The Critical Evaluation of a Comprehensive Mass Spectral Library

P. Ausloos, C. L. Clifton, S. G. Lias, A. I. Mikaya, S. E. Stein, and
D. V. Tchekhovskoi
NIST Mass Spectrometry Data Center, Gaithersburg, Maryland, USA

O. D. Sparkman
Sparkman and Associates, Antioch, California, USA

V. Zaikin
Topchiev Institute of Petrochemical Synthesis, Moscow, Russia

Damo Zhu
Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China

A description of the methods used to build a high quality, comprehensive reference library of electron-ionization mass spectra is presented. Emphasis is placed on the most challenging part of this project—the improvement of quality by expert evaluation. The methods employed for this task were developed over the course of a spectrum-by-spectrum review of a library containing well over 100,000 spectra. Although the effectiveness of this quality improvement task depended critically on the expertise of the evaluators, a number of guidelines are discussed which were found to be effective in performing this onerous and often subjective task. A number of specific examples of the particularly challenging task of spectrum editing are given. (J Am Soc Mass Spectrom 1999, 10, 287–299) © 1999 American Society for Mass Spectrometry

Prediction of the electron-ionization mass spectrum of any but the simplest molecules from first principles is generally not possible, mainly because of the complexity of the processes occurring when a molecule dissociates after being ionized by high energy electrons. For instance, complex ions can dissociate through a series of consecutive and competitive pathways, often allowing multiple paths to a single observed ion. Furthermore, these ions may rearrange before dissociation, so that an observed fragment ion may not be assigned with confidence to a distinct structural unit in the original molecule. Another problem is that relative peak abundances depend on differences in dissociation rates that cannot be predicted to a useful level of accuracy.

For these reasons, mass spectra cannot be reliably predicted, and spectra of compounds to be identified ("unknowns") are often treated simply as molecular fingerprints, for which compounds having similar spectra in a reference library are found in a "library search" and arranged in order of similarity to the unknown spectrum in a "hit list." The success of this method as a reliable aid for compound identification depends on the availability of comprehensive libraries of relevant, high quality reference spectra. While early practitioners of electron-ionization mass spectrometry often built their own specialized collections of reference spectra for comparison to unknown spectra "by hand," the increasing power and availability of computers has led to the development of large, general-purpose collections of mass spectra [1–3] and associated library search software. These systems are now widely available as an integral part of mass spectrometer data systems. The largest collections [1, 2] are composed of spectra originating from a variety of sources, including donations, purchases, or acquisition from the literature. This diversity of sources inevitably means that the spectra will be of variable quality, even when the nominal instrument operating conditions (70-eV electrons, for example) are the same.

While at one time it was hoped that computer methods might be able to play a central role in spectral quality control [4–6], serious defects in this approach were pointed out by Domokos et al. [7], and the inability of these methods to detect certain classes of

Address reprint requests to S. E. Stein, NIST Mass Spectrometry Data Center, A111/221, NIST, Gaithersburg, MD 20899.

1044-0305/99/$20.00
PII S1044-0305(98)00159-7

Received October 5, 1998
Revised December 8, 1998
Accepted December 8, 1998

serious errors recognizable by experts was documented by Zhu et al. [8]. Other studies demonstrated the inadequacy of automated methods for selecting the best of replicate spectra for the same compound [9]. Consequently, about 10 years ago a manual evaluation effort was begun with the objective of removing significant, identifiable errors from the NIST/EPA/NIH Mass Spectral Library. A comparative discussion of approaches for quality measurement and control efforts in two comprehensive libraries appeared shortly after this evaluation program began [10, 11]. The principal purpose of the present paper is to document the methods developed and applied for this large-scale evaluation project, leading to a partially evaluated version of the Library in 1992 and to a fully evaluated version in January of 1998 (NIST 98). Earlier versions of this library were released under the names NBS/EPA/MSDC Mass Spectral Database (1988), and, originally, the EPA/NIH Mass Spectral Database (1978). Each of the 129,136 electron-ionization mass spectra in NIST 98 has been evaluated according to the methods described here, by team of mass spectrometrists who are knowledgeable about ionic fragmentation processes. A large literature describes ion fragmentation processes [12], which will not be specifically covered in this article.

The need for such an evaluation has become increasingly evident as the fraction of GC/MS analysts with expertise in the application of the rules of mass spectral fragmentation has declined, partly as a result of the increasingly routine use of GC/MS. With this trend, deficiencies in library spectra are less likely to be detected by the analyst, a trend likely to continue with an increasing reliance on automated chemical identification methods. In addition to the obvious benefit to users of having higher quality reference spectra to match their spectra, a knowledge that each spectrum has been accepted by an evaluator is expected to generally enhance the confidence in results obtained by library searching.

Discussion

Sources of New Spectra

NIST 98 has 129,136 electron-ionization mass spectra of 107,886 compounds, derived from a larger collection of 175,510 spectra. The previous version (1992) of the NIST/EPA/NIH Mass Spectral library contained some 75,000 spectra of 62,350 compounds; the new release contains 55,000 additional spectra. These come primarily from other spectral collections and spectra determined specifically for the library. Other collections that have been incorporated into NIST 98, most with a specialized focus, are given in Table 1.

Spectra determined especially for the library at NIST (3110 spectra) and at NIH (9510 spectra) provided another important source of new data for NIST 98. This dedicated effort, begun eight years ago, has as its goal the determination of spectra of "useful" compounds not yet represented in the library, and also the acquisition of new spectra of compounds currently represented in the existing collection by spectra of questionable quality. "Useful" compounds are defined as compounds for which there is reason to believe that the spectrum would be of direct interest to many library users. Measurements at NIST were devoted primarily to commercially available compounds and those measured at NIH were for compounds of significance to medicinal chemistry. A practical measure of the significance of specific compounds was its presence in other indexes of compounds (Table 2).

Table 1. Data collections incorporated into NIST 98

	# spectra
Chemical Concepts[a]	31,613
ASES/MS Database, Dalian Institute[b]	4789
TNO Volatile Compounds in Food[c]	1233
Georgia[d] and Virginia[e] Crime Laboratories	1091
AAFS Toxicology Section, Drug Library[f]	835
VERIFIN[g] and CBDCOM[h] Chemical Weapons	545
Association of Official Racing Chemists[i]	186
St. Louis University Metabolic Disease[j]	131

[a] Chemical Concepts Quality Collection, Chemical Concepts GmbH, Weinheim, Germany (composed primarily of Professor Dieter Henneberg's industrial chemical collection).
[b] Library contained in the Automated Structure Elucidation System, Dalian, China.
[c] 2nd ed., TNO Food Research, Zeist, NL, 1996.
[d] Patti Price, Georgia Bureau of Investigation, Decatur, GA.
[e] Virginia Division of Forensic Science, Richmond, VA.
[f] Comprehensive Drug Library, Mass Spectrometry Database Committee, American Association of Forensic Chemists, 1997.
[g] VERIFIN Methodology Publications, Ministry of Foreign Affairs, Helsinki, Finland.
[h] Chemical and Biological Defense Command, ERDEC, Edgewood, MD.
[i] David Leung, AORC Drug Library, Association of Official Racing Chemists, Hong Kong.
[j] Dr. James Shoemaker, Metabolic Screening Laboratory, St. Louis University School of Medicine, St. Louis, MO.

Other sources of significant compounds were simply the catalogues of the major commercial suppliers of chemicals. The eventual goal is to obtain as many spectra as possible for compounds on these lists. For a compound on multiple lists, the goal is to acquire at least two independent spectra to further assure accuracy for these more significant compounds.

Few spectra were added from the largest source of available mass spectral information, the scientific literature. The vast majority of these spectra include a small number of peaks selected by the author to confirm the identity of a newly synthesized compound. In addition to the generally low significance of the compounds, these partial spectra contain insufficient information for compound identification by spectrum matching.

Evaluation Background

As long ago as 1971, when Klaus Biemann and coworkers [13] published a pioneering paper on computer searching of mass spectral collections, it was pointed out that, for maximum effectiveness in matching spectra of unknown compounds, comprehensive computer-

Table 2. Overlap in CAS registry numbers (CASRNs) between chemical indexes/collections and NIST 98

Chemical Index	# CASRNs in index	% in NIST 98	% change from 1992 version
NIST/EPA/NIH Library	69,061	100.0	33
EPA Environmental Monitoring Methods Index	1640	67.6	2
Commercially Available Fine Chemical Index	26,129	47.9	21
CRC Handbook of Data of Organic Compounds	25,584	45.4	11
NIH-NCI Inventory File	32,866	37.1	10
U.S. Pharmacopoeia/U.S.A.N. (USP)	6311	19.6	19
Toxic Substances Control Act Inventory	44,098	19.5	12
European Index of Industrial Chemical Substances	80,216	19.3	18
Registry of Toxic Effects of Chemical Substances	80,822	10.0	13

Indexes current in 1992 were used in this comparison.

searchable databases should be composed of high quality, complete spectra. Given the large size and heterogeneity of current comprehensive mass spectral libraries, insuring the quality of each spectrum in the collection is a major challenge. In 1988, after articles [7, 8] appeared in the literature pointing out errors in the version of the NIST library then available, the quality control procedures in use at that time were examined in detail [9].

The pre-1988 library was built of spectra, each of a unique compound, which were selected by computer from a larger archival collection in a fully automated selection process. Each spectrum selected had to be associated with a Chemical Abstracts Service Registry Number (CASRN); for every unique CASRN in the archive, one spectrum appeared in the library, regardless of its quality. When there were two or more spectra of the same compound, the computerized selection process was based on a so-called “Quality Index,” an algorithm designed to detect errors in mass spectra, initially proposed and implemented by Speck, Venkataraghavan, and McLafferty [4], in 1978. A modified Quality Index was adopted for quality control of the EPA/NIH Mass Spectral Data Base (the forerunner of the current NIST library) [5, 6]. The algorithm, among other things, checked for obvious errors such as peaks above the highest permissible molecular ion peak, and penalized spectra for having “illogical” neutral losses or too few peaks. The value of the Quality Index was provided with each spectrum, with the expectation that a poor or erroneous spectrum would be recognized by library users through its low numerical grade.

In 1988, the effectiveness of the Quality Index for selecting the best spectrum among several “replicate” spectra for a single compound was tested. After it was demonstrated that the Quality Index selected the better or best among replicates [9] only 50% of the time, the onerous task of evaluating each and every mass spectrum (as well as compound name, and structure) in the NIST collection was initiated. This action was obviously necessary, not only for selecting the best of among replicate spectra, but even more, for finding serious errors in spectra and taking corrective action (deletion or editing). The fundamental problem with the use of a Quality Index approach to quality control was the inability to apply fragmentation rules other than those based solely on chemical formula.

Chemical Structures

Knowledge of the chemical structure of a compound is, of course, a prerequisite for the evaluation of its mass spectrum. At the start of the evaluation process, chemical structural drawings were available only for compounds that appeared in the 1978–1982 versions of the library, and these were inconveniently located in a multivolume collection of books [14]. Therefore, an effort was made to obtain digital representations of chemical structures for all compounds in the library. The building of an auxiliary collection of molecular structure information began with the conversion to two-dimensional drawings of approximately 32,000 “connection tables” that had been used in an earlier on-line version of the library [15]. The remainder of the structures was either acquired along with the spectra of the compound, or drawn by structure-entry personnel with the assistance of nomenclature experts. Structures were drawn using commercial drawing software and converted to a format compatible with NIST chemical structure analysis software.

The availability of these chemical structures had many benefits for the evaluation process. For example, because the files could be organized and searched by structure, certain tasks could be automated:

(1) Chemical “registration” by structure, rather than Chemical Abstracts Service Registry Number (CASRN), allowed dropping the requirement that each compound to appear in the library must be associated with a CASRN. This permitted the inclusion of good quality spectra for thousands of additional compounds.

(2) With the assistance of structure-matching software, it became possible to readily identify replicate spectra not associated with a CASRN, thereby facilitating the examination of replicates with the goal of selecting the “best” spectrum among those available.

(3) Spectra for stereoisomers having different CASRNs could be identified by computer methods and compared by evaluators. Stereoisomers often possess nearly identical mass spectra.

Certain types of errors, especially those in names and formulas, could be located using the structure files:

(1) With the addition of structural information to the archival files, checking the consistency between formula, molecular weight, chemical structure, and chemical name became an important means of detecting errors in the chemical identification information. In this way, many hundreds of errors in these auxiliary data associated with the spectra have been found and corrected.

(2) Compounds with structures that could not be drawn because of problems with the names provided (ambiguous, "impossible," or unrecognizable trivial names) were discovered, leading to the removal of the spectra for such ill-characterized compounds from the archival collection.

The existence of the complete file of chemical structures allows the use of a wide range of chemical processing software for the purpose of substructure identification and substructure searching.

Finally, because the structural drawings were available in a computer format, they could be displayed in computer versions of the library distributed to the public. Library users most often find that a display of chemical structures is far easier to interpret than a list of names, particularly when dealing with complex organic compounds.

Chemical Names

The primary identifier of the compound associated with a spectrum is the chemical name as provided by the laboratory that determined the spectrum. The structure drawings are based upon these names. Understandably, for a library built up over so many decades, and with spectra originating from so many sources, the names assigned to the subject compounds are not systematic, and include common names, commercial names, IUPAC names, Chemical Abstracts names, and many other variations. When carrying out a search of the library by chemical name, there is no way to know which of the many variations on a particular name may or may not appear in the "Names" file of the library. For the most common compounds, almost any recognizable name will suffice for a search (e.g., the spectrum of methane can be retrieved by entering "marsh gas"), but for less common compounds there is a high probability that only a single name—that provided with the spectrum when it was originally added to the library—will be found.

At the time the library was managed by EPA, an attempt was made to collect as many alternative names as possible for compounds. This policy has been continued at NIST, although to date no dedicated effort has been made to provide internally consistent, systematic names for all compounds. Some effort has been expended on selecting the best (most readily recognizable) name for compounds with multiple names, although this task is not yet complete. Also, for salts which yield spectra of the corresponding "free-base" under electron ionization, efforts have been made to change the original name (and CASRN) to that of the "free-base" compound.

All names provided with the library are maintained in a separate database in which each name is associated with a CASRN or spectrum identification number (when a CASRN is unavailable) as well as a chemical formula. Errors in CASRNs were identified by a mismatch in the CAS checksum value and by inconsistency with the chemical formula associated with that CASRN in NIST internal files. Use of this library ensures that each replicate spectrum will be associated with the same chemical identification information.

Evaluation of the NIST Mass Spectral Library

Because effective computer methods for finding inconsistencies between the mass spectrum and structure of a compound are not available, evaluation was done in the traditional manner, with experienced mass spectrometrists examining each spectrum. The task was facilitated using automated procedures designed to locate spectra containing possible errors, to compute individual quality index factors for a spectrum, and to locate similar spectra. However, all actual decisions to retain, edit, or delete a spectrum have been based on detailed evaluations by scientists. Of the eight full-time or part-time evaluators who have worked on this project since 1988, the names of six appear among the list of authors of this article (Ausloos, Clifton, Lias, Mikaya, Zaikin, and Zhu). Others associated with the evaluation early on were Stephen Down (Royal Society of Chemistry, and later Downstream Data, in England), and H. Zohdi of the University of Cairo. The evaluation procedure described below draws extensively on their experience.

The evaluation procedure was designed to be as objective as possible and to concentrate on a detailed examination of each individual spectrum as well as on the collection of replicate spectra when available. To the best of our knowledge, this is the first report of any similarly structured program to critically evaluate mass spectral libraries.

Because of the unavoidable subjectivity in deciding whether a spectrum was of sufficient quality for inclusion in the library, procedurally it was decided that the decision to reject or edit a spectrum must be agreed upon by two evaluators. In the event of a disagreement about a particular spectrum, the evaluators communicate their reasoning to one another and attempt to reach a consensus. Disagreements usually involve defective spectra, and most often derive from a difference in strictness of judgment about what should or should not be retained in the library rather than a difference of scientific opinion about the details of the spectrum.

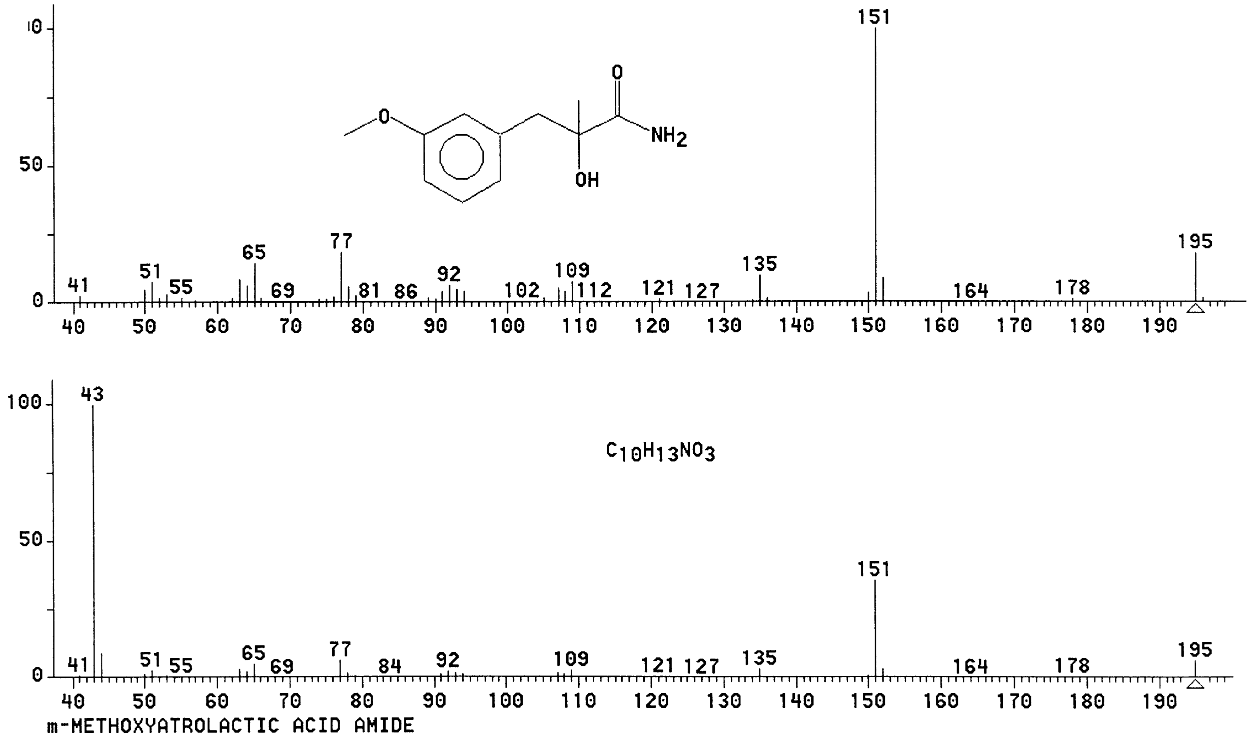

Figure 1. The original spectrum, shown on the bottom, had a large peak at *m/z* 43. This did not correspond to a fragment that logically could be formed by the compound. Comparisons to related compounds indicated that this peak could not be explained as arising from a shifted *m/z* 44 fragment. It was removed as contaminant peak.

The evaluation was carried out using an evaluation "form" consisting of a printout of the spectrum which includes both a graphic and a text representation, as well as a structural drawing. The evaluator identifies the major (and often the minor) peaks in terms of fragmentation processes, and recommends an action: Accept, Flag (as low quality or redundant), Delete, or Edit. When there is a problem, a discussion is written on the page, which can then be considered by a second evaluator. When two evaluators have agreed upon an action, this action is recorded in the master copy of the archive. The archive editing program automatically creates a permanent "log" of all edits and/or deletions carried out.

The spectrum evaluation and library restructuring program was accomplished in four distinct phases:

Preliminary cleanup. The objective of this initial phase was to find and correct or delete seriously flawed spectra suffering from errors that could be identified by computer methods. The most common error found in the preliminary cleanup was that many chemical formulas were found to disagree with the listed molecular weight for the compound, apparently because of a computer error at some time in the past which truncated many chemical formulas. For new spectra, molecular weights are now calculated from the formula, so such discrepancies can no longer occur. Other problems dealt with at this stage were spectra in which water or air peaks were predominant; these were identified by a computer search, and, where possible, the peaks (or portions of peaks) due to water or air were subtracted from the spectrum after an evaluator first ensured that the targeted peaks were indeed due to these impurities. A number of spectra with major peaks at higher mass-to-charge ratio than the molecular ion were also identified by computer. This generally led to either the discovery of a molecular weight and/or formula error or deletion of the spectrum.

Evaluation of replicate spectra. After the preliminary cleanup was completed in 1989, evaluation of all replicate spectra was undertaken as the next major task, mainly because of the need to decide which would be included in the distributed library (the Quality Index calculation was incapable of performing this task). Although fewer than 20% of the compounds in the library had replicate spectra, the importance of this subset of compounds, coupled with the added reliability due to having confirmatory spectra, ensured that this evaluation effort would noticeably enhance the overall quality of spectra in the library for average users. As this phase of evaluation proceeded, a decision was made to change the policy of releasing only one spectrum per compound, and to provide, along with the main library, a separate replicate spectrum file that

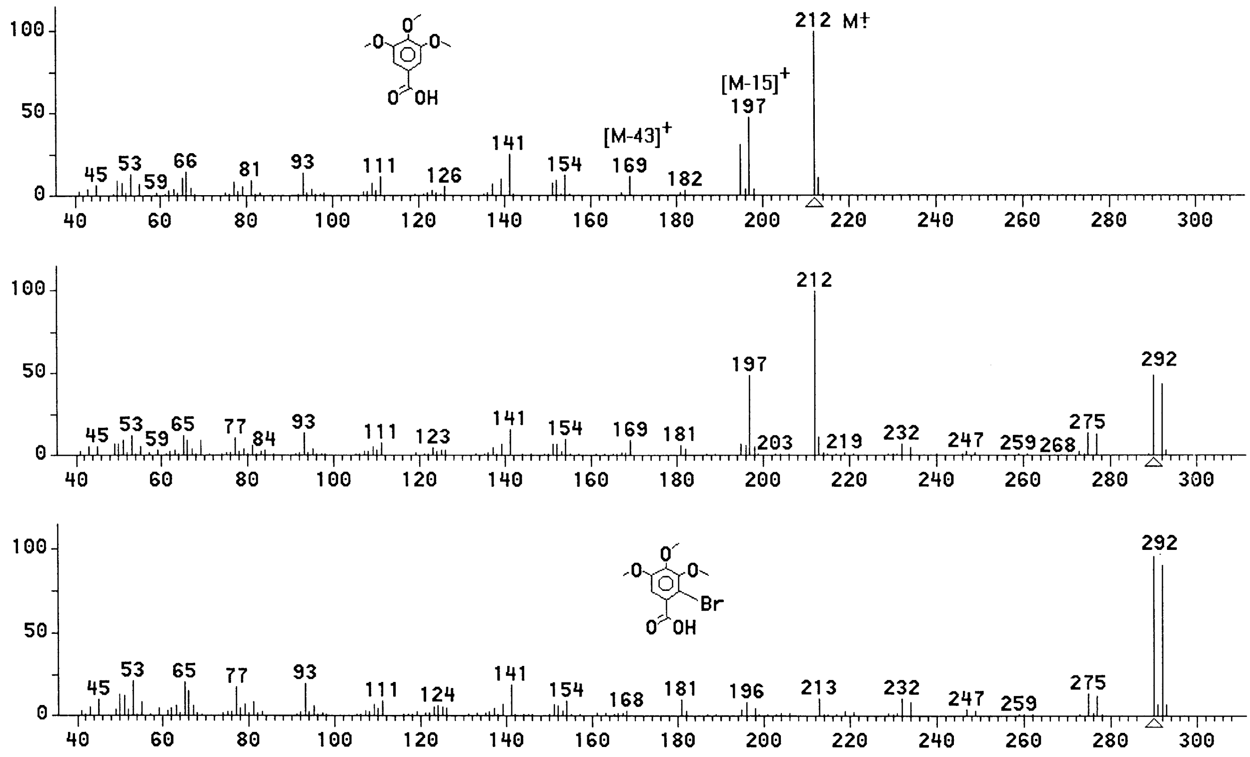

Figure 2. The subject spectrum of this brominated compound, shown in the middle panel, actually represented a mixture known to contain the brominated and nonbrominated analogs (spectra shown above and below). The spectrum of the nonbrominated analog was subtracted and rescaled, leaving a corrected spectrum of the compound, shown on the bottom.

contained a limited number of good quality spectra for compounds already represented in the main library. This change in policy was made for the following reasons: (1) it was learned that certain spectral searching/matching algorithms used in commercial instruments gave significantly better results if replicates were included in the library; (2) replicate spectra have some value for documenting typical variations in mass spectra for common compounds; (3) in a few instances, significant variations in the spectrum of a compound can occur because of different degrees of decomposition in the mass spectrometer chamber. This file, called the "Selected Replicates Library," contained approximately 12,000 spectra of 8000 compounds and was distributed for the first time with the 1992 edition of the NIST/EPA/NIH Mass Spectral Database. Represented in this release were spectra for 62,350 compounds. Efforts were made to restrict the number of replicates to two, although additional replicates were accepted when they were sufficiently unique.

Evaluation of spectra without replicates. This task was inherently more difficult than evaluation of spectra with replicates. In several thousand cases, however, stereoisomers could be compared and replicate spectra could be found for comparison in other collections.

Evaluation of newly acquired spectra. While the evaluation of the original archive was proceeding, significant numbers of new spectra were acquired. The evaluation of these spectra then became a major hurdle to the production of the new library. A particularly time consuming task was the evaluation of new replicate spectra, since each of these had to be compared to all previous spectra of the compound, and in some cases dozens of such replicates were in the archive. Because of the magnitude of this task and since nearly five years had passed since the previous release, it was decided to delay the full evaluation of a large subset of the new nonreplicate spectra because these had been evaluated previously by other groups using procedures similar to those described here. These spectra, however, were subjected to the computer tests employed for all other spectra and the suspect spectra were individually examined. The 1998 archive contained over 177,510 spectra, nearly triple the size of the 1988 archive.

Except for the preliminary cleanup, which involved selectively removing several thousand exact duplicate copies of spectra, as well as locating spectra with the most egregious errors, the actual evaluation procedures carried out during the various phases were the same and are discussed together in the next section.

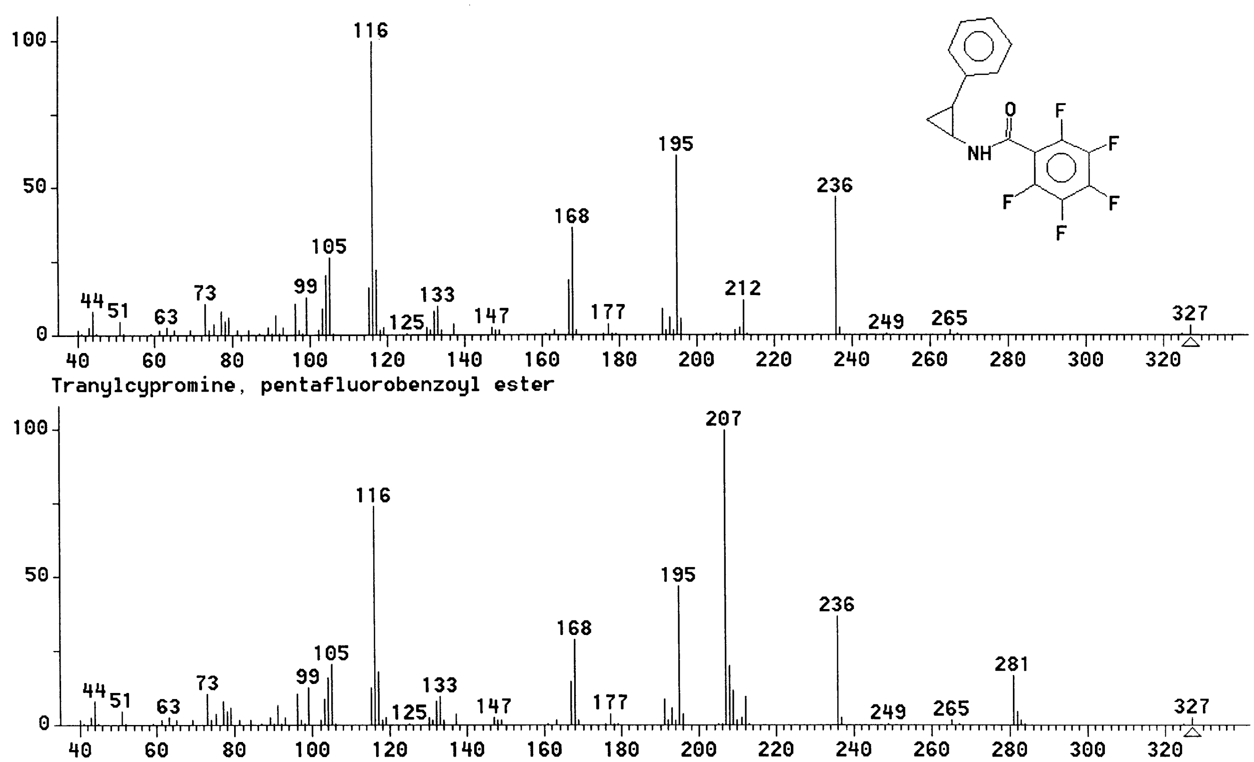

Figure 3. The original spectrum, shown on the bottom, contained peaks at 207 and 281 that had large isotope peaks that are characteristic of impurities from "column bleed" (ions containing Si). These peaks, which could not be explained in terms of the structure of the subject compound, were removed; the corrected spectrum is shown on the top. The possible bleed peak at 73 was unedited since a portion of it may have arisen from the compound under study.

Evaluation Procedures and Criteria for Editing/Deleting Spectra

The primary criterion for inclusion of a spectrum in the library is that it has been verified that the spectrum is consistent with the structure of the subject molecule, and contains its most characteristic peaks. The evaluation of each individual (or replicate) spectrum includes the following overlapping steps.

(1) Examination of the assigned name, structural drawing, and the spectrum itself to ensure that they are consistent. Inconsistencies at this stage are most often due to an error in the structural drawing, and are corrected by redrawing the structure. In some cases, this may mean that the formula and molecular weight must be corrected to conform to the name provided by the original contributor of the spectrum.

(2) Obvious problems not related to fragmentation mechanisms are also identified. These include: (A) Incompleteness: Spectra reported in the literature generally include only the major peaks, and hence are often incomplete. Since such spectra are not particularly useful for library searches, the policy has been adopted to include such spectra only if the compound is of special interest (see discussion above under "Sources of New Spectra") and no other spectrum is available. It is a long-term objective to replace each incomplete spectrum with a complete spectrum determined especially for the library. (B) Correctness of the isotope ratios for the molecular ion and major fragments: This is done with the help of a NIST-developed program for predicting isotope peaks, which has been incorporated in the library maintenance program. Except for cases of clear instrument or transcription error, isotope peaks for the molecular ion are not added. Except in such narrowly defined instances, peaks are never added in the evaluation process. (C) Detector saturation: This is most reliably determined, when possible, by an examination of the abundances of the isotope peaks associated with the base peak. In a saturated spectrum, these isotope peaks will be too large. In cases where an isotope peak in a saturated spectrum can be unambiguously attributed to a particular ion, it is possible to correct the spectrum by increasing the abundance of saturated peaks relative to other peaks in the spectrum by the appropriate amount.

(3) Ascertain whether the major peaks are reasonable for the particular molecular structure. This step may involve many operations. (A) Obviously, one starts by examining the peak due to the molecular ion (if there is one) to verify that it appears at the correct mass. (B) The evaluator then examines other peaks in the spectrum and verifies that they are reasonable for the particular

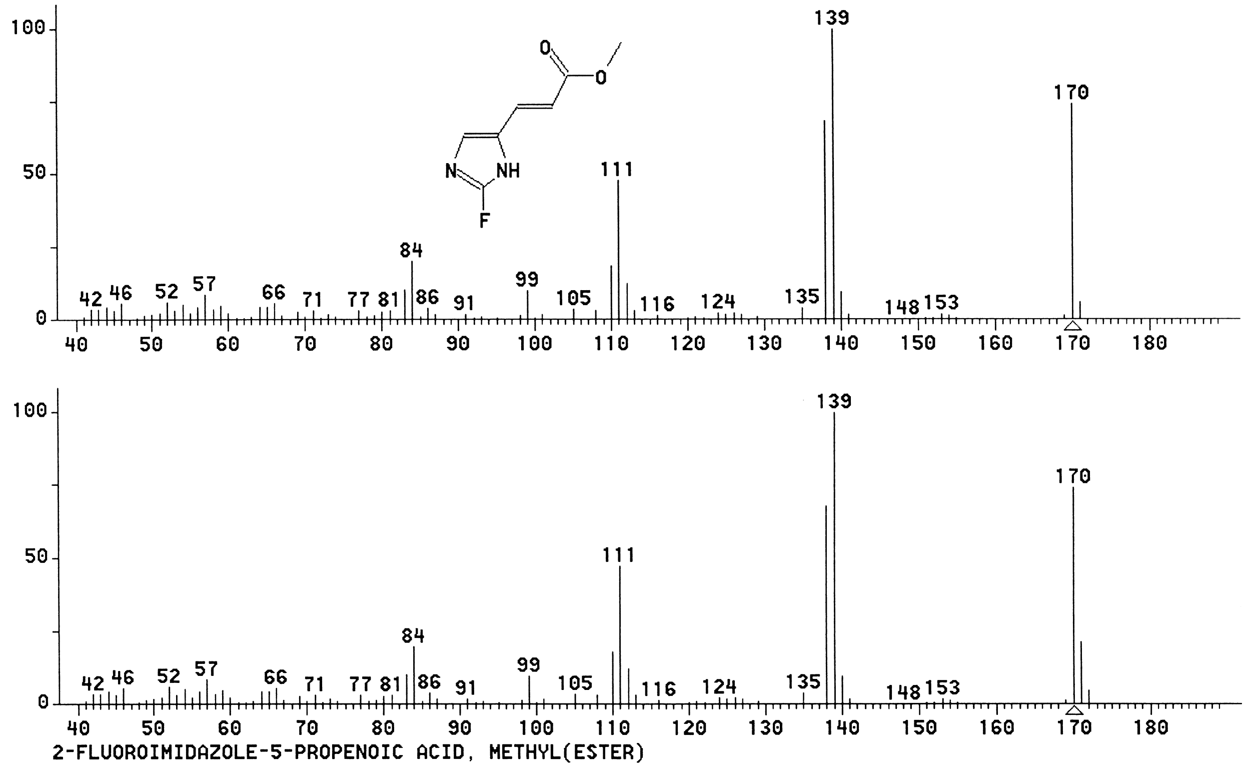

Figure 4. The original spectrum, shown on the bottom, exhibited large peaks one and two units above the parent peak with abundances that did not correspond to the predicted abundances of the isotope peaks. Their presence was due to "chemical ionization" effects (proton transfer in the ion source) and the abundances were reduced to conform to correct values for the isotope peaks. The corrected spectrum is shown on top.

molecular structure assigned to the compound. In this phase, all available spectra of the particular compound are examined, including (if any) all spectra for the compound in the NIST data archive, and also those in other mass spectral collections, especially the Wiley Registry of Mass Spectra [2]. (The Wiley Registry of Mass Spectra actually has some 36,847 spectra in common with the NIST library because both incorporated spectra from older collections.) Comparisons may also include spectra of stereoisomers, other isomers, homologues, derivatives, etc.

The evaluator must have an in-depth knowledge of established rules of fragmentation [12]; ultimately, all decisions are based on the evaluator's expertise. (Even an agreement between a spectrum and its replicate spectra or spectra of stereoisomers does not guarantee that the spectrum will be accepted; in several instances, it was found that a number of similar spectra were bad in the same way.) Of course, in many instances, there are no other spectra with which a spectrum can be compared, and the evaluator must simply assess the validity of a spectrum wholly from a knowledge of the molecular structure and fragmentation rules.

When the features of a spectrum are found to be reasonable for the particular molecular structure, then the spectrum is accepted for the library. If the library contains two or more correct spectra of a compound, one is chosen for the main library and the other(s) selected for the replicates file. Under certain conditions, certain types of low quality spectra may be included. For example, an incomplete spectrum may be included if no higher quality spectrum of the subject compound is available, and at least 10 of the most characteristic peaks are present. Similarly, "monoisotopic" spectra (spectra of compounds where no isotopic peaks were recorded) may be included, but only if no better quality spectrum of the subject compound is available or can be obtained.

In cases where corrections could be reliably made, spectral editing was done. Merely flagging these spectra would have relatively little benefit to users who depend on library searching to identify compounds. The major categories of correctable errors are:

(1) Peaks due to impurities: the most common correctable error is the presence of peaks from a foreign compound or compounds. This may result when material from the chromatographic column "bleeds" (so-called "column bleed"), when compounds previously determined in the mass spectrometer have saturated the walls of the inlet or the ionization chamber ("memory effects," Figure 1), or—particularly in the case of older spectra determined before analytical instruments were commonly coupled to chromatographic columns—spectra of samples containing impurities. Peaks (or

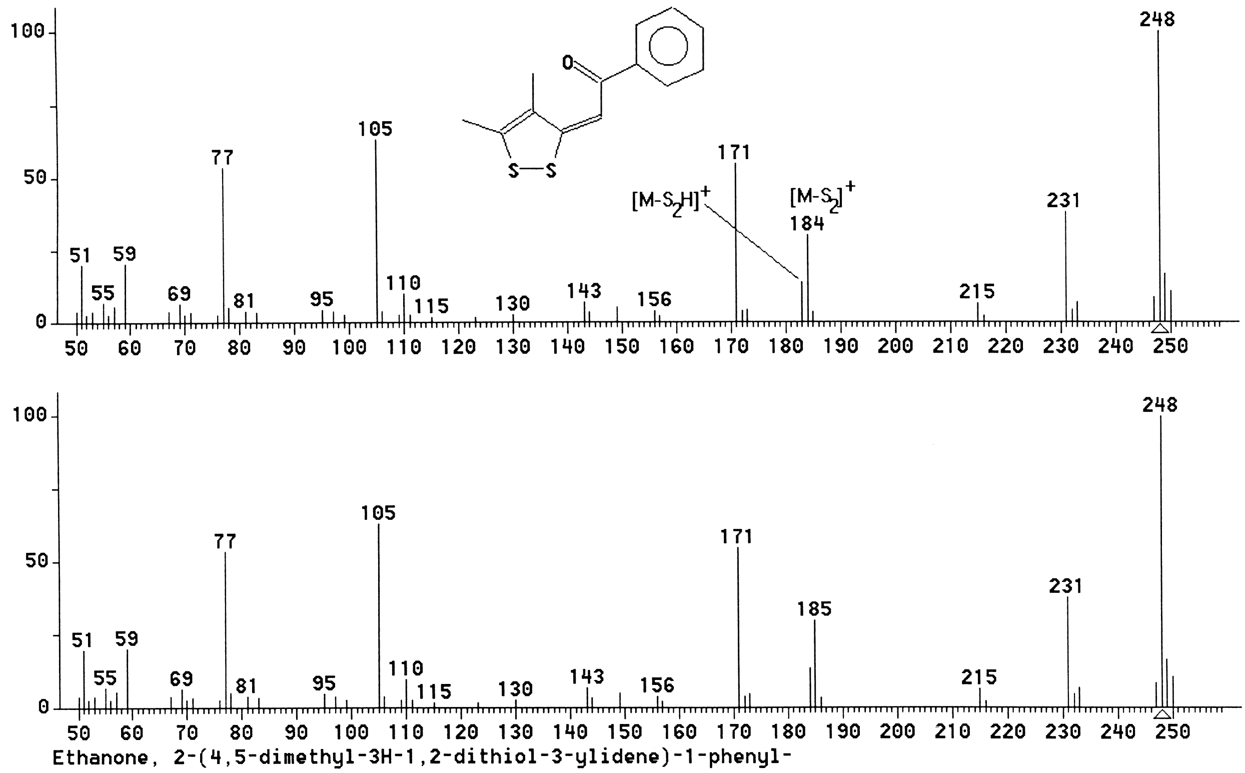

Figure 5. The original spectrum (bottom) had significant peaks at masses 184 and 185, with the peak at 185 predominating. Expected neutral losses from this molecule would be S2(32) and HS2(33), which should lead to the observation of fragment ions at masses 183 and 184. Since there was no logical way to explain the loss of a fragment of mass 31, the cluster of peaks was shifted one mass unit lower. This is a transcription error. The corrected spectrum is shown on top.

portions of peaks) from impurities can be subtracted from the spectrum provided the identity of the impurity compound can be established (Figure 2) or (in the case of column bleed) a reasonably reproducible pattern for the impurity peaks can be determined (Figure 3). Searches of the archive for spectra containing spectra of common solvents (benzene, methylene chloride, etc.) revealed dozens impure spectra which were either corrected or deleted.

(2) "Chemical Ionization" effects: when a spectrum has been determined under conditions such that the ions in the chamber undergo collisions with neutral molecules before being detected, it is possible in certain compounds that ion/molecule reactions such as proton transfer from an ionized molecule to a neutral molecule may occur. In this case, the abundance of the peak one unit higher than the parent ion peak (the "parent plus one") will be significantly elevated above that of the expected isotope peak at that position. (In the event that the parent peak is the base peak, the evaluator must be able to distinguish between this effect and detector saturation; this distinction can sometimes be made by an examination of the abundance of the "parent plus two" peak.) Since protonated molecules ("chemical ionization spectra") typically do not undergo extensive dissociation, the assumption can be made that the occurrence of chemical ionization has not significantly altered the body of the spectrum, and the abundance of the "parent plus one" peak can simply be corrected. An example of a spectrum corrected for this effect is shown in Figure 4.

(3) Transcription errors: some spectra were found to have one or more peaks that were displaced from their logical or expected position by one unit. This kind of error is most common among old spectra determined before mass spectrometers were computerized; operators visually transcribing spectra "by hand" sometimes made mistakes in correctly identifying the locations of peaks (Figure 5). In other preautomation spectra, the abundances of certain peaks were transcribed incorrectly by a factor of 10 (Figure 6). In the event that such errors can be unambiguously identified, they are corrected.

(4) Detector saturation: as discussed above, a spectrum displaying detector saturation can be adjusted, provided that isotope peaks associated with a saturated peak can be unambiguously identified. (When this is not possible, spectra which show clear evidence of detector saturation are deleted, or labeled as poor spectra, depending on the extent of the problem.)

(5) Spurious peaks: occasionally peaks appear in spectra because of instrument noise. Such peaks can sometimes be recognized because they would represent

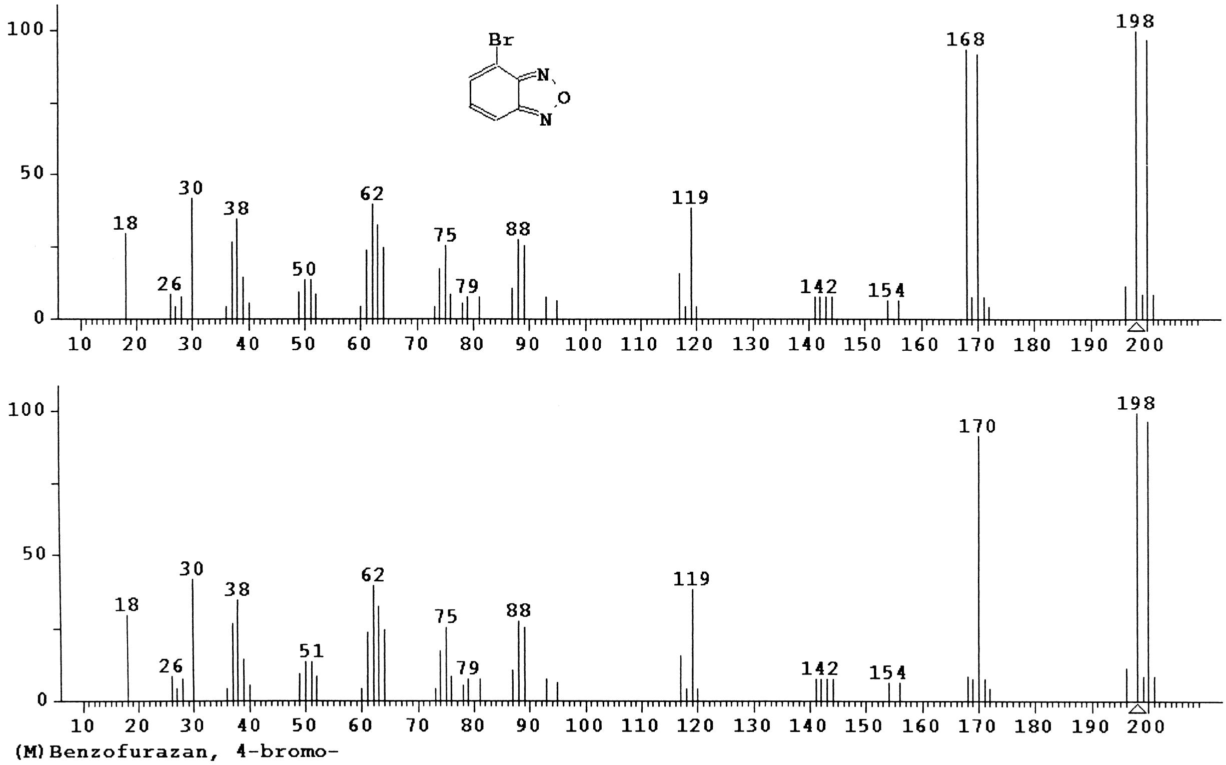

Figure 6. The peaks at *m/z* 168 and 170 result from the loss of NO (30) from the parent ion. The fragment ion retains the bromine atom, so the abundances of these two peaks should exhibit the 1:1 ratio of the two major bromine isotopes. The abundance of the peak at *m/z* 168 is one tenth of that at *m/z* 170. An examination of the spectrum of a chlorine-substituted analogue confirms that it is the *m/z* 168 peak that should be corrected upward, rather than the 170 peak being corrected downward. The abundance of the peak at *m/z* 168 was adjusted to the appropriate value. (The spectrum is indeed an old spectrum, probably transcribed "by hand.")

"illogical" losses, and have no associated isotope peaks (Figure 7). Such peaks are simply deleted.

(6) Errors in auxiliary data: as discussed above, the library contained numerous spectra with errors in the auxiliary identifying information (compound formula, molecular weight, CASRN). When the spectrum was found to be consistent with the name provided by the original laboratory, that name was taken as the primary identifier, and other information was corrected to conform. Occasionally (as pointed out in the literature by an author who later became one of the evaluators of the library [8]), the spectrum itself becomes the primary identifier, and it is the name (and other information), or sometimes the structural drawing (Figure 8), which must be corrected.

Generic Problems and Evaluation Policies

Since a mass spectrum originates from a distinct gas phase compound, it is this precursor compound which is given as the source of each spectrum. Problems can arise, however, when the gas phase compound is not identical to the starting condensed phase sample.

Such problems most often occur because of low volatility, reactivity or impurities in the sample. These problems are minimized when spectra are obtained by GC/MS, since only volatile substances can elute from the column and impurities are generally separated. Any decomposition prior to detection is usually revealed in the chromatogram. Decomposition in the injector, for instance, often generates complex mixtures while decomposition in the column leads to characteristic broad peaks. Certain organic salts are an exception. In some cases these are converted to covalent (free-base) forms prior to vaporization. For instance, good quality spectra of amines may be produced by quaternary ammonium salts such as hydrochlorides, hydrobromides, etc. The chemical names and CASRN of the free-base forms of these compounds are associated with these spectra, which generally differ from labeling of the precursor sample. Other types of onium salts may isomerize into covalent adducts with molecular weights equal to that of the cation and anion pairs. Such spectra are rejected. In addition, certain compounds may undergo chemical isomerization prior to injection or in the injector itself. It is the responsibility of the evaluator to detect such problems and, when appropriate, reject the resulting spectra.

Problems arising from low sample volatility are particularly common for direct analysis (probe) mea-

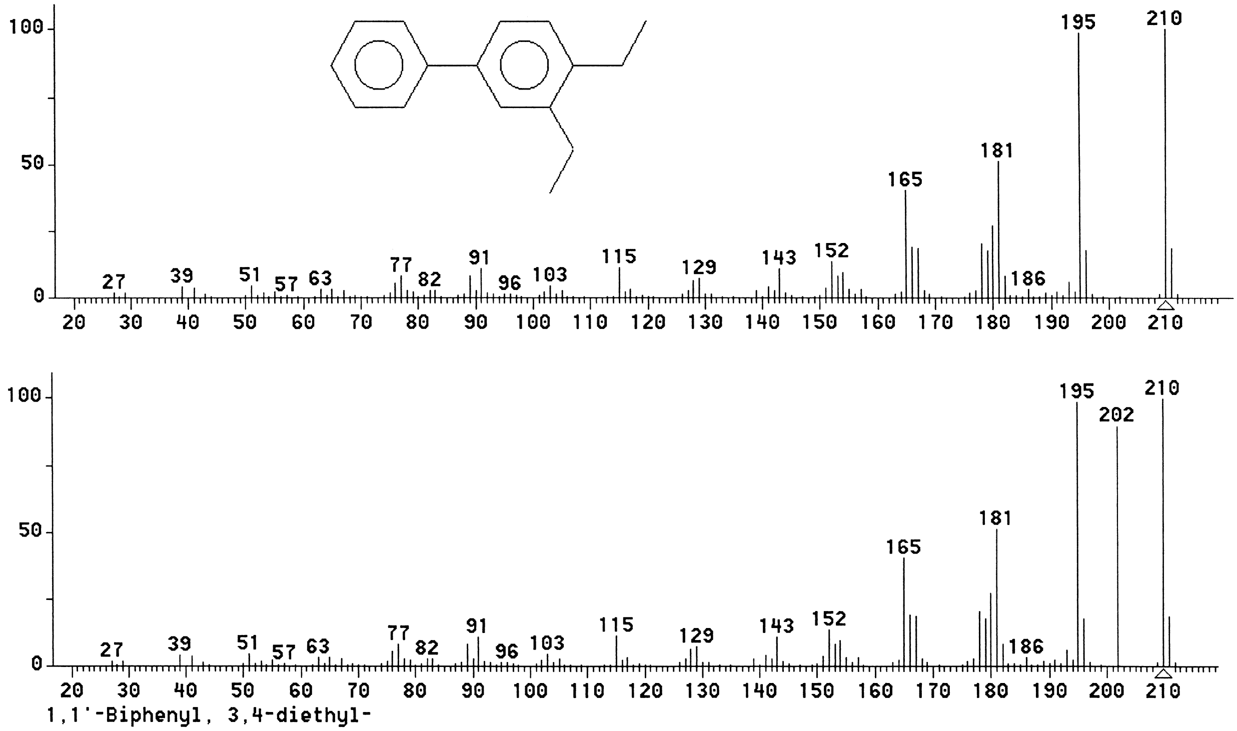

Figure 7. The submitted spectrum had a large peak at *m/z* 202. This peak had no isotope peak and did not correspond to a logical loss. It was removed as a noise spike.

surements, since with sufficient heating virtually any organic substance will give rise to volatile products. For spectra taken with a probe, five varieties of problems are common:

(1) Salts: when it is clear to the evaluator that the gas phase compound whose spectrum was determined was the free-base, rather than the precursor salt, peaks clearly originating from the anion portion of the salt are removed and the spectrum is assigned to the free-base compound. (For example, a spectrum of a compound nominally labeled as a quaternary ammonium hydrochloride will usually exhibit a good amine spectrum with a spectrum of HCl superimposed; the peaks originating from HCl are subtracted, and the spectrum is labeled as a spectrum of the amine.) If decomposition is indicated or there is no stable form of the free-base compound, the spectrum is rejected. One goal for the immediate future is to recheck all spectra attributed to salts, to ensure that the spectra have been treated in a consistent manner by different evaluators.

(2) Low volatility: for compounds of very low volatility, peaks in the vicinity of the molecular ion peak are generally required to confirm that the compound has vaporized without significant decomposition.

(3) Reactivity: because of their combined reactivity and low volatility, certain classes of compounds tend to thermally degrade prior to volatilization. Examples are polyfunctional amines, amides, carboxylic acids, and polyols. However, in some cases good quality spectra unobtainable by GC/MS may be acquired by probe methods, and these are of value for LC-MS analysis. Accepting or rejecting such spectra requires expert judgment of the evaluator.

(4) Volatile impurities: relatively volatile impurities, especially solvents used to dissolve samples for probe analysis, are common sources of spurious peaks in probe spectra.

(5) Background subtraction: efforts to remove impurities by subtracting a region thought to contain impurities from the target compound region can lead to significant distortion if the region subtracted actually contains a significant contribution from the target compound.

One problem inherent in all conventional EI mass spectrometry is the decomposition of gas phase compounds in the mass spectrometer inlet or ionization chamber. Since mass spectra of thermal decomposition products often resemble fragmentation products of the precursor ion, this problem can be very difficult to detect. Adding to the problem is the variability of spectra arising from decomposition products, which can be sensitive to surface conditions. Our general policy was to accept such spectra, selecting the spectrum showing the least decomposition for the main library when replicate spectra are available. Some evaluators have argued that spectra exhibiting such effects, while of low quality, may be of value for the users of the library whose instrument may also produce spectra for certain labile compounds with the same problem.

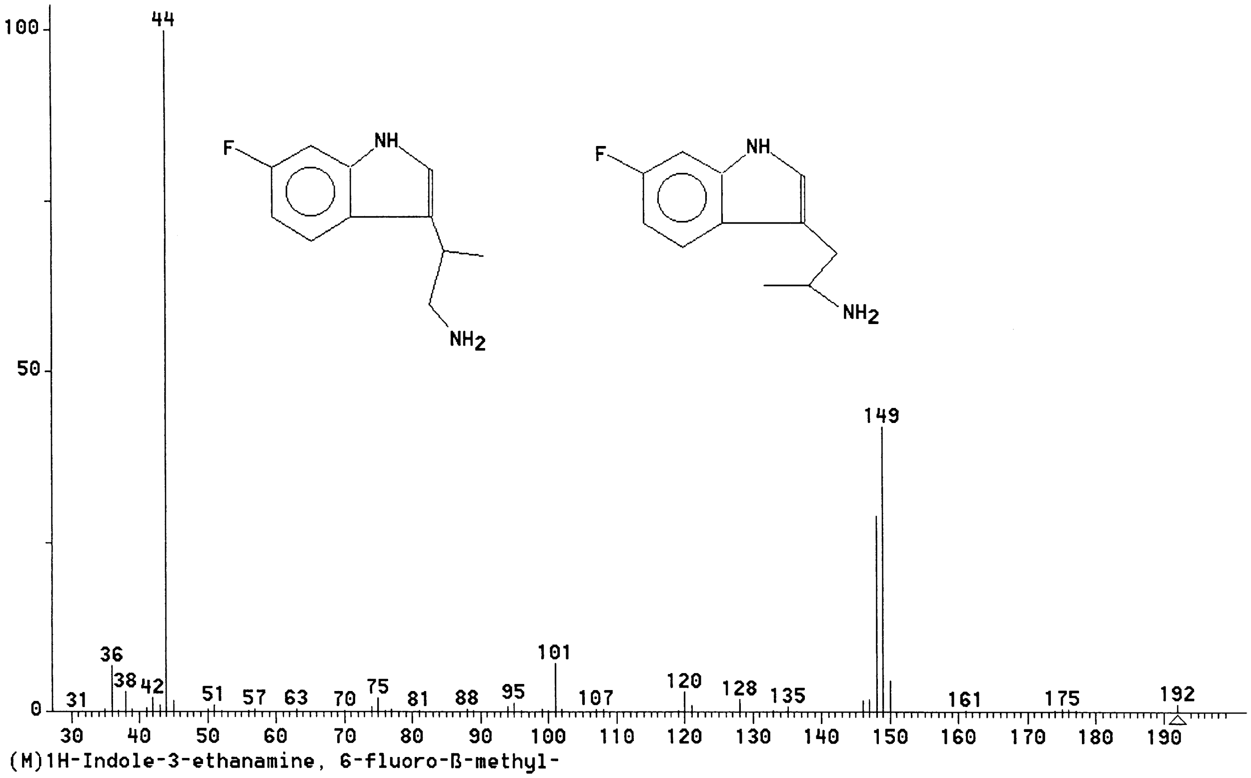

Figure 8. The features of the mass spectrum were in good agreement with the name but not with the structural drawing. The erroneous structure on the left was redrawn to the correct structure on the right.

Final Statistics

From the archive of 175,510 spectra, 129,136 spectra of 107,886 compounds were selected for inclusion in NIST 98. This included 69,061 different CASRNs, 13,205 of which were associated with 21,250 replicate spectra. Replicate spectra without an assigned CASRN are not presently included.

Of the 46,374 archival spectra excluded from the library, approximately 30% were isotopically labeled, 47% were flagged for deletion (incorrect), and the rest were either marked as being of a quality too low for inclusion, or were redundant replicate or duplicate spectra. For each of the 13,205 compounds having replicates, a manual selection of the best spectrum was made.

Peaks were deleted or otherwise edited in approximately 7% of spectra selected for inclusion. The mean and median number of peaks per spectrum was 93 and 78, respectively, significantly higher than corresponding values reported in 1991 [11] of 72 and 53. Approximately 2% of spectra have fewer than 10 peaks whereas 38% have more than 100 peaks, in contrast to the corresponding prior values of 3.5% and 23%, respectively.

Of the 62,235 spectra of different compounds in the main library of the previous version, 7929 do not appear the main NIST 98 library. Of these, 4541 were replaced by better quality spectra and spectra for 3388 compounds were deleted, including 1607 with CASRNs. Changes to the chemical name were made for 1860 compounds and CASRNs were added or revised for 5244 spectra.

The significance of a compound can be roughly assessed by the presence of its CASRN in selective chemical indexes. Percentages of compounds in these indexes that are also in NIST 98 are given in Table 2 along with the percent change from the previous version. The two principal factors determining these former values are the fractions of compounds in each index that are volatile and their commercial availability. Nearly all volatile compounds in the EPA list are included as are nearly half of all commercially available compounds (the bulk of those remaining are involatile). On the other hand only 10% of compounds in the TSCA Inventory are represented, many being mixtures, involatile substances and exotic compounds.

"Recall/reliability" plots derived from match factors in library searching have been interpreted to provide information concerning the quality of a mass spectral library [16]. We have not performed such an analysis since we are unable to separate library quality from other factors that strongly influence these plots. For instance, inclusion of replicate spectra in the library (or even exact duplicate spectra) of compounds in the test set can significantly improve recall/reliability "performance" by increasing the number of correct answers

[10], while having little effect on more widely used performance measures (position of the correct hit in the hit list) [17]. Similar effects can occur by selecting alternative methods for computing match factors [16, 18]. We therefore feel that the most effective means of describing the quality of a library is, as presented here, to provide the criteria for including and editing spectra along with statistical measures of the spectra and of the distribution of compounds.

Future Work

A considerable amount of work remains to be done to further improve library quality. Six high priority areas are listed below:

(1) Archive program: a networked data evaluation program has been developed to replace the existing archive editing program. This will provide access to the evaluation history of each spectrum and permit the entry of evaluator comments. It is planned that these will be made available to interested users, perhaps through the Internet.

(2) Fragmentation software: algorithms have been developed that identify the peaks in a spectrum that are consistent with fragmentation rules. This will be employed both to find possible errors in the library and to assist in the evaluation of new spectra.

(3) Substructure analysis: methods have been developed for reliably identifying the presence and absence of certain chemical substructures from a spectrum by analyzing results of library searches [19]. Library entries where these predictions are inconsistent with the reported structure will be examined for errors. Other methods for substructure searching will also be applied to find groups of compounds expected to have similar spectra for further analysis.

(4) Acquisition of spectra of relevant compounds: a goal is to acquire spectra for as many compounds as possible that appear in the indexes in Table 2. Further, when such compounds are commercially available and are represented by just one spectrum, a replicate spectrum will be sought.

(5) Chemical nomenclature: attention will be given to the collection of chemical names attached to the spectra, and efforts will be initiated to provide CASRNs and systematic names for all compounds.

(6) Review of spectra of salts and derivatives: all remaining of spectra still attributed to salts will be re-examined for the purpose of renaming and possibly editing their spectra. Derivatives will be linked to the CASRNs of their starting compounds.

References

1. *NIST/EPA/NIH Mass Spectral Database*, Standard Reference Database 1, 1992; NIST 98, Standard Reference Database 1, 1998, Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD.
2. McLafferty, F. W.; Stauffer, D. B. *Registry of Mass Spectral Data, 6th Electronic Edition*; Wiley: New York, 1994.
3. Several other examples are: (a) *The Eight Peak Index of Mass Spectra*; Royal Society of Chemistry: Nottingham, England; (b) *The Chemical Concepts Quality Collection*; Chemical Concepts GmbH: Weinheim, Germany; (c) *SBDS Spectral Database*; National Institute of Materials and Chemical Research: Ibaraki, Japan.
4. Speck, D. D.; Venkataraghavan, R.; McLafferty, F. W. *Org. Mass Spectrom.* **1978,** *13*, 209.
5. Milne, G. W. A.; Budde, W. L.; Heller, S. R.; Martinsen, D. P.; Oldham, R. G. *Org. Mass Spectrom.* **1982,** *17*, 547.
6. Terwilliger, D. T.; Behbehani, A. L.; Ireland, J. C.; Budde, W. L. *Biomed. Environ. Mass Spectrom.* **1987,** *14*, 263.
7. Domokos, L; Henneberg, D.; Weinamm, B. *Anal. Chim. Acta* **1983,** *150*, 37–44.
8. Zhu, D.; She, J; Hong, Q.; Liu, R.; Lu, P.; Wang, L. *Analyst* **1988,** *113*, 1261–1265.
9. Lias, S. G. J. R. *J. Res. NIST* **1989,** *94*, 25.
10. McLafferty, F. W.; Stauffer, D. B.; Loh, S. Y. *J. Amer. Soc. Mass Spectrom.* **1991,** *2*, 438–439.
11. Stein, S. E.; Ausloos, P.; Lias, S. G. *J. Amer. Soc. Mass Spectrom.* **1991,** *2*, 441.
12. See, for example: Watson, J. T. *Introduction to Mass Spectrometry, 3rd ed.*; Lippincott-Raven: Philadelphia, 1997; McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra, Fourth ed.*; University Science Books: Mill Valley, CA, 1993; Vul'fson, N. S.; Zaikin, V. G.; Mikaya, A. I. *Mass Spectrometry of Organic Compounds*; Khimiya: Moscow, 1986; Beynon, J. H.; Saunders, R. A.; Williams, A. E. *The Mass Spectra of Organic Molecules*; Elsevier: Amsterdam, 1968.
13. Hertz, H.; Hites, R. A.; Biemann, K. *Anal. Chem.* **1971,** *43*, 681.
14. Heller, S. R.; Milne, G. W. A. *EPA/NIH Mass Spectral Data Base*, NSRDS-NBS 63, U. S. Government Printing Office, Washington D. C. 1978–1983, Vols 1–3, Suppl 1–2.
15. *Chemical Information System*, an on-line collection of databases administered in the 1970s by Fein–Marquart Associates, Baltimore, MD.
16. McLafferty, F. W.; Zhang, M.-Y.; Stauffer, D. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1998,** *9*, 92–95.
17. Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 859–866.
18. Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 316–323.
19. Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1995,** *6*, 644.