

PFIT

1 Introduction

The PFIT program reads data comprising a set of responses for some molecule, each response being the potential arising at a point P due to the polarization of the molecule by a unit point charge at some other point Q . There are typically 2000 or so points, and the response file includes the response for every pair of points. The response doesn't include the direct electrostatic potential at P due to Q .

The user specifies a polarizability model, comprising dipole and possibly higher-rank polarizabilities at each of a number of sites in the molecule, usually the nuclear positions, and the program attempts to fit the values of these polarizabilities to reproduce the response data as well as possible.

The PFIT data files generated by the `localize.py` script provided with the CAMCASP package include a polarizability model, but it can be modified by the user if necessary. They also include a set of trial polarizability values to be used as a starting set for the fit, and a set of penalty definitions which prevent the fitted values from straying too far from the starting values.

2 Running the program

```
pfit < data > output
```

3 Input commands

3.1 General

Keywords in the data file are not case-sensitive, and both keywords and numerical values are read in free format. However line-breaks are significant; each command and subcommand must appear on a separate line. Blank lines and lines beginning with “!” are ignored, as is anything in a data line that is enclosed in parentheses, provided the opening parenthesis is preceded by at least one space.

Data in other files can be read at appropriate points in the data file using the command `#include file`

The specified file is read as if it had been pasted into the data file at this point. Reading from the main data file continues when end of file is encountered in the included file. `#include` commands can be nested – that is, included files may themselves include further files.

3.2 Allocation section

```
Allocate  
:  
End
```

Allocate arrays of appropriate size for the problem. This must be the first item in the data file, if it is required. The following items may appear:

```
Batches  $n$ 
```

The response data may be provided in a number of separate batches. This specifies the maximum number of batches. Default 1.

Points n

Maximum number of response points in a batch. Default 1000.

Parameters n

Maximum number of parameters to be fitted. Default 20.

Sites n

Maximum number of polarizability sites. Default 5.

Rank n

Maximum rank of polarizability components. Default 1 (dipole).

3.3 Response data

The information on the points in each batch, and the response values for those points, may be provided in text or binary form.

Data *filename*

Read data for a batch (number of points in the batch, point positions, response values) from the specified binary file.

Lattice

Read data in text form for a batch (point names and positions, response values). The data, and the preceding ‘Lattice’ keyword, will normally be in an included data file.

Any number of ‘Data’ or ‘Lattice’ commands, up to the maximum number of batches, may appear, but binary and text files may not be mixed.

3.4 Site specification

Sites

⋮

End

Specifies the position and optional rotation of the local axes for each site that will carry polarizabilities. Each site is specified on a separate data line in the following form:

name **at** $x\ y\ z$ [*rotation*]

If the rotation specification is omitted, the local axes will be parallel to the global axes. If it is present, it takes one of the forms

Rotated $\alpha\ \beta\ \gamma$

Rotated by ψ about $n_x\ n_y\ n_z$

Here α , β and γ are Euler angles. In the alternative form, (n_x, n_y, n_z) is a vector describing the axis of rotation. It need not be a unit vector. Angles are in degrees.

Sites may also be specified in bonds:

name **BOND** $s_1\ s_2\ f$ [*rotation*]

The specifies a site in the s_1 – s_2 bond, a fraction f of the distance from s_1 to s_2 . If f is omitted, the site is placed halfway along the bond.

Axes

⋮

End

Provides an alternative way to define local site axes. The site positions must have been defined previously. This command takes one or more subcommands of the form

s₀ axis-spec axis-spec

This defines local axes for site *s₀*. Each *axis-spec* takes the form

{x|y|z} [FROM *s₁*] TO *s₂*

or

{x|y|z} [FROM *s₁*] BETWEEN *s₂* AND *s₃*

If the FROM element is omitted, *s₁* = *s₀*. In the first form, the direction specified is from site *s₁* to site *s₂*; in the second it runs from *s₁* along the bisector of the angle between the *s₁–s₂* and *s₁–s₃* directions.

The first axis given is taken in the direction specified. The second axis is derived from the second direction by orthogonalisation to the first. The third axis is defined so that the *x*, *y* and *z* axes form a right-handed set. For example, to define local axes for the atoms in water from given positions for the O and H atoms:

Axes

O z between H1 and H2 x from H1 to H2

H1 z from O to H1 x from O to H2

H2 z from O to H2 x from O to H1

End

Note that the local axis frames for the two H atoms are related by symmetry. This means that their polarizability components will be identical, with no tiresome differences of sign. For a fuller discussion of this matter, see ?.

3.5 Specification of the polarizability model

Polarizabilities

⋮

End

One line of data for each polarizability component. Each line takes the form

a b t u = [–] *p₁* [* *c₁*] [± *p₂* [* *c₂*] ...]

where *a* and *b* are site names;

t and *u* are multipole names (00, 10, 11c, etc.);

p_i is a parameter name; *c_i* is an optional coefficient.

or alternatively

a b COPY a₀ b₀

The first form means that the polarizability α_{tu}^{ab} is equal to parameter *p₁*, or minus parameter *p₁* if *p₁* is preceded by a minus sign, multiplied by the coefficient *c₁* if present. If further terms are present, each introduced by ‘+’ or ‘–’, the polarizability is a linear combination of the

parameters given. Plus and minus signs must be preceded *and followed* by white space.

The second form means that the polarizability parameters for α^{ab} are the same as for $\alpha^{a_0b_0}$. Note that the sites may need to be rotated for this to be true without changes of sign.

For example, for CO₂, we might have

Polarizabilities

```
C  C  10  10  = alpha(CC)_zz
C  C  11c 11c  = alpha(CC)_xx
C  C  11s 11s  = alpha(CC)_xx
O1 O1 10  10  = alpha(OO)_zz
O1 O1 11c 11c  = alpha(OO)_xx
O1 O1 11s 11s  = alpha(OO)_xx
O2 O2 copy O1 O1
```

End

No rotation of the oxygen sites is needed in this case, but if there were dipole-quadrupole polarizabilities on the oxygen atoms, one of them would need to be rotated, so that (say) the local z axis on each oxygen pointed away from the carbon.

3.5.1 The polarizability definition file

When `PFIT` is invoked by the `CAMCASP localize.py` script, the polarizability model definition is taken from a file `name.pdef`, where *name* is the CAMCASP job name. If this file is not provided, CAMCASP will generate one automatically. It is important to inspect this file, and ensure that it is correct, and if necessary to modify it and perhaps to simplify it. CAMCASP attempts to take account of molecular symmetry, and in particular atoms that are specified to have the same type in the molecule definition will be assigned the same polarizabilities. This will only be appropriate if the local axes are defined correctly, taking the symmetry into account. It also means that atoms of a particular element should only be specified as having the same type if they are symmetry-related or expected to have similar polarizabilities. For example, in fluorobenzene, there are four types of C atom: C1 (to which the F is attached), C2 & C6 (next to C1), C3 & C5, and C4. However it might be a reasonable approximation to group some of these together as a single type.

3.6 Suppressing optimization

```
Fix
:
End
```

Specify parameters that are not to be optimized. Each entry in this section takes the form

name $\alpha(0)$ $\alpha(1)$... $\alpha(n)$

If frequency-dependent polarizabilities are to be calculated, a parameter value must be given for each frequency; otherwise only $\alpha(0)$ need be given.

If all parameters are fixed, no optimization is carried out, but the residuals and quality-of-fit statistics are calculated.

3.7 Static or frequency-dependent polarizabilities

Static

Calculate the static polarizability parameters only. It is usually sensible to explore static polarizability models first, before running a full frequency-dependent calculation.

Frequencies [n] [ω_0]

Calculate the frequency-dependent polarizability parameters. n is the number of points to be used in the quadrature, and ω_0 is the base frequency. These parameters need to correspond to the values used in CAMCASP, if the responses were generated by CAMCASP. They normally need not (and should not) be specified, as the defaults of 10 points and $\omega_0 = 0.5$ a.u. correspond to the CAMCASP defaults.

3.8 Penalties

Sometimes it is necessary to constrain the fitting process to prevent parameter values becoming unphysical. The induction energy takes the form

$$U_{\text{ind}} = -\frac{1}{2} \sum_{abtu} V_t^a \alpha_{tu}^{ab} V_t^b,$$

where V_t^a is the potential or one of its derivatives at atom a and α_{tu}^{ab} is a polarizability. This must be negative for any set of values of the V_t^a ; that is, the set of polarizability values α_{tu}^{ab} must be positive definite. The program reports any failure in positive definiteness, though it is not guaranteed to be 100% reliable in this. Such cases may occur when the data points do not provide enough information to characterise the polarizability parameters accurately. An example is where an atom is ‘buried’, i.e., surrounded by other atoms so that all the data points are some distance away from it.

To deal with this problem, penalty functions are used. At present only the simplest case is implemented: each penalty function is of the form $F = s_i(p_i - p_i^0)^2$, where p_i^0 is an ‘anchor’ value and s_i is the strength of the penalty. The penalty section of the data file takes the form

PENALTIES

strength s

name anchor [strength]

:

END

The value s is taken as the strength for all following parameters for which no strength value is supplied. The strength specification may be repeated if different values are required for different sets of parameters. The strength coefficient s may be zero, in which case no penalty is applied to subsequent parameters listed without a strength value. This means that a list of anchor values, perhaps derived from another calculation, can be included in the file without any need to delete them or comment them out if the corresponding parameters are to be unconstrained.

Suitable strength values are in the range 10^{-3} – 10^{-4} or less. The program prints the anchor values and the fitted parameters, to allow the effect of the penalties to be assessed.

When CAMCASP generates the data files for PFit it includes a suitable set of penalties.

3.9 Other control commands

SVD [THRESHOLD t] [KEEP n]

Use singular value decomposition to solve the least-squares equations. KEEP n specifies that the n largest singular values are to be retained; THRESHOLD t specifies that singular values smaller than t times the largest one are to be discarded.

This option has been found unnecessary so far.

DAMPING b

Use Tang–Toennies damped interaction functions with damping parameter b in the polarizability model.

3.10 Running the calculation

Start

Calculate the fitted polarizabilities.

Print *options*

Various printing commands. The options are:

parameters

Print the fitted parameter values.

polarizabilities

Print the polarizabilities corresponding to the fitted parameters.

total [origin O_x O_y O_z]

Print the total molecular polarizabilities, referred to the specified origin. Default origin is (0, 0, 0).

Integrate

:

End

Calculate integrals over polarizabilities at imaginary frequency, and the dispersion coefficients. Each entry in this section specifies four sites, though normally the first two will be the same, as will the last two. For CO₂, for example:

Integrate

C C C C

C C 01 01

01 01 01 01

End

Technical details

Theory

The polarizability model is

$$\tilde{\alpha}_{PQ} = - \sum_{abtu} T_{0t}^{Pa} \alpha_{tu}^{ab} T_{u0}^{bQ}, \quad (1)$$

where

$$\alpha_{tu}^{ab} = \sum_k c_{tuk}^{ab} p_k, \quad (2)$$

with c_{tuk}^{ab} a fixed coefficient and p_k the fitting parameter.

Then we minimise the sum of squared deviations from the computed α_{PQ} :

$$S = \sum_{PQ} (\tilde{\alpha}_{PQ} - \alpha_{PQ})^2, \quad (3)$$

so

$$\begin{aligned} 0 &= \frac{1}{2} \frac{\partial S}{\partial p_k} = \sum_{PQ} (\tilde{\alpha}_{PQ} - \alpha_{PQ}) \frac{\partial \tilde{\alpha}_{PQ}}{\partial p_k} \\ &= - \sum_{PQ} (\tilde{\alpha}_{PQ} - \alpha_{PQ}) \sum_{abtu} T_{0t}^{Pa} T_{u0}^{bQ} c_{tuk}^{ab} \\ &= \sum_{PQ} \sum_{a'b't'u'} T_{0t'}^{Pa'} \sum_{k'} c_{t'u'k'}^{a'b'} p_{k'} T_{u'0}^{b'Q} \sum_{abtu} T_{0t}^{Pa} T_{u0}^{bQ} c_{tuk}^{ab} + \sum_{abtu} \alpha_{PQ} T_{0t}^{Pa} T_{u0}^{bQ} c_{tuk}^{ab} \\ &= \sum_{PQ} \sum_{k'} g_{PQ}(k) g_{PQ}(k') p_{k'} + \sum_{PQ} \alpha_{PQ} g_{PQ}(k), \end{aligned} \quad (4)$$

where

$$g_{PQ}(k) = \sum_{abtu} T_{0t}^{Pa} c_{tuk}^{ab} T_{u0}^{bQ}. \quad (5)$$

We can add penalty constraints to the problem. The most general form quadratic in the parameters is

$$V = \sum_{kk'} (p_k - p_k^0) v_{kk'} (p_{k'} - p_{k'}^0), \quad (6)$$

where p_k^0 is the preferred or anchor value for parameter p_k , and the matrix $v_{kk'}$ is symmetric and positive definite. Then

$$\frac{1}{2} \frac{\partial V}{\partial p_k} = \sum_{k'} v_{kk'} (p_{k'} - p_{k'}^0), \quad (7)$$

and the equations become

$$\sum_{k'} \left(\sum_{PQ} g_{PQ}(k) g_{PQ}(k') + v_{kk'} \right) p_{k'} = - \sum_{PQ} \alpha_{PQ} g_{PQ}(k) + \sum_{k'} v_{kk'} p_{k'}^0. \quad (8)$$

File formats

The binary response data file is a sequential unformatted file with the following record structure:

- Record 1: number of points, n . (Integer*4)
- Next n records: point positions as cartesian coordinates in a.u. ($3 \times \text{real*8}$)
- Next $n(n + 1)/2$ records: response value matrix in triangle format, i.e. in the order (1,1), (2,1), (2,2), (3,1), (3,2), (3,3), Values in a.u., one record for each element, real*8.

References

Stone, A. J. (2013) *The Theory of Intermolecular Forces*, Oxford University Press, Oxford, 2nd edn.