Hacking the Electronic Wave Function via Semiempirical Hamiltonians

Hacking the Electronic Wave Function via Semiempirical Hamiltonians

By

Vitaly V. Chaban

Cambridge
Scholars
Publishing

Hacking the Electronic Wave Function via Semiempirical Hamiltonians

By Vitaly V. Chaban

This book first published 2025

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-4071-8
ISBN (Ebook): 978-1-0364-4072-5

TABLE OF CONTENTS

ACKNOWLEDGMENTS

This monograph succinctly overviews a rich body of research spanning roughly ninety years. The work could have been impossible without the world's top scientists, who supported our interest in semiempirical methodologies, shared their program codes, supplied in-home parameterizations, and clarified fundamental and technical issues. The continuous communication with such dedicated researchers and educators constituted many pleasant hours of the author's professional life.

The author sincerely thanks Dr. James J. P. (Jimmy) Stewart for fruitful communication on the semiempirical themes, dedicating lots of efforts to develop and refine semiempirical Hamiltonians, and granting us the academic license for the MOPAC line of software. The author appreciates Prof. Walter Thiel for our exciting discussions on his research of electron excitations within a semiempirical formalism. The author highly values our brief but constructive correspondence with Prof. Gerd Rocha for clarifying details and philosophy of the SPARKLE model for f-elements and possible extensions of the existing parametrizations. The author appreciates Prof. Tim Clark's clarifications regarding his implementations of the parallel semiempirical Hamiltonians.

Our in-house developments are indebted to the developers of ASE and SciPy.org's rich libraries of “scientific” functions and subroutines implementing widely used mathematical algorithms. Such teamwork dramatically hastens the development of research programs and helps test new methods.

The author acknowledges Nadezhda Andreeva for helping with some copy-editing work related to this monograph.

Last, the author of this book often uses the visual interface program called Gabedit to perform significantly different actions. The author feels delighted to publicly thank Prof. Abdul-Rahman Allouche for his twenty-year-long dedication to developing and supporting such versatile and powerful software.

PREFACE

Computational research tools are on the rise nowadays. Substantial progress in semiconductor technologies made it possible to model many properties, phenomena, and processes in various sciences. Modeling and simulation help understand complex systems that cannot be described using direct equations. Computers help build and process mathematical models that contain vital parameters of the analogous physical object. Professional researchers define the properties of the building blocks along with the equations driving their elementary behavior and observe complex effects and inapparent dependencies.

Mathematical models and simulations help avoid actual experimentation since the latter can appear time-consuming, dangerous, and costly in all senses. In-silico modeling is used to solve real-world problems cheaply and time-efficiently. Properly designed simulations facilitate understanding the properties of the exciting system or its probable behavior without probing that system in the real world.

In molecular physics and chemistry, modeling and simulation are most frequently applied to the many-atom and many-molecule problems. Statistical procedures are fundamental to processing the simulation results and linking the model system's properties to the actual system's properties.

Understanding materials and the versatile processes endlessly occurring in them is a cornerstone of humanity’s progress in virtually all fields of everyday life. Microscale and nanoscale phenomena determine the physical properties of materials. Atomistic and supra-atomistic phenomena determine the chemical properties of materials. Experimental and theoretical methods of investigation work hand in hand to reveal the mysteries of matter. Microscopies provide structural properties with a reasonably high spatial resolution. Calorimetry records thermal effects. The electronic-structure modeling approaches link structures to potential energy surfaces, charge density distributions, molecular conformations, and thermochemistry.

The author of this book has substantial experience in developing, supplementing, adapting, coding, and numerous task-specific applications of computational methods in chemistry, physics, and biology from the beginning of the 2000s. In ancient times, computational reality drastically differed from the present. While the computational molecular sciences methods were known in the professional communities, the program codes were not so much available and seldom adequately documented. The boom of computational exercises inspired in the 1990s by the drastic advances in computing was on. Many researchers in natural sciences were actively learning programming languages to implement this or that promising algorithm. The computational resources were still insufficient to routinely study the problems of natural chemical and biological interest. High-level programming using high-level languages was rising and becoming fashionable in expansive communities. The cheap processors' clock speeds ranged around 100-200 MHz and random-access memory per device rarely exceeded 64 MB. The affordable hard disk volumes were below 1 GB. The internet connection speed and traffic limits exterminated the goodwill to maintain everyday collaborations with partners overseas. At the same time, the technical capabilities of the early 2000s were enough to solve local scientific problems and nurture computational philosophies.

Someday, we will start with classical molecular dynamics and Monte Carlo simulations to describe molecular and ion-molecular systems composed of roughly a thousand interaction centers. This dedication brought us to the force field development endeavors. Thus, we started to use different generations of density functional theory, Hartree-Fock, and post-Hartree-Fock methods to derive and subsequently fit interaction energies. Next, the nanoscale carbon boom happened in the middle of the 2000s. The sizes of the coveted systems increased in terms of atoms and electrons. The necessity to handle several hundred electrons per simulation drew our attention to the semiempirical Hamiltonians.

PM3 was a method of choice then, and shortly after, James Stewart published an outstanding development called PM6. While PM6 still used a few decades-old approximations – neglecting diatomic differential overlap – the method was successfully extended to transition elements and incorporated a few corrections. PM6 had been a considerable advance over all previous semiempirical methods from the end-user's perspective since it enormously expanded the accessible model chemistry. Neither density functional tight binding nor reactive force fields (both mentioned methods aim to be computationally cheap alternatively to the ab initio methods) gave us any comparable flexibility to explore the problems of materials science. The semiempirical Hamiltonians provided us with encouraging results for carbon nanotubes, functionalized fullerenes, quantum dots, chemically modified graphene sheets, competitive ion binding to proteins, many-component ion-molecular systems, non-aqueous solvation effects, high-energy chemistry, nanoscale clusters of gold, copper, silver, and platinum.

The ultimate accuracy of the semiempirical predictions may be readily debated. In some unusual systems, our semiempirical predictions were indeed canceled by hybrid density functional theory verifications. In the meantime, we could semiquantitatively assess the thermodynamic stabilities of such “border” compounds as cubanes, polynitrocubanes, and carbon-nitrogen cubanes. Organic chemistry is traditionally better modeled by semiempirical methods. In turn, inorganic crystals and newly designed nanoscale structures exhibit less bright coincidence with the results of ab initio methods. More thermodynamically stable particles are more suitable to be simulated by the semiempirical Hamiltonians. Critical features of molecules are more straightforward to capture than those of ions.

PM6 was followed by a few fascinating extensions, each introducing significant improvements for a real-world molecular design. For instance, the extension of PM6 to include empirical correction for dispersion attraction termed PM6-D3 allowed chemical researchers like us to simulate more than one hydrophobic molecule per system. The subsequent extensions to accurately reproduce hydrogen bonding termed PM6-DH+ and PM6-DH2 made the semiempirical Hamiltonian even more competitive in simulating organic and bioorganic systems, including explicit solvation effects. Ultimately, the PM6-DH2X extension was elaborated. It added the empirical terms to describe better energetics and geometrical parameters connected to halogen-oxygen and -nitrogen interactions, i.e., halogen bonding. Even certain chemical limitations associated with applying these solutions to numerous systems of high research interest did not blur an outstanding leap forward that the semiempirical Hamiltonians have made.

The scientific paper by J. J. P. Stewart validating the performance of PM7 was published in 2013 and was branded by the author as a modified version of PM6. PM7 contained a new set of atomic and diatomic parameters and implementation improvements. PM7 could boast of generally higher accuracy than all other semiempirical Hamiltonians. The new method continued to neglect the diatomic differential overlap. We were happy that the developer shared our vision regarding a semiempirical Hamiltonian as an affordable tool to palpate large chemical and biological systems. In that way, PM7 became the only electronic-structure method that could model biochemistry and support drug-related studies.

In parallel with the electronic-structure research projects, we continued to be avid molecular dynamics simulators using classical and ab initio Hamiltonians to describe the potential energy surface of the simulated system. Therefore, we understood the importance of averaging the simulation results concerning molecular thermal motion. The simulated thermal motion allows the system to navigate the potential energy surface and sample the states that the local geometry optimization algorithms could never capture. The inclusion of the thermal motion effect is necessary to obtain a reliable theoretical description of the systems in the liquid aggregate state and inherent flexibility.

In 2014, V.V. Chaban implemented a straightforward interface between PM7 and selected molecular dynamics routines, such as integrating nuclear equations of classical motion, temperature coupling, and boundary conditions. The resulting method, termed PM7-MD, used the Born-Oppenheimer approximation, i.e., electrons and nuclei moved independently without changing energy among one another. We applied PM7-MD to numerous chemical problems and published state-of-the-art results. The principal difference between the PM7-MD and PM7 studies was the presence of nuclear kinetic energy and unrestrained or semi-restrained phase space sampling. In PM7-MD, the simulated system could wander over its entire phase space (i.e., the combinations of coordinates and momenta) to the extent of the accessible potential barriers. This feature of PM7-MD made it possible to explore multiple regions near the system's stationary points within the same simulation and generate the most realistic atomistic configurations.

In later years, we continued our intensive investigations of the possibilities that the PM7 Hamiltonian offers. In 2015, V.V. Chaban combined PM7 with a peculiar annealing procedure to range possible, stable geometries of nanoscale gold clusters, silver clusters, doped clusters, semiconducting quantum dots, liquid systems, and coordination compounds. The elaborated method shows perspectives that work together with experimental research to decrypt the geometries of rigid inorganic structures. In 2021, we published the results obtained by another simulation method using the PM7 Hamiltonian as its engine. The instantaneous kinetic energy injections were employed to locally perturb the otherwise realistic molecular trajectory and thereby identify hard-to-locate stationary points of the simulated system.

Someday, we realized we had accumulated much information regarding the semiempirical Hamiltonians, their evolution, and preferential applications that needed to be summarized and shared with the community. We do this in the hope that our vision of the semiempirical methods' place in computational physics, chemistry, and biology and our proven expertise in building techniques on top of the available Hamiltonians may inspire intensive research efforts in these directions. We honestly tried to minimize the mathematical content in our narrative. Most of the computational details in a semiempirical calculation are technical and would be better learned from computer codes than books. Instead, we opted to shed light on the chemical essence and perspectives to enhance the prediction power of the semiempirical methods.

The history of the semiempirical methods is rich, and the achievements are vivid. It is impossible to mention them in a single monograph that aims to be concise and pursue specific goals. The author apologizes for the omitted semiempirical methods and methodological breakthroughs that might have obtained less attention than they deserve. We did not give enough attention to MNDO-d, and subsequent efforts to extend the basis set since these developments were de facto summarized and included in the latest semiempirical methods, such as PM6 and PM7. We largely missed the methods to simulate excited states, such as ZINDO by M. Zerner, SINDO by K. Jug, and PPP by Pariser, Parr, and Pople. Density functional theory is a platform for simulating excited electronic states and electron relaxation processes nowadays.

We only briefly mentioned efforts to add electron correlation in the framework of the NDDO formalism. Such implementations were computationally expensive and were ultimately combated by the correctly performed parametrizations of more traditional semiempirical Hamiltonians. SAM1, developed by M. J. S. Dewar and coworkers in 1993, was reported to show superior results in those times, but its development was primarily phased out after the death of Dewar. SAM1 is not actively used nowadays.

We opted to exclude some new Hamiltonians, such as hpCADD, MNDO-F, and ad hoc modifications of the well-known methods, from our consideration. hpCADD is devoted to computing molecular electrostatic potential, but its application to real-world systems is associated with restrictions. MNDO-F extends classical MNDO with self-consistent, anisotropic, multi-center dispersion correction. More time is needed to understand the potential of these methods. An interested reader is undoubtedly referred to respective research papers.

Our discussion is inevitably biased toward the modern Hamiltonians, whose implementations allow simulating numerous chemical elements simultaneously, adding empirical corrections, and incorporating them into extended studies, such as QM/MM, MOZYME, PM7-MD, and global energy minimum search. One of the most appealing points of such methods is their computational affordability, resulting in the ability to palpate interesting superstructures and phenomena swiftly.

This monograph might serve as a succinct but transparent guide across all tangible semiempirical Hamiltonians developed over ninety years. As such, it may interest graduate students and non-professionals in quantum chemistry. Nevertheless, the monograph is about what we can do using semiempirical Hamiltonians and related numerical methods here and now.

On the following pages, we provide definite answers and exemplify robust research recipes. We discuss the analysis protocols that intersect directly with the interpretation of chemical and biological experiments. We share our real-world approach to unraveling the prediction capabilities of modern semiempirical Hamiltonians. The monograph represents professional work in computational natural sciences and addresses experts in our primary research field.

CHAPTER 1

MODELING IN THE HISTORY OF HUMANITY

It would be reasonable to state that nascent modeling as a tool for learning about the universe originated before the Common Era. Modeling was a natural supplement to the empirical exploration of the problem, wherein the latter was not complete enough. The cumbersome requirements of an experiment might readily stipulate a temptation to substitute it with a decently elaborated model irrespective of the considered problematics.

During the ancient times of Democritus and Epicurus, there were ideas about the existence of non-divisible particles ("atoms") and their shapes. Based on these models, superficial attempts were made to explain the physical properties of various then-known substances and materials. Modeling efforts and understanding of elementary particles by the ancient Greeks was a reasonably primitive exercise, in any case. Atoms, as defined by the ancient Greeks, have little in common with atoms as objects of atomic molecular theory, dating back to the 19th century. The vectorizing role of such philosophic concepts should not be overestimated. What seems truly important about the ancient Greeks is the ability of human thought to pick up analogies and extrapolate observations.

The information about ancient modeling is generally scarce, but the preserved mathematical advances associated with ancient scientists permit us to believe that they undertook specific exciting projects. Much later, Isaac Newton and his contemporaries used modeling even more consciously to verify the laws of nature and solve essential problems in astronomy. At those times, early scholars had already understood that the motion of stars and planets could be accurately predicted for decades by applying relatively simple procedures of mechanical modeling. By mathematical modeling, we now understand the search for a solution without an analytical recipe.

Modeling was inseparable from the formulations of laws in classical physics and heavily fostered the development of mathematics. One can quickly learn from the history of science that mathematical advances routinely and substantially forestalled experimental physics. These mathematical developments signify the role of this or that sort of modeling as a dominant research practice over centuries. Since it is frequently tricky to reliably differentiate between simulations and so-called thought experiments, some of the research reports of the 1700s and 1800s can be perceived as not legit modeling.

Modeling was widely used in the Renaissance when famous Italian architects and sculptors developed reduced models to design and advertise peculiar buildings and structures. Such models could be used to trial the mechanical endurance of extravagant projects. Both G. Galilei and L. da Vinci used models to learn about various aspects of nature. These researchers were looking for the limits of applicability of modeling methods and models, among other things. Modeling is traditionally exercised in modern architectural projects to advertise the idea and test its possible drawbacks.

In the 19th century, modeling was used routinely in line with the vigorous progress of sciences and technologies. It is not easy to find a field of research in which simulations have not been applied. Based on specific flavors of modeling, the works of such scientists as Kelvin, Maxwell, Kekulé, Butlerov, and many others were carried out. These scientists contributed fundamental discoveries, some of which were ultimately inspired by applying modeling techniques and considering analogies. Reading the works mentioned above is enlightening and inspiring.

As sometimes referred to in educational materials, electronic computer simulation methods were developed after a large-scale presentation during the Manhattan Project on September 17, 1943. This cornerstone simulation dealt with the nuclear detonation process. The two mathematicians John von Neumann and Stanislaw Ulam were solving a poorly understood problem: the behavior of neutrons. The nuclear breaking experiments were relatively expensive and not optimally controlled. In turn, the technical problem was complex enough. The mathematical method "Roulette" was formulated. The authors of the method suggested that using the relationship between stochastic processes and differential equations in the opposite direction is possible. The mathematicians proposed using the stochastic approach for approximating multidimensional integrals in transport equations. The latter arose in connection with the problem of neutron motion in the isotropic medium. The developed method was relatively successful in mapping the nuclear experiments and confirming technological developments from the theorist's perspective. Simulations in the electronic computer, in-silico modeling, received impetus for further development [1]. Numerous problems in physics that do not have analytic solutions but can be represented via a system of differential equations got within reach first in history.

The invention of computer prototypes in the middle of the last century played a significant role in computationally intensive modeling. The development of electronic computers that could generate pseudorandom numbers at high speed dramatically expanded the range of problems for which the application of stochastic methods could be fruitful. A notorious breakthrough was the Monte Carlo method. The methods based on the Metropolis algorithm have been developed and successfully applied to many practically significant problems [2].

The Monte Carlo method is a mathematical advance that can be applied to many scientific and semi-scientific problems in which elements of stochasticity are present. The usage of Monte Carlo goes far beyond physics and chemistry. Many processes happening in the world can be described by employing random probability distributions. Therefore, they can be analyzed statistically even though precise predictions of their corresponding constituent actions are impossible.

In modern times, computer simulations are often used as a supplement or replacement in research cases where analytical solutions are impossible. A separate large group of computational studies is optimization problems. Therein, a function of many variables is optimized iteratively to obtain a set of data best suiting some user-defined criterion. Such results are widely used in physical sciences and technological solutions to many-dimensional problems.

Many types of computer simulations exist nowadays. Their common feature is an attempt to identify a set of typical scenarios for a model, whereas the enumeration and subsequent systematic trials of all scenarios are impossible. The typical scenarios must be obtained for a particular combination of external conditions, such as an applied electric field or anisotropic pressure.

Computer simulation methods of physical and chemical processes occupy a unique niche. These are used to predict the structures and properties of molecular or submolecular systems in the gas-state and condensed phases. Computer simulation methods in physics, chemistry, and biology are powerful and sophisticated. Their prediction capabilities allow using numerical simulations instead of many routine experiments, whereas the results can typically be trusted. If used professionally, simulations always issue results that at least qualitatively agree with the analogous experiments. The due statistical processing is a cornerstone of computer simulation’s validity. Statistical mechanics is a theoretical area of physics that forms a foundation for thoughtful simulations.

Electronic-structure modeling is based on systems of electrons and nuclei assembled in atoms. Atoms form chemical bonds with adjacent atoms by communalizing their valence electrons. Electrons are particles and waves simultaneously, resulting in their permanent delocalization in space.

Electrostatic attraction and repulsion forces are solely responsible for the electron-nuclei systems' stable states (stationary points). The electronic-structure numerical simulation methods aim to compute the potential energies of every electron in the atomistically-precise system. These electron energies reflect the distance of electrons from their respective atomic nuclei. The high potential energies of the electrons imply that they are located farther from their atomic nucleus. The valence electrons possess higher potential energy than the core electrons. Modern electronic-structure methods handle the systems of hundreds and even thousands of electrons. Therefore, small-sized molecules, middle-sized molecules, and even specific molecular assemblies can be successfully simulated. An entire protein is often referred to as a landmark of modern simulation methods advance. There are plenty of descriptors apart from the blunt system's size to judge the computational cost of the computer simulation. The most essential criteria are the number of electron energy levels and sampling times in the case of simulated dynamics. Simplifications introduced into the computational procedures allow for further increases in the maximum size of the simulated system.

Semiempirical methods use the mathematical procedures of electronic-structure methods but benefit from specific parameterizations. The parameterizations can be performed based on the experimental data, high accuracy, costly ab initio calculations, and more. The results of the parametrizations are double. The procedure of the wave function calculation, including the smoothness of convergence, accelerates because some integrals do not need to be evaluated as in the ab initio methods. In turn, the accuracy of the calculations increases – at least, it must – because the supplied parametrizations reflect reliable experimental data. They implicitly include all effects that the theory misses. Although the situation is not seamless in practice, the group of semiempirical methods represents an exciting advance in computational chemistry. Semiempirical methods can be fruitfully used in chemical research and provide valuable insights into the problems studied at a reasonably cheap computational cost.

In the present compendium, we discuss the history of the electronic-structure semi-empirical methods and swiftly outline their role in chemical and physical simulations. We try to go formula-free or nearly formula-free. Instead of concentrating our narrative on the methods' anatomy, we envision, discuss, and exemplify their possibilities and perspectives. We explain why semiempirical methods are arguably the most robust tool to start palpating nearly any atomistically precise problem involving a significant number of electrons in its elementary action. We, however, never hesitate to underline that any computational method must be applied consciously, whereas its area of applicability must be appropriately justified by extensive benchmarking. Selective comparison of semi-empirical results to the analogous output of generally more accurate and mathematically precise electronic-structure methods is always praised and encouraged.

CHAPTER 2

BASIS OF THE SEMIEMPIRICAL METHODS

In 1926, Erwin Schrödinger published an equation describing the change in space and time of the quantum state of a system:

$$i\hbar\frac{d}{dt}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle. \qquad (1)$$

Here, *t* is the time, Ψ(t)> is the state vector of the quantum system, whereas $\hat{H}$ is the Hamiltonian operator.

The wave function has been thereby introduced. The derivation of Schrödinger's equation was preceded by a long period of development in physics, particularly experiments with electron excitations and scattering and debates about the nature of an atom. The solution of Schrödinger's equation by analytical means is possible only in the most straightforward cases. Such precisely soluble cases include one-electron atoms, i.e., hydrogen, helium cation, lithium doubly-charged cation, a linear oscillator, a potential well with infinitely high walls, and a few similar setups. In the case of many-electron systems, no exact solutions to Schrödinger's equation exist, being its immanent mathematical property. There is no doubt that numerous types of approximate solutions are possible. The Hartree-Fock method is one of the oldest approximate methods to find a solution for many-electron systems [3-5].

The origin of the Hartree-Fock method dates back to the late 1920s, shortly after the introduction of the Schrödinger equation in 1926. Douglas Hartree published a straightforward mathematical procedure called the self-consistent field method for approximate wave function calculation. The procedure allowed us to evaluate the atoms and ions' total energy and electron energy levels. The history of quantum chemistry began at that point. Hartree used the developments achieved in the older Bohr quantum theory [6-8].

In Bohr's model of an atom, the energy of the state with the principal quantum number "n" is given in atomic units as $E = -\frac{\mathbf{1}}{\mathbf{n^2}}$. Atomic spectra guided the development of the mathematical engine. It was known that the energy levels of the multi-electron atoms are adequately described using a modified version of Bohr's formula. By introducing the quantum defect "d" as an empirical parameter, the energy levels of an atom were described by the formula $E = -\frac{\mathbf{1}}{(\mathbf{n+d})^2}$. We note that it was possible to reproduce transitions observed in the X-ray range satisfactorily. The existence of a non-zero "quantum defect" was associated with electron-electron repulsion, which was inexistent in the isolated hydrogen atom. This repulsion results in a partial screening of the nuclear charge. Based on the mentioned early studies, other potentials were later introduced. The latter contained more empirical parameters in the hope of a better reproduction of the experimental data [9-10].

D. R. Hartree did not feel comfortable with the empirical parameters within seemingly a first-principles algorithm. The goal was to solve the multi-component time-independent Schrödinger's equation from fundamental physical principles, i.e., ab initio. The supplied parameters should not have gone beyond the electron and nucleus charges. The first solution method became known as Hartree's method. However, many of Hartree's contemporaries did not recognize the physical reasoning behind Hartree's method. The solution still contained empirical elements. Furthermore, its connection to the solution of the many-electron Schrödinger's equation was not evident. In 1928, J. Slater and J. A. Gaunt independently showed that Hartree's method could be formulated on a sounder theoretical basis by applying the variational principle to the trial wave function (ansatz) as a product of one-electron functions [9-10]:

$$\Psi = \Psi_1(1)\, \Psi_2(2) \ldots \Psi_n(n). \tag{2}$$

The above wave function could have been an exact solution to Schrödinger's equation if the Hamiltonian had no electron-electron repulsion component. Hartree took this component into account in an expression that described the repulsion of an electron from some averaged self-consistent field of other electrons. In other words, the exact electron-electron repulsion is replaced by the averaged field created by the remaining electrons. Such an assumption divides the function in the following way:

$$\left(-\frac{1}{2}\nabla_i^2 - \sum_I^N \frac{Z_I}{r_{Ii}} + V_i^{\mathrm{eff}}\right)\Psi_i = E_i \Psi_i\ . \tag{3}$$

Thus, a separate equation shall be written for each electron in the modeled system.

In 1930, J. Slater and V. Fock pointed out that Hartree's method did not account for the wave function's antisymmetry principle. In turn, Hartree's method used Pauli's exclusion principle in its older formulation, which forbids the presence of two electrons in the same quantum state. However, it turned out to be fundamentally incomplete in its neglect of quantum statistics [5,11].

J. Slater proposed to use completely asymmetric wave functions for electron systems as determinants. The latter representation is presently known as Slater's determinant [9,12]:

$$\Psi=\frac{1}{\sqrt{n!}}\begin{vmatrix}\Psi_1(1) & \Psi_1(2)\ldots & \Psi_1(n)\\ \Psi_2(1) & \Psi_2(2)\ldots & \Psi_2(n)\\ \ldots\ldots & \ldots.\ldots & \ldots\ldots \\ \Psi_n(1) & \Psi_n(2)\ldots & \Psi_n(n)\end{vmatrix} \tag{4}$$

Herein, n is the number of electrons, whereas the normalizing coefficient is $\frac{1}{\sqrt{n!}}$. It was shown that Slater's determinant, the one-electron orbital determinant pioneered by Heisenberg and Dirac in 1926, satisfies the antisymmetric property of the exact solution. It is, therefore, a suitable ansatz for applying the variational principle. The original method of Hartree can then be perceived as an approximation to the Hartree-Fock method in which the exchange is neglected. Fock's original method relied mainly on the group theory and looked somehow unphysical. The next iteration occurred in 1935 when Hartree reformulated an initial method for numerical calculations [3-4,12-14].

The ready-to-use Hartree-Fock method required significant computational efforts to describe even relatively simple chemical systems. Its usage for a couple of decades after publication was reasonably scarce. After the development of computers, the first computationally intensive works appeared in the 1950s. The researchers only applied the Hartree-Fock method to the isolated many-electron atoms for a long time. The atoms were described as spherical entities. Therefore, the system's symmetry was used to simplify equations further and decrease the associated computational expenses. The mean-field approximation plays a paramount role in such types of calculations. Manual solving the Hartree-Fock equations for an atom, like Fe, Cu, or Ag, was laborious. In turn, calculations on small molecules and ions required considerable computational resources far beyond what had been available in the 1950s.

CHAPTER 3

GRADUAL ADVANCE OF SEMIEMPIRICAL METHODS

The semiempirical electronic-structure methods make use of the Hartree-Fock equation. However, they introduce numerous approximations into the algorithm. Furthermore, some parameters have been obtained (recalculated) from the atom-level experimental data and ab initio calculations. The semiempirical methods are much quicker than the Hartree-Fock method. In turn, it is more difficult to parallelize their procedures since they became anisotropic due to the ad hoc parameterizations. Thousands of electrons per system can be simulated without divide-and-conquer approaches. The ability of the semiempirical methods to solve the Hartree-Fock equations for large systems coupled with the gradient minimization algorithms, i.e., stationary point search schemes, might be invaluable in molecular biology. Such an algorithm could find preferable geometries of proteins and folding states of other macromolecules (Figure 1). Semiempirical methods could promote research in molecular genetics by identifying and rating available docking spots. In principle, parameterization represents a continuous and infinite process that allows new semiempirical methods to emerge over time and implement more advanced parameters in their procedures.

Using empirical parameters theoretically makes it possible to consider all electronic effects simultaneously. Recall that the Hartree-Fock method is strictly limited by a one-electron approximation that significantly limits its applicability in some physical issues. Some integrals in the Hartree-Fock equation (for example, the two-electron integral) are approximate. In some cases, they are entirely omitted when a proper justification exists. The parameterization of the semiempirical methods means that the selected calculated descriptors coincide with the experimental results. In the latest generations of semiempirical methods, the results of the high-precision ab initio methods have also been used to pick up the most suitable parameterizations. Thanks to that, the parameters for a more significant part of the Periodic Table were derived even though not all parameters are of the

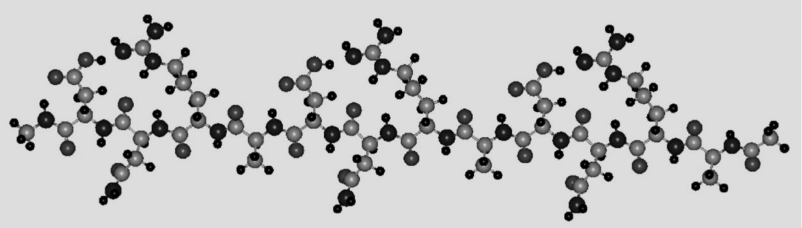

Figure 1. One of the possible stationary point geometries of the [ALA-ARG-ASN-ASP]$_4$ oligopeptide. The empirical formula of the simulated system is $C_{54}H_{91}N_{25}O_{22}$. The total number of atomic nuclei in the system, including the hydrogen nuclei, amounts to 192. The total number of valence electrons is 564. The standard heat of formation of the optimized system (oligopeptide) equals -4325 kJ mol^{-1}. We assume that folded structures of this oligopeptide exhibit lower potential energies than the unfolded oligopeptide. The PM7 Hamiltonian was employed.

same level of accuracy. Sometimes, the discrepancies connected with certain heavy elements are frustratingly significant. The newest semiempirical methods represent a convenient tool for conducting an initial assessment of the problem and outlining the research schedule using more accurate computational and experimental methods. The performance of the semiempirical methods is good in the case of low-energy states. If the system's structure is overly perturbed, convergence issues may be expected.

The same external factors boosted the progress of semiempirical methods as other computational tools in physics and chemistry. First, developing computing devices and programming languages created a natural excess of processor power at research universities and algorithm implementation opportunities. Second, the accuracy of experimental methods grew, and large sets of data were accumulated that could benefit from direct and indirect comparisons with the computational results. In the 1980s and the 1990s, the available processor power became enough to model small molecules, up to 100 electrons, quantum-chemically. Applying the Hartree-Fock and emerging density functional theory methods was still very expensive, although it was possible for selected benchmarking studies. In turn, the semiempirical calculations appeared affordable for routine usage in chemical applications.

Furthermore, the structural and thermodynamic results of the semiempirical calculations could be selectively mapped to ab initio results. Namely, the advancement of silicon technologies and the dynamics of their development at the end of the last century fostered the emergence of more sophisticated semiempirical codes like MOPAC (openmopac.net). Most of the obtained results were in reasonable agreement with the reference data and provided at least a qualitative understanding of the studied phenomena. In the 2000s, when parallel computations emerged, the semiempirical calculations partially lost their popularity in favor of density functional theory methods, which offered more significant scaling factors in the supercomputing environments and provided more accurate results in the case of unusual systems, excited-state configurations, and potential energy surface scans. The semiempirical calculations decreased their fraction in atomic and molecular physics but significantly increased the number of their applications in organic chemistry and molecular biology, where larger systems must be analyzed at once. Molecular dynamics simulations based on the semiempirical Hamiltonians, such as PM7, filled the gap between ab initio molecular dynamics for minimal systems and very short simulation times and classical molecular dynamics simulations that could treat many thousand interaction centers but failed to describe many-component systems with diverse non-bonded forces adequately.

CHAPTER 4

HÜCKEL METHOD AND EXTENDED HÜCKEL METHOD

Erich Hückel can boast of being the author of the first semiempirical method that allowed computations with accurate numerical results at the end [15-17]. In the 1930s, Hückel published an approximation of a molecular orbital ψ_μ as a linear combination of atomic orbitals χ_p, MO LCAO:

$$\psi_\mu = \sum_p c_{p\mu} \chi_p. \tag{5}$$

Due to its noteworthy simplicity, humans could do the corresponding calculations directly, even for relatively complex molecules. MO LCAO was a fascinating development at the beginning of the 1930s, first allowing us to imagine the power of quantum chemistry.

The Hückel method has been used to determine the energies of molecular orbitals of π-electrons in the conjugated hydrocarbon systems, such as ethylene, benzene, and butadiene [15-17]. Subsequently, the method was extended to conjugated molecules, such as pyridine, pyrrole, and furan. Extending the Hückel method to heteroatoms was essential in computational organic chemistry [18]. Presently, there are implementations of the Hückel methods that can deal with many elements.

The significant features of the original Hückel method are the limitation to conjugated hydrocarbons and the account of only π-electrons of molecular orbitals. Hückel correctly understood that, namely, valence electrons determine critical properties of molecules, whereas the σ-electrons might be bluntly omitted. The Hückel method provides sound results only for flat equilibrium geometries. Orbital symmetry is used as an input parameter of the Hückel method, thereby greatly simplifying necessary calculations. The matrix elements of the Fock operator are rigidly fixed. They are estimated based on external (experimental) data.

The Hückel method is limited to predicting the energies of molecular orbitals in conjugated organic-chemistry systems. The Hückel method does not provide predictions regarding the organic molecule's structure.

Contrariwise, the experimentally known structure of a molecule is used as an input parameter of the method.

As soon as the parameterization of MO LCAO is trustworthy, the relative potential energies of the valence orbitals in the systems containing s- and p-electrons can be obtained. Therefore, it is possible to identify the bands in the electronic absorption spectra of such molecules by identifying the valence orbital and transition energies.

Roald Hoffman developed the extended Hückel method in 1963 [18]. While the original method only considered π-orbitals, the extended method included σ-orbitals. It explicitly took into account the overlap integrals [19]. The extended Hückel method determines the molecular orbitals and geometries of organic molecules. It involves calculations of electronic interactions in a relatively simple way. The interelectronic and internuclear interactions are omitted, and the total energy is the sum of the terms for each electron in the molecule.

The extended Hückel method can be seen as the most straightforward, fastest, and least accurate semiempirical quantum-chemical method [18]. Its application is limited mainly to the analysis of the structure of molecular orbitals. The determination of the molecular orbital's shape and sequence is straightforward. This approach appropriately calculates systems with a uniform electronic density distribution. An evident example is hydrocarbons, which are, however, an extensive group of compounds that constitute the basis of oil and gas chemistry. Extensive approximations applied led to incidental errors. The widely exemplified case is the impossibility of benzene as a molecule whatsoever. According to the extended Hückel calculation, benzene must readily decompose into three acetylene molecules. Moreover, the discussed decomposition reaction must be strongly exothermic, as computed based on the total energy difference between the product and the reactant. These thermodynamic inaccuracies and similar flaws cast doubts about the large-scale application of the extended Hückel method to explore organic chemistry.

The inability of early semiempirical Hamiltonians to deal with peculiar organic structures fostered the development of new philosophies and approximations to Hartree's theory. The newer semiempirical methods describe the versatile properties of the benzene molecule without any problems. We computed electronic structures and optimized geometries of benzene using the six modern methods, particularly MNDO, AM1, PM3, RM1, PM6, and PM7 (Figure 2), spanning over forty years of development. The obtained standard enthalpies of formation are +89 kJ mol^{-1} (MNDO), +92 kJ mol^{-1} (AM1), +98 kJ mol^{-1} (PM3), +96 kJ mol^{-1} (RM1), +101 kJ mol^{-1} (PM6), and +96 kJ mol^{-1} (PM7). Compare all these results to the experimental value of +49 kJ mol^{-1}. Interestingly, there is no gradual improvement from MNNO to PM7, while all modern methods perform better than the extended Hückel method. In the meantime, the result differs in all methods employed. Such a trend or no trend originates from the parameterization procedures, which aim to minimize an average discrepancy of many properties from the experiment rather than bring specific results to the reference values.

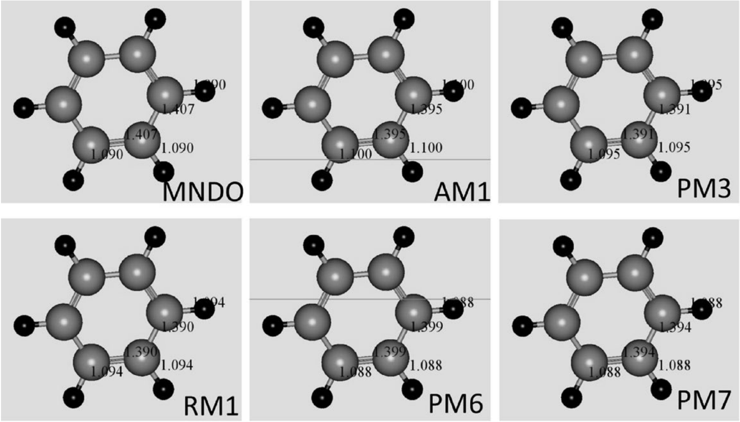

Figure 2. The optimized molecular structures of the benzene molecule for which the early semiempirical Hamiltonians provided inconsistent results on the equilibrium geometry and formation enthalpy. All electronic structure calculations used an isolated (placed in vacuo) benzene molecule. The carbon-carbon and carbon-hydrogen covalent bond lengths are given explicitly in the molecular images to compare the employed methods directly. All used semiempirical Hamiltonians adhere to the neglect of diatomic different overlap approximation.

Benzene has always been a hard-to-simulate molecule. The thermodynamic stability of benzene is still captured by semiempirical methods, which have reasonably mediocre quality. The heat of formation provided by the newest PM7 Hamiltonian is nearly two times higher than the reference property. However, the benzene molecule is presently stable. Thus, such a three-dimensional structure corresponds to the stationary point on the semiempirical potential energy surface. The cases of molecules that were not stable in the semiempirical calculations but existed in reality systematically decreased over the years in the course of the theoretical developments and more extensive parameterizations. Using less abrupt simplifications in the modern semiempirical methods compared to the extended Hückel method and implementing the d-orbital functions into the basis sets is a cornerstone of better results. The extended Hückel method was remarkably computationally cheaper than the Hartree-Fock method but suffered from systematic oversimplifications. Only the development of processor units made it possible to derive more robust methodologies and balanced approximations that were precise and affordable.

The extended Hückel method is surprisingly satisfactory for heteroatom-containing organic systems. The method accounts for the dependence of the Hamiltonian on the partial charge of the interaction center, whereas the dependence is assumed to be linear. Irrespective of the undoubted didactic benefits of the Hückel and extended Hückel methods, none offer the essential predictive potential to be applied as a sole research tool. In modern times, the routine usage of Hückel-based approximations of the electronic wave function is possible for personal records only. We strongly doubt whether any respected scientific journal agrees to accept the paper whose principal results are obtained via the Hückel methods.

CHAPTER 5

METHOD OF NEGLECT OF DIFFERENTIAL OVERLAP

As we have already stated in previous chapters, the development of semiempirical methods was modulated by the available computational power. In the case of the first theoretical attempts, the clock speeds of the electronic computers were frustratingly non-commensurate with the size of the simulation problems in chemistry and biology. Nevertheless, some problems related to the many-electron isolated atoms could be tackled.

Let us discuss some performance descriptors from those times and how they have progressed until now. The first general-purpose computing machine, ENIAC, contained a 0.1 MHz clock in a cycling unit. In other words, ENIAC was well over 10,000 times slower than any modern office PC. Each instruction took twenty cycles on ENIAC, amounting to the instruction rate of 0.005 MHz. However, many problems in natural, technical, and economic sciences genuinely benefited from the first computers. The latter solved a sort of task that was principally untractable before. A philosophical joke saying that a computer helps solve problems that do not exist without a computer emerged later. The drastic limitations in computational power provoked the development of highly efficient algorithms to explore problems with clearly absent analytical solutions. The methods associated with Monte Carlo simulations in various fields of science and technology are certainly among the most recognized achievements in applied mathematics.

Significantly later, the first mass-produced electronic computer could perform two million cycles in a second. It was a famous Altair 8800 brand developed by MITS, equipped with an 8080 central processing unit from Intel. As time proceeded, new technologies and engineering solutions swiftly boosted processor speeds. In the 1970s-1980s, computer clock speeds roughly doubled yearly. Around 1980, the best Intel processors reached a frequency of 5 MHz.

Only at the end of the 1990s, two decades later, did processor clock speed surpass the landmark of 1.0 GHz. The high-end Athlon processor