Organic and Biological Chemistry

Applications of Artificial Intelligence for Chemical Inference. IV.[1] Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra[2]

Armand Buchs,[3] A. M. Duffield, Gustav Schroll,[4] Carl Djerassi,* Allan B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg

Contribution from the Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305. *Received March 11, 1970*

Abstract: The repertoire of the computer program Heuristic DENDRAL has been extended to the solution of unknown saturated amines from their mass, and, when available, nmr spectra. From a library of 93 saturated amine mass spectra, 37 of which were supplemented by nmr data, the correct solution always appeared in the answer. The utilization of mass spectra alone resulted (Table I) in a drastic curtailment of the total search space, while if nmr data were available the final output consisted usually of one entry, the correct solution. As is apparent from Table I, the first part of the program, the PRELIMINARY INFERENCE MAKER, is capable of solving problems where the total search space far exceeds one million isomers.

Previous publications[1,5] have described the results of computer interpretation of the low resolution mass spectra of aliphatic ketones and ethers. In the case of ethers a program was added to utilize nmr data (if available). The heart of the computer program (called Heuristic DENDRAL) was the DENDRAL algorithm which constructs complete and irredundant lists of aliphatic molecules or radicals, in a linear notation, corresponding to any desired empirical formula. Our general approach to the computer interpretation of mass spectra begins with the domain of all possible structures which might *a priori* fit the experimental data. In order to expand the challenge to more complex situations we decided to approach the general solution of the mass spectra of saturated amines, since for any given number of carbon atoms, the number of possible saturated amines is considerably larger than for aliphatic ketones or ethers.[6] It should be emphasized that our purpose has been to demonstrate the feasibility of the Heuristic DENDRAL approach solely to those classes of compounds that offer new problems rather than to one functional group after another.

The basic approach to the problem of interpreting low resolution mass spectra, with the aid of nmr data if desired, is described in our earlier publication[1] dealing with saturated ethers and can be summarized in the following paragraph.

Heuristic DENDRAL is divided into three main subprograms called PRELIMINARY INFERENCE MAKER, STRUCTURE GENERATOR, and PREDICTOR. The first part finds which particular structural features are consistent with the mass spectral data and the elementary composition of the compound studied. Its output is then sent to the STRUCTURE GENERATOR which builds an irredundant and complete list of structures compatible with the information supplied by the PRELIMINARY INFERENCE MAKER and the constraints imposed by BADLIST.[1,5] Each generated structure is then given as input to the PREDICTOR which predicts significant peaks of its mass spectrum. The program either rejects the candidate or accepts it depending upon the fit of the predicted spectrum with the experimental one. Finally the accepted candidates are ranked from the most to the least plausible.[7]

The PRELIMINARY INFERENCE MAKER has now been improved by incorporating much more mass spectrometric theory about fragmentation mechanisms, and by using nmr data at an early stage. This paper will now describe how this program infers plausible substructures from mass spectra and nmr data of saturated amines. As will be shown in this paper, the efficiency achieved in the PRELIMINARY INFERENCE MAKER with this class of compounds leads to results which are in most cases, even for large molecules, precise enough such that the two other phases of Heuristic DENDRAL (STRUCTURE GENERATOR and PREDICTOR) need not be used. This represents a somewhat different application of Heuristic DENDRAL than the one which was used for

* To whom correspondence should be addressed.

(1) Part III: G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Amer. Chem. Soc.*, 91, 7440 (1969).

(2) Financial assistance from the Advanced Research Projects Agency (Contract SD-183), the National Aeronautics and Space Administration (Grant NGR-05-020-004), and the National Institutes of Health (Grants GM 11309 and AM 04257) is gratefully acknowledged.

(3) On leave of absence from the University of Geneva.

(4) Recipient of a Fulbright travel award.

(5) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Amer. Chem. Soc.*, 91, 2977 (1969).

(6) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *ibid.*, 91, 2973 (1969).

(7) An nmr PREDICTOR is also available to the user at the very end of the process. It takes as input the structures ranked by the mass spectrum PREDICTOR and the experimental nmr spectrum, provided it was completely interpreted. It also either rejects or accepts candidates in a ranked order depending on how well the predicted nmr spectrum fits the experimental data.

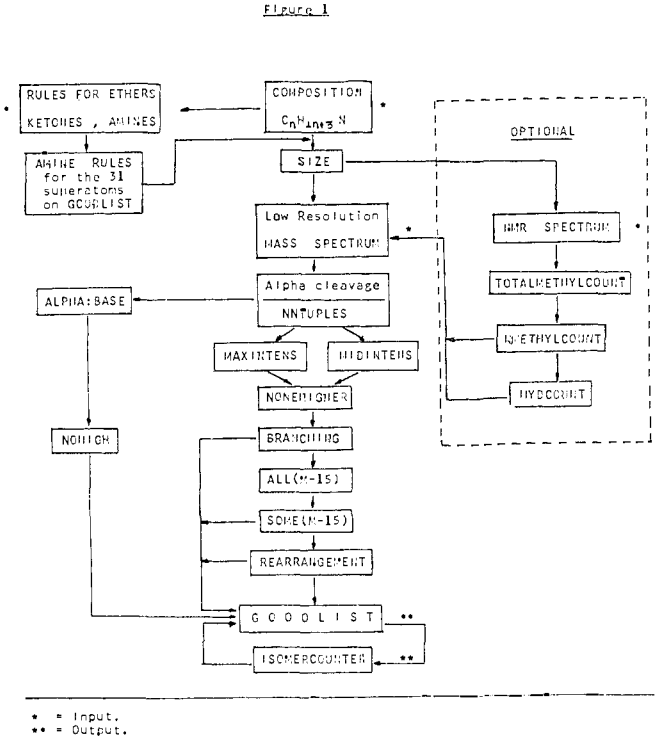
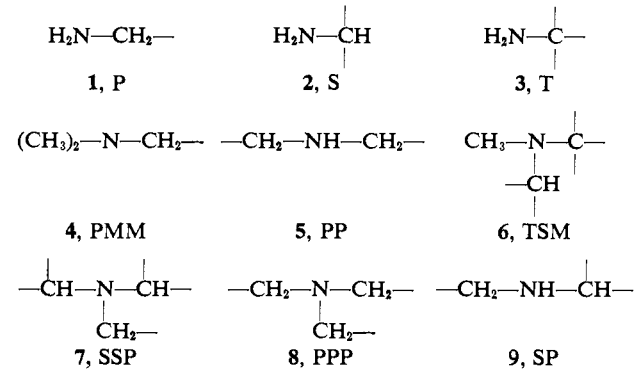


Figure 1. Sequence of the decision processes during inference phase.

ethers[1] and ketones.[5] *It should be emphasized that the approach employed previously with ketones*[5] (no complete list of superatoms and no nmr data) *or ethers*[1] (no complete list of superatoms and nmr data introduced at the end rather than the beginning of the pruning process) *would not have been sufficient to handle the problem posed by saturated amines.*

The decision processes invoked by Heuristic DENDRAL in the interpretation of amine mass spectra (supplemented by, but not dependent upon, the availability of nmr data) are schematically represented in Figure 1. If the composition of the unknown agrees with $C_nH_{2n+3}N$, the program extracts (decision "AMINE RULES," Figure 1) from memory the rules pertinent to amine mass (and nmr) spectra.

In order to approach the solution of an unknown low resolution mass spectrum of any saturated amine, it was necessary to define a complete set of possible amine subgraphs (*i.e.*, superatoms)[8] which could be inferred from the data. This is conveniently accomplished by using a combination of the four symbols T, S, P, and M for the superatom names. In this convention the number of symbols in a name refers to the number of α-carbon atoms bound to nitrogen, from one for primary amines to three for tertiary amines. The symbols themselves give the number of free valences on each α-carbon atom: P for one, S for two, and T for three free valences (see **1**, **2**, and **3**). The canonical order of the symbols is T > S > P > M. With this imposed order, PMM is the triplet chosen to represent the subgraph **4** instead of the equivalent but non-canonical names MMP and MPM. Using this nomenclature 31 superatoms[9] can be constructed from all possible combinations of the four symbols T, S, P, and M. Some additional examples of the use of this shorthand structure representation are listed below.



This notational scheme of representing the partial structure of saturated amines offers two major advantages. First, it is completely exhaustive. Thus, any saturated amine contains a subgraph which must belong to *one and only one* of the 31 superatoms. The second advantage is the ease of translation from a partial chemical structure and *vice versa*. The name of a superatom contains all the information needed for writing rules and conditions which will have to be satisfied by the data if the superatom is going to be an acceptable candidate (weight, free valences, rearrangement possibilities, etc).

In applying the proper processes to validate a particular superatom, the PRELIMINARY INFERENCE MAKER program is controlled by a table. To change the action of the program one need only change a table of superatom names and associated spectral features of molecules containing these superatoms. This way of driving the action of the program enables the user to change his mind about the properties he would expect to find in the mass spectra of compounds of any superatom class. Thus, before testing each superatom the program inquires in the property table about which processes should be used.

All 31 superatoms are initially found on a list, called GOODLIST, and each superatom is then tested for consistency with the experimental data, and either kept on GOODLIST or removed from it according to the results of the various tests. The first test is depicted in Figure 1 as "SIZE." The program compares the carbon content of the elementary composition with the number of carbon atoms required to construct the smallest molecule from the superatom by addition of only methyl groups to the free valences. If a superatom requires more carbon than is available to build the smallest molecule, that superatom is discarded from further consideration at this very early stage.

In our previous publication[1] concerning the ability of Heuristic DENDRAL to interpret low resolution mass spectra of saturated ethers, fully interpreted nmr data, if available, were used at the very end of the pruning process. This was done to test the validity of each candidate accepted by the mass spectrometric part of the program. With saturated amines, however, it was found desirable to introduce nmr data (if available) at the same time as the mass spectrum, *viz.*, in the PRE-

(8) As described in previous publications[1,5] a superatom is defined as a structural subunit having at least one free valence. In the present context only carbon atoms can be attached to the free valence(s).

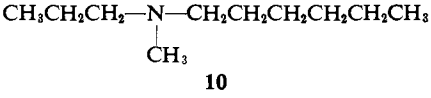
(9) In addition to these 31 superatoms there exist three more molecules, M, MM, and MMM, which translate to methylamine, dimethylamine, and trimethylamine, respectively.

LIMINARY INFERENCE MAKER program. Thus, as soon as the program has normalized the amplitudes of the peaks in the mass spectrum and removed any improbable mass points (*e.g.*, M − 4 through M − 14) it accepts the nmr data. It should be stressed that the program does not require nmr input, such that if these data are unavailable, the program bypasses the nmr subroutine and proceeds directly to examine the mass spectrum. Neither does the program require completely interpreted nmr data; it is able to accept partial information from an nmr spectrum.

The nmr spectrum can be supplied to the program in different ways; if it is easily interpreted (no doubts about the multiplicity of the different carbon-bound methyl signals and an integration curve available) the nmr spectrum is given as a list of sublists in which each sublist is composed of three elements: the chemical shift (δ),[10] the number of protons responsible for the signal (n), and one of the following symbols S, D, T, Q, or M to denote the multiplicity of the signal (singlet, doublet, triplet, quartet, or multiplet, respectively). The supplied nmr data then have the form

$$((\delta_1\ n_1\ \text{mult}_1)\cdots(\delta_i\ n_i\ \text{mult}_i)\cdots(\delta_n\ n_n\ \text{mult}_n))$$

If the multiplicity of the methyl signals in the nmr spectrum is not readily apparent, as for example in the spectrum of *N*-methyl-*N*-propyl-*n*-hexylamine (**10**), the first sublist will contain the total height of the integration curve *vs.* the height at δ = 1.1 ppm (Figure 2, run 1). The program will know that it has to calculate the number of carbon-bound methyl radicals by using the integral information and the number of hydrogen atoms in the compound (found by the program from the given empirical composition).



10

Sometimes, when no integration curve is available and the spectrum is not first order, some information can still be extracted easily from the nmr spectrum. Thus the presence of a sharp singlet in the *N*-methyl region will be used by the program and it will know that no significant information could be extracted from the *C*-methyl region of the nmr spectrum.

The program first examines the nmr spectrum to determine what kind of information was supplied and to decide how it should count the number of *C*-methyl groups. Thus the program looks first for signals at $\delta < 1.2$ ppm (either singlets, doublets, or triplets) and ensures that they originate from a number of protons exactly divisable by three. Otherwise the calculation is performed by using the two values from the integration curve and the total number of hydrogens in the compound.[11] Should neither of these two quantities be supplied, the program answers "NO INFORMATION" and no negative decision is made concerning the minimum number of methyl radicals needed to keep on GOODLIST the superatom under test.

(10) The chemical shifts are standardized against δ = 0 ppm for tetramethylsilane.

(11) If A = total height, B = height in *C*-methyl region, and N = number of hydrogens in the elementary composition, the integer value of the relation $(((B/A) \times N)/3)$ is returned.

Figure 2

```
C10H23N   *PPM*   N-methyl-n-propyl-n-hexylamine

ADJUSTED SPECTRUM = ((41 . 25)(42 . 28)(43 . 38)(44 . 63)(55 . 5)
(56 . 3)(57 . 9)(58 . 34)(70 . 4)(72 . 3)(84 . 3)(86 . 100)
(87 . 4)(128 . 28)(129 . 1)(157 . 5))

 Run 1

WAS A NMR SPECTRUM AVAILABLE                          ? YES
COULD ALL SIGNALS BE INTERPRETED                      ? NO
NMR SPECTRUM                    = ((201 54)(2.3  4  T)(2.2  3  S))
NUMBER OF CARBON-BOUND METHYLS                        = 2
NUMBER OF NITROGEN-BOUND METHYLS                      = 1
TOTAL NUMBER OF METHYLS                               = 3
MINIMUM NUMBER OF ALPHACARBON BOUND HYDROGENS = NO VALID INFORMATION

GOODLIST = PPM

MASSES OF ATTACHED RADICALS :

PPM   (71, 29)                                        1 ISOMER.

TOTAL NUMBER OF ISOMERS = 1

-------------------------------------------------------------------

 Run 2

WAS A NMR SPECTRUM AVAILABLE                          ? YES
COULD ALL SIGNALS BE INTERPRETED                      ? NO
NMR SPECTRUM                                          = ((2.2  ?  S))
NUMBER OF CARBON-BOUND METHYLS                        = NO INFORMATION
NUMBER OF NITROGEN-BOUND METHYLS                      = 1 OR 2
TOTAL NUMBER OF METHYLS                               = NO INFORMATION
MINIMUM NUMBER OF ALPHACARBON BOUND HYDROGENS = NO VALID INFORMATION

GOODLIST = SSM   TPM   PPM   SMM   TM

MASSES OF ATTACHED RADICALS :

SSM   ((15, 29)(29, 29))                              1 ISOMER.

TPM   ((15, 15, 29)(29))                              1 ISOMER.

PPM   (71, 29)                                        8 ISOMERS.

SMM   (71, 29)                                        8 ISOMERS.

TM    (71, 29, 15)                                    8 ISOMERS.

TOTAL NUMBER OF ISOMERS = 26

-------------------------------------------------------------------
```

Figure 2. PRELIMINARY INFERENCE MAKER outputs with *N*-methyl-*n*-propyl-*n*-hexylamine (**10**) as an unknown. Use of partially interpreted nmr spectrum.

The nmr program then counts the number of nitrogen-bound methyl groups by searching for a singlet signal at δ = 2–4 ppm. If the number of hydrogens responsible for the signal is exactly divisable by three, the program considers it as a signal due to nitrogen-bound methyl group(s) only after having performed some validation tests in order to ensure, for example, that the signal does not originate from 3 (or 6) hydrogens on the α-carbon atoms, all these carbon atoms being tertiary. Even if no integration curve is available, a singlet in the *N*-methyl region still means that at least one such methyl is present. In this case a question mark printed in place of the number of hydrogens causes the program to answer that there are either one or two nitrogen-bound methyl groups. The decision of keeping or rejecting a superatom will then be made on the basis of these two possibilities (see Figure 2). With our example (*N*-methyl-*n*-propyl-*n*-hexylamine) the nmr spectrum was supplied with or without an integral curve (see runs 1 and 2 of Figure 2). In the case of run 1 the program finds the correct number of *N*-methyl groups. However, when no integral curve is supplied (run 2), the presence of a sharp singlet at δ = 2.2 ppm is a clear indication of the presence of one or two *N*-methyl groups. The total number of methyl groups, provided both the number of *C*-methyls and *N*-methyls had defined values, is remembered under the name TOTALMETHYLCOUNT,[12]

Figure 3

Test *	Superatoms eliminated	Why ?
SIZE	TTT TTS	There are only 30 carbon atoms in the compound. **
TOTALMETHYLCOUNT	TTP TTM TSS TSP TSM TPP TPM TMM SSS SSP SSM SPP SMM TT TS SS TP TM	They require more methyl groups than the three which are found by the NMR subroutine.
METHYLCOUNT	PPP SP PP T S D	A N-methyl signal is in the NMR spectrum.
	PMM	Only one of the methyl groups is an N-methyl.
ALPHA:BASE	PM	m/e 44 is not the base peak.
NOHIGHS	SM	No valid multiple can be found by using the allowed α-fission peaks for SM (72, 86, 100, 114, 128 and 142) and 157 as molecular weight. m/e 100 and 114 are not in the mass spectrum, and m/e 142 cannot be used due to insufficient number of carbon atoms. The only multiple built is (72, 128) but the sum of the intensities of these two ions is less than 70%. ***

* The names of the tests refer to Figure 1.
** N-ethyl-N-propyl-N-hexylamine.
*** Rearrangement for molecular ions having SM as subunit is not a favored process.

Figure 3. Effect of the different tests upon the pruning of GOODLIST with *N*-methyl-*n*-propyl-*n*-hexylamine (see run 1 in Figure 2 for supplied data).

and the number of *N*-methyl groups under the name NMETHYLCOUNT.

Finally, the program counts the number of α-carbon bound protons, provided no nitrogen methyl signals were found. It searches the nmr spectrum for signals at $\delta > 2.2$ ppm having any multiplicity, and stores the value under the name HYDCOUNT. In our example (Figure 2), as an *N*-methyl group is already identified, the program does not search for signals originating from α-carbon hydrogens.

Having exhausted its survey of the nmr spectrum, the inference program commences its examination of the 31 superatoms on GOODLIST.

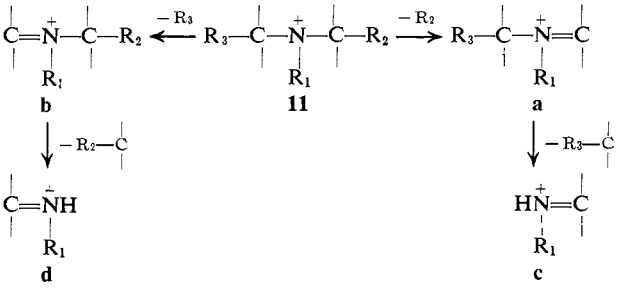
As the first necessary condition, each superatom has a number $(m_1 - 1)$ where m_1 represents the minimum number of methyl groups (1 for P, 2 for S, 4 for SS, up to 9 for TTT) which must be validated by the nmr subroutine for the superatom under test to remain on GOODLIST. The superatom passes this test (decision "TOTALMETHYLCOUNT," Figure 1) only if the number of methyl groups which the program finds exceeds $m_1 - 1$; otherwise it is deleted from GOODLIST at this very early stage, and will not be tested further. As shown in Figure 3 for *N*-methyl-*n*-propyl-*n*-hexylamine, 19 superatoms out of the remaining 29 (two were already eliminated by the test referred to as "SIZE" in Figure 1) are removed from GOODLIST by this test.

A second necessary condition, also related to the structure of each of the 31 superatoms, requires that the number m_2 of *N*-methyl groups found from the nmr spectrum must be the same as the number of M's in the superatom name. Any superatom requiring a number of nitrogen methyl groups different from that found by the program is deleted from GOODLIST. If the value of NMETHYLCOUNT is 1, for example, only superatoms with one M in their name are tested further. Should the value of NMETHYLCOUNT be partially undefined (1 or 2), only superatoms with 1 or 2 M's in the name would pass the test, and should it be totally undefined (NO INFORMATION), all superatoms would pass this test (decision "NMETHYLCOUNT," Figure 1).

If a superatom passes both these tests and has no M in its name a final nmr test is applied. For each superatom without M, a value m_3 related to its structure (2 for P, 3 for SP, etc.) represents the maximum number of α-carbon hydrogens which can be found in the spectrum in order for the superatom to be accepted for further consideration. If the value of HYDCOUNT is smaller than that of m_3, the superatom passes the test; otherwise it is removed from GOODLIST. It should be pointed out that each superatom requires an exact number of α-carbon hydrogens, but this number rapidly becomes small compared to the total number of hydrogens in the empirical formula when the size of the molecule increases. Therefore in order to avoid errors from marginal integration curves it was found safer to demand that the number found be just smaller than, or equal to, m_3.

If any of the values of TOTALMETHYLCOUNT, NMETHYLCOUNT, or HYDCOUNT are not defined, the corresponding test is passed successfully by default.

Following the nmr search the program then confronts the mass spectrum. The first condition programmed into the mass spectrometry section of the inference program relates to the well-documented[13,14] propensity of aliphatic amines to undergo α cleavage (see **11** → **a** + **b**).



For superatoms with only one free valence the first condition (decision "ALPHA:BASE," Figure 1) is that the only α-fission peak must be the base peak (respectively, 30 ($CH_2{=}N^+H_2$), 44 ($CH_2{=}N^+H{-}CH_3$), and 58 ($CH_2{=}N^+{-}(CH_3)_2$) for P, PM, and PMM), and the second condition (decision "NOHIGH," Figure 1) is that there should be no peaks with an intensity greater than 10% above the mass of one-half the molecular weight, provided *m/e* 30, 44, or 58 is not already above this limit, a fact which occurs for small molecules. This condition takes into account the possibility of β to ϵ cleavage and the rather improbable fact of finding intense peaks from cleavage occurring further away from the nitrogen atom. Since in the mass spectrum of *N*-methyl-*n*-propyl-*n*-hexylamine (**10**)

(12) The program stores parameter values under identifiable names such as this for later use.

(13) H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Mass Spectrometry of Organic Compounds," Holden-Day, San Francisco, Calif., 1967, pp 297-303.

(14) R. S. Gohlke and F. W. McLafferty, *Anal. Chem.*, 34, 1281 (1962).

m/e 44 is not the base peak (see spectrum tabulated in Figure 2), the superatom PM does not pass this test (see Figure 3).

For any other superatom to be accepted there is a definite number of α-cleavage fragments which must be located in the experimental data (decision "NTUPLES," Figure 1). The lower mass limit of the α-cleavage peaks searched for is equal to the mass of the superatom in question (referred to as "overweight") added to $((n - 1) \times 15)$ where n is the number of free valences in that superatom. For example the superatom PPP (**8**) has $n = 3$ and a mass of 56 amu, so the search will begin at mass $((2 \times 15) + 56)$, *i.e.*, mass 86. The program, to validate the presence of a PPP subgraph, must then search the mass spectrum in quest of sets of three peaks, the sum of the peaks in each set being equal to $[((n - 1) \times \text{mol wt}) + 56]$. If at least one of these sets (called ntuples)[15] is found, the α-cleavage condition is partially satisfied for that superatom. In the case of *N*-methyl-*n*-propyl-*n*-hexylamine (**10**), only SM and PPM remain as valid superatom candidates at this stage of the pruning process. Figure 3 shows why SM is eliminated by this test. For the superatom PPM the program finds (86, 128) as a valid ntuple, which translates to ($M - C_5H_{11}$, $M - C_2H_5$).

α cleavage is a major fragmentation mechanism only in those amines which cannot undergo a favored rearrangement process. This is true of amines having as subunit certain superatoms which bear this property in their α-carbon(s) and in their name. For these superatoms an ntuple will be kept only if the sum of the intensities of the α-fission peaks exceeds an empirically determined value of 70%, and only the ntuple with the highest sum of intensities[16] will be accepted (decision "MAXINTENS," Figure 1). With our example **10** the ntuple (86, 128) which is found for PPM has a sum of intensities in excess of 70% (see spectrum tabulated in Figure 2) and therefore passes the test.

With secondary and tertiary α-mono- or α-disubstituted amines (represented by superatoms with a name containing more than two symbols disregarding the M's, and with at least one S or T) α cleavage still remains a favored process but the major ion arises from a well-known rearrangement mechanism (see **11** → **a** + **b** → **c** + **d**). The ntuples are therefore tested against a less stringent condition for the total intensity of the α-fission peaks (decision "MIDINTENS," Figure 1). For these superatoms, all the ntuples with an intensity sum greater than 30% are kept for further test. If all ntuples for a superatom have been eliminated it is removed from GOODLIST.

Surviving ntuples are then submitted to a further test (decision "NONEHIGHER," Figure 1). The program requires that for each ntuple there be no intense peak at a mass higher than that of the ion with greatest m/e present in the ntuple. This rule protects the program from incorrectly identifying the highest mass α-cleavage ion, since this latter ion would be expected to have the greatest intensity of any peak found between itself and the $(M - 1)^+$ ion.[17] In the mass spectrum of **10** (tabulated in Figure 2) used to illustrate stepwise the pruning process, no peak with m/e greater than the mass of the α-fission ion $(M - C_2H_5)^+$ has an intensity above 10%. The correct superatom PPM with its associated ntuple (86, 128) is therefore not removed from GOODLIST by this test.

It is well known in mass spectrometry that at 70 eV[18] the larger alkyl group is preferentially expelled in an α-cleavage fragmentation. The program (decision "BRANCHING," Figure 1) exploits this concept to check whether it has correctly identified the ions resulting from α cleavage. In any ntuple the lowest mass value should have an intensity in excess of the next heavier α-fission fragment provided the difference in number of carbon atoms between the molecular ion and the heavier α-fission ion is less than three. Should the difference in mass between the molecular ion and the high-mass α-cleavage ion be larger than the mass of a C_2H_5 group then the program requires that the intensity of the low mass ion be in excess of $(0.5 + 0.1 \times \Delta C)$ of the intensity of the higher α-fission ion.[19] This process takes into consideration (by reducing the stringency of the condition and by even allowing the intensity of the high mass ion to be greater than that of the low mass ion) the known[13] ease of elimination in α-fission of a tertiary radical over a secondary and a secondary over a primary. In an ntuple, each α-cleavage ion is successively compared to the ion adjacent to it. The intensity of any α-cleavage ion is always normalized against the number of its occurrences on probability grounds. In the case of *N*-methyl-*n*-propyl-*n*-hexylamine (**10**), the difference between m/e 128 which is the high mass α-cleavage fragment in the ntuple (86, 128) and the molecular weight (m/e 157) represents a $C_2H_5\cdot$ radical. Since branching is impossible in an ethyl radical, the program requires that the intensity of the ion at mass 86 should be greater than that of the ion at mass 128 (see spectrum tabulated in Figure 2). However, if one considers a general molecule **12**, of molecular weight 185, the correct

$$C_4H_9—CH_2—NH—CH_2—C_6H_{13}$$
12

ntuple (100, 128) for the superatom PP (**5**), *i.e.*, ($M - C_6H_{13}$, $M - C_4H_9$), would be accepted even if the intensity of ion m/e 100 is less than that of ion m/e 128. The $C_4H_9\cdot$ group could be tertiary, secondary, or primary. In this case the program would allow the

(15) The term "ntuple" is used to refer to any set of possible α-cleavage peaks for a superatom in the context of a particular mass spectrum. For example, for the superatom SSP (7) and an amine having a molecular weight of 171 ($C_{11}H_{25}N$), the following sets would be built as *a priori* valid ntuples: $n = 5$; overweight = 54; sum of peaks $= (4 \times 171) + 54 = 738$; m/e of lowest α-fission peak possible $= (4 \times 15) + 54 = 114$; possible ntuples: (114, 156, 156, 156, 156), (128, 142, 156, 156, 156), and (142, 142, 142, 156, 156) (m/e 114 corresponds to $M - C_4H_9$, m/e 128 to $M - C_3H_7$, m/e 142 to $M - C_2H_5$, and m/e 156 to $M - CH_3$).

(16) All intensity values refer to relative abundances with intensity of the base peak = 100%. All threshold values were chosen on theoretical trends and corrected so they never eliminate the correct superatom but still give the maximum pruning effect.

(17) The $(M - 1)^+$ ion is not considered as an α-fission peak; the $(M - 15)^+$ ion, even if not present in the spectrum, is allowed to be used for building ntuples, provided its mass in conjunction with the masses of the other peaks in the ntuple satisfies the equation used for this purpose.

(18) This is not the case, however, at low ionizing voltage. See C. A. Brown, A. M. Duffield, and C. Djerassi, *Org. Mass. Spectrosc.*, **2**, 625 (1969).

(19) If this difference is C_3 or greater the possibility exists that the α-cleavage ion of higher mass can result from expulsion of a secondary or even tertiary radical. If this is possible, the program must weaken this condition such that a lower mass ion can be less intense than the higher mass ion. However, if the difference of mass between the highest mass α-cleavage ion found and the molecular weight is less than the mass of a C_3 unit, no possibility of branching exists. The expression $(0.5 + 0.1 \times \Delta C)$ was arrived at empirically.

Figure 4

```
C8H19N   *SP*      Ethyl-1,3-dimethylbutylamine

ADJUSTED SPECTRUM = ((15 . 4)(16 . 1)(17 . 1)(18 . 5)(27 . 13)
(28 . 6)(29 . 9)(30 . 8)(31 . 1)(41 . 9)(42 . 7)(43 . 9)(44 . 36)
(45 . 3)(46 . 1)(55 . 1)(56 . 3)(57 . 1)(58 . 6)(59 . 1)(69 . 1)
(70 . 1)(71 . 1)(72 . 100)(73 . 1)(83 . 1)(84 . 1)(85 . 1)
(86 . 1)(87 . 1)(98 . 1)(112 . 1)(114 . 6)(128 . 1)(129 . 1))

WAS A NMR SPECTRUM AVAILABLE                    ? NO

GOODLIST = PPM    SMM    TM    SP

MASSES OF ATTACHED RADICALS :

PPM   (57, 15)                                  4 ISOMERS.

SMM   (57, 15)                                  4 ISOMERS.

SP    ((15)(15, 57)))                           4 ISOMERS.

TM    (57, 15, 15)                              4 ISOMERS.

TOTAL NUMBER OF ISOMERS = 16
```

Figure 4. PRELIMINARY INFERENCE MAKER output with ethyl-1,4-dimethylbutylamine (**14**) as an unknown.

intensity of the ion at mass 100 to be less than that of the ion at mass 128. It would require that the abundance of *m/e* 100 must be at least equal to (0.5 + 0.1 × 2), *i.e.*, 0.7 times the abundance of *m/e* 128.

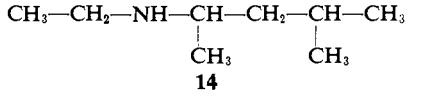
If any superatom still exists as a viable candidate to explain the experimental data and should its ntuple consist entirely of $(M - 15)^+$ α-fission ions, then it is subjected to another decision process (decision "ALL (M − 15)," Figure 1). The intensity of the $(M - 15)^+$ ion should in this condition be equal to or greater than $(1 - 1/n) \times 50$, where n is the number of times the mass of the $(M - 15)^+$ ion is present in the ntuple or number of possible α cleavages leading to an $(M - 15)^+$ ion. For example, in the case of diisopropylamine (**13**) the correct superatom SS would only be

$$(CH_3)_2—CH—NH—CH—(CH_3)_2$$
13

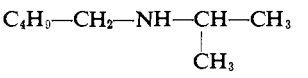
accepted if the $(M - 15)^+$ α-cleavage ion has an intensity greater than $(1 - 1/4) \times 50$, *i.e.*, greater than 38% relative abundance. If besides some $(M - 15)^+$ α fissions an ntuple contains at least one ion of lower mass (loss of a group having more than one carbon atom) the condition about the intensity of the $(M - 15)^+$ ion is much less demanding (decision "SOME (M − 15)," Figure 1). This intensity needs only to be equal to or greater than the number of (M − 15) α fissions minus one. This allows the $(M - 15)^+$ ion to be absent from the spectrum if only one α cleavage can lead to it.

After this test, superatoms with no rearrangement possibility or for which rearrangement can give rise to ions of only moderate intensity, and which still have at least one surviving ntuple, are accepted as plausible candidates. The masses of the alkyl fragments which should be attached to the free valences are then calculated by simply subtracting the mass of each α-fission peak in the ntuple from the molecular weight. These masses we refer to as "partition" because they are the masses of the alkyl groups which must be attached to the free valences of the superatom. The superatom with its list of partition is then examined by a subroutine which calculates the number of isomers compatible with both the structure of the superatom and the masses of the alkyl groups which have to be attached to this superatom (decision "ISOMERCOUNTER," Figure 1). With our example **10**, the correctly identified superatom PPM with its ntuple (86, 128) is definitely accepted at this stage of the pruning process. The program subtracts then both masses 86 and 128 from the molecular weight 157, translating the ntuple (86, 128) to the list of partition (29, 71). When no integral curve is supplied for the nmr spectrum, eight isomers can be constructed from the superatom PPM with (29, 71) as list of partition (see run 2 in Figure 2). However, when the nmr spectrum is supplemented by an integral curve, only one isomer is compatible with the information given by the PRELIMINARY INFERENCE MAKER program (see run 1 in Figure 2).

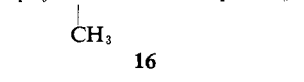
Superatoms for which rearrangement is a major process are further tested for the presence of at least one intense ion arising from rearrangement, with *m/e* in accordance with the structure of the superatom and its set of partition(s). The intensity of the rearrangement peak should be greater than 30% if the superatom is to be kept on GOODLIST. The subroutine which handles the rearrangement mechanism is programmed so that it takes into account all rearrangement possibilities for the superatom and the partitions under test. This allows the program to assign the correct alkyl fragments to each different α carbon. For example, in the case of ethyl-1,3-dimethylbutylamine (**14**), a molecule with an SP subgraph (**8**), the correctly inferred superatom will be assigned the following list of partition: "(15, 15, 57)." The alkyl groups which have to be attached to the three free valences of the SP superatom are two methyls and a C_4H_9 radical. This can be done in two ways (structures **15** and **16**).



14



15



16

With both structures **15** and **16** rearrangement ions are expected at *m/e* 30 ($CH_2{=}N^+H_2$) and 44 ($CH_3—N^+H{=}CH_2$). But, if as generally accepted, the most favored rearrangement is the one where the C–N bond is broken and a hydrogen transferred to the nitrogen atom after expulsion by α cleavage of the heavier substituent, one can postulate that for the second structure (**16**), which is the correct one, the rearrangement leading to *m/e* 44 should be favored. With the first structure (**14**) the ion of *m/e* 30 would be expected to give a more intense signal than the ion of *m/e* 44. The intensities of *m/e* 30 and 44 in the mass spectrum of **14** are, respectively, 8 and 36%. So the program chooses the second structure and, instead of simply giving "SP (15, 15, 57)" as superatom and partition, it assigns the correct distribution of the alkyl fragments between the two different α-carbon atoms by answering "SP ((15, 57) (15))." Figure 4 shows the output of the PRELIMINARY INFERENCE MAKER program with ethyl-1,3-

Table I. Curtailment of the Search Space by Heuristic DENDRAL

Amine	No. of amine isomers	Mass spectra alone: No. of superatoms on GOODLIST	Mass spectra alone: No. of possible isomers	Mass spectra + nmr spectra: No. of superatoms on GOODLIST	Mass spectra + nmr spectra: No. of possible isomers
3-Hexyl	39	1	2	1	1
1,3-Dimethylbutyl	39	2	8	1	2
2,2-Dimethyl-3-butyl	39	2	8	1	1
N-Methylethyl-*n*-propyl	39	8	15	2	2
N,*N*-Dimethyl-2-butyl	39	6	6	1	1
2-Heptyl	89	2	16	1	1
n-Propyl-*n*-butyl	89	5	10	1	1
1,3-Dimethylpentyl	89	2	16	1	4
1,5-Dimethylhexyl	211	2	34	1	9
N,*N*-Dimethyl-3-hexyl	211	3	6	1	1
N-Methyl-*n*-butylisopropyl	211	10	31	1	1
Diisopropylethyl	211	6	18	1	1
N-Methyl-*n*-propyl-*n*-butyl	211	5	24	1	1
n-Propyl-*n*-hexyl	507	7	42	1	1
3,3,5-Trimethylhexyl	507	1	89	*a*	
N-Methylisopropyl-*n*-amyl	507	5	20		
N-Methyl-*n*-propyl-*n*-hexyl	1,238	10	46	1	1
N,*N*-Dimethyl-3-octyl	1,238	8	36	1	1
n-Butyl-*n*-hexyl	1,238	3	48	1	1
N,*N*-Dimethyl-2-ethylhexyl	1,238	4	156		
n-Amyl-*n*-hexyl	3,057	9	112		
Tri-*n*-butyl	7,639	2	8		
Di-*n*-heptyl	48,865	6	510		
Triisoamyl	124,906	2	40	1	9
N-Methyl-8-hexadecyl	321,198	1	3,471		
n-Octyl-*n*-nonyl	830,219	2	6,942	1	1
N,*N*-Dimethyl-8-hexadecyl	12,156,010	4	14,418	1	1
Tri-*n*-hexyl	12,156,010	2	240	1	1
Tri-*n*-heptyl	38,649,142	2	1,938		

[a] Blanks in columns 5 and 6 indicate that no nmr spectrum was available.

dimethylbutylamine (**14**) as an unknown when only the mass spectrum is supplied.

The rearrangement process is the last test (decision "REARRANGEMENT," Figure 1) for those superatoms for which rearrangement is a favored mechanism. Each accepted candidate along with its correctly assigned list of partition(s) is then sent to the isomer counter subroutine which calculates the number of compatible isomers by taking into account the assignment of the different alkyl groups to specific positions.

Two different outputs for *N*-methyl-*n*-propyl-*n*-hexylamine (**10**) are reported in Figure 2. In the first case (run 1) the supplied nmr spectrum contains enough information to allow the program to calculate the number of *C*-methyl groups; the user made no decision about the multiplicity of the *C*-methyl signals. As shown, only the correct superatom remains on GOODLIST. The search space is curtailed from 1238 possible isomers for $C_{10}H_{23}N$ (see Table I) to one structure.

Then (Figure 2, run 2) it was assumed that no integration curve was available. Clearly, the only straightforward information from the nmr spectrum is a sharp singlet at $\delta = 2.2$ ppm. The output reflects this fact by showing additional superatoms on GOODLIST, which could not be eliminated for their required number of methyl groups, like TPM, TM, and SSM, or on the basis of the number of *N*-methyl groups needed, like SMM. The information is nevertheless sufficient to eliminate all superatoms without M's in their name.

It is interesting to note that with the aid of nmr data GOODLIST is pruned mainly by the nmr tests (decisions "TOTALMETHYLCOUNT," "NMETHYLCOUNT," and "HYDCOUNT," Figure 1). Figure 3 shows that already after the second nmr test (decision "NMETHYLCOUNT," Figure 1) only the three superatoms PM, SM, and PMM remain on GOODLIST in the case of run 1 (Figure 2). *Inserting the nmr tests at the beginning of the process is therefore very efficient in saving time.*

Even for rather large molecules and without the aid of nmr data, the PRELIMINARY INFERENCE MAKER program is able to curtail the search space in quite an impressive way. As can be seen from the outputs reported in Figure 5, only the correct superatom is inferred from the mass spectrum of *N*-methyl-8-hexadecylamine. The search space is reduced by a factor of 92 (see Table I). This factor is even larger in the case of tri-*n*-heptylamine; from 38, 649, 142 *a priori* possible amine isomers for $C_{21}H_{45}N$, the isomer counter subroutine finds 1938 structures compatible with the output of the inference phase, *i.e.*, a reduction factor of nearly 20,000.[20]

The program has been successfully tested with 93 amines;[21] for 37 of them nmr data were available. With this set of examples the program always selects the correct answer in the final output.

Some results are reported in Table I. In every case when an nmr spectrum is used only the correct superatom is found on GOODLIST and with the exception of five cases out of 37, only one structure is compatible with the output. The five exceptions all include the correct compound.

(20) For both the above mentioned compounds the problem would have been completely solved with the use of nmr data to supplement the mass spectrum.

(21) Sixty-six mass spectra were taken from McLafferty's excellent publication on amines.[14]

Figure 5

```
C16H35N    *S*    6-hexadecylamine.

ADJUSTED SPECTRUM = ((41 . 23)(42 . 7)(43 . 31)(44 . 7)(53 . 2)
(54 . 2)(55 . 14)(56 . 24)(57 . 10)(58 . 1)(67 . 2)(68 . 1)
(69 . 9)(70 . 4)(71 . 1)(81 . 1)(82 . 1)(83 . 3)(84 . 2)(85 . 1)
(97 . 2)(98 . 2)(101 . 1)(102 . 1)(126 . 1)(128 . 100)(129 . 3)
(130 . 1)(140 . 1)(141 . 94)(142 . 4)(143 . 1)(155 . 1)(157 . 1)
(173 . 1)(240 . 1)(241 . 1))

WAS A NMR SPECTRUM AVAILABLE                    ? NO

GOODLIST = S

MASSES OF ATTACHED RADICALS :

S      (113, 99)                          3471 ISOMERS.

TOTAL NUMBER OF ISOMERS = 3471

-------------------------------------------------------------

C21H45N    *PPP*    Tri-n-heptylamine.

ADJUSTED SPECTRUM = ((15 . 1)(16 . 1)(17 . 1)(18 . 2)(27 . 1)
(28 . 5)(29 . 3)(30 . 8)(31 . 1)(41 . 5)(42 . 2)(43 . 6)(44 . 9)
(45 . 1)(55 . 3)(56 . 2)(57 . 6)(58 . 4)(59 . 2)(60 . 1)(69 . 1)
(70 . 1)(71 . 1)(72 . 1)(75 . 1)(83 . 1)(84 . 2)(85 . 1)(86 . 1)
(87 . 1)(98 . 3)(99 . 1)(100 . 1)(101 . 1)(112 . 2)(113 . 1)
(114 . 1)(115 . 1)(126 . 2)(127 . 1)(128 . 7)(129 . 1)(140 . 2)
(141 . 1)(142 . 11)(143 . 1)(154 . 1)(155 . 1)(156 . 1)(183 . 1)
(169 . 1)(170 . 1)(183 . 1)(184 . 1)(185 . 1)(197 . 1)(198 . 1)
(212 . 1)(213 . 1)(225 . 1)(226 . 100)(227 . 3)(240 . 1)(253 . 1)
(254 . 1)(268 . 1)(269 . 1)(282 . 1)(283 . 1)(296 . 1)(309 . 1)
(310 . 2)(311 . 2))

WAS A NMR SPECTRUM AVAILABLE                    ? NO

GOODLIST = PPP    TMM

MASSES OF ATTACHED RADICALS :

PPP    (85, 85, 85)                       959 ISOMERS.

TMM    (85, 85, 85)                       969 ISOMERS.

TOTAL NUMBER OF ISOMERS = 1938
```

Figure 5. PRELIMINARY INFERENCE MAKER outputs for *N*-methyl-8-hexadecyl- and tri-*n*-heptylamines when only the mass spectra were supplied.

It can be concluded that with the aid of nmr data only the first subprogram of Heuristic DENDRAL needs to be used. In fact, the PREDICTOR program is unable at the present stage to choose the correct molecule among candidates containing the same superatom as a subunit along with the same weight for the radicals attached to the α-carbon(s). This is not a surprising fact; mass spectrometry just does not give much information about structural features located too far away from the charge center. When no nmr data are used the search space is still greatly reduced as can be seen from the results recorded in Table I.

Clearly, more information could easily be extracted from the nmr spectra by the PRELIMINARY INFERENCE MAKER program. It would be possible to use the multiplicity of the *C*-methyl signals and not use the nmr spectrum only as a methyl counter (and occasionally as an α-carbon hydrogen counter). This would nevertheless require that the STRUCTURE GENERATOR program be able to accept overlapping information, a fact it cannot handle for the time being.

The results which have so far been obtained are sufficiently promising so as to stimulate further research on more general and complex problems. It remains to be seen, however, whether the heuristics can be made sufficiently precise for other types of organic molecules such that the present degree of efficiency obtained with amines can be maintained.

Experimental Section

The computer program described here is part of the complete Heuristic DENDRAL program. It runs on the IBM 360/67 computer at the Stanford Computation Center using the LISP programming language. Without nmr data, the computer needs 4.26 min to interpret 93 mass spectra. When nmr data are used the process is about 30% faster. Mass spectra which had not been reported in the literature were recorded in our laboratory,[22] some with a Varian MAT CH-4 mass spectrometer, others with an AEI MS-9 mass spectrometer. The majority (26 out of 37) of the nmr spectra used was recorded in our laboratory by Dr. L. J. Durham of Stanford University.

(22) We thank Mr. R. G. Ross and Mr. R. T. Conover for recording the mass spectra.