MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
MERT SENGUL	Materials Science and Engineering	Pennsylvania State University	mys12@psu.edu

Project Title

Development of a data-driven method to predict ReaxFF force field parameters.

Project Synopsis (approx. 100 words)

The optimization of force field parameters is critical during the development of the ReaxFF potential. The initial parameter values in optimization process affects the quality of the converged force field and time required for this process. This project intends to develop a data-driven method to predict initial force field parameters to be used during the optimization of the ReaxFF potential. The target is to use the reference values (e.g. bond length, partial charges etc.) as inputs for prediction. The goal is to decrease the force field optimization time and increase the force field quality by a better selection of initial parameter values.

Identified Data-Science Collaborative Need (approx. 100 words)

The ReaxFF force field includes approximately 100 parameters for each atom type to define the inter and intra-molecular interactions between them. Each parameter has different effects on these interactions and there is a correlation between most of the parameters. There is a need of a method that can learn the relationship between force field parameters and molecular properties. The method will make predictions for the parameters of a force field when the molecular property values for specified molecular geometries are known.

Data Origin and Access (*data must be available and sharable with data science teams* – please

address: data source/origin, access privileges, sharing privileges)

The training data set is composed of learning and testing parts. The data in the testing part consists of already developed and verified values, and those in the learning part is generated by atomistic simulations. The learning data consists of different combination of parameters and corresponding molecular properties, thus has the information about the effect of change in each parameter to molecular properties. We will make a series of published ReaxFF training sets available to the data science team.

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

Atomistic simulations have been used to determine structural and thermodynamic properties of crystalline, glass or liquid systems. The quantum mechanics (QM) based atomistic simulation methods have been commonly used to investigate the reactions in materials science. The QM methods provide accurate energies, charges and reaction pathways. However, simulation times and sizes are limited due to high computational cost, which is the motivation behind the development of empirical potentials. These potentials provide fast access to forces and, in turn, dynamical evolution. However, due to unalterable connectivity between atoms, traditional empirical potentials are incapable of modeling the evolution of systems during reactive events.

The ReaxFF is a reactive force field, capable of modeling large atomistic systems including reactive events for long simulation times at a wide range of temperature values. **The ReaxFF has proven itself to be reliable with its large user population and through around 700 publications in literature** [1, 2]. The ReaxFF consists of approximately 100 parameters per element type, and the force field parameters are grouped into six sections (Figure 1A). These sections are (number of parameters in parentheses): 1) General parameters (*34*), 2) Atom parameters (*21*), 3) Bond parameters (*15*), 4) Off-Diagonal Terms (*6*), 5) Valence angle parameters (*7*), 6) Torsion angle parameters (*5*). The inter and intra-atomic interactions are defined by these parameters. The parameters are optimized to reproduce reference values, which are molecular properties (e.g. bond lengths, bond angles, charges and energies etc.) of reference systems (Figure 1B). The reference values are obtained by QM methods or experiments, and are compared with those generated by the ReaxFF during the optimization process. The goal of the parameter optimization is to search for a parameter value in a defined interval, until the error defined in Equation 1 is minimized.

$$Error = \sum_{i=1}^{n} \left[\frac{X_{i,ReaxFF} - X_{i,training}}{\sigma_i} \right]^2$$

where the denominator is the difference between the reference value and the ReaxFF generated value for molecular property, and the numerator is the relative weight of the molecular property in optimization. The parameter optimization procedure is an iterative process and for each cycle, it is done by a sequential search over all parameters. Because most of the parameters are correlated, iteration continues until the error converges. In each cycle, one of the parameters is selected and optimized by fitting a parabolic function to three data points. One of these data points is the initial parameter value, and the other two points are the deviations from the initial value in two directions. The fitted parabola is minimized to obtain the optimized parameter value. The cycling terminates once the optimization of all parameters is complete.

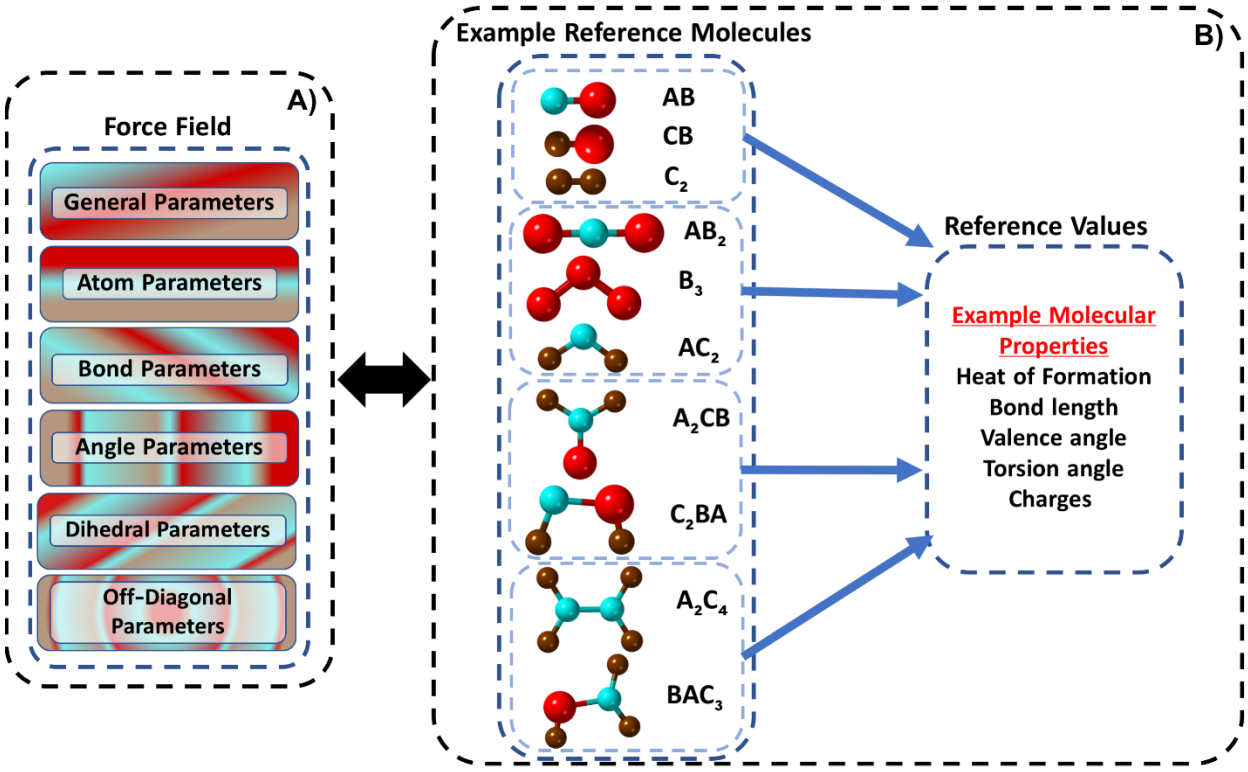

Figure 1. The representation of the training data set. (A) The subsections of a typical the ReaxFF force field. (B)(Left) Some examples for reference geometries that are used to optimize the force field parameters. Atom type-Color: A-Blue, B-Red, C-Brown. The force field parameters are optimized to reproduce the reference values of molecular properties for each reference molecule. (Right) Some examples of molecular properties.

In order to obtain a good force field, the force field parameter landscape should be explored thoroughly, but, this results in longer optimization times. In order to decrease the optimization time required for convergence of parameters, the parameter scan interval can be decreased. However, this limits the range and results in local convergence. For example, for identical reference systems, two different combinations of parameters might reproduce the same molecular properties. However, a limited scan of parameter space

detects only one of the combinations as the optimized force field. A solution to this problem without sacrificing from the optimization time and accuracy would be to start with initial parameters that can reproduce molecular property values close to reference values. **A data-driven learning method which can unveil the relation between force field parameters, and the properties of different molecular geometries could accomplish this goal.**

The collaboration involves the development of the data-driven method by data science team and the provision of the training data set by materials science team. The training data set is composed of learning and testing parts. The testing part of the data set has the same structure with the learning part, but the data are from optimized and verified values from past studies. The learning part consists of two sections. The first section is a set of combinations of different force field parameter values in defined parameter-specific intervals. The second section is a set of molecular property values computed for different molecular geometries by using each of the force field combinations in the first section. As can be seen in Figure 1B, there are several possible different molecular geometries composed of three different atom types and these molecular geometries are grouped into subsections depending on the number of atoms in the reference system. For this study three different atom types are used and it can be extended in the future.

The planned size for the learning part of the training set is approximately ten thousand different force fields and corresponding molecular properties computed for twenty different molecular geometries. Because both sections of the learning part are composed of subsections, this sectioned structure enables the ability to adjust the training set size. For example, the number of reference molecules or the number of molecular properties for each reference molecule can be increased or decreased, or some of the force field parameters can be kept constant. **These adjustments can be done with the assistance of materials science team according to the requirements of the methods applied by data science team.**

Improvement in the ReaxFF force field optimization will be beneficial to all users in the literature to develop or modify the force fields in line with requirements. **Depending on the results obtained by data science team, this study can be extended in the future in two ways:** First, the method can be used for optimization, or can be combined with other advanced optimization methods applied to the ReaxFF such as genetic algorithm [3], Monte Carlo algorithm [4] to enhance them. Second, the method can be modified to track the error evolution with the alteration of the force field parameters, which will improve the parameter search process used in optimization procedures.

References:

1. T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama and A. C. T. van Duin, npj Computational Materials **2** (1), 15011-15011 (2016).
2. S. D. Adri C. T. van Duin, Francois Lorant, and William A. Goddard III, Journal of Physical Chemistry A **105**, 9396-9409 (2001).
3. Henrik R. Larsson, Adri C. T. van Duin, Bernd Hartke, Journal of Computational Chemistry **34**, 2178–2189, (2013).
4. E. Iype, M. Hutter, A.P.J. Jansen, S.V. Nedea, C. C. M. Rindt, Journal of Computational Chemistry **34**, 1143–1154 (2013).