# Statistical Analysis of Atom Topological Neighborhoods and Multivariate Representations of a Large Chemical File

Roger Attias*,† and Michel Petitjean

Institut de Topologie et de Dynamique des Systèmes, Université Paris 7—CNRS URA 34,
1 Rue Guy de la Brosse, 75005 Paris, France

Received February 25, 1993®

Statistical results on a large file are presented for a set of topological parameters. The distribution of skeleton atom neighborhoods is analyzed; it suggests a partition on atoms and contributes to the study of generic aspects of concentric substructures. The bivariate distribution (layer depth/number of atoms) extends the study to any depth. The multivariate distributions, (number of *R*-extremal atoms)/(number of *D*-extremal atoms)/(number of centers), and (number of atoms)/(number of bonds)/(number of cycles), are then presented; they show clusters of values and also empty ranges of values.

## INTRODUCTION

The diversity of applications in chemistry has resulted in numerous topological indices and substructure systems.[1]

Conversely, statistical studies contribute to the analysis of chemical files and to the design of specific tools in the field.

A statistical study of topological parameters is therefore proposed on a large Chemical Abstracts Service (CAS) subfile comprising the structures registered from 1965 to July 1978 and from which coordination compounds, polymers, incompletely defined structures, and alloys had been removed (in order to be homogeneously processed by an initial application); the file comprises 3 424 428 compounds, resulting in 4 019 514 components (a component is a connected set of atoms, as conventionally defined by CAS; a compound may comprise more than one component, e.g. salts).

The algorithmic generation of structural moieties has led to the definition of various parameters, rooted on an atom, which describe some aspects of its neighborhood, at varying distance from the root.

The skeletons of the two-level topological neighborhoods are being considered; they have been exhaustively generated for this large file, and the results are analyzed.

The study is then extended to concentric layers of any depth.

Several univariate distributions have been computed on large files (e.g. atoms, bonds, paths, concentric layers,[2] or extensive ring studies[3,4]). Multivariate distributions are more complex to represent, but they allow detailed subclassifications, e.g. the radius–diameter distribution.[5] The following bivariate distributions are consequently presented: the distribution of the number of concentric layers in the number of atoms/layer depth plane; the distribution of the number of components in the space multiplicity of the *R*-extremal atom/multiplicity of the *D*-extremal atom/multiplicity of the center; and the distribution of the number of components in the atoms/bonds/rings space.

## ATOM TOPOLOGICAL NEIGHBORHOOD

A property which reflects the exact topological neighborhood is assigned to each atom of the file. It consists of the atom connectivity $C$ and the connectivity of each of its neighbors, $X_1, X_2, ..., X_c$. We represent the atom topological neighborhood by $C(X_1X_2...X_c)$, the $X_i$'s being in decreasing order. This notation is exemplified for the atom neighborhoods in Figure 1.

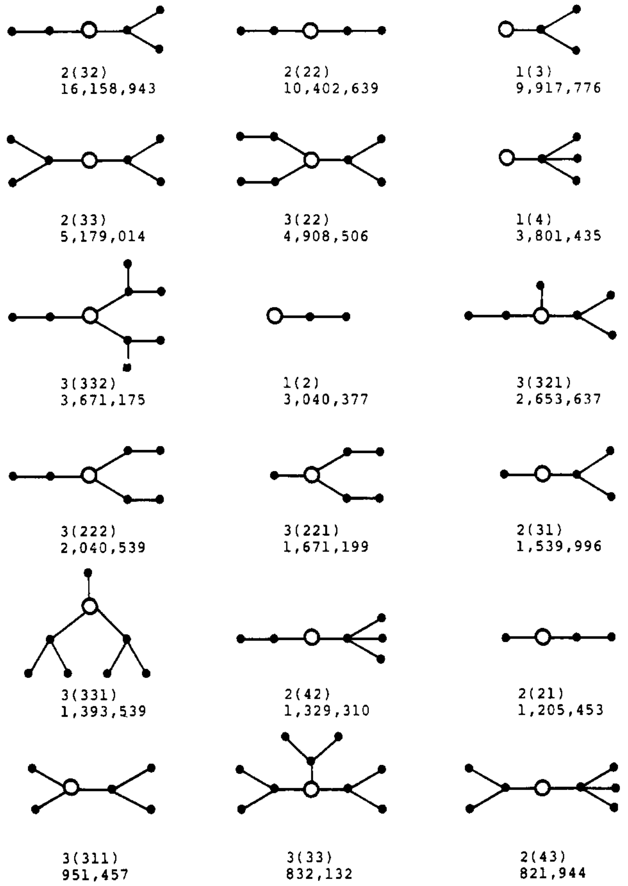


**Figure 1.** Highest occurring atom types. The first 15 atom types contribute for 88.9% of the total number of occurrences.

The notation is unique and unambiguous; it defines a lexicographic order among the atom types:

$$\text{if} \quad C_1(X_1X_2...X_{C_1}) < C_2(Y_1Y_2...Y_{C_2})$$

$$\text{then} \quad C_1 < C_2$$

$$\text{or} \quad C_1 = C_2 \text{ and } X_1X_2...X_{C_1} < Y_1Y_2...Y_{C_2}$$

The neighborhoods of the atoms with a nonzero connectivity degree less or equal to 4 (i.e., $0 < C < 5$) have been generated

[†] New address: Laboratoire de Chimie et de Biochimie, Pharmacologiques et Toxicologiques, Université René Descartes—CNRS URA 400, 45 Rue des Saints Pères, 75006 Paris, France

[®] Abstract published in *Advance ACS Abstracts*, August 15, 1993.

0095-2338/93/1633-0649$04.00/0

**Table I.** Distribution of the Atom Neighborhood Types[a]

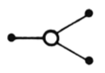
| atom type | atoms spanned | no. of atoms | no. of compounds | atom type | atoms spanned | no. of atoms | no. of compounds |
|---|---|---|---|---|---|---|---|
| Atom Degree = 1 | | | | | | | |
| 1(1) | 2 | 18 398 | 9 080 | 1(4) | 5 | 3 801 435 | 1 152 735 |
| 1(2) | 3 | 3 040 377 | 1 783 006 | 1(5) | 6 | 36 243 | 9 800 |
| 1(3) | 4 | 9 917 776 | 2 883 189 | | | | |
| Atom Degree = 2 | | | | | | | |
| 2(11) | 3 | 13 538 | 13 320 | 2(43) | 8 | 821 944 | 563 189 |
| 2(21) | 4 | 1 205 453 | 853 524 | 2(44) | 9 | 126 858 | 79 453 |
| 2(22) | 5 | 10 402 639 | 2 355 668 | 2(51) | 7 | 5 637 | 2 296 |
| 2(31) | 5 | 1 539 998 | 1 012 459 | 2(52) | 8 | 7 524 | 3 450 |
| 2(32) | 6 | 16 158 943 | 3 023 905 | 2(53) | 9 | 14 121 | 5 095 |
| 2(33) | 7 | 5 179 014 | 1 538 515 | 2(54) | 10 | 2 917 | 1 420 |
| 2(41) | 6 | 262 213 | 172 549 | 2(55) | 11 | 2 292 | 1 099 |
| Atom Degree = 3 | | | | | | | |
| 3(111) | 4 | 26 845 | 26 209 | 3(443) | 12 | 25 838 | 22 699 |
| 3(211) | 5 | 813 320 | 580 871 | 3(444) | 13 | 3 121 | 2 374 |
| 3(221) | 6 | 1 671 199 | 1 219 299 | 3(511) | 8 | 1 173 | 761 |
| 3(222) | 7 | 2 040 539 | 1 296 491 | 3(521) | 9 | 685 | 537 |
| 3(311) | 6 | 951 457 | 685 934 | 3(522) | 10 | 8 280 | 3 721 |
| 3(321) | 7 | 2 653 637 | 1 548 366 | 3(531) | 10 | 806 | 503 |
| 3(322) | 8 | 4 908 506 | 2 149 318 | 3(532) | 11 | 2 920 | 1 293 |
| 3(331) | 8 | 1 393 539 | 783 951 | 3(533) | 12 | 22 451 | 4 188 |
| 3(332) | 9 | 3 671 175 | 1 489 483 | 3(541) | 11 | 105 | 74 |
| 3(333) | 10 | 832 132 | 539 863 | 3(542) | 12 | 103 | 71 |
| 3(411) | 7 | 93 130 | 74 962 | 3(543) | 13 | 8 851 | 3 396 |
| 3(421) | 8 | 236 838 | 190 760 | 3(544) | 14 | 324 | 277 |
| 3(422) | 9 | 706 435 | 487 146 | 3(551) | 12 | 285 | 171 |
| 3(431) | 9 | 144 657 | 113 816 | 3(552) | 13 | 185 | 112 |
| 3(432) | 10 | 729 965 | 437 645 | 3(553) | 14 | 317 | 156 |
| 3(433) | 11 | 161 853 | 133 280 | 3(554) | 15 | 68 | 36 |
| 3(441) | 10 | 17 330 | 15 101 | 3(555) | 16 | 381 | 240 |
| 3(442) | 11 | 73 749 | 53 707 | | | | |
| Atom Degree = 4 | | | | | | | |
| 4(1111) | 5 | 46 595 | 34 401 | 4(5111) | 9 | 656 | 426 |
| 4(2111) | 6 | 209 737 | 163 480 | 4(5211) | 10 | 293 | 189 |
| 4(2211) | 7 | 165 217 | 126 315 | 4(5221) | 11 | 75 | 58 |
| 4(2221) | 8 | 103 613 | 91 459 | 4(5222) | 12 | 590 | 342 |
| 4(2222) | 9 | 77 661 | 65 523 | 4(5311) | 11 | 298 | 171 |
| 4(3111) | 7 | 283 278 | 205 944 | 4(5321) | 12 | 35 | 28 |
| 4(3211) | 8 | 216 406 | 188 097 | 4(5322) | 13 | 105 | 71 |
| 4(3221) | 9 | 98 661 | 90 632 | 4(5331) | 13 | 492 | 338 |
| 4(3222) | 10 | 70 026 | 64 621 | 4(5332) | 14 | 1 406 | 1 058 |
| 4(3311) | 9 | 105 891 | 96 600 | 4(5333) | 15 | 2 137 | 1 648 |
| 4(3321) | 10 | 248 714 | 187 017 | 4(5411) | 12 | 172 | 115 |
| 4(3322) | 11 | 91 275 | 84 410 | 4(5421) | 13 | 13 | 8 |
| 4(3331) | 11 | 42 538 | 40 049 | 4(5422) | 14 | 23 | 12 |
| 4(3332) | 12 | 67 716 | 62 597 | 4(5431) | 14 | 204 | 155 |
| 4(3333) | 13 | 18 955 | 17 695 | 4(5432) | 15 | 443 | 307 |
| 4(4111) | 8 | 58 426 | 36 573 | 4(5433) | 16 | 870 | 615 |
| 4(4211) | 9 | 37 350 | 27 289 | 4(5441) | 15 | 329 | 100 |
| 4(4221) | 10 | 23 606 | 20 924 | 4(5442) | 16 | 163 | 118 |
| 4(4222) | 11 | 17 548 | 14 918 | 4(5443) | 17 | 184 | 124 |
| 4(4311) | 10 | 43 720 | 33 007 | 4(5444) | 18 | 105 | 31 |
| 4(4321) | 11 | 117 926 | 81 011 | 4(5511) | 13 | 78 | 65 |
| 4(4322) | 12 | 35 734 | 30 208 | 4(5521) | 14 | 0 | 0 |
| 4(4331) | 12 | 30 891 | 24 126 | 4(5522) | 15 | 37 | 26 |
| 4(4332) | 13 | 25 409 | 22 134 | 4(5531) | 15 | 2 | 2 |
| 4(4333) | 14 | 6 063 | 4 627 | 4(5532) | 16 | 56 | 37 |
| 4(4411) | 11 | 65 397 | 20 738 | 4(5533) | 17 | 32 | 23 |
| 4(4421) | 12 | 9 467 | 7 524 | 4(5541) | 16 | 24 | 21 |
| 4(4422) | 13 | 6 388 | 5 412 | 4(5542) | 17 | 165 | 104 |
| 4(4431) | 13 | 7 268 | 5 423 | 4(5543) | 18 | 22 | 12 |
| 4(4432) | 14 | 5 231 | 4 557 | 4(5544) | 19 | 435 | 180 |
| 4(4433) | 15 | 2 133 | 1 456 | 4(5551) | 17 | 17 | 10 |
| 4(4441) | 14 | 5 245 | 2 215 | 4(5552) | 18 | 29 | 19 |
| 4(4442) | 15 | 1 324 | 956 | 4(5553) | 19 | 44 | 30 |
| 4(4443) | 16 | 737 | 546 | 4(5554) | 20 | 374 | 167 |
| 4(4444) | 17 | 549 | 299 | 4(5555) | 21 | 627 | 395 |

[a] The 125 atom types are lexicographically ordered. The occurrence, the incidence (number of compounds), and the number of atoms of the neighborhood $\sum(x_i + 1)$ are reported.

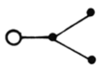
(i.e., 77 446 029 atoms among the 77 915 142 total atoms of the file).

The neighborhood of each atom defines a fragment which is a two-level tree. The root and its first level successors exhibit their *exact* connectivity degree; the peripheral atoms (second level) have, by construction, an undetermined connectivity.

The topological parameters are totally defined by the generation rules, e.g., the explicit and implicit parameters associated with fragments I and II describe clearly two very different contexts.



I



II

The distribution of the atom degrees being uneven, the connectivity degrees of the neighbors ($X_i$) which are higher than 5 have been set to 5; this value represents the range of unfrequent high connectivities (i.e., the 49 783 atoms of connectivity 5 have been aggregated with the 16 303 atoms with connectivity greater than 5). This improves the readability of the results.

Let $C$ be the maximum connectivity degree of the focus, and let $p$ be the maximum connectivity degree of its neighbors: the number $n$ of possible distinct types of atom neighborhoods is $n = (C + p)!/C!p!$ (see Appendix).

In our study, $C = 4$ and $p = 5$; it yields $n = 126$ (125 when excluding the atom with connectivity zero). This choice of $C$ and $p$ is a compromise between exhaustivity of existing topological motifs, and simplicity of use and interpretation. Thus, for $C = 5$ and $p = 5$, $n = 252$; i.e., $n$ should be multiplied by 2 in order to slightly increase the specificity.

**Atom Classification.** In this study the basic entity is the *atom*, considered as an element belonging to one class (its connectivity) and to one of its disjoint subclasses (the set of connectivity degrees of all its connected atoms), as defined by its topological context.

A simple procedure assigns each atom to one of the 125 neighborhood values. The occurrence of each atom type (i.e., the number of atoms sharing a given neighborhood) has been computed, yielding a partition of the atoms. The results (Table I) are in the lexicographical order of the notation of the atom type. The number of compounds which comprise at least one instance of the atom type (i.e., its incidence) is also reported.

Like most parameters which have been studied in the field, the distribution is strongly uneven.

The first ranking atom type is found in 20.8% (2(32)) of the total number of atoms in the file. Fourty seven percent of the atoms belong to one of the three most occurring atom types (2(32), 2(22), 1(3)) and the cumulative frequencies of the 10 and 15 first atom types is respectively 79.7 and 88.9%.

These highest occurring atom neighborhoods (Figure 1) are the few primitives which contribute to the most frequent chemical moieties. They may be part of cyclic moieties as well as part of acyclic moieties; pending chains (e.g., 1(2) and 2(21)) are identified.

The distribution of atoms of connectivity 1 suggests a classification of terminal atoms according to the connectivity of their neighbors: among the 16 814 229 atoms of connectivity 1 (21.6% of the total number of atoms), 59% are linked to an atom of connectivity degree 3, and 22.6, 18.1, 0.21, and 0.11% are linked respectively to an atom of connectivity 4, 2, 5, and 1.

**Multiple Occurrence.** The most occurring atom types have also the highest incidence; thus 2(32), 1(3), 2(22), and 3(322) occur respectively in 88.3, 84.19, 68.79, and 62.76% of the compounds of the file. A slight difference in ranking is however observed; it results from a difference in the densities of the atom types (average multiplicity of the atom type within the compounds). The atom types with the highest and lowest densities are listed in Table II.

Some frequent chemical moieties comprise multiple occurrences of the same atom type (which generates high related density values), e.g.:

(a) 2(32) occurs 4 times in a para substitued six membered ring and 2 times in a mono, ortho, or meta substituted six membered ring.

**Table II.** Atom Types with Highest and Lowest Densities (Average Multiplicity per Compound)

| highest densities | | lowest densities | |
|---|---|---|---|
| atom type | density | atom type | density |
| 1(3) | 3.44 | 2(11) | 1.02 |
| 1(4) | 3.30 | 3(111) | 1.02 |
| 1(5) | 3.70 | 4(3221) | 1.09 |
| 2(22) | 4.42 | 4(3222) | 1.08 |
| 2(32) | 5.34 | 4(3322) | 1.08 |
| 2(33) | 3.37 | 4(3331) | 1.06 |
| 3(533) | 5.36 | 4(3332) | 1.08 |
| 4(4411) | 3.15 | 4(3333) | 1.07 |
| 4(5441) | 3.29 | 4(4332) | 1.15 |
| 4(5444) | 3.39 | 4(4432) | 1.15 |

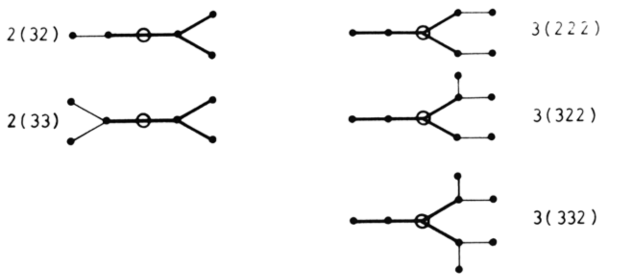


**Figure 2.** Topological overlapping. The atom neighborhoods 2(32) and 2(33) share each a five-atom motif (shown in bold lines) with 3(332) or with 3(322) or with 3(222); this overlap defines potential larger moieties.

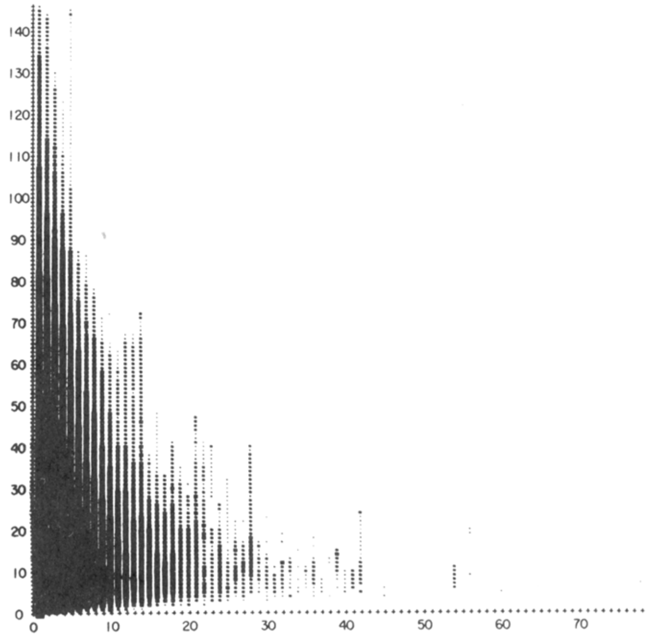


**Figure 3.** Distribution of the number of layers having a given number of atoms (abscissa) at a given depth (ordinate). The size of a (printed) point ($x$,$y$) is logarithmically related to the corresponding number of layers. The data are listed in Table III for depths 1–6 (e.g., there are 10 layers having 33 atoms at depth 4).

(b) 2(22) is a linear sequence which can be found in both cyclic and acyclic moieties. It occurs 3 times in a mono substituted six membered ring and 2 times in an ortho substituted ring. Every additional atom to a cyclic or acyclic sequence of this type generates a new occurrence of this Atom Neighborhood.

(c) Terminal atoms which are linked to the same atom generate each a new occurrence of the same atom neighborhood (1(3), 1(4), or 1(5)); e.g., 1(4) occurs at least twice for each occurrence of 4($XY$11), $X$ and $Y$ being any value.

The unexpected high density of some low-ranking atom types with fairly highly substituted neighbors may denote the

**Table III.** Distribution of the Number of Layers Having a Given Number of Atoms at a Given Depth, from Depth = 1 to Depth = 6

| atoms in a layer | depth = 1 layers | depth = 2 layers | depth = 3 layers | depth = 4 layers | depth = 5 layers | depth = 6 layers |
|---|---|---|---|---|---|---|
| 1 | 16 814 229 | 5 355 681 | 8 854 408 | 13 697 655 | 14 869 406 | 15 688 537 |
| 2 | 37 072 401 | 25 038 530 | 22 743 926 | 21 023 480 | 21 991 122 | 21 447 113 |
| 3 | 21 202 208 | 25 468 813 | 18 455 869 | 18 263 341 | 16 457 896 | 14 215 889 |
| 4 | 2 357 223 | 13 398 296 | 74 799 520 | 11 616 805 | 10 410 608 | 8 592 347 |
| 5 | 49 830 | 5 692 057 | 7 531 344 | 6 152 170 | 4 946 745 | 3 992 399 |
| 6 | 11 803 | 1 955 274 | 3 362 318 | 3 033 523 | 2 459 964 | 2 045 480 |
| 7 | 156 | 379 656 | 1 232 687 | 1 256 613 | 983 250 | 812 042 |
| 8 | 633 | 84 927 | 377 542 | 468 260 | 407 865 | 371 018 |
| 9 | 219 | 18 993 | 115 357 | 172 140 | 163 930 | 167 748 |
| 10 | 3 279 | 6 579 | 37 108 | 61 740 | 71 091 | 75 870 |
| 11 | 43 | 2 081 | 10 662 | 22 655 | 29 338 | 31 578 |
| 12 | 87 | 1 127 | 7 319 | 15 347 | 20 726 | 27 253 |
| 13 | 3 | 327 | 2 485 | 6 145 | 8 089 | 10 527 |
| 14 | 1 | 192 | 1566 | 4 380 | 6 013 | 8 530 |
| 15 | 0 | 100 | 849 | 2 853 | 3 621 | 5 009 |
| 16 | 0 | 53 | 682 | 1 586 | 2 345 | 2 896 |
| 17 | 0 | 9 | 255 | 800 | 1 328 | 1 837 |
| 18 | 0 | 21 | 309 | 703 | 1 544 | 1 994 |
| 19 | 0 | 3 | 22 | 192 | 527 | 746 |
| 20 | 0 | 18 | 49 | 225 | 527 | 830 |
| 21 | 0 | 0 | 17 | 152 | 305 | 334 |
| 22 | 0 | 9 | 1 | 57 | 112 | 133 |
| 23 | 0 | 0 | 10 | 19 | 55 | 40 |
| 24 | 0 | 0 | 13 | 98 | 149 | 196 |
| 25 | 0 | 0 | 6 | 32 | 26 | 111 |
| 26 | 0 | 0 | 0 | 1 | 64 | 72 |
| 27 | 0 | 0 | 1 | 11 | 25 | 54 |
| 28 | 0 | 0 | 0 | 0 | 35 | 23 |
| 29 | 0 | 0 | 0 | 0 | 3 | 4 |
| 30 | 0 | 0 | 4 | 9 | 5 | 73 |
| 31 | 0 | 0 | 0 | 2 | 32 | 32 |
| 32 | 0 | 0 | 0 | 2 | 2 | 23 |
| 33 | 0 | 0 | 0 | 10 | 0 | 3 |
| 34 | 0 | 0 | 0 | 0 | 2 | 1 |
| 35 | 0 | 0 | 0 | 0 | 0 | 2 |
| 36 | 0 | 0 | 0 | 1 | 6 | 4 |
| 37 | 0 | 0 | 0 | 0 | 2 | 5 |
| 38 | 0 | 0 | 0 | 1 | 0 | 0 |
| 39 | 0 | 0 | 0 | 0 | 0 | 1 |
| 40 | 0 | 0 | 0 | 0 | 4 | 4 |
| 41 | 0 | 0 | 0 | 0 | 0 | 30 |
| 42 | 0 | 0 | 0 | 0 | 10 | 0 |
| 45 | 0 | 0 | 0 | 2 | 0 | 6 |
| 54 | 0 | 0 | 0 | 0 | 0 | 10 |
| 60 | 0 | 0 | 0 | 0 | 2 | 0 |

presence of symmetrical complex moieties in the molecule.

The topological moieties 2(11) and 3(111), which span completely a component or a compound, have a low density.

**Sampling.** For some practical reasons, the files have been subdivided into five subfiles (in the Registry Number sequences), which have been processed separately before being merged. The five partial results thus available from this partition present slight variations; they are reported only in order to describe the effects of the sample size even within a homogeneous large file.

The distribution of most of the atom neighborhoods is similar in the five subfiles. However a few exceptions have been observed, mainly among the low-ranking skeletons: 2(51) and 2(53) have an occurrence in the third subfile which is 3 times the value of the expected occurrence. The successive values of the occurrence of 3(543), i.e., 3912, 2671, 2179, and 52, present a significant decrease. These variations may express the presence/absence of certain types of compounds, which correspond also, in a CAS file, to indexing periods.

The high-ranking atom neighborhoods are frequent in a wide diversity of compounds, but their frequency may slightly vary with the sample.

**Embedment. Overlapping.** The occurrences of the 125 rooted skeletons are highly interdependent; each atom is present in several moieties. An atom A, considered as the root of the neighborhood $C(X_1X_2...X_c)$ is present: in 1 fragment, at the root position; in the $C$ fragments which are rooted on its $C$ neighbors and within which A is at the first level position; and in the $\sum(X_i - 1)$ fragments which are rooted on the atoms lying at a distance 2 from $A$ ($A$ is at the second level of these fragments); i.e. in $1 + C + \sum(X_i - 1) = \sum X_i + 1$ fragments, which is also the number of atoms spanned by the atom neighborhood. This equality is an obvious consequence of the property of symmetry of the topological distance between the root and the leaves; it holds for atom neighborhoods of any size.

The following quantitative analysis provides a schematic illustration of the multiple contribution of an atom: if the 16 158 943 occurrences of 2(32) were independent (i.e., if the six atoms of this neighborhood contributed each to only one of these occurrences), and without taking into account any contribution of other atom neighborhoods, the number of distinct atoms spanned by the individual occurrences of this skeleton would be 16 158 943 × 6 = 96 953 658 atoms, which is 25% higher than the total number of distinct atoms in the file. It can be therefore globally deduced that the overlapping between two different occurrences of 2(32) (which may topologically comprise up to four atoms), is statistically frequent.

Overlapping between atom neighborhoods is examplified with some frequent fragments on Figure 2; potentially larger moieties can be deduced.

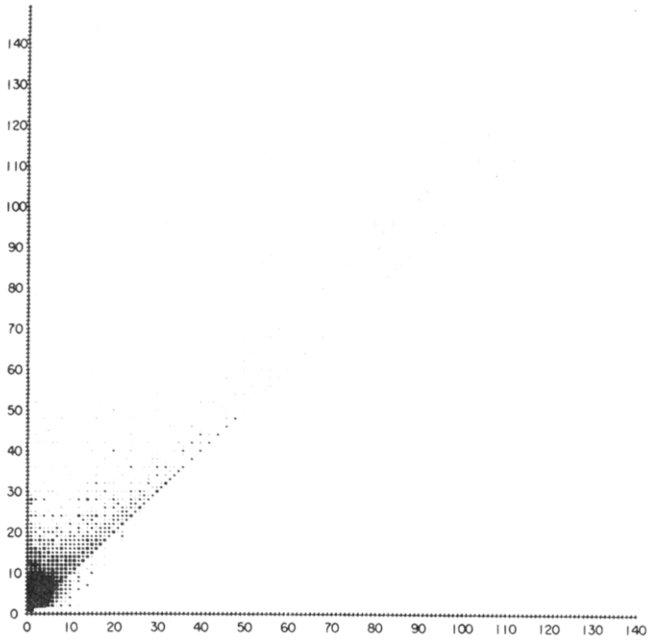


**Figure 4.** Distribution of the number of components according to the number of centers (abscissa) and to the number of $R$-extremal atoms (ordinate). The size of a printed point $(x,y)$ is logarithmically related to the corresponding number of components. The data are partially listed in Table V (e.g. there are 10 374 components having six centers and six $R$-extremal atoms).

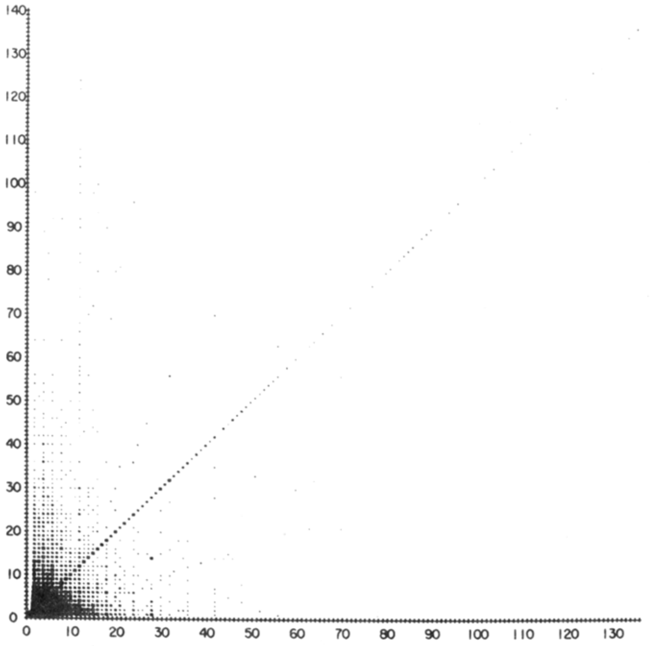


**Figure 5.** Distribution of the number of components according to the number of $D$-extremal atoms (abscissa) and to the number of centers (ordinate). The data are partially listed in Table VI.

Embedment can be exactly analyzed only for atom neighborhoods of connectivity 1: $1(X)$ is totally embedded $n + 1$ times within each of the neighborhoods of type $X(Y_1Y_2...Y_{x-1}1)$, $n$ being the number of $Y_i$'s equal to 1. Thus, for $X = 2$, the occurrence of 1(2) is the sum of the occurrences of 2(11) (twice), 2(21), 2(31), 2(41), and 2(51).

Further overlapping analysis requires additional statistical investigation.

**Measure of the Redundancy.** The number of atoms which is spanned by an atom neighborhood $C(X_1X_2...X_c)$ is $\sum X_i + 1$.

The representation of the file by means of the atom neighborhoods describes a total number of redundant atoms $T = \sum N_iA_i$ ($i$, varying from 1 to 125, is the atom neighborhood type, $N_i$ is the occurrence of $i$, and $A_i$ is the number of atoms spanned by $i$).

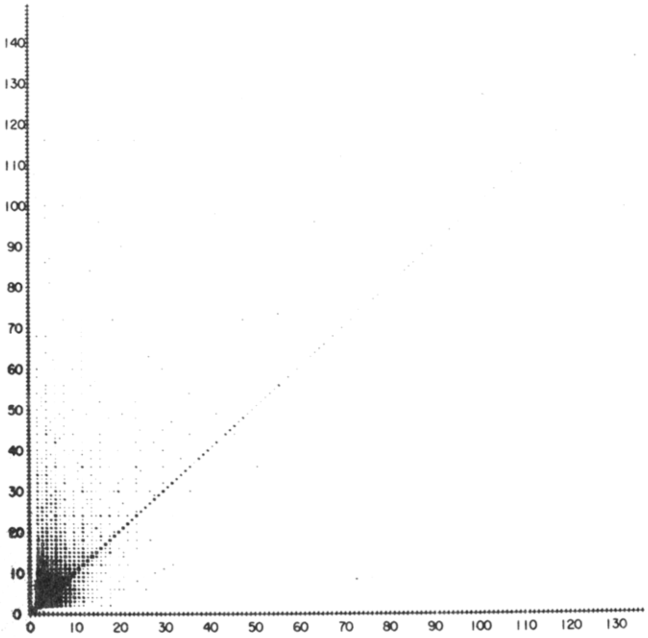


**Figure 6.** Distribution of the number of components according to the number of $D$-extremal atoms (abscissa) and to the number of $R$-extremal atoms (ordinate). The data are partially listed in Table VII.

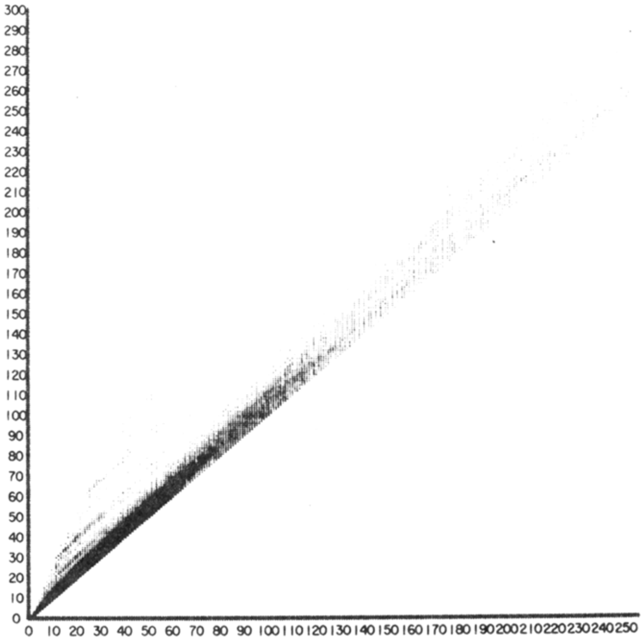


**Figure 7.** Distribution of the number of components according to the number of atoms (abscissa) and to the number of bonds (ordinate). The components lying on a parallel line to the $x = y$ diagonal have the same number of cycles.

The computed value of $T$ is 471 446 043. We call the redundancy factor of the atom neighborhood the ratio $r = T/\sum N_i$ ($\sum N_i$ being also the total number of distinct atoms in the file). For this file, $r = T/77446029 = 6.09$. It represents also the average number of atom neighborhoods to which an atom participates, or the average number of atoms within an atom neighborhood. A redundancy factor of this type is used in EURECAS.[6] The overlapping power varies with the atom connectivity and with the number of atoms which are spanned by the neighborhood. For each subclass of atoms of connectivity $c = 1, 2, 3$, or 4, we have computed the ratios (number of redundant atoms spanned by atom neighborhoods with

**Table IV.** Distribution of the Summation of the Products, Number of Layers at Depth = *d* Times Number of Atoms per Layer at Depth = *d*, and Distribution of the Mean Number of Atoms per Layer at Depth = *d*

| | atoms | | | atoms | | | atoms | |
|---|---|---|---|---|---|---|---|---|
| depth | sum | mean | depth | atoms | mean | depth | sum | mean |
| 0 | 77 915 142 | 1.000 | 50 | 904 376 | 2.566 | 100 | 4 428 | 1.665 |
| 1 | 164 357 002 | 2.120 | 51 | 822 726 | 2.490 | 101 | 3 692 | 1.555 |
| 2 | 229 244 876 | 2.962 | 52 | 764 290 | 2.460 | 102 | 3 426 | 1.560 |
| 3 | 238 692 512 | 3.090 | 53 | 726 886 | 2.502 | 103 | 2 866 | 1.433 |
| 4 | 221 359 272 | 2.920 | 54 | 666 188 | 2.467 | 104 | 2 758 | 1.501 |
| 5 | 202 636 784 | 2.782 | 55 | 611 690 | 2.420 | 105 | 2 358 | 1.404 |
| 6 | 179 934 712 | 2.666 | 56 | 583 028 | 2.474 | 106 | 2 184 | 1.421 |
| 7 | 152 706 078 | 2.542 | 57 | 525 988 | 2.401 | 107 | 1 840 | 1.296 |
| 8 | 127 099 050 | 2.461 | 58 | 482 148 | 2.364 | 108 | 1 866 | 1.413 |
| 9 | 102 337 698 | 2.379 | 59 | 452 832 | 2.402 | 109 | 1 548 | 1.309 |
| 10 | 82 488 910 | 2.356 | 60 | 406 912 | 2.334 | 110 | 1 560 | 1.414 |
| 11 | 66 382 512 | 2.355 | 61 | 371 236 | 2.298 | 111 | 1 332 | 1.307 |
| 12 | 53 557 820 | 2.361 | 62 | 353 212 | 2.355 | 112 | 1 304 | 1.344 |
| 13 | 43 501 662 | 2.373 | 63 | 317 674 | 2.284 | 113 | 1 146 | 1.257 |
| 14 | 36 087 700 | 2.416 | 64 | 292 086 | 2.260 | 114 | 1 138 | 1.317 |
| 15 | 29 671 486 | 2.420 | 65 | 276 096 | 2.307 | 115 | 1 008 | 1.252 |
| 16 | 25 029 242 | 2.453 | 66 | 247 612 | 2.243 | 116 | 972 | 1.291 |
| 17 | 21 479 812 | 2.505 | 67 | 221 624 | 2.168 | 117 | 872 | 1.247 |
| 18 | 18 251 854 | 2.513 | 68 | 208 694 | 2.218 | 118 | 840 | 1.292 |
| 19 | 15 719 910 | 2.528 | 69 | 187 798 | 2.174 | 119 | 726 | 1.222 |
| 20 | 13 859 090 | 2.581 | 70 | 176 378 | 2.196 | 120 | 694 | 1.264 |
| 21 | 12 031 596 | 2.574 | 71 | 161 236 | 2.184 | 121 | 614 | 1.226 |
| 22 | 10 584 384 | 2.583 | 72 | 144 684 | 2.133 | 122 | 578 | 1.265 |
| 23 | 9 451 346 | 2.621 | 73 | 128 382 | 2.076 | 123 | 516 | 1.268 |
| 24 | 8 251 666 | 2.586 | 74 | 119 352 | 2.134 | 124 | 456 | 1.243 |
| 25 | 7 330 208 | 2.585 | 75 | 104 168 | 2.056 | 125 | 392 | 1.202 |
| 26 | 6 692 854 | 2.648 | 76 | 95 674 | 2.052 | 126 | 348 | 1.200 |
| 27 | 5 926 080 | 2.618 | 77 | 87 854 | 2.057 | 127 | 306 | 1.195 |
| 28 | 5 320 062 | 2.615 | 78 | 78 998 | 2.029 | 128 | 258 | 1.152 |
| 29 | 4 847 706 | 2.645 | 79 | 71 760 | 2.010 | 129 | 230 | 1.139 |
| 30 | 4 314 768 | 2.609 | 80 | 68 252 | 2.079 | 130 | 204 | 1.133 |
| 31 | 3 890 326 | 2.591 | 81 | 60 444 | 2.008 | 131 | 180 | 1.125 |
| 32 | 3 634 336 | 2.662 | 82 | 55 406 | 2.021 | 132 | 166 | 1.137 |
| 33 | 3 274 204 | 2.627 | 83 | 50 674 | 2.059 | 133 | 152 | 1.152 |
| 34 | 3 002 064 | 2.627 | 84 | 45 004 | 2.021 | 134 | 134 | 1.155 |
| 35 | 2 787 416 | 2.662 | 85 | 39 322 | 1.985 | 135 | 120 | 1.176 |
| 36 | 2 498 736 | 2.604 | 86 | 35 996 | 2.044 | 136 | 110 | 1.222 |
| 37 | 2 278 066 | 2.583 | 87 | 29 642 | 1.911 | 137 | 94 | 1.205 |
| 38 | 2 168 938 | 2.668 | 88 | 25 780 | 1.920 | 138 | 84 | 1.273 |
| 39 | 1 956 388 | 2.609 | 89 | 20 948 | 1.822 | 139 | 74 | 1.321 |
| 40 | 1 786 716 | 2.571 | 90 | 17 510 | 1.790 | 140 | 80 | 1.667 |
| 41 | 1 675 338 | 2.604 | 91 | 14 002 | 1.685 | 141 | 60 | 1.429 |
| 42 | 1 521 776 | 2.551 | 92 | 12 386 | 1.771 | 142 | 64 | 1.882 |
| 43 | 1 405 354 | 2.524 | 93 | 10 258 | 1.669 | 143 | 56 | 2.000 |
| 44 | 1 362 388 | 2.617 | 94 | 9 138 | 1.687 | 144 | 64 | 2.909 |
| 45 | 1 243 010 | 2.552 | 95 | 7 750 | 1.645 | 145 | 24 | 1.500 |
| 46 | 1 160 560 | 2.536 | 96 | 7 022 | 1.678 | 146 | 4 | 1.000 |
| 47 | 1 099 932 | 2.567 | 97 | 5 848 | 1.581 | 147 | | |
| 48 | 1 003 772 | 2.506 | 98 | 5 532 | 1.675 | 148 | | |
| 49 | 932 612 | 2.478 | 99 | 4 542 | 1.563 | 149 | | |

connectivity *c*)/(number of atoms with connectivity *c*), which are respectively:

$$r_1 = 68053664/16814229 = 4.047$$

$$r_2 = 216659211/37072401 = 5.844$$

$$r_3 = 165716356/21202169 = 7.816$$

$$r_4 = 21016812/2357230 = 8.916$$

These values are also the average number of atoms within an atom neighborhood with a given connectivity.

**Table V.** Distribution of the 993 992 Components Having the Same Number of Centers and *R*-Extremal Atoms

| multiplicity | components | multiplicity | components | multiplicity | components |
|---|---|---|---|---|---|
| 1 | 403 027 | 31 | 13 | 66 | 1 |
| 2 | 416 239 | 32 | 102 | 68 | 3 |
| 3 | 97 309 | 33 | 19 | 70 | 2 |
| 4 | 50 107 | 34 | 34 | 72 | 9 |
| 5 | 9 740 | 35 | 13 | 77 | 1 |
| 6 | 10 374 | 36 | 72 | 78 | 1 |
| 7 | 1 038 | 37 | 5 | 80 | 3 |
| 8 | 1 516 | 38 | 20 | 81 | 7 |
| 9 | 637 | 39 | 7 | 83 | 1 |
| 10 | 581 | 40 | 29 | 84 | 2 |
| 11 | 263 | 41 | 2 | 85 | 1 |
| 12 | 364 | 42 | 19 | 86 | 1 |
| 13 | 189 | 44 | 27 | 88 | 4 |
| 14 | 245 | 45 | 6 | 89 | 1 |
| 15 | 151 | 46 | 11 | 90 | 1 |
| 16 | 226 | 47 | 1 | 94 | 1 |
| 17 | 133 | 48 | 25 | 96 | 2 |
| 18 | 256 | 49 | 8 | 102 | 1 |
| 19 | 71 | 50 | 6 | 104 | 1 |
| 20 | 210 | 51 | 1 | 108 | 1 |
| 21 | 72 | 52 | 5 | 110 | 1 |
| 22 | 129 | 53 | 2 | 112 | 1 |
| 23 | 46 | 54 | 6 | 118 | 1 |
| 24 | 171 | 55 | 1 | 120 | 1 |
| 25 | 32 | 56 | 7 | 126 | 1 |
| 26 | 101 | 58 | 3 | 134 | 1 |
| 27 | 34 | 60 | 7 | 136 | 1 |
| 28 | 81 | 63 | 2 | | |
| 29 | 39 | 64 | 5 | | |
| 30 | 100 | 65 | 2 | | |

**Table VI.** Distribution of the 1 053 418 Components Having the Same Number of Centers and *D*-Extremal Atoms

| multiplicity | components | multiplicity | components | multiplicity | components |
|---|---|---|---|---|---|
| 1 | 403 027 | 31 | 12 | 66 | 1 |
| 2 | 451 777 | 32 | 101 | 68 | 3 |
| 3 | 119 770 | 33 | 19 | 70 | 2 |
| 4 | 54 650 | 34 | 33 | 72 | 9 |
| 5 | 7 204 | 35 | 13 | 77 | 1 |
| 6 | 10 194 | 36 | 72 | 78 | 1 |
| 7 | 936 | 37 | 5 | 80 | 3 |
| 8 | 1 468 | 38 | 20 | 81 | 7 |
| 9 | 617 | 39 | 7 | 83 | 1 |
| 10 | 425 | 40 | 28 | 84 | 2 |
| 11 | 228 | 41 | 2 | 85 | 1 |
| 12 | 404 | 42 | 19 | 86 | 1 |
| 13 | 163 | 44 | 27 | 88 | 4 |
| 14 | 232 | 45 | 6 | 89 | 1 |
| 15 | 133 | 46 | 11 | 90 | 1 |
| 16 | 226 | 47 | 1 | 94 | 1 |
| 17 | 131 | 48 | 25 | 96 | 2 |
| 18 | 259 | 49 | 8 | 102 | 1 |
| 19 | 69 | 50 | 6 | 104 | 1 |
| 20 | 209 | 51 | 1 | 108 | 1 |
| 21 | 69 | 52 | 5 | 110 | 1 |
| 22 | 124 | 53 | 2 | 112 | 1 |
| 23 | 42 | 54 | 6 | 118 | 1 |
| 24 | 180 | 55 | 1 | 120 | 1 |
| 25 | 30 | 56 | 6 | 126 | 1 |
| 26 | 97 | 58 | 3 | 134 | 1 |
| 27 | 33 | 60 | 7 | 136 | 1 |
| 28 | 79 | 63 | 2 | | |
| 29 | 35 | 64 | 5 | | |
| 30 | 101 | 65 | 2 | | |

The use of controlled redundancy has proven to be efficient in application fields, e.g., substructure search systems. In the initial fragment search systems, a compound is selected according to the presence/absence of nonoverlapping query fragments, with no syntactical check. False drops are mainly due to compounds within which the fragments are interrelated differently than in the query. In subsequent systems the overlap which is allowed between fragments defines an implicit additional constraint and acts statistically as a syntactical process. The approach can be enhanced by defining multiple local generic points of view.[6]

## MULTIVARIATE DISTRIBUTIONS

**Distribution of the Number of Concentric Layers in the Number of the Atoms/Layer Depth Plane.** Each of the

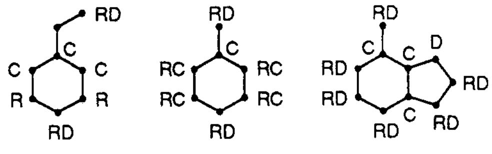
**Table VII.** Distribution of the 3 519 453 Components Having the Same Number of *R*-Extremal Atoms and *D*-Extremal Atoms

| multiplicity | components | multiplicity | components | multiplicity | components |
|---|---|---|---|---|---|
| 1 | 403 027 | 31 | 12 | 66 | 1 |
| 2 | 1 020 765 | 32 | 104 | 68 | 3 |
| 3 | 899 782 | 33 | 24 | 70 | 3 |
| 4 | 706 199 | 34 | 34 | 72 | 9 |
| 5 | 229 986 | 35 | 16 | 77 | 1 |
| 6 | 145 205 | 36 | 93 | 78 | 3 |
| 7 | 57 658 | 37 | 5 | 80 | 3 |
| 8 | 27 515 | 38 | 20 | 81 | 8 |
| 9 | 11 510 | 39 | 11 | 83 | 1 |
| 10 | 4 808 | 40 | 28 | 84 | 2 |
| 11 | 1 094 | 41 | 2 | 85 | 1 |
| 12 | 5 128 | 42 | 40 | 86 | 1 |
| 13 | 509 | 44 | 28 | 88 | 4 |
| 14 | 1 525 | 45 | 11 | 89 | 1 |
| 15 | 630 | 46 | 11 | 90 | 1 |
| 16 | 515 | 47 | 1 | 94 | 1 |
| 17 | 161 | 48 | 27 | 96 | 2 |
| 18 | 889 | 49 | 8 | 102 | 1 |
| 19 | 103 | 50 | 6 | 104 | 1 |
| 20 | 301 | 51 | 1 | 108 | 1 |
| 21 | 256 | 52 | 8 | 110 | 1 |
| 22 | 139 | 53 | 2 | 112 | 1 |
| 23 | 43 | 54 | 6 | 118 | 1 |
| 24 | 283 | 55 | 1 | 120 | 1 |
| 25 | 35 | 56 | 10 | 126 | 1 |
| 26 | 97 | 58 | 3 | 134 | 1 |
| 27 | 68 | 60 | 7 | 136 | 1 |
| 28 | 485 | 63 | 3 | | |
| 29 | 35 | 64 | 5 | | |
| 30 | 118 | 65 | 2 | | |

77 915 142 atoms of the file is in turn considered as a focus. Its neighbors are in the one-depth concentric layer, and its next neighbors are in the two-depth concentric layer, etc. The distribution of the number of concentric layers is displayed on Figure 3 (e.g., depths range from 5 to 40 for 28 atoms in a layer (abscissa)); the values are listed in Table III for the six first layers. A total of 898 037 816 concentric layers were found in the file, resulting in only 1941 different couples: depth layer/number of atoms in a layer, e.g., the cumulative relative frequency of only the 40 most occurring couples is 95.4%. This is small as compared to the potential combinatoric situations. The layers with *A* atoms have a distribution of their depths *d* varying irregularly with *A* (i.e. the standard deviation of *d* does not vary monotonically with *A*).

The mean number of atoms per layer is given in Table IV. This mean number of atoms is about three atoms for layers 2 to 4, and decreases slowly to about one atom for the deepest layers.

**Distribution of the Number of Components in the *R*-Extremal Atom Multiplicity/*D*-Extremal Atom Multiplicity/Center Multiplicity Space.** The center of a component is an atom from which the deepest concentric layer has a minimal depth. The atoms lying on this minimal depth layer are called *R*-extremal (this depth is the radius of the component). The atoms from which the deepest concentric layer has a maximal depth are called *D*-extremal (this depth is the diameter of the component). Examples are shown as follows:



R: R-extremal atom, D: D-extremal atom, C: center

The three structures comprise respectively 3,5,3 centers, 4,6,6 *R*-extremal atoms, and 2,2,7 *D*-extremal atoms.

**Table VIII.** Distribution of the Number of Components Having a Given Number of *R*-Extremal Atoms

| atoms | components | atoms | components | atoms | components |
|---|---|---|---|---|---|
| 1 | 403 027 | 51 | 7 | 101 | 0 |
| 2 | 1 028 158 | 52 | 18 | 102 | 3 |
| 3 | 985 676 | 53 | 4 | 103 | 0 |
| 4 | 823 075 | 54 | 22 | 104 | 2 |
| 5 | 333 347 | 55 | 3 | 105 | 0 |
| 6 | 241 812 | 56 | 19 | 106 | 0 |
| 7 | 95 636 | 57 | 1 | 107 | 1 |
| 8 | 48 963 | 58 | 5 | 108 | 2 |
| 9 | 20 459 | 59 | 0 | 109 | 0 |
| 10 | 11 445 | 60 | 14 | 110 | 2 |
| 11 | 3 598 | 61 | 0 | 111 | 0 |
| 12 | 8 553 | 62 | 2 | 112 | 4 |
| 13 | 1 853 | 63 | 6 | 113 | 0 |
| 14 | 3 220 | 64 | 7 | 114 | 0 |
| 15 | 1 467 | 65 | 2 | 115 | 0 |
| 16 | 1 523 | 66 | 3 | 116 | 3 |
| 17 | 563 | 67 | 0 | 117 | 0 |
| 18 | 1 663 | 68 | 8 | 118 | 2 |
| 19 | 327 | 69 | 2 | 119 | 0 |
| 20 | 672 | 70 | 4 | 120 | 2 |
| 21 | 438 | 71 | 0 | 121 | 0 |
| 22 | 357 | 72 | 15 | 122 | 0 |
| 23 | 131 | 73 | 0 | 123 | 0 |
| 24 | 879 | 74 | 0 | 124 | 0 |
| 25 | 107 | 75 | 0 | 125 | 0 |
| 26 | 261 | 76 | 1 | 126 | 1 |
| 27 | 140 | 77 | 1 | 127 | 0 |
| 28 | 644 | 78 | 4 | 128 | 0 |
| 29 | 83 | 79 | 0 | 129 | 0 |
| 30 | 249 | 80 | 4 | 130 | 1 |
| 31 | 40 | 81 | 13 | 131 | 0 |
| 32 | 194 | 82 | 0 | 132 | 0 |
| 33 | 44 | 83 | 1 | 133 | 0 |
| 34 | 85 | 84 | 3 | 134 | 2 |
| 35 | 35 | 85 | 1 | 135 | 0 |
| 36 | 172 | 86 | 3 | 136 | 2 |
| 37 | 12 | 87 | 1 | 137 | 0 |
| 38 | 33 | 88 | 6 | 138 | 0 |
| 39 | 20 | 89 | 1 | 139 | 0 |
| 40 | 69 | 90 | 4 | 140 | 0 |
| 41 | 4 | 91 | 0 | 141 | 0 |
| 42 | 77 | 92 | 0 | 142 | 0 |
| 43 | 2 | 93 | 0 | 143 | 0 |
| 44 | 56 | 94 | 2 | 144 | 0 |
| 45 | 16 | 95 | 0 | 145 | 0 |
| 46 | 22 | 96 | 6 | 146 | 0 |
| 47 | 5 | 97 | 1 | 147 | 0 |
| 48 | 50 | 98 | 1 | 148 | 0 |
| 49 | 13 | 99 | 0 | 149 | 1 |
| 50 | 12 | 100 | 4 | 150 | 0 |

The 4 019 514 components have generated the following: 790 different couples, number of centers/number of *R*-extremal atoms (Figure 4); 710 different couples, number of centers/number of *D*-extremal atoms (Figure 5); and 721 different couples, number of *R*-extremal atoms/number of *D*-extremal atoms (Figure 6).

The strong correlation between the three univariate distributions, according to each of these three parameters, is pointed out by the diagonals shown in the figures; i.e., 993 992 components have their number of centers equal to their number of *R*-extremal atoms (Table V), 1 053 418 components have their number of centers equal to their number of *D*-extremal atoms (Table VI), and 3 519 453 components have their number of *R*-extremal atoms equal to their number of *D*-extremal atoms (Table VII).

The univariate distribution of the components having a given number of *R*-extremal atoms shows that even values are slightly preferred (Table VIII). This is also true for *D*-extremal atoms (see ref 2).

**Distribution of the Number of Components in the Atoms/Bonds/Cycles Space.** The distribution of the components in

**Table IX.** Distribution of the Number of Components According to the Number of Independent Cycles for Various Number of Atoms

| | components | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| cycles | $A =$ 160 | $A =$ 170 | $A =$ 180 | $A =$ 190 | $A =$ 200 | $A =$ 210 | $A =$ 220 | $A =$ 230 | $A =$ 240 | $A =$ 250 |
| 0 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 1 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 5 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| 3 | 4 | 1 | 3 | 1 | 1 | 3 | 3 | 1 | 0 | 1 |
| 4 | 1 | 3 | 1 | 8 | 1 | 2 | 2 | 0 | 0 | 2 |
| 5 | 1 | 2 | 0 | 2 | 0 | 1 | 0 | 2 | 0 | 2 |
| 6 | 2 | 5 | 4 | 2 | 6 | 0 | 1 | 1 | 0 | 0 |
| 7 | 7 | 0 | 4 | 1 | 0 | 3 | 2 | 1 | 1 | 1 |
| 8 | 0 | 4 | 2 | 2 | 4 | 3 | 1 | 1 | 8 | 1 |
| 9 | 5 | 0 | 3 | 2 | 2 | 6 | 3 | 1 | 1 | 3 |
| 10 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 4 |
| 11 | 1 | 6 | 1 | 0 | 0 | 1 | 2 | 1 | 0 | 0 |
| 12 | 3 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 13 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 2 |
| 15 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 16 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 17 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 3 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 0 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 22 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 23 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 24 | 0 | 5 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 0 |
| 25 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 27 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 1 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 |
| 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 |
| 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 |

the atoms/bonds/cycles space is equivalent to a bivariate distribution, since the three parameters are related by $A - B + C = 1$, where $A$ is the number of atoms of the component, $B$ the number of bonds, and $C$ the number of independant cycles. The atoms/bonds distribution is shown on Figure 7; the others can be obtained by an appropriate projection. The 4 019 514 components generate only 4052 different couples: number of atoms/number of bonds. Only 239 of these couples represent more than 1000 components, and 2064 of them represent 1–4 components.

For a number $A$ of atoms less than about 60, the distribution of the number of components within each value of $A$ exhibits local modes corresponding to various cyclization ranges (Figure 7). These local modes are related to the local maxima shown by the univariate distribution of the components according to the number of cycles[2] and which occur for groups of values which are located around the multiples of 19 cycles. The highest cyclized components are typically boranes.

For $A$ greater than about 60, the distribution becomes unimodal, and it becomes bimodal approximately when $A$ is greater than 150 atoms: the components with a number of cycles belonging to the ranges around 15–20 cycles are lacking. Some examples of these bimodal distributions are given in Table IX.

The bidimensional clusters suggest therefore subclassifications which provide a new perspective on the 4 019 514 components.

## CONCLUSION

The statistical study of the file has yielded topological classifications on atoms on the one hand and on compounds on the other hand.

The two-level atom neighborhoods are the generic description of concentric substructures; their limited number allows a global analysis of their features.

Chemical graphs show specific distributions of graph theoretical parameters. In bivariate studies classes are subdivided, and clusters of values and empty classes are thus identified. For instance, the number of rings, in compounds comprising more than about 150 atoms, belongs to two distinct ranges of values, some values being skipped; this cannot be observed with the aggregated values of the monovariate studies.

## ACKNOWLEDGMENT

We are very grateful to the reviewers for their most helpful comments.

## APPENDIX

The number $n_c$ of distinct atom neighborhoods having a focus with a given connectivity degree $c$ and a maximum connectivity degree $p$ for the first neighbors can be derived from their notation: it is equal to the number of sequences of $p$ nonincreasing integers belonging to the range $(1,p)$ or, by setting $p' = p - 1$, to the range $(0,p')$. It can be shown recursively and by using the property

$$B(m,m+0) + B(m,m+1) + ... + B(m,m+p) = B(m+1,m+p+1) \quad (1)$$

This number is

$$n_c = B(c,c+p') \quad (2)$$

$B$ being the binomial coefficient: $B(n,m) = m!/n!(m - n)!$. The total number $T$ of distinct atom neighborhoods, having a maximum value $C$ for the connectivity of the focus and a maximum value $p$ for the connectivity of the first neighbors, is then a sum over $c$: $T = \sum n_c$. By using (1) and (2), we obtain $T = B(p,p+c) = B(c,p+c)$.

## REFERENCES AND NOTES

(1) Attias, R.; Substructure Systems and Structural Retrieval Systems. In *Encyclopedia of Library and Information Science*; Kent, A., Ed.; Marcel Dekker Inc.: New-York, 1992; Vol. 50, No. 13, pp 308–363. Also published in: *Encyclopedia of Microcomputers*; Kent, A., Williams, J. G., Eds.; MDI: New-York 1993; Vol. 11, pp 219–270.

(2) Petitjean, M.; Dubois, J. E. Topological Statistics on a Large Structural File. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 332–343.

(3) Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180–187.

(4) Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. VI. Substance-Related Statistics. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76–82.

(5) Petitjean, M. Applications of the Radius–Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.

(6) Attias, R. EURECAS/DARC. The Substructure in Chemistry. Contribution to the Representation of Structural Information and Application to Very Large Data Base Retrieval, Ph.D. Thesis, Université Paris 7, France, 1992.