Exploration of the Accessible Chemical Space of Acyclic Alkanes

Robert S. Paton and Jonathan M. Goodman*

Department of Chemistry, Unilever Centre for Molecular Science Informatics, Lensfield Road, Cambridge CB2 1EW, U.K.

Received July 12, 2007

Saturated acyclic alkanes show steric strain if they are highly branched and, in extreme cases, fall apart rapidly at room temperature. Consequently, attempts to count the number of isomeric forms for a given molecular formula that neglect this physical consideration will inevitably overestimate the size of the available chemical space. Here we derive iterative equations to enumerate the number of isomers (both structural and optical are considered separately) for the alkane series that take into account the inherent instability of certain carbon skeletons. These function by filtering out certain substructures from the graph representation of the molecule which have been found to be thermodynamically unstable. We use these relations to report new estimates of the size of physically accessible chemical space for acyclic alkanes and show that for large molecules there are more isomers that are structurally disallowed than there are allowed.

1. INTRODUCTION

The familiar concepts of isomerism were developed in the 1860s,[1] when it was realized that there were different possible structural formulas for the same atomic composition. This realization provided crucial evidence for the validity of structural formulas and the existence of structured molecules. Since then various methods have been devised to calculate the precise number of possible isomers (both structural and stereochemical) for a given molecular formula.

The problem of the number of isomeric hydrocarbons of the methane series has concerned mathematicians and chemists since 1874. It was in this year that Cayley first formulated iterative equations to formally enumerate the number of isomers of the alkanes.[2] This has continued to be a topic of interest, particularly with Henze and Blair making several enumerations for alkanes and homologous sequences of alkane derivatives using recursion algorithms in 1931.[3] Shortly afterward, motivated by this problem of chemical isomerism, Pólya[4,5] introduced a powerful enumeration algorithm that used molecular symmetry, weighting factors, and generating functions.

More recent applications of isomer enumeration have been seen in the field of combinatorial chemistry. Various estimates have been proposed for the size of chemical space of small organic molecules, ranging from $10^{18}-10^{200}$ compounds.[6] Enumeration techniques are used to chart the size and composition of the small molecule chemical universe, from which the structural diversity of a combinatorial library can be quantified.[7] Davies and Freyd have also considered the physical reality of the astronomical numbers involved in isomer enumeration—they calculated that $C_{167}H_{336}$ is the smallest alkane for which the number of stereoisomers exceeds the classical figure for the number of particles in the universe ($\sim 10^{80}$).[8]

In many of these approaches to isomer enumeration it is assumed that all conceivable molecules can exist. However, alkanes may be very strained, even without cyclic structures, as multiple branching pushes side chains close together. Neighboring quaternary centers are likely to lead to highly strained structures.[9] Klein[10,11] has considered self-avoidingly embedding alkanes onto a tetrahedral (diamond) lattice, thus simulating the sp^3 bond angles of unsaturated hydrocarbons. The smallest acyclic alkane isomers which cannot fit onto the lattice are shown in Figure 1.

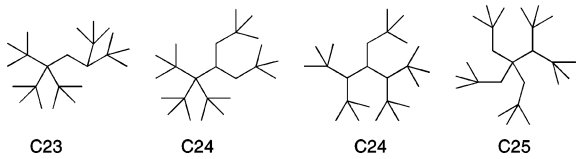

Figure 1. The smallest alkanes which cannot be embedded on a diamond lattice.

Figure 2. Highly strained branched alkanes.

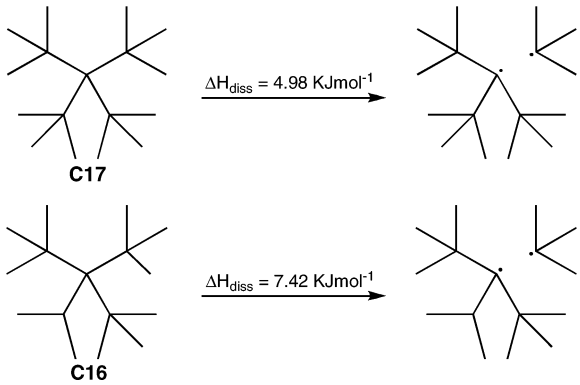

Figure 3. Fragmentation reactions of highly strained acyclic alkanes into radicals.

Any unsaturated hydrocarbon which cannot be fit on a diamond lattice is clearly very strained. However, tetra-*tert*-

* Corresponding author e-mail: jmg11@cam.ac.uk.

10.1021/ci700246b CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/20/2007

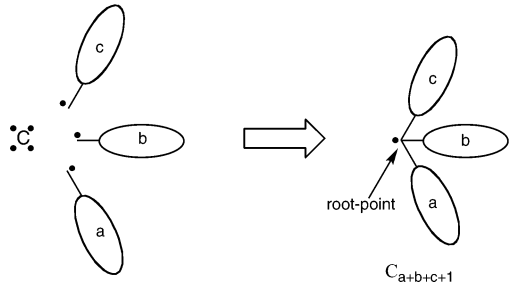

Figure 4. The iterative generation of alkyl radicals (rooted trees). Here three smaller alkyl radicals (containing a, b, and c carbon atoms) are joined to create a new radical of size a+b+c+1. Radicals can also be zero-order, consisting of a single H-atom.

butylmethane ($C_{17}H_{36}$) (Figure 2) can be fitted on a diamond lattice, and yet it has never been synthesized and theoretical studies have shown it must be extremely strained.[12]

Studies within our group on acyclic alkane conformations[13,14] led us to the question of the simplest acyclic alkane that cannot be made.[15] Certainly, extraordinary molecules can be made and analyzed at very low temperatures for a brief time before they decompose. For example, cyclobutadiene can be characterized at low temperatures provided it is embedded in a solid argon or nitrogen matrix. However, in previous studies and in this paper we concern ourselves with molecules that can be made and stored at room temperature.

Analysis of the strain energy from molecular mechanics calculations was used to calculate the activation energy of the dissociation of highly branched alkanes into radicals.[16] These results, summarized in Figure 2, suggest that tri-*tert*-butyl-isopropylmethane ($C_{16}H_{34}$) will spontaneously undergo thermolysis at −16 °C, while tetra-*tert*-butylmethane ($C_{17}H_{36}$) will undergo thermolysis at −145 °C. Additional DFT calculations suggest that the homolytic dissociation of both C16 and C17 structures into radicals to be slightly endothermic (Figure 3). However, entropic effects will favor the dissociation of one molecule into two, strongly suggesting that these two molecules are both thermodynamically unstable. On this basis we made the prediction that C17 could never be made and suggested that C16 is the smallest alkane that cannot be made.

In this paper we return to the classic example of alkane isomerism. We formally develop the mathematical relations established by Henze and Blair[3] and Davies and Freyd to enumerate the number of isomeric hydrocarbons of the methane series, while excluding any isomers that contain the unstable highly branched C16 and C17 substructures. Thus our isomer count (both structural and stereochemical isomers are considered here) represents only those molecules which are realizable/ synthesizable at room temperature. We have implemented these new mathematical relations using the Java programming language,[17] and we have enumerated the numbers of isomers for the alkanes up to $C_{2000}H_{4002}$. We then analyze the effects that these physical constraints have on the size of accessible chemical space for the saturated hydrocarbons, by comparison with a simple isomer counting scheme that takes no account of the inherent instability of certain structures.

2. ENUMERATION METHODOLOGY

In the standard approach to isomer enumeration each isomer is represented as a molecular graph.[18] The graphs are

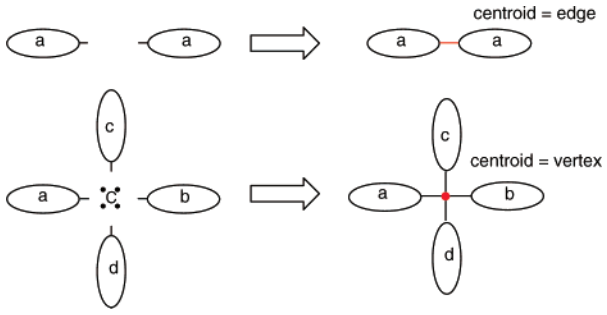

Figure 5. Two manners of joining rooted trees to obtain an alkane: two radicals of size a can join along a common bond to create an alkane of size 2a, or four radicals of size a, b, c, and d (up to two may be zero) can join at a central atom to create an alkane of size a+b+c+d+1.

"hydrogen deleted" in the sense that only carbon atoms and C−C bonds are explicitly represented. Thus structural isomers of the alkane C_nH_{2n+2} can be uniquely represented by n vertices, each of which is connected to one, two, or three edges (signifying a primary, secondary, or tertiary carbon atom, respectively). An enumeration is first made of alkyl radicals or graph-theoretically speaking "rooted trees". These trees can be iteratively generated from three smaller rooted trees (i.e., alkyl radicals) and so on as shown below in Figure 4.

The number of structural isomers $R(n)$ for a given rooted tree of size n can be calculated from the iterative relations:

$$R(0) = 1$$

$$R(1) = 1$$

$$R(n+1) = \sum_{\substack{a<b<c \\ a+b+c=n}} R(a)R(b)R(c) + \sum_{\substack{a\neq c \\ 2a+c=n}} \binom{R(a)+1}{2} R(c) + \sum_{3a=n} \binom{R(a)+2}{3} \quad (1)$$

The three terms correspond to the cases of no pairs, one pair, and three of a kind with regard to the size of smaller radicals a, b, and c. The binomial coefficient

$$\binom{m+k-1}{k} = \frac{m!}{k!(m-k)!} \quad (2)$$

is equal to the number of ways of choosing k things, each of one of m types with repeated types allowed.

The various rooted trees thus obtained can be joined together to form alkanes as shown in Figure 5. Any graph representation with n vertices has a unique "centroid", defined as either an edge that when removed yields two rooted trees with $n/2$ vertices or as a vertex that when removed yields a set of trees each with fewer than $n/2$ vertices. A centroid edge is only possible for graphs with an even number of vertices. Every structural isomer can be constructed in this fashion, so the number of structural isomers $T(n)$ for the alkane C_nH_{2n+2} can be calculated from the iterative relation

$$T(n) = \sum_{2a=n} \binom{R(a)+1}{2} + \sum_{\substack{a<b<c<d \\ a+b+c+d=n-1}} R(a)R(b)R(c)R(d) + \sum_{\substack{c<d<n/2 \\ a\neq c, a\neq d \\ 2a+c+d=n-1}} \binom{R(a)+1}{2} R(c)R(d) + \sum_{\substack{a<c \\ 2a+2c=n-1}} \binom{R(a)+1}{2}\binom{R(c)+1}{2} + \sum_{a\neq d<n/2} \binom{R(a)+2}{3} R(d) + \sum_{4a=n-1} \binom{R(a)+3}{4} \quad (3)$$

The first term counts the number of isomers with a centroid edge. The remaining terms count the number of isomers with a centroid vertex and correspond to the cases of no pairs, one pair, two pairs, three of a kind, and four of a kind with regards to rooted tree size.

The enumeration method outlined thus far counts the number of structural isomers for the alkane series. It can be further modified to count all possible *stereo*isomers for the alkane series, utilizing the notion of a stereotree.[8,18,20] A stereotree is formed from smaller rooted trees as above; however, the ordering of rooted trees is important: two such orderings are equivalent if one is obtainable from the other by an even permutation. The formulas are modified as follows:

$$R(1) = 1$$

$$R(n+1) = \sum_{\substack{a<b<c \\ a+b+c=n}} R(a)R(b)R(c) + \sum_{\substack{a\neq c \\ 2a+c=n}} F(R(a),2)R(c) + \sum_{3b=n} F(R(b),3) \quad (4)$$

Here

$$F(n,k) = n + K - 1k + nk \quad (5)$$

is the number of ways of choosing a sequence of k things each of m types with repeated types allowed, where two such sequences are equivalent if one is obtainable from the other by an even permutation.

$$T(n) = \sum_{2a=n} \binom{R(a)+1}{2} + 2\sum_{\substack{a<b<c<d \\ a+b+c+d=n-1}} R(a)R(b)R(c)R(d) + \sum_{\substack{c<d<n/2 \\ a\neq c, a\neq d \\ 2a+c+d=n-1}} F(R(a),2)R(c)R(d) + \sum_{\substack{a<c \\ 2a+2c=n-1}} \binom{R(a)R(c)+1}{2} + \sum_{a\neq d<n/2} F(R(a),3)R(d) + \sum_{4a=n-1} F(R(a),4) \quad (6)$$

Table 1. Numbers of Possible Isomers for Alkanes C_nH_{2n+2}

n	structural isomers	stereoisomers
1	1	1
2	1	1
3	1	1
4	2	2
5	3	3
6	5	5
7	9	11
8	18	24
9	35	55
10	75	136
20	366319	3396844
40	6.2481801147 E13	1.3180446189 E16
60	2.2158734536 E22	1.0425013434 E26
80	1.0564476907 E31	1.1043761194 E36
100	5.9210720381 E39	1.3734319092 E46
200	9.4304332880 E83	1.1696865263 E97
500	6.9888806978 E217	1.3134900146 E251
1000	7.3535427719 E441	5.9346473764 E508
2000	4.6055365780 E890	6.8465857595 E1024

The number of isomers for the alkane series up to $C_{2000}H_{4002}$ was calculated using these relations, and the results are summarized below in Table 1. (For larger numbers all digits are given in the Supporting Information). The asymptotic behavior for alkane enumerations approaching the many atom limit has been described in the work of Pólya,[4] Otter,[21] and more recently by Davies[8] and Bytautas and Klein.[22] It has shown that the behavior of $R(n)$ and $T(n)$ as n increases can be described by asymptotic forms, the simplest of which is of the form $AB^n/n^{2.5}$.[21] In the limit of very large values of n, the growth factor (i.e., $R(n+1)/R(n)$) is given by B in the above relationship. Applying the asymptotic analysis techniques of Otter to our data, we obtained values for A and B. Rather than a least-squares regression, we used the logarithmic relationship derived in eq 7.

$$T(n) = \frac{AB^n}{n^{2.5}}$$

$$\therefore \ln(T(n) \times n^{2.5}) = \ln A + n\ln B \quad (7)$$

A plot of $\ln(T(n) \times n^{2.5})$ against n using our data within the range $n = 100$ to $n = 1000$ gave a straight line (linear correlation coefficient > 0.999) and from the gradient and y intercept we obtained a value for A of 0.657 and a value for B (growth factor) of 2.815 for the structural isomer count, in agreement with previous work. We used these values to extrapolate outside this data range and provide an estimate for the number of isomers of $C_{2000}H_{4002}$. The resulting figure, 4.603×10^{890}, underestimates the fully enumerated value by only 0.06%, and so this fitting seems very robust and could conceivably be used with some confidence to predict even larger isomer counts without the need for explicit computation, which demands ever larger amounts of CPU time for bigger structures (An explicit enumeration for $C_{2000}H_{4002}$ takes around 48 h on a single processor machine.).

3. FILTERING UNWANTED SUBSTRUCTURES

Since the quasi-spherical C16 or C17 molecules (Figure 2) are the smallest alkanes thermodynamically unstable toward homolysis, it is logical to assume that any bulkier molecule will itself suffer the same fate. That is to say, any

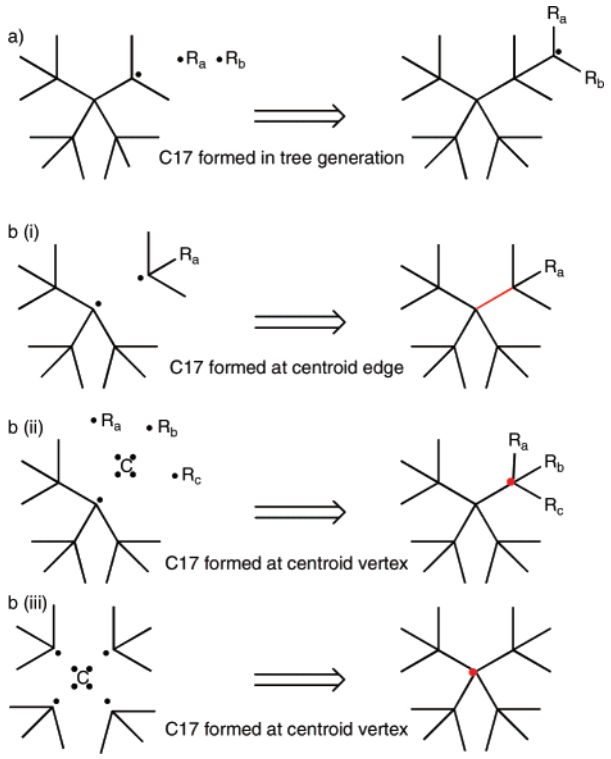

Figure 6. The undesirable C17 substructure can arise in the construction of (a) alkyl radicals or (b) bringing together certain alkyl radicals at the centroid (edge or vertex) in the context of our enumeration method.

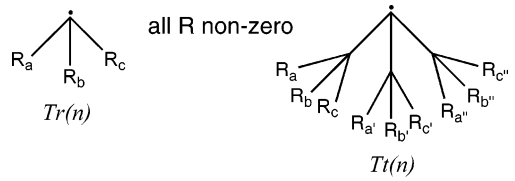

Figure 7. The structure of radicals enumerated by $T(n)$ and $Tt(n)$.

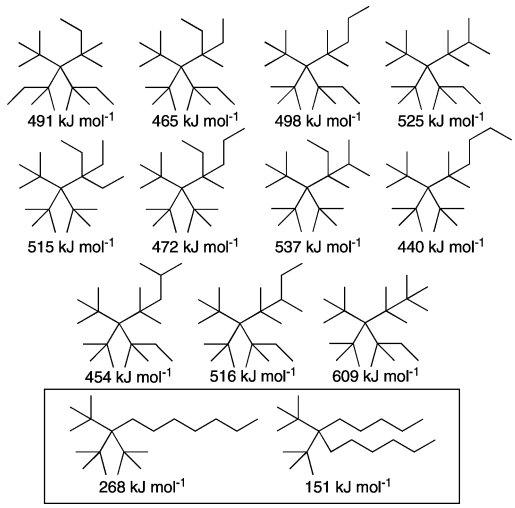

Figure 8. Strain energy (kJ mol^{-1}) of the 11 disallowed isomers of $C_{20}H_{42}$. For comparison, two sterically hindered isomers without the C17 substructure are shown at the bottom of the figure (boxed).

structural isomer explicitly containing the C16 or C17 substructural unit at some position will itself be thermodynamically unstable. Subtraction of the number of isomers containing these substructures from the number of conceivable isomers will yield a figure for the number of isomers which can, in theory, be synthesized. We treat the two cases (C16 disallowed and C17 disallowed) separately to quantitatively study the effect that a more or less stringent condition has on the isomer count.

(i) C17 Elimination. An isomer is eliminated if the C17 (tetra-*tert*-butylmethane) substructure occurs *at least once* somewhere in the entire structure. In the enumeration methodology outlined above the occurrence of this substructure could arise (a) in the formation of the rooted trees or (b) as the rooted trees come together at the centroid edge or vertex, or a combination of both (Figure 6).

In addition to the previously defined alkyl radical generating function $R(n)$, we define the terms $Tr(n)$ to enumerate those radicals tertiary at their root (i.e. with three nonzero branches) and $Tt(n)$ to enumerate those radicals tertiary at the root, with each adjacent site being quaternary. These values are given by the following relations:

$$0 \le n \le 3, Tr(n) = 0$$

$$Tr(n+1) = \sum_{\substack{0<a<b<c \\ a+b+c=n}} R(a)R(b)R(c) + \sum_{\substack{0\neq a\neq c \\ 2a+c=n}} \binom{R(a)=1}{2} R(c) + \sum_{3a=n} \binom{R(a)=2}{3} \quad (8)$$

$$0 \le n \le 12, Tt(n) = 0$$

$$Tt(n+1) = \sum_{\substack{3<a<b<c \\ a+b+c=n}} Tr(a)Tr(b)Tr(c) + \sum_{\substack{a\neq c \\ 3<a,3<c \\ 2a+c=n}} \binom{Tr(a)=1}{2} Tr(c) + \sum_{3a=n} \binom{Tr(a)=2}{3} \quad (9)$$

Thus to avoid the scenario in Figure 6a trees which are allowed, herein enumerated as $R'(n)$, must not connect a Tt radical with two nonzero radicals at any point in their iterative formation. The first term of the iteration (eq 9) represents the case for which three nonzero radicals of different size are joined together. In this case, none of these radicals must be a Tt branch so these are subtracted from the available number of radicals a, b, and c. The second term has a zero radical (i.e., a hydrogen atom) so there are no restrictions. The next three terms represent the case of a pair (two radicals the same size) in which there are no, one, and two hydrogen radicals, respectively. The final term considers the scenario of three nonzero radicals of the same size.

$$0 \le n \le 15, R'(n) = R(n)$$

$$R'(n+1) = \sum_{\substack{0<a<b<c \\ a+b+c=n}} (R'(a) - Tt'(a))(R'(b) - Tt'(b))(R'(c) - Tt'(c)) + \sum_{b+c=n} R'(b)R'(c) + \sum_{\substack{0\neq a\neq c \\ 2a+c=n}} \binom{R'(a) - Tt'(a) + 1}{2} (R'(c) - Tt'(c)) + \sum_{2a=n} \binom{R'(a)+1}{2} + R'(n) + \sum_{3a=n} \binom{R'(a) - Tt'(a) + 2}{3} \quad (10)$$

where $Tr'(n)$ and $Tt'(n)$ are the allowed versions of $T(n)$ and $Tt(n)$ respectively, constructed in an entirely analogous fashion but using allowed radicals only in their construction:

$$Tr'(n+1) = \sum_{\substack{0<a<b<c \\ a+b+c=n}} R'(a)R'(b)R'(c) + \sum_{\substack{0\neq a\neq c \\ 2a+c=n}} \binom{R'(a)+1}{2} R'(c) + \sum_{3a=n} \binom{R'(a)+2}{3} \tag{11}$$

$$Tt'(n+1) = \sum_{\substack{3<a<b<c \\ a+b+c=n}} Tr'(a)Tr'(b)Tr'(c) + \sum_{\substack{a\neq c \\ 3<a,3<c \\ 2a+c=n}} \binom{Tr'(a)+1}{2} T'(c) + \sum_{3a=n} \binom{Tr'(a)+2}{3} \tag{12}$$

Turning to the formation of disallowed structures as branched trees are brought together (Figure 6b), we then proceed to enumerate alkane structures adapting the relationship in eq 3 to include only those isomers that are disallowed at the centroid, $N'(n)$. The first two terms describe a centroid edge, which connects at least one *Tt* rooted tree with a tree which is tertiary or more branched (see Figure 6b(i**)**). All subsequent terms describe a centroid vertex. We account for the two disallowed scenarios in Figure 6b(ii**)** by summation of all centroids with four tertiary branched radicals and all centroids with at least one *Tt* branched radical and three nonzero radicals. Care is taken to avoid double counting, so for the case of four tertiary radicals, we exclude the existence of *Tt* radicals so the number of possible radicals for each branch is $Tr'(a)$-$Tt'(a)$. The second term is obtained by subtracting all isomers which contain no *Tt* radicals from all possible isomers to leave only those structures which contain at least one *Tt* radical at the centroid. This process is repeated for the different scenarios which correspond to the cases of no pairs, one pair, two pairs, three of a kind, and four of a kind with regards to rooted tree size.

$$\begin{aligned}
N'(n) = {} & \sum_{2a=n} \left(\binom{Tt'(a)+1}{2} + Tt'(a)(Tr'(a) - Tt'(a)) \right) + \\
& \sum_{\substack{a<b<c<d \\ a+b+c+d=n-1}} (Tr'(a) - Tt'(a))(Tr'(b) - Tt'(b))(Tr'(c) - Tt'(c))(Tr'(d) - Tt'(d)) \\
& + \sum_{\substack{a<b<c<d \\ a+b+c+d=n-1}} \begin{pmatrix} R'(a)R'(b)R'(c)R'(d) - (R'(a) - Tt'(a))(R'(b) \\ -Tt'(b))(R'(c) - Tt'(c))(R'(d) - Tt'(d)) \end{pmatrix} + \\
& \sum_{\substack{c<d<n/2 \\ a\neq c, a\neq d \\ 2a+c+d=n-1}} \left(\begin{array}{l} \binom{Tr'(a) - Tt'(a) + 1}{2}(Tr'(c) - Tt'(c))(Tr'(d) - Tt'(d)) \\ + \binom{R'(a)+1}{2} R'(c)R'(d) - \binom{R'(a) - Tt'(a) + 1}{2}(R'(c) - Tt'(c))(R'(d) - Tt'(d)) \end{array} \right) \\
& + \sum_{\substack{a<c \\ 2a+2c=n-1}} \binom{Tr'(a) - Tt'(a) + 1}{2} \binom{Tr'(c) - Tt'(c) + 1}{2} \\
& + \sum_{\substack{a<c \\ 2a+2c=n-1}} \left(\binom{R'(a)+1}{2} \binom{R'(c)+1}{2} - \binom{R'(a) - Tt'(a) + 1}{2} \binom{R'(c) - Tt'(c) + 1}{2} \right) \\
& + \sum_{a\neq d<n/2} \left(\binom{R'(a)+2}{3} R'(d) - \binom{R'(a) - Tt'(a) + 2}{3}(R'(d) - Tt'(d)) \right) \\
& + \sum_{4a=n-1} \left(\binom{Tr'(a) - Tt'(a) + 3}{4} + \binom{R'(a)+3}{4} - \binom{R'((a)) - Tt'((a)) + 3}{4} \right)
\end{aligned} \tag{13}$$

Subtracting $N'(n)$ from the number of isomers with all possible centroids (similarly enumerated using only allowed rooted trees $R'(n)$) then yields the number of structural isomers which do not contain the C17 substructure *anywhere* in the molecule. An analogous procedure was followed to obtain values for the number of allowed stereoisomers (full details of the enumeration are given in the Supporting Information). The results are shown in Table 2.

For $C_{20}H_{42}$, 11 structural isomers are disallowed in our enumeration method. The strain energy of these disallowed isomers was evaluated by molecular mechanics. Our earlier studies suggest that a strain energy of about 300 kJ mol^{-1} is sufficient to make the molecule unstable at room temperature.[14] In all cases the strain energy of these isomers is greater than that calculated for the isolated C17 molecule (416.05 kJ mol^{-1}), and so these isomers are also likely to be unstable. By comparison, the strain energy of two isomers not containing the C17 substructure (but with some steric crowding) is significantly less than that of both the C17 molecule and the C16 molecule (335 kJ mol^{-1}). We conclude that our counting scheme is correct to include these less-strained structures.

This problem of steric overcrowding severely limits the number of possible isomers that could exist as n becomes large. Since the number of "virtual" isomers (i.e., all those that can be drawn on paper without physical consideration, as calculated

Table 2. Numbers of Alkane Isomers That Are Allowed, i.e., Do Not Contain the Unstable C17 Substructure Anywhere in the Molecule, Also Expressed as a Fraction of the Total Possible Number Neglecting Physical Considerations

	structural isomers			stereoisomers		
n	$N_{allowed}$	$N_{allowed}/N_{possible}$	$N_{allowed}/N_{disallowed}$	$N_{allowed}$	$N_{allowed}/N_{possible}$	$N_{allowed}/N_{disallowed}$
17	2.48930 E4	1.00	24893.0	1.43254 E5	1.00	143254.0
18	6.05220 E4	1.00	60522.0	4.08428 E5	1.00	408428.0
19	1.48280 E5	1.00	37070.0	1.17377 E6	1.00	293441.5
20	3.66308 E5	1.00	33300.7	3.39683 E6	1.00	242630.7
50	1.11734 E18	1.00	2764.6	1.11430 E21	1.00	13286.7
100	5.91386 E39	1.00	819.6	1.37303 E46	1.00	3403.7
150	6.43477 E61	1.00	461.1	3.45929 E71	1.00	1837.6
200	9.40087 E83	1.00	318.0	1.16875 E97	1.00	1242.9
250	1.61399 E106	1.00	241.9	4.63809 E122	1.00	935.2
300	3.06879 E128	0.99	194.9	2.03791 E148	1.00	748.3
350	6.26071 E150	0.99	163.1	9.60645 E173	1.00	623.1
400	1.34484 E173	0.99	140.2	4.76754 E199	1.00	533.5
450	3.00481 E195	0.99	122.8	2.46094 E225	1.00	466.3
500	6.92552 E217	0.99	109.3	1.31033 E251	1.00	414.1
600	3.94966 E262	0.99	89.5	3.98784 E302	1.00	338.2
700	2.41688 E307	0.99	75.8	1.30213 E354	1.00	285.7
800	1.55718 E352	0.98	65.7	4.47655 E405	1.00	247.3
900	1.04356 E397	0.98	57.9	1.60073 E457	1.00	217.9
1000	7.21421 E441	0.98	51.8	5.90434 E508	0.99	194.8
2000	4.42848 E890	0.96	25.0	6.77474 E1024	0.99	94.3

Table 3. Numbers of Alkane Isomers Which Are Allowed, i.e., Do Not Contain the Unstable C16 Substructure Anywhere in the Molecule, Also Expressed as a Fraction of the Total Possible Number Neglecting Physical Considerations

	structural isomers			stereoisomers		
n	$N_{allowed}$	$N_{allowed}/N_{possible}$	$N_{allowed}/N_{disallowed}$	$N_{allowed}$	$N_{allowed}/N_{possible}$	$N_{allowed}/N_{disallowed}$
17	2.48910 E4	1.00	8297.0	1.43251 E5	1.00	35812.8
18	6.05140 E4	1.00	6723.8	4.08417 E5	1.00	34034.8
19	1.48256 E5	1.00	5294.9	1.17372 E6	1.00	24452.5
20	3.66231 E5	1.00	4161.7	3.39667 E6	1.00	19190.2
50	1.11402 E18	1.00	299.2	1.11321 E21	1.00	951.4
100	5.85644 E39	0.99	90.6	1.36820 E46	1.00	261.7
150	6.32555 E61	0.98	51.4	3.43696 E71	0.99	142.0
200	9.17178 E83	0.97	35.5	1.15758 E97	0.99	95.7
250	1.56266 E106	0.96	26.9	4.57909 E122	0.99	71.6
300	2.94841 E128	0.96	21.7	2.00545 E148	0.98	57.0
350	5.96874 E150	0.95	18.1	9.42242 E173	0.98	47.2
400	1.27220 E173	0.94	15.5	4.66075 E199	0.98	40.3
450	2.82042 E195	0.93	13.5	2.39783 E225	0.97	35.1
500	6.44990 E217	0.92	12.0	1.27247 E251	0.97	31.0
600	3.62113 E262	0.91	9.7	3.84672 E302	0.96	25.2
700	2.18116 E307	0.89	8.2	1.24760 E354	0.95	21.1
800	1.38320 E352	0.87	7.0	4.26020 E405	0.95	18.2
900	9.12370 E396	8.59	6.1	1.51310 E457	0.94	15.9
1000	6.20750 E441	0.84	5.4	5.54320 E508	0.93	14.2
2000	3.24047 E890	0.70	2.4	5.93893 E1024	0.87	6.5

without any substructure filtering) increases exponentially, the ratio of possible to impossible isomers will asymptotically approach zero even though the number of possible isomers will approach infinity as n increases. In the case of stereoisomerism this approach toward zero occurs more slowly. This behavior is clearly identifiable from Figure 9. Even at $C_{2000}H_{4002}$ there are considerably more allowed than disallowed structural isomers (ratio allowed:disallowed = 25:1).

(ii) C16 Elimination. An isomer is eliminated if the C16 (tri-*tert*-butyl-isopropylmethane) substructure occurs *at least once* somewhere in the entire structure. The recursion relations in this case are significantly more complex than for C17 since the undesirable substructure, being smaller, can arise in the carbon skeleton in many more ways. As before, the undesirable substructure could arise (a) in the formation of the rooted trees or (b) as the rooted trees come together at the centroid edge or vertex, or a combination of both, shown in Figure 11.

In addition to previously defined $R(n)$, $Tr(n)$, and $Tt(n)$, here we need to define the term $S(n)$ to enumerate secondary radicals and the term $Ts(n)$ which is used to enumerate those radicals which are tertiary at their root and where each branch also possesses a tertiary root *or* one where one branch is secondary and the other two are tertiary (i.e., $Tt(n)$ is a subgroup of $Ts(n)$). These values are obtained from the following recursion relations:

$$0 \le n \le 2, S(n) = 0$$

$$S(n+1) = \sum_{0<b<c}^{b+c=n} R(b)R(c) + \sum_{2a=n} \binom{R(a)+1}{2} \quad (14)$$

$$0 \le n \le 12, Ts(n) = 0$$

$$Ts(n+1) = \sum_{\substack{3<a<b<c \\ a+b+c=n}} Tr(a)Tr(b)Tr(c)Tr(d) + \sum_{\substack{3<a<b<c \\ a+b+c=n}} (S(a)Tr(b)Tr(c) + Tr(a)S(b)Tr(c) + Tr(a)Tr(b)S(c)) + \sum_{\substack{a\neq c \\ 2a+c=n}} \left(\binom{Tr(a)+1}{2}(S(c)+Tr(c)) + S(a)Tr(a)Tr(c)\right) + \sum_{3a=n} \left(\binom{Tr(a)+2}{3} + \binom{Tr(a)+1}{2}S(a)\right) \quad (15)$$

Again we are in a position to enumerate those alkyl radicals that will not lead to an undesirable alkane structure. From Figure 11a there are two possible radical structures that always form the disallowed C16 skeleton upon connection at their root position. Hence in the enumeration of allowed radicals, $R'(n)$, these substructures are excluded to give the relationship in eq 15 where $Ts'(n)$ is the version of $Ts(n)$ constructed from solely allowed radicals:

$$0 \le n \le 14, R'(n) = R(n)$$

$$R'(n+1) = \sum_{\substack{0<b<c \\ b+c=n}} (R'(b) - Tt'(b))(R'(c) - Tt'(c)) + \sum_{\substack{0<a<b<c \\ a+b+c=n}} (R'(a) - Ts'(a))(R'(b) - Ts'(b)(R'(c) - Ts'(c)) + \sum_{\substack{0\neq a\neq c \\ 2a+c=n}} \binom{(R'(a) - Ts'(a) + 1)}{2}(R'(c) - Ts'(c)) + \sum_{2a=n} \binom{(R'(a) - Tt'(a) + 1)}{2} + R'(n) + \sum_{3a=n} \binom{(R'(a) - Ts'(a) + 2)}{3} \quad (16)$$

We then proceed to enumerate alkane structures adapting the relationship in eq 3 to include only those isomers that are disallowed at the centroid, $N'(n)$, which can occur in any of five ways shown in Figure 10b. In the case of a centroid edge (Figure 10b(i)),b(ii)), the disallowed C16 substructure can arise from the connection of a *Ts* and a *Tr* radical or of a *Tt* and an *S* radical. In the case of a centroid vertex, the C16 substructure can arise in one of three ways: (i) A *Ts* radical and three nonzero radicals are brought together. (ii) A *Tt* radical and two nonzero radicals are brought together. One of the four branches around the centroid *must* be zero to avoid overlap with the previous case. (iii) Three tertiary and one secondary or four tertiary radicals are brought together. These *must not* be *Ts* radicals to avoid overlap with the first case.

These structures are enumerated for the different scenarios which correspond to the cases of no pairs, one pair, two pairs, three of a kind, and four of a kind with regards to rooted tree size. Subtracting this total from the number of isomers with all possible centroids (similarly enumerated using only allowed rooted trees $R'(n)$) then yields the number of structural isomers which do not contain the C16 substructure *anywhere* in the molecule. The results are shown in Table 3. For full details of the recursion relation refer to the Supporting information.

Since the C16 substructure is smaller than the C17, there are more ways in which it can arise in the carbon skeleton.

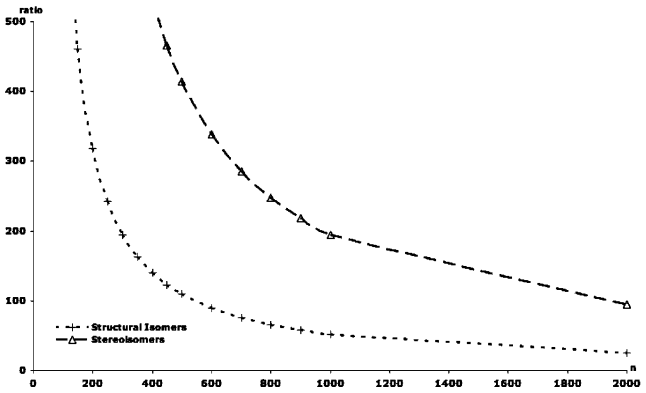

Figure 9. Plot of the ratio of thermodynamically stable to unstable alkane isomers when the C17 substructure is filtered out by our enumeration scheme. Clearly evident is the asymptotic approach to zero of the ratio of allowed to disallowed isomers with increasing alkane size (i.e., increasing proportion of disallowed isomers).

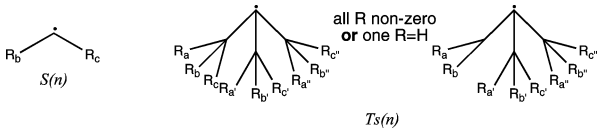

Figure 10. The structure of radicals enumerated by $S(n)$ and $Ts(n)$.

Unsurprisingly the proportion of disallowed structures increases more rapidly, and by the time $C_{2000}H_{4002}$ is reached the ratio of allowed to disallowed structures is 2.4:1. The decline in the proportion of allowed isomers is shown in Figure 12.

4. RESULTS AND DISCUSSION

Assuming all isomer counts were of the form $AB^n/n^{2.5}$ we constructed logarithmic plots using data in the range $n = 100$ to $n = 2000$ as detailed earlier in the text. These plots yielded values of A and B for our "allowed" isomer counts (i.e., those that filter out the C16 and C17 undesired substructures), and in every case the linear correlation coefficient was greater than 0.999 so the behavior is described well by the above form. For the numbers of allowed isomers smaller values of B are obtained than for the total number of "paper" isomers—this is because their growth is slowed by the increasing proportion of disallowed isomers. Hence disallowing the C16 substructure leads to a greater reduction in the growth factor as the proportion of disallowed isomers is larger. From the values of A and B thus obtained (Figure 13), we can easily calculate and make predictions of the ratio of allowed:total number of "virtual" (i.e., unfiltered) isomers or equally the ratio of allowed: disallowed isomers.

$$\frac{T_{\text{allowed}}}{T_{\text{paper}}} = \frac{A_a}{A_p}\left(\frac{B_a}{B_p}\right)^n \quad (17)$$

$$\frac{T_{\text{allowed}}}{T_{\text{paper}}} = \left(\frac{1}{\frac{A_a}{A_p}\left(\frac{B_a}{B_p}\right)^n - 1}\right) \quad (18)$$

Since the growth factor of the total number is greater than for the allowed isomer counts ($B_p > B_a$), the denominator of the ratio of allowed to disallowed isomers tends to infinity

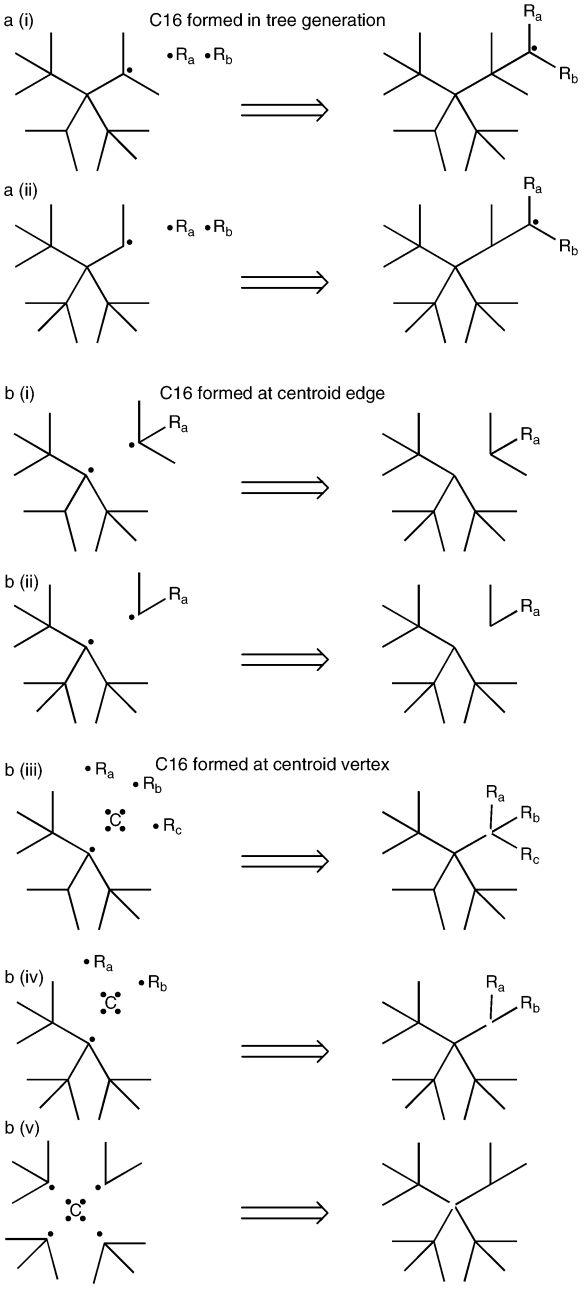

Figure 11. The ways in which the undesirable C16 substructure can arise in our chosen method of alkane isomer enumeration with a few possible scenarios.

as *n* increases, and therefore this ratio tends asymptotically to zero. That is to say, it becomes increasingly difficult to make alkanes as they get larger, since the number of unfavorable structures grows faster than does the number of favorable structures. This argument that the number of crowding-constrained structures becomes an ever smaller fraction of the total theoretical count has been made qualitatively by Klein,[10] which as is shown in Figure 14, ratios calculated by fitting to asymptotic relations are in excellent agreement with the explicitly calculated values. Hence we can use these relations to extrapolate beyond the upper limit of our data to predict that the first alkane for which there are equal numbers of allowed and disallowed structural isomers is $C_{3783}H_{7568}$ if C16 is disallowed or $C_{33\,477}H_{66\,956}$ if C17 is disallowed.

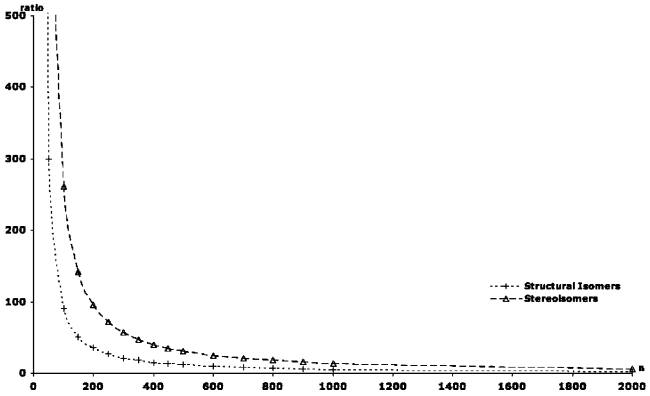

Figure 12. Plot of the ratio of thermodynamically stable to unstable alkane isomers when the C16 substructure is filtered out by our enumeration scheme. Again the ratio asymptotically approaches zero with increasing alkane size, although (as would be expected) much more rapidly than when the C17 substructure is disallowed.

	Structural Isomers		Stereoisomers	
	A	*B*	*A*	*B*
"Paper"	0.6565212295	2.8154593407	0.2868933817	3.2871196928
Allowed (C17 filtered)	0.6570958967	2.8154029983	0.2869744225	3.2871019205
Allowed (C16 filtered)	0.6622929942	2.8149549200	0.2881757332	3.2868798273

Figure 13. Values of *A* and *B* (to 10 decimal places) for our various isomer counts.

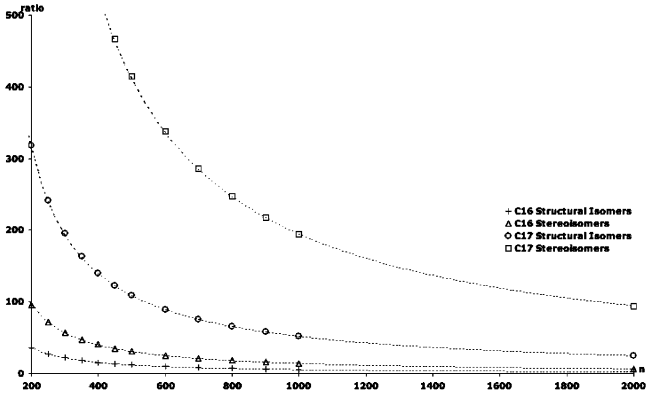

Figure 14. Graph showing the ratio of allowed to disallowed isomers, as enumerated explicitly (data points) and as calculated from fitted *A* and *B* values (curves).

5. CONCLUSION

If an acyclic alkane structure with a molecular weight above 53 kD is randomly generated, it is more likely than not that the substance cannot exist. Our enumeration strategy overestimates the odds of the successful construction of such a structure, because we consider only a C16 forbidden unit. There are many other units that will also be forbidden which do not include this substructure. For example, if the methyl groups on propane are substituted with isopropyl groups, and this process is repeated several times on the new structure, a different impossible structure will result (Figure 15). However, the smallest unstable structure is likely to be rather larger in these series. $C_{31}H_{64}$ in Figure 15 is very likely to be stable on the basis of its low strain energy, although $C_{63}H_{128}$ is unlikely to exist. Figure 14 shows that the number of structures excluded by disallowing the C16 unit is much greater than the number excluded by disallowing the C17 unit. While we cannot enumerate all impossible substructures, they must all be larger than C17 and so will exclude a smaller fraction of the possible isomers.

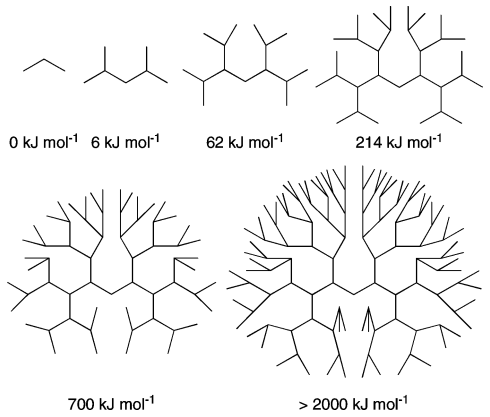

Figure 15. Isomers for which the successive addition of isopropyl groups leads to impossible structures. Strain energies of more than 300 kJ mol^{-1} are likely to be unstable.

An applet which performs the enumerations outlined in this paper, along with the Java source code, is available at http://www-jmg.ch.cam.ac.uk/data/isomercount/

These calculations demonstrate that chemical space is restricted for larger molecules. In our earlier work,[15] we reported the smallest alkane that cannot be made. We now suggest that $C_{3783}H_{7568}$ is the smallest alkane for which a randomly chosen isomer has a less than even chance of actually existing, because of the need to exclude C16 substructures. If all impossible substructures were considered, the smallest such alkane may be smaller still. The odds diminish as the number of carbon atoms increase. For $C_{20\,000}H_{40\,002}$ the odds of successfully generating an isomer are 30:1 against. For $C_{50\,000}H_{100\,002}$ the odds are less than that of being struck by lightning.[25] It might be suggested that these are moving toward molecules that are so big they cannot be made.

ACKNOWLEDGMENT

We thank the EPSRC for financial support.

Supporting Information Available: Isomer counts for $C_{1000}H_{2002}$ and $C_{2000}H_{4002}$ including all digits and the full algorithm used to enumerate isomers filtering the C16 substructure. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Crum Brown, A. On the Theory of Isomeric Compounds. *Trans. R. Soc. Edinburgh*: *Earth Sci*. **1864**, *23*, 707−719.
(2) Cayley, A. On the Mathematical Theory of Isomers. *Philos. Mag.* **1874**, *47*, 444. Cayley, A. On the Analytical Forms Called Trees with Applications to the Theory of Chemical Compounds. *Rep. Br. Assoc. Adv. Sci.* **1875**, 257−305.
(3) Henze, H.; Blair, C. M. The Number of Structurally Isomeric Alcohols of the Methanol Series. *J. Am. Chem. Soc.* **1931**, *53*, 3042−3046. Henze, H.; Blair, C. M. The Number of Structurally Isomeric Alcohols of the Methane Series. *J. Am. Chem. Soc.* **1931**, *53*, 3077−3085. Henze, H.; Blair, C. M. The Number of Structurally Isomeric Alcohols of the Ethylene Series. *J. Am. Chem. Soc.* **1933**, *55*, 680−686.
(4) Pólya, G. Combinatorial enumeration of groups, graphs and chemical compounds. *Acta Math.* **1937**, *68*, 145−254. Pólya, G. Algebraic calculation of isomers of some organic compounds. *Zeit. Kryst. A* **1936**, *93*, 414−443. Pólya, G. About the growth of the number of isomers of homologous series in organic chemistry. *Vierteljschr. Naturforsch. Ges. (Zürich)* **1936**, *81*, 243−258.
(5) For review of the devlopments in Pólya enumeration theory up to 1985, see: Pólya, G.; Read, R. C. *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*; Springer-Verlag: Berlin, 1987. A more recent review for the mathematical developments can be found in the following: Kerber, A. Enumeration Under Finite Group Action. Basic Tools, Results, and Methods. *Commun. Math. Comp. Chem. (MatCh)* **2002**, *46*, 151−198.
(6) Petit-Zeman, S. Charting Chemical Space: Finding New Tools to Explore Biology. 4th Horizon Symposium, Palazzo Arzaga, Italy, October 23−25 2005. Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.
(7) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 1504−1508.
(8) Davies, R. E.; Freyd, P. J. $C_{167}H_{336}$ Is the Smallest Alkane with More Realizable Isomers than the Observed Universe Has "Particles". *J. Chem. Educ.* **1989**, *66*, 278−281.
(9) Alder, R. W.; Maunder, C. M.; Orpen, A. G. The Conformational Effects of Quaternary Centres. *Tetrahedron Lett.* **1990**, *46*, 6717−6720.
(10) Klein, D. J. Rigorous Results for Branched Polymer Models with Excluded Volume. *J. Chem. Phys.* **1981**, *75*, 5186−5189.
(11) For a review of the importance of steric crowding in Pólya-type enumeration of alkanes, see: Klein, D. J.; Seitz, W. A. Graphs, Polymer Modes. Excluded Volume & Chemical Reality. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier, 1963; pp 430−445. For related work also, see: Kennedy, J. W.; Quintas, L. V. Extremal f-trees and embedding spaces for molecular graphs. *Disc. Appl. Math.* **1983**, *5*, 191−209.
(12) Irloff, L. D.; Mislow, K. Molecules with T Symmetry. Conformational Analysis of Systems of Type $M[C(CH_3)_3]_4$ and $M[Si(CH_3)_3]_4$ by the Empirical Force Field Method. *J. Am. Chem. Soc.* **1978**, *100*, 2121−2126.
(13) Goodman, J. M. What is the Longest Unbranched Alkane with a Linear Global Minimum Conformation? *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 876−878.
(14) Nair, N.; Goodman, J. M. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317−320.
(15) Goodman, J. M.; de Silva, K. M. N. What Is the Smallest Saturated Acyclic Alkane that Cannot Be Made? *J. Chem. Inf. Model.* **2005**, *45*, 81−87.
(16) Rüchardt, C.; Beckhaus, H.-D.; Hellman, G.; Weiner, S.; Winiker, R. Observation of a Linear Relationship between Thermal Stability and Strain Energy of Alkanes. *Angew. Chem., Int. Ed. Engl.* **1977**, *16*, 875−876.
(17) Sun Microsystems, Inc. http://java.sun.com/ (accessed Aug 20, 2007).
(18) For a review of the use of molecular graphs in chemistry, see: Balaban, A. T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 334−343. Ivanciuc, O.; Balaban, A. T. Graph Theory in Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., et al., Eds.; Wiley: 1999; Vol. 2, pp 1169−90. Ivanciuc, O.; Graph Theory in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: 2003; pp 103−138.
(19) Read, R. C. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic: New York, 1976; pp 24−61.
(20) Robinson, R. W.; Harary, F.; Balaban, A. T. The Number of Chiral and Achiral Alkanes and Monosubstituted Alkanes. *Tetrahedron* **1976**, *32*, 355−361.
(21) Otter, R. The Number of Trees. *Ann. Math.* **1948**, *49*, 583−599.
(22) Bytautas, L.; Klein, D. J. Chemical Combinatorics for Alkane-Isomer Enumeration and More. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 1063−1078. Bytautas, L.; Klein, D. J. Alkane Isomer Combinatorics: Stereostructure Enumeration and Graph-Invariant and Molecular-Property Distributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 803−818.
(23) Chang, G.; Guida, W. C.; Still, W. C. An Internal Coordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.
(24) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.
(25) U.S. National Weather Service, National Oceanic & Atmospheric Administration (NOAA). http://www.lightningsafety.noaa.gov/medical.htm (accessed Aug 20, 2007).

CI700246B