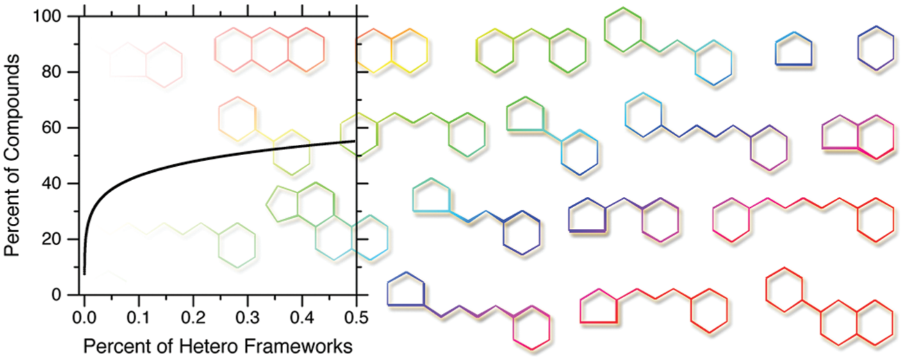
Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry

Alan H. Lipkus,* Qiong Yuan, Karen A. Lucas, Susan A. Funk, William F. Bartelt III, Roger J. Schenck, and Anthony J. Trippe

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

alipkus@cas.org

Received January 18, 2008



By analyzing the scaffold content of the CAS Registry, we attempt to characterize in a comprehensive way the structural diversity of organic chemistry. The scaffold of a molecule is taken to be its framework, defined as all its ring systems and all the linkers that connect them. Framework data from more than 24 million organic compounds is analyzed. The distribution of frameworks among compounds is found to be top-heavy, i.e., a small percentage of frameworks occur in a large percentage of compounds. When frameworks are analyzed at the graph level, an even more top-heavy distribution is found: half of the compounds can be described by only 143 framework shapes. The most significant finding is that the framework distribution conforms almost exactly to a power law. This suggests that the more often a framework has been used as the basis for a compound, the more likely it is to be used in another compound. This may be explained by the cost of synthesis: making a new derivative of a framework is probably less costly if many other derivatives are known. We believe this power law is evidence that the minimization of synthetic cost has been a key factor in shaping the known universe of organic chemistry.

Introduction

The analysis of chemical diversity has become a topic of considerable interest in recent years. This interest has been stimulated largely by the challenge of discovering new and novel small-molecule pharmaceuticals. The development of technologies such as combinatorial synthesis and high-throughput screening has made it possible to explore druglike regions of chemistry space in relatively short times. Chemistry space is vast, however, and the problem of selecting which regions of that space to explore remains a key issue in drug discovery.[1] In this context, the analysis of chemical diversity has emerged as a way to guide the exploration of chemistry space.

Assessing chemical diversity requires that each structure be characterized by one or more descriptors.[2] These can be molecular descriptors such as physicochemical properties or topological indexes. They can also be specific structural features. For example, the substructural fragments used to calculate similarity are often used for diversity assessment.[3,4] Larger substructures such as ring systems can also be used. A number of studies have analyzed diversity based on the ring systems in structures.[5–11] An advantage of using large features like rings

[1] Lipinski, C.; Hopkins, A. *Nature* **2004**, *432*, 855–861.

[2] Brown, R. D. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31–49.

[3] Willett, P. *Curr. Opin. Biotechnol.* **2000**, *11*, 85–88.

[4] Downs, G. M. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P. Eds.; Marcel Dekker: New York, 2004; pp 515−537.

[5] Nilakantan, R.; Bauman, N.; Haraki, K. S. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447–452.

10.1021/jo8001276 CCC: $40.75 © 2008 American Chemical Society
Published on Web 05/28/2008

is that structures having such features in common often belong to the same chemical family.

One type of large structural feature that is often associated with a specific chemical family is the molecular framework. This concept was proposed by Bemis and Murcko[12] as a way to help understand the common features in drug molecules. By their definition, the framework of a structure consists of all the ring systems and all the linkers, which are acyclic fragments that connect the ring systems. The framework is obtained by pruning all side-chain atoms, i.e., nonring atoms not on a direct path between two ring systems. By this definition, only cyclic structures have a framework. Typically, the framework describes only molecular topology, i.e., contains no three-dimensional or stereochemical information. Part of the reason this concept is useful in medicinal chemistry is that it describes the arrangement of rings in a structure, and rings are key building blocks in the design of drugs. The framework can be viewed as one possible definition of the molecular scaffold, a term which is widely used in medicinal chemistry but is not precisely defined.[13]

The framework content of a database can be taken as an indicator of its structural diversity. Bemis and Murcko used this idea to analyze the diversity of a commercial drug database.[12] When they considered only the shapes of frameworks, ignoring element and bond information, they found that half of the 5120 drugs were described by the 32 most frequently occurring framework shapes. They concluded that the shape diversity of the set of known drugs is very low. This kind of diversity analysis based on scaffolds has been used to study other chemical structure databases.[14–20]

The database of chemical substances used in the present study is the CAS Registry. This database began as an in-house tool to support the indexing of chemical substances found in the scientific literature.[21] It replaced the time-consuming effort of naming substances as a means of identification with a computer-based technique utilizing a unique and unambiguous representation of molecular structure.[22] Due to the explosive growth of the scientific literature, the Registry itself has grown to a collection encompassing more than 33 million organic and inorganic small-molecule compounds and is recognized as an authoritative resource for substance identification.[23]

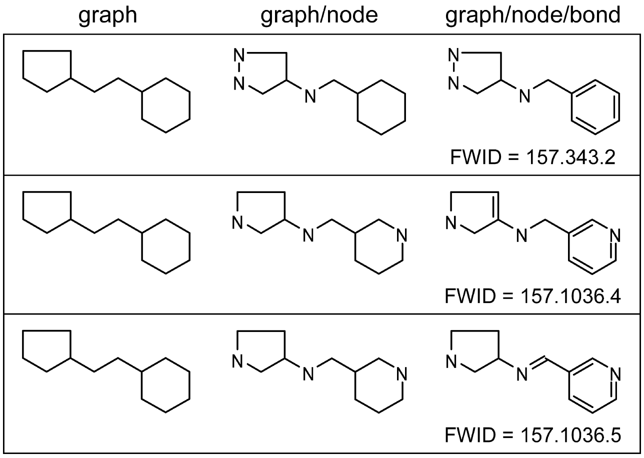


FIGURE 1. Framework Identifier (FWID) for three frameworks. The frameworks shown are all identical at the graph level and thus have the same graph id (157). Two of them are identical at the graph/node level and thus have the same node id (1036).

We recently extracted framework data from the CAS Registry. In this paper, we present the results of an analysis of that data. The analysis will focus on the frameworks associated with the organic subset of the Registry. Because of its size and coverage, the Registry is the best available representation of the "known universe" of chemistry. Hence, an analysis of this framework data offers a unique opportunity to take a comprehensive look at the structural diversity of organic chemistry.

Methods

The process used to extract frameworks from Registry substances involves two stages. The first is a simple iterative algorithm that finds the framework of each structure. The second is a procedure that looks for a match between each new framework and all previously found frameworks; this procedure builds a nonredundant file of frameworks each of which is assigned a unique identifier. The framework extraction process is not applied to every substance in the Registry. Frameworks are extracted only from Registry substances that have explicit representations at the atom level (such substances have a size limit of 253 non-hydrogen atoms). This excludes sequences from consideration. Acyclic substances are ignored because the framework definition is not applicable. Multicomponent substances and polymers are also ignored.

The framework of each structure is found using the following algorithm: (1) Flag all terminal atoms. (2) Flag every atom adjacent to a flagged atom unless it is adjacent to more than one unflagged atom. (3) Repeat (2) until no more atoms can be flagged. When finished, the unflagged atoms and the bonds between them constitute the framework. The framework does not retain any information indicating the attachment sites of the pruned side chains. It therefore has no information about whether a side chain was attached by a single bond or by a double bond; these two cases lead to different geometries at the attachment site, but such differences are ignored in the framework. Any information about stereochemistry, charges, uncommon isotopes, or the three-dimensional shape of the structure is not included in the framework.

When the framework of the current Registry substance is found, it is compared against a file of previously seen frameworks. If there is no match, it is added to that file and is assigned a new identifier. The format chosen for the Framework Identifier (FWID) is the same as that of the CAS Ring Identifier, which is used to uniquely identify ring systems in the Registry.[24] The FWID is a faceted number. It combines three identification numbers each of which signifies one aspect of framework structure. There is a graph id (which denotes

[6] Lee, M.-L.; Schneider, G. *J. Comb. Chem.* **2001**, *3*, 284–289.

[7] Lipkus, A. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 430–438.

[8] Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; McLay, I. M.; Bradshaw, J. *J. Med. Chem.* **2003**, *46*, 3257–3274.

[9] Kho, R.; Hodges, J. A.; Hansen, M. R.; Villar, H. O. *J. Med. Chem.* **2005**, *48*, 6671–6678.

[10] Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. *J. Chem. Inf. Model.* **2006**, *46*, 553–562.

[11] Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. *J. Med. Chem.* **2006**, *49*, 4568–4573.

[12] Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887–2893.

[13] Brown, N.; Jacoby, E. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

[14] Xue, L.; Bajorath, J. *J. Mol. Model.* **1999**, *5*, 97–102.

[15] Xu, J. *J. Med. Chem.* **2002**, *45*, 5311–5320.

[16] Liu, B.; Lu, A.; Zhang, L.; Liu, H.; Liu, Z.; Zhou, J. *Internet Electron. J. Mol. Des.* **2004**, *3*, 143–149.

[17] Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272–17277.

[18] Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. *Mol. Divers.* **2006**, *10*, 389–403.

[19] Krier, M.; Bret, G.; Rognan, D. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

[20] Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.

[21] Leiter, D. P., Jr.; Morgan, H. L.; Stobaugh, R. E. *J. Chem. Doc.* **1965**, *5*, 238–242.

[22] Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107–113.

[23] Weisgerber, D. W. *J. Am. Soc. Inf. Sci.* **1997**, *48*, 349–360.

TABLE 1. **Framework Statistics from the CAS Registry**

category	number
compounds[a]	24 282 284
frameworks, graph level	836 708
frameworks, graph/node level	2 594 176
frameworks, graph/node/bond level	3 380 334

[a] Single-component, cyclic organic compounds registered as of the end of June 2007.

the underlying connectivity), a node id (which denotes the pattern of elements), and a bond id (which denotes the pattern of bond types).

The FWID is designed to be hierarchical. It represents frameworks at three levels of structural information, as shown in Figure 1. The graph level has connectivity information but ignores element and bond types. The graph/node level has connectivity and element information but ignores bond types. The graph/node/bond level has connectivity, element, and bond type information. Because of the hierarchical nature of the FWID, it can be used to quickly organize frameworks for comparison and analysis at these different levels.

The procedure to determine whether the current framework matches a previously found framework also looks for partial matches; this is necessary to create hierarchical FWIDs. The procedure uses the Morgan algorithm[22] to put the connection table of the current framework into a canonical form. From this connection table are derived three hash codes based on the graph, graph/node, and graph/node/bond levels. The graph/node/bond hash code is used to identify possible exact matches among the previously found frameworks. If one of these frameworks does match exactly the current framework, its FWID is associated with the current substance.

If no exact match is found, the procedure searches for possible partial matches at the graph/node level or the graph level, using the appropriate hash codes. If a partial match is found, the FWID of the partially matching framework is used as the basis for creating a new FWID for the current framework. If no partial match is found, a completely new FWID is created. The connection table of the new framework is stored, and its FWID is associated with the current substance.

Results and Discussion

The framework extraction process described above was run against all substances entered into the Registry as of the end of June 2007. It found frameworks in 25 956 900 substances. These substances consist of both organic and inorganic compounds. As already noted, we want to focus our analysis on frameworks from organic compounds. For this reason, the only compounds included in our analysis were those that (1) contain carbon and (2) do not contain any element other than H, B, C, Si, N, P, As, O, S, Se, Te, and the halogens (unless that element is present only as an ion of a salt, in which case the ion is removed and the compound is included in our analysis). This filtering left a set of 24 282 284 compounds.

Table 1 shows the numbers of frameworks at different levels of structural information in this set of compounds. According to these numbers, there are 3.1 frameworks at the graph/node level for every one at the graph level. This suggests that a significant amount of structural diversity is introduced in going from the graph to the graph/node level. In contrast, there are only 1.3 frameworks at the graph/node/bond level for every one at the graph/node level, which means that the graph/node/bond level adds much less diversity. For this reason, and for simplicity, our diversity analysis will focus on the graph/node and graph levels.

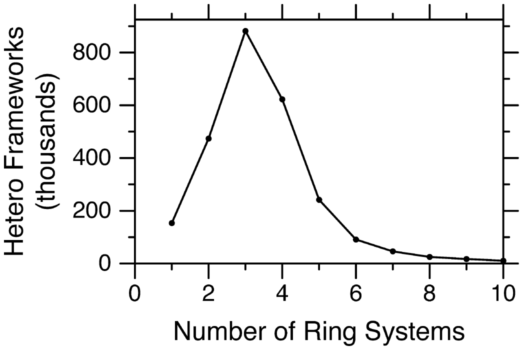


FIGURE 2. Histogram of the number of ring systems in hetero frameworks.

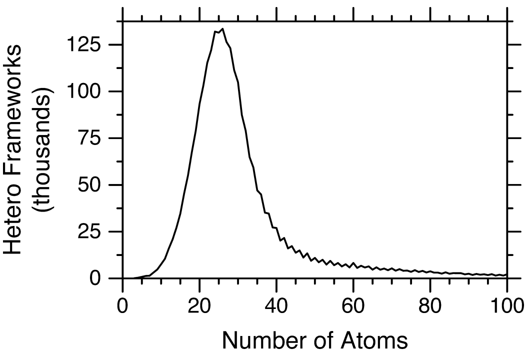


FIGURE 3. Histogram of the number of atoms in hetero frameworks.

A. Diversity Analysis of Hetero Frameworks. The pattern of heteroatoms in a molecule can play a crucial role in its chemical and biological properties. Frameworks at the graph/node level include element types and thus contain information about the pattern of heteroatoms. Since almost all of the frameworks at this level (98.6%) contain at least one heteroatom, these frameworks will be called *hetero frameworks*.

A.1. Size and Heteroatom Content. As would be expected, the frameworks found among organic compounds exhibit a wide range of sizes. There are a few frameworks that contain more than 40 ring systems and a few that have more than 250 atoms. However, such extreme values are exceptional. For the vast majority of frameworks, these quantities are relatively small and fall into a narrow range.

As shown in the histogram of Figure 2, hetero frameworks usually have only a few ring systems: 95.0% of the frameworks contain six or fewer ring systems. Only 1.2% contain more than 10. There are 153 021 hetero frameworks that consist of a single ring system. This number represents only 5.9% of the frameworks, but it indicates the very high degree of ring diversity in this compound set.

Figure 3 is a histogram of the atom count in hetero frameworks. This data shows that 50.0% of the frameworks have between 20 and 30 atoms, and 89.7% have 50 or fewer atoms. Only 1.8% have more than 100 atoms. A noticeable feature of this histogram is a small even/odd alternation: frameworks with an even number of atoms are slightly more common than those with an odd number. Previous studies of several databases have revealed a similar alternation effect for chemical structures.[25,26] This has not been fully explained, but it has been suggested

[24] Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.

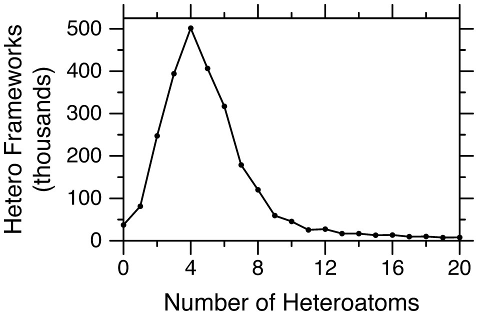


FIGURE 4. Histogram of the number of heteroatoms in hetero frameworks.

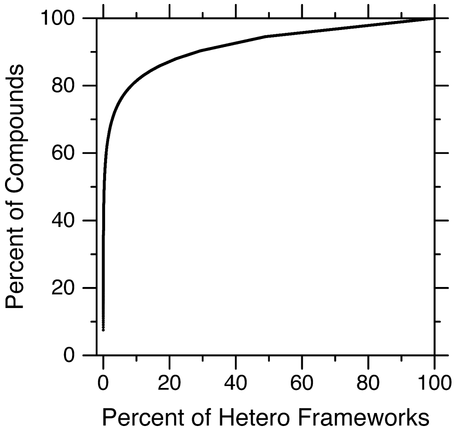


FIGURE 5. Percentage of compounds containing a particular percentage of hetero frameworks.

the effect is due to the synthetic methods (e.g., dimerization) used to build large molecules from smaller ones.[27]

A histogram of the number of heteroatoms is shown in Figure 4. This data shows that 50.2% of the hetero frameworks contain three, four, or five heteroatoms. Only 2.2% contain more than 20 heteroatoms. A small percentage (1.4%) of hetero frameworks are composed entirely of carbon atoms. Overall, 17.6% of the atoms in the hetero frameworks are heteroatoms. Those that occur most frequently are N (66.2% of heteroatoms), O (22.8%), S (8.7%), P (1.1%), B (0.5%), and Si (0.5%).

A.2. Distribution of Hetero Frameworks. A central question with regard to structural diversity is how frameworks are distributed among organic compounds. There is a simple way of graphing the data that can help illustrate this distribution. We first order the hetero frameworks by their frequency of occurrence among organic compounds (most to least common). We then plot percentage of (most common) frameworks on the *x* axis and the percentage of compounds that contain those frameworks on the *y* axis. The result is shown in Figure 5.

For a perfectly even distribution (i.e., each framework occurring in the same number of compounds) the points would fall on the diagonal. The curve obtained is quite different: it rises very steeply and then levels off. This indicates that a very small percentage of frameworks are found in a large percentage of compounds. For instance, the most common 5.0% of the hetero frameworks are found in 75.5% of the compounds. This kind of distribution can be described as top-heavy. A more detailed view of the curve in Figure 5 is shown in Figure 6, where the *x* axis scale has been greatly expanded. This plot shows how very top-heavy the framework distribution actually is. For instance, 0.25% of the hetero frameworks are found in 49.6% of the compounds.

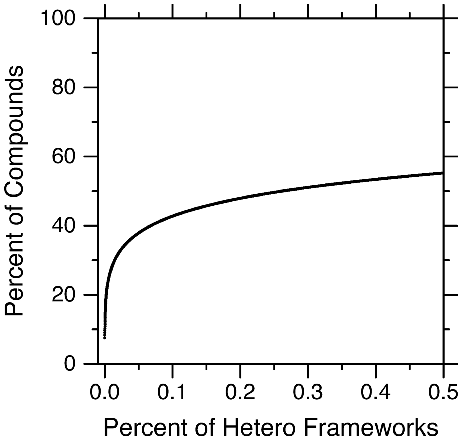


FIGURE 6. Percentage of compounds containing a particular percentage of hetero frameworks. This is the leftmost part of Figure 5 on a greatly expanded *x* axis.

The curve in Figure 5 turns sharply at around 10% on the *x* axis. This point separates the region in which 10% of hetero frameworks account for more than 80% of compounds and the region in which 90% of hetero frameworks account for less than 20% of compounds. In other words, this marks a transition from frameworks that are very common to frameworks that are relatively uncommon. The most uncommon hetero frameworks are those that occur in a single compound. There are 1 323 013 of these, which means that 51.0% of all hetero frameworks occur once. These unique frameworks account for only 5.4% of all compounds. (The point in Figure 5 where the unique frameworks begin is actually visible as a subtle discontinuity in slope at 49% on the *x* axis.)

The 30 hetero frameworks that occur most frequently are shown in Figure 7. These frameworks are smaller than most hetero frameworks. They also have fewer heteroatoms, but their heteroatom content [N (65.4%), O (23.1%), S (11.5%)] is similar to hetero frameworks as a whole. These frameworks are associated with large numbers of compounds. They are present in 17.2% of organic compounds. This suggests that a significant portion of organic chemistry is based on a very limited range of framework diversity. In fact, 12.7% of organic compounds are based on just the first 10 frameworks in Figure 7.

A.3. Power-Law Distribution. There is another way to examine how frameworks are distributed among organic compounds. We associate with each hetero framework a frequency value, i.e., the number of organic compounds it occurs in. The distribution of this value over the set of all hetero frameworks is shown in Figure 8. Frequency is plotted on the *x* axis, and the number of hetero frameworks with that frequency is plotted on the *y* axis. Each axis is on a logarithmic scale.

The left side of this plot shows that most hetero frameworks occur infrequently; the leftmost point corresponds to the 1.3 million of them that occur in only a single compound. As the frequency value increases, the number of frameworks that occur with that frequency drops. The rightmost point represents the hetero framework with the largest frequency, which occurs in 1 831 672 compounds. This plot illustrates, more directly than Figure 5, how hetero frameworks are distributed such that most of them occur with very low frequency but a few of them occur with extremely high frequency.

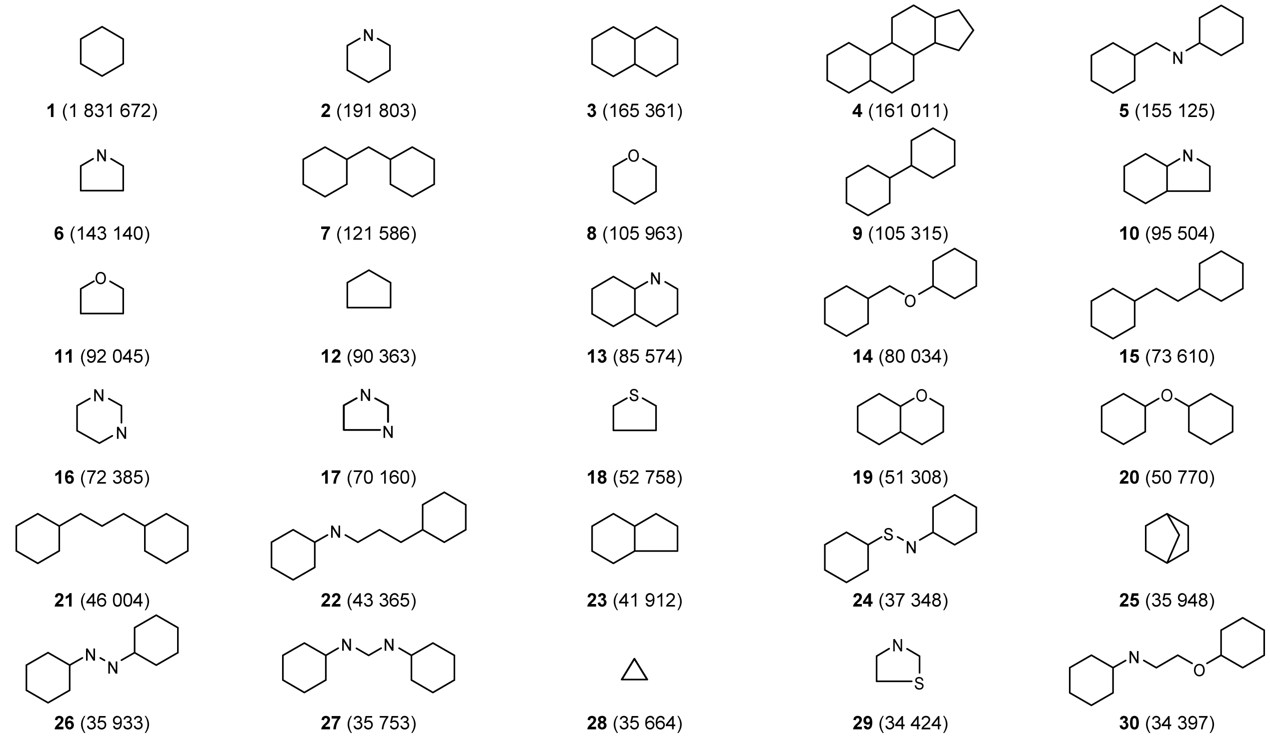


FIGURE 7. Most frequently occurring hetero frameworks. Numbers of compounds in which they occur are shown in parentheses.

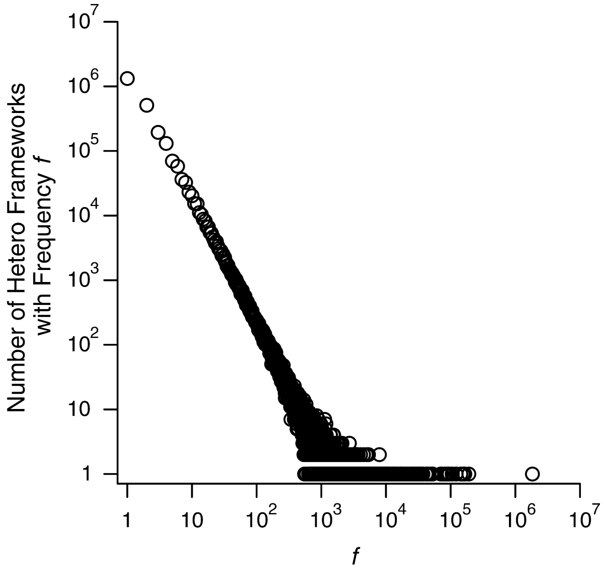


FIGURE 8. Distribution of the frequency with which hetero frameworks occur. Each axis is on a logarithmic scale.

The distribution in Figure 8 is linear, at least over the range of 1−10^2. When this kind of log−log plot is linear, it implies that the distribution obeys a power law. This has the form

$$p(x) = Cx^{-\alpha} \qquad (1)$$

where $p(x)$ is the probability that the value x occurs and C is a normalization constant. A distribution exactly obeying this law will appear as a straight line on a log−log plot since

$$\ln p(x) = -\alpha \ln x + \ln C \qquad (2)$$

The absolute slope of the line corresponds to the power-law exponent α. Note that we prefer to plot counts on the *y* axis, as in Figure 8, rather than probabilities, but this does not change the slope.

The distribution in Figure 8 becomes quite noisy in the region beyond 10^2. In this region, smaller and smaller numbers of frameworks are associated with each frequency, and fluctuations in this number result in increasing statistical noise. This kind of noisy tail is typical in power-law distributions. One way to correct for noise in a power-law distribution is to use a cumulative distribution function. Instead of plotting $p(x)$, we plot $P(x)$, defined as the probability that the value x or greater occurs. It can be shown from eq 1 that[28]

$$P(x) = C'x^{-(\alpha-1)} \qquad (3)$$

where $C' = C/(\alpha - 1)$. This distribution is also a power law and will thus appear as a straight line on a log−log plot.

In Figure 9, the frequency data for hetero frameworks has been plotted as a cumulative distribution. In this log−log plot, frequency is given on the *x* axis and the number of hetero frameworks with that frequency or higher is given on the *y* axis. The linearity is more pronounced than in Figure 8 and is seen to extend much further. This straight line implies that the distribution of frameworks over all of organic chemistry conforms almost exactly to a power law.

A large number of quantities from the physical and social sciences are believed to follow a power-law distribution. These quantities arise in a wide variety of scientific fields: physics, biology, linguistics, bibliometrics, sociology, economics, computer science. The best known example of a power-law

[25] Petitjean, M.; Dubois, J.-E. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 332–343.

[26] Sarma, J. A. R. P.; Nangia, A.; Desiraju, G. R.; Zass, E.; Dunitz, J. D. *Nature* **1996**, *384*, 320.

[27] Desiraju, G. R.; Dunitz, J. D.; Nangia, A.; Sarma, J. A. R. P.; Zass, E. *Helv. Chim. Acta* **2000**, *83*, 1–15.

[28] Newman, M. E. J. *Contemp. Phys.* **2005**, *46*, 323–351.

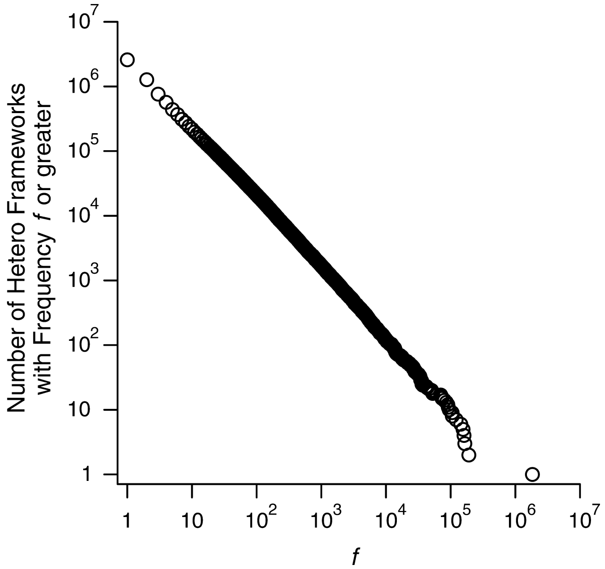


FIGURE 9. Frequency distribution in Figure 8 plotted as a cumulative distribution.

distribution is that of word frequencies, which came to be known as Zipf's law.[29] In organic chemistry, there have also been observations of power-law behavior. A study of several specialized data sets of 500−1200 compounds suggested that scaffolds follow a power law.[20] Another study found evidence of power-law behavior in a database of chemical reactions.[30]

The slope of the distribution in Figure 9 can be used to calculate the power-law exponent α. To estimate the slope, this distribution was fitted to a line using least-squares linear regression. Since this is a cumulative distribution, the fitted line must take on the value log(total number of hetero frameworks) at a frequency of 1. In order to incorporate this constraint, the distribution was translated to put this point at the origin, and the technique of regression through the origin[31] was used. An estimate of −1.07 was obtained for the slope. It is apparent from eq 3 that this slope should equal $-(\alpha - 1)$, and so the calculated value for the exponent is $\alpha = 2.07$. This value for α is comparatively low. A list of 12 power-law distributions found in the physical and social sciences[28] shows that only three of them have an exponent less than 2.07. Since the lower the value of α, the more top-heavy the distribution, it appears that the distribution of frameworks is unusually top-heavy.

A.4. Growth Model for Power-Law Behavior. There is a long history of theoretical work aimed at explaining the occurrence of power-law distributions. Because such distributions occur in a wide range of data, it has been assumed that the mechanisms which underlie this behavior must be of a general nature. Over the years, a number of theoretical models have been proposed. One of the most general models is a particular type of growth mechanism. In its simplest form, this model consists of a set of objects, each of which has some property associated with it (e.g., frequency of occurrence). It is assumed that new objects appear at some rate (and have some small value of the property when they appear). It is also assumed that the property associated with each object grows at a rate proportional to its current value. Over time, the distribution of this property among the objects tends to approach a power law.

This kind of growth is sometimes described as a "rich-get-richer" process. It was first proposed by Yule[32] to explain the power-law distribution of species among genera. Simon[33] later gave an improved mathematical analysis of this process. Among the examples he discusses is that of word frequencies, and this example is of particular interest with regard to molecular frameworks. Viewing a body of text as a stream of words, the next word in the stream is either new (not seen before) or old (already seen). Simon's model assumes that new words appear with constant probability. It also assumes that the more often a word has been seen, the greater the likelihood that it is the next word in the stream. The steady-state solution of this model yields the power-law distribution of word frequencies (Zipf's law).

In a similar fashion, the Registry can be viewed as a stream of compounds, ordered by time of registration. Every compound in this stream is new because the Registry is a nonredundant database. However, if each compound is replaced by its framework, what results is a stream of both new and old frameworks. This situation is analogous to the previous example of words in a text. As with word frequencies, the frequencies of hetero frameworks are distributed according to a power law. This suggests that the same assumption made by Simon concerning words applies also to frameworks, i.e., the more often one has been used, the more likely it is to be used again.

It seems plausible to expect that the more often a framework has been used as the basis for a compound, the more likely it is to be used in another compound. If many compounds derived from a framework have already been synthesized, these derivatives can serve as a pool of potential starting materials for further syntheses. The availability of published schemes for making these derivatives, or the existence of these derivatives as commercial chemicals, would then facilitate the construction of more compounds based on the same framework. Of course, not all frameworks are equally likely to become the focus of a high degree of synthetic activity. Some frameworks are intrinsically more interesting than others due to their functional importance (e.g., as a building block in drug design), and this interest will stimulate the synthesis of derivatives. Once this synthetic activity is initiated, it may be amplified over time by a rich-get-richer process.

B. Diversity Analysis of Graph Frameworks. Frameworks at the graph level will be called *graph frameworks*. These are interesting because they describe the basic shape of a framework (we are referring here to "topological" shape rather than three-dimensional shape). The analysis of graph frameworks can give insight into the shape diversity of organic compounds. It can also give insight into the diversity of heteroatom patterns since a graph framework can be associated with more than one hetero framework.

B.1. Distribution of Framework Shapes. The distribution of graph frameworks among organic compounds can be illustrated by the same method used for hetero frameworks in which percentage of compounds is plotted as a function of percentage of frameworks. The resulting plot is presented in Figure 10. This plot, like Figure 6, shows the curve on an expanded *x* axis. The initial rise of this curve is extremely steep,

[29] Zipf, G. K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Cambridge, MA, 1949.

[30] Fialkowski, M.; Bishop, K. J. M.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. *Angew. Chem., Int. Ed.* **2005**, *44*, 7263–7269.

[31] Weisberg, S. *Applied Linear Regression*, 3rd ed.; Wiley-Interscience: Hoboken, NJ, 2005; pp 42−43.

[32] Yule, G. U. *Philos. Trans. R. Soc. London* **1925**, *B213*, 21–87.

[33] Simon, H. A. *Biometrika* **1955**, *42*, 425–440.

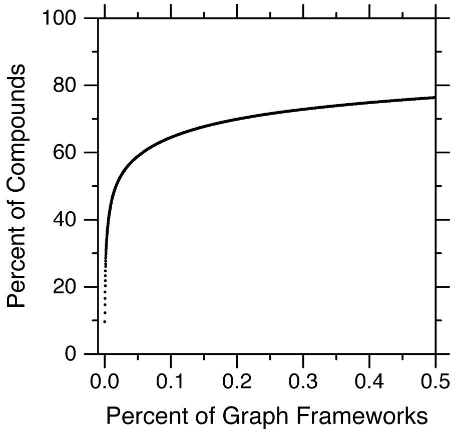


FIGURE 10. Percentage of compounds containing a particular percentage of graph frameworks. The *x* axis is greatly expanded and is on the same scale as Figure 6.

more so than the distribution of hetero frameworks. This means that the distribution of framework shapes is extremely top-heavy.

As expected for such a distribution, relatively few framework shapes are needed to describe the shapes of large numbers of organic compounds. For instance, the curve in Figure 10 shows that the shapes of 70.0% of the compounds are described by only 0.20% of the graph frameworks. In fact, to describe the shapes of just over 50% of the compounds requires only 143 graph frameworks, which represent just 0.017% of the total. This has important implications for structural diversity. It means that a small number of framework shapes play a dominant role in organic chemistry.

The extent to which these shapes dominate organic chemistry can be dramatically illustrated as follows. Imagine our set of organic compounds is split into two equal parts. We know this can be done in such a way that the shapes of the compounds in one half are described by only 143 graph frameworks. It follows that the shapes of the compounds in the other half are described by the other 836 565 graph frameworks. This implies that the set of known organic compounds can be divided in half in such a way that the shape diversity of one half is 5850 times greater than the shape diversity of the other half.

The 30 graph frameworks that occur most frequently are shown in Figure 11. These framework shapes are extremely common: 35.7% of the organic compounds have one of these 30 framework shapes; 32.3% have one of the first 20 shapes; 26.1% have one of the first 10. (The first few points in Figure 10 are distinctly visible because each of the top few graph frameworks occurs in such a high percentage of compounds.) This reinforces the view that much of organic chemistry is based on an extremely limited range of shapes. The vast majority of framework shapes, however, are associated with very few compounds. In fact, 393 144 graph frameworks occur in only a single compound. These constitute 47.0% of all graph frameworks but describe the shapes of only 1.6% of all compounds.

Not only do the graph frameworks in Figure 11 describe a large fraction of organic compounds, they also appear to describe a large fraction of the set of known drugs. As already noted, Bemis and Murcko found a set of 32 framework shapes that described half of the compounds in a drug database.[12] The first 17 shapes in Figure 11 are in this set of 32, and these 17 shapes account for 41% of the compounds in that drug database. In total, 24 of the shapes in Figure 11 are in the set of 32 drug shapes. This suggests there is considerable overlap between the most common shapes of drugs and the most common shapes of organic compounds in general.

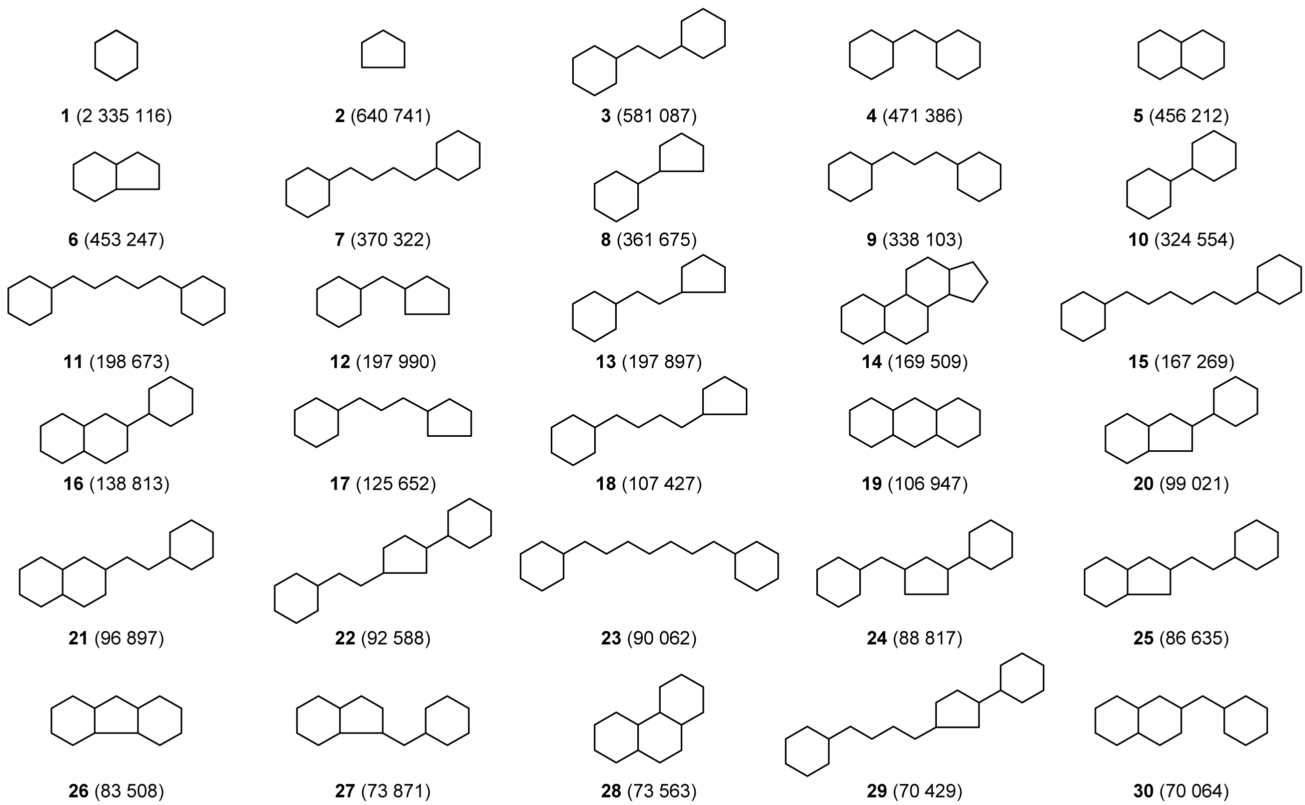


FIGURE 11. Most frequently occurring graph frameworks. Numbers of compounds in which they occur are shown in parentheses.

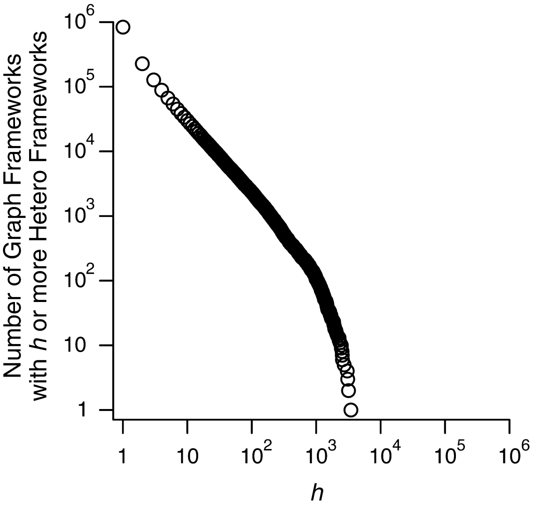


FIGURE 12. Distribution of the number of hetero frameworks per graph framework. This is plotted as a cumulative distribution.

B.2. Heteroatom Diversity of Framework Shapes. To this point, our diversity analysis has examined graph and hetero frameworks separately, but these two levels are related. Each graph framework, which describes a framework shape, can be associated with all the hetero frameworks having that shape. In other words, there is a mapping between graph and hetero frameworks, and this mapping tells us something about the diversity of heteroatom patterns within framework shapes. This is an aspect of structural diversity worth examining.

In going from the graph level to the hetero level, the number of frameworks grows by a factor of 3 (Table 1). As might be expected, this increase in diversity is not uniform over all graph frameworks. The number of hetero frameworks per graph framework varies widely. The distribution of this quantity is shown in Figure 12. Like Figure 9, this is a cumulative distribution plotted on a log−log scale. Most graph frameworks (72.9%) have only one hetero framework. However, as this plot shows, many framework shapes are associated with a large number of hetero frameworks; some shapes are associated with more than 3000 hetero frameworks.

The distribution of hetero frameworks per graph framework is definitely top-heavy: 1.0% of the graph frameworks give rise to 38.5% of all hetero frameworks. Nevertheless, the distribution in Figure 12 does not appear to follow a power law. Its tail shows a cutoff that is uncharacteristic of a true power-law distribution. This cutoff is likely due to factors that sharply limit the number of heteroatom patterns that can actually be incorporated within a given framework shape. These factors include chemical stability and synthetic accessibility. The size of the framework itself is also a limiting factor: larger frameworks have more possible heteroatom patterns, but they are harder to synthesize.

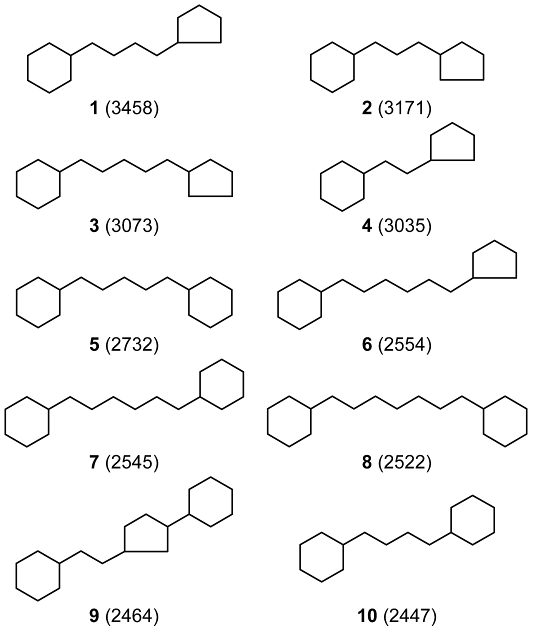


FIGURE 13. Graph frameworks with the most hetero frameworks. Numbers of hetero frameworks are shown in parentheses.

The framework shapes associated with the greatest numbers of hetero frameworks are shown in Figure 13. Most of these are found among the commonly used shapes in Figure 11. An unexpected finding is that almost all of these shapes have the same general motif: a pair of five- or six-membered rings linked by a chain of two or more atoms. The fact that chemists have created more hetero frameworks with these shapes than any others may be due in part to a combinatorial effect. Frameworks with this motif can be synthesized from a few simple reactants (e.g., one acyclic and two monocyclics), and by selecting different combinations of these from a relatively small pool of reactants, a very large number of hetero frameworks can be generated.

Conclusions

This work has attempted to characterize the structural diversity of organic chemistry through a framework analysis of the organic subset of the CAS Registry. The use of frameworks as the basis for diversity analysis has certain limitations: it excludes acyclic compounds and ignores that part of the structural diversity associated with acyclic groups (side chains) attached to the framework.[34] Nevertheless, this is a conceptually simple way to assess diversity and is easy to apply to an extremely large structure database. The present analysis did not try to decompose organic chemistry into more specialized classes, e.g., drugs or natural products. Further insights into diversity might be obtained by focusing on such classes, but the objective of the present analysis was to consider organic chemistry as a whole.

The most significant finding of this work is that the distribution of frameworks over all of organic chemistry conforms almost exactly to a power law. This is important because it tells us something about how chemists explore chemistry space. It suggests that the exploration of chemistry space is governed to a large extent by a rich-get-richer process. In other words, chemists are more likely to use a particular framework to make a compound the more often that framework has been used in the past. This type of process will tend to amplify the synthetic activity associated with a framework. This results in the proliferation of certain frameworks and leads to an overall distribution that obeys a power law.

It is not surprising that some frameworks occur much more frequently than others. However, the extreme unevenness in the way frameworks are distributed among organic compounds is somewhat surprising. This is particularly true at the graph level,

[34] Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1999**, *42*, 5095–5099.

where it is found that only 143 framework shapes can describe half of the compounds. The fact that both graph and hetero frameworks have very top-heavy distributions tells us that the exploration of organic chemistry space has tended to concentrate on relatively small numbers of structural motifs.

The exploration of chemistry space depends on decision making about what to synthesize. A significant concern in these decisions is the cost of synthesis, as measured in materials and time. The cost of making a new derivative of a framework is probably lowered if many other derivatives are known since this increases the chances of finding an appropriate precursor that can be purchased or made by a published synthesis. This gives rise to the rich-get-richer process, whose signature is the power-law framework distribution. We believe the presence of this power law is quantitative evidence that the minimization of synthetic cost has been a key factor in shaping the known universe of organic chemistry.

This work has shown that the study of chemical diversity can reveal interesting patterns among tens of millions of organic compounds, the collective output of many decades of synthetic chemistry. This kind of large-scale study could have practical value, especially for the field of drug discovery. A lack of structural diversity among test compounds has been cited as a potential bottleneck in the drug discovery process.[35] By identifying regions of chemistry space that are underexplored, large-scale diversity studies might play a helpful role in guiding future synthetic efforts.

JO8001276

[35] Burke, M. D.; Berger, E. M.; Schreiber, S. L. *Science* **2003**, *302*, 613–618.