ParFit Manual

ParFit 1.2

Manual Updated:

July 9, 2017

I. Contents

II. 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 MM calculations produce more accurate results for the particular 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. ParFit has been developed to partially automate the time-consuming parameter fitting process.

ParFit is a Python program designed to determine MM3 or MMFF94 parameters of bond lengths, bond angles, and dihedral angle rotations, 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, to within a certain threshold, a user supplied quantum mechanically (QM) derived energy profile. The ParFit derived parameters can then be used to calculate MM3 or MMFF94 properties of larger molecules by supplying the new parameters to an MM program such as PCModel (http://www.serenasoft.com/).

This manual contains the necessary information to run ParFit including: 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.

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 the QM, MM, and Python modules.

III. 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.
- torsion index: these indices refer to the order of the torsions listed in the ParFit input file when optimizing parameters using multiple torsions.

IV. Getting Started

The following programs and modules that are required to run ParFit:

*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
Mengine	included with ParFit under BSD license. To install, run: `make` in the ParFit/Mengine directory, where the MM ENGINE source code is located.

Optional codes and libraries:

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

Mengine is a proprietary molecular mechanics software, the source code is included with ParFit under a BSD license and copyright is held by Serena Software, Bloomington, IN 47402. GAMESS is a freely available electronic structure code. Matplotlib is a free optional Python program that can be used to plot data. Alternatively, one can use other software such as Microsoft Excel to plot data.

V. 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), reside in the Data/Gamess directory. The Data/Engine directory stores background data relating to MM calculations during runs, and Data/ParFit is where output data for each ParFit run is stored. This manual is included in the Doc subdirectory. The Mengine directory contains the molecular mechanics calculation source code. The ParFit subdirectory contains the executable file: ParFit.py along with example input files that start with: “scan_inp_”. The utilities found in the Utils directory are described in Section VIII. For more information regarding the GAMESS input and output files, please visit the dedicated website http://www.msg.ameslab.gov/gamess/.

VI. Running ParFit

ParFit takes directives from an input file called dih_scan_inp located in the ParFit directory where the ParFit executable file, ParFit.py, is also located. Run ParFit by navigating to the ParFit directory and use the following command:

```
      ./ParFit.py
```

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

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

A properly formatted input file can be generated by using the PFinp.py program located in the Utils directory. Alternatively, the input file may be generated using a text editor and following the format detailed in Section VI.

It may be necessary to end the program using the standard Unix command, Ctrl+c, such as if the program is not advancing (no printout to the terminal or log file), or if iterations of parameter fitting are not reaching the threshold value. If the calculation is exited, the parameters from the last step are found in the ParFit output files, Data/ParFit/. The input and output files are further explained in Section IX.

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.

VII. Input File

The parameters for three properties can be calculated with ParFit, the are: bond length, bond angle, and torsion angle parameters. The input file formats for calculating the parameters for each property are described below in detail.

A. Bond Lengths

Parameters are fit for one bond length at a time. The input file will necessarily have 6 lines and they are:

- Line 1 -

[**full**/***comp**], [*root_filename], [atom index 1] [atom index 2], [starting length] [final length] [length step size], *bond

> Choose the type of QM data available for the bond length potential energy curve, “comp” for a compact file, or “full” if you have separate QM files for each bond length. (*When selecting the “comp” format, the atom indices and bond length information is contained in the data file and thus it is not necessary here.) The “root filename” must match the root filename used for the QM data single compact data file or the series of files (please see section VII for more information on QM data formats that ParFit can read). The molecular geometry is read from the QM files, thus, the atom indices and the position of the atom in the list of atomic coordinates from the input file must match with the QM data file. The bond length range and step size (measured in angstroms) is also

included in line 1. The last keyword in line one directs the program on the property.

- Line 2 -

[Mexecutable path]

Indicate the entire path where the Mengine executable file is located.

- Line 3 -

[**mm3/mmff94**]

Indicate the MM type, mm3 or mmff94.

- Line 4 -

[**ga/fmin/hybr**] [**min**/index]

Choose the algorithm that will be used to calculate 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.

The optional second parameter choses the reference point for the RMSE minimizing computation. The MM energy curve is defined up to an arbitrary additive constant, which is fixed by equating the QM and MM energy curves at some value of the bond length that is being fit (reference point). The reference point is selected by either choosing the keyword "min", which keyword selects the global minimum, or by explicitly specifying an index of the array of bond-length values. If the second parameter is not specified, the "min" keyword is the default. The keyword "index" should be replaced by the point along the curve that is to be used as the reference point for the RMSE calculation. Please note, python uses the zero as the first index point.

- Line 5 -

[parameter line number] p p

This line instructs ParFit to modify the two bond length parameters in order to fit the QM data. The "parameter line number" refers to the line containing the bond length parameters in the default MM parameter file, "add_MM3.prm_orig" or "add_MMFF94.prm_orig" located in the ParFit_root_dir/Data/Engine/ directory. For more information on the parameters, please see the references.

- Line 6 -

[**csv_on/csv_off**] [step_int]

Choosing "csv_on" directs ParFit to print comma separated value (csv) files, for every "step_int" (see below) step in the parameter calculation. The files contain the QM energies for each fixed bond length and the corresponding MM energies calculated with the parameters at that step. The files are printed to the

Data/ParFit/[input filename]/[root_filename]/ subdirectory. The files are named “opt_[Y].csv” where Y is the step number.

The above “step_int” value might be specified by an optional second parameter. By default, it is 10.

B. Bond Angles

Like with bond lengths, there are 6 lines in the input file with two lines modified slightly. The first difference is in line 1 where 3 atoms need to be listed to define the bond angle, measured in degrees. The other optional change is in line 4 where you can choose the reference point on bond-angle MM energy curve to be one other than the default of the minimum energy point.

- Line 1 -

[**full**/***comp**], [*root_filename], [atom index 1] [atom index 2] [atom index 3], [starting angle] [final angle] [angle step size], *angl

Choose the type of QM data available for the bond angle potential energy curve, “comp” for a compact file, or “full” if you have separate QM files for each bond angle. (*When selecting the “comp” format, the atom indices and angle information is contained in the data file and thus not necessary here.) The “root filename” must correspond to the filename used for the QM data single compact data file or the series of files for that torsion (please see section VII for more information on QM data formats that ParFit can read). The molecular structure is read from the QM files. The input file atom indices, the position of the atom in the list of atomic coordinates, must match with the QM data file(s). For full QM data files, the angle information, starting, final and step size, are to be included in line one. Finally, the final keyword indicates the type of property to calculate parameters for, in this case “angl” for bond angle.

- Line 4 -

[**ga/fmin/hybr**] [**min**/index]

Choose the algorithm that will be used to calculate 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.

The optional second parameter choses the reference point for the RMSE minimizing computation. The MM energy curve is defined up to an arbitrary additive constant, which is fixed by equating the QM and MM energy curves at some value of the bond angle that is being fit (reference point). The reference point is selected by either choosing the keyword “min”, which keyword selects the global minimum, or by explicitly specifying an index of the array of bond-angle values. If the second parameter is not specified, the “min” keyword is defaulted.

C. Torsions

ParFit can be used to optimize MM parameters using one or multiple torsions. A ParFit torsion angle 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 are descriptions of the input file format for fitting a single torsion or multiple torsions.

- Line 1 -

mult, [number of torsions]

Enter the integer corresponding to 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. (*When selecting the “comp” format, only the root file name is needed. The atom indices and angle information is contained in the data file and thus not necessary here.) The “filename root” must be the root filename used for 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 read 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 must correspond to the values used to create the QM energy profile.

- Line N + 1 -

[Mexecutable path]

Indicate the entire path where the Mengine executable file is located.

- Line N + 2 -

[**mm3/mmff94**]

Indicate the MM type, mm3 or mmff94.

- Line N + 3 -

[**ga/fmin/hybr**] [**min**/index]

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.

The optional second parameter choses the reference point for the RMSE minimizing computation. The MM energy curve is defined up to an arbitrary additive constant, which is fixed by equating the QM and MM energy curves at some value of the torsion that is being fit (reference point). The reference point is selected by either choosing the keyword “min”, which keyword selects the global minimum, or by explicitly specifying an index of the array of torsion values. If the second parameter is not specified, the “min” keyword is defaulted.

- 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.” It is possible to fit coupled torsion PESs by making two parameters equal to each other by labeling them as “pN” where N is a number. The lines would then be:

[parameter line x] p1 c c

[parameter line y] p1 c c

In this case, parameter V1 would be adjust to the same number for two torsion parameters located on lines x and y. For a further explanation of a case where this may be useful, please see the How To document located in the /Doc subdirectory

- Line M + 1 (last line) -

[**csv_on/csv_off**] [step_int]

Choosing “csv_on” directs ParFit to print comma separated value (csv) files, for each torsion, comparing QM and MM energy profiles at every “step_int” (see below) optimization step in the Data/ParFit/[input filename]/ [root_filename]/ subdirectory. The files are named “opt_[Y].csv” where Y is the step number. For multiple torsion runs, there will be subdirectories under the input filename with the respective root filename for each torsion.

The above “step_int” value might be specified by an optional second parameter. By default, it is 10.

To generate a series of fixed torsion GAMESS input files, the user may run the GAMESS input generating program using the command:

```
./Ginp.py
```

There are example input files for torsions, bond angles, and bond length named ginp_inp, ginp_inp_angl, and ginp_inp_bond, respectively. The examples are located in the ParFit/ directory. Below is the format for the Ginp.py input files:

[root_filename], [atom indices separated by a space], [starting value] [final value] [value step size], [**/angl/bond**]

> The root filename should correspond to the name of a template GAMESS input file. An example GAMESS input file template is provided in the GAMESS subdirectory, it is called: opmmm-mp2-popt-dd.inp. The atom indices are 2, 3, or 4 integers that define the bond length, bond angle, or torsion, respectively. The next three values are the starting, final, and value step size, where the values for bond length are measured in angstroms and the values for angles are measured in degrees. The final keyword is the property type. If generating the default, torsion input files, no keyword is necessary. If generating bond angle or bond length input files, choose "angl" for bond angle or "bond" for bond length.

To generate a compact file from a series of fixed bond length, bond angle, or torsion GAMESS output files, the user may run the Gout program using the command:

```
./Gout.py
```

The program will generate a compact file named "[root_filename]-scan" and will be saved in the Data/Gamess/ subdirectory. The series of files must be named in the following format:

> [root_filename][fixed property value].log

where property is either "bond", "angl", or "diha" for bond length, bond angle, and torsion angle, respectively. The input file for Gout.py follows the exact format as the Ginp.py input file but needs to be named "gout_inp".

VIII. 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 fixed bond length, bond angle, or torsion angle is needed. The following naming scheme must be followed:

[root_filename]-[fixed property value].log

the fixed property value is the fixed bond length, bond angle, or torsion angle rounded to an integer.

B. Compact

When the “comp” keyword is given in the input file, all of the QM data (each structure, energy and fixed torsion 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 indices]

The atom indices corresponding to the 2, 3, or 4 atoms that make up the bond length, bond angle, or torsion angle, respectively.

- Line 2 -

[initial length/angle/torsion] [final length/angle/torsion] [length/angle/torsion step size]

The initial, final, and step size values for the length, angle, or torsion that is varied to generate the corresponding to the structures 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 values, respectively, and $\mathrm{n_s}$ is the step size.

- Line 4[¥] -

[root_filename]-[identifier] [number of atoms in the structure] [actual torsion] [structure energy in Hartrees]

Line 4 contains the file name root with a unique identifier after a dash (this can be the fixed length/angle or torsion value), 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 value.

IX. Output

Upon running ParFit, output 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 root-mean-square error (RMSE), standard deviation, minimum and maximum RMSEs. The genetic algorithm parameterization is complete once the individual does not improve for five consecutive generations. For further information on genetic algorithms, please refer to the DEAP documentation at: https://pypi.python.org/pypi/deap. Upon completion of the genetic algorithm run, the lowest root-mean-square error 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 RMS error, and the parameters, listed in the same order as they are found in the input file. The run is complete once the RMS error reaches the threshold value, 0.2.

If running ParFit with the hybrid option, the genetic algorithm will be used to determine a rough estimate of the parameters and will then be followed by the simplex algorithm to fine tune the fit. The output for this run follows the genetic algorithm output and automatically switches to the simplex algorithm once there is no improvement in an individual for five generations.

Along with the standard output, several output files are generated while running ParFit. Starting in ParFit version 1, a snapshot of the parameters is printed to the Data/ParFit/ subdirectory. If the user has the “csv_on” keyword, snapshots comparing the MM energy profiles to the QM energy profiles will be printed every 10 steps. The user may then plot the MM and QM energy profiles to visually evaluate the fit.

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. The MM and QM energy profile comparisons are printed to the Data/ParFit/[input filename]/[root_filename] subdirectory and have the filename format:

opt_[Y].csv

where Y is the algorithm step number. For bond length and bond angle runs, there is no X value since only one bond length or angle can be fit per run.

Temporary files generated by running ParFit are located in the Data/Engine subdirectory. These files end in “pcm” and may be deleted by the user at the end of the parameterization.

X. Atomic Database File

The file ~/ParFit/atomic.db contains data that specifies the default atomic charge (default_charge), MM3 type (default_mm3_type), MMF94 type(default_mmff94_type), covalent radius (cov_radii), and bond lengths corresponding to single, double, and triple bond orders (bond_ords). The user can change or/and extend the values in this database. See the comments inside the database file for details.

XI. Utilities

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

A. 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
- Mengine 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

B. 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.

XII. Contacts

A. Principal Investigators

Mark S. Gordon, mgordon@iastate.edu

Theresa L. Windus, twindus@iastate.edu

B. Developers

Federico Zahariev, fzahari@iastate.edu

Marilu Dick-Pérez, marilu@iastate.edu

XIII. References

A. 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.

B. MMFF94

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