Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 23 (2005) 51–58
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

# Selection of heterocycles for drug design

Howard B. Broughton [a,*], Ian A. Watson [b]

[a] Lilly S.A., Avda. de la Industria, 30 Alcobendas, 28108 Madrid, Spain
[b] Eli Lilly Corporate Center, 355 E Merrill Street, Indianapolis, IN 46225, USA

Received 4 September 2003; received in revised form 24 February 2004; accepted 4 March 2004
Available online 10 May 2004

**Abstract**

A method has been devised to obtain heterocyclic ring systems suitable for use in drug design and library design, with an emphasis on the selection of systems with good absorption, distribution, metabolism, excretion and toxicity (ADMET) properties in man. This has been achieved by extraction of the ring systems found in drugs that have reached Phase II or later stages of drug development and launch. Properties have been calculated for these ring systems to enable them to be rationally selected from the database, including descriptors based on molecular size, shape, hydrogen bonding and orbital properties. In many cases, the properties have been calculated for different attachment points of the same heterocycle. Principal components analysis has been used to enable visualization of the set of heterocycles in a useful ''chemical space''. Using this space, it is possible to select heterocycles for drug design to explore specific aspects of the properties of the heterocycle, such as size or hydrogen bonding, while maintaining other parameters near constant, or to select heterocycles with extreme values of these properties but which are nonetheless likely to be acceptable in a drug. The differences between the properties calculated for the most- and least-frequently used heterocycles from the late-phase drug set have been analyzed, and may suggest that heterocycles in successful drugs are more likely to have calculated quantities associated with lower chemical reactivity.


*Keywords:* Heterocycle; Diversity; Similarity; Library design; Descriptors; Properties; Known drugs; Ring systems; ADME; ADMET; Absorption; Distribution; Metabolism; Excretion; Toxicity

## 1. Introduction

The critical role played by heterocycles in drug design cannot be denied. Even where the natural substrate or ligand for a biological target does not contain a heterocycle, drugs—whether of natural or man-made origin—that act on that target frequently contain heterocyclic groups. These may mimic the heterocycles found in the natural ligands or substrates, or they may mimic other functional groups, such as the amides of a peptide ligand. In this latter case, the heterocycle often confers stability on a ligand which would otherwise be rapidly degraded in vivo, or, equally important, enables the ligand to be absorbed from the gut and/or to penetrate the blood–brain barrier.

Modern synthetic chemistry provides the drug designer with a vast range of possible heterocycles—even consideration of the simplest mono- and di-nuclear aromatic species with a limited vocabulary of atoms—C, N, O and S—provides tens of thousands of ''chemically reasonable'' systems (exactly how many depending upon the definition of ''reasonable'' used), most of which can be substituted at various positions to produce a bewildering array of possible structures. Very many of these possible systems are already known to synthesis, and much of the remainder would appear amenable to synthesis by adaptation of existing techniques. Thus, the choice of the drug designer is not significantly constrained by synthetic considerations. Where a heterocycle is being introduced into a ligand as a bioisostere, the key considerations in choice of heterocycle are thus to maintain or improve the activity of the ligand while improving its absorption, brain penetration and metabolic stability characteristics. This is increasingly a critical aspect even in the early stages of drug design [1].

In addition to their use in the ''evolutionary'' drug design process in which existing molecules are modified by

* Corresponding author. Tel.: +34-91-623-3337; fax: +34-91-663-3411.
*E-mail address:* broughton_howard@lilly.com (H.B. Broughton).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2004.03.016

introduction or exchange of heterocyclic moieties, heterocycles are also frequently the ''core'' structure of combinatorial library designs [2]. In this case, it is important for the library designer to select a viable core that has a good chance of providing good ADMET properties to the final library. It may also be important for the library designer to select a heterocyclic core that is either very similar to a known core, or to select heterocycles as cores that are diverse with respect to one another.

The importance of ring systems in drugs has been widely recognized, and has led to the analysis of ring systems (and other components of a drug molecule) in the context of understanding what is a drug-like molecule [3,4]. However, though such analyses have been found to be valuable in the construction of combinatorial libraries [5], they have neither continued to consider which heterocycles are more likely to have good ADMET properties, nor have they been able to quantitate the similarity and diversity of the heterocycles found beyond the aspect of overall shape [6], or topology [7]. Nonetheless, the importance of heterocyclic ring systems in drug design is clear from the priority given to them by chemists. This paper describes a method for identification of ring systems suitable for use in drug design and library design, and how these heterocycles have been associated with descriptors to enable suitable choices to be made. In essence, drugs that have reached Phase II or higher in clinical trials may be considered to have been demonstrated to have ADMET properties worthy of further interest, and therefore, any ring system contained within such drug molecules, at least in some contexts, is probably associated with good ADMET properties. By the term ''good ADMET properties'' we explicitly wish to include properties such as low toxicity or long plasma half-life, even for drugs not designed for oral administration. Our methods could be adapted to the design exclusively of oral drugs, but that was not our objective: not all targets of interest require oral drugs, and indeed oral activity is not a prerequisite during all the stages of drug discovery and research. Descriptors previously described by McGuire and co-workers [8] were adapted to determine the similarity or diversity of heterocycles and can be applied to this set of compounds to enable suitable choices to be made for drug and library design purposes, in conjunction with the frequency with which each heterocycle is found in the late-phase drug molecules studied here.

## 2. Methods

Compounds were selected using ISIS/Base from the MDDR database (MDDR-3D 2001.1 (23.05)) [9] that have reached at least Phase II of clinical trials by searching the ''PHASE'' database field with the query like ''*%phase%II%*'' and combining the resulting list of compounds with that obtained with the query like ''*%launched%*'', used against the same field. Structures were then selected using the query

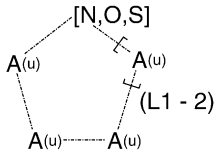


Fig. 1. Query used to retrieve heteroaromatic-containing molecules.

shown in Fig. 1, such that only molecules containing five- or six-membered heteroaromatic systems were retrieved. The MDDR may not be a complete reference set, since it dates only from the mid-1980s, but nonetheless for this purpose was felt to be a suitable source of molecules. Many heterocycles used in older drugs have been reused in more recent (and often, from the ADMET perspective, better) compounds.

The structures of the molecules thus identified were exported in SD file format and converted to SMILES format [10] with dbtranslate [11], using the options +chiral -translate daysmiles - type maccs, or with the in-house program fileconv with equivalent options. With the molecules thus in hand in an appropriate format, we proceeded to identify the ring systems present in each molecule, where a ring system may be defined as a group of atoms whose bonds are all in a ring topology. The in-house program smi2rings was used to identify such ring systems, excluding systems with more than four subrings as being probably too complex to use for drug design purposes. Smi2rings was originally developed as part of Lilly's third party small molecule acquisition efforts [12]. The corporate collection was profiled and a database of rings and ring systems built with smi2rings. As molecules proposed for purchase are considered, the rings and ring systems in the new molecules are computed and looked up in the database. Molecules containing completely new rings or ring systems are given a significantly higher probability of being selected, while those with rings or ring systems that are comparatively rare in the existing collection are given proportionately enhanced probabilities of being selected. Some manual intervention in this process is often needed to avoid purchasing molecules containing novel, but nevertheless undesirable heterocycles.

Smi2rings operates in the following way:

1. Read and parse options.
2. Read in SMILES and convert to internal connection table format.
3. Apply smallest-set-of-smallest-rings [13] (SSSR) algorithm to locate rings.
4. Check for further rings fused to the first ring found and join together into a ring system. Mark rings thus considered as ''done''.
5. If appropriate options were given, add atoms in substituents to the ring system definition.
6. Check whether there are any remaining rings not marked as ''done''. Pick one of these and repeat 4, 5 and 6 until no rings remain.

7. Write completed ring systems to a gdbm [14] database file.

As indicated above, in addition to identifying ring systems, smi2rings is able to detect substituents and optionally include some of these still attached to the output ring systems. A query file was used to define those substituents to be retained in this way, and was set to retain $CF_3$, $NO_2$, CN, Cl, F, Br, OMe, OH, $NMe_2$, NHMe, $NH_2$, $CO_2Me$, $CO_2H$, $CONH_2$, $SO_2Me$, SMe, Me, $SO_2NH_2$ and the first atom of any other substituent attached through a non-ring bond (including a double bond). The gdbm output of smi2rings was then converted to SMILES format (in the process, the number of times each ring system was found in the molecules from the MDDR was appended as identifier to the SMILES) and filtered through a substructure search to select only those ring systems that contained a heteroaromatic ring system, defined using the SMARTS query [10]:

$$[!c;!C;R]@[R]@;=,:;[R]@[R]@;=,:;[R]$$

The resulting structures were converted to three-dimensional structures using CONCORD v. 4.06 [11]; where chirality was not known for the few chiral structures in the input data, CONCORD was allowed to assign an arbitrary chirality. The descriptors previously reported by McGuire and co-workers [8] suppose that each heterocycle has a point of attachment to the remainder of the molecule, and this point was represented in their work as a methyl group. In principle, each heterocycle could be substituted at any vacant position with a methyl group to provide all possible orientations. However, the purpose of this study was to identify heterocycles that could be incorporated into possible drug molecules with an increased probability of having good ADMET properties. The ADMET properties of heterocycles are known, in some cases, to show variability depending on the substitution pattern around the ring. The method of extracting rings from the original molecules included both the point of attachment to the remainder of the molecule and any methyl substituent present. These were felt to be better choices of attachment point than simply using any vacant position on the core heterocycle. Thus, the 580 ring systems were read into Sybyl v. 6.8 [11] and each of the possible substituent points—a methyl group attached to a heterocycle—was identified using a call to the search2d expression generator of the form:

$$\%\text{search2}d(\%\text{sln}(\$\text{mol\_area})\text{Hev[r]} \sim \text{Hev[r]} - C[\text{HAC} = 1\&!R]\text{NoDup 0 } Y)$$

Each hit was considered in turn, taking care to retain only one of any pair of hits that identified the same attachment point. The Sybyl ''orient'' command was used with the atoms identified in each hit to place the carbon atom of the methyl group corresponding to the attachment point (C1) at the origin, the atom in the ring system to which it was attached (*X*2) along the *X*-axis, and one of the neighbours of *X*2 (*X*3) in the *X*–*Y* plane. Each reoriented molecule was renumbered so that C1 was atom 1, *X*2 atom 2 and *X*3 atom 3, and was then given a unique name by appending the hit number to the molecule name as provided in the input. Renumbering the molecule in this way ensured that during MOPAC geometry optimization the orientation of the molecule would be maintained, and a unique name was generated for each orientation of each heterocycle.

Geometry optimization and initial property calculation was carried out using MOPAC93 [15], with the MNDO Hamiltonian. In addition to the new geometry, the ESP and DIPOLE keywords were used to obtain electrostatic information, the MMOK keyword was used to correct the geometry of any amide moieties present, and the convergence criteria were set to GNORM = 1.0*D*−3 SCFCRT = 1.0*D*−9. The molecular geometry and ESP charges were retrieved from the output and applied to the molecule, and the components of the dipole vector and the HOMO and LUMO energies were also obtained from the output files. The molecular volume was calculated using the Sybyl %volume() expression generator on the optimized geometry. A further section of spl was used to identify the extreme points of the molecule. *X* and *Y* coordinates were obtained for each atom, and the van-der-Waals radii (as defined by the %atom_info($atom_id vdw) expression generator call) were added and (for *Y* coordinates) subtracted from each atom coordinate. In this way, the maximum and minimum *Y* extent of each ring system was identified (vdw width) and the maximum *X* extent (vdw length, since the attachment point was by definition at (0, 0, 0) in the coordinate space) obtained similarly. These correspond closely to the *Vdw_w* and *Vdw_l* parameters from McGuire and co-workers [8], though they are not exactly identical when computed for the heterocycles from that paper. The difference is presumably due to slight differences in the values of van-der-Waals radii used in this work compared to those used previously. The dipole angle used in the previous paper was computed by temporary addition of an atom (*XX*) to the molecule structure at the coordinates corresponding to the *X*, *Y* and *Z* components of the dipole vector from the MOPAC calculation, and then using the Sybyl %angle() expression generator to return the *X*2–C1–*XX* angle.

The remaining descriptors were *C* log *P* (calculated using the standard Daylight software, v 4.71) [11] and the hydrogen-bonding parameters $L_{CHA}$ and $H_{CHA}$, respectively, the lowest charge on a heteroatom in the ring or directly attached to the ring, and the highest charge on a hydrogen attached to a ring. In the present study, the definition is restricted to heteroatoms in or attached to the ring, a restriction that was not necessary in the McGuire and co-workers [8] paper since in that study the only heteroatoms present were in the ring system being studied, while in the present work heteroatom-containing substituents were permitted. $L_{CHA}$ and $H_{CHA}$ were calculated using a method similar to that employed by McGuire and co-workers [8], but

CNDO calculations were done in Gaussian 98 [16] using the CNDO and DIIS keywords.

Data from all of the above calculations were combined into a single text file and read into a sybyl molecular spreadsheet, associated to a database containing the structures (appropriately oriented) of the ring systems. A principal components analysis with four components was then carried out using the Sybyl Factor Analysis tool with no rotation and standard scaling.

## 3. Results

The initial database query identified 918 molecules containing suitable heterocycles that were in Phase II or later stages of development and launch. These 918 molecules contained 580 unique ring systems, the most frequently-identified ring systems being 3-indolyl and 3-pyridyl with 43 hits each. After allowing for different possible points of attachment, the properties of 721 heterocycle orientations were calculated. Note that these also include different tautomers of the same heterocycle—no attempt was made to eliminate duplicate structures of this type, on the basis that the structure provided in the MDDR could, at least in some cases, be drawn with a particular tautomer based upon experimental evidence or theoretical considerations. It is also worth noting that no attempt has been made to cluster the parent structures so as to eliminate the dominance of "me-too" structures, which may skew the frequency distribution of heterocycles—for example, 14 of the 26 tetrazoles found are in highly similar angiotensin AII antagonist molecules. Even with this proviso, the most frequently-found heterocycles are still those that have been most extensively tested and found satisfactory in ADME terms in the human. The top 30 structures found are given in Fig. 2, along with their calculated properties in Table 1 and the four principal components in Table 2. The number in parentheses following each structure in Fig. 2 is the number of different "activity indices" recorded in the MDDR for drugs containing the (substituted) heterocycle in question. Thus, CPD4674_32_R_1, though found in 32 drugs, is only associated with 20 activity indices. Since many compounds in the MDDR are associated with several activity indices (e.g. an insulin sensitizer would be associated not only with index 43130—insulin sensitizers—but also with antidiabetics, index 43100), just 20 activity indices for 32 drugs suggests that this heterocycle only works well in a relatively small class of drugs (in this case, mostly beta lactam antibiotics). Obviously, before selecting such a heterocycle it would be necessary to check that the class(es) of drugs in which it has been used are relevant to the work being undertaken.

It should be noted that the heterocycles shown in Fig. 2 include, in some cases, more than one orientation of the

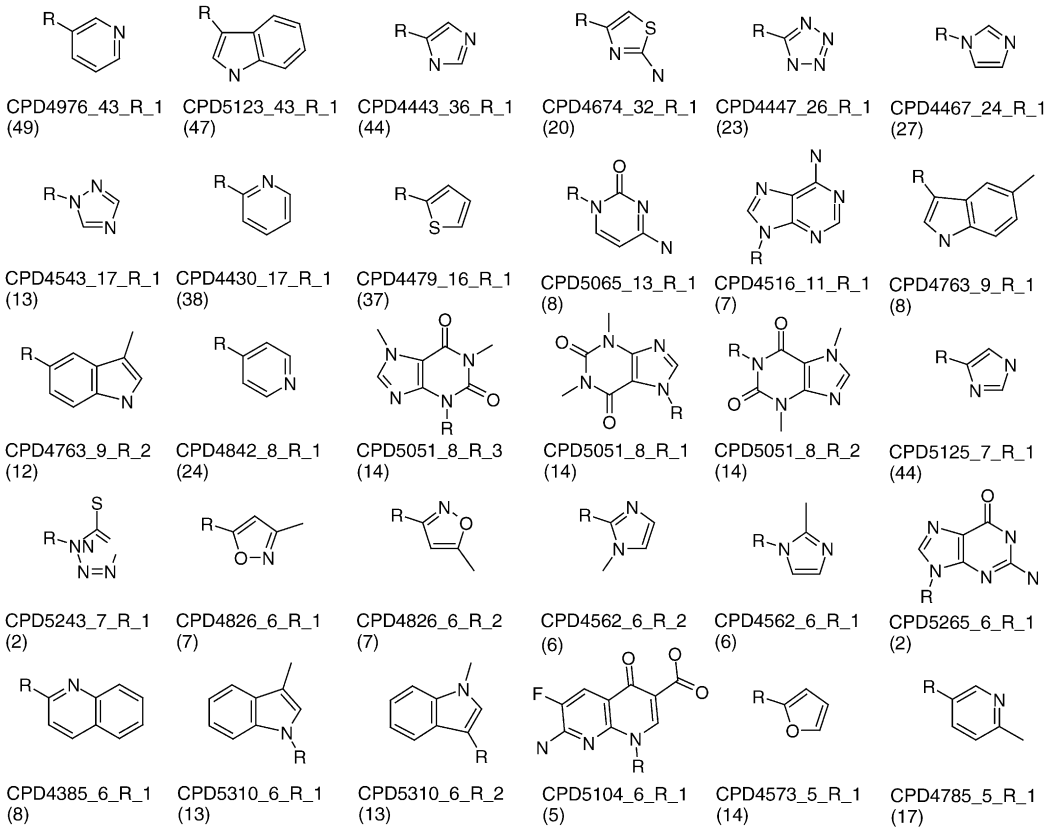


Fig. 2. The 30 most frequently-found heterocycles.

Table 1
Calculated descriptors for the 30 most frequently-found heterocycles

| Compound_ID | Dipole | Dipole angle | Volume | HOMO energy | LUMO energy | $L_{CHA}$ | $H_{CHA}$ | $C \log P$ | VDW Width | VDW Length |
|---|---|---|---|---|---|---|---|---|---|---|
| CPD4976_43_R_1 | 1.93 | 117.50 | 86.30 | −9.58 | −0.09 | −0.20 | 0.03 | 1.14 | 6.52 | 6.45 |
| CPD5123_43_R_1 | 1.90 | 59.12 | 116.60 | −8.31 | 0.16 | −0.30 | 0.16 | 2.63 | 8.85 | 6.93 |
| CPD4443_36_R_1 | 3.35 | 126.87 | 74.30 | −8.99 | 0.76 | −0.29 | 0.17 | 0.24 | 6.37 | 5.49 |
| CPD4674_32_R_1 | 1.56 | 69.20 | 90.40 | −9.30 | −0.32 | −0.33 | 0.06 | 0.73 | 7.28 | 6.22 |
| CPD4447_26_R_1 | 5.19 | 139.22 | 65.70 | −11.18 | −0.22 | −0.21 | 0.17 | −0.83 | 5.71 | 5.08 |
| CPD4467_24_R_1 | 3.64 | 165.08 | 75.00 | −9.05 | 0.85 | −0.22 | 0.03 | −0.01 | 6.44 | 5.61 |
| CPD4543_17_R_1 | 2.98 | 157.09 | 70.00 | −10.16 | 0.48 | −0.23 | 0.04 | −0.26 | 5.84 | 5.38 |
| CPD4430_17_R_1 | 1.85 | 58.70 | 86.50 | −9.57 | −0.06 | −0.22 | 0.03 | 1.14 | 6.61 | 6.45 |
| CPD4479_16_R_1 | 0.55 | 70.12 | 82.60 | −9.38 | −0.07 | 0.05 | 0.05 | 2.29 | 6.35 | 5.85 |
| CPD5065_13_R_1 | 5.45 | 87.94 | 100.60 | −9.43 | −0.38 | −0.40 | 0.05 | −1.32 | 6.94 | 7.20 |
| CPD4516_11_R_1 | 2.74 | 126.84 | 115.50 | −9.06 | −0.34 | −0.36 | 0.03 | −0.40 | 8.64 | 7.77 |
| CPD4763_9_R_1 | 1.84 | 62.86 | 132.70 | −8.38 | 0.13 | −0.29 | 0.16 | 3.13 | 9.77 | 6.92 |
| CPD4763_9_R_2 | 1.84 | 25.67 | 132.70 | −8.38 | 0.13 | −0.29 | 0.16 | 3.13 | 8.67 | 8.00 |
| CPD4842_8_R_1 | 2.03 | 179.84 | 86.00 | −9.68 | −0.01 | −0.21 | 0.03 | 1.14 | 6.50 | 5.83 |
| CPD5051_8_R_3 | 3.89 | 76.79 | 150.30 | −9.13 | −0.39 | −0.38 | 0.04 | −0.04 | 9.76 | 7.00 |
| CPD5051_8_R_1 | 3.89 | 125.60 | 150.40 | −9.13 | −0.39 | −0.38 | 0.04 | −0.04 | 9.64 | 7.93 |
| CPD5051_8_R_2 | 3.89 | 45.35 | 152.70 | −9.13 | −0.39 | −0.38 | 0.04 | −0.04 | 9.55 | 7.89 |
| CPD5125_7_R_1 | 3.50 | 44.61 | 73.90 | −9.00 | 0.86 | −0.27 | 0.17 | 0.24 | 5.91 | 5.60 |
| CPD5243_7_R_1 | 4.49 | 144.35 | 81.40 | −10.20 | −0.33 | −0.13 | 0.00 | 0.59 | 7.00 | 5.09 |
| CPD4826_6_R_1 | 2.70 | 114.75 | 85.50 | −10.12 | 0.00 | −0.20 | 0.05 | 0.66 | 5.72 | 6.79 |
| CPD4826_6_R_2 | 2.70 | 86.31 | 86.00 | −10.12 | 0.00 | −0.20 | 0.05 | 0.66 | 5.87 | 6.75 |
| CPD4562_6_R_2 | 3.55 | 92.79 | 90.10 | −8.96 | 0.76 | −0.24 | 0.03 | 0.26 | 6.89 | 5.57 |
| CPD4562_6_R_1 | 3.55 | 166.00 | 90.30 | −8.96 | 0.76 | −0.24 | 0.03 | 0.26 | 7.54 | 5.59 |
| CPD5265_6_R_1 | 5.58 | 150.69 | 121.70 | −8.80 | −0.40 | −0.39 | 0.19 | −1.38 | 9.53 | 7.33 |
| CPD4385_6_R_1 | 1.71 | 57.35 | 125.50 | −8.99 | −0.53 | −0.23 | 0.03 | 2.53 | 7.97 | 8.48 |
| CPD5310_6_R_1 | 2.02 | 159.28 | 133.00 | −8.29 | 0.17 | −0.22 | 0.03 | 3.10 | 9.42 | 6.83 |
| CPD5310_6_R_2 | 2.02 | 58.74 | 133.50 | −8.29 | 0.17 | −0.22 | 0.03 | 3.10 | 9.35 | 6.93 |
| CPD5104_6_R_1 | 5.63 | 171.53 | 164.90 | −9.15 | −0.93 | −0.35 | 0.07 | 0.13 | 11.51 | 6.98 |
| CPD4573_5_R_1 | 0.42 | 64.53 | 73.70 | −9.03 | 0.55 | −0.27 | 0.04 | 1.82 | 6.09 | 5.77 |
| CPD4785_5_R_1 | 1.81 | 115.02 | 102.20 | −9.47 | −0.16 | −0.22 | 0.03 | 1.64 | 6.51 | 7.33 |

same heterocycle. This is due to the way in which the orientations were constructed (described above) since it was not possible to distinguish between a methyl substituent and the true point of attachment of the heterocycle to the remainder of the drug molecule. As can be seen from the tables, the position of attachment does affect the parameters calculated for the heterocycle, and therefore it is necessary to illustrate each orientation separately.

## 4. Discussion

The database of ring systems found contains a wide structural variety—even among the 30 most frequently-found species, there is considerable variability in size, structure and polarity, as can be seen from Fig. 2 and the data in Table 1. A total of 204 ring systems of the 721 orientations found were found more than once in the late-phase compounds identified from the MDDR, with 39 structures being hit five times or more.

Study of the structures of the frequently-hit compounds is revealing. For example, there are five late-phase or launched molecules in the MDDR that contain an otherwise unsubstituted 2-furyl moiety identified as CPD4573_5_R_1. Using this substructure to search the MDDR (in conjunction with the PHASE search terms explained in the *methods* section of this work) reveals that there are three compounds containing a 2-acylfuran (prazosin, mometasone and mirfentanil) while two contain a 2-alkylfuran (lafutidine and frusemide). These latter two, both launched drugs, are especially interesting because of the widely-held view that unsubstituted furans are too labile to be used in drug molecules. Similar observations can be made with many of the other heterocycles found.

While there were no statistically significant differences between the means of the descriptors used for the most-frequently found heterocycles as compared to the entire set, there are a number of interesting trends. Unsurprisingly, the size-related parameters (Volume, VDW Length and VDW Width) show larger means and maxima for the entire set as compared to the 30 most frequently found—evidently, the entire set includes complex ring systems of up to four rings that would not be expected to be common, but will skew the distribution of these parameters upward. This may also be the reason for the slightly higher $C \log P$ values found for the entire set as compared to the most frequently found ring systems. Interestingly, the HOMO energy was generally slightly lower (more negative) and the LUMO energy slightly more positive for the most-frequently-found set than for the entire set. This may reflect a lower reactivity among these heterocycles, which could be reflected in better metabolism parameters and hence more frequent successful

Table 2
Principal Component Scores for the 30 most frequently-found heterocycles

| Compound_ID | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| CPD4976_43_R_1 | −1.60 | 0.07 | 0.12 | −0.24 |
| CPD5123_43_R_1 | 0.07 | −1.21 | −1.50 | 0.69 |
| CPD4443_36_R_1 | −1.47 | 0.05 | −1.61 | 2.29 |
| CPD4674_32_R_1 | −1.02 | 0.15 | −1.00 | −0.21 |
| CPD4447_26_R_1 | −2.14 | 2.65 | −0.58 | 1.14 |
| CPD4467_24_R_1 | −1.94 | 0.06 | −0.03 | 1.90 |
| CPD4543_17_R_1 | −2.35 | 0.89 | 0.08 | 1.06 |
| CPD4430_17_R_1 | −1.51 | −0.05 | −0.52 | −0.91 |
| CPD4479_16_R_1 | −2.07 | −1.12 | 0.48 | −1.12 |
| CPD5065_13_R_1 | −0.72 | 1.86 | −1.29 | 0.30 |
| CPD4516_11_R_1 | −0.39 | 0.66 | −0.39 | 0.35 |
| CPD4763_9_R_1 | 0.36 | −1.32 | −1.04 | 0.68 |
| CPD4763_9_R_2 | 0.36 | −1.37 | −1.76 | −0.32 |
| CPD4842_8_R_1 | −1.77 | 0.25 | 0.78 | 0.79 |
| CPD5051_8_R_3 | 0.17 | 0.84 | −0.37 | 0.22 |
| CPD5051_8_R_1 | 0.25 | 0.92 | −0.07 | 0.49 |
| CPD5051_8_R_2 | 0.30 | 0.78 | −0.83 | −0.52 |
| CPD5125_7_R_1 | −1.52 | −0.13 | −2.47 | 1.21 |
| CPD5243_7_R_1 | −1.97 | 1.31 | 1.40 | 0.27 |
| CPD4826_6_R_1 | −1.82 | 0.70 | −0.19 | −0.42 |
| CPD4826_6_R_2 | −1.78 | 0.64 | −0.42 | −0.74 |
| CPD4562_6_R_2 | −1.60 | −0.16 | −0.55 | 1.00 |
| CPD4562_6_R_1 | −1.54 | −0.04 | 0.26 | 2.00 |
| CPD5265_6_R_1 | 0.32 | 1.64 | −1.32 | 2.47 |
| CPD4385_6_R_1 | −0.22 | −0.59 | −0.25 | −1.69 |
| CPD5310_6_R_1 | −0.20 | −1.34 | 1.08 | 1.03 |
| CPD5310_6_R_2 | −0.12 | −1.52 | 0.09 | −0.28 |
| CPD5104_6_R_1 | 0.79 | 1.64 | 1.10 | 1.58 |
| CPD4573_5_R_1 | −1.74 | −1.15 | −1.03 | −0.13 |
| CPD4785_5_R_1 | −1.22 | −0.12 | 0.07 | −0.65 |

use in drug design. This will be further checked against a much larger selection of possible heterocycles whose descriptors will be compared with sets similar to those examined here. Table 3 shows statistics for the entire molecule set and Table 4 for the top 30 heterocycle orientations of Fig. 2.

The easiest method to view the diversity of the set is via the principal components analysis. The eigenvalues show that the cumulative fraction of variance explained by each component is 29, 49, 61 and 72%, respectively. The eigenvectors are given in Table 5.

Examination of the eigenvectors suggests that the principal components broadly represent:

- Component 1: Largely molecular size.
- Component 2: Overall polarity, orbital properties.
- Component 3: Hydrogen bonding.
- Component 4: Orientation of polar groups relative to attachment.

Graphs of one component versus another are given in Fig. 3. The graphs show PC1 versus PC2, PC1 versus PC3, PC1 versus PC4 and PC2 versus PC3.

The above graphs can be used to select heterocycles that are similar to a given heterocycle, or that are different, or that differ in specific aspects. For example, the highlighted compounds (grey point marker) in Fig. 3 (structures shown in Fig. 4) are similar in PC1 (size), PC2 (polarity, orbital

Table 3
Statistics for the descriptors of the entire set of heterocycle orientations

| Statistic | Dipole | Dipole angle | Volume | HOMO energy | LUMO energy | $L_{CHA}$ | $H_{CHA}$ | $C \log P$ | VDW Width | VDW Length |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 2.79 | 104.64 | 142.82 | −9.04 | −0.33 | −0.30 | 0.06 | 1.65 | 8.92 | 7.49 |
| S.D. | 1.47 | 43.25 | 39.38 | 0.57 | 0.57 | 0.10 | 0.06 | 1.37 | 1.60 | 1.27 |
| Max. | 8.07 | 179.84 | 289.80 | −7.59 | 1.17 | 0.14 | 0.37 | 6.33 | 14.29 | 11.91 |
| Min. | 0.00 | 0.13 | 65.70 | −12.18 | −2.22 | −0.70 | 0.00 | −2.08 | 5.29 | 4.45 |

Table 4
Statistics for the descriptors of the top 30 heterocycle orientations

| Statistic | Dipole | Dipole angle | Volume | HOMO Energy | LUMO Energy | $L_{CHA}$ | $H_{CHA}$ | $C \log P$ | VDW Width | VDW Length |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 2.94 | 103.99 | 104.33 | −9.24 | 0.03 | −0.26 | 0.07 | 0.87 | 7.62 | 6.57 |
| S.D. | 1.38 | 44.41 | 28.36 | 0.64 | 0.45 | 0.09 | 0.06 | 1.30 | 1.58 | 0.93 |
| Max. | 5.63 | 179.84 | 164.90 | −8.29 | 0.86 | 0.05 | 0.19 | 3.13 | 11.51 | 8.48 |
| Min. | 0.42 | 25.67 | 65.70 | −11.18 | −0.93 | −0.40 | 0.00 | −1.38 | 5.71 | 5.08 |

Table 5
Eigenvectors (loadings) of the four principal components derived

| Component | Dipole | Dipole angle | Volume | HOMO energy | LUMO energy | $L_{CHA}$ | $H_{CHA}$ | $C \log P$ | VDW Width | VDW Length |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.280 | −0.093 | 0.901 | 0.457 | −0.502 | −0.577 | 0.399 | 0.392 | 0.793 | 0.546 |
| 2 | 0.711 | 0.150 | −0.047 | −0.711 | −0.610 | −0.276 | 0.064 | −0.678 | −0.043 | −0.016 |
| 3 | 0.107 | 0.520 | 0.239 | −0.204 | −0.191 | 0.448 | −0.580 | 0.340 | 0.366 | −0.243 |
| 4 | 0.165 | 0.568 | −0.044 | 0.278 | 0.355 | −0.187 | 0.380 | −0.200 | 0.219 | −0.470 |

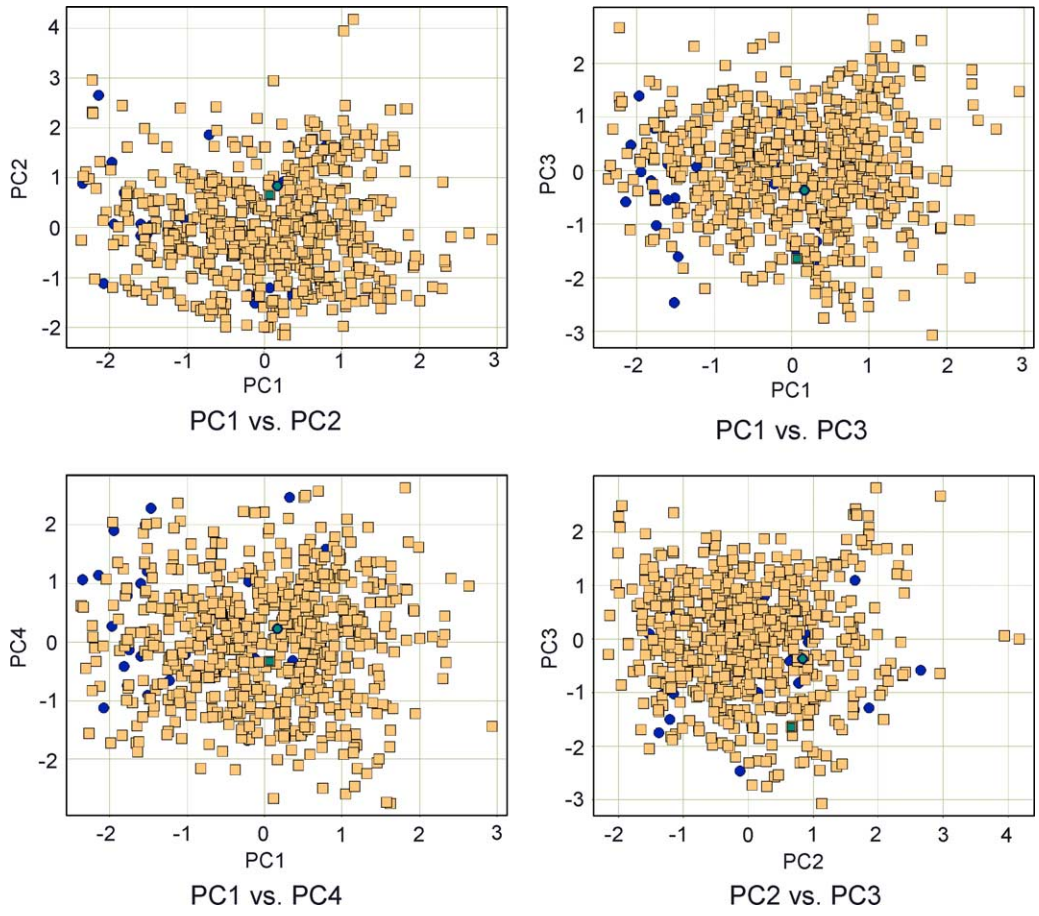


Fig. 3. Graphs of principal components. Compounds in the most-frequently hit 30-heterocycle list are marked in dark blue circles, the remainder in light orange squares. Two compounds (CPD5051_8_R_3 from the frequently-hit structures above, and CPD4629_1_R_2 which was found only once in the late-phase MDDR set) are highlighted with in grey within the marker.

properties) and PC4 (orientation of polar groups relative to attachment), but CPD4629_1_R_2 shows a notably more negative value in PC3 (hydrogen bonding). Thus, CPD4629_1_R_2 and close analogues might make interesting replacements for CPD5051_8_R_3. While it is often the case that near neighbors in the principal component space defined in this paper are also highly similar structures as viewed by a chemist, there are also many cases, such as that illustrated by the compounds of Fig. 4, where the replacement is not such an obvious choice.

Clearly, either directly in the descriptor space or in the principal component space, it is possible to define a distance between any pair of compounds, and thus any of the standard techniques such as cluster analysis or statistical experimental design may be applied to select a suitable diverse subset for the construction of a combinatorial library. Even more simply, outliers in these spaces can readily be identified—for example, the 2-trifluoromethyltetrazol-1-yl heterocycle lies in the extreme upper left of the PC1 versus PC2 and PC1 versus PC3 plots—and can thus be used to establish the behavior of a structure–activity relationship at the extremes. Equally, a novel heterocycle of interest in vitro can be projected into the principal components space of this model and thus it can be seen whether it lies close to the heterocycles found in late-phase drug molecules or not. This can help to prioritize one series of heterocycles over another for investigation as potential drug candidates.

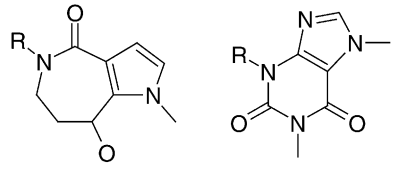


Fig. 4. Structures of the highlighted compounds.

## 5. Conclusions

A method has been devised to extract the heterocycles found in Phase II or later stages of development from the MDDR database, and these heterocycles have been characterized and their similarities and differences examined using a range of descriptors, simplified into four principal components. Based on such analyses, it is possible

to identify suitable replacements for heterocycles in existing molecules. These replacements can be chosen to mimic to the greatest extent possible the properties of the original system, or to vary specifically certain properties such as hydrogen-bonding while maintaining other properties such as size and overall polarity as constant as possible. The method can also be used to select a diverse subset of heterocycles for combinatorial library generation, or to select ring systems with properties near the extremes of those study to evaluate the behavior of a structure-activity relationship under such circumstances.

## References

[1] S. Ekins, B. Boulanger, P.W. Swaan, M.A.Z. Hupcey, Towards a new age of virtual ADME/TOX and multidimensional drug discovery, Mol. Diversity 5 (2002) 255–275.
[2] D.A. Horton, G.T. Bourne, M.L. Smythe, The combinatorial synthesis of bicyclic privileged structures or privileged substructures, Chem. Rev. 103 (2003) 893–930.
[3] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, J. Med. Chem. 39 (1996) 2887–2893.
[4] X.Q. Lewell, A.C. Jones, C.L. Bruce, G. Harper, M.M. Jones, I.M. Mclay, J. Bradshaw, Drug Rings Database with Web Interface: a tool for identifying alternative chemical rings in lead discovery programs, J. Med. Chem. 46 (2003) 3257–3274.
[5] W.H.B. Sauer, M.K. Schwarz, Molecular Shape Diversity of Combinatorial Libraries: a prerequisite for broad bioactivity, J. Chem. Inf. Comput. Sci. 43 (2003) 987–1003.
[6] M. Bohl, J. Dunbar, E.M. Gifford, T. Heritage, D.J. Wild, P. Willett, D.J. Wilton, Scaffold searching: automated identification of similar ring systems for the design of combinatorial libraries, Quant. Struct. - Activity Relat. 21 (2002) 590–597.
[7] A.H. Lipkus, Exploring chemical rings in a simple topological-descriptor space, J. Chem. Inf. Comput. Sci. 41 (2001) 430–438.
[8] S. Gibson, R. McGuire, D.C. Rees, Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments, J. Med. Chem. 39 (1996) 4065–4072.
[9] Maccs Drug Data Report, MDL Information Systems, Inc., San Leandro, CA, http://www.mdli.com.
[10] Daylight Chemical Information Systems Inc., Mission Viejo, CA; http://www.daylight.com for full details of SMILES, SMARTS, etc.
[11] dbtranslate, sybyl 6.8, concord 4.06, Tripos Inc., St. Louis, MO, USA.
[12] R.E. Higgs, K.G. Bemis, I.A. Watson, J.H. Wikel, Experimental designs for selecting molecules from large chemical databases, J. Chem. Inf. Comput. Sci. 37 (1997) 861–870.
[13] M.J. Plotkin, Mathematical basis of ring-finding algorithms in CIDS [Chemical Information and Data System], J. Chem. Document. 11 (1971) 60–63.
[14] J. Downs, P. Nelson, P. Gaumond, Free Software Foundation, http://www.gnu.org.
[15] J.J.P. Stewart, MOPAC 93, QCPE #455.
[16] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O.; Farkas, J. Tomasi, V.; Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, K.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, A.G. Baboul, B.B. Stevanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, B. Johnson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle, J.A. Pople, Gaussian 98, Revision A.9, Gaussian, Inc., Pittsburgh, PA.