Historical Overview of Molecular Modeling

outtake from <u>Molecular Modeling on the PC</u>

"By copying, the ancient models should be perpetuated."
—Hsieh Ho, fl 500, from *Spirit of the Brush*

Drawn representations of chemical structures are among the earliest forms of molecular model, having arisen in the mid-nineteenth century in the course of the development of the structural theory of organic chemistry. From the concept of the molecular formula and a knowledge of functional groups, or "radicals," molecules were depicted as "rational formulas." These followed a variety of conventions in grouping the "radicals" together with brackets or parentheses, but essentially failed to provide a sense of molecular structure as we know it today. Between 1858 and 1861, the three chemists Archibald Scott Couper, Friedrich August Kekulé von Stradonitz, and Aleksandr Mikhailovich Butlerov independently introduced the general rules of valence for organic chemistry, and the first written structures involving chains of carbons with lines drawn as "bonds" to substituent atoms and groups. The term "chemical structure" was first used at this time.

Also in 1861, Johann Josef Loschmidt produced an amazingly prescient collection of 368 molecular structures as graphic displays with atomic domains. These included the first "correct" structure of benzene, showing a uniform cyclic arrangement of the ring carbon atoms. Also described were 120 other aromatic compounds, cyclopropane, and representations of the vinyl and allyl moieties. Loschmidt's structures depicted relative atomic size, and varied bond distance with bond multiplicity. Although we can appreciate its value from our modern vantage point, Loschmidt's contribution was published obscurely and never properly reached the attention of the chemical community of his day. In his seminal 1865 paper on the structure of benzene, Kekulé mentions Loschmidt only in passing, and decades would pass before structural theory would catch up.

Starting in 1860, Louis Pasteur embellished the theory of molecular structure with the third spatial dimension as a result of his work on optically active compounds. The first recorded use of a physical molecular model in organic chemistry was by August Wilhelm Hofmann in 1865. In a lecture entitled, "On the Combining Power of Atoms," before the Royal Society of Great Britain, he used the metaphor of croquet balls joined by sticks to describe methane, chloroform and other compounds of carbon.

The color scheme which he established is still widely used today, with occasional variations: white for hydrogen, black for carbon, red for oxygen, blue for nitrogen and green for chlorine. The Institute of Physics has attempted to promulgate a uniform color code for molecular models which, in addition to the Hofmann usage mentioned above stipulates yellowish-green for fluorine, light green for chlorine, mid-green for bromine, dark green for iodine, yellow for sulfur, purple for phosphorus, black or gray for silicon, and brown, silver or gold for metals.

Two-dimensional drawn versions of ball and stick models were employed by Crum-Brown in 1865, and Sir Edward Frankland and B. F. Duppa in 1867, although they continued to equivocate on the significance of such structural representations. Further elaborations on three-

dimensional molecular structure came with the suggestions in 1869 by E. Paterno, and in 1874 by Jacobus Henricus van't Hoff and Joseph Achille LeBel that carbon compounds have tetrahedral geometry. Paterno had published structural drawings depicting rotational isomers of 1,2-dibromoethane, but did not anticipate interconversion. C. A. Bischoff suggested energy barriers to rotation about carbon-carbon bonds in 1890, and in 1898 van't Hoff elaborated upon this with the idea of there being a "favored configuration" about a torsional angle, which we would understand today as the most stable conformation.

The concept of the **force field** originated in the first half of the twentieth century from vibrational spectroscopy, which considered the forces acting between every pair of atoms in the molecule, or in a lattice in the case of ionic crystals. A formulation which later had a significant effect on molecular modeling was that of Urey and Bradley, in which they wrote quadratic **Hooke's Law** potential equations to describe some of the harmonic vibrations in simple molecules, but found the **Morse potential** to give the best fit to empirical data for bond stretching. This early history of force fields is covered well in Burkert and Allinger.

The force field concept did not have a great impact upon the non-physical chemical community until the year 1946, when the first suggestions for the use of a new method for modeling molecules in a more quantitative way arose. It was based upon a combination of steric interactions and a Newtonian mechanical model of bond stretching, angle bending, and torsional vibrational modes. Three separate groups proposed versions of what would be later known as the **molecular mechanics** method, or the empirical force field method, for calculating molecular structures. Hill's force field had a Lennard-Jones 6-12 potential to handle the repulsive and attractive nonbonding terms, and quadratic terms for bond stretch and angle bend. Dostrovsky, Hughes and Ingold focused on the repulsive and attractive nonbonding terms, in a study on substitution and elimination reactions. Westheimer and Mayer analyzed the conformations of hindered biphenyls, formulating the potential energy as a function of the displacements of individual atoms from their equilibrium positions (to handle bond stretch and angle bend), along with a nonbonding interaction potential which included repulsive and attractive terms. From our modern vantage point, the Hill force field is the most recognizable when compared to current formulations. However, the work of Frank Westheimer became more widely appreciated, and this approach was for a while referred to as the "Westheimer method."

Despite the inspired work in the nineteenth century on the three-dimensional nature of molecules, this concept did not yet widely influence the thinking of the chemical community. This changed abruptly in 1950 with the publication of Barton's short note on how the conformations of steroids affect their chemistry, which laid the foundation of **conformational analysis**. From then onward, an appreciation of the three-dimensional aspect became crucial to understanding structure, stability, conformation and reactivity. In the following year, Pitzer described the torsional potential as a cosine function with structure-dependent periodicity, and presented an analysis of the structural features in simple hydrocarbons which should govern the energetics of this interaction.

Occasionally, appreciation of a scientific discovery will expand beyond the scientific community to be shared by the entire world. Such was the case with Watson and Crick when in 1953, building upon the work of at least a dozen other contemporary scientists, they presented

their structure of DNA. This structure arguably constitutes the most famous physical molecular model of all time.

In the same year, a group of scientists from Los Alamos published their studies of, "Equation of State Calculations by Fast Computing Machines." This work, carried out on the advanced MANIAC computer, laid the groundwork for computer-based **Monte Carlo** methods, established the **Metropolis algorithm** (from the first-named author) for **simulated annealing**, and was the ancestor of **molecular dynamics** calculations.

In 1959, Corey and Bailar reported the mathematical derivation of models for isomeric cobalt bis- and tris-diamine complexes from bond length and angle span input, using a vector analysis method and admittedly ignoring solvation effects. Assuming that the potential energy differences between conformers was due primarily to nonbonded interactions between the ligands, they calculated the relative steric energies of the isomers by applying the Mason-Kreevoy nonbonding H-H repulsion potential function to both the H-H and C-H nonbonding interactions present in their models.

Applications of nonbonding potentials to organic structure modeling began to appear from Kitaigorodsky in 1960. In his model, "hard sphere" nonbonding potentials were used to calculate interactions between any proximate atom pairs not directly bonded to each other. These interactions were superimposed upon initial structures having ideal bond lengths and angle spans, with results which were of good quality for that period.

The first published use of a computer for empirical force field calculations of molecular structure was in 1961 by Hendrickson, who examined the conformations of medium-sized rings. His force field contained a quadratic potential for angle bending, a three-fold cosine potential for torsional strain, and a modified Buckingham potential for nonbonding steric interactions. He reasoned that changes in the bond length would be negligible, so these were frozen at the equilibrium values. Hendrickson applied his technique widely to the problem of conformational analysis.

In 1963, Schachtschneider and Snyder reported the derivation of a set of transferable parameters for a valence force field for hydrocarbons, from a vibrational analysis of the n-alkanes. Transferability is crucial to the development of a general force field with a limited number of parameters.

Wiberg employed computer force field calculations to address conformational analysis in 1965. He developed a **steepest descent algorithm** as a scheme for geometry optimization. Wiberg also published algorithms for transforming Cartesian coordinates to internal coordinates, to calculate bond lengths, bond angles, and torsional angles from these coordinates. He also provided a geometric algorithm to predict the location of the hydrogens on a methylene carbon.

In a paper in 1968 in which he reported force field calculations on some cyclophanes, Boyd presented the modified **Newton-Raphson method** for strain energy minimization, which reached convergence more rapidly than the Wiberg algorithm.

The force field treatment of metal-containing compounds was evolving from the pioneering work of the previous decade, when in 1969 the first in a series of publications from the Snow group appeared. They described the results of careful geometry optimization in cobalt pentaamine chelates using Boyd's method and constants culled from the literature.

Over the latter half of the twentieth century, force field calculations of molecular structure have developed largely in pace with the development of fast computing machines to carry out the extensive calculations required. Early growth was slow, though, due to unfamiliarity of the average chemist with the programming and use of mainframe computers, and also due simply to lack of access. Faced with laborious trigonometric calculations, many researchers would still employ carefully constructed scale models to study molecular structure. U.S. government scientists did have access to computers, and made seminal contributions in this area, such as the Monte Carlo work mentioned above, and the archetypal molecular drawing program, **ORTEP** (Oak Ridge Thermal Ellipsoid Plotter), developed at Oak Ridge National Laboratory in 1965. Also with the government's support, the first networked computer (the ancestor of the Internet) was demonstrated in October 1969, by establishing a linkage between a computer at UCLA in Los Angeles and another at SRI in Menlo Park. This latter event went virtually unnoticed at the time, but subsequent advances in this area would revolutionize the study of molecular structure, and nearly every other human endeavor which depends upon the transfer of information.

By the early 1970's, the major force field formulations were being reported: the work leading up to the **ECEPP**, and later **UNICEPP** force fields by Scheraga <u>et al.</u>, the Consistent Force Field (**CFF**) by Lifson and Warshel, the **MMI** force field by Allinger <u>et al.</u>, the **EAS** force field by Schleyer <u>et al.</u>, Boyd's force field, and the **MUB** force field by Bartell. This stage of the history of molecular modeling is well documented elsewhere, and some details will be discussed in the case of specific force fields in Chapter 5. These force field models were based on different assumptions, and though they used many of the same or similar potential functions, they were weighted and parametrized differently. For example, Allinger's MMI and Engel, Andose and Schleyer's EAS force field would both replicate the gauche butane interaction well, but for different reasons: MMI has 'hard' hydrogens which gave rise to greater repulsions in the gauche form, whereas the EAS field with 'softer' hydrogen and 'harder' carbons was still able to reproduce the desired interactions.

In 1971, Lee and Richards described the **molecular surface** in the context of protein structure, and provided an algorithm to derive it. This involved constructing spheroids about each atom scaled by the van der Waals radius, and considering as the surface those portions of these spheroids which were not contained within any other such spheroids. This analysis proceeded through the molecule in a series of planar sections, upon each of which the corresponding projection of the surface was "plotted." In those days before the advent of modern computer graphics, the surface projection for each planar section was imprinted on a transparency sheet, and these were stacked atop one another to visualize the molecular surface. Expanding upon this work two years later, Shrake and Rupley came up with the first dot surface formulation applied to protein structure, finding that 92 points would adequately define an atomic spheroid for the purposes of determining a surface.

In 1972 Wiberg and Boyd presented an algorithm for exploring conformational interconversions based on systematic modifications torsional angle. This technique subsequently enjoyed broad usage as the **dihedral driver** method.

1974 saw the first report of computer-aided molecular modeling of oligosaccharides from the Lemieux group. They began with crystal diffraction structures for the individual saccharide units, employed the Kitaigorodsky hard sphere nonbonding potentials, and developed torsional potentials for the **ϕ-angles** and **ψ-angles** which were parametrized to account for the <u>exo</u>-anomeric effect. This hard sphere, exo-anomeric (**HSEA**) formulation was far simpler than the other force fields of the time. However, it was computationally rapid and gave quite satisfactory correlations with the conformational equilibria derived from NMR studies, and continues to be popular with carbohydrate chemists.

Around this same time, force field formulations for synthetic macromolecules began to appear, such as was used in the conformational analysis of polypropylene by Suter and Flory.

The early 1980's saw the beginning of the personal computer industry, and the consequent increase in the accessibility of computers in general, and the use of the **graphical user interface** (GUI). As force field modeling was married to graphic display on desktop units, the era of personal molecular modeling for the average chemist had begun. A convenient demarcation is the year 1983, when the distribution of the Model (later MacroModel) program by the Still group at Columbia began. This program could be configured to run on a mainframe with a variety of desktop platforms as the user interface. At about the same time, the development of high-quality real-time color graphics and Connolly's molecular surface program contributed to the rapid evolution of this technique. The graphic presentation of color-coded molecular surfaces which displayed at the same time shape, charge and hydrophobicity provided the modeler with a much higher information density.

In 1986, M. Saunders refocused attention on the central force field model with his STRFIT molecular modeling program. In this method, bond angles and torsional angles are not considered; only forces between pairs of individual atoms are calculated. The resulting structures are quite close to those obtained from MM2. In the following year, Saunders presented the "random kick" algorithm for conformation searching, which was especially helpful in searching over ring conformations.

The history of the World Wide Web begins in 1993, when the first graphical browsers became available. Non-graphical Internet applications, such as electronic mail and file transfer protocol (FTP), had been developing since 1969, but the vast potential of the graphical user interface can encompass and integrate these functions, and provide the additional enhancements of graphic images.

Within the past ten years, so much has occurred that it is difficult to sort out which advances will truly stand out over time. One is the advent of the so-called **Class 2 Force Fields**, which contain anharmonic potentials, and utilize explicit off-diagonal terms from the force constant matrix. A not unrelated development is the use of structural data obtained from high-end *ab*

initio calculations to parametrize these new force fields. The quantum mechanically-derived potential surface is then scaled to reproduce experimental data.

The prediction has made that the **Class 3 Force Fields** will be able to model the influence of chemical effects, electronegativity, and hyperconjugation on molecular structure and properties.

Another ongoing trend is the development and implementation of virtual reality for molecular modeling. Basically, this will entail enhancing the visual third dimension, further developing and enhancing audio sensation, and introducing tactile sensation. It is unlikely that this type of modeling will be implemented on the desktop platform any time soon.

The best (in the author's opinion) demonstration of virtual reality molecular modeling has been achieved through a group collaboration. In this Virtual Biomolecular Environment (**VIBE**) system, a massively parallel computational engine runs an interactive molecular dynamics simulation under the CHARMM force field, and this is interfaced with a 3 m cubic "theater," called the **CAVE** (Cave Automatic Virtual Environment). The modeler enters this theater and interacts with the three-dimensional simulation by means of a wand-link remote control device. The modeler's head and hand movements are tracked with a six-degree-of-freedom electromagnetic sensor to maintain the correct perspective in the virtual environment. The wand may be used to dock a ligand into a binding site, and bumps and other interactions are experienced as sounds.

Before leaving the subject of the history of modeling, some additional comments on graphical (written) and physical models may be made. In drawn molecular structures today, a relatively uniform international standard exists with the common two-dimensional line and polygon orthography of organic chemistry, despite numerous embellishments and variations. The stylistic features of orientation and the depiction of stereochemistry can still vary a great deal between sub-fields within organic chemistry (for example carbohydrates and terpenoids), as well as from group to group and from scientist to scientist. A number of conventions have been developed to enhance the information conveyed by drawings, for example the Newman and sawhorse projections to indicate dihedral geometry, the Haworth depiction of carbohydrates, and the myriad of techniques for depicting peptide chains. Each of these emphasizes certain features of the molecule, while distorting or ignoring others. For example, a conformational analysis which depends upon Newman projections can succeed only at a very crude, qualitative level, and is highly dependent upon the graphic skill and perspective of the artist.

Connolly has reviewed the physical molecular models from the mid-twentieth century onward. A comprehensive treatment of the subject of physical molecular models for both organic and inorganic substances is found in the monograph by Walton, and a more compact compilation of information on the commercially available molecular model sets is provided by Gordon and Ford. The better physical molecular models are somewhat more successful, in that they normally give at least one type of information accurately. Bond lengths, bond angles, internuclear distances, and torsional angles (but not torsional barriers!) can be modeled semi-quantitatively with skeletal stereomodels, such as the Dreiding, Darling or Fieser type, or any of the many varieties of ball-and-stick, or ball-and-spoke model (e.g. Cenco-Petersen, HGS-Maruzen, etc.). The Dreiding stereomodels have perhaps found the widest usage. Molecular

models need not be commercial or costly to prepare, and it is possible to produce useful models from inexpensive materials, such as colored wire. Molecular surfaces, some torsional barriers, and crude docking and molecular recognition phenomena can be modeled with CPK or other kinds of space-filling models. These models all have the advantage that they can provide simultaneous visual and tactile information about the subject molecule.

As the field of molecular modeling continues to evolve, ways are being developed to couple the higher accuracy of the computer-based methods with the intuitive feel of physical models. In one recent report, **stereolithography** is used to translate computer-designed structures into physical molecular models. The data from molecular orbital calculations or from neutron diffraction studies was used to guide the fabrication of plastic models by laser curing of an acrylate resin blend. The stereolithography apparatus works by building the model iteratively through a series of "slices", from the bottom up. The ideally spherical atoms are approximated by polyhedrons, and the number of facets can be varied to differentiate atom types while remaining low enough to give manageable data file sizes. The scales used were $1\text{Å} = 0.19\text{-}0.89$ cm, depending upon the overall size of the molecule. Molecules, transition state complexes, and "molecular impressions" or shape-complementary surfaces were prepared. The fabrication process, including a post-curing, takes from 10-27 h depending on the complexity of the model. The authors point out that the availability of such custom-made models can serve a general need, and can be especially helpful to the visually impaired. Another recent patent claims the preparation by stereolithography of a physical model of an inhibitor plus part of the receptor, guided by the use of modeling software.