ParFit Manual

ParFit 1.0

Manual Updated:

January 8, 2015

Contents

I. Introduction 3

II. Conventions 3

III. Getting Started 4

IV. File Structure 4

V. Running ParFit 4

VI. Input File 5

VII. QM Data Files 6

A. Full 7

B. Compact 7

VIII. Output 8

IX. Utilities 9

Input Generator 9

Energy Profile Plotting 9

X. Contacts 9

Principal Investigators 9

Developers 10

XI. References 10

MM3 10

MMFF94 10

I. Introduction

MM3 and MMFF94 molecular mechanics (MM) methods use empirically fitted parameters to calculate molecular properties including energies and geometries. However, the parameters for a particular set of atom connectivity may not exist or the default parameters may not produce the desired level of accuracy. The parameters may be generated or their accuracy may be improved by manually adjusting parameters such that they accurately describe a subset of molecules. A good set of parameters is transferable; they can be used to accurately calculate molecular properties of molecules that resemble the set of molecules used to fit the parameters. The parameter fitting process can be time consuming and ParFit has been developed to partially automate the process.

ParFit is a Python program designed to determine MM3 or MMFF94 parameters for dihedral angle rotations about a bond, referred to as torsions in this document. ParFit generates parameters by adjusting a set of user identified MM parameters until the MM energy profile matches, within a certain threshold, a user supplied quantum mechanics (QM) derived energy profile. The ParFit derived parameters can then be used to calculate MM3 or MMFF94 properties for larger molecules. The current version can fit multiple torsion parameters in a single run; future versions will also fit bond length and bond angle parameters.

The manual contains: conventions used; where one can find the modules, programs and tools to get started; a description of the file structure; and instructions on how to compile and run ParFit including a description of the input files and utilities.

It is assumed that the user is familiar with QM and MM calculations, running Linux command line programs, and has basic knowledge of Python. Please refer to the websites in the *Getting Started* and *References* sections for more information relating to QM, MM, and the Python modules.

II. Conventions

- Pay close attention to commas. If a comma is needed, it will be included in the example input line.
- [ ]: square brackets should be replaced with the values described within.
- **Bold font format**: keyword options separated by "/", choose one.
- `fixed width format`: indicates a command to be entered in the terminal command line.
- atom index: indices refer to the position of an element in the list of coordinates (found in the Gamess output files or in the compact file) **starting from one**.
- torsion index: these indices **start from zero** and refer to the order of the torsions listed in the ParFit input file when optimizing parameters using multiple torsions.

III. Getting Started

In order to run ParFit, the following programs and modules are required:

*Python	http://www.python.org/	*Version 2.7.0 or higher but not version 3
Numpy	http://www.numpy.org/	
Scipy	http://scipy.org/scipylib/index.html	
Deap	https://pypi.python.org/pypi/deap	
MM Engine	http://serenasoft.com/	

Optional codes and libraries:

Gamess	http://www.msg.ameslab.gov/gamess/	
*Matplotlib	http://matplotlib.org	*optional Python curve plotting program

MM Engine is proprietary molecular mechanics software that may be obtained by contacting Serenasoft directly. Gamess is a freely available electronic structure code. Matplotlib is an optional Python program that can be used to plot data; alternatively, one can use other software such as Microsoft Excel.

IV. File Structure

The ParFit parent directory contains the subdirectories: Data, Doc, ParFit, and Utils. The Data subdirectory contains three subdirectories: Data/Engine, Data/Gamess, and Data/ParFit. The QM calculation data, in the form of Gamess output files or a compact file (data format described in Section VII), should be put in the Data/Gamess directory. The Data/Engine directory is where data relating to MM calculations are stored during runs, and Data/ParFit is where output data for each ParFit run is stored. This manual is included in the Doc subdirectory. The ParFit subdirectory contains the executable file: ParFit.py, the ParFit library, and example input files that start with: dih_scan_inp. The utilities found in the Utils directory are described in Section VIII. For more information regarding the input and output files from MM Engine and Gamess, please visit the dedicated websites, http://serenasoft.com/ and http://www.msg.ameslab.gov/gamess/.

V. Running ParFit

Navigate to the ParFit subdirectory where the ParFit.py file is located. From that directory, run ParFit by using the command:

```
./ParFit.py
```

or, to direct the output to a file, using the standard Unix command format:

```
./ParFit.py > [output_filename]
```

It may be necessary to end the program using the standard Unix command, Control+c. If this is the case, the parameters in the last step are found in the output files. The output and output files are further explained in Section IX.

The program takes its directives from an input file located in the same directory. The input file can be created in a text editor, or generated by using the PFinp.py program that is located in the Utils directory. Use of the PFinp.py program is recommended, as it generates the input file in the proper format. The input file format is further explained in Section VI.

ParFit fits parameters such that the MM energy profiles correlate to the quantum mechanically derived energy profiles. ParFit can read the QM data from one of two formats: a compact file that includes the energy of the molecule for a series of fixed torsion angle values or separate Gamess output files, one for each torsion angle value. The QM compact data file format is further described in Section VII.

VI. Input File

ParFit can be used to optimize MM parameters using one or multiple torsions. A ParFit input file can be created using the interactive PFinp.py program or a user may choose to create or modify an input file using a text editor. Following is the description of an input file, should the user choose to create one without the PFinp.py program.

An input file for optimizing torsion parameters has the form:

- Line 1 -

mult, [number of torsions]

> Enter the number of torsions that will be used to fit parameters.

- Line 2 through N -

[**full/comp**], [root filename], [atom index 1] [atom index 2] [atom index 3] [atom index 4], [starting torsion] [final torsion] [torsion step size]

> For each torsion, choose the type of QM data available, “comp” for a compact file, or “full” if you have separate QM files for each fixed torsion angle. The “filename root” should be the root filename of the QM data single compact data file or the series of files for that torsion (See section VII for more information on QM data formats that ParFit can read). The molecular structure is taken from the QM files; the atom indices in the input file and the QM files must match. The atom index is the position of the atom in the list of atomic coordinates starting from one. The four indices in each line must correspond to the four atoms making up the torsion angle that is varied to create an energy profile: starting, final, and step size torsion angle (in degrees). The torsion angles should correspond to the values used to create the QM energy profile.

- Line N + 1 -

[MM engine executable path]

Indicate the entire path where the MM engine executable file is located.

- Line N + 2 -

[**mm3/mmff94**]

Indicate the MM type, mm3 or mmff94.

- Line N + 3 -

[**ga/fmin/hybr**]

Choose the algorithm to be used to fit parameters. The options are: the "ga" genetic algorithm, the "fmin" Nedler-Mead simplex algorithm, or the "hybr" hybrid genetic algorithm followed by Nedler-Mead simplex algorithm.

- Line N + 4 through M, repeated for each torsion set of parameters to be fit -

[parameter line number] [**c**/**p**] [**c**/**p**] [**c**/**p**]

This block of lines instructs ParFit on which parameters to modify in order to fit the QM data. The "parameter line number" refers to the torsion line number in the default MM parameter file, "add_MM3.prm_orig" or "add_MMFF94.prm_orig" located in the ParFit_root_dir/Data/Engine/ directory. Each MM3 or MMFF94 torsion is fit by three parameters. The user can choose which of these parameters to adjust by selecting "p" or to keep constant by selecting "c."

- Line M + 1 (last line) -

[**csv_on/csv_off**]

Choosing "csv_on" directs ParFit to print comma separated value (csv) files, for each torsion, comparing QM and MM energy profiles at every 10th optimization step. See Section VIII for more details about the output.

To generate Gamess input files, the user may run the Gamess input generating program using the command:

```
./Ginp.py
```

The program requires a one line input file with the form:

ginp, [root filename], [atom index 1] [atom index 2] [atom index 3] [atom index 4], [starting torsion] [final torsion] [torsion step size]

The root filename should correspond to the name of a template Gamess input file. An example template Gamess input file is provided in the Gamess subdirectory, it is called: opmmm-mp2-popt-dd.inp.

VII. QM Data Files

ParFit requires that the QM energy profile be supplied in a particular file and filename format. ParFit can read QM data in either full Gamess output files, one for each fixed torsion angle, or a compact file that includes the geometry at various fixed torsion angles and its corresponding energy. The formats for the QM data files are outlined below.

A. Full

When the “full” keyword is given in the ParFit input file, ParFit expects full Gamess output files in the /Data/Gamess subdirectory. One output file for each torsion angle should be supplied. The following naming scheme must be followed:

[root filename]-[torsion angle (rounded to an integer)].log

B. Compact

When the “comp” keyword is given in the input file, all of the QM data (each structure, energy and fixed torsion angle value) must be contained in a single file. The “comp” keyword may be used with data from any electronic structure program as long as the molecular properties are contained within a file with the following name and format. The file name should be:

[root filename]-scan

and should have the following format:

- Line 1 -

[atom index 1] [atom index 2] [atom index 3] [atom index 4]

The atom indices correspond to the atoms making up the torsion angle being used to fit MM parameters.

- Line 2 -

[initial torsion] [final torsion] [torsion step size]

The initial torsion, final torsion, and torsion step size should correspond to the torsion values for which structures are included in the file.

- Line 3 -

[number of structures included in the comp file]

The total number of structures (N) included in the file should be:

$$\mathrm{N} = \frac{\mathrm{n_f} - \mathrm{n_i}}{\mathrm{n_s}} + 1\ .$$

where $\mathrm{n_f}$ and $\mathrm{n_i}$ are the final and initial torsion values, respectively, and $\mathrm{n_s}$ is the step size.

- Line 4[¥] -

[root filename]-[torsion angle (rounded to integer)] [number of atoms in the structure] [actual torsion] [structure energy in Hartrees]

Line 4 contains the root filename with the rounded integer torsion value appended after a dash, the number of atoms in the molecule, the actual torsion value used in the QM calculation, and the QM energy in Hartrees.

- Line 5[¥] -

[chemical symbol] [x-coordinate] [y-coordinate] [z-coordinate] [atomic number]

Line 5 starts a block of lines containing the atomic coordinates of the molecule.

[¥] Lines 4 and 5, should be repeated to include the geometry and energies of the molecule at each torsion angle value.

VIII. Output

The status of the ParFit run will be printed to the terminal. The print out will depend on the fitting algorithm. If the genetic algorithm is running, the generation information will print: generation, number of evaluations, average difference, rmse, minimum and maximum rmse. The run is complete once the optimal parameters do not change throughout five generations. Upon completion of the genetic algorithm run, the lowest rmse and the corresponding parameters, which are listed in the same order as they are found in the input file, are printed. When running a Nedler-Mead simplex algorithm, the output is: the step number, the rmse, and the parameters, listed in the same order as they are found in the input file. The run is complete once the root mean square error, rmse, reaches the threshold value, 0.2.

If running ParFit with the hybrid option, ParFit will automatically switch to the simplex algorithm after completing a genetic algorithm optimization. The output for this run follows the genetic algorithm output and automatically switches to the simplex algorithm.

Along with the standard output, several output files are generated while running ParFit. Starting in ParFit version 1, a snapshot of the parameters and energy profile fits (provided the user chooses to have csv file printing turned on) comparing QM and MM energy are printed to the Data/ParFit/[input filename]/[root filename] subdirectory while ParFit is running. These files allow the user to visualize the MM and QM energy profile for every tenth step in the optimization process as well as the corresponding parameters.

The parameters are printed to the Data/ParFit/[input filename] subdirectory with filenames of the form:

add_[MM]_[N].prm

where MM is either MM3 or MMFF94 and N is the fitting algorithm step number. If the user chooses to print the comparison of MM and QM energy profiles, they are printed to the Data/ParFit/[input filename] subdirectory and have the filename format:

opt_[Y].csv

where X is the torsion index in a multiple torsion run and Y is the algorithm step number.

Other files generated by running ParFit are located in the Data/Engine subdirectory. These are the MM fixed dihedral angle geometry optimizations calculated during the parameter fitting process. These files end in “pcm” and may be deleted by the user at the end of the parameterization.

IX. Utilities

In the utilities directory are scripts that may be useful for running ParFit or collecting the data needed to run ParFit.

Input Generator

The input generating program is run by using the command:

```
./PFinp.py
```

from the Utils directory. The information needed is:

- number of torsions used to fit parameters
- QM data format – file(s) type, root file name
- atom indices making up the varied torsion angle in the QM calculations
- initial, final and step size of the torsion angle that was varied in the QM calculations
- parameter line numbers in the add_MM3.prm_add or add_MMFF94.prm_add file and which of the 3 parameters on that line are to be adjusted
- MM engine full path location
- MM method, either MM3 or MMFF94 for which you wish to get parameters
- algorithm used to fit the MM to the QM energy profile, “ga” for genetic algorithm or "fmin" for the simplex algorithm

Energy Profile Plotting

If matplotlib is installed and you’d like to use it to plot the QM and MM energy profiles, open the Python program file /Utils/QM_vs_MM_energies.py with a text editor, change the file name to the csv file to be displayed and run it using the command:

```
./QM_vs_MM_energies.py
```

The user may choose to change other variables in the file, such as labels, where indicated. Once the file has been modified, run it to get a plot of the QM and MM energy profile comparison. Note, the csv file must be in the same directory as QM_vs_MM_energies.py program.

X. Contacts

Principal Investigators

Mark S. Gordon, mgordon@iastate.edu

Theresa L. Windus, twindus@iastate.edu

Developers

Federico Zahariev, fzahari@iastate.edu

Marilu Dick-Pérez, marilu@iastate.edu

XI. References

MM3

Allinger, N. L., Yuh, Y. H., & Lii, J-H. (1989) Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. J. Am. Chem. Soc. 111, 8551-8565.

MMFF94

Halgren, T. A. (1996) Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94