Molecular Complexity Analysis of de Novo Designed Ligands

Krisztina Boda and A. Peter Johnson*

ICAMS, School of Chemistry, University of Leeds, LS2 9JT, U.K.

The de novo approach to structure-based rational drug design can provide a powerful tool for suggestion of entirely novel potential leads. However, programs for structure generation typically generate large numbers of putative ligands; therefore, various heuristics (such as estimation of binding affinity and synthetic accessibility) have to be adopted to evaluate and prune large answer sets with the goal of suggesting ligands with high binding affinity but low structural complexity. A novel method for complexity analysis is described. This method provides a rapid and effective ranking technique for elimination of structures with complicated molecular motifs. This complexity analysis technique, implemented within the SPROUT de novo design system, is based on the statistical distribution of various cyclic and acyclic topologies and atom substitution patterns in existing drugs or commercially available starting materials. A novel feature of the technique that distinguishes it from other published methods is that the matching takes place at various levels of abstraction, so that it can evaluate complexity scores, even for structures which contain atoms with unspecified atom type, which is sometimes the case with the initial output of de novo structure generation systems.

1. Introduction

The de novo approach to rational drug design provides a powerful tool for the construction from smaller components of entirely novel molecules that satisfy a set of user-defined constraints, such as shape and electrostatic complementarity to a protein binding site. Many de novo design methods are able to suggest very large numbers of diverse putative ligands, and tools for navigating the answer sets are used to select a limited number of candidates for synthesis and biological testing. Predicted binding affinity is obviously an important parameter which is used in this context, but despite intensive research, the currently available functions for predicting binding affinity often perform poorly, and it would be unwise to carry out a significant amount of synthesis just on the basis of a high predicted binding affinity. On the other hand, much less attention has been devoted to the development of methods for evaluation of the synthetic accessibility of hypothetical ligands, an important consideration, since many of the structures suggested by de novo systems may be structurally too complex to be worthy of further synthetic and biological studies.

A related problem concerns the selection of the most "drug-like" compounds from a diverse set of structures. The extent of interest in this area is reflected by the large number of published works on this topic. (For comprehensive reviews about drug-likeness techniques and their applications in various stages of drug design, see refs 1−3). The methods deployed range from simple counting methods[4−5] to structural descriptor-based analysis[6] and pharmacophore point[7] and functional group filter[8] techniques, utilizing various computational techniques such as decision trees,[9] genetic algorithms,[10] and neural networks,[11−12] with the ultimate aim of classifying a compound as either drug-like or non-drug-like.

The most frequently cited heuristic guide, devised by Lipinsky,[4] and usually referred to as the "rule of five", was derived from an analysis of compounds from the WDI (World Drugs Index) database, designed to define the requirements for molecules to be successful as orally available drugs. It consists of limits on the number of hydrogen bond donors and acceptors, relative molecular weight, and lipophilicity (logP).

While the "rule of five" provides a set of pharmaceutically and biologically relevant "global" properties expressing drug-likeness, other methods have focused on the topological characteristics and "local" structural features of the molecule. For example, the CMC (Comprehensive Medicinal Chemistry) database has been analyzed in order to identify common drug-like frameworks,[13] side-chains,[14] and frequently occurring functional groups.[15]

The structural features of WDI structures have also been examined by a fragmentation program, called RECAP,[16] with the objective of extracting high quality building blocks for combinatorial library design by exhaustively cleaving structures into fragments by simply destroying bonds which are formed via common chemical reactions. The MDDR (MDL Drug Data Report) database has also been extensively analyzed in in order to identify bioisosteres[17] and local structural motifs.[18]

The complexity analysis method described in this study demonstrates that analysis of local structural motifs and their frequency of occurrence in databases of existing drugs and starting materials can provide a clear indicator not only of synthetic accessibility but perhaps also of drug-likeness.

1.1. Structure Generation in SPROUT. SPROUT[19,20] is an interactive de novo molecular structure design program that consists of several modules offering automatic methods for solving a number of problems associated with structure-based de novo design process. These include the analysis of a protein structure to permit the identification of potential interaction sites, the fragment-based generation of novel structures that fit steric and electrostatic constraints, and, in the final phase, scoring and clustering the solutions by various techniques including estimated binding affinity.

The structure generation process of SPROUT involves the construction of 3D generic molecular graphs called skeletons which satisfy the requirements of the receptor site. This process starts by docking molecular fragments to hydrogen bonding or hydrophobic target sites. These target sites are small, continuous geometric regions of space within the receptor cavity in which potential ligand atoms can be placed to achieve favorable

---

* To whom all correspondence should be addressed. Phone: +44 113 343 6515. Fax: +44 113 217 8559. E-mail: a.p.johnson@chem.leeds.ac.uk.
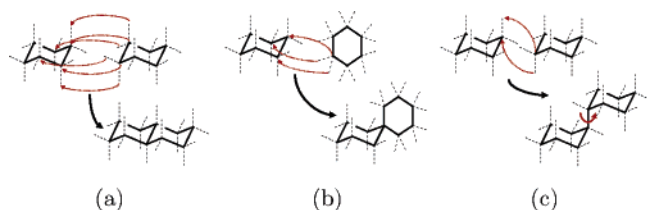
Figure 1. The three types of joining operation performed in SPROUT: (a) the fuse join, (b) the spiro join, and (c) the new bond join.
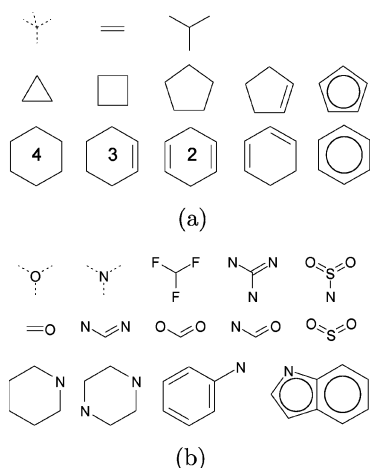

(a)

(b)

Figure 2. (a) Generic and (b) specific building blocks of SPROUT (Numbers inside rings indicate the number of different conformations of the fragment in the library).

interactions between the ligand and the receptor, thereby providing strong constraints for structure generation due to the highly directional nature of hydrogen bonding interactions.

Fragments docked to various regions of the receptor site are then linked together in a stepwise manner using a library of generic and specific fragments. In this sequential building up process, fragments are joined together in various ways such as fusing a pair of ring bonds together, spiro joining two ring atoms, and forming a new bond between any two fragment atoms (see Figure 1). This structure generation exhaustively explores the molecular search space in a deterministic manner, thereby leading to a wide diversity of generated structures.

The utilization of some generic building blocks provides one of the ways in which SPROUT tackles the inherent combinatorial nature of the structure generation process, which derives from the fact that there are a huge number of potential structures which can be generated from combinations of even a limited number of building blocks. While the user can easily augment the set of building blocks used by SPROUT, there is an initial default set, which currently contains the fragments shown in Figure 2.

SPROUT uses the generic fragment approach to moderate the space and the time-complexity of the problem by distinguishing the atoms of generic fragments by their hybridization state alone and not by their precise atom type. By this abstraction, a single generic fragment can be used as a surrogate for a whole family of potential "real" fragments that have the same connectivity and approximate geometry (see Figure 3). In practice, the generic fragments, called templates, are constructed from carbon atoms using the average carbon−carbon bond length and bond angle for the given hybridization.

Flexible fragments are represented in the library with multiple low energy conformations to introduce some ligand flexibility into the structure generation process (see Figure 2). Different
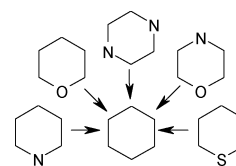

Figure 3. The fragment generalization approach in SPROUT. A class of geometrically similar "real" fragments can be represented by a generic "carbon" template.

applications may require different fragment sets; therefore, the program provides tools for managing the existing fragment libraries and building new ones.

Having assembled molecular graphs that satisfy requirements of the binding site, we can modify the structures by assigning specific atom types to any generic atoms, whereby molecular graphs are converted into "real" molecular structures which fulfill the electrostatic and the hydrophobic characteristics of the receptor. After atom substitution, the final structures can be ranked according to their estimated binding affinity.

2. Methodology

The current version of SPROUT has been tested on a variety of proteins and usually generates a large number of potential ligands with high predicted binding affinity. However, assembling ligands in a stepwise manner from simple building blocks does not guarantee synthetically feasible solutions. Therefore, it often requires a considerable amount of work by the chemist to evaluate the synthetic feasibility of the proposed structures or close relatives of them. Such manual evaluation of synthetic feasibility is possible for a small to medium number of candidates but is quite impractical for a larger answer set. This is a general problem for structures generated by de novo design, and a number of approaches have been explored which aim to provide computational solutions to the problem.

CAESA[19] is a rule-based expert-system which attempts to overcome this problem by scoring and ranking the generated structures according to an estimate of synthetic accessibility. However, the sophisticated retrosynthetic analysis used by CAESA means that it is relatively slow (seconds to minutes per structure) and could be used to analyze a final answer set within a reasonable time frame, but is not suitable for evaluating the synthetic complexity of the many thousands of intermediate structures generated in a SPROUT run. This ability would be practically useful, since it would allow pruning at an early rather than a later stage of structure generation, which is always desirable in situations where a combinatorial explosion has to be avoided.

In this connection, it is worth noting that pruning by synthetic accessibility is permissible at any stage of the structure-growing process because further growth will usually increase the difficulty of synthesis. This is not true for pruning by estimated binding affinity since a weakly binding part structure may produce a strongly binding ligand after further growth.

The inspiration for the complexity analysis method presented here comes from manual analysis of synthetic feasibility of many de novo structures generated over the years, where it was observed that in many cases synthetic complexity was caused by the presence of uncommon substitution patterns in rings and chains rather than from the presence of more obvious complex features such as stereocenters. The method is based upon the assumption that if a molecular structure contains only chain and ring structural motifs which occur frequently in commercially available starting materials or in previously (easily) synthesized
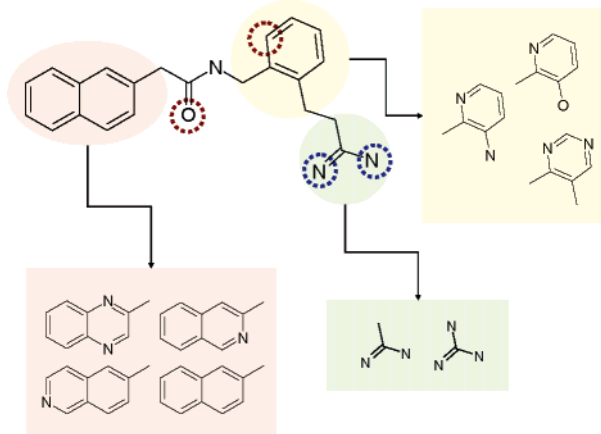
Figure 4. Example of a partially substituted structure, in which generic (indicated here as carbon) atoms can represent any atom type. (Blue and red dotted cycles symbolize donor and acceptor binding target sites, respectively).

structures, then the target structure is likely to be readily synthesizable.

To assess the frequency of occurrence of a particular substitution pattern, it is necessary to construct a complexity database which is generated by enumerating the local substitution patterns found in the source compound database and keeping counts of their frequency of occurrence. This statistical distribution of the substitution patterns can then be utilized to assess the synthetic accessibility (complexity) of structures generated by any de novo design program.

One of the advantages of this "local-structure" method is that if the source compound database comprises drug-like molecules, then the method might provide an indication of the "drug-likeness" of designed structures as well as their synthetic complexity, since structures largely or wholly composed of structural fragments from known drugs might also be "drug-like" in character. Clearly this "concept of drug-likeness" takes no account of global properties such as lipophilicity, and these should be assessed by other methods.

When considering a method for complexity-based prioritization of structures generated by de novo design, it is necessary to allow for the presence of generic atom types, i.e., the structures subjected to complexity analysis might possess atom types as yet undefined (Figure 4).

Accordingly the applied complexity analysis needs to be carried out at two levels (topology level and atom substitution level). The initial matching phase includes comparison of the topology of substitution patterns of the generated structure against the database of topological motifs found in drug-like structures and is performed taking into account bond order and atom hybridization. If there are matching topological motifs, then atom type matching is performed at the 2nd level of the analysis. In this second matching phase, specific atom types can be matched only with the same atom type whereas generic atoms can be matched against any atom type except when the atom is docked to a hydrogen bonding target site, in which case it can match only with atoms with the same binding properties (such as donor or acceptor).

2.1. Preparation of the Complexity Databases. Two separate complexity databases have been constructed for use in the estimation of complexity of de novo designed structures. The first one is used to quantify synthetic complexity as well as drug-likeness and it is built by extracting local structural motifs found in structures in the MDDR database, which contains over 100 000 biologically relevant compounds. The

Table 1. Number of Structures Eliminated from MDDR (itemized by applied filters)

	number of structures
total number of structures (initially)	113842
without 2D structural information	2676
with undesirable atom type[a]	864
molecular weight less than 100	66
molecular weight higher than 700	7185
without suitable therapeutic activity[b]	9469
total structures removed[c]	15673
remaining structures	95499

[a] Only H, Li, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, and I atom types are allowed. [b] Altogether 74 drug activity classes (with keywords such as "Sweetener", "Vitamin", "Mineral", and "Radiopharmaceutical") are excluded from the 615 activity classes found in MDDR. [c] The number of the structures removed by each filter represents the number of structures eliminated if the filters were applied independently of each other in order to illustrate their real pruning power independently of the order in which they are employed.

Table 2. Number of Structures Eliminated from Combined Starting Material Databases (Maybridge + Aldrich + Lancaster) Itemized by Applied Filters

	number of structures
total number of structures (initially)	169550
without 2D structural information	107
with undesirable atom type[a]	991
identical[b]	5723
total structures removed	6821
remaining structures	162729

[a] Only H, Li, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br and I atom types are allowed. [b] Starting materials are considered identical if they differ only in salt counter ion or stereo configuration.

second one, which is utilized to assess just synthetic accessibility, is constructed from the combined Aldrich, Maybridge, and Lancaster starting material databases (SM henceforth), which together contain almost 170 000 compounds.

Because MDDR contains a wide variety of molecules, some of which possess biological properties that are not regarded as "drug-like" along with drugs which are still under development, various filters are employed in order to eliminate any molecules with non "drug-like" characteristics from the final set that participates in the complexity database construction process. Consequently, inclusion of structures was based on a molecular weight range (100−700), absence of undesirable atom types, and membership of appropriate therapeutic classes. In contrast to other studies, structures marked as being in "Biological Testing" phase (∼92% of the structures in the MDDR database are labeled as being in the "Biological Testing" phase) were not removed, as it was felt that novel but synthetically accessible structures should not be penalized. The screening effect of the applied filters and the number of remaining structures are summarized in Table 1.

A similar filtering process was performed for the structures of the combined starting material database. If a compound occurred more than once in the united data set, then only one instance was kept. The number of structures removed by various filters is summarized in Table 2.

After undesirable compounds are screened out of the initial sets, hydrogen atoms are removed from each remaining structure along with any inorganic counterions or other residues. For the sake of consistency, the stereo information of input structures has to be ignored because of the large number of undefined stereocenters present in the overall input data set. This is followed by ring perception to determine the topology of each
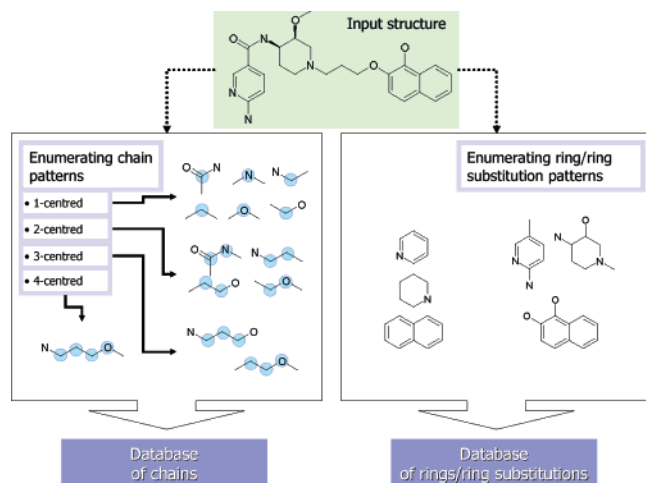
Figure 5. Example of the enumeration process to obtain chain and ring patterns (The centers of the chain patterns are highlighted by blue cycles).

Table 3. The 10 Most Frequent Substitution Patterns of Naphthalene in MDDR with the Frequency of Their Occurrences

| 3748 | 3288 | 1608 | 782 | 504 |
| 486 | 459 | 403 | 397 | 362 |

structure and detect rings (up to nine-membered), and by atom perception to assign essential properties (such as hybridization, aromaticity, number of attached hydrogens, and binding properties (acceptor, donor, both, neither)) for each atom.

In the next step, the complexity database is constructed by exhaustive and systematic enumeration of acyclic and cyclic substructures (i.e., patterns) present in each molecule (see example in Figure 5). The various types of analyzed patterns are defined as follows:

1-Centered Chain Pattern. Any nonterminal chain atom with its nearest heavy atom neighbors. There are six such patterns in the example structure (Figure 5).

2-, 3-, and 4-Centered Chain Pattern. Any 2, 3, or 4 adjacent nonterminal chain atoms with their nearest heavy atom neighbors. (The 1-, 2-, -3-, and 4-centered local structural concept is adopted from ref 18; however, in our method they are used only for acyclic portions of the examined structure).

Ring Pattern. Contiguous atoms participating in a ring or ring system. Fused, spiro, and bridged ring combinations are considered as one coherent unit.

Ring Substitution Pattern. Ring atoms together with their immediate bonded neighbors. This pattern type is particularly useful for estimation of synthetic accessibility since many ring substitutions occur quite infrequently. (See the frequency of occurrences of various side-chain substitutions of naphthalene in Table 3.)

The total number of substructural motifs (unique topologies, atom substitution patterns) for the MDDR and the starting material dataset are summarized in Table 4.

The efficient use of these large numbers of ring and chain motifs in complexity analysis requires a highly efficient procedure for exact structure matching. Canonical names which unambiguously encode molecular structures independently from their atom numbering are routinely used in chemoinformatics for exact structure matching. In our case, canonical topology names are constructed from the perceived hybridization of the substitution pattern atoms and the types of bond connecting them. The algorithm for canonical name generation was implemented by customizing the original Morgan algorithm[21] and SEMA[22] variant. By these means, the computationally expensive atom-by-atom matching required to determine equivalent topology is replaced by rapid string comparison.

For ease of search, the complexity database is implemented in the form of a set of hierarchies. Each detected topological graph (with an associated canonical topology name) corresponds to an atom substitution hierarchy (Figure 6). In the hierarchy, each node represents a unique atom substitution pattern for the given topology along with a counter that stores its frequency of occurrence. The hierarchy is constructed in such a way that the most generic atom substitution is located at the root of the hierarchy and more specific atom substitutions (i.e., patterns with more heteroatoms) are found by navigating deeper into the hierarchy.

The hierarchies associated with the topological graphs expand as more and more patterns are found during the complexity database construction. If a new heteroatom substitution pattern occurs for a topology, then a new node is inserted into the corresponding hierarchy by introducing relationships between the new node and the existing ones. If the atom substitution pattern is already present in the hierarchy, then the occurrence counter of the appropriate node is incremented.

Analysis of the retrieved patterns reveals structural differences between the MDDR and SM databases. Figure 7 illustrates the correlation between the frequency of occurrence of specific rings found in MDDR and their frequency of occurrence in the starting materials database. From the 7388 various ring substitution patterns, 2309 are not present in the MDDR and 4048 are missing from the SM. Figure 7 also depicts some ring motifs that are present in MDDR with high frequency, but absent or occur only once in SM, and vice versa.

Table 4. Number of Overall and Unique Enumerated Topologies and Atom Substitution Patterns in the Filtered MDDR and Combined Starting Material (Maybridge + Aldrich + Lancaster) Databases

	MDDR			SM			MDDR + SM		
	TO[a]	UT[b]	US[c]	TO[a]	UT[b]	US[c]	TO[a]	UT[b]	US[c]
1-centered chain	619682	144	937	752856	185	1359	1372538	202	1559
2-centered chain	424218	659	3524	457727	801	4453	818945	927	5942
3-centered chain	338982	2392	9108	303738	2609	9059	642720	3498	14796
4-centered chain	291357	5918	16931	208207	5454	12646	499564	8891	25862
ring systems	233572	2689	5085	360683	1853	3340	594255	3895	7388
ring subs	233572	14987	25926	360683	10270	19212	594255	22565	41047

[a] TO = Overall total occurrences of enumerated patterns. [b] UT = Number of unique topologies (considering hybridization, connectivity, and bond order). [c] US = Number of unique substitution patterns.
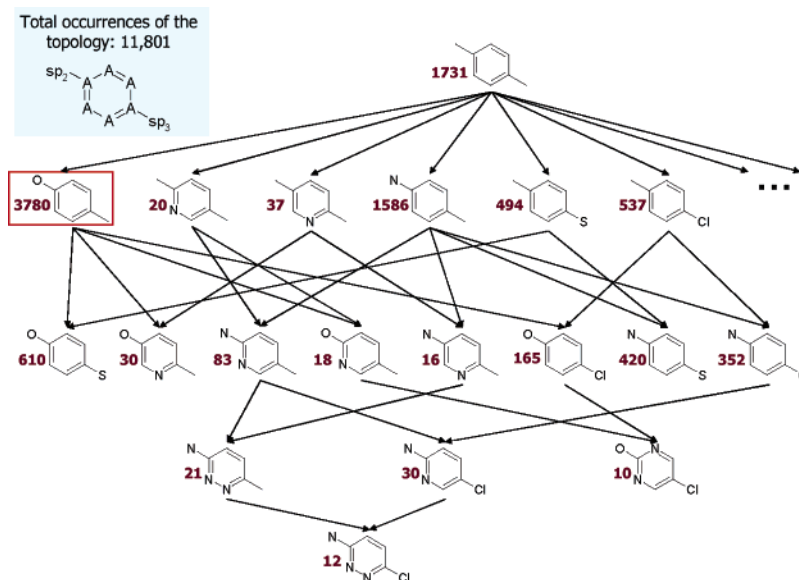
Figure 6. Example for the atom substitution hierarchy. The number next to structure indicates its frequency of occurrence in the screened MDDR. The most common atom substitution is in the red box. For the sake of clarity and simplicity, only 19 atom substitution patterns out of 64 found are depicted in the hierarchy. These occur more than 10 times and contain only C, O, N, S, or Cl atoms. The structure in the blue box represents the topology associated with the hierarchy. (sp_2, sp_3, and A indicate atom hybridization and aromatic atom type, respectively.)
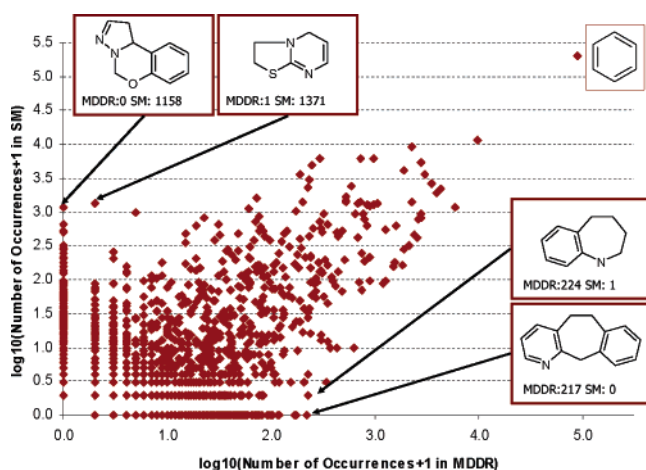
Figure 7. Frequency of occurrences of ring motifs (considering atom types and hybridization) found in MDDR versus ring motifs of SM.

Figure 8 shows the relative frequency of rings in MDDR and SM databases in percentages. For the sake of clarity, two rings which occur significantly more frequently than any other ring in both MDDR and SM have been omitted from Figure 8. These are benzene which represents 38.23% of the ring types in MDDR and 55.01% in the SM, and pyridine which represents 4.14% of the rings in MDDR and 3.24% in SM.

2.2. Estimating Structural Complexity. After the two complexity databases are constructed, they are used to score the drug-likeness and synthetic complexity of structures generated by SPROUT (or any other de novo design method). The complexity scoring system which has been adopted is designed to penalize structural motifs that are infrequent or absent from the complexity database.

The hierarchical architecture of the constructed database is designed to facilitate rapid multilevel complexity analysis. By this means, the first investigation takes place at the topology
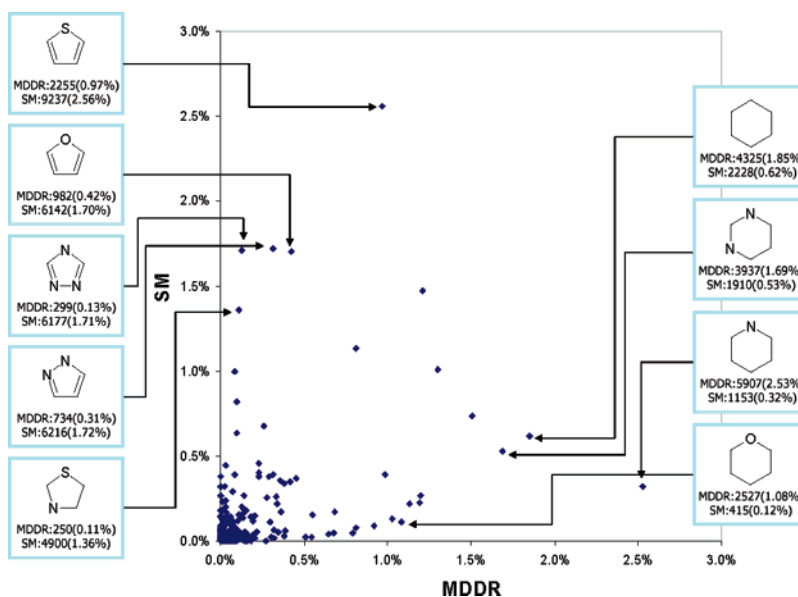
Figure 8. Relative frequency of ring motifs in percentages (considering atom types and hybridization) found in MDDR versus ring motifs of SM.

	Topology occurred	Topology *score*	Atom subs. occurred	Atom subs. *score*	Atoms
1st	2,341	*(8.785)*	154	*(6.998)*	6 8 5 7
2nd	57,915	*(2.080)*	31,008	*(0.000)*	7 6 9 10
3th	52,313	*(2.293)*	31,824	*(0.000)*	9 11 7
4th	24,526	*(3.876)*	7,468	*(0.000)*	20 15 22 23
5th	804	*(8.139)*	102	*(3.543)*	6 7 8 5 9 10
6th	27,535	*(1.874)*	17,319	*(0.000)*	7 9 6 10 11
7th	355	*(6.806)*	71	*(1.145)*	6 7 8 5 9 10 11
8th	104,749	*(0.000)*	90,498	*(0.000)*	5 4 0 3 1 2
9th	1,537	*(14.609)*	1,351	*(0.000)*	13 14 18 11 12 15 17 16
10th	20,636	*(0.000)*	16,857	*(0.000)*	5 4 0 6 3 1 2
11th	992	*(9.165)*	760	*(0.000)*	14 13 12 15 11 18 19 16 20 9 17 21

Total (normalised) score: 6.301

Figure 9. Example of complexity score calculation for cephalexin The scores are calculated by using the complexity database generated from MDDR. (Only exact atom type matching performed since this is not a de novo designed structure).

level, while on the second level the probability of heteroatom substitutions are examined.

The process starts by retrieving all chain and ring patterns of the de novo designed structure, in the previously detailed manner (Figure 5). The canonical topology names of the obtained substructures are then matched against the set of topologies stored in the complexity database. If the topology is present in the database, then a topology score is calculated using empirically derived equation 1 which considers the correlation of its frequency of occurrences with the occurrences of the most frequent topology for the given pattern type i.e., ring or chain.

$$\text{Score}_{\text{Topology}} = \begin{cases} \left(1 - \dfrac{ln(\text{No. occur. of matched topology})}{ln(\text{No. occur. of most common topology})}\right) * \textbf{Penalty} \\ \qquad \text{, if topology exists in database} \\ 2 * \textbf{Penalty} \\ \qquad \text{, if topology missing from database} \end{cases} \quad (1)$$

The penalty values used in the examples presented here are 25, 20, 15, and 10 for 1-, 2-, 3-, and 4-centered chain patterns, and 40 and 30 for rings and ring substitutions. These are not immutable constant values, and the user can alter the penalty values for individual pattern types in order to tailor the system for different applications. The current set of penalty values was derived by a process of trial and error and incorporates the heuristic that setting higher penalty values for smaller patterns emphasizes their importance.

Figure 9 itemizes the complexity score calculation for the cephalexin. In the table, each line represents an enumerated ring or chain pattern which contributes to the complexity analysis. The atom numbers in the table correspond to the atom numbering in the structural diagram on the right. The table details how many times each topology and atom substitution pattern occurs in the database and the partial score generated for the given pattern. This example shows that the 8th pattern (benzene ring) is the most frequent ring in the MDDR and therefore is not penalized. The topology of the fused ring (9th), however, occurs only 1537 times in the database; therefore, it increases the overall complexity of the structure.

If the topology is found in the database, then the second level of the analysis is performed to establish the frequency of the appropriate heteroatom substitution. This process involves traversing through the atom substitution hierarchy of the
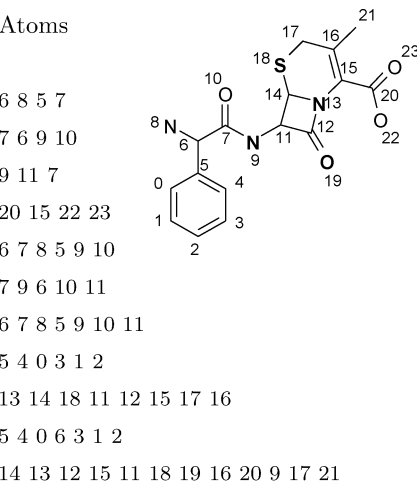
topology and identifying adequate atom substitution pattern(s) (Figure 6). The program is designed to perform two types of atom matching:

Exact matching finds the exact atom substitution pattern, if it is present, in the hierarchy of the topology.

Generic matching is implemented for partially substituted SPROUT structures. Generic atoms can be matched with any atom type. If a binding property (such as acceptor, donor) is defined for an atom, then it can match only with atoms exhibiting the same property.

In the latter matching mode, if there is more than one corresponding substitution pattern, then the one with the highest occurrence is used to calculate the atom substitution score (equation 2).

In case of the cephalexin example (Figure 9), there are quite a few chain (such as 2nd, 3rd, 4th, and 6th) and ring (8th, 9th, 10th, and 11th) substitution patterns which are the most common ones for the given topology. (Only exact atom matching is performed since this is not a de novo designed ligand.)

$$\text{Score}_{\text{AtomS}} = \begin{cases} \left(1 - \dfrac{ln(\text{No. occur. of best matched atomsubs.})}{ln(\text{No. occur. of most common atomsubs.})}\right) * \textbf{Penalty} \\ \qquad \text{, if matching atom substitution exists} \\ \qquad \text{ in database} \\ 2 * \textbf{Penalty} \\ \qquad \text{, if topology or atom substitution} \\ \qquad \text{ missing from database} \end{cases} \quad (2)$$

The total complexity score (Equation 3) is a normalized score which is composed of the individual topology and atom substitution scores divided by the number of patterns. In the

Table 5. Number of Enumerated Patterns of the 50 Top-Selling Drugs with the Number of Absent Topologies and Heteroatom Substitutions (itemized by pattern types)

	number of patterns	number of absent topology	number of absent atom substitution
1-centered chain	273	0	0
2-centered chain	186	0	2
3-centered chain	144	3	5
4-centered chain	123	6	9
ring systems	97	0	0
ring substitutions	97	3	6

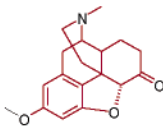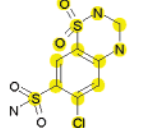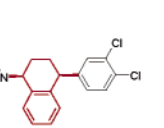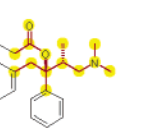**Table 6.** The 50 Top-Selling Drugs Ranked by Descending Order of Their Complexity Score[a]
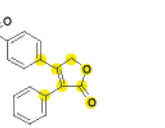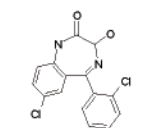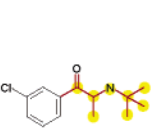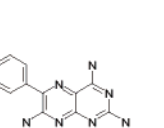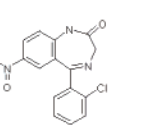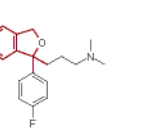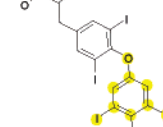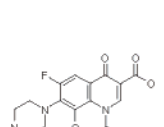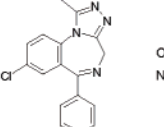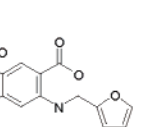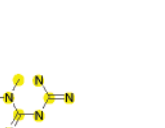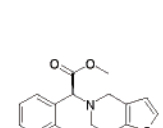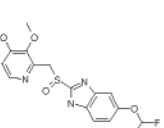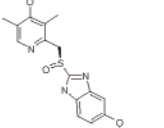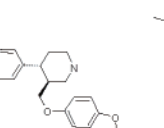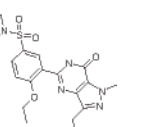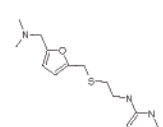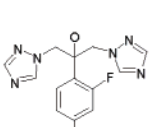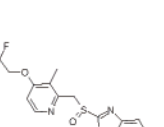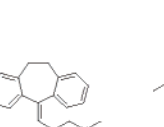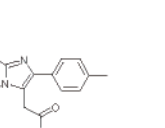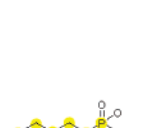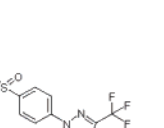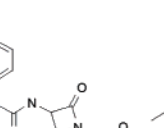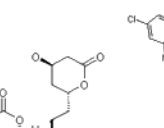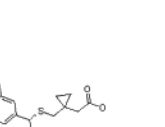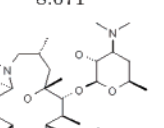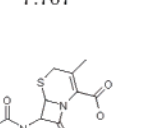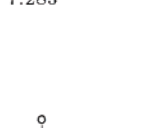
| (1) Hydrocodone 51.252 | (2) Hydrochlorothiazide 44.955 | (3) Sertraline 29.545 | (4) Propoxyphene 21.578 | (5) Rofecoxib 17.366 |

| (6) Lorazepam 15.237 | (7) Bupropion 14.357 | (8) Triamterene 13.845 | (9) Clonazepam 13.843 | (10) Citalopram 13.333 |

| (1) (11) Levothyroxine 13.260 | (2) (12) Levofloxacin 12.655 | (3) (13) Alprazolam 12.613 | (4) (14) Furosemide 12.302 | (5) (15) Metformin 11.586 |

| (1) (16) Clopidogrel 10.908 | (2) (17) Pantoprazole 10.238 | (3) (18) Esomeprazole 10.048 | (4) (19) Paroxetine 9.835 | (5) (20) Sildenafil 9.007 |

| (1) (21) Ranitidine 8.923 | (2) (22) Fluconazole 8.913 | (3) (23) Lansoprazole 8.834 | (4) (24) Amitriptyline 8.800 | (5) (25) Zolpidem 8.698 |

| (26) Alendronate 8.324 | (27) Celecoxib 8.071 | (28) Amoxicillin 7.707 | (29) Simvastatin 7.285 | (30) Montelukast 6.980 |

| (31) Venlafaxine 6.566 | (32) Azithromycin 6.550 | (33) Cephalexin 6.301 | (34) Albuterol 6.278 | (35) Gabapentin 6.188 |

Table 6 (Continued)

(36) Cetirizine 5.956
(37) Amlodipine 5.510
(38) Ramipril 5.483
(39) Valsartan 5.474
(40) Ibuprofen 5.329

(41) Quinapril 5.213
(42) Fluoxetine 5.181
(43) Pravastatin 5.177
(44) Atorvastatin 4.993
(45) Fexofenadine 4.922

(46) Metoprolol 4.910
(47) Atenolol 4.633
(48) Lisinopril 4.088
(49) Acetaminophen 3.811
(50) Salmeterol 3.362
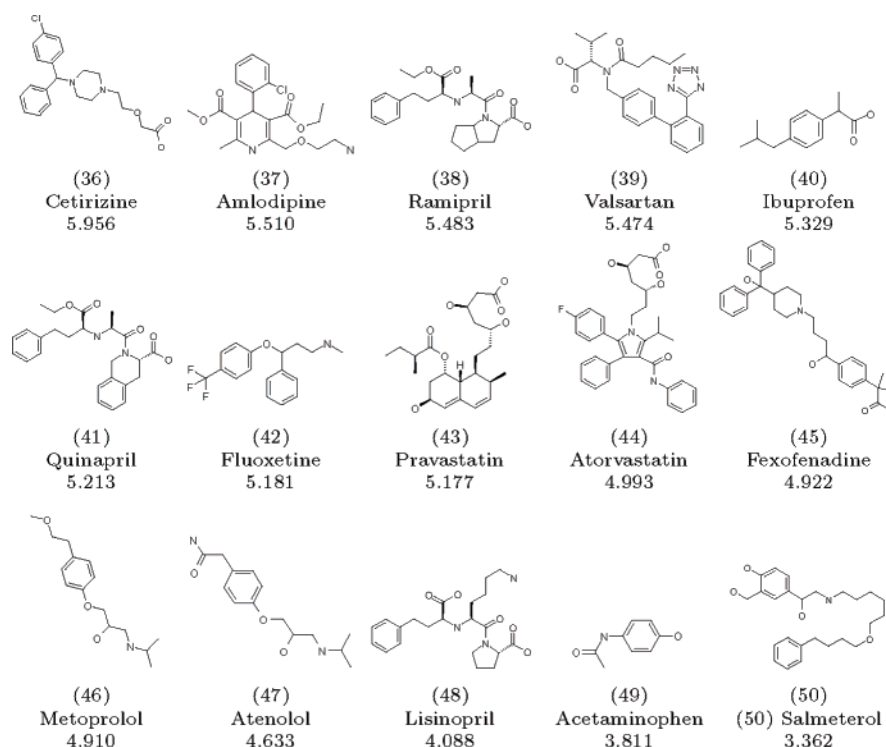
[a] Red highlights topologies and yellow cycles indicates heteroatom substitutions that are not present in the complexity database generated from MDDR.

Figure 10. The distribution of the estimated drug-likeness of 50 top selling drugs.

cephalexin example, shown in Figure 9, the total normalized score is 6.301.

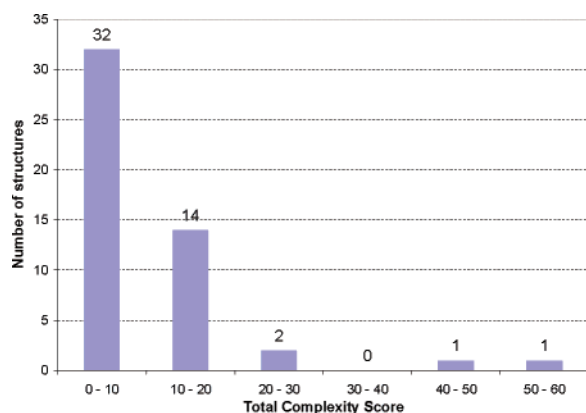The system also allows the user to invoke penalties for the presence of perceived stereo centers (P_s) and rotatable bonds (P_{rb}) which also contribute to poor synthetic accessibility and lower entropy of the structure, respectively.

$$\text{Score}_{\text{Total}} = \frac{\sum \text{Score}_{\text{Topology}} + \sum \text{Score}_{\text{AtomS}}}{\text{Number of Patterns}} + P_s + P_{rb} \quad (3)$$

3. Results and Discussion

The performance of the complexity analysis technique detailed above has been investigated by applying it to 50 top selling (nonsteroid) drugs.[23] All of the test drug structures were removed from the MDDR (the generic names of the drug structures are utilized to identify them in the MDDR database), and then a new validation complexity database was constructed from the limited set (using the previously detailed process). This was followed by the calculation of complexity score of each of

Table 7. The 50 Top-Selling Drugs Ranked by Ascending Order of Their CAESA Prediction Together with the Complexity Score

IN[a]	drugs	CP[b]	CS[c]	IN[a]	drugs	CP[b]	CS[c]
32	Azithromycin	10%	54.53	36	Cetirizine	66%	10.51
43	Pravastatin	16%	30.74	18	Esomeprazole	66%	14.76
29	Simvastatin	24%	38.0	59	Clonazepam	67%	12.95
1	Hydrocodone	25%	101.52	28	Amoxicillin	68%	30.31
38	Ramipril	32%	24.02	37	Amlodipine	69%	8.52
41	Quinapril	43%	15.40	16	Clopidogrel	69%	24.88
12	Levofloxacin	45%	19.59	50	Salmeterol	69%	6.96
44	Atorvastatin	46%	17.68	34	Albuterol	71%	12.17
33	Cephalexin	47%	19.14	46	Metoprolol	72%	8.41
20	Sildenafil	47%	26.19	23	Lansoprazole	73%	12.71
4	Propoxyphene	49%	29.16	7	Bupropion	73%	20.03
13	Alprazolam	51%	70.12	22	Fluconazole	74%	25.08
3	Sertraline	51%	35.99	26	Alendronate	80%	27.56
48	Lisinopril	52%	14.90	19	Paroxetine	83%	21.62
30	Montelukast	52%	18.89	49	Acetaminophen	83%	2.01
10	Citalopram	55%	17.07	5	Rofecoxib	87%	23.09
39	Valsartan	55%	18.03	21	Ranitidine	94%	15.80
17	Pantoprazole	57%	12.02	35	Gabapentin	95%	7.62
6	Lorazepam	59%	17.26	14	Furosemide	96%	13.12
27	Celecoxib	60%	5.73	11	Levothyroxine	100%	12.20
2	Hydrochloro-thiazide	61%	26.93	47	Atenolol	100%	7.78
25	Zolpidem	61%	16.16	40	Ibuprofen	100%	7.79
31	Venlafaxine	62%	13.56	8	Triamterene	100%	17.56
45	Fexofenadine	64%	11.24	15	Metformin	100%	10.35
42	Fluoxetine	65%	15.75	24	Amitriptyline	100%	19.06

[a] IN = Index of the structure in Table 6. [b] CP = CAESA prediction of synthetic accessibility. [c] CA = Complexity analysis score using SM complexity database.

the 50 drug structures. In this process only exact atom type matching was performed.

Table 5 summarizes the details of the study. Analysis of the results indicates that the method successfully identifies high percentages of topologies and heteroatom substitutions. The distribution of the scores is shown in Figure 10. The molecular structures of the analyzed drugs together with their generic names and their complexity scores are presented in Table 6.
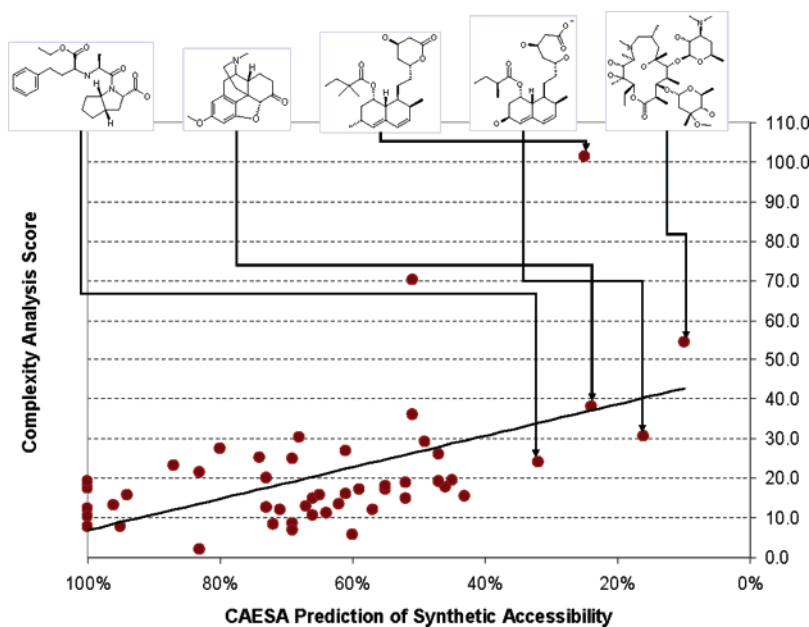
Figure 11. CAESA prediction of synthetic accessibility versus score calculated by complexity analysis. The black line indicates the linear trendline.
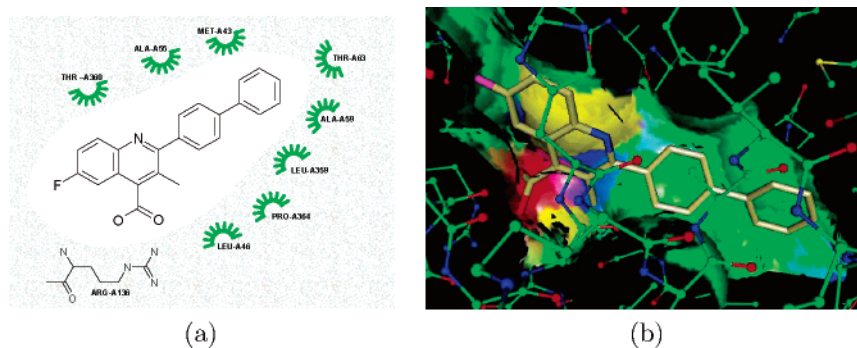

Figure 12. The key interaction region of 1D3G with the brequinar analogue: (a) the schematic representation of the complex; (b) snapshot of the inhibitor in the binding pocket.

The explanation for the surprisingly low score of azithromycin is that the program detects rings only up to nine-membered; therefore, the macrocyclic ring of the structure is matched against chain patterns rather than rings. Of the high scoring compounds, hydrocodone (**1**) has a genuinely complex structure. On the other hand, hydrochlorothiazide (**2**) is given a high score because of the unusual fused ring combination. The system does not currently recognize that retrosynthetic cleavage of the heterocyclic ring present would give a much simpler structure.

To validate the performance of the method as a predictor of synthetic accessibility, a comparison of these scores with CAESA predictions for the same set of compounds was carried out. CAESA[19] is an expert system-based program that is designed to estimate synthetic accessibility of individual members of a series of hypothetical drug candidates by performing a comprehensive retrosynthetic analysis, establishing synthetic routes between available starting material compounds and the target structure and also examining structural features, such as topology, stereochemistry, and functional groups contained within the structure that give rise to synthetic difficulty or complexity.

Complexity scores for these 50 drugs were calculated utilizing the complexity database constructed from just the starting material catalogues since the comparison only involved synthetic accessibility. In this process, exact atom type matching was performed again with a penalty value of 2.0 added to the total complexity score for each identified stereocenter in the structure.

The latter was introduced in order to compensate for the fact that stereochemistry had been ignored when the complexity database was constructed.

The CAESA predictions for synthetic accessibility and the complexity analysis scores for the top-selling drugs are detailed in Table 7. CAESA evaluates the synthetic accessibility in percentages i.e., the lower the percentages the more complex the structure. 100% indicates that the compound is commercially available.

Even though compared to CAESA the complexity analysis is a crude method for predicting synthetic accessibility, the analysis of the results reveals a surprisingly high correlation between the CAESA predictions and complexity analysis scores. Figure 11 displays the results and highlights the five most complex structures of the drug set according to CAESA.

Furthermore, the complexity analysis operates significantly faster than CAESA. The CAESA calculation took 703 s, whereas the elapsed time of loading the entirely complexity database into the memory and calculating the complexity scores were carried out in 8 s. Accordingly, the current method is more suitable for prioritization of thousands of structures within a reasonable time frame and provides an acceptable compromise between the speed of the analysis and the accuracy of calculated scores.

After validation of our method by analyzing its performance to predict complexity of known drugs, the technique was utilized to evaluate the complexity of structures constructed by a de
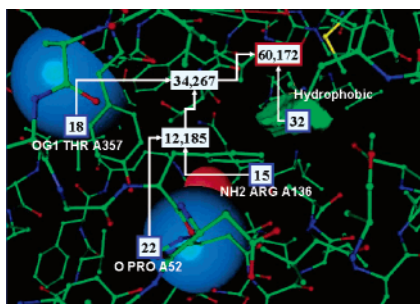
Figure 13. The connection of the selected target sites. Arrows indicate the steps of the connection. Numbers in boxes show the number of partial and final structures.
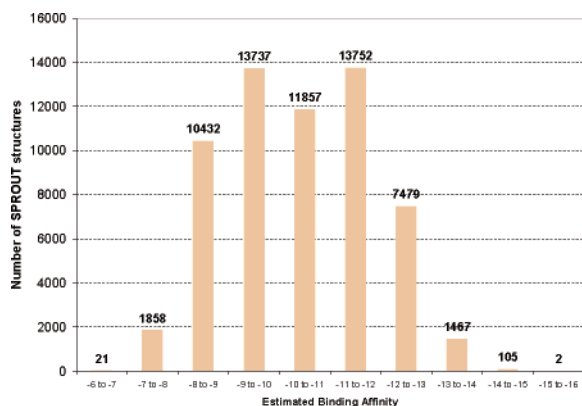

Figure 14. The distribution of the estimated binding affinity of the molecules generated.
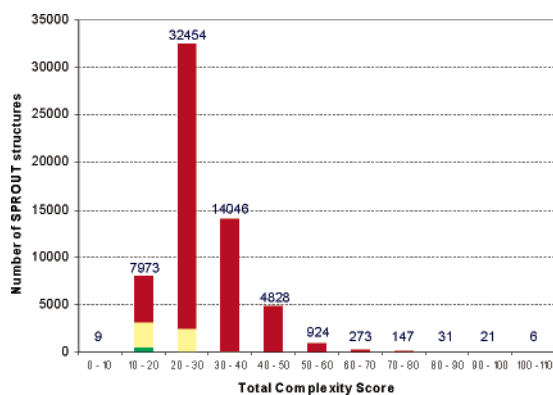

Figure 15. The distribution of complexity score of the molecules generated (red and yellow color indicate structures with topology or atom substitution problem(s), respectively. The green color represents structures for which every enumerated topology and substitution pattern in the examined structure is present in the complexity database.

novo design process. First, potential inhibitors were generated by SPROUT for the enzyme dihydroorotate dehydrogenase, the *P. falciparum* variant of which is an attractive target for the development of new antimalarial drugs.[24]

The de novo drug design process of SPROUT starts with importing the human variant of the enzyme complexed with a potent inhibitor, brequinar (PDB code 1D3G) into the SPROUT system (Figure 12), followed by the identification of hydrophobic regions and hydrogen bonding interaction sites.

In the next step, small fragments are docked to each selected target site followed by the structure generation phase which involves connecting the already anchored fragments together in a sequential building up process using both generic and specific building blocks in order to generate a diverse set of solutions.
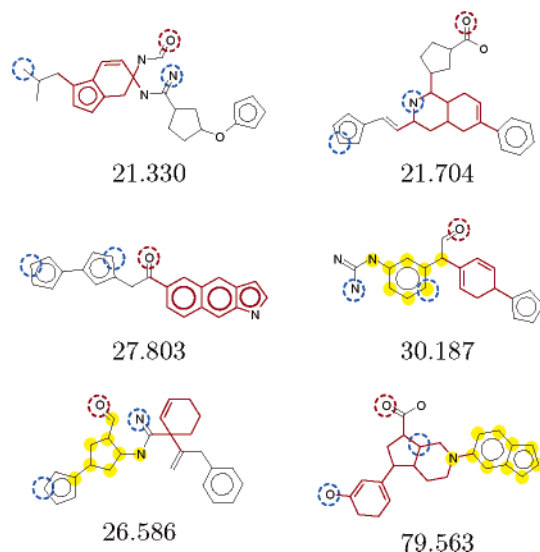

Figure 16. Examples of structures with high complexity scores (red color highlights topologies and yellow cycles indicate heteroatom substitutions that are absent in the used complexity database).
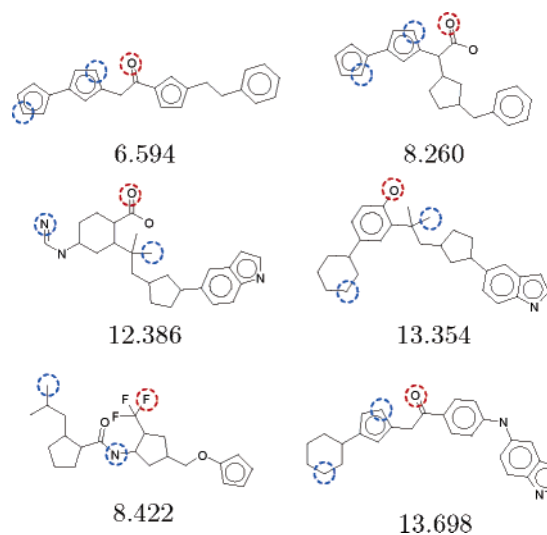

Figure 17. Examples of structures with low complexity scores (blue and red dashed cycles symbolize donor and acceptor binding interaction sites, respectively).

The numbers of docked fragments and the steps of the structure generation are shown in Figure 13. A total of 60 712 structures, which satisfy all of the selected target sites and the boundary constraint, were generated. These structures were then subjected to both the binding affinity and structural complexity estimation procedures. In the process of complexity analysis, generic atom type matching was performed using a combined complexity database of MDDR and SM. Figure 14 and Figure 15 show the distribution of binding and complexity scores of the generated structures, respectively.

According to the complexity analysis, only 5718 of the 60 712 structures (9.42% of the generated ones) have all their structural topology matched in the complexity database and only 495 structures (0.82%) with matching substitution patterns also. Figure 16 and Figure 17 depict structures with high and low complexity scores, respectively.

The user of the system has to make a judgment of where to apply a cutoff when using this complexity analysis. In this particular example a cutoff value of 20 would seem reasonable and would allow 87% of the answer set to be discarded (Figure 15).

In this example 1000−1200 structures (depending on size) were analyzed per minute on a 2.8 GHz Linux PC, which provides an indication of the efficiency of the process and a justification for the use of rapid topological name matching and the hierarchical structure of the complexity database.

4. Conclusion

A novel method for complexity analysis is described. This method matches structural motifs present in de novo generated structures against those found in compounds of drugs/starting materials databases and provides a quantitative structural complexity measure that can be used for prioritisation.

However, it should be emphasized that that there is a potential problem with any method which uses databases of existing structures, in that these compounds have not exhausted all possible local structural motifs. Therefore, structures with novel (but synthetically accessible) structural features may be incorrectly penalized as being complex. One way to ameliorate this problem would be to use a very large database, such as the Chemical Abstracts Registry file, as the data source.

The method described here can also be utilized in the reverse sense, for the analysis of catalogues of external suppliers in order to identify structures with novel structural motifs, the inclusion of which could enhance the chemical diversity of in-house databases.

References

(1) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.
(2) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002**, *54*, 255−271.
(3) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening − an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(5) Fecik, R. A.; Frank, K. E.; Gentry, E. J.; Menon, S. R.; Mitscher, L. A.; Telikepalli, H. The search for orally active medications through combinatorial chemistry. *Med. Res. Rev.* **1998**, *18*, 149−185.
(6) Xu, J.; Stevenson, J. Drug-like index: A new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(7) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(8) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.
(9) Wagener, M.; van Geerestein, V. J. Potential drugs and nondrugs: Prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(10) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(11) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "druglike" and "nondruglike" molecules?. *J. Med. Chem.* **1998**, *41*, 3314−3324.
(12) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(13) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(14) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(15) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(16) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP − Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(17) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.
(18) Wang, J.; Ramnarayan, K. Toward designing drug-like libraries: A novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* **1999**, *1*, 524−533.
(19) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34−50.
(20) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. Sprout − recent developments in the de-novo design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.
(21) Morgan, H. L. Generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(22) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.
(23) RxList LLC, The Top 200 Prescriptions for 2003 by Number of US Prescriptions Dispensed. url: http://www.rxlist.com/top200.html.
(24) Baldwin, J.; Farajallah, A. M.; Malmquist, N. A.; Rathod, P. K.; Phillips, M. A. Malarial dihydroorotate dehydrogenase. *J. Biol. Chem.* **2002**, *277*, 41827−41834.

JM050054P