

Drug Rings Database with Web Interface. A Tool for Identifying Alternative Chemical Rings in Lead Discovery Programs

Xiao Qing Lewell,^{*,†} Andrew C. Jones,[†] Craig L. Bruce,[†] Gavin Harper,[†] Matthew M. Jones,[†] Iain M. Mclay,[†] and John Bradshaw[‡]

Medicines Research Centre, GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K., and Daylight Chemical Information Systems, Inc., Sheraton House, Castle Park, Cambridge, CB3 0AX, U.K.

Received January 27, 2003

This paper describes the development of a drug rings database and Web-based search tools. The database contains ring structures from both corporate and commercial databases, along with characteristic descriptors including frequency of occurrence as an indicator of synthetic accessibility and calculated property and geometric parameters. Analysis of the rings in several major databases is described, with illustrations of applications of the database in lead discovery programs where bioisosteres and geometric isosteres are sought.

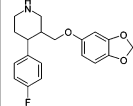
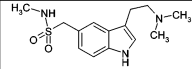
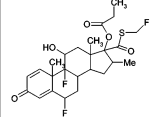
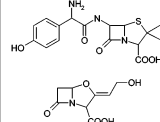
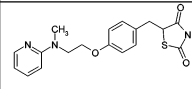
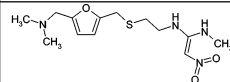
Introduction

There is a continuous need in the pharmaceutical industry to generate new drug candidate molecules. This need is typically addressed by one of several means: high-throughput screening of corporate collections, active acquisition and screening of external suppliers' offerings to enhance chemical diversity and desired chemotypes, and knowledge-based drug design and synthesis including use of combinatorial chemistry techniques. All of these methods are aimed at enhancing the chance of success in identifying novel and potent molecules for candidate selection.

A typical drug molecule usually consists of a combination of chemical rings, chains, and functional groups. Among these, rings can form a large proportion of a drug molecule. This can be illustrated by the number of different rings present in the top-selling drugs from one of the major pharmaceutical companies (GlaxoSmithKline) in 2001 (Table 1) and the proportion of rings in development compounds in PJB's pharmaproject,¹ a database containing a collection of preclinical to clinical phase candidates and in some cases marketed drugs. Of the 10K development compounds, 96% of the structures contain rings, and among these, 56% of the molecular weight is the weight of the rings. This argues in favor of devoting a large effort to designing rings, whether they form scaffolds for positioning the functional groups in the correct orientation and position to interact with receptors, or help to reduce the unfavorable loss of conformational entropy upon receptor binding, or simply, possess intrinsic bioactive properties.

The efforts to replace rings form a large part of medicinal chemistry practice. A typical lead optimization program often involves a trial and error process of replacing chemical functionalities including rings. Often this intellectual process is based on a chemist's knowledge and recollection of his or her past experiences. While this has been effective in the past, with the

Table 1. Six Top-Selling GSK Drugs in 2001

Structure	Name (Mechanism of Action)
	Paxil ^a /Seroxat ^a (5HT uptake Inhibitor)
	Imigran ^a (5HT1d Agonist)
	Flixotide ^a (Glucocorticoid activity agonist)
	Augmentin ^a (Beta-lactamase inhibitor)
	Avandia ^a (PPAR gamma agonist)
	Zantac ^a (H2-antagonist)

^a Trademark of the GlaxoSmithKline group of companies.

advent of combinatorial chemistry technology and the ever-increasing number of molecular structures that can be readily synthesized, there is an increasing number of potential new structural ring templates. It has become impossible for any one chemist to recall them all.

There have been many efforts to describe and categorize ring-containing chemical systems in the past, dating from 1990 when Nilakantan et al. described their ring-based chemical structure query system.² In a related development, Lipkus from Chemical Abstracts Service developed ring-topology-based descriptors to categorize and search ring-containing systems in the Chemical Abstracts database.^{3–5} Such concepts, along with other graph-theoretical based methods such as the ring-cluster concept developed by Nilakantan,⁶ are useful

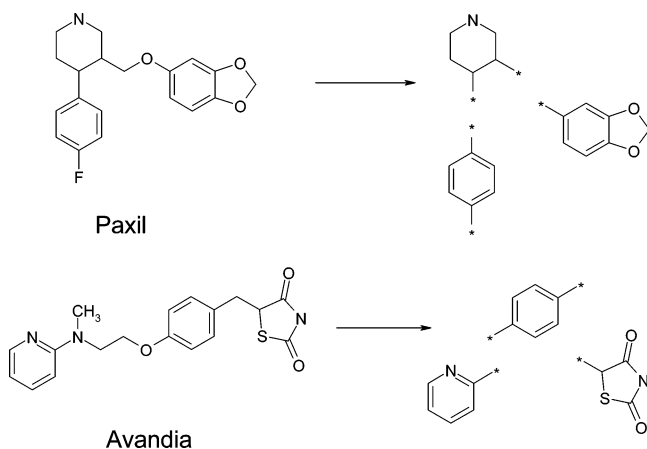
* To whom correspondence should be addressed. Phone: +44 (0) 1438 745745. Fax: +44 (0) 1438 764918. E-mail: xiao.q.lewell@gsk.com.

[†] GlaxoSmithKline Research and Development.

[‡] Daylight Chemical Information Systems, Inc.

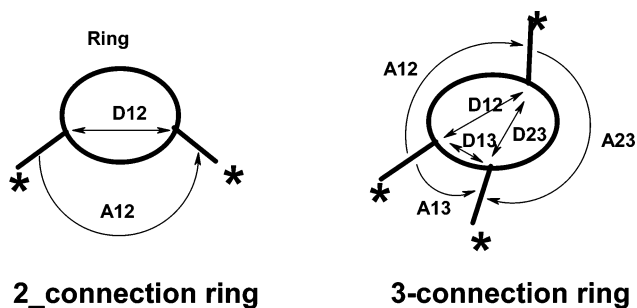
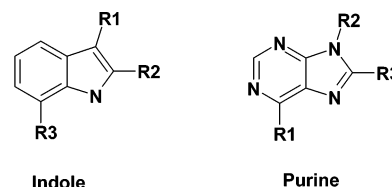
Table 2. Drug Rings Analysis and Ring Utilization Diversity

database	no. of structures analyzed	no. of unique rings	ring utilization diversity (%) [(no. of unique rings)/(no. of structures)]
Corporate Registry	1425099	57231	4.0
BACD	3562146	56467	1.6
PJB2000	10761	4462	41.5
NCI96	30379	9696	31.9
DNP102	113900	27788	24.4
WDI014	65171	13813	21.2
Medchem02	43850	6060	13.8
ACD014	287764	15254	5.3

**Figure 1.** Bond cleavage examples.

aids in assessing database diversity based on their chemical ring content and in identifying voids in corporate collections with a view to ensuring that future compound acquisition will enhance database diversity.

At the medicinal chemistry project level, bioisosteres have been used to replace specific functionalities and rings in many individual situations. Recent examples are given in the areas of antitumor agents⁷ and nicotinic,⁸ endothelin,⁹ and dopamine¹⁰ agents where functional groups including rings are replaced by bioisosteres. At the generic level, Sheridan¹¹ has described

**Figure 2.** Geometric variables for two- and three-connection rings.

Randomised Order of Geometric Variables

Indole	1.34 (d12)	3.62(d13)	3.62(d23)	71.6(a12)	160.3(a13)	88.8(a23)
Purine	3.57 (d12)	3.6 (d13)	1.36 (d23)	162.3 (a12)	91.2 (a13)	71.1 (a23)

Canonicalised Order of Geometric Variables

Indole	1.34 (d12)	3.62(d13)	3.62(d23)	71.6(a12)	160.3(a13)	88.8(a23)
Purine	1.36 (d23)	3.57 (d12)	3.6 (d13)	71.1 (a23)	162.3 (a12)	91.2 (a13)

Figure 3. Illustration of canonicalization of three-connection geometric variables.

an algorithm to identify typical bioisosteres in the drug collection MDDR.¹² Bemis and Murko have classified typical molecular frameworks¹³ and side chains¹⁴ in drug collections into several major classes to aid drug design efforts.

Although many methods and efforts have been described in the literature for categorizing ring-containing chemical systems, to date there have been no commercial software or database available to us that documents the knowledge that could be gained from a

Table 3. Calculated Properties for the Unique Set of Rings

calcd/stored property	description of the property
text parameter	
SMILES	a string representation of a chemical structure (see ref 15)
ring name	an artificial name given to the ring
counts	
acid	no. of acid functional groups in the ring system (defined by SMARTS notation)
base	no. of basic groups in the ring system (defined by SMARTS notation)
rings	no. of rings in the ring system (for example, naphthalene has two rings)
aromatic rings	no. of aromatic rings in the ring system (for example, tetrahydronaphthalene has one aromatic ring)
H-bond donor	no. of hydrogen bond donors in the ring system (defined by SMARTS notation)
H-bond acceptor	no. of hydrogen bond acceptors in the ring system (defined by SMARTS notation).
N	no. of nitrogen atoms in the ring system
O	no. of oxygen atoms in the ring system
P	no. of phosphorus atoms in the ring system
S	no. of sulfur atoms in the ring system
connections	no. of cleavage points in the ring when it was derived from a molecule
logical parameter	
fuse	whether the ring system contains fused rings (i.e., two rings share the same bond, defined by SMARTS notation)
spiro	whether the ring system contains a spiro junction (defined by SMARTS notation)
numerical parameter	
MW	molecular weight of the ring system
complexity	structural complexity of the ring system (explained in the text and Appendix)
two-connection geometry	geometric parameters for two-connection ring systems (explained in the text)
three-connection geometry	geometric parameters for three-connection ring systems (explained in the text)
frequency in individual database	number of times a particular ring appears in a particular database
overall scaled frequency	measure of overall frequency that accounts for the absolute number of occurrences of rings in a database and the relative database size (defined in the text)

The screenshot shows the DAYBASE web interface in Microsoft Internet Explorer. The browser address bar shows the URL: <http://ukw3s3.ggr.co.uk:8000/gvhtml/acj13545/daybase2.html>. The page title is "DAYBASE new_drug_rings web interface".

The interface is divided into several sections:

- Search Query:** A table with columns "Properties:", "Min.", and "Max.".

Properties:	Min.	Max.
Complexity:	<input type="text"/>	<input type="text"/>
Connections:	<input type="text" value="3"/>	<input type="text" value="3"/>
Ring Count:	<input type="text" value="2"/>	<input type="text" value="2"/>
Arom. Ring Count:	<input type="text"/>	<input type="text"/>
MW:	<input type="text"/>	<input type="text"/>
Base:	<input type="text"/>	<input type="text"/>
Acid:	<input type="text"/>	<input type="text"/>
HBA:	<input type="text"/>	<input type="text"/>
HBD:	<input type="text"/>	<input type="text"/>
Oxygen:	<input type="text"/>	<input type="text"/>
Sulfur:	<input type="text"/>	<input type="text"/>
Phosphorus:	<input type="text"/>	<input type="text"/>
Nitrogen:	<input type="text"/>	<input type="text"/>
Spiro:	-None-	<input type="checkbox"/>
Fuse:	-None-	<input type="checkbox"/>
- Frequencies:** A table with columns "Overall Freq:", "Min.", and "Max.".

Overall Freq:	Min.	Max.
BACD_lite:	<input type="text"/>	<input type="text"/>
gwr:	<input type="text"/>	<input type="text"/>
pjb_2000:	<input type="text"/>	<input type="text"/>
sb2gw:	<input type="text"/>	<input type="text"/>
acd011:	<input type="text"/>	<input type="text"/>
affymax:	<input type="text"/>	<input type="text"/>
dnp102:	<input type="text"/>	<input type="text"/>
medchem02:	<input type="text"/>	<input type="text"/>
nci96:	<input type="text"/>	<input type="text"/>
wdi014:	<input type="text"/>	<input type="text"/>
- Search type and Options:**
 - Sketch structure (ISIS Draw - Double click below to open if ISIS Draw is available)
 - Chemical structure diagram of a 1,3,5-substituted indole core.
 - Or tick to enter string below:
 - Range Search: Sim. Coeff. Use
 - Similarity search:
 - Superstructure search:
 - SMARTS search:
 - Lookup datatree(s):
- Geometry:**
 - Use Corina: (Concord is used by default)
 - 2pt Geometry:

	Min.	Max.	View
Angle:	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>
Distance:	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>
 - 3pt Geometry:

	Min.	Max.	
Angle:			
AB:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
BC:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
AC:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
 - Distance:

	Min.	Max.	
AB:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
BC:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
AC:	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>
- Sorting Options:**
 - Sort by:
 - Display Results Per Page
 - Sort: Ascending Descending
 - Numeric sort:
 - Length of string:
 - Save hitlist:
 - Load hitlist:
 - Show smiles:

At the bottom, there are four buttons: Refresh, Search, New Search, and General Help.

Figure 4. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: inputting the 1,3,5-substituted indole core as query.

comprehensive analysis of the drug rings and that make searching for replacement rings a smooth process for medicinal chemists.

This paper describes our effort at GSK to address this deficiency. The objective of this work was to analyze and extract drug rings occurring in molecules from both corporate and commercial databases and to document the knowledge gained in a Web-searchable format. Value-added knowledge for these drug rings was generated including the development of a novel way of describing and storing geometric information. In this way, medicinal chemists, as well as computational

chemists, can use the composite information as a tool for idea generation for ring isosteres and geometric scaffolds in lead generation and optimization programs.

Method

The essence of the method is the following. Molecules in different databases are collected. Ring cleavage methodology was developed to cleave chemical structures and retain rings. Rings are then collated, and properties for the rings are calculated, including the frequency of occurrence. The structures of the rings and their associated properties are stored in a Daylight¹⁵

Hit	Connection	Ring Count	3pt Angle 1	3pt Angle 2	3pt Angle 3	3pt Distance 1	3pt Distance 2	3pt Distance 3	SMILES	Similar
3	3	2	141.2	80.6	138.3	2.22	3.87	4.19		1
4	3	2	60	80.7	140.7	1.4	3.87	4.25		1
5	3	2	70.2	18.2	88.4	1.36	2.58	3.64		1
6	3	2	141.2	18.3	159.4	2.22	2.58	3.63		1
7	3	2	60	88.4	148.4	1.4	3.64	4.63		1
8	3	2	120	18.2	138.2	2.42	2.58	4.19		1
9	3	2	60	18.2	78.2	1.4	2.58	3.8		1

Figure 5. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: similarity search results for 1,3,5-substituted indole core (hit 3).

database. A Web search engine using a combination of Daylight Merlin Control Language and PERL with a Chime Pro plug-in is then used, allowing both query input and results to be displayed on the Web. Each of the steps will be discussed below.

Starting Databases for Analysis. Several databases each possessing a significant number of structures were analyzed. These include both corporate and commercially available databases (Table 2).

Corporate Registry contains proprietary compounds, whereas BACD is a compilation database containing most commercially available compounds. The remaining databases are readily available and contain late-stage development compounds (PJB2000¹), National Cancer Institute compounds (NCI96¹⁶), natural products in Derwent (DNP102¹⁷), the World Drug Index (WDI014¹⁸), Medicinal Chemistry database (Medchem02¹⁹), and Available Chemicals database (ACD014²⁰).

Bond Cleavage and Ring Frequency Analysis. Each structure in the databases was processed as follows. Single and olefinic noncyclic bonds were cleaved, and the cleavage point was marked in the SMILES representation with a "*", an any-atom notation in the Daylight SMILES language. The retention of this information is useful for further interrogation of the rings and for scaffold replacements as discussed in the geometric section of this paper. Figure 1 illustrates examples of cleavage in the case of Paxil and Avandia.

Ring occurrence frequencies were analyzed for each database and used as an indication of the probable synthetic accessibility. These frequencies are not absolute indications of synthetic accessibility because some

databases contain natural product based chemistries and some are biased toward particular kinds of chemistries and historical drug discovery projects. Nonetheless, it was felt that the analysis of this information would be interesting in itself in relation to the understanding of particular database contents and also as a guide to chemistries that have been used frequently in the past and potential chemistries that could be explored more in the future.

Calculated Property Descriptors. Table 3 shows a list of calculated properties for a unique set of rings generated. Standard parameters were computed using the Daylight Toolkit, while the nonstandard ones were computed using the schemes described below.

Complexity Descriptor. The complexity descriptor is a descriptor developed on the basis of the density of the Daylight fingerprint, taking the position of each bit in the fingerprint into account (see Appendix for details). It has the advantage that similar functional groups that are turned on in similar positions could be grouped more closely to each other than using the simple count of bits on. In practice, it sorts structures in increasing order of structural complexity.

Overall Scaled Frequency. The overall scaled frequency is defined using the formula

$$\text{overall scaled frequency} = \frac{\text{sum of the ratio of individual database frequency} \times \text{scale factor}}{\text{frequency} \times \text{scale factor}}$$

where the ratio of individual database frequency is the frequency of ring occurrence in a database divided by

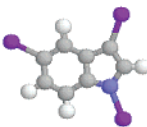
DAYBASE | Lookup in new_drug_rings@ukwsv11 for *c1cn(*)c2ccc(*)cc12 - Microsoft Internet Explorer

Address http://ukw333.ggr.co.uk:8000/gwcp/acq/13545/lookup.cgi?user=rg4406&smiles=2a6331636e282a296332636363282a2963633132

Links Best of the Web Customise Links Free Hotmail Microsoft Product News Today's Links Web Gallery Windows

THOR LOOKUP

Find drugs containing this ring in the following database: Retrieve
 Check to search for exact structure (click here for more information)



MDL

SMILES string is *c1cn(*)c2ccc(*)cc12

Distance1	2.22
Distance2	3.82
Distance3	4.13
Angle1	141.7
Angle2	80.5
Angle3	137.9
Distance1	2.22
Distance2	3.87
Distance3	4.19
Angle1	141.2
Angle2	80.6
Angle3	138.3
Overall DB Freq	35839
Fingerprint	128 bytes of binary data
Orig nbits	1024
Orig nbits set	74
Nbits	1024
Nbits set	74
Version	1
Timestamp	200208081714.21
Graph	*c1cn(*)c2ccc(*)cc12
Base	0
Acid	0
HBD	0
HBA	0
Sulfur	0
Nitrogen	1
Oxygen	0
Phosphorus	0
Connections	3
Ring Count	2
Aromatic Ring Count	2
Spiro	FALSE
Fuse	TRUE
Molecular Weight	114.13
Complexity	39564
Database	BACD_lite
BACD_lite_Frequency	221
BACD_lite_Ring Name	BACD_LITE_1222
Database	GWR
GWR_Frequency	831
GWR_Ring Name	GWR_198
Database	PJB_2000
PJB_2000_Frequency	9
PJB_2000_Ring Name	PJB_2000_221
Database	SB2GW
SB2GW_Frequency	497
SB2GW_Ring Name	SB2GW_249
Database	acd011
acd011_Frequency	19
acd011_Ring Name	ACD011_1208
Database	affymax
Affymax_Frequency	3
Affymax_Ring Name	AFFYMAX_371
Database	dnp102
dnp102_Frequency	5
dnp102_Ring Name	DNP102_3042
Database	Medchem02
medchem02_Frequency	16
medchem02_Ring Name	MEDCHEM02_351
Database	wdi014
wdi014_Frequency	18
wdi014_Ring Name	WDI014_520

Figure 6. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: looking up data stored in the drug rings database for 1,3,5-substituted indole.

the number of structures in the database. The scale factor is chosen to make the lowest occurring ring frequency into an integer. The overall frequency gives a quick indication of the rings that occur frequently in the overall databases, although without any guarantee that the ring is evenly distributed in the databases or occurs in only one database. The descriptor is useful for

the initial screening of those rings that occur most frequently.

Connection. The number of connection points is the number of “*” symbols that represent the points of cleavage in the original molecules.

Fuse, Spiro, and Other Properties. The computation of “fuse”, “spiro”, and other properties such as acids,

Hit	Connection	Ring	3pt Angle	3pt Angle	3pt Angle	3pt Distance	3pt Distance	3pt Distance	p1R	N1R 96	WDI	SMILES
6	3	2	142.6	78.6	138.8	2.26	3.8	4.25	1	~	3	
7	3	2	141.2	78.3	140.6	2.22	3.8	4.25	5	~	15	
8	3	2	144.5	76.3	139.2	2.2	3.67	4.08	1	~	~	
9	3	2	145.1	76.5	138.4	2.23	3.67	4.09	1	~	1	
10	3	2	143	74.7	142.3	2.2	3.67	4.19	2	~	2	
11	3	2	141.2	80.6	138.3	2.22	3.87	4.19	9	~	18	

Figure 7. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: using the geometry of the 1,3,5-indole to search for alternative ring scaffolds. Hit 11 is the original target of 1,3,5-indole. Hit 7 is 1,3,6-indole as a geometric alternative to the 1,3,5-indole. Only a fraction of the hits are shown here.

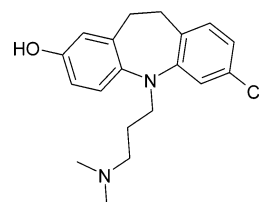
bases, and H bonds is carried out on the basis of predefined SMARTS definitions.

Geometric Descriptors. Geometric scaffolds are of clear importance in drug design, both in lead optimizations, where a scaffold is often desired to be replaced by other scaffolds for scientific as well as other reasons, and in the selection of scaffolds for combinatorial library design. A seminal paper by Bartlett's group²¹ has attempted to address the issue of geometric scaffolds in 3D databases by focusing on the relationship between bonds. Boyd²² has extended the idea discussed by Bartlett to study geometric diversity of functional groups in chemical databases. Recently Wild²³ has described an atom-based fitting effort for locating potential ring scaffolds.

We have set out to describe the geometry of two- and three-connection rings for the following reasons. The two- and three-connection rings are the simplest ring templates to test the geometric storage and the searching methodology being developed. It has been shown in the past by the Farmer hypothesis²⁴ that in the design of active analogues the optimal number of side chains to consider would be three. This will be borne out later in this paper, where we show that the number of four-connection rings and beyond falls off in occurrence frequency in the databases. The reason for this may be due to their synthetic complexity rather than their probable bioactivity. However, to address the geometric searching capability for four-connection rings and beyond, a different methodology needs to be developed. In this paper, we address the two- and three-connection

rings because they are the most frequently occurring and support the geometric methodology being developed.

The geometric descriptors developed should be simplistic yet sufficiently accurate to describe, store, and search for potential rings using these descriptors. 3D coordinates for two- and three-connection rings are generated using both CONCORD²⁵ and CORINA,²⁶ each method providing a single structure. The reason for using both 3D-generation programs was to give the user a choice of which method he/she prefers for 3D structures. CONCORD and CORINA also generated different conformations for certain rings, for example, in the case of the tricyclic core of the 8-hydroxyclopiramine molecule. CONCORD gives a flat tricyclic structure, whereas



8-Hydroxyclopiramine

CORINA gives a puckered tricycle. This results in somewhat different geometries for certain rings from the two methods and thus opens up the possibility to use a single method for consistency in searches or to use both methods for greater geometric coverage.

For two-connection rings, two simple geometric parameters, distance and bond vector angle, were calcu-

Searching PJB_2000@ukwsv11 for *c1cn(*)c2cc(*)ccc12 - Microsoft Internet Explorer

Address <http://ukw3s3.ggr.co.uk:8000/gwgc/acj13545/getoriginal.cgi>

Please be patient...searching other databases may take a few minutes

hits 1 to 31 of 31 in hitlist "xq14406", database "PJB_2000@ukwsv11.thor.xq14406"

Superstructure search using *c1cn(*)c2cc(*)ccc12 first prev 1 - 100 next last 100 per page

hit	Name	Smiles	Similarity
1	NSCD659687		0.3394
2	BM211298		0.2902
3	SINALFA		0.2879
4	NSP307		0.2569
5	SDZ208912		0.2418
6	AG555		0.2403

Figure 8. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: following hit 7 (1,3,6-indole) in the PJB2000 database for original molecules containing this core.

lated. For the three-connection rings, three distances and three bond vector angles were computed (Figure 2).

Distance is defined as the Euclidean distance between the two atoms connecting the cleaved atoms. The vector angle is defined by

$$\text{vector angle} = \cos^{-1}\left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}\right)$$

where **a** and **b** are the bond vectors between the connecting atom and the cleaved atom, “*”.

The six geometric variables in the three-connection rings were canonicalized to allow a geometric search using these variables. The canonicalization was done by sorting the distances in increasing order and carrying over the associated vector angles for the canonicalized distances. This way, a geometric search is a straightforward process using the canonicalized order of geometric variables. Figure 3 illustrates the concept of canonicalization. Consider two geometrically equivalent core templates as shown. Geometric searching on a randomly assigned order of geometric variables would be meaningless because the random order does not give correspondence between geometrically equivalent variables. However, if we canonicalize the order of these variables, the geometric search could then be performed on this order. The reason the method suggested here works in practice is facilitated by the nature of the molecular connections. Chemical bond lengths and angles take certain discrete values (e.g., a typical bond length of 1.5 Å and a typical tetrahedral bond angle of

109°), thus making the geometric variables discrete rather than continuous, particularly at the smaller distance range where most of the druglike molecules fall.

The simplicity of the geometric variables arranged in such a way allows users to understand and interpret the search results. A caveat from the canonicalization of geometric variables for the three-connection geometry is the potential mismatch to the target geometry where a small change in the canonicalization order may result in a better fit to the template. For example, a small change in a close distance pair (say from 2.1 to 2.2 Å) might result in a large change of angle (say from 90° to 50°), and thus, a much better fit is obtained if the canonicalized distance order is switched in this instance. This is solved by postprocessing using systematic comparisons of all the vector pairs in the noncanonicalized combinations to obtain and sort according to the rms fit of the geometric variables. In practice, these simple geometric descriptions, stored in such a canonicalized way, work well in identifying closely related geometric scaffolds.

Database Generation and Web Interface. A Daylight database was built containing the set of unique rings extracted and the associated calculated properties. To enable medicinal chemists to access the database easily, a Web interface to the database was developed using PERL and Merlin control language. The interface allows the user to input all search criteria in one step, although behind the scene searches are sequential. The

THOR LOOKUP

Find drugs containing this ring in the following database:

Check to search for exact structure ([click here for more information](#))

MDL

SMILES string is <chem>CC(C)N1CCN(CCN2CCC(CC2)c3cn(c4ccc(F)cc4)c5cc(Cl)ccc35)C1=O</chem>	
Fingerprint	128 bytes of binary data
Orig nbits	1024
Orig nbits set	308
Nbits	1024
Nbits set	308
Version	1
Timestamp	200011011513.55
Graph	<chem>CC(C)N1CCN(CCN2CCC(CC2)C3CN(C4CC(Cl)CCC34)C5CCC(F)CC5)C1=O</chem>
Isomeric SMILES string is <chem>CC(C)N1CCN(CCN2CCC(CC2)c3cn(c4ccc(F)cc4)c5cc(Cl)ccc35)C1=O</chem>	
Parent Avg molecular weight	483.03
Parent Molecular Formula	C27H32ClFN4O
Isomer	<chem>CC(C)N1CCN(CCN2CCC(CC2)c3cn(c4ccc(F)cc4)c5cc(Cl)ccc35</chem>
2D-coordinates	5.76,1.78,5.06,1.34,5.06,0.53,4.34,1.76,4.34,2.57,2.93
Molecular Formula	C27 H32 Cl F N4 O
Accession Number	18949
Init Date	131993
Name	Lu042
Development Status	D
Last Update	JAN
Originator	Lundbeck
Therapeutic Use	Antidepressant
Pharmacological Activity	5 Hydroxytryptamine 2 antagonist
CAS number	1397737
Avg molecular weight	483.03
Molecular Formula	C27H32ClFN4O
CAS Number	1397737
Name	AG555

Figure 9. Illustrative work flow for identifying alternative cores to 1,3,5-substituted indole core: checking for data in PJB2000 for AG555.

final results are displayed back on the screen once all the searches are done. The search results are typically fed back to the user within a matter of seconds.

Figure 4 shows the query entry page for the drug rings database. Search capabilities (Figures 5–9) include the following:

- (1) range search where a range in the various calculated descriptors can be searched for, including geometric variables,
- (2) structural similarity search using Tanimoto similarity index,
- (3) structural SMARTS search where the user can input his/her own specific SMARTS query,
- (4) superstructure search for rings containing the input query ring,

(5) data lookup to retrieve stored data associated with the query ring or on subsequent whole molecular structures from other databases.

Search results can be sorted numerically by a user-chosen descriptor.

Analysis of the Drug Rings in Corporate and Commercial Databases

Frequency of Rings. The accumulation of knowledge about drug rings from various databases makes it possible to study distribution patterns. Overall, some 120K unique drug rings, accounting for alternative connection patterns, were obtained from ca. 5.6 million structures. This corresponds to 2.1% of “ring utilization diversity”, a figure obtained from dividing the number

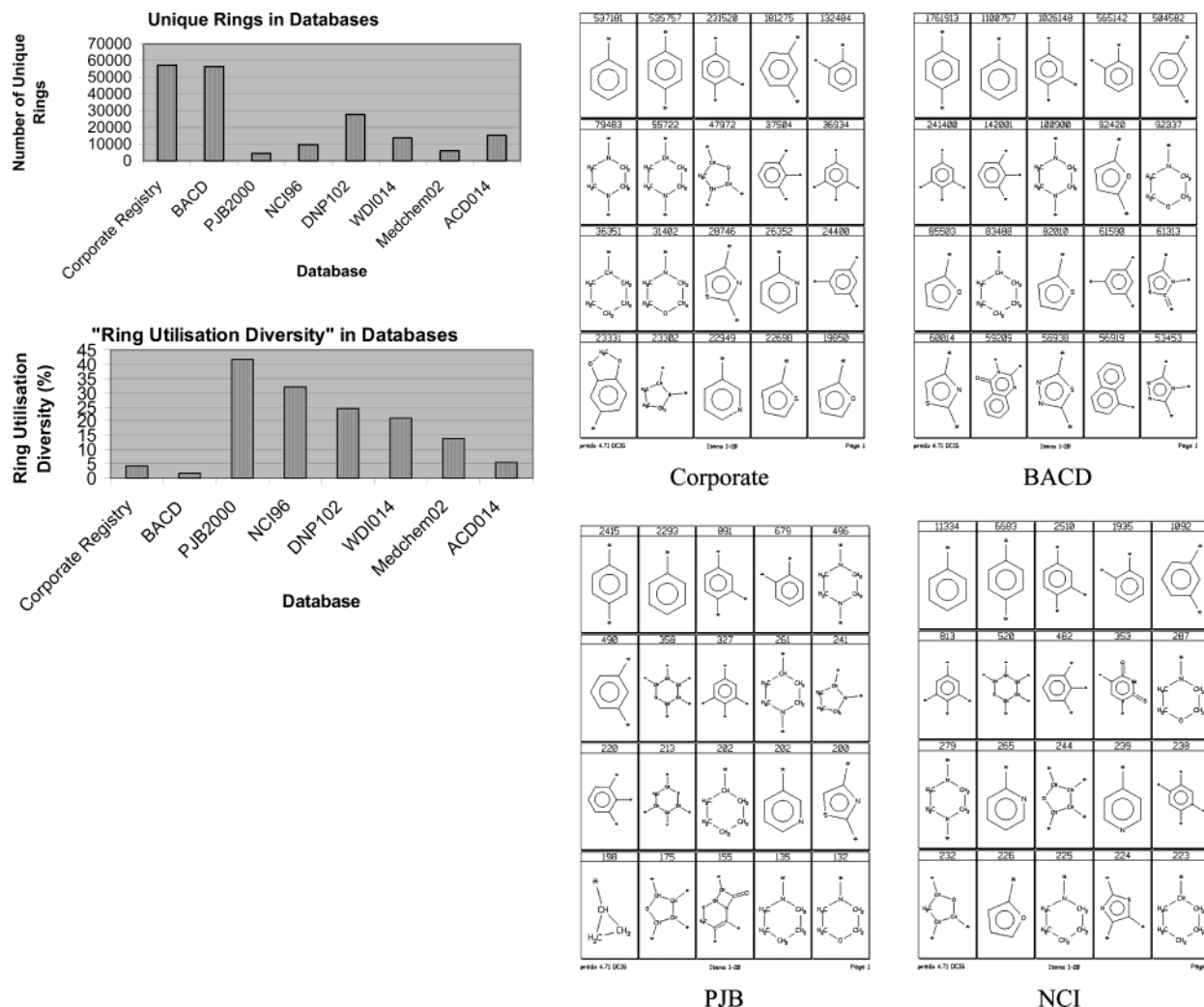


Figure 10. Number of unique rings and "ring utilization diversity" in different databases.

of unique rings by the number of structures. The ring utilization diversity, however, varies considerably in different databases, ranging from 41.5% in PJB2000, a database containing development compounds, to 1.6% in BACD, a database containing a composite repository for commercial compound offerings, as shown in Figure 10.

The percentage ring utilization diversity is influenced by both the database size and database contents. As the database size increases, the rate of increase of novel rings is expected to decrease because of the nature of the medicinal and combinatorial chemistry programs that affects both corporate and large commercial collections. This is reflected by the Corporate Registry and BACD database figures (Table 2). On the other hand, the database contents also influence the diversity figure. In the case of PJB2000, which is a database containing a collection of reputedly highly optimized key compounds in late preclinical and clinical development stages from different research groups, the database is expected to contain diverse structures. Thus, a 41.5% figure reflects both the nature of the database and its small size. It is worth noting that the absolute number of unique rings is by far the largest with large databases such as in the Corporate Registry and BACD.

The most frequently occurring rings also differ, reflecting the nature of the database. For example, in

the case of the natural product database DNP, the top occurring rings are mostly oxygen-containing rings including tetrahydropyran, furan, and tetrahydrofuran, chromone, and benzodioxole, in addition to the frequently occurring phenyl ring. In the more traditional "druglike" databases such as the Corporate Registry, PJB, and WDI, we observe rings such as piperazine, piperidine, and thiazole occurring more frequently. Figure 11 shows the top 20 occurring rings in each of the eight databases.

Phenyl, with different connection patterns, is the only common ring across the top 20 rings in all eight databases. Combining the top 20 occurring rings from all eight databases produced 48 rings. If the normalized frequency distribution of individual databases is plotted for the 48 rings, we see variations in distribution patterns, particular across DNP compared to other more druglike databases (Figure 12). For example, the peak at ring 33 represents the 5-substituted tetrahydropyran ring, which occurs most frequently in DNP but not so often in other databases.

Table 4 shows the overlap of rings between eight databases. Thus, for example, 5.3% of the Corporate Registry rings are covered by PJB while 67.3% of the PJB rings are covered by the Corporate Registry. This gives the opportunity to identify those rings that occur in other databases, for example, PJB, which contains

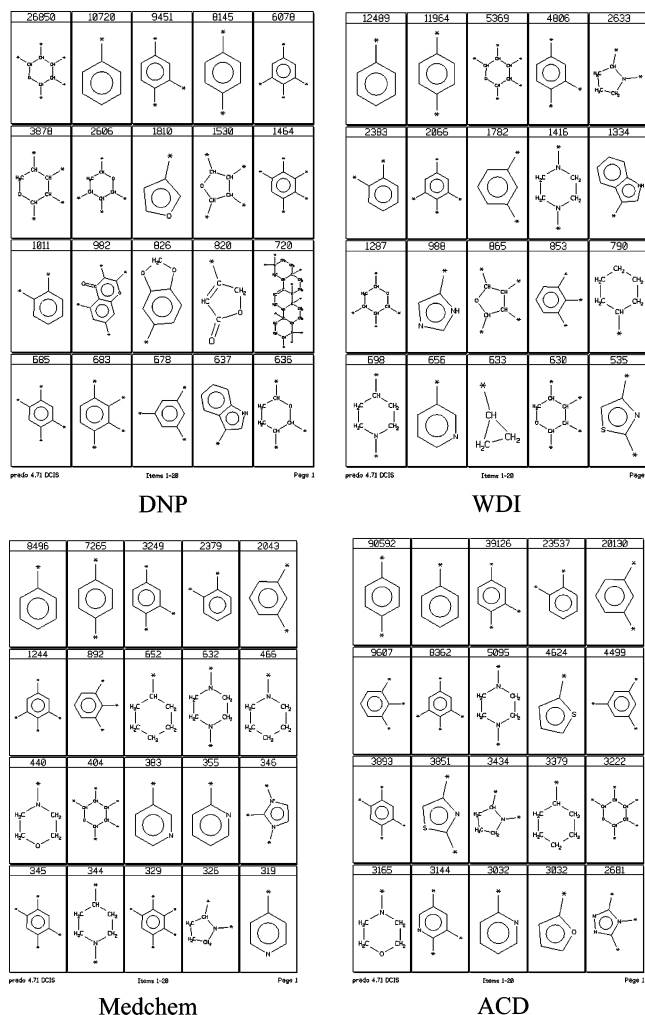


Figure 11. Top 20 occurring rings in eight databases.

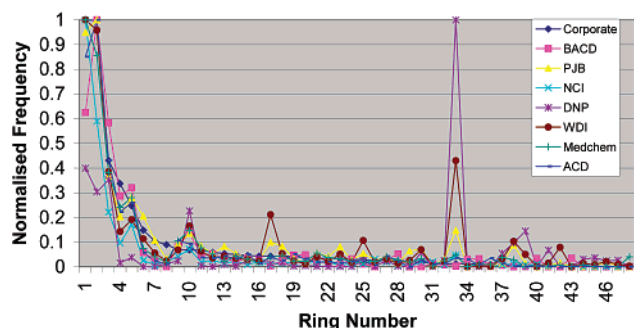


Figure 12. Frequency distribution of top occurring rings in eight databases.

late-stage development compounds, but do not occur in Corporate Registry and to evaluate feasibility to incorporate these rings in drug discovery projects, where a ring fulfills geometric as well as other requirements of the projects.

Ring Complexity. The frequency of occurrence of the rings could be used as an indication of synthetic accessibility. Figure 13 plots the occurrence frequency against structural complexity for the Corporate Registry rings. As expected, the frequency of the ring decreases as complexity increases, since preparation of such rings usually is harder. However, there are also some simple rings that do not occur frequently, suggesting that these have not been explored further because of project history

or because of other reasons such as synthetic difficulties or chemical instability.

The drug rings database should enable complex rings, which would usually be avoided, to be targeted for synthesis and the less complex rings, which are unexplored previously by research groups, to be highlighted.

Ring Geometry. Figure 14 shows a plot of the number of connection points (cleavage points in the original structures) versus the frequency of occurrence. It can be seen that the two- and three-connection rings are the most frequently occurring.

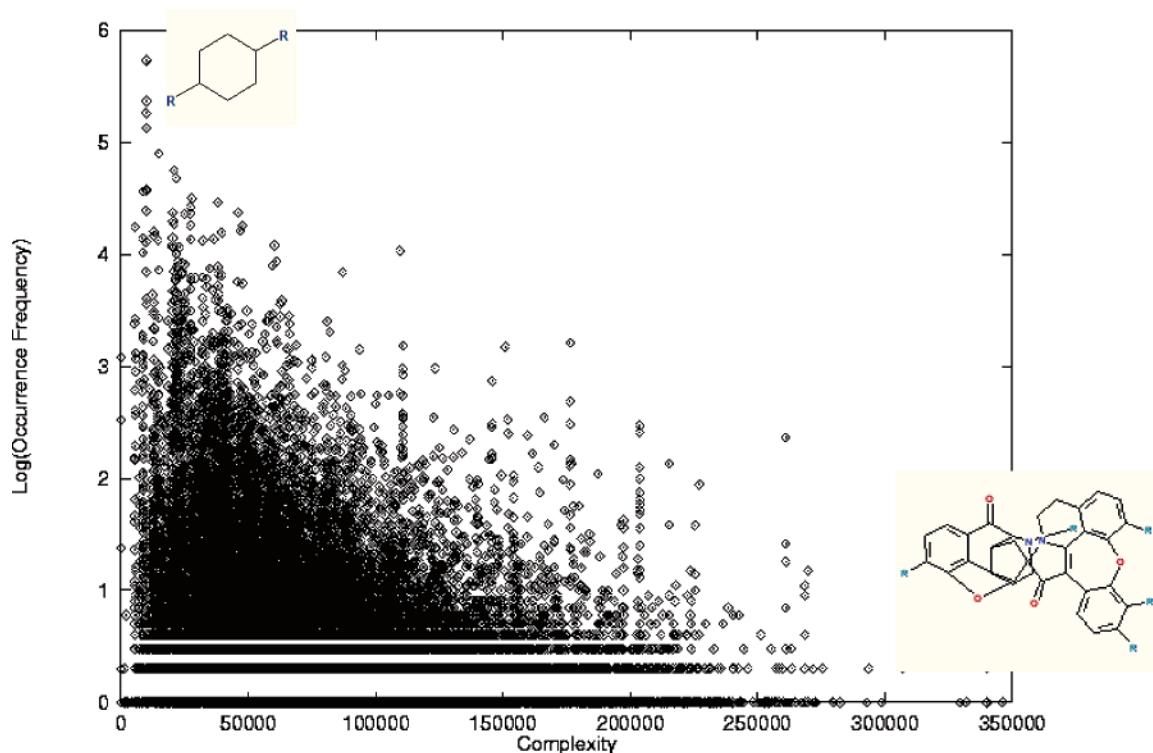
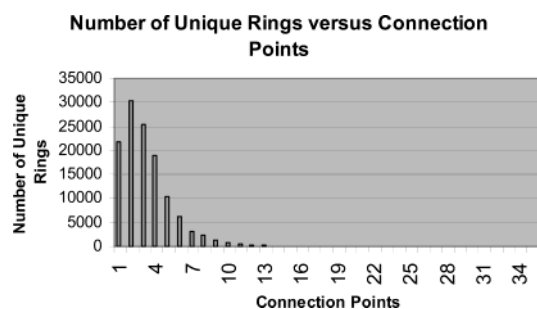
For the reasons stated earlier, the two- and three-connection rings are most likely to be useful as geometric scaffolds for alternative ring replacement. Both CONCORD and CORINA were used to calculate the geometric parameters for these rings to give the user a choice of method for ring geometry. In practice, CONCORD handled fewer ring structures in 3D generation than CORINA, failing at rings where there are special atom types or flexible macrocycles. Table 5 shows the conversion rate in both CONCORD and CORINA.

Inspection of the geometries generated by both methods showed that the parametrizations are slightly different for bond lengths and bond angles and significantly different for torsion angles, resulting in significant differences for vector angles for some of the rings. This is shown in Figure 15 for CONCORD versus CORINA in connection distances and vector angles for two-connection rings. The three-connection geometry correlation also shows a similar trend in that there are significant differences when the vector angles generated by the two methods are compared (data not shown here). The implication of this for scaffold selection is that the user could choose to use the same method for consistency, or use both methods to expand the geometric coverage.

One useful analysis on the geometric descriptors obtained is the comparison of different databases where the nature of the collection may dictate the types of rings and the associated geometries. In particular, a comparison of rings in the Corporate Registry with those of typical drug collections such as PJB may shed light on differences and geometric regions where we could focus our ring design efforts. For example, for two-connection rings, there are high-concentration regions where rings are made in both Corporate and PJB databases. Figure 16 shows a plot of the distance versus vector angle for Corporate Registry and PJB. Two observations are made. First, the Corporate Registry covers a larger geometric range than PJB and covers all of the ring geometries occupied by PJB. Second, the normalized frequency distributions (the normalized frequency is obtained by binning the frequency of rings in the two-dimensional space of distance versus vector angle and then normalizing the largest binned frequency value to 1.0 for each database) comparison between Corporate Registry and PJB suggests a higher relative population for PJB rings in several regions with distance and vector angle of, for example, around 4 Å and 80°, 120°, 140°, and 5 Å and 150°. Further visual inspection of these regions may indicate rationales for the higher relative proportion of rings in PJB compared to Corporate Registry and may provide a steer to the types of rings that are potentially favored as druglike.

Table 4. Overlap of Eight Databases for Rings

	Corporate	BACD	PJB	NCI	DNP	WDI	Medchem	ACD
% of Corporate rings:	100	41.3	5.3	7.4	9.5	11.5	7.6	18.2
% of BACD rings:	41.9	100	4.2	7.0	7.3	8.6	6.4	22.3
% of PJB rings:	67.3	52.3	100	30.1	31.5	78.8	53.6	44.2
% of NCI rings:	43.6	40.7	13.8	100	22.5	26.8	17.9	28.3
% of DNP rings:	19.6	14.8	5.1	7.8	100	25.6	6.3	9.6
% of WDI rings:	47.7	35.4	25.5	18.8	51.5	100	27.4	30.0
% of Medchem rings:	71.6	59.6	39.4	28.5	29.1	62.4	100	51.0
% of ACD rings:	68.3	84.0	12.9	18.0	17.5	25.3	20.3	100

**Figure 13.** The log of the frequency of ring occurrence in the Corporate Registry against ring complexity.**Figure 14.** Number of unique rings versus connection points.**Table 5.** 3D Conversion Rate for Rings Using CONCORD and CORINA

no. of rings	CONCORD conversion (success rate)	CORINA conversion (success rate)
two-connection: 30388	27948 (92.0%)	30261 (99.6%)
three-connection: 25291	23484 (92.9%)	25194 (99.6%)

(It is emphasized that the absolute number of rings in the regions described is still higher in the Corporate Registry because of the much larger database size of the Corporate Registry and the larger absolute numbers of rings in the Corporate Registry). Likewise, the similarities and differences for the three-connection

geometries could also be explored to provide opportunities for the design of ring scaffolds. In the case of the three-connection rings, the six canonicalized geometric parameters could be analyzed by data reduction techniques such as principal component analysis. Alternatively, each pair of the three sets of distance/angle pair could be treated independently from each other for graphical visualization and comparison with two-connection geometries (Figure 16). In this case, one point on the graph represents the presence of one geometric paring for a ring. There are also two other points somewhere on the graph representing the same ring.

Application of the Drug Rings Database

The drug rings database can be viewed as an idea generator for use in lead generation and optimization programs where a lead ring is to be replaced by an alternative ring for potency, ADME, or other reasons. We use two retrospective examples to illustrate the use of the database, sometimes in conjunction with other techniques, to arrive at potential ring or scaffold replacements.

Case 1. Geometric Templates for Indole Replacement. We first demonstrate the geometric capability of the drug rings database for searching for ideas for geometric templates.

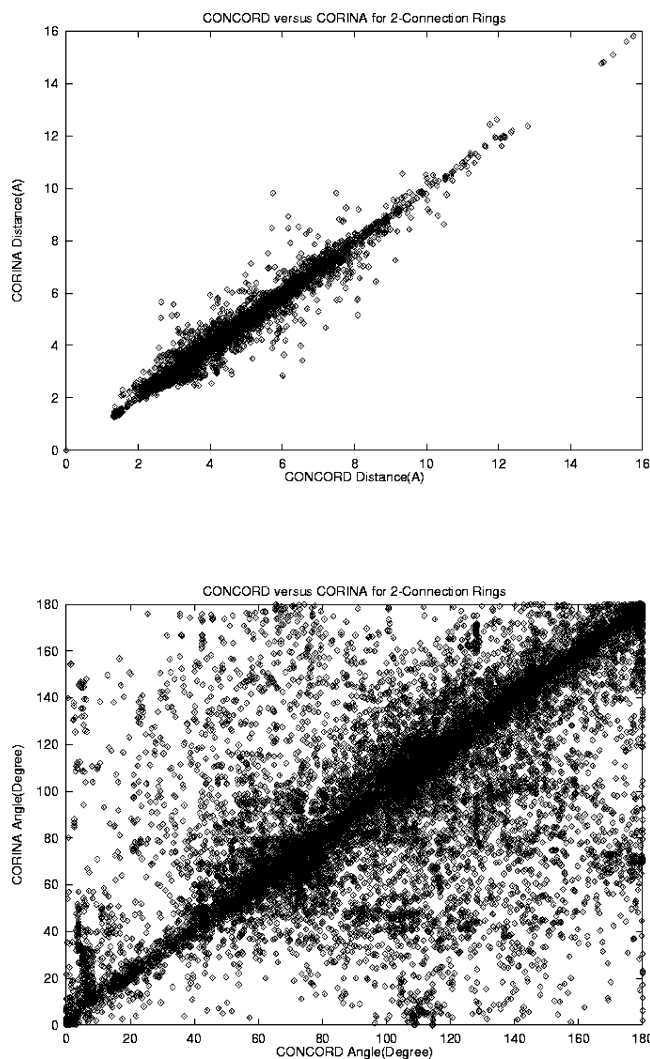


Figure 15. Correlation between CONCORD and CORINA connection distances and vector angles for the two-connection rings.

Indole is a commonly occurring motif among drug molecules, typified by, for example, ligands acting at the 5-HT receptors. The 1,3,5-substituted indole-containing molecules have been claimed to act via several pharmacological mechanisms, including antagonizing 5HT₂ and leukotriene D₄ receptors (Figure 17). Since the mechanisms are diverse and unrelated, one can postulate that the indole core was acting as a template providing the right geometric extensions for the side chains. If this is the case, is it possible to identify alternative templates from the drug rings database to replace the 1,3,5-indole as alternative cores when designing ligands acting on these receptors?

Figures 4–9 show such a hypothetical exercise. Starting from the 1,3,5-substituted indole, we step through a similarity search to obtain stored information on the starting template, including the geometric parameters, to search for those templates that contain the same geometry as the starting template, and then to retrieve the original molecules containing these alternative templates from the individual original databases. The idea here is that once the alternative templates have been identified, it is desirable to look at the molecular environments of the templates and derive further data

on, for example, potential synthetic routes. Note that we have restricted the search to templates that must occur in the PJB2000 database. This has identified 11 templates that contain the right geometry within the tolerance given. If we extend the search to rings occurring in any of the originating databases, we can identify 113 templates satisfying the geometry.

From this exercise, the 1,3,6-substituted indole was identified as a likely candidate for geometric replacement of the 1,3,5-indole. Recalling that the objective of the exercise was to identify geometric replacements in the hope that the core merely acts as a geometric linker, the measure of success of such an assumption in a real situation would be to make the compound and test for activity. Should this be the case in the 1,3,5-indole case, we would be looking for activities in the 1,3,6-substituted compounds. This indeed is the case.

The 1,3,5-substituted indole, 1-(3-{4-[5-chloro-1-(4-fluorophenyl)-1*H*-indol-3-yl]piperidin-1-yl}propyl)imidazolidin-2-one, is active as a 5HT₂ antagonist.²⁷ Its geometric isostere, 1,3,6-substituted indole, 1-(3-{4-[6-chloro-3-(4-fluorophenyl)indol-1-yl]piperidin-1-yl}propyl)imidazolidin-2-one, is also a 5HT₂ antagonist,²⁸ suggesting that the indole is acting as a geometric core rather than the property of the indole per se being important (Figure 18a).

In the case of leukotriene D₄ antagonist, the indole moiety also potentially acts as a geometric core. This is supported by the fact that the 3,5-disubstituted indole, {3-[2-methoxy-4-(toluene-2-sulfonylaminocarbonyl)benzyl]-1*H*-indol-5-yl}carbamic acid cyclopentyl ester,²⁹ and its geometric equivalent, the 1,6-disubstituted indole {1-[2-methoxy-4-(toluene-2-sulfonylaminocarbonyl)benzyl]-1*H*-indol-6-yl}carbamic acid bicyclo[2.2.1]hept-2-yl ester,³⁰ both act as leukotriene D₄ antagonists (Figure 18b).

Since the occurrence frequencies in the original databases were recorded, we can get an indication of the likely synthetic accessibility and any potential intellectual property barriers. Once an alternative core is identified to be of interest, the Web interface takes a user directly to the original database, retrieving structural and other associated information for those molecules from which the ring was originally derived. Chemists can then choose to follow up on the synthetic route of the ring from these original molecular data. Alternatively, other prioritization methods can be employed outside the drug rings database to rank order potential hits for synthetic priorities.

Case 2. Bioisosteres for Endothelin Antagonists.

As well as identifying new core templates, the drug rings database can be used for traditional isosteric replacement of rings. This is illustrated with an example of the endothelin antagonists.

Endothelins are among the most potent endogenous peptide vasoconstrictors known. Intense efforts have been devoted to the discovery of non-peptide antagonists with therapeutic potentials in vasoconstrictive diseases. Mederski et al.⁹ have described a series of endothelin antagonists where the methylenedioxy group was replaced by several bioisosteres such as benzothiadiazole (Figure 19). Benzothiadiazole as a suitable bioisostere for the methylenedioxy group was arrived at from a study performed on a subset of potential isosteres by

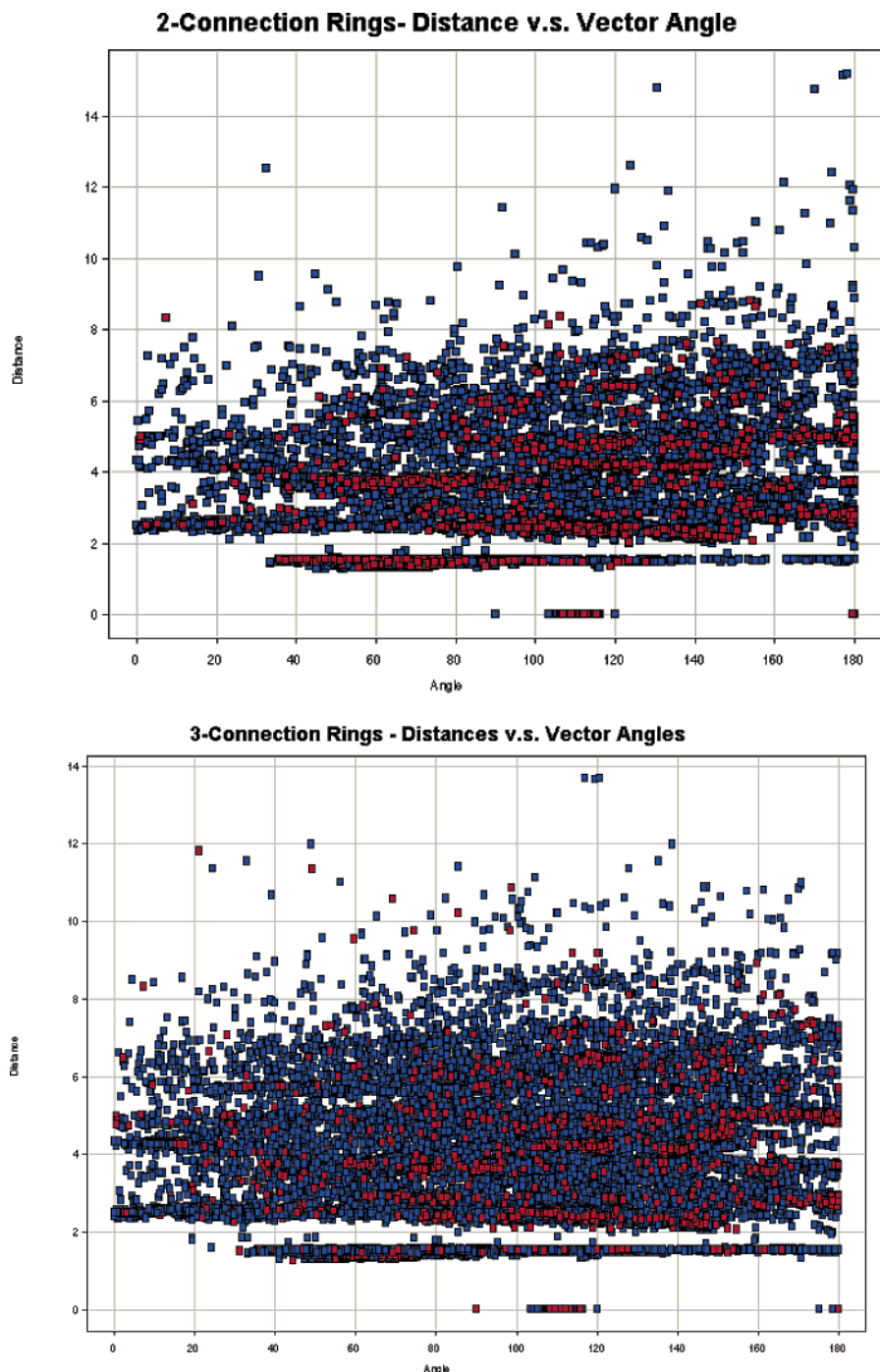


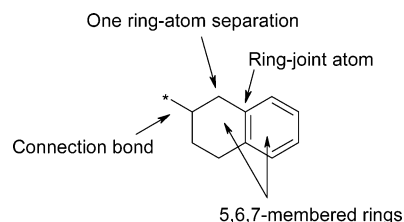
Figure 16. Distance versus vector angle for two- and three-connection rings occurring in Corporate Registry and PJB: (red) PJB; (blue) Corporate Registry.

the original authors.³¹ A Kohonen network was used to rank/prioritize the potential set shown in Table 6. These authors did not have access to all the ring systems that may be suitable for such programs.

Armed with the enhanced resource available through our drug rings database, we have undertaken an analysis to identify alternative potential isosteres to the methylenedioxyphenyl moiety.

Starting from a list of 21 688 potential rings with one connection from the drug rings database, filters of bicyclic templates containing one to two aromatic rings, a “meta” connection to the ring-joint atom (i.e., the connection bond needs to be separated from the ring-

joint atom of the bicycle by one ring-atom),



and already synthesized in the GlaxoSmithKline Corporate Registry were applied. This has filtered down to 567 bicyclic rings that potentially could replace the

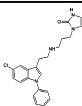
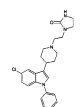
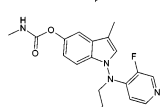
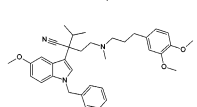
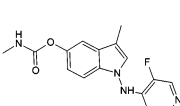
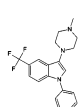
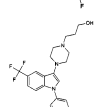
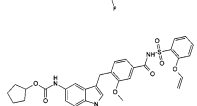
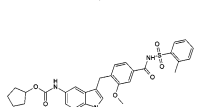
Structures	Pharmacological Activity	Originator	Name
	5-Hydroxytryptamine 2 antagonist	Lundbeck	CERULEN
	5-Hydroxytryptamine 2 antagonist	Lundbeck	S1991
	Adrenoreceptor agonist	Hoechst	BELATIN
	Calcium channel antagonist	Beaufour-Ipsen	BIIP20XX
	Cholinesterase inhibitor	Hoechst Marion Roussel	FR172357
	Dopamine D2 antagonist	Lundbeck	SCH34164
	Dopamine antagonist	Lundbeck	AY30468
	Leukotriene D4 antagonist	Ciba-Geigy	AG555
	Leukotriene D4 antagonist	AstraZeneca	ACCOLEIT

Figure 17. 1,3,5-Substituted indole-containing molecules that act via several pharmacological mechanisms. Data derived from the PJB2000¹ database.

methylenedioxyphenyl ring. The list is then ranked according to combined electronic, shape, and lipophilic similarity to the target ring using the ASP³² method within TSAR.³³

The ASP method essentially computes similarity among a set of molecules with prealigned orientation and conformation or aligns molecules such that the similarity score is maximized and computed. The similarity score is computed according to the formula

$$C_{AB} = \frac{w_S S_{AB} + w_Q Q_{AB} + w_L L_{AB}}{w_S + w_Q + w_L}$$

where S_{AB} , Q_{AB} , L_{AB} are shape, electrostatic, and lipophilicity similarity indices calculated when molecules A and B are compared. w_S , w_Q , w_L are user-defined weighting factors for shape, electrostatic, and lipophilicity, respectively. Algorithms for calculating shape, electrostatic, and lipophilicity similarity scores can be found in ref 32.

When equal weighting is applied to shape, electrostatic, and lipophilicity, we obtain a combined similarity

score from the three individual properties for the potential isosteres. Examples of top and bottom ranking isosteres for the methylenedioxyphenyl ring are shown in Table 7. Table 7 suggests further isosteres in addition to those identified by Anzali (Table 6). This is not surprising because we have a much larger pool of potential isosteres to start with derived from the drug rings database. In general, the highly functionalized isostere is predicted to be a poor substitute, while isosteres with more lipophilic moieties are predicted to be better substitutions. It is also of interest and somewhat surprising to observe that several perceived commonly occurring rings such as benzofuran, shown in the top ranking isosteres list, actually occur less frequently than expected. In real medicinal chemistry experiments, it is still important to evaluate whether any of the top ranking rings would be a suitable substitution for the methylenedioxyphenyl to provide equivalent or better biological activities in any given project (and we have not done that!).

It is of interest to compare and contrast the different results from our current evaluation of isosteres with the

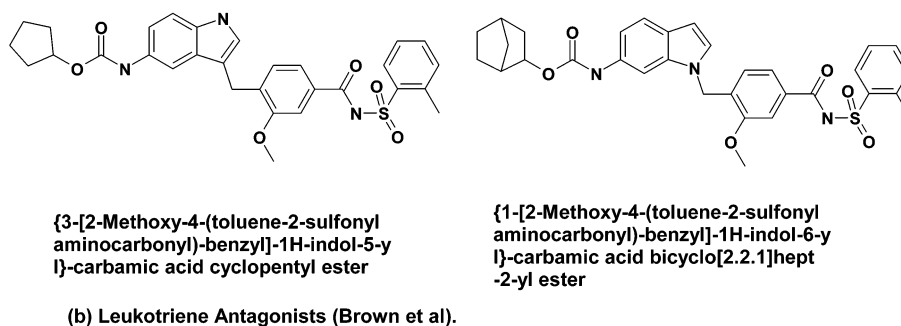
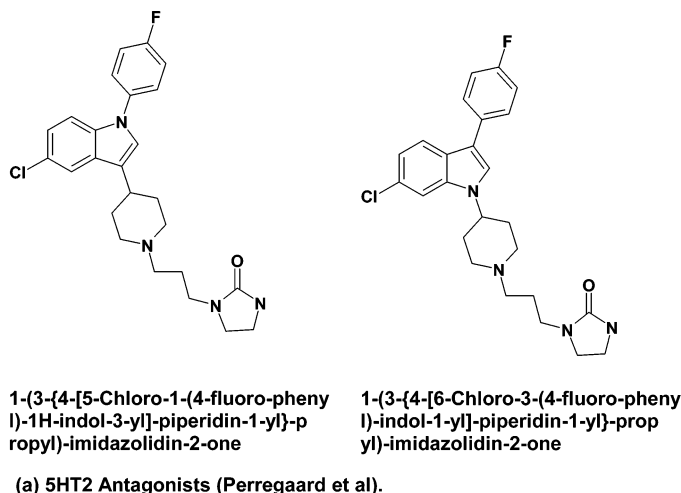


Figure 18. Indole acts as a geometric core in 5HT2 and leukotriene D4 antagonists.

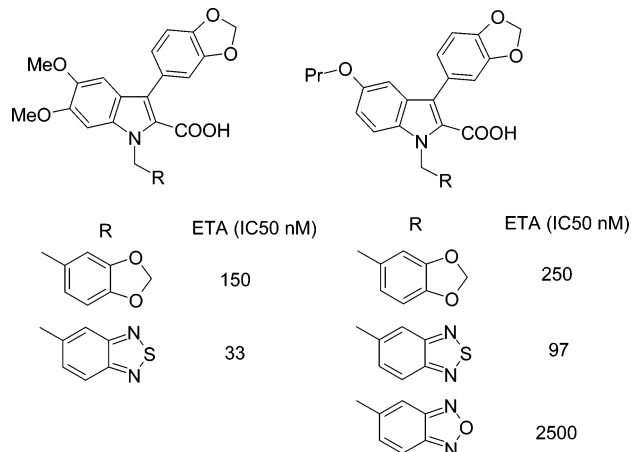


Figure 19. Endothelin antagonists with ETA potency (Mederski et al.). Kohonen network was used by Anzali to suggest relative rank orders within the six isosteres. TSAR rank is derived from this work, which ranks among a set of 567 potential isosteres (see text for details).

results from Anzali. Anzali used the Kohonen network method on a limited subset of potential isosteres. The method is a neuronetwork method involving training and data reduction from the property of the isosteres. Only the electronic similarity using PM3 charges was considered in their calculations. The nature of the Kohonen network also permits the nonalignment of the isosteres. We have used AM1 charges and aligned molecules to a fixed orientation to the target. The properties considered were electronic, shape, and lipophilic property values at grid points around molecules. Thus, not surprisingly, the results are somewhat dif-

Table 6. Similarity Ranking to the Methyleneedioxy Phenyl Group^a

Isostere	Kohonen Rank (Anzali et al)	Isostere	TSAR Rank from this work
	Identity (Highest similarity)		1 (Identity)
			30
			62
			168
			196
	(Lowest similarity)		442 (Lower similarity)

^a Similarity decreases going down the list.

ferent. Both approaches were expected to provide complementary and alternative views on potential isosteres.

The lessons from the case studies are the following. The drug rings database provides a valuable knowledge base to seek ideas for alternative rings that either act as a geometric scaffold or act as potential isosteres to a lead motif. Although geometric fit of core templates discussed in case 1 is a good start for ring replacement, the consideration of shape, electronic, and lipophilic similarities will further enhance our chance of success for such isosteric replacement experiments. The past frequency of occurrence of rings for a corporate collection provides both opportunities and challenges to explore those motifs that have not been explored very often thus far.

Table 7. Examples of Top and Bottom Ranking Isosteres to Methyleneedioxyphenyl Moiety Using TSAR Rank^a


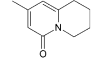

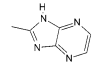

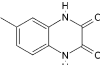

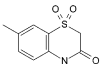

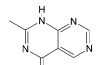

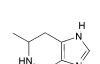

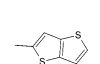

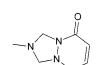

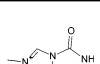

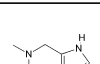

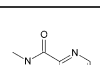

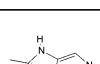

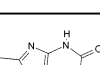

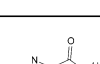

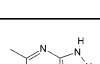




Top Ranking Isosteres	Database Frequency	Similarity to Methylene dioxyphenyl	Bottom Ranking Isosteres	Database Frequency	Similarity to Methylene dioxyphenyl
	23331	1		2	0.456085
	27	0.924931		11	0.444748
	315	0.924096		9	0.440992
	250	0.920501		2	0.421613
	24	0.912438		2	0.318911
	199	0.910108		2	0.260619
	4	0.897681		155	0.259356
	29	0.895655		5	0.222106
	1114	0.894562		10	0.2161
	2	0.893568		12	0.197168
	33	0.892489		2	0.185298
	3	0.890094		20	0.167272
	10	0.875132		2	0.144172
	411	0.874482		50	0.117055
	17	0.868954		10	0.09869

Table 7. (Continued)

Top Ranking Isosteres	Database Frequency	Similarity to Methylene dioxyphenyl	Bottom Ranking Isosteres	Database Frequency	Similarity to Methylene dioxyphenyl
	41	0.866446		8	0.094799
	2	0.862301		3	0.072595

^a Frequency is the occurrence in the Corporate Database, and similarity is a score derived from combining electronic, shape, and lipophilic similarities for the target.

Conclusions and Future

This paper describes the development of a chemist-friendly, value-added knowledge base of drug rings. The 120 K rings were obtained from a variety of major sources including both corporate and external databases. Simple yet effective geometric descriptors were developed to enable an easy search for geometric isosteres. Frequencies of occurrence in the original databases gave an indication of synthetic accessibility and potential intellectual property barriers. Other computed descriptors were also generated for more targeted searching.

Retrospective examples of substituted indole as geometric scaffolds in the 5HT₂ antagonist area and methylenedioxyphenyl isostere replacements in the endothelin antagonist area are given for potential applications of such a database. We believe that the database and its Web interface will prove to be a valuable idea generator for medicinal chemists for lead generation and optimization programs where a ring is required to be replaced by an alternative.

One caveat of simple geometric parameters described in this paper is that they cannot completely specify the geometry of the ring. Further parameters such as the improper torsion angle would help to describe the geometry more precisely at the expense of simplicity. We anticipate future development will include a more complete geometric description and a geometric fingerprint to enable similarity searching on all geometries instead of just two- and three-connection rings. It is also desirable to include all isomers of the ring to extend geometric coverage because different isomers would potentially result in different 3D geometry. With the availability of the Daylight Oracle cartridge technology, the plan is now to migrate the database into the relational Oracle environment, allowing close integration with other in-house systems. This would allow a more automatic updating mechanism for continued enhancement of the database to capture and utilize the knowledge created in our lead discovery programs.

Acknowledgment. We acknowledge the support provided by our chemistry and computational chemistry colleagues in the evaluation of the system, Drs. Mike Hann and Drake Eggleston for comments and support, and Jeremy Yang and Daylight for providing support and guidance relating to Daylight tools.

Appendix

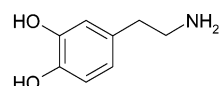
Structural Complexity Descriptor Based on Daylight Fingerprint. The concept of structural complexity may be represented by the density of its structural fingerprint if such a fingerprint encodes the information content of the structure, as is the case with the Daylight fingerprint. Thus, if we use the DAYLIGHT fingerprint of 1024 bits, each unique atom path in a molecule corresponds to several bits that are turned on. The more bits that are turned on, the more unique atom paths occur in a structure that should correspond to an increase in molecular complexity. By simple summation of the number of bits on, it is possible to describe how complex a molecule is.

We have developed an extension to this simple concept of complexity descriptor that sums the positions of the bits turned on as defined below, the reason being that structures sorted by this descriptor group locally similar structures together and make the results easier to visualize.

In the case of the DAYLIGHT fingerprint, the complexity descriptor is defined as

$$\text{complexity descriptor} = \text{sum of bit positions where bits on turned on}$$

The following example illustrates this.



↓ Daylight Fingerprint

0 1 0 1 1 1 0 0 1 01 Binary Bits
1 2 3 4 5 6 7 8 9 10.....1024 Position of Bits

$$\text{Complexity Descriptor} = 2+4+5+6+9+\dots+1024 = 1200 \text{ (e.g.)}$$

In practice, the complexity descriptor thus obtained correlates well with the number of bits on and is a good indicator of structural complexity.

References

- (1) PJB Publications Ltd., London. Web site: <http://www.pjbpubs.com/>.
- (2) Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. A ring-based chemical structural query system: use of a novel ring-complexity heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (1), 65–68.
- (3) Lipkus, A. H. A Ring-Imbedding Index and Its Use in Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 92–97.

- (4) Lipkus, A. H. Mining a large database for peptidomimetic ring structures using a topological index. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (3), 582–586.
- (5) Lipkus, A. H. Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 430–438.
- (6) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: new ideas, concepts, and tools. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 447–452.
- (7) Hazeldine, S. T.; Polin, L.; Kushner, J.; White, K.; Bougeois, N. M.; Crantz, B.; Palomino, E.; Corbett, T. H.; Horwitz, J. P. Synthesis and Biological Evaluation of Some Bioisosteres and Congeners of the Antitumor Agent, 2-[4-[(7-Chloro-2-quinoxalinyloxy]phenoxy]propionic Acid (XK469). *J. Med. Chem.* **2002**, *45* (14), 3130–3137.
- (8) Gohlke, H.; Guendisch, D.; Schwarz, S.; Seitz, G.; Tilotta, M. C.; Wegge, T. Synthesis and Nicotinic Binding Studies on Enantiopure Diazine Analogues of the Novel (2-Chloro-5-pyridyl)-9-azabicyclo[4.2.1]non-2-ene UB-165. *J. Med. Chem.* **2002**, *45* (5), 1064–1072.
- (9) Mederski, W. W. K. R.; Osswald, M.; Dorsch, D.; Anzali, S.; Christadler, M.; Schmitges, C.-J.; Wilm, C. 2. Endothelin antagonists: evaluation of 2,1,3-benzothiadiazole as a methylenedioxyphenyl bioisostere. *Bioorg. Med. Chem. Lett.* **1998**, *8* (1), 17–22.
- (10) van Vliet, L. A.; Rodenhuis, N.; Dijkstra, D.; Wikstrom, H.; Pugsley, T. A.; Serpa, K. A.; Meltzer, L. T.; Heffner, T. G.; Wise, L. D.; Lajiness, M. E.; Huff, R. M.; Svensson, K.; Sundell, S.; Lundmark, M. Synthesis and pharmacological evaluation of thiopyran analogues of the dopamine D3 receptor-selective agonist (4a*R*,10b*R*)-(+)-*trans*-3,4,4a,10b-tetrahydro-4-*n*-propyl-2*H*,5*H* [1]benzopyrano[4,3-*b*]-1,4-oxazin-9-ol (PD 128907). *J. Med. Chem.* **2000**, *43* (15), 2871–2882.
- (11) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 103–108.
- (12) *Molecular Design Drug Data Report*, Version 99.1; Molecular Design Ltd.: San Leandro, CA.
- (13) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.
- (14) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42* (25), 5095–5099.
- (15) Daylight Chemical Information Systems Inc. Web site: <http://www.daylight.com>.
- (16) NCI96, National Cancer Institute Database. Web site: <http://www.nci.nih.gov/>. NCI96 is a version supplied via Daylight Inc.
- (17) DNP102, Derwent Dictionary of Natural Product. Web site: <http://www.derwent.com/>.
- (18) WDI, Derwent World Drug Index. Web site: <http://www.derwent.com/worlddrugindex/index.html>.
- (19) Mechem02, Medchem database, Pomona College and BioByte Corp., Claremont, CA. Database distributed by Daylight.
- (20) ACD014, MDL's Available Chemicals Directory. Web site: <http://www.mdli.com/products/acd.html>.
- (21) Lauri, G.; Bartlett, P. A. CAVEAT: a program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8* (1), 51–66.
- (22) Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterizing the geometric diversity of functional groups in chemical databases. *J. Comput.-Aided Mol. Des.* **1995**, *9* (5), 417–424.
- (23) Bohl, M.; Dunbar, J.; Gifford, E. M.; Heritage, T.; Wild, D. J.; Willett, P.; Wilton, D. J. Scaffold searching: Automated identification of similar ring systems for the design of combinatorial libraries. *Quant. Struct.-Act. Relat.* **2002**, *21*(6), 590–597.
- (24) Farmer, P. S., Ed. *Arens Drug Design*; Academic Press: New York, 1980; Vol. 10, pp 119–143.
- (25) Rusinko, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *CONCORD: a program for the rapid generation of high quality approximate 3D structures* (developed at University of Texas at Austin); Version 4.0.4, Tripos Associates Inc., St. Louis, MO 63144.
- (26) Corina-Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3* (6c), 537–547.
- (27) Perregaard, J.; Arnt, J.; Boegesoe, K. P.; Hyttel, J.; Sanchez, C. Noncataleptogenic, centrally acting dopamine D-2 and serotonin 5-HT2 antagonists within a series of 3-substituted 1-(4-fluorophenyl)-1*H*-indoles. *J. Med. Chem.* **1992**, *35* (6), 1092–1101.
- (28) Andersen, K.; Perregaard, J.; Arn, J.; Nielsen, J. B.; Begtrup, M. Selective, centrally acting serotonin 5-HT2 antagonists. 2. Substituted 3-(4-fluorophenyl)-1*H*-indoles. *J. Med. Chem.* **1992**, *35* (26), 4823–4831.
- (29) Brown, M. F.; Marfat, A.; Antognoli, G.; Chambers, R. J.; Cheng, J. B.; Damon, D. B.; Liston, T. E.; McGlynn, M. A.; O'Sullivan, S. P.; Owens, B. S.; Pillar, J. S.; Shirley, J. T.; Watson, J. W. *N*-Carbamoyl analogs of Zafirlukast: potent receptor antagonists of leukotriene D4. *Bioorg. Med. Chem. Lett.* **1998**, *8* (18), 2451–2456.
- (30) Brown, M. F. Preparation of bicyclic carbamates as LTD4 antagonists. U.S. Patent 5,439,929, 1995 (A 19,950,808; CAN 124:55946; AN 1995:793003 CAPLUS).
- (31) Anzali, S.; Mederski, W. W. K. R.; Osswald, M.; Dorsch, D. 1. Endothelin antagonists: search for surrogates of methylenedioxyphenyl by means of a Kohonen neural network. *Bioorg. Med. Chem. Lett.* **1998**, *8* (1), 11–16.
- (32) *ASP 3.2 User Guide*, 1997, Oxford Molecular Group (now Accelrys Inc.), Oxford, England. See references contained therein. For example, see the following. Burt, C.; Richards, W. G.; Huxley, P. The Application of Molecular Similarity Calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.
- (33) TSAR, a QSAR product from Accelrys Inc. Web site: <http://www.accelrys.com/>.

JM0300429