A Tutorial for using the CRYOFF (Create Your Own Force Field) code to develop a force field for 1,4 butanediol in water.

Written by: Raymond Weldon (rjweldon@uark.edu)

Advisor: Feng Wang (fengwang@uark.edu)

Department of Chemistry and Biochemistry,

University of Arkansas,

Fayetteville, AR, 72701, USA

1. Introduction

You will be guided through the process of using the Adaptive Force Matching (AFM) algorithm to develop a force field for a solute in water. We will use 1,4 butanediol as an example.[1]

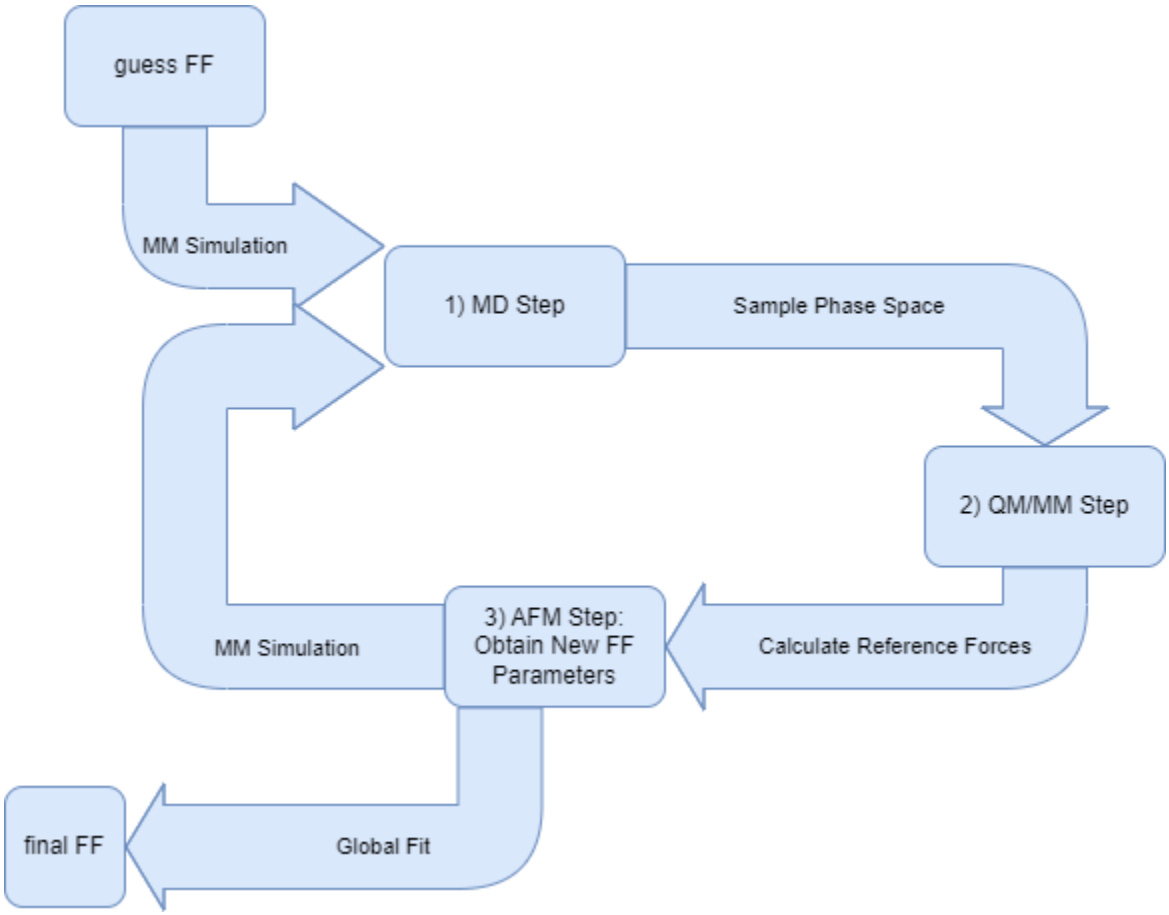

Figure 1: Complete AFM cycle, showing all stages of FF development. The first step of the cycle is the MD step to sample phase space; the second step is the QM/MM step where forces will be calculated for the QM region; the third step is the FM Step where the CRYOFF code will be utilized to fit the force field parameters to the QM forces. Once new parameters are found, the cycle can be repeated until convergence, and a global fit will be performed after convergence to obtain the final force field.

We assume the readers of this tutorial have experience doing MD simulations with Gromacs. We note that Gromacs versions released after 2020 no longer support tabulated potentials. Thus, we have to use earlier versions of Gromacs. We will begin with a gromacs xtc file produced by a previous MD simulation using a guess force field. With the help of AFM tools, we will iterate through the three steps of AFM. The QM calculations will be performed with ORCA, although the current AFM tools support other QM packages, including Gaussian, PQS, and GAMESS. We highly recommend reading the CRYOFF manual as well as the AFMtools manual, before starting any force field development on your own. Note that the AFM tools were written with perl under Linux. Thus, a working experience with Linux is assumed.

2. Setting Up the Tutorial Package

All files needed for this tutorial are located in the gzipped archive, 14bud_tutorial.tar.gz on the group webpage or can be cloned from github at https://github.com/uark-wanglab/AFM-Tutorial. Move the gzipped file where you would like to perform the tutorial on your Linux computer, and unzip it using the command

tar -xvzf 14bud_tutorial.tar.gz

Three directories listed below should have been extracted

AFMtools

CRYOFF

Tutorial

The AFMtools directory contains the 1.2.2 version of our group's AFM tools scripts to facilitate the AFM iterations; CRYOFF contains the code used to perform the force matching; Tutorial contains a complete example of one generation of force matching along with a directory that can be used to fit dispersion.

Please create a directory to perform the AFM using the Tutorial directory as an example.

Setting up AFMtools

To use AFMtools throughout this tutorial, please source the setenv.afm script.

Assuming the file is under /home/alex/14bud_tutorial/AFMtools

Type:

source /home/alex/14bud_tutorial/AFMtools/setenv.afm

This will put the AFMtools directory in your PATH. For frequent use of AFMtools, it may be convenient to put the above command in your .bashrc file.

Compile and Install CRYOFF

The best compiler for the CRYOFF executable is the intel Fortran compiler. For parallel execution, we will assume you have the intel MPI installed. Instructions for compiling and installing the CRYOFF fitting code is included in the CRYOFF manual.

Assuming the intel compilers and MPI environment is installed under

/usr/local/intel/oneapi

One possibility to compile CRYOFF is

source /usr/local/intel/oneapi/setvars.sh intel64

mpiifort -fpp cry3.0.0.f90 -DMPI -O3 -mavx2 -axAVX,core-AVX2,core-AVX512 -lmkl_lapack95_lp64 -lmkl_intel_lp64 -lmkl_intel_thread -lmkl_core -liomp5 -lpthread -lmpi -o cry.300.x

This should produce the CRYOFF run time executable cry.300.x

The source code for version 3.0.0 of CRYOFF and a precompiled executable has been included in the tutorial files. The precompile executable may or may not work for your system depending on system libraries. It is recommended to compile a version of CRYOFF from the source code.

Once the code is compiled. You may need to include it in the PATH so you can run it to perform the fit.

The atom typing for the 1,4-butanediole is shown in Figure 2, which is useful to understand the rest of the tutorial.

3. Fitting of Dispersion

In AFM, we find it is better to fit dispersion to either SAPT or Grimme's empirical dispersion correction for DFT before the AFM iterations. In this tutorial, we will fit to the Grimme D3 dispersion with Beck-Johnson damping (D3(BJ)).[2] Since fitting to D3 dispersion is somewhat similar to the workflow from the MD to the FM steps and involves the same set of input files, we will present the fitted parameters in Table 1 but will postpone the discussion of the dispersion fitting to Section 8.

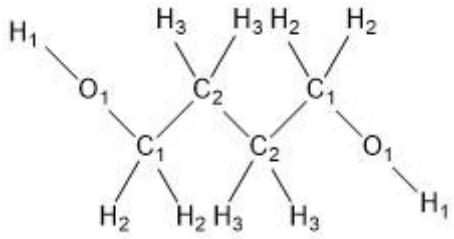

Figure 2: Atom types for the 1,4-butanediol molecule. With AFM, generally no two atoms use the same atom type unless they are equivalent by symmetry or share almost identical chemical environment. .

Atom1	Atom2	C6 (kcal/(mol Å^6))	R_0 (Å)
C1	O1	-840.73401	1.9970000
C2	O1	-660.18209	1.9970000
O1	O1	-339.99103	1.9494990
C1	C1	-1256.8644	2.0447500
OW	O1	-386.64814	1.9494990
OW	C2	-995.00435	1.9970000
OW	C1	-1354.3569	1.9970000

Table 1: C6 and R_0 for different atom pairs. These are the parameters for the short-range-damped (SRD) dispersion. *,A description of the SRD interaction can be found in the CRYOFF manual.*

4. Creating the input files for QM/MM calculations

We will start from a previous Gromacs simulation with the trajectory saved in the traj_comp.xtc file. While this initial guess simulation was run with an AFM model, this trajectory can be obtained with any force

field. From this file, conformations can be extracted as .gro files with the 01_runme.sh (Figure 3) script also in the Tutorial/01_QM_MM/298/conf directory.

```
1 #!/bin/bash
2
3 echo "2 0" | gmx_d trjconv -f traj_comp.xtc -s ../topol.tpr -n ../index.ndx -split 80 -dt 80 -o
  final.gro -nzero 3 -center -pbc mol
4
5 ls -1 final* | xargs ./BUU-master
6
7 mv final*gro grofiles
```

Figure 3: Script "01_runme.sh" to be run under the "Tutorial/01_QM_MM/298/conf" directory.

Line 3 calls the gromacs trjconv function to save 101 .gro files from 2 to 10 ns in the trajectory. Each configuration is centered with the 1,4-butanediol molecule in the middle of the box (the "echo 2 0" selects the 1,4 butanediol to be centered) and each water molecule is made whole if it is broken by the PBC. (-pbc mol) This is important as not all AFM scripts understand PBC, and many AFM scripts assume the center of the QM region is close to the center of the QM/MM configuration.

Line 5 passes the name of each of the configurations produced by the trjconv to BUU-master script to produce .pxyz files. A .pxyz file is an internal format used by AFM scripts that contains mark information to facilitate QM/MM region selection.

```
1 872
2 21.8503   21.8503   21.8503
3 C1   9.19   11.79   11.56   1BUU   4
4 H1   8.76   10.82   11.78   1BUU   4
5 H2   8.51   12.45   11.04   1BUU   4
6 C2   10.47   11.53   10.7   1BUU   4
7 H3   10.33   12.05   9.68   1BUU   4
8 H4   11.24   11.94   11.3   1BUU   4
9 C3   10.84   10.05   10.4   1BUU   4
```

Figure 4: General structure of a .pxyz file. Line 1 is an integer containing the number of atoms in the .pxyz file. Line 2 contains the box dimension (in Angstroms), and each line from Lines contain: Atom name, X, Y, Z, name of the molecule, and a mark value. (note that the .pxyz file can be viewed as a standard xyz file using graphical programs such as VMD.)

```
#!/usr/bin/perl

#mark values
$valBUU=4;  #BUU
$valup=3;   #solvation factor 1 water
$valqm=2;  #solvation factor 0 water
$valmm=1;    #MM water


$rcudio=3.0; #cut off for DIO
$rcuboundary=2.6; #cutoff for boundary
$rcumm=9.0; #cutoff for MM

foreach $file (@ARGV)
{
$basename=substr $file,0,-4;
@base_path=split(/\//, $basename);
$filename=@base_path[@base_path-1].".pxyz";
$xyzname=@base_path[@base_path-1].".xyz";

#gro to pxyz
$cmd1="gro2pxyz $file|mark_byname BUU $valBUU > $filename";

#mark DIO and water aroud DIO
$cmd2="mark_within_range 1 16 $rcudio $valqm $filename";

#randomly mark up five  water and their boundary
$cmd3="markup_random $valqm $valup $filename|markup_random 2 3|markup_random 2 3 |markup_random
2 3 |markup_random 2 3 > $filename._$$; mv $filename._$$ $filename";
$cmd4="mark_boundary $rcuboundary $valup $valqm $filename";

#change mark 2 waters to mark 3 waters if they are surrounded by waters with a mark greater than
 or equal to 2
$cmd5="markup_mol $rcuboundary  $valqm  $valup $filename > $filename._$$; mv $filename._$$ $file
name";

#mark mm
$cmd6="mark_within_range 1 16 $rcumm $valmm $filename";

#dropoff and sort
$cmd7="pxyz_dropoff 0 $filename | pxyz_sort > $filename._$$; mv $filename._$$ $filename";
$cmd8="pxyz_2vxyz $filename > $xyzname";

#execute
system("$cmd1");
system("$cmd2");system("$cmd3");
system("$cmd4");system("$cmd5");
system("$cmd6");system("$cmd7");
system("$cmd8");
}

```

Figure 5: "BUU-master" script as found in "Tutorial/01_QM_MM/298/conf"

The BUU-master script takes a gromacs configuration file and produces a .pxyz file that contains the QM/MM information. In AFM, the system is divided into a QM region and MM region. The QM region is further divided into a central zone and a buffer zone. Only the forces from the central zone are used for fitting. These regions and zones are distinguished by mark values. The central zone will have a higher mark value than the buffer zone, which in turn has a higher mark value than the MM region. Note that the AFM markup scripts (such as markup_random, markup_mol) will only increase the mark value and

will never decrease the mark value. Thus, if one molecule has been marked to be in the QM region, which has higher mark values, subsequent operations with a markup script will not lower its mark.

Line 22 converts a .gro file into a .pxyz file, and pipes the .pxyz file to the script “mark_byname” which will mark the molecule named “BUU” with a mark value of 4.

Line 25 calls “mark_within_range” script to mark all molecules within 3 Å of 1,4-butanediol (atoms 1 through 16 to a mark of 2.

Line 28 randomly selects 5 molecules from the first hydration shell of the 1,4-butanediol to be in the central zone of the QM region. (solvation factor 1). Thus, the markup_random script is called 5 times. In each call of markup_random one molecule with a mark of 2 will be marked up to a mark of 3.

Line 29 create the buffer zone(mark of 2) by marking any water within 2.6 Å ($rcuboundary) of any molecule in the center zone (mark of 3).

Line 32 will promote a molecule in the buffer zone to the central zone if there are no molecules with lower marks than 2 within 2.6 Å of such a molecule. Since we only fit forces for atoms in the central region, this step will allow more water to be fit as a buffer region molecules will be moved to central region if there is no nearby MM atoms.

Line 35 marks all molecules within 9 Å of 1,4-butanediol to 1, which will be MM water.

Line 38 cleans up the .pxyz file by removing all molecules with a mark of 0 or less and then sorts the .pxyz file from highest to lowest mark.. (Note the default mark created by gro2pxyz in Line 22 is -9)

Line 39 generates a .xyz file to facilitate the examination of the conformation by VMD. In this file, the atom types and the marks are combined in way to make visualization of different regions easier. This file is also used by the ref_gen_step1_cord script.

We generally refer these types of scripts as master scripts. The master scripts call other AFM tools scripts to generate QM/MM conformations for each configuration. The best approach to create a master script for a new system is to run the AFMTools scripts on a sample conformation to select the desired QM/MM region. After the sequence of scripts are tested, then the commands can then be put in the master script. We designed the AFM Tools scripts such that each step of the QM/MM selection procedure described in our publications can be accomplished in one line with an AFMTools script.

An example of the QM/MM configurations generated is shown as Figure 6. All the generated configurations can be found in the directory Tutorial/01_QM_MM/298/conf in the tutorial package.

After executing the “01_runme.sh” command 101 .pxyz files should have been created with names final000.pxyz to final100.pxyz.

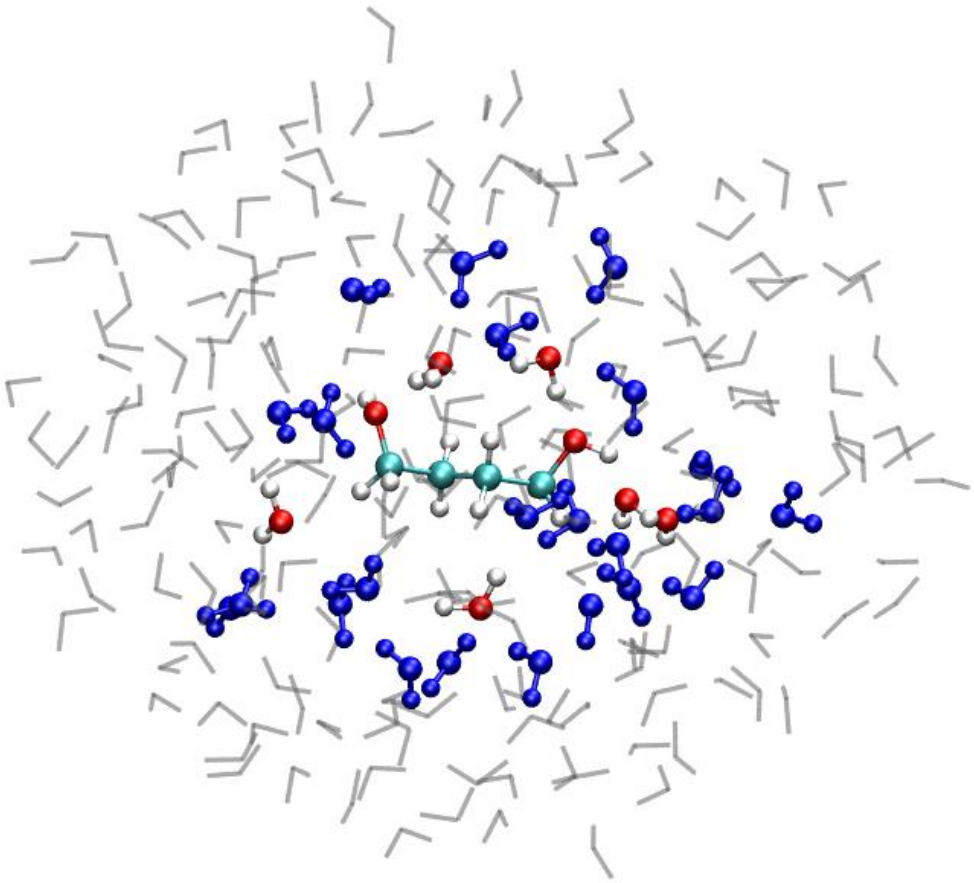

Figure 6: QM/MM configuration for 1,4-butanediol in water with the QM region in CPK and the MM region as lines. The blue waters are in the buffer region, shielding the QM region from the MM charges.

4.1 Running QM/MM calculations to get Gradients.

In this tutorial, ORCA will be used to perform the QM/MM calculations. We assume your system already has ORCA installed. The super computer cluster at University of Arkansas is named pinnacle. The job submission was accomplished with the runme.pinnacle script. You may need to write a similar script to run these jobs on your system.

The script that generates input files for ORCA is 01_orca_inp_master.pl

```
 1 #!/usr/bin/perl
 2 # the inputs are the pxyz files generated from 01_runme.sh
 3 foreach $file (@ARGV)
 4 {
 5 $basename=substr $file,0,-5;
 6 @base_path=split(/\//, $basename);
 7 $orca_input=@base_path[@base_path-1].".orca.inp";
 8 $dirname=@base_path[@base_path-1];
 9
10 ### Create QM Region Orca File
11 $cmd1="pxyz_select_n $file 4 3 2 | grep -v 'MW' | xyz_add_linenu | pxyz_orca_upd_xyz MyMol.templ
   .inp nam_translation > hold";
12 ### Update Charges in Orca File
13 $cmd2="pxyz_select 1 $file | grep -v 'OW' | xyz_add_linenu | pxyz_orca_upd_chg hold chginfo > $o
   rca_input && rm hold";
14 $cmd3="mkdir $dirname";
15 $cmd4="mv $orca_input $dirname";
16 $cmd5="cp -a runme.pinnacle $dirname";
17 system("$cmd1");
18 system("$cmd2");
19 system("$cmd3");
20 system("$cmd4");
21 system("$cmd5");
22 }
```

Figure 7: Script "01_orca_inp_master.pl" as can be found in the directory "Tutorial/01_QM_MM/298/orca_inp"

We assume a set of OCRA input files has been created and tested using some sample conformation before using the AFM scripts.

The AFM scripts are designed to update an existing input file of a QM package (such as ORCA) using information from .pxyz file and won't write an QM input file from scratch. In this example, the update of the ORCA input file is accomplished with lines 11 and 13 of the 01_orca_inp_master.pl.

Line 11, select atoms with marks 4, 3, and 2, which is the QM region of QM/MM and creates a new input file using the MyMol.templ.inp file as template. The new input file retained everything in the MyMol.templ.inp file except update the configuration for the QM region. The temporary input file is named hold. The script that update the coordinate of an ORCA template input file is pxyz_orca_upd_xyz.

The nam_translation file translates atom types in the pxyz file to the atom types for ORCA.

Line 13, selects atoms with mark 1 and uses the pxyz_orca_upd_chg script to update the partial charges portion of the temporary ORCA input file (hold) and write the output to the file defined by the variable $orca_input file.

The input files and a runme.pinnacle file are placed in a new directory ($dirname) for each QM/MM calculation.

Since the BLYPSP-4F model in the simulation is a 4-site model, in line 11, one can see the MW lines in the input are removed when listing the QM atoms. Line 13 also removed the OW line as there is no charge on the OW site.

The 01_orca_inp_master.pl script can take pxyz files as input and can be executed with the command

./01_orca_inp_master.pl ../conf/*.pxyz

in Tutorial/01_QM_MM/298/orca_inp directory.

The “02_run_orca.sh” script in the tutorial files takes the directory names where the ORCA simulations is to be performed as arguments, for example “./02_run_orca.sh final{000..100}”.

4.2 Produce the ref file for CRYOFF

After the ORCA calculations, the ref file will be created for the AFM fitting using the gradients computed by ORCA.

A 03_Copy_Engrad.sh script can be provided in the conf directory to copy the output gradient files from OCRA to a separate directory for creating the .ref files.

In the Tutorial/01_QM_MM/298/genref directory, the coordinates of each configuration are in the .xyz files copied from the selection from the Tutorial/01_QM_MM/298/conf directory. The 01_gen_ref.pl (Figure 8) script will create the .ref files and place them in the ref_files directory.

```
 1 #!/usr/bin/perl
 2 # the input files are *.xyz files
 3 foreach $file (@ARGV)
 4 {
 5 $basename=substr $file,0,-4;
 6 @base_path=split(/\//, $basename);
 7 $name=@base_path[@base_path-1];
 8 $gradname=$name.".orca.engrad";
 9 $cmd1="chunk 2 9999 $file | grep -v ' MW' | ref_gen_step1_cord molinfo | ref_upd_orca_grad $grad
   name | ref_upd_net | xyz_add_msite Ow Hw Hw M | ref_fix_msite M | xyz_add_msite Omm Hmm Hmm Mmm
   | ref_fix_msite Mmm | xyz_fix_linenu > ref_files/$name.ref";
10 system("$cmd1");
11 }
```

Figure 8: Script "01_gen_ref.pl" as can be found in Tutorial/01_QM_MM/298/genref

The first step in generating a ref file is to run ref_gen_step1_cord (Line 9, Figure 8) using a molinfo file (Figure 9).

An in-depth discussion of the molinfo file can be found in the AFMTools manual. In short, each molinfo file contains a header, which is followed by the information for each molecule type. The information for each molecule type includes a line with the number of atoms, the molecule name, and the solvation factor, which is followed by one line for each atom of the molecule. The number in the atom line is the weight for computing the center for computing the torque of the molecule. If the weight is 1.0 as in the example, the centroid location will be put in the NetF and Torq field in the output pxyz file to compute molecular torque relative to this location. The “next” keyword indicates the end of the definition of this molecule.

In the example, (line 9 of Figure 8) the ref_gen_step1_cord takes its coordinates from standard input, the chunk 2 9999 removes the first line from the file defined by the variable $file. The Msite information in the file is removed by the grep command

The output of the ref_gen_step1_cord is a ref file without any gradient information. The gradient information is added by the ref_upd_orca_grad scripts taking them from the file defined by $gradname. Note that it is the user's responsibility to make sure there is a one-to-one correspondence between the gradient in the $gradname file and the corresponding atom in the ref file generated by ref_gen_step1_cord. The ref_upd_orca_grad file does not check if the atom names match.

Also, the ref_upd scripts replace the gradient information in the ref file by adding the gradients in the gradient file to those in the ref file. The ref_gen_step1_cord, will write the ref file with the gradients of all atoms being zero.

Also on line 9, the ref_upd_net script computes the net force and net torques, the xyz_add_msite adds the msite back into the .ref file. The ref_fix_msite script puts the correct solvation factor and molecular name information on the msite.

The reason that the Msite information has to be first removed and then added back in is that the gradient computed by ORCA does not have the M site. One has to first remove the Msite before the gradient information is updated to ensure the atom order in the ref file and in the gradient file match.

```
 1 BUU in water 298 K
 2 16     BUUqm     1.0
 3 C1            1.0
 4 H2            1.0
 5 H2            1.0
 6 C2            1.0
 7 H3            1.0
 8 H3            1.0
 9 C2            1.0
10 H3            1.0
11 H3            1.0
12 C1            1.0
13 H2            1.0
14 H2            1.0
15 O1            1.0
16 H1            1.0
17 O1            1.0
18 H1            1.0
19 next       OW3
20 3      H2Oqm    1.0
21 Ow            1.0
22 Hw            1.0
23 Hw            1.0
24 next        OW2
25 3      H2Oqm    0.0
26 Ow            1.0
27 Hw            1.0
28 Hw            1.0
29 next        OW1
30 3       H2Omm      0.0
31 Omm           0.0
32 Hmm           0.0
33 Hmm           0.0
34 next
```

Figure 9: The "molinfo" file used in the creation of the reference force (.ref) file

Each .ref file contains the reference forces of each frame that will be used to fit your desired model. An illustrative sample of a .ref file can be seen in Figure 10. NetF and Torq lines are the net force and net torque of each molecule. The first line specifies the number of atoms (counting NetF and Torq lines), and the second line is a comment line.

After these two lines, each line will contain the atom name, the location (x, y, z, in angstroms) of the atom, the forces (x-force, y-force, z-force, in kcal/(mol·Å)), the solvation factor, and the molecule name.

These files will be concatenated together to create the final .ref file (total.ref) for the fit.

```
932
BUU in water 298 K
C1      9.19000      11.79000      11.56000      27.9679250432935      5.02354514287434      -3.1040328925688          1.0      1BUUqm
H2      8.76000      10.82000      11.78000      -10.9218033971737      1.16084909526253      10.5397779987498          1.0      1BUUqm
H2      8.51000      12.45000      11.04000      -6.64828580591228      -5.68456807657981      -5.8062098636739          1.0      1BUUqm
C2      10.47000      11.53000      10.70000      -38.2102427892304      -0.385830975820612      -36.8892694365289          1.0      1BUUqm
H3      10.33000      12.05000      9.68000      12.2807267284405      -9.00730866781931      39.9115128799998          1.0      1BUUqm
H3      11.24000      11.94000      11.30000      25.1112406271545      15.4872754122114      -3.027639317246          1.0      1BUUqm
C2      10.84000      10.05000      10.40000      18.8164714000793      20.695619175228      18.3370609204001          1.0      1BUUqm
H3      10.03000      9.51000      9.94000      -7.87207579975789      -1.15273968334467      -11.0817842817267          1.0      1BUUqm
H3      10.98000      9.56000      11.39000      6.60439242083744      0.798842062241502      -15.3674527968264          1.0      1BUUqm
C1      12.16000      9.96000      9.52000      -40.658325390753      -10.8484692699382      -6.3195032306508          1.0      1BUUqm
H2      12.31000      10.86000      8.86000      0.74783853771785      -13.3282103552185      25.5959634229106          1.0      1BUUqm
H2      11.98000      9.15000      8.78000      14.9771375329727      3.72669699438063      15.1804141594377          1.0      1BUUqm
O1      13.32000      9.76000      10.34000      8.50317101489419      6.87674596175928      -7.56436988187044          1.0      1BUUqm
H1      13.34000      8.89000      10.78000      -7.76590209790349      2.74562727649702      -11.1711866358897          1.0      1BUUqm
O1      9.56000      12.52000      12.80000      1.5877118386051      -26.7147432579822      -25.4913573682516          1.0      1BUUqm
H1      8.76000      12.96000      13.07000      -10.9894398449341      10.8197230596391      7.61652058158583          1.0      1BUUqm
NetF      10.7362500      10.8625000      10.7462500      -6.4694600      0.2130539      -8.6415557      1.0      1BUUqm
Torq      10.7362500      10.8625000      10.7462500      -3.9737883      -2.2452537      5.1741773      1.0      1BUUqm
Ow      13.85000      12.10000      11.68000      -14.1067698013111      -45.2283035603648      17.4479476445738          1.0      2H2Oqm
Hw      14.06000      12.69000      10.99000      13.2590919476955      29.6202132744314      -31.0182288122094          1.0      2H2Oqm
Hw      13.65000      11.22000      11.24000      4.4286336743913      17.5301190573208      7.77809261156508          1.0      2H2Oqm
M      13.8520000      12.0420000      11.4540000      0      0      0      1.0      2H2Oqm
NetF      13.8533333      12.0033333      11.3033333      3.5809558      1.9220288      -5.7921886      1.0      2H2Oqm
Torq      13.8533333      12.0033333      11.3033333      1.7218471      -1.6984060      -1.5640115      1.0      2H2Oqm
Omm      4.73000      4.86000      5.38000      0      0      0      0.0      32H2Omm
Hmm      4.59000      5.16000      4.42000      0      0      0      0.0      32H2Omm
Hmm      4.20000      5.50000      5.81000      0      0      0      0.0      32H2Omm
Mmm      4.5960000      5.0480000      5.2740000      0      0      0      0.0      32H2Omm
Omm      8.35000      5.86000      2.49000      0      0      0      0.0      33H2Omm
Hmm      8.81000      6.48000      3.01000      0      0      0      0.0      33H2Omm
Hmm      8.98000      5.36000      1.89000      0      0      0      0.0      33H2Omm
Mmm      8.5680000      5.8840000      2.4740000      0      0      0      0.0      33H2Omm
Omm      5.28000      10.61000      7.54000      0      0      0      0.0      34H2Omm
Hmm      6.03000      11.07000      7.11000      0      0      0      0.0      34H2Omm
Hmm      5.59000      10.02000      8.29000      0      0      0      0.0      34H2Omm
Mmm      5.4920000      10.5840000      7.6040000      0      0      0      0.0      34H2Omm
Omm      1.06000      9.06000      9.55000      0      0      0      0.0      35H2Omm
Hmm      0.61000      8.24000      9.67000      0      0      0      0.0      35H2Omm
Hmm      0.58000      9.74000      10.03000      0      0      0      0.0      35H2Omm
Mmm      0.8740000      9.0320000      9.6700000      0      0      0      0.0      35H2Omm
```

Figure 10: First few lines of a .ref file. The second to last column is the solvation factor. The CRYOFF code only fit lines with solvation factor greater than 0. However, coordinates of all the atoms must be in the .ref file for proper calculation of gradients on solvation factor 1 atoms by CRYOFF. NetF and Torq are the net force and torque of each molecule. They are not included for MM molecules in this example.

To execute this script, first move the .xyz files from the conf/ directory to the genref/ directory, and then type the command

./01_gen_ref.pl *.xyz

This command will populate the directory “ref_files” with individual ref files for each configuration.

From within the directory “ref_files”, to create the final .ref file, just run the command

cat final* > total.ref

The total.ref file needs to be copied to the directories “02_CRYOFF/inter” and “02_CRYOFF/intra” for the fitting.

5. Performing the CRYOFF Fitting

Once the reference forces have been produced and concatenated to create one large .ref file, the CRYOFF fitting must be performed to generate the parameters for the next generation of AFM. This is accomplished through the use of two files: inter.ff and intra.ff, which direct the fitting of intermolecular and intramolecular parameters, respectively. We will use a two-step fit procedure and fit the intermolecular forces before fitting the intramolecular forces. Several scripts have been designed to aid

the process in production of the intra.ff file using the output parameters after performing the intermolecular fit.

It is worth noting that the dispersion parameters were already fitted before AFM iterations (In Section 3 and 8). This section only covers the fitting of other parameters during an AFM iteration.

5.1 Intermolecular fit.

In the tutorial, all the files related to the intermolecular fit are in the Tutorial/02_CRYOFF/inter directory.

The fit can performed by the command (assuming cry.300.x is in your PATH)

cry.300.x inter.ff

or in parallel over 4 CPUs.

mpirun -np 4 cry.300.x inter.ff

The precompiled version uses the intel MPI.

The inter.ff file can be found in this directory. Please refer to the CRYOFF manual for a detailed discussion of this file.

```
1 [file]  total.ref  inter.off
2 [optimiz_control]  simplex  conv=1E-8 maxit=2000
3 [keywords]  Inter  norm=w
```

Figure 11: First three lines of the inter.ff file as found in directory "Tutorial/02_CRYOFF/inter"

The first three lines of the inter.ff file are shown in Figure 11. In the first line, the [file] section specifies the name of the .ref file and the name of the output file. The CRYOFF code will not overwrite the output file. Thus, it will quit with an error if the output file already exists.

The [opt] section (line 2) sets parameters related to the optimization. In the example, the word “simplex” will trigger the simplex algorithm to do a non-linear optimization. Note that if there are no nonlinear parameters defined when the force field terms are specified later in the file, a linear optimization will be performed regardless of whether non-linear optimization is requested in the opt section.

The [keywords] section requested an intermolecular fit (Inter) and the default weight (norm=w) for each force questions. The default weight is the optimal in most cases. When the initial guess force field is very poor, the REL weight may work better.

After the first 3 lines, the molecular specification and intermolecular terms follow. For detailed information regarding these sections, we will defer to the CRYOFF manual.

The molecular definition has an exclusion list [Exc]. The CRYOFF code will compute all non-bonded pairs within the molecule unless it is excluded in this list explicitly. While these intramolecular non-bonded

interactions won't affect net force and torque and does not affect intermolecular fit, it is important to make sure this list is correct for intramolecular fit.

```
 97      [Cou]   31
 98    Hw    Hw    FixT    0.4415602500    1
 99    Hw    M    FixT    -0.8831205000    2
100    M    M    FixT    1.7662410000    3
101    Hw    Hmm    FixT    0.4415602500    4
102    Hw    Mmm    FixT    -0.8831205000    5
103    M    Hmm    FixT    -0.8831205000    6
104    M    Mmm    FixT    1.7662410000    7
105    H1    Hw    Fit    0.3087503200    8
106    H1    M    Fit    -0.6171023100    9
107    H1    Hmm    Fit    0.2592798700    10
108    H1    Mmm    Fit    -0.5183988100    11
109    H2    Hw    Fit    -0.0671224030    12
110    H2    M    Fit    0.1341522000    13
111    H2    Hmm    Fit    -0.0793730110    14
112    H2    Mmm    Fit    0.1587458800    15
113    H3    Hw    Fit    0.0707187850    16
114    H3    M    Fit    -0.1415676100    17
115    H3    Hmm    Fit    0.0345317850    18
116    H3    Mmm    Fit    -0.0692217950    19
117    C1    Hw    Fit    0.4592152700    20
118    C1    M    Fit    -0.9185030300    21
119    C1    Hmm    Fit    0.4532994400    22
120    C1    Mmm    Fit    -0.9065761600    23
121    C2    Hw    Fit    -0.1968024800    24
122    C2    M    Fit    0.3935362800    25
123    C2    Hmm    Fit    -0.1113023700    26
124    C2    Mmm    Fit    0.2225538100    27
125    O1    Hw    Fit    -0.5784369400    28
126    O1    M    Fit    1.1569173000    29
127    O1    Hmm    Fit    -0.5116249500    30
128    O1    Mmm    Fit    1.0233621000    31
```

Figure 12: [Cou] section of inter.ff file as found in "Tutorial/02_CRYOFF/inter"

Since CRYOFF only fits charge products instead of having atomic partial charges as parameters, all charge products are specified under the [Cou] section (Figure 12). Note that some charge products such as those between solvation factor 0 water and MM water are ignored as those won't affect the fit of the solvation factor 1 molecules.

```
154 [Cstr]   18
155 6: 1  8    2   12    2   16   1  20   1 24  1 28  0.0  1E3
156 6: 1  9    2   13    2   17   1  21   1 25  1 29  0.0  1E3
157 6: 1 10    2   14    2   18   1  22   1 26  1 30  0.0  1E3
158 6: 1 11    2   15    2   19   1  23   1 27  1 31  0.0  1E3
159 2: 2  8    1   9                                0.0  1E3
160 2: 2  10   1  11                                0.0  1E3
161 2: 2  12   1  13                                0.0  1E3
162 2: 2  14   1  15                                0.0  1E3
163 2: 2  16   1  17                                0.0  1E3
164 2: 2  18   1  19                                0.0  1E3
165 2: 2  20   1  21                                0.0  1E3
166 2: 2  22   1  23                                0.0  1E3
167 2: 2  24   1  25                                0.0  1E3
```

Figure 13: [Cstr]/Charge constraint section of the inter.ff file as it is found in "Tutorial/02_CRYOFF/inter"

The neutrality of the molecules is enforced in the [CSTR] (Figure 13) section. The weight of the constraint is 10^3. Note that the weight is squared in the charge restraint equations. Thus, it is not advisable to use very large weights. On the other hand, the parameters are relatively insensitive to the choice of the weight. Our tests show that any weight between 10^6 to 10^2 should be fine. Please see the CRYOFF manual for more discussion.

5.2 Intramolecular Fit

The intramolecular fit directory is in Tutorial/02_CRYOFF/intra. The intramolecular fit is performed in a similar fashion to the intramolecular fit, with the command

cry.300.x intra.ff

However, the intra.ff file must be updated to include parameters obtained from the intermolecular fit. This can be accomplished with the off2ff scripts.

5.2.1 The Intramolecular Template File

The off2ff scripts are designed to modify/update an existing intra_template.ff file with intermolecular parameters obtained in the first step (Sec. 5.1). The easiest approach in practice is to write an intra.ff file with sections to be auto populated with the off2ff scripts. In the example, an intra_template.ff file is provided for this purpose.

```
224 [EXPINTER]  0
225
226 [EXPINTRA] 9
227    H1  H2    Fit      267412  3.6
228    H1  C2    Fit      267412  3.6
229    H1  O1    Fit      267412  3.6
230    H2  H3    Fit      267412  3.6
231    H2  C1    Fit      267412  3.6
232    H3  O1    Fit      267412  3.6
233 ;   C1  C1    Fit      267412  3.6
234    C1  O1    Fit      267412  3.6
235    C2  O1    Fit      267412  3.6
236    O1  O1    Fit      267412  3.6
```

Figure 14: Portion of the intra_template.ff that shows the EXP sections.

Figure 14 shows a section of the intra_template.ff file where the EXPINTER section is empty and EXPINTRA has some initial values. The off2ff script can be used to either populate a section (e.g. EXPINTER) or update a section (EXPINTRA). Note that the CRYOFF program will treat both sections as EXP sections. The AFMTools will recognize these as two different sections to allow different actions to be taken for each section.

5.2.2 protocol files.

The off2ff script follows the protocol specified in the protocol files (proto)to update the template intra.ff file to create the final intra.ff for fitting.

off2ff proto.copy inter.off intra_template.ff intra.ff

where the parameter will be taken from the inter.off file following the protocols defined in the proto.copy file using the intra_template.ff file. The intra.ff file is the file that will be created for the intramolecular fit step and is provided in the tutorial directory.

The proto.copy file (Figure 15) is provided with the tutorial

```
1 copy COU,col4,ln1-31        COU,col4,ln1-31
2 copy EXP                    EXPinter,fix
3 charge  COU,atm1,0.66450,ln8,ln12,ln16,ln20,ln24,ln28            COU,col4,ln32-49
```

Figure 15: Procotol file "proto.copy" as found in "Tutorial/02_CRYOFF/inter"

The first line instructs the off2ff script to copy the parameters of COU section from the .off file to the .ff file. It states the fourth column of line 1 through 31 in the .off will be copied to the COU section, where the parameter is in the fourth column and the line number is also from 1 through 31.

Line 2 in Figure 15 states that the EXP section should be copied in whole to the [EXPINTER] section of the intra_template.ff file, and that those parameters should be fixed.

Note that the [COU] section in the intra.ff contains charge products between pairs of solute atoms. These solute-solute intramolecular Coulombic pairs do not have to be included in the intermolecular fit

step as they do not affect the net force and net torque. However, these are important for the intramolecular fit. Line 3 of the proto.copy file will make sure all the charge products are properly computed using the information from the inter.ff file.

For details of the off2ff script and the protocol files, please read the AFM manual.

6. Next Generation Input Files

After the fitting has been performed using CRYOFF, the next iteration of AFM starts from the MD step. We provide the off2top script to automatically update a template Gromacs topology file using the AFM parameters.

Note that due to the limited support for the Buckingham type potential in Gromacs, tabulated potentials are used to simulate the final force fields. Unfortunately, Gromacs has dropped support for tabulated potentials in newer releases. Only gromacs 2019.6 and earlier can be used.

6.1 off2top

A detailed explanation of the off2top script can be seen in the AFMTools manual. Essentially, the off2top script updates the .top or .itp files used by GROMACS with parameters from the intra.off file obtained following the intramolecular fit step in Section 5. All of the files necessary for producing the topology files and tabulated potentials are located in the directory “Tutorial/02_CRYOFF/Tabpot_Topology”. To generate all required GROMACS input files, one can execute the runme.sh script in this directory.

```
1 #!/bin/bash
2 offget_charge COU,atm1,0.6645,ln8,ln12,ln16,ln20,ln24,ln28 intra.off  | grep COU | awk '{print $1, $4}' >
   temp_chgfile
3 adjust_charge temp_chgfile QQequations > chgfile && rm temp_chgfile
4 off2top protocol.mol intra.off template_BUU.itp BUU.itp
5 off2top protocol.nonbonded_list intra.off template_nonbonded_BUU_water.itp temp_nonbonded.itp
6 off2top protocol.nonbonded.para intra.off temp_nonbonded.itp nonbonded.itp && rm temp_nonbonded.itp
7
8 off2tab protocol.tab intra.off nonbonded.itp && cp -a BUU_OW_OW.xvg BUU.xvg
9 cp -a BLYPSP-4F_b0.xvg BUU_b0.xvg
```

Figure 16: "runme.sh" script as found in "Tutorial/02_CRYOFF/Tabpot_Topology/"

The commands related to production of the .top/.itp files are in lines 2-6. For example, Line 4 writes a new itp file, BUU.itp, using template_BUU as example. The output file, BUU.itp will take parameters from the intra.off file by following the protocol spelt out by the protocol.mol file.

The partial charges produced by CRYOFF may not make 1,4 butanediol exactly neutral. This is fixed in Line 3 with the adjust_charge script. The adjust_charge script reads the QQequations file.

```
1 4 H2 + 4 H3 + 2 H1 + 2 O1 + 2 C1 + 2 C2 = 0
2 2 C1
```

Figure 17: QQequations file as found in "Tutorial/02_CRYOFF/Tabpot_Topology"

Line 1 states that the sum of charges of 4 H2, 4 H3, etc, as seen on the left side of the equation is zero.

Within the runme.sh script, a “chgfile” will be produced containing the final charges to be used in the next generation of gromacs simulations.

Below is the “protocol.mol” file, the first column specifies what to be updated, the section column specifies the source of the parameters, and the third column is the location of parameters in the top or itp file. The 4th column provides a unit conversion between .off and GROMACS units, while the last column provides some fine control. The details of the file formats can be found in the AFM manual.

```
 1 ######molecule definition#####################
 2 #                      chgpam/offpam                                                    toppam
          unitcov       control
 3 mol.charge file=chgfile BUU,atoms,col7,ln1-16   1   default
 4
 5 #                offpam                    toppam                          unitcov              #control
 6 mol.bonded         BUUQM,col4,ln1-6        BUU,bonds,col4,ln1-15             1.0               type,molinfo
 7 mol.bonded         BUUQM,col5,ln1-6        BUU,bonds,col5,ln1-15             1.0               type,molinfo
 8 #
 9 mol.bonded         BUUQM,col4,ln7-15       BUU,angles,col5,ln1-26            1.0               type,molinfo
10 mol.bonded         BUUQM,col5,ln7-15       BUU,angles,col6,ln1-26            1.0               type,molinfo
11 ##
12 mol.bonded         BUUQM,col4,ln16-18      BUU,dihedrals,col6,ln1-5          1.0               type,molinfo
13 mol.bonded         BUUQM,col5,ln16-18      BUU,dihedrals,col7,ln1-5          1.0               type,molinfo
14 mol.bonded         BUUQM,col6,ln16-18      BUU,dihedrals,col8,ln1-5          1.0               type,molinfo
```

Figure 18: protocol.mol file as found in "Tutorial/02_CRYOFF/Tabpot_Topology"

In the runme.sh script, the short-range repulsion and van der Waals interactions are generated with Line 8 (Fig. 16) . In GROMACS, the tabulated potentials are scaled by the “C12” and “C6” parameters, thus there are more than one way to model the same short-range potentials with different choices of this scaling factor. If long range correction to energy and stress were to be used during Gromacs simulation, it is important that the C6 in the topology file is correct.

The C6 parameters are copied from the intra.off file in Line 6 (Figure 16) and the corresponding tabulated potential are created in Line 8.

```
1 #######parameter file#############################
2 #                          grid                        #prefix
3 tab.nonbond          5.0,0.0005,ln1-20                 scale=C6,type=atomtype,prefix=BUU
4
```

Figure 19: protocol.tab file as found in "Tutorial/02_CRYOFF/Tabpot_Topology"

Note that the scale=C6 in the protocol.tab file instruct the scripts to use the C6 in the topology file.

Line 9 of the runme.sh script copies the intramolecular tabulated bond potential to a new name expected by Gromacs during execution of mdrun.

6.2 GROMACS Simulation

To perform the next generation of GROMACS simulations, copy all the necessary files from the Tabpot_Topology directory into the directory designated for the MD step of the next generation:

cp -a BUU* nonbonded.itp ../../03_Next_Gen/298

From there you may run GROMACS, and perform the sampling step, QM/MM step, and fitting step all over again.

7. The Global Fit

Some AFM parameters may fluctuate from generation to generation. Thus, some less stable parameters may not seem to converge even after several generations. One way to judge convergence is to monitor selected radial distribution functions to ensure the simulated structures no longer change. Once convergence has been reached, the global fit may be performed to generate the final model. This fit is typically done by using the .ref files generated by the most recent few generations. 4 generations were used when we developed the 1,4 butanediol potential. The .ref file for the global fit can be created by concatenating the .ref file of all the converged generations. From here, inter and intra CRYOFF fitting is performed to obtain the final model.

8. Fitting of D3 dispersion

In AFM, dispersion parameters are generally fit using SAPT or D3 calculations before the AFM iterations. We will fit D3(BJ) dispersion for the 1,4-butanediol. The D3(BJ) parameters are chosen to be consistent with the B3LYP exchange-correlation functional in the QM/MM step.

We prefer to fit the dispersion using dimer conformations from an MD simulation. While fitting to SAPT will be done using SAPT energies with the FITE keyword of CRYOFF, the fitting of D3(BJ) dispersion can be done by fitting forces due to the availability of D3 gradients.

In the case of 1,4-butanediol, dimers were selected within a certain range of distances. We generally do not include dimers with nearest atom distances shorter than 5 Å to avoid excessive contributions from higher order terms. The C6 (or sometimes C6 and C8) fit in this step is designed to capture the long-range asymptotic behavior of dispersion.

8.1 Fitting of Intramolecular Dispersion

Intramolecular dispersion parameters were actually fitted as intermolecular parameters between two 1,4-butanediol molecules. Assuming the dimer geometries are in an xyz file, the AFM tools provide the "getd3force" script, which calls the dftd3 program of Grimme. (not included in the tutorial but can be obtained for free from Stefan Grimme's website)

The getd3force script writes the gradient in a .d3grad file.

The ref file can be created in a similar fashion of other ref files using the .xyz files and .d3grad files. In the case of D3, the forces are added to the .ref file with the "ref_upd_d3_grad" script.

```
 1 [MolTyp]    BUT  4
 2   [ atoms ]  18
 3  1        C1                C1
 4  2        H2                H2
 5  3        H2                H2
 6  4        C2                C2
 7  5        H3                H3
 8  6        H3                H3
 9  7        C2                C2
10  8        H3                H3
11  9        H3                H3
12  10       C1                C1
13  11       H2                H2
14  12       H2                H2
15  13       O1                O1
16  14       H1                H1
17  15       O1                O1
18  16       H1                H1
19  17       NetF             NetF
20  18       Torq             Torq
21
22
23   [ Bonds ]    0
24   [ Angles]    0
25   [Dihedral]   0
26   [ Exc ]     41
27     2     1  0  0
28     3     1  0  0
29     4     1  0  0
30     5     4  0  0
31     6     4  0  0
32     7     4  0  0
33     8     7  0  0
34     9     7  0  0
35    10     7  0  0
36 ...
37 all exclusions included
38 ...
39
40 [ SRD ]    6
41   O1   O1    fit 1    -6     1.94949  7407420
42   O1   C1    fit 1    -6     1.99700  7407420
43   O1   C2    fit 1    -6     1.99700  7407420
44   C1   C1    fit 1    -6     2.04474  7407420
45   C1   C2    fit 1    -6     2.04474  7407420
46   C2   C2    fit 1    -6     2.04474  7407420
```

Figure 20: Sample input .ff file for fitting intramolecular dispersion parameters

In this tutorial, the completed .ref file for the intramolecular dispersion fit has been provided as "intra_dimers.ref". The .ff file for the intramolecular fit is provided as intra_D3.ff (Figure 20) A sample output file is provided as intra_D3.off

The only terms being fit in this case will be dispersion. In the example, we used the short-range-damped form[3] with the SRD keyword.

8.2 Fitting Intermolecular Dispersion

In dispersion terms between 1,4-butanediol and water is fitted similarly as the 1,4-butanediol intramolecular terms, except the dimer involves a 1,4-butanediol and a water. The fitted parameters have been provided previously in Table 1. The files for this fit will be omitted.

References:

1. D. Zheng and F. Wang, ACS Physical Chemistry Au **1** (1), 14-24 (2021).
2. S. Grimme, S. Ehrlich and L. Goerigk, J. Comput. Chem. **32** (7), 1456-1465 (2011).
3. Y. Yuan, Z. Ma and F. Wang, J. Phys. Chem. B **125** (6), 1568-1581 (2021).