SMFA

Systematic Molecular Fragmentation by Annihilation

Table of Contents

1. Purpose, Capabilities and Requirements

1.1 Purpose

The purpose of the SMFA package is to provide the means to calculate the energy, structure and properties of moderate to large organic, organo-metallic, inorganic and biological molecules using modern electronic structure methods. The structure of both energy minima and saddle points ("transition states") can be determined.

SMFA works by decomposing a molecule into relatively small molecular fragments. Ab initio quantum chemistry calculations (or DFT) on these fragments, combined with perturbation theory, are used to accurately estimate the structure, energy and properties of the whole molecule. The largest fragments typically contain about 4 - 7 chemical functional groups. The computational time required is determined by the time required to perform quantum chemistry calculations on these fragments. The number of fragments is proportional to the number of chemical functional groups in the molecule. If the user has just one computing unit available, then the total computer time required is linearly proportional to the size of the molecule. If the user has enough computing units available, the quantum chemistry calculations for the fragments can be performed in parallel, and the total computer time required is *independent* of the size of the molecule. In this way, the ultimate purpose of SMFA is to make accurate quantum chemistry calculations of large molecules feasible.

1.2 Capabilities

The package can be used in conjunction with any one of the following quantum chemistry program packages: GAMESS (US), GAUSSIAN, NWChem and QChem. The user can therefore employ any quantum chemistry method that is available in these program packages, for example, Hartree Fock, Móller Plesset perturbation theory, coupled cluster methods, or some flavour of density functional theory. At present, SMFA can calculate the

• Energy,

• Energy gradient,

• Hessian and frequencies,

• Minimum energy structures (with or without constraints),

• Saddle point (transition state) structures (with or without constraints),

• Scan of minimum energies on a specified path.

• Various properties and isoenergetic reaction schemes

SMFA can evaluate the above properties with explicit solvent molecules included in the structure. Implicit solvent approximations are NOT presently available.

Utility programs also provide the means to calculate various properties:

(a) Infra-red spectra using a peak resolution specified by the user.

(b) Infra-red spectra for molecules with isotopic substitutions.

(c) The electrostatic potential on the solvent accessible surface.

(d) a qualitative description of the molecule as a participant in isodesmic, homodesmotic, and corresponding higher order iso-energetic chemical reactions.

(e) The dipole polarizability tensor.

(f) The hyperpolarizability tensor.

(g) Values of selected internal coordinates.

(h) Add hydrogen atoms apparently missing from available structures.

Importantly, Section 6 of this manual also contains instructions for how to obtain any property which is calculated by the quantum chemistry packages.

Although SMFA has been applied to crystals[1-3], crystal surfaces[4-6], and fluids [7] under periodic boundary conditions, the periodic version of SMFA has NOT been implemented in this release.

The methods incorporated in SMFA have been published previously. This manual does not describe the methodology of the fragmentation approach in detail; you will have to read the literature [8] [1,2,7,9-15], perhaps starting with some reviews[16,17]. The manual is intended solely as a guide for how to use SMFA. However, some additional implementation details (not previously published) are included in the Appendices of this manual.

1.3 Requirements

(a) The user must have installed at least one of GAMESS (US), GAUSSIAN, NWChem or QChem on their computer.

(b) The quantum chemistry program DALTON is used by SMFA to provide some essential calculations that are not available on some or all of the four packages above; so DALTON must be installed.

(c) Some Fortran routines in SMFA employ routines from the LaPack library, so this library must be installed.

(d) Perl must be installed.

We note that GAMESS (US), NWChem and DALTON are all available free-of-charge under licence.

1.4 Installing SMFA on your computer

See Appendix A.

1.5 Acknowledgements

Part of the energy of a large molecule arises from interactions between segments of the molecule that are separated by large distances. In SMFA, the dispersion, electrostatic and induction contributions to these interactions are evaluated using perturbation theory. [18]

When using GAMESS-US, GAUSSIAN or QChem, the electrostatic interactions are evaluated (in large part) using charges, and higher order multipoles distributed on the atoms. GAMESS-US calculates these distributed multipoles, using its own version of the GDMA method which was developed by Anthony Stone. When using GAUSSIAN or QChem, the SMFA program directly uses the GDMA program version 2.1. [19]

The authors gratefully thank Professor Anthony Stone for the use of GDMA.

The authors also gratefully acknowledge the assistance of our colleague Dr Andrew Gilbert for his assistance with "Cmake" and the preparation of SMFA for installation.

2. Running SMFA

2.1 Overview

SMFA is intended to be easy for the non-expert quantum chemist to use. However, in order to carry out any reliable quantum chemistry calculation, the user needs to specify an appropriate quantum chemistry method and a basis set for the electronic wavefunction or density. If you don't have even that much knowledge, then ask advice from someone who does. Someone who is familiar with using one of GAMESS (US), GAUSSIAN, NWChem or QChem will be able to use SMFA easily.

In addition to a method and basis set, you will need a molecule (possibly including solvent molecules to solvate it). In practice, you will need the Cartesian (xyz) coordinates stored in a file on your computer (details below). This structure might come from a crystal structure, NMR determined structure, or via a chemical structure drawing program (eg Spartan, Jmol, QMol). The structure can be optimised to that of an energy minimum or a saddle point by SMFA.

SMFA stands for "Systematic Molecular Fragmentation by Annihilation". The method can be used with a systematic series of values of the input variable "Level". Larger values of Level decompose the molecule into larger fragments which require more computer time for the electronic structure calculations, but give a more accurate estimate of the molecular property. A sensible user runs SMFA with a secession of increasing values of Level to see convergence of the value of the property they are interested in.

2.2 Preliminaries

During its operation, SMFA produces many intermediate files that have the same name for whichever molecule is being evaluated, so you should create a separate subdirectory for each molecule if you want to investigate more than one molecule at the same time. Put the file containing the Cartesian coordinates of the molecule into that subdirectory. The format of this file is that of a standard "xyz" format file. This is

line 1: The number of atoms in the molecule (eq 472)

line 2: a comment (eg the chemical/biological name of the molecule) or blank line

line 3: The chemical element symbol and x, y, z coordinates of the first atom (Å)

For example,

C   2.634578   -0.012476   10.426925

and so on for each atom in the molecule. This is a free format read, except that 2 characters should be allowed for the elemental symbol. SMFA requires correct elemental symbols. For example, magnesium is Mg not MG or mg. (Note: files from the Protein Data Base may not observe this rule)

If you have the molecular structure in some other format than xyz, then the OpenBabel program may provide the means to convert from that format to xyz format.

NOTE: There are many (free) molecular graphics programs that accept xyz format files as input (eq iMol, MacMolPlt, IQMol, VMD) and produce a graphical representation of the molecule (eg ball and stick). It will prove very useful to have such a graphics program installed somewhere convenient. Some aspects of SMFA (eg optimisation subject to constraints) will require the user to know the number of some individual atoms in the sequence of input coordinates. Graphics will make this easy if you can look at a graphic of the molecule with the atom numbers shown. If you will have to specify constraints, then the "Internal coordinates utility" in the Utilities option (in the main menu) will help you find values of atom-atom distances, valance angles and dihedral angles in the input geometry.

2.3 Getting started

We start SMFA by simply typing SMFA (RETURN).

A menu appears:

```
            main menu options

->   1) Set up input
     2) SMFA examines the input and prints comments or queries to OUT_SMFA
     3) Fragment the molecule
     4) Time limit, memory, etc
     5) Run all the electronic structure calculations
     6) Restart the electronic structure calculations
     7) Utilities
     8) Exit SMFA
```

We use the arrow keys to move up and down the menu and the RETURN key to choose.

Let's start with **item 1** "Set up input"

2.4 Set up input

Choosing

```
1) Set up input
```

sends us to the place where we input quantities needed by SMFA and the quantum chemistry programs.

This choice brings up another menu

```
            submenu: Input control

->   1) Specify the Level of fragmentation
     2) Specify the cartesian coordinates file
     3) Specify the electronic structure calculation
     4) Specify hydrogen bonding and any unusual bonding
     5) Specify metal charges (optional)
     6) Exit input control
```

We will now examine each of these parts of the required input.

NOTE: If you have completed any part of these submenu selections once, then you normally do not have to re-enter these selections if you want to change some other part of the input.

2.4.1 Submenu item 1   Specify the Level of fragmentation

This choice brings two questions to answer

`Enter the Level of fragmentation`

The response can be 1 or 2 or 3 or ....

("1" is possible but would not produce reliable results). This is the important parameter denoted as "Level" (see Section 4 of this Guide). A second question appears

`Enter the cutoff value for non-bonded interactions`

We denote this parameter as d_{tol} (see Appendix B). A typical response would be "1.1"

(or a larger value for more accurate gradients (see Appendix B). If two functional groups in the molecule are not close in terms of bonded connections, and the shortest atom-atom distance between them is greater that d_{tol} times the sum of the atom Van der Waals radii, then the interaction between these groups is evaluated using perturbation theory.

NOTE: If Level = 2, it is recommended that you set $d_{tol} = 0$.

NOTE: If Level = 1, you MUST set $d_{tol} = 0$.

The program returns to the submenu.

2.4.2 Submenu item 2   Specify the cartesian coordinates file

This choice brings a single question

`Name of the file containing the Cartesian coordinates.`

The response is the name of a file with "xyz" format in the current subdirectory.

NOTE: If you have entered a new coordinate file name, then all the input information requested below must be re-entered. That is, any previous input in the submenu items below will have been deleted.

The program returns to the submenu.

2.4.3   Submenu item 3   Specify the electronic structure calculation

This choice brings up a series of questions. Many require a "Y" or "N" answer. SMFA is not case sensitive in such cases, so "y" or "n" will do.

The first question asks the user to choose one of the four supported quantum chemistry program packages.

What quantum chemistry program package will you use?

The current choices are:

1 GAMESS US

2 Gaussian09

3 NWChem

4 QChem

Choose a number

You enter an integer, 1, 2, 3, or 4. Depending on the choice, a series of questions appear that ask for the input required for that particular quantum chemistry program. Some responses will require a detailed knowledge of the format used by that program to specify the ab initio method and basis set. This section also asks the user to accept or reject the use of embedded charges as part of the means to model polar solvent molecules. We consider each quantum chemistry package in turn.

2.4.3. A   GAMESS(US)

Having chosen GAMESS, you are prompted:

```
Enter the calculation type as an integer:
Energy     (0)
Force      (1)
Frequency (2)
Optimize   (3)
Find TS    (4)
Scan       (5)
```
        You enter 0, 1, 2, 3, 4 or 5.

This prompts a request to enter the relevant items in the $CONTRL section of GAMESS input:

```
                    $CONTRL group variables


Here you enter necessary parts of the $CONTRL group, ONE item per line
For example
SCFTYP=RHF
MPLEVL=2

Omit the RUNTYP variable
Omit the charge (ICHARG) and multiplicity (MULT)
Omit $BASIS
Omit COORD as COORD=UNIQUE is used by SMFA
Omit the $DATA section
The format MUST BE EXACTLY as GAMESS requires
Begin now and end with a RETURN
```

Much of the $CONTRL section is automated by SMFA, but you must enter the information that sets

the ab initio quantum chemistry method. For example, you might enter

SCFTYP=RHF

if you want a restricted Hartree Fock calculation, or

SCFTYP=RHF

MPLEVL=2

if you want an MP2 calculation, or

SCFTYP=RHF

MPLEVL=B3LYP

if you want a B3LYP calculation.

Each $CONTRL variable is on a separate line, and a blank line (RETURN key only) terminates this

response and elicits the next prompt:

```
Here you enter other GAMESS variable groups (if any), eg $SYSTEM, $SCF, etc
The format MUST BE EXACTLY as GAMESS requires (don't leave out compulsory spaces)
Begin now and end with a RETURN
```

Typical input might be

```
 $SYSTEM MWORDS=4 $END
```

or other bits of obscure jargon well known to GAMESS initiates. Note $SYSTEM above is prefaced by a single space. The blank line (RETURN only) elicits the next question :

```
Does the chosen electronic structure  method specified above account for
electronic correlation leading to dispersion (see the user's manual)?
Do you therefore want to account for dispersion at long range?
Answer Y or N
```

You enter Y or N.

SMFA attempts to approximate the electronic energy that would be calculated for the whole molecule at the given level of theory. The exact molecular energy includes the dispersion interaction between parts of the molecule that are far apart. However, some electronic structure methods ignore this dispersion interaction, so SMFA also ignores it. For example, the Hartree-Fock method and many common DFT methods (eg B3LYP) ignore dispersion, so "N" would be the appropriate response in those cases. In contrast, for MP2, CCSD and "dispersion-corrected" DFT methods, "Y" would be the appropriate response. This response elicits the next question:

```
GAMESS allows a different basis set for each chemical element
Do you want all elements to have the same basis (Y or N) ?
```

You enter Y or N.

Generally, one uses a single choice of basis set for all atoms in the molecule, eg a Pople basis set like 6-31G(d,p), or a Dunning basis set like aug-cc-pvTZ, in which case the response is "Y". However, for example, if there are both metal atoms and first-row atoms in the molecule, one might want different basis sets for the metals and for the other atoms, in which case the response is "N".

If the answer is "Y", this elicits the next prompt:

```
Enter the common basis for all atoms
You MUST follow the format required by GAMESS
in 'card sequence U' of the $DATA group
Enter one or more lines and finish with a (blank) RETURN
```

You will need to know the precise format for GAMESS basis sets, as entered in `'card sequence U'`. So, you will need the GAMESS user's manual. For example, for a Pople 6-31G basis set in `'card sequence U'` you would enter

 N31 6

NOTE: It seems to be very difficult to find the correct format for more complicated basis sets than "N31 6". Probably, the simplest approach for GAMESS is to answer "N" to the question above (ie multiple basis sets), and proceed as detailed immediately below.

If you want multiple basis sets, the answer is "N", and this elicits the next prompt:

SMFA knows which elements are present in the molecule, and asks you enter the appropriate basis set after each element is shown. So, your screen might look like:

H

N21 3

C
N31 6

O
N31 6

For more complicated basis sets, a not-too-difficult approach is as follows.

      (i) Launch the EMSL basis set exchange website https://bse.pnl.gov/bse/portal

      (ii) Choose the basis set in the table on the left

      (ii) Choose Format: GAMESS-US

      (iii) Click the elements you want in the periodic table

      (iv) Click "Get Basis Set"

      (v) Cut and paste the basis set definition for each element from the EMSL output into the

      SMFA input after each element prompt. Do not include the element name in the EMSL output.

Your screen then might look like

H

```
     S    3
     1      18.7311370                0.03349460
     2       2.8253937                0.23472695
     3       0.6401217                0.81375733
     S    1
     1       0.1612778                1.0000000
     P    1
     1       1.1000000                1.0000000
```

C

```
     S    6
     1    3047.5249000                0.0018347
     2     457.3695100                0.0140373
     3     103.9486900                0.0688426
     4      29.2101550                0.2321844
     5       9.2866630                0.4679413
     6       3.1639270                0.3623120
     L    3
     1       7.8682724               -0.1193324            0.0689991
     2       1.8812885               -0.1608542            0.3164240
     3       0.5442493                1.1434564            0.7443083
     L    1
     1       0.1687144                1.0000000            1.0000000
     D    1
     1       0.8000000                1.0000000
```

Do NOT include the $END line.

SMFA will then go to the next prompt:

SMFA often needs to calculate the static dipole polarizabilty of each functional
group. SMFA uses the DALTON program, rather than GAMESS to do this.
The 6-311+G(d,p) basis set is considered adequate for this purpose, and is the default.
If you want to over-ride this default, you must enter a basis set in a
format that is recognised by DALTON.
Enter the chosen basis set, or hit the RETURN key to accept the default:

NOTE: you need to use the DALTON format for basis functions here [eq 6-31G(d,p) is 6-31G** in

DALTON parlance]. If you are using a single basis set for the energy calculation, and that basis is

available in DALTON, then the appropriate choice here is to use the same basis for the polarizability.

So,

your response here might be

6-31G**

This completes the input for GAMESS, unless you have chosen Optimize (3)

or Find TS (4) or Scan (5) above.

For Optimize (3) or Find TS (4), SMFA asks for additional information as follows:

The geometry will be optimised in a number of discrete steps (geometry changes).
It is prudent to limit the number of steps employed to limit the total cpu time
consumed. The optimisation can always be restarted from the last step reached.
Enter the maximum number of steps allowed

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then asks:

The geometry optimisation can be carried out with constraints on bond lengths,
valence bond angles and dihedral angles, if requested.

Do you want constraints?

You enter "Y" or "N". If you enter "Y", then SMFA responds:

Enter the number of bond lengths to be constrained

You enter an integer, say 2 (or 0). If the number of bonds is greater than zero, SMFA responds:

Enter the atom numbers and bond lengths for each constraint

You enter data that looks like

12 63 1.09

63 79 1.53

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the

corresponding bond length (to be constrained) is given in Angstrom.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 3 atom numbers and angle in degrees for each angle

You enter data that looks like

12 63 79 104.5

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the

corresponding angle is in degrees. The middle atom number (here 63) is at the vertex of the angle.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 4 atom numbers and dihedral angle in degrees for each angle

You enter data that looks like

12 63 79 53 60.0

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding dihedral angle is in degrees.

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, $-\pi$ to π. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA then terminates the input for geometry optimisation.

For Find TS (4), SMFA asks for additional information as follows:

```
Many saddle points may exist on the potential energy surface
of large molecules. To help identify the saddle point (TS) of interest,
you must enter the atom numbers of a few atoms (say 3), that you expect to
change position as the molecule passes through the TS.

Enter the number of atoms you will denote as relevant to the TS
```

You enter an integer. This might be (say) 3 for the atoms involved in the breaking and forming bonds at the TS, but can be any positive integer.

SMFA then prompts:

```
Enter each atom number, one per line
```

You enter data that looks like

14
15
33

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the number of atoms equals the integer above.

For Scan (5), SMFA asks for additional information as follows:

```
The geometry will be optimised for a set of constraints.
You must specify a set of configurations defined by a path, a sequence of
constraints.
These constraints apply to bond lengths, angles, and dihedral angles, which
are kept fixed while all other degrees of freedom are optimised.

See the Manual for details.

Enter the number of geometries in the scan
```

You enter an integer, say 10.

SMFA then responds:

```
For each geometry in the scan, the unconstrained degrees of freedom will be
optimised in a set of steps. You must specify the maximum number of such
optimisation steps allowed.
Enter the maximum number of steps.
```

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then responds:

```
Enter the number of bond lengths to be constrained on the path
```

You enter an integer (which may be 0 or more). If you enter 1 or more,

SMFA then responds:

```
Enter the atom numbers, initial bond length, and increment for each bond constraint
```

You enter as many lines of input as the number of bond constraints specified above. Each line might

look something like:

12 47 1.53 0.05

In this example, the bond between atoms 12 and 47 will be constrained to a length of 1.53 (Å) at the

first geometry in the scan, 1.58 (Å) at the second geometry, and so on. A negative increment (eg -0.05)

would decrease the bond length at each step on the scanned path.

SMFA then responds:

```
Enter the number of angles to be constrained
```

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

```
Enter the 3 atom numbers, initial angle and increment in degrees for each angle
```

For each constrained angle, you enter the 3 atom numbers that define the angle, the initial value of the

angle on the path, and the increment for the angle (both in degrees). The integers n1, n2, n3 define the

angle that the vector from n2 to n1 makes with the vector from n2 to n3. An input line might look like:

10 17 54 104.5 5.0

Angles are always positive numbers, increments can be positive or negative.

SMFA then responds:

```
Enter the number of dihedral angles to be constrained
```

You enter an integer (which may be 0 or more).

If the number of dihedral angles is 1 or more, SMFA responds:

```
Enter the 4 atom numbers, initial angle and increment in degrees for each dihedral
```

For each constrained dihedral angle, you enter the 4 atom numbers that define the dihedral angle, the

initial value of the dihedral angle on the path, and the increment for the dihedral angle (both in

degrees). The integers n1, n2, n3, n4 define the angle that the plane containing atoms n1, n2 and n3 makes with the plane containing atoms n2, n3 and n4. An input line might look like:

23 567 32 17 60.0 10.0

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, -180 to 180. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

This completes the input specific for GAMESS. SMFA returns to the Input control submenu.

2.4.3. B   GAUSSIAN

Having chosen GAUSSIAN09, you are asked:

```
Enter the type of calculation as an integer:
Energy     (0)
Force      (1)
Frequency (2)
Optimize   (3)
Find TS    (4)
Scan       (5)
```

You enter 0, 1, 2, 3, 4 or 5. This prompts a request to enter the method used:

```
Enter the calculation method [eg HF or MP2 or other]
```

You enter simply HF or MP2, CCSD(T) or B3LYP or whatever method GAUSSIAN recognises.

This prompts the question:

```
Does this method account for electronic correlation leading to dispersion?
(See the user's manual). Do you want to account for dispersion at long range?
Answer Y or N
```

You enter Y or N.

SMFA attempts to approximate the electronic energy that would be calculated for the whole molecule at the given level of theory. The exact molecular energy includes the dispersion interaction between parts of the molecule that are far apart. However, some electronic structure methods ignore this dispersion interaction, so SMFA also ignores it. For example, the Hartree-Fock method and many common DFT methods (eg B3LYP) ignore dispersion, so "N" would be the appropriate response in those cases. In contrast, for MP2, CCSD and "dispersion-corrected" DFT methods, "Y" would be the appropriate response. This response elicits the next question:

```
GAUSSIAN allows a different basis set for each chemical element.
Do you want all elements to have the same basis (Y or N) ?
```

You enter Y or N.

Generally, one uses a single choice of basis set for all atoms in the molecule, eg a Pople basis set like 6-31G(d,p), or a Dunning basis set like aug-cc-pvTZ, in which case the response is "Y". However, for example, if there are both metal atoms and first-row atoms in the molecule, one might want different basis sets for the metals and for the other atoms, in which case the response is "N".

If the answer is "Y", this elicits the response:

```
Enter the basis for all atoms
```

You enter 6-31G(d,p) or cc-pVQZ or whatever basis GAUSSIAN recognises.

If the answer is "N", SMFA knows which elements are present in the molecule, and asks you to enter the appropriate basis set after each element is shown. So, your screen might look like:

```
Enter the basis (check availability in the GAUSSIAN library)
for the following elements
H
3-21G
C
6-31G

O
6-31G
```

where you have entered each appropriate basis set after the prompt for each element.

Completing the basis set specification elicits the response:

```
Enter any other keywords (optional), or hit RETURN:
```

The normal response might well be a blank line. There are many possibilities, such as SCF=QC or

MaxDisk=XXXMB.

Moreover, if you have programmed a means to extract some property from the output (see Section 6),

you would enter the keyword for that property here.

This response elicits the next question:

```
Very polar solvent molecules can induce polarisation in both
solute and other solvent molecules.
SMFA can evaluate the energy of polar solvents and solutes, at a
lower Level of Fragmentation, if embedded charges are used to describe
the solvent environment. In order to identify the solvent, SMFA needs the
chemical composition of the solvent.
If your system contains polar solvent molecules,
do you want to specify the use of embedded charges? (Y/N)
```

You enter Y or N.

The answer "Y" is recommended if polar solvent molecules, like H_2O, are present in the input

geometry, as SMFA will produce a more accurate estimate of the energy at the same cost in cpu time as

"N".

If the answer is "Y", then SMFA will ask you some questions to help it identify the solvent molecules

in the total structure. SMFA will ask you to

```
Enter the number of atoms in the solvent molecule
```

You enter an integer. In the case of H_2O, you enter 3; for H_2CO you enter 4, etc.

This prompts SMFA to write

```
You have to enter each element in the solvent. For example, for water
you would enter
H
H
O
Now enter the elemental symbols for the atoms in the solvent:
```

For H$_2$CO, you would enter (the order is not important)

H

H

C

O

This completes the input for GAUSSIAN, unless you have chosen `Optimize (3)` or `Find TS (4) or Scan (5)` above.

For `Optimize (3)` or `Find TS (4)`, SMFA asks for additional information as follows:

`The geometry will be optimised in a number of discrete steps (geometry changes). It is prudent to limit the number of steps employed to limit the total cpu time consumed. The optimisation can always be restarted from the last step reached. Enter the maximum number of steps allowed`

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then asks:

> `The geometry optimisation can be carried out with constraints on bond lengths, valence bond angles and dihedral angles, if requested.`

`Do you want constraints?`

You enter "Y" or "N". If you enter "Y", then SMFA responds:

`Enter the number of bond lengths to be constrained`

You enter an integer, say 2 (or 0). If the number of bonds is greater than zero, SMFA responds:

`Enter the atom numbers and bond lengths for each constraint`

You enter data that looks like

12 63 1.09

63 79 1.53

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding bond length (to be constrained) is given in Angstrom.

SMFA then responds:

`Enter the number of angles to be constrained`

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

`Enter the 3 atom numbers and angle in degrees for each angle`

You enter data that looks like

12 63 79 104.5

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding angle is in degrees. The middle atom number (here 63) is at the vertex of the angle. SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 4 atom numbers and dihedral angle in degrees for each angle

You enter data that looks like

12 63 79 53 60.0

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding dihedral angle is in degrees.

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, -180 to 180. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA then terminates the input for geometry optimisation.

For Find TS (4), SMFA asks for additional information as follows:

Many saddle points may exist on the potential energy surface
of large molecules. To help identify the saddle point (TS) of interest,
you must enter the atom numbers of a few atoms (say 3), that you expect to
change position as the molecule passes through the TS.

Enter the number of atoms you will denote as relevant to the TS

You enter an integer. This might be (say) 3 for the atoms involved in the breaking and forming bonds at the TS, but can be a larger integer.

SMFA then prompts:

Enter each atom number, one per line

You enter data that looks like

14
15
33

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the number of atoms equals the integer above.

For Scan (5), SMFA asks for additional information as follows:

> The geometry will be optimised for a set of constraints.
> You must specify a set of configurations defined by a path, a sequence of constraints.
> These constraints apply to bond lengths, angles, and dihedral angles, which are kept fixed while all other degrees of freedom are optimised.
>
> See the Manual for details.

Enter the number of geometries in the scan

You enter an integer, say 10.

SMFA then responds:

> For each geometry in the scan, the unconstrained degrees of freedom will be optimised in a set of steps. You must specify the maximum number of such optimisation steps allowed.

Enter the maximum number of steps

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then responds:

Enter the number of bond lengths to be constrained on the path

You enter an integer (which may be 0 or more). If you enter 1 or more.

SMFA then responds:

Enter the atom numbers, initial bond length, and increment for each bond constraint

You enter as many lines of input as the number of bond constraints specified above. Each line might look something like:

12 47 1.53 0.05

In this example, the bond between atoms 12 and 47 will be constrained to a length of 1.53 (Å) at the first geometry in the scan, 1.58 (Å) at the second geometry, and so on. A negative increment (eg -0.05) would decrease the bond length at each step on the scanned path.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 3 atom numbers, initial angle and increment in degrees for each angle

For each constrained angle, you enter the 3 atom numbers that define the angle, the initial value of the angle on the path, and the increment for the angle (both in degrees). The integers n1, n2, n3 define the angle that the vector from n2 to n1 makes with the vector from n2 to n3. An input line might look like:

10 17 54 104.5 5.0

Angles are always positive numbers, increments can be positive or negative.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 4 atom numbers, initial angle and increment in degrees for each dihedral

For each constrained dihedral angle, you enter the 4 atom numbers that define the dihedral angle, the initial value of the dihedral angle on the path, and the increment for the dihedral angle (both in degrees). The integers n1, n2, n3, n4 define the angle that the plane containing atoms n1, n2 and n3 makes with the plane containing atoms n2, n3 and n4. An input line might look like:

23 567 32 17 60.0 10.0

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, $-\pi$ to π. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA returns to the Input control submenu.

This now completes the input specific for GAUSSIAN.

2.4.3. C   NWChem

Having chosen NWChem, you are asked:

```
Enter the type of calculation as an integer:
Energy     (0)
Force      (1)
Frequency (2)
Optimize   (3)
Find TS    (4)
Scan       (5)
```

You enter 0, 1, 2, 3, 4 or 5.

This prompts a request to enter the input that NWChem needs to carry out the ab initio calculations:

```
Enter each line of NWChem input needed for the tasks.
For example, for an SCF, DTF, MP2 or TCE-based calculation.
Include the TASK command
Omit the charge,
but include 'singlet' in the SCF or DFT section.
Begin now and end with a RETURN
```

Obviously, you will need to be familiar with NWChem commands and syntax. For example, for a simple Hartree Fock calculation of the energy you might enter

SCF

direct

singlet

END

TASK SCF ENERGY

For an MP2 calculation, you would replace TASK SCF ENERGY by TASK MP2 ENERGY. For mp2 frequencies, you might put "task mp2 freq" (it's not case sensitive).

NOTE: The "TASK" must be consistent with the "type of calculation" entered above.

NOTE: If you have chosen Optimize (3), Find TS (4) or Scan (5), then you must set TASK to OPTIMIZE. For example, TASK SCF OPTIMIZE or TASK MP2 OPTIMIZE, etc.

NOTE: When you ask NWChem to calculate the frequencies or hessian (eg TASK SCF FREQ), this program does not automatically calculate the gradients, as do some other programs. In fact, SMFA is expecting to read both the gradients and the hessian, when the frequencies (hessian) are requested. Hence, when calculating the frequencies (hessian) with NWChem, you must put (for example, using SCF):

TASK SCF GRADIENT

TASK SCF FREQ

Aside: If more than the default memory is needed, you may need to include an allocation of memory here. By default, NWChem assigns the available memory with only 50% assigned to "global". Often one needs more global memory than stack or heap, so an assignment such as (for example)

memory stack 100 heap 100 global 1300 mb

might be appropriate.

When you complete this entry, SMFA asks:

<pre style="color:#c00">Does the calculation method account for electronic correlation leading to dispersion?
(See the user's manual). Do you want to account for dispersion at long range?
Answer Y or N</pre>

You enter Y or N.

SMFA attempts to approximate the electronic energy that would be calculated for the whole molecule at the given level of theory. The exact molecular energy includes the dispersion interaction between parts of the molecule that are far apart. However, some electronic structure methods ignore this dispersion interaction, so SMFA also ignores it. For example, the Hartree-Fock method and many common DFT methods (eg B3LYP) ignore dispersion, so "N" would be the appropriate response in those cases. In contrast, for MP2, CCSD and "dispersion-corrected" DFT methods, "Y" would be the appropriate response.

The next question is then:

<pre style="color:#c00">NWChem allows a different basis set for each element
Do you want all elements to have the same basis (Y or N) ?</pre>

You enter Y or N.

Generally, one uses a single choice of basis set for all atoms in the molecule, eg a Pople basis set like 6-31G(d,p), or a Dunning basis set like aug-cc-pvTZ, in which case the response is "Y". However, for example, if there are both metal atoms and first-row atoms in the molecule, one might want different basis sets for the metals and for the other atoms, in which case the response is "N".

If the answer is "Y", this elicits the response:

<pre style="color:#c00">Enter the basis for all atoms</pre>

You enter 6-31G(d,p) or cc-pVQZ or whatever basis NWChem recognises.

If the answer is "N", SMFA knows which elements are present in the molecule, and asks you to enter the appropriate basis set after each element is shown. So, your screen might look like:

<pre style="color:#c00">Enter the basis (check availability in the NWChem library)
for the following elements
H
3-21G
C
6-31G</pre>

where you have entered each appropriate basis set after the prompt for each element.

When the basis set input is complete, NWChem responds with:

```
Very polar solvent molecules can induce polarisation in both
solute and other solvent molecules.
SMFA can evaluate the energy of polar solvents and solutes, at a
lower Level of Fragmentation, if embedded charges are used to describe
the solvent environment. In order to identify the solvent, SMFA needs the
chemical composition of the solvent.
If your system contains polar solvent molecules,
do you want to specify the use of embedded charges? (Y/N)
```

You enter Y or N.

The answer "Y" is recommended if polar solvent molecules, like H_2O, are present in the input geometry, as SMFA will produce a more accurate estimate of the energy at the same cost in cpu time as "N".

If the answer is "Y", then SMFA will ask you some questions to help it identify the solvent molecules in the total structure. SMFA will ask you to

```
Enter the number of atoms in the solvent molecule
```

You enter an integer. In the case of H_2O, you enter 3; for H_2CO you enter 4.

This prompts SMFA to write

```
You have to enter each element in the solvent. For example, for water
you would enter
H
H
O
Now enter the elemental symbols for the atoms in the solvent:
```

For H_2CO, you would enter (the order is not important)

H

H

C

O

This completes the input for NWChem, unless you have chosen `Optimize (3)` or `Find TS (4) or Scan (5)` above.

For `Optimize (3)` or `Find TS (4)`, SMFA asks for additional information as follows:

```
The geometry will be optimised in a number of discrete steps (geometry changes).
It is prudent to limit the number of steps employed to limit the total cpu time
consumed. The optimisation can always be restarted from the last step reached.
Enter the maximum number of steps allowed
```

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then asks:

The geometry optimisation can be carried out with constraints on bond lengths, valence bond angles and dihedral angles, if requested.

Do you want constraints?

You enter "Y" or "N". If you enter "Y", then SMFA responds:

Enter the number of bond lengths to be constrained

You enter an integer, say 2 (or 0). If the number of bonds is greater than zero, SMFA responds:

Enter the atom numbers and bond lengths for each constraint

You enter data that looks like

12 63 1.09

63 79 1.53

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding bond length (to be constrained) is given in Angstrom.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 3 atom numbers and angle in degrees for each angle

You enter data that looks like

12 63 79 104.5

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding angle is in degrees. The middle atom number (here 63) is at the vertex of the angle.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 4 atom numbers and dihedral angle in degrees for each angle

You enter data that looks like

12 63 79 53 60.0

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding dihedral angle is in degrees.

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, -180 to 180. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA then terminates the input for geometry optimisation.

For Find TS (4), SMFA asks for additional information as follows:

    Many saddle points may exist on the potential energy surface
    of large molecules. To help identify the saddle point (TS) of interest,
    you must enter the atom numbers of a few atoms (say 3), that you expect to
    change position as the molecule passes through the TS.

    Enter the number of atoms you will denote as relevant to the TS

You enter an integer. This might be (say) 3 for the atoms involved in the breaking and forming bonds at the TS, but can be a larger integer.

SMFA then prompts:

    Enter each atom number, one per line

You enter data that looks like

14
15
33

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the number of atoms equals the integer above.

For Scan (5), SMFA asks for additional information as follows:

    The geometry will be optimised for a set of constraints.
    You must specify a set of configurations defined by a path, a sequence of
    constraints.
    These constraints apply to bond lengths, angles, and dihedral angles, which
    are kept fixed while all other degrees of freedom are optimised.

    See the Manual for details.

Enter the number of geometries in the scan

You enter an integer, say 10.

SMFA then responds:

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then responds:

Enter the number of bond lengths to be constrained on the path

You enter an integer (which may be 0 or more). If you enter 1 or more.

SMFA then responds:

Enter the atom numbers, initial bond length, and increment for each bond constraint

You enter as many lines of input as the number of bond constraints specified above. Each line might look something like:

12   47   1.53   0.05

In this example, the bond between atoms 12 and 47 will be constrained to a length of 1.53 (Å) at the first geometry in the scan, 1.58 (Å) at the second geometry, and so on. A negative increment (eg -0.05) would decrease the bond length at each step on the scanned path.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 3 atom numbers, initial angle and increment in degrees for each angle

For each constrained angle, you enter the 3 atom numbers that define the angle, the initial value of the angle on the path, and the increment for the angle (both in degrees). The integers n1, n2, n3 define the angle that the vector from n2 to n1 makes with the vector from n2 to n3. An input line might look like:

10   17   54   104.5   5.0

Angles are always positive numbers, increments can be positive or negative.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 4 atom numbers, initial angle and increment in degrees for each dihedral

For each constrained dihedral angle, you enter the 4 atom numbers that define the dihedral angle, the initial value of the dihedral angle on the path, and the increment for the dihedral angle (both in degrees). The integers n1, n2, n3, n4 define the angle that the plane containing atoms n1, n2 and n3 makes with the plane containing atoms n2, n3 and n4. An input line might look like:

23 567 32 17 60.0 10.0

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, -π to π. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA returns to the Input control submenu.

This now completes the input for NWChem.

2.4.3. D   QChem

Having chosen QChem, you are asked:

```
Enter values for the $rem QChem input
The JOBTYPE can be SP, FORCE, FREQ, OPT, TS or SCAN
JOBTYPE
```

You enter one of the allowed values for the JOBTYPE variable, for example

JOBTYPE OPT

SMFA then asks for the quantum chemistry method:

METHOD

and you make an appropriate response, such as

METHOD MP2

SMFA then responds with:

```
Enter any additional values for the $rem QChem input
Do not enter a BASIS value
Begin now, and end with a RETURN
```

In (most) simple cases, no additional $rem variables may be necessary. Note that if the JOBTYPE is TS or SCAN, this will appear as OPT in the OUT_SMFA output file (so don't worry).

If you have programmed a means to extract some property from the output (see Section 6), you would enter the instruction for the calculation of that property here.

NOTE: You don't enter the regular basis set here. However, for an RI-MP2 or RI-CC calculation, you must enter the auxiliary basis here, eg

AUX_BASIS RIMP2-CC-PVDZ

SMFA then asks:

```
Does the calculation method account for electronic correlation leading to dispersion?
(See the user's manual). Do you want to account for dispersion at long range?
Answer Y or N
```

You enter Y or N.

SMFA attempts to approximate the electronic energy that would be calculated for the whole molecule at the given level of theory. The exact molecular energy includes the dispersion interaction between parts of the molecule that are far apart. However, some electronic structure methods ignore this dispersion interaction, so SMFA also ignores it. For example, the Hartree-Fock method and many common DFT methods (eg B3LYP) ignore dispersion, so "N" would be the appropriate response in those cases. In contrast, for MP2, CCSD and "dispersion-corrected" DFT methods, "Y" would be the appropriate response.

The next question is:

QChem allows a different basis set for each element
Do you want all elements to have the same basis (Y or N) ?

You answer Y or N.

Generally, one uses a single choice of basis set for all atoms in the molecule, eg a Pople basis set like 6-31G(d,p), or a Dunning basis set like aug-cc-pvTZ, in which case the response is "Y". However, for example, if there are both metal atoms and first-row atoms in the molecule, one might want different basis sets for the metals and for the other atoms, in which case the response is "N".

If you answer "Y", then SMFA responds:

Enter the basis for all atoms

You enter 6-31G(d,p) or cc-pVQZ or whatever basis QChem recognises.

If the answer is "N", SMFA knows which elements are present in the molecule, and asks you to enter the appropriate basis set after each element is shown. So, your screen might look like:

Enter the basis (check availability in the QChem library)
for the following elements
H
3-21G
C
6-31G

where you have entered each appropriate basis set after the prompt for each element.

When the basis set input is complete, QChem responds with:

Very polar solvent molecules can induce polarisation in both
solute and other solvent molecules.
SMFA can evaluate the energy of polar solvents and solutes, at a
lower Level of Fragmentation, if embedded charges are used to describe
the solvent environment. In order to identify the solvent, SMFA needs the
chemical composition of the solvent.
If your system contains polar solvent molecules,
do you want to specify the use of embedded charges? (Y/N)

You enter Y or N.

The answer "Y" is recommended if polar solvent molecules, like H_2O, are present in the input geometry, as SMFA will produce a more accurate estimate of the energy at the same cost in cpu time as "N" (at a given value of Level).

If the answer is "Y", then SMFA will ask you some questions to help it identify the solvent molecules in the total structure. SMFA will ask you to

Enter the number of atoms in the solvent molecule

You enter an integer. In the case of H_2O, you enter 3; for H_2CO you enter 4, etc.

This prompts SMFA to write

You have to enter each element in the solvent. For example, for water
you would enter
H
H
O
Now enter the elemental symbols for the atoms in the solvent:

For H_2CO, you would enter (the order is not important)

H

H

C

O

If you specified JOBTYPE OPT or TS or SCAN above, then SMFA asks for additional information.

For JOBTYPE OPT or TS, SMFA asks for additional information as follows:

The geometry will be optimised in a number of discrete steps (geometry changes).
It is prudent to limit the number of steps employed to limit the total cpu time
consumed. The optimisation can always be restarted from the last step reached.
Enter the maximum number of steps allowed

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then asks:

The geometry optimisation can be carried out with constraints on bond lengths,
valence bond angles and dihedral angles, if requested.

Do you want constraints?

You enter "Y" or "N". If you enter "Y", then SMFA responds:

Enter the number of bond lengths to be constrained

You enter an integer, say 2 (or 0). If the number of bonds is greater than zero, SMFA responds:

Enter the atom numbers and bond lengths for each constraint

You enter data that looks like

12 63 1.09

63 79 1.53

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the

corresponding bond length (to be constrained) is given in Angstrom.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 3 atom numbers and angle in degrees for each angle

You enter data that looks like

12 63 79 104.5

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding angle is in degrees. The middle atom number (here 63) is at the vertex of the angle.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer, say 1 (or 0). If the integer is greater than zero,

SMFA then responds:

Enter the 4 atom numbers and dihedral angle in degrees for each angle

You enter data that looks like

12 63 79 53 60.0

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the corresponding dihedral angle is in degrees.

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, -180 to 180. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

SMFA then terminates the input for geometry optimisation.

For JOBTYPE TS, SMFA asks for additional information as follows:

        Many saddle points may exist on the potential energy surface
        of large molecules. To help identify the saddle point (TS) of interest,
        you must enter the atom numbers of a few atoms (say 3), that you expect to
        change position as the molecule passes through the TS.

        Enter the number of atoms you will denote as relevant to the TS

You enter an integer. This might be (say) 3 for the atoms involved in the breaking and forming bonds at the TS, but can be a larger integer.

SMFA then prompts:

        Enter each atom number, one per line

You enter data that looks like

```
14
15
33
```

where the atom numbers refer to the order of the atoms in the coordinate file (submenu item 2), and the

number of atoms equals the integer above.

For JOBTYPE SCAN, SMFA asks for additional information as follows:

> The geometry will be optimised for a set of constraints.
> You must specify a set of configurations defined by a path, a sequence of
> constraints.
> These constraints apply to bond lengths, angles, and dihedral angles, which
> are kept fixed while all other degrees of freedom are optimised.
>
> See the Manual for details.

Enter the number of geometries in the scan

You enter an integer, say 10.

SMFA then responds:

> For each geometry in the scan, the unconstrained degrees of freedom will be
> optimised in a set of steps. You must specify the maximum number of such
> optimisation steps allowed.

Enter the maximum number of steps

You enter an integer for the maximum number of steps (eg 20 or 100 or whatever)

SMFA then responds:

Enter the number of bond lengths to be constrained on the path

You enter an integer (which may be 0 or more). If you enter 1 or more.

SMFA then responds:

Enter the atom numbers, initial bond length, and increment for each bond constraint

You enter as many lines of input as the number of bond constraints specified above. Each line might

look something like:

12   47   1.53   0.05

In this example, the bond between atoms 12 and 47 will be constrained to a length of 1.53 (Å) at the

first geometry in the scan, 1.58 (Å) at the second geometry, and so on. A negative increment (eg -0.05)

would decrease the bond length at each step on the scanned path.

SMFA then responds:

Enter the number of angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 3 atom numbers, initial angle and increment in degrees for each angle

For each constrained angle, you enter the 3 atom numbers that define the angle, the initial value of the angle on the path, and the increment for the angle (both in degrees). The integers n1, n2, n3 define the angle that the vector from n2 to n1 makes with the vector from n2 to n3. An input line might look like:

10 17 54 104.5 5.0

Angles are always positive numbers, increments can be positive or negative.

SMFA then responds:

Enter the number of dihedral angles to be constrained

You enter an integer (which may be 0 or more).

If the number of angles is 1 or more, SMFA responds:

Enter the 4 atom numbers, initial angle and increment in degrees for each dihedral

For each constrained dihedral angle, you enter the 4 atom numbers that define the dihedral angle, the initial value of the dihedral angle on the path, and the increment for the dihedral angle (both in degrees). The integers n1, n2, n3, n4 define the angle that the plane containing atoms n1, n2 and n3 makes with the plane containing atoms n2, n3 and n4. An input line might look like:

23 567 32 17 60.0 10.0

NOTE: Definitions of dihedral angles differ in sign and magnitude; here, dihedrals are taken in the range, $-\pi$ to π. The utility program Internal Coordinates (see Section 5), reports bond lengths, bond angles and dihedral angles for the current molecular coordinates. See Appendix C for the definition of the dihedral angle.

Following any optimization/constraint questions, SMFA returns to the Input control submenu.

This completes the input for QChem.

2.4.4 Submenu item 4   Specify hydrogen bonding and any unusual bonding

SMFA fragments a molecule according to the bonded connections between atoms. An algorithm decides what is a single bond, and what is a multiple bond (that is not broken in a fragmentation). However, SMFA allows the user to over-ride some features of the standard algorithm. For example, the input structure might describe some "snapshot" of a chemical reaction, where bonds are breaking and forming. The user can demand that some atoms be considered as bonded, even though the atom-atom distance may be much larger than a normal bond. In this way, the same fragmentation will be applied to a sequence of "snapshots" of the reaction, giving a "smooth" energy profile for the process. Hence, if you are using the SCAN facility to calculate the energy on a path where a bond is stretched beyond "breaking point", you might want to specify this bond to exist no matter what its length.

Similarly, SMFA treats hydrogen bonds like single bonds by default, but you can override this if desired.

If you are happy with all the default settings regarding bonding, then you can skip this item.

If not, select this item, and the following questions will appear:

Question 1

```
Do you want to specify some bonds as multiple bonds,
that would normally be taken as single bonds (Y/N
or hit RET to skip all)?
```

Enter Y or N

and proceed to the next question, or hit the RETURN key to skip this entire submenu item. Multiple bonds are never broken in the fragmentation process. If you enter Y, this prompts another question:

```
Enter the number of (unusual) user-specified double bonds (usually 0)
```

The response is an integer. An integer greater than 0 prompts

```
Enter the atom numbers for these double bonded atoms (eg 27    243)
```

The atom numbers refer to the order of the atoms in the coordinate file (submenu item 2).

You must enter as many pairs of atom numbers (one pair per line) as the integer entered above.

Question 2

```
Do you want to specify some bonds as single bonds,
that would normally be taken as multiple bonds (Y/N)?
```

Enter Y or N

Y prompts:

Enter the number of (unusual) user-specified single bonds (usually 0)

The response is an integer. An integer greater than 0 prompts

Enter the atom numbers for these single bonded atoms (eg 86   97)

The atom numbers refer to the order of the atoms in the coordinate file (item 2).

You must enter as many pairs of atom numbers (one pair per line) as the integer

entered above.

Question 3

Do you want to specify that single bonds exist
that would normally not be considered as bonds (Y/N)?

Y prompts:

Enter the number of (unusual) user-specfied extra bonds (usually 0)

The response is an integer. An integer greater than 0 prompts

Enter the atom numbers for the atoms connected by these extra bonds (eg 147   15)

The atom numbers refer to the order of the atoms in the coordinate file (item 2).

You must enter as many pairs of atom numbers (one pair per line) as the integer entered above.

Question 4

By default SMFA assumes that the amide moiety is treated as a single group
Do you want to treat the Amide CN bond as a single bond only (Y/N)?

Enter Y or N.

If you enter "Y", SMFA will produce slightly smaller fragments in molecules that

contain amide groups (eg proteins) and hence reduce the cpu time, but at the cost of

lower accuracy.

Question 5

By default SMFA assumes that CX2 or CX3 (X=F,Cl,..) is a single group.
Do you want to treat these CX bonds as single bonds only (Y/N)?
Enter Y or N.

If you enter "Y", SMFA will produce slightly smaller fragments in molecules that

contain such groups and hence reduce the cpu time, but at the cost of lower accuracy.

Question 6

By default SMFA assumes that hydrogen bonds will be treated as regular bonds
Do you want to ignore hydrogen bonds (Y/N)?

Enter Y or N.

If you enter "Y", SMFA will produce slightly smaller fragments in molecules that

contain hydrogen bonds and hence reduce the cpu time, but at the cost of lower accuracy.

<u>Question 7</u>

By default SMFA assumes that if two charged groups are close together
then they should be considered to be bonded
Do you want to ignore this default (Y/N)?

Enter Y or N.

If you enter "Y", SMFA will produce slightly smaller fragments in molecules that

contain such "adjacent" charged groups and hence reduce the cpu time, but at the cost of

lower accuracy. The idea is that the two parts of something like a "salt bridge" should be

consider as bonded.

If you enter "N", this prompts a further question

Two charged groups are considered close together if the distance
between them is less than the sum of the Van der Waals radii of
the closest atoms multiplied by a factor
Enter the value of this factor (the recommended value is 1.5)

Enter a value.

The program returns to the submenu.

2.4.5 Submenu item 5   Specify charges for metals & other atoms (optional)

SMFA determines whether or not a chemical functional group is neutral or charged based on the valency of the chemical elements (accounting for alternative possibilities, eg Phosphorus might have a valency of 3 or 5). However, SMFA cannot know, *a priori*, what is the charge of a metal atom (eg Fe^{2+} or Fe^{3+} ?). If the molecule contains metal atoms, the user must specify the formal charge on each metal. If you forget to do this, SMFA will identify such "unspecified" metals, and query this in the output file OUT_SMFA, when the input is checked. You will be asked to modify any default charges assigned by SMFA.

In addition, the valency of some atoms may be unusual in some circumstances, for example if the molecule is a radical, or the geometry is very perturbed, as in a transition state. For example, a reaction may involve abstracting a hydrogen atom from a carbon atom, leaving a group like this

in the molecule. It is impossible to know a priori if such a group (eg CH_3) is an anion, a cation, or neutral. The user must specify the charge of such an atom (the carbon, in this case). NOTE: This is likely to be an issue during a scan or in the search for a transition state.

If you select this item, SMFA asks a series of questions.

Question 1

```
Are there metals or other atoms whose charge must be specified
or modified (Y/N, or hit RETURN to skip)?
```

Enter Y or N (and proceed to the next question, or hit the RETURN key to skip this question. NOTE: if you have previously specified some metal charges, or if SMFA has assigned default charges to some metals, AND you hit the return key, then the earlier assigned charges will be implemented. "Y" prompts

Question 2

```
Enter the number of metal atoms
```

The response is an integer. An integer greater than 0, prompts

Question 3

```
Enter the atom number(s) and associated charge(s), eq 34  2
```

Enter each metal (or unusual) atom number and charge, one atom per line, as many atoms as entered above. That is, you enter lines that look like

```
63   2
103  0
```

The atom numbers refer to the order of the atoms in the coordinate file (submenu item 2) and the number of lines corresponds to the integer entered above. In the example above, "103 0" might signify

that atom number 103 is a carbon radical, while "63   2" might signify that an Fe atom at position 63 in the input file is Fe^{2+}.

2.4.6   Submenu item 6   Exit input control

Self explanatory.

In general, once you have answered all the questions in one of the submenu items once, you do not need to repeat that submenu item again, if you want to make a change elsewhere in the input. The exception is that if you change the molecule coordinate file, the earlier input is removed, and you must complete the input again.

2.5 Input review and preparation

Main menu item 2 is

2) SMFA examines the input and prints comments or queries to OUT_SMFA

It is ESSENTIAL to choose this item after completing the input or if the input has been changed in any way. Having read the input parameters, SMFA prepares certain files needed by the fragmentation process, including a description of the bonding between atoms and the formal charges of functional groups. The output of this process and a record of the values of input parameters are printed to the main output file, OUT_SMFA. You should look at OUT_SMFA to check for mistakes in the input. If the input is unsatisfactory, then you return to the input menu and adjust the input values. Else, you go to item 3 in the main menu. See Section 3.1 for a guide to the output in OUT_SMFA.

2.6 Fragmentation

Item 3 in the main menu is

3) Fragment the molecule.

It is ESSENTIAL to choose this item before proceeding to the electronic structure calculation. The process is automated, no input is required except a RETURN key when the process is completed. The theoretical background for the fragmentation is described in the references. The coordinates for the fragments produced at the requested value of Level are written to the file seefrags. This file is in a standard xyz format which can be read by many molecular graphics programs (eg MacMolPlt or VMD). A brief summary of the number of fragments, and the number of electrons in the largest fragment is appended to OUT_SMFA.

SMFA also inspects the fragmentation output and decides what would be an optimum number of cpus to use to run the ab initio calculations of the fragments in parallel (in order to minimise the "walltime" for the calculation). The recommended number of cpus is written to OUT_SMFA. The output file, OUT_SMFA, also provides information on the largest fragment produced by the fragmentation process. The size of this molecule determines the degree of computational difficulty for the ab initio calculations. The user should consider what computational resources (memory, scratch disk space, time, etc) would be necessary for the desired basis set and level of ab initio theory on this largest fragment. This will determine your choices for some "system variables" in the next section. An input file for the ab initio calculation for this largest fragment is contained in a file named LARGE.com (or LARGE.inp or LARGE.nw, depending on the quantum chemistry package employed). If the fragment in LARGE.com is indeed very large, you might need to run some trial calculations with

LARGE.com to see what memory, disk space, time etc is needed, and if you need to include additional items in the ab initio program input (Section 2.4), such as memory requirements.

2.7 System variables

Item 4 in the main menu is

```
4) Time limit, memory, queue etc
```

Here you are asked to create a, or modify an existing, submission script for the execution of the ab initio calculations and other functions of SMFA. This code has been written under the assumption that a PBS (https://en.wikipedia.org/wiki/Portable_Batch_System) controls the execution of all jobs and queues.

When you choose this item, SMFA prints

```
A file called pbsfile must contain the instructions
for the PBS that controls job queues, memory requirements,
time limits, the number of cpus requested, etc.
This file also loads the quantum chemistry program packages.
```

If a pbsfile exists in the current directory, SMFA uses this file. Otherwise, a pbsfile has been saved in the SMFAPAC/bin subdirectory of the main SMFA directory, denoted pbsfile_standard, and SMFA uses this file. If a pbsfile exists in the current directory, SMFA prints

```
The current file of PBS instructions contains the following:
```

followed by the current contents of the pbsfile, for example

```
#!/bin/bash
#PBS -P a00
#PBS -N gau
#PBS -l mem=2gb
#PBS -l ncpus=1
#PBS -l jobfs=2gb
#PBS -l walltime=1:00:00
#PBS -l wd
#PBS -q normal


# enter instructions that make the quantum chemistry package available    (do not change this line)
# and finish with a blank line                                            (do not change this line)
module load qchem/4.3

# enter instructions that make the DALTON  package available              (do not change this line)
# and finish with a blank line                                            (do not change this line)
module load dalton
```

Otherwise, SMFA prints

```
The standard file of PBS instructions contains the following:
```

followed by the contents of a file similar to that above.

SMFA then provides a means to edit the pbsfile. You are prompted as follows:

```
To add a new line, Enter A
To change any line in the PBS instructions, Enter C
To remove any line in the PBS instructions, Enter R
To terminate editing this file, simply hit the RETURN key
```

The A, C and R above are not case sensitive. Following these and successive prompts, allows you to edit the file. When editing is complete, SMFA saves the modified file in the current directory. Alternatively, you can copy pbsfile_standard in the SMFAPAC/bin/ subdirectory into the current directory for this molecule as pbsfile, and use any text editor to directly edit pbsfile.

The ab initio calculations on all the fragments can be executed sequentially or in parallel. The choice of sequential versus parallel is made in the next Section/Menu item.

NOTE: In addition to resource requirements (memory, disk space, time limit), pbsfile contains a directive for which queue the calculation is submitted to.

PLEASE NOTE: You must enter the instructions for loading the quantum chemistry packages as in

enter instructions that make the quantum chemistry package available   (do not change this line)
and finish with a blank line   (do not change this line)
module load qchem/4.3

enter instructions that make the DALTON package available   (do not change this line)
and finish with a blank line   (do not change this line)
module load dalton

The instruction regarding "do not change this line " is important because SMFA uses these lines to locate the instructions for loading the quantum chemistry packages.

The instruction to "load" or otherwise make available a quantum chemistry package,
 eg module load qchem/4.3, may occupy one or more lines, but you must finish with a blank line (RETURN only, no spaces).

Unless you want to change the quantum chemistry package, you need only enter these instructions once.

2.7.1 Sequential execution

If you plan to run the calculations **sequentially**, the core memory and scratch space memory should be set as appropriate for the ab initio calculation on the largest fragment. The time limit should be set to allow all calculations to complete (carried out one after the other). The number of atoms and electrons in this largest fragment were reported in the OUT_SMFA file after the fragmentation has been completed. A sample input file, LARGE.com, for this fragment was also

created. You MUST set #PBS -l ncpus=1 above in this case. Generally, even the largest fragment is too small to benefit significantly from the use of multiple processors.

2.7.2 Parallel execution

If you plan to run the calculations in parallel, then you must set the number of cpus here (call it ncpus), say

#PBS -l ncpus=100

This can be the number of cpus recommended in the OUT_SMFA file or more or less. You can specify more cpus, if that is convenient for your system. Then (to be safe) you should set the memory and scratch disk space to be appropriate for ncpus calculations, each as large as that in LARGE.com, running independently. The time limit should be set larger than the time required for one calculation on LARGE.com (as there are a lot of other things for SMFA to do).

Note again that you must include instructions in pbsfile to load the relevant ab initio quantum chemistry package. This package may be one of GAMESS(US), GAUSSIAN, NWChem or QChem. In addition, the DALTON package must always be loaded.

2.8 Electronic Structure Calculations

Item 5 in the main menu is

5) `Run all the electronic structure calculations`.

When you choose this item, the following submenu appears:

```
                    submenu: Run Options

    1) Run all calculations sequentially
    2) Run all calculations in parallel
    3) Exit run options
```

You choose one of these options. Then all *ab initio* or DFT calculations are prepared and submitted for execution. The process is automatic. The SMFA program will terminate with the message that (when the ab initio calculations are complete) the output will be appended to OUT_SMFA.

2.9 Restart electronic structure calculations

The multitude of electronic structure calculations may have been aborted prematurely for a number of reasons: For example, a hardware fault occurred; or insufficient memory or disk space was requested, causing the quantum chemistry program to abort; or the user chose to abort the process for some reason.

Many individual calculations may have been executed successfully before the job aborted, and the user may choose to restart the calculations, beginning with the first incomplete task in the sequence.

The restart facility DOES NOT APPLY to geometry optimisation or TS search or a scan, but only to calculation of the energy, gradient or frequency (and any properties requested). For optimisation etc, you should store any optimised energies and geometries (see Section 3.6), and re-run the optimisation or scan using the last 'Newcoords.xyz' geometry as the starting geometry.

You CANNOT restart the quantum chemistry calculations in SMFA, if you have changed (or want to change) anything related to the molecular structure or bonding, or the fragmentation parameters, or the electronic structure method. You must change the input and go through the usual sequence of steps to run the electronic structure calculations.

If the partially complete job was running in sequential mode, then you MUST restart in sequential mode; if originally in parallel mode, then you MUST restart in parallel mode.

To restart the calculation, launch SMFA, and then go straight to Item 6 in the main menu, which is

```
6) Restart the electronic structure calculations
```

When you choose this item, SMFA responds with

```
You can restart the SEQUENTIAL or PARALLEL quantum chemistry calculations in SMFA
if you have suffered a hardware failure, or similar catastrophe.

Before you restart, you can change the time, memory or disc space requests.

For SEQUENTIAL calculations, you can change the time, memory or disc space requests,
by editing the pbsfile directly, or via main menu item 4, and then come back here
to perform the restart.

For PARALLEL calculations, you can change the time, memory or disc space requests,
BUT NOT THE NUMBER OF PROCESSORS, by editing the pbsfile directly, or via main menu
item 4, and then come back here to perform the restart.


You CANNOT restart the SEQUENTIAL or PARALLEL quantum chemistry calculations in
SMFA if you have changed anything related to the molecular structure or bonding,
or the fragmentation parameters, or the electronic structure calculation.
```

You CANNOT restart a GEOMETRY OPTIMISATION or SCAN using this facility
You should store any optimised energies and geometries, and re-run the
optimisation or scan using the last 'Newcoords.xyz' geometry as the
starting geometry.

To exit this menu item, hit RETURN

To continue, enter RESTART

If you do not want to proceed with a restart at this time, you simply hit the RETURN key.

If you wish to initiate a restart then, you enter RESTART (actually not case sensitive).

SMFA will respond with:

If the calculation was running SEQUENTIALLY, you must continue that way.
If the calculation was running in PARALLEL, you must continue that way.
Enter S or P to continue as SEQUENTIAL or PARALLEL

If you enter S or P, then SMFA will continue with the restart, otherwise the restart will abort.

If you are running sequentially, SMFA will respond by first re-examining the input and printing comments to OUT_SMFA (the previous OUT_SMFA file is overwritten). You have to hit the RETURN key when this step is complete. SMFA then recalculates the fragmentation. You have to hit the RETURN key when this step is complete. SMFA then submits the ab initio calculations, and exits. When the calculations are complete, the output (OUT_SMFA) should be indistinguishable from that of an un-interrupted job.

If you are running in parallel, then SMFA submits a job which simply restarts the ab initio calculations beginning with the jobs that were not complete and then proceeds to append the output to OUT_SMFA.

3. Output Guide

The principal output file for SMFA is called OUT_SMFA.

3.1 Checking the input

The first section of this file contains a summary of the input parameters entered by the user. This is followed by a description of any formally charged groups that SMFA has found within the input structure. If metals were discovered, that had not been specified in the input (section 2.4.5), then SMFA warns the user that these charges should be specified. The coordinates of all formally charged groups are written to OUT_SMFA.

NOTE: The user should examine the geometries of any such charged groups. The coordinates printed in OUT_SMFA are in a format easily "pasted" into an appropriate graphics program like Jmol or MacMolPlt (or loaded into VMD). SMFA identifies formal charges on the basis of bonding and valency. If SMFA identifies unphysical charges, this is likely due to a defect in the input geometry. For example, if the structure has been obtained from a crystallography data base, then hydrogen atoms may be missing from the structure and spurious charges will then result. Other typographical errors may have resulted in unphysically short or long atom-atom distances which will result in mis-assignment of the bonding and hence mis-assignment of formal charges. If the structure is not "normal", in the sense that some atom-atom distances are neither typical for bonds nor for "non-bonded interactions" (for example the geometry signifies some intermediate point on a reaction path), then the charge of some atoms may need to be specified explicitly by the user (see Section 2.4.5). The user should check this part of the output at least once.

3.2 Fragmentation output

The coordinates for the fragments produced at the requested value of Level are written to the file seefrags. This file is in a standard xyz format which can be read by many molecular graphics programs (eg MacMolPlt or VMD). A brief summary of the number of fragments, and the number of electrons in the largest fragment is appended to OUT_SMFA.

An ab initio input file for the largest fragment is created with the file name "LARGE.com" or LARGE.inp or LARGE.nw, whichever suffix is usual for the quantum chemistry package. This file can be used to run trials to estimate the cpu time, memory and disk space requirements. Note, if you are running a geometry optimization or scan, then the type of job specified in LARGE.com will be an optimization. You should change this to a gradient calculation for the purpose of a trial calculation.

3.3 Output Energies

When the ab initio calculations have been completed, SMFA collects the multitude of fragment energies, the results of long-range interactions (calculated using perturbation theory) and prints the resultant total electronic energy to OUT_SMFA.

<u>Error messages</u>

The total electronic energy is evaluated from many individual fragment calculations. One or more of these component electronic structure calculations may have failed; either due to a hardware or system error, or for example, inadequate memory or disk space. If SMFA thinks a calculation has failed to complete satisfactorily, it prints a message to OUT_SMFA like

```
calculation may have failed for charge.49.grp.com
```

or

```
calculation may have failed for FRAG34.com
```

or some other ".com" filename, such as ab.19.85.com, nb.140.0.com, or nb.34.0-polar.com, where the numerals in the filename can take many different values. If such a message appears in OUT_SMFA, the user should inspect each of the corresponding ".log" files (eg `FRAG34.log`) to check if the electronic structure calculation did indeed fail.

If an individual calculation did FAIL, then the value of the energy reported in OUT_SMFA is INCORRECT.

In this case, you may need to repeat the entire calculation, if the error was caused by an error in the input. However, if the error was caused by a hardware or system error, or if an increase in the allowed memory or disk space in the pbsfile file (see Section 2.7) would solve the problem, then it may be possible to repeat only those component calculations that failed. See Section 2.9 on restarting calculations.

3.4 Output Gradients

If you have chosen to evaluate the forces or gradients, the gradients are written to a file called combinedderivs, under the heading "First derivatives". The order of the derivatives is x,y,z for atom 1, then x,y,z for atom 2, and so on. This file also contains the atomic masses, elemental symbols, energy, coordinates, and energy second derivatives (if they were requested). A brief note is appended to OUT_SMFA.

NOTE: Any subsequent gradient, frequency or optimisation process will overwrite combinedderivs.

3.5 Output Frequencies

If you have chosen to calculate the frequencies, these frequencies are written to a file called FREQUENCIES, along with the Infrared intensity for each mode. The Cartesian displacements of the atoms for each normal mode are written to a file called NORMAL_MODES. A simulated Infrared spectrum is written to a file called SPECTRUM. The Cartesian hessian is written to a file called combinedderivs, under the heading "Upper triangle of the second derivatives". The rows and columns of the hessian are in the order: xyz for atom 1, then xyz for atom 2, and so on.

The simulated spectrum assumes Lorentzian peaks with a FWHM of 5 cm^{-1}. You can produce another simulated spectrum with a different FWHM using the "Frequencies" utility program (see Section 5). Using this utility, you can also evaluate frequencies, intensities and spectra for **isotopic substitutions**.

NOTE: The zero-point energy is written at the bottom of the FREQUENCIES file and is also written to the OUT_SMFA file.

3.6 Optimised Structures

A brief description of the methodology used to find minimum energy and transition state (saddle point) geometries is provided in Appendix C. The criteria for convergence are also described therein.

Converged structures

When a structure has converged to a minimum or saddle point, the geometry is written to a file called CONVERGEDCOORDS. This file is in standard xyz format (as are input coordinate files). A summary of gradients and energies at each step in the optimisation process is contained in a file called optout. A brief statement noting convergence is appended to the OUT_SMFA file.

NOTE: At present, frequencies are not automatically calculated for the converged structure. Hence, if you want these frequencies, you must carry out a separate calculation for frequencies with the converged structure as the input geometry.

Unconverged structures

If the structure has not converged by the maximum allowed number of steps, the current (final) structure is written to a file called Newcoords.xyz. Moreover, the sequence of structures at each step in the optimisation process are written to files called Newcoords.xyz.step where "step" is an integer denoting the step number.

The optimisation process can be easily restarted by putting any of these "Newcoord.xyz" files as the initial geometry and running SMFA again.

NOTE: All files with names beginning with "Newco" are removed at the beginning of the optimisation process, so you must rename any such files you wish to retain, including that nominated as the new initial structure.

NOTE: As described in Appendix C, the hessian, and energy and gradient, are calculated at the initial geometry at the Hartree Fock level of theory, while at subsequent steps the energy and gradient are evaluated at the chosen level of theory. Hence, you should not be perturbed by the fact that the energy (reported in optout and OUT_SMFA) changes abruptly between the initial step (step 0) and the first step.

3.7 Scans

A scan consists of a sequence of geometry optimisations at discrete points along a path in the molecular coordinate space. The output consists of

(1) A summary of the input to SMFA is included in OUT_SMFA, including a description of the sequence of constraints which define the path.

(2) If the geometry optimisation completes within the input maximum number of steps, the geometry is saved in a file denoted SCANcon$k$, where k denotes the kth point along the path. If the convergence criteria have not been met within the input maximum number of steps, the geometry is saved in a file denoted SCANuncon$k$.

(3) Values of the molecular energy and constraint conditions during geometry optimisations at each point on the scanned path are contained in OUT_SMFA.

(4) A summary of the energy at the final geometry at each point along the path is given a the end of the OUT_SMFA file.

(5) A summary of gradients, displacements and energies at each step in the optimisation process at point k is contained in a file called optout.k.

The sequence of intermediate geometries during each optimisation (in files denoted Newcoords.xyz.step are overwritten by the corresponding files at the next point along the path.

4. Suggested Procedures

It is important to remember that SMFA is not exact for the calculation of molecular energies or other properties. However, the systematic character of the method allows the user obtain a good indication of the accuracy obtained, relative to a calculation on the whole molecule.

<u>Feasible, reliable calculations</u>

Of course, an MP2/cc-pVDZ or even a CCSD(T)/pV5Z calculation of the molecular energy is not exact. Indeed, we care little about the exact total energy of some molecular structure, we are almost always interested in the relative energies of two or more structures (eg reactants and products of some reaction). One tries to judge convergence towards the exact relative energy (or other property) from a sequence of calculations using progressively higher levels of theory and progressively larger basis sets, in the usual practice of quantum chemistry. Established practice would normally apply this systematic approach to a some moderately small system containing the chemistry of interest; a system small enough for larger basis sets and higher levels of theory to be <u>feasible</u>. This allows us to "benchmark" a more modest method that is feasible for larger systems, but that gives <u>reliable</u> results compared to the most reliable methods for the small system.

This established procedure is still applicable using SMFA. Generally, the SMFA estimates of relative energies (and other properties) can be expected to be <u>reliable</u> for Level = 3 or 4 (more about this later). For Level = 3, the largest fragments (the largest "molecules" for which ab initio calculations are performed) can be expected to contain between 4 and 6 chemical functional groups. For Level = 4, the largest fragments can be expected to contain between 5 and 7 chemical functional groups. The size of the largest fragment establishes what level of ab initio theory and basis set is <u>feasible</u> using SMFA. You can find information about the largest fragment in the OUT_SMFA file **after** you have fragmented the molecule (main menu item 3), that is **before** any ab initio calculations have been performed. A sample quantum chemistry input file is provided for this fragment in a file denoted LARGE.com (or LARGE.inp or LARGE.nw).

Hence, before any ab initio calculations have been carried out, you can see what the largest calculation will be and therefore you could estimate what level of ab initio theory and size of basis set will be <u>feasible</u> for your application of SMFA, given the available computational resources. You should consider what is feasible for Level = 3 and Level = 4. Hopefully (but not certainly) you will not need to use much higher values of Level.

You use different values of Level to establish convergence of energies and properties with respect to the fragmentation approximation.

Extensive testing indicates (thus far) that the reliability of SMFA at a given value of Level does not depend significantly on the level of ab initio quantum chemistry or the size of the basis set. This means that if SMFA estimates the relative energy of two molecular structures to within a few kJ mol^{-1} of the calculated whole molecule value at (say) HF/6-31G, then one can expect about the same accuracy at CCSD(T)/aug-cc-pV5Z.

Hence, you can test convergence with respect to Level, by running SMFA with (say) HF/6-31G for Level = 2, 3, 4, and so on. You should see the reported values of the relative electronic energies converging in the series of results (albeit not monotonically).

Once you have established the smallest value of Level that gives results that are reliable to within some desired tolerance, then calculations can be carried out at this value of Level with larger basis sets and higher levels of ab initio theory, in the usual way.

Hence, a sensible approach to the study of a large molecule would be:

(i) Choose a low level of theory and small basis set, say HF/6-31G.

(ii) Set the cutoff value for non-bonded interactions, denoted d_{tol}, to the default value of 1.1 (or 0.0 for Level 2, see Section 2.4.1 and Appendix B)

(iii) Calculate the energy for Level = 2, 3 and 4 to see if the energy has converged (try Level = 5 to be certain).

(iv) Repeat step (iii) with larger values of d_{tol}, say as large as 1.5. Record the computation times to understand the impact of increasing the values of Level and d_{tol}.

(v) Determine from these calculations what is a reliable and computationally efficient value of Level and d_{tol}.

(vi) You can now calculate the energy with these values of Level and d_{tol}, using more reliable combinations of quantum chemistry method and basis set. Using the largest fragment as a guide, you can determine what quantum chemistry methods are feasible with the available resources.

(vii) The approach to optimization of the geometry or searching for a saddle point is the same as for small molecules. It is more efficient to determine the structures with modest basis sets and levels of theory first, then to refine the structure with more reliable methods.

(viii) Having found a stationary point, evaluate the vibrational frequencies, infrared spectrum, and any other properties of interest (see Sections 5 and 6).

It should not be necessary to repeat steps (i) to (v) at another structure of the same molecule. However, one might want to confirm the convergence (with respect to Level and d_{tol}) of relative

energies, by increasing the value of Level by 1, compared to the "reliable" value determined above. Fairly extensive experience suggests that the fragmentation approximation may be converged by Level = 3 (not always), but that calculations at Level = 4 are necessary to confirm this.

Increasing the value of d_{tol} from 1.1 to 1.5 generally does improve the accuracy of calculated energy gradients [14] (and frequencies), and final optimization of minimum energy or saddle point structures (and frequencies) might be carried out with d_{tol} = 2. SMFA normally uses the default values for convergence of the SCF calculation (for the particular quantum chemistry package), but employs a tighter convergence criterion if gradients or frequencies are requested.

There have been a number of published reports of the accuracy achieved by SMFA [9,10] [14] for the calculation of energies for moderate sized molecules. Note, once again, SMFA is not exact (but impressively accurate).

In applications of quantum chemistry to <u>large</u> molecules, one should ask oneself: what is a sensible level of accuracy to aim for? [17] Under normal circumstances, the molecule of interest will be in (or near) thermal equilibrium, with an average total energy (including vibration and rotation) far above the energy of the equilibrium geometry. An attempt to determine the equilibrium structure and energy with higher and higher accuracy is likely to make less and less sense.

Saddle points and Scans

Both the optimisation and saddle point searches in SMFA [options Optimise (3) and Find TS (4) in section 2.4.3] are carried out using a local quadratic approximation to the molecular potential energy surface.

In the case of optimisation, this will lead to an energy minimum "close" to the initial geometry. It will not necessarily lead to the global energy minimum.

In the case of saddle point searches, common experience suggests that locating the saddle point that defines the transition state of interest is difficult unless the initial geometry is quite close by. In large molecules, it may be very difficult to "guess" an appropriate initial geometry. The **scan** facility [Option Scan (5) in section 2.4.3] provides a means to explore the molecular potential energy surface and determine a suitable initial guess for a saddle point structure.

For example, if one knows a bond that breaks in the transition from reactants to products, then one can run a scan with increasing values of this bond. One might find an energy maximum along this scanned path, and the geometry at that maximum could be a suitable initial guess at the saddle point geometry. NOTE: In such a scan, the "breaking bond" will eventually exceed the length that SMFA accepts for a bond to exist. At this point the fragmentation of the molecule will be different from that in the previous scan step. Hence, the calculated energy will change "discontinuously". To prevent this,

you can use the option `4) Specify hydrogen bonding and any unusual bonding` in the input control submenu to specify that the "breaking bond" should always be taken as a bond, regardless of whether SMFA would normally consider this to be a bond.

5. Utilities Guide

In the main menu, choosing

`6) Utilities"`

brings up another submenu

```
                    submenu: Utility Programs

->      1) Frequencies with isotopic substitution
        2) Isodesmic, homodesmotic and analogous reactions
        3) Combining isodesmic, homodesmotic etc reactions
        4) Electrostatic potential on the solvent-accessible-surface
        5) Dipole Polarizability
        6) Dipole Hyperpolarizability
        7) Internal Coordinates
        8) Add H atoms
        9) Exit
```

5.1 Frequencies with Isotopic substitutions

The purpose of this utility is to allow the user to recalculate the vibrational frequencies with isotopic substitutions for some atoms. The utility of this feature is that Infrared spectra of large molecules are rather diffuse and unresolved into individual peaks, due to the large number of vibrational transitions of similar frequencies. However, if the spectrum of an isotopically substituted molecule is also available, then the difference between the substituted and unsubstituted spectra may provide useful information.

NOTE: The user must first have run a frequency calculation for the molecule (that is job type 2 in the input submenu item 3). A normal frequency job produces the force constant matrix (hessian) in a file called `combinedderivs` and the electric dipole moment derivatives in a file named `combDipDerivs`. This utility needs these files, so the utility should be used in the subdirectory where the frequency job was executed.

The atomic masses can be changed from their default values in two ways:

(i)   You can choose to change the masses of all atoms that have the same chemical element (as chosen by the user), or

(ii)   You can choose to change the masses of particular atoms (one or more atoms).

The new mass must be given in atomic mass units (gm mol^{-1}), so for example ^{13}C would be given a mass of 13.00335.

Choosing

1) Frequencies with isotopic substitution

then brings up the question:

You can choose to change the isotope for every atom of a chosen element, or you can change the isotope for particular chosen atoms (by number).
Do you choose a particular element (Y/N)?

Enter Y or N

If you enter "Y", this produces the prompt:

Enter the elemental symbol and mass (in atomic mass units)

Then for example, you enter

C   13.00335   (or simply 13.)

You must leave at least one space between the element and the mass. Note that SMFA requires correct elemental symbols, eg magnesium is Mg, not MG or mg.

If you enter "N", this produces the prompt:

For each atom to be substituted by an isotope, you must enter the atom number and the isotope mass (in atomic mass units). Enter the number and mass for each atom on one line, and finish with a RETURN (blank line):

You enter, for example

27   13.0

54   13.0

123   2.0

which would mean that atoms 27 and 54 were replaced by ^{13}C and atom 123 by deuterium. The atom number refers to the order in the input coordinate file. Either answer to the question above will then produce the following prompt:

The IR spectrum will be simulated with Lorentzian peaks with a given
full width at half maximum (FWHM). The default value of this width is 5 cm-1
Enter your chosen width or hit RETURN for the default:

You enter the peak width you wish (in cm^{-1}) eg

10.0

The value chosen should reflect the appropriate resolution for the experimental data with which you wish to compare the calculated spectrum. The program then calculates the data and reports that

The isotope substituted frequencies are in file FREQUENCIES
The Cartesian displacements for the normal modes are in file NORMAL_MODES
The simulated spectrum is in file SPECTRUM

NOTE: The files FREQUENCIES, NORMAL_MODES and SPECTRUM will overwrite pre-existing files with these names, so you should have saved the "normal isotope" results under some other name.

The vibrational frequencies are reported in cm^{-1}. The intensities for each vibrational mode are calculated as the square magnitude of the derivative of the dipole moment vector with respect to the normal coordinate, multiplied by the transition frequency. The spectrum is simulated simply as a sum of Lorentzian peaks with a height proportional to the intensity (as discussed above) and the FWHM chosen above.

The format of the SPECTRUM file is two columns, frequency (cm^{-1}) and intensity. It is a simple matter to paste this data (or import the file) into simple plotting programs or spread sheets (eg Excel). Once you have established this data in the spreadsheet, one can simply calculate the "difference spectrum" and plot any of the original or difference spectra.

NOTE: You can get spectra for the unsubstituted molecule with different FWHM, simply by entering "N" to the first question; hitting RETURN in response to "For each atom to be substituted..." and then entering the FWHM value of choice.

NOTE: The vibrational frequencies and intensities are calculated in the harmonic approximation for ground to first vibrationally excited state transitions only. Therefore, both the reported frequencies and the simulated spectrum do not account for anharmonicity or for overtone and combination bands which will almost certainly be present in any experimental spectrum.

5.2 Isodesmic, homodesmotic and analogous reactions

Background

An isodesmic reaction [20] is one in which the number and type of chemical bonds is the same for both reactants and products. For example, reactants and products have the same number of C-C, C-H, C=O bonds, etc. Since, the heat of formation of molecules is mostly determined by the number and type of bonds, the heat of reaction is near zero for an isodesmic reaction. This fact allows one to estimate the heat of formation of one species in the reaction, if the heat of formation of all other species is known.

A homodesmotic reaction [21] is similar, except that in addition to the same number and types of bonds in both reactants and products, the neighbouring substitutents of those bonds are the same. Given an even closer correspondence in bonding for reactants and products, a homodesmotic reaction has a heat of formation which is even more likely to be near zero.

As it turns out, SMFA naturally produces instances of such reactions. In SMFA, the fragmentation of a molecule M is written as

$$M \rightarrow \sum_{n=1}^{N_{frag}} c_n F_n \qquad (5.2.1)$$

where the c_n are integer coefficients and the F_n are fragment molecules. Some of the coefficients are negative. If we subtract these terms from both sides of (5.2.1), we can write

$$M + \sum_{c_n < 0} (-c_n) F_n \rightarrow \sum_{c_n > 0} c_n F_n \qquad (5.2.2)$$

In SMFA, all the structures in the fragments are the same as their corresponding structure in the molecule. However, if we let all structures in (5.2.2) take their equilibrium distributions, then Eq. (5.5.2) represents a chemical reaction at equilibrium.

If the fragmentation is performed at Level = 1, then the reaction is isodesmic; if Level = 2, then the reaction is homodesmotic. Some years ago [1], we labelled the corresponding reaction at Level = 3 "isoperiochic", but the label hasn't stuck.

Implementation

SMFA allows the user to derive an isodesmic, homodesmotic, or higher analogue reaction for their molecule of interest, even without doing any quantum chemistry calculations.

However, SMFA relies on the "free" software Openbabel to convert the Cartesian coordinates of a molecule to its unique InchI (the IUPAC International Chemical Identifier). So you will need to have installed and loaded Openbabel on your computer (eg `module load openbabel/2.3.2`).

You then simply launch SMFA and in the main menu:

Select

`->    1) Set up input`

    (a) Enter the Level of Fragmentation (if you have not already done so);

    (b) Enter the filename containing the molecule coordinates (if you have not already done so);

    (c) Enter something under "Specify the electronic structure calculation", anything will do, as we are not performing any ab initio calculations

    (d) Under "`Specify hydrogen bonding and any unusual bonding`",

        answer "N" to all the questions except,

        answer "Y" to "Do you want to ignore hydrogen bonds (Y/N)?", and

        answer "Y" to " By default SMFA assumes that if two charged groups...."

    (e) Exit input set up.

```
    Select

2) SMFA examines the input and prints comments or queries to OUT_SMFA

    Select

3) Fragment the molecule

    Select

5) Utilities
    Select

2) Isodesmic, homodesmotic and analogous reactions
```

SMFA will complete the task and append the results to the OUT_SMFA file. The reactant and product molecules will be identified by their fragment number in the seefrags file, and by their InchI. The file seefrags contains the Cartesian coordinates of all fragments, while seefrags.inchi contains the corresponding InchI.

In addition, SMFA produces 8 other files to describe the reaction. Let's suppose the file containing the molecule coordinates (see step 1b above) was called molecule.xyz. The 8 new files created by SMFA are:

molecule.xyz.lhs.inchi

molecule.xyz.lhs.coeff

molecule.xyz.lhs.coords

molecule.xyz.lhs.svg

molecule.xyz.rhs.inchi

molecule.xyz.rhs.coeff

molecule.xyz.rhs.coords

molecule.xyz.rhs.svg

Here molecule.xyz.lhs.inchi contains a list of the Inchi for each compound on the left-hand-side of the reaction (5.5.2). The first Inchi is that of the original molecule. The file " molecule.xyz.lhs.coeff" contains the coefficients, $-c_n$, of the compounds on the left-hand-side of the reaction, " molecule.xyz.lhs.coords" contains the Cartesian coordinates of these compounds, and " molecule.xyz.lhs.svg" is a file that you can open with a web browser like FireFox. The web browser will display a table of 2D chemical structure drawings for the compounds on the lhs of reaction (5.5.2).

The corresponding files labelled with ".rhs" contain the same information for the right-hand-side of the reaction.

NOTE: The 2D chemical structure drawings (.svg files), created by OpenBabel and viewed in FireFox, should be treated with some caution. It appears that these drawings show double bonds in the wrong places and radical sites where no such radicals exist (due to misplacing the double bonds). Nonetheless, with the aid of some knowledge of chemistry, the 2D drawings provide a helpful illustration of the reactants and products.

Additional manipulations of near-isoenergetic reactions can be carried out using the utility in the next section.

IMPORTANT CAVEAT: The basis of isodesmic, homodesmotic, etc reactions ignores long range interactions between functional groups in a molecule (assuming that "bonded" and "near bonded" effects are the major contributors to heats of formation). If long range electrostatic interactions are large then this approach will not provide an accurate estimate of the heat of formation.

5.3 Combining isodesmic, homodesmotic etc reactions

If you have created the files for more than one isodesmic, homodesmotic (etc) reaction, you can use this utility to 'subtract' one reaction from another. For example, you may have found an isodesmic or homodesmotic etc reaction for each of two isomers. Since both reactions are near isoenergetic, a new reaction that we form by subtracting one reaction from the other is also near isoenergetic. In this example, this new reaction could be used to estimate the energy difference between the two isomers. If the isomers have many similarities in structure, then 'subtracting' one reaction from the other will result in the cancelation of many components of the reactions. Thus a simpler reaction for estimating the energy difference between the two isomers will be obtained. Many other applications for constructing near isoenergetic reactions may come to mind.

Choosing

```
3) Combining isodesmic, homodesmotic etc reactions
```

will result in some comments and two prompts:

```
The isodesmic (etc) utility created several files for the reactions with names like
name1.lhs.inchi and name2.lhs.inchi
where name1 and name2 were the names of the coordinate files for these molecules.

All you have to do is enter these two file names, eg name1 and name2

Enter the first file name
```
You enter the filename for molecular coordinates previously used in Section 5.2 (say name1)
```
Enter the second file name
```

You enter another filename for molecular coordinates previously used in Section 5.2 (say name2)

The files for the reaction formed by subtracting the reaction for name2 from the reaction for name1 are written to files with the prefix name1-name2. These files, like "name1-name2.lhs.inchi" have the same form as corresponding files created in Section 5.2. So, these output files can form the input for further manipulation of reactions via this utility.

5.4 Electrostatic potential on the solvent-accessible-surface

SMFA will generate the data to allow you to visualise the electrostatic potential (ESP) on the solvent accessible surface surrounding the molecule. The electrostatic potential is generated from a charge, and dipole and quadrupole moments distributed onto every atom in the molecule. These charges and electrostatic moments are generated from the electron densities that were generated for Level = 1 fragments, using Stone's method applied to the output from GAUSSIAN or QChem. Similarly, Stone's method is integrated into GAMESS, and the corresponding moments are used. For NWChem, only distributed charges on the atoms are available for use. See Appendix E for more details about the implementation of distributed multipole moments. If you have used SMFA to carry out a calculation of the energy, gradient or hessian (at any value of Level), then the necessary data is available to SMFA.

The solvent accessible surface is defined as the surface that is separated from the nearest atom in the molecule by the Van der Waals radius of that atom plus a radius associated with the solvent molecule.

This utility program provides two output files for graphics.
The first produces a file called SOLACCSURFACE. This is an 'xyz' format file which can be loaded by programs like VMD or MacMolPlt to show you the solvent-accessible-surface surrounding the molecule. This graphic is useful in conjunction with the second file, called ESP.cube. This file has the format of a GAUSSIAN 'cubegen' file, and can be read by VMD (and other programs) to show the electrostatic potential on the solvent-accessible-surface.

NOTE: To use this utility, the current directory must be that in which you have just carried out an energy, or gradient or hessian calculation for the molecule of interest. The electron density for the molecule which generates the ESP is a Level =1 approximation to the electron density at the level of ab initio theory and basis set which was employed. For QChem and GAMESS, only the SCF density is used. If you are only interested in the ESP, you can save cputime by choosing Level = 1 and d_{tol} = 0.0 in the SMFA input, and choosing only an energy calculation.

In the Utilities subdirectory, choosing

3) Electrostatic potential on the solvent-accessible-surface

brings up three prompts:

You must enter a value for this solvent radius (the default is 1.4 Angstrom)

You enter a value (such as 1.4, which is commonly used for H_2O).

You must enter the desired density of points on the surface (per square Angstrom)
 The default value is 1.0

You enter a value such as 1.0.

When you look at the graphics (see below), you might decide to try different values.

You must enter the number of grid points for the 'cube' of ESP data,

The recommended value is 100 (per axis direction)

You enter a value such as 100.

The program will produce a rectangular prism of data points, with 100 points on the longest axis, and proportionally less on shorter sides. This box will enclose the molecule by a clear margin on all sides. The resultant output will be some MB in size. The larger integer you select above, the longer will be the calculation time (expect minutes), the larger will be the ESP.cube file, and the better will be the graphical resolution.

A brief note on the output is appended to OUT_SMFA.

Output

The file SOLACCSURFACE is a simple xyz format file which you can open in free-ware such as VMD, MacMolPlt, iMol or similar programs for representing molecules. If you choose a "ball and stick" representation, then this file shows the molecule surrounded by small off-white balls which lie on the solvent-accessible-surface for the radius of solvent molecule that you have chosen.

The file ESP.cube can be opened by VMD. It will show a coloured surface with represents the electrostatic potential on the solvent-accessible-surface (blue for positive, red for negative). To see this in VMD, you should

(i) Launch VMD. A graphics window will open and also a window labeled "VMD Main".

(ii) Under the "File" menu in the "VMD Main" window, choose "New Molecule". A new window will appear called "Molecule File Browser". In this window, Click "Browse". Locate the "ESP.cube" file, and choose it.

(iii) In the "Molecule File Browser" window, under "Determine file type", choose "Gaussian Cube" (if VMD has not already done so). Then click "Load". A new line will appear in the "VMD Main" window, with the name of your file (eg ESP.cube).

(iv) In the "VMD Main" window, choose "Representations" under the "Graphics" menu. A new window will appear, entitled "Graphical Representations".

(v) In this "Graphical Representations" window: under "Drawing Method", choose "Solvent"; next to "Representation Method" choose "Mesh"; next to "Probe Radius" choose a value like 2.8. The coordinates in ESP.cube are in atomic units (Bohr). The ESP is also in atomic units. The "Graphical Representations" window should look like this:
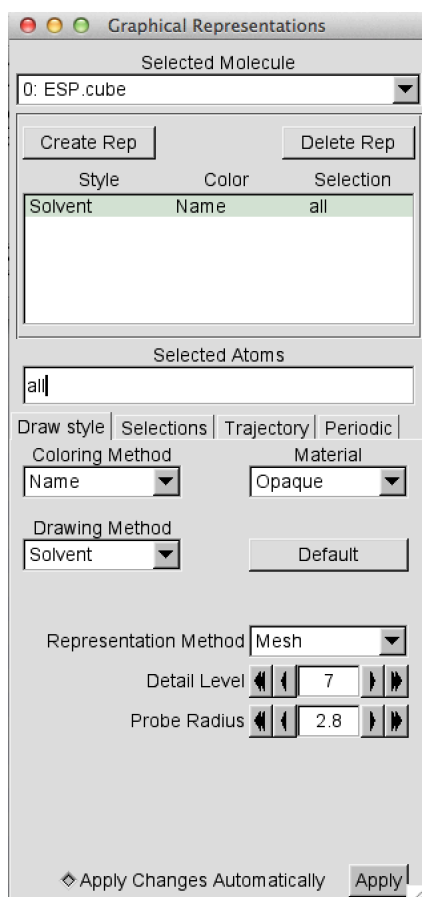
You can click "Apply".

(vi) A coloured mesh representation of the ESP on the surface (with the solvent molecule or "probe" radius chosen) will appear in VMD's graphics window (blue for positive, red for negative). You can fiddle with the settings in VMD to find the image you like best. You can change the value of the "Probe Radius" to see what effect that has on the ESP.

The image that is produced from the SOLACCSURFACE file is useful for comparison with the ESP surface, as SOLACCSURFACE shows the molecule itself within the solvent-accessible-surface.

5.5 Dipole Polarizability

SMFA estimates the electric dipole polarizability of the molecule using two methods, from data that may already have been evaluated by SMFA.

<u>Method 1</u>

If the molecule is neutral, or if long-range dispersion has been accounted for in the calculation of the energy, then the polarizability of the *individual chemical functional groups* will have been evaluated by SMFA. For GAMESS and NWChem the polarizabilities are actually evaluated using DALTON. The total molecular polarizability is then simply estimated as the sum of the group polarizabilities.

This is likely a significant overestimate of the actual value, as many "capping" hydrogen atoms have contributed to the sum (atoms that are not in the actual molecule). However, this method is included here, because it is available for all quantum chemistry programs interfaced with SMFA.

When you choose this utility the results for Method 1 are appended to OUT_SMFA.

<u>Method 2</u>

For GAUSSIAN, the dipole polarizability is automatically evaluated if the frequencies have been calculated. Alternatively, if the keyword POLAR (not case sensitive) has been included in Section 2.4.3.B (see `Enter any other keywords (optional), or hit RETURN:`) then the dipole polarizability is calculated.

In these two cases, the molecular polarizability can be obtained as a sum over the fragment polarizabilities. This should be more accurate than Method 1, since the contributions of capping hydrogens approximately cancel in the sum over fragments, F_n:

If $\quad M \rightarrow \sum_{n=1}^{N_{frag}} c_n F_n$,

then $P \approx \sum_{n=1}^{N_{frag}} c_n P(F_n)$, for any property, P, including the polarizability.

The results are appended to OUT_SMFA.

For QChem, you must explicitly request calculation of the dipole polarizability by including

MOPROP 2

in the $rem input (see Section 2.4.3.D). Then the molecular polarizability can be obtained as a sum over the fragment polarizabilities. The results are appended to OUT_SMFA.

5.6 Dipole Hyperpolarizability

If frequencies have been evaluated (before hyperpolarizabilities are requested here and without some other subsequent calculation, eg of the energy only), then the molecular hyperpolarizability can also be obtained as a sum over the fragment hyperpolarizabilities (as above) using GAUSSIAN. This property has not been programmed for QChem as yet.

The results are appended to OUT_SMFA.

5.7 Internal Coordinates

This utility program enables the user to find the values of atom-atom distances, angles formed by three atoms and dihedral angles defined by four atoms (see Appendix C for the definition of dihedral angles). Atom-atom distances are reported in Angstrom, while bond angles and dihedral angles are reported in degrees. The geometry investigated is that contained in the coordinate file specified in Section 2.4.3.

NOTE: You can use this utility to find the current value of internal coordinates, when using constrained optimization or the Scan facility.

If you choose

7) Internal Coordinates

SMFA prompts

```
To find a bond length, enter the 2 atom numbers
To find a bond angle, enter the 3 atom numbers
To find a dihedral angle, enter the 4 atom numbers
To exit, hit RETURN
```

You enter a sequence of integers (atom numbers). For example, your input, and the consequent output, might look like:

13   12

```
       13          12     1.230775
```

or

```
   13  12  11
     13          12          11   120.844332
```

or

```
   13  12  11  14
     13          12          11          14  -117.820407
```

In these examples, the angle is the angle that the vector 12 --> 13 makes with the vector 12 --> 11; the dihedral angle is the angle between the planes (13 12 11) and (12 11 14).

5.8 Add H atoms

Sometimes coordinates which are reported from diffraction or NMR experiments have hydrogen atoms "missing". This may be due to the difficulty in determining the positions of H atoms in X-ray crystallography. Alternatively, an NMR based report may contain only a segment of the total structure and "terminal" atoms are reported with an incomplete valence - so it appears that a H atom should be appended to this terminus. The quantum chemistry calculations would be drastically perturbed if this "incomplete" structure were employed, so we naturally want to add the "missing" H atoms.

If there are hundreds of missing H atoms, you probably need to find another structure, or locate a program that can fix it.

If only a few H atoms are missing, then SMFA has a utility that can be used to add H atoms, one at a time.

IMPORTANT NOTE: You may only become aware of the fact that there are missing H atoms when OUT_SMFA reports the presence of charged groups at the time that it checks the input. You should use a graphics program (eg iMol, MacMolPlt, IQMol, VMD) to look at all the charged groups that OUT_SMFA reports, to check for missing H atoms. For each of these charged groups, OUT_SMFA reports the atom number for the first atom in each group. This will allow you to locate the position of the charged group (with a missing H atom) in a graphic of the total structure. To insert a H atom, you will need to look at this structure and estimate where the additional H atom should go (approximately).

If you choose

```
8) Add H atoms
```

SMFA responds with

```
To get coordinates for a H atom to replace one that is missing,
you will need to enter the atom numbers for three atoms in the structure:
The H atom will be bonded by 1 Angstrom to the second atom;
the first atom is taken to lie on an imaginary x axis from the first atom;
the third atom is taken to lie in an imaginary xy plane,
```

You enter the three appropriate atom numbers. As you can see, you are being asked to imagine an xyz axis system with the second atom at the origin. You will need to estimate (guess) the spherical polar angles that would define the position of the new H atom in this imaginary axis system.

You enter three atom numbers, eg

23 105 44

SMFA then responds with

Enter the two angles in degrees

you enter two angles, eg

-25.0 -80.0

Then SMFA responds with (eg)

H   1.395596   -0.671719   -0.109015

 Enter the 3 atom numbers, or enter RETURN to finish

You can enter RETURN to finish, or enter atom numbers for a second additional H atom, and so on.

Before you enter RETURN, you should copy the line above, which contains the H label and coordinates, so you can "paste" this into the coordinate file.

NOTE: When you have finished adding additional H atoms, you should repeat main menu item 2 to check that the geometry is now appropriate.

6. Write your own property

If you can write Unix, Perl or other scripts (SMFA is a mixture of Perl, Unix and Fortran), you can use SMFA to evaluate any property that GAMESS(US), GAUSSIAN, NWChem or QChem can produce. A little reminder of how SMFA evaluates the molecular energy is useful for explaining what you have to do. In SMFA, the molecular energy E is given by

$$
\begin{aligned}
E = \sum_{n=1}^{N_{frag}} c_n E(F_n) \qquad &\text{(A)} \\
+ \sum_i \sum_j c_{ij} \left[E\left(F_i^{(1)} F_j^{(1)}\right) - E\left(F_i^{(1)}\right) - E\left(F_j^{(1)}\right) \right] \qquad &\text{(B)} \qquad\qquad \text{(5.5.1)} \\
+ \mathrm{E}_{pert}
\end{aligned}
$$

where the c_n are integer coefficients and the F_n are fragment molecules, evaluated for the value of Level requested. The c_{ij} are also integer coefficients and $F_i^{(1)}$ are fragments evaluated for Level = 1. Term (B) accounts for non-bonded, but relatively close, interactions between functional groups, and E_{pert} is a small correction evaluated using perturbation theory (see Appendix E).

Any molecular property can be thought of as a derivative of the total energy in the presence of some applied field. So, any property can be written (to a good approximation) as

$$
\begin{aligned}
P = \sum_{n=1}^{N_{frag}} c_n P(F_n) \qquad &\text{(A)} \\
+ \sum_i \sum_j c_{ij} \left[P\left(F_i^{(1)} F_j^{(1)}\right) - P\left(F_i^{(1)}\right) - P\left(F_j^{(1)}\right) \right] \qquad &\text{(B)}
\end{aligned} \qquad\qquad \text{(5.5.2)}
$$

The first term (A) in (5.5.2) is overwhelmingly the largest contribution to the property value. Using term (A) alone was the method adopted in (5.5) and (5.6) for the polarizabilities. Details for implementing part (B) are contained in Appendix D.

To implement part (A) of (5.5.2) for any property, you simply have to

(i)   Include the appropriate "keyword" or other instruction in the input for the ab inito calculation, requesting the property you want (see Section 2.4.3 above).

(ii)   Follow the SMFA process down to and including running the ab initio calculations (Section 2.7)

(iii)   The output of these calculations for fragments F_1, F_2, and so on, are written to files called FRAG1.log, FRAG2.log, and so on. You must write a Unix script, Perl script, or similar, to extract the value of the property you want from these "log" files.

(iv)   The coefficients in (5.5.2 A), c_1, c_2, and so on, are contained in the file signs.out. Your script

simply has to take the property value from each FRAGn.log file, multiply it by the corresponding c_n and add up the result.

If you care to write your script in Perl, it can be incorporated into SMFA for everone to use!

Note also, that the derivatives of properties with respect to the atomic coordinates can be estimated using SMFA. From (5.5.2),

$$\frac{\partial P}{\partial x_\alpha} = \sum_{n=1}^{N_{frag}} c_n \frac{\partial P(F_n)}{\partial x_\alpha} \qquad \text{(A)}$$

$$+ \sum_i \sum_j c_{ij} \left[\frac{\partial P\left(F_i^{(1)} F_j^{(1)}\right)}{\partial x_\alpha} - \frac{\partial P\left(F_i^{(1)}\right)}{\partial x_\alpha} - \frac{\partial P\left(F_j^{(1)}\right)}{\partial x_\alpha} P\left(F_j^{(1)}\right) \right] \qquad \text{(B)}$$

$$(5.5.3)$$

If derivative properties, $\dfrac{\partial P(F_n)}{\partial x_\alpha}$, are evaluated by the quantum chemistry program for each fragment,

then the corresponding quantities for the whole molecule can be estimated. This requires knowing which atoms in the whole molecule are present in each fragment. This information is contained in the file frags.out. The details needed to evaluate (5.5.3) are also contained in Appendix D.

Properties that involve higher order coordinate derivatives can be evaluated in the same way. The NMR chemical shift for all the atoms in a molecule is an example of a property that can be evaluated from the fragment calculations. [15]

7. Test Cases/Examples

7.1 Relative energies of two protein conformers

The solution structure of a small peptide was investigated by NMR and reported in Ref. [22]. A total of 20 structures were deposited in the Protein Data Base (RCSB PBD) with the identifier (PDB ID) 1xv4. The 1st and 20th structures have been used in this test case, with Cartesian coordinates in files 1xv4_1.xyz and 1xv4_2.xyz, respectively, included in the directory SMFA/doc/testcases/1xv4. This small peptide contains 224 atoms.

The energy of these two structures, and the corresponding energy difference, has been determined at HF/6-31G(d,p) using GAUSSIAN09 with

Level	d_{tol}
2	0.0
3	1.1
4	1.1

The results are shown in Table 7.1.

Table 7.1 Energies and relative energies of the two 1xv4 conformers, at HF/6-31G(d,p).

Level	Energy (au) 1xv4_1	Energy (au) 1xv4_2	Relative Energy (kJ mol^{-1})
2	-5143.290365	-5143.227675	-164.59
3	-5143.292222	-5143.233856	-153.24
4	-5143.294725	-5143.236096	-153.93

The file OUT_SMFA_1xv4_1_Level3 is included in the test case directory as an example of the OUT_SMFA file. From this file, you can see that the amide bond has NOT been taken as a multiple bond, otherwise all parameters are standard. No dispersion is included for this Hartree-Fock calculation.

Note that the conformer energy difference is converged by Level = 3, although each conformer total energy varies between Level = 3 and Level = 4.

From the OUT_SMFA file, one sees that 14 cpus were recommended for parallel execution. Using 14 cpus on an Intel Xeon E5-2670, this calculation required 2002 secs total cpu time, and 234 secs walltime.

7.2 Energy gradient

The GAMESS-US program was used to calculate the B3LYP/cc-pVDZ gradient at a geometry of c-cyclotridecene. The molecular coordinates (in file R389225-3d.xyz), the OUT_SMFA file and the combinedderivs file are contained in the SMFA/doc/testcases/c-cyclotridecene subdirectory. The gradients (in Hartree/Bohr) are contained in the combinedderivs file. The cc-pVDZ basis set was implemented using the procedure outlined in Section 2.4.3A, following the phrase "For more complicated basis sets".

The cc-pVDZ basi set was also used to evaluate (using DALTON) the polarizabilities which are needed to evaluate the induction energy (negligible in this case).

7.3 Geometry optimization

A structure for c-cyclotridecene was optimized at HF/cc-pVDZ with Level = 3 and d_{tol} = 1.1. The initial geometry (file R389225-3d.xyz) together with the OUT_SMFA and optout files are contained in the SMFA/doc/testcases/opt_c-cyclotridecene/HF directory, along with the converged geometry in file CONVERGEDCOORDS.

The CONVERGEDCOORDS file was renamed optL3dtol1.1c-cyclotridecene.xyz and used as the initial geometry for optimization at MP2/cc-pVDZ. This file, along with the OUT_SMFA and optout files and the converged geometry in file CONVERGEDCOORDS, are contained in the SMFA/doc/testcases/opt_c-cyclotridecene/MP2 directory.

All MP2 calculations were carried out using the QChem program. 16 processors were employed in a parallel calculation (although the OUT_SMFA files shows that 23 processors were recommended).

For comparison, c-cyclotridecene was also optimized at HF/cc-pVDZ with Level = 3 and d_{tol} = 1.1, using the GAMESS(US) program. The corresponding files are contained in the SMFA/doc/testcases/opt_c-cyclotridecene/GAM directory.

7.4 Frequencies

The CONVERGEDCOORDS file for the MP2/cc-pVDZ (Level = 3, d_{tol} = 1.1) optimization above were copied to a file entitled optMP2L3dtol1.1c-cyclotridecene.xyz.

The MP2/cc-pVDZ frequencies for c-cyclotridecene were then evaluated using the QChem program. 16 processors were employed in a parallel calculation (although the OUT_SMFA files shows that 23 processors were recommended). The OUT_SMFA, FREQUENCIES, NORMAL_MODES, and SPECTRUM files are contained in SMFAPAC/doc/testcases/Freq/.

SMFA creates the combinedderivs file, which contained the hessian (inter alia), and the combDipDerivs file, which contains the Cartesian derivatives of the molecular dipole moment. Both these

files are needed to evaluate the vibrational spectrum. These two additional files are also contained in SMFA/doc/testcases/Freq/.

A deuterated spectrum

   To test the Utility program 5.1 "Frequencies with isotopic substitutions", this utility was used to substitute a mass of 2.0 for the two hydrogen atoms (atom numbers 36 and 37) which are attached to the "ene" group in c-cyclotridecene. Having chosen the Utilities in the main menu and "1) Frequencies with isotopic substitution" in the submenu, the following prompts from SMFA and responses were performed:

```
This utility evaluates the vibrational frequencies, intensities,
and spectrum for isotopically substituted molecules

The force constant matrix (contained in combinedderivs) and the
dipole derivatives that were previously evaluated for this molecule
in this subdirectory are used to evaluate the frequencies.

You can choose to change the isotope for every atom of a chosen element,
or you can change the isotope for particular chosen atoms (by number).
Do you choose a particular element (Y/N)?
n
For each atom to be substituted by an isotope, you must enter the atom number
and the isotope mass (in atomic mass units)
Enter the number and mass for each atom on one line, and finish with a
RETURN (blank line):
36  2.0
37  2.0


The IR spectrum will be simulated with Lorentzian peaks with a given
full width at half maximum (FWHM). The default value of this width is 5 cm-1
Enter your chosen width or hit RETURN for the default:
5.0

The isotope substituted frequencies are in file FREQUENCIES
The Cartesian displacements for the normal modes are in file NORMAL_MODES
The simulated spectrum is in file SPECTRUM
```
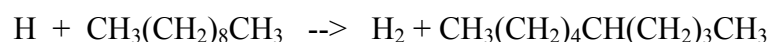
   The output files FREQUENCIES, NORMAL_MODES and SPECTRUM are contained in the SMFA/doc/testcases/Freq/deuterated subdirectory, renamed with a "_d2" suffix.

7.5 Scan

   To illustrate some features of the scan procedure, we consider the reaction of a hydrogen atom abstracting a hydrogen atom from a CH_2 group in decane:

$$H \; + \; CH_3(CH_2)_8CH_3 \; \; --> \; \; H_2 + CH_3(CH_2)_4CH(CH_2)_3CH_3$$

We begin with an initial structure in which a H atom approaches decane along the direction of a CH bond, as shown below.
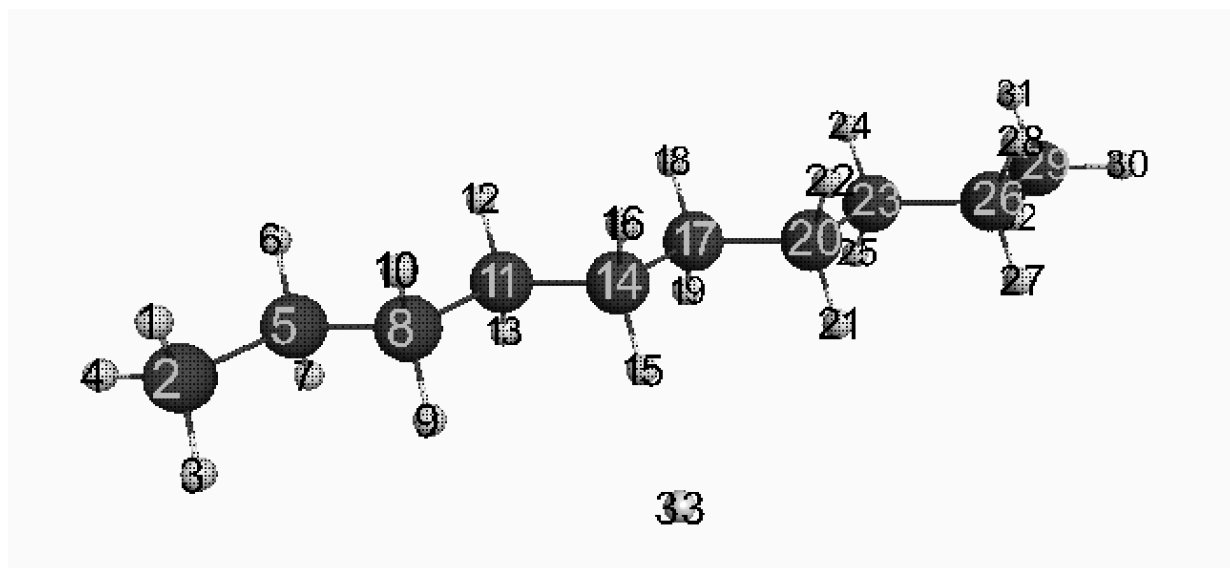
The coordinates (file H+decane_react.xyz) are contained in SMFA/doc/testcases/scan/reactants/. In this initial geometry, the incoming H atom is 1.644 Å away from the closest H atom in decane.

The scan will optimize the structure for a succession of decreasing H....H distances, in order to obtain an energy profile as a function of the H...H distance.

Reactant optimization

Before carrying out the scan, we have optimized the reactant geometry subject to the constraint that the H...H distance is 1.644 Å. The initial geometry, optimized coordinates, OUT_SMFA and optout files are contained in SMFA/doc/testcases/scan/reactants/. The two hydrogens involved in the constraint are atom numbers 15 and 33 in the coordinate file, and the OUT_SMFA file shows this constraint. The optimization was carried out using the NWChem program at HF/cc-pVDZ with Level = 3, and d_{tol} = 1.1. The input shown in OUT_SMFA has another feature to note.

When this reaction proceeds, the product will be a radical, containing a -HC$^{\bullet}$- group. In considering the bonding in such molecules, SMFA cannot, a priori, know whether a carbon with unsatisfied valence is a radical or an anion (or even a cation in some circumstances). Hence, the user has to inform SMFA (in the input) that this carbon has a charge of 0. Hence, in the input submenu, under

5) Specify charges for metals & other atoms (optional)

the user must specify that atom number 14 has charge 0.

Under the constraint that the H..H distance is 1.644 Å, it is unlikely that optimization will lead to breaking the C-H bond, and the formation of a carbon radical. Hence, it is probably not necessary to include this charge specification here. However, during the following scan, where the H..H distance is reduced, this charge specification will be essential.

The following files have been included in the SMFA/doc/testcases/scan/reactants/ directory: OUT_SMFA, optout, CONVERGEDCOORDS, and the initial geometry file H+decane_react.xyz.

Reaction path scan

Beginning with the converged geometry, H+decane_react.xyz, a scan of the minimum energy as a function of the H...H distance was carried out using the NWChem program at HF/cc-pVDZ with Level = 3, and d_{tol} = 1.1.

The following files are included in SMFA/doc/testcases/scan/reaction_scan:

H+decane_react_opt.xyz, OUT_SMFA, SCANcon1, SCANcon2, SCANcon3, SCANcon4, SCANcon5, SCANcon6, SCANcon7, SCANuncon8, SCANcon9, SCANcon10, and optout.1,...., optout.10.

The OUT_SMFA file shows that the scan includes 10 points, on a path which is defined by the atom-atom distance between atoms 15 and 33 having an initial value of 1.664 Å, reducing by 0.1 Å at each step. The optimized geometry at each step is contained in the file labeled SCANcon1, 2 etc. Note that although 100 steps were allowed for each optimization, the geometry did not reach convergence for step 8 on the path. At the tail of OUT_SMFA, you can see that the energy profile on the path reached a maximum at step 7. Hence, we can presume that the geometry in SCANcon7 is in the vicinity of the transition state (saddle point) for the hydrogen abstraction reaction (at HF/cc-pVDZ).

7.6 TS search

Starting from the SCANcon7 geometry above, optimization to a saddle point, lead (in 32 steps) instead to a geometry which was well towards the reactants, H + decane.

Starting from the SCANuncon8 geometry above, optimization to a saddle point, lead instead to a geometry well towards the products, H_2 + a decane radical.

Hence, the scan process was carried out again with the SCANcon7 geometry as a starting point, but only reducing the H..H distance by 0.025 at each step. This located a better estimate of the maximum energy on the path, at a geometry labeled step4.xyz (where the H..H distance is 0.969) in subdirectory SMFA/doc/testcases/TS/.

With step4.xyz as the input coordinate file, optimization to a TS converged. This geometry, CONVERGEDCOORDS, the OUT_SMFA and optout files are also contained in SMFA/doc/testcases/TS/.

Both the scan and TS optimization calculations were carried out using GAUSSIAN at HF/cc-pVDZ with Level = 3, and d_{tol} = 1.1.

7.7 Isodesmic reactions

Two molecules, denoted MOGQOO and TAXYIA in the Cambridge Structure Database, are isomers. The Cartesian coordinates for these are in the SMFA/doc/testcases directory in subdirectories MOGQOO (the file is MOGQOO.xyz) and TAXYIA (the file is TAXYIA.xyz).

The procedure of Section 5.2 in the User's Manual was followed, with Level = 1, using the Cartesian coordinate file MOGQOO.xyz. This gives an isodesmic reaction involving MOGQOO. The output files created by this utility are also contained in the MOGQOO subdirectory. This includes the OUT_SMFA file.

The same procedure was carried out using TAXYIA.xyz, to give an isodesmic reaction involving this molecule. The output files created by this utility are also contained in the TAXYIA subdirectory. This includes the OUT_SMFA file.

7.8 Combining isodesmic (etc) reactions

The utility of Section 5.3 (Combining isodesmic, homodesmotic etc reactions) was then carried out using MOGQOO.xyz for the first file name requested, and TAXYIA.xyz for the second file name. The output files from this procedure are contained in the subdirectory MOGQOO_minus_TAXYIA.

7.9 Electrostatic potential on the solvent accessible surface.

A small protein (PDB ID 5tlr), of 557 atoms, has been chosen to illustrate this utility. Twenty structures for this protein were reported ["Spider peptide toxin HwTx-IV engineered to bind to lipid membranes has an increased inhibitory potency at human voltage-gated sodium channel hNaV1.7", Agwa, A.J., Lawrence, N., Deplazes, E., Cheneval, O., Chen, R.M., Craik, D.J., Schroeder, C.I., Henriques, S.T., (2017) Biochim. Biophys. Acta 1859: 835-844], and we have chosen the first of these, denoted 5tlr_1.xyz in SMFA/doc/testcases/solventsurface/. Following the instructions in Section 5.4, we have carried out a HF/6-31G(d,p) energy calculation for this structure with Level = 1 and d_{tol} = 0. Once the energy calculation was completed, we continued to follow the instructions in Section 5.4, and obtained the files OUT_SMFA, SOLACCSURFACE and ESP.cube, which are also contained in SMFA/doc/testcases/solventsurface/.

7.10 Dipole Polarizability

In Section 7.3, the structure of c-cyclotridecene was optimized at MP2/cc-pVDZ. This structure, in the file optMP2L3dtol1.1c-cyclotridecene.xyz, has been included in the directory SMFA/doc/testcases/polarizability/.

Using GAUSSIAN with the keyword POLAR, the energy of this structure was evaluated at HF/cc-pVDZ with Level = 3 and d_{tol} = 1.1. Utility 5.6 was then used to evaluate the dipole polarizability and the resultant OUT_SMFA file was included in SMFA/doc/testcases/polarizability/ as OUT_SMFA_GAU_HF.

Using QChem, with the $rem instruction MOPROP 2, the energy of this structure was also evaluated at HF/cc-pVDZ with Level = 3 and dtol = 1.1. Utility 5.6 was then used to evaluate the dipole polarizability and the resultant OUT_SMFA file was included in SMFA/doc/   Finally, both the GAUSSIAN and QChem calculations were repeated for MP2/cc-pVDZ, and the corresponding files, OUT_SMFA_GAU_MP2 and OUT_SMFA_QCH_MP2 are included in testcases/polarizability/.

7.11 Dipole Hyperpolarizability

We again use the structure of c-cyclotridecene which was optimized at MP2/cc-pVDZ in Section 7.3 (file name optMP2L3dtol1.1c-cyclotridecene.xyz).

For GAUSSIAN, a frequency calculation was carried out using HF/cc-pVDZ with Level = 3 and d_{tol} = 1.1. Using option

```
6) Dipole Hyperpolarizability
```

in the Utilities submenu, results in the hyperpolrizability tensor appended to the OUT_SMFA file. This file is labeled OUT_SMFA_GAU_HF in the SMFA/doc/testcases/hyperpolarizability/ directory.

7.12 Internal coordinates

We again use the structure of c-cyclotridecene which was optimized at MP2/cc-pVDZ in Section 7.3 (file name optMP2L3dtol1.1c-cyclotridecene.xyz), as an example. Using (for example) MacMolPlt, the structure was represented graphically as
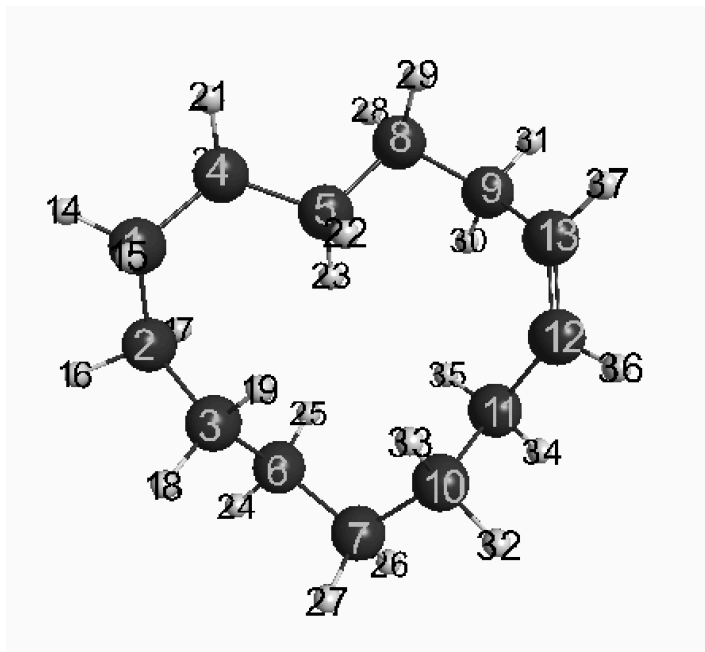

Then, with optMP2L3dtol1.1c-cyclotridecene.xyz chosen as the coordinate file, choosing

7) Internal Coordinates in the Utilities submenu, SMFA responds with

To find a bond length, enter the 2 atom numbers
To find a bond angle, enter the 3 atom numbers
To find a dihedral angle, enter the 4 atom numbers
To exit, hit RETURN

In order to measure the structure around the double bond, we might enter the following atom numbers and receive the responses shown

```
    12   13
            12          13      1.354582
    36   12   13
            36          12         13    117.178690
    36   12   13   37
      36          12          13          37    -0.829765
```

For example, we could look at a distance across the ring:

```
    23   35
      23          35      2.294300
```

All distances are in Å and angles are in degrees.

80

7.13 Adding H atoms

As an example, a molecule was downloaded from the Protein Database [PBS ID 2mvh] [23]. The first of 20 structures was used in this example, and is contained in the file 2mvh_1_original.xyz in the directory SMFA/doc/testcases/addHatoms.

This structure was input into SMFA with some simple additional inputs (value of Level and ab initio method, etc) to allow SMFA to examine the input. The resultant OUT_SMFA file is also contained in SMFA/doc/testcases/addHatoms. The OUT_SMFA file lists charged groups that it finds in the structure. For example, the first charged group in the list looks like
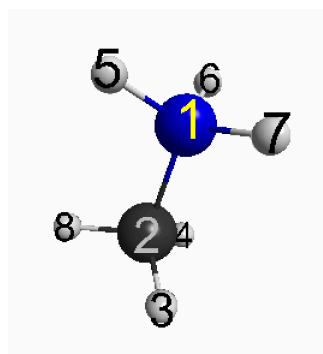
which is clearly a protonated amine group. However, the second last charged group looks like
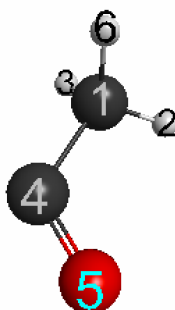
in which a carbon atom is clearly missing a bond, for whatever reason. We can add a H atom to restore full valence to this carbon atom as follows. First we load the coordinate file, 2mvh_1_original.xyz, into a graphics program (here MacMolPlt), so we can see the structure. From OUT_SMFA, we see that the carbon atom numbered 1 above is atom 436 in the original structure. So, secondly, we expand our view of the structure to locate and see the atoms near number 436. It might look something like
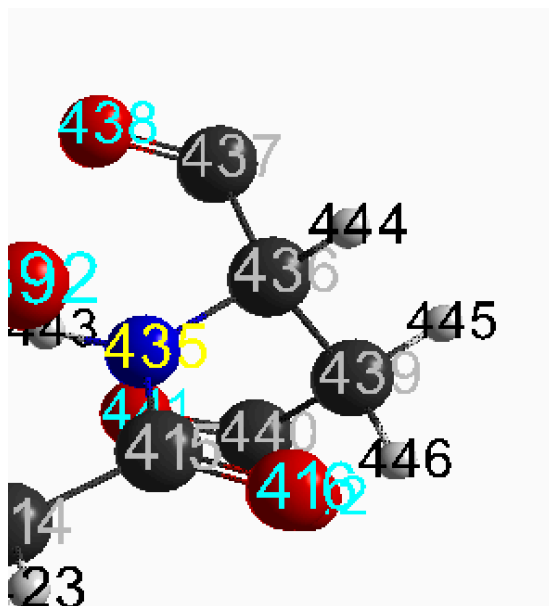
Now, we can choose

8) Add H atoms

in the Utilities submenu, and add a hydrogen atom to atom 437. SMFA responds

To get coordinates for a H atom to replace one that is missing,
you will need to enter the atom numbers for three atoms in the structure:
The H atom will be bonded by 1 Angstrom to the second atom;
the first atom is taken to lie on an imaginary +x axis from the second atom;
the third atom is taken to lie in an imaginary xy plane,
in the +y direction from the second atom.
You will then enter two spherical polar angles, theta and psi
(your best guess), and the program will output the coordinates,
which you can paste into a modified coordinate file.

Enter the 3 atom numbers, or enter RETURN to finish

We could enter (note that 437 is the second atom number)

436 437 438

SMFA responds

Enter the two angles in degrees. We entered (an easy guess for this sp2 carbon)

90.0 -120.0

and SMFA responded

H   -26.566643   -2.608096   -18.328955

We add these atomic coordinates to the original structure to get a new coordinate file, labeled

2mvh_1_corrected.xyz (also in SMFA/doc/testcases/addHatoms). NOTE: We must change the number of atoms at the top of this file from 446 to 447. The structure in this region now looks like this
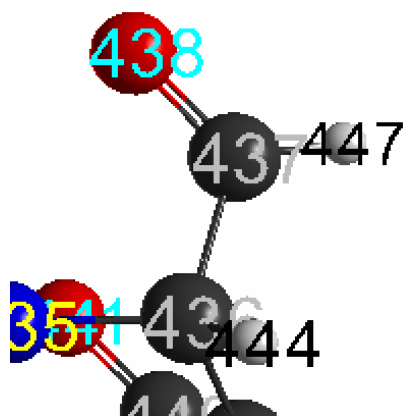
where the hydrogen atom number 447 has been located in a plausible position.

7.14 Input checks

5kbo

This is a small protein of 2991 atoms, downloaded from the RSC Protein DataBase, with ID 5kbo. The downloaded file, 5kbo.cif, contained 20 structures, and in this example, we use the first of these, denoted 5kbo_1.xyz (in SMFA/doc/testcases/inputchecks/5kbo. We completed the input and ran main menu choice

2) SMFA examines the input and prints comments or queries to OUT_SMFA

If you exit SMFA and look at the OUT_SMFA file, you will see two points of interest.

(i) SMFA identified two H--S bonds that are abnormally short, and supplies the atom numbers for these atoms. If you use a molecular graphics program to look at the structure in these two locations, you will see that the S--H bonds are short, but probably not unphysically short, so no action is necessary.

(ii) SMFA lists many charged groups which it has identified (61 in fact). If you look at the first group in a graphics program, you will see that a H atom has been omitted from the stucture at a nitrogen atom (a --NH group is not plausible under the normal conditions wherein this structure was obtained experimentally). Hence, to get physically sensible calculations with this structure, you will need to add a hydrogen at this nitrogen, using Utility 5.8.

5tp6

This is a small protein of 2769 atoms, was downloaded from the RSC Protein DataBase, with ID 5tp6. The downloaded file, 5tp6.cif, contained 20 structures, and in this example, we use the first of these, denoted 5tp6_1.xyz (in SMFA/doc/testcases/inputchecks/5tp6. We completed the input and ran main menu choice

2) SMFA examines the input and prints comments or queries to OUT_SMFA

If you exit SMFA and look at the OUT_SMFA file (also in SMFA/doc/testcases/inputchecks/5tp6), you will see that despite having a net charge of -13, there are no warnings of unusual bonding, and all the (many) charged groups are sensible. The structure is physically reasonable.

5ce4

This is a small protein of 2690 atoms, was downloaded from the RSC Protein DataBase, with ID 5ce4. The downloaded file, 5ce4.cif, was converted to "xyz" format using the obabel program, in file SMFA/doc/testcases/inputcheck/5ce4/. We completed the input and ran main menu choice

2) SMFA examines the input and prints comments or queries to OUT_SMFA

If you look at the resultant output in OUT_SMFA (contained in SMFA/doc/testcases/inputcheck/5ce4/), you will see that SMFA has reported a number of anomalies in the bonding, in which possibly unphysical atom-atom distances have been found. If you look at the structure in a graphics program, near the atom numbers indicated in OUTSMFA, you will see several instances in which a carbon atom and an oxygen atom (inter alia) have been nearly superimposed in this structure. Clearly the structure is unphysical (due to reasons unknown) and should not be used for electronic structure calculations.

8. Citation

If SMFA is used in publications, you should cite

The SMFA program for quantum chemistry calculations on large molecules,
R. Kobayashi, M. A. Addicoat, R. Amos and M. A. Collins

Appendices

Appendix A. Installing the code

A README file is contained in the SMFAPAC/ directory with instructions for installing the code. For later reference, the contents of this README file are reproduced here.

```
Installation Instructions for SMFA on Linux systems.


(1) The SMFAPAC/bin directory must be on the user's path, which can be
achieved by adding the following to the appropriate rc file:

For ~/.cshrc:

set path = ( $path /path/to/SMFAPAC/bin)

For ~/.bashrc:

export PATH=$PATH:/path/to/SMFAPAC/bin

Ensure /path/to is replaced with the actual path to the directory.

(2) The following two compilers and one library must be on the user's path:
      intel-fc
      intel-mkl
      intel-cc

(3) The user must ensure that the Cmake program is available.

(4) Perl must be installed.

(5) The user has to provide access to at least one of
GAMESS(US),
GAUSSIAN,
NWChem, or
QChem,

and also to DALTON.

      The SMFA_Users_Guide.pdf (Section 2.7) in SMFAPAC/doc contains
information about how SMFA needs to be told (in Main Menu Item 4) how to
"load" each (at least one) of these quantum chemistry packages.

(6) One of the optional utility programs in SMFA requires the openbabel
program (see Section 5.2 in SMFA_Users_Guide.pdf), so you will need to
install openbabel if you want to use this feature. You can install openbabel
at any time (the build below does not require it).

(7) Once the above preliminaries are complete, the user simply executes the
following commands

cd SMFAPAC
mkdir build
cd build
cmake ../
```

```
make
make install
```

Comment
      The build is out-of-source, which means that all the object files are created in subdirectories of the build directory. You can 'make clean' or simply delete the entire build directory if you want to start from scratch again. When you 'make install' all the compiled binaries are moved to the exe directory in the SMFAPAC directory.

(8) If you have not already done so, copy the SMFA_Users_Guide.pdf in SMFAPAC/doc to somewhere you can read pdf files. The User's Guide contains detailed instructions on how to use SMFA.

Appendix B. The parameter d_{tol}

The energy of a molecule is given by

$$E = \sum_{n=1}^{N_{frag}} c_n E(F_n) \qquad \text{(A)}$$

$$+ \sum_i \sum_j c_{ij} \left[E\left(F_i^{(1)} F_j^{(1)}\right) - E\left(F_i^{(1)}\right) - E\left(F_j^{(1)}\right) \right] \qquad \text{(B)} \qquad \text{(B.1)}$$

$$+ \mathrm{E}_{pert}$$

The first term (A) is the dominant contribution, arising from the energy of the fragments evaluated for a given value of the parameter Level. For a given value of Level, every functional group G_i in the molecule will be in at least one fragment with any group G_k that is separated from G_i by no more than Level bonds. As these "bonded" interactions between groups contribute strongly to the energy of the molecule, term (A) is the largest component of the rhs of (B.1). However, there may be groups that are relatively close in space to some group G_i even though they may be well separated from G_i in terms of bonded connections. These "nonbonded" interactions can be significant in energy on the scale that is important for chemistry, and must be accounted for. Term (B) and the third term E_{pert} account for these effects. These bonded interactions are treated as follows:

The molecule is fragmented at Level = 1. The fragments are pairs of adjacent groups and single groups. Each of these Level = 1 fragments is then interacted with every other fragment in the Level = 1 set of fragments, except that interactions between groups that have already been accounted for in term (A) are discarded. Most of these interactions between Level = 1 fragments are very small because they are well separated in space. However, we define d:

$$d = \frac{\text{the minimum atom-atom distance}}{\text{the sum of the Van der Waals radii for the two atoms}} \qquad \text{(B.2)}$$

for each pair of Level = 1 fragments. Then, if

$$d \leq d_{tol} \qquad \text{(B.3)}$$

the interaction between these fragments is evaluated as in term (B). The coefficients c_{ij} arise from the coefficients of each fragment in the Level = 1 fragmentation. Conversely, if $d > d_{tol}$, then the interaction between these fragments is evaluated (from first principles) using perturbation theory. The sum of all such interactions is given by E_{pert}. These perturbations include electrostatic interactions, dispersion and induction. The details are contained in Ref. 4.

The derivatives, with respect to the atomic coordinates, of both terms A and B are obtained from the electronic structure calculations. The corresponding derivatives of E_{pert} are only available approximately. See Ref. 9 for details. Not surprisingly then, tests [9] show that the energy derivatives are

more reliably accurate for larger values of d_{tol}. Nonetheless, values of d_{tol} as small as 1.1 return quite accurate derivatives.

There are a number of practical reasons for using a modest value of d_{tol} (say 1.1 - 1.5). First, the number of ab initio calculations in term B increases rapidly with increasing d_{tol}. Secondly, term B will be subject to "basis set superposition" error (BSSE) if small basis sets are used. (Term A does not appear to be subject to significant BSSE, due probably to the larger sizes of the fragments and the overlaps between fragments.) E_{pert} is not subject to BSSE. Unless high accuracy is required for gradients and hessians, $d_{tol} \approx 1.1$ - 1.5 is best. In fact, $d_{tol} = 1.1$ is normally recommended. Finally, as the value of Level increases, the number of non-bonded interactions for which (B.3) holds declines, and the total non-bonded contribution to the energy declines in magnitude.

There is one exception to this general recommendation. When Level = 2, each group only has "non-bonded" interactions with groups that are separated by three bonds (Level = 2 fragments only contain next-nearest neighbour interactions between groups). The capping hydrogens that are used to terminate the Level = 1 fragments are relatively close to capping hydrogens on a group that is only three bonds distant. Term B then contains spurious interactions between capping hydrogens that are often less than 2.5 Å apart. These spurious interactions can be avoided by setting $d_{tol} = 0$, so that term B is avoided, and we rely only on E_{pert}. Term A alone for Level = 2 rarely provides an adequate approximation to the energy, except for relatively straight chain molecules. However, it is useful to evaluate the energy with Level = 2 as part of the attempt to see convergence with increasing values of Level. For Level = 2, one might as well use $d_{tol} = 0$.

Appendix C. Structure optimization methods

(i) Optimisation algorithm

The geometry optimisation (minimum energy) and transition state search are carried out by the program ROGEROPT within SMFA. The geometry optimisation is carried out in Cartesian coordinates.

The displacement at each step in the energy minimisation is evaluated using the "rational function optimisation (RFO)" approach [Ajit Banerjee, Noah Adams, Jack Simonsand Ron Shepard, J. Phys. Chem. 1985,89, 52-57; Adam B. Birkholz1 , H. Bernhard Schlegel, Theor Chem Acc (2016) 135:84]

The hessian is evaluated ab initio only at the initial geometry. This hessian is evaluated at the Hartree Fock level, using the basis set requested at input. At each subsequent geometry, an approximate "updated" hessian is calculated using the "flowchart" method described in Section 2.2 of A. B. Birkholz and H. B. Schlegel, Theor Chem Acc (2016) **135**, 84. A projection operator is used to remove translational and rotational components from the updated hessian.

The energy gradient at each step of the optimization is evaluated using the level of theory and basis set which were requested at input.

The eigenvalues of the "augmented Hamiltonian [see Eq.(7) of A. B. Birkholz and H. B. Schlegel, Theor Chem Acc (2016) **135**, 84] are shifted up by 0.05 plus the lowest eigenvalue.

The calculated displacement at each step is moderated by a "trust radius". If the norm of the calculated displacement exceeds the trust radius, the displacement is scaled to have a norm equal to the trust radius. The trust radius is initialized with a value of 0.3 Bohr. At each subsequent step, we compare the expected change in energy from the last step with the actual change. If the next calculated displacement exceeds the trust radius and the expected change in energy from the last step is within 25% of the actual change, then the trust radius is increased by 50% before adjusting the norm of the displacement. Alternatively, if the next calculated displacement exceeds the trust radius and the expected change in energy from the last step is NOT within 25% of the actual change, then the trust radius is reduced by a factor of 2 before adjusting the norm of the displacement.

Convergence: At each step in the optimisation, SMFA measures the average magnitude of the elements of the gradient vector (Hartrees/Bohr), the magnitude of the largest component of that vector, the average magnitude of the elements of the next displacement vector (Bohr) and magnitude of the largest component of that vector. Each of these quantities is taken to satisfy a convergence criteria when less than 0.00015, 0.0004, 0.0012, and 0.0018, respectively. When any three of these criteria are met, the geometry is taken to be converged. These convergence criteria were defined by reference to

standard quantum chemistry packages.

  Constraints: Bond lengths, bond angles, and dihedral angles can be "constrained" during an optimisation. These are not strict constraints. SMFA includes a penalty function, P, which is added to the total electronic energy. It is the sum of these two components which is minimised. The penalty function is given by P:

$$P = \frac{1}{2} \sum_i^{bonds} p_r \left[r_i - r_i^0 \right]^2 + \frac{1}{2} \sum_j^{angles} p_\theta \left[\cos(\theta_j) - \cos(\theta_j^0) \right]^2$$

$$+ \frac{1}{2} \sum_k^{dihedrals} p_\varphi \left[\cos(\varphi_k) - \cos(\varphi_k^0) \right]^2 + p_\varphi \left[\sin(\varphi_k) - \sin(\varphi_k^0) \right]^2$$

where we have set $p_r = p_\theta = 1$ Hartree, $p_\varphi = 0.5$ Hartree. r_i^0, θ_j^0 and φ_k^0 denote the "preferred" values of the bond lengths, angles and dihedral angles, respectively. The dihedral angles are defined as follows.
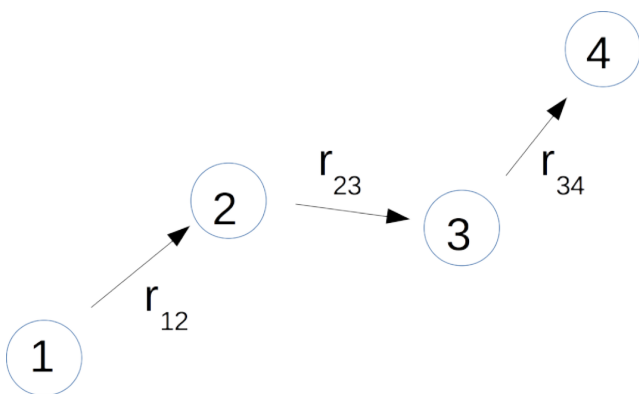
Using the definitions of atom numbers 1, 2, 3, and 4, and the vectors $\mathbf{r}_{12}$, $\mathbf{r}_{23}$, and $\mathbf{r}_{34}$, we define the vectors

$\mathbf{A} = \mathbf{r}_{34} \times \mathbf{r}_{23}$,
$\mathbf{B} = \mathbf{r}_{23} \times \mathbf{r}_{12}$,
$\mathbf{C} = \mathbf{r}_{23}$.

Then

$$\cos(\varphi) = \frac{\mathbf{A} \bullet \mathbf{B}}{|\mathbf{A}||\mathbf{B}|},$$

$$\sin(\varphi) = \frac{\mathbf{C} \bullet (\mathbf{A} \times \mathbf{B})}{|\mathbf{A}||\mathbf{B}||\mathbf{C}|}$$

φ is defined as $\arccos[\cos(\varphi)]$ if $\sin(\varphi) > 0$, and $- \arccos[\cos(\varphi)]$ if $\sin(\varphi) < 0$.

(ii) Scan algorithm

The scan program simply executes the optimisation algorithm for a sequence of constraints. The optimised geometry at one step along the sequence of constraints is the initial configuration at the next step.

(iii) Finding a saddle point (TS)

First order saddle points are found using the algorithm of J. M. Bofill, Journal of Computational Chemistry, **Vol.** 15, No. 1, 1-11 (1994), denoted B1994. The hessian is evaluated at the initial geometry (using Hartree Fock), and updated at each subsequent step using Eq. (10) of B1994. At each step, the molecular geometry is updated using the algorithm denoted by processes (a) to (k) in B1994 [see text below Eq. (21) in B1994].

Appendix D. Implementation of property and property gradient methods

Eqns (5.5.2) and (5.5.3) are reproduced here:

$$
\begin{aligned}
P &= \sum_{n=1}^{N_{frag}} c_n P(F_n) \qquad\qquad\qquad\qquad\text{(A)} \\
&+ \sum_i \sum_j c_{ij} \left[P\left(F_i^{(1)} F_j^{(1)}\right) - P\left(F_i^{(1)}\right) - P\left(F_j^{(1)}\right) \right] \qquad\text{(B)}
\end{aligned}
\qquad\text{(D.1)}
$$

$$
\begin{aligned}
\frac{\partial P}{\partial x_\alpha} &= \sum_{n=1}^{N_{frag}} c_n \frac{\partial P(F_n)}{\partial x_\alpha} \qquad\qquad\qquad\qquad\qquad\text{(A)} \\
&+ \sum_i \sum_j c_{ij} \left[\frac{\partial P\left(F_i^{(1)} F_j^{(1)}\right)}{\partial x_\alpha} - \frac{\partial P\left(F_i^{(1)}\right)}{\partial x_\alpha} - \frac{\partial P\left(F_j^{(1)}\right)}{\partial x_\alpha} P\left(F_j^{(1)}\right) \right] \qquad\text{(B)}
\end{aligned}
\qquad\text{(D.2)}
$$

where the property P (which may be simply the energy) is calculated by the ab initio program for each fragment.

It may be that the ab initio package outputs the value of P and/or the derivatives, $\left\{ \dfrac{\partial P}{\partial x_\alpha} \right\}$, in some file other than the usual "log" file; for example in a checkpoint file. If so, you must include instructions in the

ab initio input for this file to be created. If necessary, you must arrange for this file to be readable. In the instructions below, we will assume that the output is in the "log" file.

The quantum chemistry output for the fragment F_n (see Eq. D.1) is in file FRAGn.log. The coefficient c_n is the nth integer in file signs.out.

The quantum chemistry output for the fragment $F_i^{(1)}$ (see Eq. D.1) is in file nb.i.0.log.

The quantum chemistry output for the fragment $F_i^{(1)} F_j^{(1)}$ (see Eq. D.1) is in file ab.i.j.log.

The coefficient c_{ij} (see Eq. D.1) is in ab.i.j.log immediately following the characters "Isg_coeff=".

The format in which P or $\left\{ \dfrac{\partial P}{\partial x_\alpha} \right\}$ is written in these log files varies from one quantum chemistry package to the next, and unfortunately, even varies within one quantum chemistry package depending on what quantum chemistry method was used. So, be warned, finding the data you want may depend on whether it was calculated at HF or MP2 or CCSD or some other method. Authors of such programs have apparently not bothered with a consistent output style.

Perl scripts (very amateurish ones) to extract data for P or $\left\{ \dfrac{\partial P}{\partial x_\alpha} \right\}$ can be found in the files

SMFAPAC/bin/ SMFA_gam.pl, SMFA_gau.pl, SMFA_nwc.pl and SMFA_qch.pl for the four quantum chemistry programs supported by SMFA. For example, within SMFA_gau.pl the subroutine extract_gau extracts quantities like the fragment atomic coordinates, the energy, the forces and hessians from the FRAGn.log, ab.i.j.log and nb.i.0.log files, using other subroutines like getcoords_gau. You will need to write scripts analogous to extract_gau and getcoords_gau for the property of interest.

In the case of simple properties, P, you need only evaluate the product c_n P(FRAGn.log) and add up the results to get term A in (D.1), and evaluate the product c_{ij} [P(ab.i.j.log) - P(nb.i.0.log) - P(nb.j.0.log)] and add up the results to get term B in (D.1).

Derivative properties

Evaluating Eq. (D.2) requires you to know how the coordinates in the various fragments correspond to the coordinates in the whole molecule. For example, the first two atoms in FRAG1.log might be atoms 235 and 57 in the whole molecule. Moreover, the fragments will contain "capping" hydrogen atoms which are not present in the whole molecule, and the values of $\dfrac{\partial P}{\partial x_\alpha}$ associated with these caps must be assigned to atoms in the whole molecule. The positions of the caps are determined by the positions of the atoms, n1 and n2, in the bond that was broken to form the fragment:

$$x_\alpha(cap) = (1-f) * x_\alpha(n1) + f * x_\alpha(n2) \hspace{2cm} (D.3)$$

where the factor f was determined by SMFA from the covalent radii of the atoms [Elemental_Radii.xlxs from https://ccdc.cam.ac.uk].

Term A in (D.2)

For quantities like $\left\{ \dfrac{\partial P(\text{FRAGn.log})}{\partial x_\alpha} \right\}$, all the additional information you will need to evaluate term A in

(D.2) is contained in the file frags.out. This file contains the number of FRAGn.log files and then the number of atoms from the whole molecule that are in each fragment, call it nat0(n). The identity of these atoms is then listed for each fragment; showing, for example, that atoms 1, 2, 3... in fragment k are atoms 23, 164, 15 etc in the original molecule. frags.out then lists the data you need for Eq. (D.3). For example, after the lists of atoms in each fragment, frags.out contains a section that looks like this:

```
The capping atoms are attached to
        1          21        13   0.669117648347851
        1          18        16   0.669117648347851
        1          18        19   0.669117645279965
        2           5         7   0.669117645279965
        2           5         6   0.669117645279965
        3           7         8   0.669117645279965
     .........
     .........
```

The first number in each line refers to the fragment number, FRAG1, FRAG2 etc. In this example there are 3 caps for fragment 1, 2 caps for fragment 2, etc. The second and third numbers in each line refer to the atoms in the original molecule and the last number is the factor f. The hydrogen caps are always the last atoms in the fragment. Lets take fragment 2 as an example. Say fragment 2 has nat0(2) "real atoms", and in the example above, it has two caps. Then, the gradient of the property for the nat0(2) + 1 atom in the second fragment must be assigned as follows (as you might write the code)

$$\frac{\partial P}{\partial x_\alpha(5)} = \frac{\partial P}{\partial x_\alpha(5)} + sign(2) * (1 - 0.669117..) * \frac{\partial P(\text{FRAG2.log})}{\partial x_\alpha(nat0(2) + 1)}$$

$$\frac{\partial P}{\partial x_\alpha(7)} = \frac{\partial P}{\partial x_\alpha(7)} + sign(2) * 0.669117... * \frac{\partial P(\text{FRAG2.log})}{\partial x_\alpha(nat0(2) + 1)} \hspace{1cm} (D.4)$$

where sign(2) refers to the sign for the second fragment (see signs.out).

You can now evaluate term A in Eq (D.2).

An example of this whole process can be found in SMFA_gau.pl where subroutine extract_gau gets the fragment coordinates, energy, forces and hessians from the FRAGn.log files and concatenates the data to a file call `FragDerivatives`. The fortran program SMFA/src/`Derivs_smfa` then uses frags.out

and signs.out to allocate the energy gradients and hessians from the fragments to the atoms in the whole molecule.

Term B in (D.2)

Again, the coefficient c_{ij} (see Eq. D.1) is in ab.i.j.log immediately following the characters "Isg_coeff=". The information required to allocate derivative data from ab.i.j.log and nb.i.0.log is located in the file OUT_Lev1_ATOMALLOCATION. The top portion of this file might look like this

```
        The number of final L1 frags from L1L1_mac.f
              133
        For each of              133   fragments:the number of atoms
        followed by the allocation of each atom
                   1           13
         1      230     1.000000
         1      243     1.000000
         1      244     1.000000
         1      231     1.000000
         1      242     1.000000
         1      210     1.000000
         1      211     1.000000
         1      228     1.000000
         1      229     1.000000
         1      235     1.000000
         2      210     0.330882       209     0.669118
         2      228     0.330882       234     0.669118
         2      229     0.330882       232     0.669118
                   2            6
         1        8     1.000000
         1       15     1.000000
         1       16     1.000000
         1       17     1.000000
         1        7     1.000000
         2        7     0.264706         6     0.735294
                   3            6
         1      198     1.000000
         1      205     1.000000
        ......
        ......
```

This tells us that 133 files of the type nb.i.0.log are relevant to Eq.(D.2). Line 5

```
              1           13
```

tells us that the Level = 1 fragment number 1, that gives rise to nb.1.0.log, has 13 atoms. The first ten lines below that (beginning with 1) tell us the atom numbers in the molecule corresponding to the first ten atoms associated with nb.i.0.log. Hence there are contributions to the property derivatives from nb.1.0.log such as that from the first atom:

$$\frac{\partial P}{\partial x_\alpha(230)} = \frac{\partial P}{\partial x_\alpha(230)} - c_{1j} * \frac{\partial P(\text{nb.1.0.log})}{\partial x_\alpha(1)}, \text{ for all values of } j,$$

and so on. The next three lines above (beginning with 2) identify hydrogen caps in the fragment and the atoms in the molecule that determine their positions. Hence, the 11th atom associated with nb.i.0.log. contributes to the derivative of the property with respect to two atoms in the original molecule:

$$\frac{\partial P}{\partial x_\alpha(210)} = \frac{\partial P}{\partial x_\alpha(210)} - c_{1j} * 0.330882 * \frac{\partial P(\text{nb.1.0.log})}{\partial x_\alpha(11)}$$

$$\frac{\partial P}{\partial x_\alpha(209)} = \frac{\partial P}{\partial x_\alpha(209)} - c_{1j} * 0.669118 * \frac{\partial P(\text{nb.1.0.log})}{\partial x_\alpha(11)}$$

, for all values of j. So it continues for the other two caps.

Then the next line tells us that the second Level = 1 fragment, that gives rise to nb.2.0.log, has 6 atoms, including one cap. And so it goes on.

This data describes the atom allocations for all relevant nb.i.0.log and nb.j.0.log files and so also describes the atom allocations for all ab.i.j.log files, as the order of the atoms in ab.i.j.log is that of nb.i.0.log followed by nb.j.0.log.

As an example, the subroutine extract_gau in SMFA_gau.pl creates files called abforces and abhessians which contain the energy gradients and hessians from all the ab.i.j.log, nb.i.0.log and nb.j.0.log files. The fortran program `ABNBderivatives.f` (with the executable stored as `ABNBderivatives`) reads this data and allocates the derivative data as indicated above.

  This completes the description of terms A and B in both (D.1) and (D.2). Higher order derivatives of properties (eg the hessian of the energy) can be constructed by considering the higher derivatives of D.2, and using the data in frags.out, signs.out and `OUT_Lev1_ATOMALLOCATION`.

Appendix E. Use of perturbation theory

The energy, E_{pert}, first referred to in Eq.(5.5.1) has been described in great detail in the papers listed in the References. It is composed of electrostatic, induction and dispersion interactions that are not accounted for in the first two terms in Eq. (5.5.1). Based on the value of the parameter d_{tol}, these are interactions between parts of the molecule that are well separated from each other in terms of both bonded connectivity and distance.

The approach to approximating these interactions has evolved over time, and the method programmed in SMFA is that described in the paper noted for citation. Two different approaches are adopted by SMFA:

(i) If the molecule has formally charged groups, or if the user adopts the use of embedded charges to describe polar solvent molecules, then we use Method1;

(ii) in the absence of such charges, we use Method2.

The interaction of parts of the molecule at long distances is comprised of electrostatic interactions, the associated induction effect, and dispersion.

Method1

If there are formal charges in the molecule, or if polar solvent molecules are present, then induction makes a significant contribution to the molecular energy, even when interactions take place over long distances. In this case, the charge distribution of all formally charged groups and all solvent molecules are evaluated. For GAUSSIAN and QChem this is achieved using the natural population analysis (NPA) approach [A. E. Reed and F. Weinhold, J. Chem. Phys. **78**, 4066–4073 (1983); A. E. Reed, R. B. Weinstock, and F. Weinhold, J. Chem. Phys. **83**, 73 (1985)].

For NWChem, NPA charges are not available, and charges on the atoms in these groups are evaluated using NWChem's "ESP" approach. The calculation of energies for all fragments (bonded and nonbonded) are carried out in the presence of "embedded" or "background" charges on those atoms (in charged groups or solvents) which are not involved in the fragment. GAMESS does not permit the use of embedded charges, so Method 1 is not employed for GAMESS.

The use of embedded charges accounts quite well for both the electrostatic interaction of these charged groups (or solvent molecules) with the rest of the molecule, and accounts well for induction which is large in such systems.

Electrostatic interactions also occur between formally neutral groups at a distance. These interactions are evaluated using distributed multipole moments on all the atoms, as previously described [4] for GAMESS, GAUSSIAN and QChem. For NWChem, distributed charges only are employed, obtained from NWChem's ESP approach. For GAMESS, the distributed multipole moments are only available at

the SCF level of ab initio theory. Hence, even though the other terms in Eq. (5.5.1) might be evaluated using MP2 or CCSD or other correlated method, the long range electrostatics are evaluated with SCF-based charge distributions for GAMESS. Similarly, the electron density at post-SCF levels is not made available by QChem, so the long range electrostatics are evaluated with SCF-based charge distributions only for QChem. GAUSSIAN provides the correlated electron density for MP2 and coupled cluster methods as well as for SCF. Hence, the best available density is used for the long range electrostatics when GAUSSIAN is employed.

Finally, if the ab initio method employed accounts for dynamic electron correlation at a distance, then there is a dispersion interaction between such groups. This effect is normally small or relatively small molecules, but can be very significant in cases of high mass density (eg in crystals such as diamond [5]) or folded proteins. These dispersion interactions between groups at a distance are evaluated as previously described.[4] The DALTON program is necessary to provide the ab initio calculation of the imaginary frequency polarizability needed to evaluate these interactions.

Method2

If there are no formally charged groups and no use of embedded charges for solvent molecules, or if GAMESS is used, then Method2 is adopted.

The long range electrostatic interactions for *all* well-separated groups are evaluated using the method outlined in Method1.

Induction (a small effect if there are no significant charge distributions) is accounted for perturbatively using the static polarizability of each group (evaluated ab initio), as previously described.[4]. DALTON is also used to evaluate these polarizabilities when either GAMESS or NWChem are otherwise employed. Long range dispersion interactions are evaluated as previously described.[4]

Appendix F. Additional implementation details

Hybrid Fragmentation Scheme

     The systematic molecular fragmentation by annihilation (SMFA) algorithm has been modified to reduce the computational time. The original SMFA algorithm [M. A. Collins, Phys. Chem. Chem. Phys. **14**, 7744–7751 (2012)] is employed in conjunction with a method which was originally developed to update a fragmentation as the molecular structure changes [M. A. Collins, J. Phys. Chem. A, **120**, 9281−9291 (2016)].

     The original SMFA algorithm is slow whenever the molecular geometry involves dense connections (chemical bonds or hydrogen bonds) between functional groups. In most molecules, each functional group is connected to two other groups (as in a chain) with occasional branches, in which some groups are connected to three or four other groups. By contrast, in crystalline or liquid water, for example, almost all functional groups are connected to four other groups. In the latter case, the original SMFA algorithm is slow. Similarly, many protein structures are densely connected due to a high number of hydrogen bonds. While executing the SMFA algorithm for highly connected structures, very many large fragments persist over long cycles of the repetitive algorithm, in which groups are sequentially eliminated. This greatly slows down the reduction of the set of fragments to the final (relatively small) set.

To significantly speed up the fragmentation of a given structure, we simply "eliminate" bonds (pretend they don't exist) that make some functional groups highly connected. [Aside: If a bond exists between two functional groups that both have more than two connections, then that bond is "eliminated"]. The resultant geometry, with reduced bonding, is then fragmented (relatively rapidly) using the SMFA algorithm. The "eliminated" bonds are then re-established as follows. We use the fact that no final fragment can contain groups that are separated by more than Level bonds. Consequently, the role of a functional group in the set of fragments cannot depend on the presence or absence of a

bond that is more than Level bonded connections away from that group. To re-establish a "neglected" bond, we simply add one new fragment to the set and subtract one fragment from the set. The added fragment contains all groups that are within Level bonds from the neglected bond, with the current set of connections plus the "neglected" bond. The fragment subtracted from the set contains the same groups and connections, except for the "neglected" bond. The set of connections between groups is updated to contain the previously neglected bond. If there was more than one neglected bond, then the procedure above is simply repeated until all bonds have been "restored" to the structure. This addition and subtraction approach is described in more detail in [M. A. Collins, J. Phys. Chem. A, **120**, 9281−9291 (2016)]. The SMFA algorithm is then reapplied to the full set of fragments (the original SMFA fragments plus the additions and subtractions). Cancellations between new and old fragments leads to the same set of final fragments as would have been obtained if the original SMFA procedure had been applied without any adjustments to the bonding. However, this composite procedure is much faster because very large fragments are not retained for long periods during the SMFA procedure itself.

Appendix G. Quantum Chemistry Foibles

(a) GAUSSIAN post CCSD/QCISD electrostatic calculations

GAUSSIAN only calculates the electron density for some post-SCF methods (eg MP2, CCSD). It does not allow calculation of the post-SCF density for CCSD(T), for example, so only the SCF density is available. SMFA uses the density to evaluate the charge distribution in a molecule and hence to calculate the `"long range electrostatic component of"` the energy, as reported in the OUT_SMFA file. Therefore, if you do a CCSD(T) calculation, the OUT_SMFA file reports a CCSD(T) value for the ab initio energy under the heading

`"This energy is composed of an ab initio component of"`

but only reports a SCF value for the long range electrostatic component. The best available approximation to the correct CCSD(T) energy would be to perform both CCSD and CCSD(T) calculations, and to combined the CCSD(T) ab initio component with the values of the `"long range electrostatic component"` and `"induction energy"` (if given) from the CCSD calculation.

(b) Post Hartree Fock electrostatic calculations with QChem and GAMESS

The electron density is only available at the SCF level for QChem. SMFA uses the density to evaluate the charge distribution in a molecule and hence to calculate, via a distributed multipole analysis, the `"long range electrostatic component of"` the energy, as reported in the OUT_SMFA file. GAMESS also uses the SCF density to evaluate the distributed multipoles. Therefore, the long range electrostatic interaction energy (and the induction energy, if appropriate) is only evaluated at the Hartree Fock level for QChem and GAMESS.

(c) Electrostatic calculations with NWChem

For NWChem, as noted elsewhere in this manual, the charge distribution in a molecule is only evaluated as a sets of point charges on each atom (via a natural population analysis). Hence, the electrostatic interaction energy (and the induction energy, if appropriate) is only evaluated via an interaction of these charges.

(d) Point charges and induction

When a molecule contains formal charges, or when a very polar solvent is present, the induction energy may be significant. SMFA accounts for this induction effect for GAUSSIAN, NWChem and

QChem, by carrying out the quantum chemistry calculations of the fragments in the presence of embedded charges. [14] These charges are evaluated using the natural population analysis charges on each functional group in the molecule which are themselves evaluated in the presence of the point charges on all other groups (iterated three times).

GAMESS does not allow for embedded (partial) charges in quantum chemistry calculations, so in this case induction is always estimated using perturbation theory. [10]

(e) Dalton Polarizabilities

In some situations, the *static polarizability* of each functional group must be calculated for use in the estimation (by perturbation theory) of the long range induction and dispersion energies. For GAMESS and NWChem, these calculations are actually carried out using the DALTON program. Unfortunately, if the function group is a doublet (spin multiplicity = 2), DALTON crashes for MP2 calculations. Hence, for doublets, we have made the approximation of replacing the MP2 polarizability by the Hartree Fock polarizability (for groups with spin multiplicity = 2, only).

(f) Loading Quantum Chemistry Packages

The somewhat cumbersome instructions in Section 2.7 for "loading" or otherwise making accessible the quantum chemistry packages is due to the fact that we found GAMESS(US) and NWChem do not run correctly on our system if the DALTON module is loaded at the same time (for reasons unknown).

References

(1)   Netzloff, H. M.; Collins, M. A. Ab initio energies of non-conducting crystals by systematic fragmentation. *J. Chem. Phys.* **2007**, *127*, 134113.

(2)   Collins, M. A. Ab initio lattice dynamics of nonconducting crystals by systematic fragmentation. *Journal of Chemical Physics* **2011**, *134*, 164110.

(3)   D'Arcy, J. H.; Jordan, M. J. T.; Frankcombe, T. J.; Collins, M. A. H2 Adsorption in a Porous Crystal: Accurate First-Principles Quantum Simulation. *J. Phys. Chem. A* **2015**, *119*, 12166−12181.

(4)   Frankcombe, T. J.; Collins, M. A. Potential energy surfaces for gas-surface reactions. *Physical Chemistry Chemical Physics* **2011**, *13*, 8379.

(5)   Frankcombe, T. J.; Collins, M. A. Growing Fragmented Potentials for Gas-Surface Reactions: The Reaction between Hydrogen Atoms and Hydrogen-Terminated Silicon (111). *J. Phys. Chem. C* **2012**, *116*, 7793.

(6)   Frankcombe, T. J.; Collins, M. A.; Zhang, D. H. Modified Shepard interpolation of gas-surface potential energy surfaces with strict plane group symmetry and translational periodicity. *J. Chem Phys.* **2012**, *137*, 144701.

(7)   Collins, M. A. Can Systematic Molecular Fragmentation Be Applied to Direct Ab Initio Molecular Dynamics. *J. Phys. Chem. A* **2016**, *120*, 9281.

(8)   Deev, V.; Collins, M. A. Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

(9)   Collins, M. A.; Deev, V. A. Accuracy and efficiency of electronic energies from systematic molecular fragmentation. *J. Chem. Phys.* **2006**, *125*, 104104.

(10)   Addicoat, M. A.; Collins, M. A. Accurate treatment of non-bonded interactions within systematic molecular fragmentation. *J. Chem. Phys.* **2009**, *131*, 104103.

(11)   Collins, M. A. Systematic fragmentation of large molecules by annihilation. *Physical Chemistry Chemical Physics* **2012**, *14*, 7744.

(12)   Pruitt, S. R.; Addicoat, M. A.; Collins, M. A.; Gordon, M. S. The fragment molecular orbital and systematic molecular fragmentation methods applied to water clusters. *Physical Chemistry Chemical Physics* **2012**, *14*, 7752.

(13)   Reid, D. M.; Collins, M. A. Molecular electrostatic potentials by systematic molecular fragmentation. *J. Chem. Phys.* **2013**, *139*, 184117.

(14)   Collins, M. A. Molecular forces, geometries and frequencies by systematic molecular fragmentation including embedded charges. *J. Chem Phys.* **2014**, *141*, 094108.

(15)   Reid, D. M.; Collins, M. A. Calculating nuclear magnetic resonance shieldings using systematic molecular fragmentation by annihilation. *Phys. Chem. Chem. Phys.* **2015**, *17*, 5314.

(16)   Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2776.

(17)   Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115*, 5607.

(18)   Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon: Oxford, 1996.

(19)   Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128.

(20)   Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular orbital theory of the electronic structure of organic compounds. V. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796.

(21)   George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. *Theoret.*

Chim. Acta **1975**, *38*, 121.

(22) Japeli, B.; Pristovsek, P.; Majerle, A.; Jerala, R. Structural origin of endotoxin neutralization and antimicrobial activity of a lactoferrin-based peptide. *J.Biol.Chem.* **2005**, *280*, 16955.

(23) Gill, R. L.; Castaing, J. P.; Hsin, J.; Tan, I. S.; Wang, X.; Huang, K. C.; Tian, F.; Ramamurthi, K. S. Structural basis for the geometry-driven localization of a small protein. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E1908.