Principal Components Describing Biological Activities and Molecular Diversity of Heterocyclic Aromatic Ring Fragments

Samuel Gibson, Ross McGuire,* and David C. Rees

Department of Medicinal Chemistry, Organon Research Laboratories, Newhouse, Scotland ML1 5SH, U.K.

Received April 1, 1996[⊗]

Ten physicochemical variables have been calculated for each of 100 different aromatic rings. These variables were selected because of their potential involvement in the molecular recognition of drug−receptor binding interactions, and they include size, lipophilicity, dipole magnitude and orientation, HOMO and LUMO energies, and electronic point charges. A total of 59 different aromatic ring systems were studied including monocyclics and [5.5]-, [6.5]- and [6.6]-fused bicyclics. A principal components analysis of these results generated four principal components which account for 84% of the total variance in the data. These principal components provide a quantitative measure of molecular diversity, and their relevance for structure−activity relationships is discussed. The principal components correlate with the *in vitro* biological activity of heterocyclic aromatic fragments within a series of previously reported HIV-1 reverse transcriptase inhibitors (Saari, W. S.; *et al. J. Med. Chem.* **1992**, *35*, 3792−3802).

Introduction

Principal components analysis (PCA) is a statistical technique which transforms a set of partially cross-correlated data into a smaller set of new, orthogonal variables called the principal components (PC's) which still retain much of the descriptive power of the original data.[1] PCA has been applied to several series of chemical structures and fragments including α-amino acids[2] and benzene ring substituents.[3,4] A different technique with similar aims, nonlinear mapping, has recently been used to select test series and derive structure−activity relationships for aromatic substituents[5] and aliphatic fragments.[6] Hierarchical cluster analysis has also been used to offer a rational method for the selection of substituents.[7]

This particular study focuses on heterocyclic aromatic ring systems. Previously, descriptors of the *aromaticity* of a set of 24 monocyclic and benzo-fused heteroaromatics have been used to generate principal properties,[8] and in separate investigations Langer[9,10] has studied 16 aromatic monocycles in a total of 37 regioisomers, using comparative molecular field analysis (CoMFA)-generated descriptors in a principal components analysis. A recently published[11] study of 40 heteroaromatic ring systems used GRID[12]-derived descriptors to generate a set of principal properties. These principal properties were then used as the basis for series design using cluster analysis and D-optimal design.

The aims of this investigation were (i) to study a wider range of heterocyclic aromatic ring systems and physicochemical descriptors than had been previously reported, (ii) to identify principal components which correlate the chemical structures with biological activities, and (iii) to enable medicinal chemists to rationally select which heterocyclic rings to synthesize in order to optimize biological activities. A total of 59 different monocyclic and fused bicyclic aromatic ring systems are described, and including various regioisomers of these gives the 100 fragments shown in Chart 1. The physicochemical properties were selected on the basis of known involvement in molecular recognition, and the calculated descriptors were subjected to PCA. Thus, it is possible to represent much of the molecular diversity of the fragments (**1**−**100**) in a 2-dimensional plot. In order to investigate whether the PC's thus generated correlate with biological activity, a retrospective analysis of published data, with a suitably large dataset covering a significant range in biological activity, was performed.

Descriptor Generation and Principal Components Analysis

In the present study, a set of 100 fragments, consisting of monocyclic and fused bicyclic heterocyclic aromatic rings with various substitution points (Chart 1), was selected from among those included in *Comprehensive Heterocyclic Chemistry*.[13] Five- and six-membered monocyclics plus [6.5]-, [6.6]-, and [5.5]-fused bicyclic systems, together with phenyl and naphthyl ring fragments, were included. A CH_3 group attached to the appropriate atom in each case was used to represent the point of ring substitution. In each case the CH_3 group is attached to carbon atoms only, and all such regioisomers of the different monocyclic rings have been included with the exception that in cases where annular tautomerism is possible, only structures where one tautomer is a mirror image of the other were considered. For bicyclic systems, substitution points in only one ring have been considered, and for 3-methyl-1*H*-indazole, the structure represented as **89** was studied as this tautomer is known[14] to predominate. Although the methyl group exerts an influence on the calculated values of the descriptors, this influence should be similar in each case and should have little effect on the differences between values across the set of aryl rings.

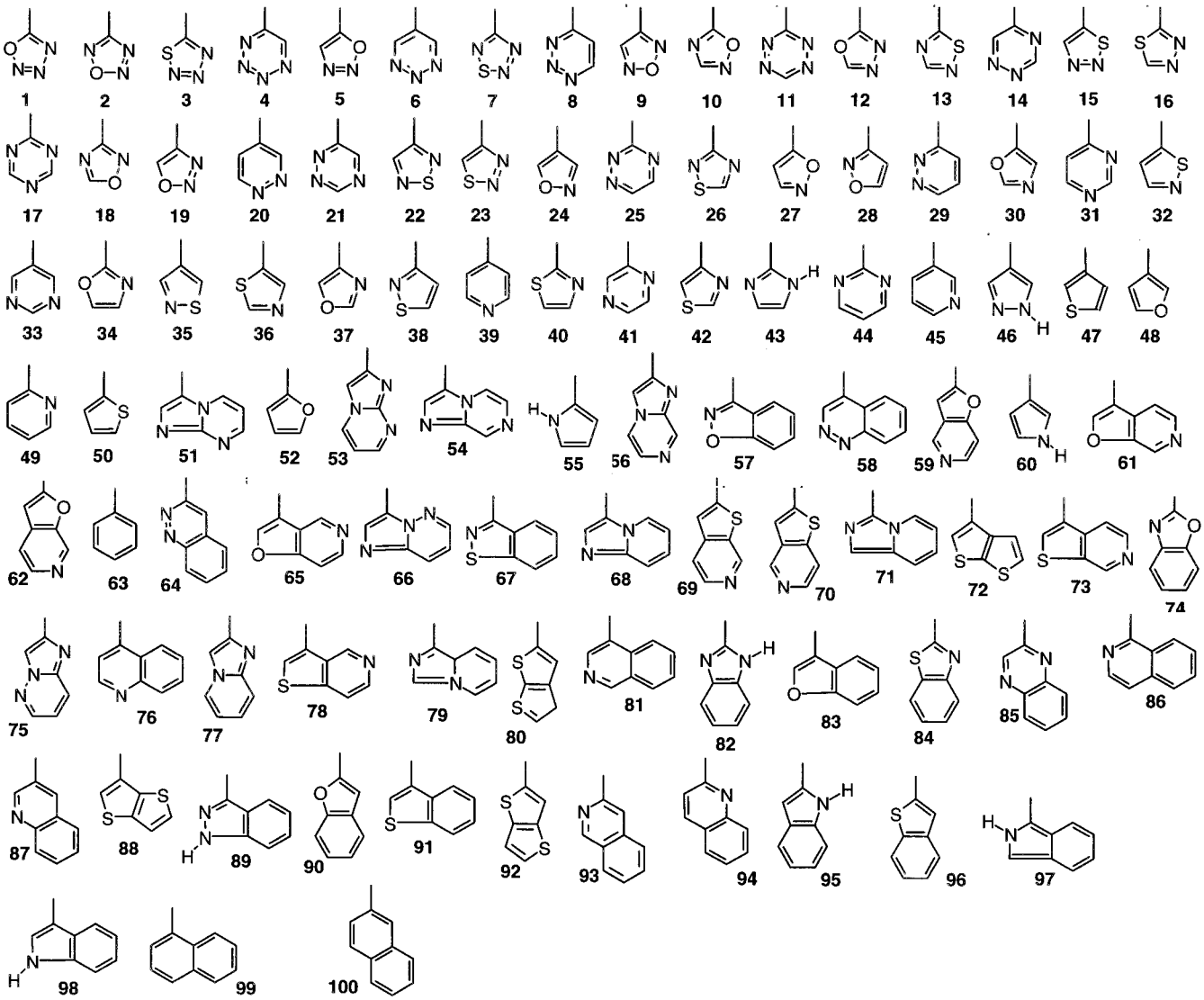
Various physicochemical properties describing steric, electronic, and lipophilic properties considered relevant to protein−ligand interaction and previously used in quantitative structure−activity relationship (QSAR)[15] were then calculated. A complete listing of the descriptor values used for each of the 100 fragments is given in the Supporting Information. Each structure was

* To whom correspondence should be addressed. Telephone: (0) 1698 732611. Fax: (0) 1698 732878. E-mail: r.mcguire@organon.akzonobel.nl.

[⊗] Abstract published in *Advance ACS Abstracts,* August 15, 1996.

S0022-2623(96)00058-1 CCC: $12.00

Chart 1



constructed within Chem-X[16] and optimized by MOPAC[17] MNDO using the Pulay converger and with PRECISE, DIPOLE, and ESP as keywords. These calculations furnished the dipole moment magnitude (Dip) and the dipole angle (Dip-Ang). The dipole angle is the angle between the dipole vector and the 'axis of attachment' (the bond between the ring carbon and CH_3 carbon). For the purposes of this study, all structures were aligned along this axis of attachment, but a deconstruction of the dipole vector into, *e.g.*, the *x* and *y* components presents problems—the *x* component could vary according to how one ring was overlaid with the others. The use of the dipole angle as described avoids this problem as this dipole angle will always be the same and always within the 0−180° range, regardless of the orientation of the ring about the axis of attachment.

As expected, large dipole magnitudes are associated with ring fragments containing groups of electronegative heteroatoms, especially when the heteroatoms are distributed asymmetrically. 4-Methyl-1,2,3 triazine (**8**) has a large dipole moment (4.65 D), for example, while its isomer 2-methyl-1,3,5-triazine (**17**), which has the heteroatoms arranged symmetrically around the ring, has a much smaller dipole magnitude (0.36 D). The MOPAC MNDO-derived highest occupied molecular orbital and lowest unoccupied molecular orbital energies (E_{homm} and E_{lumm}, respectively) were also employed as descriptors.

Calculated point charges provide a further measure of the electronic differences between various heterocycles. CNDO[18] charges have been suggested to be useful descriptors of hydrogen-bonding ability.[19] Charge descriptors were generated by taking the lowest negative charge on a heteroatom (L_{cha}) and the highest positive charge on a hydrogen (H_{cha}) for each molecule. CNDO properties were calculated for the MNDO-optimized structures. Compounds which may behave as H-bond donors (*e.g.*, **43**, **46**, and **55**) have high H_{cha} values (0.103−0.107, total range for H_{cha} 0−0.111). In the present study however, the CNDO point charge maxima and minima are used as a measure of similarity between ring fragments, not purely as a measure of H-bonding ability. Calculated point charges, dipoles, and molecular orbital energies have all previously been employed in generating QSAR's.[4,20−22]

The steric property van der Waals (VdW) volume (Vol) was also calculated. For these planar aromatic rings, the calculated VdW surface area is, as expected, extremely highly correlated with VdW volume, and the surface area measure was not included in this study. Two further steric descriptors were calculated for each fragment. The 'axis of attachment' (see above) of each

Table 1. Principal Component Eigenvalues

PC[a]	eigenvalue[b]	proportion[c]	cumulative[d]
PC1	4.29	0.43	0.43
PC2	1.87	0.19	0.61
PC3	1.30	0.13	0.74
PC4	0.96	0.10	0.84

[a] Principal Component. [b] The size of the eigenvalue is related to the proportion of variance in the descriptor correlation matrix accounted for by the principal component. [c] Division of the eigenvalue by the number of descriptors in the correlation matrix gives the proportion of the variance accounted for by the PC. [d] The sum of the variance explained by the PC's.

heterocycle was aligned with the *y*-axis of Cartesian space, and the maximum VdW width along the *x*-axis, VdW_w, and maximum VdW length along the *y*-axis, VdW_l, were calculated for each structure. These two descriptors are closely related to the STERIMOL parameters.[23,24] Although the VdW_w and VdW_l descriptors will be correlated to the total molecular volume, they describe the differences in overall shape and are especially useful in describing the variation in shape of differently substituted examples of the same heterocycle. Hydrophobicity estimates (Clog *P*) of each ring fragment were also calculated.[25]

Although in previous studies tabulated descriptors such as Hammet substituent constants[26] and *F* and *R*, the inductive and resonance parameters of Swain and Lupton,[27] etc. have been used, for fragments in the present study such values are not readily available. This situation necessitates a reliance on alternative descriptor properties which can be calculated fairly readily.

The set of calculated values thus generated represents a classification of the similarity/dissimilarity among the structures under study, according to descriptors often employed in QSAR.[15] However, with 10 different descriptors for each compound, difficulties arise in using these data in a productive manner. For example, attempting to explain the variation in biological activity in a series of compounds bearing different aromatic rings or attempting to rationally design such a series would be problematic, given the number of descriptor variables. In order to make use of the descriptors generated, in a more practicable and intuitive sense, the calculated data (10 descriptors for each of 100 fragments) were subjected to PCA.

PCA is a statistical technique which reduces a set of partially cross-correlated data into a smaller set of new, orthogonal variables (the PC's), which still retain much of the descriptive power of the original data. Although cross-correlation of variables is normally undesirable for statistical techniques such as QSAR, PCA uses the cross-correlation among a set of variables to produce the principal components.

An advantage of PCA is that no assumptions are made regarding the probability distributions of the original variables,[28] and data derived from different sources may be treated together. In this case, PCA extracted four PC's, and the results in Table 1 show that these four PC's describe 84% of the original variance.

Table 2 shows the eigenvectors for each PC. This analysis shows that PC1 is not dominated by any one original descriptor variable, as significant contributions are made by E_{homm}, Vol, Clog *P*, VdW_w and VdW_l, which, as would be expected, share significant cross-correlation. PC1 is clearly related to the size and lipophilicity of the heterocycles. PC2 is similarly complex, with variance in E_{lumm} and H_{cha} contributing to this PC, indicating that electronic and hydrogen-bonding factors may be represented by PC2.

Table 2. Eigenvectors of Principal Components

	eigenvectors[a]			
descriptor	PC1	PC2	PC3	PC4
Dip[b]	−0.176	−0.236	0.563	0.344
Dip-Ang[c]	−0.229	−0.184	−0.059	0.664
E_{homm}[d]	0.452	0.113	0.142	0.089
E_{lumm}[e]	0.135	0.626	0.178	0.008
Vol[f]	0.423	−0.326	0.015	0.107
L_{cha}[g]	−0.031	0.190	−0.660	0.390
H_{cha}[h]	0.125	0.512	0.332	0.311
Clog *P*[i]	0.433	0.072	−0.272	0.096
Vdw_w[j]	0.397	−0.189	0.062	0.349
Vdw_l[k]	0.393	−0.245	0.057	−0.204

[a] Eigenvectors are a measure of the contribution each descriptor makes to each PC. [b] Dipole moment (debyes). [c] Dipole angle (degrees, see text). [d] Energy of the HOMO calculated by MOPAC MNDO (kcal mol^{-1}). [e] Energy of the LUMO calculated by MOPAC MNDO (kcal mol^{-1}). [f] van der Waals volume (Å^3). [g] Lowest point charge on a heteroatom, calculated by CNDO (absolute electron charge). [h] Highest point charge on a hydrogen atom calculated by CNDO (absolute electron charge). [i] Calculated log octanol/water partition coefficient. [j] Maximum VdW width (Å). [k] Maximum VdW length (Å).

Factor analysis is a statistical technique closely related to PCA. In particular, factors can be extracted from a set of partially cross-correlated data and then modified (rotated) to produce new factors which are more closely identified with some of the original descriptors and less closely related to others. Such rotations often reduce the complexity of the newly extracted descriptors. However, in this case factor analysis followed by five different orthogonal rotation techniques did not significantly reduce the overall complexity of the analysis.

Table 3 shows the principal component scores for the first four principal components for each structure. The four PC's describe 84% of the original variance (see Table 1), and if the variance so described is related to factors influencing drug−receptor interactions, then plots of, *e.g.*, PC1 versus PC2 might be useful in studying series of compounds where one aromatic ring has been substituted for others in an attempt to alter the biological activity of the compounds.

Correlation between Principal Components and Biological Activity

Figure 1 shows the distribution of aromatic rings represented by PC1 and PC2. This plot may be regarded as a representation of the molecular diversity of these aromatic rings, according to the physicochemical descriptors employed. The structure corresponding to each point in Figure 1 can be determined by reference to Table 3. If, for example, PC1 and PC2 are related to properties important for biological activity, then 'active' compounds might be expected to occupy different regions in PC plots from 'less active' compounds.

In order to investigate whether the PC's described above are related to biological data, a recent literature study of HIV-1 reverse transcriptase (RT) inhibitors,[29] with a suitably large dataset, was analyzed. Compounds that differ *only* in the type of aromatic ring present at one position were taken and studied with

Table 3. Principal Component Scores for Aryl Rings **1**−**100**

	score					score			
ring no.	PC1	PC2	PC3	PC4	ring no.	PC1	PC2	PC3	PC4
1	−4.816	−1.070	0.503	−0.018	**51**	0.280	−1.923	2.894	1.132
2	−4.597	−0.650	−0.893	0.101	**52**	0.342	2.280	−0.427	−1.388
3	−4.051	−1.488	−0.233	0.268	**53**	0.491	−1.675	3.136	−1.030
4	−3.872	−1.861	0.033	0.888	**54**	0.508	−1.773	1.779	0.978
5	−3.211	0.152	0.524	0.774	**55**	0.729	4.357	0.945	1.151
6	−3.178	−1.588	0.241	1.627	**56**	0.831	−1.506	2.094	−0.887
7	−3.055	−0.594	−0.600	−0.893	**57**	0.880	−0.849	0.131	0.336
8	−3.027	−1.088	0.644	0.360	**58**	0.935	−2.034	−0.208	1.707
9	−2.947	0.502	−0.552	0.896	**59**	1.066	−1.030	0.430	0.622
10	−2.901	0.480	−1.213	−0.225	**60**	1.070	4.744	1.088	0.241
11	−2.897	−0.930	−1.943	−0.283	**61**	1.152	−1.154	0.390	0.939
12	−2.791	0.642	0.822	−0.472	**62**	1.192	−1.041	0.745	0.227
13	−2.477	−0.514	−0.286	−0.235	**63**	1.225	1.907	−2.584	−1.100
14	−2.438	−1.087	−0.022	0.650	**64**	1.272	−1.771	−0.248	−0.034
15	−2.407	−0.444	−0.871	1.473	**65**	1.297	−0.579	0.321	0.055
16	−2.400	−0.120	−0.338	−0.016	**66**	1.395	−1.105	1.286	−0.328
17	−2.386	−0.573	−0.368	−0.487	**67**	1.437	−1.256	0.384	−0.023
18	−2.191	0.930	0.130	−2.126	**68**	1.481	−1.065	1.946	0.506
19	−2.149	0.622	−0.157	0.619	**69**	1.524	−1.498	−0.018	0.336
20	−1.996	−0.490	−0.191	1.183	**70**	1.582	−1.390	−0.687	0.628
21	−1.973	−0.837	−0.389	−0.522	**71**	1.591	−0.697	1.178	0.113
22	−1.914	−0.027	−0.191	−0.005	**72**	1.598	0.099	−1.613	1.924
23	−1.909	0.488	−0.620	0.754	**73**	1.621	−0.743	−1.074	2.004
24	−1.887	0.983	−0.326	1.234	**74**	1.654	−0.768	0.454	−1.458
25	−1.734	−0.575	−0.040	−1.502	**75**	1.693	−0.852	1.582	−2.143
26	−1.684	0.415	0.100	−2.232	**76**	1.717	−1.536	−0.204	1.376
27	−1.614	1.131	0.005	0.094	**77**	1.779	−0.873	2.223	−1.325
28	−1.485	1.379	−0.080	−0.226	**78**	1.816	−0.350	−1.209	0.839
29	−1.485	−0.247	−0.034	−0.197	**79**	1.821	−0.551	1.231	−0.508
30	−1.437	1.083	0.254	−0.053	**80**	1.976	0.156	−1.520	−0.190
31	−1.275	−0.021	0.406	−0.325	**81**	2.020	−1.406	−0.272	0.521
32	−1.261	0.166	0.155	−0.024	**82**	2.098	1.070	2.759	0.641
33	−1.241	−0.605	−0.082	0.385	**83**	2.152	−0.365	−0.552	0.793
34	−1.240	1.560	0.121	−1.434	**84**	2.154	−0.917	−0.441	−1.320
35	−1.192	0.554	0.189	0.809	**85**	2.181	−1.230	−1.004	−1.138
36	−1.041	0.392	−0.788	0.103	**86**	2.273	−1.307	0.031	−0.115
37	−0.982	1.591	0.580	−1.336	**87**	2.280	−1.432	−0.276	−0.109
38	−0.949	0.783	0.555	−0.718	**88**	2.453	0.178	−1.835	0.152
39	−0.877	0.270	−0.123	0.785	**89**	2.470	1.906	0.592	0.318
40	−0.751	1.034	−0.119	−1.043	**90**	2.519	−0.113	−0.199	−0.798
41	−0.716	−0.060	−1.199	−0.327	**91**	2.601	−0.165	−1.143	1.131
42	−0.632	1.145	−0.247	−0.724	**92**	2.615	0.187	−1.779	−1.325
43	−0.363	3.501	2.989	0.397	**93**	2.624	−1.382	0.040	−1.067
44	−0.283	0.346	0.489	−2.320	**94**	2.646	−1.249	0.072	−1.026
45	−0.201	0.269	−0.246	−0.209	**95**	2.671	1.654	1.050	1.382
46	−0.124	3.696	1.227	0.840	**96**	2.846	−0.402	−1.237	−0.776
47	−0.046	1.742	−2.306	1.113	**97**	2.860	1.924	0.551	2.377
48	0.124	2.103	−0.609	−0.200	**98**	3.120	2.090	1.136	1.035
49	0.165	0.460	0.108	−1.211	**99**	3.441	0.176	−2.766	−0.178
50	0.265	1.774	−1.420	−0.619	**100**	3.552	−0.067	−2.757	−0.566

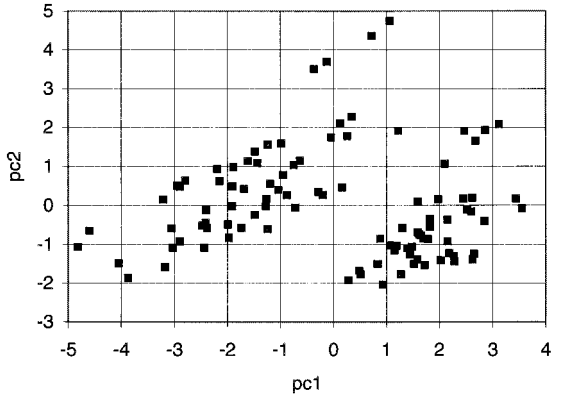


Figure 1. Plot of PC1 vs PC2 for all ring fragments in Table 3/Chart 1. Each point (■) represents an aryl ring from Table 3/Chart 1.

respect to the distribution of the aromatic rings in plots of PC1 versus PC2. The HIV-1 RT dataset contained 18 compounds and a range in activity of over 3 log units.

It should be noted that in the study described below, straightforward integer values have been used to devise the activity ranges in Figures 2 and 3—no selection of the 'best' cutoffs for activity ranges has been performed. Additionally, the 'actives' have been separated from the 'inactives' for the purposes of using cluster significance analysis[30] (CSA) which requires such binary data. For thoroughness CSA has been employed using two different cutoff values to generate the 'active' and 'inactive' sets.

Figure 2 shows the distribution of ring structures present in the homologous set of HIV-1 RT inhibitors. Points have been coded to represent the measured biological activity of the corresponding HIV-1 RT inhibitor. These activities are also shown in Table 4. Figure 2 shows a clustering of the most active compounds ($pIC_{50} > 6$) in an area of the plot (roughly PC1 > 1.6 *and* PC2 < 0.0) distinct from the areas occupied by most of the less active compounds. One point close to the cluster of most active compounds which has a $pIC_{50} <$

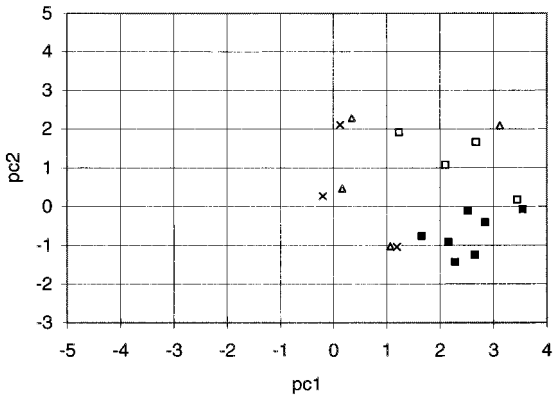


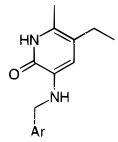
Figure 2. Plot of PC1 vs PC2 for aryl ring fragments present in the previously reported HIV-1 RT inhibitors in Table 4 (taken from ref 29): ×, pIC_{50} (of corresponding HIV-1 RT inhibitor) < 4; Δ, 4 < $pIC_{50} \leq 5$; □, 5 < $pIC_{50} \leq 6$; and ■, $pIC_{50} > 6$.

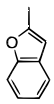
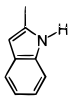
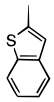
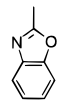
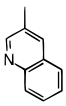
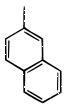
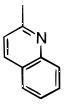
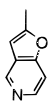
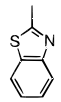
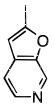
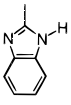
6 belongs to the 1-methylnaphthalene ring, which has a $pIC_{50} = 5.68$, the highest pIC_{50} in the range 5−6. Two points, relating to the ring structures (**59** and **62**), are associated with lower activity and have PC scores reasonably close to the group of points associated with higher activity. These two rings are distinct from the rest of the ring structures considered in this study in that they are bicyclic fragments with heteroatoms in both rings. Evidently the presence of nitrogens in these positions is not conducive to good biological activity in this case.

In order to obtain a measure of the statistical significance of the observed correlations, CSA was employed. CSA calculates the 'tightness' of a cluster of *n* compounds and then calculates the tightness of all other possible clusters of *n* compounds present. The CSA *p*-value is the number of clusters at least as tight as the 'active' cluster divided by the total number of possible clusters of *n* compounds.

In this test of the correlation between PC1 and PC2 versus biological activity, separation of the 18 compounds into 7 'actives' ($pIC_{50} > 6.0$) and 11 'inactives' ($pIC_{50} < 6.0$) gives a CSA *p*-value of 0.000 16. If $pIC_{50} > 5.5$ is taken as the cutoff value, identifying 8 'actives' and 10 'inactives', then the CSA *p*-value is calculated to be 0.000 05. A plot of PC1 vs PC4, as shown in Figure 3, gives a CSA *p*-value of 0.000 06 using the $pIC_{50} > 6.0$ cutoff to separate 'actives' from 'inactives' and a CSA *p*-value of 0.000 05 using the $pIC_{50} > 5.5$ cutoff.

Table 4. PC Scores of Aryl Ring Fragments and Corresponding pIC_{50} Values from a Series[29] of HIV-1 RT Inhibitors



Ring No.	Ar[a]	PC1 Score[b]	PC2 Score[b]	PC3 Score[b]	PC4 Score[b]	pIC_{50}[c]	Ring No.	Ar[a]	PC1 Score[b]	PC2 Score[b]	PC3 Score[b]	PC4 Score[b]	pIC_{50}[c]
45		-0.201	0.269	-0.246	-0.209	3.52	90		2.519	-0.113	-0.199	-0.798	6.48
49		0.165	0.460	0.108	-1.211	4.82							
48		0.124	2.103	-0.609	-0.200	3.84	95		2.671	1.654	1.050	1.382	5.36
63		1.225	1.907	-2.584	-1.100	5.27	96		2.846	-0.402	-1.237	-0.776	6.3
74		1.654	-0.768	0.454	-1.458	6.68	98		3.120	2.090	1.136	1.035	4.65
87		2.280	-1.432	-0.276	-0.109	6.47	100		3.552	-0.067	-2.757	-0.566	6.34
52		0.342	2.280	-0.427	-1.388	4.54	99		3.441	0.176	-2.766	-0.178	5.68
94		2.646	-1.249	0.072	-1.026	6.28	59		1.066	-1.030	0.430	0.622	4.49
84		2.154	-0.917	-0.441	-1.320	6.46	62		1.192	-1.041	0.745	0.227	3.98
82		2.098	1.070	2.759	0.641	5.12							

[a] Bond indicates the position of attachment. [b] PC scores from Table 3. [c] $-\log_{10} IC_{50}$ from ref 29.

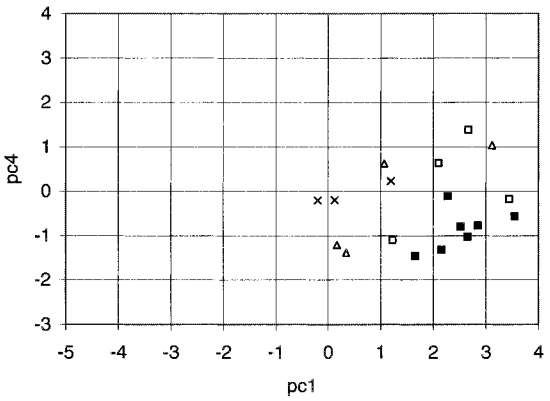


Figure 3. Plot of PC1 vs PC4 for aryl ring fragments present in the previously reported HIV-1 RT inhibitors in Table 4 (taken from ref 29): ×, pIC_{50} (of corresponding HIV-1 RT inhibitor) < 4; Δ, 4 < $pIC_{50} \leq 5$, □, 5 < $pIC_{50} \leq 6$; and ■, $pIC_{50} > 6$.

In both cases described above, there exists a reasonably clear distinction between areas in the PC plots associated with good activity and other areas where activity is poorer. Ring systems in areas associated with higher activity levels and not yet synthesized would be prime synthetic targets. Cluster significance analysis suggests that the use of plots of PC1 versus PC2 or PC4 can produce statistically significant clusters associated with good biological activity. The use of the PC plots in helping to determine synthetic targets represents a rational approach to this area of ligand design.

Comparison of the data presented here with those previously published by Musumarra *et al.*,[8] Clementi *et al.*,[11] and Langer[9] is not straightforward. The studies conducted by Musumarra *et al.* and Clementi *et al.* concerned complete heterocycles, with no point of attachment indicated. This means that a pairwise comparison of heterocycle descriptors is not possible. Langer used two sets of biological data to derive QSAR's, via partial least squares (PLS). However, one dataset had a range in biological data of only 0.28 log unit, while the other set of six compounds contained only five heterocycles considered here. Langer included heterocyles which are attached to the rest of the molecule via a ring nitrogen. Although our descriptors are not inconsistent with the biological data used by Langer, little or no conclusions can be drawn from datsets with such a limited spread in biological response or with only five data points.

Discussion

The correlation between the PC's generated in this study and the set of *in vitro* enzyme inhibition data is represented in Figures 2 and 3. Inspection of Figure 1 also suggests that the distribution of ring fragments in plots of PC1 vs PC2 is related to the physical characteristics and chemical functionality of the aryl rings. For example, monocyclic rings which may act as H-bond donors (**43**, **46**, **55**, and **60**) have PC1 scores in the range −0.4 to 1.1 and PC2 scores between 3.5 and 4.8, occupying a region of Figure 1 distinct from other rings. The bicyclic H-bond donors (**82**, **89**, **95**, **97**, and **98**) are also distributed in an area of Figure 1 distinct from other rings (PC1 > 2 *and* PC2 > 1.0).

It is also evident that a broad trend of 'more *hetero*aromatic' to 'less *hetero*aromatic' exists on going from low PC1 scores (*e.g.*, five-membered rings with four heteroatoms) to high PC1 scores (*e.g.*, nine- or ten-membered rings with less than two heteroatoms). More subtle differences in aryl ring characteristics are also conveyed in Figure 1. For example, ring fragments with PC1 scores less than −3 and PC2 scores less than −1 (structures **1, 3, 4, 6**, and **8**) are all associated with large dipole magnitudes (>3.75 D, total dipole magnitude range 0−5.2), low (negative) Clog *P* values (<−1, total Clog *P* range −2.12 to 3.81), and low E_{homm} values (<−11 kcal mol^{-1}, total E_{homm} range −12.00 to −7.75). Fragments with higher PC1 scores generally have less extreme dipole, Clog *P*, and E_{homm} values. Many descriptors make significant contributions to PC1 and PC2, as shown in Table 2, yet these complex relationships are represented in a rational and practically useful manner in Figures 1 and 2.

In some cases, statistically significant regression equations may been generated relating the biological activity to the PC's. For example, multiple linear regression using two principal components with the HIV-1 RT data in Table 4 gives eq 1:

$$pIC_{50} = 0.71(PC1) - 0.68(PC4) + 3.80 \qquad (1)$$

$$n = 18\ r^2 = 0.79\ F = 27.7\ s = 0.51$$

In eq 1, *n* is the number of observations, *r* is the correlation coefficient, *F* is the *F*-value, and *s* is the standard error.

However, since an 'active' region on a PC plot may be surrounded by 'inactive' areas, *i.e.*, biological activity may not be *linearly* related to the PC's, a regression on the PC's alone will not always be an appropriate indicator of a correlation with biological activity. The use of the PC plots appears to offer a rational, practical method for the design of 'active' molecules where activity is affected by the type of aromatic ring incorporated into a homologous series.

The PC plot in 2- or 3-dimensions also provides a basis for experimental design where synthesis is at an early stage and biological activity has not been established. For example, it may be used as a measure of molecular diversity, or factorial design[31] can be applied using the principal component scores, and a small number of 'representative' rings can be used to effectively explore the available variance. One method would be where three principal components are used as the axes of a cube which is then divided into eight smaller cubes or octants. A small number of representative rings could then be selected from each octant so that synthesis would explore an optimally diverse set of aryl rings.

In the literature example described in Figures 2 and 3, of the four principal components extracted, PC1, PC2, and PC4 give the most significant correlations with the observed biological activity. The use of PC3 in principal component biplots did not improve on the correlations obtained by PC1 and PC2. PC1 and PC2 account for a large part of the variance in the original data and might be expected to correlate with biological activity. Why PC4 should also give excellent correlation is less clear. Table 2 shows that dipole magnitude and orientation (Dip and Dip-Ang) as well as a calculated point charge (L_{cha}) contribute to this PC, and it may be that such properties in combination with those which contribute to PC1 are important in the literature case studied here.

The technique of principal components has previously been applied to reduce the dimensionality of descriptor data as outlined in the Introduction. However, it should be pointed out that in some circumstances, PCA may harbor intrinsic weaknesses. For example, if in a homologous series of compounds one single factor associated with the aryl rings has a pre-eminent importance for biological activity, then the PCA approach may serve to obscure this importance. In the present analysis 1-methylisoquinoline (**86**) and 3-methylisoquinoline (**93**) are calculated to be very similar, yet a situation can easily be envisaged where an isoquinolin-3-yl substituent could be accommodated in a binding pocket while a isoquinolin-1-yl substituent might not be tolerated because of the spatial requirements of the pocket. In such cases the PC plots would not give a clear distinction in observed activities.

The original descriptors chosen are not presented as a definitive set of initial data and the characteristics of the rings in Chart 1 will be affected to some extent by the atoms bonded to the ring in the complete molecule. Given prior knowledge of the target binding site (or supposition), different descriptors could be employed with the aim of generating a more relevant set of PC's and thence a more useful PC plot. In the hypothetical case above, for example, the inclusion of more specific shape descriptors might be advisable. Similarly, the inclusion of additional aromatic rings or the inclusion of variously substituted aromatic rings could increase the relevance of this approach to a particular chemical series.

Conclusions

Ten different physicochemical properties describing lipophilic, steric, and electronic parameters have been calculated for each of 100 different monocyclic and fused bicyclic heterocyclic aromatic ring systems (compounds **1**−**100**). A principal components analysis of these data generates two principal components which account for a total of 61% of the variation in the original data. A third PC accounts for another 13% of this variance, while a fourth PC accounts for a further 10% of the original variance. A 2-dimensional plot of PC1 vs PC2 produces a visual representation of the molecular diversity of this dataset.

In a retrospective analysis of published medicinal chemistry results, these PC's appear to be relevant descriptors for *in vitro* biological activities of HIV-1 RT inhibitors. The results of this study provide a rational basis (i) for optimizing heterocyclic aromatic fragments with novel series of biologically active compounds and (ii) for series design of heterocyclic aromatic rings for inclusion in a synthesis program.

Experimental Section

Each ring structure was constructed within Chem-X,[16] with a CH_3 group at a variety of positions representing the point of attachment of the heterocycle to the remainder of a larger compound. Each structure was optimized by MOPAC[17] (version 6.0) MNDO, using the Pulay converger and with PRECISE, DIPOLE, and ESP as keywords. The dipole magnitude and dipole angle, *i.e.*, the angle between dipole vector and the bond between CH_3 and ring structure, were then extracted from the MOPAC output using Chem-X Chem-QM procedures (descriptors Dip and Dip-Ang, respectively). The MOPAC orbital energies for the HOMO and LUMO were similarly extracted for each structure (E_{homm} and E_{lumm}, respectively).

The CNINDO[18] program was also used via its Chem-X Chem-QM interface and Chem-Stat routines to calculate the highest CNDO point charge on a hydrogen attached to the ring and the lowest (most negative) CNDO point charge on a heteroatom (H_{cha} and L_{cha}, respectively). Rings with no hydrogens attached were given H_{cha} values of zero, and similarly rings with no heteroatoms were given L_{cha} values of zero.

The steric property VdW volume (Vol) was calculated from Chem-X VdW volume maps at 3.0 contours/Å resolution. Prior to the calculation of maximum VdW width and maximum VdW length (Vdw_w and VdW_l in Table 1, respectively), each structure was positioned in Chem-X such that the bond between the CH_3 carbon and the aryl ring was aligned with the *y*-axis of Cartesian space. The maximum VdW width was calculated as the maximum difference in *x* coordinate values for any pair of atoms plus the VdW radii of the two atoms. The maximum VdW length was calculated as the maximum difference between the *y* coordinate of the CH_3 carbon and the *y* coordinate of any other atom in the ring plus the VdW radii of the two atoms. Chem-X uses the VdW radii taken from published[32] VdW parameters.

Lipophilicity values for each heterocycle (Clog *P*) were calculated using the Daylight CLOGP[25] program, version 4.34, except in nine cases [4-methyl-1,2,3-triazine (**8**), 5-methyl-1,2,3-triazine (**6**), 5-methyl-1,2,3,4-tetrazine (**4**), 4-methyl-1,2,3-oxadiazole (**19**), 5-methyl-1,2,3-oxadiazole (**5**), 5-methyl-1,2,3,4-oxatriazole (**1**), 4-methyl-1,2,3,5-oxatriazole (**2**), 5-methyl-1,2,3,4-thiatriazole (**3**), and 4-methyl-1,2,3,5-thiatriazole (**7**)] where no CLOGP value could be determined. For these structures, the Chemicalc[33] method was used. Although using data from two methods for log *P* calculation is not an ideal solution, it is preferable to having missing data. An input error in the Clog *P* program led to 5-methyl-1,2,3-thiadiazole (**15**) being determined with an estimated thiadiazole ring fragment value while the 4-methyl-1,2,3-thiadiazole (**23**) used a measured fragment value. The Clog *P* calculated value for **23** was therefore also used for **15**.

Principal components analysis was performed using the SAS[34] suite of statistical analysis programs using the SAS PRINCOMP procedure, which by default standardizes the variables prior to analysis. Factor analysis was performed using the SAS FACTOR procedure. Factors were extracted using the default principal components method and then rotated using the ORTHOMAX, QUARTIMAX, PARSIMAX, VARIMAX, and EQUAMAX options. In each case the rotated factor pattern did not significantly improve the analysis.

Cluster significance analysis was performed using the published[35] CSA1 program. The program was compiled and run on a Vax 4000. The program was tested using the test data[35] and gave the correct test results. Plots and regression equations were generated using Microsoft Excel.[36]

Acknowledgment. The authors thank Dr. D. Rae and Dr. D. Jaap in our laboratories for helpful discussions.

Supporting Information Available: Values of the 10 original descriptor variables for each of the 100 structures, the correlation matrix of the original descriptors, the initial and rotated factor patterns, the full outputs from the CSA1 program, and further statistical details on regression eq 1 (17 pages). Ordering information is given on any current masthead page.

References

(1) Franke, R.; Gruska, A. Chapter 4. Multivariate Data Analysis of Chemical and Biological Data. In *Methods and Principles in Medicinal Chemistry Vol.* 2 *Chemometric Methods in Molecular Design;* Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993; pp 113−163.

(2) Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. Multivariate Parameterization of 55 Coded and Non-Coded Amino Acids. *Quant. Struct.-Act. Relat.* **1989**, *8*, 204−209.
(3) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. Principal Properties for Aromatic Substituents. A Multivariate Approach for Design in QSAR. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32−38.
(4) Stjernlöf, P.; Ennis, M. D.; Hansson, L. O.; Hoffman, R. L.; Ghazal, N. B.; Sundell, S.; Smith, M. W.; Svensson, K.; Carlsson, A.; Wikstrom, H. Structure-Activity Relationships in the 8-Amino-6,7,8,9-tetrahydro-3*H*-benz[e]indole Ring System. 1. Effects of Substituents in the Aromatic System on Serotonin and Dopamine Receptor Subtypes. *J. Med. Chem.* **1995**, *38*, 2202−2216.
(5) Domine, D.; Devillers, J.; Chastrette, M. A Nonlinear Map of Substituent Constants for selecting Test Series and Deriving Structure-Activity Relationships. 1. Aromatic Series. *J. Med. Chem.* **1994**, *37*, 973−980.
(6) Domine, D.; Devillers, J.; Chastrette, M. A Nonlinear Map of Substituent Constants for selecting Test Series and Deriving Structure-Activity Relationships. 2. Aliphatic Series. *J. Med. Chem.* **1994**, *37*, 981−987.
(7) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16*, 1217−1222.
(8) Caruso, L.; Musumarra, G.; Katritzky, A. R. "Classical" and "Magnetic" Aromaticities as new Descriptors for Heteroaromatics in QSAR. Part 3 [1]. Principal Properties for Heteroaromatics. *Quant. Struct.-Act. Relat.* **1993**, *12*, 146−151.
(9) Langer, T. Molecular Similarity Determination of Heteroaromatics using GRID, CoMFA and Multivariate Data Analysis. *Quant. Struct.-Act. Relat.* **1994**, *13*, 402−405.
(10) Langer, T. Molecular Similarity Determination of Heteroarmoatics Using GRID, CoMFA and Multivariate Data Analysis. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; J. R. Prous Science Publishers: Barcelona, 1995; pp 388−391.
(11) Clementi, S.; Cruciani, G.; Fifi, P.; Valigi, R.; Musumarra, G. A new set of Principal Properties for Heteroaromatics Obtained by GRID. *Quant. Struct.-Act. Relat.* **1996**, *15*, 108−120.
(12) Goodford, P. J. A Computational Procedure fro Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(13) Katritzky, A. R.; Rees, C. W. In *Comprehensive Heterocyclic Chemistry, Volume 8, Subject, Author, Ring and Data Indexes;* Drayton, C. J., Ed.; Pergamon Press: Oxford, 1984; pp 927−1111.
(14) Elguero, J.; Marzin, C.; Katritzky, A. R.; Linda, P. The Tautomerism of Heterocycles, Suppl. 1. In *Advances in Heterocyclic Chemistry*; Katritzky, A. R., Boulton, A. J., Eds.; Academic Press: New York, 1976; Chapter 4, pp 291−292.
(15) The physicochemical descriptors used in this study are the same or closely related to those used in many previous QSAR investigations. For a review/discussion on these properties, see: Kubinyi, H. Chapter 3. Parameters. In *Methods and Principles in Medicinal Chemistry Vol. 1, QSAR: Hansch Analysis and Related Approaches*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993; pp 21−36.
(16) The Chem-X suite of programs, developed and distributed by CDL, Chipping Norton, England.
(17) Stewart, J. J. P. QCPE 455. *QCPE Bull.* **1989**, *9* (1), 10.
(18) Philips, A. C. QCPE 274. *QCPE* **1975**, *11*, 274.
(19) Dearden, J. C.; Ghafourian, T. A Hydrogen Bond Parameter from Molecular Orbital Theory: Comparison with α-Values. *J. Pharm. Pharmacol.* **1993**, *45* (Suppl. 2), 1143.
(20) Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; Hansch, C. Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital Energies and Hydrophobicity. *J. Med. Chem.* **1991**, *34*, 786−797.
(21) Lukovits, I. Quantitative Structure-Activity Relationships Employing Independant Quantum Chemical Indices. *J. Med. Chem.* **1983**, *26*, 1104−1109.
(22) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. J.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.
(23) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Medicinal Chemistry, 11 (VII, Drug Design);* Ariëns, E. J., Ed.; Academic Press: New York, 1976; pp 165−207.
(24) Verloop, A. *The STERIMOL Approach to Drug Design;* Marcel Dekker: New York, 1987.
(25) CLOGP3, PCmodels version 4.34, Daylight Chemical Information Systems Inc., Irvine, CA.
(26) Hansch, C.; Leo, A. *Substituents Constants For Correlation Analysis in Chemistry and Biology*; John Wiley & Sons, Inc.: New York, 1979.
(27) Swain, C. G.; Lupton, E. C. Field and Resonance Components of Substituent Effects. *J. Am. Chem. Soc.* **1968**, *90*, 4328−4337.
(28) Chatfield, C.; Collins, A. J. *Introduction to Multivariate Analysis*; Chapman and Hall: London, New York, 1980; pp 57−58.
(29) Saari, W. S.; Wai, J. S.; Fisher, T. E.; Thomas, C. M.; Hoffman, J. M.; Rooney, C. S.; Smith, A. M.; Jones, J. H.; Bamberger, D. L.; Goldman, M. E.; O'Brien, J. A.; Nunberg, J. H.; Quintero, J. C.; Schleif, W. A.; Emini, E. A.; Anderson, P. S. Synthesis and evaluation of 2-Pyridinone Derivatives as HIV-1-Specific Reverse Transcriptase inhibitors. *J. Med. Chem.* **1992**, *35*, 3792−3802.
(30) McFarland, J. W.; Gans, D. J. On the Significance of Clusters in the Graphical Display of Structure-Activity Data. *J. Med. Chem.* **1986**, *29*, 505−514.
(31) Austel, V. A Manual Method for Systematic Drug Design. *Eur. J. Med. Chem.-Chim. Ther.* **1982**, *17*, 9−16.
(32) Del Re, G.; Gavuzzo, E.; Giglio, E.; Lelj, F.; Mazza, F.; Zappia, V. Conformational Study of a Methylating Agent: The Crystal Structures of S-Methyl-L-methionine Chloride.HCl (Vitamin U Hydrochloride). *Acta. Crystallogr. Sect. B.* **1977**, *B33*, 3289−3296.
(33) Suzuki, T. J. QCPE 608. *QCPE Bull.* **1991**, *11* (4), 73.
(34) The SAS suite of statistical programs, SAS Institute Inc., Cary, NC.
(35) McFarland, J. W.; Gans, D. J. Linear Discriminant Analysis and Cluster Significance Analysis. In *Comprehensive Medicinal Chemistry*, 1st ed.; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990; pp 675−689.
(36) Microsoft Excel Version 5.0, Microsoft Corp.

JM960058H