J Comput Aided Mol Des (2007) 21:311–325
DOI 10.1007/s10822-006-9099-2

ORIGINAL PAPER

Structure and reaction based evaluation of synthetic accessibility

Krisztina Boda · Thomas Seidel · Johann Gasteiger

Received: 27 July 2006 / Accepted: 15 December 2006 / Published online: 9 February 2007

Abstract *De novo* design systems provide powerful methods to suggest a set of novel structures with high estimated binding affinity. One deficiency of these methods is that some of the suggested structures could be synthesized only with great difficulty. We devised a scoring method that rapidly evaluates synthetic accessibility of structures based on structural complexity, similarity to available starting materials and assessment of strategic bonds where a structure can be decomposed to obtain simpler fragments. These individual components were combined to an overall score of synthetic accessibility by an additive scheme. The weights of the scoring function components were calculated by linear regression analysis based on accessibility scores derived from medicinal chemists. The calculated values for synthetic accessibility agree with the values proposed by chemists to an extent that compares well with how chemists agree with each other.

Keywords *De novo* design · Reaction center · Similarity search · Synthetic accessibility

K. Boda · T. Seidel · J. Gasteiger (✉)
Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, D 91052 Erlangen, Germany
e-mail: gasteiger@chemie.uni-erlangen.de

K. Boda
e-mail: krisztina.boda@chemie.uni-erlangen.de

Present Address:
T. Seidel
Institut für Angewandte Synthesechemie, Technische Universität Wien, A 1060 Vienna, Austria
e-mail: thomas.seidel@chemit.at

Introduction

The growing number of protein 3D structures becoming available has triggered significant efforts to develop various computational methods that can exploit this structural information in order to suggest structures that bind tightly to biomolecular targets.

In the *de novo* design process, hypothetical structures are assembled from smaller components within the steric and electrostatic constraints of the binding pocket. As a consequence of the combinatorial nature of the build-up process, typical *de novo* design program can produce large numbers of proposed solutions even for a tight binding pocket. Therefore, various heuristics have to be adapted to reduce the large answer sets. One approach is to eliminate solutions with low binding score. However, a successful lead candidate is subjected to more objectives than high binding affinity. Apart from good ADMET properties (absorption, distribution, metabolism, excretion and toxicity), synthetic accessibility clearly has key importance.

Among other reasons, this is due to the fact that the ability to accurately predict binding affinity is still limited. Therefore the suggested structure has to be synthesized in order to undergo experimental in vitro testing to validate the activity predicted in silico.

Several methods already exist for predicting synthetic accessibility. These can be classified into complexity-based, starting material-based or retrosynthetic-based classes [1].

Complexity based approaches [2–4] are empirically derived heuristics that provide a rapid way to quantify molecular complexity and can be applied in order to limit combinatorial explosion of the synthesis tree in synthetic planning systems. However on their own, they are not sufficient to assess synthetic accessibility, since similarity to available starting materials, which is a very important factor, is not incorporated into this elementary approach.

This deficiency is addressed along with predicting drug-likeness in a complexity analysis method [5] that is based on the statistical distribution of various cyclic and acyclic topologies and atom substitution patterns present in starting material catalogs and drug databases.

An alternative approach to tackle the problem is to utilize synthesis design systems. There are several programs, such as WODCA [6, 7], LHASA [8] and CAESA [9], that perform a comprehensive retrosynthetic analysis. These highly interactive programs are designed to estimate synthetic accessibility and identify readily available starting materials of individual members of a series of drug candidates.

While these programs provide sophisticated tools for assessing synthetic accessibility, their complexity causes a serious drawback, namely, that they can only be used to analyze a few individual compounds within a reasonable time frame.

Therefore, one of the primary concerns in the design of a novel synthetic accessibility scoring method was to provide a reasonable compromise between the speed of the analysis and the reliability of the estimations.

Accordingly, the method described here can be utilized to rapidly assess synthetic accessibility of thousands of structures. Structures with high predicted synthetic accessibility then can be further analyzed by more comprehensive but time-consuming synthesis design programs.

Methodology

Our synthetic accessibility scoring function consists of five components. The first three components are based on structural features of the target structure and can be rapidly calculated.

1. The molecular graph complexity score [3] is a general indicator of topological and atom type composition complexity of the target structure. The method is based on graph and information theories [10] and takes account of the size, symmetry, branching, rings, multiple bonds and heteroatoms of the target molecule.
2. The ring complexity [11] component of the scoring function penalizes bridged and fused ring systems that would give rise to synthetic difficulty.
3. The stereochemical complexity is a simple counter of tetrahedral stereo centers of the target structure that contribute to poor synthetic accessibility.

The last two components of the multivariable scoring function utilize the immense structural and synthetic knowledge stored in starting material catalogs and reaction databases.

4. A structure with a complex structural motif can still easily be synthesized if its highly complex part is covered by available starting materials. Therefore, in order to reliably estimate synthetic accessibility various similarity search queries are performed to identify precursors that can be successfully mapped to the target structure with high coverage. Assessing similarity to starting materials is described in Sect. ‘Similarity to a available starting materials’.
5. Synthesis design programs perform comprehensive retrosynthetic analysis in order to transform the synthetic target structure to a sequence of progressively simpler structures along a retrosynthetic pathway, which ultimately leads to simple or commercially available starting materials [12, 13]. Accordingly, synthetic accessibility can be approximated by analyzing structural motifs where the target molecule can be decomposed into smaller components. The estimation of retrosynthetic reaction fitness is explained in Sect. ‘Retrosynthetic reaction fitness’.

The overall synthetic accessibility score (1) of a target structure is calculated by summing the weighted five individual components.

$$\text{Synthetic Accessibility} = \sum_i w_i \text{Component}_i \tag{1}$$

Similarity to available starting materials

As earlier emphasized, the synthetic accessibility of a target structure highly depends on the degree of resemblance between the structure and available starting materials. However, the concept of molecular similarity is not well-defined because it is application dependent; therefore diverse similarity measures are applied through the various stages of drug design process [14].

Most of the published similarity search methods are devised to retrieve structures from either 2D or 3D databases that possess similar biological activity with the query structure. Assessing synthetic proximity between a target structure and a set of starting materials, however, demands a different approach. Consequently, we applied transformation-based similarity criteria, which have specifically been developed for synthesis design and reaction planning [15]. By definition, two compounds are considered similar by a similarity search criterion, if their transformed structures are identical. (See examples shown in Fig. 1)

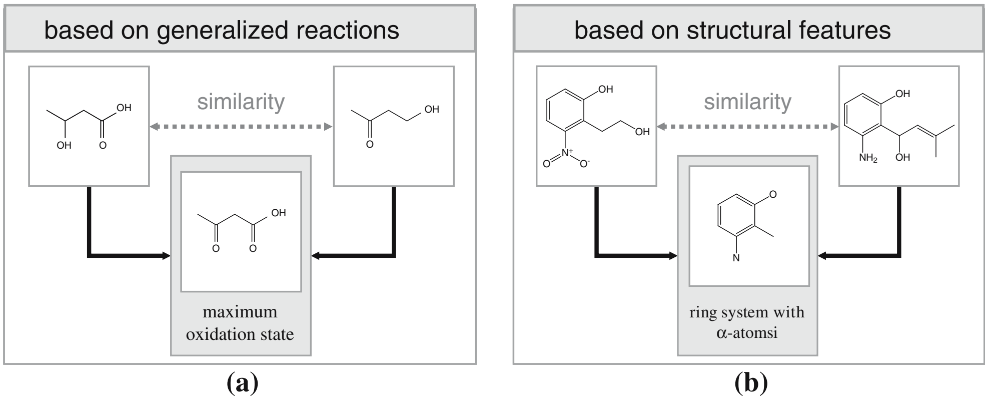

based on generalized reactions
similarity
maximum oxidation state
based on structural features
similarity
ring system with α-atomsi
(a)
(b)

Fig. 1 Concept of transformation-based similarity search (**a**) similarity criterion based on a generalized reaction, (**b**) similarity criterion based on a structural feature

Currently, 24 different similarity criteria have been implemented defining similarity based on usefulness in synthesis. Similarity search transformation can be based on generalized reactions (such as oxidation, reduction) or they can be based on topological characteristics of the structure. The latter often corresponds to the result of substructure search, for example, taking the largest ring system of a query structure. Furthermore, there are similarity definitions that combine reaction type and substructure characteristics (such as ring system with substitution pattern). Figure 2 shows examples for such similarity search transformations in order of their specificity.

The similarity search process involves modifying the target structure according to transformation associated with a certain similarity criterion. The transformed structure is then compared with each transformed compound from the catalogs of starting materials, which was derived in advance for each transformation rule. The process is illustrated in Fig. 3 where ‘‘aromatic ring system including alpha atoms’’ transformation is applied. If the transformed target structure and the transformed catalog compound are identical, then the unchanged compound from the catalog of chemicals is a potential starting material for the target compound.

For each transformed structure a unique 64-bit hash code is generated [16], thereby the computationally expensive atom-by-atom comparison required to determine structural equivalence is replaced by rapid comparison of integer numbers.

Similarity search transformations are performed for the entire starting materials catalog in advance and the transformed structures are stored in a database in the form of hash codes along with counters that store the number of frequency of occurrences of hash codes for a specific transformation. By this means, only the target structure has to be subjected to various transformations at run time achieving rapid identification of possible precursors.

Generation of starting material similarity database

A starting material similarity database is constructed from the combined Fluka, Acros and Maybridge

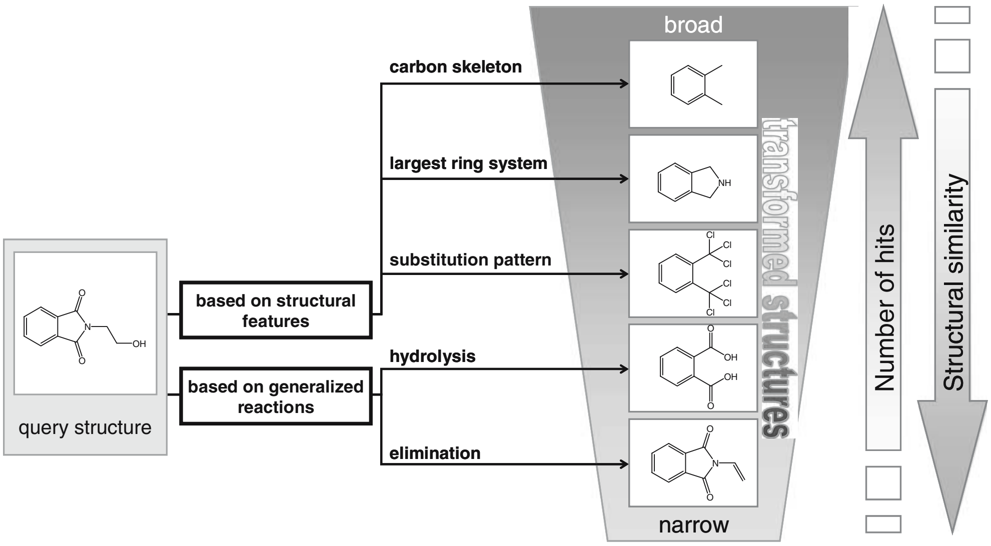
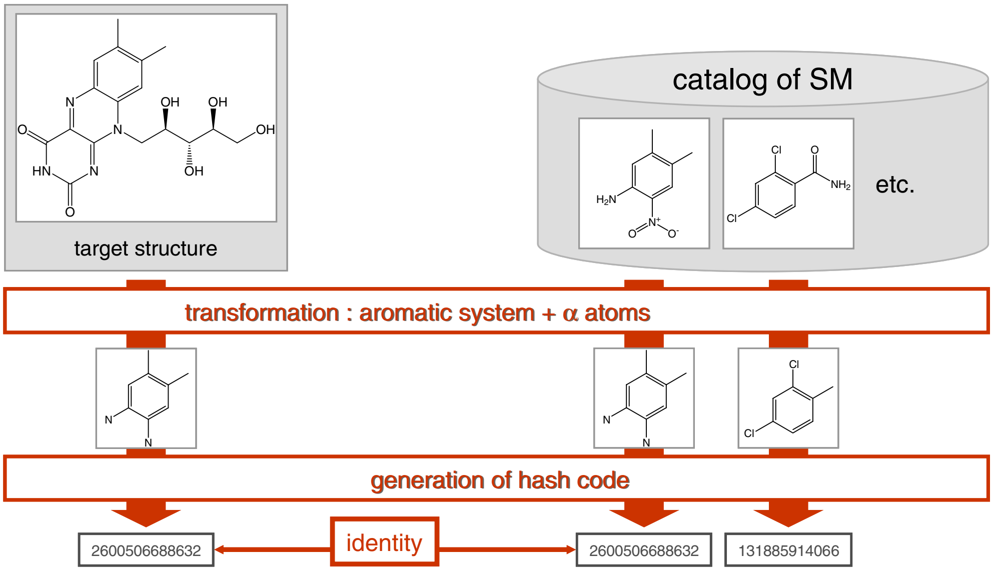


Fig. 2 Specificity of various similarity search transformations

Fig. 3 Similarity search process applying the "aromatic ring system including alpha atoms" transformation. The target (query) and all catalog compounds are transformed by the same criterion



catalogs. If a starting material occurred more than once in the united data set, then only one instance was kept. The number of starting materials removed by various filters is summarized in Table 1.

The remaining starting materials are then subjected to all available similarity search transformations [15], followed by generating hash codes and keeping account of the frequency of occurrences.

Two examples for similarity search transformations are illustrated in Fig. 4. The transformation of a staring material (1) by similarity search criterion "A" starts with identifying the aromatic atoms (2) and then taking carbon atom side chains that are attached to aromatic ring systems (3). This is followed by

Table 1 Number of structures eliminated from the combined starting material databases (Fluka + Acros + Maybridge)

	Number of structures
Starting materials (initially)[a]	84,724
Less than 3 heavy atoms	3,376
More than 50 heavy atoms	449
Identical[b]	11,254
With undesirable atom type[c]	4,301
Remaining structures	65,344

[a] Fluka, Acros and Maybridge

[b] Starting materials are considered identical if they differ only in salt counter ion or stereo configuration

[c] Only H, B, C, N, O, F, P, S, Cl, Br and I elements were considered

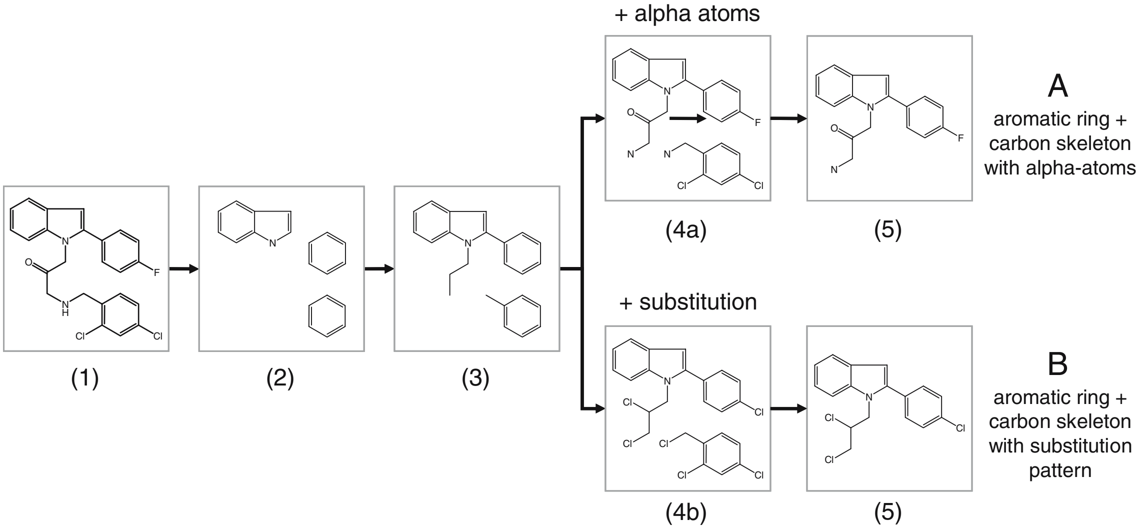

+ alpha atoms
+ substitution
(1)
(2)
(3)
(4a)
(5)
(4b)
(5)
A
aromatic ring + carbon skeleton with alpha-atoms
B
aromatic ring + carbon skeleton with substitution pattern

Fig. 4 Examples for similarity search transformations. Steps of the transformations: (1) intact starting material, (2) taking atoms that are part of any aromatic ring system, (3) taking carbon atom side chains of aromatic ring systems, (4a) considering alpha heteroatoms, (4b) considering substitution pattern (heteroatoms are substituted by chlorine atoms), (5) selecting the largest fragment

taking into consideration of the alpha heteroatoms (4a). Because this process can break the starting material into smaller unconnected fragments, the largest of these fragments is selected as a result of the similarity search transformation (5) for which a hash code is generated. The similarity transformation ‘‘B’’ differs from ‘‘A’’ by converting the alpha heteroatoms into chlorine atoms (4b) thereby marking the possible substitution sites. The distributions of structures generated by these two transformations are shown in Fig. 5.

The specificity of the similarity search criteria is directly proportional to the number of unique structures generated by the corresponding transformations. The reduction in the number of unique molecules expresses the degree of generalization inherent in a transformation.

Consequently, similarity search criterion ‘‘A’’ (in Fig. 4) is more generic than criterion ‘‘B’’. By applying the similarity transformation corresponding to criterion ‘‘A’’, 19,838 unique transformed structures are obtained from the combined starting material database. By the latter transformation, where heteroatoms are substituted with chlorine atoms, the number of retrieved structures is reduced by more than 7,000, to 12,413 unique structures.

more potential precursors are identified for the target structure. In this way, the atom scores reflect the possible coverage of starting materials on the target structure.

If a similarity search transformation identifies potential precursors for the target structure (examples shown in Fig. 8(d)–(l)), then a transformation score is calculated by Eq. 2. The transformation score ($\mathrm{SCORE}_{\mathrm{Ti}}$), which ranges between 0.0 and 1.0, is devised to penalize structural motifs that are infrequent in the similarity database. The larger is the set of potential precursors retrieved by a similarity search criterion, the smaller is the corresponding transformation score (Fig. 7).

The first three similarity search criteria (oxidation, reduction, framework) shown in Fig. 8(a)–(c) for the example structure (Fig. 6) are unable to retrieve any hits. On the other hand, the structure generated by the ‘‘aromatic ring with carbon skeleton’’ transformation (Fig. 8(l)) corresponds to the most common structure generated by that transformation; as a result, the transformation score is 0.0.

$$\mathrm{SCORE}_{\mathrm{Ti}} = 1 - \frac{\log_{10}(\text{number of hits for } \mathrm{T_i} + 1)}{\log_{10}(\text{max. number of occurrences for } \mathrm{T_i} \text{ in the database } + 1)} \tag{2}$$

The transformed structure is then mapped onto the target structure and the individual atom scores of the covered atoms are reduced, if the transformation score ($\mathrm{SCORE}_{\mathrm{Ti}}$) is smaller than the existing atom score (Eq. 3). The score of the atom that cannot be covered by the transformed starting material remains unchanged.

$$\mathrm{SCORE}_{\mathrm{Aj}} = \begin{cases} 1 & \text{, if atom can not overlapped by any transformed starting material (SM)} \\ \min\{\mathrm{Score}_{\mathrm{Ti}}\}, & \text{where atom}(\mathrm{A_j}) \text{ is covered by SM obtained from similarity transformation } (\mathrm{T_i}) \end{cases} \tag{3}$$

Evaluating similarity to starting materials

After preprocessing the starting material databases, a similarity score is calculated by mapping starting materials onto the target structure. Each atom of the target structure is associated with an atom score that is initially set to 1.0 (see target structure example in Fig. 6). These atom scores are then reduced by degrees as more and more potential precursors are identified for the target structure.

For efficiency reason, the transformed structures (Fig. 8) are mapped onto the target structure (Fig. 9) rather than the original compounds of the starting material library. The latter would result in greater coverage, however, would require computationally expensive maximum common substructure search for each hit retrieved by the similarity search transformations.

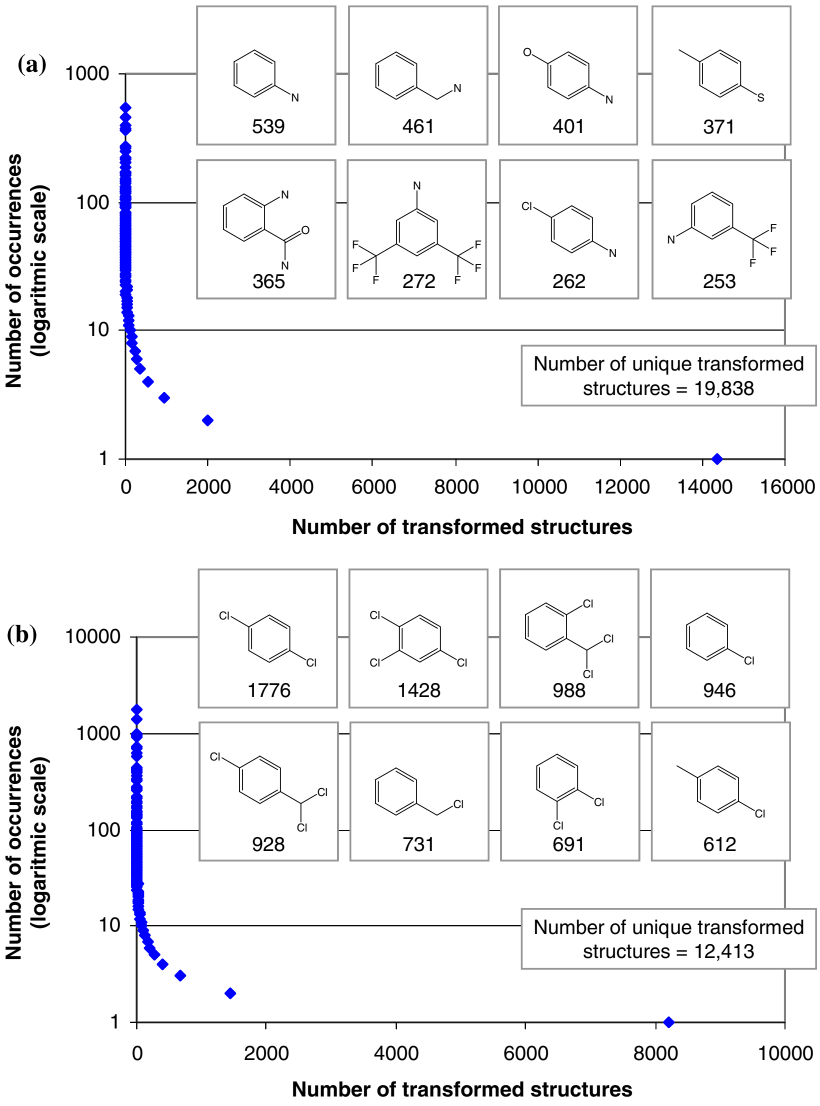


Fig. 5 Distribution of frequency of occurrences of structures resulting from similarity search transformations: (**a**) "aromatic ring + carbon skeleton with alpha-atoms" and (**b**) "aromatic ring + carbon skeleton with substitution pattern" (heteroatoms are replaced by chlorine atoms to mark the sites and degree of substitution). In both cases, the eight most frequent structures obtained by the transformation are depicted with their number of occurrences

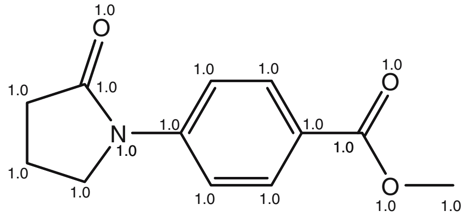


Fig. 6 Target structure with initial atom scores

Finally, the molecular starting material similarity score is calculated by summing the individual atom scores and normalizing the sum with the number of atoms. Figure 9 shows the final individual atom scores and the normalized starting material similarity score of the target structure.

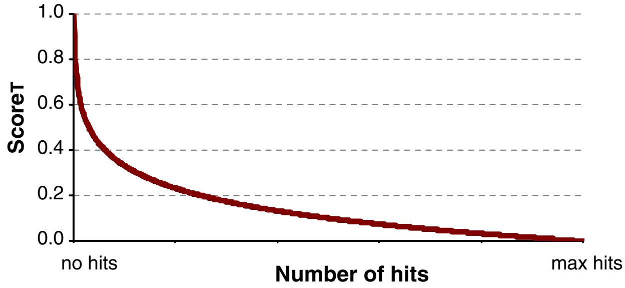
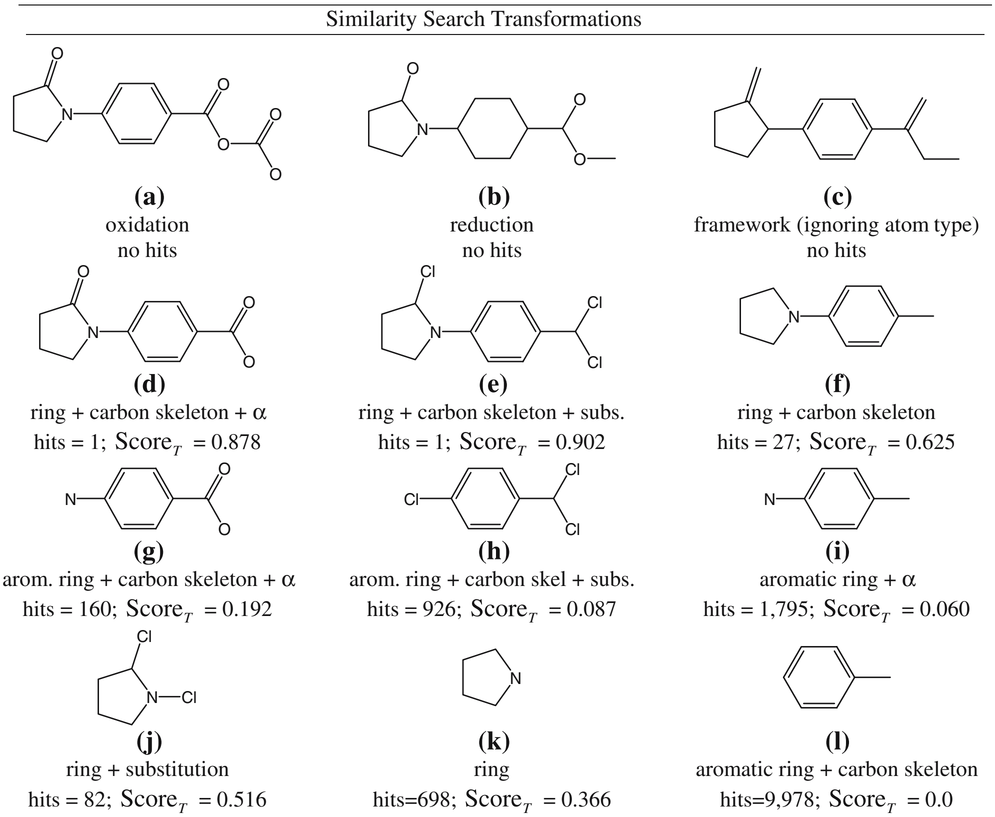


Fig. 7 The curve of the transformation score function (SCORE$_{Ti}$) that penalizes structural motifs which are infrequent or absent from the similarity database for the specific similarity search transformation

Retrosynthetic reaction fitness

The purpose of retrosynthetic reaction fitness is to estimate how easily the target structure can be

Fig. 8 Results for various similarity search transformations that are applied for the target structure shown in Fig. 9. For successful similarity searches (**d**)–(**l**) the number of hits along with corresponding transformation score are indicated



(a) oxidation no hits

(b) reduction no hits

(c) framework (ignoring atom type) no hits

(d) ring + carbon skeleton + α hits = 1; $\mathrm{Score}_T = 0.878$

(e) ring + carbon skeleton + subs. hits = 1; $\mathrm{Score}_T = 0.902$

(f) ring + carbon skeleton hits = 27; $\mathrm{Score}_T = 0.625$

(g) arom. ring + carbon skeleton + α hits = 160; $\mathrm{Score}_T = 0.192$

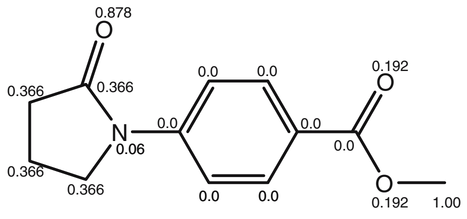
(h) arom. ring + carbon skel + subs. hits = 926; $\mathrm{Score}_T = 0.087$

(i) aromatic ring + α hits = 1,795; $\mathrm{Score}_T = 0.060$

(j) ring + substitution hits = 82; $\mathrm{Score}_T = 0.516$

(k) ring hits=698; $\mathrm{Score}_T = 0.366$

(l) aromatic ring + carbon skeleton hits=9,978; $\mathrm{Score}_T = 0.0$



final starting material similarity score = 0.237

Fig. 9 Target structure with final atom scores

dissected at strategic structural motifs to smaller and simplified precursors. The more of such key patterns are present in the target structure, the higher the probability that reasonable synthetic routes exist for the target compound. These strategic retrosynthetic motifs can be automatically extracted from available reaction databases, since a retrosynthetic transform is the explicit reverse of a synthesis reaction.

In reaction databases, transformation characteristics of reactions are automatically identified by marking the bonds directly involved in a reaction either as "change bond order" or "make/break" bond (see example in Fig. 10). Such bonds are called reaction centers. This reaction center information (RC henceforth) is utilized to define a reaction center substructure (RCSS henceforth) that consists of the atoms belonging to product reaction centers along with their direct neighbors i.e., alpha atoms. By considering the alpha atoms, the influence of the chemical environment of the reaction center can be taken into account.

According to this definition, a RCSS merges adjacent reaction center environments into one single substructure. Figure 10 shows two examples for extracting the product RCSS of synthesis reactions both having two reaction centers. A hash code is generated for each identified product RCSS, that is then inserted into a database keeping account of frequency of occurrences of unique reaction center substructures.

The product RCSS database constructed by this process can be utilized to assess retrosynthetic reaction fitness by mapping the substructures onto the target compound (see details in Sect. 'Evaluating retrosynthetic reaction fitness').

Generation of product reaction center substructure database

The Theilheimer reaction database, that contains around 47,000 high-yield functional group transformations and synthetic methods, has been utilized to build the product RCSS database.

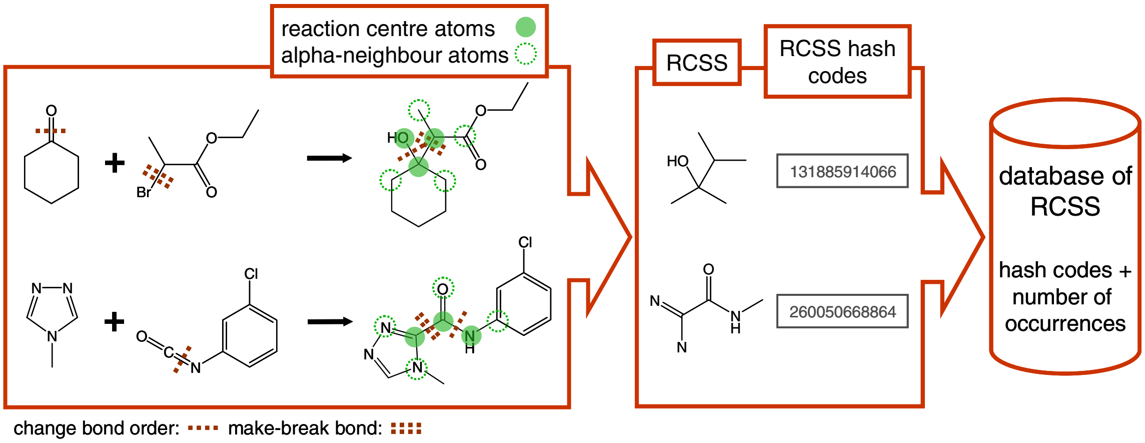


Fig. 10 Extracting product reaction center substructures from synthetic reaction

First, reactions with unspecified reaction center information or multiple product reaction centers are removed. The latter is required in order to avoid the retrieval of unconnected reaction center substructures (see examples in Fig. 11). Furthermore, reactions are also eliminated when having undesired atom types at the product reaction center. Table 2 details the number of reactions that are removed by this initial filtering process.

For each RCSS, a hash code is generated by considering the topology of the bonds at the reaction center (along with atom type, atom connection and bond order). The topology of a bond is either acyclic, aromatic or cyclic non-aromatic. By this discrimination, topologically incorrect reaction center substructure mapping onto the target structure can be avoided. Figure 12(b) shows an example where the RCSS extracted from the product of the Wittig reaction can not be mapped onto the aromatic ring due to a topological mismatch i.e., the RCSS found in the acyclic environment in the reaction database can not be mapped onto an aromatic ring.

The distribution of the 14,112 unique RCSS retrieved from the Theilheimer reaction database is shown in Fig. 13. It reveals that a significant number of RCSS occur very infrequently. Three quarter of the unique RCSS are present only once and only 2.9% of the unique RCSS occur more than 10 times.

The 4th and the 5th of the most frequent RCSS are identical substructures, however their hash codes are different due to the incorporated topology information. The keton functional group occurs as product RCSS 330 times in a chain and 324 times in a ring environment. Synthesis reactions for this topological discrimination are shown in Fig. 14.

Figure 15 illustrates the distribution of the number of heavy atoms of the retrieved substructures along with the maximum frequency of occurrences for the given size. The majority of reaction center substructures fell into the range of 4–12 heavy atoms. The line representing the maximum frequency of occurrence peaks at a size of four heavy atoms, and reaction center substructures exceeding a size of seven heavy atoms have a low frequency of occurrence.

Evaluating retrosynthetic reaction fitness

The process of calculating the retrosynthetic reaction fitness is analogous to the calculation of the starting material similarity score (Sect. 'Evaluating similarity to starting materials'). Each atom of the target structure is

Fig. 11 Examples for multiple product reaction center sites

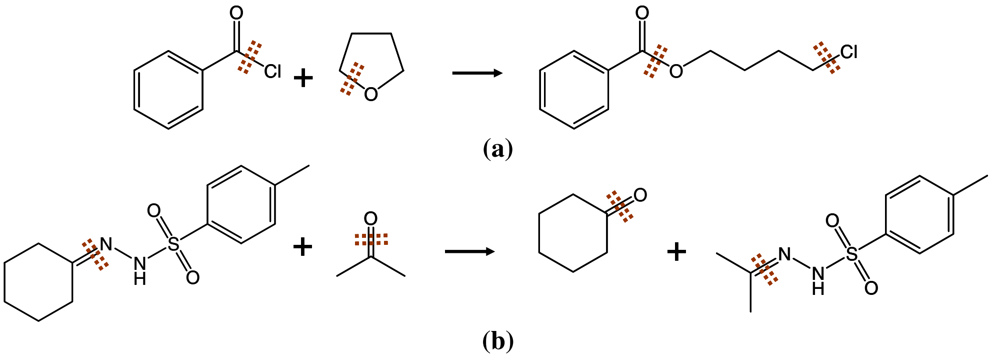


Table 2 Number of reactions eliminated from the Theilheimer reaction database

	Number of reactions
Theilheimer reactions (initially)	46,785
Missing reaction center information	3,831
Multiple product reaction center sites	4,385
With undesirable atom type[a]	1,887
Remaining reactions	36,682

[a] Only H, B, C, N, O, F, P, S, Cl, Br and I elements were considered at reaction centers

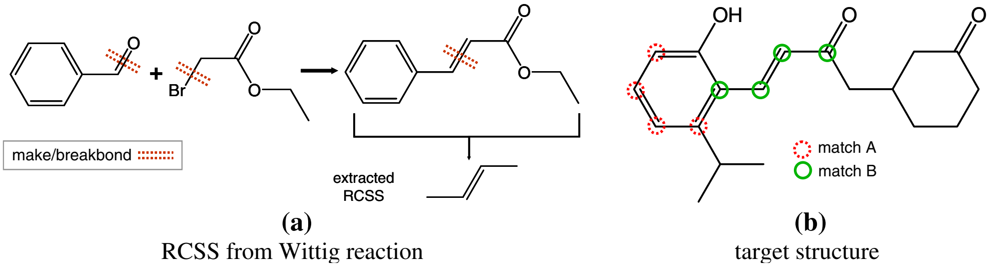
associated with an atom score that is initialized to 1.0 and gradually reduced as substructures from the RCSS database are mapped onto the target structure.

This can be achieved by performing a substructure search for each RCSS that is present in the database. However, considering the size of the database (~14,000 unique RCSS), the execution of thousands of substructure searches is inefficient, even if a fast search engine is employed.

In order to avoid this computational burden, substructures, which can coincide with real RCSS, are exhaustively enumerated in the target structure. This enumeration process is illustrated in Fig. 16. By definition, a RCSS can consist of one reacting bond with its alpha atoms or more than one reacting bonds can be merged into one RCSS. For the target structure displayed in Fig. 16, 111 potential reaction center substructures are identified when the enumeration process is limited to four potential reaction centers. The size of the enumerated RCSS is restricted in order to avoid the identification of those RCSS, which have negligible effect on the overall retrosynthetic reaction fitness score due to their low frequency of occurrences. (See chart of size versus frequency of occurrence in Fig. 15)

After the enumeration process, each substructure has to be verified, i.e., it has to be confirmed that it corresponds to a genuine product RCSS stored in the database. Therefore, a hash code is generated for the potential substructure (taking into account its topology) that is then matched against the hash codes of the product RCSS database. If the hash code is present in the database, a score is calculated based on the retrieved frequency of occurrences on the matched RCSS and the most frequent RCSS (Eq. 4) in the database. In case of the database generated from the Theilheimer database, the most frequent RCSS is the carboxylic acid group attached to a carbon atom that occurs 624 times (Fig. 13). By considering the frequency of occurs of the verified RCSS, privilege is

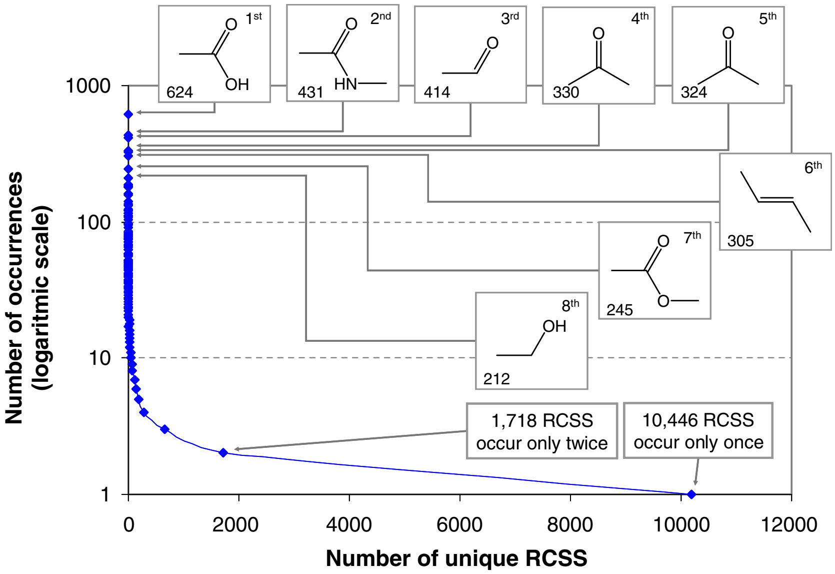
Fig. 12 Example for reaction center substructure mapping (**a**) RCSS extracted from the product of Wittig reaction (**b**) RCSS mapped to target structure. Match "A" indicates topologically incorrect, while match "B" topologically correct mapping



(**a**) RCSS from Wittig reaction

(**b**) target structure

Fig. 13 Distribution of the frequency of unique reaction center substructures retrieved from the Theilheimer database. The eight most frequent reaction center substructures with their frequency of occurrence are displayed



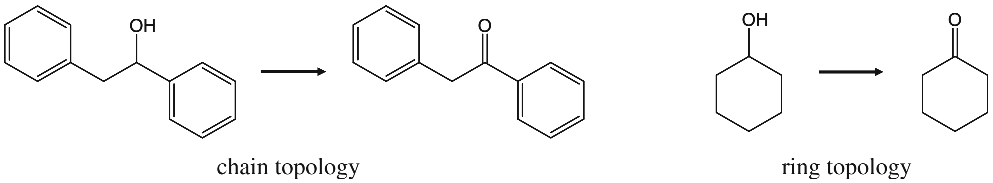


Fig. 14 Example for identical product RCSS in different topological environment

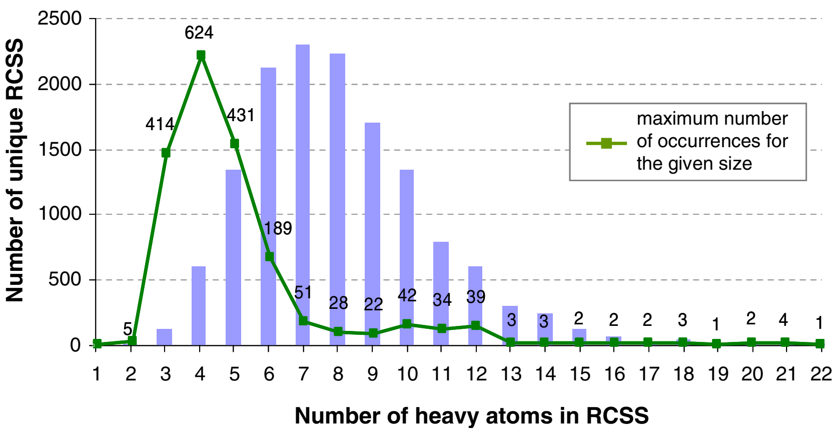


Fig. 15 Distribution of the number of heavy atoms of unique reaction center substructures retrieved from the Theilheimer database. The line represents the maximum frequency of occurrence for the given heavy atom size

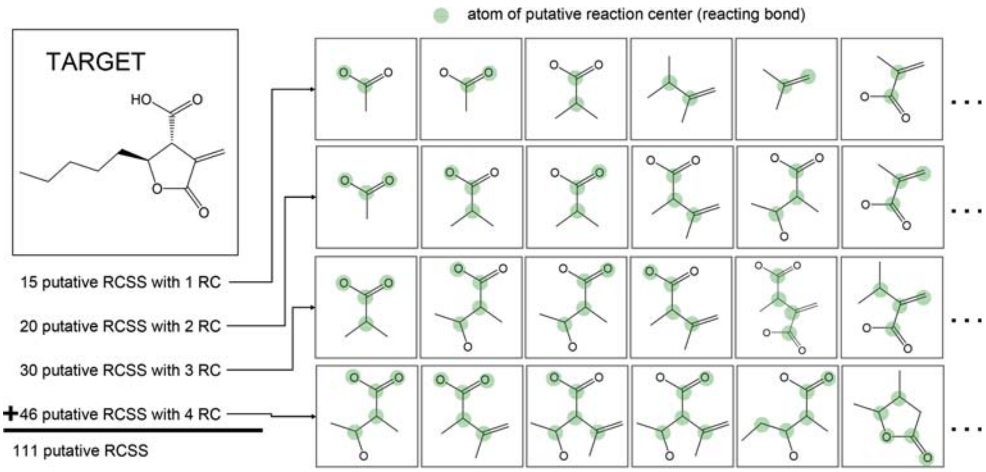

atom of putative reaction center (reacting bond)
TARGET
HO
O
O
O
15 putative RCSS with 1 RC
20 putative RCSS with 2 RC
30 putative RCSS with 3 RC
+46 putative RCSS with 4 RC
111 putative RCSS
. . .
. . .
. . .
. . .

Fig. 16 The exhaustive enumeration of potential reaction center substructures up to four adjacent reaction centers

given to structural motifs that correspond to common retrosynthetic *retrons*.

$$\mathrm{Score}_{\mathrm{RCSS}} = 1\ \mathrm{minus;}\ \frac{\log_{10}(\text{number of occurrences of matched RCSS})}{\log_{10}(\text{max . number of occurrences of RCSS} + 1)} \qquad (4)$$

At the end of the process, the overall retrosynthetic reaction fitness score is calculated by totaling up the individual atom scores and normalizing the sum with the number of heavy atoms. Figure 17 shows the final individual atom scores and the normalized retrosynthetic reaction fitness score of the target structure. The overall retrosynthetic reaction fitness ranges between 0.0 and 1.0. A smaller value indicates that the target structure can be more easily synthesized, because of the identification of more fitting retrosynthetic reactions can be found for the target structure with high frequency of occurrences.

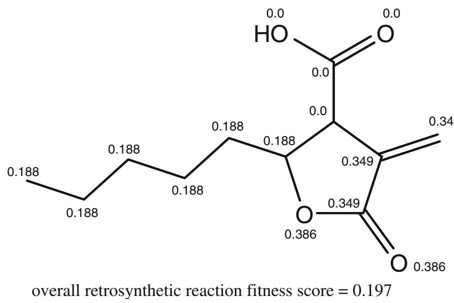


Fig. 17 Target structure with retrosynthetic reaction fitness atom scores

Analysis of synthetic accessibility estimates by medicinal chemists

In order to weight the various components of the synthetic accessibility scoring function and validate our method, we persuaded five medicinal chemists of our pharmaceutical partners to evaluate the synthetic accessibility for a set of 100 structures. Structures of the dataset have been collected from the Journal of Medicinal Chemistry and they vary in size and complexity. The distributions of various molecular properties of the dataset are shown in Fig. 18.

The chemists have been asked to score the compounds in a scale from 1 to 10, assigning score 1 to structures that have straightforward synthesis and score 10 to those structures that could be synthesized only with great difficulty.

Figure 19 shows estimations obtained for the compound of the dataset ranked by ascending order of the average estimated synthetic accessibility values. For the sake of clarity, only the average, the minimum and the maximum scores are shown in the chart. The chart reveals that chemists seem to agree on synthetic accessibility of very simple and quite complex compounds. However, there are a few structures with significant score divergence in the middle range (see examples in Fig. 20). These considerable score deviations are probably due to differences in experience and knowledge of the chemists.

A correlation coefficient is calculated for each estimate pair, these values are presented in Table 3. The coefficients, which range between 0.73 and 0.84, express an acceptable degree of agreement for such an elusive and complex concept as synthetic accessibility.

Results and discussion

After generating the two databases required for the estimation, the weights (w_i in Eq. 5) of the individual components of the synthetic accessibility function are determined by linear regression analysis based on the average accessibility scores provided by chemists.

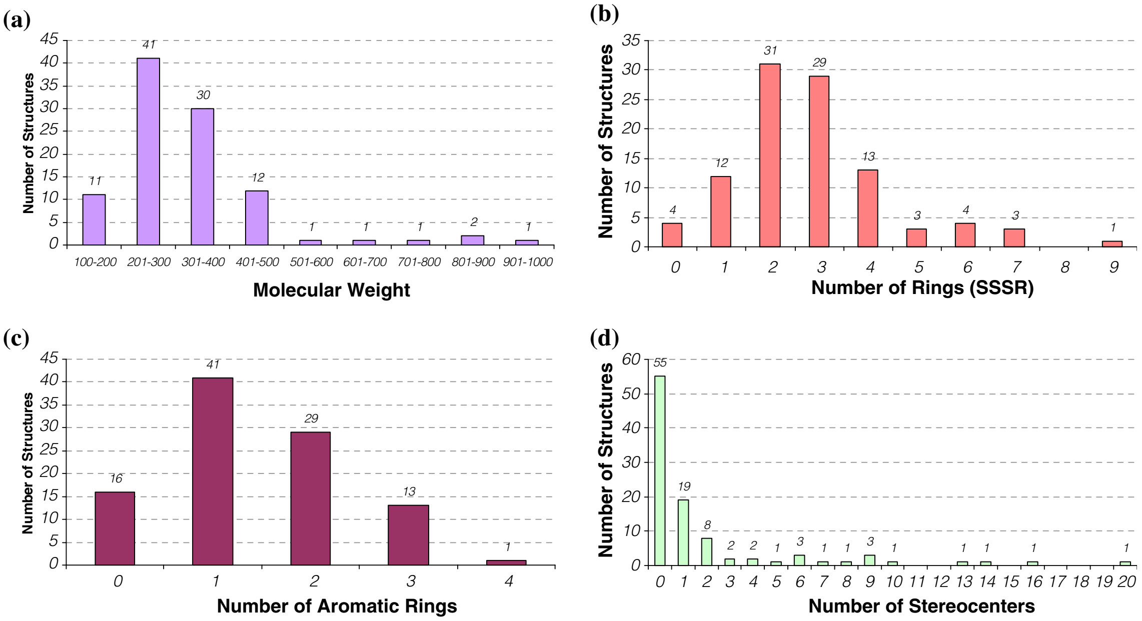


Fig. 18 The distribution of (**a**) molecular weights (**b**) number of rings (**c**) number of aromatic rings and (**d**) number of tetrahedral stereo centers of the dataset used to evaluate synthetic accessibility

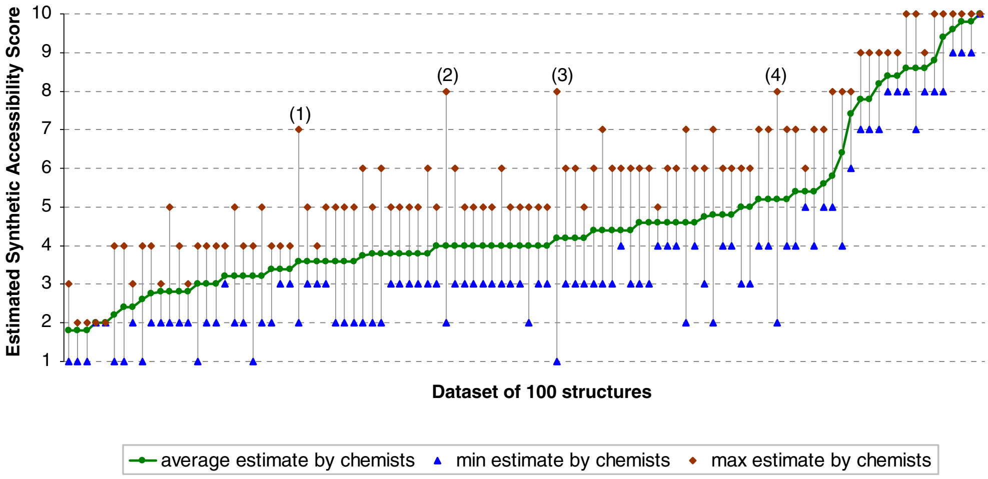


Fig. 19 The minimum and maximum of the synthetic accessibility estimates of five medicinal chemists. Structures are sorted by their average estimated values that are indicated by the line in the chart. Four structures with significant deviation are marked and presented in Fig. 20

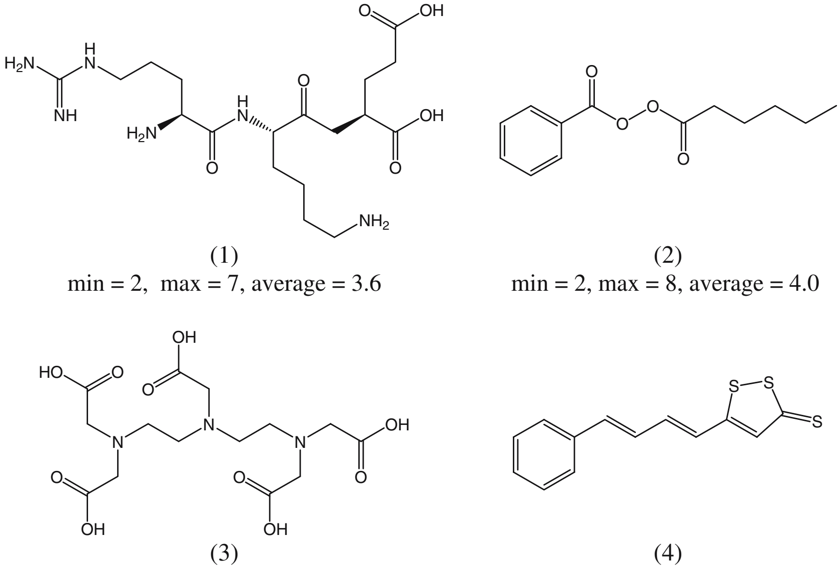


Fig. 20 Examples for significant deviations in the evaluated synthetic accessibility scores

(1)
min = 2, max = 7, average = 3.6

(2)
min = 2, max = 8, average = 4.0

(3)
min = 1, max = 8, average = 4.2

(4)
min = 2, max = 8, average = 5.2

$$SA = w_{mc}\, C_{mc} + w_{rc}\, C_{rc} + w_{sc}\, \log_2 (C_{sc} + 1) + w_{sm}\, C_{sm} + w_{rf}\, C_{rf} \qquad (5)$$

C_{mc} = molecule complexity
C_{rc} = ring complexity
C_{sc} = number of stereo centers
C_{sm} = starting material similarity
C_{rf} = retrosynthetic reaction fitness
w_i = corresponding weights

The logarithmic scaling of the stereo-center complexity component of the function was introduced in order to reduce its large range. As Fig. 18 shows, the number of stereo-center in the dataset ranges between 0 and 20. This scaling also signifies that the discrimination between compounds with small number of stereo-centers is more important than distinguishing between compounds with large number of stereo centers.

Having determined the weights of the components, the synthetic accessibility scores are recalculated for the dataset. The contributions of individual components of the scoring function for the 100 compounds are displayed in Fig. 21. The more prominent is the contribution of a component, the larger is the corresponding colored region in Fig. 21. The correlation coefficients of the individual components of the scoring function are listed in Table 4.

The results reveal that the retrosynthetic reaction fitness has the worst correlation with the chemist estimations; therefore it has moderate impact on the

Table 3 Correlation coefficients of synthetic accessibility estimations obtained from five medicinal chemists for the dataset of 100 structures

	Chemist 1	Chemist 2	Chemist 3	Chemist 4	Chemist 5
Chemist 1	–	0.75	0.77	0.84	0.74
Chemist 2		–	0.78	0.73	0.74
Chemist 3			–	0.82	0.75
Chemist 4				–	0.81
Chemist 5					–

Table 4 Correlation coefficients between individual components of the scoring function and the average value of the chemists' estimates

Components	Correlation
Molecular complexity	0.78
Ring complexity	0.32
Stereo complexity	0.82
Starting material similarity	0.67
Retrosynthetic reaction fitness	0.24

overall synthetic accessibility score. One reason behind this low contribution maybe that simply counting the number of strategic points, where the structure can be decomposed is not sufficient enough to express the quantity and complexity of the possible synthetic routes to target structures. This problem can only be addressed adequately by comprehensive retrosynthetic analysis that takes into account further important factors such as yield of synthetic reactions and total number of synthetic steps from target structure to starting materials.

The correlation coefficients between computational scores and chemist estimations are detailed in Table 5 (see also Fig. 22(b)). These correlations are comparable with correlations between chemists' estimates (Table 3), which confirms the reliability of our method. The computational predictions of synthetic accessibility versus average chemists' scores are shown in Fig. 22(a).

Figure 23 shows two structures of which the estimated synthetic accessibility significantly differs from the average estimate of the chemists. Structure (a) has a higher estimated synthetic accessibility than the score obtained from chemists' estimates. This is due to the fact that the stereo-, ring- and molecular complexity of the structure are quite high. Furthermore, there is a relatively low number of compounds in the utilized combined starting material database that contain this complex steroid ring system. Therefore the starting material similarity score cannot significantly moderate the high scores derived from the other components.

In contrast, structure (b) is quite difficult to synthesize according to the chemists. The reason for its smaller estimated synthetic accessibility lies in its relatively small number of stereo centers, but the specific arrangement of stereo centers will be quite difficult to synthesize. As visualized in Fig. 21, the stereo complexity has a significantly high contribution to the overall synthetic accessibility.

The speed of the calculation of synthetic accessibility is a crucial factor, since thousands of structures have to be analyzed in a short period of time. The evaluation of 100 structures of our dataset took 25 s on a 2.8GHz Linux PC, this elapsed time includes the loading of both the RCSS and the similarity databases into the memory and the calculation of synthetic accessibility scores. This means that on average ~200–

Fig. 21 Contribution of individual complexity components

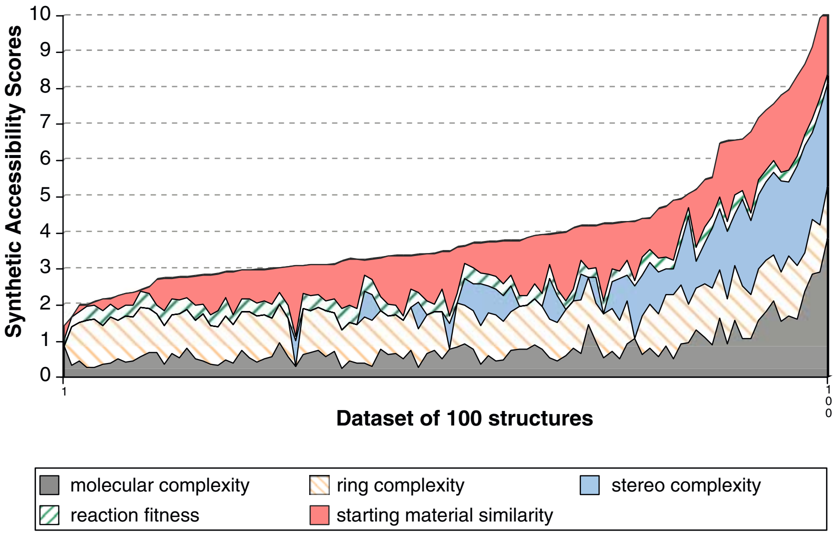
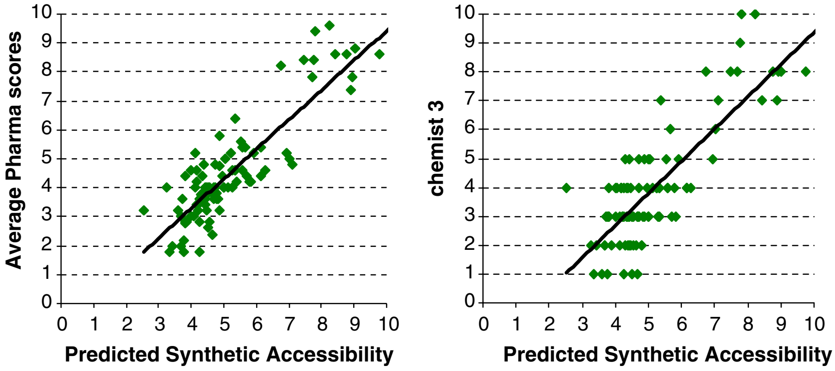


Table 5 Correlation coefficients between computational synthetic accessibility estimation and medicinal chemists' estimates

	Chemist 1	Chemist 2	Chemist 3	Chemist 4	Chemist	Average[a]
Computational estimations	0.810	0.791	0.860	0.840	0.789	0.893

[a] The weights of the synthetic accessibility function (Eq. 5) are determined to maximize the correlation coefficient between the computational estimation and the average scores provided by chemists



(a) correlation = 0.893

(b) correlation = 0.860

Fig. 22 Computational prediction of synthetic accessibility (**a**) versus average scores (**b**) versus one of the chemist estimations. Black lines indicate the linear trendlines

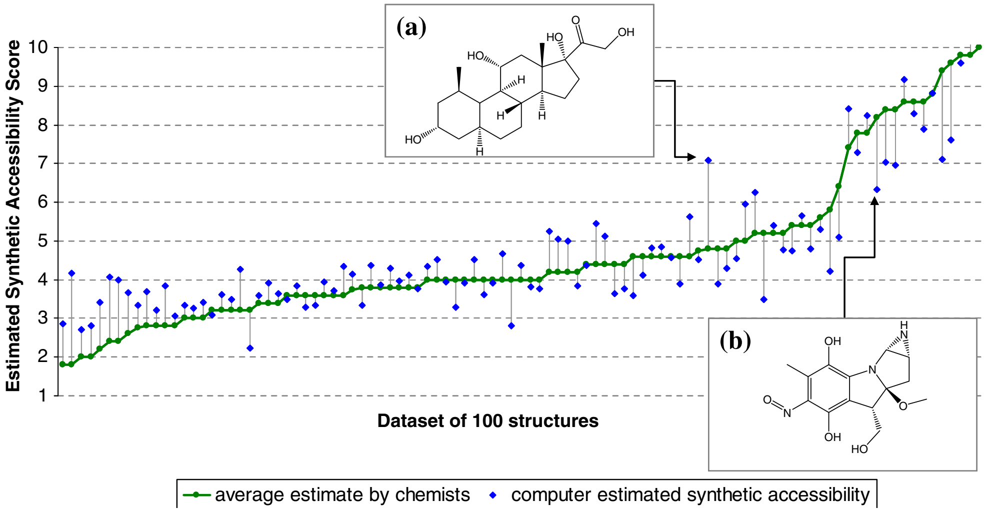


Fig. 23 Examples for significant deviations between computer-estimated synthetic accessibility and average chemists estimate

300 structures (depending on size) can be analyzed per minute.

If rapid estimation is essential, the calculation of retrosynthetic reaction fitness can be eliminated, since its contribution to the overall synthetic accessibility score is the lowest, however it is the most time-consuming operation. Table 6 details the calculation time share of individual components of synthetic accessibility scoring function.

The synthetic accessibility estimation can be used in a variety of ways. It can be employed to retrieve the *n* most synthetically accessible compounds from a dataset or it can be combined with a cut-off value to screen out compounds with high predicted synthetic accessibility. In this latter case, the choice of the actual cut-off value highly depends on the application. In case of evaluating structures, which are generated by *de novo* design programs, four might be a reasonable cut-off value according to computational chemists of our pharmaceutical partners. Combining the estimation of synthetic accessibility and binding affinity in the *de novo* design process could guarantee the high quality of the suggested structures.

The program also provides facilities to employ user-defined starting material and reaction databases in the

Table 6 Share of calculation time of individual components of the scoring function

Components	Correlation
Molecular complexity	1.2%
Ring complexity	Negligible
Stereo complexity	Negligible
Starting material similarity	41.5%
Retrosynthetic reaction fitness	57.3%

estimation process, thereby adding more flexibility to the system.

Implementation notes

The synthetic accessibility scoring function is implemented in C++ based on the MOSES software library [17] that is designed to provide a wide range of functions to handle chemical structures and synthesis reactions.

Summary and conclusion

The synthetic accessibility estimation, presented here, provides a powerful method to rapidly prioritize thousands of structures generated by *de novo* design systems. The main advantage of this method is that it integrates various features that are contemplated when chemists manually evaluate the synthetic accessibility of a set of compounds. For example, identifying the similarity of target structures to available starting materials and assessing strategic retrosynthetic fragmentation points are all part of the mental process of synthesis planning.

Furthermore, the weights of the components of the synthetic accessibility function are calibrated using estimations obtained from experienced medicinal chemists.

Because of the fast calculation process, this estimation can effectively be incorporated into virtual screening tools in order to rank thousands of candidate molecules by their synthetic complexity.

The synthetic accessibility function can easily be adapted to be applicable at the fragment level in order to suggest high quality building blocks for either *de novo* design systems or combinatorial library design. For example, building blocks retrieved from a fragmentation process, such as in the RECAP cleaving process [18], can be subjected to our synthetic accessibility calculation in order to select motifs with high synthetic availability. Structures that are constructed in the *de novo* process from a library of synthetically available building block are likely to be synthetically available themselves.

Acknowledgments The authors would like to thank the medicinal chemists of our pharmaceutical partner companies (4SC, Altana, Lilly) for evaluating the synthetic accessibility of our dataset. We are also grateful to Elsevier MDL for making available the Theilheimer reaction database for us. We are also grateful to our partners at BioSolveIT and the Center for Bioinformatics (ZBH) of the University of Hamburg for their contributions. Finally, we would like to acknowledge the work and support of the MOSES development team, especially the contribution of Achim Herwig and Oliver Sacher. This work has been financed by the German Federal Ministry of Education and Research (BMBF) as part of the NovoBench project (Project number: 0313324A).

References

1. Baber JC, Feher M (2004) Mini-Rev Med Chem 4:681
2. Barone R, Chanon M (2001) J Chem Inf Comput Sci 41:269
3. Bertz SH (1981) J Am Chem Sci 103:3599
4. Hendrickson JB, Huang P, Toczko AG (1987) J Chem Inf Comput Sci 27:63
5. Boda K, Johnson AP (2006) J Med Chem 49:5869
6. Pförtner M, Sitzmann M, Gasteiger J (eds) (2003) Computer-assisted synthesis design by WODCA. Handbook of chemoinformatics. Wiley-VCH, Weinheim, pp 1447–1507
7. Molecular Networks GmbH, WODCA Synthesis design by retro-synthetic analysis, http://www.molecular-networks.com/software/wodca/
8. Johnson AP, Marshall C, Judson PN (1992) J Chem Inf Comput Sci 32:411
9. Myatt G (1994) Computer aided estimation of synthetic accessibility. PhD thesis, School of Chemistry, University of Leeds, Leeds
10. Shannon C, Weaver W (1949) Mathematical theory of communications. University of Illinois Press, Urbana, IL
11. Gasteiger J, Jochum C (1979) J Chem Inf Comput Sci 19:43
12. Corey E, Cheng X-M (1989) The logic of chemical synthesis. John Wiley & Sons Ltd., New York
13. Warren S (1982) Organic synthesis: the disconnection approach. John Wiley & Sons Ltd., New York
14. Willett P (1998) J Chem Inf Comput Sci 38:983
15. Gasteiger J, Ihlenfeldt WD, Fick R, Rose JR (1992) J Chem Inf Comput Sci 32:700
16. Ihlenfeldt WD, Gasteiger J (1994) J Comp Chem 15:793
17. Molecular Networks GmbH, MOSES C++ software library, http://www.molecular-networks.com/software/moses/
18. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) J Chem Inf Comput Sci 38:511