Predicting Synthetic Accessibility: Application in Drug Discovery and Development

J.C. Baber[*] and M. Feher[†]

SignalGene Inc., 335 Laird Road – Unit 2, Guelph, Ontario, N1G 4P7 Canada

Abstract: The estimation and use of synthetic accessibility in the drug discovery process is discussed. A distinction is drawn between synthetic feasibility and accessibility and the practical uses of an accessibility score are examined. Various techniques used in the estimation of accessibility are described and their utility and potential accuracy compared.

Keywords: Synthetic accessibility, synthetic feasibility, drug discovery, drug development, screening, prioritization, retrosynthetic analysis.

INTRODUCTION

Although often mentioned in works on drug discovery, and *de novo* design in particular, little has been published on the assessment of synthetic feasibility and accessibility. This review examines the estimation and uses of synthetic accessibility, with emphasis on the drug discovery process. Following a brief background section, the concept of synthetic accessibility is discussed in relation to the more widely known synthetic feasibility and a number of uses for a synthetic accessibility score are proposed. Since in current practice medicinal chemists often assess synthetic accessibility manually and informally, their performance is then examined. Finally, methods of automatically estimating synthetic accessibility are examined and the advantages and disadvantages of each approach discussed. We conclude with a summary of the current state of the art and make some suggestions regarding the possible future direction of development in the field.

BACKGROUND

There are numerous techniques available to identify large numbers of compounds that may be active for almost any given active site or pharmacophore model. Those compounds that are identified through the screening of commercial or internal databases are generally simple to obtain. In contrast, those generated by a *de novo* design process or other methods that introduce novelty, such as the modification of existing leads, must be synthesized before they can be tested. Since many of these methods are capable of producing large sets of compounds it is often impractical to synthesize all possible structures that fit the model. Therefore, the size of the set is usually reduced by estimating properties, such as logP, and using these values to screen out unfavorable compounds. Another screening step that may be employed is the identification of structural features predicted to introduce toxicity, instability or metabolic liabilities. The identification of key structural features is also commonly used to exclude compounds that this process generally does a good job at removing the most are likely to be particularly difficult to synthesize. Although synthetically difficult structures, it does not guarantee that every compound that passes the screen will be easy to synthesize. Due to the large number of structures under consideration, manual examination of the entire set is often impractical and it has been recognized that the assessment of ease of synthesis is a significant bottleneck when designing new compounds [1]. While this problem arises with any structure not obtained from a database of easily available compounds, it is particularly applicable to many of the current computational approaches to *de novo* ligand design. This has been noted by a number of the producers of *de novo* design systems who have modified their algorithms in order to increase the ease of synthesis of generated structures.

The generation of difficult to synthesize structures is a particular problem for programs that use an atom-by-atom sequential buildup approach such as Legend [2,3], Genstar [4], GrowMol [5] and MA [6]. In this approach structures are generated one atom at a time from seed atoms or fragments. Structures may be grown either as carbon skeletons or substituted compounds and a variety of rules are employed to ensure that generated structures are chemically reasonable. Other systems using a sequential buildup approach such as GROW [7], GroupBuild [8] and SPROUT [9,10,11], grow structures by adding fragments rather than individual atoms. As with atom-based methods, the added fragments may be either carbon skeletons or fully substituted depending on the program. In the latest releases of fragment-based systems such as SynSPROUT and Skelgen [12] a limited set of connection rules are employed to join fragments. These rules are derived from known synthetic chemistry in an attempt to increase the ease of synthesis of generated structures. Some *de novo* design programs, particularly those that build carbon skeletons, also employ a post processing stage where generated structures are substituted with atom and bond types based on a list of preferred fragments [11,13]. The substituted fragments are generally selected on the grounds of synthetic ease, yet must be small enough to ensure that a diverse set of combinations is possible. Consequently preferred fragments are often common functional groups with known syntheses, such as amides and ethers.

[*] Address correspondence to this author at the Neurocrine Biosciences Inc., 10555 Science Center Drive, San Diego, California 92121, USA; E-mail: jcbaber@quoth.co.uk

[†] Current Address: Neurocrine Biosciences inc., 10555 Science Center Drive, San Diego, California 92121, USA

1389-5575/04 $45.00+.00

Ease of synthesis is generally less of a problem for structures generated by *de novo* design techniques based on fragment connection or docking. In these techniques programs such as GRID [14], HSITE [15,16] or MCSS [17] are used to position fragments within a target site in places where they are likely to have a strong interaction with the receptor. These fragments are then joined using bridging groups to form complete compounds. Bridging groups may be found either through database searching, as with programs such as LUDI [18,19] and CAVEAT [20,21], or by combining small spacer fragments as used by NEWLEAD [22]. In the case of database searching programs there is generally little problem obtaining or synthesizing the generated structures [23,24]. However, when small spacer fragments are used this is not necessarily the case. In order to address this problem a process called combinatorial docking has been developed [25]. In programs capable of combinatorial docking such as LUDI [18,19], TOPAS [26] and MOLOC (Lego 3) [27] structures are grown in the binding site by combining building blocks according to rules derived from well-behaved chemical reactions. Since chemistry is directly taken into account when building the structures, compounds suggested by such programs are likely to be relatively easy to synthesize – often by a combinatorial or parallel approach. However, such programs generally employ a highly restrictive set of chemical connection rules, which results in the generation of a narrow series compounds [12] rather than the diverse, but often difficult to synthesize, compounds generated by sequential buildup procedures.

The terms synthetic accessibility and synthetic feasibility have both been used previously to describe ease of synthesis. In the next section we will clarify the meaning of these terms and describe ways in which an estimate of ease of synthesis can be used in the drug discovery process.

Feasibility, Accessibility and their Uses

In this review we would like to draw a distinction between synthetic feasibility and synthetic accessibility. Synthetic feasibility will be used to describe whether or not it is possible to synthesize a compound given a specified set of conditions. In contrast, synthetic accessibility is defined as the ease of synthesis under a specified set of conditions.

The limitation of specifying conditions is necessary since most chemically reasonable compounds are synthetically feasible given infinite resources, even though synthesis of very complex compounds such as natural products [28] may take a number of years. Thus the inclusion of conditions in the definition of synthetic feasibility effectively means that a synthetically feasible compound is one for which synthesis is practical rather than just theoretically possible. An example of reasonable conditions for synthetic feasibility would be defining a synthetically feasible compound as one possible to synthesize in less than one month given the resources available in a particular lab. Although synthetic accessibility measures ease of synthesis rather than whether or not synthesis is possible the specification of conditions is still necessary. This is the case since compounds that may be considered relatively easy to synthesize in one situation may be assessed as very difficult to synthesize in another. An example of this occurs in the scale-up of syntheses since processes that are simple in the laboratory may not be practical on an industrial scale.

A major use for estimated feasibility is the screening of compounds. This may theoretically be carried out on any number of structures ranging from individual compounds, for example to ensure that particular modifications are practical, to large sets, as part of an enrichment process. This screening is particularly useful in the case of large sets of structures from *de novo* design programs. *De novo* programs are often based on graph theory, with atoms treated as nodes and bonds as vertexes. Until recently such programs have had little built-in knowledge of synthetic chemistry. Consequently structures generated by *de novo* design may often be synthetically infeasible. Synthetic feasibility is a useful tool in reducing the size of a set of structures, whether from a *de novo* design program or another source, since structures assessed as synthetically infeasible can simply be removed from the set. In the case of small sets of compounds an estimation of synthetic feasibility may be carried out manually by experienced medicinal chemists. However this is clearly not practical for large sets of structures. For large sets the screening process is often carried out computationally by searching each structure for chemical features that are known to be difficult to synthesize such as stereo centers or spiro unions. Any structure containing more than a specified number of occurrences of a given feature, for example more than two stereo centers, can then be rejected as synthetically infeasible. It is possible to modify the number of occurrences required for a structure to be flagged as infeasible, or even to require combinations of features to be present, in order to customize the assessment for a given situation.

An alternative, and potentially more accurate, estimation of feasibility could be made using a synthesis planning program such as LHASA [29], WODCA [30] or SECS [31]. Such programs have been studied extensively [30,32-34] and it is not within the scope of this review to provide detailed information on them, although a brief overview is given later. These systems attempt to discover a synthetic route for a target compound, generally by employing large knowledge bases of known chemistry and available starting materials. Consequently, a synthetically feasible compound could be defined as one for which a synthetic route covering the entire structure was identified by a given synthesis planning program. However, regardless of how it is calculated, the binary nature of a synthetic feasibility assessment significantly limits its use as part of a more complex screening process than that described above.

Unlike feasibility, synthetic accessibility is specified as a score, which may either be unbounded or within a given range such as a percentage. In its simplest form a synthetic accessibility score can be used to generate an estimate of feasibility by defining a score below which compounds are considered infeasible. However, as well as allowing screening in the same manner as feasibility, it is possible to combine a synthetic accessibility score with other properties in a more complex screening system. A more complex screening system of this type would be particularly applicable to large sets of compounds with a number of computed properties such as logP, solubility or estimated binding energy. The availability of a synthetic accessibility score also allows a variety of different selection procedures to be used. For example if too many structures pass a screen which requires a minimum score of 70 for synthetic accessibility it would be simple to increase the required level to 80. Alternatively it would be possible to retain the *n* most synthetically accessible compounds – with *n* either defined as an absolute number or as a percentage of the set – something that is not possible using an estimate of synthetic feasibility.

Any number of factors, such as time to synthesize, cost of reagents etc, may be considered in the estimation of synthetic accessibility. The contribution from each factor, and even the method used to calculate accessibility, may be modified depending on the circumstances under consideration. By using different levels of contribution from each factor, lists of available starting materials and allowed chemical reactions in the calculation, it is possible to generate multiple accessibility estimates for a given compound. An example of this is the analysis of collections of compounds with the aim of designing combinatorial libraries around the types of structures represented in the target set. To do this it is necessary to consider only chemical reactions and starting materials amenable to parallel synthesis in the analysis. Depending on the method of estimation used, it may also be possible to increase the accessibility score of those target compounds with identified chemistry, starting materials and intermediates similar to other compounds in the set. If these factors are included in the estimation then the structures likely to be most useful as the basis of a combinatorial library will be assigned the highest accessibility scores.

The availability of a synthetic accessibility score is also useful for comparing and ranking structures. This ranking of structures allows the synthesis of compounds to be prioritized, with those compounds scoring highly for ease of synthesis attempted first. Since synthesis is often the rate-determining step in early drug discovery, concentrating resources on those structures determined to be easier to synthesize will result in more compounds being tested. In turn this gives both a greater likelihood of a useful lead being identified and more information to pass back to the next round of designs. As with screening it is possible to prioritize synthesis either using a synthetic accessibility score alone or combined with other computed properties in order to target resources towards those compounds most likely to be useful leads.

Finally, depending on the method used to estimate synthetic accessibility there is often additional information produced by the analysis, which may be especially useful when dealing with individual or small sets of compounds. Many of the methods of estimating synthetic accessibility described below identify parts of the structure that are particularly difficult to synthesize. In some cases it may be possible to simplify the structure to reduce this difficulty – for example removing an unessential methyl group to reduce the number of stereocentres present. Some techniques also provide useful information about known chemistry or available starting materials that may be applicable to the synthesis of the analyzed compound. These techniques may help to suggest possible analogues by identifying available reagents that are similar to substructures in the target compound. This additional chemical information can also be useful for grouping together compounds with common intermediates or starting materials – which may allow a number of similar compounds to be synthesized simultaneously or even small libraries of compounds to be produced with little additional effort. More detail on these additional features of synthetic accessibility assessment is given in the appropriate parts of the Computational Assessment of Synthetic Accessibility section, below.

MANUAL ESTIMATION OF SYNTHETIC ACCESSIBILITY

In practice it is generally medicinal chemists that make the final decision about which compounds will be synthesized. This decision is usually supported by computational chemists, often by providing a selection of compounds that have passed a given set of screens. Since all of the compounds being considered will have good predicted binding and other properties and will usually have been screened to exclude compounds that are obviously difficult to synthesize, the medicinal chemists will generally be most interested in the ease of synthesis of each the alternatives.

The ability of experienced medicinal chemists to estimate synthetic accessibility for a variety of drug-like compounds has previously been examined [35]. In the study, eight practicing medicinal chemists from a major pharmaceutical company were asked to score a set of 100 drug-like compounds on a scale of 0 to 10 - with a score of 10 indicating that the compound is commercially available or would be straightforward to synthesize and a score of 0 being assigned to almost impossible to synthesize compounds. The chemists were asked to carry out this exercise quickly, spending no more than 3 minutes on each compound, and without using any books, databases or other form of reference. These limits were intended to be very restrictive and are in that they would not allow any detailed retrosynthetic analysis to be performed. On the other hand 3 minutes is actually a long time to spend examining each compound when limitations in human concentration and number of hours worked per day are considered.

The synthetic accessibility scores produced by the chemists were initially examined using non-parametric statistics in order to test the theory that there was in fact a ‘real’ correct value for synthetic accessibility. Since the ranking of compounds from each chemists showed a high degree of correlation the study drew the conclusion that a ‘real’ synthetic accessibility score probably existed and it was valid to assume that the estimates supplied by individual chemists were normally distributed around this value. This assumption allowed the accuracy of individual chemists to be examined. Overall it was found that the mean absolute error in chemists’ estimates was a little over 10% (with synthetic accessibility estimated on a percentage scale). However, for some compounds there was a variation of up to 70% in the estimates produced by individual chemists. A number of reasons appear to exist for these deviations but in main they appear to be due to differences in experience and knowledge.

Experience tended to show most when a chemist had previously synthesized a compound similar to the one being

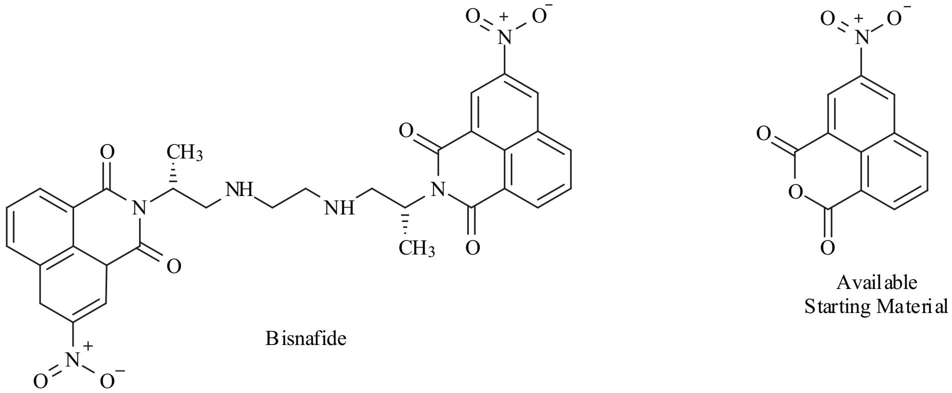


Fig. (1). Large available starting material for complex compound.

scored. In these cases the chemist usually knows the chemistry that would be used in the synthesis and, more importantly, the general types of starting materials that are likely to be available. This sort of experience is difficult to replace computationally although similarity and substructure-based searching techniques do go some of the way to replicating the chemists' thought process.

Large deviations in estimates were also seen when one or more of the chemists were missing a specific piece of knowledge. This was particularly the case when complex starting materials were available yet not known to all of the chemists in the study. An example of this is shown in Fig. (**1**). In this case the estimates of synthetic accessibility obtained from chemists were in two clusters, 7 and 8 from those chemists that realized that the Anhydride was available and 4 and 5 from those that did not. Fortunately this problem is relatively simple to rectify through the use of starting material catalogues and databases, for example CHEMCATS [36]. However ensuring that all of the available information is used in an analysis still adds a substantial amount to the effort required to obtain an accurate estimate of synthetic accessibility, even when modern tools are employed.

One important feature of the manual estimation of synthetic accessibility is that the process is essentially the first step in planning the synthesis. As such it often results in the identification of possible starting materials and chemistry that could be used in a real synthesis. For the same reasons an experienced medicinal chemist performing a manual assessment may well be able to suggest changes to a structure that would increase its synthetic accessibility. As with some computational methods of estimation, this may even extend to the identification of a set of possible analogues, or small library of similar compounds, that could be synthesized with little additional effort.

COMPUTATIONAL ESTIMATION OF ACCESSIBILITY

Although little work has been carried out directly on the computational estimation of synthetic accessibility, a number of approaches are possible. Some techniques, such as methods based on the identification of starting materials, play to the strengths of computers by accessing large databases of information, whereas others attempt to duplicate the process used by medicinal chemists. The different approaches are often inter-related, with many methods resorting to a complexity-based technique to provide data for those parts of a target structure that would otherwise be ignored by the analysis. For the purposes of this review these techniques have been divided into four main classes as described below.

Complexity-Based Estimation of Synthetic Accessibility

Complexity-based analysis is probably the most widely used technique when it comes to the automatic assessment of synthetic feasibility. In the screening process described above, the location of difficult to synthesize features effectively identifies synthetic complexity present in a target structure. It is this identified complexity that is then used to assess feasibility. As an extension to this basic screening it is possible to assign scores, and/or categories, to each complex feature. These scores may then be combined based on the number of occurrences of each substructure identified, in order to obtain an overall estimate of accessibility. A wide range of methods may be used to generate the final estimate and simple example of this process that considers both scores and assigned categories is shown in Fig. (**2**) below.

In this example the number of ring and chain stereocentres are counted separately and combined to give a score for stereochemical complexity. A similar procedure occurs with the number of spiro unions and ring fusions that are combined to give a ring complexity score. These two factors can then be used to calculate a final estimate for synthetic accessibility normalized to whatever range is required. The actual calculation of scores can be achieved in a wide range of ways ranging from simple summing to complex Bayesian reasoning.

Examples of a complexity driven synthetic accessibility score can be found in some commercial *de novo* design packages. In LeapFrog [37] synthetic difficulty scores, inversely proportional to synthetic accessibility, in kcal/mole can be calculated for any structure. At its heart the technique used by LeapFrog to estimate synthetic difficulty is relatively simple, with each atom, heteroatom, ring, rotatable bond and stereocentre assigned weights that are summed to obtain a score for the structure as a whole. However, the system does include the facility to specify a

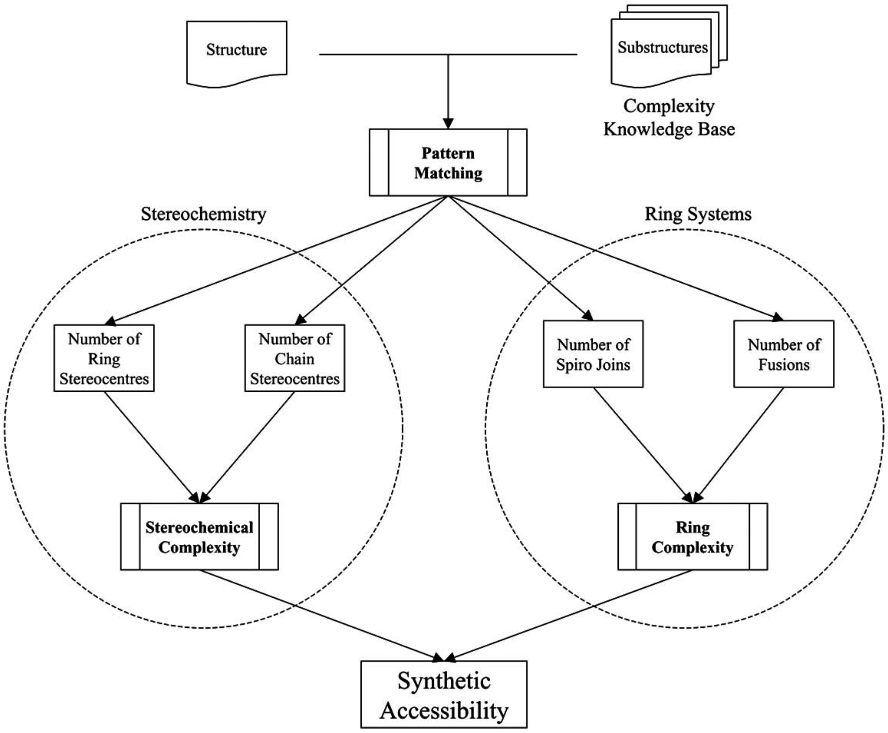


Fig. (2). Simple scoring scheme for compexity.

file containing other structural features, using the Sybyl Line Notation (SLN [38]), along with associated energies. These associated energies may either be positive for features that add to synthetic difficulty or negative for those that increase synthetic accessibility. The energy associated with each identified feature is then added to the score obtained from the initial calculation to provide a final estimate of synthetic difficulty. ChemX [39] uses a process called Novel Molecule Scoring to assess synthetic accessibility. In this system molecules are initially assigned a score of 100, which is then reduced if unfavorable bonds, many rotatable bonds or uncommon frameworks are identified. Each of these factors is calculated separately and then multiplied together to give the final score. A minimum acceptable score may be set to allow the automatic rejection of low scoring molecules.

Complexity driven estimation is fairly simple to implement and is computationally inexpensive, relative to other techniques, so is capable of producing estimates for a large set of compounds in a short space of time. However, this technique has the major disadvantage that it does not consider the starting materials that may be available. This can cause wildly inaccurate accessibility estimates if a starting material that covers a highly complex part of the compound is available, as in the case of the example shown in Fig. (**1**).

Chemistry-Based Estimation of Synthetic Accessibility

In some ways the chemistry-based approach to the estimation of synthetic accessibility is very similar to the complexity-based approach and is most useful when combined with an assessment of complexity in order to improve the accuracy of estimates. Rather than attempting to identify difficult to synthesize features, the chemistry-based approach uses a similar knowledge-based technique to identify substructures for which a synthetic route is known. Identified features with known synthetic routes may then be excluded from the parts of the compound considered by complexity analysis and therefore increase the final estimate of synthetic accessibility. This is similar to the approach used by LeapFrog [37], which allows features with a negative contribution to complexity to be specified thus effectively excluding those substructures from the complexity analysis.

As with complexity-based techniques this approach tends to be relatively computationally inexpensive and is thus capable of scoring large numbers of compounds in a short space of time. However synthetic chemistry is both extensive and constantly expanding and significant effort is required to keep a knowledge base of features with known synthetic routes up to date, more so than with a knowledge base of complex features. Large reference and reaction databases, such as CASREACT [40], are available and are regularly used by medicinal chemists when planning synthetic routes. Methods of automatically extracting information from these database have been developed [41-44] and could be used both to increase the coverage of knowledge bases and reduce their maintenance requirements. As with simple complexity analysis, a chemistry-based approach does not directly consider available starting materials – although often if a synthetic route is known for a particular class of structures then that is a good indication that a number of suitable starting materials will be commercially available.

Additionally, the chemistry-based approach is similar to that used by experienced medicinal chemists who will often identify parts of a compound for which a synthesis is known before examining the remaining areas to see if anything particularly difficult to synthesize is present. The synthetically accessible substructures identified by the chemistry-based approach may therefore be used to provide a starting point for medicinal chemists if the decision is made to progress with a given compound and a full plan synthesis is required. Knowledge of possible reactions used in the synthesis of a compound is also very useful when the synthesis of libraries of analogues is considered since substructures key to the reactions are generally known and can be used as the basis of a starting material search.

Starting Material-Based Estimation of Synthetic Accessibility

As has been shown above, the availability of suitable starting materials is often a very important factor in determining the synthetic accessibility of a given compound. It is therefore possible to get a reasonable estimate of synthetic accessibility by assessing how much of a target compound is covered by available starting materials. In its simplest form coverage may be defined as the number of common atoms in the identified starting material and target structure. However, it is of greater use to consider a more refined measure of coverage when assessing the quality of starting materials. An example of such a measure would be a score based on the number of atoms in the starting material that correspond directly to atoms in the target structure with additions for any complex features in the target structure, such as stereocentres, that are completely covered by starting material atoms

With all starting material-based techniques the choice of which databases to use as sources of starting materials is very important. Available reagents are often relatively reactive which is generally not a good feature in drugs so the simple matching of substructures without any consideration of the chemistry involved can produce misleading scores. It is possible to use sets of products, for example drugs, as a source database in starting material-based analysis. However, this introduces a bias where target structures similar to those present in the database are given the highest scores. This may actually be a welcome bias in some cases – for example if a set of drugs is used as the source database then high scoring structures are likely to be somewhat *drug-like* in nature. Although the use of drugs as a source database may be useful, it will generally result in lower scores for the more novel compounds and moves the entire process away from the estimation of synthetic accessibility.

There are a variety of different substructure searching techniques available [45,46] but they are all relatively computationally expensive and must be carried out many times in order to identify all possible starting materials. Consequently starting material-based techniques are generally slow when compared to complexity or chemistry-based methods of estimating synthetic accessibility. As with chemistry-based techniques, methods based on the identification of starting materials are most useful when applied along with complexity analysis to assess the synthetic accessibility of those parts of the structure not covered by available starting materials. This analysis results in a residual complexity score that may then be combined with a contribution based on the quality of the identified starting materials to obtain a final estimate of synthetic accessibility. As stated above, care must be taken when selecting source databases and, additionally, some effort is required to keep these databases up-to-date. However, the estimates of accessibility obtained using starting material-based methods are generally more accurate than those from purely complexity-based techniques – particularly when large starting materials are available.

Starting material-based methods also have the advantage that reagents identified using these techniques can often provide a useful starting point for medicinal chemists when planning a full synthesis. However, structural and functional changes caused by the various steps in the synthesis mean that identification of the precise starting material that would be used in a synthesis is rare, with the usual result being the identification of general classes of potential reagents. For this reason, suggested starting materials are also useful when designing combinatorial libraries or considering analogues to be synthesized. Recently, the number of commercially available compounds has increased significantly. For example the CHEMCATS Database [36] grew in size by over 40% in the final 6 months of 2002 and now contains over 5.6 million records. With this current rapid expansion in the number of available reagents it is becoming increasingly difficult for synthetic chemists to keep up-to-date with exactly which starting materials are available. Therefore, information on the general types of available reagents that may be useful in a synthesis is becoming more and more helpful.

A variety of different approaches are available for the identification of possible starting materials. These can be split in to two main camps: methods based on the exact matching of substructures and those based on similarity – either of substructures or the target structure as a whole.

Exact Matching of Starting Materials

These techniques work on the basis of identifying parts of the target structure that are also present in available starting materials. The size of sections matched may range from as large as the whole structure (giving perfect synthetic accessibility if it is found to be available in a starting material containing no additional atoms) to something as small as a simple ring system. Since substructures present in the target structure are matched with substructures present in starting materials it is often useful to employ extended patterns to allow the search to be more accurate. Extended patterns include some additional information for each atom, such as the number of connections to the rest of the structure or number of adjacent heteroatoms. Extended patterns allow a search to consider the relationship between a substructure and the structure as a whole as well as just the substructure's own properties. Thus, the use of extended patterns ensures that starting materials are only identified if they have the key substructure in a similar environment to the target structure. This is particularly helpful in the case of relatively small substructures and ring systems where substitution patterns can be identified. It is also possible to perform the search both with and without this additional information to give basic and extended matches for any given substructure.

Once matches have been found synthetic accessibility scores may be generated in a number of manners. The most simple of these is the calculation of an accessibility score based on the proportion of atoms in the target structure covered by identified starting materials. This score can be modified depending on whether a match for a basic or extended pattern was found for a given substructure, the number of matches of each type found and the number of additional atoms in the identified starting materials that do not match atoms in the target structure. However, this simple score primarily considers the coverage of identified starting materials and ignores whether they cover simple or complex parts of the target structure. This technique is therefore most useful when combined with an assessment of residual complexity. Residual complexity scoring follows a similar procedure for assessing complexity as described above. However, with residual complexity the contributions from individual difficult to synthesize features present in the target structure are weighted according to the number and quality of identified starting materials that covering that part of the structure. Thus the availability of starting materials covering identified complex features is taken into account. Following the identification of both available reagents and residual complexity, a synthetic accessibility score can be obtained by combining the results using a wide range of methods, as described above.

Compared to the methods described earlier in this review the estimation of synthetic accessibility by exact matching of starting materials is computationally expensive. The actual time taken to generate an estimate is heavily dependant upon the number of substructures considered and the size of the database searched. Substructures may be identified either automatically, such as disconnecting individual bonds in turn, or through a more intelligent system, for example using a knowledge base to specify the general patterns to match. This method requires the source database of available reagents to be kept up to date and while most suppliers will now provide their catalogue electronically, the number of reagents available and variety of different sources makes this a complex task. As with the chemistry-based techniques this approach is similar to the method used by medicinal chemists when assessing synthetic accessibility and identified reagents can provide a useful starting point for synthesis planning.

Although ameliorated by the use of smaller substructures and extended patterns, the fact that slight modifications of the original pattern are not found is a major drawback of exact matching. Excluding the trivial case of the target structure being commercially available, the reactions involved in a synthesis will generally result in a change to the starting materials used. This means that exact matching is unlikely to correctly identify a starting material that will completely match a substructure present in the target compound and not have any additional atoms. In most cases matched starting materials will be larger than the corresponding target substructure used to identify them and sometimes, particularly in the case of small substructures, substantially so. Although this makes starting materials identified by exact matching less useful for synthesis planning, as long as care is taken selecting the substructure patterns and extended properties to be matched, the frequency of substructures identified by this technique is still a useful indication of synthetic accessibility.

Similarity-Based Matching of Starting Materials

Instead of exactly matching sections of the compound it is also possible to identify potential starting materials by searching for available reagents similar to either all or part of the target structure. This approach reduces some of the problems associated with exact matching but is somewhat more difficult to apply to substructures rather than the target compound as a whole and is significantly more computationally expensive.

The field of chemical similarity has been well studied [47-49] and it is not the aim of this review to duplicate that work. For the purpose of identifying possible starting materials it is necessary to find those compounds in a database containing substructures similar to specified substructures present in the target compound. Amongst other methods, it is possible to achieve this by generalizing the base substructure and searching for matches of this generalized structure. Two examples of this generalization process, the first designed to identify structurally similar substructures and the second chemically similar ones, are given below.

Structural generalization is useful for identifying starting materials with a similar connectivity to the target structure. This can be achieved by progressively generalizing the search by removing properties such as atom and bond types as shown in Fig. (**3**), below.

This process is particularly useful with ring systems since it allows common substituent positions to be identified. Structural generalization is also very appropriate when identifying starting materials for building combinatorial libraries or sets of analogues. This is due to the fact that this type of similarity searching is capable of identifying starting materials with a similar shape and, depending on the level of generalization, possibly electronic properties to the target structure.

Another useful approach to similarity-based matching uses chemical generalization to identify compounds that are chemically similar to the target structure or substructures thereof. This method is usually more appropriate for the assessment of synthetic accessibility than purely structural similarity-based techniques, since it retains more information about the chemical properties of the target structure. Chemical generalization enables equivalent, or

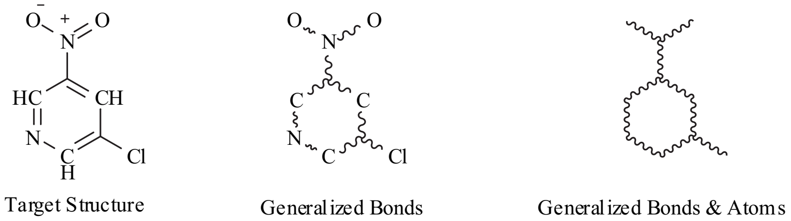
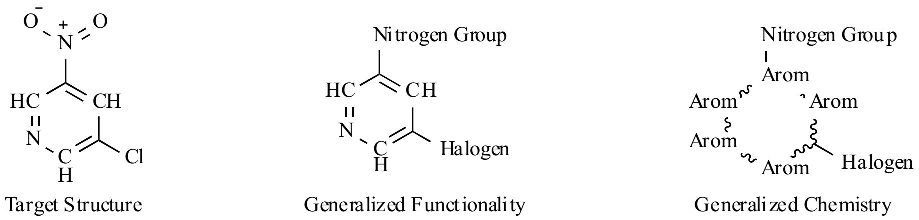


Fig. (3). Structural generalization.



Target Structure

Generalized Functionality

Generalized Chemistry

Fig. (4). Chemical generalization.

easily interconverted, functional groups to be specified, thus allowing the identification of possible analogues and, to some extent, taking the chemistry required to synthesize the target structure into account.

Each method of similarity searching has various advantages and disadvantages and the choice of which is most appropriate will depend on the exact reason that the assessment of synthetic accessibility is being carried out. For example if discrete compounds are to be synthesized then chemical similarity is often the most appropriate whereas if the goal is to design a combinatorial library of compounds then an assessment based on structural similarity may be more useful.

Similarity-based matching is generally more computationally expensive than exact matching, although this can be reduced significantly by various preprocessing techniques. It is also possible to use more than one type of similarity measure – or more than one level of generalization. In this case the number of matches of each type found could be combined, with the weighting given to each level dependant on the exact application, in order to determine a score for each substructure. The scores for all of the identified substructure can then be used either without any additional information or combined with a complexity score to determine a final value for synthetic accessibility in a similar manner to that described above.

Retrosynthesis-Based Estimation of Synthetic Accessibility

When experienced medicinal chemists assess synthetic accessibility the three techniques described above are generally all applied. Even when the assessment is purely mental the target structure is usually examined for easy to synthesize features – taking both known chemistry and available starting materials into account – and then the remainder of the compound examined for difficult to synthesize features, or residual complexity. The process of deconstructing a target structure in a search for possible reagents has been formalized for synthesis planning purposes into a technique called retrosynthetic analysis [50].

A number of systems that assist in the synthesis design process are available and generally fall under the umbrella of Computer Assisted Organic Synthesis (C.A.O.S.). Several reviews and other papers have been written on the subject [30,32,33,34] and it is not the purpose of this paper to repeat that work. However, since many of the techniques used by such systems are very closely related to the retrosynthesis-based method of estimating synthetic accessibility, a brief overview is given here.

C.A.O.S. systems cover areas ranging from reaction and substructure-based reference searching to synthesis planning programs that help chemists design a complete synthetic route. Reference and database searching programs, such as Chemical Abstracts [51], CrossFire Beilstein [52] and CHEMCATS [36], are useful when making a manual assessment of synthetic accessibility. However, the C.A.O.S. systems most closely linked to the retrosynthetic approach are synthesis planners, with examples being LHASA [29], WODCA [30] and SECS [31]. These programs take a target structure and attempt to identify a synthetic route using retrosynthetic analysis – the iterative disconnection and rearrangement of the target structure according to sets of transforms that represent the reversal of chemical reactions. At each stage in this disconnection process it is possible to check reagent databases to see if the generated precursor is available. Once available starting materials have been found for every part of the target structure a complete synthetic route has been identified.

In general these programs tend to be interactive – requiring expert intervention at each step in the process to select the retrosynthetic transform or reaction class to be applied – and are therefore most useful when applied to individual target compounds. However, synthesis-planning programs generally contain a great deal of chemical knowledge and modification to perform the retrosynthetic analysis automatically should be possible. Following the analysis, identified synthetic routes and available starting materials could be scored and an estimate of synthetic accessibility obtained. One of these systems, LHASA, has already been modified to produce the program LCOLI [53], which is able to perform retrosynthetic analysis and starting material identification automatically. Although the actual aim of this system was to design libraries based around a lead compound, it would be possible to use the output of LCOLI as the basis of an automatic estimate of synthetic accessibility.

As mentioned above, in retrosynthesis-based techniques the target compound is disconnected, or rearranged, according to known chemistry. This is very different from the approach used in starting material-based methods where substructures are selected with little or no consideration of the chemistry involved in generating them. The disconnections are carried out by identifying chemical features, in a similar manner to the chemistry-based approach described above, and then modifying the target structure so that the synthetic precursors of the identified reaction are produced [54, 55]. A database of available reagents can then be searched for these precursors – using either exact or similarity-based matching – and any available starting materials stored. For precursors that are not found in the starting material database the process can be repeated, gradually reversing the series of reactions carried out in a multi-step synthesis.

This type of automatic deconstruction of a target structure is generally very time consuming – especially if the database used contains many retrosynthetic transforms – and results in the identification of a large number of possible starting materials. The number of poor starting materials identified can be reduced and the speed of analysis increased by including one or more retrosynthetic strategies in the analysis [56,57]. In synthesis planning programs these strategies are generally specified by the expert user and are used to determine where synthetic effort will be concentrated – for example forming stereocentres or ring systems. It is possible to include some of these strategies in an automatic system by identifying those parts of the structure that, when disconnected, are likely to produce available starting materials and applying retrosynthetic transforms preferentially in that part of the compound. One example of this is based on symmetry and requires the identification of bonds that, when broken, would result in two identical precursors. Such bonds are then preferentially targeted since such a disconnection would significantly simplify the synthesis.

Following retrosynthetic analysis it is necessary to score the identified starting materials and collect them into complementary groups – where each member in the group covers a different part of the target structure or, better still, was generated by the application of a single retrosynthetic transform to a given precursor. Starting materials can be scored based on a variety of factors including the number, and difficulty of retrosynthetic transforms applied to generate them, their coverage of the target structure (both in terms of number of atoms and of chemical features that may otherwise be difficult to synthesize) and how complementary they are to other identified starting materials. An assessment of the quality and coverage of the starting materials and information on the difficulty of the retrosynthetic transforms performed can then be combined, possibly along with a residual complexity score, to obtain a final estimate for synthetic accessibility.

An example of a retrosynthesis-based system is the CASEA program [35,58,59]. CAESA performs an analysis of a set of target structures using a number of retrosynthetic knowledge bases and a database of available starting materials. The search space is reduced through the use of retrosynthetic strategies that encourage disconnections that result in symmetrical precursors and aim to identify a convergent synthesis [57]. Performance is further enhanced by searching for the presence of available reagents using an intermediates database containing generalized structures. This database of intermediates is generated from a list of available starting materials by generalizing functional groups and applying simple synthetic transformations, such as functional group interconversions. Identified starting materials are arranged in groups, initially by the retrosynthetic route used to identify them and then based on coverage of the target structure. A complexity analysis of the target structure is then performed using a knowledge base of difficult to synthesize features in a manner similar to that described above. Following this the quality of each identified reagent is assessed based on its coverage of the target structure (in terms of both the number of atoms and complex features covered), the difficulty rating of the generating retrosynthetic transforms and complementarity to other available starting materials. Poor starting materials are discarded and the grouping process repeated. Finally, information from the complexity analysis is combined with the starting material data using Bayesian reasoning to obtain a final estimate of synthetic accessibility specified as a percentage.

The CAESA system has been thoroughly tested with the generated accessibility estimates compared with those produced by experienced medicinal chemists [35]. Mean absolute deviation between the CAESA estimates and the background synthetic accessibility score were found to be 9.9%, which compares well with the 10.6% mean absolute deviation between individual chemists' estimates and the same background score. As would be expected with a retrosynthesis-based system, CAESA suggests possible starting materials and synthetic routes for each of the target structures. While these suggestions are rarely exactly what would be used in a synthesis of any given target, they provide a useful starting point from which to develop a synthesis plan. Although in CAESA the retrosynthetic analysis is carried out using a convergent strategy, rather than the divergent strategy more usually employed in combinatorial synthesis [57], the program also retains a list of all of the precursors generated during the analysis of a set of compounds. This allows the identification of those structures with common intermediates, which may be included as a factor in the final assessment of synthetic accessibility. By including this factor, the results of a CAESA analysis may be used to assist in the design of combinatorial libraries containing multiple representative compounds from the initial set, particularly if knowledge bases containing retrosynthetic transforms corresponding to reactions amenable to parallel synthesis are used.

Retrosynthesis-based systems are the most complex methods used to assess synthetic accessibility detailed here and are likely to be the most computationally expensive. Such systems also suffer from the same disadvantages as chemistry and starting material-based techniques with respect to the building and maintaining of databases of chemical reactions and available starting materials. However, since starting materials are identified by the application of chemically meaningful disconnections the problems associated with using reagent databases in starting material-based techniques are significantly reduced. Retrosynthesis-based systems most closely duplicate the thought processes of experienced medicinal chemists and probably have the greatest chance of generating an estimate that would agree with one produced manually. Since both chemistry and available reagents are considered, retrosynthesis-based techniques are also likely to provide additional information that can be a useful starting point for synthesis planning, library design and the identification of possible analogues.

Neural Network-Based Estimation of Synthetic Accessibility

The use of neural networks to assess synthetic accessibility, in a similar manner to that used to estimate *drug likeness*, has been proposed. This process involves defining a set of descriptors that can be used to specify a structure and training a network based on these descriptors as an input and synthetic accessibility as the output. Neural network-based systems for the estimation of synthetic accessibility could use the same type of techniques as existing programs designed to measure the *drug likeness* of compounds [60-62]. Although synthetic accessibility is one factor of *drug likeness* these systems tend to be trained using databases such as the World Drug Index [63] to provide examples of *drug like* compounds and the Available Chemicals Directory (ACD [64]) as examples of nondrugs. All of the compounds in the ACD are by definition either currently available or have been in the past and thus would generally be considered synthetically accessible. The World Drug Index contains a wide range of drugs, including natural products, which would not necessarily be assigned high synthetic accessibility scores if they were not known to be available. Therefore systems trained using these databases are likely to differentiate between drugs and nondrugs on grounds other than synthetic accessibility.

A serious problem with this method is obtaining accurate estimates of synthetic accessibility for use in a training set, since a large number of examples are required in order to allow a neural network to generalize successfully. Manual generation of a training set has been considered. However it generally takes an experienced chemist at least a minute to produce a reasonably accurate estimate of accessibility and even then estimates of individual chemists are of limited reliability (see Manual Estimation of Synthetic Accessibility section, above). The manual assessment of a large training set would therefore be very resource intensive. It should be possible to build a training set using compounds that had previously been synthesized. In this case each compound could be assigned a synthetic accessibility score based on the number of steps in the published synthesis – with more steps resulting in a lower accessibility score. It would even be possible to include stable intermediate structures by counting the number of previous steps required to generate the intermediate. While this would not be completely accurate, since some steps are far easier to perform than others, it would embed some knowledge of synthetic chemistry into the system and should give a reasonable score that future estimates could be based upon. A database of available compounds, such as CHEMCATS [36], could also be used to provide examples of structures with high synthetic accessibility – with the exact score assigned modified by the source, purity and/or cost of the compound if required. The inclusion of compounds from starting material databases in the training set would allow the network to learn the types of reagents generally available and should result in estimates of increased accuracy. It is worth noting, however, that the use of databases of previously performed reactions and starting materials would result in a training set severely biased towards easier to synthesize compounds and containing few or no examples of structures that are very hard or impossible to synthesize. A combined approach with a training set consisting of both automatic and manually generated estimates would probably be most appropriate but would still require a significant commitment of resources.

The definition of a set of suitable descriptors to use as input to a neural network is also a non-trivial problem. A number of fingerprints are available and their performance in a variety of situations has been published [65-67]. However, none of these fingerprints have been developed specifically for use in the assessment of synthetic accessibility and thus the relevance of individual bits is debatable. In general existing fingerprints have shown to be very versatile and have been successfully used for a wide range of applications. It is therefore likely that they could either be used as is or adapted for purpose of assessing synthetic accessibility.

A successfully trained network should generate estimates that take into account all of the main factors considered by the methods described above – chemistry, starting materials, complexity – assuming that each of these factors were represented in the training set. Additionally, neural networks are likely to be significantly faster than any other technique except simple complexity-based estimation. The majority of the time taken to process a target compound would, in most cases, be spent generating the descriptor to be used as an input to the neural network. While the network is likely to need re-training occasionally, particularly to include newly available starting materials, this could probably be carried out relatively easily since examples of new chemistry and starting materials are likely to be present in published syntheses. Thus the addition of new examples of easy to synthesize compounds could be handled using the automatic assessment method detailed above. However, as mentioned previously, care would have to be taken to ensure that an appropriate number of representative examples with low synthetic accessibility were also included in any training set.

CONCLUSIONS

This paper has reviewed the assessment of synthetic accessibility and its uses in the drug discovery process. It is the authors' opinion that while its use is currently uncommon, synthetic accessibility potentially has greater utility than the often-assessed synthetic feasibility. However, the manual assessment of synthetic accessibility is a time-consuming process requiring expert knowledge and is more likely to result in differences in opinion than the simpler yes/no of synthetic feasibility. Little work has been carried out on the automatic estimation of accessibility but a number of alternative approaches are possible each with their own advantages and disadvantages and these were reviewed in this article.

Well-designed computational systems to estimate synthetic accessibility should have the advantage of being more consistent than manual techniques. A direct comparison of values produced at different times and on different machines would be possible as long as the same method of analysis and data were used. This is not necessarily the case with estimates from experienced medicinal chemists where estimates vary both between chemists and with a given chemist's experience. Although keeping computational systems up-to-date requires some effort, the rapid rate of expansion in both possible chemistry and available reagents means that it is impractical to expect synthetic chemists to be able to match the information available to database-based systems.

An automatic estimation of synthetic accessibility has a variety of uses in the drug discovery process including the testing of possible structural modifications and prioritization of syntheses. However, as computational power and the amount of raw information available increase systems that automatically generate an estimate of synthetic accessibility are likely to become more useful as a tool for enriching sets of potential leads. This will particularly be the case as the size of those sets increase due to improvements in the algorithms used to generate or select compounds.

Retrosynthesis-based techniques appear to be the most promising approach for the future since such systems attempt to duplicate the thought process employed by experienced medicinal chemists. Consequently such methods are well understood and generally accepted by medicinal chemists who are likely to be major users of the systems and responsible for the synthesis of compounds assessed by them. Although retrosynthetic techniques are relatively resource intensive, the speed of processing is likely to increase as more computational power becomes available [45]. Additionally, the effort required to maintain such systems (in terms of available starting materials and chemistry) should reduce as improved knowledge acquisition and extraction methods are included [41,42,43,44]. A key advantage that retrosynthesis-based techniques have over other methods of assessing synthetic accessibility is that they provide a very useful, and understandable, basis for manual starting material and reaction searching. Consequently, in addition to allowing resources to be prioritized, retrosynthesis-based techniques can aid in further development of compounds by assisting the synthesis planning of both individual targets and analogues.

ACKNOWLEDGEMENT

Jonathan Schmidt for his helpful interactions and proofreading of this manuscript.

ABBREVIATION

C.A.O.S = Computer Assisted Organic Synthesis

REFERENCES

[1] Böhm,H.-J. In *Rational Approaches to Drug Design*: Proceedings of the 13th European Symposium on Quantitative Structure-Activity Relationships; Höltje, H.-D.; Sippl, W. Ed.; Prous Science, S.A.: Barcelona, **2001**; pp. 367-371.

[2] Nishibata, Y.; Itai, A. *Tetrahedron*, **1991**, *47*, 8985-8990.

[3] Nishibata, Y.; Itai, A. *J. Med. Chem.*, **1993**, *36*, 2921-2928.

[4] Rotstein, R.A.; Murcko, M.A. *J. Comput. Aid. Mol. Des.*, **1993**, *7*, 23-43.

[5] Bohacek, R.S.; McMartin, C. *J. Am. Chem. Soc.*, **1994,** *116,* 5560-5571.

[6] MA: Mollecular Assembler, Part of the Evolutionary Molecular Design Technology, SignalGene Inc., Guelph, Ontario, Canada. Schmidt, J.M. Computational Method for Designing Chemical Structures Having Common Functional Characteristics, *U.S. Patent #5699268*, **1997**.

[7] Moon, J.B.; Howe, W.J. *Proteins: Struct. Funct. Genet.*, **1991**, *11*, 314-328.

[8] Rotstein, S.H.; Murcko, M.A. *J. Med. Chem.*, **1993**, *36*, 1700-1710.

[9] Gillet, V.; Johnson, A.P.; Mata, P.; Sike, S.; Williams, P. *J. Comput. Aid. Mol. Des.*, **1993**, *7*, 127-153.

[10] Gillet, V.J.; Newell, W.; Meta, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A.P. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 207-217.

[11] Mata, P.; Gillet, V.J.; Johnson, A.P.; Lampreia, J.; Myatt, G.J.; Sike, S.; Stebbings, A.L. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 479-493.

[12] Stahl, M.; Todorov, N.P.; James, T.; Mauser, H.; Boehm, H.-J.; Dean, P.M. *J. Comput. Aid. Mol. Des.*, **2002**, *I16, 459-478.*

[13] Roe, D.C.; Kuntz, I.D. *J. Comput. Aid. Mol. Des.*, **1995**, *9*, 269-282.

[14] Goodford, P.J. *J. Med. Chem.*, **1985**, *28*, 849-857.

[15] Danziger, D.J.; Dean, P.M. *Pro. R. Soc. Lond.*, **1989**, *B236*, 101-113.

[16] Danziger, D.J.; Dean, P.M. *Pro. R. Soc. Lond.*, **1989**, *B236*, 115-124.

[17] Miranker, M.; Karplus, M. *Proteins: Struct. Funct. Genet.*, **1991**, *11*, 29-34.

[18] Böhm, H.-J. *J. Comput. Aid. Mol. Des.*, **1992**, *6*, 61-78.

[19] Böhm, H.-J. *J. Comput. Aid. Mol. Des.*, **1992**, *6*, 593-606.

[20] Bartlett, P.A.; Shea, G.T.; Telfer, S.J.; Waterman, S. In *Molecular Recognition: Chemical and Biological Problems*, Roberts, S.M. Ed.; Royal Society of Chemistry: London, U.K., **1989**, pp. 182-196.

[21] Lauri, G.; Bartlett, P.A. *J. Comput. Aid. Mol. Des.*, **1994**, *8*, 51-66.

[22] Tschinke, V.; Cohen, N.C. *J. Med. Chem.*, **1993**, *36*, 3863-3870.

[23] Böhm, H.-J. *J. Comput. Aid. Mol. Des.*, 1994, *8*, 623-632.

[24] Ludi/ACD, Accelrys Inc., San Diego, California, U.S.A., Information may be found at the following website: http://www.accelrys.com/insight/LudiACD.html, last accessed 29th January 2003.

[25] Böhm, H.-J. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*, van de Waterbeemn, H.; Testa, B.; Folkers, G., Ed.; Verlag Helvetica Chemica Acta: Basel; Wiley-VCH, **1997**, 125-132.

[26] Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. *J. Comput. Aid. Mol. Des.*, **2000**, *14*, 487-494.

[27] Lego Tutorial, Gerber Molecular Design, Amden, Switzerland, **2002**. Manual given at the following website: http://www.msg.ucsf.edu/local/programs/moloc/lego.html, last accessed 24th January 2003.

[28] Jiang, W.; Wanner, J.; Lee, R.J.; Bounaud, P.-Y.; Boger, D.L. *J. Am. Chem. Soc.*, **2002**, *124*, 5288-5290.

[29] Pensak, D.A.; Corey, E.J. In *Computer Assisted Organic Synthesis*, Wipke, W.T.; Howe, W.J. Ed.; American Chemical Society, **1977**; ACS Symposium Series *61*, pp 1-32.

[30] Gasteiger, J.; Ihlenfeldt, W.-D.; Röse, P. *Recl. Trav. Chim. Pays-Bas*, **1992**, *111*, 270-290.

[31] Wipke, W.T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. In *Computer Assisted Organic Synthesis*, Wipke, W.T.; Howe, W.J. Ed.; ACS Symposium Series 61, American Chemical Society, **1997**, 97-127.

[32] Dengler, A.; Fountain, E.; Knauer, M.; Stein, N.; Ugi, I. *Recl. Trav. Chim. Pays-Bas*, **1992**, *111*, 262-269.

[33] Ihlenfeldt, W.-D., Gasteiger, J. *Angew. Chem. Int. Ed. Engl.*, **1995**, *34*, 2613-2633.

[34] Gasteiger, J.; Pfortner, M.; Sitzmann, M.; Hollering, R.; Sacher, O.; Kostka, T.; Karg, N. *Perspectives in Drug Discovery and Design*, **2000**, *20*, 245-265.

[35] Baber, J.C. CAESA: Computer-Aided Estimation of Synthetic Accessibility – Improved Algorithms for the Identification of Starting Materials, Ph.D. Thesis, University of Leeds, **1998**.

[36] CHEMCATS, available from Chemical Abstracts Service, Columbus, Ohio, U.S.A.

[37] LeapFrog, part of the Sybyl Modeling Suite, Tripos Inc., St. Louis, Missouri, U.S.A.

[38] Ash, S.; Cline, M.A.; Homer, R.W.; Hurst, T.; Smith, G.B. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 71-79.

[39] ChemNovel, part of the Chem-X Modeling Suite, Oxford Molecular Ltd., Oxford, U.K., **2000**.

[40] Blake, J.E.; Dana, R.C. *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 394-399.

[41] Gasteiger, J.; Ihlenfeldt, W.-D.; Fick, R.; Rose, J.R. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 700-712.

[42] Chen, L., Gasteiger, J.; Rose, J.R. *J. Org. Chem.*, **1995**, *60*, 8002-8014.

[43] Nakayama, T. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 885-893.

[44] Chen, L.; Gasteiger, J. *J. Am. Chem. Soc.*, **1997**, *119*, 4033-4042.

[45] Willett, P. *J. Chemometr.*, **1987**, *1*, 139-155.

[46] Barnard, J. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 532-538.

[47] Downs, G.M.; Willett, P. *Rev. Comput. Chem.*, **1995**, *7*, 1-66.

[48] Willett, P. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

[49] Sheridan, R.P.; Kearsley, S.K. *Drug Discov. Today*, **2002**, *7:17*, 903-911.

[50] Corey, E.J.; Cheng, X.-M. *The Logic of Chemical Synthesis*, John Wiley & Sons, Inc: New York, USA, **1989**.

[51] Chemical Abstracts, available from Chemical Abstracts Service, Columbus, Ohio, USA.

[52] CrossFire Beilstein, available from MDL Information Systems, Inc., San Leandro, California, USA.

[53] LCOLI: LHASA for Combinatorial Libraries, Part of the LHASA package, Harvard University, Cambridge, Massachusetts, U.S.A., Information may be found at the following website: http://lhasa.harvard.edu/combichem.htm, last accessed 27th January **2003**.

[54] Corey, E.J.; Cramer, R.D.; Howe, W.J. *J. Am. Chem. Soc.*, **1972**, *94*, 440-459.

[55] Bone, R.G.A.; Firth, M.A.; Sykes, R.A. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 846-860.

[56] Corey, E.J.; Long, A.K., Rubenstein, S.D. *Science*, **1985**, *228*, 408-418.

[57] Warren, S.G. *Organic Synthesis: The Disconnection Approach*, John Wiley & Sons Ltd., **1982**.

[58] Myatt, G.J. *Computer Aided Estimation of Synthetic Accessibility*, Ph.D. Thesis, University of Leeds, **1994**.

[59] Gillet, V.J.; Myatt, G.; Zsoldos, Z.; Johnson, A.P. *Perspectives in Drug Discovery and Design*, **1995**, *3*, 34-50. Synthetic Feasibility Project - Unpublished Results, Nanodesign, Inc., Guelph, Ontario, Canada, **2001.**

[60] Ajay; Walters, A.P.; Murcko, M.A. *J. Med. Chem.*, **1998**, *41*, 3314-3324.

[61] Sadowski, J.; Kubinyi, H. *J. Med. Chem.*, **1998**, *41*, 3325-3329.

[62] Brüstle, M.; Beck, B., Schindler, T.; King, W.; Mitchell, T.; Clark, T. *J. Med. Chem.*, **2002**, *45*, 3345-3355.

[63] Derwent World Drug Index, Derwent Information, Thomson Derwent, London, U.K.

[64] Available Chemicals Directory, MDL Information Systems, San Leandro, California, USA.

[65] Brown, R.D.; Martin, Y.C. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 1-9.

[66] Matter, H. *J. Med. Chem.*, **1997**, *40*, 1219-1229.

[67] Durant, J.L.; Leland, B.A., Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273-1280.