1 Chemistry

Chemistry—the science of molecules—is the science of the everyday world. For the world at large, molecules are the fundamental units of matter: smaller than this, matter loses its macroscopic identity and, hence, its defining characteristics. An understanding of the structure, interactions, and reactions of molecules is thus of critical importance to a wide range of phenomena, from the fate of contaminants in the environment, through the production of plastics from crude oil, to the occurrence and treatment of genetic diseases. In a world attempting to balance energy usage, environmental quality, human health, and economic prosperity, chemistry is, and will continue to be, the key science required to meet the challenge.

The fundamental physical laws defining molecular behavior are known, and embodied in quantum mechanics as the Schrödinger and Dirac equations. As the Nobel Prize winner Paul Dirac stated in 1929, *"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."* Chemists have spent the last 75 years developing accurate solutions to these equations with much success.

Chemists are now about to cross a remarkable threshold and expect a dramatic expansion in their ability to make reliable predictions about molecular structure and processes. This is due to the confluence of advances in theoretical, computational, and experimental capabilities, allowing chemists to understand and characterize matter at the detailed atomic and molecular levels. By integrating chemists' capabilities in the areas of synthesis and characterization with computational modeling and simulation, it will soon be possible to use computation to design molecules to do what we want, and to control how we make them.

1.1 Impact on Science and Society

Chemistry is the central science. Not only does chemistry have intellectual merit in its own right, but it is a critical element of many of the other sciences important to the U.S. Department of Energy. For example, with reference to other chapters, chemical processes are responsible for the energy produced and pollutants released by an automobile engine (combustion science). Catalytic chemical processes are used to remove these pollutants from the tailpipe emissions. Likewise, an understanding of chemistry is important for predicting and mitigating the spread of pollutants in underground plumes (subsurface science), processing high-level radioactive wastes, predicting and alleviating the long terms effects of greenhouse gases and stratospheric ozone depletion (climate science), tailoring the properties of nano-materials for a broad range of applications (material and nanoscale sciences), and for understanding and manipulating the biochemical processes on which life is based (biology).

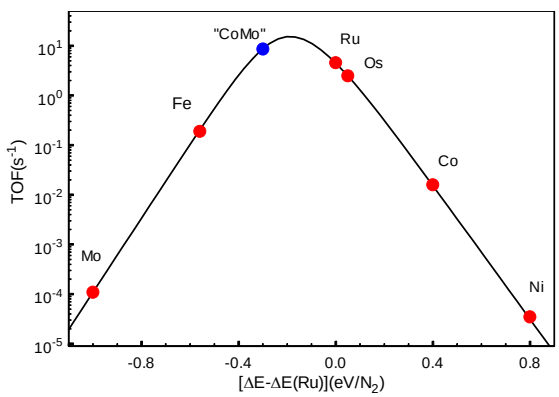
The U.S. chemical industries play a major role in the national economy and contribute directly to the high quality of life enjoyed by its citizens. There is little in everyday life that does not involve these core U.S. industries. The U.S. petroleum, chemical and pharmaceutical industries are the world's largest producers of chemicals, ranging from "wonder" drugs to paints, from cosmetics to plastics, from fuels to fertilizers. The chemical industry represents 10% of all U.S. manufacturing, employing more than one million Americans. It is also one of the few industries

that possess a favorable balance-of-trade. The petroleum and chemical industries contribute approximately $500 billion to the GNP of the U.S. These industries maintain their global competitive position by their ability to produce new products using energy-efficient, low-cost, and environmentally clean processes. An understanding of and an ability to predict the structures, energetics, and reactions of molecules are essential to achieving these goals.

1.2 Scientific Opportunities

Computational chemistry has already had an enormous impact on fundamental chemical science. It has provided new insights into the structure and properties of molecules, especially those molecules whose existence is fleeting, *e.g.*, free radicals, yet whose role in many chemical processes is critical. Computational chemistry has also dramatically advanced chemists' understanding of how chemical reactions occur – in the energetic changes that drive chemical reactions, in the structural changes that occur as reactions proceed, and in the detailed dynamics of the collision complex. These insights were initially qualitative, but as computational methods, software, and computer hardware have advanced, chemists are now able to make accurate predictions about a wide variety of chemical phenomena, sometimes literally replacing experiment by computation. With the additional investments envisioned in this report, the capabilities of computational chemistry will dramatically expand, enveloping a far broader range of chemical phenomena than possible today.

Heavy element and environmental chemistry, which are special responsibilities of the DOE, are discussed in a sidebar.



Calculated trends in ammonia synthesis rates (TOF) shown as a function of the dissociative chemisorption energy of nitrogen. The figure shows, in agreement with experimental evidence, that Ru, Os and Fe are the best catalysts and how a CoMo alloy should be as good as the best elementary catalysts.

Chemical Catalysis. The chemical industries take a cheap feedstock (*e.g.*, a hydrocarbon) and convert it into a higher value product by rearranging the atoms and/or adding functional groups. Catalytic processes are directly involved in the synthesis of 20% of all industrial products. A catalyst's role is to make a desired chemical reaction proceed much more efficiently than it otherwise would. To date, catalysts have been largely designed using trial and error, *e.g.*, synthesizing and testing a potential new catalyst to determine if yields are improved, unwanted byproducts reduced, and/or energy requirements decreased. This process is both expensive and time-consuming, and it rarely leads to novel new catalysts. Advanced experimental tools are still unable to provide data on all of the steps involved in catalytic processes. Computational modeling and simulation can fill this gap, enabling the design of catalysts from first principles. For example, *ab initio* catalyst design requires quantitative information about transition states for key reactions. This information is only accessible by computational methods. In one successful example (see Fig. 1), a combined simulation/experimental approach was used to design the first new ammonia catalyst since Haber

and Bosch's work in the early 1900's. Computational chemistry will also play a critical role in designing new materials and catalysts for hydrogen storage for the transportation sector. In order to make a major step forward in catalyst design, we must gain at least a factor of 1000 increase in the performance of simulations.

Combustion Chemistry. The chemical reactions involved in combustion determine the rate at which fuel is consumed and energy is released, as well as the quantity of pollutants produced. Thus, a thorough understanding of the chemistry of combustion systems is required to meet two important design goals for tomorrow's advanced combustion systems: reducing unwanted emissions and improving system performance. The hydrocarbon fuels involved in everyday combustion systems are complex mixtures of large hydrocarbon molecules. For example, gasoline primarily consists of iso-octane and n-heptane as well as performance enhancing additives. It is impossible for laboratory studies to obtain all of the data needed to model combustion. Many species have only a fleeting existence, yet are critical for sustaining the combustion process; many reactions only occur at the high temperatures attained in flames, yet the harsh conditions of combustion systems are difficult to reproduce in the laboratory. Recent advances in computational chemistry and the efficient implementation of these advances on terascale computers have provided a new level of capability for characterizing the chemical species and reactions involved in the combustion of the key components of gasoline, iso-octane (C_8H_{18}) and n-heptane (n-C_7H_{16}). With an increase of a factor of 100 in computing power, it will be possible to begin studies of n-hexadecane ($C_{16}H_{34}$), one of the major components of diesel fuel. This capability could revolutionize the design of new combustion systems on which the nation depends for 85% of its energy needs.

Nanoscale Science. Nanoscale structures and processes exhibit qualitatively new behavior – the systems are too large to be treated by quantum mechanics today and may be governed by microscopic collective phenomena, yet are too small to behave according to the laws of continuum mechanics. Chemistry is at the heart of nanoscale science, much of which is the study of very large molecules. Although a number of experimental techniques for characterizing nanoscale systems have been developed, many details of nanoscale structures and processes remain unobserved or unobservable. Computational chemistry methods are already being used to simulate the properties of carbon nanospheres (buckyballs) and nanotubes, and the flow of fluids confined in spaces just a few nanometers in extent. These simulations clearly establish the connection between the structure of the molecules involved in nanoscale processes, *e.g.*, hydrocarbon lubricants, and the unusual processes often observed, *e.g.*, alternate "freezing" and "thawing" behavior in fluid flow. A 100 to 1000x increase in computational capabilities will enable improved design of nanomaterials, a molecular understanding of friction, the design of new lubricants, and the development of a new generation of molecular electronic devices.

Biochemistry. Although biological systems have many unique features, at the most fundamental level, they are molecular machines. Proteins and enzymes, molecules that are essential to life, have long been the province of experimental and computational chemists and biochemists. Computational chemists pioneered protein simulations, continue to be leaders in this field, and are now developing new approaches for molecular simulations that hold great promise for elucidating the mechanisms of enzymatic reactions, for understanding and characterizing protein-protein interactions, and for understanding the intimate relationship between molecular structure and biological function. Carbohydrates, drug molecules, neurotransmitters, the permeabilities

and structural properties of cell membranes, and nucleic acids (the material of genetics) are equally important, and chemical computations will be critical keys to their ultimate understanding. Advances in computational chemistry are critical to understanding biological systems at their most fundamental level, and reliable prediction of protein chemistry will require a factor of at least 1000 increase in computational performance, for example, to improve their use in energy production or environmental remediation .

1.3 Research issues

Significant theoretical, mathematical, and computational obstacles must be overcome to realize the full potential of computational modeling and simulation in understanding and controlling chemical change. However, we are confident that these problems will be solved with appropriate investments in basic research, although it is, of course, not possible to predict exactly when the requisite breakthroughs will occur. Our confidence is based upon both experience and recent theoretical developments that, in selected areas, point a way forward. In this section, we summarize some of these research issues.

One of the key needs for the improved design and control of chemical processes is accuracy. It is now possible to routinely make reliable predictions of the heats of formation of small – to medium–size gas phase molecules (10-20 atoms) (errors of less than 1 kcal/mol) with reasonable computational effort. In addition, for some simple gas-phase reactions, it is possible to make predictions of reaction rates over broad temperature ranges to within a factor of two to three. However, the computational cost of these calculations scales as N^7, where N is related to the number of atoms in the molecule. This means that we cannot rely solely on increases in computational power if we want to make predictions about the much larger molecules that are involved in many chemical processes important to the DOE. We must expand the progress that has already been made in reducing the scaling of molecular simulations. In fact, expected increases in applications software capabilities during the next few years will finally allow calculations on molecules that are large enough to realize a significant reduction in the scaling exponent. In addition, we need to include and quantify the effects of relativity to predict reliably the properties and behavior of (radioactive) heavy-metal compounds.

The ability to treat the environment in which a reaction is taking place, *e.g.*, a solvent, requires new theoretical and algorithmic developments in order to make reliable predictions about the reaction rates in the condensed phases that are so important in chemical production systems, on atmospheric particles, and in biological systems. Modeling chemistry at the nanoscale, also introduces new challenges, requiring new boundary conditions and also multiscale methods. Excited states are also an area in need of new theoretical developments. Electronically excited states play a key role in many important chemical processes, including combustion and photosynthesis, yet we currently have no efficient computational approach that can quantitatively describe such excited states for other than the smallest molecules.

There is also a critical need to extend atomistic simulations, such as molecular dynamics, to much longer length and time scales. Reaching macroscopic time scales in molecular dynamics simulations will not be solved solely by increases in hardware capability, since there are fundamental limitations on how many time steps can be executed per second on a computer, whether parallel or serial. At present, it is routinely possible to simulate atomistic systems, or systems represented as interacting atoms such as proteins and polymeric systems, for periods of

the order of tens of nanoseconds. However, far longer time scales—microseconds, milliseconds and beyond—are needed to simulate such processes as phase transitions, rare events, enzyme kinetics, and protein folding. New methods are also required for effective sampling of infrequent events, exploring "rough" energy surfaces, and modeling reaction dynamics on surfaces with many minima.

Although data management and analysis has not posed a problem in computational chemistry in the past, on current generation supercomputer systems, computational simulations are creating as-yet-unresolved data management and analysis problems. For example, a single time step of a million-atom simulation produces tens of megabytes of data, which, integrated over billions to trillions of time steps, yields terabytes to petabytes of data. It is not possible to use tradition "store and retrieve" techniques to analyze these simulations. Computational chemistry algorithms for data analysis "on-the-fly" must be developed and refined.

We have not come to the end of innovation in theory, mathematics, and algorithms just as we have not come to the end of improvements in computational speed and memory capacity. These innovations will extend the power and utility of computational chemistry well beyond the bounds envisioned today.

1.4 Technology barriers to research

Applied and numerical mathematics: It is impossible to overstate the impact of improved numerical methods on computational chemistry. Examples include the standardized robust and fast linear algebra routines (*e.g.*, BLAS, (SCA)LAPACK), and the recent applications in chemistry of the fast-multipole method (FMM). These accomplishments were, in large part, supported by the DOE. Current major problems include: simulating multiple length scales in non-periodic systems, long-time simulations, and efficient sampling of infrequent events; dynamics and statistical mechanics of high-dimensional systems; and reduced scaling of electronic structure methods with molecular size and computational precision. Other necessary developments include improved FMM to reduce both the break-even point and the cost for high precision, numerical methods (*e.g.*, multiresolution, multigrid, low-separation rank) suitable for efficient and reliable computation in three, six and higher dimensions, and solving inverse and optimization problems such as commonly appear in either process design or the interpretation of experimental data.

Algorithms: Modern numerical methods trade precision for speed, and to do so reliably requires robust and sophisticated algorithms. For instance, the FMM algorithm describes how to compute the potential due to a collection of charges to a specified precision in a time that increases linearly with the number of charges. On a parallel computer, an algorithm must also describe the distribution and movement of data and work between the processors. Additional algorithms and techniques are necessary to accommodate the latency and bandwidth of memory, which are not keeping pace with processor speed. Anticipating future computers with tens to hundreds of thousands of processors, we need algorithms that exploit many levels of parallelism and include dynamic load balancing. Algorithms that self-adapt to the available architecture are also very desirable. Finally, the complexity of modern chemical theories is exceeding the ability of humans to implement them on modern computers. Increasingly important will be automated generation of near-optimal algorithms and software from high-level mathematical expressions.

Software: Essentially all of the complexity of the theories, algorithms, computer hardware, and programming models are visible at the software level. Even now, it is only by using sophisticated design and software engineering techniques that this complexity is manageable in computational chemistry applications. If we are to make successful use of next generation computers and algorithms, we must aggressively seek new and yet more sophisticated software approaches. Improved and fully interoperable parallel programming tools supporting multiple levels of parallelism are essential, and must provide higher abstraction levels than current tools. The poor quality of hardware vendor software and the increasingly user-*un*friendly environments of large computers must be improved. In addition, performance analysis and debugging tools must be capable of dealing with large and long-running simulations. Domain specific frameworks could ease management of workflow and resources. Finally, fault tolerance is a large concern, as long running computations on large numbers of processors are unlikely to finish before an interruption occurs in the software or hardware.

Hardware: Chemical algorithms are very diverse, and it is not possible to identify a single machine architecture that meets all of our needs. Current requirements are for architectures balanced in terms of the performance and capacities of processor, memory, interconnect and I/O. Future algorithms are expected to be less regular and not as dominated by dense linear algebra operations as are many current methods. Processor support to tolerate high memory latency is clearly important. Larger caches might enable classical molecular dynamics to become fully cache resident on big parallel computers. This will dramatically increase the speed of each time step in these simulations. However, to achieve long simulated times, we must use as many processors as possible; such scalability requires low-latency and high-bandwidth inter-processor communications. Electronic structure calculations require high-performance for matrix multiplication and fast Fourier transforms. Fault tolerance is again a major concern.

1.5 Resources required

In this section, we estimate the resources (theory, software, and hardware) required by the DOE Office of Science to advance computational chemistry in support of DOE's missions. We do not include resources needed to overcome the technology barriers.

Advancing theory is crucial to the future of computational chemistry. This places a high premium on innovation and will largely be accomplished by research groups of single principal investigators, which represent a largely under-realized opportunity for the DOE Office of Science. However, their vitality is threatened by diminishing funding and the relatively small size of university single-PI grants. The Office's investment in basic theoretical research including statistical mechanics, electronic structure, reaction dynamics, and classical dynamics must be expanded significantly and new PIs brought into the program. The rapid and effective transfer of ideas from research to production computing must also be encouraged and supported.

We must also create and support a full suite of computational chemistry simulation software that efficiently implements our models on computers ranging from PCs to building-sized ultra-scale computers with millions of processors. In a side bar, we analyze the required resources with reference to the known development costs of just one code. The expected complexity of future computers and algorithms, the breadth of the science applications, and our experience to date lead us to the following findings.

- The major applications software systems will best be developed with broad, community-based collaboration including computational chemists, computer scientists, mathematicians, and end users.
- A disciplined approach to software development is vital although computers become obsolete in three to four years, scientific applications have lifetimes of decades.
- National computational chemistry facilities to coordinate, develop, and support key applications should be considered. The U.K. Collaborative Computational Projects (CCPs) are successful, long-lived role models for the envisioned effort, although their funding level is inadequate for the current challenges.
- Additional software development resources are needed: about 20 full time staff in electronic structure, 20 in molecular dynamics/statistical mechanics, and 10 in reaction dynamics, plus postdoctoral fellows and graduate students.
- Development of expert systems and domain-specific problem-solving and collaborative environments will require 15-20 additional full-time staff. Such tools will become essential as the scale of data management and computation increase (not just the size but the complexity and number of calculations).
- In the long-term, as much as one third of the manpower must be devoted to software support and maintenance.

Whereas a large fraction of resources will be required to enable capability computing on ultrascale computers, capacity computation on smaller machines must not be neglected. Chemists need a large high performance computing capacity because of the vast number of molecules of interest, the complexity of the calculations, and the large number of chemists using computational tools. In turn, capability computing is required for large, high-fidelity simulations, to ensure accurate predictions of critical molecular properties, and to provide highly accurate benchmark results against which to test more approximate methods.

No single machine is appropriate for all of the algorithms represented in chemistry and chemists emphasize *balance* in all aspects of the hardware, with our temporary disk I/O requirements perhaps differentiating us from many other scientific disciplines. This diversity suggests opportunities to improve cost effectiveness by matching the computer architecture to the algorithms, as illustrated by the recent MD-GRAPE and WINE processors – hardware specifically designed to speed up MD simulations. Beyond 5-10 years, it is difficult to anticipate the details of computer architectures, and it is essential that chemists have early access to advanced architectures (*e.g.*, FPGAs and PIMs), including the use of simulators, to evaluate and influence the designs, and to optimize our software.

We estimate that about 40 (peak) teraflop/s-years are annually dedicated to chemistry applications associated with non-ASCI DOE work, including DOE's supercomputer centers, laboratory and group clusters, and desktop workstations. The discussion in previous sections concerning 100x and 1000x increases in resources is relative to this figure. As soon as possible, computing resources need to be increased 10-fold, by 100-fold within 5-10 years, and by 1000-fold as soon as feasible after that. In combination with concurrent advances in theoretical and computational methods, this expansion of compute power will enable truly revolutionary scientific breakthroughs. We cannot emphasize enough that without resources for the actual computations, we will fail to realize the scientific discoveries that the entire infrastructure is intended to enable.

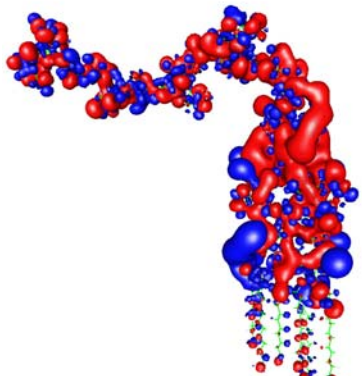
For instance, the limited computer and manpower resources presently available have discouraged most computational chemists from investing the effort necessary to develop massively parallel software.

1.6 Metrics of success

Science-based metrics should be used to measure improvements in the performance, predictions and impact of computational chemistry. Combined with measured execution rates on specific hardware, these metrics may also provide some measure of the efficacy of that hardware for selected applications. The reduced dependence of computational chemistry calculations on accuracy, molecule size, and/or time span may be measured and compared against past performance. Community-defined chemistry benchmarks *(i.e.*, the time to solution of specific chemical problems with prescribed levels of theory, accuracy, and precision) can quantify these advances. Another family of increasingly large, community-defined chemical benchmarks could measure increased capability. The expanded simulation capabilities will also be demonstrated by new scientific discoveries. Another measure of the broader impact of computational chemistry is the increased uptake of computation by industry, which typically lags use in national laboratories and universities by several years. Finally, approximately 15% of journal publications in chemistry currently employ simulation in some form; for example, this percentage could be tracked.

What is required to develop the next generation of computational chemistry software?

Solution of significant chemical problems requires the use of multiple theoretical models. As a result, computational chemistry packages contain a broad range of functionality and are large, complex software systems. Consider some of the major U.S. chemistry packages for solving the electronic Schrödinger equation: GAMESS, GAUSSIAN, NWChem, and QChem. These packages each have about one million lines of code, and each is particularly capable for certain types of simulation, each has a large following in the chemistry community (e.g., GAMESS at 10,000 sites worldwide, and NWChem at 1000 sites), and each required substantial investments to develop and is long-lived (e.g., GAUSSIAN is now over 30 years old).



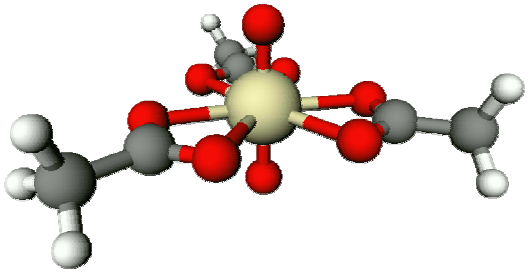
Electrostatic potential of lipopolysaccharide and counter-ions (936 atoms) from a constrained RESP fit to HF (6-31G* basis, 7943 functions).

To develop a new generation of software packages for solving the electronic Schrödinger equation on the computers being considered in this report will require a substantial investment of resources. As an example, consider NWChem, a software package that was designed from the ground up for massively parallel computers. It uses the latest software engineering practices and aims to run efficiently on 100–1,000 processors. Some tasks were recently re-designed for up to 10,000 processors. The project, which began in 1992, involved a core group of five computational chemists, a computer scientist, and an applied mathematician, augmented by a worldwide group of twenty contributors. In total, over 100 man-years of effort were invested. Similar levels of effort were expended to parallelize legacy codes such as GAMESS.

Successfully confronting the 1,000-fold increase in the complexity of future computers, algorithms, theory, and scientific problems, as well as supporting a much larger base of users, cannot be accomplished with the small core team plus contributors such as that which developed NWChem. *There is a need in the next decade for the following resources defined relative to the size of the NWChem team: electronic structure (3x), molecular dynamics (3x), reaction dynamics (1x), and problem-solving and collaborative environments (2x).*

Heavy element and environmental chemistry – special responsibilities of the DOE

Four decades of nuclear weapons production at U.S. Department of Energy facilities has resulted in the interim storage of millions of gallons of highly radioactive mixed wastes in hundreds of underground tanks, extensive contamination of the soil and groundwater at thousands of sites, and hundreds of buildings that must be decontaminated and decommissioned. The single most challenging environmental issue confronting the DOE, and perhaps the nation, is the safe and cost-effective management of these wastes. What is the physical and chemical form of the wastes in the tanks and in the ground? How can the radioactive wastes be safely processed? How can the processed waste be safely stored? Understanding the chemistry of radionuclides has long been a special responsibility of DOE's Office of Science, and, given the difficulty of handling radioactive material and the loss of experimental expertise in heavy-element chemistry in the U.S., improved ability to model the chemistry of these species is becoming a critical national need.



Uranyl triacetate. The structure, thermo-chemistry, and vibrational spectra of *model* radioactive species can be reliably predicted with relativistic quantum chemistry on high-performance computers. Fully detailed models, and other properties such as electronic spectra, are currently not feasible, in general.

Computational chemistry can provide fundamental information on the complex interactions of radioactive materials with other species, often enabling the replacement or curtailment of expensive experiments. Interactions in the soil determine how contaminant plumes spread and transform. Interactions within the complex chemical mixtures inside the waste tanks determine how to transfer, process and safely store the wastes for thousands of years. Such information is needed to solve DOE's cleanup problems in a comprehensive, cost-effective, permanent way. *True reliable predictions require the combination of the theory of relativity with quantum mechanics and will require 100 to 1000 times current computational capabilities.*