

A Proposal Submitted to

*Pew Summer Research Program
Union University
For Summer 2009*

A PHP-based Web Frontend for a Molecular Dynamic Computational Software Package

Haifei Li

Abstract:

Molecular dynamic (MD) simulations of complex gas-phase reactions are a very important research tool for the petroleum and automobile industries. Dr. Michael Salazar has done extensive research in this area for many years. His research proposal for creating computational tools has recently been approved by American Chemical Society (ACS)'s Petroleum Research Fund. These computational tools currently under development are collectively called **Accelerated Molecular Dynamics with Chemistry (AMoDC)**, and I have collaborated with him on designing an efficient module in *AMoDC* over the past few years. The collaboration was successful and the preliminary codes (written in C language) have been distributed to other universities for research purposes. *AMoDC* is an extremely efficient set of computational tools but currently has as its main target the advanced users that has both strong computer science and chemistry backgrounds. Since it was written in C and runs under a Linux environment, users need to be comfortable with not-so-user-friendly command line interfaces. Its strong requirements on computer background limit its huge potentials for wide adoptions and distributions. This Pew summer research proposal is to develop a web frontend to the computational tool set and find a way to host the web program for unlimited public access. Since this proposal depends on the ACS proposal written by Dr. Salazar, close collaboration will be necessary.

I. Background

Gas-phase reactions are ubiquitous in the chemical world. Perhaps it is due to the familiarity of these reactions that chemists are often led to not give proper attention to the fundamental majesty of them. Consider, for example, the complexity associated with gas-phase combustion processes, which happen to power our cars everyday. These gas-phase processes have very simple net reaction, $\text{fuel (l)} + \text{O}_2 \text{ (g)} \rightarrow \text{CO}_2 \text{ (g)} + \text{H}_2\text{O (g)}$. However, these combustion reactions involve $\sim 10^2$ simultaneous elementary reactions and $\sim 10^3$ different chemical species in leading from reactants to products and this process takes on the order of 10s-100s of nanoseconds.^{1,2,3} Due to this complexity is it that only recently has effort gone into performing computer simulations of combustion processes.^{4,5,6} However, this effort focuses on following the dynamics of a *single* step – of the $\sim 10^2$ possible elementary steps – in isolation. Indeed, there is much room for improvement for following the dynamics of complex combustion processes. Another field in which performing simulations is very valuable is the chemistry of the atmosphere, where, similar to combustion processes, many simultaneous reactions occur that involve a large number of chemical species.

MD simulation is the process of following the motion of atoms within a molecule by the direct integration of Newton's equations of motion, with the result being that one knows the exact location of the atoms (\vec{X} , a vector of coordinates) and their associated velocities (\vec{V} , a vector of velocities). The power of MD simulations is that, in principle, one may follow the chemistry of a reaction from the original reactants on through the formation of products. If one is able to accomplish this task, one would also know all the information about the chemical reaction, i.e., how much energy it takes to initiate the reaction, how the the energy is distributed among the products, what products are formed and what are the reactions forming the products and the product branching ratios, etc. There are, however, great difficulties to overcome to perform MD simulations and this difficulty is greatly exacerbated in the case of reactions. Two of the primary difficulties are the number of internal degrees of freedom of the chemical systems ($3N-6$, where N is the number of atoms)⁷ and the corresponding complexity and computational costs of obtaining accurate *ab initio* potential energy surface (PES) of reactive systems⁸ (*ab initio* means “first principles”, i.e., the potential energy of how atoms

interact with one another are calculated from the underlying physics). First, the $3N-6$ internal degrees of freedom can grow very large and will, further, sample very different properties. During the course of a reaction some internals may change little from an equilibrium value, while some may undergo drastic changes (i.e., those degrees of freedom that contribute significantly to the intrinsic reaction coordinate). Due to these difficulties, the chemical systems often chosen for study by *ab initio* MD simulations are small, often involving 5 or fewer atoms. The ACS proposal by Dr. Michael Salazar would expand the number of atoms in the simulation cell by at least 2 – and, perhaps, 3 – orders of magnitude. Secondly, the computational cost of the *ab initio* PESs is large because explicit methods of calculating electron correlation will be necessary for reactive processes. These methodologies have computational scaling factors that range from $O(M)$ to $O(M^7)$, where M is the number of electrons in the system, for a single energy evaluation. Furthermore, these *ab initio* methods will be called $\sim 10^5$ times for simulations of $\sim 10^2$ picosecond lengths. Thus, in order to accomplish MD simulations of larger systems for longer simulation times, there must be some mitigation of the computationally expensive *ab initio* calculations in favor of another way of calculating the energy and gradients of the energy.

Contained in Fig. 1 is a flow diagram of the *AMolDC* program.^{9,10} Briefly, the simulation begins with input coordinates (X), velocities (X_{dot}), and simulation temperature (T). The *MakeGroups* module takes those input geometries and makes all the spatially resolved groups. The MD routine then seeks to propagate the coordinates once all the forces on the atoms are calculated. In order to calculate these forces, the code will first check the *PESDatabase* (diamond: Is there sufficient data to interpolate?). If there is sufficient data, *AMolDC* will interpolate the gradient to produce the force, if not, *AMolDC* will call a reactive FF and the gradient will be used and stored in the *PESDatabase*. The simulation will continue until all the forces have been calculated and an MD step will be performed.

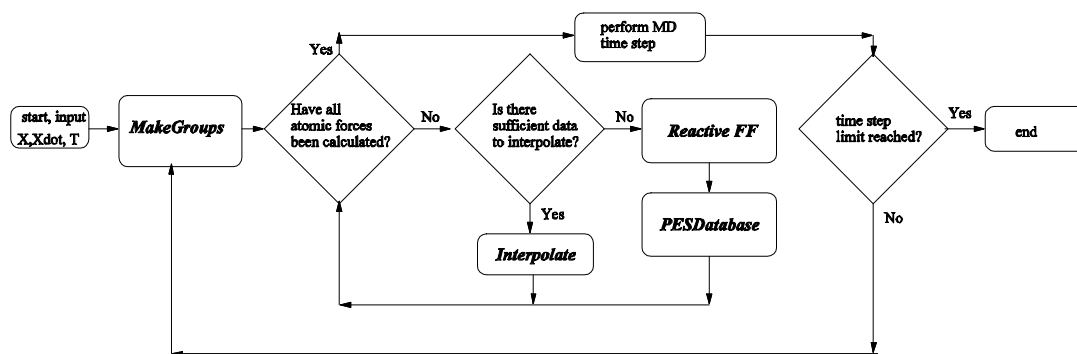


Fig. 1 The flow diagram for the *AMolDC* program

The simulation will continue along the flow diagram until the number of time steps has exceeded input criteria. The result of the MD simulation will be coordinates and velocities of all the atoms at all times of the simulation. One can take these data and analyze the reaction by plotting the coordinates or writing additional codes to extract whatever information the user needs. One way to view the data is through charts, and the other way to view the data is through animations.

II. Web-based Approach to Molecular Dynamics Simulations

The method described in Section I has been well-studied and has been implemented in C language based on SUSE Linux. However, it is not geared towards ordinary chemists. The C programs coming from the ACS proposal need to be compiled in order for chemists to use. It is a stretch to require chemists to know the details of a C compiler with its numerous options. Even though Linux has gained popularity among power users, it is still a long way before it is widely used by a diverse user base. Even if these C programs would have been ported to Windows or Mac environments, end users would still need to have a detailed knowledge of computer compiler in order to use it. A good approach would be to have a Web frontend so that chemists are completely shielded from the computer-related details associated with *AMolDC* program.

PHP^{11, 12, 13} is a widely used open source scripting language for web development. It has been successfully used by many scientific projects. The following three steps are proposed to “webify” the *AMolDC* program.

Step 1: Replace the input of the *AMolDC* program with PHP-based HTML form input.

This step requires the detail analysis of the *AMolDC* program and set a specification for the input. Some input parameters are simple options that users can select from the pull-down menu but some input parameters require the keyboard entries.

Step 2: Replace the output of the *AMolDC* program with graphs created by PHP functions.

Currently the output from *AMolDC* program is in the text form, columns of numbers (simulation time, atom number, coordinates, and velocities). Later on, users need to draw graphs by themselves based on the data in text form. In our view, it is a primitive way of presenting scientific results. We would like to directly create graphs based on the simulation results.

Step 3: Replace the C-programmed *PESDatabase* with MySQL^{14, 15, 16} database.

This is exploratory in nature because *PESDatabase* has a very strict requirement on performance. Based on our previous experience, commercial databases are too slow for the rigors of MD simulations. In this step, we need to experiment with using MySQL and if it is possible to save the most import subsets of the *PESDatabase*.

Step 4: Replace the C codes with PHP codes.

This is exploratory in nature because we don't know the potential performance penalty. The rule of thumb is that C codes are much faster than PHP code and we want to explore the difference between these two approaches.

III. Literature Survey

VMD (Visual Molecular Dynamics, <http://www.ks.uiuc.edu/Research/vmd/>) is an open source software funded by the National Institutes of Health. GROMACS (Groningen Machine for Chemical Simulations, <http://www.gromacs.org/>) is another software package for biochemical molecular simulation. CHARMM (Chemistry at HARvard Macromolecular Mechanics, <http://www.charmm.org/>) is a generic simulation software originally developed by the Chemistry

Department of Harvard University. All three of them are widely used in the chemical community, but none of them are web-based applications. In order for chemists to use, many prerequisites are required in order to setup the environment and run the system. In my view, the burden of learning computer skills in order to use the packages is not necessary for ordinary chemists. The work proposed in this proposal is to create a web-based interface to the research results generated from Dr. Salazar's ACS research proposal.

IV. Plan for Completion and Dissemination

Since I have worked with Dr. Salazar before, it is relatively easy to setup the research environment for this project. In addition to the laptop that I mainly use for development, I have access to the new B-20 lab dedicated to computer science department. I also have access to the SUSE Linux cluster setup by our Computing Services Department.

The project will start at the end of Spring 2008 and finish at the end of Summer 2009. The project will be developed according to the agile method of software engineering practices. The project is divided into four phases: analysis, design, implementation and testing. In the analysis phase, the input / output of the system will be identified by analyzing the C programs provided by Dr. Salazar. In the design phase, the functions for modules will be clearly specified. In the implementation phase, an Eclipse-based PHP development environment will be used to write PHP codes to implement functional modules. In the testing phase, an external web site will be setup so that outsiders can run the simulation through the web frontend. Invitations will be sent out to researchers all over the world that are interested in using the simulation tools for their research.

The result from the project will be a research paper suitable for conference / journal publication. Since this project is inter-disciplinary in nature, it is unknown at this moment whether the paper is going to be submitted to a Chemistry-related venue or Computer Science-related venue.

V. Budget Proposal

From my research collaboration with Dr. Salazar, I am able to get all necessary equipments (laptop, software, mouse, etc.) for conducting the research from his ACS grant. Since PHP, MySQL, and Eclipse are open source software, it does not cost money to get them. Union's existing computing infrastructure is enough for me to work on the project. The requested budget is \$4500 dollars for Dr. Haifei Li's work in the summer of 2009.

VI. Integration of Faith

One can easily see that the proposed work is right at the crossroads between computer science, chemistry, physics, and mathematics. Unfortunately, Christian Studies is not, at least at a first glance, also at this cross section. This proposal does engage the Christian faith, albeit indirectly. It is relevant to the Christian faith in so far as academic research in any scientific field is relevant and that by faculty proclaiming that Christ is Lord of all. There is not 1 square inch (or in this case, not one cubic nanometer) where what happens in that place is not controlled by God and, therefore, able to be used as a means for the glorification of God. The proposal is completely about understanding the workings of God in Christ in the fundamental processes of complex chemical reactions, the kind that propels our cars down the road every day. Nobody knows the exact chemistry of these reactions, but many research faculty members at very prestigious universities and laboratories attempt to guess the chemistry. The proposal seeks to build utilities on top of the *AMolDC* code for more general applications, but the *AMolDC* code itself is simply attempting to discover the fascinating way in which God works His wonders at the atomic level. Certainly at Union University we proclaim God's sovereign handiwork in history, biology, mathematics, and also in the fundamental details of how ubiquitous chemical reactions actually take place.

References:

1. Susnow, R.G.; Dean, A.M.; Green, W.H.; Peczak, P.; and Broadbelt, L.J. "Rate-Based Construction of Kinetic Models for Complex Systems" *J. Phys. Chem. A*, **1997**, 101, 3731-3740.
2. Gardiner, W.C. *Combustion Chemistry*; Springer-Verlag: New York, 1984.
3. Oran, E.S.; Boris, J.P. *Numerical Approaches to Combustion Modeling*; Progress in Astronautics and Aeronautics Series, Vol. 135; American Institute of Aeronautics and Astronautics: Washington, DC, 1991.
4. Moriarty, N.W.; Frenklach, M. *Proceedings of the Combustion Institute* 2000, 28, 2563.
5. Scheutz, C.A.; Frenklach M. *Proceedings of the Combustion Institute* 2002, 29, 2307.
6. Brown, A.; McCoy, A.B.; Braams, B.J.; Jin, Z.; Bowman, J.M. *J. Chem. Phys.* 2004, 121, 4105.
7. Schatz, G.C.; Horst M.T.; Takayanagi T. In *Methods for Multidimensional Dynamics Computations in Chemistry*; Thompson, D.L., Ed.; World Scientific: Singapore, 1998, p 1.
8. Shatz, G.C. In *Reaction and Molecular Dynamics*; Lecture Notes in Chemistry, Vol. 14; Lagana, A; Riganelli, A.; Springer: Berlin, 2000, p 15.
9. Salazar, M.R. "Molecular dynamics of complex gas-phase reactive systems by time-dependent groups", *J. Phys. Chem. A*, **2005**, 109, 11515-11520.
10. Salazar, M.R., Johnson, M., Duncan, W., Soller J., "Methods for Performing Molecular Dynamic Simulations Using Ab Initio Quantum Mechanical Potentials", United States Patent and Trademark Office, Application # 60/732,957.
11. Lerdorf R., Tatroe K., and MacIntyre P. *Programming PHP*, 2 edition, O'Reilly Media, Inc., 2006.
12. Zervaas Q., *Practical Web 2.0 Applications with PHP*, Apress, 2007.
13. Zandstra , M., *PHP 5 Objects, Patterns, and Practice*, Apress, 2004.
14. Tahaghoghi S.M.M., and Williams H., *Learning MySQL* O'Reilly Media, Inc. 2006.
15. Kruckenberg M., Pipes J., *Pro MySQL*, Apress, 2005.

16. DuBois P. *MySQL, 4th Edition*, Addison-Wesley Professional, 2008