Computational Methods in Chemistry

Edited by

Joachim Bargon

IBM Research Laboratory
San Jose, California

PLENUM PRESS · NEW YORK AND LONDON

Library of Congress Cataloging in Publication Data

Symposium on Computational Methods in Chemistry, Bad Neuenahr, Ger., 1979.
Computational methods in chemistry.

(The IBM research symposia series)
"Proceedings of the Symposium on Computational Methods in Chemistry, held in Bad Neuenahr, German Federal Republic, September 17–19, 1979."
Sponsored by IBM Germany.
Includes index.
1. Spectrum analysis—Data processing—Congresses. 2. Quantum chemistry—Data processing—Congresses. I. Bargon, Jaochim. II. IBM Deutschland. III. Title. IV. Series: International Business Machines Corporation. IBM research symposia series.
QD95.S93 1979 542'.8 80-14881
ISBN 0-306-40455-9

Proceedings of the International Symposium on Computational Methods in Chemistry, sponsored by IBM Germany and held in Bad Neuenahr, German Federal Republic, September 17–19, 1979.

PREFACE

The papers collected in this volume were presented at an international symposium on Computational Methods in Chemistry. This symposium was sponsored by IBM Germany and was held September 17-19, 1979, in Bad Neuenahr, West Germany.

According to Graham Richards [Nature 278, 507 (1979)] the "Third Age of Quantum Chemistry" has started, where the results of quantum chemical calculations have become so accurate and reliable that they can guide the experimentalists in their search for the unknown.

The particular example highlighted by Richards was the successful prediction and subsequent identification of the relative energies, transition probabilities and geometries of the lowest triplet states of acetylene. The theoretical predictions were based chiefly upon the work of three groups: Kammer [Chem. Phys. Lett. 6, 529 (1970)] had made qualitatively correct predictions; Demoulin [Chem. Phys. 11, 329 (1975)] had calculated the potential energy curves for the two lowest triplet states (^{3}B and ^{3}A) of acetylene; and Wetmore and Schaefer III [J. Chem. Phys. 69, 1648 (1978)] had determined the geometries of the cis (^3B_u and ^3A_u) and the trans ($^3B_2$ and $^3A_2$) isomers of these two states. In a guided search, Wendt, Hunziker and Hippler [J. Chem. PHys. 70, 4044 (1979)] succeeded in finding the predicted near infrared absorption of the cis triplet acetylene (no corresponding absorption for the trans form was found, which is in agreement with theory), and the resolved structure of the spectrum confirmed the predicted geometries conclusively.

This convincing success of quantum chemical predictions triggered our curiosity to assess the extent to which experimentalists, working in different fields of chemistry, could benefit from recent achievements of theoretical methods. At the same time, we wanted to inform the theoreticians about the current needs of the experimentalists.

The focus of this symposium was chiefly on two areas of chemistry, namely the computational progress in various kinds of spectroscopy (NMR, IR, ESR, PES, and Mass Spectrometry), and the recent achievements of quantum chemistry, sometimes called "Computer Chemistry" by the lectureres in this volume. Considerable time was spent during discussion periods to compare the strengths and weaknesses of semi-empirical versus ab initio methods when applied to problems of varying size and to properties of a different nature. In the chairman's opinion, it seems that both approaches will continue to co-exist, each taking their specific role. In the case of theoretical studies of chemical reactions, for example semi-empirical methods can provide an economical way to map out potential energy surfaces in a preliminary and crude form, perhaps by automatically forming energy gradients combined with automated geometry optimization, followed by a refinement of all parameters with sophisticated ab initio methods in the areas of highest interst.

A great number of theoretical and computational methods, customary in chemistry nowadays, had to be omitted or seriously neglected. Thus no Monte Carlo calculations were included not as the consequence of parochial thinking due to the casino next door, but rather because of the hopeless task to consider all of the known computational methods in chemistry within the framework of a coherent but limited symposium. The topics chosen here were considered to be of interest to a group of scientists whose bandwidth would overlap in the area of physical organic chemistry.

Joachim Bargon
IBM San Jose Research Laboratory
Symposium Chairman

Paul Schweitzer
IBM Germany, Sindelfingen
Symposium Manager

CONTENTS

SETTING UP, USING, AND MAINTAINING COMPUTER-READABLE SPECTRA COMPILATIONS

J.T. Clerc
University of Berne
Switzerland

H. Könitzer
Swiss Federal Institute of Technology
Zürich, Switzerland

INTRODUCTION

Identification and structure elucidation of organic compounds is today mostly done with spectroscopic methods. Even though the theoretical basis of most methods in molecular spectroscopy is quite advanced, the analyst still has to rely heavily on the use of various semi-empirical interpretation methods based on previously collected reference data. Traditionally, these reference data are compiled into spectra collections or digested and condensed into correlation tables familiar to every analytical chemist. There are three basic limitations with this traditional approach. Firstly, only those features initially selected for indexing can be used for retrieval. Secondly, multidimensional searches using logical combinations of several descriptors are hardly possible. Thirdly, manipulation of the data becomes tedious even with data collections of moderate size. If modern data processing equipment and methods are used to manipulate traditionally organized data collections, the afore-mentioned restrictions are somewhat relaxed but by no means completely removed. To index the data with additional features becomes possible but is still difficult and limited to predefined features. The same holds true for the multidimensional search capabilities. Handling of the data, however, is definitely improved by the use of computers.

Transforming conventional and traditional spectra collections into computer-readable form brings some advantages for the user. However, philosophy and concepts, often originating from the "data processing stone age", are thereby perpetuated in disguised form preventing optimal use of today's powerful hardware. Thus, it seems appropriate to reconsider the various concepts, methods and goals in the context of the possibilities offered by modern computers and analytical instruments. It is worth-while to do this very carefully. In setting up computer readable spectral data collections test studies quite often lead to solutions that, even though they exhibit grave and obvious shortcomings, will completely block any new development and/or approach. Once the data types and their computer representation are fixed and an experimental collection of some thousand spectra is set up, the activation energy for a fundamental change becomes very high. Thus, the experimental collection starts to grow despite its inherent limitations, becomes established and sets the standard for other compilations of spectral data. However, even well designed data collections tend to become obsolete due to technical advancements.

WHICH DATA SHOULD BE INCLUDED?

Collections of spectral data of organic compounds are expected to satisfy various needs. The practical analyst wants to confirm the identity of a tentatively identified compound by comparing the measured spectra with reference spectra from the data collection. This implies a unique and unambiguous identification of chemical compounds suitable for computer storage and manipulation. Using this canonical identification he retrieves the respective data set from the collection. He expects the result to be ready within a reasonable time and presented in a form suitable for direct use. If the chemical compound he searches for is not represented in the data collection, the system should offer him those entries most similar to the requested one. In the general case these will be those compounds having a high number of substructural elements in common with the structure in question. Thus, substructure search capabilities are called for. If the analyst deals with a compound on which only little and/or conflicting prior information is available, he may try to find a matching spectrum in the library. If this approach is not successful, he may wish to retrieve reference compounds with spectral characteristics similar to the recorded one. The structures of the compounds retrieved may then give valuable hints regarding the structure of the unknown. This operation mode requires access to the library through spectral data. Furthermore, routines for comparing spectra and assessing a degree of similarity have to be available. For easy interpretation of the results the structures of

the retrieved reference compounds should be put out in a form easily perceived by the chemist. As spectral data do not depend only on the structure but are also influenced by the sampling technique and various recording parameters, the data sets have to be complemented with all parameters relevant in this context.

No matter how the reference data collection is accessed, the data are used mainly to resolve ambiguities in spectral interpretation. It is therefore of utmost importance that they are reliable. Users generally consult a reference data collection when in doubt about values and significance of spectroscopic parameters, and so will normally not be in a position to judge the quality of the retrieved reference data sets directly. Thus, the data sets collected in the spectra library should be verified with greatest care.

An important field of application of spectroscopic reference data collections is the investigation of new correlations between structural and spectral features and the development of automated classification routines. Here again a high degree of reliability is of paramount importance. Many of the powerful new algorithms for uncovering hidden correlations between structure and spectra are extremely sensitive to outlayers. Thus, it is not uncommon that the parameters of a classification procedure are primarily determined by a few erroneous spectra in the training set, rather than by the bulk of correct spectra. Furthermore, hitherto unknown correlations will involve spectral and/or structural features not yet in common usage. Thus, the data collection must not be restricted to include features of known significance only. It rather should contain all available structural and spectral data items, stored in such a way that extraction of new features, possibly relevant to the solution of future, still undefined problems becomes possible.

In summary, a reference spectra library should therefore include the following pieces of information. Firstly, the complete structure of every compound has to be recorded in computer readable form. The structure representation should be unique and unambiguous, it should give unrestricted substructure search capabilities and allow for structure output easy to read by the chemist. Secondly, the spectral data have to be stored with adequate precision, accuracy and resolution. The curve trace reconstructed from the stored data should be virtually identical to the original curve trace, as judged by an experienced analyst. Then the requirement for a potentially unlimited set of spectral features is best fulfilled. Thirdly, supplemental information should include all parameters necessary to rerun the spectrum. Above all, the data should be of the highest reliability attainable at reasonable costs.

WHICH REFERENCE COMPOUNDS SHOULD BE INCLUDED?

The answers obtainable from a reference data collection are obviously limited to data sets included in the collection. A universal spectra library including every known chemical compound would be optimal in this context. However, it is obvious that such a collection can never be realized. Furthermore, a data collection of excessive size becomes very expensive to use and to maintain. Thus, a careful analysis of the problem of which compounds to include is worth-while.

The analyst expects useful results from the reference data collection. Results will be useful if they answer the user's queries at least partially. As the set of answers the system can provide is given by the contents of the data collection, it is the range of user queries that defines the optimum contents of a spectroscopic reference data collection. Since universality can never be attained, the user will, in general, not be presented with a reference compound identical to the unknown at hand and he will have to be satisfied with a set of compounds similar to his unknown. We should therefore concentrate on including carefully chosen simple model compounds rather than highly complex molecules. Furthermore, large series of homologous compounds should not be included; a few examples will do. A spectra compilation of limited size, if modelled along these lines, can provide useful reference data for a broad range of problems at acceptable costs. The variety of the model compounds to be included in the compilation depends on the user's field of interest. The compound classes on which a user group does research should certainly be represented by many reference compounds of closely similar structure. Other compound classes of lesser significance to the users may be typified by a smaller number of models. Consequently, different user groups will need different reference libraries. In the general case, it will thus hardly be possible to acquire a ready-made data collection that optimally suits one's own needs. However, a base collection covering a wide range of chemical classes with a limited number of model compounds per class is expected to be common to most applications. This base collection will then have to be enriched with compounds from the user's field of interest. The person best qualified for the selection of these compounds and having easiest access to them is the user himself. Consequently, the user of a spectroscopic reference data collection will have to acquire his own reference spectra and incorporate them into the base collection. Convenient procedures for this operation have to be available. As a further consequence, optimal results cannot in general be expected from general-purpose data collections. Either they are (implicitly or explicitly) specialized to a given field or they contain a great many reference compounds of low utility. Most

currently available data collections belong to the second type. They tend to include all spectra available without proper regard to a well balanced content. The performance of the data bank system is affected in two ways. First, the costs for using the collection are too high, and second, the large number of homologous compounds present often results in a hit list that is monopolized by one compound class. Thus, the user loses an important part of the information the system could supply and, even worse, he is lured into false confidence by the seemingly consistent answers. It is, however, not primarily the data bank designers that are to be blamed for this. As long as a potential user's first question is "how many spectra do you have?", quantity will rank before quality.

MAN-MACHINE COMMUNICATIONS

Even today a surprisingly high percentage of analytical chemists have little or no experience with computers. The "activation energy" for getting started with computers and becoming moderately efficient proves to be quite high. There are several reasons for this.

The language gap between computer people and chemists has become so wide that communication is severely restricted due to an entirely different terminology (computer-Chinese). Furthermore, the various rules and conventions for using many of the larger computer installations are true folklore in the sense that they rely on oral transmission from generation to generation. In order to overcome these and other barriers, the language in which the user has to formulate his queries should be simple and modelled along the patterns and concepts of the analyst's natural language. Our own experience with remote batch systems, however, shows that the real difficulties are not primarily arising from the command language of the spectroscopic data bank, but rather from the computing centre's operating system language and its associated terminology. This problem could probably be overcome with smaller dedicated systems that are optimized to one single task or to a few closely related applications.

Another equally important point is concerned with the presentation of the results. Here again the user expects output formats compatible with his line of thinking. In particular, to make visual comparison easy, spectra should be displayed as curve traces in a format very closely matching the standard output format of the spectra recorded in the laboratory. The output of a spectroscopic information system is often put to use in two rather different ways. In a first step, the analyst wants to check quickly on various preliminary hypotheses. Most of them can be discarded as useless at

once after a quick look at a reference data set. Quite often a new or modified hypothesis is formulated which the analyst also wants to test on the spot. For this application, an interactive system is best suited. Virtually immediate response to the user's queries is most important at this stage. The output of the system is preferably displayed on a video screen. The graphic quality of the output has not to meet highest aesthetic standards; heavy emphasis should rather be placed on a format easy to perceive. In a later stage, when (hopefully) only a few tentative solutions survive, these will have to be checked more carefully and in detail. For this, data have to be available as hard copy. Finally, to document his findings and conclusions, the analyst will probably include various reference data sets into his report. Here, beside a clear and concise presentation, aesthetic aspects become important.

As shown in previous sections, a spectroscopic data collection will be accessed by structural and spectral features. Thus, the user will have to input (partial) structures and spectral data. Input of chemical structures poses no fundamental problems, but a convenient and easy to use system requires large and sophisticated programs and/or expensive hardware. Spectral data should be input as a digital image of the complete curve trace to avoid introduction of artifacts and biases by the user. Consequently, analytical instruments delivering digital output are required, and a convenient procedure to transfer the digitally acquired data from the instrument to the data bank system has to be available.

HOW MANY SPECTRA SHALL BE STORED?

To discuss the question of how many spectra should be included in a reference data collection the storage requirements for various spectra types have to be considered. In principle, one would like to digitally store the data in such a way that all information contained in the analog spectrum is fully retained. However, this leads to excessive storage space requirements. For example a modern routine infrared spectrometer allows a resolution of the wave number scale of better than 1 cm^{-1}. Thus, for the standard range about 4'000 intensity values have to be recorded. The digital resolution of the intensity scale should be of the same order of magnitude as the reproducibility of the instrument. If a resolution of about 1 in 250 is assumed, 8 bits are necessary. Total storage space thus amounts to 32'000 bits per infrared spectrum. To store compound identification, chemical structure and technical parameters require another 8'000 bits. So the total storage space amounts to about 40'000 bits. Therefore, an average disk cassette, as today commonly used with popular minicomputers, will hold a few hundred infrared spectra only.

For bar type spectra as e.g. noise decoupled ^{13}C-NMR spectra or mass spectra, where the line shape conveys little or no useful information, storage space requirements are somewhat less demanding.

The reference data collection should as a base contain representative model compounds covering the full range of organic substances. It is very difficult to give a realistic estimate as to the number of compounds necessary. An order of magnitude may be specified, however. Textbooks treating organic chemistry on an advanced level are expected to give an overview of the full field. They generally have around 1'000 pages where on the average some 2 to 3 compounds are discussed. Thus, a realistic estimate for the size of the base collection is some thousand spectra. These will have to be enriched with spectra of compounds having special relevance to the users. This may result in a doubling of the size of the collection. We thus arrive at a spectra library containing between some 5'000 to 20'000 entries.

It is, therefore, obvious that even for a relatively small data collection a huge amount of storage space is needed if the full information content shall be retained. To relax the requirement for the full information is not an acceptable solution. If only the information currently known to be relevant is stored, all further development is blocked, and the data collection's fate is programmed for premature obsolescence. It is, however, appropriate to delete all data that is known to contain no useful information, and all tricks and gimmicks for data compression should be used.

IMPLEMENTATION

In the foregoing sections various aspects of computer readable compilations of spectroscopic reference data have been discussed. The identified requirements result in severe conflicts.

The data have to include a representation of the chemical structure of the reference compounds. The code should be compact, easy to manipulate, and simple to perceive on output. CAS registry numbers are very compact but completely unsatisfactory in all other respects. Connectivity tables are relatively easy to manipulate but are quite voluminous and difficult to interpret. Pictorial diagrams can provide optimal output but are otherwise unsatisfactory. In our experience the Wiswesser Line Notation (WLN) presents an acceptable compromise in that it is reasonably compact and directly readable. Furthermore, conversion into connectivity tables as well as into pictorial diagrams is in most cases not unduely difficult.

The spectral data have to be recorded in full. This requires adequate instrumentation on the spectrometer side. On the data system side appropriate input ports have to be provided and enough mass storage facilities have to be supplied. To output the data in a format that is optimally matched to the analyst's line of thinking, versatile and expensive computer peripherals are necessary. All this calls for a large and powerful computer system. However, the reference spectra collection and its associated programs will utilize only a small part of the available computing power, making a dedicated system highly inefficient. If the computer system has to be shared with many other users from widely different fields to justify the costs, chances are high that communication and various logistic problems will prevent a significant percentage of potential users from exploiting the data bank. This might be a temporary problem since serious attempts are made today to introduce the subject of digital computers and data processing into the chemistry curriculum. However, for the time being we have to live with the fact that there are psychological barriers to the widespread and general use of computers in chemistry. These barriers are definitely lower with dedicated mini systems, however.

We believe that large dedicated systems are not cost efficient for spectroscopic data banks. For local operation, investment and operation costs are excessively high. If operated in time share mode with a large common data base accessed by many different user groups, the data collection will in general contain mostly spectra of minor or no relevance to the individual user, with detrimental effects on quality and costs. For user groups with closely similar ranges of interest this solution might however be acceptable. The main disadvantages of large general purpose computing centres are communication problems and large turn-around times. As stated before, the former problem is probably a temporary one. However, at present we have to live with it and take it into consideration. In his daily work the analyst often needs almost instant answers to his spectroscopic problems. The result should become available when the sample is still at hand in the laboratory to allow for rerunning some measurements without going through all sample preparation steps. Not all large computer installations can provide such quick service. Furthermore, a direct link between a remote spectroscopic instrument and a large central computer also is not without problems. Regarding communications, system availability and turn-around time, small dedicated systems are optimal. However, the costs to equip a mini system with the necessary peripherals, and in particular with adequate mass storage capacity, are out of proportion. It is possible that the fulminant evolution in the field of personal computers will improve the feasibility of the "small but mine" approach.

At present the most appropriate solution seems to be a dedicate satellite connected to a large centralized system. We have, however, no experience with such a system yet. Our own implementation is realized on a large general purpose computer installation which, besides some rudimentary on-line capabilities, offers fast remote batch processing with turn-around times on the order of some 10 minutes. In this environment it is possible to arrive at an acceptable performance level.

ORGANIZATION OF THE OCETH-SYSTEM

The data compilations incorporated in the OCETH-System currently comprise about 6'500 mass spectra, 3'500 ^{13}C-NMR spectra, and 1'200 UV spectra, originating from various sources including our own laboratories. Initially, the number of compounds documented was almost twice as high, but elimination of duplicates and reference compounds of low utility have reduced the number to the present value. In addition, some 50 infrared spectra are also included to allow for developing and testing of the respective program segments.

The main objective of our implementation is to provide almost unlimited possibilities for combined processing of all data items without regard to the data type or to the spectroscopic method involved. This is realized by having a rigorously standardized format common to all spectroscopic methods presently represented and suitable for future expansion. For practical reasons the data related to different spectroscopic methods are kept on separate files. At present there are but few compounds documented with data from more than one spectroscopic method. Thus, overlap between the different spectra files is rather limited. Furthermore, as our collections are still growing fast, we have to give due consideration to the input side. New data tend to become available in batches of spectra from one method. Updating is thus more easy with specialized files. Finally, separate files for the different spectroscopic methods provide the specialized analyst with specialized data compilations for his special pet method.

This set of files, encompassing the full spectroscopic and supplemental information, are referred to as Library Files. As their detailed structure is to some extent influenced by the computer hardware and the operating system, the following description is limited to system-independent aspects.

Each spectrum documented corresponds to one data record in the library file. Each record consists of three segments. The first segment is of fixed length and holds all information related to the

sample identity, e.g. identification number, chemical name, CAS registry number, structure code (WLN notation), empirical formula, and nominal molecular mass. Furthermore, it gives the key for interpreting the second segment. The format of the first segment is identical in all library files. The second segment includes the data relating to the spectrum registration. It is again of fixed length. Here, some entries have different meanings for different spectroscopic methods; the key being given in the first segment. In addition to instrumental data, the key for interpreting the third segment is given. The third segment is of variable length and holds the spectroscopic data in highly compressed form. The length of the third segment is specified in the second segment as well as the mode of compression used. The file is headed by a header record that identifies the file, gives its length, source, and history as well as other data necessary and/or convenient for processing the file. This data structure, where each part contains the information necessary to correctly process and interpret the following parts, makes it possible to write a unified set of programs that can handle the data from all files. To retrieve any data item from any file, the user just accesses the central processing program which will take care of the various codes and data compression schemes used with the different spectroscopic methods. The central processing program consists of many subprograms designed for various applications. These include routines to output full or partial data sets in standardized formats on various periferals, programs to generate images of the library in a format suitable for data exchange with other institutions, and for generating various index files and subfiles.

Even though the data in the library files are highly compressed, the files are still rather voluminous. The mass spectra file, with 6'500 data sets, has a length of roughly 5 megabytes, or 800 bytes per compound. The ^{13}C-NMR data require somewhat less storage space, namely about 500 bytes per compound. The length of the corresponding file with about 3'500 documented compounds is thus 1.75 megabytes. UV spectra require the same number of bytes per spectrum. IR spectra, however, require about 2'500 bytes per compound.

For most standard applications the library files are not accessed directly. Rather, a specialized index file comprising an appropriate subset of the full data is used. For example the chemical name, permuted WLN code, empirical formula, and nominal molecular mass are directly available via index files. As experience has shown, most chemists at our institutions still prefer to deal with a hard copy index rather than doing a computer search. Consequently, we supply the index files in printed form or on microfiches whenever this is economically and technically feasible. Furthermore, we also provide the user with truncated spectral data in hard copy form when

this seems appropriate. This somewhat conservative approach is justified by the fact that computer terminals are not yet ubiquitous and that a large proportion of all queries involve the search for a fully specified entry, where a manual "telephone directory" search is quite adequate.

More complex search problems are done with the computer. The most common applications include the retrieval of reference compounds exhibiting spectra similar to the spectrum of a sample of unknown structure, substructure search, and multidimensional searches. For the retrieval of spectra similar to a given model we use a special search file where the spectral attributes are encoded in binary form. The spectral attributes are selected so as to emphasize structural similarities rather than individual differences between reference spectra. The system accepts the spectrum of the sample as input, compares it to all reference spectra in the file using a self-optimizing search strategy. On output a list of the 20 compounds believed to be structurally most similar to the sample is produced. In addition, the spectra of the retrieved reference compounds may be plotted either on hard copy with a digital plotter or on a video screen. The plotting routines may, of course, also be used directly for processing spectral data from other sources. Substructure search programs are still in the planning stage. However, programs for the conversion of WLN codes into connectivity tables and pictorial diagrams have been acquired recently and will be integrated into the system. Multidimensional search problems, where entries meeting several conditions at the same time have to be retrieved, arise mainly in connection with the study of structure/spectra correlations. For these applications we use inverted files. The respective programs are currently under development.

The system has been well accepted by the chemists. At present, it is most heavily used for retrieval of reference spectra of specified structure and for plotting spectra. Substructure search and retrieval of spectra similar to a given model are used less often. This is most probably due to the fact that most of our chemists do purely synthetic work where one is only very rarely confronted with compounds of completely unknown structure.

The spectroscopic data bank also provides the base for various research projects. For example, one project aims at a spectroscopy oriented classification scheme for organic compounds. The classification schemes universally used today date from the beginning of our century. They are based primarily on the reactivity of compounds and functional groups, which makes them optimal for the discussion of e.g. syntheses and reaction mechanisms. The modern analytical chemist, however, sees the compounds he deals with, rather, from a spectros-