force field parametrization

INSTALLATION AND USER GUIDE

Version 3.0

Contents

1 General purposes

The energy and its first and second geometrical derivatives obtained by Quantum Mechanical (QM) calculations for a number of conformations of a single molecule in its ground- or excited state are used to specifically parameterize the intramolecular term of a QM derived force-field (**QMD-FF**), [1–4] suitable for computer simulations based on classical physics as MC or MD. [5, 6] As most general-purpose FFs, [7–11] a QMD-FF is built up by essentially three ingredients, namely:

i) A set of selected generalized (or redundant) internal coordinates (RICs), such as bond lengths, angles, dihedrals, or nonbonded distances, that completely define the molecular geometry.

ii) A set of model potential functions associated with each RIC.

iii) A set of parameters (force constants and RIC's equilibrium values) which complete the definition of the model functions, settling molecular chemical specificity onto the FF functional.

The Joyce program reads a starting topology file in which all selected RICs and the associated model functions that define the intramolecular potential are specified. This file can be automatically generated by the Joyce code (see for instance Section 3.4), created by the user through the scripts distributed at the Joyce website, [12] or built using popular webservers. [13–16] The supported format for this input topology file is the same as used in the Gromacs [17] package, usually referenced as ***.top*** (see Section 5.3 for further details). The third ingredient, which consists in the final **QMD-FF** parameters, is created from the database purposely calculated at the QM level, **specifically** for the chosen molecular target molecule **T**. As discussed in more detail in Sections 3.3 and 5.2, such a QM database is read by Joyce from external files, which contain the calculated QM data. The supported formats are: *i)* a formatted (*.fcc*) file, compatible, with the $\mathcal{FC}classes3$ code [18], and *ii)* a formatted checkpoint file (*.fchk*) produced by the Gaussian16 package [19] (and previous releases). All QMD-FF parameters are retrieved based on this QM database calculated at first principle level. That is, the FF equilibrium values of all selected RICs are automatically extracted by Joyce from the optimized QM geometry, while the force constants are calculated through a linear fitting procedure, [1, 20] as detailed in Section 7.

The JOYCE code is developed by Giacomo Prampolini, [1] Javier Cerezo, [2] Samuele Giannini, [3] Ivo Cacelli and Nicola De Mitri. The current version of the code is available within the JOYCE3.0 package, maintained and distributed at the JOYCE website [12] by Giacomo Prampolini, Javier Cerezo, Samuele Giannini, J. Pablo Galvez, Pablo M. Martinez, Daniele Padula, Anna Piras, Abderrahmane Semmeq, and J.-Guillherme Vilhena. The whole package is open-software: it can be redistributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation (version 3). This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program. If not, see http://www.gnu.org/licenses/.

For further information type, once installed:

> *go.joyce -lic*

Please visit the JOYCE3.0 website [12] for further tutorials, templates, and downloads.

[1] giacomo.prampolini@cnr.it

[2] javier.cerezo@uam.es

[3] samuele.giannini@cnr.it

2 Installation

The Joyce package runs on any Linux platform. The *go.joyce* script, which is devised for easily handling the program, is intended to work under *tcsh* shell. Notwithstanding *bash* users can find in the source directory an equivalent *go.joyce.sh* script, in the following all examples will be referred to the original *go.joyce*.

The only requirement for Joyce installation is a Fortran compiler, such as gfortran, ifort or any other. Please note that the currently implemented compilers are:

gfortran – free GNU compiler (default)

ifort – Intel compiler, available for academic use

pgf90 – Portland compiler

To install the program you should follow the next steps:

Memo: check the correct path and version when preparing the tarball

1. Download the tarball *Joyce.v3.0*

2. Unzip and untar the Joycepackage *Joyce.v3.0* with the command:
   > *tar -xzvf Joyce.v3.0*

3. Set the environmental variable *JOYCE* to where the program was unpacked, *e.g.* if the program was unpacked in the user home directly write (in tcsh shell):
   > *setenv JOYCE /home/username/Joyce.v3.0*
   Alternatively, you can add the aforementioned variable definition in your login script.

4. Copy the Joyce main script executable (*go.joyce* in your bin directory (/home/$USER/bin) or in any other directory contained in your PATH:
   > *cp $JOYCE/bin/go.joyce /home/$USER/bin/.*

5. Compile the program with the command
   > *go.joyce -gfortran*
   to use gfortran as compiler. Just type
   > *go.joyce*
   to see other compiling options.

3 Overview

3.1 General scheme

A chart of the JOYCE worflow is reported in Figure 1. By looking at the top left corner, we see that the only *a priori* information required is the knowledge of the target molecule **T**. Starting from **T**'s chemical structure,

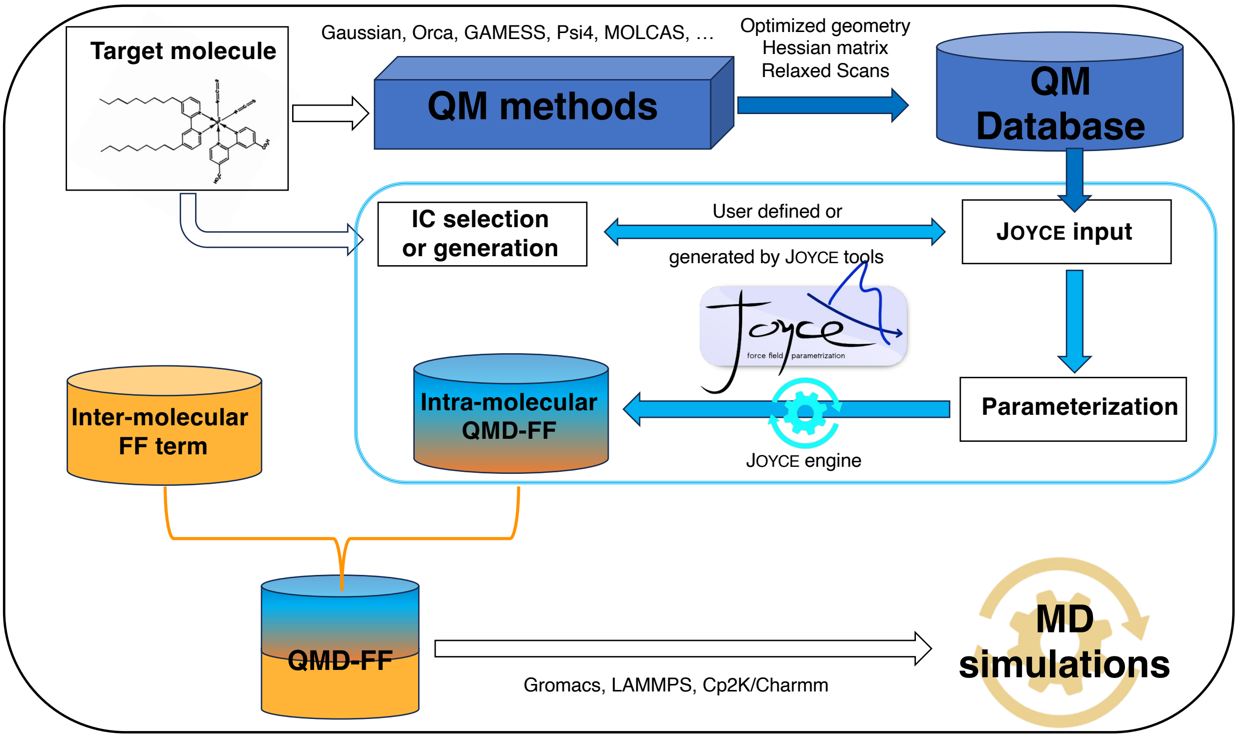

Figure 1: Flow chart of the JOYCE3.0 program.

the JOYCE code is devised to simultaneously derive all parameters of a specific intramolecular FF term (see Section 7) from QM data purposely calculated on the target molecule. Three kinds of information are required by JOYCE to carry out the parameterization, each stored in a separate file. The primary input file contains all the main instructions and specifics to run JOYCE, and is discussed in some detail in Section 5.1. Prior to parameterization (see top panels of Figure 1), the structure and flexibility of **T** are fist investigated at a proper QM level. This can be based on density functional theory (DFT), its time dependent extension (TD-DFT) or a wave-function (WF) based methods as HF, MP2 or CASPT2. The computed QM descriptors are then stored in a database (see Section 3.3), which serves as JOYCE 's secondary input. Briefly, JOYCE3.0 can then be used in two ways:

i) run a QMD-FF parameterization based on the collected QM database in terms of a well defined set of RICs and associated model potential functions, which is goven in the third input file containing such molecular topology as detailed in Section 3.4;

ii) generate a proper collection of RICs and connected model functions (based on the QM connectivity), and store all information in a output topology file, to be used in a successive JOYCE parameterization run.

Once the definition of the IC set has been read and the QM data retrieved, JOYCE3.0 is able to perform the parameterization, according to the theory reported in Section 7. The final output of a JOYCE parameterization consists in a complete collection of ICs, associated model functions and related FF parameters, stored in the output topology file as discussed in Section 3.5. The latter can be used as such for gas phase simulations or completed with supplementary information concerning the inter-molecular parameters (which can be for instance taken from literature databases or refined on the basis of additional QM descriptors [3,21,22]). This step makes QMD-FF suitable for condensed phase simulation containing molecule **T**.

3.2 Input files

There are three types of input files required to run a JOYCE parameterization on a molecular target **T**, namely:

a) the **main JOYCE input file**, usually named *joyce.<label>.inp*, where *label* is usually the name of **T**. All the parameterization details and paths to the auxiliary input files are contained herein (all details are given in Section 5.1);

b) the **QM** database (see Sections 3.3 and 5.2), that is a single file, when only one optimized structure is considered, or a collection of QM outputs, when torsional scans are taken into account;

c) the **topology file**, *<label>.stepN.top*, which contains a list of the employed ICs and their associated model potential functions (see Section 5.3);

3.3 QM database

Once the target molecule **T** has been chosen, the two main molecular descriptors required for a JOYCE-based QMD-FF parameterizations are its **optimized geometry** and **Hessian matrix**. Additionally, if **T** presents flexible torsions around specific bonds, relaxed and/or rigid **torsional energy profiles** and related geometries should also be included among the reference descriptors. Finally, in specific cases, [20] the energy gradients along the aforementioned scans can be included to complete JOYCE3.0 reference database.

All structures, their energies, gradients and Hessian matrix should be computed at QM level specifically for the traget compound **T** and with a **unique reference method**, being it WF or DFT based. In older JOYCE's releases, all QM data were stored in a formatted checkpoint file (*.fchk*), native for the GAUSSIAN16 [19] package. JOYCE3.0 code now also reads the QM info from the $\mathcal{FC}$*classes*3native *.fcc* format. [18]
All details about the *.fcc* format, the codes for which it supplies an interface, the tools to use it and their distribution can be found at thwe JOYCE3.0 website, [12] or [...]

G: @Javier: could you please complete this paragraph, adding, as Samuele was mentioning, also the url of your Fcclasses tools and related pages?

3.4 Internal Coordinates & Model functions

The RIC set used in the fitting procedure is read from an **input topology file**, usually named, *<label>.stepN.top* (*vide infra*). As discussed in some detail in Section 5.3, such a topology can be purposely edited by the user or taken from existing database, if available. The supported format for the input topology is the GROMACS native ***.top* format**. [17] The input topology file may also contain intermolecular parameters such as point charges and Lennard-Jones (LJ) parameters (see Section 7 for details), as well as other specifications that are not directly related with the JOYCE fitting procedure. In this case, JOYCE simply re-writes this information in the **output topology file**, labeled *joyce.new.top*, without performing any changes.

A complete set of RICs may be automatically generated by JOYCE3.0 in a preliminary run and stored to an auxiliary formatted text file (*.txt*), named *generated.IC*. Such a complete set of RICs is created by JOYCE3.0 from the only information (molecular geometry, connection table, bond order, *etc.*) retrieved from the QM descriptors. The content can be pasted in a new topology file (*e.g.* *<label>.step1.top*) and used in a second run. We stress that although the automatically generated harmonic potential functions might be confidently employed for stretching and bending RICs in most cases, the automated assignment to dihedral coordinates of either harmonic or Fourier-like potentials (see Section 7 for a more complete discussion) is delicate and might require user intervention. For instance, when the molecular structure hides a peculiar coordination (*e.g.* in presence of transition metals), the user can manually refine such automatic choice, or resort to other tools, as those now distributed at the JOYCE3.0 website [12] or in similar repositories [23] add Javier's page).

3.5 Output files

When the parameterization finishes normally, JOYCE3.0 produces two main output files and some optional auxiliary files. The main outputs are a log and a topology file, whose defaults names are *joyce.<label>.out* and *joyce.new.top*, respectively.

- The log file contains all information about the fitting procedure, with a detailed analysis of employed RIC, normal modes, involved vibrational frequencies, residues and standard deviation.
- The default output topology file **always gets the default name** *joyce.new.top* **and contains all parameters of the intramolecular term of the QMD-FF**. It can be used as such in gas-phase simulations with the GROMACS [17] engine.

As reported in more detail in Section 7, JOYCE3.0 QMD-FF parameters can in principle be employed with any MD code, provided the same model potential functions are implemented in the MD engine. Within the current distribution, two tools can be found [12] to automatically convert the final JOYCE3.0 topology in formats compatible to MD software other than GROMACS , namely the popular CHARMM [11] and LAMMPS [24] engines. Yet, **users are strongly encouraged to resort if possible to the GROMACS engine**, as only the *.top* format supports the full list of JOYCE3.0 model potential functions, hence allowing to fully exploit QMD-FF accuracy, in particular when dealing with large flexible targets. Furthermore, other auxiliary files are produced by the JOYCE run, namely a generated.IC, an assign, and dependence file. All these are intended as work scratch files, and can be "copied & pasted" in the main input files for successive runs.

- The *generated.IC.txt* file, contains a list of all IC which can be naturally retrieved from the **T** molecular geometry and its connection table. Note that it also contains by default all the non-bonded intramolecular non-bonded interactions (see Section 7), constructed from the optimized geometry and the intermolecular parameters (charges and LJ) found in the input topology.
- The *assign.dat* files contains the values of all parameterized force constants, and it can be pasted in the JOYCE main input file and used in a successive fitting, in which, for instance, some force constants may be fixed during the parameterization procedure. As detailed in Section 4, this assignment is routinely exploited when applying the two-step JOYCEÂ¥procedure. Further details can be found at the JOYCE3.0 website. [12]
- The *suggdeps.txt* file contains suggested dependencies between FF parameters, dictated by symmetry, chemical equivalence or transferability. This file can be pasted in the JOYCE main input, for a successive fitting where the values of selected force constants will not be independently varied during the parameterization and their ratio is kept constrained (see JOYCE tutorials at the website for further details. [12]
- Several *.dat* files, containing several data to be plotted for inquiry and validation purposes, such as the QM and QMD-FF torsional energy profiles.
- Upon request, JOYCE3.0 can also prepare some graphs, directly saved as *.agr* files compatible with the XmGrace graphical software, [25] containing comparison of relevant data computed either with the QMD-FF or with the reference QM method. The interested reader can visit the JOYCE3.0 website [12] for further information and templates.

All the aforementioned JOYCE's input and output files are described in some detail in the next sections.

4 Running Joyce

When all input files (Joyce's main input, topology and QM data files) have been produced, the program can be launched simply through the following command:

```
> go.joyce label -e -v
```

where *label* is a user defined name related to target molecule **T**, and a Joyce3.0 input file named *joyce.<label>.inp* exists in the working directory. This input file will be automatically edited (through the *-e* option) with the vim editor, and changes to the commands contained therein can be made. When this file is saved and closed, Joyce3.0 starts the parameterization. When the procedure is over, the *.log* file is again opened by the vim editor, so that results can be checked. All other files illustrated in the previous sections are created automatically in the same working directory.

The Joyce parameterization protocol can be applied to a large variety of molecular targets (see main text), resorting to the general step-wise procedure briefly outlined in the following (more details in the rest of the manual).

Step 0 **0.a]** By taking the optimized QM geometry, Joyce creates a list of all RICS of the molecular target, storing them into the *generated.IC.txt* file in the proper format *.top* for successive steps.

0.b] In addition to all bonded RICs, automatically generated by the Joyce3.0 code in **0.a]**, specific atom pairs required to represent specific intra-molecular non-bonded interactions can be included in the RICs collection.

Step 1 The set of selected RICs (either created in **Step0** or imported from other sources) is used to retrieve all harmonic parameters at once, by minimizing the term of the objective function involving QM and FF Hessian matrices (see Section 7 for details).

Step 2 The final QMD-FF is eventually retrieved by assigning all harmonic constants derived in **Step 1** and applying the FIRA [1, 20] while minimizing the term of the objective function involving QM and FF energies, as computed along the QM torsional scan. Note that the combined use of specific non-bonded terms (see **0.b]**) and FIRA is only possible since the Joyce3.0 release. [26].

Please see templates where these steps are carried out at the Joyce3.0 website. [12]

5 Input Files

5.1 Joyce's Main Input File

The Joyce input file is usually named as

joyce.<*label*>.inp

where *label* is an user defined name (usually related to target molecule **T**). It contains all specifications needed for the parameterization, including location of other input/output files, constraints on parameters, file formats, etc.

Each command is given by a keyword, which is activated by a $ symbol before it. The removal of $ deactivates the key. Here, it follows an example of joyce.*molname*.inp file, with a list of most of the supported keys.

```
$title  Parameterization of molecule T

$print  0

$equil   ../QMdata/opt+freq.fchk

$forcefield   gromacs   example.top

$generate

$zero    1.d-12

$whess   5000.   2500.0

$geom
../QMdata/scan1.0.fcc      5.0  0.0  0.0  0.0  ;  1 2 3 4
../QMdata/scan1.30.fcc     5.0  0.0  0.0  0.0  -f
../QMdata/scan1.60.fcc    10.0  0.0  0.0  0.0  -f
../QMdata/scan1.90.fcc     5.0  0.0  0.0  0.0  -f
[...]
$end

$assign
 1   =   2241.69     ; C1-C2
 2   =   2143.14     ; C2-C2
 34  =    732.30     ; C1-C2-C1
$end

$dependence   1.2
 43 =   42*1.d0   ; C2-C2-C1-H1 = H1-C1-C2-C2
$end

$keepff    38 - 103

$LJassign

$scan   joyce.scan1.dat
   1 2 3 4 ; -180. 180. 1.
$end

$gracefreq  joyce.molname_freqchk.agr

$gracetors        molname_torschk
```

$title

Title key simply sets a title for your job, that will appear in the joyce.*molname*.out output file. This is an optional key.

$print

Print key sets the verbosity in the output log file. It has to be followed from a number in the -1,4 range, being -1 the less verbose mode, whereas 4 should be used in debugging sessions only. This is an optional key (the default is 0).

$equil

Equil **keyword is required.** It must be followed by the exact path pointing to the *.fcc* or *.fchk* file, containing target **T**'s optimized geometry and Hessian matrix.

In the above example, a GAUSSIAN16 *.fchk* file is located in *QMdata* directory and is named *opt+freq*.

$forcefield

$Forcefield$ **keyword is also required**. It specifies both format type and location (via its path) of the topology input file. As already noted, the supported input format is the *gromacs* [17] *.top* topology file software packages.

In the above example the input topology file is located in the working directory and named *example.top*.

$generate

This keyword should be only employed in Step 0. When the *$generate* key is activated, the program does not perform any fitting and instead it creates a new text file, named *generated.IC.txt*, where all the internal coordinates of the target molecule are written in the GROMACS *.top* format and can be used as starting topology file in subsequent runs. RIC's are suggested considering bond order and geometrical information retrieved from the QM optimized geometry. Note that after release 3.0, the generated RICs for dihedrals are also automatically associated with the suggested model function, either harmonic or periodic, depending on the IC expected stiffness (see Section 7).

$zero

This keyword is optional, and sets the lower limit of the numerical significance of the computed eigenvalues. In other words, all eigenvalues lower than this threshold will be considered as null, and discarded. It is optional, and its default is $1{\cdot}10^{-10}$.

$whess

Whess **keyword is required for performing any QMD-FF parameterization**. It sets the diagonal and off diagonal weights of the Hessian matrix elements to be used in the objective function **I** (see Section 7). In the example above, such weights are set to standard values commonly used ion most JOYCE parameterizations. Note that Hessian diagonal elements are normally weighted twice with respect to off-diagonal ones.

$geom

Geom **keyword is required for Step 2**. This keyword is needed when performing QMD-FF parameterizations that involve target molecules with flexible coordinates, for which QM relaxed torsional scans were carried out. In these cases, this keyword reads the additional QM files, corresponding to partial optimizations with the flexible IC constrained. All these files have to be listed in the *$geoms* field, which is ended by *$end*. For each row, the path and filename of each *.fchk* file should be specified. This is followed by four numbers, indicating the energy, gradient, diagonal and off diagonal Hessian weights to be applied to the objective function **I**. Note that FIRA is applied to these RICS, and the quadruplet(s) of atoms defining the involved torsion(s) is(are) specified afterwards (preceded by a ";"). Clearly, for each scan performed along a torsional coordinate, all sampled points (resulting in separate *.fcc* or *.fchk* files) refer to the same quadruplet. Starting from the second sampled point, the quadruplet(s) complete declaration in the $geoms environment can be omitted for the sake of brevity, and substituted by the symbol "*-f*". This indicates that the FIRA declarations should be read from the previous line.
In the example reported above, two torsional profiles are considered into the fitting. The first dihedral is defined by atoms 1 to 4, and the energy, gradients and Hessian matrixes relative to all sampled points are retrieved from the GAUSSIAN16 *scan1.*.fchk* (* = 0, 30, 60 and 90) files, located in the same directory as *opt+freq.fchk* file. The second dihedral is defined by atoms 3-2-1-12, and the QM information will be extracted from the files named *scan2.*.fchk* (* = 0, 30, 60 and 90) and again located in the *G09* directory. For all these conformations, only the energy will be considered in the fitting, as all other weights are set to 0. In particular, the weight will be 5 for all energies but those referring to files *scan1.60.fchk* and *scan2.60.fchk*, which are set to 10.

$assign

Assign **keyword is required for Step 2**, as it allows for safely applying FIRA [1,20], by assigning all other harmonic QMD-FF constants to the values obtained in the previous Step 1, based on the Hessian matrix. This keword can also be useful in Step 1, for instance when some of the force constant of the FF have to be constrained to a definite value, chosen by the user or taken from other FF's. The list of such constants is declared in the *$assign* field, where the IC number (consistent with the numbering found in the topology file and reported in the joyce.*molname*.out log file) is first declared, followed by a "=" symbol and the value of the force constant, relative to that IC, to be constrained. Comments can be added after a ";" symbol, to explicitly declare the coordinate type one is dealing with.

Units are kJ/(mol Å^2), kJ/(mol rad^2) and kJ/mol for stretching, bending and torsions, respectively. In the framed example of JOYCE3.0 input file reported above, three force constants are assigned. As it can be deduced by the comments (indicated with the ";" symbol, the first two refers to stretching coordinates (IC number 1, between atoms labels C1 and C2 and IC number 4, between C2 and C2 types), while the last refer to a bending coordinate (IC number 34, defined by C1-C2-C3 sites). These force constants will not be fitted but are constrained to 2241.69 kJ/(mol Å^2), 2143.14 kJ/(mol Å^2), and 732.30 kJ/(mol rad^2), respectively.

$dependence

Some IC's may be equivalent, either for symmetry reasons or because a similar chemical behavior is expected. In these cases it is desirable to describe such IC's with the same equilibrium distances and force constants. In other words, given two equivalent IC's, the parameter values of the second are dependent on those of the first. JOYCE can find dependencies automatically, based on symmetry considerations. A number between 0.5 and 1.0 should be specified as argument of the *$dependence* field, which sets JOYCEsensitivity to IC's equivalence: lower values correspond to less strict criteria, while when this number is set to 1.0 (default), only highly dependent coordinates are considered. Values higher than 1.0 **disable** the automatic selection of dependent coordinates, and a list of all equivalent IC's may be given manually. This can be done by specifying the number consistent with the numbering found in the topology file and reported in the joyce.*molname*.out log file) of the dependent IC, followed by a "=" symbol, the number of the reference IC, a "*" symbol and the number (usually 1 for equivalent IC's) for which the force constant of the reference IC should be multiplied to give the one of the dependent IC.

In the framed example of JOYCE3.0 input file above-reported, the force constant of IC number 43 is forced during the fitting to take the same value of the one relative to IC 42. The reason for this choice is noted in the comment: both dihedrals are defined by the same quadruplet of atom types. Since 1.2 is indicated as sensitivity number, no other dependencies will be imposed.

$keepff

$Keepff$ **keyword is requisite for Step 2, *i.e.* when parameterizing flexible coordinates through model periodic functions**. As for the *$assign* field, also the equilibrium values (r^0_μ, θ^0_μ, *etc*, see Section 7) of the FF can be constrained to the value declared in the parameter **input** file. The list of the RICs, for which the equilibrium values should be taken from the parameter file and **not** from the GAUSSIAN16 equilibrium geometry, is declared as argument of the *$keepff* field. Here, the IC number (consistent with the numbering found in the topology file and reported in the joyce.*molname*.out log file) can be given separately (*e.g.* 3, 6, 12, 54) or grouped (*e.g.* 3 - 10). While it is advisable to force JOYCE to read the γ^μ_j phase constants of the periodic functions, **the equilibrium values of the "stiff" IC should be, as far as possible, taken from the QM equilibrium geometry**.

In the example, the geometrical parameters related to IC number 38 to 103 will be taken from the *exam-*

ple.top topology file and not from the QM equilibrium geometry read from the *opt+frq.fchk* GAUSSIAN16 file.

$LJassign

This keyword is optional, yet recommended in Step 2 when dealing with very flexible targets, which required the use of non-bonded intramolecular terms. *LJassign* was introduced since release 3.0, and activates the so called "Route II" [26] to handle FIRA and internal LJ. Concretely, with this keyword all intramolecular LJ interactions declared in the input topology file are included in the QMD-FF, and the torsional terms fitted while taking into account their contribution.

$gracefreq

This keyword requires JOYCE3.0 to print a *.agr* file, compatible with the popularXMGRACE graphic software, [25] which contains information on the quality of the Hessian fitting in terms of FF vs. QM vibrational frequencies and normal modes.

$gracetors

Similar to the previous keyword and implemented since the 3.0 release, it allows to print an *.agr* file containing information on the quality of the relaxed scan fitting. JOYCE3.0 automatically produce separate *.agr* files for each torsion considered in the following *scan* section.

$scan

When the *$geom* key is activated, *i.e.* when the target molecule has very flexible degrees of freedom, it may be desirable check the capability of the parameterized QMD-FF to reproduce the QM computed torsional profiles. This can be done by comparing the QM energy curves, computed with respect of the minimum energy geometryfrom the files specified in *$geoms* field (and collected in the *qmscan.*.dat* JOYCEoutput files), with those obtained with the parameterized FF. The first argument of the *$scan* keyword is the name of the file in which the FF energy profile should be printed. This should be followed by the declaration of IC number for which an energy scan is requested and the specifications of the scan. The IC's is identified by specifying its defining quadruplet (or triplet for bending) of atoms followed by a ";" symbol. The scan directive are given by specifying the initial and final value that the IC should assume during the scan and the required scan step. The *$scan* keyword **can be repeated** for each IC for which the scan is requested. In the framed example file, only a scan is requested, to be printed in the working directory in a file named *joyce.scan1.dat.* The scan refers to the dihedral defined by atoms 1 to 4, and the FF energy will be computed for all values of this dihedral from -180^o to 180^o in steps of 1^o.

Moreover, the next commands are not included in the above reported framed sample input file, however they activate optional useful features of the JOYCE program.

$rearr

It may happen that the atom numbering employed in the QM calculations differs from the one chosen in the parameter input file. In some cases it may be rather cumbersome to change the order atoms appear, since all the computed informations (Hessian matrices in *.fchk* and IC definitions in FF files) depend on atom numbering. The *rearr* command solves this problem by rearranging the atom numbering in the QM data.

```
$rearr  1  2  3  4  5  6  7  8  12 13  \
        14 11 10  9
```

In the above framed example the numbering of the first eight atoms is unchanged, whereas atoms 12 to 14 are moved in positions 9 to 11, atom 11 is placed in 12th positions, *etc.*

$UnitedAtom

This keyword introduces the UA approximation. The theory is described in some detail in section 7.2.4. The *&UnitedAtom* environment is closed by a *$end* command. Between these keys all UA sites are specified by indicating the number of the atom whose position will be the center of the UA site, followed by the numbers of all atoms that should be grouped to it. An example is framed in the following.

```
$UnitedAtom
8  17 18
10 21 22
11 23 24 25
$end
```

Here atoms 17 and 18 are considered as a unique interaction site with atom 8, with the latter being the atom center of the new UA site. The same procedure is applied to atoms 21,22 and 10, whereas four atoms (23, 24, 25 and 11) are included in the third UA site.

$fitLJ

This key allows JOYCE to fit the Lennard-Jones force constants (ϵ's) defined for non-bonded intramolecular interactions, as reported in equation (11). **It is important to note that for default LJ interactions are not fitted by JOYCE, but they are included as constants in the FF energy expressions. Conversely, when the *fitLJ* key is activated, all employed ϵ's values may be parameterized, so the user needs to explicitly specify those values that he wants to keep constant in the *assign* section.** Note also that geometrical LJ values (σ's) are always taken form the input FF parameter file.

$sep_ el

This key allows Joyce to be coupled with polarizable models or, more generally, with electrostatic models other than standard Coulomb interaction (3). In this case, a second argument is needed in the ***$equil*** section, indicating a path to a *.fchk* file containing the Hessian elements computed for only the electrostatic part of the Hamiltonian. This implies a slight modification of the merit functional (35). Further details can be found in Joyce 2013 paper [27].

$mass

This key is followed by a number, which is used to set all atoms to a unique mass. It is intended for particular applications and its standard use is deprecated.

$normal

This keyword sets the computational route to get vibrational normal modes. With no arguments (default or without using the *$normal* keyword) Joyce computes vibrational normal modes by working in a space orthogonal to the translational and rotational modes (by using the Graham-Schmidt orthogonalization). Conversely, if *zero* is used as argument of the *$normal* keyword the normal modes are computed directly by analyzing the null eigenvectors.

$wfreq

This keyword should be followed by a scaling factor to empirically adjust the computed vibrational frequencies, to correct QM systematic errors when present.

$boltz

A Boltzmann weight can be introduced with this keyword, to weight the energies of the geometries considered. Note that this is alternative to the weights assigned in the *$geoms* section, which is instead recommended.

5.2 QM Database

5.2.1 Native $.fcc$ format

G: I've pasted this part from the SI, I think Javier wrote it. Do we want to change something? I do not think it is required by copyright since it is SI.

Since Joyce3.0, [26] we extended the supported file formats with molecular information, including geometry, energy, gradient and Hessian. Namely, in addition to *.fchk* data files from Gaussian16 (*vide infra*), the code can now extract such information from *.fcc* state files, the native format used in $\mathcal{FC}classes3$. [18] In this way, we can take advantage of tools developed to generate these state files from different electronic structure codes, which are available from a GitHub repository [28].

The $.fcc$ file contains sections, that can appear in any order in the file. Depending on the type of calculation, some of these sections can be skipped. The possible sections are (other section in the file will be ignored, i.e., *INFO*):

- `GEOM` (optionally specifying the units, i.e., `(UNITS = ANGS)`) This section includes the molecular geometry in XYZ format. Namely, the first line of the section indicates the number of atoms, followed by a line that generally contains a comment. Then, the element name and x, y, z coordinates (in Å) are given for each atom in the structure.

- `ENER` (optionally specifying the units, i.e., `(UNITS = AU)`) This section contains the absolute energy in atomic units.

- `GRAD` (optionally specifying the units, i.e., `(UNITS = AU)`) This section includes the elemenets of the gradient, ordered by atom including the x,y,z components for each one. The number of elements by line is arbitrary, but it is usually limited by 5 per line.

- `HESS` (optionally specifying the units, i.e., `(UNITS = AU)`) This section contains the elements of the lower triangular part of the Hessian matrix, in atomic units.

In the following a sample state file in `.fcc` format for water is given. note that not all sections are required in a run.

```
INFO
State file generated from file: water.fchk (format: fchk)
GEOM      UNITS=ANGS
     3

Geometry from water.fchk in xyz format (with filter: all)
O         0.00000000    0.00000000   -0.06661080
H         0.00000000   -0.75914368    0.52858090
H        -0.00000000    0.75914368    0.52858090
ENER      UNITS=AU
```

```
-7.63256737E+01
GRAD      UNITS=AU
 7.81563942E-28  0.00000000E+00 -7.78708105E-06  1.29381262E-23 -4.22604034E-07
 3.12721587E-07 -1.29389077E-23  4.22604034E-07   3.12721587E-07
 HESS      UNITS=AU
 3.85407173E-06  0.00000000E+00  6.75100308E-01  0.00000000E+00  0.00000000E+00
 4.68278502E-01 -1.92703587E-06  0.00000000E+00  0.00000000E+00  1.71308411E-06
 0.00000000E+00 -3.37550154E-01  1.95839781E-01  0.00000000E+00  3.68859016E-01
 0.00000000E+00  2.64651356E-01 -2.34139251E-01  0.00000000E+00 -2.30245568E-01
 2.20815038E-01 -1.92703587E-06  0.00000000E+00  0.00000000E+00  2.13951761E-07
 0.00000000E+00  0.00000000E+00  1.71308411E-06  0.00000000E+00 -3.37550154E-01
-1.95839781E-01  0.00000000E+00 -3.13088619E-02 -3.44057875E-02  0.00000000E+00
 3.68859016E-01  0.00000000E+00 -2.64651356E-01 -2.34139251E-01  0.00000000E+00
 3.44057875E-02  1.33242132E-02  0.00000000E+00  2.30245568E-01  2.20815038E-01
```

5.2.2 Gaussian16 *.fchk* format

The Joyce code alos directly reads all QM data from formatted checkpoint files created by the Gaussian16 package. [19] The formatted checkpoint file (*.fchk*) can be easily created from the more standard unformatted checkpoint file (*.chk*) through the utility *formchk*, available with the Gaussian16 suite of programs.

If a *molname.chk* unformatted checkpoint file has been created by the Gaussian16 program, this can be thus transformed into a formatted *molname* file through the command

% > $GAUSS_EXEDIR/formchk molname.chk

where the environment variable *$GAUSS_EXEDIR* should have been previously defined installing Gaussian16 .

5.3 Topology Input File

The topology input file contains the most important information required for QMD-FF parameterization, namely the molecular topology of target **T**, and consists in

i) the list of atom types to be employed in parameterization

ii) the collection of RICs defining the system

iii) the specification of the model potential functions associated to each coordinate

As already mentioned, the format required for such input file is the *.top* employed by the Gromacs engine. [17] Therefore, only a quick explanation of the main features will be given here. Detailed information about the format of a *.top* topology file can be found in the Gromacs manual. [17]
The main function of the *.top* topology file is divided in three sections:

1. the intermolecular section, which contains the description of the parameters describing intermolecular interactions
2. the intramolecular section which contains the description of the intramolecular FF
3. the system section, which is divided into a ***[system]*** and a ***[molecules]*** environment.

It is only the second part containing the IC specifications that will be used by the Joyce program. At the end of the fitting procedure, Joyce writes the final values of the parameterized FF into a new output *.top* file (called *joyce.new.top*, see Section 6) suitable for MD simulations performed with the Gromacs code. The intermolecular section is instead simply copied and pasted from the input to the output topology *.top* file, except when it is used within the *$generate* framework, as indicated in the previous Section. An example of *.top* topology file for the simple *n*-butane molecule is reported in the following. Users are warmly encouraged to visit the Joyce3.0 website where further templates and tutorials can be found. [12]

```
                           Starting .top file: n-Butane

;                                ---------------------------
;                                 1)  Inter-molecular part :
;                                ---------------------------

[ defaults ]
; nbfunc   comb-rule   gen-pairs  fudgeLJ  fudgeQQ
    1          3          no        1.0      0.0

[ atomtypes ]
; name    mass      charge ptype   sigma(nm)   epsilon (kJ/mol)
 C1      12.0110   -0.18     A       0.350       0.27612
 C2      12.0110   -0.12     A       0.350       0.27612
 H1       1.0079    0.06     A       0.250       0.12555
 H2       1.0079    0.06     A       0.250       0.12555
```

```
;                              -----------------------------
;                                2) Intra-molecular part :
;                              -----------------------------
[ moleculetype ]
; Name      nrexcl
  but         3

[ atoms ]
; nr type   resnr residue  atom   cgnr charge  mass
   1 C1       1     but      C1       1 -0.180 12.0110
   2 C2       1     but      C2       2 -0.120 12.0110
   3 C2       1     but      C2       3 -0.120 12.0110
   4 C1       1     but      C1       4 -0.180 12.0110
   5 H2       1     but      H2       5  0.060  1.0079
                      [...]
  14 H1       1     but      H1       6  0.060  1.0079

[ bonds ]
;  ai    aj    type     r0 (nm)   ks (kJ/(mol nm^2))
   1     2     1        0.0       0.0              ;  1   C1-C2
   2     3     1        0.0       0.0              ;  2   C2-C2
                  [ ... ]
   1    14     1        0.0       0.0              ; 13   C1-H1

[ angles ]
;  ai    aj    ak type  th0 (degr)   kb (kJ/(mol rad^2)
   1     2     3    1      0.0        0.0          ; 14  C1-C2-C2
   2     3     4    1      0.0        0.0          ; 15  C1-C2-C2
                     [ ... ]
   2     1    13    1      0.0        0.0          ; 36  C2-C1-H1

   2     1    14    1      0.0        0.0          ; 37  C2-C1-H1

[ dihedrals ]
;  ai    aj    ak    al type   gam  kd (kJ/mol) n
   1     2     3     4     1   0.0000   0.0      0  ; 38 C1-C2-C2-C1
   1     2     3     4     1 180.0000   0.0      1  ; 39 C1-C2-C2-C1
   1     2     3     4     1  90.0000   0.0      2  ; 40 C1-C2-C2-C1
   1     2     3     4     1   0.0000   0.0      3  ; 41 C1-C2-C2-C1
   1     2     3     4     1   0.0000   0.0      4  ; 42 C1-C2-C2-C1
  12     1     2     3     1   0.0000   0.0      3  ; 43 H1-C1-C2-C2
   2     3     4     9     1   0.0000   0.0      3  ; 44 C2-C2-C1-H1

 [ pairs ]
;  ai   aj        f_qq    qi     qj   sigma (nm)  epsilon (kJ/mol)
;  1-4  C1--C1
   1   4    2      0.0    0.0    0.0   0.3700        0.1

[ exclusions ]
;  ai    aj
   1     2
   1     3
    [...]
  13    14
;                                      ----------------------------
;                                       3)  System definition :
;                                      ----------------------------

[ system ]
; Name
n-butane

[ molecules ]
; Compound   #mols
but            1
```

Intermolecular section

The intermolecular section has always to be the first in a *.top* file. It contains the atom labels definition and it is made up of two environments.

[defaults]

The *[defaults]* environment defines the potential functions used to compute non-bonded interactions, reported in equation (11) in Section 7. The outline of this section has always to be the following:

```
    [defaults]

          $1    $2    $3    $4    $5
```

Variable $1 defines the form of the non-bonded model potential function, and $1 = 1 sets the LJ form, shown in equation (4). Variable $2 sets the combination rules to be used to compute ϵ_{ij} and σ_{ij} from the single site values ϵ_i and σ_i: $2 = 2 selects Lorentz-Berthelot mixing rules, while OPLS ones are chosen by setting $2 = 3. The third variable ($3 = yes/no) forces Gromacs to compute non-bonded intramolecular interactions by using the same ϵ_i and σ_i given for the intermolecular ones. Variables $4 and $5 define the scaling factors that will be respectively applied to LJ and Coulomb terms of the non-bonded intramolecular interactions. Note that Joyce ignores the latter three variables (but they are copied and pasted into the final output topology file), so variable $3 should be switched off all the time, to avoid confusion. **If intramolecular non-bonded LJ interaction are to be used within the Joyce3.0 procedure, the intramolecular ϵ_{ij} and σ_{ij} values, together with the chosen scaling factors, should be separately defined in the *[pairs]* section (see the following).** Conversely, if the *$generate* key is activated, the LJ intramolecular parameters will be automatically created through the appropriate mixing rules form the ϵ_i and σ_i values read from the *[atomtypes]* section, illustrated in the next section. In the sample file above, OPLS combination rules are chosen for the n-butane molecule's inter-molecular term, whereas LJ intramolecular pairs are specifically defined in the *[atomtypes]* section (*vide infra*).

[atomtypes]

In this environment, all atom types of the adopted model are defined, by specifying for each type i its mass, charge and ϵ_i and σ_i **intermoilecular** LJ parameters. **It is important to stress that the JOYCE procedure allows for defining any number of atom types, ranging from the few provided by general purpose FFs to more specific descriptions (dictated by symmetry or chemical equivalence).** The user is encouraged to visit JOYCE3.0 website [12] for a gallery of different examples. The outline of the *[atomypes]* section is the following:

```
[atomypes]

        $1    $2    $3    $4    $5    $6
```

Variable $1 defines the atom type label, $2 its mass in g mol^{-1}, $3 its charge in |e|. Variable $4 defines the kind of particle atom type refer to: in GROMACS three *ptypes* are possible, namely atoms ($4 = A), shells ($4 = S) or virtual sites ($4 = V). The last two variables, $5 and $6, set the values of the atom type LJ parameters σ_i and ϵ_i, in nm and kJ/mol, respectively. In the framed sample file, three atom types are defined namely H, C1 and C2 for Hydrogen, CH_3 and CH_2 carbon atoms, respectively.

Intramolecular section

The molecular topology is defined here in a number of subsections, which are discussed in the following.

[moleculetype]

This subsection is required, because, like for atom types, any molecule has to be specified in GROMACS by a label. In the intramolecular section, target **T**'s molecule-type is defined in the *[moleculetype]* subsection.

```
[moleculetype]

$1   $2
```

Variable $1 is used to identify molecular type by a string which can be up to 10 characters long, while $2 defines the number of bonds which have to separate two sites for the intramolecular non-bonded interactions to be computed. In the example file framed at the beginning of this Section *but* is the label assigned to the *n*-butane molecule, whereas $2 = 3 stands for excluding non-bonded interactions between atoms that are closer than 3 bonds.

[atoms]

The *[atoms]* subsection is the most important environment, as it defines the molecule, matching each atom with its atom-types, defined in *[atomtypes]*. **In Joyce parameterizations this section must follow the *[moleculetype]* subsection, because all other subsections use the order of sites declared here.**

```
[atoms]

        $1      $2      $3      $4      $5      $6      $7      $8
                 ⋮
```

Variable $1 specifies the atom number, while $2 specifies its atom type. Arguments $3 and $4 are used by Gromacs to specify the residue or molecule to which the atoms belong, by giving its number and label, respectively. Variable $5 specifies the atom name, while $6 indicates the charge group (see Gromacs manual for further details). Finally, last two variables set atomic charge and mass, which can be varied by the user with respect to the one given in the *[atomtypes]* environment.

In the Joyce program, it is fundamental that the order of the atoms (specified by $1) matches exactly the order used in the QM calculations (this in all *.fcc* or *.fchk* files). Although severe tests are automatically performed by Joyce3.0 to verify this feature, the user should always control the numbering of atoms in both files.

In the example reported in frame, the label C1 is assigned to Carbon atoms of the CH_3 groups, while C2 label indicate internal Carbons. Atoms 4 to 14 are all Hydrogen atoms, corresponding to a unique atom type, H.

[bonds]

Within the *[bonds]* environment Gromacs allows for several distinct types of bonded pair interactions, whose detailed description can be found in the online manual. [17] At the moment only harmonic stretching is implemented in Joyce3.0, so only these interactions will be described. Indeed, each harmonic bond between two sites is defined by writing the following line.

```
[bonds]
$1    $2    $3    $4    $5
             ⋮
```

Variables $1–$5 are defined according to equation (7) in Section 7, where $1 and $2 define the **atom numbers** (referred to the ordering given in the *[atoms]* environment) of the pair μ involved in the bond, $4 is the equilibrium distance r^0_μ in nm and $5 the stretching force constant k^s_μ in kJ/(mol nm^2). Finally argument $3 indicates the type of function to be used for the stretching potential: $3 $=$ 1 sets the harmonic form. In the sample *.top* file, framed at the beginning of this section, the definitions of selected bonds for the butane molecule are shown. As noted in the comment, the first bond refers to bond between the first $C1$ and $C2$ atoms, while the second line describes the $C2 - C2$ IC. The last line describes a C1-H type bond, between atoms 1 and 14. The IC numbering in the Joyce procedure is thus taken from this ordering: distance between atoms 1 and 2 (C1-C2) will be IC number 1, distance between atoms 2 and 3 (C2-C2) IC number

2, distance between 3 and 4 (C2-C1) number 4, and so on. **Note that in the *.top* file produced by JOYCE3.0 the IC numbering is explicitly declared in the comment as reported the template above.**

[angles]

As for stretching, more than one functional form is implemented in GROMACS to describe angle bending. [17] However the following discussion will be limited to harmonic form. A harmonic angle bending interaction is defined in the *[angles]* section according to

```
[angles]
      $1   $2   $3   $4   $5   $6
            ⋮
```

where the parameters $1–$6 indicate the triplet μ of atoms defining the angle ($1–$3), the equilibrium angle θ^0_μ in degrees ($5) and the bending force constant k^b_μ in kJ/(mol rad^2) ($6) that enter in equation (8) in Section 7. As for stretching, argument $4 indicates the type of function to be used for the bending potential: $4 $=$ 1 sets the harmonic form. In the sample file some of the butane angles are defined. In such way, bending IC are added to the stretching IC included in the FF by defining them in the previous *bond* environment. IC numbering just increases sequentially: if for instance 13 is the number of the last defined bond, the first angle here reported will be IC number 14, and so on.

[dihedrals]

GROMACS provides several functional forms for the description of the potential with respect to a dihedral angle. [17] In particular, both harmonic and Fourier forms, defined respectively in equation (9) and (10) in Section 7, are supported and have to be defined in the same*[dihedrals]* section according to the following format:

- Stiff dihedrals: each harmonic dihedral is given by a line in the following form:

  ```
  [dihedrals]
  $1   $2   $3   $4   $5   $6   $7
                ⋮
  ```

  Variable $5 is set to 2 to activate harmonic torsions. All other variable ($1–$4 and $6–$7) refer to equation (9) and are are defined in the usual way: parameters $1–$4 indicate the atom quadruple μ which defines the dihedral δ_μ, $6 the equilibrium dihedral angle ϕ^0_μ in degrees and $7 the harmonic torsion force constant k^t_μ in kJ/(mol rad^2).

- Soft dihedrals: the parameters defining the Fourier-like torsional potentials are defined as:

  ```
  [dihedrals]
  $1   $2   $3   $4   $5   $6   $7   $8
                ⋮
  ```

At difference with harmonic torsions, variable $5 is set to 1 to activate Fourier like torsions. An anharmonic dihedral potential has to be defined in terms of a Fourier series with an, in principle, arbitrary number of cosine–terms. The number of terms specifying the interaction for the site quadruple $(\kappa - \lambda - \omega - \tau)$, where $(\lambda - \omega)$ represents the central bond, is therefore also technically not restricted. With reference to equation (10), arguments $1–$4 indicate, as in the previous case, the atom quadruplet μ which defines the dihedral. Variables $6–$8 set the values for the phase angle γ_j^μ in degrees ($6), the force constant (in kJ/mol) $k_{j\mu}^d$ ($7) and the multiplicity n_j^μ (($8) for the j-th term in the Fourier series.

Please note that if both harmonic and anharmonic forms have to be considered (usually referred to different dihedrals) only one *[dihedrals]* environment is necessary, since both function types are defined by variable $5. In the butane *.top* example file, all defined dihedrals are anharmonic. Note that 5 cosine terms are used to describe the potential function for the dihedral formed by the four carbon atoms (1-4), while only one function is employed for the methyl dihedrals. Please note that, despite 5 cosine terms are used for the potential function (see equation (10), only one IC arises from the same atom quadruplet. In other words, only three new IC (dihedrals) are added to the IC list in the aforementioned example.

[pairs]

In almost all general purpose force-fields, [7, 8, 10, 13, 29] the intramolecular non-bonded interactions are automatically included in the FF by accounting for the interactions among **all** atoms of the target molecule, except those that are less than three bonds away. More specifically, the interactions within all considered pairs are computed through the usual sum of LJ and Coulomb terms, employing **as such** the same parameters defined for intermolecular interactions. The only exception being the 1–4 terms, that are usually scaled by some empirical factor. However, GROMACS enables *any* intramolecular interaction to be modified specifically. This feature is exploited by JOYCE in order to:

1. select only specific atom pairs to interact, while excluding all the rest
2. employ for such pairs user-defined LJ parameters
3. turn off all charge-charge interaction **within** the molecule if not necessary (as in most neutral molecules)

All such specifics are given in the *[pairs]* section as follows:

```
[pairs]
      $1    $2    $3    $4    $5     $6    $7    $8
             ⋮
```

where $1 and $2 are the numbers of the involved pair, variable $3 sets the functional form to be employed: $3 = 1 uses only LJ term of equation (11), while $3 = 2 also adds Coulomb contributions. In the former case ($3 = 1), $4 sets the value of σ_{ij} in nm and $5 the force constant ϵ_{ij} in kJ/mol. Conversely, when $3 = 2, $4 indicates the linear scaling factor to apply to charge-charge interactions, $5 and $6 are the charges assigned to $1 and $2 atoms. Finally $4 sets the value of σ_{ij} in nm and $5 the force constant ϵ_{ij} in kJ/mol.

It is important to underline once more that LJ parameters (and charges) concerning intramolecular interactions are defined here and may in principle differ from the ones employed for intermolecular interaction and defined in the *[atoms]* section. Note also that in the standard Joyce3.0 procedure, the charges (**for intra-molecular interactions**) are set to zero and hence not employed. Nonetheless, the default topology created by Joyce through the *$generate* command (see Section 5) lists all possible intramolecular pairs, yet assigning their parameters by applying Lorentz-Berthelot combination rules on the intermolecular parameters defined for each atom. In the butane framed sample file reported at the beginning of this section, only a single non-bonded intramolecular IC is activated. The selected pair considers only the two methyl carbons, which interact through the sole LJ terms, with a user define σ of 3.7Å and and ϵ of 0.1 KJ/mol.

[exclusions]

To consistently apply Joyce3.0 default parameterization route [26] it is necessary to first switch off all intramolecular non-bonded interactions explicitly, and then declare only the selected active pairs through the *[pairs]* section as detailed above. For this purpose all pairs have to be declared in the *[exclusions]* sections as follows:

```
[exclusions]
      $1      $2
               ⋮
```

SG: We should probably mention the priorities with which Gromacs handles pairs and exclusions. Will discuss with Giacomo.

@Javier: I'm afraid we should add here a brief section of couplings ...

System section

This section is only used by Gromacs during MD and ignored by Joyce (but copied and pasted into the final output *.top* file). It contains a title for the simulated system and the number of molecules for each species in the simulation. In the framed sample file, the system is simply named $n - butane$ and only one molecule is expected.

It is important to stress that this section allows to **extend the specific QMD-FF refined by Joyce from gas-phase simulations to condensed phase systems**. For instance, bulk liquids mad up of N_{mol} target molecules or solute-solvent systems, where solute target **T** is surrounded bu one or more different species. In the case of systems with the same kind of molecules, only their number should be specified in the *molecules* section. This allows Gromacs to use the QMD-FF description for each of them. Conversely, if target **T** has to be solvated, a new intramolecular section should be prepared for the solvent and appended to the one created by Joyce.

6 Output Files

6.1 Main output file

The Joyce standard output file is usually named as

joyce.<*label*>.out

where joyce.<*label*>.inp it is the user defined name of the related input file. This file contains printouts and detailed information about all operations performed by the Joyce code during the parameterization procedure. The verbosity level of the printouts is set by the value assigned to the $*print* keyword in the joyce.*molname*.inp file. Because of the many information it contains, the output file cannot be described in much detail in this manual. Only some of the most important sample sections will here be illustrated. **A careful reading of the output log file is highly recommended to Joyce new users**.

In the following a short explanation of the most important sections of the output file is given.

```
================================================
         JOYCE  Parameterization starts
================================================

------------------------------
   A)  Scanning input file
------------------------------

      input file ok

Frequency plot required: joyce.but_freqchk.agr
Torsions plot required:    but_torschk

Title: | Butane - Step 2 |


------------------------------
   B) Reading QM training data
------------------------------

QM INPUT FILE for geom-0 (QM):  ../../QMdata/opt+freq.fcc

============ reading FCC data ===========
 title: Butane opt geom + freq
 n.atoms ..........................    14
 E(tot) ...........................  -158.45877100
```

In the above frame, the first two sections of the joyce.*molname*.out log file are reported. In section *A*), Joyce3.0 scans the input file and checks its validity. Thereafter, prints some info on the required plot files (if any) and on the project title. In the second section, Joyce prints some information retrieved from the QM database (here in .*fcc* format). In particular, the QM reference energy (E_0) appearing in equation (35) is recovered and printed (*E(tot)*). From the QM optimized geometry Joyce3.0 also recovers the connection table of the target molecule **T**. It reconstructs the molecular connectivity from the atomic covalent radii implemented in the code and the related bond orders. Note that this information can be printed in the output by increasing the print level through the ***print*** key.

In the same section, as reported in the frame below, Joyce3.0 recovers, based on the bond orders computed for the optimized geometry, **all** intramolecular coordinates (bonds, angles and dihedrals), called natural IC or

NIC, printing their labels (according to the atom names found in the QM database) and equilibrium values.

```
====== INTERN COORD analysis:  NATURAL-ICs: style=FF  ======
   TYPE     NAME            EQUIL.VALUE
                            Angs or deg
      1  distance  C1-C2                      1.5302       1   2
      2  distance  C2-C3                      1.5325       2   3
             [...]
     19  angle     C1-C2-H9                 109.5439      1   2   9
     20  angle     C2-C3-C4                 113.3938      2   3   4
             [...]
     38  dihedral  C1-C2-C3-C4             -179.9998       1   2   3   4
              [...]
     64  dihedral  H11-C3-C4-H14           -178.0147      11   3   4  14
```

In the last columns, the numbers (which refer to the atom order found in *.fcc* file) of the atoms involved in the definition of each NIC are also indicated. All the recovered NIC are printed in the *generated.IC.txt* output file (see subsection 6.2).

In the third section, JOYCE3.0's output shows instead the set of ICs defined in the GROMACS topology file, are printed together with their equilibrium value (computed from the QM optimized geometry), as shown in the following.

```
------------------------------
  C) Reading FF & IC definition
------------------------------

Gromacs input file ..........: but.step2.top

1) Atom Types
Site  Name       Charge        Mass        Sigma        Epsilon
 1     C1         0.000       12.011        0.350         0.276
 2     C2         0.000       12.011        0.350         0.276
 3     H1         0.000        1.008        0.250         0.126
 4     H2         0.000        1.008        0.250         0.126

 2) Stretching parameters

 Bond      Atoms            k_s       r0    FF term     Atoms
 1       C1    C2         2254.51    1.530    1          1    2
 2       C2    C2         2130.51    1.533    2          2    3
                  [...]
13       C1    H1         3100.08    1.104   13          4   14

3) Bending parameters

Angle       Atoms          k_b     theta0   FF term       Atoms
1     C1    C2    C2     710.63  113.390     14          1    2    3
2     C2    C1    H1     346.67  111.620     15          2    1    5
           [...]
22     H1    C1    H1     313.34  107.580     35        12    4   13
23     H1    C1    H1     313.34  107.580     36        12    4   14
24     H1    C1    H1     313.34  107.330    37        13    4   14
```

```
4.2) Fourier torsions
Dihedral        Atoms        Ncos   K_d      n    gamma    FF term      Atoms
1        C1  C2  C2  C1   5    0.0000   0      0.00      38     1   2   3   4
                               0.0000   1     0.00       39
                               0.0000   2      0.00      40
                               0.0000   3      0.00      41
                               0.0000   4      0.00      42
2        H1  C1  C2  C2   1    0.0000   3      0.00      43     5   1   2   3
3        H1  C1  C2  C2   1    0.0000   3      0.00      44     6   1   2   3
4        H1  C1  C2  C2   1    0.0000   3      0.00      45     7   1   2   3
5        C2  C2  C1  H1   1    0.0000   3      0.00      46     2   3   4  12
6        C2  C2  C1  H1   1    0.0000   3      0.00      47     2   3   4  13
7        C2  C2  C1  H1   1    0.0000   3      0.00      48     2   3   4  14
The following functions keep the R0/Ang0/Gamma values as given in FF file
38   C1-C2-C2-C1_n=0
39   C1-C2-C2-C1_n=1
40   C1-C2-C2-C1_n=2
41   C1-C2-C2-C1_n=3
42   C1-C2-C2-C1_n=4
43   H1-C1-C2-C2_n=3
44   H1-C1-C2-C2_n=3
45   H1-C1-C2-C2_n=3
46   H1-C1-C2-C2_n=3
47   H1-C1-C2-C2_n=3
48   H1-C1-C2-C2_n=3
```

If the *$keepff* keyword is activated, a list of the selected IC equilibrium values (constrained to the value read from the topology file) is also given at this point, as shown above. It is important to stress once again that the set of IC defining the FF and effectively employed during the parameterization **is the one read from the GROMACS topology file**. This set can be arbitrarily chosen by the user, depending on the characteristics of the target molecule **T** (as illustrated in section 7.1.2 and Figure 6), and the number of IC that compose it can exceed 3N-6, N being the number of atoms of the molecule. For these reason, as mentioned in section 7.1.2, **the IC selected in the topology file are named as redundant IC (RIC)**.

Once the RIC set has been defined the parameterization procedure starts. First of all, JOYCE3.0 associates a model function to each RIC, as assigned in the topology file. Thereafter, the program starts retrieving information from the QM training database, concretely by considering **T**'s optimize geometry and Hessian matrix. From such info, QM vibrational modes and frequency are analyzed in temrs of the defined RICS, and the results printed as follows:

```
-----------------------------
D) Working with the QM Hessian
-----------------------------

INCREMENT Alpha, Beta matrices for ABSOLUTE MINIMUM GEOMETRY
Energy weight .....................    0.0000
Gradient weights ..................    0.0000
diag Hessian weights ............... 5000.0000
off-diag Hessian weights ........... 2500.0000
further freq dep. weight for Hessian   -1.0000

  ====================================================
     N O R M A L   V I B R A T I O N A L   M O D E S
                       GEOMETRY 0
  ====================================================

  Compute the Mass weighted Hessian
  Diagonalize the Mass weighted Hessian
  Eigenvalues of the (M-1/2)*F*M(-1/2) matrix (mH)GEOMETRY 0
  1 0.3141E-03   2 0.1080E-02   3 0.1377E-02   4 0.1490E-02    5 0.3745E-01
  6 0.1143E-01   7 0.1360E-01   8 0.1482E-01   9 0.1895E-01  10 0.1980E-01
  11 0.2179E-01  12 0.2396E-01  13 0.2809E-01  14 0.2975E-01  15 0.3388E-01
  16 0.3566E-01  17 0.3623E-01  18 0.3996E-01  19 0.4058E-01  20 0.4063E-01
  21 0.4425E-01  22 0.4444E-01  23 0.4478E-01  24 0.4489E-01  25 0.4537E-01
  26 0.4577E-01  27 0.1871      28 0.1879      29 0.1893      30 0.1895
  31 0.1899      32 0.1927      33 0.1978      34 0.1983      35 0.1989
  36 0.1990

            Frequencies in 1/cm   GEOMETRY 0
   1    123.003   2    228.124   3    257.531   4    267.883   5    424.756
   6    741.881   7    809.383   8    844.930   9    955.322    10    976.484
  11   1024.469  12  1074.392  13  1163.236  14  1197.042    15  1277.571
  16   1310.553  17  1320.966  18  1387.318  19  1398.026    20  1399.016
  21   1459.989  22  1463.041  23  1468.742  24  1470.551    25  1478.275
  26   1484.835  27  3002.068  28  3008.730  29  3019.914    30  3021.358
  31   3024.177  32  3046.798  33  3087.033  34  3090.517    35  3095.316
  36   3096.069
```

Joyce recovers the Cartesian Hessian matrix from the QM optimized geometry and computes mass weighted Hessian, normal modes and frequencies according to equations (27)-(30). Note that translation and rotations are not considered. As shown in the above frame, it prints all the information gained, numbering the frequencies (and the corresponding normal modes) from the lowest to the highest. For each normal mode, Joyce3.0 computes its projection over the set of selected RIC, printing them in a matrix form, where each column contains the coefficients of the normal mode corresponding to the reported frequency projected over the RIC. Each row contains up to 10 columns. In the example framed here below, we show how this simple analysis gives a quick snapshot on how the vibrations distribute themselves over the RIC. As could be expected, the lowest frequencies involve many RIC, essentially dihedrals, whereas high frequency ones are more localized in stretching or bending RIC. **It is crucial to stress here that such analysis gives hints about which IC should be considered as flexible (and as such represented through periodic functions) or stiff (to be accounted for via harmonic potentials).**

```
                  VIBRATIONAL NORMAL MODES: Int.Coord. displac.s GEOMETRY 0
                   123.2  228.4  257.7  267.9  424.7 ...
                     1      2      3      4      5   ...
   1 C1-C2         0.000  0.000  0.003  0.000 -0.021 ...
   2 C2-C2         0.000  0.000  0.000  0.000 -0.019 ...
   3 C2-C1         0.000  0.000 -0.003  0.000 -0.021 ...
  14 C1-C2-C2      0.000  0.000 -0.064  0.000 -0.051 ...
  15 C2-C2-C1      0.000  0.000  0.064  0.000 -0.051 ...
  16 C2-C1-H       0.000  0.000  0.017  0.000  0.024 ...
                                  [...]
  38 C1-C2-C2-C1 -0.207  0.043  0.000  0.000  0.000 ...
  39 H-C1-C2-C2  -0.039 -0.177  0.000 -0.197  0.000 ...
  40 C2-C2-C1-H  -0.039 -0.177  0.000  0.197  0.000 ...
  41 C1--C1        0.000  0.000  0.000  0.000 -0.159 ...
  42 H--C1         0.169 -0.044 -0.173  0.002 -0.119 ...
  43 H--C1        -0.169  0.044 -0.173 -0.002 -0.119 ...

                                  [...]

                  3024.1 3046.8 3087.0 3090.5 3095.4 3096.1
                    31     32     33     34     35     36
   4 C1-H          0.000  0.000  0.000  0.000 -0.265 -0.279
   5 C1-H          0.040  0.088 -0.273 -0.290  0.108  0.118
   6 C1-H         -0.040 -0.088  0.273  0.290  0.107  0.118
   7 C1-H          0.000  0.000  0.000  0.000 -0.263  0.281
   8 C1-H         -0.040  0.088  0.273 -0.290  0.107 -0.119
   9 C1-H          0.040 -0.088 -0.273  0.290  0.107 -0.119
  10 C2-H         -0.234 -0.238 -0.046 -0.108 -0.015 -0.022
  11 C2-H          0.234  0.238  0.046  0.108 -0.015 -0.022
  12 C2-H         -0.234  0.238 -0.046  0.108 -0.015  0.022
  13 C2-H          0.234 -0.238  0.046 -0.108 -0.015  0.022
  16 C2-C1-H       0.000  0.000  0.000  0.000  0.013  0.014
  17 C2-C1-H      -0.010 -0.014  0.013  0.012 -0.007 -0.007
  18 C2-C1-H       0.010  0.014 -0.013 -0.012 -0.007 -0.007
  19 H-C1-H       -0.002 -0.003  0.012  0.014  0.006  0.006
  20 H-C1-H        0.002  0.003 -0.012 -0.014  0.006  0.006

                       End Normal Modes Calculation
                       Computed Alpha,Beta for HESSIAN points     1  666
                       Computed  666  new points for geom:
                       ABSOLUTE MINIMUM GEOMETRY
```

If flexible ICs are present, and a QM relaxed scan is given to JOYCE in the *geoms* section of the input file, the code starts a loop on all given geometries, printing the results in section *E*) of the output file. All scanned geometries go through the following operations:

i) Cartesian coordinates and the total energy are read for each geometry from the QM database

ii) FIRA is applied to the current geometry by performing a rigid rotation of the scanned dihedral, starting from the fully optimized geometries and displacing the chosen dihedral to the value computed from the QM current constrained optimization.

iii) The QM relaxed energy read from the training database is associated with the new geometry obtained through FIRA.

The following analysis is only performed on geometries other than the absolute minimum, thus it will only be activated if an energy scan has been performed on some selected RIC.

```
-------------------------------
E) Working with relaxed QM scans
-------------------------------
========================================
            GEOMETRY n.  1
========================================
QM input file: ../../QMdata/Scan1/butane.delta_000.fcc
============ reading FCC data ===========
title: State file generated from file: butane.delta_000.fchk (format: fchk)
n.atoms .........................    14
E(tot) ..........................  -158.44960700
Hessian not found on file fcc
E(tot) - E(reference geom) ........    24.06008200
The geometry is changed according to the FROZEN options
1. the follow. RICs are obtained by the current QM geom
2. the reference geom is changed accordingly
      and the obtained geom is used in the follow

RIC     --- atoms ----      current    refer   change
1       1   2   3   4      -0.000 -180.000  180.000
WARNING: the following RICs are changed too much
     FrozGeo   RelaxGe    change                    allowed
14   113.394   116.838    -3.444      1   2   3   0   3.000    1.148
20   113.394   116.838    -3.444      2   3   4   0   3.000    1.148
----- GEOMETRY from frozen changes (angstrom) -----
                x           y           z        Nucl.ch
1  C1     -0.4740072  -2.1583107   0.0012238     3.2
2  C2      1.0561133  -2.1764456   0.0021622     3.2
3  C3      1.6478863  -3.5900924  -0.0001270     3.2
4  C4      0.5869913  -4.6928581  -0.0028806     3.2
5  H5     -0.8678333  -1.1289935   0.0029178     0.5
                           [...]
```

As an example, in the above frame is reported the information printed by Joyce, concerning the first of the n-butane non equilibrium geometries (given in the *$scan* section). In this geometry, the carbon backbone dihedral was displaced to 0° , *i.e.* in a *cis* conformation, but similar information is printed in the output for all other scanned geometries. In the first block, information about the QM level of theory is printed, together with the computed absolute energy and the energy difference (*E(scf) - E(reference geom)* in kJ/mol) with the absolute minimum (geometry 0 in the *trans* conformation). In this case the latter is $\sim$ 24 kJ/mol. Next, the scanned RIC is specified together with some information about the employed FIRA approximation, explained in some detail in section 7.2.5. If the value of any RIC, other than the scanned one, results to be much different in the current scanned geometry with respect of the reference (absolute minimum) one, Joyce gives a warning and prints the difference between the two values. **Particular attention should be**

paid to this info in case an atom pair interacting through nonbonded interactions appears in this list. If so, please resort to the *LJassign* keyword as detailed in Section 5.1. In the previous frame, for example, RICs number 14 and 20, *i.e.* the bending angles the two Carbon atoms triplets (1,2,3 and 2,3,4) change together with the scanned dihedral, passing from ∼ 113° (in the *trans* minimum energy reference geometry) to ∼ 117° (in the current *cis* conformation). Finally, the geometry created through the FIRA and effectively employed in the parameterization is printed for reference.

When the loop on all geometries is over, JOYCE3.0 starts the QMD-FF parameterization, solving the system of equations to find the best linear parameters as reported in equations (35)-(45). The results are summarized in section F) of the output as follows.

```
     ------------------------------
     F) QMD-FF parameterization
     ------------------------------
     ==============================================
     S O L V E   T H E   L I N E A R   S Y S T E M
     F O R   T H E   B E S T   P A R A M E T E R S
     ==============================================

     n. of parameters .............   48
     n. of points .................  680

     ----------- read dependences -----------
    1          param  44   =   param  43  *    1.0000
    2          param  45   =   param  43  *    1.0000
    3          param  46   =   param  43  *    1.0000
    4          param  47   =   param  43  *    1.0000
    5          param  48   =   param  43  *    1.0000

    ------- read assigned parameters -------
    ------- Expected input units:    -------
    ------- [L] =A ; [E] = kJ/mol    -------
    - all nonbonded prms have been assigned -

         ------- Expected input units:    -------
         ------- [L] =A ; [E] = kJ/mol    -------

          1   param   1   =      0.24046    input =  2254.5152
          2   param   2   =      0.22723    input =  2130.5117
          3   param   3   =      0.24046    input =  2254.5152
                                 [...]
         35   param  35   =      0.11934    input =   313.3363
         36   param  36   =      0.11934    input =   313.3363
         37   param  37   =      0.11934    input =   313.3363
```

First, JOYCE looks whether assigned parameters or dependencies were given in the input, through the *$assign* or *$dependencies* commands. If so, all read assignments are printed out, as shown in the frame below. Please note that the **parameter number always refers to the order given in the topology file**, where the RIC spanning the FF are defined. The linear system is solved using the Single Value Decomposition (SVD) algorithm, [1] after JOYCE has constructed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ matrices (see section 7.2.3) from the information reported in all above frames, considering the imposed constraints and/or dependencies.

Further details about the numerical procedure are also printed by the program, as shown in the next frame. After the number of free parameters has been written, the effective threshold (as set in the *$zero* environment in the input file) and the eigenvalues consequently discarded are reported in some detail, which can be increased by setting the proper print level through the *print* keyword.

```
Number of free parameters .........     6
given threshold for null eigenvalues ....     0.100D-11 * Max(eigenvalue)
maximum value of the A matrix ...........   0.357D+01
both A and B are multiplied by ..........   0.100D+01
sum (over points) of the weights ........   0.100D+02
first kept metric eigenvalues from    43
7.749D-02  1.081D-01  1.081D-01  3.691D-01  1.773D+00

metric eigensolution n.  42   0.000000
0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
                      [...]
0.00   0.00   0.00   0.00   0.00   0.00   0.00   1.00

metric eigensolution n.  43  0.7749045E-01
0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
                       [...]
0.00   0.00   0.00   0.00   0.00   0.00   0.00  -0.76   0.27   0.37
0.27   0.37   0.00   0.00   0.00   0.00   0.00   0.00

Max err in the matrix inversion     1.00
for the matrix element               1    1
```

Due to the empirical nature of this threshold, it is **requisite**, at least in the preliminary runs, to check if the selected values is adequate to discard only the null eigenvalues. To aid this task, as appears in the above frame, Joyce prints the first discarded eigenvalue and the first one kept.

As displayed in the next frame, Joyce3.0 prints a summary of the best parameters final values (both in atomic and standard units), specifying if the constants was assigned, dependent or free to vary. In the latter case the program also computes and writes a "sensitivity index" (which is set to zero in the former cases), that gives some indication whether the selected RIC is connected to the scanned one.

```
     ==== VALUE OF THE BEST LINEAR PARAMETERS ====
         # param.s   48        #. of free param.s   11
         sensitivity index:    =1 OK,  =0 useless function
         ---------------------------------------------

         #    value     Sens Index
         1   0.240460      0.000   C1-C2
         2   0.227234      0.000   C2-C2
         3   0.240460      0.000   C1-C2
         4   0.330646      0.000   C1-H1
         5   0.330646      0.000   C1-H1
         6   0.330646      0.000   C1-H1
         7   0.319339      0.000   C2-H2
```

```
   8   0.319339       0.000   C2-H2
   9   0.319339       0.000   C2-H2
                 [...]
  13   0.330646       0.000   C1-H1
  14   0.270664       0.000   C1-C2-C2
  15   0.132040       0.000   C2-C1-H1
                 [...]
  37   0.119343       0.000   H1-C1-H1
  38  -0.000750       1.000   C1-C2-C2-C1_n=0
  39   0.001676       1.000   C1-C2-C2-C1_n=1
  40   0.000672       1.000   C1-C2-C2-C1_n=2
  41   0.002879       1.000   C1-C2-C2-C1_n=3
  42   0.000059       1.000   C1-C2-C2-C1_n=4
  43   0.000820       1.000   H1-C1-C2-C2_n=3
                 [...]
  48   0.000820       1.000   H1-C1-C2-C2_n=3
  TOTAL STANDARD DEVIATION 0.189D-04   atomic units
```

Once they have been determined, all the best parameters (*i.e.* force constants and equilibrium coordinates) are also printed in a **output topology** file. This file is named by default *joyce.new.top*, and as already mentioned is directly printed in the GROMACS *.top* format. **This file is the main JOYCE output and it can be used by the GROMACS engine [17] to start a MD simulation on the target molecule**. When the fitting procedure is ended, the above mentioned information is also printed in the JOYCE output as follows:

```
RESULTS OF THE LINEAR FITTING (kJ/mol L=angst)
     exact     computed   residue    Chisq
1    0.0029     0.0034   -0.0004   0.00000   hes 0  1  1     0.0271
2    0.0000     0.0002   -0.0002   0.00000   hes 0  2  1     0.0135
3    0.0101     0.0113   -0.0011   0.00000   hes 0  2  2     0.0271
4    0.0000     0.0000   -0.0000   0.00000   hes 0  3  1     0.0135
5    0.0000    -0.0000    0.0000   0.00000   hes 0  3  2     0.0135
                            [...]
665     0.0000    -0.0000    0.0000    0.00018   hes 0 36 35     0.0135
666     1.8658     1.8734   -0.0076    0.00018   hes 0 36 36     0.0271
667    24.0601    23.8180    0.2420    0.00039   Energy- 1       0.0360
668    14.1016    14.5547   -0.4532    0.00113   Energy- 2       0.0360
669     3.9908     3.6221    0.3687    0.00162   Energy- 3       0.0360
670     8.0787     8.3319   -0.2533    0.00185   Energy- 4       0.0360
671    14.4770    14.3392    0.1378    0.00192   Energy- 5       0.0360
                       [...]
679     6.1122     6.3630   -0.2508    0.00258   Energy-13       0.0360
680     0.0026    -0.0972    0.0998    0.00262   Energy-14       0.0360

TOTAL STANDARD DEVIATION 5.118D-02  kJ/mol L=angst
```

The above frame shows the results obtained for the n-butane molecule, where the standard deviation with respect to JOYCE's objective function (35) is reported together with a detailed list of the fitting results. With respect to the notation of equation (35), the point number is reported in the first column, while in the sixth and seventh column the point type (energy, gradient or Hessian) and the g geometry in which is computed are reported. In the next two columns (8^{th} and 9^{th}) the K,L components are specified, while the given weights (W_g, W'_{Kg} and W''_{KLg}) are shown in the last column in their normalized value. The second to the fifth column report respectively the QM (*exact*) energy (or gradient or Hessian), its FF value (*computed*), the difference (*residue*) between the formers and the partial mean square deviation (*Chisq*). Note that, starting form the JOYCE3.0 release, these results can also be saved in the graphical *.agr* format. Futher information can be found at the JOYCE website. [12]

After the fitting procedure is complete, the program computes the QMD-FF mass weighted Hessian matrix, its eigenvalues and the resulting frequencies based on the optimized force constants and the selected RIC. Next, it prints the data obtained according to the same format previuoly employed for the QM Hessian and frequencies, as appears from the next frame.

```
=====================================================
N O R M A L   V I B R A T I O N A L   M O D E S
by the optim. FF (EQUIL GEOMETRY)
=====================================================

Compute the Mass weighted Hessian
Diagonalize the Mass weighted Hessian

Eigenvalues of the (M-1/2)*F*M(-1/2) matrix (mH) by the optim. FF (EQUIL GEOMETRY)
1 0.2312E-03      2 0.1031E-02      3 0.1271E-02      4 0.2473E-02      5 0.3947E-02
6 0.1045E-01      7 0.1399E-01      8 0.1449E-01      9 0.1664E-01     10 0.1793E-01
11 0.1916E-01     12 0.1943E-01     13 0.2167E-01     14 0.2414E-01     15 0.3255E-01
16 0.3306E-01     17 0.3905E-01     18 0.4184E-01     19 0.4193E-01     20 0.4245E-01
21 0.4384E-01     22 0.4401E-01     23 0.4416E-01     24 0.4603E-01     25 0.4683E-01
26 0.5164E-01     27 0.1851         28 0.1861        29 0.1870         30 0.1873
31 0.1932         32 0.1938         33 0.2005         34 0.2006         35 0.2006
36 0.2007

Frequencies in 1/cm   by the optim. FF (EQUIL GEOMETRY)
1    105.525     2    222.870     3    247.413     4    345.160     5    436.034
6    709.555     7    820.973     8    835.406     9    895.274    10    929.307
11    960.774    12    967.534    13   1021.706    14   1078.266    15   1252.237
16   1261.962    17   1371.433    18   1419.699    19   1421.226    20   1429.891
21   1453.150    22   1455.915    23   1458.527    24   1489.027    25   1501.971
26   1577.162    27   2986.119    28   2993.868    29   3001.550    30   3004.034
31   3050.301    32   3055.482    33   3107.927    34   3108.391    35   3108.760
36   3109.380
```

Similarly to the QM normal analysis carried out in the previous sections, the QMD-FF normal modes projection over the RICS is printed in the same format used for their QM counterparts, as briefly shown in the next frame.

```
 VIBRATIONAL NORMAL MODES: Int.Coord. displac.s by the optim. FF (EQUIL GEOMETRY)
                105.5    222.9    247.4    345.2    436.0   ...
                  1        2        3        4        5     ...
  1 C1-C2       0.000    0.000    0.000    0.003   -0.033   ...
  2 C2-C2       0.000    0.000    0.000    0.000   -0.037   ...
  3 C2-C1       0.000    0.000    0.000   -0.003   -0.033   ...
 14 C1-C2-C2    0.000    0.000    0.000   -0.066   -0.043   ...
 15 C2-C2-C1    0.000    0.000    0.000    0.066   -0.043   ...
 16 C2-C1-H1    0.000    0.000    0.000    0.016    0.003   ...
                                                      [...]

                3050.3   3055.5   3107.9   3108.4   3108.8   3109.4
                 31       32       33       34       35       36
  4 C1-H1       0.000    0.000    0.000    0.000   -0.383    0.036
  5 C1-H1      -0.012    0.010   -0.392   -0.101    0.193   -0.018
  6 C1-H1       0.012   -0.010    0.392    0.101    0.193   -0.018
  7 C1-H1       0.000    0.000    0.000    0.000    0.039    0.357
  8 C1-H1      -0.008   -0.009    0.110   -0.360   -0.019   -0.179
  9 C1-H1       0.008    0.009   -0.110    0.360   -0.019   -0.179
                                                      [...]
```

Finally, a brief comparison between QM and FF computed vibrational frequencies is printed, as shown in the next frame. It is worth noticing that the indicated standard deviation is only for comparison purposes, as the fitting target is the minimization of equation (35). Note also that the comparison between QM and FF frequencies is done on the base of the overlap of the corresponding normal modes (QM and FF), and not on the nearest value of the frequency itself.

```
==================================================
      Compare Norm Modes from QM and FF
==================================================

iFF Freq/FF   ig03 freq/g03       overl     err
35   3108.8    36   3096.1        0.765     12.6
36   3109.4    35   3095.4        0.764     14.0
34   3108.4    34   3090.5        0.806     17.9
33   3107.9    33   3087.0        0.851     20.9
31   3050.3    32   3046.8        0.922      3.5
32   3055.5    31   3024.1        0.975     31.4
30   3004.0    30   3021.4        0.847    -17.3
29   3001.5    29   3019.9        0.866    -18.4
27   2986.1    28   3008.7        0.913    -22.6
28   2993.9    27   3002.0        0.910     -8.1
23   1458.5    26   1484.8        0.656    -26.3
18   1419.7    25   1478.3        0.668    -58.6
21   1453.2    24   1470.6        0.983    -17.4
22   1455.9    23   1468.8        0.991    -12.9
19   1421.2    22   1463.1        0.674    -41.8
24   1489.0    21   1459.9        0.842     29.1
25   1502.0    20   1399.0        0.686    102.9
```

```
       26    1577.2    19    1398.0        0.779    179.2
       20    1429.9    18    1387.4        0.693     42.5
       15    1252.2    17    1320.7        0.947    -68.5
       17    1371.4    16    1310.5        0.934     60.9
       16    1262.0    15    1277.5        0.852    -15.5
       13    1021.7    14    1197.1        0.986   -175.4
       14    1078.3    13    1163.3        0.977    -85.0
       10     929.3    12    1074.5        0.925   -145.2
        9     895.3    11    1024.4        0.794   -129.2
       11     960.8    10     976.4        0.785    -15.7
       12     967.5     9     955.1        0.888     12.5
        7     821.0     8     844.8        0.923    -23.9
        8     835.4     7     809.4        0.967     26.0
        6     709.6     6     741.8        0.984    -32.2
        5     436.0     5     424.7        0.978     11.3
        3     247.4     4     267.9        0.995    -20.5
        4     345.2     3     257.7        1.000     87.4
        2     222.9     2     228.4        0.991     -5.5
        1     105.5     1     123.2        0.998    -17.7
                 Standard deviation (cm-1)       64.68
```

Note that if the *$gracefreq* keyword has been activated in the input file, the above results are also printed in an output file in the *.agr* format for a straightforward visualization. Eventually, if the fitting was successfully performed, JOYCElog file should always end with a *NORMAL EXIT* sentence.

6.2 Generated IC file

As already discussed, the *generated.IC.txt* contains a list of all NIC, automatically retrieved by JOYCE from the QM optimized geometry Its format is identical to the intramolecular section of a *.top* file, whose details are given in Section 5.3. **Note that the *generated.IC.txt* can be edited and modified by the user**, and thereafter combined with the intermolecular part to be used for subsequent fittings.

6.3 Suggested dependencies file

All dependencies found by JOYCE, either on the base of the given labels or on the redundancy of the employed functions, are written in a file named *suggdeps.txt* in the following format:

```
$dependence 1.2
      3 =     1*1.d0    ; C1-C2                 = C1-C2
      5 =     4*1.d0    ; C1-H1                 = C1-H1
      6 =     4*1.d0    ; C1-H1                 = C1-H1
     11 =     4*1.d0    ; C1-H1                 = C1-H1
     12 =     4*1.d0    ; C1-H1                 = C1-H1
     13 =     4*1.d0    ; C1-H1                 = C1-H1
      8 =     7*1.d0    ; C2-H2                 = C2-H2
      9 =     7*1.d0    ; C2-H2                 = C2-H2
     10 =     7*1.d0    ; C2-H2                 = C2-H2
```

```
    20 =   14*1.d0   ; C1-C2-C2           = C1-C2-C2
    16 =   15*1.d0   ; C2-C1-H1           = C2-C1-H1
    17 =   15*1.d0   ; C2-C1-H1           = C2-C1-H1
                       [...]
    37 =   30*1.d0   ; H1-C1-H1           = H1-C1-H1
    34 =   33*1.d0   ; H2-C2-H2           = H2-C2-H2
    44 =   43*1.d0   ; H1-C1-C2-C2_n=3    = H1-C1-C2-C2_n=3
    45 =   43*1.d0   ; H1-C1-C2-C2_n=3    = H1-C1-C2-C2_n=3
    46 =   43*1.d0   ; H1-C1-C2-C2_n=3    = H1-C1-C2-C2_n=3
    47 =   43*1.d0   ; H1-C1-C2-C2_n=3    = H1-C1-C2-C2_n=3
    48 =   43*1.d0   ; H1-C1-C2-C2_n=3    = H1-C1-C2-C2_n=3
```

Note that this file can be straightforwardly **copied and pasted into a JOYCEmain input file**, to refine a previous fitting performed with incomplete dependencies. In the above framed sample file, for instance, in the first line IC number 3 is set as dependent on number 1. The reason for doing this is suggested by JOYCE and automatically written: the distances *C2-C1* (IC number 3) and *C1-C2* (IC number 1) should be described by the same FF parameters for their symmetry.

6.4 Force constants assign file

All the optimized parameters at the end of a JOYCE run are also written in a text file named *assign.dat.* As for the suggested dependencies file, also the content of *assign.dat* can also be copied and pasted into a new JOYCE input file. **Note that If this is done without any previous editing, all the force constants will be constrained to their previously optimized values, and no fitting will be performed in the second run.** On the contrary, if the constraints on some IC are removed from the *assign.dat* file before pasting it into a new input, the new fitting will be performed only on those IC removed from the list, keeping all the force constants found in the *$assign* environment fixed to the values optimized in the previous fitting.

```
$assign
 1 =   2254.5152345     C1-C2
 2 =   2130.5117241     C2-C2
 3 =   2254.5152345     C1-C2
 4 =   3100.0797022     C1-H1
          [...]
37 =    313.3362781     H1-C1-H1
38 =     -1.9686612     C1-C2-C2-C1_n=0
39 =      4.3993865     C1-C2-C2-C1_n=1
40 =      1.7642241     C1-C2-C2-C1_n=2
41 =      7.5582350     C1-C2-C2-C1_n=3
42 =      0.1556409     C1-C2-C2-C1_n=4
43 =      2.1534692     H1-C1-C2-C2_n=3
44 =      2.1534692     H1-C1-C2-C2_n=3
45 =      2.1534692     H1-C1-C2-C2_n=3
46 =      2.1534692     H1-C1-C2-C2_n=3
47 =      2.1534692     H1-C1-C2-C2_n=3
48 =      2.1534692     H1-C1-C2-C2_n=3
$end
```

Although this procedure may seem a bit cumbersome, **it can be very useful in the case of a well-different set of coordinates, such as stiff and soft IC**. In this case it may be useful to fit in a first run all IC, only on the base of the optimized geometry and Hessian. Since the information used is insufficient to accurately characterized soft coordinates, a second run is necessary. In this second run, QM scans on soft IC are also read by Joyce. Thus, to simplify the fitting procedure, the second run (*i.e.* the one with the QM scan data) can be performed keeping all stiff constants constrained to the values optimized in the first run. This is done by pasting into the second input file the *assign.dat* produce by Joyce in the first run, after removing all soft IC from the list therein. SG: Please refer to a specific example on the website, so that people can go and look it up It is worth noticing that, despite Joyce adds at the end of the assigned list also the non-bonded intramolecular constants (ϵ_{ij}, σij and the product $f_{ij} * q_i * qj$), **the LJ parameters and the atomic charges are not optimized by the default settings of the program**. These should be considered as assigned parameters, even if they do not appear in the assigned list. The ϵ_{ij} parameters can be varied in the fitting procedure if the *fitLJ* key is activated in the Joyceinput. In this case, those ϵ_{ij} that the user wants to keep constrained, should be explicitly listed in the *assign* environment.

6.5 QM and FF scan files

When soft ICs are present among the selected RICs, Fourier-like potential should be employed and energy scans along such coordinates are required. This can be done by activating the *$geom* key (see section 5.1). In these cases, Joyce automatically computes the energy profiles along each of the given coordinates (from the given list of the files containing the QM training information). The energies are printed out in text files, named *qmscan.XX.dat* (where XX is the number of the scanned soft IC), with the following format:

```
#  coordinate   DE (kJ/mol)
      0.000        24.0431
     30.000        14.0819
     60.000         3.9559
     90.000         8.0613
    120.000        14.4950
    150.000         6.8780
    180.000         0.0011
```

In the above framed sample file the scanned RIC (a dihedral angle in this case) is reported in the first column, whereas in the second column is printed the difference (ΔE in kJ/mol) between the energy of the scanned geometry and the absolute energy minimum (taken from the QM optimized geometry indicated in the *$equil* keyword).

Additionally, if the *scan* keyword is activated in the input file (see section 5.1), an energy profile scan, along the same RIC scanned with QM methods, will be performed by Joyce with the optimized FF. The results is printed into a text file, easily plotted by any graphic software, whose format is reported in the following.

```
# RESULTS OF JOYCE SCAN
#  38  C1-C2-C2-C1      -180.000  180.000
#  first  row =C1-C2-C2-C1
    -180.000      -0.0308
    -179.000      -0.0216
    -178.000       0.0061
    -177.000       0.0521
    -176.000       0.1165
    -175.000       0.1989
             [...]
     175.000       0.1989
     176.000       0.1165
     177.000       0.0521
     178.000       0.0061
     179.000      -0.0216
     180.000      -0.0308
```

To check how the parameterized FF performs along such energy profiles, the QM and FF scans can be compared by plotting them into the same graph, as discussed in the following sections.

6.6 Frequency plots and energy scans

Starting from the JOYCE3.0 release, two additional output files may be created through a JOYCE run. Concretely, if the *$gracefreq* and *$gracetors* are activated in the main input file, the results of the vibrational analysis and the torsional energy scans will be plotted in the *.agr* format, suitable for the XMGRACE graphical interface. [25] Further info and examples can be found at the website. [12]

7 Theory and methods

7.1 Force-fields

7.1.1 Introduction

The total energy (E_{tot}) of a system of M_{mol} molecules in classical simulation is usually [5, 6] computed as a sum of two contributions, namely

$$E_{tot} = E_{inter} + E_{intra}^{M_{mol}} \tag{1}$$

where E_{inter} and $E_{intra}^{M_{mol}}$ are the interaction energy among different molecules and the sum of the internal energy of each molecule. In standard FFs, E_{inter} is computed as a sum of pairwise contributions among all the N_{sites} interaction sites used to model the system. In particular,

$$E_{inter} = E_{LJ} + E_{Coul} \tag{2}$$

where the long-range electrostatic term is

$$E_{Coul} = \sum_{i=1}^{N_{sites}} \sum_{j=1}^{N_{sites}} \frac{q_i q_j}{r_{ij}} \tag{3}$$

whereas the short range 12-6 Lennard-Jones term is

$$E_{LJ} = \sum_{i=1}^{N_{sites}} \sum_{j=1}^{N_{sites}} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{4}$$

where i and j are interaction sites **belonging to a pair of different molecules**.

The total intramolecular term $E_{intra}^{M_{mol}}$ is the sum of the molecular internal energies (E^{intra}) of all the M_{mol} molecules composing the system, *i.e.*

$$E_{intra}^{M_{mol}} = \sum_{K=1}^{M_{mol}} V_K^{intra} \tag{5}$$

where V_K^{intra} **is the internal energy of molecule** K.
Given a model potential function V_K^{intra}, the main goal of the JOYCE program is to find, with respect of reference QM computed data, the best parameters to represent the intramolecular energy for a chosen target molecule K, hence parameterizing the intra-molecular term of a QMD-FF..

7.1.2 Internal Coordinates

Broadly speaking, the FF model intramolecular potential V_K^{intra} is a function of a set $\{Q_1, Q_2, ..., Q_{N_K}\}$ of generalized (nuclear) internal coordinates (IC) describing target molecule K. Within the JOYCEprocedure, the selection of this set is either automatically performed by the program or defined by the user, once the model (*i.e.* the definition of all interaction sites) mimicking the target molecule has been chosen. It is worth noticing that the selected IC are not necessarily required to be linearly independent, but redundant ($N_K > 3N_{sites}$ - 6) set of internal coordinates (RIC) can also be employed.

As an example, let's consider the n-butane molecule. For the sake of clarity, let's also suppose to adopt an united atom (UA) representation, where all Hydrogen atoms are grouped in an unique interaction site with the Carbon atom bonded to them. Within this model, sketched in Figure 2, the model n-butane molecule is composed by $N_{sites} = 4$ interaction sites, whose geometry is completely described by ($3N_{sites}$ - 6)=6 IC. A natural choice for the minimal set of IC is reported in Figure 3, where the three bond lengths (R$_{1-2}$, R$_{2-3}$ and R$_{3-4}$), two bond angles (θ_{1-2-3} and θ_{1-2-3}) and the torsional dihedral ($\delta_{1-2-3-4}$) are evidenced with different colors. Among these IC, it is convenient to distinguish between "stiff" and "soft" degrees of freedom, [1, 20, 30] highlithed in Figure 3 with reddish and blueish colors, respectively. Bond

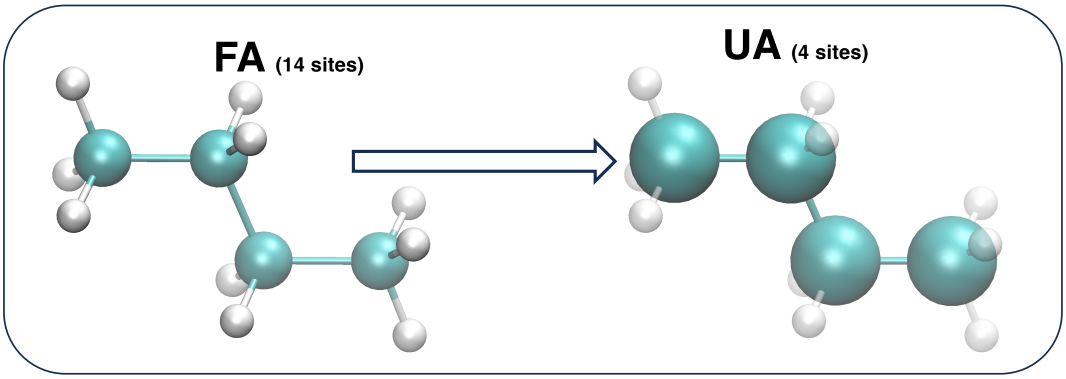

Figure 2: UA model of the n-butane molecule. The number of interaction sites is reduced to four, which are completely described by six natural internal coordinates.

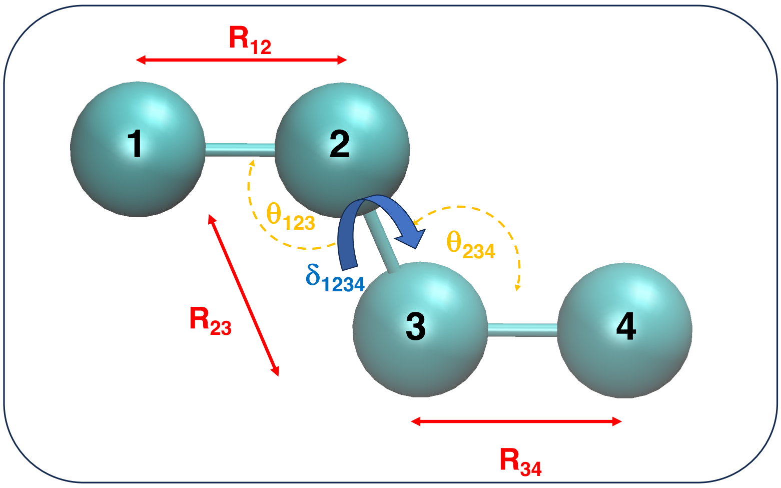

Figure 3: 6 natural IC for n-butane molecule: three bond lengths (R$_{12}$, R$_{23}$ and R$_{34}$), two bond angles (θ_{123} and θ_{123}) and the torsional dihedral (δ_{1234}) are evidenced by red, orange and blue lines, respectively.

lengths and angles are intrinsically connected to stretching and bending motions, which are relatively high energy motions, usually characterized by small displacements from their equilibrium values. These type of coordinates will be classified as stiff. Conversely the rotation around the central bond, can be considered as

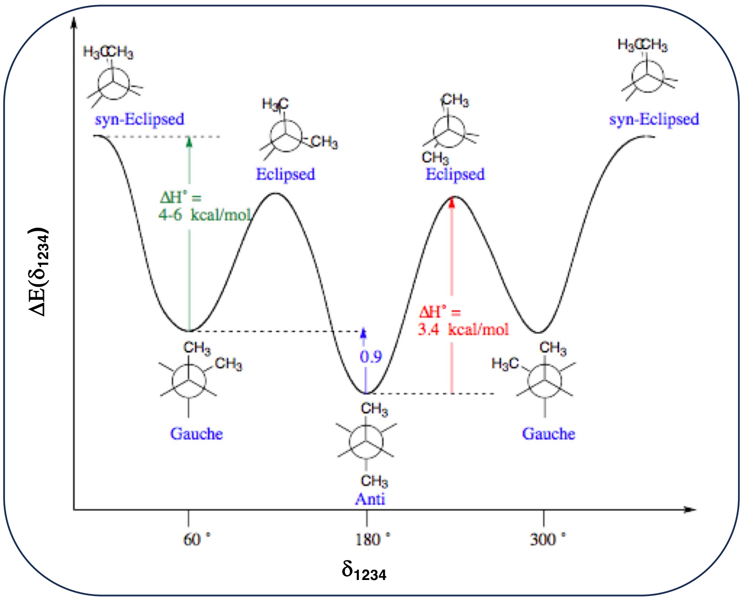

Figure 4: Torsional energy profile for the n-butane molecule.

a soft coordinate, since the internal energy profile, as a function of δ, is characterized (see Figure 4) by the presence of local minima, separated by relatively low energy barriers. This coordinate can be subjected to large amplitude oscillations, eventually leading to the population of different *trans* and *gauche* minima, even at room temperature [1]).

Another type of soft coordinate is the intramolecular distance between a pair of atoms connected by more than two bonds. Despite internal distances are intrinsically redundant (in that they can be expressed as a function of bond lengths, angles and dihedrals defining the relative position of the two atoms defining them), they turn out to be very useful in the case of large molecules, when important interactions between two different molecular regions take place after some geometrical rearrangement. Conversely, their use for smaller molecules does not seem to yield any particular advantage. As an example let's consider two different molecules, namely the PMME-H molecule and n-butane. In the latter case, the R_{14} intramolecular distance could be chosen as an adjunctive generalized coordinate. As shown in Figure 5, this distance is strongly dependent on the torsional dihedral δ: it reaches its maximum value in the *trans* ($\delta = 180°$) conformation, its minimum in the totally eclipsed one ($\delta = 0°$), and intermediate values in the *gauche* conformers ($\delta = \pm 60°$). Notwithstanding is undoubtable utility in rationalizing the origin of the three butane local minima, this RIC can be however neglected in the set defining the FF, as the torsional potential can be easily described by the combination of periodic functions only dependent on the dihedral.

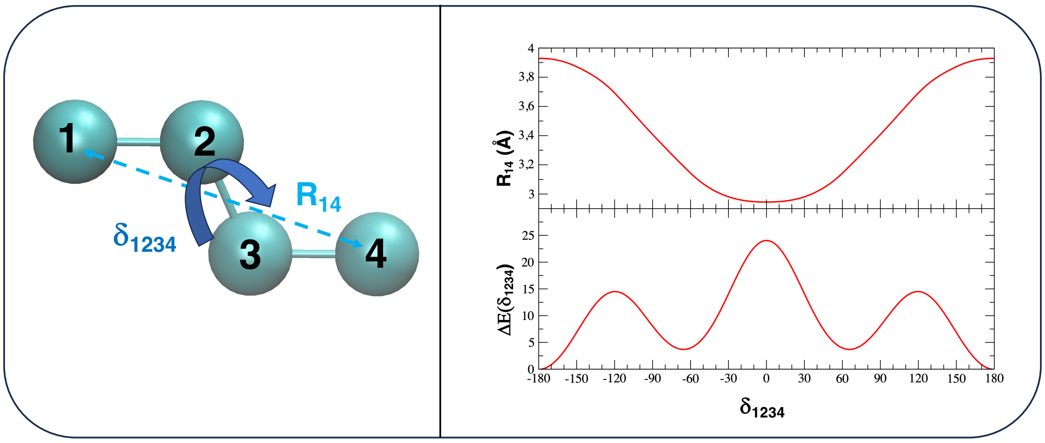

Figure 5: Torsional energy profile and 1-4 internal distance as a function of torsional dihedral δ for the *n*-butane molecule.

A different situation arises for a slightly larger molecule, the PMME-H, shown in Figure 6. One of the lowest energy conformers is characterized by an intramolecular hydrogen bond, evidenced in Figure 6 with a dotted line, between the H atom of the carboxyl group and the Oxygen of the neighboring carboxylate. As both of these groups can easily change their relative orientation by varying (at least) either δ and χ dihedrals, the internal O-H distance depends on these soft ICs in rather complex fashion. In such cases, despite not necessary, it is convenient to include intramolecular distances in the RIC set defining the FF.

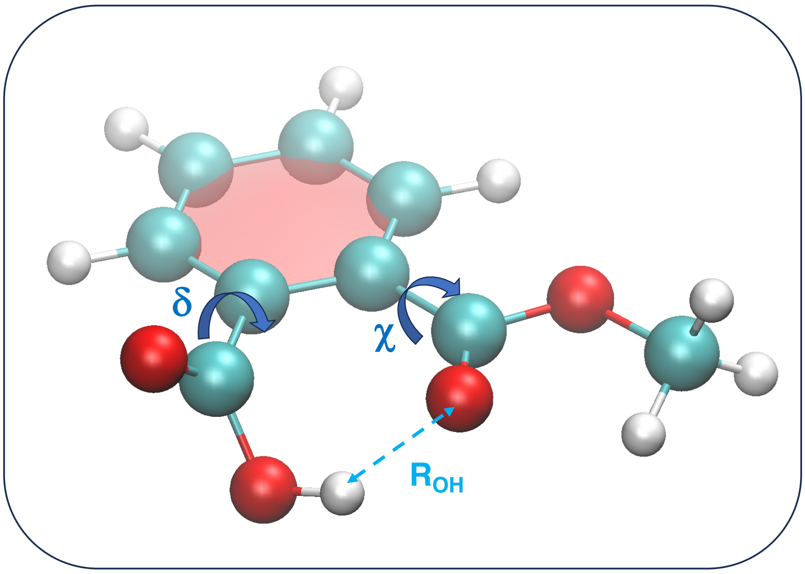

Figure 6: Some soft IC characterizing the PMME-H molecule.

7.1.3 Force-field model potential functions

Once the set of RIC to be employed has been chosen, the analytical form of the model potential energy functions depending on the selected RIC must be declared. In the JOYCEcode, the intramolecular FF is thus expressed as a sum of different terms, namely

$$V^{intra} = E_{stretch} + E_{bend} + E_{Rtors} + E_{Ftors} + E_{nb}^{intra} + E_{Coupl} \tag{6}$$

The first three terms count for the stiff IC, *i.e.* bond stretchings, angle bendings and stiff angle dihedrals (Rdihedrals), as those that drive the planarity of aromatic rings and are expressed with harmonic potentials:

$$E_{stretch} = \frac{1}{2} \sum_{\mu}^{N_{bonds}} k_{\mu}^{s} (r_{\mu} - r_{\mu}^{0})^2 \tag{7}$$

$$E_{bend} = \frac{1}{2} \sum_{\mu}^{N_{angles}} k_{\mu}^{b} (\theta_{\mu} - \theta_{\mu}^{0})^2 \tag{8}$$

$$E_{Rtors} = \frac{1}{2} \sum_{\mu}^{N_{Rdihedrals}} k_{\mu}^{t} (\phi_{\mu} - \phi_{\mu}^{0})^2 \tag{9}$$

Conversely, the model functions employed for soft, flexible dihedrals (Fdihedrals) are sums of periodic functions, namely

$$E_{Ftors} = \sum_{\mu}^{N_{Fdihedrals}} \sum_{j=1}^{N_{cos_{\mu}}} k_{j\mu}^{d} \left[1 + cos(n_{j}^{\mu} \delta_{\mu} - \gamma_{j}^{\mu})\right] \tag{10}$$

where $N_{cos_{\mu}}$ is the number of cosine function employed to describe the potential of the δ_{μ} dihedral. Finally, if any internal distance between atoms not directly bonded to each other is defined, the non-bonded intramolecular contribution is computed as

$$\begin{aligned} E_{nb}^{intra} &= \sum_{i=1}^{N_{nb}} \sum_{j=i+1}^{N_{nb}} 4\epsilon_{ij}^{intra} \left[\left(\frac{\sigma_{ij}^{intra}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}^{intra}}{r_{ij}} \right)^{6} \right] \\ &+ \sum_{i=1}^{N_{nb}} \sum_{j=i+1}^{N_{nb}} \frac{q_i q_j}{r_{ij}} \end{aligned} \tag{11}$$

Note that, at difference with intermolecular energy terms reported in equations (3) and (4), i **and** j **indexes run over atoms of the same molecule, therefore** $j \neq i$**, to avoid self interaction. Furthermore, the interaction parameters** ϵ_{ij}^{intra} **and** σ_{ij}^{intra} **do not have to be necessarily the same employed in the intermolecular LJ interaction reported in equation (4), although some MD software do not allow to use different LJ parameters for intra and inter-molecular interactions**.

Memo: add here some comments once the LAMMPS & AMBER issues have been solved or settled once for all

All of the aforementioned FF terms are expressed as sums of contribution, each depending on a single IC (even if, as before stated, r_{ij} can always be expressed as a function of other generalized IC's). If the last term of equation (6) is not included, such FF is often termed as diagonal.

As a matter of fact the E_{Coupl} term, takes explicitly into account the coupling of two (or more) RIC. Since 2018, [20] the JOYCEcode has been equipped to handle two-variable functions designed to explicitly account for the coupling between the considered ICs. As the detailed in the original paper, [20] a more convenient formalism can be adopted to clarify the coupling functions that can be used within the JOYCEparameterization procedure. In fact, when the couling term E_{Coupl} is missing, the intramolecular energy of a purely diagonal FF can be re-written as

$$V^{intra}(\mathbf{q}) = \sum_{a=1}^{N_{func}} p_a f_a(q_\mu) \tag{12}$$

where q_μ stands for any internal coordinate $r_\mu, \theta_\mu, \phi_\mu, \delta_\mu$ reported in equations (7)-(10), p_a refers to the force constant (either k^s_μ, k^b_μ, ...), while f_a is the diagonal potential function ($[r_\mu - r^0_\mu]^2$, $[\theta_\mu - \theta^0_\mu]^2$, $[1+cos(n_j\delta_\mu - \gamma_j)]$, etc.) assigned to q_μ. Within this framework, the generalized coupling introduced in JOYCEconsists in a sum of N_{Coupl} pairwise linear terms, defined as the product between two functions, f and g, each depending only on one of the coupled ICs:

$$E_{Coupl} = \sum_{a}^{N_{Coupl}} p_a f_a(q_\mu) g_a(q_\nu) \tag{13}$$

Following this notation, the coupling terms between two stiff ICs, are expressed by a product of linear terms involving each considered coordinate, as for instance:

$$E^{str-str}_{\mu\nu}(r_\nu, r_\mu) = k^{ss}_{\mu\nu}(r_\mu - r^0_\mu)(r_\nu - r^0_\nu) \tag{14}$$

$$E^{bnd-bnd}_{\mu\nu}(\theta_\mu, \theta_\nu) = k^{bb}_{\mu\nu}(\theta_\mu - \theta^0_\mu)(\theta_\nu - \theta^0_\nu) \tag{15}$$

$$E^{str-bnd}_{\mu\nu}(r_\mu, \theta_\nu) = k^{sb}_{\mu\nu}(r_\mu - r^0_\mu)(\theta_\nu - \theta^0_\nu) \tag{16}$$

$$E^{str-Rtors}_{\mu\nu}(r_\mu, \phi_\nu) = k^{srt}_{\mu\nu}(r_\mu - r^0_\mu)(\phi_\nu - \phi^0_\nu) \tag{17}$$

While similar expressions can be easily derived for all couplings involving stiff ICs, for the sake of clarity, we describe in more detail the effect of coupling a stiff coordinate with a soft one, as for instance a selected dihedral angle $\delta\prime$ and a neighboring stretching distance, r_1. If we adopt a linear form for the term related to the stiff coordinate, $f(r_1) = (r_1 - r^0_1)$, and merge the coefficient and the function related to the soft coordinate into a generic function $G(\delta\prime)$, equation (13) leads to

$$E_{Coupl} \equiv E^{Coupl}(r_1, \delta\prime) = (r_1 - r^0_1)G(\delta\prime) \tag{18}$$

The shape of $G(\delta\prime)$ should be flexible enough so as to reproduce the reference profiles obtained by QM calculations. In this sense, a Fourier series can again provide the required flexibility, together with other desirable features. [20] All requisites can be verified using the functional form:

$$G(\delta\prime) = \sum_i k_i^c[1 + \sin(n_i\delta\prime - \gamma_i^c)] \tag{19}$$

which leads, to the following explicit coupling terms:

$$E_{\mu\nu}^{str-tors}(r_\mu, \delta_\nu) = \sum_{j=1}^{N_{sin_\nu^i}} k_{j\mu\nu}^{st}(r_\mu - r_\mu^0)(sin(n_j^\nu\delta_\nu - \gamma_j^\nu) \tag{20}$$

$$E_{\mu\nu}^{bnd-tors}(\theta_\mu, \delta_\nu) = \sum_{j=1}^{N_{sin_\nu^i}} k_{j\mu\nu}^{bt}(\theta_\mu - \theta_\mu^0)(sin(n_j^\nu\delta_\nu - \gamma_j^\nu) \tag{21}$$

$$E_{\mu\nu}^{Rtors-tors}(\phi_\mu, \delta_\nu) = \sum_{j=1}^{N_{sin_\nu^i}} k_{j\mu\nu}^{rtt}(\phi_\mu - \phi_\mu^0)(sin(n_j^\nu\delta_\nu - \gamma_j^\nu) \tag{22}$$

$$E_{\mu,\nu}^{tors-tors}(\delta_\mu, \delta_\nu) = \sum_{j=1}^{N_{sin_\mu^i}} \sum_{k=1}^{N_{sin_\nu^i}} k_{jk\mu\nu}^{tc} sin(n_j^\mu\delta_\mu - \gamma_j^i) sin(m_k^\nu\delta_\nu - \gamma_k^i) \tag{23}$$

7.2 Parameterization Procedure

7.2.1 Notation

To make the formulae easier to be understood, the following notation will be adopted for the summation indexes and symbols

i, j are used for the Cartesian coordinates (CCs) x or mass weighted Cartesian coordinates $(1 \div 3N)$

μ, ν indicate the redundant internal coordinates [31, 32] (RICs) $q \quad (1 \div N_{RIC})$

K, L run over the normal coordinates (NCs) Q $(1 \div 3N - 6)$ $(3N - 5$ for linear molecules)

g run over the considered molecular geometries $(0 \div N_g)$

a, b indicate the functions f used to represent the empirical FF and/or the number of linear parameters of the FF $(1 \div N_{func})$

s, t run over the quantities to be represented by the FF (energies, energy gradients and Hessian) for the considered geometries $(1 \div N_{points})$

The target FF, to be used in molecular dynamics or molecular mechanics, is expressed through the linear combination of functions f_a of a set of RICs as in equation (12), with or without the coupling term (13). The functions $f(q)$ entering these equations may conveniently be expressed in terms of displacements with respect to a given reference geometrical conformation identified by the vector q^0

$$\Delta q_\mu = q_\mu - q^0_\mu \tag{24}$$

Usually the RICs consist in all bond stretches, angle bendings and dihedral torsions that can be obtained from a given connectivity criteria referred to the reference conformation. The inversion coordinate [33] can be included for atoms bonded to three other atoms. Non-bonded intramolecular interactions can also be added in order to make the FF more accurate. In usual FFs the number of RICs exceeds $3N$-6 and therefore they form a redundant set of coordinates. Although equation (6) has been written in general form, each function f_a only depends on one or two RICs, as reported in detail n equations (7)–(23).

7.2.2 Internal coordinates transformations

Since the Hessian and gradients are computed in CCs, whereas the FF is usually expressed through RICs, some coordinate transformation is required. For infinitesimal displacements with respect to a given geometrical conformation, the RICs are related to the nuclear CCs x through a non invertible transformation

$$\delta q = B \ \delta x \tag{25}$$

where δq and δx are colum vectors. The Wilson rectangular B matrix

$$B_{\mu i} = \left(\frac{\partial q_\mu}{\partial x_i} \right) \tag{26}$$

is related to the geometry the displacements are referred to, and can be accurately computed both in analytical [34] and numerical ways.

The normal coordinates are computed from the Hessian matrix in CCs

$$H_{ij} = \left(\frac{\partial^2 E}{\partial x_i \partial x_j} \right) = E''_{ij} \tag{27}$$

obtained by a QM calculation at a given geometry. H is transformed to the mass weighted CCs form and diagonalized by a unitary matrix C

$$M^{-1/2}\ H\ M^{-1/2} C = C\Lambda \tag{28}$$

The matrix M is diagonal and for each CC contains the mass m of the related atom. The columns of the C matrix are the linear combinations of the mass weighted CCs that correspond to the NCs displacements

$$\delta Q_K = \sum_{i=1}^{3N} \sqrt{m_i} C_{iK} \delta x_i \tag{29}$$

or in matrix form

$$\delta Q = \widetilde{C}\ M^{1/2} \delta x \tag{30}$$

where δQ and δx are column vectors. In the case the geometry corresponds to an absolute or local energy minimum, $3N - 6$ eigenvalues Λ_K are positive and refer to vibrations, whereas the 3 translational and 3 rotational modes are identified by zero eigenvalues. In other cases negative eigenvalues can occur and these do not correspond to vibrational modes. If all the NCs are retained, the transformation of equation (29) is fully invertible

$$\delta x = M^{-1/2} C \delta Q \tag{31}$$

The relation between the RICs and the NCs can be easily obtained exploiting the completeness of the CCs basis set. Using equations (25) and (31)

$$\delta q = B M^{-1/2}\ C \delta Q = T \delta Q \tag{32}$$

where the T matrix is defined as

$$T_{\mu K} = \left(\frac{\partial q_\mu}{\partial Q_K} \right) \tag{33}$$

Thus the RICs may be expressed in terms of the NCs and the inclusion or not of the rotational and translational modes is uninfluential since they leave the RICs unchanged.

7.2.3 The optimal parameters of the Force Field

The best FF parameters in order to represent the internal molecular motion are obtained by minimizing the following objective function, written as a sum over the considered molecular geometries

$$I = \sum_{g=0}^{N_g} I_g \tag{34}$$

where

$$I_g = W_g \left[(E_g - E_0) - V_g\right]^2 + \sum_{K=1}^{3N-6} \frac{W'_{Kg}}{3N-6} \left[E'_{Kg} - V'_{Kg}\right]^2 + \sum_{K \le L}^{3N-6} \frac{2W''_{KLg}}{(3N-6)(3N-5)} \left[E''_{KLg} - V''_{KLg}\right]^2 \tag{35}$$

The indexes K, L (capital letters) run over the normal coordinates and include all the modes except for the rotational and translational ones. E_g is the total energy obtained by a QM calculation and E_0 is the same at the reference geometry ($g = 0$). E'_{Kg} (E''_{KLg}) is the energy gradient (Hessian) at a given geometry with respect to the NC evaluated at the same geometry. V, V' and V'' are the corresponding quantities calculated by the FF in equation (6). The constants W, W' and W'' weight the several terms at each geometry and can be chosen in order to drive the results depending on the circumstances. The energy, gradient and Hessian terms are normalized in order to account for the different number of terms and to make the weights independent from the number of atoms in the molecule.

To compute the energy derivatives entering the merit function (35) we have to perform some transformations since no derivative is originally expressed with respect to the NCs. Indeed standard quantum chemistry programs provide derivatives E' and E'' with respect to CCs. Using the above relations and exploiting the completeness of the CCs, the transformation is simple

$$E'_K = \left(\frac{\partial E}{\partial Q_K}\right) = \sum_{i=1}^{3N} \left(\frac{\partial E}{\partial x_i}\right)\left(\frac{\partial x_i}{\partial Q_K}\right) = \sum_{i=1}^{3N} E'_i \, m_i^{-1/2} \, C_{iK} \tag{36}$$

or, in matrix form

$$[E']_{NC} = \widetilde{C} M^{-1/2} [E']_{CC} \tag{37}$$

where the square parentheses indicates column vectors of energy gradients computed with respect to the NCs and the CCs. The FF energy gradients at a given geometry

$$V'_K = \sum_{a=1}^{N_{func}} p_a \left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{a=1}^{N_{func}} p_a \, f'_{aK} \tag{38}$$

can be conveniently computed using the derivatives of the basis function with respect to the RICs, that is

$$\left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{RIC}} \left(\frac{\partial f_a}{\partial q_\mu}\right)\left(\frac{\partial q_\mu}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{RIC}} \sum_{i=1}^{3N} \left(\frac{\partial f_a}{\partial q_\mu}\right) T_{\mu K} \tag{39}$$

or in matrix form

$$[f'_a]_{NC} = \widetilde{T} [f'_a]_{RIC} \tag{40}$$

The Hessian matrix of the QM calculation in NCs

$$E''_{KL} = \left(\frac{\partial^2 E}{\partial Q_K \partial Q_L}\right) \tag{41}$$

is obtained from the Hessian matrix in the CC basis according to

$$[E'']_{NC} = \widetilde{C} M^{-1/2} [E'']_{CC} M^{-1/2} C \tag{42}$$

The second derivatives of the FF are a bit more complicated since they involve derivatives of the B matrix and are conveniently expressed in explicit form

$$\left(\frac{\partial^2 f_a}{\partial Q_K \partial Q_L} \right) = \sum_{\mu\nu=1}^{N_{RIC}} T_{\mu K} \left(\frac{\partial^2 f_a}{\partial q_\mu \partial q_\nu} \right) T_{\nu L} + \sum_{\mu\nu=1}^{N_{RIC}} T_{\mu K} \left(\frac{\partial f_a}{\partial q_\nu} \right) \left(\frac{\partial T_{\nu L}}{\partial q_\mu} \right) \tag{43}$$

As shown in equation (6), the FF is linear in the p parameters, thus the least squares minimization of functional (35) can be written as

$$\sum_a^{N_{func}} \sum_s^{N_{point}} \alpha_{bs} \, W_s \, \alpha_{as} \, p_a = \sum_s^{N_{point}} \alpha_{bs} \, W_s \, \beta_s \tag{44}$$

where the index s runs over the collections $[g]$, $[Kg]$ and $[KLg]$ defined in equation (35) for energy, gradient and Hessian, respectively. Following this notation the matrix α and the vector β are defined as

$$\alpha_{as} = f_{as} \text{ or } f'_{as} \text{ or } f''_{as} \quad ; \quad \beta_s = E_s \text{ or } E'_s \text{ or } E''_s$$

and

$$W_s = W_s \text{ or } \frac{W'_s}{3N-6} \text{ or } \frac{W''_s}{(3N-6)(3N-5)}$$

where f's are the functions of equation (6), E, E', E'' the QM data and W, W', W'' the weights of the merit function (35). Thus, defining

$$A = \alpha W \widetilde{\alpha}$$
$$b = \alpha W \beta$$

one has to solve a standard linear equation in the form

$$Ap = b \tag{45}$$

where A is a symmetric matrix.

In usual FF it is convenient for practical purposes, to employ functions of the RIC that will be in general redundant over the considered points. The scalar product between the FF functions is defined as

$$f_a \cdot f_b = \sum_{s=1}^{N_{point}} W_s f_{as} \, f_{bs} \tag{46}$$

and the redundancy strongly depends on the number and type of points included in the fitting. However in general the f set might not be linearly independent. This leads to a singular A matrix and the direct inversion method can not be used to solve the linear system (45). On the contrary, the Singular Value Decomposition method [35, 36] adapted to symmetric matrices is adequate and provides a stable solution of the linear system.

7.2.4 United Atom Theory

In many molecular simulations a group of atoms whose individual behavior is considered not to be crucial for the properties to be investigated, can be grouped in a single interaction site. This approach, henceforth named United Atom (UA), allows for saving computational time and simultaneously removes some high frequency vibrational modes which can limit the integration time step in MD simulations. The most common example concerns aliphatic chains where each CH_2 group is treated as a single interaction site (C_2) with FF parameters accounting for the effect of the hydrogen atoms both in the non-bonded interactions and electrostatic charge. Despite recent work has been done for some torsional potentials, usually the intramolecular FF parameters of "stiff" IC are not changed in the UA approach, thus the parameters driving the C_2-C_2-C_2 stretching and bending motion in the aliphatic chains are the same as those commonly employed in the FA description.

In the UA approximation the involved atoms are considered to move as a single point with the consequence that the translational movements with respect to the rest of the molecule can be somehow taken into account, but the relative rotational movements are irreparably lost. In other words a three dimensional object described by 6 coordinates is transformed into a single point described by 3 coordinates. Even in the (non realistic) hypothesis that there exists some local vibrational modes much faster than those involving the atoms close to the UA, this approximation affects the motion of the neighboring atoms. Thus the remaining vibrational frequencies are altered by the UA approach and it is convenient focusing on the representation of the intra-molecular potential energy rather than on the vibrational analysis.

In the JOYCE program the UA atom approach, consistently with the previous FA approach, is treated on the basis of *ab initio* calculation of energies, gradients and Hessian. The main problem concerns with the transformation of the gradient vector and Hessian matrix in equation (35) in the case the number of effective atoms is less than than the number of true atoms in the molecule. Let consider for simplicity the case of a single UA in which N_{UA} atoms are grouped together. We use the indeces μ, ν for the Cartesian coordinates referred to the atoms involved in the UA and the indeces a, b for those of the remaining atoms not involved in the UA (in this section we are forced to change the previous notation). For simplicity we suppose that only one atom in the UA group is linked to the unaltered atoms. The first order energy expansion around a given geometry is

$$E^{(1)} = \sum_a \sum_s^{x,y,z} E'_{as} \; \delta t_{as} + \sum_\mu \sum_s^{x,y,z} E'_{\mu s} \; \delta t_{\mu s} \tag{47}$$

where t_{as} represents the s-th component of the CC of the a-th atom. The new gradient vector of the united atom U for a given geometry is transformed according to the simple expression

$$E'_{Us} = \sum_\mu E'_{\mu s} \qquad\qquad (s = x, y, z) \tag{48}$$

where E'_U represents the energy gradient with respect to the UA displacements. This expression is consistent with the hypothesis that the UA represents a set of internally frozen atoms: $\delta t_{Us} = \delta t_{\mu s}$ ($\mu = 1...N_{UA}$) and holds for simultaneous translations but not for rotations of the grouped atoms.

The second order energy is

$$E^{(2)} = \frac{1}{2}\sum_{ab}\sum_{sr}^{x,y,z} E''_{as,br}\delta t_{as}\delta t_{br} + \frac{1}{2}\sum_{\mu\nu}\sum_{sr}^{x,y,z} E''_{\mu s,\nu r}\delta t_{\mu s}\delta t_{\nu r} + \sum_{a\mu}\sum_{sr}^{x,y,z} E''_{as,\mu r}\delta t_{as}\delta t_{\nu r} \tag{49}$$

Defining the UA Hessian matrix as

$$E''_{Us,Ur} = \sum_{\mu\nu} E''_{\mu s,\nu r} \tag{50}$$

$$E''_{as,Ur} = \sum_{\mu} E''_{as,\mu r} \tag{51}$$

the energy expression becomes

$$\begin{aligned} E^{(2)} &= \frac{1}{2}\sum_{ab}\sum_{sr} E''_{as,br}\delta t_{as}\delta t_{br} + \frac{1}{2}\sum_{sr} E''_{Us,Ur}\delta t_{Us}\delta t_{Ur} + \sum_{a}\sum_{sr} E''_{as,Ur}\delta t_{as}\delta t_{Ur} \\ &= \frac{1}{2}\widetilde{\delta t}E''\delta t \end{aligned} \tag{52}$$

It is easy to verify that such a transformation of the Hessian matrix will preserve the three null eigenvalues due to translations, whereas the rotational modes of a molecule with UA included may lead to small (unphysical) energy contributions with the further undesirable consequence of small mixing between rotational and vibrational modes.

The two other quantities of the UA to be defined are the mass and the position. For most of standard UAs (*e.g.* methylene and methyl groups) the mass is taken as the sum of the involved atoms. In the case only one atom of the grouped atoms forms bonds with the rest of molecule, the natural choice for the position seems to make the UA coincident with that atom. However other choices are possible, for example the UA may be placed in the center of mass of the grouped atoms at the equilibrium geometry and/or its mass may be chosen in order to preserve the original inertia moments. Taking as criteria the magnitude of the rotational eigenvalues and the perturbation of the vibrational modes, these attempts do not lead to any improvement and were rejected. With the original choice the rotational eigenvalues at the equilibrium geometry are found to be much lower than the low frequency vibrational modes and the contamination is very small.

In summary the UA approach preserves some of the original atom-atom interactions contained in the Hessian matrix and leads to a useful simplification of the intra-molecular energy hyper-surface but does not allow conserving the rigorous implementation of the all-atom force field presented in this paper.

7.2.5 Frozen Internal Rotation Approximation (FIRA)

Let us briefly turn to the UA butane model shown in Figure 2. As already noted, this model can be described [1] by only six ICs: three bond distances, two bond angles and one dihedral, where only the latter as a soft IC. Let us suppose that the JOYCEfitting includes two conformations, namely the *trans* equilibrium (C_1–C_2–C_2–C_1 = 180°) and totally eclipsed (C_1–C_2–C_2–C_1 = 0°) The central C_2–C_3 distance is different for the two conformations (2.90 Å and 2.95 Å , respectively for the staggered and eclipsed [1]), whereas the two bending angles change by about 3° . Despite the dihedral angle is by far the most evident geometrical change on going from staggered to eclipsed conformation, relevant energy contributions occur even for the small changes of the other ICs: bond lengths and angles were found [1] to account for 3.4 and 2.9 kJ/mol, respectively. Consequently the torsional energy term of eq. (10) accounts for about 75% of the relative energy E(eclip)-E(stagg). Therefore the resulting pure torsional potential (eq. (10)) describes a lower barrier (18 rather than 24 kJ/mol), being the remaining gap accounted for the energy terms of the bond distances and angles.

This (rather obvious) finding has the unpleasant consequence that a good description of the large amplitude torsional geometrical movements cannot be achieved with high accuracy by simple FFs. Indeed, by using a class I FF (*i.e.* no coupling term), the fraction of the torsional energy connected with the changes of the other IC is completely loss, because there is no reason the bond lengths and angles change during the internal rotation (frozen rotation). In fact the information linking the dihedral to the other ICs in QM calculation is completely lost, since in central FFs the motion of one IC is independent from the position of the other ICs. The straightforward remedy for this problem would require the inclusion of a relevant number of coupling functions in equations (14) - (23), as done for example in the QMFF procedure [37], with the consequence of increasing the number of functions in the FF.

A more simple and direct solution is to ignore the changes of most of the ICs not directly involved in the internal rotation and, in case, retaining the changes of few pertinent ICs whose coupling term with the dihedral are included in the FF. This route has the effect of ascribing the torsional energy to the torsional term (10) only, whereas in the QM calculation it is distributed on several ICs since all the ICs are in principle coupled to each other. This method, which is implicitly adopted in partial parameterization of flexible molecules will be called FIRA: frozen internal rotation approximation.

References

[1] Cacelli, I.; Prampolini, G. Parametrization and Validation of Intramolecular Force Fields Derived from DFT Calculations *J. Chem. Theory Comput.* **2007**, *3*, 1803–1817.

[2] Prampolini, G.; Livotto, P. R.; Cacelli, I. Accuracy of Quantum Mechanically Derived Force-Fields Parameterized from Dispersion-Corrected DFT Data: The Benzene Dimer as a Prototype for Aromatic Interactions. *J. Chem. Theory Comput.* **2015**, *11*, 5182–96.

[3] Vilhena, J. G.; Greff da Silveira, L.; Livotto, P. R.; Cacelli, I.; Prampolini, G. Automated Parameterization of Quantum Mechanically Derived Force Fields for Soft Materials and Complex Fluids: Development and Validation *J. Chem. Theory Comput.* **2021**, *17*, 4449–4464.

[4] Prampolini, G.; Silveira, L. G. D.; Vilhena, J. G.; Livotto, P. R. Predicting Spontaneous Orientational Self-Assembly: In Silico Design of Materials with Quantum Mechanically Derived Force Fields *J. Phys. Chem. Lett.* **2022**, *13*, 243–250.

[5] Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids; Clarendon: Oxford, 1987.

[6] Frenkel, D.; Smith, B. Understanding Molecular Simulations; Academic Press: San Diego, 1996.

[7] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713; PMID: 26574453.

[8] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[9] Jorgensen, W. L.; Maxwell, D. S.; Tirado-rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[10] Jorgensen, W. L.; Tirado-Rives, J. Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6665–70.

[11] Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program *J. Comp. Chem.* **2009**, *30*, 1545–1614.

[12] Martinez, P.; Piras, A.; Giannini, S.; Semmeq, A.; Galvez, J.; Padula, D.; Cerezo, J.; Vilhena, J.; Prampolini, G.; JOYCE website or change tile, add url http://www.xxxxx, last consulted Dec 2024; 2024.

[13] Wang, J.; Wang, W.; P. A., K.; D. A., C.; Tirado-Rives, J. Automatic atom type and bond type perception in molecular mechanical calculations *J. Mol. Graph. Model.* **2006**, *25*, 247260.

[14] Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands *Nucleic Acids Res.* **2017**, *45*, W331–W336.

[15] Procacci, P. PrimaDORAC: A Free Web Interface for the Assignment of Partial Charges, Chemical Topology, and Bonded Parameters in Organic or Drug Molecules *J. Chem. Inf. Mod.* **2017**, *57*, 1240–1245.

[16] Morado, J.; Mortenson, P. N.; Verdonk, M. L.; Ward, R. A.; Essex, J. W.; Skylaris, C.-K. ParaMol: A Package for Automatic Parameterization of Molecular Mechanics Force Fields *J. Chem. Inf. Mod.* **2021**, *61*, 2026–2047.

[17] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers *SoftwareX* **2015**, *1-2*, 19 – 25.

[18] Santoro, F.; Cerezo, J.; *$\mathcal{FC}$classes*3, a code for vibronic calculations. Available from `http://www.iccom.cnr.it/en/fcclasses`; 2019.

[19] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.; GAUSSIAN16 Revision D.02; 2016; Gaussian Inc. Wallingford CT.

[20] Cerezo, J.; Prampolini, G.; Cacelli, I. Developing accurate intramolecular force fields for conjugated systems through explicit coupling terms *Theor. Chem. Accounts* **2018**, *137*, 80.

[21] Cacelli, I.; Cimoli, A.; Livotto, P. R.; Prampolini, G. An Automated Approach for the Parameterization of Accurate Intermolecular Force-Fields: Pyridine as a Case Study. *J. Comp. Chem.* **2012**, *33*, 1055.

[22] Prampolini, G.; A., C.; Cacelli, I.; Picky3.0, a Fortran 77 code for inter-molecular force field parameterization, available free of at http://www.iccom.cnr.it/en/picky-en/, last consulted May 2022; 2020.

[23] Padula, D.; A program to automatically select and group internal coordinates given a geometry. The output is a GROMACS topology ready to be used with the Joyce program, to parameterise a Force Field., `https://github.com/dpadula85/SelIntCoords`; 2024.

[24] Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales *Comp. Phys. Comm.* **2022**, *271*, 108171.

[25] Turner, P. XMGRACE, Version 5.1. 19 *Center for coastal and land-margin research, Oregon Graduate Institute of Science and Technology, Beaverton, OR* **2005**, *2*, 19.

[26] Giannini, S.; Martinez, P.; Semmeq, A.; Galvez, J.; Piras, A.; Landi, A.; Padula, D.; Santoro, F.; Vilhena, J.; Cerezo, J.; Prampolini, G. Joyce3.0: A General Purpose Protocol for the Specific Parameterization of Accurate Intramolecular Quantum Mechanically Derived Force-Fields *J. Chem. Theory Comput.* **2024**, *submitted.*

[27] Barone, V.; Cacelli, I.; De Mitri, N.; Licari, D.; Monti, S.; Prampolini, G. Joyce and Ulysses: Integrated and User-Friendly Tools for the Parameterization of Intramolecular Force Fields from Quantum Mechanical Data. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3736–51.

[28] Cerezo, J.; Tools to interface $\mathcal{FC}$*classes*3 with quantum chemistry codes, visit: `https://github.com/jcerezochem/fcc_tools`, last consulted September; 2022.

[29] Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

[30] Cerezo, J.; Aranda, D.; Avila Ferrer, F. J.; Prampolini, G.; Santoro, F. Adiabatic-Molecular Dynamics Generalized Vertical Hessian Approach: A Mixed Quantum Classical Method to Compute Electronic Spectra of Flexible Molecules in the Condensed Phase *J. Chem. Theory Comput.* **2020**, *16*, 1215–1231.

[31] Pulay, P.; Fogarasi, G. Geometry optimization in redundant internal coordinates *J. Chem. Phys.* **1992**, *96*, 2856.

[32] Peng, C.; Ayala, P.; Shlegel, H.; Frisch, M. Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States *J. Comp. Chem.* **1996**, *17*, 49.

[33] Dasgupta, S.; Goddard III, W. Hessian biased force fields from combining theory and experiment *J. Chem. Phys.* **1989**, *90*, 7207.

[34] Bakken, V.; Helgaker, T. The efficient optimization of molecular geometries using redundant internal coordinates *J. Chem. Phys.* **2002**, *117*, 9160.

[35] Dasgupta, S.; Yamasaki, T.; Goddard III, W. The Hessian biased singular value decomposition method for optimization and analysis of force fields *J. Chem. Phys.* **1996**, *104*, 2898.

[36] Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. Numerical Recipies in Fortran 77; Cambridge University Press: Cambridge, 1992.

[37] Maple, J.; Hwang, M.-J.; Stockfish, T.; Dinur, U.; Waldman, M.; Ewig, C.; Hagler, A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 *J. Comp. Chem.* **1994**, *15*, 162.

List of Figures