Heteroaromatic Rings of the Future

William R. Pitt,* David M. Parry,† Benjamin G. Perry,‡ and Colin R. Groom§

UCB Celltech, Granta Park, Great Abington, Cambridge CB15 6GS, United Kingdom

Received December 2, 2008

Small aromatic ring systems are of central importance in the development of novel synthetic protein ligands. Here we generate a complete list of 24847 such ring systems. We call this list and associated annotations VEHICLe, which stands for virtual exploratory heterocyclic library. Searches of literature and compound databases, using this list as substructure queries, identified only 1701 as synthesized. Using a carefully validated machine learning approach, we were able to estimate that the number of unpublished, but synthetically tractable, VEHICLe rings could be over 3000. However, analysis also shows that the rate of publication of novel examples to be as low as 5−10 per year. With this work, we aim to provide fresh stimulus to creative organic chemists by highlighting a small set of apparently simple ring systems that are predicted to be tractable but are, to the best of our knowledge, unconquered.

Introduction

There have been a number of reports in the literature where the authors have estimated the size or have even enumerated parts of chemical space. The famous mathematician Arthur Cayley (1875)[1] was perhaps the first to calculate the size of a particular part of chemical space. He calculated the number of alkanes of the general formula C_nH_{2n+2} up to $n = 13$ (799). Perhaps the most quoted paper on the subject of chemical space dimensions is by Bohacek et al.,[2] where the authors estimate the size of the universe of organic molecules to be 10^{60}. Weininger[3] carried out an illustrative calculation of the number of *n*-hexane derivatives optionally decorated at each free valence with one of a possible 150 substituents. The result of 10^{29} is small in comparison to 10^{60}. However, when compared to estimates of the number of seconds since the big bang, of the order of 10^{16}, its immensity becomes apparent.[3] The size of the chemical universe grows exponentially with the number of atoms used for the obvious reason that as molecules get larger they can be further extended in an ever-increasing number of ways. A practical illustration of the combinatorial nature of the problem was provided by Ertl.[4] Computationally, he broke a set of three million commercially available compounds into substituents of less than 12 atoms. He then calculated the number of molecules with less than 36 heavy atoms that could be generated by combining these substituents, leading to an estimate of the number molecules that could be prepared using currently known synthetic methods of between 10^{20} and 10^{24}. Cramer et al.[5] also employed a fragment/reagent-based approach to explore a similarly sized virtual library of synthetically accessible compounds. They calculated that 100 or so reactions and 7000 building blocks could be combined to produce about 10^{20} potential products. They then went on to provide a means to virtually screen this entire chemical space by rapid filtration of reactants using either cost or shape similarity criteria. Google is rumored to process 20 petabytes (10^{15} bytes) of data per day, and petabyte data storage is now a (albeit extremely expensive) reality. However, assuming 10 bytes/molecule, one quickly runs into limitations of disk space and processing time when enumerating virtual libraries of orders bigger than 10^{16} molecules.

If molecules are decomposed to sets of reaction schemes and reagents, then problems of storage are greatly reduced. Similarly, the Markush representation of combinatorial libraries is additive in nature and therefore a far more efficient means of storing virtual libraries of this sort.[6] Where calculated properties are also additive, for example molecular weight, these can also be computed with the same improved efficiency.[6]

Within drug discovery there is currently considerable interest in fragment-based approaches[7] for the identification of novel ligands. One of the great attractions of this approach to drug design is that it leads to a far more efficient search of chemical space.[8,9] Taken to its logical conclusion, this reduces the problems of exploring chemical space and drug design to the larger of the number of possible reagents/fragments and the number of compatible reactions. Intuitively, and using the Cramer study[5] as a guide, it is likely that the former number will be much larger than the latter. As a demonstration of this, Fink et al. have created a virtual library of all possible organic molecules of up to 11 atoms containing only carbon, nitrogen, oxygen, and fluorine.[10,11] This library contains just over 26 million compounds; a number well within the capabilities of a hypothetical Google-sized virtual screening company (see above).

Complete enumeration of a virtual chemical library allows the mapping of that part of chemical space. Thus, a consistent frame of reference, as was employed in the ChemGPS[12] approach, can be realized. This completeness also allows for exhaustive similarity comparisons, as might be carried out when searching for isosteric scaffold replacements. Another advantage that completeness provides is in the field of inverse QSAR.[13] With a complete virtual library, one is assured of finding the globally optimal molecule or molecules in terms of a QSAR prediction or fit to a pharmacophore.

Current methods of generating virtual libraries can be divided into three categories: one-dimensional, string-based methods (e.g., Ho and Marshall[14] and Clark et al.[15]), graph theory-based methods (e.g., Fink et al. [10,11] and Brown et al.[16]), and reaction-based methods (e.g., Cramer et al.[5]); our method falls into this last category. This approach involves taking a set of reactants and a set of rules for combining them into virtual products. In our research, the "reactants" and "reactions" bear no relationship

* To whom correspondence should be addressed. Phone: 44-(0)1753 534655. Fax: +44(0)1753 447603. E-mail: will.pitt@ucb.com. Current address: UCB Celltech, 208 Bath Road, Slough SL1 3WE, United Kingdom.
† Current address: Addex Pharmaceuticals.
‡ Current address: Cyclofluidic.
§ Current address: Cambridge Crystallographic Data Centre.

10.1021/jm801513z CCC: $40.75
Published on Web 04/06/2009

to those that would be employed in a real laboratory. This freedom from the restrictions of chemistry in the real world allows the creation of purely imaginary molecules and opens the doors to some interesting computational experiments. However, only molecules that have been made or could be made are of any lasting interest, particularly to a pharmaceutical company! In this work, our aim was to produce a virtual library of synthetically tractable chemical structures through the careful design of "starting materials" and "reaction" schema. However, even after putting thought into the design process, many of the molecular structures generated were thought to be extremely unlikely to be synthesizable, necessitating a selection process. Calculated metrics of synthetic feasibility or complexity are of increasing utility as one way of evaluating virtual libraries.[17] The novel method described here uses empirical rules, generated automatically from comparisons of substructures that have been synthesized with those that do not occur in published sources of chemical structures. Its purpose is to ensure ring systems that are unlikely to be synthesizable are removed from further consideration and that any tractable novel examples are highlighted.

The virtual exploratory heterocyclic library (VEHICLe[a]) is a complete set of heteroaromatic ring systems. Heteroaromatic rings are common in synthetic bioactive small molecules and serve as a good model set for exploring work of this kind; in addition, they act as the focal point of the molecules within a series being worked on by medicinal chemists. The reasons are for this are: (a) they are often capable of highly efficient binding[18] to proteins because of their shape and hydrophobic nature, (b) their lack of flexibility, combined with hydrogen bonding potential from their heteroatoms, can provide a level of target selectivity, (c) rapid exploration of the effect of adding different substituents is facilitated by applicability of parallelizable reactions such as aryl−aryl bond formation (Suzuki, Stille) or similar palladium-mediated coupling reactions,[19] (d) two or more substitution positions can be explored without the complication of introducing a stereo center, and (e) an unusual heteroaromatic ring system or substitution pattern can provide some of the novelty required for a patent application. However, some of these advantages also lead to disadvantages: (i) their hydrophobic nature and flat shape can result in low aqueous solubility, (ii) a lack of flexibility, combined with hydrogen bonding, leads to restricted structure−activity relationships (SAR), (iii) rapid exploration of a chemical series can lead to large molecules, ultimately making drug discovery more difficult, and (iv) the most synthetically accessible ring systems have often appeared in many previous patents, making novelty more difficult to achieve. Because of the importance of rings within bioactive molecules, they have also been a focus for analyses and computational tools have been developed to help chemists select or replace them.[20,21]

Our initial aim was to discover the full set of heteroaromatic ring systems and then to assess how many potentially biologically active, synthetically tractable, and drug-like heteroaromatic ring systems could exist. This knowledge should help to predict whether novel chemical matter based on these types of structures will be increasingly hard to identify in future. If one subscribes to the theory that the druggable genome is relatively small and already well exploited,[22,23] then the search for novel anchor fragments could become increasingly difficult in the future, leading to increased patent congestion.

This paper describes the method of construction of a complete set of aromatic ring systems. We also analyze the contents in terms of general trends in usage of these ring systems within medicinal chemistry. One novel aspect of this paper is the use of a virtual library of this kind for the prediction of synthetic tractability. We propose that there are many simple and tractable aromatic heterocycles that are yet to be mentioned in the literature. Examples of these unconquered synthetic challenges are revealed in this paper in the hope that this will stimulate further progress in this important area of synthetic organic chemistry.

[a] Abbreviations: VEHICLe, virtual exploratory heterocyclic library.

Table 1. Number of Unique Molecules in the Selected Compound Datasets (see *Methods*) and the Number of Unique Ring Systems[a]

	M	R	VR	VR%
launched + phases II and III	2461	950	120	13
phase I	730	494	86	17
Derwent	44367	7910	388	5
vendor catalogues	2991988	24073	708	3

[a] M: molecules; R: distinct ring systems; VR: VEHICLe ring systems; VR%: the number VEHICLe ring systems expressed as a percentage of the total set of ring systems.

Results

The pipeline that constructs VEHICLe runs in about 3 min on a single core, 3 GHz Intel Xeon workstation. Without the inclusion of the carbonyl building block, only 2986 ring systems are produced, when included, the number increases to 24867. Within this number, 1744 pairs of tautomeric equivalents exist, in 772 clusters. This leaves 23895 unique ring systems. Table 1 shows how many of these ring systems occur within the compound data sets.

Heteroaromatic rings are important constituents of drugs and bioactive molecules. Of the 2461 drug molecules extracted from the MDDR (Symyx Technologies), there are over 1000 occurrences of VEHICLe rings (almost 2000 if you include phenyl). However, these are drawn from only 120 types. Compounds tested in man tend to have a much higher likelihood of containing a VEHICLe ring system compared to those in patent and commercial catalogues. However, the majority of ring systems in these molecules are not included in VEHICLe. This is usually because they contain more than two rings (286 and 117 for the lauched/phase II/III and phase I sets respectively) or they were partially saturated (496 and 275). For further analysis of ring system usage, see Lewell et al.[20] and Ertl et al.[21]

Synthetic Tractability Prediction: Results for the Training Set. To assess synthetic tractability, a random forest machine learning method, trained using empirical data on published ring systems, was employed. After training a random forest of 100 trees, the training data was fed back in and the synthetic tractability predicted. Figure 1 (upper) shows the very top of the tree generated from all the input data. Figure 1(lower) shows the contents of an example node or bucket. This example illustrates how tautomers are treated as separate structures. Depending on the substitution patterns present in the input data, one tautomer can be labeled "good" (i.e., present in a known compound) and another "bad". Unsubstituted examples in the training compound data sets were entered as drawn and the assumption made that these represented the most stable form. The classification results are shown in Table 2. As expected, none of the known ring systems were predicted to be intractable and a relatively small number (9%) of the unknown ring systems were predicted to be tractable.

The distribution of p(good) values produced is shown in Figure 2. As can be seen from this plot, only a few ring systems

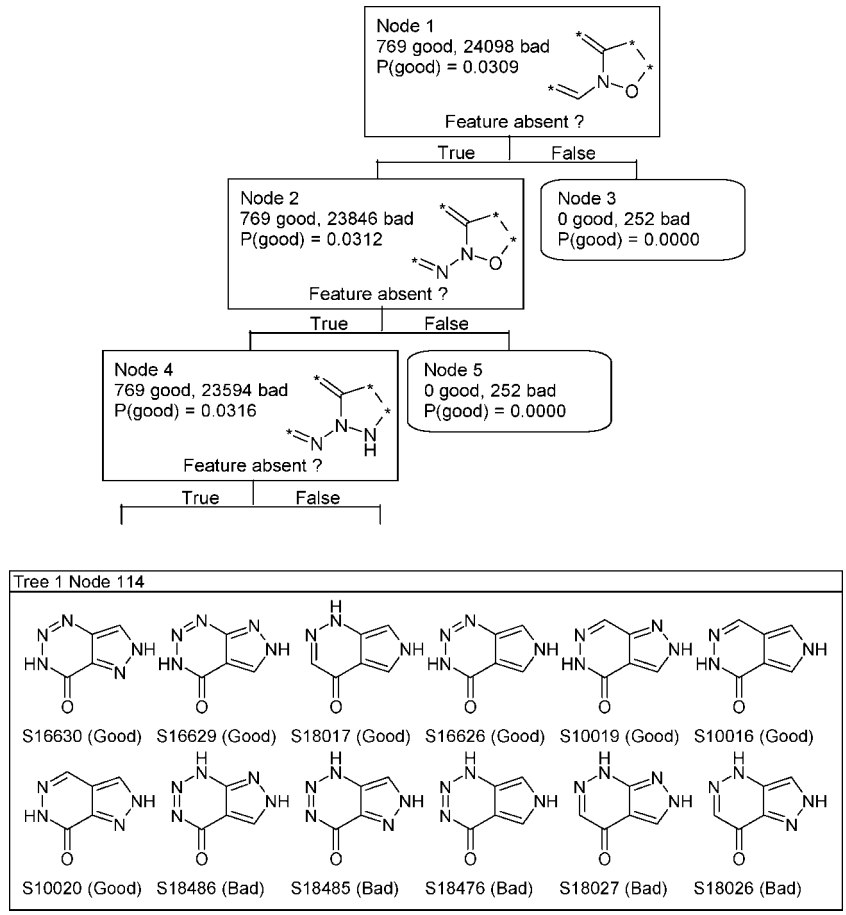

Figure 1. (upper) Top three of the 200 layers of the decision tree trained with all the data in Table 1. (lower) Example contents of a node. This node is classified as "good" on the basis that more than 20% of its constituents are from known molecules. In this case, the *p*(good) value is 0.37 (7/19), as 7 out of the 12 ring systems are known.

Table 2. Results of the Classification of the Training Set

input[a]	prediction[b]	*N*[c]
good	good	769
good	bad	0
bad	good	2185
bad	bad	21913

[a] Good: known ring system in training compound dataset. Bad: unknown. [b] Predicted tractability (*p*(good) > 0.5). [c] Number of ring systems.

are in the gray area 0.1 < *p*(good) < 0.9. All the known ring systems had a *p*(good) of over 0.9.

The method described above highlighted 2185 ring systems as potentially synthetically tractable but that had not appeared in the training compound data sets. The next step was to assess whether there was any validity to these predictions.

Synthetic Tractability Prediction: Validation Using Results for Unseen Data Sets. To test whether the decision tree classification method produced useful results, additional databases of synthesized compounds were searched. If the method is predictive, then significantly more ring systems that were previously labeled as "bad" (i.e., not present in the original compound data sets) but occur in one or more of these addition databases should be classified as "good" (true positives) than classified as "bad" (false negatives) by the random forest. The false positive rate is harder to assess, as synthetic tractability remains an open problem.

A set of 36 VEHICLe heteroaromatic rings from the UCB corporate collection, which were not in the original compound training set, were classified using the random forest. To assess

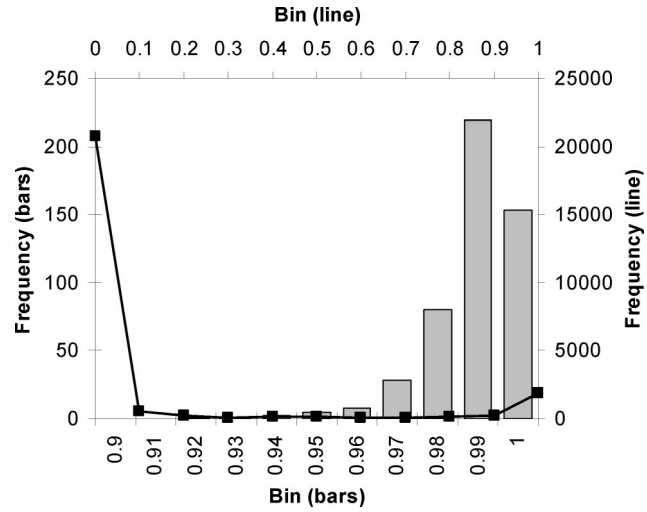

Figure 2. Histogram of the random forest output binned by *p*(good) value. The line graph shows the frequency within VEHICLe as a whole. The bar chart shows the frequencies for only the "good" molecules in training set.

the closest similarity of these structures to any of the training set "good" molecules, a nearest neighbor analysis was carried out. The nearest Tanimoto similarities using MDL public keys (Symyx Technologies) were in the range 1.0−0.68. However, this type of similarity measure is not very informative for such small compounds. Of the 36 structures, 12 were regioisomers, 10 had nitrogen insertions, 6 were tautomers, 2 had miscel-

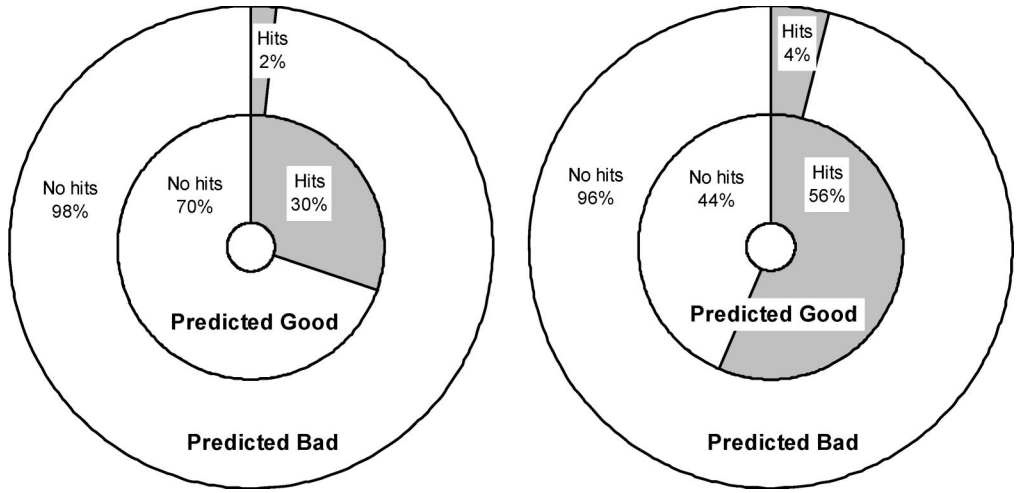

Figure 3. Random forest predictions of synthetic tractability shaded by the percentage of ring systems with at least one substructure hit in the Beilstein database. The numbers at the top of each pie are the total number of rings systems included. Left pie chart shows the result if good is defined as having a p(good) > 0.5. The right pie chart shows the result if only ring systems with a p(good) of 0.0 (bad) or 1.0 (good) are selected.

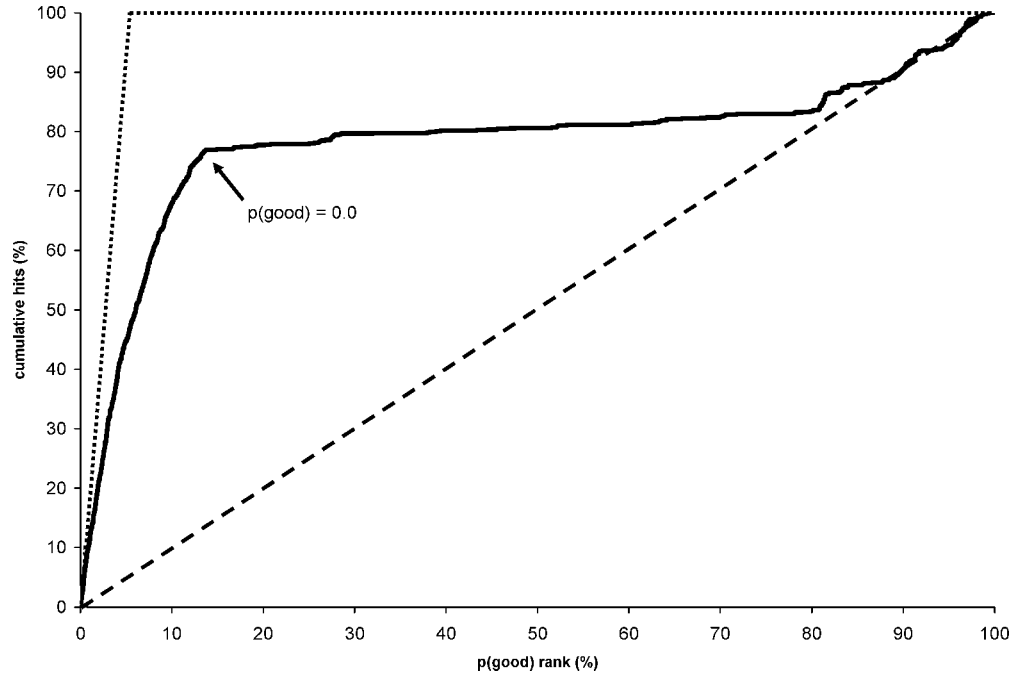

Figure 4. Enrichment plot of ring systems with Beilstein hits ordered by p(good). The dashed line indicates random performance and the dotted line perfection. The rank after the marked p(good) = 0 is arbitrary.

laneous changes, 2 had two changes, and 3 had no real similarity to any training set "good" molecules. This entire set of UCB ring systems were successfully classified as "good" with a p(good) ≥ 0.95 by the random forest method, i.e., we predicted that they could be made and subsequently discovered that they had been.

A more extensive validation was carried out employing the Beilstein chemical literature database. Only ring systems with no substructure hits within our original compound data set were included. This left 24098 rings systems out of a total of 24867, of which 21913 were classified "bad" and 2185 "good" by the random forest. Of the "bad" ring systems, only 374 (a minimum 2% false negative rate) were found as substructures of one or more compounds in the Beilstein database, and 663 (a minimum 30% true positive rate) of the "good" ring systems were found (see Figure 3). This represents a 15-fold improvement over a random classification. If only ring systems that had a classification p(good) of 1.0 were selected, this ratio increases to 56 (56/1, see Figure 3). Of these 56 ring systems, five had tautomers in the training set, two had at least one tautomeric hit in Beilstein, and two were different tautomeric forms of the same structure. This leaves 48 ring systems that are predicted to be synthesizable with p(good) = 1.0 that are not present in any of the compound databases (a total of over 10 million compounds) searched so far.

Figure 4 shows that about 75% of the ring systems with at least one Beilstein hit were found in the first 15% selected. This plot also shows that even predictions with a $0 < p(\text{good}) < 0.5$ are good indications of ring system tractability. There are, however, six ring systems with more than 30 Beilstein hits classified as "bad" with p(good) = 0.0 among the misclassifications. These are shown in Figure 5. These must be misclassified because they contain substructures that are underrepresented in the training data.

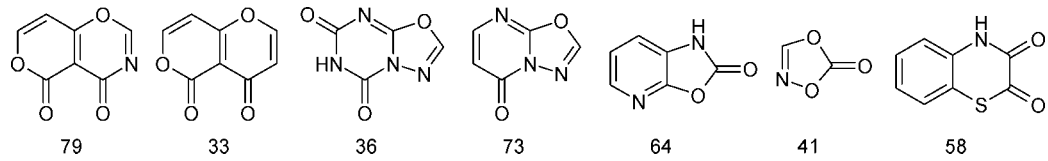

Figure 5. Ring structures that were misclassified as synthetically intractable. Shown below each structure is the number of times each was found in a molecule in the Beilstein database.

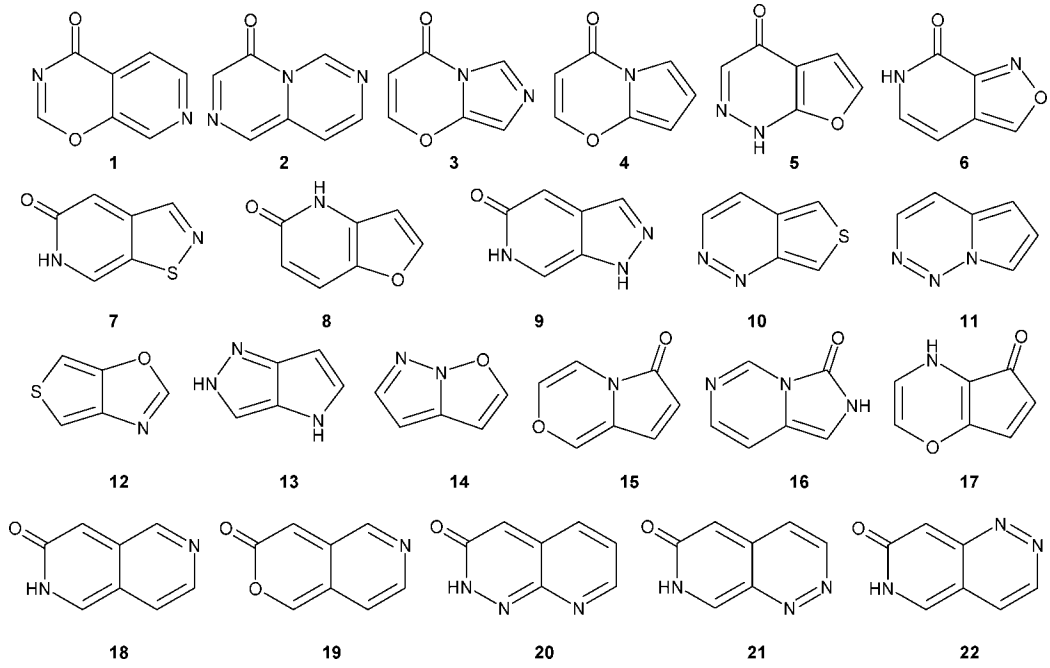

Figure 6. The 22 ring systems selected by eye from those with four or fewer heteroatoms and with $p(\text{good}) > 0.95$ calculated using the final random forest generated with Beilstein data included. All had zero synthetic papers found in SciFinder at the initial time of writing, including tautomers. Note added in proof: **8**,[24] **9**,[25] and **13**[26] were published for the first time in 2008; this satisfying observation was discovered during a SciFinder search carried out at the end of the reviewing process for this article.

A further search of the literature was done to check how many of the 48 ring systems that are predicted to be synthesizable with a calculated $p(\text{good}) = 1.0$ could be found. Individual substructure searches were done in SciFinder (Chemical Abstracts Service), which contains more than 36 million chemical substances. The substructure search engine of this product will not only find exact matches but also other tautomeric forms of the query. Of the 48, 15 were found to be part of published compounds, with papers usually dating back to the 1960s or 1970s, leaving 33 remaining unknown.

We then went on to generate a new random forest using the searches of Table 1 data sets and the Beilstein database combined. As might be expected, this tree predicted even more ring systems to be tractable. In fact, it predicted 3288 unique (after tautomer duplicates and tautomers of the training set are removed) unknown ring systems to be tractable with $p(\text{good}) > 0.5$. Of these, 232 unique ring systems had less than five heteroatoms and $p(\text{good}) > 0.95$. For illustration, we selected a small sample of 35 of these ring systems and searched for them in SciFinder. There were 13 that had substructure hits, and the remainder are shown in Figure 6. These 22 ring systems are displayed as a challenge to creative organic chemists to either make or explain why they cannot be made!

Ring System Complexity and Usage Statistics. Once all database searches were complete, a better analysis of the trends within the subset of known ring systems could be carried out. One of the most striking and perhaps predictable observations was the scarcity of ring systems containing a double nitrogen bridge. Of the 5841 examples containing this moiety within VEHICLe, only seven known examples were found.

As one might expect, the ring systems that have been made and incorporated in organic compounds tend to be the least complex of the complete set found in VEHICLe. Molecular complexity or synthetic difficulty can be calculated in a number of different ways, but one component can be the number of heteroatoms;[17] other factors that have been used including number of chiral centers, rings size, etc., are not relevant here. The VEHICLe database is dominated by ring systems with a high proportion of heteroatoms. The number of ring systems in the database peaks at 6−7 heteroatoms (see Figure 7) and between 4−6 carbons. If all these ring systems were equally synthetically tractable and desirable, then their occurrence in known compounds would follow the same trend. However, of those that are found in one or more of the existing compound databases, most have 4 heteroatoms and 5−6 carbons. The significance of the deviation of ring system usage from random can be calculated by the following equation:

$$\text{significance}_{i,j} = \frac{(\text{nobs}_{i,j}/\text{nobs}_j)}{(\text{ntot}_{i,j}/\text{ntot}_j)}$$

where $\text{significance}_{i,j}$ is the significance of the frequency of observation a ring system with i heteroatoms in set j; nobs is the frequency of observation, and ntot is the total number of ring systems in VEHICLe. For each ring system, an observation on the compound data sets in Table 1 is counted only once. It

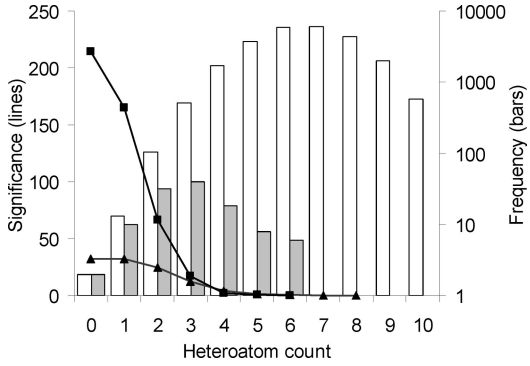

Figure 7. Histogram of the frequency of ring system occurrence in VEHICLe, binned by number of heteroatoms for all ring systems (white bars) and those found in molecules labeled as launched or in clinical trials in the MDDR (gray bars). Also shown is significance (see equation above) of the number of hits found for all compounds in the original training set (triangles) and for the MDDR drugs set (squares).

can be seen in Figure 7 that this significance is highest for 0 (benzene) or 1 heteroatom and drops off rapidly, especially for the set of ring systems that occur as substructures of drugs in clinical trials or launched on the market. In terms of information theory, this is data that could be described as having low entropy, the order being imposed in this case by the limitations of synthetic tractability or to "the principle of least effort".[27] In drug discovery, for example, a novel or rare complex ring system will not be used unless absolutely necessary. This effect could be compounded by the use of a common pool of inexpensive commercially available reagents. Novelty of chemical matter can often be achieved more easily by adding particular substituents at different positions around a ring acting as a scaffold for other chemical functionality.

The search of the Beilstein database was extended to all VEHICLe ring systems that had less than 100 hits in original compound database searches. This set of 182 most common ring systems were omitted in order to speed up the database search. A total of the 1519 of the remaining ring systems generated at least one hit in Beilstein, bringing the total of known rings to 1701. Plotting the frequency of occurrence versus rank on a log−log scale shows that VEHICLe ring system frequency does not follow Zipf's power law, which would lead to a straight line (see Figure 8). This plot clearly shows the overwhelming prevalence of the phenyl ring, occurring over six million times in the three million compound data set in Table 1. Power−law distributions have been observed in chemical substructures, for example in Bemis and Murcko assemblies[28,29] and other substructure types.[30] Here, only the ring systems that occur less than 500 times in the compounds follow a power−law straight line. This is similar to the observations of Lipkus et al.[29] of rigid segments and ring systems combined. They found that 10% of these types of substructure account for 90% of occurrences in their compound data set. Here the effect is even more marked with 2% (15 ring systems) accounting for 90% of occurrences, with 70% by phenyl alone. This could be in part due to the fact that commercial catalogues give rise to the bulk of the compounds studied here, which may well be less diverse than the CAS Registry sample used by Lipkus et al.[29] The reuse of ring systems probably also reflects factors such as the phylogenetic nature of drug discovery[31,32] and the conservative, risk minimizing nature of the business rather than the unsuitability of the majority of ring systems.

Novel Ring System Publication Statistics. When predicting whether novel ring systems have synthetic potential, it is informative to examine the rate at which new ring systems are published in the literature. This information gives some indication of the chances of these predictions being realized and at what rate.

A sample of VEHICLe was selected to give a historical insight into our coverage of heteroaromatic space. The ring systems that had been previously used as Beilstein queries were added to those that had less than 100 hits in the original compound search to give a total of 24595 queries. These queries produced 79456 hits, with 1519 queries having at least one hit. The distribution of the number of hits per query is shown in Figure 9. This plot shows that many of these hits came from a relatively small number of queries. The 1196 queries that gave between 1 and 40 hits generated a total of 10285 hits and data on these hits could be obtained by doing just 11 separate searches. Excluding relatively few commonly published ring systems (272 out of a possible 1791) in this way is therefore unlikely to affect an analysis of publication rates of novel heteroaromatic rings over recent years.

The publication date of the oldest hit for each of the 1196 ring systems was extracted and results plotted in Figure 10 (bar chart). According to the records stored in Beilstein for this sample, the publication of compounds containing entirely novel ring systems reached a peak rate of 41 per year in late 1970s and declined to a rate of 6−10 a year in the first five years of this century. However, this distribution could simply reflect the distribution of records within the Beilstein database as a whole. The total number of records binned by publication year is not available, but the histogram of all hits generated by the 1196 queries shows this structure to some extent (see Figure 10 line chart). It can be seen that there is indeed a similar distribution, again showing a peak in the late 1970s. However, unlike the first publication dates, the rate does not decrease thereafter. This shows that the drop-off in the publication of novel ring systems for final quarter of the last century is likely to be a genuine trend. This observation is in line with our initial assumption that the majority of the VEHICLe ring systems that could be made have already been made.

The 73 ring systems for which the Beilstein database contained no reference to earlier than the year 2000 were searched individually using SciFinder. Of these, 41 were confirmed as being first published this century. Some were substructures of bioactive molecules, including described in COX-2 and p38 inhibitor patents. The p38 inhibitors were novel analogues of VX-745, a p38 inhibitor from Vertex, which reached phase II clinical trials. The heteroaromatic core of this molecule was first seen in the literature in 1998 (see Figure 11).

Discussion and Conclusions

Within UCB, VEHICLe has been used for several years to aid scaffold replacement design on medicinal chemistry projects. This has been achieved in conventional ways similar to those published by Lewell et al.[20] using only ring systems that have published synthetic routes. As such it has been a complete, indexed, and annotated reference library that has provided input for virtual library production and screening. Perhaps the best known example where this approach has born fruit is the Vardenafil (Bayer) analogue of Sildenafil (Pfizer). Both are "blockbuster" drugs, and their main structural difference is in the core heteroaromatic ring system. The syntheses of these ring systems were first published in 1966 and 1979, respectively. As well as aromatic ring systems, a library of half-aromatic, half-saturated ring systems was also generated, as these were thought more likely to produce more soluble compounds.

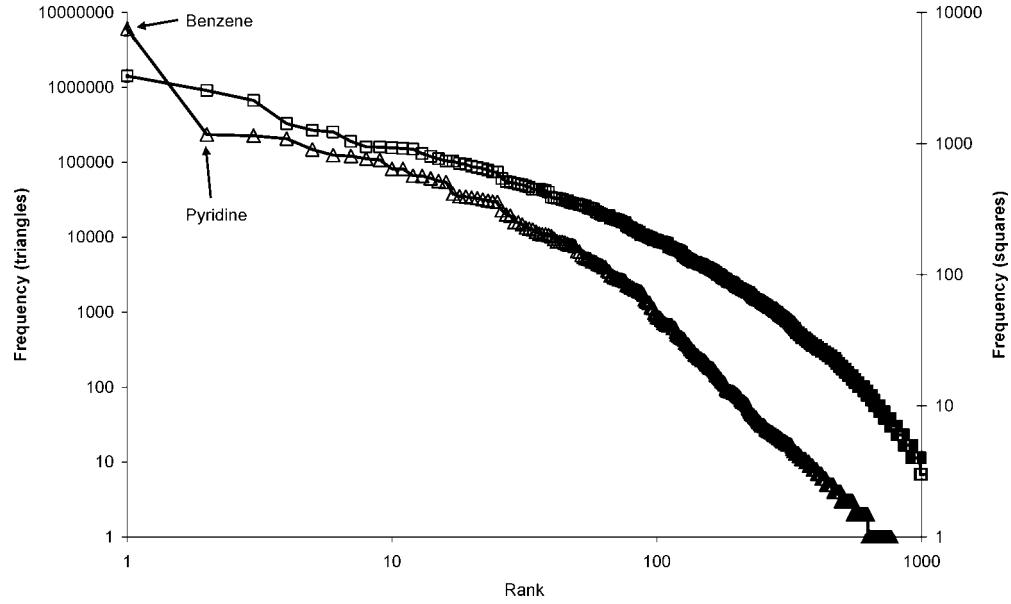

Figure 8. Zipf plot of the frequency of occurrence of ring systems plotted in rank order on a log−log scale. Triangles for the frequency of occurrence in the original compound data set; squares for the number of hits in the Beilstein database (only ring systems with less than 100 substructure hits in the original data set).

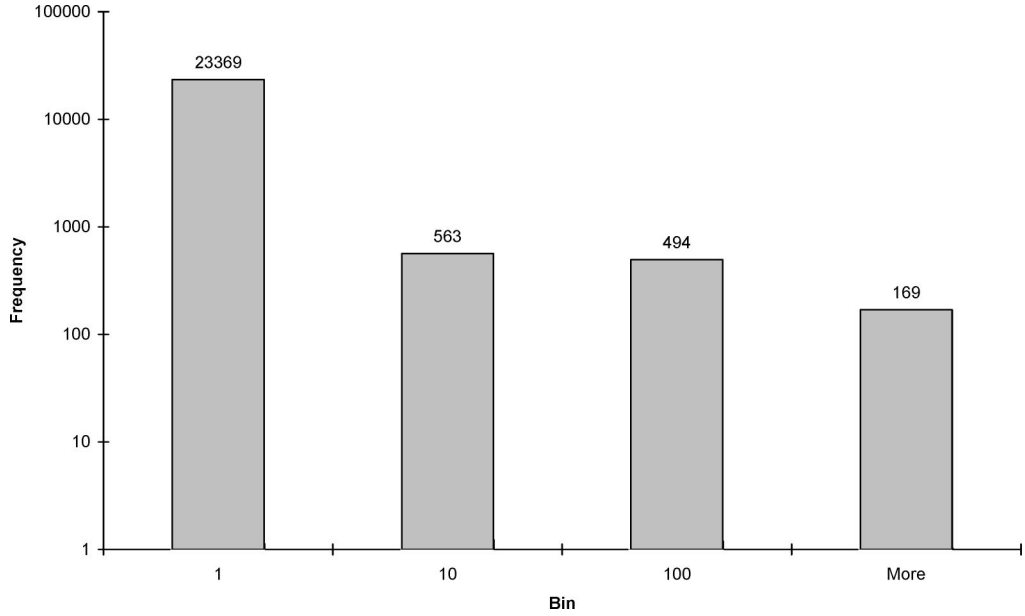

Figure 9. Distribution of the number of hits (x) found in the Beilstein database for the sample of 24098 VEHICLe ring systems with less than 100 substructure hits in the original compound data sets.

In summary, we have presented the enumeration and analysis of all possible aromatic monocyclic and bicyclic ring systems (within certain restrictions). After extensive searches of compound and literature databases, we find that only 1701 out of a possible 24867 have ever been made and published. Of these known ring systems, very few are routinely used when constructing drug-like molecules. This finding is in agreement with the work of Ertl et al.[21] The fewer heteroatoms they contain the more likely they are to be utilized, a fact also highlighted by Lipkus et al.[29] Could it be that the "principal of least effort" leads to a very low level of diversity or information entropy or is it due to restrictions derived from the requirements of biological compatibility? Alternatively, the predominance of inhibitors that are designed to be competitive with endogenous small molecule ligands of a small number of gene families could have led to this lack of diversity. This idea is in keeping with finding by Ertl et al.[21] that ring systems from bioactive molecules cluster together in a small number of islands in property space. In the future, it is possible that different drug discovery strategies will predominate such as targeting protein−protein interactions. If this is the case, the requirements for novelty, potency, and drug-like properties may lead to broader usage patterns.

We used knowledge of the substructure building blocks of the known ring systems to predict the possibility that over 3000 novel ring systems could be synthesized in the future. A subset of about 200 low complexity examples are predicted to be more likely to appear in the literature in the near future. However, we also estimate that the current publication rate of novel small heteroaromatic ring systems is between 5 and 10 per year and falling. We hope that the publication of this work will help extend the horizons of chemists into unexplored regions of

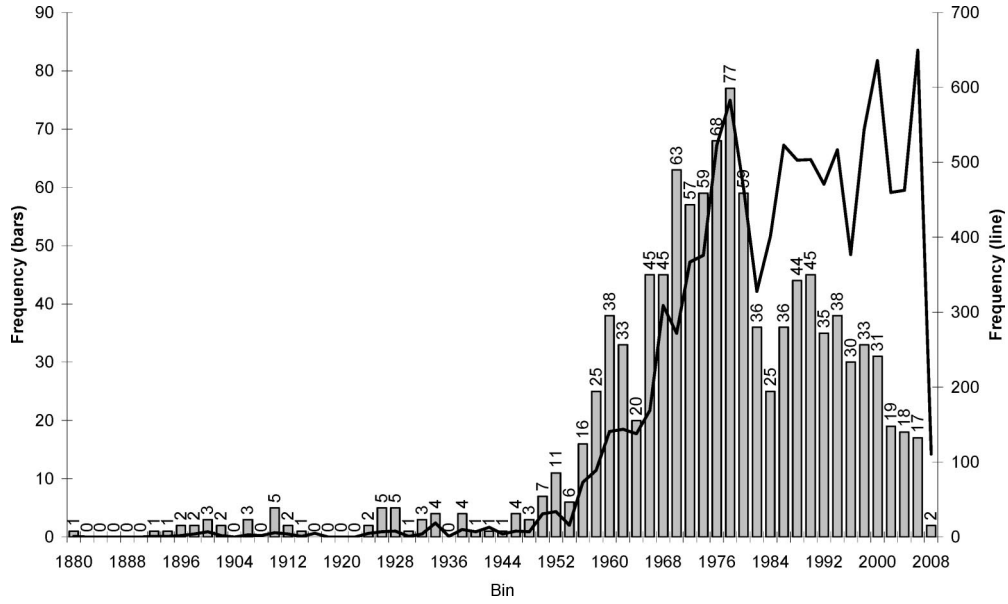

Figure 10. First publication year for VEHICLe ring systems substructure hits in the Beilstein [ref] database for a sample of 963 ring systems[1] (bars). Dates of publication of all the Beilstein hits containing one of the 963 selected ring systems (line).

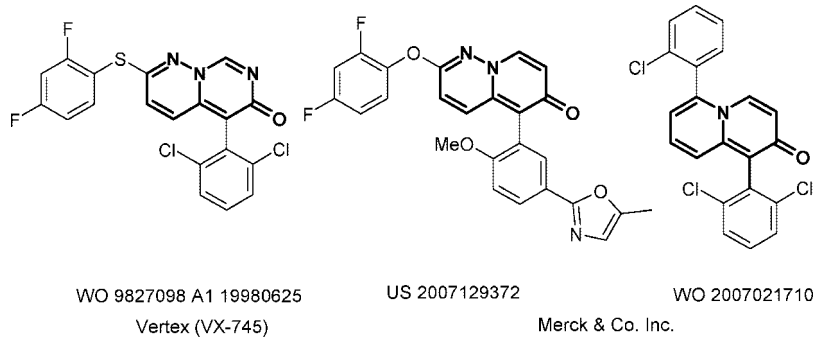

Figure 11. Examples of recently published for the first time VEHICLe ring systems that have been incorporated into drug molecules.

chemical space. Furthermore, we challenge organic chemists everywhere by highlighting some apparently simple and interesting ring systems that, to the best of our knowledge, have yet to be exemplified but should be within our capabilities to create.

Methods. To focus this work on chemical structures of interest to the pharmaceutical industry, VEHICLe was designed to contain all possible heteroaromatic rings systems with the following constraints:

(a) Only mono and bicyclic rings,
(b) Only 5 and 6 membered rings and combinations thereof,
(c) Only containing C, N, O, S and H,
(c) All neutral,
(d) All obey Hückel's 4n+2 rule of aromaticity,
(e) Only exocyclic carbonyls,
(f) Further constraints are implicit in the construction method (see below).

Tautomeric forms are treated as separate ring systems. This was done to facilitate database substructure searching because the tautomeric nature of an aromatic ring system can be lost when they are part of a larger molecule. Also, some database search engines will only return an exact match so that separate tautomers have be searched for separately. Our virtual library should have significant overlap with that produced by Ertl et al.,[21] but there are some important differences from their approach. First, they include tricyclic structures, whereas we do not. Second, they place restrictions on the usage of certain chemical building blocks. It was important for us not to do this during the construction phase of the library because we aimed to assess the usefulness of such filters by analyses of the library contents.

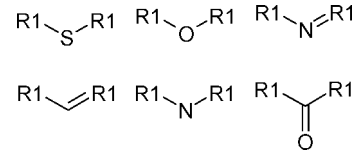

Figure 12. Building blocks.

VEHICLe Construction. The overall schema for the construction of VEHICLe is very simple. First, six atomic building blocks are combined into all possible chains of length 5 and 6. These chains are then closed into all combinations of five- and six-membered rings from which the aromatic rings are stored to form the monocycle set of VEHICLe. All rings are then fused into every combination of bicycles, of which the aromatic ones are stored. These molecular structure manipulations are carried out in Pipeline Pilot (Accelrys Software Inc.) using the Enumerate Combinatorial Reaction component.

The six different building blocks that were the basis for building up the ring systems in VEHICLe are shown in Figure 12. They are input into the pipeline as separate MDL mol files.[33] These atom/bond order building blocks, together with the rules used for chain formation, ensure neutral and stable valences.

These six building blocks built into all compatible combinations of chains of 5 and 6 atoms using rules encoded as

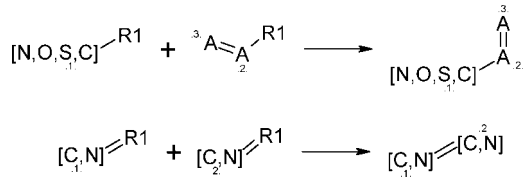

Figure 13. Chain forming rules.

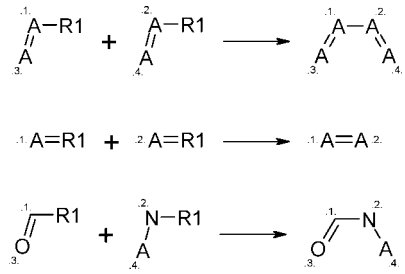

Figure 14. Ring closure transformations.

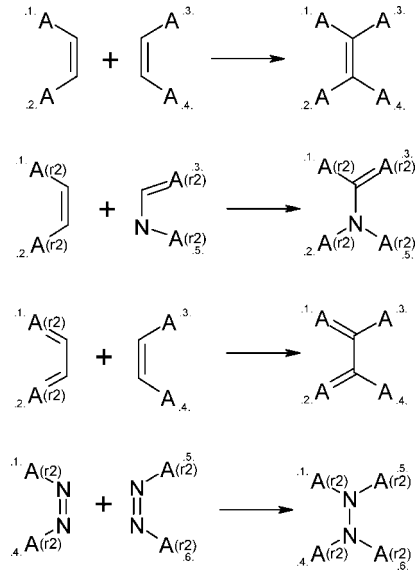

Figure 15. Ring fusion rules. Note that the fourth fusion rule could be omitted, as very few known ring systems contain a double nitrogen bridge of this sort.

"reactions" in MDL REACCS rxn format files;[33] however, these reactions have no real synthetic meaning. These rules are shown in Figure 13, one for a single and one for a double-bond formation. Two atom chains are reacted further with the single atom building blocks to make three atom chains and two atom and three atom chains combined to give chains of the desired length. At each stage, duplicate molecules are removed by comparison of their canonical SMILES strings (Daylight Chemical Information Systems Inc.). No limitation was put upon the number of times a building block could be used.

These chains are then closed into rings using the rules shown in Figure 14. At this stage, aromatic rings are identified by using the standard method implemented in Pipeline Pilot (Accelrys Software Inc.) (Num_AromaticRings = 1) and a further set identified that obey Hückel's $4n + 2$ rule. The Pipeline Pilot aromatized rings are dearomatized so that the ring fusion procedure works properly. This was achieved by the insertion of a dummy atom into the ring (we use polonium; anything is possible in virtual chemistry!) to be removed later. This procedure ensures that single and double bonds are fixed in position in an obvious way.

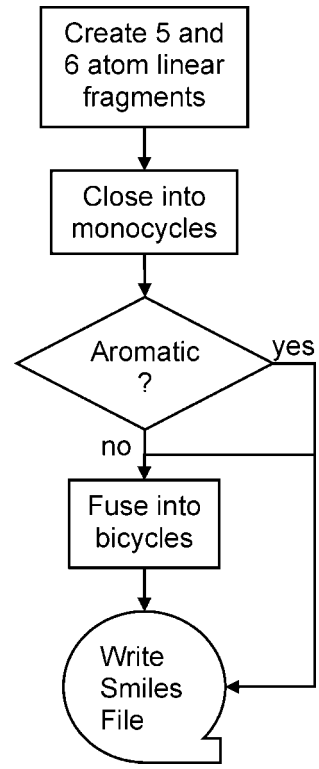

Figure 16. Flowchart of how VEHICLe was constructed in Scitegic Pipeline Pilot.

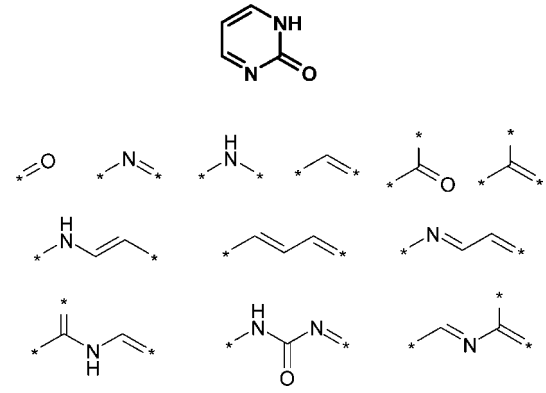

Figure 17. Illustration the fragments recorded by the ECFP_2 fingerprint (Accelrys Software Inc.) for the ring system showing in the top center.

Both aromatic and antiaromatic rings are then fused together by using the rules shown in Figure 15. The resulting bicycles have all polonium atoms excised, tested for aromaticity again, and the unique aromatic examples stored in VEHICLe. A flowchart representation of the top-level pipeline is shown in Figure 16.

Prediction of Synthetic Tractability. Largely due to fact that no limitation was put on the number of times a particular atom type could occur in a single ring system, many structures in VEHICLe are outlandish and would obviously be either very difficult or impossible to make. For example, many were constructed almost entirely of nitrogens. It was felt that useful synthetic tractability rules might be derived by using a statistical discrimination technique.

VEHICLe was classified into two sets: those ring systems that are substructures of known molecules, the "good" set, and the remainder as the "bad" set. A molecule was considered good (i.e., known) if it occurred in a collection of over three million

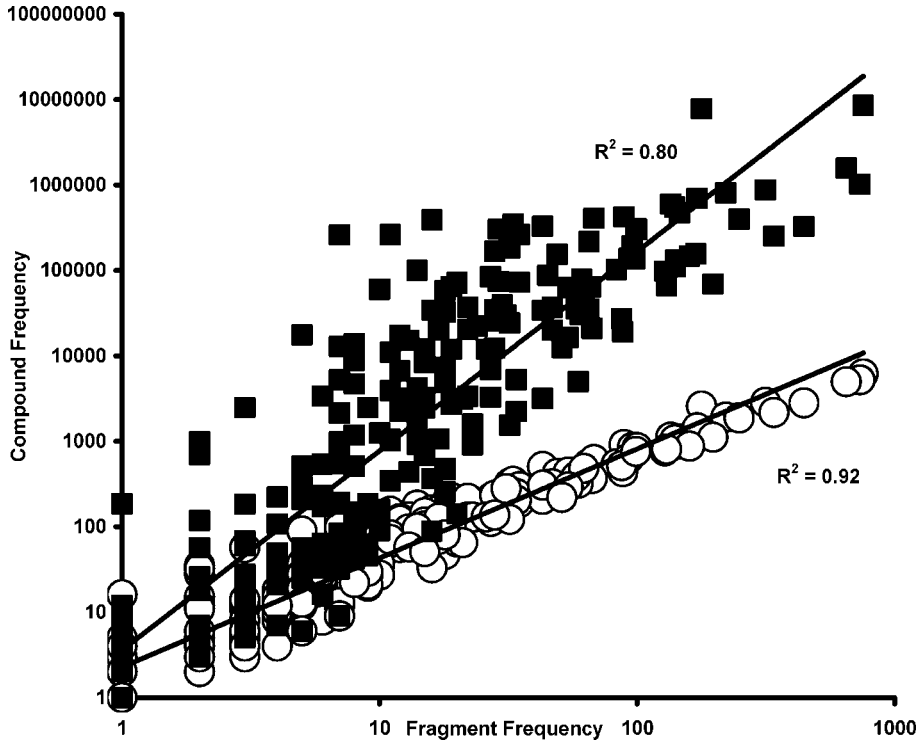

Figure 18. ECFP_2 fragment frequency of occurrence in available ring systems is correlated with the number of available compounds and the number of different compound data sets. The number of available "good" VEHICLe ring systems per fragment plotted against the number of sources of those ring systems (filled squares) and plotted against the number of compounds containing those ring systems (circles) (log−log plot).

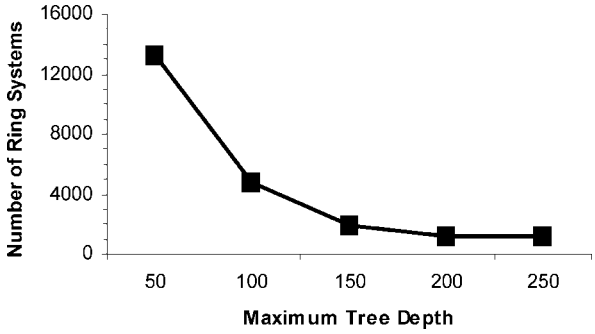

Figure 19. Number of VEHICLe ring systems not present in the literature training set that were classified as "good" by the decision tree as a function of maximum tree depth.

known molecules, collated in 2006. These molecules included commercially available compounds, drawn from 25 vendors screening and building block collections plus the contents of the ChemACX (CambridgeSoft Corporation) database. A set of about 40000 example structures from medicinal chemistry patents, collected from the Derwent World Drug Alerts Plus database (Thomson Scientific) 2001−2006, and about 3000 molecules labeled as being in clinical trials or "launched" drugs extracted from the MACCS-II Drug Data Report (MDDR) (Symyx Technologies) were also added. The combined compound data set contained 769 different VEHICLe rings: i.e., the "good" set, leaving 24098 in the "bad" set. Given the small size of this ratio, the assumption was made that the vast majority of these unknown/bad ring systems are difficult or impossible to make using existing techniques. In addition, the presence or absence of certain indicator constituent fragments might also be a good guide to synthetic tractability. For instance, certain atomic configurations may well be unstable or no known synthesis may have been developed to enable their construction. Conversely, other configurations might be commonly utilized due to the availability of certain starting materials or ease of construction, such as nonbridgehead N−N bonds or N−O bonds installed synthetically through the use of hydrazine or hydroxylamine condensation, respectively.

Scitegic ECFP_2 circular fingerprints (Accelrys Software Inc.) were used as descriptors of the ring systems. In this way, the presence or absence of each type of small substructure fragment was recorded for each ring system. The type of fragments recorded is illustrated in Figure 17. There are 346 unique fragments of this type in the VEHICLe ring systems.

To try and divide off the subset of the VEHICLe ring systems that have not yet been synthesized but may be in the future, a decision tree method was employed. Specifically, the NovoD ArborPharm (NovoDynamics, Inc.) decision tree software, as implemented in Scitegic Pipeline Pilot, was used. Various splitting methods are available for tree construction. The method that was chosen as best suited to the task in hand was the Buja pure bucket split method.[34] This method was developed to "cherry pick" interesting subsets of the data instead of trying to model the whole data set equally well. Interesting data sets in our case could be, for instance, large sets of ring systems that contain a certain fragment fingerprint that has never or rarely occurred in a known ring system compared to the unknown/"bad" set. Pure subsets of this sort are split off the tree one at a time leading to long, thin (unbalanced) trees. Buja and Lee find these sorts of trees to be "highly expressive and interpretable" and therefore well suited for data mining. Here we employ decision tree methodology not only for data mining but also for prediction. The "EnrichmentThreshold" parameter was set to 0.2, meaning that if at least 20% of the molecules in a node are classified as "good", then the whole node is classified as such. The "GoodBias" was set to 32 (the ratio of "bad" to "good" examples in VEHICLe, i.e., 24867/769); this variable

is the bias applied to the correct prediction of observations in the good class compared to the correct prediction of nodes in the bad class. This was done because the aim was to have a very low false negative rate. The false positive rate in this study is harder to assess because it is assumed that not all synthetically tractable ring systems have already been made and published. A very high GoodBias parameter was found to have detrimental effect on the variability between trees in the random forest (see below), resulting in a loss of the fine grain uncertainty values produced. The predicted uncertainty is expressed as *p*(good) values. If the GoodBias is set too high, these values come out as either 0.0 or 1.0 with nothing in between. The maximum number of levels was varied from 50 to 250 in steps of 50 and a maximum of 200 was chosen for all further studies (see below). A minimum of 10 molecules per node of the tree (bucket size), representing 0.04% of the full data set, was also specified.

The frequency of usage of each ring system within the training data set might possibly be used as additional training data. However, it was found that this frequency is positively correlated with number of times an ECFP_2 fragment occurs within the data set (see Figure 18), i.e., the more common ring systems tend to be constructed from the more common constituent fragments. Thus ring usage frequency is already implicitly encoded within the input data as it stands. Incidentally, this information provides evidence that the presence/absence of such fragments could be used to provide a good indication of the synthetic tractability of a given ring system.

Varying the Maximum Tree Depth. Before proceeding with the final training of the complete random forest, the effect of changing the maximum tree depth was tested on a single tree. This parameter was systematically varied from 50 to 250 in steps of 50. In all cases, ring systems that were labeled as "good" in the training set were correctly classified, i.e., there were no false negatives. Presumably this was achieved by setting the GoodBias parameter to 32. Of the library members that that were labeled as "bad", the number of these that were classified as "good" decreased with increasing tree depth up to a maximum depth of about 200 (see Figure 19). More and more ring systems were classified as "bad" as more and more fingerprint fragments were included in the classification until no more significant fingerprints are left. As stated above, our assumption was that most ring systems that could be made have already been made. These were included in the training data set as "good" ring systems. Thus, a low number of predicted "good" molecules (from the "bad" set) was deemed desirable. On the other hand, smaller trees are easier to interpret and less likely to produce random predictions. As a result of this analysis, a maximum tree depth of 200 was chosen for the training of the random forest.

For the final prediction, a forest of 100 trees was produced. Nodes near the bottom of the tree tend to be more arbitrary in the nature of the classification due to the small number of remaining data (minimum 10 data items per node) and the fact that the most discriminative descriptors are selected first. The random forest method can mitigate this effect to some degree by averaging. The *p*(good), which is the likelihood that particular ring system is in the "good" category is the calculated proportion of "good" training molecules in its leaf node. For a forest of trees, this is averaged over all trees where the ring system is included. To create 100 different trees, a small random selection of the training data is removed each time. The "PreserveMinority" parameter was set to "true" as the known synthesized molecules in the training set represent only 3% of VEHICLe. This meant that the training data selected for exclusion was chosen exclusively from the structures labeled as "bad". The power of the resulting predictions was tested using unseen data. This additional data on known ring systems was extracted from the Beilstein database (Beilstein GmbH) using Crossfire (Symyx Technologies).

Acknowledgment. Thanks to James Turner for collecting together the Derwent patent alerts and Alicia Higueruelo for helping with the database searches and for her encouragement throughout.

References

(1) Cayley, A. On the analytical forms called Trees, with application to the theory of chemical combinations. *Rep. Br. Assoc. Adv. Sci.* **1875**, *45*, 275–305.
(2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
(3) Weininger, D. Combinatorics of small molecular structures. In *Encyclopedia of Computational Chemistry, Vol. 1*; Schleyer, P. v. R., Schreiner, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H. F., III, Eds.; John Wiley & Sons, Ltd.: Chichester, 1998; pp 425−430.
(4) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.
(5) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: generating and searching 10(20) synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 341–350.
(6) Barnard, J.; Downs, G.; Scholley-Pfab, A.; Brown, R. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graph. Model.* **2000**, *18*, 452–463.
(7) Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: leads by design. *Drug Discovery Today* **2005**, *10*, 987–992.
(8) Rotstein, S. H.; Murcko, M. A. GroupBuild: a fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700–1710.
(9) Rarey, M.; Kramer, B.; Lengauer, T. Time-efficient docking of flexible ligands into active sites of proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 300–308.
(10) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
(11) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504–1508.
(12) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *32*, 157–166.
(13) Cho, S.; Zheng, W.; Tropsha, A. Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259–268.
(14) Ho, C. M.; Marshall, G. R. DBMAKER: a set of programs to generate three-dimensional databases based upon user-specified criteria. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 65–86.
(15) Clark, D.; Firth, M.; Murray, C. MOLMAKER: de novo generation of 3D databases for use in drug design. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 137–145.
(16) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087.
(17) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982–7989.
(18) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.
(19) Larsen, R. Palladium catalysis in the synthesis of medicinal agents. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 651–667.
(20) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; McLay, I. M.; Bradshaw, J. Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46*, 3257–3274.
(21) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the rings. In silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573.

(22) Hopkins, A.; Groom, C. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.
(23) Overington, J.; Al Lazikani, B.; Hopkins, A. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.
(24) Li, X.; Xu, H.; Tachdjian, C.; Werner, S.; Zhang, F.; Zlotnik, A.; Zoller, M.; Klebansky, B.; Fine, R.; Kang, X.; Patron, A. Modulation of chemosensory receptors and heterobicyclic ligands associated therewith and their preparation. Patent WO20080306076, 2008.
(25) Tsikouris, O.; Bartl, T.; Tousek, J.; Lougiakis, N.; Tite, T.; Marakos, P.; Pouli, N.; Mikros, E.; Marek, R. NMR study of 5-substituted pyrazolo[3,4-*c*]pyridine derivatives. *Magn. Reson. Chem.* **2008**, *46*, 643–649.
(26) Sparey, T.; Abeywickrema, P.; Almond, S.; Brandon, N.; Byrne, N.; Campbell, A.; Hutson, P. H.; Jacobson, M.; Jones, B.; Munshi, S.; Pascarella, D.; Pike, A.; Prasad, G. S.; Sachs, N.; Sakatis, M.; Sardana, V.; Venkatraman, S.; Young, M. B. The discovery of fused pyrrole carboxylic acids as novel, potent D-amino acid oxidase (DAO) inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3386–3391.
(27) Zipf, G. K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley Press: Cambridge, MA, 1949.
(28) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(29) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., III.; Schenck, R. J.; Trippe, A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443–4451.
(30) Benz, R. W.; Swamidass, S. J.; Baldi, P. Discovery of power-laws in chemical space. *J. Chem. Inf. Model.* **2008**, *48*, 1138–1151.
(31) Sneader, W. *Drug Discovery: The Evolution of Modern Medicines*; John Wiley & Sons Inc.: New York, 1985.
(32) Sneader, W. *Drug Prototypes and Their Exploitation*; John Wiley & Sons Ltd.: Chichester, 1996.
(33) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.
(34) Buja, A.; Lee, Y.-S. Data mining criteria for tree-based regression and classification. In *KDDD-2001: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: New York, 2001; pp *27*−*36*.

JM801513Z