Data Processing in Chemistry

dpc '80

Zdzisław Hippe
(Editor)

Data Processing in Chemistry

A Collection of Papers Presented at the Summer School, Rzeszów, Poland, August 26–31, 1980

Edited by

Zdzisław Hippe

I. Łukasiewicz Technical University
Rzeszów, Poland

PWN–Polish Scientific Publishers
Warszawa

Elsevier Scientific Publishing Company
Amsterdam–Oxford–New York
1981

Distribution of this book is being handled by the following publishers

for the U.S.A. and Canada
ELSEVIER/NORTH-HOLLAND, INC.
52 Vanderbilt Avenue, New York, N.Y. 10017

for Albania, Bulgaria, Chinese People's Republic,
Cuba, Czechoslovakia, German Democratic Republic,
Hungary, Korean People's Democratic Republic, Mongolia,
Poland, Romania, the U.S.S.R., Vietnam
and Yugoslavia
ARS POLONA
Krakowskie Przedmieście 7, 00-068 Warszawa, Poland

for all remaining areas
ELSEVIER SCIENTIFIC PUBLISHING COMPANY
335 Jan van Galenstraat, P.O. Box 211
1000 AE Amsterdam, The Netherlands

Graphic design: Zygmunt Ziemka

ISBN 83-01-03064-X

Printed in Poland

PREFACE

The volume Data Processing in Chemistry contains papers presented at the International Summer School held in the I. Łukasiewicz Technical University in Rzeszów (Poland) during August 26-31, 1980. The main aim of the DPC-School was to initiate discussion on research devoted to specialized, high-level software for chemistry. The carefully selected papers cover the latest achievements in the application of computers to chosen fields of chemistry and chemical sciences, and are written by leading experts of chemical informatics. They have been prepared in such a way that the information can also be extrapolated readily to similar problems in other areas of research. They will thus serve the entire scientific community.

This volume is subdivided into four parts. The first three discuss the three basic models in chemical informatics (calculational, morphological and semantic). The last part contains various topics generally connected with hardware problems. It should be pointed out that this division of topics is one of convenience, since several of the example problems in each specific category may belong simultaneously to more than one model. A certain amount of overlap in the material is thus unavoidable.

The first part deals with applications of computers to quantum-chemical and related problems, to a generalized approach to designing experiments and processing data, chemical

kinetics, graph theory and to Monte-Carlo calculations of some structural properties of polymeric chains. The topics all represent calculational models (CM) in chemical informatics. Those models are created essentially when features of the problem to be solved are represented by means of digits or numerals, and connections between the features may be approximated with satisfactory accuracy by known mathematical relations.

The basic attributes for the morphological models (MM) are mutual links between objects, the manner of bringing together simple objects into complex objects (for instance, to generate structure from a set of substructures), and the common features of such aggregation. Here, in the second part of the volume, morphological models are well exemplified by discussion of large data bases and library search methods, as well as methods of artificial intelligence, applied to the elucidation of chemical structures.

In the most sophisticated models, the semantic ones (SM), their elements are neither numerical (as in CM) nor structural (as in MM) values, but concepts (notions) and connections between them. The third part of the volume deals with semantic models in chemistry: with computer-aided planning of organic reactions and multi-step syntheses. Both approaches (application of known reactions stored in computer memory and usage of mathematical models of constitutional chemistry) are discussed in considerable depth.

The last part of the volume briefly discusses some specific problems of machine representation of chemical structures, some questions of interfacing analytical instruments with minicomputers, and recent results, very useful in chemistry, in parallel programming.

The excellent contributions of the individual authors to this book are gratefully acknowledged. Their special efforts to adhere strictly to the schedule allowed prompt publication of a coherent volume.

Rzeszów, November 1980 Zdzisław Hippe

CONTRIBUTORS TO THIS VOLUME

H. ABE Toyohashi University of Technology, Toyohashi, Japan

O. ACHMATOWICZ, Jr. Agricultural University, Warsaw, Poland

J. BAUER Technical University, Munich, F.R.G.

J. BRANDT Technical University, Munich, F.R.G.

J.T. CLERC University of Bern, Switzerland

P.A.D. deMAINE The Pennsylvania State University, Pennsylvania, U.S.A.

J. DUGUNDJI University of Southern California, Los Angeles, Cal., U.S.A.

L. EDSBERG Royal Institute of Technology, Stockholm, Sweden

R. FRANK Technical University, Munich, F.R.G.

J. FRIEDRICH Technical University, Munich, F.R.G.

I. FUJIWARA Toyohashi University of Technology, Toyohashi, Japan

J. GASTEIGER Technical University, Munich, F.R.G.

L.A. GRIBOV Institute of Geochemistry and Analytical Chemistry, U.S.S.R. Academy of Sciences, Moscow, U.S.S.R.

H.J. HARWOOD University of Akron, Ohio, U.S.A.

S.R. HELLER U.S. Environmental Protection Agency, Washington, D.C., U.S.A.

K. HERZOG Technical University, Dresden, G.D.R.

R. HIPPE I. Łukasiewicz Technical University, Rzeszów, Poland

Z. HIPPE I. Łukasiewicz Technical University, Rzeszów, Poland

A.M. JANICKI Institute of Mathematical Machines, Warsaw, Poland

J.W. KENNEDY University of Essex, Colchester, Essex, U. K.

H. KOENITZER Technical University, Zurich, Switzerland

M. MARSILI Technical University, Munich, F.R.G.

B. PAULUS Technical University, Munich, F.R.G.

A.J. SADLEJ Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland

S. SASAKI Toyohashi University of Technology, Toyohashi, Japan

A. von SCHOLLEY Technical University, Munich, F.R.G.

W. SCHUBERT Technical University, Munich, F.R.G.

E. STEGER Technical University, Dresden, G.D.R.

I. UGI Technical University, Munich, F.R.G.

T. YAMASAKI Toyohashi University of Technology, Toyohashi, Japan

CONTENTS

PART I. CALCULATIONAL MODELS

COMPUTATIONAL QUANTUM CHEMISTRY: ADVANCES, PERSPECTIVES AND LIMITATIONS

A. J. Sadlej

Institute of Organic Chemistry, Polish Academy od Sciences, Warsaw, Poland

A wide popularity of quantum-mechanical approaches to chemical problems [1-3] is certainly based on the anticipation that quantum mechanics is the right theory for describing, explaining, and predicting chemical phenomena [4]. It means that, in principle, chemical problems could be solved by processing a relatively small set of the fundamental data, which are necessary to set up the hamiltonian of a given system [5]. These data are to be processed via the Schrödinger equation [5], whose solutions provide a complete information about all quantum states of the system under consideration.

However, a general quantum treatment of most systems of chemical interest suffers from tremendeous mathematical difficulties [6]. For these systems, which are built of a number of electrons and nuclei, several approximations must be introduced and the analytic methods of pure mathematics are replaced by the appropriate methods of numerical analysis. In consequence, the exact solution of the Schrödinger equation is replaced by the computation of approximate wave functions and energies. Once the approximate functions are known, they can be either analyzed or further processed to obtain chemically useful information. This step usually involves some additional computations. Hence, the progress in computational methods and computer technology becomes a decisive factor which determines the range of chemical applications of quantum theory.

The present review is addressed primarily to non-specialists and is intended to provide a brief account of data processing problems in computational methods of quantum chemistry. Moreover, the considerations will be limited to the methods which are commonly referred to as the non-empirical or ab initio approaches. This terminology is a little confusing, for it implies that ab initio calculations are carried out starting with very first principles of quantum theory and that they employ only the most fundamental set of data. It could be certainly so, if we knew the exact techniques for processing these data. In practice, however, we need some further approximations. This, in turn, requires introducing some new data.

Before discussing the data processing problems it appears worth-while to give a brief summary of the most characteristic approximations that are introduced in quantum chemistry of molecular systems. Above all, it should be realized that molecular calculations are almost exclusively performed by using the non-relativistic quantum theory [7]. Fortunately enough, the relativistic effects are frequently negligible and, at least for organic molecules, the non-relativistic approximation works perfectly well.

By neglecting the relativistic effects we are left with the ordinary molecular hamiltonian [5] for electrons and nuclei. A considerable reduction of the system dynamics complexity is then achieved by introducing what is known as the Born-Oppenheimer approximation [8,9]. According to this approximation the massive particles are assumed to be at rest, so that the motion of electrons is considered in the field of some fixed nuclear framework. Consequently, one can define the electronic hamiltonian of a molecule $H(\boldsymbol{R})$ in which the nuclear coordinates $\boldsymbol{R}$ appear as parameters:

$$H = H(\boldsymbol{R}) = \sum_i h(i) + \sum_{i<j} \sum v(i,j). \qquad (1)$$

The so-called one-electron terms $h(i)$ are given by the sum of the i-th electron kinetic energy operator $t(i)$ and the appropriate nuclear attraction terms $v_\alpha(i) = v_\alpha(i; \boldsymbol{R}_\alpha)$, i.e.,

$$h(i) = h(i;\boldsymbol{R}) = t(i) + \sum_{\alpha} v_{\alpha}(i;\boldsymbol{R}_{\alpha}) \ , \qquad (2)$$

where $\boldsymbol{R}_{\alpha}$ is the position of the α-th nucleus and the summation is performed over all nuclei of a given molecule [5]. The second sum in eqn. (1) corresponds to the electron-electron repulsion operators.

The electronic hamiltonian depends on the positions $\boldsymbol{R}$ of all nuclei, $\boldsymbol{R} = \{\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots\}$, but the dynamics of the whole system is effectively reduced to the dynamic problem for electrons. Thus, in nearly all problems of molecular quantum chemistry, what we are trying to solve is the electronic Schrödinger equation

$$H(\boldsymbol{R})\Psi(\boldsymbol{x};\boldsymbol{R}) = E(\boldsymbol{R})\Psi(\boldsymbol{x};\boldsymbol{R}) \qquad (3)$$

whose eigenfunctions $\Psi(\boldsymbol{x};\boldsymbol{R})$ depend on space and spin variables $\boldsymbol{x} = \{x_1, x_2, \ldots\}$ of all electrons. Both the eigenfunctions and the energy eigenvalues $E(\boldsymbol{R})$ of the electronic problem (3) have a parametric dependence on $\boldsymbol{R}$. In practice, before trying to solve eqn. (3) we have to assume some nuclear configuration, and thus, nuclear coordinates enter the molecular problem as the additional set of data.

The separation of nuclear and electronic motions through the Born-Oppenheimer approximation brings about a very useful, though approximate, concept of the potential energy hypersurface for the motion of nuclei [5,10]. This is the approximate total molecular energy $E_{tot}(\boldsymbol{R})$ expressed as the sum of the electronic energy $E(\boldsymbol{R})$ and the nuclear repulsion energy $E_{nucl}(\boldsymbol{R})$,

$$E_{tot}(\boldsymbol{R}) = E(\boldsymbol{R}) + E_{nucl}(\boldsymbol{R}) \ . \qquad (4)$$

The calculation of the potential energy hypersurfaces for polyatomic molecules is certainly one of primary goals of quantum chemistry. Having no room for a more detailed discussion of this subject, let us only mention that modern chemical kinetics and the theory of chemical reactions are based on the knowledge of these hypersurfaces [10-12].

Also we shall not discuss here the validity of the Born-Oppenheimer approximation [13]. For the present purposes the most important result is that molecular calculations are to be

performed for some fixed geometries of the nuclear skeleton. On repeating these calculations for different geometries one obtains a point-wise approximation to the potential energy hypersurface (4). Thus, from now on, we shall focus our attention on the methods of solving the electronic Schrödinger equation (3). In spite of the simplifications which have led to this equation, its application in calculations for many--electron molecules makes some further approximations necessary. Most of the present molecular calculations are based on what is called the orbital approximation for many-electron systems.

1. THE ORBITAL APPROXIMATION AND THE SCF HFR METHOD

In principle, one could try to obtain directly some approximate solutions of eqn. (3) by using the variation method [14] with trial electronic wave functions of sufficient flexibility. This, however, brings about enormous mathematical complications, mostly because of intractable multi-dimensional integrations over all electronic variables. These problems can be to some extent circumvented by assuming that each electron in a given many-electron system is described by its own function, which is called the spin-orbital and depends on the space ($\boldsymbol{r}$) and spin (σ) variables, $u(x) = u(\boldsymbol{r},\sigma)$ [5, 15]. Most frequently the spin-orbitals are represented as a product of the one-electron function $v(\boldsymbol{r})$ and either α or β spin function

$$u(x) \begin{cases} v(\boldsymbol{r})\ \alpha \equiv u(x) \\ v(\boldsymbol{r})\ \beta \equiv \bar{u}(x)\ . \end{cases} \tag{5}$$

The function $v(\boldsymbol{r})$ is referred to as the orbital. The approximate wave function of the many-electron system is then built as the antisymmetrized product (determinant) of different spin-orbitals for each electron. This function is referred to as the Slater determinant [5, 15]. For the ground electronic state of most molecules with 2n electrons a valid approximation to the solution of eqn. (3) is given by a single Slater determinant of the form:

$$\Psi(\boldsymbol{X};\boldsymbol{R}) = \Psi(\boldsymbol{X}) = N \det|u_1(x_1)\bar{u}_1(x_2)\ldots u_n(x_{2n-1})\bar{u}_n(x_{2n})| \tag{6}$$

where N is the normalization factor. For the sake of simplicity

the explicit identification of the assumed nuclear geometry will be omitted.

The function (6) refers to the model in which each orbital is doubly occupied. Though this is the most useful and popular model in molecular applications of quantum chemistry, its approximate nature must be carefully recognized. By introducing the orbital approximation we neglect the dynamical correlation between electronic motions (5). The improvements over the orbital model and the so-called correlation energy problem will be discussed in one of the subsequent sections.

Now, what we need to know are the best orbitals which should be used in constructing the approximate wave function (6). By applying the variation principle [14] to the energy formula obtained for the function (6) [5] we arrive at the set of integro-differential equations of the form

$$F(i)u_k(i) = e_k u_k(i) \ , \tag{7}$$

which are known as the Hartree-Fock (HF) equations [5, 15]. The Hartree-Fock operator $F(i)$ is the same for all electrons and the numbers e_k are called the orbital energies. The most important feature of the HF operator is that it depends on all occupied orbitals of a given system. Thus, eqns. (7) must be solved by using some iterative scheme.

According to eqn. (7) the many-electron problem (3) is effectively reduced to the one-electron problem. Nonetheless, the explicit integration of molecular HF equations is prohibitively difficult. A practical method for the solution of these equations was proposed in 1951 by Roothaan [16] and consists in using some analytic approximations to the HF orbitals.

Suppose that we know a set of functions $\{\chi_\alpha\} = \{\chi_1, \chi_2, \ldots, \chi_m\}$ which satisfy only some fairly general assumptions with regard to their analytic form [5]. The set $\{\chi_\alpha\}$ will be called the basis set. As proposed by Roothaan each HF orbital is expanded in the set $\{\chi_\alpha\}$, i.e.,

$$u_k = \sum_{\alpha}^{m} c_{\alpha k} \chi_\alpha \tag{8}$$

and in this way the integration of eqns. (7) is replaced by the

algebraic problem

$$\boldsymbol{F}\,\boldsymbol{c} = \boldsymbol{S}\,\boldsymbol{c}\,\boldsymbol{e} \tag{9}$$

where $\boldsymbol{c}$ is the m × m square matrix of the expansion coefficients $c_{\alpha k}$ arranged in columns for each k = 1, 2, ..., m and $\boldsymbol{e}$ is the diagonal matrix of orbital energies e_k. The square matrices $\boldsymbol{F}$ and $\boldsymbol{S}$ are defined by

$$F_{\alpha\beta} = h_{\alpha\beta} + \sum_{\mu}^{m}\sum_{\nu}^{m} R_{\mu\nu}[2(\alpha\beta|\nu\mu) - (\alpha\mu|\nu\beta)] , \tag{10}$$

$$S_{\alpha\beta} = (\alpha|\beta) = \int \chi_\alpha^*(1)\, \chi_\beta(1)\, dv_1 , \tag{11}$$

where

$$h_{\alpha\beta} = (\alpha|h|\beta) = \int \chi_\alpha^*(1)\, h(1)\, \chi_\beta(1)\, dv_1 , \tag{12}$$

$$(\alpha\beta|\lambda\sigma) = \int\int \chi_\alpha^*(1)\chi_\beta(1)\, r_{12}^{-1}\, \chi_\lambda^*(2)\chi_\sigma(2)\, dv_1 dv_2 , \tag{13}$$

and

$$R_{\mu\nu} = \sum_{i}^{n} c_{\mu i} c_{\nu i}^* \tag{14}$$

The algebraic equations (9) are referred to as the Hartree-Fock-Roothaan (HFR) equations [5,15]. Since the matrix $\boldsymbol{F}$ depends on their solutions through eqn. (14) they are solved iteratively starting with some numerical values of the expansion coefficients or the density matrix elements $R_{\mu\nu}$. By solving the matrix equation (9), which is set up by using the initial matrix $\boldsymbol{R}$ and the known values of all integrals (11) - (13), we obtain new expansion coefficients. Then, new $\boldsymbol{R}$ and $\boldsymbol{F}$ are formed and the whole process is repeated until the two subsequent density matrices differ by less than some assumed threshold. The whole procedure is referred to as the self-consistent field (SCF) Hartree-Fock-Roothaan (HFR) method [5, 15] and results in analytic approximations for the HF orbitals of a given many-electron system. This method can be generalized for systems of arbitrary number of electrons and for multideterminantal forms of the trial wave function [17-19]. However, the general processing scheme remains nearly the same as described above.