JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

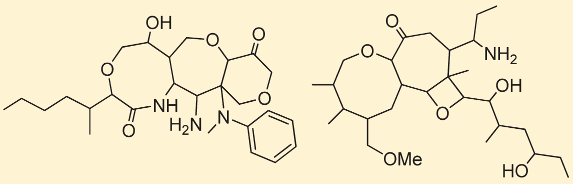
pubs.acs.org/jcim

Natural Product-Like Virtual Libraries: Recursive Atom-Based Enumeration

Melvin J. Yu

Eisai, Inc. 4 Corporate Drive, Andover, Massachusetts 01810, United States

S Supporting Information



ABSTRACT: A new molecular enumerator is described that allows chemically and architecturally diverse sets of natural product-like and drug-like structures to be generated from a core structure as simple as a single carbon atom or as complex as a polycyclic ring system. Integrated with a rudimentary machine-learning algorithm, the enumerator has the ability to assemble biased virtual libraries enriched in compounds predicted to meet target criteria. The ability to dynamically generate relatively small focused libraries in a recursive manner could reduce the computational time and infrastructure necessary to construct and manage extremely large static libraries. Depending on enumeration conditions, natural product-like structures can be produced with a wide range of heterocyclic and alicyclic ring assemblies. Because natural products represent a proven source of validated structures for identifying and designing new drug candidates, mimicking the structural and topological diversity found in nature with a dynamic set of virtual natural product-like compounds may facilitate the creation of new ideas for novel, biologically relevant lead structures in areas of uncharted chemical space.

INTRODUCTION

Natural products from both marine and terrestrial sources play an important role in the discovery of new chemical entities used to treat a variety of human diseases.[1,2] Having been optimized over millennia for interaction with biological macromolecules (e.g., protein function modulation), secondary metabolites in particular represent a demonstrated source of both material and inspiration for identifying novel agents with useful pharmacological activity.[3,4] From these laboratories, for example, the structurally complex marine natural product halichondrin B (HB)[5,6] served as the starting point for a drug discovery program that ultimately led to eribulin (E7389, ER-086526, NSC 707389),[7−11] a totally synthetic macrocyclic ketone analogue that recently completed a phase III clinical trial for locally advanced or metastatic breast cancer.[12]

Following upon this discovery, a second generation effort was initiated to further explore the utility of halichondrin-based derivatives as anticancer agents. Although structurally simplified relative to the natural product that inspired it, eribulin nevertheless represents the most structurally complex drug or drug candidate to be prepared by total synthesis. As a result, each new analogue had to be carefully selected to address a specific structure−activity relationship (SAR) question before synthesis and biological evaluation. Given the practical limitations associated with molecules of this structural complexity, we considered a number of complementary approaches to help expand our understanding of the series, one of which was virtual compound library screening. However, to accomplish this, we needed access to virtual compounds that more closely resemble natural products in terms of their structural and topological diversity rather than what may be considered more traditional small molecules derived from commercially available fragments or starting materials. Although there are a number of molecular enumerators capable of assembling the latter, we are not aware of any that could be used to construct virtual natural product-like libraries of the type we envisioned.

To meet this need, we developed a new enumerator capable of rapidly generating chemically and architecturally diverse sets of virtual compounds in a recursive manner. While atom-based enumeration[13] and evolutionary algorithms[14−17] in the context of drug discovery have been described in the literature for some time, to our knowledge a successful merging of these approaches to generate natural product-like structures has not been reported.

Existing enumeration methods rely on a variety of techniques such as SMILES string mutation and crossover,[18] reaction vectors,[19] and extended valence signature matching.[20] For our particular application, however, we wanted to generate compounds built from a single carbon atom as the starting point. We envisioned such an approach might be modeled after the way natural products such as secondary metabolites arise in response to environmental stress. For example, if a carbon atom (or a collection of privileged core structures)[21,22] was randomly decorated with a variety of atoms or simple fragments that could either propagate or terminate a growing chain (or create new rings), then complex structures could arise from a relatively small set of standard building blocks. Molecules that pass through cutoff filters derived from one-dimensional molecular descriptors (e.g., MW or Lipinski's rules[23]) could then be submitted to a series of fitness functions (e.g., QSAR equation, CoMFA model, pharmacophore model, ADME

Received: May 25, 2010
Published: March 09, 2011

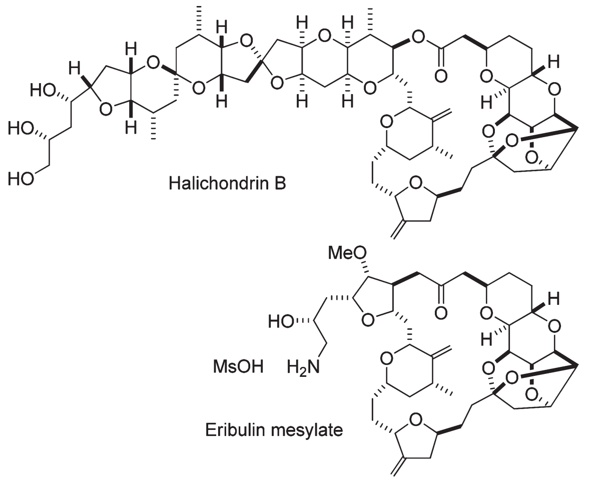


Figure 1. Chemical structure of halichondrin B and eribulin mesylate.

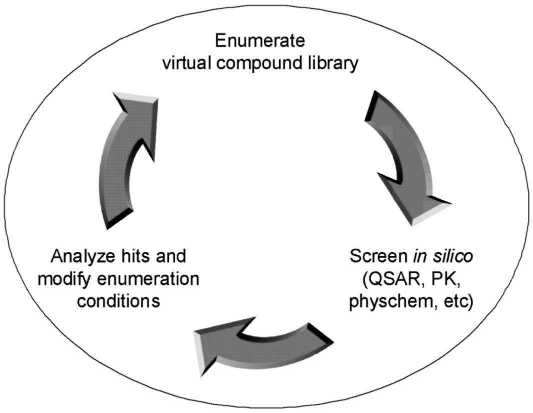

Enumerate virtual compound library
Screen *in silico* (QSAR, PK, physchem, etc)
Analyze hits and modify enumeration conditions

Figure 2. Recursive enumeration, screening, and analysis cycle.

filter, etc.), where those predicted to exhibit the desired biological, toxicological, or physicochemical properties would "survive." A machine-learning algorithm with or without expert user input could then, in principle, be used to weight the fragment probability selection process as it proceeds and thereby refine the enumeration output to generate molecules with increasing levels of interest (Figure 2). The underlying assumption is that the enumeration−survival process would eventually converge to a set of virtual molecules that would be suitable for synthesis and biological evaluation.

This manuscript describes a successful realization of this approach with a new atom-based molecular enumerator capable of generating natural product-like structures for in silico screening with possible application to scaffold hopping,[24] and de novo ligand design.[25]

METHODS

Enumeration Engine. An algorithm was devised that allowed either propagating or terminating fragments to be randomly selected and attached to a user-defined set of connection points (nodes) on the core structure. The core could be as simple as a single atom or as complex as a polycyclic ring system with one or more nodes. As fragments are added, the resulting chains could either continue to grow via propagation nodes, terminate via introduction of a terminating fragment, or cyclize to form a ring. The process would continue until either all nodes were filled or user-defined termination criteria were satisfied (e.g., maximum molecular weight was achieved, Figure 3).

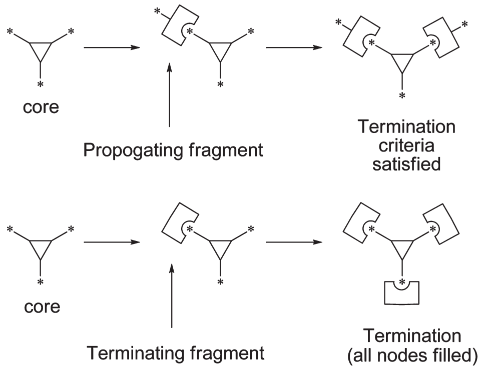

core
Propogating fragment
Termination criteria satisfied
core
Terminating fragment
Termination (all nodes filled)

Figure 3. Connection points for added fragments (nodes) are designated by an asterisk. Termination occurs when user-defined criteria are satisfied (e.g., MW achieved) or when all nodes are filled.

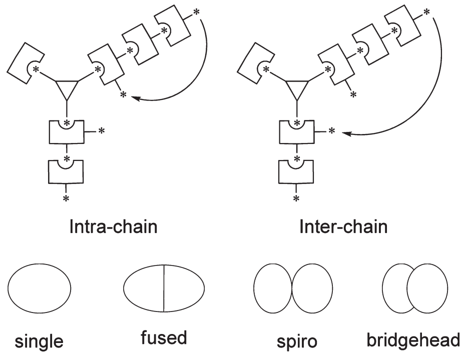


Figure 4. Possible ring forming modes.

Table 1. Canonical Connectivity Rules

1. If the added fragment is hydrogen − proceed.
2. If the added fragment node is heteroatom and flag indicates presence of adjacent heteroatom in the core/propagating node − do not add.
3. If the added fragment node is heteroatom and core/propagating node indicates heteroatom attachment not allowed − do not add.
4. If the core/propagating node is heteroatom and the node of the added fragment indicates heteroatom attachment not allowed − do not add.
5. If none of the above applies − proceed.

As depicted in Figure 4, ring structures could be formed by cyclizing two nodes within a growing chain (intrachain) or across two chains (interchain). In this manner, single, fused, bridgehead, or spirocyclic rings could be generated giving rise to a structurally and topologically diverse set of compounds.

To minimize formation of chemically invalid structures, the addition of heteroatoms are tracked within each molecule and flags are created indicating the presence of adjacent heteroatoms in a growing chain. In addition, a set of five simple canonical connectivity rules were created (Table 1). Used together, these two eliminate repeat heteroatom strings or unstable ketal chains from forming, while still allowing cyclic ketals to be generated (see Figure 5, for examples).

A current limitation to this approach is the inability to create unsaturation points de novo. However, this is easily overcome by simply including olefinic, carbonyl, or aromatic groups in the fragment set for incorporation in the enumeration process.

Although the enumeration algorithm was written as a standalone application, it is intended to be used in conjunction with commercial software (e.g., Pipeline Pilot[26]) for molecular property calculations (e.g., clogP, polar surface area, number of hydrogen bond donors/acceptors, etc). Because Pipeline Pilot is capable of removing duplicate structures, custom code to avoid duplicates during the enumeration process was deemed unnecessary. Thus, the raw enumerated structures were directly submitted to Pipeline Pilot. Duplicates (if any) were eliminated using the appropriate module, and the target number of unique structures (e.g., 50,000) was then passed to the fitness function for processing and subsequent fragment frequency analysis.

Fragment Library. As is the case for all molecular enumerators, the fragment library is key to the types of molecules to be generated. In the present study, only carbon, hydrogen, nitrogen, and oxygen atoms were considered. Other atoms such as chlorine, fluorine, phosphorus, or sulfur could be added by simply appending them to the standard fragment file as desired. There are two basic types of fragments, those that propagate a growing chain, and those that terminate. If designated, the propagation fragments can also form rings in either a donor or acceptor mode. A ring donor node is one that can initiate a ring by forming a bond with an acceptor node. In general, these follow chemical reactivity rules. For example, nitrogen and oxygen atoms can only be ring donors, whereas carbon can be either a donor or acceptor. With regard to polyatomic fragments (e.g., CH_2N), the propagation node need not be the same atom as the ring forming node. In addition to the node type, each fragment has an attendant flag that designates whether or not heteroatom attachment is allowed. Used in concert with the canonical connectivity rules (Table 1), these flags prevent invalid molecules and unstable polyheteroatom attachments from forming. Two example enumerations beginning with a four node carbon atom core structure are provided in Figure 6.

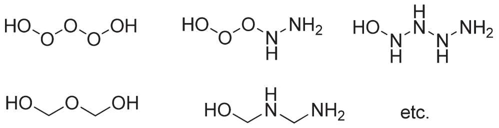


Figure 5. Representative invalid or unstable structures.

For this study, we selected a 21 member fragment set that included both atoms and simple polyatomic fragments (Table 2). The latter fragments were chosen to allow methyl, carbonyl, or phenyl groups to be readily incorporated in the enumerated molecules. We felt that these represent the fundamental building blocks needed to generate a sufficiently wide range of biologically relevant compounds appropriate for in silico screening. Additional fragments could be added to the basic set depending on the nature and specific needs of a project.

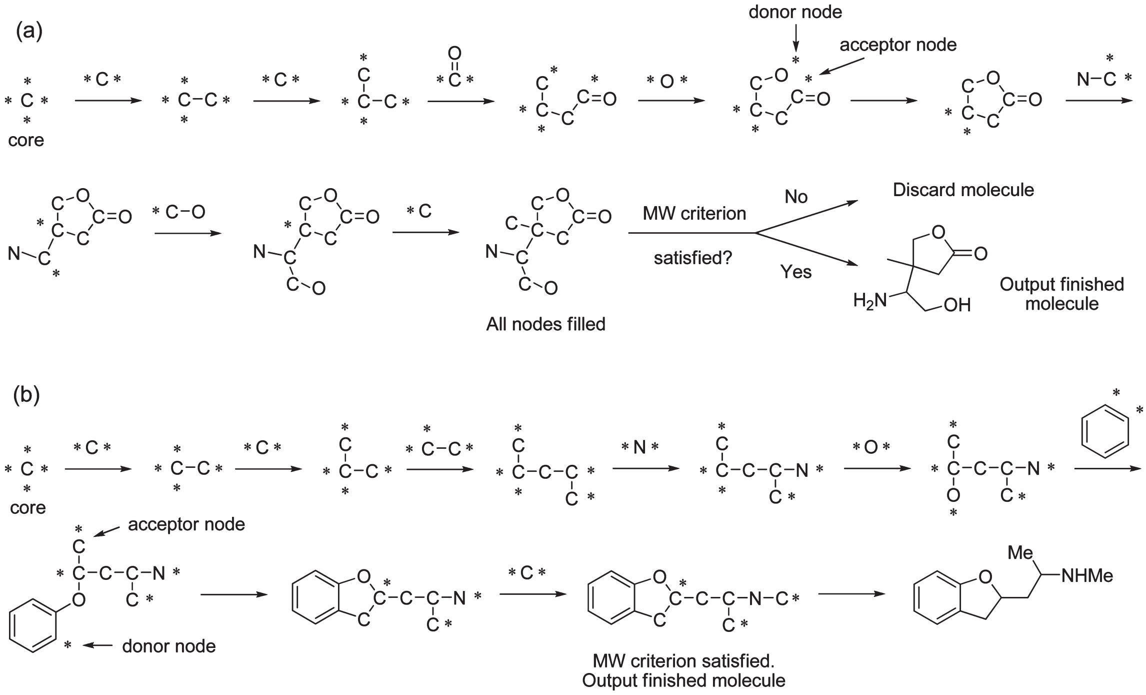


Figure 6. Two example enumerations beginning with a four node carbon atom as the core. Fragments, which can consist of a single or multiple atoms, are sequentially added in random fashion to nodes designated by an asterisk. The added fragments can either propagate a growing chain, cyclize to form a ring, or terminate depending on the added fragment (see Table 2 for fragment types). At each step (designated by an arrow) the enumerated compound is checked to determine if it meets the user-specified MW. If it falls within the desired MW range, then the growth process is terminated. The molecule is written to file, and construction of a new molecule is initiated beginning with the original core structure (in this case a four node carbon atom). If the MW criterion is not met, then the growth process is allowed to continue. If, however, all of the nodes become filled and the molecule can no longer accept any additional fragments, then the molecule is either accepted or rejected on the basis of its molecular weight. The enumeration process proceeds until the target number of virtual compounds is reached.

Table 2. Standard Fragment Library[a]

number	fragment	allow heteroatom attachment
1	$*C^{T}$	yes
2	$*C^{P}$	yes
3	$*C^{P,D}$	yes
4	$*C^{P,A}$	yes
5	$*C^{P,A,A}$	no
6	$*N^{T}$	no
7	$*N^{P}$	no
8	$*O^{T}$	no
9	$*O^{P}$	no
10	$CH_3N^{*,P}$	no
11	$^{P,*}CHN$	no
12	$^{P,*}CHN^{D}$	no
13	$^{A}CH_2N^{*,P}$	no
14	$^{P,*}CHO^{D}$	no
15	$^{D}CH_2O*$	no
16	$^{A}CH_2O*$	no
17	$^{A*}C{=}O$	yes
18	$^{P,*}CHC^{A}{=}O$	yes
19	$^{P,*}CHCH_2{}^{D}$	no
20	$^{P,*}CHCH_2{}^{A}$	no
21	phenyl ($*C1,C2^{D}$)	yes

[a] Note that a single atom can serve more than one function. It can be (1) the point of initial attachment in a growing chain, (2) the point of propagation, (3) a termination atom, or (4) a donor or acceptor ring forming node. *Location of initial attachment to growing chain. ^{P}Propagating node. ^{T}Terminating node. ^{D}Ring forming donor node. ^{A}Ring forming acceptor node.

The fragment selection and addition process proceeds in a Monte Carlo fashion. By default, the fragment probability weights are all assigned a value of 1, making the probability of selecting any particular fragment equal. Modifying the fragment weights will consequently focus the type of molecules enumerated (e.g., more oxygen atoms, more aromatic rings, fewer nitrogen atoms, etc), resulting in a biased compound set. By analyzing the fragment patterns associated with molecules of interest (i.e., compounds predicted to be active in a quantitative model or to exhibit desired physicochemical properties), the fragment probability weights can be adjusted through selective pressure for the next enumeration cycle. With each successive iteration the fragment probability weights change and should begin to converge to an optimal set of values.

The enumeration algorithm requires two input sd files (one containing the core structure and the other containing the fragments) and a text file of fragment probability weights. The program tracks the attached fragments for each node and writes the enumerated molecular and fragment information to an output sd file. The chemical structures can be cleaned using a Pipeline Pilot protocol for viewing or submitted directly for molecular descriptor calculation and in silico screening.

Fragment Analysis Algorithm. The current prototype fragment analysis algorithm operates in a rudimentary manner by simply counting how frequently a particular fragment was added to each node in the original core structure for compounds that both meet the user-specified molecular weight criterion and pass the fitness function. The frequency values are normalized and reported as an integer fragment probability weight (Figure 7). If a fragment did not appear at a specific node, then its weight is assigned a zero. For example, a weight of 100 signifies the probability of selecting that fragment is 100 times greater than selecting a fragment with a probability weight of one. The resulting weight array is then used directly for enumerating the next generation of compounds, after which the weight array is updated. An example of how fragment weights can change from the first generation to the fifth for a core structure consisting of a single carbon atom with four nodes appears in the Supporting Information. The enumeration and fragment weight calculation flowchart as a function of generation is illustrated in Figure 8.

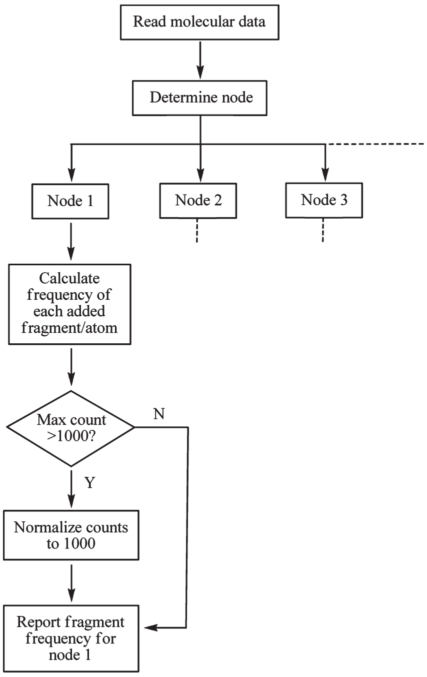


Figure 7. Fragment analysis and weighting algorithm.

For example, if three molecules were enumerated (see Table 3) from a four node carbon atom core with the fragments listed in Table 2, the resulting fragment weight file would simply be the number of times each fragment appeared as a function of node summed across all three molecules. In this particular case, fragment 1 appeared a total of two times at node 1 (once in molecule A and once in molecule B). Thus, its weight would be assigned a two. Fragment 2 appeared a total of three times at node 1 (once in molecule B and twice in molecule C). Thus, its weight would be assigned a three. This process would be repeated for all fragments across all four nodes to give the results summarized in Table 4.

The enumeration and fragment analysis algorithms were written in C and compiled using a GNU C compiler. Program execution is rapid with a typical run time of less than 30 s on a Windows laptop computer to fully enumerate and output 50,000 structures from a four node carbon atom core, the standard fragment set, default fragment weights, and a target MW of 100−300.

Drug-Likeness and Natural Product-Likeness Models. In silico models to calculate drug- and natural product-likeness scores for the enumerated molecules were constructed as follows. Compounds from the ACD and the MDL/SYMYX Drug Data Report (MDDR release 2007.2) databases were used as the nondrug and drug sets, respectively. Compounds from the DrugBank[27] not found in the MDDR were used as an external

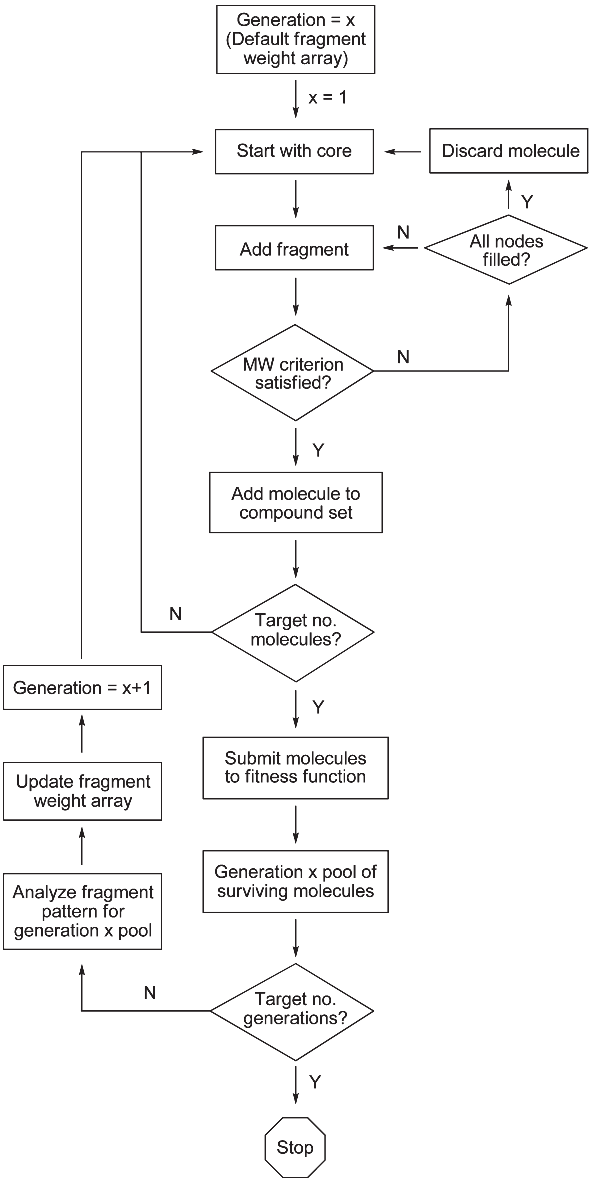


Figure 8. Enumeration and fragment weight analysis flowchart. For each generation, a specified number of virtual compounds (e.g., 50,000) with a desired MW (e.g., 200−500) are enumerated. The compounds are then passed to a fitness function. Compounds that "survive" comprise the pool of compounds for that generation. The fragment pattern for the surviving compounds in that generation is analyzed, and the fragment weight array is updated. The cycle is repeated until the target number of generations is achieved.

test set. Chemical structures from the Dictionary of Natural Products (DNP),[28] the MDDR, and ACD data sets were prepared as follows. (1) Blank entries were removed by requiring MW > 12. (2) The molecules were recursively deglycosylated using Pipeline Pilot. (3) Counterions, if any, were removed, and the residual charges were neutralized. (4) Duplicate entries and structures that contain metals (e.g., Pt) were removed. With respect to the MDDR data set, 120 compounds had a MW > 5,000. Because some of these caused problems for the fingerprint calculation, they were removed. Using this general procedure, the three data sets outlined in Table 5 were obtained.

Table 3. Three Enumerated Molecules with Various Fragments Added to a Four Node Carbon Atom Ccore

	molecule A				molecule B				molecule C			
node	1	2	3	4	1	2	3	4	1	2	3	4
fragment no.[a]	9	2	3	17	1	3	2	4	2	2	2	1
	1	1	4	2	2	1			2	3		
			2		6							

[a] Fragment numbers correspond to those listed in Table 2.

Table 4. Fragment Weight File Calculated Using the Three Molecules Listed in Table 3

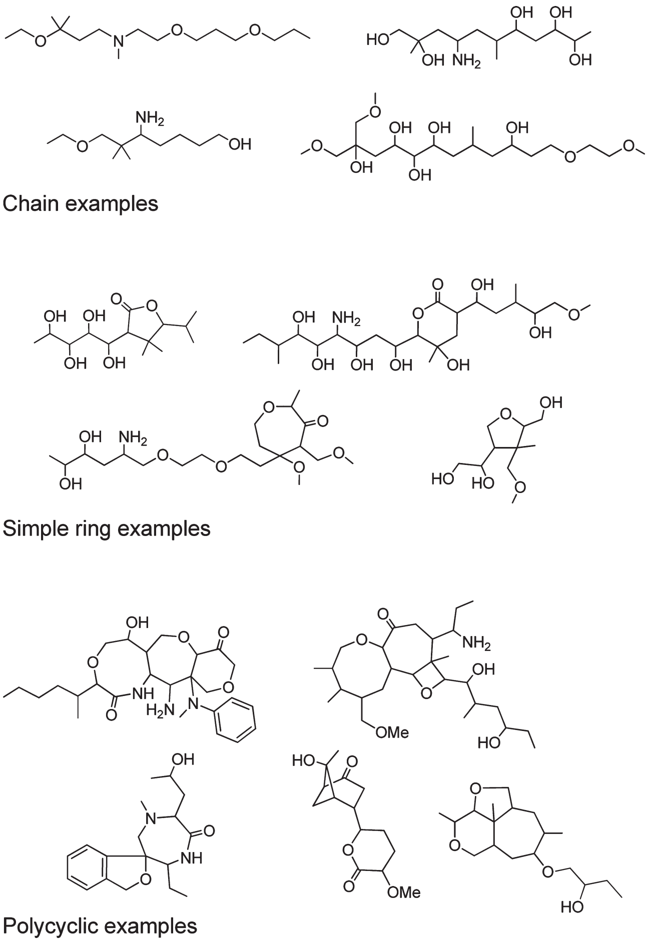
fragment no.	node 1	node 2	node 3	node 4
1	2	2	0	1
2	3	2	3	1
3	0	2	1	0
4	0	0	1	1
5	0	0	0	0
6	1	0	0	0
7	0	0	0	0
8	0	0	0	0
9	1	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	0	0	0	1
18	0	0	0	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0

Table 5. Data Sets Used to Train the Naïve Bayesian Classification Models

data set	original no. of entries	final no. of structures
ACD[a]	661,820	545,377
MDDR	180,784	157,023
DNP	230,107	141,613

[a] Final no. of structures represents the basic set before removal of either natural product or drug structures.

Prior to building the natural product-likeness model, we examined the ACD for natural product structures. Any that were common to both the DNP and ACD were removed from the latter. Two sets were then constructed as follows. First, the DNP data set was randomly divided into two halves. A matched number of randomly selected ACD structures was then added to each half to afford sets A and B consisting of 141,589 and 141,637 compounds, respectively. These two were then used to build the Bayesian natural product-likeness model using Pipeline Pilot.


Chain examples

Simple ring examples

Polycyclic examples

Figure 9. Example enumeration output from a four node carbon atom core and the standard fragment library. MW range = 200−500. Fragment weights were adjusted to minimize phenyl ring selection.

Construction of the drug-likeness model proceeded similarly. Compounds in the ACD common to either the MDDR or the DrugBank were removed from that database prior to use. The MDDR data set was randomly divided into two halves, and a matched number of randomly selected ACD structures were added to each half to afford sets C and D consisting of 157,248 and 156,798 compounds, respectively. These two were then used to build the Bayesian drug-likeness model using Pipeline Pilot.

■ RESULTS AND DISCUSSION

Depending on the desired output, the enumerator can generate chains, simple rings, more complex polycyclic structures, or a combination of all three. For example, by using a four node carbon atom as the core with the standard fragment library listed in Table 2, a MW threshold acceptance criterion of 200−500, and fragment weights adjusted to suppress phenyl ring selection, a structurally diverse array of virtual compounds was enumerated (see Figure 9 for examples). The distribution of oxygen to nitrogen atoms could be altered by simply modifying the fragment weights or the ratio of chains to rings could be adjusted by either engaging or disengaging the various ring forming toggles. It is interesting to note that under certain enumeration conditions, the output compounds visually resemble natural product-like structures in terms of their "steric complexity" as defined by Henkel and co-workers.[29]

In another example, 50,000 compounds were enumerated from a four node carbon atom and the standard fragment library using default fragment weights with a MW range of 100−300. The distribution of AlogP, number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds appears in Figure 10. Approximately half of the enumerated structures had a MW between 290 and 300, with the remainder distributed across the range of 110 to 290. This is likely a reflection of terminating fragments/atoms capping all of the nodes, thereby ending the growth process. The AlogP histogram exhibited a bell shape curve with a maximum at 1.0. Similarly, the number of hydrogen bond donors and acceptors plot also exhibited a bell shape curve with a maximum at 2 and 4, respectively. Interestingly, the number of rotatable bonds, which ranged from 0 to 13 for this compound set, was centered around 2. This relatively low number is due to a majority of cyclic structures being formed relative to the number of chains.

After altering the fragment weights to lower the number of nitrogen atoms, the same number of compounds was enumerated (see Table 6 for fragment weights). In this case, the AlogP distribution was shifted to the right with a maximum around 2.0−2.5. As expected, the number of rotatable bonds remained unchanged with a reduction in the number of hydrogen bond donor and acceptor groups (Figure 11). Thus, altering the fragment weights resulted in a biasing of the enumerated compound set in a predictable manner.

Limiting the number of ring substituents is easily accomplished by not allowing cyclization to occur either within a growing chain or across chains. Running the enumeration as described above, but with both intrachain and interchain ring options disengaged gave a set of 50,000 chain compounds with a range of rotatable bonds centered around 7 (Figure 12).

Initial Proof of Concept. Secondary metabolites can serve a number of important host functions depending on the nature and environment of the producing organism. In certain instances, for example, these may represent basic defense mechanisms against invaders where the secondary metabolites may exhibit varying degrees of toxicity. To test how selective pressure (e.g., either increasing or decreasing need for toxicity) would alter the predicted properties of enumerated compounds as a function of generation, we employed the protocol depicted in Figure 2. Because natural products have been proposed to target proteins more essential to an organism, i.e., for defense,[30] the initial fitness function we chose for illustration purposes and initial proof of concept is the TOPKAT rat oral LD_{50} toxicity prediction model integrated with Pipeline Pilot. We felt that this particular in silico model would appropriately demonstrate the approach because compounds predicted to be toxic must not only be pharmacologically active (in order to be toxic), but also orally bioavailable and therefore biologically relevant. In this example, virtual compounds that meet a specified predicted threshold value (used to define either toxic or nontoxic) would be considered "active". For the enumeration step, we used a carbon atom with four attachment nodes as the core with the standard 21 member fragment library listed in Table 2. To minimize the number of variables that might confound the analysis, the MW and number of enumerated compounds were fixed. For each generation, 50,000 nonduplicate structures with MW 299−301 were generated.

First, we examined the recursive enumeration process using selective pressure to generate compounds predicted to be *nontoxic*

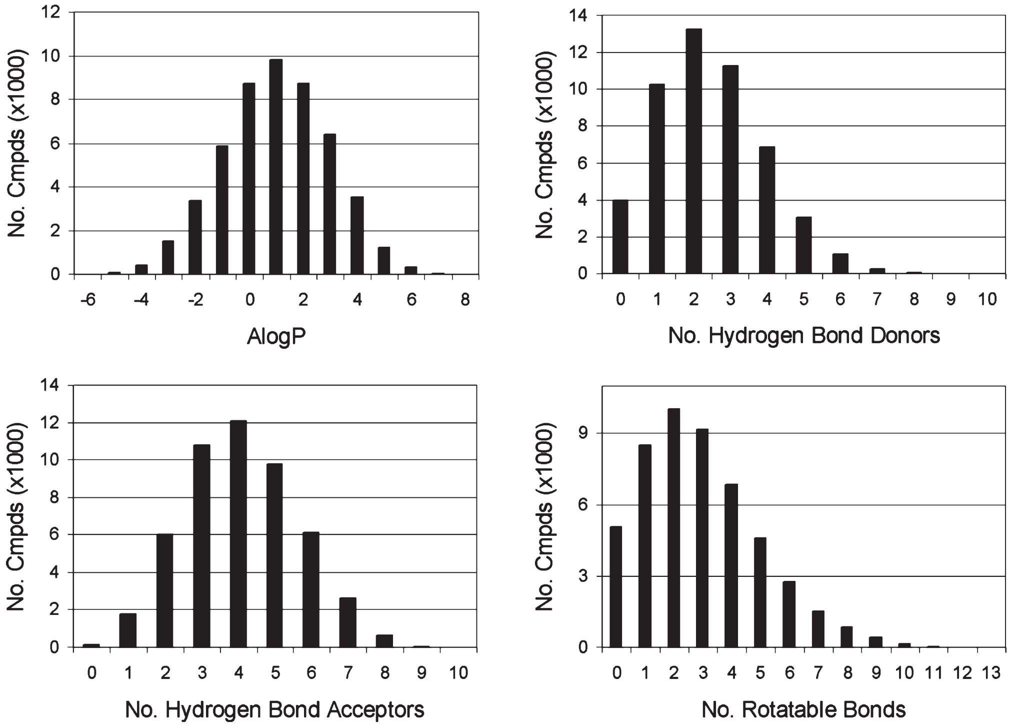


Figure 10. Histogram analysis for 50,000 virtual compounds enumerated from a four node carbon atom and the standard fragment library with default fragment weights. MW = 100−300.

Table 6. Modified Fragment Weights Used to Enumerate 50,000 Compounds with MW 100−300 Beginning with a Four Node Carbon Atom Core[a]

fragment	node 1	node 2	node 3	node 4
1	10	10	10	10
2	10	10	10	10
3	10	10	10	10
4	10	10	10	10
5	10	10	10	10
6	1	1	1	1
7	1	1	1	1
8	10	10	10	10
9	10	10	10	10
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	10	10	10	10
15	10	10	10	10
16	10	10	10	10
17	10	10	10	10
18	10	10	10	10
19	10	10	10	10
20	10	10	10	10
21	10	10	10	10

[a] See Table 2 for the identify of fragments 1−21.

(rat oral LD_{50} >10 g/kg). Those predicted to meet this criterion were flagged as "active" and the fragment pattern analyzed using the algorithm outlined in Figure 7. Fragment weights were adjusted accordingly, and the process was repeated for the next generation. At generation one using default fragment weights, 112 compounds (0.22% of total) were identified that met the predicted rat oral LD_{50} criterion of nontoxicity. After five generations, the number jumped over 20-fold to 2385 (4.8% of total), and after 10 generations, the number reached 2843 (5.7% of total). Plotting the number of compounds that meet the oral LD_{50} criterion by generation reveals a nonlinear relationship (Figure 13). It is interesting to note that the initial enumeration conditions generated a majority of compounds that were predicted by the TOPKAT model to be toxic below 1 g/kg body weight in rat. Although it is tempting to speculate, it is not clear why such a large percentage of virtual compounds randomly built from carbon, nitrogen, and oxygen atoms would be predicted to be toxic. However, modification of the fragment weights through selective pressure shifted the ratio to include a higher percentage of compounds predicted to be less toxic and distributed across a wider range of values.

The next step was to reverse the selective pressure direction and enumerate compounds predicted to be toxic with a rat oral LD_{50} value less than 0.05 g/kg. Beginning with generation one (default fragment weights), 190 compounds out of 50,000 (0.38%) were identified that satisfy the target oral LD_{50} criterion. By generation five that number was 2619 (5.2%), and by generation 10, the number of compounds reached 3766 (7.5%, Figure 14). Thus, even though the prototype fragment analysis algorithm operates at a rudimentary level, it was able to appropriately modify the fragment weights and thereby enrich the next generation of compounds with analogues that meet target threshold values. Chemical structures for the virtual compounds predicted by TOPKAT to be the 10 most "active" (i.e., toxic) from generation 10 appear in Figure 15.

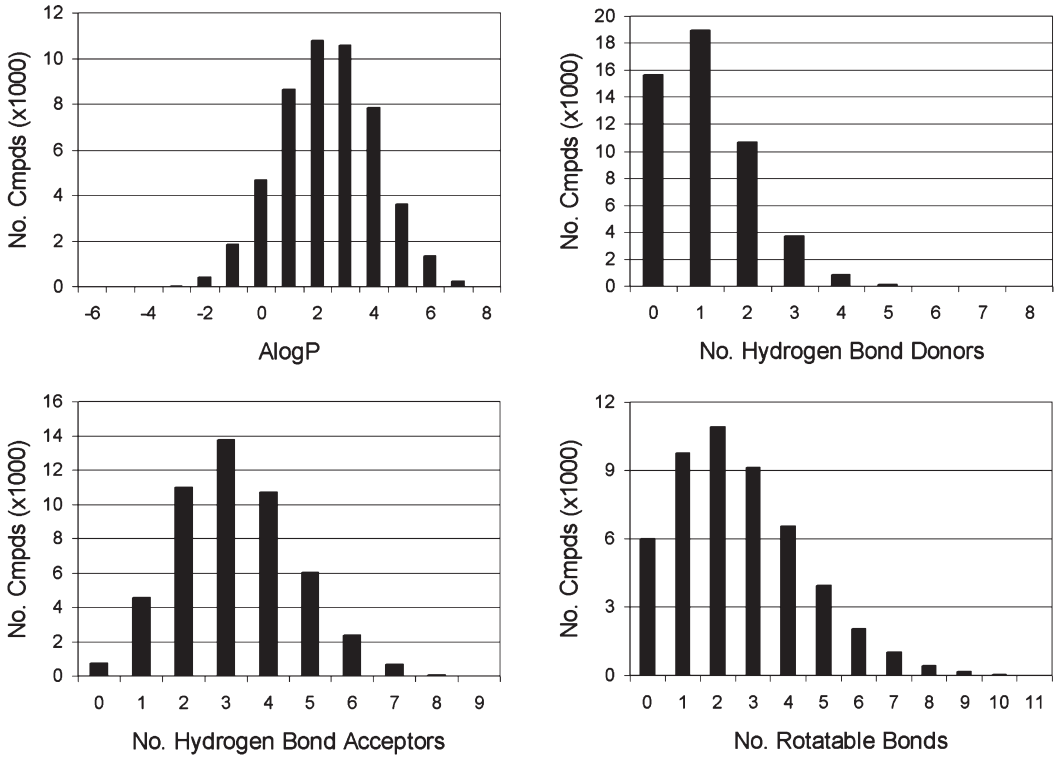


Figure 11. Histogram analysis for 50,000 virtual compounds enumerated using modified fragment weights.

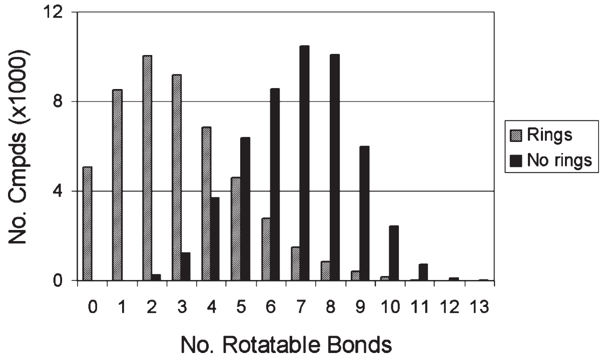


Figure 12. Number of rotatable bonds for 50,000 enumerated compounds with and without ring option engaged.

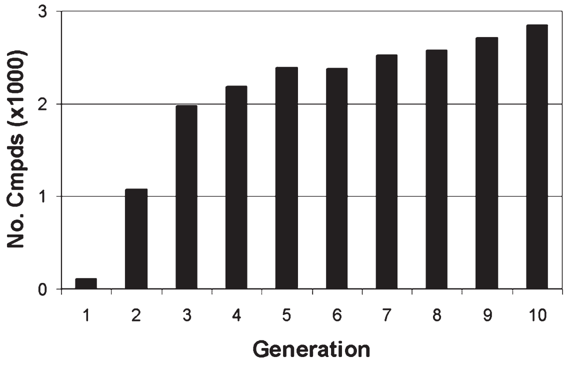


Figure 13. Number of virtual compounds (from 50,000 enumerated structures) predicted to be nontoxic (LD_{50} >10 g/kg body weight) as a function of generation.

Of particular interest are virtual compounds **1** and **2**, whose basic 6,7-benzomorphan skeleton is not only a substructure of morphine (Figure 16), but is also common to a number of clinically used analgesic drugs (e.g., pentazocine and phenazocine). Given this observation, we examined generations 1–9 for appearance of similar structures and found that the number of virtual 6,7-benzomorphan derivatives that meet the predicted target oral LD_{50} criterion (<0.05 mg/kg) steadily rose from 3 to 28 as a function of generation (Figure 17). Thus, the recursive process of enumerating a single carbon atom using the standard set of fragments gave rise to virtual compound libraries increasingly enriched in a particular class of agents known to be pharmacologically active.

A structurally related series of tricyclic benzazepine derivatives (e.g., **11**) possessing the same scaffold as the marine natural product aphanorphine[31] was also found among the enumerated compounds (Figure 16). Like the 6,7-benzomorphan analogues, the number of virtual tricyclic benzazepine derivatives predicted to be "active" in each generation increased as the recursive process proceeded. For generation one there were only two, but by generation 10 that number grew to 24 (Figure 17).

Other benzomorphan related structures found among the "actives" from generation 10 include virtual compounds **12**–**15** (Figure 18). The ring systems for all of these are known, with the parent structures **17**[32] and **19**[33] reported to exhibit various levels of analgetic activity in vivo. In addition, the N-methyl derivative of **17** was described as being "relatively toxic".[33] While derivatives of **18** are claimed to be devoid of antinociceptive activity, several have been reported to be "very toxic" in mouse pharmacological studies.[34] Derivatives of B-norbenzomorphan (e.g., **20**) are also reported in the literature to be biologically active with anticholinesterase activity

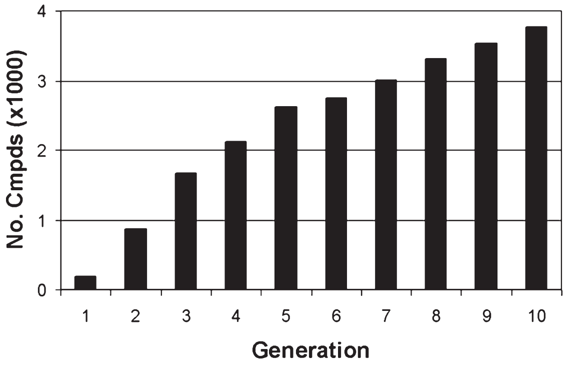


Figure 14. Number of virtual compounds (from 50,000 enumerated structures) predicted to be toxic ($LD_{50} < 0.05$ g/kg body weight).

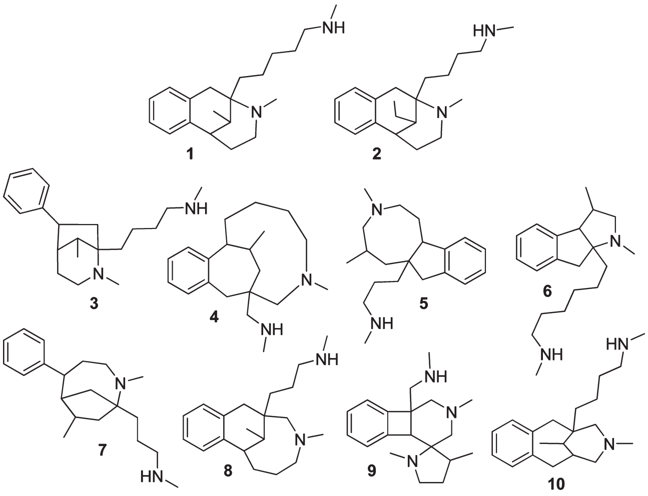


Figure 15. Top 10 virtual compounds from generation 10 predicted by TOPKAT to be the most "active" (i.e., toxic) from 50,000 enumerated structures (predicted rat oral LD_{50} = 6−8 mg/kg).

demonstrated both in vitro and in vivo.[35] In addition, the parent 2-azabenzonorbornene skeleton **16** found in virtual compound **12** has been studied as a conformationally constrained benzylamine adrenergic agent.[36]

The fused hexahydroindeno[2,1-b]pyrrole ring system found in structure **6** (see Figures 15 and 19) is also biologically validated with derivatives claimed in the patent literature as NMDA[37] and neuronal calcium channel antagonists.[38] In addition, a series of compounds containing this skeleton was studied as congeners of the natural alkaloid physostigmine in the context of acetylcholinesterase inhibition.[39] Thus, the molecular enumerator is capable of generating biologically validated structures with the potential for scaffold hopping across ring systems.

Also enumerated are scaffolds that to our knowledge have not yet been reported in the literature (e.g., **23** from compound **3**, Figure 20). Consequently, the molecular enumerator can potentially be used to suggest new templates for creating novel lead structures by the addition of appropriate functionality as guided by the fitness function output.

As mentioned earlier, the TOPKAT model was employed as the fitness function for initial proof of concept because many natural products are believed to function as defense agents. Use of more project specific in silico models (e.g., protein−ligand interactions, pharmacokinetics, and physicochemical properties) as the fitness function would be expected to provide a more refined set of structures for consideration by drug discovery scientists during the lead identification process.

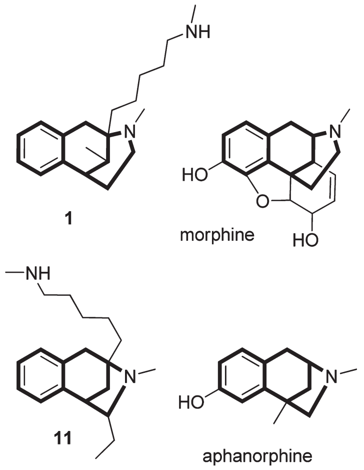


Figure 16. Substructure comparison between representative virtual compounds **1** and **11** from generation 10 and the natural products morphine and aphanorphine.

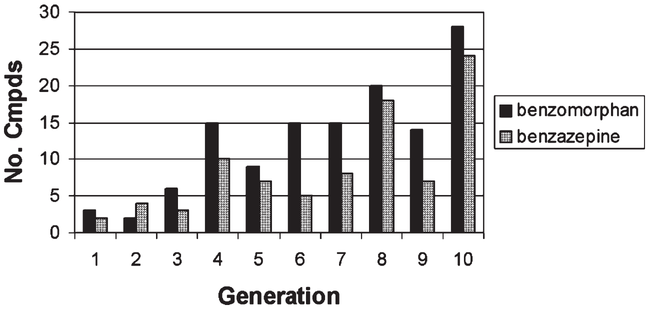


Figure 17. Number of virtual 6,7-benzomorphan and tricyclic benzazepine derivatives predicted to meet the target oral LD_{50} criterion (<0.05 mg/kg) as a function of generation.

Second Proof of Concept. For a second proof of concept, we used the same general experimental conditions as described above, but with a series of two fitness functions to optimize the enumerated compounds for natural product-likeness (NP-likeness) and drug-likeness. To accomplish this, we built two naïve Bayesian classification models for implementation in Pipeline Pilot. The first was used to define NP-likeness and the second to define drug-likeness.

Collections of natural products and "synthetic" molecules can be differentiated on the basis of molecular property distributions and structural characteristics.[29,40−42] For example, Ertl and coworkers reported a natural product-likeness scoring algorithm that could be used to help prioritize compound libraries.[43] Their method used atom centered fragments as HOSE codes[44] to characterize structural features and is in principle an application of a naïve Bayesian statistic. Because we did not have access to those particular codes, we explored whether other fingerprint methods could be used as an alternative. Using Pipeline Pilot, the molecular features of each molecule were represented by their atom environment fingerprints FEFP_2 and FEFP_6;[45] extended-connectivity fingerprints ECFP_2, ECFP_6, FCFP_2, FCFP_6;[46]

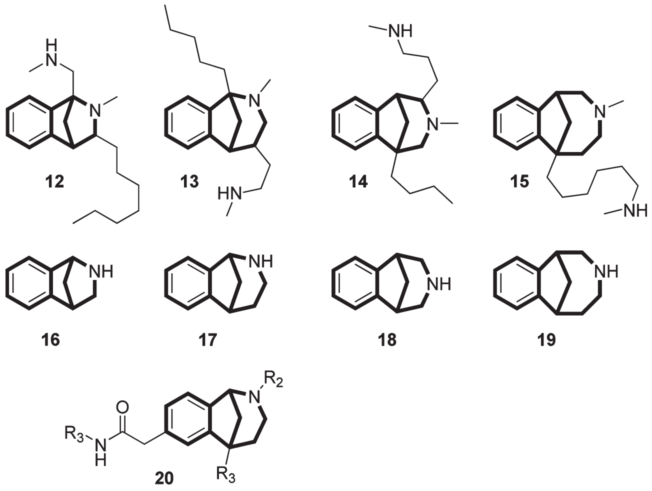


Figure 18. Comparison of virtual benzomorphan related compounds **12**−**15** with structures from the literature.

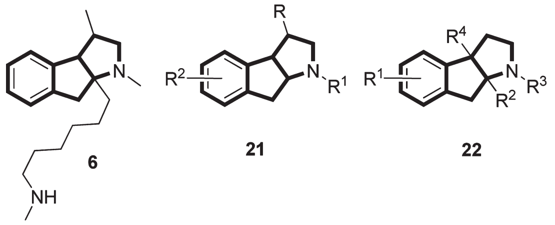


Figure 19. Substructure comparison between virtual compound **6** from generation 10 and generic structures **21** and **22** claimed in the patent literature as CNS active agents.

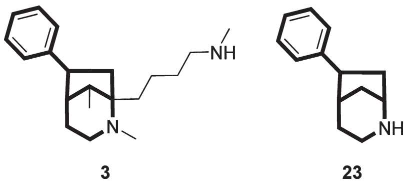


Figure 20. Example of an enumerated scaffold not yet reported in the literature.

Table 7. Bayesian NP-Likeness Model Performance and AUC for the Validation Set ROC Curve as a Function of Fingerprint Method

fingerprint Method	AUC
FEFP_2[a]	0.929
FEFP_6[a]	0.971
ECFP_2[b]	0.970
ECFP_6[b]	0.983
FCFP_2[b]	0.936
FCFP_6[b]	0.978
MDL public keys	0.925

[a] Functional class atom environment fingerprint. [b] Extended connectivity fingerprint.

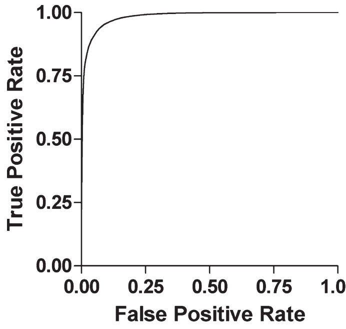


Figure 21. Bayesian NP-likeness model validation set ROC curve (ECFP_6 fingerprint method). The AUC is 0.983.

and the MDL public keys. The MDL/SYMYX Available Chemicals Directory (ACD release 2007.3) and the CRC Dictionary of Natural Products (DNP) were used to train a naïve Bayesian NP-likeness classification model (see Methods Section for details). During the course of this study, an update for the DNP was received (DVD release 192). This updated set was prepared as described in the Methods section and was found to contain an additional 2941 unique natural product structures not present in the original version used to construct the training and validation sets. As a result, the complement compounds from the two DNP versions could be used as an external test set for the NP-likeness model.

With respect to drug-likeness, a variety of machine-learning tools, methods, and approaches have been reported to differentiate drug-like from nondrug-like molecules.[47−61] Most recently, for example, a model-free (rule-based) drug-likeness filter was developed by Ursu and Oprea using molecular fragments selected from extended connectivity atom environments.[62] While methods to discriminate drugs from nondrugs have been reported in the literature for some time, we needed a rapid scoring filter for our enumerated structures that could be implemented in Pipeline Pilot. As a result, we investigated the utility of using extended connectivity fingerprints in a manner similar to Good and Hermsmeier.[63]

Performance of the natural product-like Bayesian model as a function of fingerprint method was assessed by calculating the area under the curve (AUC) from the validation set ROC plot. Using this metric, the extended-connectivity fingerprints slightly outperformed the atom environment and the MDL public keys, although all of the methods gave excellent results (Table 7). The AUC values are comparable to that reported by Ertl and coworkers[43] for differentiating compounds in the DNP from those in the Novartis compound registry using HOSE codes. The ROC plot for the ECFP_6 method appears in Figure 21.

The ECFP_6 fingerprint was selected for advancement, and the second Bayesian model was built using the sets reversed (i.e., set B used as the training set and set A used as the validation set). Performance of this model was identical to the first (validation set AUC = 0.984). In scoring test compounds for NP-likeness, the probability values from the two ECFP_6 models were averaged and used as the final NP-likeness score. Under these conditions, test compounds that more closely resemble the ACD set (less NP-like) would exhibit probability values approaching zero, and those that more closely resemble the DNP set (more NP-like) would exhibit values approaching one.

Using this model, a histogram of NP-likeness scores for the DNP and ACD sets was generated that clearly shows a difference between the two (Figure 22). A histogram of NP-likeness scores

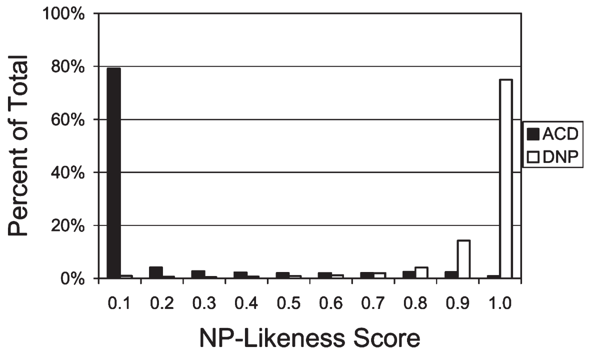


Figure 22. Histogram of NP-likeness scores for the ACD (N = 534,985) and DNP (N = 141,613) data sets.

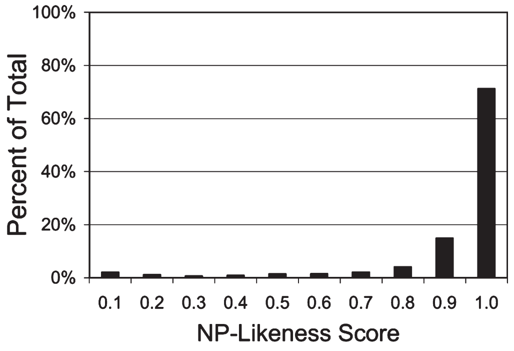


Figure 23. Histogram of NP-likeness scores for the external test set derived from the complement of the DNP DVD versions 191 and 192 (N = 2,941). Approximately 85% of the structures exhibited an NP-likeness score ≥0.90.

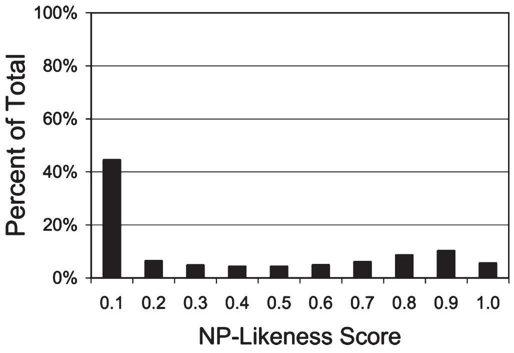


Figure 24. Histogram of NP-likeness scores for the public NCI compound data set (N = 265,196).

for the external test set appears in Figure 23. Gratifyingly, 85% of the recently added natural products to the DNP exhibited an NP-likeness score ≥0.90. Thus, on the basis of the validation and external test set results, we felt that the NP-likeness model was sufficiently accurate for use by the enumerator. For comparison, scores were also calculated for the public NCI compound data set,[64] which contains 265,196 structures (Figure 24). Although somewhat dominated by low scoring compounds, approximately 40% of the structures in the NCI set exhibited an NP-likeness score of 0.5 or greater. At the 0.90 or greater level, the number was 16%.

By way of comparison, the first six examples from the MDPI collection[65] presented by Ertl and co-workers as having a high calculated NP-likeness score using their HOSE-based algorithm is presented in Figure 25. All of these structures are also predicted by the ECFP_6 method to possess a high calculated NP-likeness.

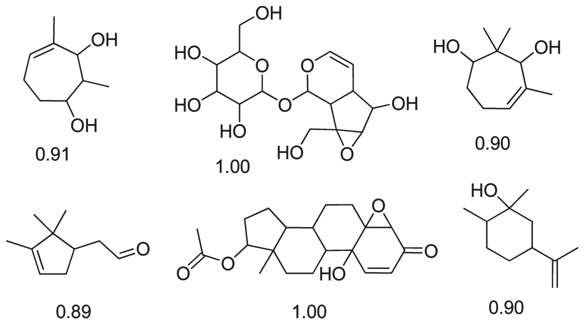


Figure 25. Comparative example of structures from the MDPI collection. These represent the first six compounds presented by Ertl and co-workers[43] as having a high calculated NP-likeness using their HOSE-based scoring method. Shown under each structure is the ECFP_6 NP-likeness score.

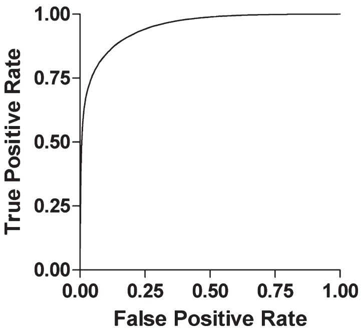


Figure 26. Drug-likeness model validation set ROC curve for the ECFP_6 fingerprint method. The AUC is 0.949.

Construction of the drug-likeness model proceeded similarly. Using ECFP_6 as the independent parameter, the first Bayesian model was built with set C as the training set and set D as the validation set (see Methods section for details regarding the two sets). The AUC for the validation set ROC curve was found to be 0.949 (Figure 26). The sets were then reversed (i.e., set D used as the training set and set C used as the validation set), and the second Bayesian model was built. Performance of the second model was comparable to that of the first (validation set AUC = 0.947). As described earlier for the NP-likeness model, the drug-likeness score for test compounds was simply calculated as the average probability from the two models. A histogram of drug-likeness scores for the MDDR and ACD sets appears in Figure 27. The model was tested using the DrugBank data set as an external test set (the DrugBank data set was prepared as previously described and any structures common to the MDDR were excluded). A histogram of drug-likeness scores for the 3162 compounds from the external test set appears in Figure 28. Approximately 84% of these were calculated to have a drug-likeness score of 0.7 or greater. For comparison, a histogram of drug-likeness scores for the public NCI data set appears in Figure 29. On the basis of the validation and external test set results, we felt that the model was sufficiently accurate to help guide the enumeration output.

The NP-likeness and drug-likeness models were linked in series and incorporated as the fitness function in the enumeration cycle depicted in Figure 2. In this manner, structures that are both

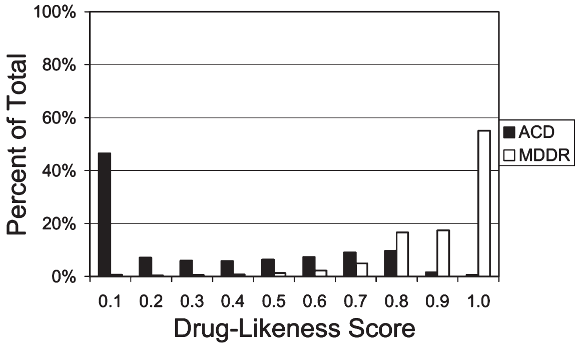


Figure 27. Histogram of drug-likeness scores for the ACD (*N* = 540,723) and MDDR (*N* = 157,023) data sets.

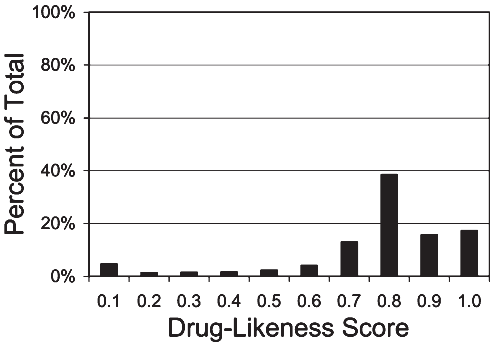


Figure 28. Histogram of drug-likeness scores for the external test set (DrugBank, *N* = 3,162). Approximately 84% of the compounds exhibited a drug-likeness score ≥0.7.

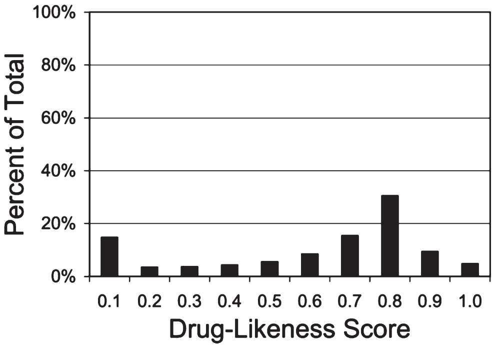


Figure 29. Histogram of drug-likeness scores for the public NCI compound data set (*N* = 265,196).

NP-like and drug-like would be expected to be generated. As before, the enumeration step used a carbon atom with four nodes as the core with the standard 21 member fragment library listed in Table 2. For each iteration, 50,000 nonduplicate structures were generated with a MW range of 299−301. For both NP-likeness and drug-likeness, threshold acceptance values of 0.90 were applied (i.e., only those structures with a calculated NP-likeness and drug-likeness score ≥0.9 were allowed to survive). As described earlier, the fragment pattern for the surviving compounds were analyzed using the algorithm outlined in Figure 7. Fragment weights were adjusted accordingly and the process was repeated for the next generation.

Table 8. Number of Enumerated Compounds from 50,000 (MW 299-301) That Met the Threshold Acceptance Criteria of 0.90 or 0.95 for Both NP-Likeness and Drug-Likeness

	threshold acceptance	
generation	≥0.90	≥0.95
1[a]	1,052	3
2[a]	9,935	28
3[a]	15,887	98
4[a]	19,526	158
5[a]	21,066	144
6[a]	21,562	142
7[b]	–	242
8[b]	–	388
9[b]	–	494
10[b]	–	521

[a] Threshold acceptance value of 0.90 was used to determine compound survival. [b] Threshold acceptance value of 0.95 was used to determine compound survival.

Table 9. Diversity Metrics[a] for Enumerated 50,000 Compound sets 1−10 and the DNP

generation	no. fingerprint features[b]	structural diversity[c]
1	13.3	0.87
2	12.1	0.86
3	10.9	0.85
4	9.86	0.85
5	8.82	0.85
6	8.00	0.85
7	8.14	0.84
8	8.16	0.84
9	7.82	0.84
10	7.55	0.84
DNP	5.05	0.91

[a] ECFP_6 fingerprint method. [b] Calculated as the total number of fingerprint features divided by the total number of molecules. [c] Calculated using eq 1.

Using this procedure, the number of compounds that met both the calculated NP-likeness and drug-likeness score thresholds rose rapidly from 1052 to 21,562 by generation 6 (Table 8). Although this represents a 20-fold increase, the number of compounds that met the 0.95 threshold value plateaued by generation 4 and is likely due to the large number of surviving compounds used for the fragment analysis step. As a result, the threshold acceptance criteria were increased to 0.95 for generations 7−10. Under these more stringent conditions, the number of compounds that met the 0.95 threshold continued to steadily increase with each successive generation. By generation 10 the number reached 521, which represents a 2 orders of magnitude increase from generation 1.

As might be expected, structural diversity as defined by the average number of ECFP_6 fingerprint features per molecule gradually declined with each successive generation (Table 9). Calculated structural diversity as defined by eq 1,[66] on the other hand, remained relatively constant suggesting that diversity can be achieved through both the combination as well as the absolute

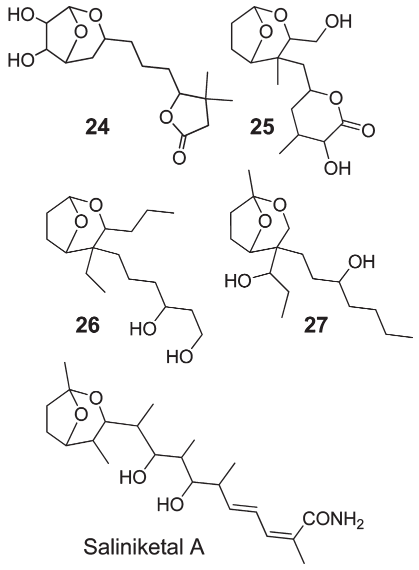


Figure 30. Structures of virtual compounds **24**−**27** from the expanded generation 10 set and saliniketal A.

number of fingerprint features. For example, compared to the enumerated series, the DNP has the fewest average number of fingerprint features per molecule and yet has the highest calculated structural diversity.

$$\text{Structural diversity} = 1 - \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} T(j,k)}{N^2} \qquad (1)$$

where T represents the Tanimoto coefficient between compounds j and k using ECFP_6, and N represents the number of compounds in the set.

Using the fragment weights employed in generation 10, an additional 300,000 virtual compounds were enumerated under the same experimental conditions. These were combined with the original generation 10 set to afford a total of 350,000 nonduplicate structures. With this expanded set, a series of cyclic ketal structures was identified from which the substituted 2,8-dioxabicyclo[3.2.1]octane analogues **24**−**27** emerged (Figure 30). This cyclic ketal ring system is common to a number of natural products such as the marine actinomycete-derived saliniketals A and B.[67] These two natural products are reported to inhibit ornithine decarboxylase (ODC) induction and may therefore represent a possible lead in the design of anticancer chemotherapeutic or chemopreventive agents. Other compounds from the expanded set include a series of 11-oxabicyclo[6.2.1]undecane analogues structurally related to a number of sesquiterpenes (e.g., 1,4-epoxy-10(14),11-germacradiene-3,5-diol, Figure 31) and a series of hexahydrofuro(2,3-b)furans (Figure 32). Substituted analogues of the latter have been claimed in the patent literature as platelet-activating factor (PAF) antagonists.[68] In addition, this ring system can be found in a number of compounds including the mycotoxin asteltoxin,[69] the fungal metabolite communiols A-D,[70] and certain HIV protease inhibitors (e.g., darunavir).[71] Also found among the ring assemblies generated by the enumerator is the tricyclic lactone ring exemplified by virtual compounds **37** and **38** (Figure 33). This ring is common to the fungal metabolite ampullicin, which is reported to exhibit plant growth regulatory activity.[72]

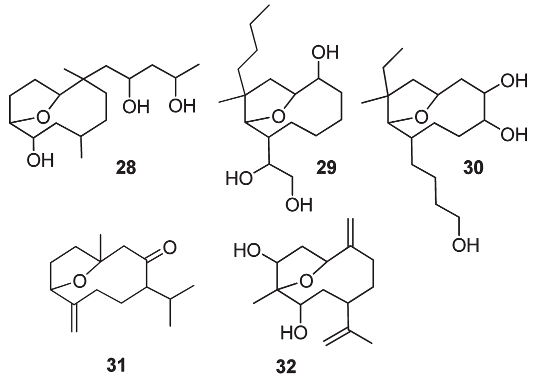


Figure 31. Virtual 11-oxabicyclo[6.2.1]undecane analogues **28**−**30** from the expanded generation 10 set and two examples from the DNP (**31**−**32**).

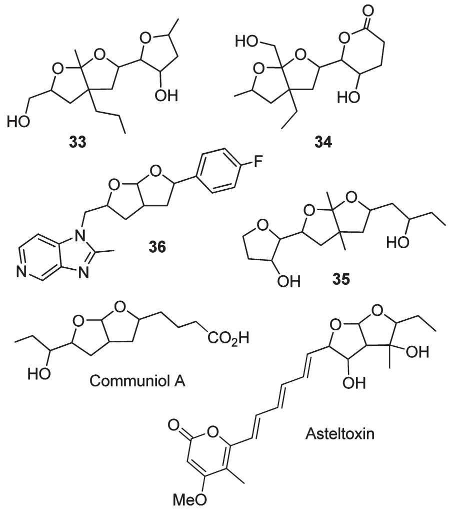


Figure 32. Virtual hexahydrofuro(2,3-b)furans **33**−**35** from the expanded generation 10 set, a compound claimed in the patent literature as a PAF antagonist (**36**), communiol A, and asteltoxin.

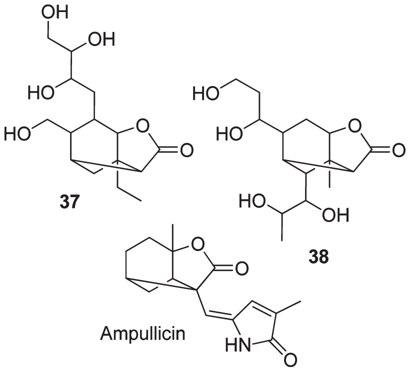


Figure 33. Virtual tricyclic lactones **37**−**38** from the expanded generation 10 set and the fungal metabolite ampullicin.

Of greater interest, perhaps, are scaffolds generated by the enumerator that are not yet represented in the DNP. These include a series of simple spirolactones exemplified by virtual compounds **39**−**41** (Figure 34). Although spirolactones exist in a number of

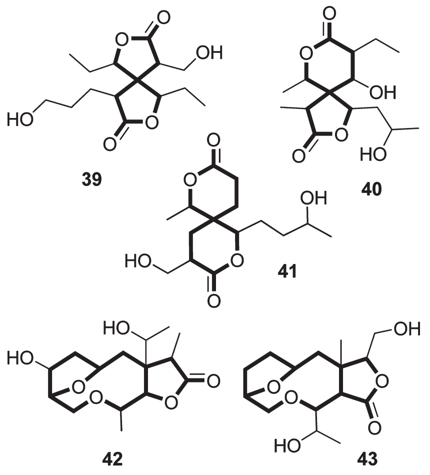


Figure 34. Virtual spirolactones **39**−**41** and fused lactones **42**−**43** from the expanded generation 10 set. Highlighted bonds represent the parent skeleton (scaffold).

Table 10. Fragment Weights Used to Enumerate 150,000 Virtual Compounds from a Four Node Carbon Atom in the Absence of a MW Constraint

fragment no.	node 1	node 2	node 3	node 4
1	2	2	2	2
2	5	5	5	5
3	5	5	5	5
4	5	5	5	5
5	5	5	5	5
6	2	2	2	2
7	2	2	2	2
8	2	2	2	2
9	5	5	5	5
10	5	5	5	5
11	5	5	5	5
12	5	5	5	5
13	5	5	5	5
14	5	5	5	5
15	5	5	5	5
16	5	5	5	5
17	5	5	5	5
18	5	5	5	5
19	5	5	5	5
20	5	5	5	5
21	5	5	5	5

natural products as part of a more complex ring system (e.g., ememogin, vilmorinine C, haperforin F, and ailantinol A), the parent was not found in our version of the DNP. Synthesis of the parent spirolactone **39** (see highlighted bonds) has been reported in the literature,[73] but to our knowledge no biological activity has been reported for it or for related derivatives. Other ring systems generated by the enumerator include a series of (8R)-3,11-dioxabicyclo[6.2.1]undecane structures (e.g., **42**−**43**) that have no counterpart in our version of the DNP.

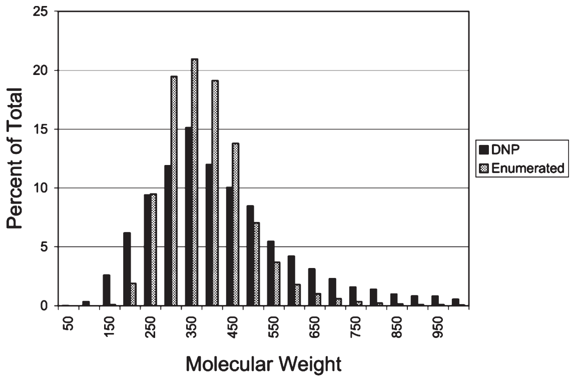


Figure 35. Comparative MW distribution of virtual molecules enumerated from a four node carbon atom in the absence of any MW constraints (N = 150,000) and the DNP (N = 141,163).

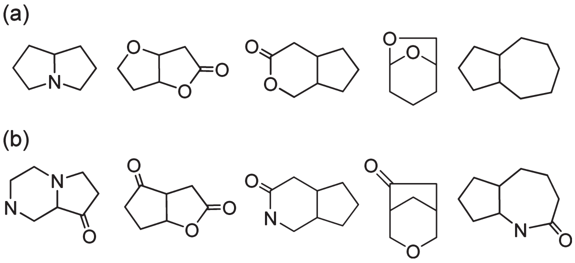


Figure 36. (a) Examples of ring assemblies common to both the enumerated set and the DNP. (b) Examples of ring assemblies found only in the enumerated set.

Ring Assembly Analysis. Rings constitute an important feature of most biologically active agents and therefore represent one of the critical factors that impact the design of compound screening libraries and lead identification strategies in general. Collectively, natural products contain a number of diverse ring systems,[74] which in contrast to most biologically active synthetic molecules are primarily aliphatic.[75] In our Pipeline Pilot analysis of the DNP, over 16,000 different ring assemblies were identified, where a ring assembly is defined as the fragment that remains after removal of all nonring bonds. As demonstrated in the previous proof of concept study, the enumerator is capable of generating a variety of ring types with a high degree of calculated NP-likeness. Analysis of 150,000 structures randomly selected from the expanded generation 10 set with an NP-likeness score ≥0.90 identified over 47,000 different ring assemblies, of which 243 were common to the DNP. This particular set, however, was limited to compounds with MW 299−301. To investigate the ability of the enumerator to generate ring containing structures over a range of molecular weights, we modified the enumeration conditions to afford virtual compounds in the absence of a MW constraint. Under these conditions, growth of each individual virtual molecule was allowed to continue unchecked until all nodes became filled either by the addition of a terminating fragment or through cyclization to form a ring. Starting with a four node carbon atom and using the standard fragment library with the fragment weights listed in Table 10, 150,000 nonduplicate structures containing at least one ring were enumerated. In this way, the virtual compounds exhibited a MW distribution similar to that of the DNP (Figure 35).

Analysis of this enumerated set identified 7132 ring assemblies, of which 187 were common to the DNP. Representative examples appear in Figure 36. Some of the ring assemblies

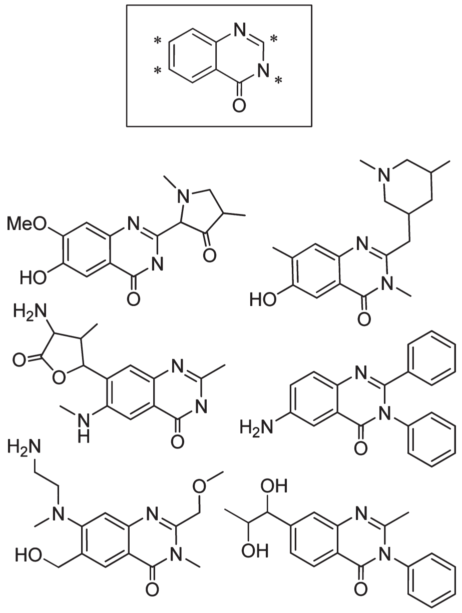


Figure 37. Representative virtual 4(3*H*)-quinazolinone derivatives enumerated from the indicated core structure and the standard 21 member fragment library (MW 100−300). The four asterisks on the 4(3*H*)-quinazolinone core mark the connection points (nodes) for added fragments.

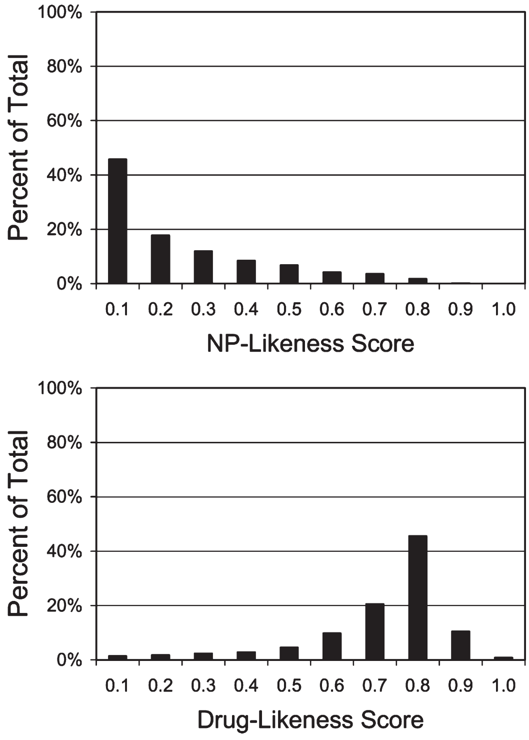


Figure 38. Histogram of NP-likeness and drug-likeness scores for the enumerated quinazolinone set, MW 100−300 (*N* = 10,000).

found only in the enumerated set are simple derivatives of those that are common to both (e.g., lactone to lactam), while others represent completely different ring systems. Although some of the latter may be synthetically challenging to access, a significant number of others may find utility in helping to either generate new ideas for novel molecular scaffolds, identify bioisosteric replacements for existing scaffolds, or provide the starting point for further elaboration and in silico studies (e.g., docking).

Other Core Structures. The core structure used by the molecular enumerator is not limited to just single atoms. If desired, aromatic rings could be assigned attachment points (nodes) and decorated with a variety of substituents. For example, building a 4(3*H*)-quinazolinone library of MW 100−300 using the standard 21 member fragment library listed in Table 2 furnished the representative virtual compounds shown in Figure 37. While the enumerated quinazolinones scored low in terms of NP-likeness, they scored high in terms of drug-likeness (Figure 38). Thus, the enumerator is capable of generating a wide variety of chemotypes with a range of calculated properties depending on the initial core structure and experimental conditions.

■ CONCLUSIONS

Recursive atom-based enumeration has the potential to generate a wide range of chemically and architecturally diverse virtual compounds from a relatively small set of standard building blocks. In this manner, core structures as simple as a single carbon atom or as complex as a polycyclic aromatic ring system can be rapidly and easily functionalized to provide drug-like structures for in silico screening. Minimizing formation of chemically invalid or unstable structures is readily accomplished through a simple but very effective set of canonical connectivity rules. Coupled with rudimentary machine-learning capability, the enumerator has the ability to produce biased virtual libraries enriched with analogues predicted to have targeted biological, toxicological, and physicochemical properties.

Depending on enumeration conditions, natural product-like structures can be produced. Because natural products represent a proven source of validated structures for identifying and designing new drug candidates,[76−78] mimicking the structural and topological diversity found in nature with a dynamic set of virtual compounds may facilitate the creation of new ideas for novel, biologically relevant lead structures in as yet unexplored areas of chemical space.

Large public virtual compound libraries currently exist to support in silico screening.[79] However, the ability to dynamically generate and focus relatively small libraries in a recursive manner could reduce the computational time and infrastructure necessary to construct and process extremely large static libraries. Thus, the ability to quickly enumerate a vast range of chemically diverse structures in silico may help the medicinal and computational chemist explore uncharted regions of chemical space with new scaffolds directed by biological or properties-based fitness functions.

■ ASSOCIATED CONTENT

S Supporting Information. Fragment weights for the second proof of concept enumeration run (generations 1−10). This material is available free of charge via the Internet at http://pubs.acs.org.

■ ACKNOWLEDGMENT

The author thanks Dr. Yoshihiro Takashima for preparing the Pipeline Pilot deglycosylation protocol.

REFERENCES

(1) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461–477.

(2) Harvey, A. L. Natural products in drug discovery. *Drug Discovery Today* **2008**, *13*, 894–901.

(3) Newman, D. J.; Cragg, G. M. Natural product scaffolds as leads to drugs. *Future Med. Chem.* **2009**, *1*, 1415–1427.

(4) Haustedt, L. O.; Mang, C.; Siems, K.; Schiewe, H. Rational approaches to natural-product-based drug design. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 445–462.

(5) Hirata, Y.; Uemura, D. Halichondrins-antitumor polyether macrolides from a marine sponge. *Pure Appl. Chem.* **1986**, *58*, 701–710.

(6) Aicher, T. D.; Buszek, K. R.; Fang, F. G.; Forsyth, C. J.; Jung, S. H.; Kishi, Y.; Matelich, M. C.; Scola, P. M.; Spero, D. M.; Yoon., S. K. Total synthesis of halichondrin B and norhalichondrin B. *J. Am. Chem. Soc.* **1992**, *114*, 3162–3164.

(7) Towle, M. J.; Salvato, K. A.; Budrow, J.; Wels, B. F.; Kuznetsov, G.; Aalfs, K. K.; Welsh, S.; Zheng, W.; Seletsky, B. M.; Palme, M. H.; Habgood, G. J.; Singer, L. A.; DiPietro, L. V.; Wang, Y.; Chen, J. J.; Quincy, D. A.; Davis, A.; Yoshimatsu, K.; Kishi, Y.; Yu, M. J.; Littlefield, B. A. In vitro and in vivo anticancer activities of synthetic macrocyclic ketone analogues of halichondrin B. *Cancer Res.* **2001**, *61*, 1013–1021 .

(8) Wang, Y.; Habgood, G. J.; Christ, W. J.; Kishi, Y.; Littlefield, B. A.; Yu, M. J. Structure−activity relationships of halichondrin b analogues: Modifications at C.30−C.38. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1029–1032.

(9) Seletsky, B. M.; Wang, Y.; Hawkins, L. D.; Palme, M. H.; Habgood, G. J.; DiPietro, L. V.; Towle, M. J.; Salvato, K. A.; Wels, B. F.; Aalfs, K. K.; Kishi, Y.; Littlefield, B. A.; Yu, M. J. Structurally simplified macrolactone analogues of halichondrin B. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5547–5550.

(10) Zheng, W.; Seletsky, B. M.; Palme, M. H.; Lydon, P. J.; Singer, L. A.; Chase, C. E.; Lemelin, C. A.; Shen, Y.; Davis, H.; Tremblay, L.; Towle, M. J.; Salvato, K. A.; Wels, B. F.; Aalfs, K. K.; Kishi, Y.; Littlefield, B. A.; Yu, M. J. Macrocyclic ketone analogues of halichondrin B. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5551–5554.

(11) Yu, M. J.; Kishi, Y.; Littlefield, B. A. Discovery of E7389, a Fully Synthetic Macrocyclic Ketone Analog of Halichondrin B. In *Anticancer Agents from Natural Products*, 1st ed.; Cragg, G. M., Kingston, D. G. I., Newman, D. J., Eds.; Taylor & Francis/CRC Press: Boca Raton, FL, 2005; pp 241−265.

(12) Eisai, Inc. http://www.eisai.com/view_press_release.asp?ID=129&press=231 (accessed March 1, 2010).

(13) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545–552.

(14) Globus, A.; Lawton, J.; Wipke, T. Automated molecular design using evolutionary techniques. *Nanotechnology* **1999**, *10*, 290–299.

(15) Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494.

(16) Pegg, S. C.-H.; Haresco, J. J.; Kuntz, I. D. A genetic algorithm for structure-based de novo design. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 911–933.

(17) Vinkers, M. H.; De Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; Van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765–2773.

(18) Douguet, D.; Thoreau, E.; Grassy, G. A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449–466.

(19) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-based approach to de novo design using reaction vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163–1184.

(20) Faulon, J.-L.; Churchwell, C. J.; Visco, D. P., Jr. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721–734.

(21) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfieldt, J. Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

(22) Patchett, A. A.; Nargund, R. P. Privileged structures: An update. *Annu. Rep. Med. Chem.* **2000**, *35*, 289–298.

(23) 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.

(24) Schneider, G.; Schneider, P.; Renner, S. Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(25) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opinion Drug Discovery Dev.* **2008**, *11*, 365–374.

(26) *Pipeline Pilot*, version 7.5.2; Accelrys: San Diego, 2008.

(27) DrugBank. http://www.drugbank.ca/ (accessed October 7, 2010).

(28) *Dictionary of Natural Products*; DVD release 191; Chapman & Hall/CRC Press: London, 2010.

(29) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Int. Ed.* **1999**, *38*, 643–647.

(30) Dančík, V.; Seiler, K. P.; Young, D. W.; Schreiber, S. L.; Clemons, P. A. Distinct biological network properties between the targets of natural products and disease genes. *J. Am. Chem. Soc.* **2010**, *132*, 9259–9261.

(31) Gulavita, N.; Hori, A.; Shimizu, Y. Aphanorphine, a novel tricyclic alkaloid from the blue-green alga. *Tetrahedron Lett.* **1988**, *29*, 4381–4384.

(32) Mokotoff, M.; Jacobson, A. E. Azabicyclo Chemistry II. Synthesis of 1,5-methano-2,3,4,5-tetrahydro-1H-2-benzazepines. B-norbenzomorphans (1). *J. Heterocycl. Chem.* **1970**, *7*, 773–778.

(33) Mazzochi, P. H.; Harrison, A. M. Synthesis and analgetic activity of 1,2,3,4,5,6-hexahydro-1,6-methano-3-benzazocines. *J. Med. Chem.* **1978**, *21*, 238–240.

(34) Mazzochi, P. H.; Stahly, B. C. Synthesis and pharmacological activity of 2,3,4,5-tetrahydro-1,5-methano-1*H*-3-benzazepines. *J. Med. Chem.* **1979**, *22*, 455–457.

(35) Chen, Y. L.; Liston, D.; Nielsen, J.; Chapin, D.; Dunaiskis, A.; Hedberg, K.; Ives, J.; Johnson, J., Jr.; Jones, S. Syntheses and anticholinesterase activity of tetrahydrobenzazepine carbamates. *J. Med. Chem.* **1994**, *37*, 1996–2000.

(36) Grunewald, G. L.; Sall, D. J.; Monn, J. A. Conformational and steric aspects of the inhibition of phenylethanolamine N-methyltransferase by benzylamines. *J. Med. Chem.* **1988**, *31*, 433–44.

(37) Hayashibe, S.; Yamasaki, S.; Shiraishi, N.; Hoshii, H.; Tobe, T. Preparation of Fused Indane Compounds as NMDA Receptor Antagonists. Patent WO 2009069610 A1, June 4, 2009.

(38) Harling, J. D.; Orlek, B. S. Preparation of Hexahydroindeno-[2,1-b]pyrroles as Neuronal Calcium Antagonists. Patent WO 9710210 A1, March 20, 1997.

(39) Ul-Haq, Z.; Mahnood, U.; Jehangir, B. Ligand-based 3D-QSAR studies of physostigmine analogues as acetylcholinesterase inhibitors. *Chem. Biol. Drug Des.* **2009**, *74*, 571–581.

(40) Feher, M.; Schmidt, J. M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

(41) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245–1252.

(42) Rosén, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel chemical space exploration via natural products. *J. Med. Chem.* **2009**, *52*, 1953–1962.

(43) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68–74.

(44) Bremser, W. HOSE: A novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355–365.

(45) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

(46) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(47) 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.

(48) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.

(49) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.

(50) Frimurer, T. M.; Bywater, R.; Nrum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non druglike" compounds. *J. Chem. Inf. Model.* **2000**, *40*, 1315–1324.

(51) Wagener, M.; van Geerestein, V. J. Potential drugs and nondrugs: Prediction and identification of important structural features. *J. Chem. Inf. Model.* **2000**, *40*, 280–292.

(52) Xu, J.; Stevenson, J. Drug-like index: A new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Model.* **2000**, *40*, 1177–1187.

(53) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841–1846.

(54) Brustle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, physical properties, and drug-likeness. *J. Med. Chem.* **2002**, *45*, 3345–3355.

(55) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Model.* **2003**, *43*, 1882–1889.

(56) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Model.* **2003**, *43*, 2048–2056.

(57) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302–321.

(58) Muller, K.-R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'drug-likeness' with kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–253.

(59) Zheng, S. X.; Luo, X. M.; Chen, G.; Zhu, W. L.; Shen, J. H.; Chen, K. X.; Jiang, H. L. A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model.* **2005**, *45*, 856–862.

(60) Li, Q. L.; Bender, A.; Pei, J. F.; Lai, L. H. A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776–1786.

(61) Schneider, N.; Jackels, C.; Andres, C.; Hutter, M. C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628.

(62) Oleg Ursu, O.; Oprea, T. I. Model-free drug-likeness from fragments. *J. Chem. Inf. Model.* **2010**, *50*, 1387–1394.

(63) Good, A. C.; Hermsmeier, M. A. Measuring CAMD technique performance. 2. How "druglike" are drugs? Implications of random test set selection exemplified using druglikeness classification models. *J. Chem. Inf. Model.* **2006**, *47*, 110–114.

(64) DTP/ 2D and 3D Structural Information, National Cancer Institute. http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html (accessed October 1, 2010).

(65) MDPI Compound Collection, v46. http://www.mdpi.org/molmall/.

(66) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.

(67) Williams, P. G.; Asolkar, R. N.; Kondratyuk, T.; Pezzuto, J. M.; Jensen, P. R.; Fenical, W. Saliniketals A and B, bicyclic polyketides from the marine actinomycete *Salinispora arenicola*. *J. Nat. Prod.* **2007**, *70*, 83–88.

(68) Whittaker, M. Hexahydrofuro[2,3-b]furans as Platelet-Activating Factor Antagonists. Patent WO 9308194 A1 19930429, 1993.

(69) Kruger, G. J.; Steyn, P. S.; Vleggaar, R. X-ray crystal structure of asteltoxin, a novel mycotoxin from *Aspergillus stellatus curzi*. *J. Chem. Soc. Chem. Commun.* **1979**, 441–442.

(70) Che, Y.; Gloer, J. B.; Scott, J. A.; Malloch, D. Communiols A-D: New mono- and bis-tetrahydrofuran derivatives from the coprophilous fungus *Podospora communis*. *Tetrahedron Lett.* **2004**, *45*, 6891–6894.

(71) Molina, J. M.; Hill, A. Darunavir (TMC114): A new HIV-1 protease inhibitor. *Expert Opin. Pharmacother.* **2007**, *8*, 1951–1964.

(72) Kimura, Y.; Nakajima, H.; Hamasaki, T.; Matsumoto, T.; Matsuda, Y.; Tsuneda, A. Ampullicin and isoampullicin, new metabolites from an ampulliferina-like fungus sp. No. 27. *Agric. Biol. Chem.* **1990**, *54*, 813–814.

(73) Harrison, T.; Pattenden, G.; Myers, P. L. Radical cyclisations onto 2(5H)-furanone and maleate electrophores leading to spiro- and linear-fused γ-lactone ring systems. *Tetrahedron Lett.* **1988**, *29*, 3869–3872.

(74) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Nat. Acad. Sci.* **2005**, *102*, 17272–17277.

(75) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenbauer, 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.

(76) Newman, D. J. Natural products as leads to potential drugs: An old process or the new hope for drug discovery. *J. Med. Chem.* **2008**, *51*, 2589–2599.

(77) Kinghorn, A. Drug Discovery from Natural Products. In *Foye's Principles of Medicinal Chemistry*, 6th ed.; Lemke, T. L., Williams, D. A., Eds.; Lippincott Williams & Wilkins: Philadelphia, 2008; pp 10−25.

(78) Koehn, F. E.; Carter, G. T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 206–220.

(79) Blum, L. C.; Reymond, J.-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.