# Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds†

David J. Cummins,*,‡ C. Webster Andrews, James A. Bentley, and Michael Cory

Divisions of Medicinal Chemistry and Information Technology, Glaxo Wellcome, Five Moore Drive, Research Triangle Park, North Carolina 27709

Received October 31, 1995⊗

A molecular descriptor space has been developed which describes structural diversity. Large databases of molecules have been mapped into it and compared. This analysis used five chemical databases, CMC and MDDR, which represent knowledge bases containing active medicinal agents, ACD and SPECS, two databases of commercially available compounds, and finally the Wellcome Registry. Together these databases contained more than 300 000 structures. Topological indices and the free energy of solvation were computed for each compound in the databases. Factor analysis was used to reduce the dimensionality of the descriptor space. Low density observations were deleted as a way of removing outliers, which allowed a further reduction in the descriptor space of interest. The five databases could then be compared on an efficient basis using a metric developed for this purpose. A Riemann gridding scheme was used to subdivide the factor space into subhypercubes to obtain accurate comparisons. Most of the 300 000 structures were highly clustered, but unique structures were found. An analysis of overlap between the biological and commercial databases was carried out. The metric provides a useful algorithm for choosing screening sets of diverse compounds from large databases.

## INTRODUCTION

The availability of new high throughput biological screening technology, in particular the availability of large numbers of cloned receptors and enzymes, have provided a vast array of target systems for drug discovery screening. Important drug therapies may result from inhibition of these receptor systems, and extensive effort is currently being directed toward development of large libraries and arrays of synthetic inhibitors. Combinatorial chemistry approaches to drug design provide a wide range of molecular diversity for drug discovery. The automation of biological assays has also allowed the development of high throughput screening systems that can investigate thousands of compounds per week.

In response to these demands for the screening of structurally diverse compounds many industrial and academic groups have made sets of synthetic compounds available for purchase and insertion into screening systems. These sets can contain thousands of structures, not all of which are desirable for screening. Selection of structures from large compound sets or databases requires the development of automated selection methods. This paper reports a method for superimposition of databases and selection of compounds based on descriptor space. Two available knowledge bases with medicinal activity (CMC and MDDR)[1,2] were used to define desirable molecular descriptor space. Two databases of commercially available compounds (ACD and SPECS)[3,4] were used as the source of available compounds. The Wellcome Registry (notation: WR), a proprietary database representing compounds prepared as part of research efforts at the Wellcome Foundation Ltd. laboratories over the past 40 years, was included to allow investigation of the relationship between commercial compounds and the Wellcome compounds. One of the goals was to find commercial structures that filled medicinally active structural space that had not been explored by previous research efforts.

The general strategy has the following steps. First, define a descriptor space for the database structures. Then form a superpopulation by mapping the five databases into the descriptor space. Next, subdivide the descriptor space into volume elements and compare the databases on the basis of volume elements. In this way the similarities and differences between databases can be defined. Finally, commercial compounds falling within the medicinally active volume domain defined by MDDR and CMC can be selected for screening.

Conceptually the data can be viewed from the perspective of a volume analysis problem. In practice, two difficulties were faced. There was a limit to how finely the descriptor space could be gridded. This limit was not imposed by computational considerations as one might suppose but was rather a sampling issue, related to sample size and dimensionality (i.e., number of descriptors used in the analysis). The second difficulty was that the data contained a small percentage of very extreme observations that complicated the volume analysis.

## PREPARATION OF DATABASES

The five databases were converted into SMILES format. Each database was output as SD files and converted to SMILES using the dbtranslate utility program available in the UNITY Chemical Information Software from Tripos Associates.

Text processing macros were used to remove counterions and convert unusual isotopes to normal isotopes. Inorganic compounds were removed since they do not represent areas of interest for this study. In addition, polymeric and large

† Keywords: volume, factor analysis, outliers, hypercube, chemical descriptor, database mining.

‡ cummins@stat.ncsu.edu, wa20097@glaxo.com.

⊗ Abstract published in *Advance ACS Abstracts,* April 1, 1996.

S0095-2338(95)00168-5 CCC: $12.00 

oligomeric compounds were removed since they represent difficult problems in property prediction.

The free energy of solvation (GSOLV) for each compound was computed using an algorithm developed in-house[5] that utilizes the GCL/SMARTS language of the 1989 Medchem software[6] to identify functional groups and compute a free energy of solvation model that includes additive and proximity terms. The free energy of solvation is effective in distinguishing molecules on the basis of polarity. For instance, hydrocarbons can be distinguished from ketones and monoanions from dianions. Aliphatic hydrocarbons have a free energy of solvation that is near zero or slightly positive, while polar molecules have values that are strongly negative, meaning that they are stabilized in an aqueous environment.

Also, topological indices were computed with version 2 of Molconn-X.[7] Topological indices encode the connectivity or 2D structure of each molecule in numerical form. The topological indices used in the analysis were selected from a large set computed by Molconn-X. Many of the indices are highly collinear, some have zero variance (i.e., constant), and some are highly discrete.

One of the selection criteria was to identify unique descriptors. An ideal variable was one that was not highly correlated ($r > 0.65$) to any other variable and moderately correlated ($r > 0.45$) with just a few other variables. Variables that were not correlated with a large number of other variables were more desirable. For example, GSOLV was correlated with no other descriptors with a correlation above 0.65 (CORRS = 0). There were 12 other descriptors with this desirable quality. The remaining descriptors had an increasing number of correlations. At the other extreme, one of the descriptors used (nxp3) was moderately correlated with 55 other descriptors, reflective of the fact that the correlation was not the only criterion used for descriptor selection.

In addition, weight was given to the desirability of near-Gaussian data. The principal factor analysis that was performed makes an assumption of normality. Highly discrete variables were thrown out. Desirable features were near-continuity, high coefficient of variation (CV), low skewness, and low kurtosis. Also descriptors that are physically interpretable were kept whenever possible. A partial listing of this analysis is given in Table 1.

The coefficient of variation is the standard deviation divided by the mean. The CV adjusts for different measurement scales across descriptors, putting them on an "equal footing". As an example in Table 1 WT has a huge standard deviation and XVCH3 has a very small standard deviation, but these two descriptors have about the same CV. For any descriptor a larger CV suggests a more uniform distribution; hence, a large CV is desirable to spread the observations over the range of the descriptor.

The final set of descriptors selected included the free energy of solvation (GSOLV) and 60 topological indices, which are listed in Table 2.

Together, one physical property and 60 topological indices form the descriptor space for analysis of the databases. This set was limited by the constraint that any descriptor used must be computable for every compound in the superpopulation. There were other desirable descriptors, for example clogP, that were not included because they were not computable for large percentages of the superpopulation. [Among topics for future work is the use of imputation methods for filling in missing data. For the clogP descriptor 10−15% of structures had missing fragment values, but these can be measured or estimated to eliminate the problem.] Nonetheless, in the sense that this descriptor space represents a large number of available compounds and a large number

**Table 1.** Correlation Analysis: Standard Deviations and Number of Correlations Greater than 0.65

| OBS | variable | std_dev | mean | CV | CORRS |
|---|---|---|---|---|---|
| 1 | NE44 | 0.26 | 0.04 | 6.39403 | 0 |
| 2 | TM3 | 0.32 | 0.07 | 4.83005 | 0 |
| 3 | TG3 | 1.52 | 0.81 | 1.87451 | 0 |
| 4 | NE12 | 1.07 | 0.89 | 1.19819 | 0 |
| 5 | GSOLV | 157.86 | −140.28 | 1.12530 | 0 |
| 6 | XVCH6 | 0.02 | 0.03 | 0.68839 | 0 |
| 7 | MULDIAM | 1.34 | 3.33 | 0.40375 | 0 |
| 8 | NXCH6 | 0.25 | 0.93 | 0.27348 | 0 |
| 9 | NELEM | 0.78 | 4.46 | 0.17600 | 0 |
| 10 | XVCH3 | 0.05 | 0.01 | 4.99920 | 1 |
| 11 | NXCH3 | 0.20 | 0.04 | 4.82111 | 1 |
| 12 | XVCH8 | 0.00 | 0.00 | 4.29085 | 1 |
| 13 | XVCH4 | 0.02 | 0.01 | 4.11889 | 1 |
| 14 | XVCH7 | 0.01 | 0.00 | 3.88779 | 1 |
| 15 | NXCH4 | 0.24 | 0.06 | 3.87353 | 1 |
| 16 | NXCH8 | 0.25 | 0.07 | 3.78884 | 1 |
| 17 | NXCH7 | 0.29 | 0.09 | 3.18114 | 1 |
| 18 | XVCH9 | 0.00 | 0.00 | 2.19305 | 1 |
| 19 | XVCH10 | 0.00 | 0.00 | 2.05207 | 1 |
| 20 | KNOTP | 7.09 | −4.09 | 1.73165 | 1 |
| 21 | NXCH10 | 0.45 | 0.28 | 1.61270 | 1 |
| 22 | NXCH9 | 0.45 | 0.28 | 1.59199 | 1 |
| 23 | KNOTPV | 3.75 | −2.54 | 1.47861 | 1 |
| 24 | XVCH5 | 0.04 | 0.03 | 1.15524 | 1 |
| 25 | NXCH5 | 0.49 | 0.59 | 0.83617 | 1 |
| 26 | MULRAD | 1.05 | 1.83 | 0.57114 | 1 |
| 27 | ISHAPE | 0.10 | 0.92 | 0.10380 | 1 |
| 28 | NE24 | 1.23 | 0.63 | 1.97352 | 2 |
| 29 | NTPATH | 45196.58 | 4859.45 | 9.30077 | 3 |
| 30 | WT | 221509.22 | 50140.62 | 4.41776 | 3 |
| 31 | NE34 | 1.16 | 0.55 | 2.09012 | 3 |
| 32 | TM | 2.35 | 2.19 | 1.07666 | 3 |
| 33 | XVC4 | 0.11 | 0.05 | 2.05613 | 4 |
| 34 | NE14 | 1.60 | 0.88 | 1.81890 | 4 |
| 35 | NE22 | 4.75 | 7.21 | 0.65916 | 4 |
| 36 | NCIRC | 2.18 | 2.84 | 0.76790 | 6 |
| 37 | ND4 | 0.88 | 0.54 | 1.63829 | 7 |
| 38 | NXC4 | 0.88 | 0.54 | 1.63829 | 7 |
| 39 | NUMHBD | 2.81 | 2.34 | 1.20322 | 8 |
| 40 | TETS2 | 51.31 | 33.75 | 1.52008 | 9 |
| | [partial omission due to space considerations] | | | | |
| 85 | SUMDELI | 20.12 | 25.72 | 0.78226 | 45 |
| 86 | ND3 | 4.93 | 9.42 | 0.52326 | 45 |
| 87 | NXP4 | 37.23 | 77.62 | 0.47960 | 46 |
| 88 | SI | 0.17 | 1.40 | 0.12259 | 46 |
| 89 | K2 | 6.58 | 11.47 | 0.57390 | 47 |
| 90 | NXC3 | 6.23 | 11.57 | 0.53836 | 47 |
| 91 | KA1 | 11.29 | 22.53 | 0.50085 | 48 |
| 92 | K1 | 11.98 | 24.61 | 0.48689 | 48 |
| 93 | K0 | 26.28 | 44.29 | 0.59323 | 49 |
| 94 | DXV0 | 3.75 | −7.25 | 0.51741 | 49 |
| 95 | IDWBAR | 1.09 | 8.33 | 0.13067 | 49 |
| 96 | NCLASS | 12.52 | 27.70 | 0.45191 | 50 |
| 97 | XV0 | 7.84 | 18.00 | 0.43540 | 50 |
| 98 | NVX | 13.10 | 30.29 | 0.43247 | 50 |
| 99 | FW | 186.06 | 430.50 | 0.43219 | 50 |
| 100 | SUMI | 34.99 | 73.30 | 0.47732 | 51 |
| 101 | XV1 | 4.74 | 10.84 | 0.43715 | 51 |
| 102 | XVP3 | 2.86 | 6.16 | 0.46515 | 52 |
| 103 | XV2 | 3.90 | 8.58 | 0.45468 | 52 |
| 104 | NXP1 | 13.86 | 32.58 | 0.42526 | 53 |
| 105 | DXV1 | 3.50 | −6.42 | 0.54537 | 54 |
| 106 | NXP3 | 26.98 | 59.77 | 0.45139 | 55 |
| 107 | WP | 26.98 | 59.77 | 0.45139 | 55 |
| 108 | NXP2 | 19.96 | 45.90 | 0.43481 | 56 |
| 109 | PF | 39.92 | 91.80 | 0.43481 | 56 |

**Table 2.** Variables Retained in the Analysis

| descriptor | position | property measured |
|---|---|---|
| gsolv | 1 | free energy of solvation |
| nvx | 2 | number of vertices |
| xv0 − xv2 | 5 | connectivity valence indices |
| xvp3 − xvp10 | 13 | connectivity valence path indices |
| xvc3 | 14 | connectivity valence cluster indices |
| xvpc4 | 15 | connectivity simple path/cluster-4 index |
| dxv0 − dxv2 | 18 | difference connectivity valence indexes |
| dxvp3 − dxvp10 | 26 | difference connectivity valence path indexes |
| k0 − k3 | 30 | kappa zero, kappa simple indexes |
| ka1 − ka3 | 33 | kappa alpha indices |
| si | 34 | Shannon information index |
| totop | 35 | total topological index |
| sumI | 36 | sum of intrinsic state values I |
| sumdelI | 37 | sum of delta-I values |
| tets2 | 38 | total topological index based on E-state |
| phia | 39 | flexibility index |
| idw | 40 | Bonchev−Trinajstic information index |
| idwbar | 41 | Bonchev−Trinajstic information index, norm. |
| idc | 42 | Bonchev−Trinajstic information index |
| idcbar | 43 | Bonchev−Trinajstic information index, norm. |
| W | 44 | Wiener W number |
| Wp | 45 | Wiener P number |
| pf | 46 | Platt f number |
| nxp1 − nxp10 | 56 | count of path subgraphs |
| nxpc4 | 57 | count of path/cluster-4 subgraphs |
| tg | 58 | terminal groups |
| diam | 59 | graph diameter |
| rad | 60 | graph radius |
| nd1 | 61 | number of vertexes for which delta = 1 |

**Table 3.** Database Size and Duplicate Entries

| database | size | size, no dups | % duplicates |
|---|---|---|---|
| CMC | 5 285 | 4 708 | 11 |
| MDDR | 50 387 | 48 181 | 4 |
| SPECS | 54 932 | 43 281 | 21 |
| ACD | 142 163 | 121 825 | 14 |
| WR | 106 408 | 99 857 | 6 |
| superpopulation | 359 175 | 317 852 | 11 |

of descriptors, the descriptor space defines a diversity space for typical organic molecules.

Duplicate structures can occur within or between databases because there are duplicate entries and because counterion removal creates identical entries. Additionally, the use of nonstereochemical SMILES meant that enantiomers and diastereomers became identical. Duplicates were eliminated at the beginning of the analysis when the superpopulation was formed. Since identical structures could have different registry numbers in different databases, duplicates were eliminated on the basis of the topological indices. Table 3 lists the number of compounds before and after duplicates were removed.

**Treatment of Descriptor Space and Volume.** After the descriptors were chosen, a factor analysis[10] was performed to reduce the dimensionality of the data. Four factors were sufficient to explain 90% of the variation in the data, while seven factors were needed to explain over 95% of the variation. A criterion often used for selecting the optimal number of factors is to use the eigenvalues as a cutoff; the point at which the eigenvalues are less than 1.0 represents a "shrinking away" from that dimension. This criterion would suggest six factors as the optimal since the sixth eigenvalue was 1.08 and the seventh eigenvalue was 0.796. Analyses were done with four, five, and six factors, but, for reasons to be explained later, the dimensionality issue made four factors the best choice.

The factor analysis resulted in each of the factors being standardized, with a mean of zero and a standard deviation equal to one. However, they do not have the same ranges and can be very different with respect to skewness.

The factors were rotated orthogonally so as to maximize the loadings of a few variables on each factor. The result of this seen in Table 4 was that GSOLV overwhelmingly loaded into factor 8 by itself with a score of 94. Because it is a physical property describing polarity, it was desirable to keep GSOLV in any diversity analysis. Thus for example in the 5-factor studies, factors 1−4 and factor 8 were used. This is not an unusual practice; for example, in regression analyses based on principal components it is not always the earliest components that are the most significant in the regression. In previous work[8] in which 10 principal components were computed, it was discovered that the three most important PCs were the eighth, second, and tenth in that order. The eighth PC accounted for only 3% of the variation in $x$ but explained 24% of the variation in the dependent variable in the regression.

The factors in Table 4 can be interpreted and show that certain kinds of indices can load uniquely into one factor. Factor 1 is related to many of the $\chi$ indices which are size dependent, whereas factor 2 is related to the difference $\chi$ indices which are designed to be size independent. Factor 3 is related to the flexibility and shape indices. Factor 4 is related to the number of terminal groups. Factor 5 is related to the total topology indices TOTOP and TETS2. Factor 6 has relatively weak loadings and is hard to interpret, but factor 7 is related to the Bonchev−Trinajstić information indices (these are information indexes like the Shannon index; the reader is referred to Hall[7]). Factor 8 is uniquely the free energy of solvation.

**Volume Scaling.** Given that the volume of hyperspace balloons as the number of dimensions increases and given that it was necessary to experiment with different dimensionalities (numbers of factors) and compare results, the volume measurements have been made independent of dimensionality by using "scaled" volumes. Thus, the Euclidean volume is raised to the power of 1 divided by the number of factors used. If it is necessary to use six factors instead of eight, for instance, the resulting six-factor volume measures will be on an "equal footing" with the same measure using eight factors. In effect, the Euclidean volume measure becomes a geometric mean.

## THE ALGORITHM

The algorithm was coded using the SAS statistical software package.[9]

**Gridding of Descriptor Space.** As discussed above factor analysis reduces the 61 variables to four factors. The occupied volume of the resulting factor space was estimated in the next step. To get accurate volume estimates of the occupied regions of descriptor space, a Riemann style approach[11] was used. The descriptor space was partitioned (gridded) into subhypercubes (a subdivision of hyperdimensional space), each of equal size. This volume was computed as a unit volume. The number of occupied subhypercubes was then counted and multiplied by the unit volume to get an estimate of the total occupied volume. In this way, much of the empty "inner space" (subhypercubes that contain no observations) was prevented from inflating the volume estimate.

**Table 4.** Factor Patterns: Rotated Factor Pattern (Rotation Method: Quartimax)[a]

| descriptor | (size) factor1 | (diff $\chi$) factor2 | (flex/shape) factor3 | (term grps) factor4 | (totop) factor5 | (??) factor6 | (bonchev) factor7 | (gsolv) factor8 |
|---|---|---|---|---|---|---|---|---|
| PF | 99* | −6 | −9 | 2 | 0 | 2 | 0 | −1 |
| NXP2 | 99* | −6 | −9 | 2 | 0 | 2 | 0 | −1 |
| NXP1 | 99* | −11 | 4 | −5 | −2 | 0 | 1 | 0 |
| NVX | 98* | −14 | 11 | 0 | −3 | −1 | 2 | −1 |
| NXP3 | 97* | 1 | −22 | −4 | 3 | 1 | −1 | 1 |
| WP | 97* | 1 | −22 | −4 | 3 | 1 | −1 | 1 |
| K0 | 96* | −17 | 12 | 1 | −2 | −2 | 10 | −2 |
| XV0 | 96* | −6 | 21 | 1 | −4 | 10 | 2 | 0 |
| XV1 | 95* | 4 | 23 | −3 | −5 | 13 | 2 | −3 |
| K1 | 94* | −20 | 25 | 10 | −5 | −1 | 4 | −1 |
| NXP4 | 94* | 8 | −31 | −9 | 7 | 1 | −1 | 1 |
| KA1 | 93* | −16 | 29 | 13 | −5 | 2 | 4 | −2 |
| IDWBAR | 93* | −11 | 3 | −7 | −10 | 1 | −28 | 1 |
| XV2 | 92* | 18 | 17 | 8 | −6 | 24 | 2 | −4 |
| NXP5 | 90* | 13 | −35* | −11 | 12 | −1 | −2 | 2 |
| SUMI | 90* | −26 | 4 | 26 | −4 | −13 | 2 | −5 |
| XVP3 | 90* | 34 | 8 | 1 | −7 | 22 | 2 | −2 |
| NXP6 | 87* | 18 | −38* | −10 | 18 | −2 | −3 | 2 |
| IDC | 85* | −19 | 21 | 6 | 2 | −1 | 42* | −3 |
| NXPC4 | 84* | 14 | −39* | 23 | 8 | 6 | −3 | 2 |
| NXP7 | 83* | 21 | −38* | −12 | 23 | −4 | −3 | 2 |
| XVP4 | 83* | 48* | 3 | 0 | −8 | 20 | 3 | −2 |
| K2 | 81* | −24 | 50* | −4 | −5 | −4 | 6 | 1 |
| SI | 81* | −17 | 0 | −12 | −11 | −6 | −28 | −1 |
| RAD | 81* | −20 | 43* | −17 | −7 | −1 | −14 | −1 |
| NXP8 | 80* | 23 | −37* | −14 | 29 | −6 | −3 | 2 |
| DIAM | 80* | −20 | 43* | −18 | −8 | −1 | −15 | −1 |
| XVP5 | 79* | 57* | 0 | −2 | −6 | 13 | 2 | 0 |
| KA2 | 79* | −19 | 56* | −1 | −4 | −1 | 6 | 1 |
| SUMDELI | 78* | −26 | −2 | 47* | −3 | −16 | 9 | −7 |
| NXP9 | 77* | 24 | −36* | −14 | 35* | −7 | −4 | 3 |
| IDCBAR | 77* | −20 | 35* | −19 | −11 | −2 | −38* | 0 |
| W | 76* | −20 | 26 | 5 | 1 | −1 | 54* | −4 |
| NXP10 | 75* | 24 | −33 | −13 | 41* | −8 | −3 | 3 |
| IDW | 72* | −19 | 25 | 6 | 1 | −1 | 58* | −4 |
| TOTOP | 72* | −3 | −12 | 0 | 62* | −3 | 1 | 0 |
| XVP6 | 71* | 68* | 2 | 3 | −5 | 7 | 1 | 0 |
| PHIA | 69* | −18 | 66* | 6 | −3 | 1 | 7 | 0 |
| XVPC4 | 65* | 50* | −15 | 30 | −5 | 38* | 0 | −2 |
| XVC3 | 58* | 32 | −2 | 45* | −2 | 46* | −3 | −5 |
| DXV0 | −83* | 34 | 12 | −15 | 3 | 29 | 0 | 6 |
| DXV1 | −84* | 42* | 9 | −18 | 2 | 24 | 0 | 0 |
| DXVP7 | −18 | 97* | −8 | −2 | 1 | −3 | −1 | 1 |
| DXVP8 | −19 | 96* | −5 | −1 | 3 | −12 | −1 | 1 |
| DXVP6 | −18 | 95* | −9 | −1 | −2 | 10 | 0 | 2 |
| DXVP5 | −19 | 93* | −12 | −8 | −3 | 20 | 0 | 2 |
| DXVP9 | −24 | 93* | −4 | −1 | 7 | −21 | −3 | 1 |
| DXVP10 | −30 | 89* | −2 | 1 | 9 | −26 | −3 | 1 |
| DXVP4 | −28 | 88* | −11 | −7 | −5 | 31 | 1 | 0 |
| DXVP3 | −44* | 79* | −7 | −8 | −3 | 37* | 0 | 0 |
| XVP8 | 63* | 75* | 4 | 2 | 0 | −13 | 0 | 0 |
| XVP9 | 61* | 75* | 5 | 1 | 4 | −22 | −1 | 0 |
| XVP10 | 57* | 74* | 7 | 2 | 6 | −28 | −1 | 0 |
| XVP7 | 68* | 72* | 2 | 1 | −2 | −4 | 1 | 0 |
| DXV2 | −62* | 63* | 4 | 0 | 0 | 43* | −1 | −2 |
| KA3 | 56* | −20 | 77* | 6 | 0 | 0 | 6 | −3 |
| K3 | 60* | −25 | 73* | 4 | −1 | −4 | 6 | −2 |
| TG | 62* | −21 | 9 | 72* | −2 | 2 | 4 | −4 |
| ND1 | 62* | −21 | 9 | 72* | −2 | 2 | 4 | −4 |
| TETS2 | 55* | 8 | −9 | 0 | 75* | 1 | 4 | −1 |
| GSOLV | −32 | 9 | −4 | −6 | 0 | −1 | −2 | 94* |

[a] Printed values have been multiplied by 100 and rounded to the nearest integer. Values greater than "34" have been flagged by an *.

This was necessary because the superpopulation of structures fills only a minor part of the descriptor space. Most of the structures fall into a rather small, densely populated region of space (as shown in Figure 1). A small percentage fall into outlying regions where the density of structures is low. Thus, there is a great deal of descriptor space with no chemical representatives.

Additionally, it is difficult to see how the databases compare because they are so highly overlapping. A strategy to remove this difficulty is to focus the analysis on the highly populated part of descriptor space. This is done by removing the outlying structures from the analysis.

Subhypercubes containing few observations represent low density space; these observations can be defined as outliers.

A robust volume analysis was performed in which these outlier observations were removed iteratively to "trim" the superpopulation in a data-driven fashion. The effect of removing low density observations is to focus in on the "core" of the diversity space so that databases can be compared without the "squashing" effect that outliers have on the main core of diversity space. Multiple iterations are needed, since when the worst outliers are removed, the core is allowed to "breathe", expanding slightly. But this expansion may reveal additional outliers that were masked by the more extreme ones, and so it is necessary to iterate until this phenomenon subsides. The outliers are of interest in their own right and are listed in a separate file for each iteration. A study of the outliers should not be ignored but is different from the superimposition approach that is the focus of this paper.

In practice, low density space is defined in terms of the number of observations per subhypercube. For this application it was decided that any subhypercube containing seven or fewer observations would be labeled as a low density subvolume; the observations were removed and the space contracted. It was observed that four iterations of the algorithm were sufficient to remove the serious outliers. After four iterations all of the extreme outliers were eliminated (compare Figures 1 and 2), with minimal loss of data: 310 866 observations remained out of the 317 852 total. At that point, the databases could be compared in detail. [The determination of what constitutes low-density is very subjective. We did various studies in which we varied the criterion from subcubes containing one observation up to as many as 20 observations. Qualitatively, results such as presented in Table 9 are not sensitive to the outlier criterion. The most useful diagnostic was observing plots like Figure 2 for various outlier criteria, coupled with a prior belief that less than 10% of the data were outliers (our results flagged 2.2% as outliers). The best choice will be unique to each application.]

**Determination of the Optimal Number of Subhypercubes.** This technical development has been placed in the Appendix.

**Removal of Outliers.** The presence of a relatively few outlying observations had a significant effect on this analysis. They inflated the descriptor space used in the analysis and, in the process, caused 95% of the observations to be highly clustered. As a result, most of the observations fell into a few subhypercubes, and most subhypercubes were empty. Since most observations fell into a small number of subhypercubes, this densely occupied space is not adequately subdivided, and the analysis becomes trivial. For example, if all five databases fell into one subhypercube, all that can be said is that all five databases occupy the same space. Adequate database comparisons can only be achieved by spreading the observations over many subhypercubes.

The problem was solved by removing the outlying observations. As mentioned previously observations were deleted in subhypercubes containing seven or fewer observations. This had the effect of reducing the range and volume of descriptor space. The gridding was repeated. (Since the same number of subintervals is always used to subdivide each axis, the volume of each subhypercube is smaller as the overall volume decreases.) New subhypercubes resulted that contained less than seven observations; these were deleted. The gridding was repeated until the volume of descriptor space changed minimally from iteration to iteration, or when the ranges of descriptor space ceased to change rapidly.

**Table 5.** Ranges for the Descriptor Space, by Iteration[a]

| iter | min1 | max1 | min2 | max2 | min3 | max3 | min4 | max4 |
|---|---|---|---|---|---|---|---|---|
| 1 | −2.22 | 9.50 | −6.18 | 50.56 | −5.47 | 26.08 | −64.12 | 4.10 |
| 2 | −2.22 | 8.79 | −5.38 | 21.0 | −5.28 | 11.74 | −23.01 | 2.74 |
| 3 | −2.22 | 8.68 | −3.96 | 11.7 | −4.41 | 10.85 | −21.52 | 2.38 |
| 4 | −2.22 | 7.56 | −3.96 | 10.01 | −4.41 | 9.14 | −21.52 | 2.38 |
| 5 | −2.22 | 6.58 | −3.84 | 9.21 | −4.41 | 9.14 | −21.10 | 2.38 |
| 6 | −2.22 | 6.58 | −3.84 | 9.21 | −4.39 | 9.14 | −21.10 | 2.38 |
| 7 | −2.22 | 6.58 | −3.84 | 9.21 | −4.39 | 9.14 | −21.10 | 2.38 |

[a] Listed are the min. and max. values for the four factors used. Generated using 20 Riemann subintervals. The definition of an outlier is any observation in a subcube with seven or fewer total observations. One can see that even after removing the outliers (iters 4+) the ranges of the data show skewness reflecting the fact that the data are not Gaussian distributed.

**Table 6.** Volume of Superpopulation, by Iteration and Database, Based on Four Factors[a]

| iteration | volume | % change | CMC | MDDR | SPECS | ACD | WR |
|---|---|---|---|---|---|---|---|
| 1 | 10.23 | | 6.34 | 8.52 | 7.85 | 9.19 | 8.45 |
| 2 | 6.656 | −34.9 | 4.30 | 5.70 | 4.98 | 5.93 | 5.70 |
| 3 | 5.527 | −17.0 | 3.72 | 4.85 | 4.28 | 4.96 | 4.89 |
| 4 | 4.987 | −9.76 | 3.51 | 4.46 | 4.00 | 4.53 | 4.49 |
| 5 | 4.646 | −6.84 | 3.37 | 4.22 | 3.78 | 4.27 | 4.25 |
| 6 | 4.220 | −9.18 | 3.32 | 3.99 | 3.70 | 4.04 | 4.04 |
| 7 | 4.164 | −1.32 | 3.32 | 3.97 | 3.68 | 4.01 | 4.03 |

[a] This volume is a "scaled" volume, which minimizes the volume changes. The raw Euclidean volume can be obtained by raising the scaled volume to the fourth power.

**Table 7.** Outlier Report, by Iteration and Database

| iteration | observ | total | CMC | MDDR | SPECS | ACD | WR |
|---|---|---|---|---|---|---|---|
| 1 | 317 852 | 1415 | 19 | 366 | 241 | 549 | 240 |
| 2 | 316 437 | 2963 | 51 | 875 | 327 | 1001 | 709 |
| 3 | 313 474 | 2608 | 32 | 811 | 286 | 785 | 694 |
| 4 | 310 866 | 2290 | 48 | 729 | 262 | 638 | 613 |
| 5 | 308 576 | 1702 | 31 | 568 | 145 | 493 | 465 |
| 6 | 306 874 | 132 | 4 | 32 | 24 | 45 | 27 |
| 7 | 306 742 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 1 shows the original descriptor space at iteration 1, and Figure 2 shows the descriptor space at iteration 4. The extreme outliers present at iteration 1 are removed at iteration 4, with the consequence that the descriptor space is focused on the high density observations.

Table 5 shows that the min. and max. values of the descriptor space, in terms of factors, cease to change after iteration 4. Table 6 shows that the volume of descriptor space also ceases to change drastically after four iterations. The volume decreased by a factor of 16 from iteration 1−4 ($10.23^4/4.987^4$). Thus the outliers exacerbated the volume problem in high dimensional space. Table 7 shows the number of outliers deleted from each database at each iteration; most of them were deleted from the bigger databases, MDDR, ACD, and the Wellcome Registry. Less than 10 000 observations total were deleted as outliers. Table 8 shows statistics on the number of occupied subcubes by iteration. The median number of observations per subcube increases steadily, and the number of observations in the most populated subcube decreases steadily, as the outliers are progressively removed.

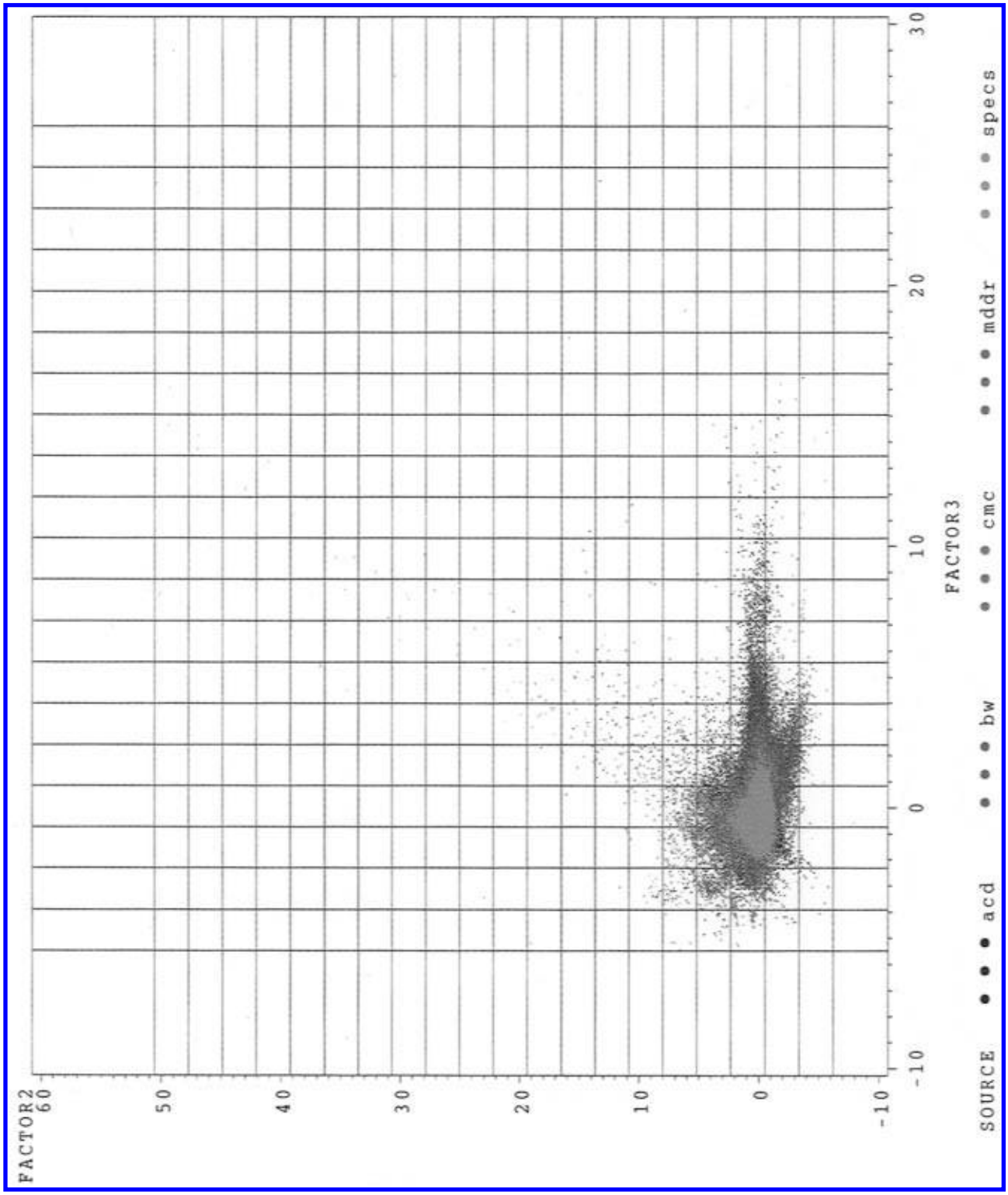


**Figure 1.** Comparison of five databases: iteration 1. Outlier is defined as seven observations or less in a subhypercube.

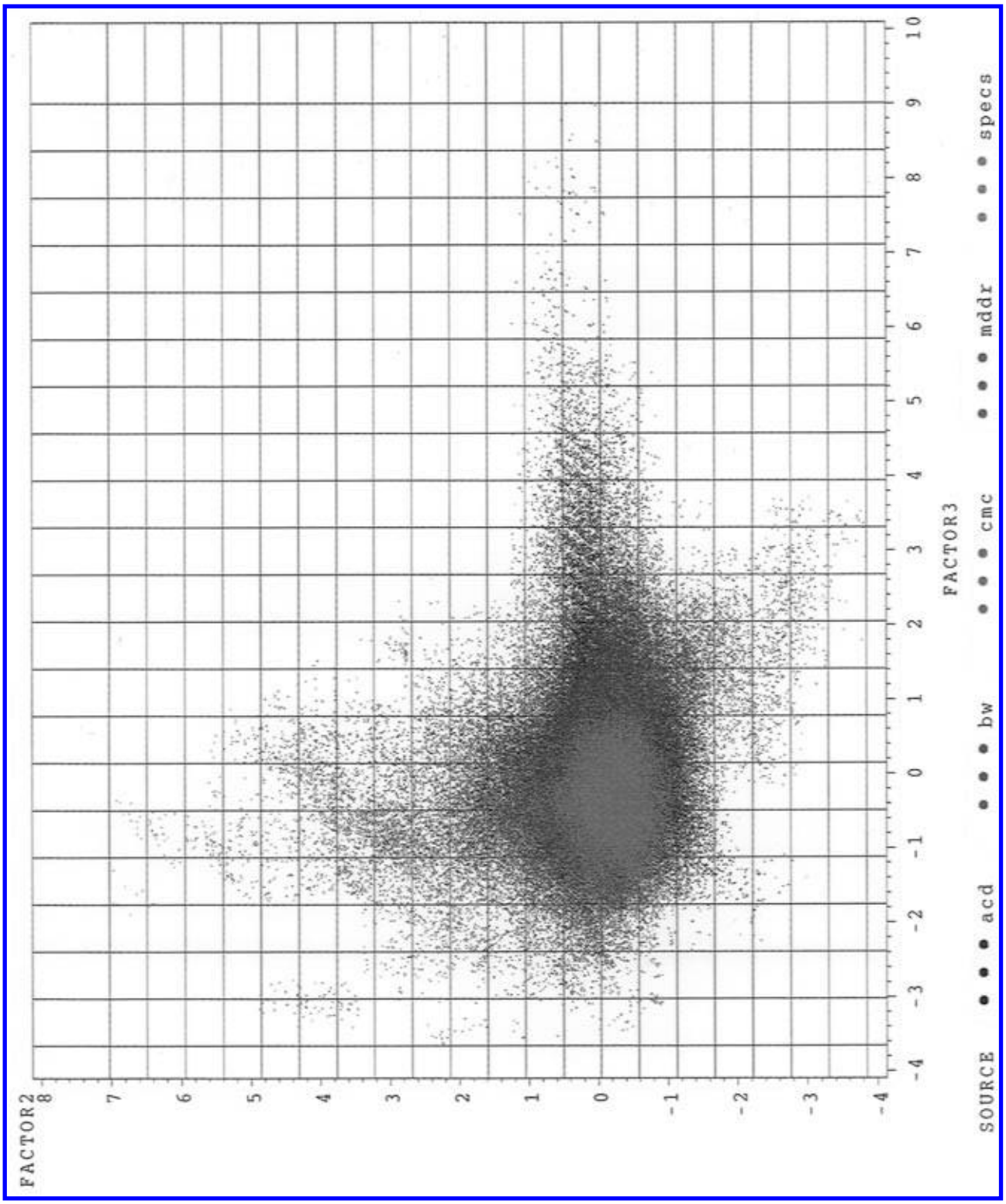


**Figure 2.** Comparison of five databases: iteration 4. Outlier is defined as seven observations or less in a subhypercube.

**Table 8.** Statistics on Subcubes Occupied, by Iteration

| iteration | no. of subcubes occupied | median no. of observations | no. of observations in most populated subcube |
|---|---|---|---|
| 1 | 1223 | 3 | 65 458 |
| 2 | 3048 | 3 | 14 428 |
| 3 | 2266 | 8 | 13 598 |
| 4 | 2059 | 11 | 15 199 |
| 5 | 2331 | 7 | 15 281 |
| 6 | 2149 | 12 | 12 065 |
| 7 | 1592 | 22 | 10 213 |
| 8 | 1240 | 37 | 10 213 |

**Table 9.** Matrix of Pairwise Database Overlaps for Iteration 4

| | CMC | WR | MDDR | SPECS | ACD |
|---|---|---|---|---|---|
| | | Raw Count of Occupied Subcubes | | | |
| CMC | **563** | 506 | 515 | 460 | 516 |
| WR | | **1480** | 1069 | 797 | 1186 |
| MDDR | | | **1423** | 736 | 1058 |
| SPECS | | | | **939** | 835 |
| ACD | | | | | **1494** |
| | | Volume Percentages of Superpopulation Based on Occupied Subcubes | | | |
| CMC | **27%** | 90% | 91% | 82% | 92% |
| WR | 34% | **72%** | 72% | 54% | 80% |
| MDDR | 36% | 75% | **69%** | 52% | 74% |
| SPECS | 49% | 85% | 78% | **46%** | 89% |
| ACD | 35% | 79% | 71% | 56% | **72%** |

In Table 8 is the output showing subhypercube memberships by iteration. By the eighth iteration the largest subcube contains 10 213 compounds as compared to 65 458 in the first iteration. These numbers seem very large, especially when one considers that the smallest database (CMC) contains only 4708 compounds. This is a result of the combined influence of three factors. First, the data are highly clustered. The diversity space is divided into equal subsections, but the data are far from being uniformly spread throughout. Thus many subhypercubes will be empty, and this is actually the point of doing the subhypercube divisions, so that the empty space is not counted. Secondly, outliers have a squashing effect on the core of the diversity space, and, as the output below shows, the removal of those outliers greatly relieves that problem. The third factor influencing this is the number of subdivisions taken along each factor (discussed in the Appendix).

One can see from the median that in iteration 1 the typical subcube contained three observations and by iteration 8 the typical subcube contained 37 observations, showing that the low density subcubes have been removed in a fairly gradual fashion. The effect is that the volume is reduced more quickly than the number of observations, so that the median number of observations per subvolume increases. It is also interesting to note how the median dips down at iteration 5 and then goes back up. This is due to the effect of the database "breathing", described in the section on the gridding of descriptor space.

## RESULTS

**Database Comparisons.** An initial assumption of the study was that the MDDR database represents the universe of drug activity. As discussed above, it was necessary to remove a small number of extreme compounds from each database including MDDR. Outlier compounds removed from the other databases could be active compounds but were not further investigated. Their structural uniqueness alone does not imply biological activity. However, these outliers represent relatively few compounds, and their distribution in MDDR space could be separately evaluated.

Comparative statistics were accumulated on a per subvolume basis. Thus, the occupants of each subvolume were identified by database of origin and counted. Pairwise database overlaps were then computed, meaning the number of subvolumes that the two databases shared. Table 9 shows the pairwise overlaps for the five databases in the superpopulation.

The MDDR and CMC databases are knowledge bases of biological activity data and, for the purpose of this study, can be combined to represent a grand knowledge base. The subvolumes containing this knowledge base represent the known "universe" of biological activity in descriptor space. Commercial compounds from the ACD or SPECS databases can be classified as "active" or "inactive" by whether the commercial compound falls into a subvolume of the knowledge base. These classifications are predictions based on the information in the knowledge base. An active classification is certainly not a guarantee of biological activity, since it is well-known that the substitution of a methyl group at the proper site can convert an active drug into an inactive compound. Such a small change would change the position of the structure in descriptor hyperspace but may leave it in the same subhypercube where activity is predicted. Nonetheless, the classifications are a rational way of selecting structures for testing.

In addition, unique compounds are identified using the subhypercubes as well. Any subhypercube for which all the observations come from only one database triggers a uniqueness flag for each compound in that subhypercube. The number of compounds in the subhypercube is not relevant, only the fact that no other database has compounds in that subvolume of diversity space. These unique compounds are listed in separate files for each iteration. It will almost certainly be the case that some compounds are both unique and an outlier at the same time, since a subhypercube with seven or fewer observations, all of which are from the same database, would fit both the definition of unique and the definition of outlier.

Table 9 gives the fourth iteration database overlap results. The first matrix gives the raw counts of number of subhypercubes jointly occupied. The diagonals of this matrix give the number of subcubes each database occupies, e.g., CMC compounds lie in 563 subhypercubes. Reading across the first row one sees that of those 563 subcubes, 506 are also occupied by WR compounds, 515 are also occupied by MDDR compounds, etc.

The second matrix presented in Table 9 converts the subcube numbers into percentages. The diagonal elements, given in bold print, represent how much volume (i.e., number of subcubes) the database occupies as a percentage of the whole superpopulation. The off-diagonal elements compare the databases pairwise. These can be interpreted as the numerator being obtained from the column element and the denominator from the row element. For example, row 1 shows what percentage of CMC subcubes the other databases co-occupy, and column 1 shows what percentage CMC co-occupies with the other databases. As another example, the Wellcome Registry (WR) occupies 85% of the subcubes occupied by the SPECS database whereas SPECS only

**Table 10.** Number of Commercial Structures Overlapping with the Biologically Active Space (MDDR + CMC)

| database and logic | no. of subhypercubes | % of biological space occupied | no. of compds |
|---|---|---|---|
| SPECS | 758 | 52 | 41 527 |
| SPECS, NOT WR | 44 | 3 | 248 |
| ACD | 1093 | 74 | 116 640 |
| ACD, NOT WR | 134 | 9 | 383 |

occupies 54% of WR subcubes. WR occupies 72% of the superpopulation subcubes, and SPECS occupies 46% of those subcubes.

CMC occupies the smallest descriptor volume (27% of the occupied subcubes), in line with the fact that it contains far fewer compounds than the other databases but also due to the fact that CMC represents a subset of MDDR that became marketed drugs. This is consistent with the fact that CMC occupies only 36% of the MDDR database volume.

MDDR, the largest biological knowledge base, and ACD, the largest commercial database, yield an interesting comparison. MDDR and ACD occupy descriptor volumes of roughly equal size (69% versus 72%), despite the fact that MDDR contains roughly half the number of compounds as ACD. The two databases overlap rather extensively: MDDR occupies 71% of the ACD subcubes and ACD occupies 74% of the MDDR subcubes. This is reasonable since many of the commercial compounds are of biological interest and vice versa. It is also reasonable that MDDR is slightly smaller than ACD in terms of descriptor volume; even though ACD contains biological compounds, it contains many compounds of synthetic interest, as well as other unique types of compounds.

**Comparisons with Biologically Active Space (CMC + MDDR).** There is a large amount of overlap between the commercial databases and the biologically active space. Figure 3 shows the commercial compounds (ACD + SPECS) in black and the medicinal knowledge databases (MDDR + CMC) in red and shows the high degree overlap between the two sets. Shown in Table 10 are the number of commercial compounds in SPECS and ACD predicted to be biologically active by virtue of overlapping with the volume occupied by the union of MDDR and CMC.

There are 41 527 SPECS compounds predicted to have biological activity, but only 248 of those do not overlap with the volume occupied by the Wellcome Registry. Also, there are 116 640 ACD compounds predicted to have biological activity, but only 383 of those do not overlap with the volume occupied by the Wellcome Registry. The column labeled "% of biological" shows what percentage of the biologic descriptor space is represented, e.g., the portion of SPECS that overlaps with the biological space occupies 52% of that space, whereas the portion of SPECS that overlaps with the biological space but is not represented by Wellcome compounds occupies 3% of the biological space.

**Uniqueness Volumes.** Table 11 gives the volumes of descriptor space that each database occupies uniquely, i.e., areas that each database occupies which no other database occupies. These volumes are also expressed as a percentage of volume of the total diversity space. Thus for example line 1 shows that at the first iteration the MDDR database contributed 6.1% of the total superpopulation volume uniquely, so that the superpopulation volume would be 6.1% lower if the MDDR data were absent. Also, by the seventh iteration the uniqueness contribution for MDDR went down to 1.8%. This effect is seen for each of the databases. The iterative removal of low density subhypercubes allows one to "zoom in" on the core areas where, as it turns out, the databases largely overlap and uniqueness is harder to observe.

**Table 11.** Uniqueness Volumes Expressed as Percentage

| iteration | CMC | MDDR | WR | SPECS | ACD |
|---|---|---|---|---|---|
| 1 | 2.4 | 6.1 | 5.3 | 5.5 | 6.9 |
| 2 | 1.8 | 4.0 | 3.7 | 3.1 | 4.3 |
| 3 | 1.2 | 3.3 | 3.0 | 2.4 | 3.2 |
| 4 | 1.3 | 2.9 | 2.7 | 2.0 | 2.7 |
| 5 | 1.0 | 2.6 | 2.4 | 1.6 | 2.4 |
| 6 | 0.9 | 1.9 | 1.7 | 1.3 | 1.7 |
| 7 | 0.0 | 1.8 | 1.5 | 0.9 | 1.3 |

Of particular note is the fact that CMC has no uniqueness at all by the seventh iteration, which reflects the restrictive nature of the CMC database. The other databases thoroughly explore the space occupied by CMC, as they should.

At the fourth iteration of outlier removal from the superpopulation, 142 subhypercubes were identified which contained only ACD structures. The structures in these subcubes therefore represent compounds in ACD that are not present in the other databases, including the biological databases. Figure 4 shows structures from nine of these subhypercubes chosen from different structural classes. The subcube number is found in the upper left of each header. Several points are of note. First, many of the molecules are obviously unique, particularly the perfluorinated or long chain alkyl compounds. Secondly, the structural similarity within each subcube is quite high. One reason for this is the use of the free energy of solvation descriptor which has the effect of classifying molecules by functional groups, i.e., polarity.

## CONCLUSION

Five molecular databases were superimposed and compared with each other pairwise and this comparison was summarized with a single metric (Table 9). Each database was compared with the superpopulation as well to give an idea of how well each database represents the structural diversity of the whole collection. The subspace of the superpopulation that represents biologically active drugs was defined and identified in descriptor space (Figure 3). Commercial compounds that overlap with this biologically active space were identified and marked for possible purchase (Table 10). As a matter of special interest, compounds that were unique to each database were identified (Figure 4). These were the cases in which a subhypercube was occupied by compounds from only one database.

In obtaining these results two major technical problems were overcome. Outlying structures expanded the volume of descriptor space by more than an order of magnitude and caused the bulk of the structures to be severely clumped together. This was handled with an iterative removal of low-density observations. The problem of how to accurately measure the occupied volume in descriptor space was solved by discarding the unoccupied areas of descriptor space with a Riemann style gridding. This gridding avoided inflated volume estimates, while limiting the resolution of the gridding avoided deflated volume estimates.

This work was done as a volume analysis. Another approach that has been considered is a nearest-neighbor

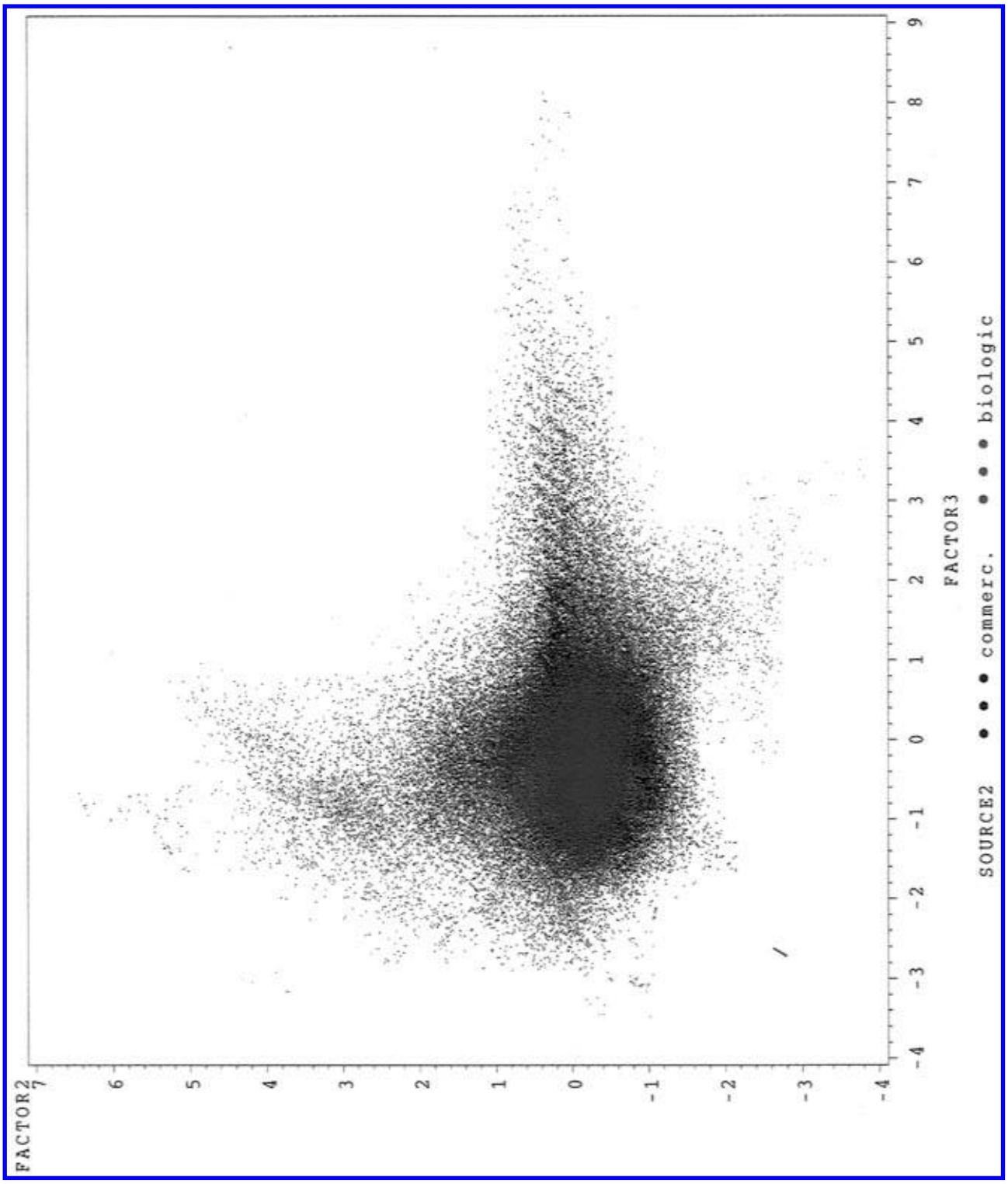


**Figure 3.** Iteration 4. Outlier is defined as seven observations or less in a subhypercube.

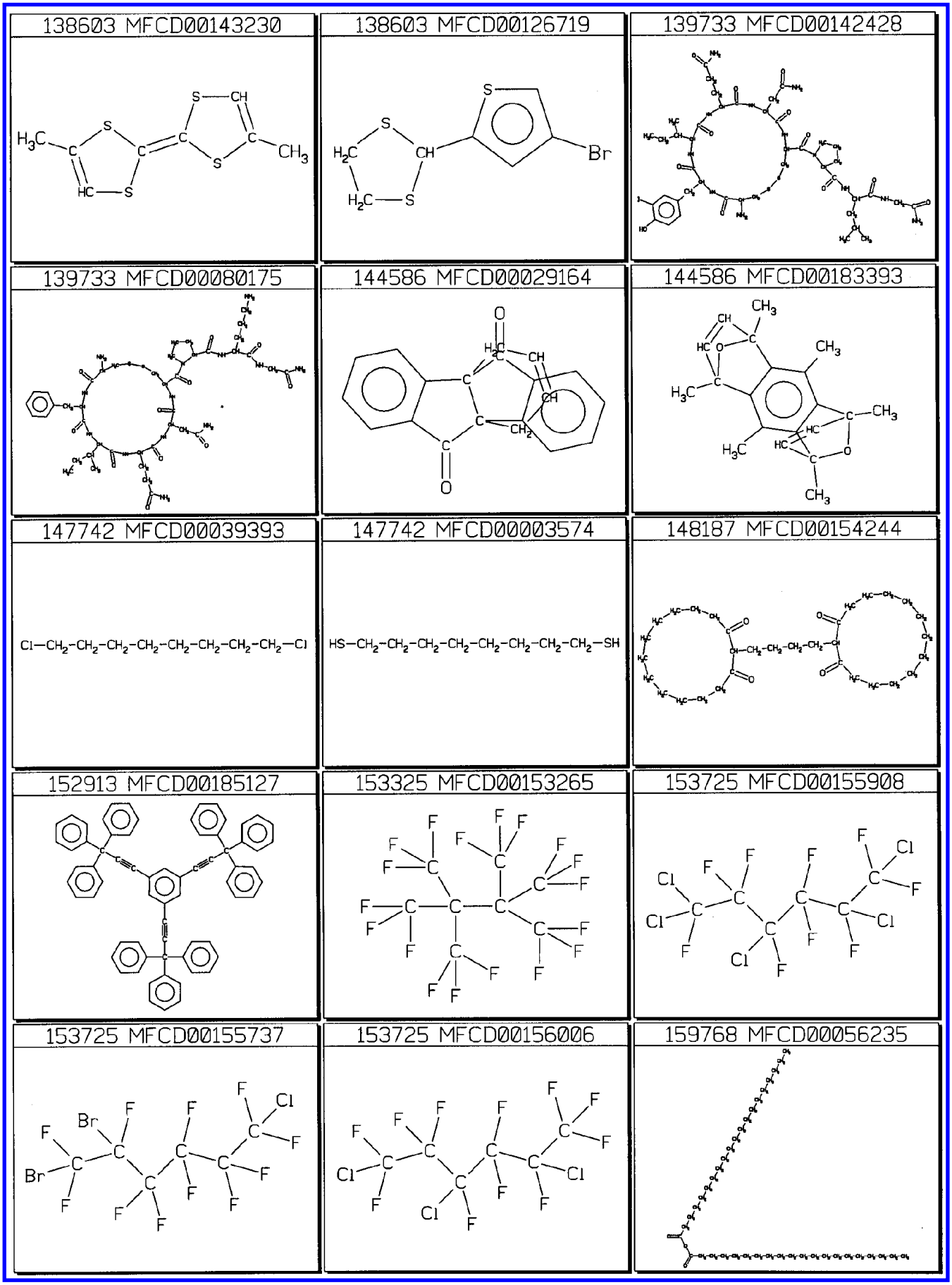


**Figure 4.** Structures unique to ACD. Structures from 9 of 142 subcubes are shown.

approach. The first step of this approach would be to create a grid of points that gives a good representation of the descriptor space occupied by the superpopulation. A space-filling design would be a valuable tool for this; an alternative

would be a density-based clustering. With the grid in place, the procedure would simply take each observation in each database, find the grid point closest to the observation, and record the distance from the observation to the grid point (using Euclidean distance or $L_1$ or similar distance metric). This can be used to get an idea of how "close" each database can get to each of the representative grid points. The benefit of doing a nearest-neighbor analysis is that it is free of the difficulty of getting accurate volume estimates. The problem with the approach lies in constructing useful summary measures. It is also true that outliers are as much of a problem for the nearest-neighbor approach as they are for the superimposition approach used here.

## APPENDIX

**Determination of the Optimal Number of Subhypercubes.** The iterative "Riemann" algorithm employed herein uses a fixed number of subhypercubes which did not change with iteration. The number of subhypercubes is computed by eq 1.

$$\#\text{subhypercubes} = r^d \quad (1)$$

where $r$ is the number of subdivisions taken along each factor axis and $d$ is the number of factors used.

**Table 12.** Volume of the Superpopulation (Pilot Study of 110 000) of Five Databases as a Function of Number of Dimensions ($d$) and Number of Subintervals ($r$) of each Dimension

| $r$ | vol, $d = 3$ | vol, $d = 5$ | vol, $d = 8$ |
|---|---|---|---|
| 1 | 30.77 | 30.18 | 48.10 |
| 2 | 26.31 | 25.21 | 34.27 |
| 5 | 19.74 | 15.57 | 18.67 |
| 10 | 15.22 | 10.79 | 11.09 |
| 20 | 10.72 | 6.98 | 6.55 |
| 50 | 6.96 | 3.93 | 3.28 |

Equation 1 shows that the number of subhypercubes is a function of the number of factors and the number of subintervals taken along each axis in hyperspace. Creating more partitions would mean smaller subhypercubes and therefore each subcube would contain fewer compounds. This would seem to be beneficial in that the data, which are highly clustered, would not be clumped into a small number of volume elements. Furthermore, the Riemann approach to approximation prescribes that accuracy increases as the intervals get smaller and more numerous. There is however, a limit to how many subintervals one can take, and this limit is not due simply to computational considerations. As the descriptor space is subdivided more and more finely, each observation will at some point be in a subcube by itself. At the limit, the occupied volume will go to zero (the volume of a point in space is zero). To illustrate the point, Table 12 shows the computed volume of the superpopulation using three, five, and eight factors, and various numbers of partitions. The volumes reported are scaled and represent the occupied volume, that is, the sum of the occupied subvolumes. The data in Table 12 were obtained from a pilot study using a random sample of 110 000 out of the 317 852 compounds.

It is evident that the volume estimate goes down as the number of subintervals goes up, in accord with expectation. The fact that the volume goes down is reflecting the fact that more and more empty space is eliminated from the volume estimate as the space is divided into finer subportions.

**Table 13.** Simulated Superpopulation Volumes (size = 110 000), Where $d$ = Number of Dimensions and $r$ = Number of Subintervals Taken Along Each Dimension

| $r$ | $d = 3$ vol | $d = 5$ vol | $d = 8$ vol |
|---|---|---|---|
| 1 | 17.20 | 20.37 | 20.38 |
| 2 | 17.20 | 20.37 | 20.38 |
| 5 | 17.20 | 20.37 | 16.92 |
| 10 | 17.20 | 18.58 | 8.59 |
| 20 | 17.20 | 10.15 | 4.30 |
| 30 | 17.06 | 6.79 | 2.87 |
| 50 | 14.09 | 4.07 | 1.72 |
| 100 | 7.85 | 2.04 | 0.86 |

This raises the question: how many subintervals are optimal, which is equivalent to the asking how many subhypercubes are optimal? To address this issue, a Monte Carlo study was used to determine the optimal number of subdivisions to take. The study began with an exploration of the effect of increasing the number of dimensions ($d$) and the number of Riemann-style subintervals ($r$) taken along each dimension. The number of dimensions is equal to the number of factors used. Uniform random variates were generated with sampling properties that were similar to the actual superpopulation data. The number of variates was equal to the size of the superpopulation and the min. and max of each variate matched the min. and max of the corresponding factor used in the superpopulation analysis. Thus, the simulated data represents what the superpopulation would look like if it filled the occupied space uniformly. The Monte Carlo results in Table 13 show occupied volumes as a function of the number of subintervals ($r$) used to grid each axis and dimensionality ($d$, or the number of factors).

Table 13 shows that the occupied volume of the uniform data is independent of the number of gridding intervals r, but only to a point. This "breakdown point" occurs at $r = 30$ for $d = 3$, $r = 10$ for $d = 5$, and $r = 5$ for $d = 8$. The conclusion is that even for data which uniformly "span" a descriptor space, there is a limit to how fine a grid can be placed for volume estimation, and finer gridding results in a loss of volume accuracy. Furthermore, the breakdown point occurs for smaller values of r as d increases, so higher dimensional spaces must be gridded less finely. The entry for $d = 8$ and $r = 100$ illustrates dramatically that extreme gridding causes loss of accuracy.

Comparing Table 12 with Table 13 shows the huge effect that outliers have. The observed data (Table 5) had several thousand huge outliers which caused the inflated volume estimates of 30.77, 30.18, and 48.10 in row 1 of Table 12. The volume breaks down immediately ($r = 2$), in contrast to the Table 6 results. This is the effect shown in Figure 5, where the dotted line represents the observed data and the solid line represents the simulated uniform data.

It should also be noted that in Table 12 the analysis was done with all outliers included. The fact that the $d = 8$ volume tracks the $d = 5$ volume for high values of $r$ indicates that outlier removal is needed in addition to subdividing and using the Riemann approach.

Riemann style volume estimation should be like Riemann integration, in which the accuracy is improved by taking smaller and smaller subintervals of the real number line, and in the limit the number of subintervals approaches infinity and the width of the subintervals approaches zero. Thus the

question arises, why is there a point where finer gridding causes a loss of accuracy? The problem is that in this application the volume is being estimated empirically, using a sample dataset which in effect gives a discrete gridding of the real number line in multidimensional space. The real numbers are dense; if one chooses any two real numbers, no matter how close they are to each other a gap exists between them. With only five factors and ten Riemann partitions there are already 100 000 subhypercubes ($10^5$, eq 1) which is the number of observations in the random uniform dataset. Thus, the breakdown point in Table 13 for the $d = 5$ case ($r = 10$) happened at the point where there was, on average, one subhypercube for each observation. Beyond this, either increasing the number of subintervals or increasing the dimensionality, there are actually more subhypercubes than observations! The issue here is the cost of volume accuracy: to obtain more accuracy, one must obtain a larger sample of observations.

A rough rule-of-thumb was that there certainly should be more observations than subhypercubes, and a rule of approximately two observations for every subhypercube was chosen. This rule leads to the following formula for the number of subintervals

$$r = (n/2)^{1/d} \qquad (2)$$

where $r$ is the number of Riemann subintervals the algorithm will take along each dimension, $n$ is the number of observations in the dataset, and $d$ is the dimension or the number of factors used in the analysis. The equation provides a useful upper limit to the number of Riemann subintervals to be used without losing accuracy in the volume measurements.

Now applying this to the Monte Carlo study in Table 13 one can see that this formula is consistent with the observed data. Given $d = 3$ (3 factors) and a dataset of 110 000 compounds, the formula prescribes that $r = 38$ is the optimal number of subintervals to use. Table 13 shows that the volume did remain constant up to and including $r = 30$, consistent with expectation. If one wanted to use more subintervals ($r > 38$), too many subhypercubes would result. Table 13 shows that this occurs for $r = 50$ and $r = 100$ in that the occupied volume is falling off as the number of subdivisions increase. In other words, using too many subintervals (or equivalently, subhypercubes) results in volume estimates that are downwardly biased.

Figure 5 illustrates the breakdown in volume accuracy for $d = 2$ through $d = 5$. The solid line shows the volume of the simulated uniform data, and the dashed line shows the volume of the actual superpopulation data. The plots are log−log scale so that the $x$ axis is the log of the number of Riemann subintervals and the $y$ axis is the log of the volume. In each case, the point at which the solid line begins to slope downward is very close to the point which the "rule-of-thumb" formula prescribes as the optimal number of subintervals (after exponentiating back to original scale). In looking at Figure 5, one should understand that the fact that the volumes of the observed data are decreasing at a rapid rate is not a bad thing since the data are both highly clustered and contain severe outliers. The problem is deciding when is enough. The simulated uniform data provide an objective stopping point, seen in the figure as the vertical line. This is the point at which any further gridding is unwise even for an ideal dataset with uniform spanning and no outliers.

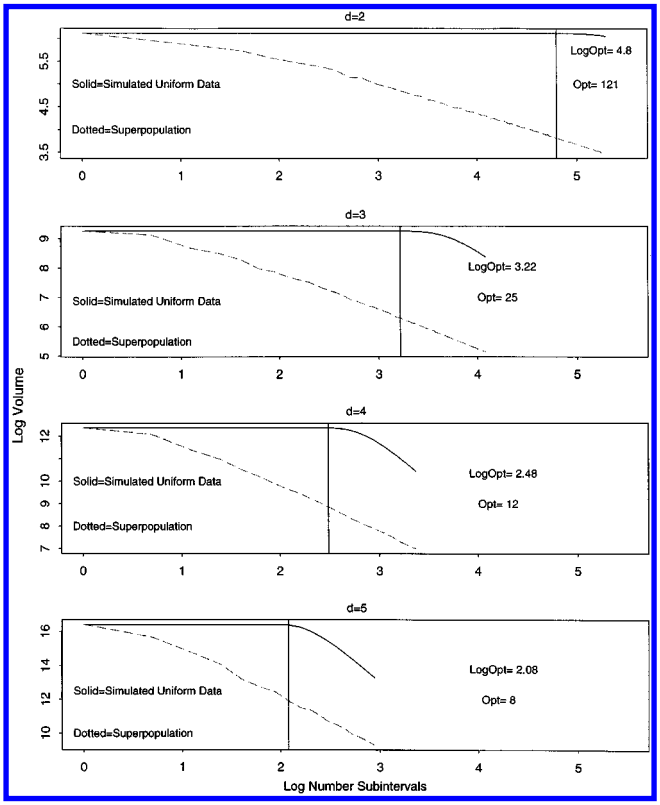


**Figure 5.** Volume estimate as a function of number of Riemann subintervals.

**Table 14.** Optimal Number of Subdivisions ($r$) as a Function of Number of Factors and Number of Observations Using Eq 2

| no. of factors | $n = 110\,000$ | $n = 317\,852$ | var. explained (%) |
|---|---|---|---|
| 2 | 235 | 399 | 80 |
| 3 | 38 | 54 | 87 |
| 4 | 15 | 20 | 90 |
| 5 | 9 | 11 | 93 |
| 6 | 6 | 7 | 94 |
| 7 | 5 | 6 | 96 |
| 8 | 4 | 4 | 97 |
| 9 | 3 | 4 | 97 |
| 10 | 3 | 3 | 98 |

Equation 2 can now be used to predict the maximum number of subdivisions to take for the superpopulation. Table 14 shows the optimal $r$ prescribed by eq 2 for different numbers of factors and for two datasets of different sizes, where n equals the number of observations in each dataset. The second dataset, $n = 317\,852$ is the same size as the superpopulation of five databases. Thus column 3 of Table 14 provides information relevant to the choice of parameters for use in the gridding analysis of the superpopulation. It is desirable to maximize the degree of grid partitioning since this enhances the resolution of the volume analysis (sampling accuracy). It is also desirable to use as many factors as possible (descriptor accuracy) so as to maximize the variance explained. Equation 2 prescribes a trade-off between the two criteria. Hence, a compromise was necessary, and four factors and 20 subdivisions were chosen to construct the grid for the superpopulation.

One can also use eq 2 to predict the dataset size necessary to obtain an accurate volume estimate given some fixed value of $r$. Suppose one wants to test the formula for $r = 100$ and $d = 3$. Working backwards, the formula states that in order to take 100 subintervals, the sample size needed is $100^3 * 2 = 2\,000\,000$ observations! Or, looking at Table 6, the

breakdown in volume accuracy that occurs at $d = 5$ and $r = 10$ could be avoided by using $10^5 * 2 = 200\,000$ observations (instead of the 110 000 that were used in the simulation).

**Supporting Information Available:** A plot using an alternate factor, primarily GSOLV, is available (1 page). This material is contained in many libraries on microfiche, immediately follows this article in the microfilm version of the journal, can be ordered from the ACS, and can be downloaded from the Internet; see any current masthead page for ordering information and Internet access instructions.

## REFERENCES AND NOTES

(1) Comprehensive Medicinal Chemistry (CMC); Molecular Design Limited: San Leandro, CA 94577. An electronic database version of the Drug Compendium that is Volume 6 of *Comprehensive Medicinal Chemistry* published by Pergamon Press in March 1990. Contains drugs already on the market.
(2) MACCS-II Drug Data Report (MDDR), version 94.1; Molecular Design Limited: San Leandro, CA 94577. An electronic database version of the Prous Science Publishers journal *Drug Data Report*, extracted from issues starting mid-1988. Contains biologically active compounds in the early stages of drug development.
(3) Available Chemicals Directory (ACD), version 94.1; Molecular Design Limited: San Leandro, CA 94577. Contains specialty and bulk chemicals from commercial sources.
(4) SPECS/BioSPECS Database, version 94.5; Brandon Associates: Merrimack, NH 03054. Contains chemicals from private sources.
(5) Andrews, C. W. Manuscript in preparation.
(6) MedChem Version 3.54; Daylight Chemical Information Systems, Inc.: Claremont, CA 91711.
(7) Hall, Lowell, H.; Kier, Lemont, B. Molconn-X, Version 2.0, A Program for Molecular Topology; Hall Associates Consulting: 2 Davis Street, Quincy, MA 02170.
(8) Sutter, J. M.; Kalivas, J. H.; Lang, P. M. Which Principal Components to Utilize for Principal Components Regression. *J. Chemometrics* **1992**, *6*, 217−225.
(9) SAS version 6.09; SAS Institute, Cary, NC 27512.
(10) Afifi, A. A.; Clark, V. Computer Aided Multivariate Analysis; Wadsworth: copyright 1984.
(11) The approximation of the volume of occupied hyperspace by partitioning the space into subintervals is analogous to Riemann integration. See: Thomas, G. B. *Calculus and Analytic Geometry*, 4th ed.; Addison-Wesley Publishing Company: Reading, MA, 1968.
(12) Silverman, B. W. Density Estimation; Chapman & Hall: copyright 1986.
(13) Hoaglin, D. C.; Mosteller, F.; Tukey, J. Understanding Robust and Exploratory Data Analysis; Wiley: copyright 1983.

CI950168H