Convergence DCL / Centers for Chemical Innovation - Summary for Discussion

Understanding, predicting, and designing molecular properties and behavior is a grand challenge of the chemical sciences. In particular, much progress has been made in understanding and modeling the fundamental physics that govern intermolecular interactions and conformational dynamics. Advancing and applying this understanding to predict molecular properties accurately and quickly is a crucial modern challenge that would enable the design of new classes of important molecules such as solvents, chemical probes, polymers, medicines, and surfactants. Free energies of binding, adsorption, and solvation are examples of properties that are important to predict accurately, due to their close associations with molecular design criteria such as binding affinity, selectivity, and solubility.

Meeting the property prediction challenge involves building molecular mechanics force fields that produce accurate simulation results compared to experiments, are inexpensive to compute, and transferable across diverse chemical structures. Current approaches fall short of meeting this challenge due to the lack of unified databases, methods, and infrastructure for systematically improving force fields as new data is continually made available. The core of most existing force fields was built by hand using very small data sets and limited computer power in the 1980s and has been little changed since then due to these deficiencies. Breaking this impasse requires a convergent effort of experimentalists and theoreticians across the mathematical, physical, chemical and life sciences because:

1. Simulating molecular properties precisely requires expertise in **molecular simulation, statistical mechanics, and physical chemistry**;
2. Parameterizing force fields accurately requires expertise in **statistical inference**, **optimization**, and **machine learning**;
3. Harnessing experimental and computational datasets to synthesize predictive models, one of NSF's Ten Big Ideas (Harnessing the Data Revolution), requires expertise in **data science, quantum chemistry**, and **experimental physical chemistry**;
4. Generation of high-quality data for parameterizing and validating force fields requires expertise in **experimental synthetic and analytical chemistry**;
5. Identification of the most important chemical problems requires expertise in **chemistry**, **biochemistry**, and **materials science**;
6. Comprehension and representation of the diverse chemical space requires expertise in **chemistry and cheminformatics**;
7. Reliable and reproducible execution of the whole project at large scales requires expertise in **computer science, software engineering,** and **distributed computing**.

We have assembled a correspondingly intellectually diverse team of experts with strong shared interest in this problem. The team members have wide-ranging departmental affiliations and cover a broad space of intellectual traditions and methodologies (synthetic chemistry, Bayesian inference, cheminformatics, quantum chemistry, optimization, artificial intelligence / machine learning, thermochemistry, modern software engineering), and have individually and jointly published in a wide range of journals spanning disciplines including organic and physical chemistry, physics, biophysics, chemical biology, cheminformatics, and theoretical modeling.

Together we created the **Open Force Field Initiative**, whose mission is to build open, accurate force fields for molecular simulation to enable quantitative molecular design supported by open, modern software infrastructure. Our initiative will develop advanced methods to synthesize new and existing experimental and quantum chemical datasets to build force fields for predictive molecular design across the chemical sciences, ensuring they can be widely applied to a broad range of chemical systems in major software packages. The convergence of our disciplines has already produced key innovations, such as a modern, flexible force field specification infrastructure that overcomes the limitations of atom typing in traditional force fields through the use of standardized direct chemical perception trees (DOI:10.1101/286542). In this project, we aim to develop the theory, algorithms, datasets, and infrastructure necessary to rapidly produce iteratively optimized force fields that can predict properties of molecular liquids and polymer precursors relevant to molecular design, including densities, enthalpies of mixing, vapor pressures and partition coefficients. We expect this will represent a fundamental and potentially transformative advance in predictive molecular simulation that spans multiple scientific disciplines and NSF directorates.

Our Initiative is founded on the principle of openness, and the results and tools that result from our research will be released with permissive open source licenses allowing both academic and industrial research groups to benefit. In coordination with the NSF-funded Molecular Sciences Software Institute (MolSSI), we have fully adopted best practices and tools of the modern Python software ecosystem for engineering, testing, and deploying code—such as package managers, continuous integration, code coverage reports, and documentation—ensuring our software is easy and install and produces reproducible results in diverse research computing environments. We will coordinate with developers of major simulation packages (e.g. Amber, CHARMM, GROMACS, NAMD, LAMMPS) to work toward native support of force fields and infrastructure generated by this effort, though translators will be available in the interim. Further dissemination, training, and community engagement will be carried out in coordination with MolSSI, which sponsors both workshops focusing on interoperability and training for students entering the discipline. Through these workshops we will introduce the infrastructure for enabling force field science, enabling students to bring these new concepts and tools back to their home institutions.

The member PIs of the proposed collaboration are as follows:

- John Chodera, Computational & Systems Biology, Sloan-Kettering Institute
- Michael Gilson, Pharmaceutical Chemistry, University of California at San Diego
- Shantenu Jha, Electrical & Computer Engineering, Rutgers University
- David Mobley, Pharmaceutical Sciences, University of California at Irvine
- Michael Shirts, Chemical Engineering, University of Colorado at Boulder
- Zhiqiang Tan, Statistics, Rutgers University
- Lee-Ping Wang, Chemistry, University of California at Davis