



ATOM TYPER AND ANALYSER (D_ATA)

USER MANUAL

VERSION 1.1, MAY 2025

Chin W Yong

Computational Materials and Molecular Science,
UK Research and Innovation,
Science and Technology Facility Council,
Daresbury Laboratory,
Warrington WA4 4AD,
Cheshire, UK

D_ATA is the property of STFC (under the body of UK Research and Innovation (UKRI)), Daresbury Laboratory and is issued free under licence to academic institutions pursuing scientific research of a non-commercial nature. Commercial organisations may be permitted a licence to use the package after negotiation with the owners. Daresbury Laboratory is the sole centre for distribution of the package. Under no account is it to be redistributed to third parties without consent of the owners.

The D_ATA manual assumes readers possess at least a basic knowledge in molecular simulation and general chemistry knowledge. The manual mainly describes the functionality of D_ATA but does not describe how and when to use the features offered by D_ATA. For latter case, please refer to DL_Software Digital Guide web resources (dl-sdg.github.io)

D_ATA is a computer program package written in C that primarily serves as a post-analysis software tool for data containing atomistic information, especially molecular trajectories produced from computational packages that carry out molecular dynamics (MD) simulations.

If you use D_ATA in your work, please include the following reference in your publication:

D_ATA – Atom Typer and Analyser, version 1.1, https://www.ccp5.ac.uk/D_ATA

Disclaimer

While extensive tests have been made to ensure smooth working of D_ATA and the accuracy of the supplied data parameters, neither the UKRI, STFC, EPSRC, CoSeC, CCP5, HEC-MCC nor the author of the D_ATA package or its derivatives guarantee that the software package is free from error. We disclaim any responsibility for any failure, inaccuracy, harm and damage to your projects, theoretical or experimental works because of using D_ATA.

Table of Contents

1. Introduction	5
1.1 Functions	5
1.2 Quick Start Guide	5
1.3 Program Structure	6
1.4 Program Operations	7
1.5 Release History and Features	7
2. D_ATA Input.....	9
2.1 Path Entry File (<i>d_ata_path</i>)	9
2.2 Master Input Configuration File (INPUT_FILE).....	9
2.3 D_ATA Control File (CONTROL_FILE)	10
2.4 Configuration Template File	12
2.4.1 PDB and GROMACS gro template files	12
2.4.2 PDB Input files without template	13
2.4.3 Molecule-labelling template file, <i>d_ata.template</i>	13
2.5 Interaction mapping library file (<i>DLF_map</i>)	14
3. D_ATA Output.....	16
3.1 The Chemical Structure File (<i>csf</i>)	16
3.2 The general output file, <i>d_ata.output</i>	17
3.3 General results output file, <i>d_ata.results</i>	18
3.4 Secondary results output file, <i>d_ata.results2</i>	21
3.5 Count-Time profiles (<i>count_XX.results</i>)	23
4. Atom Configuration File	24
4.1 The xyz format	24
4.2 The Gromacs gro format	25
4.3 PDB format	26
4.4 DL_POLY HISTORY format	27
5. Atom typing (DL_F Notation).....	28
5.1 The DL_F Atoms (Atoms).....	28
5.2 Chemical Group (CG) and Chemical Group Index (CGI).....	28
5.3 Primary DL_F Tokens	29
5.4 Secondary DL_F Tokens.....	29
5.5 Notation Rules	30
5.6 Notation Format.....	30
5.7 Organic Molecules Examples.....	31

5.8 Further information	35
6. Atomic Interactions (DANAI).....	36
6.1 Macro-interactions.....	36
6.2 Micro-interactions.....	37
6.3 Notation rules	38
6.4 Further information	40
7. Interaction counts	42
7.1 [L2] Interactions	42
7.2 [L3] Interactions	42
7.3 [R3] Interactions.....	44
7.4 [L4] Interactions	44
8. Non-bonded Interactions	46
8.1 Hydrophobic Interactions	46
8.2 Hydrogen bond Interactions.....	48
9. Analysis and Calculations	49
9.1 Pearson correlation coefficient, R	49
9.2 Linearity parameter, l	49
9.3 Triangular parameter, t_r	50
10. Example Structures	53
11. Glossary.....	55

1. Introduction

In Chemistry, the non-chemical function and behaviour of all molecular materials are underpinned by the strength and directivity of the constituent inter-atomic interactions. D_ATA is developed to identify these interactions, and from such, to classify, annotate and quantify them to provide statistical and structural information of a molecular systems.

D_ATA contains two unique features that are implemented in [DL FIELD](#) and [DL ANALYSER](#), respectively: DL_F Notation to describe the chemical characteristics of atoms and DANAI, which is a chemical language construct to annotate atomic interactions. Both features are complimentary to each other to create a universal atomic expression syntax that are independent of how the molecular systems are derived, either from simulations or experimental measurements.

The D_ATA program enables users to describe atomic interactions that is searchable and discoverable without resorting to traditional pictorial or diagrammatic illustrations. The program can rationalise complex molecular structural interactions to carry out cheminformatic, comparative and statistical data analytics.

1.1 Functions

In summary, D_ATA contains the following features:

- (1) Reads atomic configurations expressed in DL_POLY HISTORY, xyz, PDB or Gromacs *gro* format.
- (2) Identification and annotation of the chemical nature of constituent atoms in molecular systems (Typer).
- (3) Identification and annotation of atomic and molecular non-bonded interactions in the system (Typer and Analyser).
- (4) Quantification of non-bonded interactions to rationalise the roles these interactions play in the molecular systems (Analyser).
- (5) Characterisation of geometrical orientation of interactions (Analyser).

1.2 Quick Start Guide

The following instructions show you how to run D_ATA to demonstrate its capabilities.

Step 1: Install D_ATA - Once it is registered, users will be sent a file attachment called *d_ata_1.1.tar.gz*, which contains the Program's source code.

To extract and install the program, type:

```
tar -xvf d_ata_1.1.tar.gz
```

This creates a new directory called *d_ata_1.1/* which contains all D_ATA file components.

Step 2: Compile D_ATA – go to the *source/* directory. Type 'make'. This will produce the *d_ata.exe* (or just *d_ata*) executable file in the D_ATA home directory.

Step 3: Open and inspect *dl_a_path* file. This is the first entry point for D_ATA to locate various file components to run the program. Make sure you are using the correct control and master input files.

```
control = d_ata.control  
input = d_ata.input
```

Step 4: Save (if require) and close *dl_f_path* and run D_ATA, by simply typing:

```
./d_ata
```

D_ATA will read the configuration file defined in the *d_ata.input* file and carry out atom typing and analysis accordingly.

1.3 Program Structure

As the name suggests, the D_ATA program source is broadly separated into two main components: atom typing (Typer) and interaction analysis (Analyser).

D_ATA contains the following file structures:

- (1) Source codes together with a *Makefile*, in the *source/* folder.
- (2) Data library files located in the *lib/* directory. These files are
 - Interaction reference data file, *DLF_map*
 - Chemical Group reference file, *DLF_notation*
- (3) D_ATA control file (CONTROL_FILE), *d_ata.control*. It contains all the available options to operate D_ATA.
- (4) D_ATA input file (INPUT_FILE), *d_ata.input*. This is the master input file which contains at least one input configuration file or a collection of molecular trajectory files.
- (5) The path entry file, *d_ata_path*, which specifies the directory paths and names of various file components. The *d_ata_path* must always locate at the D_ATA home directory (the folder where the D_ATA executable file is located).

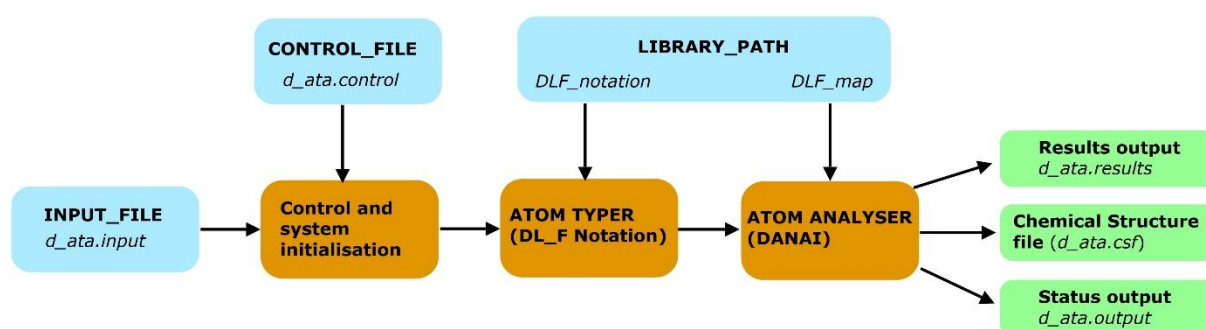
Upon a successful run, D_ATA will produce the following output files:

- (1) The status output file, *d_ata.output* in the OUTPUT_PATH (as define in the *d_ata_path* file). By defaults, this is *output/* folder, which provides general information and analysis status.
- (2) The results output file, *d_ata.results* in the OUTPUT_PATH folder.
- (3) Secondary results output file, *d_ata.results2* in the OUTPUT_PATH.
- (4) The results output files, named as *count_XX.results*, that list the count versus time data for various interaction modes. The XX refers to the macro-interaction.
- (5) Chemical structure file, *d_ata.csf* in the OUTPUT_PATH folder.

1.4 Program Operations

D_ATA is operated via several [input files](#) without any scripting capability. The location of each file is defined in a special *path entry* file called *d_ata_path* which must be located at the D_ATA home directory, or the location where the main executable D_ATA program is located.

Below shows a flowchart of the program operations.



1.5 Release History and Features

The following lists the development history of D_ATA. Collectively, it is a complete list of D_ATA features shown in a timeline order.

D_ATA version 1.0, released November 2024 contains the following features:

- Read xyz or Gromacs *gro* structures as input configurations.
- Read PDB or Gromacs *gro* file as a template to determine amino acid residues involved in interactions.
- Determine and annotate chemical nature of atoms in DL_F Notation.
- Detect and identify binary type [L2] interactions.
- Detect and identify ternary type [L3] linear interactions.
- Detect and identify [L4] linear interactions.
- Detect and identify [R3] ring enclosure interactions.
- Able to detect hydrogen bond (HB) interactions.
- Able to detect hydrophobic (HP) interactions for alkyl carbons, including those from amino acid residues.
- Carry out averages and Pearson correlation analysis for various modes of interactions detected.
- Measure linearity parameter for [Lx] conformation, where $x > 2$.
- Measure triangular order parameter for [R3] conformation.

D_ATA version 1.1, released May 2025 contains the following additional features or changes:

- Reads *d_ata.template*, a special template file, to define labels for internally identified molecules in the system.
- Produce secondary results file (*d_ata.results2*), to indicate detailed interactions among various molecular groups.
- Reads PDB files as input configurations.

- Reads DL_POLY HISTORY trajectory files as input configurations.

2. D_ATA Input

This chapter describes various file components that are needed to operate D_ATA.

2.1 Path Entry File (*d_ata_path*)

This file specifies the directory paths and names of various file components, including the D_ATA *control* input filename (CONTROL_FILE) and results output files. The *d_ata_path* cannot be changed and must always locate at the D_ATA home directory (the folder where the D_ATA executable file is located).

When D_ATA is run, *d_ata_path* is the first file to read, to provide information how to read and write files. Each file precedes with a FILE type and an equal '=' sign. The default values for each FILE type are shown below:

```
# Directory paths for D_ATA

CONTROL_FILE = d_ata.control
INPUT_FILE = d_ata.input
OUTPUT_PATH = output/
LIBRARY_PATH = lib/
```

For the first two types, CONTROL_FILE and INPUT_FILE, the default filenames are defined without specific directory path. This means they are in the D_ATA home directory where the D_ATA executable file is located. Of course, these files can be renamed and located elsewhere where location path can be specified.

The third type is the OUTPUT_PATH, which defines **only** the *path* where output files (such as *d_ata.output*, *d_ata.csf* etc.) will be located. This also includes the general results file, *d_ata.results*, and count-profile files, *count_XX.results*.

Finally, the LIBRARY_PATH defines the path where the library files, *DLF_notation* and *DLF_map* are located.

2.2 Master Input Configuration File (INPUT_FILE)

Before D_ATA program can be operated, at least one input configuration file must be provided and defined in the INPUT_FILE (*d_ata.input* by default), which is defined in *d_ata_path* file.

The INPUT_FILE is a master text file contains a collection of configuration files to be read and analysed as shown below:

```
3
usr/name/conf1.xyz
usr/name/conf2.xyz
usr/name/conf3.xyz
```

In this example, the INPUT_FILE lists three configuration files, *conf1.xyz*, *conf2.xyz* and *conf3.xyz*. Each file can contain either a single atomic configuration or a collection of configurations such as atomic trajectories obtained from molecular dynamics simulations.

D_ATA can recognise different types of input configurations. For more details, please see [Chapter 4 Atomic Configuration File](#).

When D_ATA is run, it will extract the first configuration in the first file entry, which is *1.xyz* in this example. It will then carry out the atom typing procedure and annotate every atom in the system with [DL F Notation](#). This information is written out into a chemical structure file called *d_ata.csf* in the OUTPUT_PATH folder.

If Program Mode is set to 2 (from Option **2** of the CONTROL_FILE), D_ATA will assume all subsequent configurations contain the same number of atoms arranged in the same sequence as the first configuration. In other words, they belong to the same system configuration. In addition, D_ATA assumes there is no chemical reaction, or there is no breaking and forming of covalent bonds in the trajectory files.

If Program Mode is set to 1, then D_ATA will read the files in succession and carry out independent atom typing only. In this case the input files do not have to be similar configurations.

2.3 D_ATA Control File (CONTROL_FILE)

Program operations are controlled by a single CONTROL_FILE (defined with a filename called *d_ata.control* by default), which is defined in the *d_ata_path* file. The CONTROL_FILE is a normal text file which can contain information up to 120 columns for each line.

All possible options are listed in the file. Each option contains a brief description how it is used and possible input values for each option. In addition, the sequence of these options is fixed and read sequentially by D_ATA. This is also the file that will access most frequently by users. Above shows an example of a CONTROL_FILE.

```

1 Title for control file.
2 2      * Program mode (1=Typer, 2 = Typer and Analyser)
3 1      * All-inclusive interaction mode (1=yes, 0=no)
4 none   * Template file (PDB or gro only)
5 auto   * Periodic boundary condition type
6 40.0 0.0 0.0 * Cell vector a (x, y, z)
7 0.0 40.0 0.0 * Cell vector b (x, y, z)
8 0.0 0.0 40.0 * Cell vector c (x, y, z)
9 all     * Group A: Atom range. Min Max or 'all'
10 1      * Molecule-based analysis (0 = off, 1=between molecule, 2=within molecule).
11 3      * No of configuration to skip.
12 1      * DANAI tier level.
13 0.005  * Average count fraction threshold
14 2.5    * Distance threshold for H...D in HB interactions
15 120.0  * Angle threshold for HB interactions
16 4.5    * Distance threshold for C...C in HP interactions
17 0      * HB analysis
18 1      * HP analysis
  
```

The bold red numerical values do not form part of the CONTROL_FILE. They are shown as Option labels for purposes of illustrations in this Manual. The general features of the control file are as follows:

Option **1** Title for the CONTROL_FILE.

Option **2** Program Mode Switch to indicate D_ATA is used as either a Typer (1), or Typer and Analyser (2). If the value is 1 it means D_ATA only carried out atom typing and write out the information into chemical structure files (*csf*). No atomic interaction analysis will be carried out. This means D_ATA can read more than one configuration files of different formats in the INPUT_FILE, each with a different total number of atoms in the system. The typing information will be written out to several *csf* files (*d_ata_X.csf*), one for each input configuration file and $X = 1$ to total number of configuration files specified.

If the value is 2, then D_ATA will carry out, firstly, the atom typing, and, secondly, the analysis process. In this case, all configuration files must belong to the same system. That is, they contain the same number of atoms and arranged in the same sequence. Furthermore, only one *csf* file, *d_ata.csf*, will be written out.

Option **3** All-inclusive interaction mode. If this is turn on (1), then D_ATA **will not** distinguish atoms whether they are interacted in isolation or not, according to the DANAI statement. This means Atoms in all [DANAI](#) statements will be expressed in capital letters without small letters. If it is turn off (0), then D_ATA will separate similar interactions to different DANAI expressions, which can contain a mixture of capital letters and small letters.

Option **4** Optional template file, in PDB, Gromacs gro format, or a special *d_ata.template* file. If template file is not supplied or not available, put *none*. See [Section 2.4](#) for more details.

Option **5** Periodic boundary condition flag. The value represents the type of periodic boundary condition for the molecular system of interest. Zero means no (open) boundary. If 'auto' is inserted, this instructs D_ATA to obtain the periodic boundary information from the xyz configuration files that contain the cell parameter information (CRYST statement).

Option **6-8** The following three rows of number define the cell vectors **a**, **b**, and **c** according to DL_POLY notation. If the keyword 'auto' or zero value is defined in Option **3**, then these cell vectors will be ignored. Otherwise, the periodic boundary condition flag defined in the Option **4** must match with the simulation box type.

Option **9** Atom index range defines atoms to be analysed, from minimum to maximum. To account for all atoms, then put the word *all*.

Option **10** This option instructs D_ATA how to analyse the atomic interactions: The value 1 means carry out analysis for atoms originated from *different* molecules. The value 2 means only carry out analysis for atoms situated *within* a molecule. The value 0 deactivates this option, which means carry out analysis for all atoms and do not consider their molecular origins.

Option **11** This indicates every number of configurations to skip before carrying out analysis on the next configuration. Example above instructs D_ATA to ignore three configurations before carry out analysis on the next configuration. Skipping of configurations are reported in the [d_ata.output](#) file.

Option **12** DANAI tier level. This instructs D_ATA to carry out analysis to the extent defines by the tier levels. This is not being used for now.

Option **13** Define significance of DANAI interaction counts. This is defined as a ratio:

$$r_n = \frac{\text{interaction count } n}{\text{Maximum count}}$$

Where n is some DANAI interaction statement. The Maximum count is the largest average count among all DANAI interactions detected for a given class of interaction. If r_n is less than the threshold specified, then the average count for that interaction statement will be considered as insignificant and ignored. To show all counts, insert the value 0.0. Note that for value of 1.0, no count will be shown. Please refer to [Section 3.3](#) as an example.

Option **14** and **15** Set distance and angle criteria for D_ATA to recognise an HB interaction. See [Section 8.2](#) for more details.

Option **16** Set distance threshold for alkyl C...C distance in an HP interaction.

Option **17** Carry out hydrogen bond interaction detection and analysis. The values *1* or *0* respectively activates or deactivates this option.

Option **18** Carry out hydrophobic detection and analysis among the alkyl carbons. The values *1* or *0* respectively activates or deactivates this option.

2.4 Configuration Template File

Option **4** of the CONTROL_FILE allows user to provide more information about the atoms and molecules in the system. This is particularly useful if additional information is needed to indicate molecule groups or identity of molecular residues that participate in interactions.

If such additional information is available, D_ATA will write out additional analysis information in a secondary results file, *d_ata.results2*, to provide details such as residue labels that involve in the interactions. Otherwise, D_ATA will only write out results in a *d_ata.results* file which, by default, will always be produced.

2.4.1 PDB and GROMACS gro template files

The atom sequence of a template file must be the same as those of input configuration file. **Only PDB and Gromacs gro files can be used as a template file.** These template files are useful especially when xyz or DL_POLY HISTORY formats are used as the input configurations. Note that total atoms in a template file can be smaller than the input configuration file but atom sequence in the template file must match with that of input configuration files. This means that the atom index in the template file must always start at index 1, or the first atom in the sequence.

For a PDB template file, it should include the residues and their ID number sequence and, if possible, chain identifier or molecular group. For a GROMACS gro template file, only the residues and their ID sequences will be provided but does not include other details such as molecular group.

D_ATA will use this information to determine the relevant residues (such as those of amino acids in proteins) involved in interactions (such as HB or HP interactions).

2.4.2 PDB Input files without template

If no PDB or gro template file is supplied, D_ATA will extract additional information at the very first frame of the input configuration file and use this as a PDB template and automatically create the secondary results file, *d_ata.results2*. Obviously, it is assumed all other subsequent PDB configurations have identical index sequences. For more information about *d_ata.results2* please refer to [Section 3.4](#).

Note that D_ATA will give extract additional information from the PDB input files even if *d_ata.template* file (see below) is called in the CONTROL_FILE.

2.4.3 Molecule-labelling template file, *d_ata.template*

In addition to PDB and gro template files, D_ATA also accepts a special molecule-labelling template file, called the *d_ata.template* in Option **4** of the CONTROL_FILE. This file is particularly useful when using along with matching PDB input configuration files that **do not contain chain identifier or molecular group label**.

For example, consider a PDB system configuration that contains several proteins that do not form covalent bonds with one another. D_ATA can identify them as different Molecules and **internally** label them accordingly as shown in the *d_ata.output* file as A, B and C:

```
...
...
-----SETUP MOLECULES-----

Identified types of molecules as follows:

Molecule: A
Total molecules: 1
Total atoms per molecule: 10270
Molecular weight: 74236.504670
Composition: C3278 H5051 O1007 N902 S32

Molecule: B
Total molecules: 1
Total atoms per molecule: 3721
Molecular weight: 26224.274380
Composition: C1160 H1874 O373 N312 S2

Molecule: C
Total molecules: 1
Total atoms per molecule: 7732
Molecular weight: 55363.771860
Composition: C2491 H3818 O744 N669 S10
...
...
```

These internal labels will not be used in interaction analysis and subsequent results output in *d_ata.results2* file will only indicate Atoms from which residues that meet the interaction criteria but not the protein molecules from which they belong to.

However, the following *d_ata.template* file can be provided, to indicate the actual labels for these Molecules as follows:

```
# D_ATA standard template file
# Reference list for molecular internal label and the
# actual user-defined name

A = SpeA
B = FkpA
C = OpgG
```

This information instructs D_ATA to label the protein molecules as follows:

```
...
...
-----SETUP MOLECULES-----

Identified types of molecules as follows:

Molecule: A = SpeA
Total molecules: 1
Total atoms per molecule: 10270
Molecular weight: 74236.504670
Composition: C3278 H5051 O1007 N902 S32

Molecule: B = FkpA
Total molecules: 1
Total atoms per molecule: 3721
Molecular weight: 26224.274380
Composition: C1160 H1874 O373 N312 S2

Molecule: C = OpgG
Total molecules: 1
Total atoms per molecule: 7732
Molecular weight: 55363.771860
Composition: C2491 H3818 O744 N669 S10
...
...
```

Subsequent *d_ata.results2* file will then provide a more complete picture of atomic interactions. That is, interacting pairs between Atoms from which residues and protein molecules to which they belong to.

2.5 Interaction mapping library file (*DLF_map*)

Located in the LIBRARY_PATH, *DLF_map* file is the master index file that lists different interaction types, the CGs involve and the atomic species that participate in interactions. Below show as an example, for the [HB interactions](#):

```

---HB
water      O H *
alcohol    O H *
ketone     O * *
ester      O * *
carboxylic O H *
amine      N H *
aniline    N H *
amide      N O H
phenol     O H *
enol       O H *
serine     O H *
threonine  O H *
tyrosine   O H *

```

D_ATA will use this file to search for relevant interactions present in the system and look for Atomic species that involve in such interactions. If new *CGs* are present in atomic configurations but not present in the *DLF_map* file, then they must be included in the file for D_ATA to detect any interactions involving these new *CGs*.

Note that, for HB interactions, some of the *CGs* listed above only contain the acceptor species, namely, the oxygen atoms. For instance, the *ketone* and *ester CGs*. D_ATA will automatically look for other *CGs* that contain the polar hydrogen that participates in an HB interaction.

3. D_ATA Output

This chapter describes how results are produced from D_ATA written to various file components.

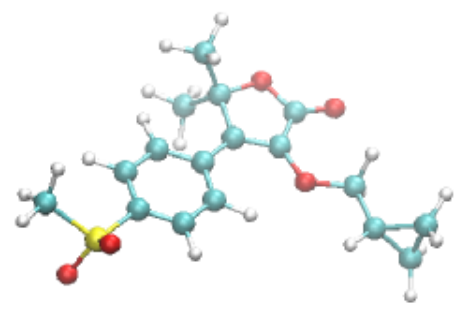
3.1 The Chemical Structure File (csf)

The *csf* is a file that defines the chemical identity in DL_F Notation for every atom identified in the system. Each time when D_ATA is run, a *csf* will be generated, called *d_ata.csf* and located in the OUTPUT_PATH directory.

Below shows a section of *csf* generated for a drug molecule called firocoxib, an anti-inflammatory drug for animals.

Chemical structure file.
Created from D_ATA version 1.00
Atomic information extracted from example_structures/firocoxib.xyz

ATOM_TYPE	imp_flag	neighbour_number	neighbour1	neighbour2 ...
1	C1_benzene	1 3	2	6 8
2	C2_benzene	1 3	1	3 24
3	C3_benzene	1 3	2	4 25
4	C4_benzene	1 3	3	5 7
5	C5_benzene	1 3	4	6 26
6	C6_benzene	1 3	1	5 27
7	CE_alkene	1 3	4	9 12
8	S_sulphone	0 4	1	21 22 23
9	Cq_cycloalkane	0 4	7	10 19 20
10	OL_ester	0 2	9	11
11	C_ester	1 3	10	12 13
12	CE_alkene	1 3	7	11 14
13	OE_ester	0 1	11	
14	O_ether	0 2	12	15
15	Cs_alkane	0 4	14	16 28 29
16	CH_cyclopropyl	0 4	15	17 18 30
17	CH_cyclopropyl	0 4	16	18 31 32
18	CH_cyclopropyl	0 4	16	17 33 34
19	Cp_alkane	0 4	9	35 36 37
20	Cp_alkane	0 4	9	38 39 40
21	Cp_alkane	0 4	8	41 42 43
22	O_sulphone	0 1	8	
23	O_sulphone	0 1	8	
24	HC_benzene	0 1	2	
25	HC_benzene	0 1	3	
26	HC_benzene	0 1	5	
27	HC_benzene	0 1	6	
28	HC_alkane	0 1	15	
29	HC_alkane	0 1	15	
30	HC_cyclopropyl	0 1	16	
31	HC_cyclopropyl	0 1	17	
32	HC_cyclopropyl	0 1	17	
33	HC_cyclopropyl	0 1	18	
34	HC_cyclopropyl	0 1	18	
35	HC_alkane	0 1	19	
36	HC_alkane	0 1	19	
...				
...				



configuration file of the molecule

connectivity information

atom types in DL_F Notation

atom index number

If the Program Mode is set to 1 (Option **2** in the CONTROL_FILE), then the *csf* would be the only output file produced. If there are more than one configuration files in the INPUT_FILE, then one *csf* file will be produced for each configuration file, named as *d_ata_XX.csf* where XX is the configuration number.

3.2 The general output file, *d_ata.output*

This file provides general information and analysis status, including control options selected by the user. The file is located at the OUTPUT_PATH folder.

The file also includes system information such as types of molecules identified, relevant Chemical Groups present in the system and types of interaction would be identified. An example of the output is shown below:

```
...
...
-----SETUP MOLECULES-----
Identified types of molecules as follows:
Molecule: A
Total molecules: 2808
Total atoms per molecule: 17
Molecular weight: 102.136000
Composition: C5 H10 O2

-----END SETUP MOLECULES-----

-----ATOM ANALYSER INITIALISATION-----

The following Chemical Groups (CGIs) identified:
alkane (1)
carboxylic (20)

*** Expected interaction types in the system ***

1 - Hydrogen bond (HB)
Atoms involve:
O...H (CGI: 20)

Detection criteria:
H..D distance <= 2.500000   Angle A--H...D >= 120.000000

2 - Hydrophobic (HP)
Atoms involve:
C...C (CGI: 1)

Detection criteria:
C(alkyl)...C(alkyl) distance <= 4.500000

...
...
```

In addition, *d_ata.output* file also provides information on analysis progress as shown below:

```

Read from Frame 1 (Time: 0.001500)
Total atom in system = 47736
Total atom for analysis = 47736

XX - Skip frame 1 (MD time = 0.001500) and not account for analysis.
XX - Skip frame 2 (MD time = 3.001500) and not account for analysis.
XX - Skip frame 3 (MD time = 6.001500) and not account for analysis.

-----BEGIN ANALYSIS-----

Frame 4 (Time: 9.001500)
Periodic cell information (from input file):
Cell vectors:
  79.76400  0.00000  0.00000
 -0.00000  79.76400  0.00000
 -0.00000 -0.00000  79.76400
Cell parameters:
  79.76400  79.76400  79.76400  90.00000  90.00000  90.00000
imcon value: 1

XX - Skip frame 5 (MD time = 12.001500) and not account for analysis.
XX - Skip frame 6 (MD time = 15.001500) and not account for analysis.
XX - Skip frame 7 (MD time = 18.001500) and not account for analysis.

Frame 8 (Time: 21.001500)
Periodic cell information (from input file):
Cell vectors:
  79.86000  0.00000  0.00000
 -0.00000  79.86000  0.00000
 -0.00000 -0.00000  79.86000
Cell parameters:
  79.86000  79.86000  79.86000  90.00000  90.00000  90.00000
imcon value: 1

```

D_ATA reports the system size along with the MD simulation time. This information is extracted from the input xyz files, to indicate the data are correctly read. Some of the configurations are skipped, according to the value set for Option **11** in the CONTROL_FILE. Example above indicates three frames are skipped for every frame that has been analysed.

3.3 General results output file, *d_ata.results*

When Program Mode 2 is set (Option **2**), D_ATA will carry out atomic interaction analysis and the results will be shown in *d_ata.results*, located in the OUTPUT_PATH folder.

In general, this file reports, firstly, types and modes of interactions identified for each analysed time frame and, secondly, the overall averages, including correlations and structural order parameters, if applicable.

Example below illustrates a section of the results output for a molecular system consists of pure pentanoic acid derived from MD simulations.

-----ANALYSIS MODE-----

Carry out interaction analysis between molecules only.

According to control file options selected, the following interaction will be scanned:

Hydrogen bond (dipole-dipole)

Hydrophobic interactions (between alkyl groups)

-----END ANALYSIS MODE-----

HB_20_20 MD_time = 9.001500

```
1. [L2]O20E:H20O    = 1935
2. [L2]O20E:h20O    = 123
3. [L2]O20L:h20O    = 157
4. [L2]H20O:O20L    = 436
5. [L2]o20E:H20O    = 116
6. [L2]o20E:h20O    = 36
7. [L2]o20L:H20O    = 1
8. [L2]o20L:h20O    = 1
9. [L3]H20O:O20E:h20O = 36
10. [L3]H20O:O20E:H20O = 40
11. [L3]O20E:H20O:O20E = 2
12. [L3]O20L:H20O:O20E = 6
```

...

HP_1_1 MD_time = 9.001500

```
1. [L2]c1p:c1s      = 7396
2. [L2]c1p:C1s      = 619
3. [L2]c1p:c1p      = 2327
4. [L2]c1s:c1s      = 6146
5. [L2]c1s:C1s      = 744
6. [L2]c1s:C1p      = 31
7. [L2]c1p:C1p      = 33
8. [L2]C1s:C1s      = 36
9. [L2]C1p:C1s      = 8
10. [L3]c1s:c1p:c1p = 8039
11. [L3]c1s:c1p:c1s = 6487
```

...

HB_20_20 MD_time = 21.001500

```
1. [L2]O20E:H20O    = 1914
2. [L2]h20O:O20L    = 142
3. [L2]O20L:H20O    = 418
4. [L2]h20O:O20E    = 123
5. [L2]o20E:H20O    = 123
6. [L2]h20O:o20E    = 38
7. [L2]H20O:o20L    = 3
```

...

...

The outcome of the analysis results depends very much on the selected analysis mode in the CONTROL_FILE.

D_ATA will identify interaction types and their occurrences in terms of number of counts for each analysed configuration.

This is macro-interaction in DANAI syntax. In this case, HP_1_1 means hydrophobic interactions between alkyl carbons.

These are micro-interaction expressions, or different modes of interactions under HP_1_1, along with the corresponding number of unique counts for the relevant interactions.

Note that elements are expressed in either capital or small letters, indicating it is either participating in exclusive (capital letter) or non-exclusive (small letter) interactions. For more details, please see [Chapter 6.3](#).

If all-inclusive interaction mode (Option **3**) is switched on, then D_ATA will lump all similar interactions involving similar chemical species into one DANAI statement and **Atoms are expressed in capital letters only**. That is, D_ATA no longer distinguishes interactions that are either exclusive or non-exclusive for an Atom. Below shows analysis outcome for the same configuration as above.

```

...
Hydrophobic interactions (between alkyl groups)
-----END ANALYSIS MODE-----

HB_20_20 MD_time = 9.001500
1. [L2]O20E:H200 = 2210
2. [L2]O20L:H200 = 595
3. [L3]H200:O20E:H200 = 76
4. [L3]O20E:H200:O20E = 3
5. [L3]O20L:H200:O20E = 9
6. [L3]O20L:H200:O20L = 3
7. [L3]H200:O20L:H200 = 1
8. [L4]H200:O20E:H200:O20L = 1
9. [L4]O20E:H200:O20E:H200 = 1
...

```

When all-inclusive mode is switched on in the CONTROL_FILE, then sum of the interaction counts will consist of various interaction components of Atoms that participate in both exclusive interactions and non-exclusive interactions.

For example, the total counts for this interaction mode is the sum of $1935+123+116+36 = 2210$ (compare with above).

Once all simulation frames have been processed, D_ATA carries out time averaging over all analysed configurations as follows:

```

=====
AVERAGE COUNTS
=====
Total Samples: 75

HB_20_20
-----
Unique interactions: 2783.120000 deviation = 16.039086
DANAI
mean deviation
[L2]O20E:H200 = 1930.706667 19.429718
[L2]O20E:h200 = 117.666667 10.395298
[L2]O20L:h200 = 139.213333 11.949107
[L2]H200:O20L = 433.213333 15.304286
[L2]o20E:H200 = 123.320000 12.733326
[L2]o20E:h200 = 34.626667 5.447381
[L2]o20L:H200 = 2.640000 2.017523
[L3]H200:O20E:h200 = 32.506667 5.164942
[L3]H200:O20E:H200 = 45.266667 5.985167
[L3]O20E:H200:O20E = 4.226667 2.069289
[L3]O20L:H200:O20E = 7.533333 2.664999
[L3]O20L:H200:O20L = 2.240000 1.504128
[L3]O20E:H200:o20E = 3.666667 1.975404
[L3]o20E:H200:O20L = 2.280000 1.238924
[L4]O20E:H200:O20E:H200 = 2.106667 1.456693

Interactions with an average ratio count less than 0.001000 are not shown.
(ratio = average_count/max_average_count)

HP_1_1
-----
Unique interactions: 17134.933333 deviation = 150.042957

DANAI
mean deviation
[L2]c1p:c1s = 7255.653333 93.116485
[L2]c1p:C1s = 650.480000 24.249459
[L2]c1p:c1p = 2259.466667 36.950538
[L2]c1s:c1s = 6038.306667 96.858243
[L2]c1s:C1s = 792.093333 28.267849
...

```

Results are obtained by averaging over 75 configuration frames.

Average number of different interactions identified, in this case, for HB_20_20, or hydrogen bonds among carboxylic groups.

Only interactions with significant counts are shown, according to the threshold specified in the CONTROL_FILE.

After that, correlations and other quantities such as order parameters will be shown, if applicable.

3.4 Secondary results output file, *d_ata.results2*

When Program Mode 2 is set (Option **2**) and a template file is also supplied (Option **4**), D_ATA will carry out further analysis to identify the origin [L2] binary interactions. The results are then written out to a secondary results output file called *d_ata.results2*, which describe [L2] binary interactions in terms of residues and molecules to which such interaction belong. This information is useful to investigate molecular interactions in proteins.

Take for example, consider a complex molecular system consists of several interacting proteins. An MD simulation is run to produce a trajectory file which then translated into PDB format. D_ATA is used to carry out the analysis using a *d_ata.template* (See [Section 2.4.3](#)) file to label the proteins as follows:

```
# d_ata.template for the protein system.
# Reference list for molecular internal label and the actual user-defined name.

A = Ecotin
B = Tpx
C = TolB
D = OpgG
```

Then, select appropriate analysis options in the CONTROL_FILE and the resulting analysis mode is shown near towards the beginning *d_ata.results* and *d_ata.results2* as follows:

```
-----ANALYSIS MODE-----

Carry out interaction analysis between molecules only.

All-inclusive mode switched ON.
This means D_ATA does not distinguish between exclusive (isolated) and non-exclusive
interactions. All relevant interactions will be counted together for a given DANAI
expression.
All DLF Atoms will be expressed in capital letter cases.

According to control file options selected, the following interaction will be scanned:
Hydrogen bond (dipole-dipole)
Hydrophobic interactions (between alkyl groups)

-----END ANALYSIS MODE-----
```

Note that D_ATA will bypass any interactions occur within a molecular protein and consider only interactions that occur **between** the proteins. The matching *d_ata.results* and *d_ata.results2* are shown below, at time = 1900 ps (1.9 ns):

```
MD_time = 1900.0000: Total HP contacts for HP_912_914 = 41
MD_time = 1900.0000: Total HP contacts for HP_912_902 = 49
MD_time = 1900.0000: Total HP contacts for HP_917_917 = 23
MD_time = 1900.0000: Total HP contacts for HP_917_915 = 27
MD_time = 1900.0000: Total HP contacts for HP_917_908 = 47
MD_time = 1900.0000: Total HP contacts for HP_917_914 = 30
MD_time = 1900.0000: Total HP contacts for HP_917_902 = 56
MD_time = 1900.0000: Total HP contacts for HP_915_915 = 96
MD_time = 1900.0000: Total HP contacts for HP_915_908 = 10
MD_time = 1900.0000: Total HP contacts for HP_915_914 = 14
MD_time = 1900.0000: Total HP contacts for HP_915_902 = 3
MD_time = 1900.0000: Total HP contacts for HP_908_908 = 20
MD_time = 1900.0000: Total HP contacts for HP_908_914 = 31
MD_time = 1900.0000: Total HP contacts for HP_908_902 = 20
MD_time = 1900.0000: Total HP contacts for HP_914_914 = 8
MD_time = 1900.0000: Total HP contacts for HP_914_902 = 14
MD_time = 1900.0000: Total HP contacts for HP_902_902 = 223
```

For each analysed frame, D_ATA lists out number of unique interactions between two CGs. For example, HP_915_908 has a total of ten HP contacts.

From *d_ata.results*

```
HP_922_917 MD_time = 1900.000000
1. [L2]C922:C917 = 3
```

```
HP_915_908 MD_time = 1900.000000
1. [L2]C915:C908A = 1
2. [L3]C915:C908A:C915 = 1
```

```
HP_908_914 MD_time = 1900.000000
1. [L2]C908A:C914 = 1
```

Even then, D_ATA only picks a number of interaction as per the analysis mode specified. For example, for HP_915_908, there are only two different modes of such interactions that occur across two different proteins.

From *d_ata.results2*

Format display:

```
MD_time - Macro-interaction ([L2]micro-interaction) = readidue_A : residue_B
```

```
1900.0000 - HB_44_44 ([L2]H44N:O44) = MET397-TolB : LYS38-ecotin
1900.0000 - HP_922_917 ([L2]C922:C917) = VAL412-OpgG : PRO11-Tpx
1900.0000 - HP_915_908 ([L2]C915:C908A) = MET397-TolB : GLY39-ecotin
1900.0000 - HP_908_914 ([L2]C908A:C914) = GLY396-TolB : LYS38-ecotin
```

This file lists all inter-protein interactions that occur at a given time. For example, at t=1.9 ns, there is one HB interaction and three HP interactions.

This is HB interaction between two amide groups of the backbone chain. Specifically between the hydrogen amide from MET397 and and carbonyl O of LYS38.

Recall the illustrated example. This shows the HP interaction occurred between MET397 and GLY39 from TolB and Ecotin, respectively.

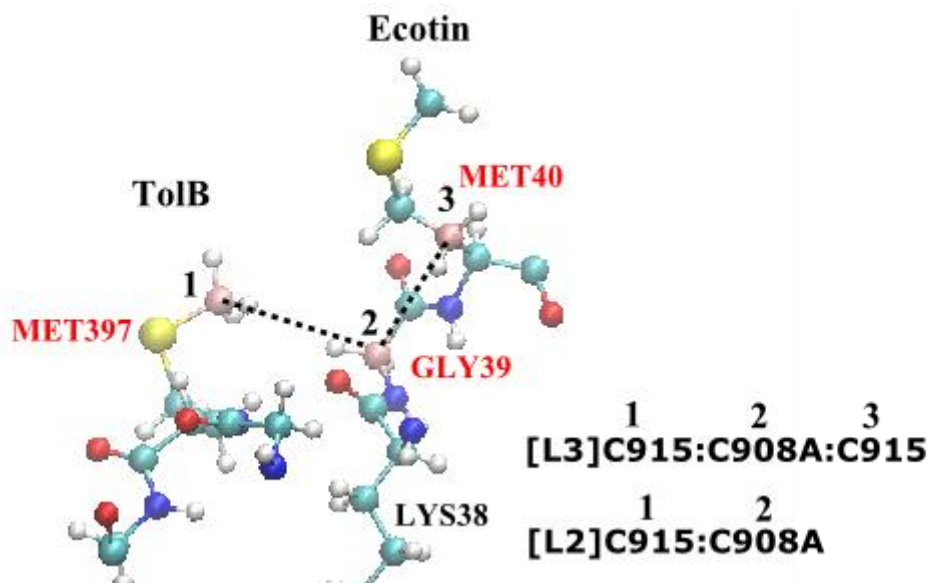
Recall that, from *d_ata.results*, at time=1.9 ns, there are two types of interactions for HP_915_908 (CGI 915 is methionine and CGI 908 is glycine):

```
HP_915_908 MD_time = 1900.000000
1. [L2]C915:C908A = 1
2. [L3]C915:C908A:C915 = 1
```

And yet *d_ata.results2* file shows there is only one unique interaction between two proteins:

```
1900.0000 - HP_915_908 ([L2]C915:C908A) = MET397-TolB : GLY39-ecotin
```

This means interaction 1 and 2 are related. It turns out [L2] is just a portion of the more extended [L3] interaction. The configuration involving the relevant residues are shown below.



3.5 Count-Time profiles (*count_XX.results*)

The *d_ata.results* file shows counts for all identified interactions for all analysed configurations. However, time-averages are only obtained for significant interactions as mentioned in Section 3.3. The count versus time data for these interactions are written out in a file named as *count_XX.results*, where *XX* is a macro-interaction. All relevant time-profile interactions are recorded in the file, located in OUTPUT_PATH folder.

Below shows an example of a small section of the file, *count_HP_1_1.results*. The first column is the MD time with the corresponding count value at the second column.

```
Counts-time profiles for HP_1_1
Only DANAI profiles with count threshold above 0.001000 are shown.
[L2]c1p:c1s
9.001500 7396
21.001500 7176
33.001500 7203
45.001500 7336
57.001500 6993
...
...
```


4. Atom Configuration File

For D_ATA to work, at least one atomic configuration file must be supplied to the [INPUT_FILE](#). Depending on the file format, atomic configuration files can contain either one system configuration, or multiple configurations. For the latter case, the molecular simulation time can be provided for each configuration, so that a time profile information can be obtained.

D_ATA recognises the following file formats:

4.1 The xyz format

For D_ATA to recognise this file format, filenames must be ended with an extension, *xyz*. This format can either contain a single or multiple configurations. For the latter case, atoms and their sequences must be identical for all configurations. This file format is the least informative that only contains the atom labels and the corresponding x, y and z coordinates. An example format is shown below:

```

No_of_atoms      time= 3.403 ps          step = 0.002
[Space, or some title text]
C  -168.977917  0.083913  -172.855802
O  -166.101003  0.011045  -172.847280
O  -163.233789  0.058050  -172.845844
H  -160.365398  0.091437  -172.887635
C  -157.509796  0.036537  -172.905057
...
...
No_of_atoms      time= 3.503 ps          step = 0.002
[Space, or some title text]
C  -168.947712  0.073760  -172.905155
O  -166.111531  0.142515  -172.867775
O  -163.193092  0.076674  -172.869657
H  -160.299232  0.134099  -172.840889
C  -157.412480  0.106132  -172.815243
...
...

```

This example shows two trajectory frames of MD time = 3.403 ps and 3.503 ps. This information (with the keyword **time=**) is needed for D_ATA to produce DANAI counts time profile in the results output file (*d_ata.results*). Both frames show atoms are arranged in the same sequence. D_ATA assumes there is no change in the atom sequences and the number of atoms for all frames.

The second line of each frame is the title section which can be left blank or contain some information which is ignored by D_ATA. However, if the title section contains the CRYST1 statement, which specifies the cell parameters for each frame, then the program will convert these parameters into corresponding cell vectors and apply periodic boundary effect. Note that Option **5** in the CONTROL_FILE must be set to *auto* for this to work.

Example below shows an atomic configuration that uses the CRYST1 statement.


```

451    time= 34.5000 ps  step = 0.001
CRYST1  83.908  20.7302253  17.9859383759  90.00  90.00  90.00 (P 1)
C  10.000000  1.390000  0.000000
C  10.000000  0.695000  1.204000
C  10.000000 -0.695000  1.204000
...
...

```

Note that usually element symbols are used to describe Atoms in molecules. For isolated Atoms, such as those of anion and cations that do not form bond with other Atoms, the charge symbols must also be specified. For example, Cl⁻ instead of Cl that is connected to other atoms in molecules. However, for Atoms with formal charges and connected to other Atoms, then no charge should be specified. For example, the N Atom in ammonium should be expressed as 'N', rather than 'N⁺'.

For neutral Atoms, a zero-charge value can be specified. For example, H0, Na0, refer to neutral hydrogen and sodium atoms respectively.

The xyz files can be used alongside with a PDB [template](#) file, to provide information that is not available in xyz format such as residue labels for amino acids in protein systems.

4.2 The Gromacs gro format

For D_ATA to recognise this format, filenames will be ended with the *.gro* extension. This format can only contain one configuration. It is an input configuration file for GROMACS molecular simulation package. To use this file format, the periodic boundary condition must be set to auto at Option **5** of the CONTROL_FILE. D_ATA will read the system size defined at the last line of the *gro* file.

Example below shows a portion of configuration for a protein.

```

Title for the system.
25490
1NMET  N   1 12.934  7.751 12.226 -0.0962  0.1537  0.0532
1NMET  H1  2 12.880  7.760 12.311 -0.1165 -1.9812  0.3239
1NMET  H2  3 13.016  7.810 12.225  1.1709 -1.4033  2.2349
1NMET  H3  4 12.959  7.654 12.213 -2.7833 -0.4698 -1.1706
1NMET  CA  5 12.849  7.790 12.110 -0.0842 -0.1902 -0.4425
1NMET  HA  6 12.904  7.753 12.024 -0.9901 -2.6443 -0.0290
1NMET  CB  7 12.719  7.708 12.100 -0.6320  0.0479  0.5699
1NMET  HB2 8 12.739  7.601 12.094  0.1472 -0.0481  3.2307
...
...
1NMET  HE3 17 12.275  7.664 12.094  0.2690 -3.1075 -1.1527
1NMET  C   18 12.825  7.942 12.107 -0.4894  0.4157 -0.2353
1NMET  O   19 12.836  8.007 12.210 -0.6360  0.0284 -0.0199
2SER   N   20 12.792  8.003 11.994 -0.2236  0.0818  0.4013
2SER   H   21 12.788  7.946 11.910 -1.3019  0.2833  0.3102
2SER   CA  22 12.769  8.147 11.974  0.3153  0.8000 -0.4856
2SER   HA  23 12.854  8.197 12.019  0.7583  1.2493 -1.7832
2SER   CB  24 12.774  8.181 11.826  1.0027  0.6630 -0.1939
2SER   HB2 25 12.692  8.131 11.775 -0.6032  1.2965  1.6499
2SER   HB3 26 12.765  8.287 11.800  0.7110  0.3532 -1.4050

```

Since atom labels rather than the actual element symbols are frequently defined in the gro file (as the example shows above), D_ATA will attempt to make a guess and convert these labels to element symbols before carrying typing and analysis.

4.3 PDB format

The PDB, or Protein Databank format is the most popular format for biomolecules such as proteins and nuclei acids (DNA, RNA). D_ATA adheres to strict PDB standard format and different information must contain within the appropriate columns in a PDB file (see below). However, not all information will be required by D_ATA and only those relevant to the conversion process will be discussed here.

Column range (inclusive)	Data	Remark
13-16	<i>Atom label</i>	If <i>element symbol</i> is not defined, then this data will be treated as the element symbol for the atom. Any numerical characters will be ignored.
18-20	<i>Residue names</i>	Residue names such as those of amino acids. Equivalent to the MOLECULE_KEYS.
21	<i>Reserves for residue names</i>	Some residue names such as those of carbohydrates may contain more than three characters. The fourth one will be located here.
22	<i>Chain ID</i>	If <i>Molecular Group name</i> is not defined, this information will be used as the Molecular Group for the atom.
23-26	<i>Residue sequence</i>	Numerical sequence for molecules such as amino acids or carbohydrates.
29-56	<i>X,y,z coordinates</i>	Location of the atom in x,y,z coordinates.
70-76	<i>Molecular group name</i>	Molecular group name definition. This will override <i>Chain ID</i> if it is defined.
77-79	<i>Element symbol</i>	The element symbol of the atom. If this is not defined, then <i>Atom label</i> will be used to determine the element symbol of the atom.

Unless element symbols are provided in column 77-79 of the PDB file, D_ATA will attempt to extract this information from the atom label columns. D_ATA could terminate if elements obtained from atom label columns are unsuccessful.

4.4 DL_POLY HISTORY format

This is the trajectory file produced from DL_POLY MD package. It is recognised as such by D_ATA if the input filenames contain the word 'HISTORY'. Below shows the top portion of a HISTORY file, which is a molecular system consists of ethyl acetate molecules in water.

```

Ethyl acetate with water.
      0      1    6400           11      140846
timestep      0    6400 0 1      0.001000      0.000000
      40.4583644273      0.0000000000      0.0000000000
      0.0000000000      40.4583644273      0.0000000000
      0.0000000000      0.0000000000      40.4583644273
CO4      1    12.011500      0.510000      0.000000
      15.19224985      -15.24682479      17.56645027
O      2    15.999400      -0.430000      0.000000
      14.11710013      -15.48526151      16.95452605
CT      3    12.011500      -0.180000      0.000000
      16.04805711      -14.01504849      17.25205615
OES      4    15.999400      -0.330000      0.000000
      15.75022824      -16.02734264      18.47192028
CT      5    12.011500      0.190000      0.000000
      15.05227963      -17.08090301      19.08499389
CT      6    12.011500      -0.180000      0.000000
      15.66431126      -17.64383846      -20.07989719
HAE      7    1.007970      0.060000      0.000000
      15.51528383      -13.56327308      16.38032485
HAE      8    1.007970      0.060000      0.000000
      16.17456885      -13.39596911      18.09814820
...
...

```

Atoms are described in terms of atom keys that are characteristics of the force field scheme, which in this example, is that of OPLS2005 generated by DL_FIELD. However, D_ATA can determine the element types based on the corresponding atomic weights.

A matching [PDB template](#) can be supplied along HISTORY files, to provide additional information such as residues such as those of protein systems.

5. Atom typing (DL_F Notation)

Molecules are made up of atoms connected by bonds and are represented by some molecular formulas expressed in terms of elemental symbols. However, in Chemistry, an element can have different chemical behaviour due to the different electronic environments depending on the arrangements of neighbouring atoms. In classical molecular simulations such as the molecule dynamics (MD), it is important to identify correctly an atom that is subjected to these various environments before a correct set of force field parameters can be assigned to it. The procedure to carry out such activities is called *atom typing*.

Usually, atom typing involves some form of symbolic annotations and atoms of the same element are distinguished from one another by these annotations called the *atom types*. However, there is no consensus what standard to use for atom types: different force field (FF) schemes use different atom typing of some arbitrary notation syntax, which is often cryptic in nature.

DL_F Notation is a universal atom typing notation which is the unique feature implemented in DL_FIELD. It is a consistent atom typing scheme that is contiguous across different force field schemes. The Notation avoids the use of cryptic symbols, and all atom types are described in self-explanatory, easy to understand form, while at the same time can precisely indicate the chemical nature of each atom in a molecule. This smoothens data transitions when converting molecular models among different FF schemes in DL_FIELD.

In D_ATA, DL_F Notation is used to annotate the chemical nature of every atom in the system. Only a minimum amount of information is needed from the atomic configuration files – the element symbols and the corresponding xyz coordinates.

D_ATA use the same atom typing engine implemented in DL_FIELD to determine atoms expressed in DL_F Notation. This information is written out into a [chemical structure file](#) (csf), *d_ata.csf*.

5.1 The DL_F Atoms (Atoms)

DL_F notation only applicable to a selection atoms in the Periodic Table of the Elements. These atoms are called DL_F Atoms, or simply Atoms within the DL_F Notation context. The following atomic elements are defined as DL_F Atoms:

- (1) Elements that are commonly found in organic molecules: H, C, O, N, S and P.
- (2) Halogens: F, Cl, Br, I
- (3) Noble gases: He, Ne, Ar, Kr, Xe
- (4) Alkali metals: Li, Na, K
- (5) Alkaline earth metals: Be, Mg, Ca, Sr, Ba, Ra
- (6) Transition metals: Fe, Co, Ni, Pd, Ag, Au
- (7) Silicon, as in organosilicon compounds. Si

D_ATA will give an error if an element is not an DL_F Atoms.

5.2 Chemical Group (CG) and Chemical Group Index (CGI)

Each atom is associated with a chemical identity called the Chemical Group (CG). A CG consists of a specific group of *DL_F Atoms* and bonds within a molecular structure that

exhibits a characteristic chemical behaviour. CGs are quite similar to functional groups in organic chemistry. Each CG is also represented by the unique CG index number, or *CGI* ranged between 1-9999.

Some examples of CG are *alkane*, *aldehyde*, *alcohol*, *acylhalide*, etc, with the corresponding unique *CGI* of 1, 17, 15, 18, respectively. For a complete list please refer to *DLF_notation* file in the *LIBRARY_PATH*.

In addition, CGs can also include groups of atoms, which can consist of several functional groups. These complex CGs are so named because they can have significant chemical behaviour such as those found in biological chemistry. For example, *hydantoin*, *xanthine* and amino acid CGs, such as *alanine*, *leucine*, etc.

In some cases, Atoms can be assigned to different CGs if they are located in different topological structures within a molecule. For example, *cyclopropyl* and *cyclobutyl* are CGs that contained saturated carbon atom in a 3-member and 4-member ring, respectively. On the other hand, carbon atoms from other cycloalkanes such as cyclopentane and cyclohexane are generally assigned to the *cycloalkane* CG, instead of the usual *alkane* CG.

5.3 Primary DL_F Tokens

If needed, DL_F Atoms would need additional information called tokens, to locate or to differentiate one another within a CG. The available primary tokens are shown below:

- (1) Degree of substitution, *p*, *s*, *t*, *q* - refer to the degrees of substitution on either carbon (C) and nitrogen (N) atoms, depending on the number of connected hydrogen (H) atoms. The C and N atoms can be classified into primary (*p*), secondary (*s*), tertiary (*t*) and quaternary (*q*) atoms.

For N atoms, *Np*, *Ns* and *Nt* refer to two, one and zero H atoms attach to an N atom. The *Nq* applies to cases such as the *ammonium* CG, which contain four bonds with no H atom.

- (2) Positional tokens. *E* refers to a terminal position within a CG (the 'end' atom). *L* refers to a linked position within a CG (the 'link' atom).
- (3) Serial tokens. This is represented in numerical values (1, 2, 3, etc), to indicate the Atomic positions that are usually connected serially one with the other. These tokens are usually applied to aromatic structures and some complex cyclic compounds.

Atoms charge states. The positive (+) and negative (-) signs refer to the charge state of an Atom. They are usually used for Atoms belong to the *cation* and *anion* CGs as can be found in solutions.

5.4 Secondary DL_F Tokens

This is a group of optional supporting tokens for Atoms, to indicate the way or the type of atom to which it is connected to. The available tokens are shown below:

R - next to an aryl carbon atom.

U - next to an unsaturated carbon atom (such as alkene, alkyne type).

A - alpha, refers to the first Atom that connects respect to an Atom within a *CG*.

B - beta, refers to the second Atom that connects with respect to an Atom within a *CG*.

G - gamma, refers to the third Atom that connects with respect to an Atom within a *CG*.

D - delta, refers to the fourth Atom that connects with respect to an Atom within a *CG*.

E - epsilon, refers to the fifth Atom that connects with respect to an Atom within a *CG*.

Z - zeta, refers to the sixth Atom that connects with respect to an Atom within a *CG*.

Element symbols of the neighbouring Atom itself.

5.5 Notation Rules

The following lists the rules in using the DL_F Notation.

- (1) Each DL_F Atoms in a system model must associate with a *CG* and the corresponding *CGI* to which it belongs.
- (2) Every DL_F Atom can only associate with not more than one token from **each group** of the DL_F token. This means that there can be only one token, either from the Primary DL_F Token or the Secondary DL_F Token that associates with a DL_F Atom. However, two tokens are permissible only if each is originated from both Primary and Secondary Token groups.
- (3) Single-valence DL_F Atoms such as hydrogen and halogens must always express along with a Secondary DL_F Token. This is usually the element symbol of a neighbouring DL_F Atom.

5.6 Notation Format

D_ATA will carry out atom typing and determine the atom type of each atom in the system and expressed it in DL_F Notation. Each DL_F atom type is referenced to an atom key. An atom key can be regarded as the abbreviation of the atom type. D_ATA will generate atom keys internally to determine atomic interactions.

The DL_F atom types are expressed as follows:

$$A[t]_{CGname}[\#n]$$

Where *A* is the element symbol of an DL_F Atom, *[t]* is the optional DL_F token, which can be either primary or secondary types, or both. The *CGname* is the name of the Chemical Group to which the DL_F Atom *A* belongs.

The corresponding DL_F atom keys are expressed as follows:

$$ACGI[t][\#n]$$

Where *A* is the element symbol of a DL_F Atom, *CGI* is the Chemical Group index number.

The optional expression, *#n*, is called the *version number*. If there are more than one version of identical atom types, but may share different sets of potential parameters, then

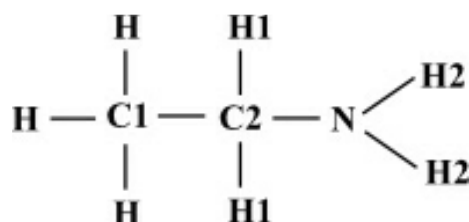
#*n* would be needed, where *n* is an arbitrary integer. Different values of *n* differentiate one version of atom type from the others. However, version labelling is only relevant in DL_FIELD for setting up force field models. In D_ATA, this quantity will be ignored.

For example, a primary alkyl carbon atom has the DL_F atom type of *Cp_alkane* with the corresponding atom key of *C1p*. In this example, *C* is the carbon atom, *p* is a primary token refers to primary carbon and *alkane* is the CG, which has the unique CGI of 1. The alkyl hydrogen atoms are expressed as *HC_alkane* with the atom key *H1C*. In this case, the *C* is a Secondary DL_F Token which is the neighbouring DL_F Atom to which the hydrogen atom is connected to.

5.7 Organic Molecules Examples

The following lists several example structures together with the assignment of atom types in the DL_F Notation.

(a) Ethylamine

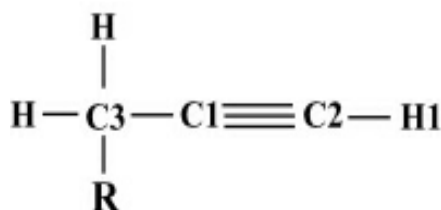


This is a primary amine and D_ATA will assign the atom types (atom keys in bracket) for each atom as follows:

C1 = *Cp_alkane* (*C1p*)
 C2 = *Cs_alkane* (*C1s*)
 H = *HC_alkane* (*H1C*)
 H1 = *HC_alkane* (*H1C*)
 N = *Np_amine* (*N45p*)
 H2 = *HN_amine* (*H45N*)

The letters *p* and *s* refer to the *primary* and *secondary* atoms, respectively, referring to the number of neighbouring hydrogen atoms, or more generally, the degree of substitutions. Similarly, a *tertiary* and *quaternary* species will be represented by the standard chemistry notation of *t* and *q*, respectively. The values 1 and 45 are the CGI for the *alkane* and *amine* CGs, respectively (See *DLF_notation* file in LIBRARY_PATH folder). Note also that, for a single-valence atom, such as the H atom, the atom to which it is bonded to must always be indicated. For example, the alkyl hydrogen (H1) is shown as *HC_alkane*, instead of 'H_alkane'. It is also equivalent to a more specific annotation, *HCS_alkane*, or the hydrogen atom connected to a *secondary* alkyl carbon.

(b) Propyne

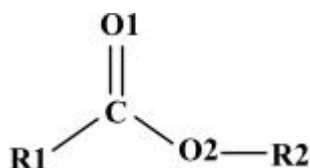


This is an alkyne and D_ATA will assign the ATOM_TYPES (ATOM_KEYS in bracket) for each atom as follows:

C1 = CL_alkyne (C3L)
 C2 = CE_alkyne (C3E) - terminal alkyne carbon
 H1 = HC_alkyne (H3C)
 C3 = Cs_alkane (C1s)
 H = HC_alkane (H1C)

Note that the positions of C1 and C2 are distinguished from each other by the position tokens *E* and *L*.

(c) *An ester molecule*

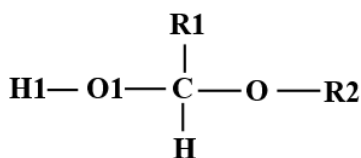


Atom type assignment:

C = C_ester (C19)
 O1 = OE_ester (O19E)
 O2 = OL_ester (O19L)

The CG contains two O atoms, O1 and O2. Both atoms are distinguished from each other by the positional token labels *E* and *L*, which refer to the 'end' position and 'linked' position, respectively. Therefore, O1 is the 'terminal' carbonyl oxygen with the atom type *OE_ester*. Whereas the O2 that forms links with C and R2 is assigned with the atom type *OL_ester*. For the carbon atom (C), no supporting token is needed because it is unique within the *ester* group. Once again, the value 19 is the CGI for the *ester* CG.

(d) *A hemiacetal molecule.*



A hemiacetal group is a functional group that contained both an alcohol group and an ether group linked to a common carbon atom. More specifically, hemiacetal has a general formula of R1-HC(OH)-O-R2, where R1 and R2 are some substituents.

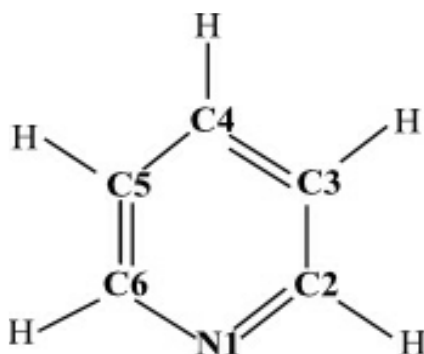
Atom type assignment:

C = C_hemiacetal (C25)

H = *HC_hemiacetal* (H25C)
 O = *O_hemiacetal* (O25)
 O1 = *OH_hemiacetal* (O25H)
 H1 = *HO_hemiacetal* (H25O)

For the oxygen atom with the label 'O1', the hydrogen neighbour atom is shown, to distinguish it from the other ether oxygen with the label 'O'.

(e) *Pyridine*



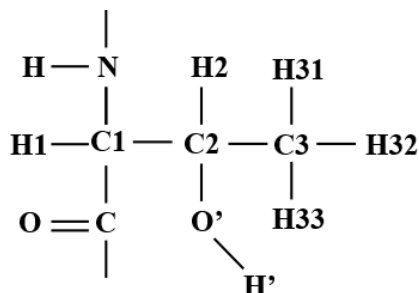
The aromatic carbon atoms are distinguished from one another by a numerical token assigned serially around the ring. In the case of hydrogen atoms, they are assigned with the generic atom type *HC_aromatic*. Take example for a pyridine molecule, D_ATA will assign the atoms as follows:

N1 = *N1_pyridine* (N501-1)
 C2 = *C2_pyridine* (C501-2)
 C2 = *C3_pyridine* (C501-3)
 C2 = *C4_pyridine* (C501-4)
 C2 = *C5_pyridine* (C501-5)
 C2 = *C6_pyridine* (C501-6)
 H = *HC_aromatic* (H11C)

Where *CGI* of 501 and 11 refer to pyridine and the generic aromatic *CGs* respectively. Note that the numerical tokens are expressed as -1, -2, etc, to distinguish them from the *CGI*.

(f) *Threonine, an amino acid.*

D_ATA can distinguish amino acids from other types of compounds with similar structures. For example, the threonine (THR) is a hydrophilic amino acid that contains the *alcohol CG*. The amino acid residue can form part of a protein structure via the usual amide links:



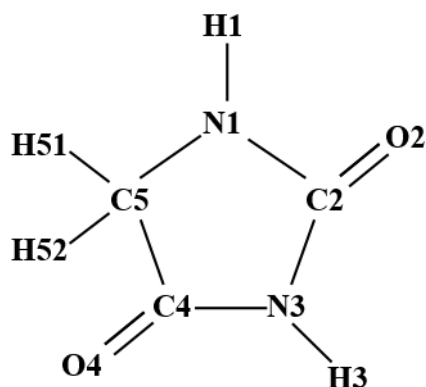
D_ATA can detect the structure as an amino acid and assign the atoms as follows:

N = *Ns_amide* (N44s)
 H = *HN_amide* (H44N)
 C1 = *CA_amino_acid* (C900A)
 H1 = *HCA_amino_acid* (H900CA)
 C = *C_amide* (C44)
 O = *O_amide* (O44)
 C2 = *CB_threonine* (C919B)
 H2 = *HCB_threonine* (H919CB)
 O' = *O_threonine* (O919)
 H' = *HO_threonine* (H919O)
 C3 = *CG_threonine* (C919G)
 H31, H32, H33 = *HCG_threonine* (H919CG)

D_ATA assigns atoms from the side group that belong to the *threonine* CG, instead of the usual *alkane* and *alcohol* CGs. Note the position tokens *B* and *G* are used to indicate the *beta* carbon and the *gamma* carbon relative to the *alpha* carbon that is connected to the *amide* Atoms. Instead of labelling C1 as *Ct_alkane* and H1 as *HC_alkane*, these atoms are assigned to the special *amino_acid* CG due to their significance in biology and to indicate the whole structure is in fact an amino acid residue that formed part of the amide linkage in a protein.

(g) Hydantoin

Some complex heterocyclic compounds can be of biological or pharmaceutical importance, and hence, a special name is given to these structures. They can be made up of more than one CG. For example, hydantoin is an organic compound with the ring structure as shown below.



Hydantoin can also refer to a class of compounds with the same ring structure, of which the hydrogen atoms can be substituted. D_ATA can recognise such a moiety and assign the atoms, with a standard numbering sequence as follows:

N1 = N1s_hydantoin (N600-1s)
 H1 = HN1_hydantoin (H600N1)
 C2 = C2_hydantoin (C600-2)
 O2 = OC2_hydantoin (O600C2)
 N3 = N3s_hydantoin (N600-3s)
 H3 = HN3_hydantoin (H600N3)
 C4 = C4_hydantoin (C600-4)
 O4 = OC4_hydantoin (O600C4)
 C5 = C5s_hydantoin (C600-5s)
 H51, H52 = HC5_hydantoin (H600C5)

5.8 Further information

For more information about the DL_F Notation and how the conversion works please refer to the following references:

Describe conversion and annotation procedures of DL_F Notation
 C. W. Yong, *J. Chem. Inf. Model.* **56**, 1405-1409 (2016)

Information about standard and universality behaviour of DL_F Notation
 C. W. Yong, SSRN (2022), <http://dx.doi.org/10.2139/ssrn.4254942>

Atom typing in molecular simulations and use of DL_F Notation in DL_FIELD: Available in online resources, in DL_Software Digital Guide (DL_SDG),

https://dl-sdg.github.io/RESOURCES/TUTORIALS/dlf_8.html

6. Atomic Interactions (DANAI)

Once atomic structures are annotated with DL_F Notation, D_ATA can carry out nonbonded interaction analysis by identifying types of interactions and the extent (or modes) of such interactions between atoms from up to two different CGs. Once similar type and modes of interactions are identified, it will be counted, and the interaction mode will be annotated by using DANAI notation.

The DANAI (DL_ANALSER Notation for Atomic Interactions) is a Chemistry-based language construct for interatomic interactions. It is expressed based on the DL_F Notation to indicate the non-bonded topological interaction type and structure between two CGs, bypassing the use of diagrammatic or pictorial illustrations.

DANAI syntax is standardised and applicable to a wide variety of atomistic-based models and configurations derived from both simulations and experimental observations. From such, data analysis can be carried out to quantify and rationalise interactions between atoms and molecules in the system.

In DANAI notation, the full description of any given interactions must always be expressed in terms of the *macro-interactions* and the corresponding *micro-interactions*: The former describes an interaction in a general sense, while the latter describes specific interactions that can occur within the context of the macro-interaction. A micro-interaction describes actual Atomic species involve, including the number of CGs and the way they interact.

6.1 Macro-interactions

The format to describe a general interaction between two CGs is shown as follows:

$$I_CGI1_CGI2$$

Where *I* is the interaction type and *CGI* is the *Chemical Group Index* which is the unique numerical value for a given *Chemical Group* (CG) in the DL_F Notation.

Below lists the complete list of interaction types:

DD – dipole-dipole interactions.
HB – hydrogen bonding. This is a special case of the *DD* interaction.
ID – induced dipole interactions.
HP – hydrophobic interactions, a special case of *ID* interactions for alkyl groups.
EI – electrostatic interactions, such as between cations and anions.
CD – Charge-dipole interactions, such as between ions and polar atoms.
PS – parallel π - π stacking interactions between benzene rings.
PD – parallel displaced π - π stacking between benzene rings.
PH – between C-H... π interactions.
PT – T-shaped π - π stacking interactions (special case of PH).
PO – polar atom- π interaction
PM – metal- π interaction
PI – Ion- π interactions (as between cations, anions and an aromatic π -delocalisation system).

For example, the macro-interaction HP_1_1 means hydrophobic interactions between two alkyl groups, where the CGI value 1 refers to the *alkane* CG. On the other hand, HB_15_20 means hydrogen bond interactions between *alcohol* (15) and *carboxylic* (20) CGs.

Macro-interactions are defined in *DLF_map* file located in the *LIBRARY_PATH*. This file can be edited to include interaction types and Atoms involve for new CGs. By referencing to this file, D_ATA will look for macro-interactions for the system configuration. From such, the corresponding micro-interactions will be identified and quantify for subsequent data analysis.

6.2 Micro-interactions

This is defined as a set of interaction configuration mode between two CGs as defined in the macro-interactions. Micro-interactions describe a variety of way how atoms from up to two different CGs interaction with each other. The general format is as follows:

[Sa]interaction_notation

Where *S* is the topological structure of interactions, *a* is the number of **distinct** CGs involve in the micro-interaction that form such interaction structure. Examples of *S* are shown below:

J - a junction or network interaction

R - a ring structure

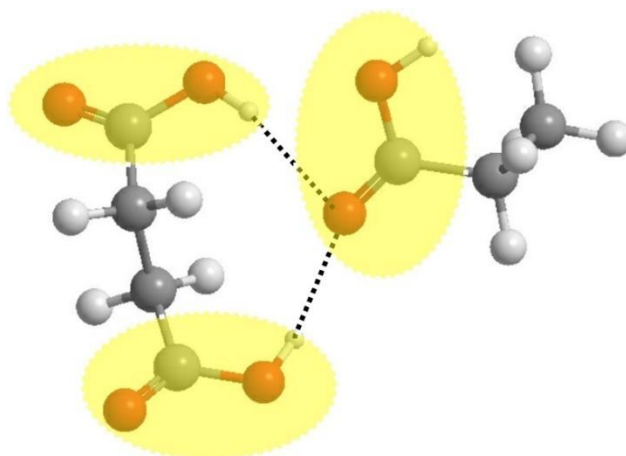
L - a linear structure

C - complex structure contains some of the above-mentioned structures.

For example, *[L3]* means a micro-interaction involves atoms from three distinct CGs in a linear fashion. *[R2]* means a micro-interaction involves atoms from two distinct CGs, forming a ring structure.

The *interaction_notation* consists of a line of text that annotates the Atomic species involved in the interaction. These atomic species are expressed in DL_F Notation atom keys.

A distinct CG refers to a group of member Atoms that span across a molecular connectivity until other Atoms belong to a different CG is encountered. Diagram below shows three distinct *carboxylic* CGs (shaded in yellow) from two molecules. The molecule on the left contains two distinct *carboxylic* CGs separated by a group of Atoms belong to *alkane* CG.



Interactions shown in dotted lines are expressed in DANAI as *[L3]H2O0:O20E:H2O0*.

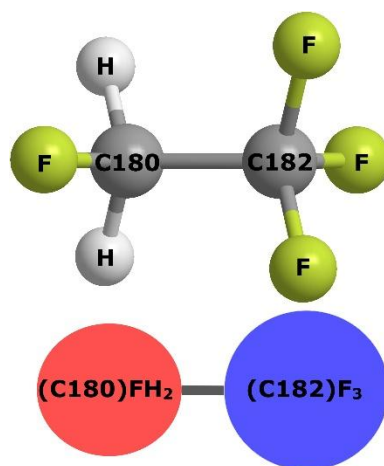
6.3 Notation rules

This section describes notation rules by making an example reference to HFA-134a (1,1,1,2-tetrafluoroethane) molecule, interacting with one another in a condensed phase system.

Diagram below shows a ball-and-stick representation of the HFA-134a molecules, showing the elemental symbols fluorine and hydrogen atoms. For the carbon atoms, they are labelled with the DL_F Notation, where the unique numerical values of 180 and 182 indicate they are of the *monohaloalkane* and *trihaloalkane* CGs, respectively.

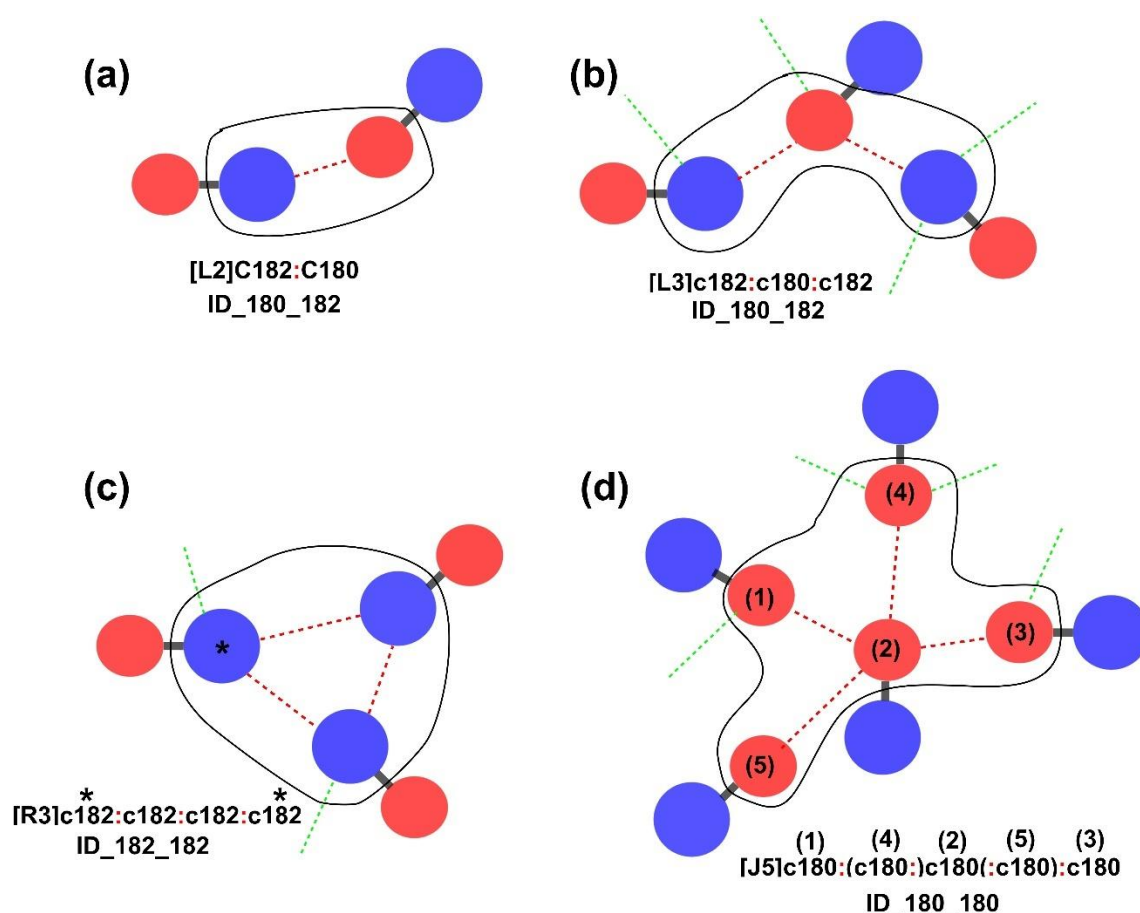
The HFA-134a molecule can be categorized into two different groups centred at the respective carbon atoms: the monofluoroalkyl group represented by a red sphere, and trifluoroalkyl group represented by a blue sphere.

The molecules are predominantly interacting with one another via the intermolecular induced-dipole (ID) interactions and the significance of such interactions are identified based on the distance criteria between any two non-bonded spheres centred round the carbon atoms.



The distance criteria are so chosen (5 Å) to make sure only significant interactions are accounted between two species that are at the direct 'line of sight' without any atom sandwiched in between the two carbon centers.

D_ATA will identify these interactions and construct a global interaction map of HFA134a. The molecules interact with one another in a variety of ways and let's consider some of these interactions (a), (b), (c) and (d), as shown below:



The diagrams illustrate four different molecular interactions together with the corresponding DANA I statements. Interpretations of these DANA I statements are summarized as follows:

(i) The colon (:) represents the non-bonded interaction between two chemical species. This is shown as the red dotted lines in the diagram, which also correspond to the red colons shown in the *interaction_notations*.

(ii) Information contains within the square brackets indicates the topological structure and the number of participating chemical species that formed such a structure in the interaction. For example, [L3] means three chemical species interacting serially, forming a linear structure **(b)**, and [J5] means branched structure involving five species **(d)**.

(iii) For interactions forming an enclosure (ring), the first and the last chemical species refer to the same species, which indicates the extent of the ring enclosure. For instance, in **(c)**, the chemical species marked with * corresponds to the marked DANA I *interaction_notation*.

(iv) A chemical species enclosed within a bracket means it is a branched species. For instance, consider that species (1), (2) and (3) are interacting with one another, forming a linear interacting chain, as shown in Figure **(d)** above. Chemical species that are labelled (4) and (5) are regarded as the two branched species that interact with the member species along the chain. Subsequently, they are enclosed within brackets in the DANA I statement. Of note are the locations of species (4) and (5) and their corresponding : symbols that are positioned in the statement. These symbols indicate which member species located along the chain they are interacting with. The DANA I statement indicates

both species (4) and (5) are interacting with species (2). For example, species (4) is placed between species (1) and species (2) in the DANAI statement. The `:` symbol within the bracket is placed to the right, to indicate the interaction is with species (2) and not species (1). Similarly, species (5) is placed between species (2) and species (3). The `:` symbol within the bracket is placed to the left, to indicate species (5) interacts with species (2) and not species (3).

(v) If a chemical species is specified in the DANAI statement in the uppercase letter, it means the only interaction involve with this species is what is indicated in the *interaction_notation* statement, which is an ID interaction in this case. For example, the DANAI statement in Figure **(a)** is expressed as '`C182:C180`'. Since both are expressed in capital letter 'C', this means the only detected ID interaction at the C182 species is with a C180 species and *vice-versa*. No additional *similar* (ID) interaction is detected with other chemical species (as with other C180 or C182 atoms). In other words, it is an *isolated interacting* pair of chemical species, in terms of ID interactions. These atoms are called *exclusive atoms*.

(vi) If a chemical species is specified in the DANAI statement in the lowercase letter, it means the chemical species may interact with more than one other chemical species via ID interactions, *including* those species that are not shown in the *interaction_notation* statement. These additional ID interactions are shown as green dotted lines in Figure 2. Consider Figure **(c)** for ID_182_182, where three C182 species interact with one another, forming a ring structure. These species are expressed as '`c182`', implying that there may be other ID interactions detected with other C182 species, as indicated by the green dotted line, apart from the said species that involved in the ring formation. In the case of Figure **(b)** for ID_180_182, all chemical species are express in small letter c, implying the said species are interacting with other species, *either* C180 or C182, via ID interactions. These atoms are called the *non-exclusive* atoms.

Please note: If Option **3** in the CONTROL_FILE is to switch 'all-inclusive' interactions to 'on' (1), then D_ATA will not distinguish among atoms either they participate in exclusive or non-exclusive interactions. These interactions will be counted and assign to a single DANAI statement that contains the relevant interactions. All participating atoms will be expressed in capital letters in DANAI statements

6.4 Further information

For more information about DANAI and how to interpret the expression please refer to the following references:

Describe DANAI syntax and rules:

C.W. Yong and I.T. Todorov *Molecules* **23**, 36 (2018)

Describe DANAI syntax and rules, with a haloalkane as an example.

C. W. Yong *et. al. Data in Brief*, **50**, 109485 (2023).

Examples of DANAI expressions for HB interactions between two carboxylic groups, available in DL_Software Digital Guide, DL_SDG.

https://dl-sdg.github.io/RESOURCES/TUTORIALS/dla_9.html

Note: DANAI is first implemented in DL_ANALYSER (hence the name) for several specific sets of CGs. However, it is recommended to use D_ATA for atomic interaction analysis because D_ATA covers a broader interaction combination for all CGs.

7. Interaction counts

Once Atoms are assigned to DL_F Notation, D_ATA will carry out interaction analysis and then annotate each set of interactions with a matching DANAI statement.

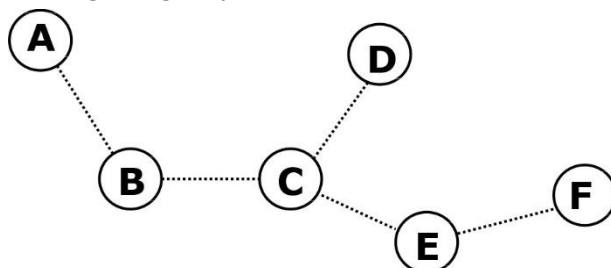
Note that current D_ATA version can only detect a few topological interactions as shown below. More will be added in subsequent versions.

7.1 [L2] Interactions

This is the simplest interactions involving two atoms from two **distinct** CGs. For exclusive, isolated binary interactions, Atom A can only have one interaction (shown as ':') with Atom B and vice-versa.

[L2]A:B

For interactions involving a group of Atoms, D_ATA looks for all possible unique



combination pairs by going through all Atoms serially, and that the first Atom index (A) must be smaller than the second one (B). Consider an interaction network as shown below:

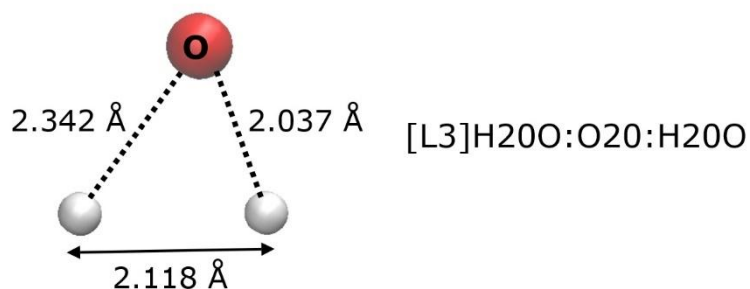
Assume there is no other similar interaction involves with other Atoms other than what is shown above, D_ATA will identify the following [L2] interactions:

[L2]A:b
[L2]b:c
[L2]c:D
[L2]c:e
[L2]e:F

7.2 [L3] Interactions

This type of interaction involves three successive Atoms from three **distinct** CGs. However, the first and the third Atoms do not interact significantly with each other. For ID and HP interactions, these Atoms must be at a distance that is larger than the threshold specified. Otherwise, the three interacting atoms would be considered under the [R3] classification: a ring structure.

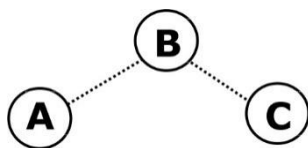
For HB interactions involving, say two H atoms with an O atom, the H...H distance can be less than the threshold distance specified. This is because D_ATA does not consider interactions between the two H atoms fall under the HB classification. Diagram below shows an example of such structure, extracted from a simulated system contained *carboxylic* CGs.



Here, two hydrogen atoms formed HB interactions with an O atom. The HB interaction threshold is set to 2.5 Å and the interactions with the oxygen atom are therefore considered significant. D_ATA will consider the topological structure of such interaction to be [L3] type, even though the distance between the two H atoms is less than the threshold distance. In fact, it is even less than one of the interacting distances between O and H.

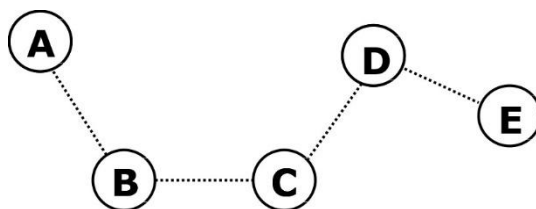
The interacting structure can be characterised by calculating the [linearity parameter](#), l , which indicates it is in fact a highly puckered structure ($l = 2.068$).

For an exclusive, isolated [L3] interaction, Atom A and Atom C each must interact with the common Atom B and there is no additional interaction among the Atoms or any other Atoms that are outside the scope (see below).



The corresponding DANA statement is [L3]A:B:C or the equivalent [L3]B:A:C where A and C positions can be inter-changed. This means Atom B must have two neighbours and Atoms A and C can only have one each (neighbouring to Atom B).

For an interaction network chain involving more than three atoms, D_ATA will scan through all Atoms with neighbour numbers that are two or more and determine its neighbouring Atoms, whether they are of *exclusive* or *non-exclusive* types. Consider a chain of five interacting Atoms as shown below:

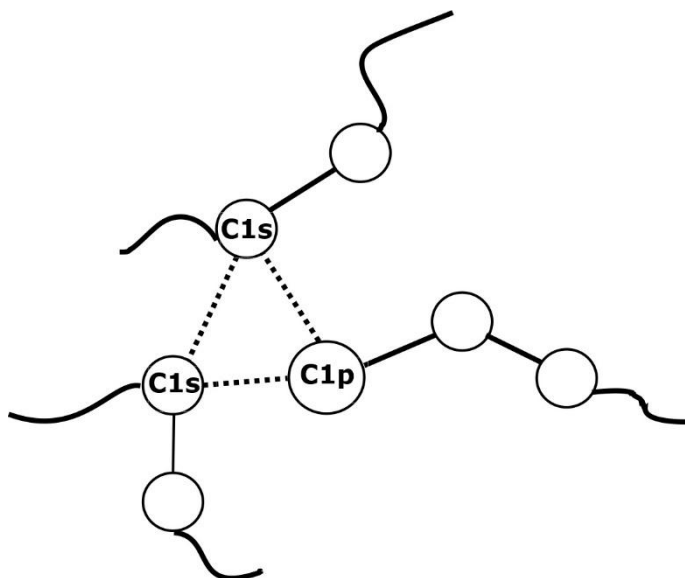


Assume there is no other similar interaction involves with other Atoms other than what is shown above, D_ATA will identify the following [L3] interactions:

[L3]A:B:c
[L3]b:C:d
[L3]c:D:E

7.3 [R3] Interactions

This type of topological interaction involves three Atoms from three **distinct** CGs, forming a triangular ring enclosure. Diagram below shows an example sketch of three alkane molecules (only the relevant atoms are labelled), with a primary alkyl group (centered at C1p carbon) interact with two secondary alkyl groups (centered at C1s carbon) from two distinct alkane CGs.



The corresponding DANAI statement is either [R3]C1p:C1s:C1s:C1p, or the equivalent [R3]C1s:C1p:C1s:C1s. Note the first and the last Atoms are referring to the same Atom.

7.4 [L4] Interactions

This non-bonded interaction structure involving four Atoms from four **distinct** CGs in succession, forming a chain of interactions. Consider an isolated [L4] interactions in succession as shown below:

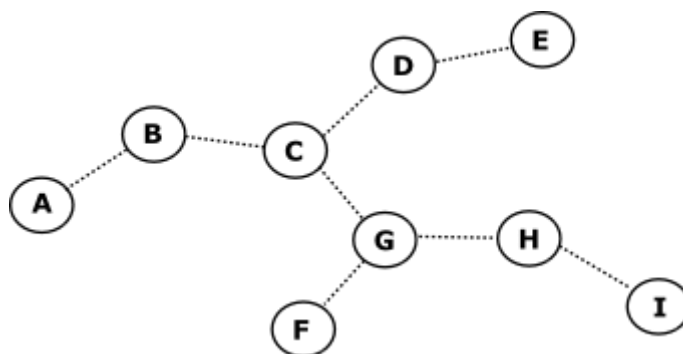


The corresponding DANAI expression is also shown. For all-exclusive interactions, Atoms B and C must have exactly two interacting neighbours: Atom B has A and C as neighbours, while Atom C has B and D as neighbours. Furthermore, Atom A can only have one neighbour (B) and Atom D also can only have one neighbour (C).

To qualify interactions involving four species adopting the [L4] structure, D_ATA will check to ascertain they do not form ring structure. That is, there is no interaction linkage between A...C, B...D and A...D.

For an interaction network chain involving more than four Atoms, D_ATA will look for all interacting pairs $i-j$ with $i < j$ and with respective neighbour numbers that are two or more. After that, their respective neighbouring Atoms, h and k , will be identified. If all-inclusive interaction mode is switched off (Option **2**), then the resulting chain of atoms $h-i-j-k$ will

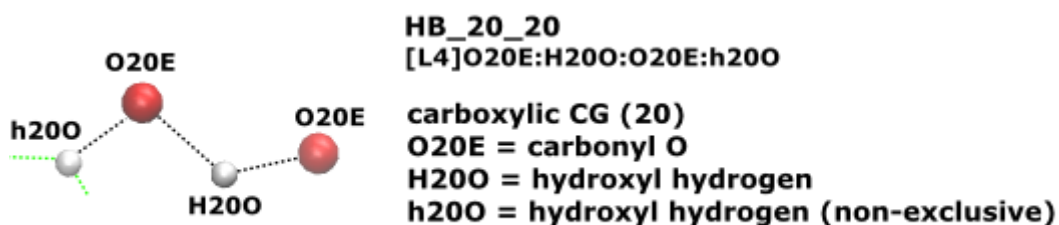
go through additional checks to decide whether they are of *exclusive* or *non-exclusive* types. Consider a network chain of interacting Atoms as shown below:



Assume there is no other similar interaction involves with other Atoms other than what is shown above, D_ATA will identify the following [L4] interactions:

[L4]A:B:c:d
 [L4]b:c:D:E
 [L4]b:c:g:h
 [L4]d:c:g:h
 [L4]d:c:g:F
 [L4]b:c:g:F
 [L4]F:g:H:I
 [L4]c:g:H:I

Diagram below show an actual example of an [L4] interactions involving four Atoms from four different carboxylic CG with the CGI value of 20. This structure is extracted from an actual molecular simulation of ethanoic acid liquid model.



Only the Atoms that participate in the interactions are shown. Note that there are three exclusive atoms, participating in the HB formation as per the DANAI expression. However, the fourth Atom, h200, is a non-exclusive Atom that also participate in other HB with other Atomic species that are not shown in the DANAI expression (show as green dotted lines).

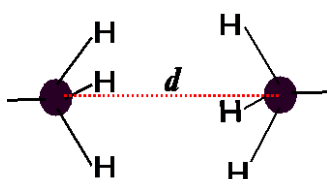
8. Non-bonded Interactions

The current version (1.0) of D_ATA can only detect two types of non-bonded interactions: the hydrophobic and hydrogen bond interactions. Future versions will include different types of interactions.

8.1 Hydrophobic Interactions

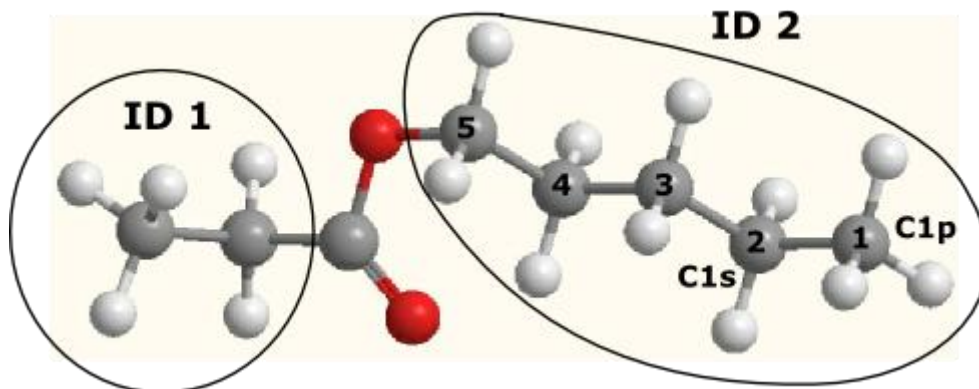
The hydrophobic (HP) interaction is a type of induced-dipole (ID) interaction that is specifically reserved for alkyl carbon (CG *alkane*) including those of amino acid residues.

A non-bonded pair of HP interaction is counted based on the critical distance of d between two alkyl groups centred at the alkyl carbon atoms (see below). By default, this distance is set to 4.5 Å and can be changed from the CONTROL_FILE (Option **16**).



Unlike most other CGs, the *alkane* CGs can vary in sizes depending on the molecular system. For instance, a system consists of alkane chains of varying lengths.

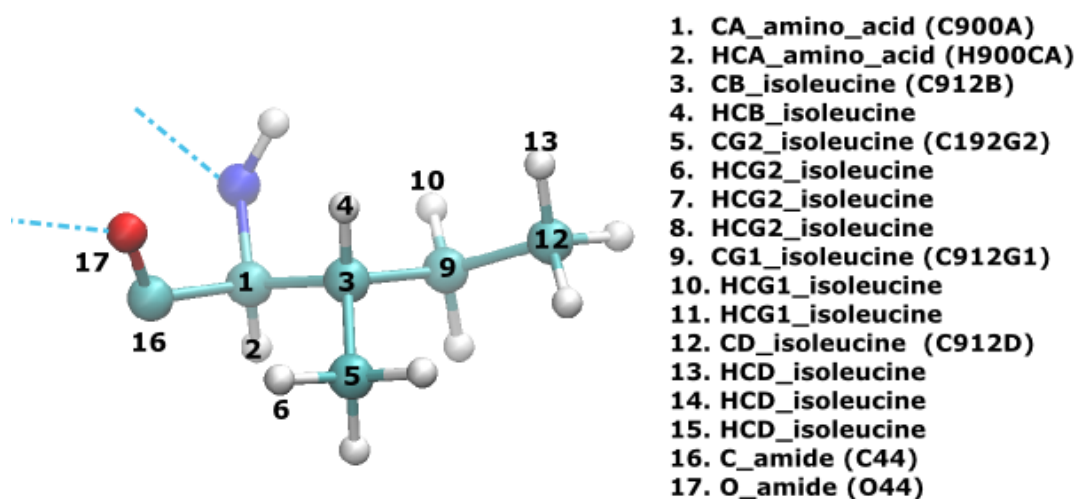
In D_ATA, each Atom group within a CG is collectively distinguished from one another by the unique label assigned to each group of CG Atoms, even for groups of Atoms belonging to the same CG. For example, consider the molecule shown below.



This is an ester molecule containing two groups of *alkane* CGs, labelled as ID1 and ID2, and are separated by the *ester* CG. Both groups are considered as separate groups of the same CG type (*alkane*). If the ester group is substituted with, say, a methylene CH₂ group, then the whole molecule would consist of purely *alkane* Atoms and assign to a single ID label.

Now, consider groups of *alkane* Atoms arranged in a series for ID2. The 1-2 Atoms are bonded and will be ignored. In addition, D_ATA will also ignore any 1-3 and 1-4 non-

bonded interactions. However, D_ATA will look for any HP interactions for 1-5 Atoms or beyond if the analysis mode, [Option 10](#), is set to 0 (consider all) or 2 (within molecule).



For hydrophobic parts of amino acids such as the alkyl groups, D_ATA distinguishes these from the normal 'alkyl' groups by assigning them to a specific amino acid CG. For example, consider an isoleucine amino acid residue fragment (extracted from a protein structure) as shown below. The D_ATA Typer assigns the molecular fragment to the *isoleucine* CG. The characters A, B, D and G refers to the *alpha*, *beta*, *gamma* and *delta* carbon positions according to the standard amino acid assignments.

Unlike the normal alkyl carbon, these carbon atoms in isoleucine are distinguishable from one another and the amount of identified DANAI expressions for interactions involving these atoms would be intractable. For this reason, except for the alpha C, D_ATA will assign these alkyl carbon atoms with the single type *C_isoleucine* with the corresponding atom keys as C912. The reassigned list of atom types is shown as follows:

1. CA_amino_acid (C900A)
2. HCA_amino_acid (H900CA)
3. C_isoleucine (C912)
4. HCB_isoleucine
5. C_isoleucine (C192)
6. HCG2_isoleucine
7. HCG2_isoleucine
8. HCG2_isoleucine
9. C_isoleucine (C912)
10. HCG1_isoleucine
11. HCG1_isoleucine
12. C_isoleucine (C912)
13. HCD_isoleucine
14. HCD_isoleucine
15. HCD_isoleucine
16. C_amide (C44)
17. O_amide (O44)

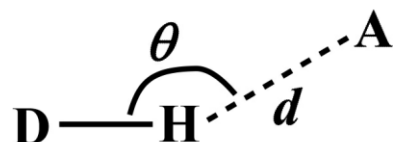
Note that the original assignments for the alkyl hydrogen atoms are retained, indicating the positions of the C atoms to which they are connected to.

The same concept also applies to other hydrophobic amino acid residues such as alanine, leucine, valine and proline. For polar amino acids such as asparagine, aspartic, etc., only

the non-alkyl carbon will be indicated with the positions. For example, for asparagine, - not sure to put this.

8.2 Hydrogen bond Interactions

The hydrogen bond (HB) interaction is a special case of dipole-dipole (DD) interaction and is so named because the interaction involves a hydrogen atom connected to a large electronegative atom such as oxygen (a donor atom, D). This makes the hydrogen atom



particularly attractive to an acceptor (A) species, or another atom with a large electronegative and contains lone pair of electrons.

D_ATA will only identify an HB interaction if it satisfies both criteria: the distance between H and A is less than d and the angle is θ or more. By defaults, these thresholds are set to 2.5 Å and 120° and can be changed from the CONTROL_FILE (Options **13** and **14**).

9. Analysis and Calculations

Once atomic interactions are accounted, D_ATA will determine averages and the associated deviations for every interaction statement. However, only those 'significant counts' will be shown in the results file. The amount of information shown in the results file is determined based on the average count threshold define in Option **13** in the CONTROL_FILE.

D_ATA can also out further analysis as follows:

9.1 Pearson correlation coefficient, R

The inter-relationships among the identified interaction modes can be determined from the correlation calculations as follows:

$$R_{x-y} = \frac{\langle \Delta C_x \cdot \Delta C_y \rangle}{\sqrt{\langle C_x^2 \rangle \langle C_y^2 \rangle}} \quad \Delta C_i = C_i - \mu_i$$

Where x and y are two different types, C is the associated count number and μ is the average count of that interaction. The correlation coefficient R is essentially a normalised covariance of xy that measures the linear correlation between the interaction modes. The R value will always fall within the range:

$$-1.0 \leq R \leq +1.0$$

With the following interpretation:

- (1) A positive value gives a positive correlation, that is, the formation of one interaction also corresponds to the formation of the second interaction and *vice versa*.
- (2) A negative value gives a negative correlation, that is, the formation of one interaction is at the expense of the reduction of the other interaction and *vice versa*.
- (3) The zero correlation means both interactions are completely uncorrelated to each other: formation of one interaction is not directly dependent upon the other interaction.
- (4) The correlation value of ± 1.0 means they are completely dependent on each other.

Therefore, the magnitude of R indicates the strength of interaction correlation, ether they are positively correlated or negatively correlated. Note that self-correlation R_{xx} always gives a value of +1.0.

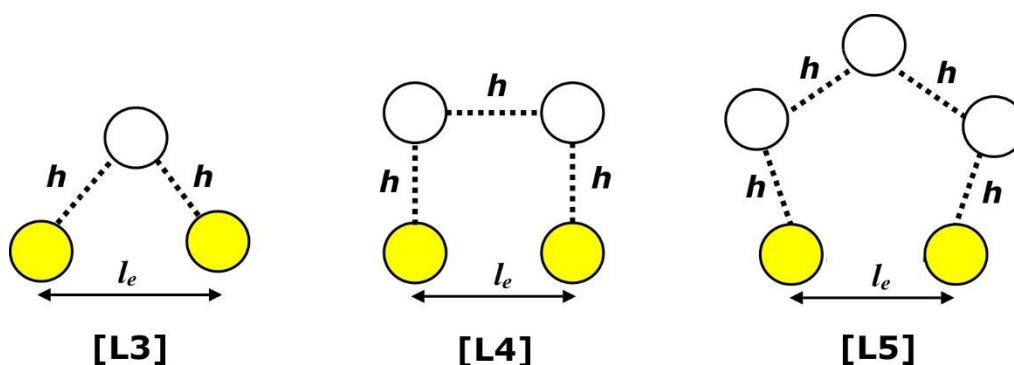
9.2 Linearity parameter, l

This quantity measures the contour shape of a group of atoms interact serially forming an $[Lx]$ topological structure where $x > 3$ and is the number of distinct CGs, with one Atom from each CG. Linearity parameter is simply defined as

$$l = \frac{d}{l_e}$$

Where d is the segmental contour length, or the sum of the distance between two successive Atoms along the interaction structure and l_e is the end-to-end distance or the distance between the first and the last interacting Atom in the chain segment that made up the $[Lx]$ structure. For a perfectly linear structure, $d = l_e$ and l is always 1.0. Any value that is larger than 1.0 is a measure of the departure from linearity.

Consider interactions involving same type of Atoms, such as the induced dipole (ID), or the hydrophobic interactions (HP) that are specific to alkyl carbons. Below shows three extreme-folded interacting structures for $[L3]$, $[L4]$ and $[L5]$ with the Atoms at the interacting distance threshold h . The yellow spheres indicate the first and the last Atoms and their end-to-end l_e values are just above h .



The value of l is $O(x-1)$ as $l_e \rightarrow h$. The larger the value of x , the larger l becomes, or the more circular (less linear) the overall structure becomes. Therefore, the value of l adopts the following range:

$$1 \leq l < (x - 1)$$

This means $[Lx]$ adopts a more 'clustering' or folded conformation as $l \rightarrow (x-1)$.

In cases of interactions involving different Atomic species, such as those of dipole-dipole (DD), or hydrogen bond (HB), l_e can be smaller than h and therefore $l > \sim(x - 1)$ because D_ATA carries out structural calculations on Atoms based on the type of interaction involved (See [Chapter 7.2](#) $[L3]$ interactions).

In fact, the quantity l can also be considered as a measure of *tortuosity*, or the 'roughness' of a linear topological structure.

9.3 Triangular parameter, t_r

The [\[R3\] interaction](#) involves three interacting Atoms, forming a triangular shape, which can be quantified using the triangular parameter t_r :

$$t_r = \left[1 - \frac{3 \sum_i^3 (\theta_i - 60^\circ)^2}{2 (\sum_i^3 \theta_i)^2} \right] \left[1 - \frac{2 \sum_{i>j}^3 (h_i - h_j)^2}{(\sum_i^3 h_i)^2} \right]$$

and

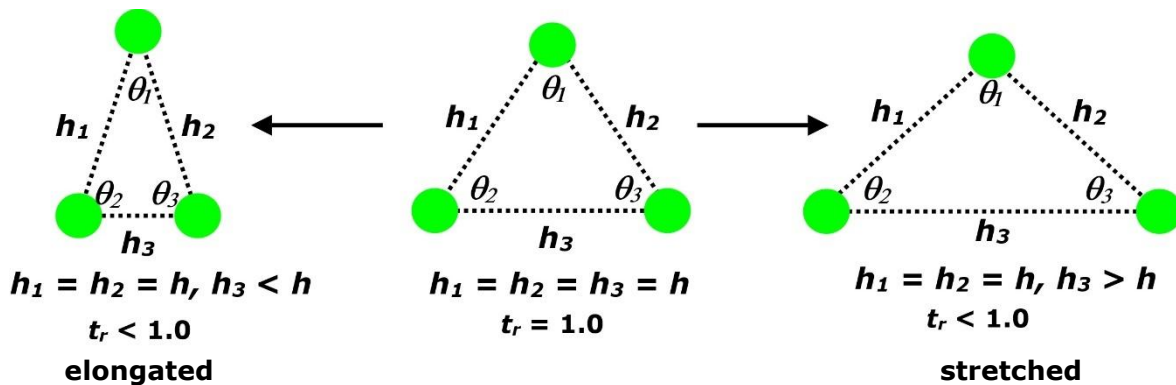
$$1 \leq t_r < 0$$

where θ_1 , θ_2 and θ_3 are the three inner angles of a triangle and h_1 , h_2 and h_3 are the corresponding sides of the triangle. The equation above shows t_r depends on two factors: the extent of the angles' deviations from 60° (angular component) and the extent of the length differences among the three sides of the triangle (length component).

If an equilateral triangle is defined to be the 'perfect' triangle, then both components = 1 and $t_r = 1$ with $\theta = 60^\circ$ for all angles and $h_1 = h_2 = h_3$.

Consider an equilateral triangle as shown below. If the triangle is deformed with h_3 decreases and h_1 and h_2 remain the same, then θ_1 will decrease ($< 60^\circ$) while θ_2 and θ_3 increase, forming a 'squashed' (or elongated) isosceles triangle (diagram on the left). For a highly elongated triangle, as $h_3 \rightarrow 0$, $\theta_1 \rightarrow 0^\circ$ and $\theta_2, \theta_3 \rightarrow 90^\circ$, tending towards a folded linear-like structure with $t_r \rightarrow 0$.

Conversely, if the triangle is deformed with h_3 increases and h_1 and h_2 remain the same, then θ_1 will increase ($> 60^\circ$) while θ_2 and θ_3 decrease, forming a stretched isosceles triangle (diagram on the right). For a highly stretched triangle, as $h_3 \rightarrow 2h$, $\theta_1 \rightarrow 180^\circ$ and $\theta_2, \theta_3 \rightarrow 0^\circ$, tending towards a stretched linear-like structure with $t_r \rightarrow 0$ again.

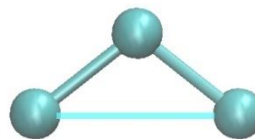


The value of t_r is therefore a measure of triangularity between both extremes. Note that an elongated triangle tends to have a higher t_r value (more 'triangular') than a stretched counterpart. This is because the angular deviations from 60° for an elongated triangle is smaller when compared to that of a stretched triangle. This, however, is partially compensated by the length component that is contributing to a larger t_r value for a stretched triangle.

Diagram below shows two isosceles triangles. Note how the angular and length components balanced out each other, and show the stretched triangle is in fact slightly more 'triangular' than the elongated one.



$t_r = \mathbf{0.693}$ (angular = 0.909, length = 0.763)
 $\theta = 23.797^\circ, 78.101^\circ, 78.101^\circ$
 $h_1 = h_2 = 1.500 \ h_3 = 0.619$

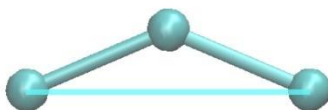


$t_r = \mathbf{0.781}$ (angular = 0.870, length = 0.898)
 $\theta = 103.33^\circ, 38.336^\circ, 38.336^\circ$
 $h_1 = h_2 = 1.500 \ h_3 = 2.353$

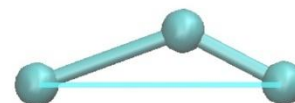
Diagram below shows a few more different triangles, each with $t_r \sim 0.5$, to provide a general sense about a middle threshold above which the structure becomes 'more triangular' and below which the structure becomes 'less triangular' and moving towards a linear structure.



$t_r = \mathbf{0.500}$
 $\theta = 15.308^\circ, 82.346^\circ, 82.346^\circ$
 $h_1 = h_2 = 1.500 \ h_3 = 0.400$



$t_r = \mathbf{0.501}$
 $\theta = 134.011^\circ, 22.994^\circ, 22.994^\circ$
 $h_1 = h_2 = 1.500 \ h_3 = 2.762$



$t_r = \mathbf{0.501}$
 $\theta = 132.46^\circ, 27.747^\circ, 19.791^\circ$
 $h_1 = 1.086 \ h_2 = 1.494 \ h_3 = 2.367$

Diagram on the left and middle show an elongated and stretched isosceles triangles. Diagram on the right is a scalene triangle (no equal sides). However, in most cases, interacting Atoms should adopt more triangular configurations ($t_r \gg 0.5$) due to the distance thresholds criteria and steric repulsions between Atoms.

10. Example Structures

D_ATA software package includes several example structures for user to try out how the software operates. These structures are in the *examples/* folder.

By default, D_ATA is set to read the first example (*example1.xyz*) once the program is compiled and run (see [Section 1.2](#) how to do this).

To run another example structure, just edit *d_ata.input* file and change the filename accordingly. By default, the all-inclusive analysis mode is turned on (Option **3**). Try to switch this off and run D_ATA again and see the results output differences between the modes.

Below only shows two example structures. Additional examples can be found in D_ATA tutorial exercises in DL_Software Digital Guide:

https://dl-sdg.github.io/RESOURCES/EXERCISES/exercises_data.html

Example 1 (*example1.xyz*) – ethanoic acid liquid

Pure ethanoic acid liquid obtained from molecular dynamics simulations. The system consists of 674 molecules which gives a total of 5392 atoms. The file contains five configuration frames.

D_ATA will detect two types of interactions: HB and HP, between the *carboxylic CGs* and *alkane CGs*, respectively. The follow shows a list of interaction counts:

```
HB_20_20
-----
Unique interactions: 693.200000  deviation = 6.046487
DANAI          mean      deviation
[L2]O20E:H20O      = 579.400000 11.689311
[L2]H20O:O20L      = 113.800000 7.467262
[L3]O20L:H20O:O20E  = 5.400000 1.356466
[L3]H20O:O20E:H20O  = 22.600000 4.498889
[L3]O20E:H20O:O20E  = 5.800000 2.712932
[L3]O20L:H20O:O20L  = 0.600000 0.800000
[L4]O20E:H20O:O20E:H20O = 3.600000 2.727636
[L4]H20O:O20E:H20O:O20L = 1.000000 1.095445

Interactions with an average ratio count less than 0.001000 are not shown.
(ratio = average_count/max_average_count)

The following count-time profile files created:
output/count_HB_20_20.results

HP_1_1
-----
Unique interactions: 946.600000  deviation = 13.602941

DANAI          mean      deviation
[L2]C1p:C1p      = 946.600000 13.602941
[R3]C1p:C1p:C1p:C1p = 122.200000 6.764614
[L3]C1p:C1p:C1p   = 1741.200000 59.479072
[L4]C1p:C1p:C1p:C1p = 3063.200000 145.447448
...
...
```

Example 2 (*example2.xyz.gz*) – propanone-water system

Note this is a zipped file. D_ATA can read both zipped and unzipped files.

Two component system in liquid phase: acetone and water. The exact composition is shown in *d_ata.output* file as follows – there are 30 propanone molecules and 2036 water molecules.

```
-----SETUP MOLECULES-----  
  
Identified types of molecules as follows:  
  
Molecule: A  
Total molecules: 30  
Total atoms per molecule: 10  
Molecular weight: 58.081720  
Composition: C3 H6 O1  
  
Molecule: B  
Total molecules: 2036  
Total atoms per molecule: 3  
Molecular weight: 18.015340  
Composition: H2 O1
```

D_ATA expects the following interactions exist in the system (from *d_ata.results*):

*** Interaction profile composition ***

Macro-interactions:

HB_800_800
HB_800_16

Number of water CG (800) = 2036
Number of ketone CG (16) = 2066

Hydrogen bond (HB) interactions
between water - water and water-
ketone.

*** Interaction profile composition ***

Macro-interactions:

HB_1_1

Number of alkane CG (1) = 60

Hydrophobic (HP) interactions
between alkyl carbons (methyl
groups in this case).

11. Glossary

This chapter lists the meaning of terms within the framework of D_ATA.

All-inclusive interaction mode – This means D_ATA **does not** distinguish an *Atom* that either participates in an *isolated interaction* or not. All Atoms that participate in a given interaction mode within a macro-interaction will be considered. This option can be set in the CONTROL_FILE.

Atom – See *DL_F Atom*.

Chemical Group (CG) - A group consists of at least one or more connecting atom members that made up the chemical characteristic behaviour that is distinguished from other group atoms. For example, **carboxylic** is a Chemical Group, which consists of four atoms: carbonyl carbon (C20), carbonyl oxygen (O20E), hydroxyl oxygen (O20L) and hydroxyl hydrogen (H20O). The value '20' is called the Chemical Group Index (CGI), which is the unique numerical value for the CG.

DANAI – Acronym for DL_ANALYSER Notation for Atomic Interactions. It is a standard expression or Chemistry-based language construct to describe nonbonded interactions at a local level. A DANAI statement contains information about the Atomic species involve, the extent of interactions and the overall topological structure of Atoms involve in these interactions.

DL_F Atom – Or simply called *Atom*. It is a normal atom in every sense, but it is so coined within the context of *DL_F Notation*, to indicate such atom is recognisable within the Notation. The list of recognisable Atoms will eventually encompass other elements in future D_ATA versions.

DL_F Notation – Acronym for DL_FIELD Notation. It is a standard notation for an Atom that contains information about its chemical characteristics within a molecule.

Exclusive atom – DL_F Atom expressed in a DANAI statement that only participates in a non-bonded interaction with other DL_F Atoms as indicated in a DANAI statement. There is no other similar type of interaction occurs with other DL_F Atoms that are not indicated in the DANAI statement. The participating Atoms are expressed in capital letters. For instance: **C1p**, **O20E**, etc. See also *non-exclusive* atom.

Exclusive interaction – Also called an *isolated interaction*. A type of interaction that involves at least two or more *exclusive* DL_F Atoms as described in a DANAI statement. There are no additional interaction of the same type occurs with other DL_F Atoms other than what is described in the DANAI statement. The participating Atoms are expressed in capital letters. For instance: **C1p**, **O20E**, etc.

Isolated interaction – Also called an *exclusive interaction*.

Macro-interaction – A general, overall interaction between two Chemical Groups. For example, to describe collectively hydrogen bond interactions between carboxylic and alcohol CGs is expressed as **HB_15_20**, where 15 is the CGI for *alcohol* and 20 is the CGI for *carboxylic*.

Micro-interaction – A specific mode of interaction within a *macro-interaction*. In DANAI, a micro-interaction essentially consists of topological description and number of CGs involve

in the interaction and the Atomic species (expressed as DL_F Notation atom keys) involve in such interaction.

Non-exclusive atom - DL_F Atom that participates in a type of non-bonded interaction with other DL_F Atoms as indicated in a DANAI statement, **and** with other DL_F Atoms that are not shown in the statement. The participating Atoms are expressed in small letters. For instance: **c**1p, **o**20E, etc. See also exclusive atom.

Non-exclusive interactions - Interactions among Atoms according to a DANAI expression. In addition, these Atoms also form similar interactions with other Atoms not shown in the DANAI expression. Atoms that formed multiple interactions in this way are called non-exclusive atoms.

End of D_ATA version 1.1 user manual

C W Yong, May 2025

Please quote the following references in your publication:

For D_ATA Software:

D_ATA – *Atom Typer and Analyser, version 1.1*, https://www.ccp5.ac.uk/D_ATA

For DL_F Notation:

C W Yong, 'Descriptions and Implementations of DL_F Notation: A Natural Chemical Expression System of Atom Types for Molecular Simulations', *J. Chem. Inf. Model.* **56**, 1405–1409 (2016)

For DANAI:

C.W. Yong and I.T. Todorov, 'DL_ANALYSER Notation for Atomic Interactions (DANAI): A Natural Annotation System for Molecular Interactions, Using Ethanoic Acid Liquid as a Test Case', *Molecules* **23**, 36 (2018)