1/ΔlogP) is taken as the contribution of separation/purification to the overall RASA-score. This is reasonable since a small value of ΔlogP (indicating difficult to separation/purification) means a large value of 1/ΔlogP. The difference of logP (ΔlogP) between the target molecule and its precursors is calculated for each step along the most favorable reaction path. Then the contribution of separation/purification is calculated by summing up all the reciprocals of the absolute values of ΔlogP. In RASA, ALogP was used.[29]

Implementation of RASA. The source code for the implementation of RASA was written in C programming language (gcc) under the UNIX/LINUX operating system. The RASA program is available free of charge to not-for-profit institution upon request from the corresponding author.

RESULTS AND DISCUSSION

Establishment of the Scoring Function RASA-Score for the Synthetic Accessibility. The proposed scoring function RASA-score for the synthetic accessibility is a sum of three weighted individual components: contribution from the optional effective synthetic routes, the complexity of the most favorable synthetic route, and the difficulty of separation/purification associated with the most favorable synthetic route (see eq 8). The weights of the individual components in eq 8 were determined by linear

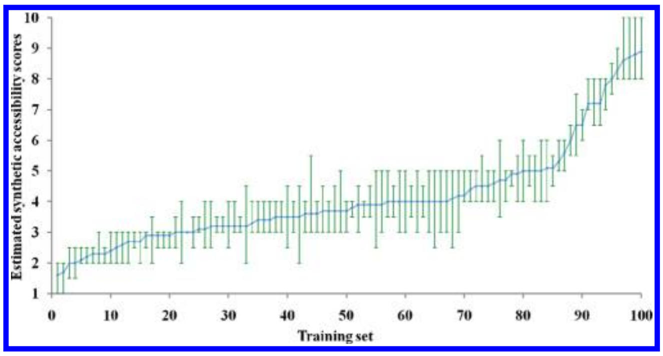


Figure 7. The minimum and maximum values of the estimated synthetic accessibility scores by five medicinal chemists for the training set compounds. Compounds are sorted by their average estimates that are indicated by the line.

Table 1. Correlation Coefficients between Synthetic Accessibility Estimates Given by Five Individual Medicinal Chemists for the Training Set Compounds

	chemist 1	chemist 2	chemist 3	chemist 4	chemist 5	average
chemist 1	-	0.830	0.763	0.763	0.726	0.898
chemist 2		-	0.837	0.777	0.735	0.918
chemist 3			-	0.836	0.813	0.924
chemist 4				-	0.819	0.920
chemist 5					-	0.886

regression analysis based on the estimated synthetic accessibility scores of training set compounds.

100 compounds were selected from the CMC (Comprehensive Medicinal Chemistry) database to form the training set. Five experienced medicinal chemists (group G1) were invited to evaluate the synthetic accessibility. Sufficient time (a period of one month) was given, and they were allowed to look for any resources, including journals and SciFinder Scholar database. They were asked to act independently and give their scores in the same scale from 1 to 10, with smaller values assigned to compounds that they think to easier to synthesize, and larger values to compounds that they consider to difficult to synthesize. Figure 7 presents the scores of the training set compounds ranked by the average estimated synthetic accessibility values (All the training set compounds together with their synthetic accessibility scores estimated by the medicinal chemists are presented in the Supporting Information). Table 1 shows the correlation coefficient (r^2) for each estimated score pair. The correlation coefficients are between 0.726 and 0.837, indicating that the chemists have a considerable degree of agreement for the synthetic accessibility of the training set compounds. On the other hand, we have also noticed that there are still some differences among the estimated scores for the training set compounds. The average estimated scores are anticipated to be more able to reflect the synthetic accessibility than scores estimated by individual medicinal chemists and can be ratified by more medicinal chemists; the correlation coefficients among the scores estimated by individual medicinal chemists and average ones are between 0.886 and 0.924. Thus, the average estimated scores were used to train the synthetic accessibility scoring function RASA-score.

After the weights of components in RASA-score were determined, the synthetic accessibility scores for the training set compounds were calculated by the established RASA-score. The calculated scores together with the average estimated scores for the 100 training set compounds are shown in Figure 8a. The calculated scores correlate very well with the average estimated scores with a correlation coefficient of 0.866 (see Table 2 and Figure 8b); this value is also comparable with the correlation coefficients between estimated scores by individual chemists and the average estimated scores (0.886−0.924, see Table 1). All of these indicate that the synthetic accessibility scores calculated by RASA have an acceptable degree of agreement with chemists.

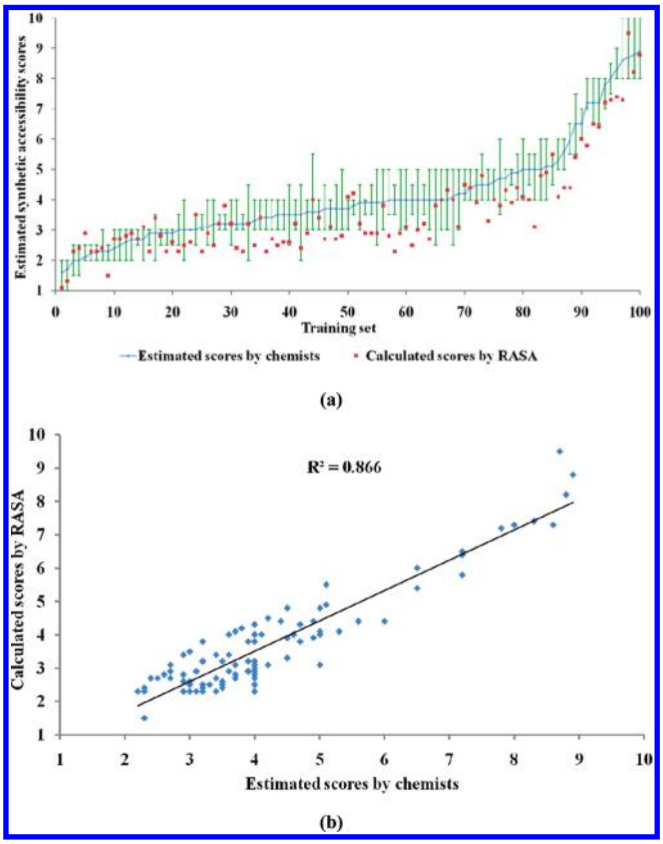


Figure 8. (a) Calculated values of RASA-score (red) together with the estimated synthetic accessibility scores by five medicinal chemists (blue) for the training set compounds. Error on blue points indicates standard error of mean of estimations by 5 chemists. (b) Correlation analysis of the calculated scores by RASA and average estimated scores by five medicinal chemists for the training set compounds.

Table 2. Correlation Coefficients between the Calculated Scores by RASA and the Estimates by Medicinal Chemists for the Training Set Compounds

	chemist 1	chemist 2	chemist 3	chemist 4	chemist 5	average
RASA	0.806	0.782	0.804	0.786	0.773	0.866

Figure 9 presents the contributions of individual components of RASA-score for the 100 compounds. Obviously, all three of the components of RASA-score have considerable contributions to the total RASA-score, and the biggest contribution comes from the complexity of the most favorable synthetic route. The correlation coefficients of the individual components with the total RASA-score values are listed in Table 3. Clearly all the components correlate positively with the total RASA-score

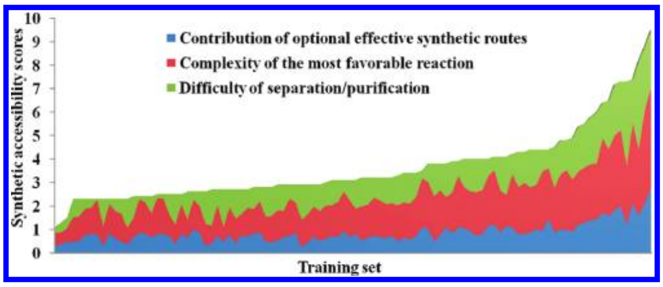


Figure 9. Contributions of individual components of RASA-score to the overall RASA-score for the training set compounds.

Table 3. Correlation Coefficients between the Values of Individual Components of RASA-Score and Overall RASA-Score Values for the Training Set Compounds

component	correlation
contribution of optional effective synthetic routes	0.772
complexity of the most favorable reaction	0.658
difficulty of separation/purification	0.528

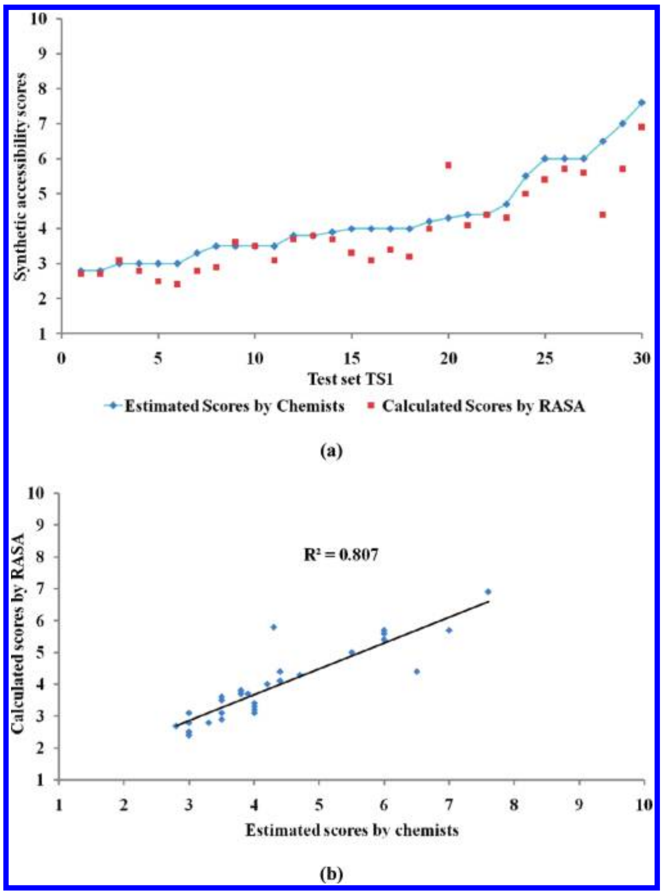


Figure 10. (a) Calculated values of RASA-score together with the estimated synthetic accessibility scores by the group G1 medicinal chemists for the test set TS1 compounds. (b) Correlation analysis of the calculated scores by RASA and average estimated scores by the group G1 medicinal chemists for the test set TS1 compounds.

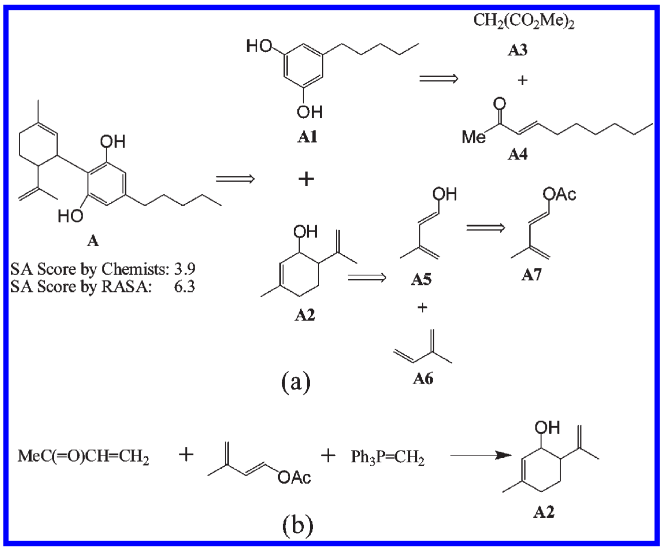


Figure 11. (a) The most favorable reaction route for compound A derived from the retrosynthetic analysis. (b) A practical synthetic route for the intermediate A2.

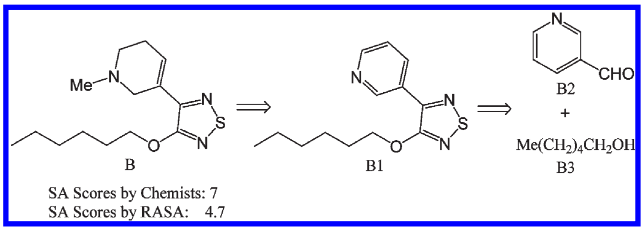


Figure 12. The most favorable reaction route for compound B derived from the retrosynthetic analysis.

values, and the first two components in eq 8, namely the optional effective synthetic routes and the complexity of the most favorable synthetic route, have the best correlation with the total RASA-score; the correlation coefficients are 0.772 and 0.658, respectively.

Validation of RASA-Score by External Independent Test Sets. Two external test sets, TS1 and TS2, were adopted to validate RASA-score. TS1 contains 30 compounds, which were taken from the CMC database. The group G1 medicinal chemists were asked again to score the TS1 compounds on a scale from 1 to 10 as before. The average estimated scores obtained from the five chemists together with the calculated scores by RASA are shown in Figure 10a. The correlation coefficient (r^2) between them is 0.807 (see Figure 10b), which is comparable with the correlation coefficients between estimated scores by individual chemists and the average estimated scores (0.837−0.932, see the Supporting Information).

From Figure 10 (a) and (b), we can notice two outliers, for which there are big differences between the calculated scores and the average estimates by medicinal chemists. The chemical structures of the two outliers are shown in Figure 11 (compound A) and Figure 11 (compound B), respectively. For compound A, the calculated score by RASA is 6.3, which is noticeably larger than the average estimated score (3.9). A careful analysis shows that the most favorable reaction route derived from the retrosynthetic analysis corresponds to a three-step reaction (see Figure 11a). In this reaction route, one intermediate, namely A2, is supposed to be synthesized by a two-step reaction, in which one of the intermediates, i.e. A5 (see Figure 11a), is relatively unstable. However, when consulting the medicinal chemists who gave the estimated scores, we were told that compound A2 could be synthesized through one-step condensation reaction[30] (shown in Figure 11b) and in this way the unstable A5 is avoided. This could be used to interpret why RASA gives a larger score and a smaller estimated score was given by the medicinal chemists. Contrary to compound A, the calculated synthesis accessibility score for compound B by RASA is 4.7, which is smaller than the average estimated score (7) by medicinal chemists. Figure 12 shows the most favorable reaction route of compound B derived from the retrosynthetic analysis, which is a two-step reaction. In the most favorable reaction route, the intermediate B3 (nicotinaldehyde) is actually a starting material. Nevertheless, most of the medicinal chemists did not see B3 as a starting material, and they continued to disconnect B3 into smaller fragments. This could be the reason why they gave a larger estimated score for compound B.

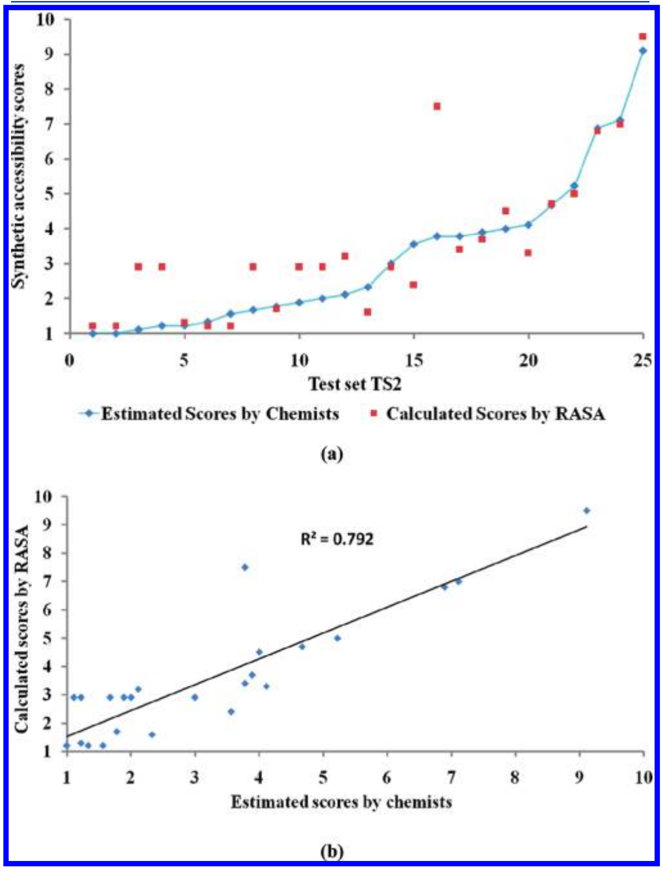


Figure 13. (a) Calculated values of RASA-score together with the estimated synthetic accessibility scores by the group G2 medicinal chemists for the test set TS2 compounds. (b) Correlation analysis of the calculated scores by RASA and average estimated scores by the group G2 medicinal chemists for the test set TS2 compounds.

To further validate RASA-score, another test set TS2 that contains 25 compounds was also adopted. The TS2 compounds and their estimated scores (by other medicinal chemists, here called group G2 medicinal chemists) were all taken from ref 13; the TS2 was built based on the original test set in ref 13 by removing the natural product-like compounds and those violating Lipinski rule of 5. The calculated scores by RASA and average estimated scores by chemists are detailed in Figure 13(a). The correlation coefficient (r^2) between them is 0.792 (Figure 13b), which further confirms the reliability of RASA.

There is also an outlier (see Figure 13), whose calculated score by RASA is 7.5, but the average estimate is 3.78. The chemical structure of the outlier is shown in Figure 14 (compound C). A careful analysis shows that the most favorable reaction route derived from the retrosynthetic analysis corresponds to a three-step reaction (see Figure 14). One of the main reasons which result in a larger score given by RASA may relate to the difficulty of separation. The calculated ΔlogP between compound C and the intermediate C1 is 0.58, and that between the intermediate C3 and C5 is 0.03. The very small values of ΔlogP lead that RASA assigns the difficulty of separation a large contribution to the overall RASA-score. In addition, the unstable chemical structure in the intermediate C2 and C3, namely α-hydroxy ketone, which could be easily dehydrated to α,β-unsaturated ketone, could be another reason that RASA assigns a larger score for compound C.

Evaluation of the Running Speed of RASA. The CMC database that contains 8895 drug molecules was used to evaluate the running speed of RASA. It took a total of 78 h for all of these compounds on a 2.5 GHz (Intel Xeon E5420 CPU) Linux PC. The average time used for each compound is about 31 s. Route Designer, which is a famous automatic synthetic route generator based on retrosynthetic analysis of target molecules, needs an average time of five minutes to perform a retrosynthetic analysis for one molecule. The time used for retrosynthetic analysis is just a part of the running time of RASA. This demonstrates that RASA has an obvious

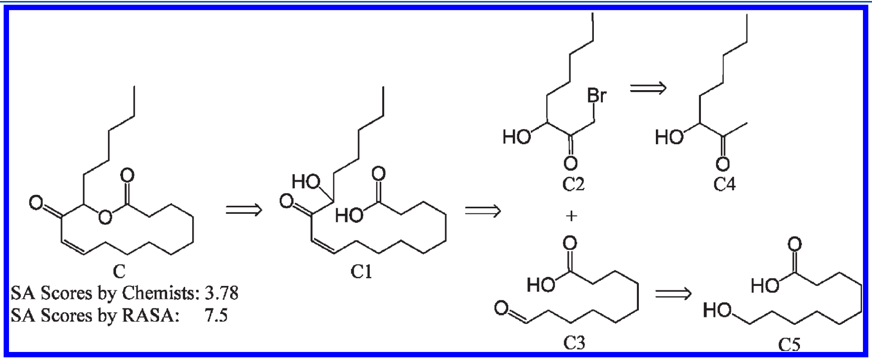


Figure 14. The most favorable reaction route for compound C derived from the retrosynthetic analysis.

advantage in running speed compared with similar automated retrosynthetic analysis packages.

CONCLUDING REMARKS

A new rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules, called RASA, is described here. RASA first constructs a synthesis tree for the target molecule based on retrosynthetic analysis; in this process a series of strategies are used for limiting combinatorial explosion of the synthesis tree. A scoring method for the assessment of synthetic accessibility is then suggested based on the optional effective synthetic routes, the complexity of reactions, and the difficulty of separation/purification associated with the most favorable synthetic route. These individual components are combined to an overall score of synthetic accessibility by an additive scheme, namely RASA-score. The contributions of individual components are calibrated by linear regression analysis based on the synthetic accessibility scores of a set of selected training set compounds given by the group G1 medicinal chemists. Two external test sets, TS1 and TS2, whose synthetic accessibility scores were provided by the group G1 and G2 medicinal chemists, respectively, were adopted for the evaluation of RASA. The correlation coefficient between RASA-score values and average estimates for TS1 is 0.807 and that for TS2 is 0.792, showing that the calculated scores by RASA agree considerably well with estimated scores given by medicinal chemists.

Finally, the characteristics of RASA can be summarized as follows. (1) RASA attempts to mimic the thought process employed by medicinal chemists, making it acceptant to experimental medicinal chemists in principle. (2) A series of strategies are adopted to reduce the synthesis tree, which can help to mitigate combinatorial explosion and save time for the computation of synthetic accessibility. (3) The scoring function RASA-score involves contributions from not only the complexity of reaction but also the difficulty of separation/purification associated with the most favorable synthetic route as well as the contribution from optional effective synthetic routes. This further increases the reliability of RASA. (4) RASA not only can be used to calculate the synthetic accessibility but also is capable of suggesting possible reaction routes based on retrosynthetic analysis. In summary, the validity and reliability as well as the high speed of RASA and its capability of suggesting synthetic routes enable it a useful tool in drug discovery.

ASSOCIATED CONTENT

Ⓢ **Supporting Information.** All of the compounds used in the training set and the test set TS1 as well as the scores assigned by invited medicinal chemists and ten examples of transforms as well as detailed information including the complexity scores and reactivity penalty models. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

Phone: +86-28-85164063. Fax: +86-28-85164060. E-mail: yangsy@scu.edu.cn.

ACKNOWLEDGMENT

This work was supported by the National S&T Major Project (2009ZX09501-001, 2009ZX09103-133), the National Natural Science Foundation of China (20872100).

REFERENCES

(1) Shoichet, B. J. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(2) Ren, J. -X.; Li, L. -L.; Zou, J.; Yang, L.; Yang, J. -L.; Yang, S. -Y. Pharmacophore modeling and virtual screening for the discovery of new transforming growth factor-β type I receptor (ALK5) inhibitors. *Eur. J. Med. Chem.* **2009**, *44*, 4259–4265.

(3) Xie, H. -Z.; Li, L. -L.; Ren, J. -X.; Zou, J.; Li, Y.; Wei, Y. -Q.; Yang, S. -Y. Pharmacophore modeling study based on known Spleen tyrosine kinase inhibitors together with virtual screening for identifying novel inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 1944–1949.

(4) Pierce, A. C.; Jacobs, M.; Moody, C. S. Docking study yields four novel inhibitors of the protooncogene Pim-1 kinase. *J. Med. Chem.* **2008**, *51*, 1972–1975.

(5) Shen, J.; Tan, C. -F.; Zhang, Y.-Y.; Li, X.; Li, W.-H.; Huang, J.; Shen, X.; Tang, Y. Discovery of Potent Ligands for Estrogen Receptor β by Structure-Based Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5361–5365.

(6) Schneider, G.; Fechner, U. Computer-based de novo design of drug like molecules. *Nat. Rev. Drug. Discovery* **2005**, *4*, 649–663.

(7) Schneider, G. Trends in virtual combinatorial library design. *Curr. Med. Chem.* **2002**, *23*, 2095–2101.

(8) Baber, J. C.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681–692.

(9) Kim, H. J.; Choo, H.; Cho, Y. S.; No, K. T.; Pae, A. N. Novel GSK-3βinhibitors from sequential virtual screening. *Bioorg. Med. Chem.* **2008**, *16*, 636–643.

(10) Huang, Q.; Li, L.-L.; Yang, S.-Y.; PhDD A New Pharmacophore-based de novo Design Method of Drug-like Molecules Combined with Assessment of Synthetic Accessibility. *J. Mol. Graphics Modell.* **2010**, *28*, 775–787.

(11) Myatt, G. Computer aided estimation of synthetic accessibility. Ph.D. Thesis, 1994, School of Chemistry, University of Leeds, Leeds.

(12) Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–325.

(13) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(14) Selzer, P.; Roth, H.; Ertl, P.; Schuffenhauer, A. Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* **2005**, *9*, 310–316.

(15) Barone, R.; Chanon., M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272.

(16) Allu, T. K.; Oprea, T. I. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1237–1243.

(17) Ihlenfeldt, W.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613–2633.

(18) Todd, M. W. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.

(19) Corey, E. J.; Jorgensen, W. L. Computer-Assisted Synthetic Analysis. Synthetic Strategies Based on Appendages and the Use of Reconnective Transforms. *J. Am. Chem. Soc.* **1976**, *98*, 189–203.

(20) Johnson, A. P.; Marshall, C.; Judson, P. N. Starting material oriented retrosynthetic analysis in the LHASA program. 1. General description. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 411–417.

(21) 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.

(22) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602.

(23) Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **1967**, *14*, 19–37.

(24) Corey, E. J.; Chen, X. M. *The Logic of Chemical Synthesis*; Wiley: New York, 1989.

(25) Tubert, I. Computer-Assisted Organic Synthesis: Development of an Educational Software Package. B.Sc. thesis, 2001, National Autonomous University of Mexico, Mexico City.

(26) Hantzsch, A. Condensationsprodukte aus Aldehydammoniak und ketonartigen Verbindungen. *Ber.* **1881**, *14*, 1637–1638.

(27) Diels, O.; Alder, K. Synthesen in der hydroaromatischen Reihe. *Liebigs Ann. Chem.* **1927**, *460*, 98–122.

(28) Veith, G. D.; Austin, N. M.; Morris, R. T. A rapid method for estimating log*P* for organic chemicals. *Water Res.* **1979**, *13*, 43–47.

(29) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.

(30) Serra, S.; Brenna, E.; Fuganti, C.; Maggioni, F. Lipase-catalyzed resolution of p-menthan-3-ols monoterpenes: preparation of the enantiomer-enriched forms of menthol, isopulegol, trans- and cis-piperitol, and cis-isopiperitenol. *Tetrahedron Asymmetry* **2003**, *14*, 3261–3423.