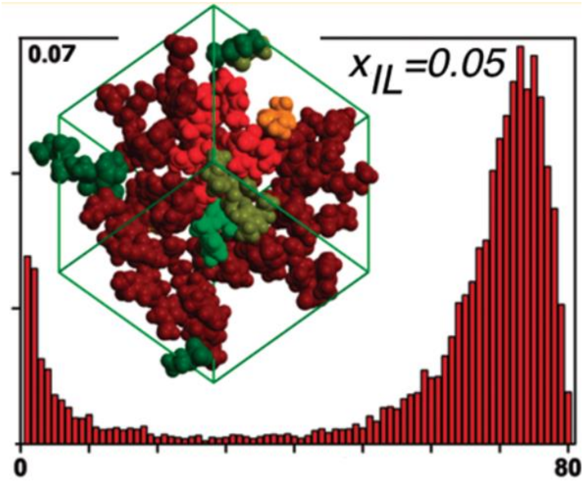
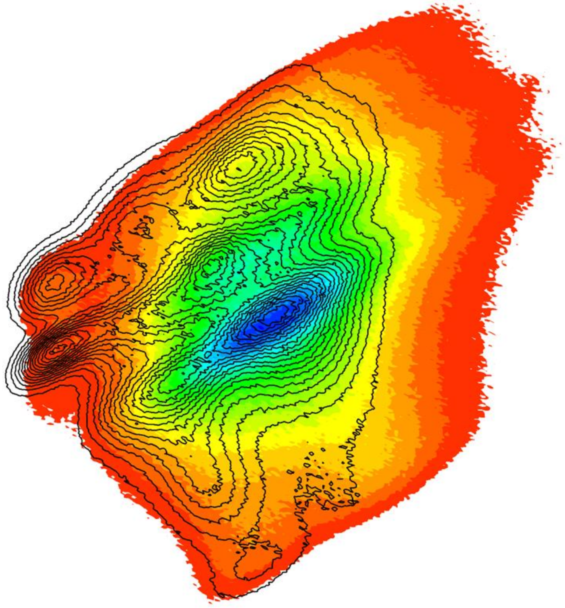
AGGREGATES

Studying the Aggregation of Molecules in Trajectories of Molecular Dynamics





Carlos Bernardes
(cebernardes@ciencias.ulisboa.pt)

10/1/2024

UNIVERSITY OF LISBON

Version 3.3.1

** INDEX **

1. DESCRIPTION

AGGREGATES is a Fortran code developed to find and analyze molecular structures in computer-simulated trajectories. The program can read trajectory files from several MD packages namely, DLPOLY, CHARMM, LAMMPS, and GROMACS. Additionally, PDB files can also be used. The program output consists of a series of statistical functions that can be used to characterize the size, shape, and organization of molecules in a network. An extensive description of the program functionalities can be found in the following publication:

C.E.S. Bernardes; *J. Comp. Chem.* **2017**, *38*, 753-755

If you find this software useful for your research, please cite the previous reference. To cite this User Manual please use:

C.E.S. Bernardes; "Aggregates User Manual", *Zenodo*, **2024**, DOI: http://doi.org/10.5281/zenodo.10465776.

The version currently available is working, however, it is not free from bugs. Any problems or questions do not hesitate to send an email to cebernardes@ciencias.ulisboa.pt. Additional information about the program and release notes can be found at http://webpages.fc.ul.pt/~cebernardes/index.html.

2. INSTALLING AND RUNNING

After download, the program should be extracted by typing:

```
$ tar -xvf agg_3_3.tar
```

The program is installed from the "src" directory, using the following steps:

(1) Change to directory "src":

```
$ cd agg_source_3_X_X/src
```

(2) Pre-compile the "libgmxfort" and "libxdrfile" libraries using:

```
$ sudo .pre-install.sh
```

The earlier script installs the libraries in the "/usr/local" folder, a step thatnormally requires superuser privileges.

(3) Finally, to install the program, typing:

```
$ make
```

This installs the program using Gfortran, and places the program executable in the folder "agg_source_3_X_X/EXE". This directory should be added to the system PATH, or the file copied to a folder already included in the PATH (e.g., /usr/local/*bin* or ~/*bin*). If necessary, check the available folders by typing:

```
$ echo $PATH
```

You can also edit the "BINROOT" line of the makefile file in the "src" directory, to install the program directly in the desired place.

The program runs by typing:

```
$ aggregates [arguments]
```

The following arguments are available:

```
-dcd     [file]    Read trajectory from a CHARMM DCD file.
-dcddir   [dir]    Reads all DCD trajectory files in a folder.
-h                 Help information.
-hstdir   [dir]    Reads all HISTORY trajectory files in a folder.
-lmps    [file]    Read trajectory from a Lammps output.
-lmpsdir  [dir]    Reads all LAMMPS trajectory files in a folder.
-pdbdir   [dir]    Read PDB files from a directory.
-pdbtrj  [file]    Read trajectory from a PDB file.
-psf     [file]    Read molecular information from a CHARMM psf file.
-top     [file]    Read the topology file.
-v                 Verbose mode.
--version          Output version information.
-xtc     [file]    Read the XTC trajectory file from GROMACS
-xtcdir   [dir]    Reads all GROMACS XTC trajectory files in a folder.
```

After starting the program, it checks the available topology files and retrieves the necessary information. If no information is introduced regarding the trajectory and topology files to use, AGGREGATES will look, by default, for DLPOLY's FIELD and HISTORY files. The user will then be guided by a series of questions to set the analysis.

If the calculation is to be repeated for other identical systems or to run the program in the background, an input file can be used. For this, at the beginning of the questionnaire, the user is prompted if wants the answers to be recorded. This generates a file named "INPUT_XX", where "XX" is an incremental index that avoids overwriting pre-existing input files. To use these files, run the program a:

```
$ aggregates [Arguments] < INPUT_XX
```

These archives are human-readable. Thus, it is possible to edit them before running the program. The transferability of input files generated by different versions of the program is not ensured.

3. UNITS

The default distance units used by the program in the starting questionnaire and output files is Angstrom (10^{-10} m). GROMACS uses nanometers (10^{-9} m) so that, AGGREGATES automatically converts the necessary data into Angstrom. Thus, no conversion of the input/output files (i.e., topology and *.XTC files) is necessary.

4. INPUT FILES

AGGREGATES needs two INPUT files: i) one having the information about the simulated system and ii) a trajectory file. For the first case, by default, the FIELD file of DLPOLY is searched. In the case of CHARMM and GROMACS, it is possible to use the PSF and TOP topology files, respectively. However, to use LAMMPS or (a series of) PDB files, a topology file with the following structure can be used:

```
types          2
name         NTF
nummols      400
atoms          5
C   12.011
F   18.998
F   18.998
F   18.998
S   32.066
finish
name        mimo
nummols      400
atoms          4
N   14.007
C   12.011
N   14.007
C   12.011
finish
```

In this file “types” refers to the number of types of molecules in the simulation box; “name” the molecule name; “nummols” the number of molecules in the system; “atoms” the number of atoms in the molecule; and “finish” indicates that all entries were given for the molecule. Immediately after the “atoms” entry, the name of each atom followed by the corresponding atomic mass should be given. Note that: i) **the order of appearance of the molecules and atoms should match that used in the trajectory file** (the program will not check this); ii) **the initial 2 characters of the atom's**

name will be used to recognize the element. The program is case sensitive, i.e., an atom that starts with "NA" is recognized as a Nitrogen, and one that starts with "Na" is recognized as a sodium atom (this also applies while reading other types of topology files). **Always check if the information retrieved from the topology files is correct**. This will not be checked by the program and, wrong data may lead to abnormal results. For this assessment, check the information initially printed by the program.

In the case of LAMMPS trajectory files, only the following options are supported:

- Boundary conditions

```
xx yy zz
dd dd dd
ff ff ff
```

- Atomic coordinates

The following must be present in any order: element id x y z.

It is also possible to use unwrapped coordinates: xu yu zu.

5. QUESTIONNAIRE

After starting the program, a series of questions prompt the user to set the calculations. All calculations are independent, so that, any combination of analysis is, in principle, possible. Although not bulletproof, the user will be guided to avoid unsupported options.

A discussion of the calculation options can be found in the paper [1]. In brief, considering two interaction centers A and B, it is possible to consider:

1. **A** = **B** are equal and found in the same molecule so that aggregates of the type -A-A-A-A- are searched by the program.
2. **A** and **B** are in different molecules, with the added restriction that **A** must be connected to **B**'s and vice-versa. Aggregates of the type -A-B-A-B- are searched by the program.
3. **A** and **B** belong to the same molecule but are formed by different atoms. Aggregates of the type -A-A-A-A- are searched by the program.
4. **A** and **B** are placed in different molecules and interaction is considered in any order (i.e., an **A** is considered to belong to the aggregate even if it only connects another **A**). Aggregates of the type -A-B-A-B- are searched by the program.
5. Do not perform aggregation analysis.
6. Same as option two, but molecules of **B** can be excluded from the aggregate depending on the number of **A** molecules in their neighborhood. Aggregates of the type -A-B-A-B- will be computed by the program.

6. OUTPUT FILES

All runs generate a file that has information about the calculations and the statistical analysis. This file is named "AGGOUT_XXX.txt", where "XXX" is an incremental index that avoids overwriting output files of consecutive runs. Details regarding the statistical functions can be found elsewhere [1].

Depending on the selected calculation setup, other files are also recorded by the program. A list of the files recorded during a single run is always appended at the end of AGGOUT_XXX.txt and may include:

- ***FRM_COLOR_1_XXXX.pdb*** PDB files of each configuration, with the molecules that compose the aggregates. The code in the file after "COLOR", depicts the coloring mode used: 1- aggregates; 2- the number of neighbors (see the color code at the beginning of the PDB files). The PDB files are recorded in the folder "frms" which is created by the program in the directory where the program is run.
- ***AGG_TRAJECTORY.pdb*** PDB file located in the "frms" folder, which contains the aggregates found in all trajectory configurations and colored as in the case of FRM_COLOR_1_XXXX.pdb. The trajectory can be opened with, e.g., VMD. However, because the number of atoms in each configuration may vary, VMD may fail to open the file properly.
- ***CDF_big_xyz_XX.txt*** Text file with the results of combined distribution functions (CDF). "big" (or "small") refers to results obtained for aggregates with more (or less) elements than the threshold defined by the user in the questions section. XYZ refers to a file with the results in the form of three columns. It is also possible to record the data as a matrix (index "MTX"), to produce 3D surfaces.
- ***SDF_big_at_nm3.cube*** Spatial distribution function recorded as a Gaussian cube file format, with units of nm^{-3}. "big" (or "small") refers to results obtained for aggregates with more (or less) elements than the threshold defined by the user. "at" is the atom name used as a probe in the calculation. **Caution:** in this case, files with the same name may be overwritten.

- ***SDF_big_at_dens_norm.cube*** Spatial distribution function recorded as a Gaussian cube file format, with units normalized considering uniform particle density. “big” (or “small”) refers to results obtained for aggregates with more (or fewer) elements than the threshold defined by the user. “at” is the atom name used as a probe in the calculation. These files are only generated in the case of type 5 calculation (i.e., no aggregation analysis is performed). **Caution:** in this case, files with the same name may be overwritten.
- ***1st_Shell_B_Total_XX.txt*** First shell analysis results, with the data in three columns: D1, D2, and value. “B” (or “S”) refers to results obtained for aggregates with more (or less) elements than the threshold defined by the user in the questions section. Total stands for the global analysis results combining all types of interactions.
- ***1st_Shell_B_Partial_XX.csv*** Similar to the global first shell analysis files, but with data for the individual interactions. In this case, the data is recorded as a comma-separated values file (CSV). Due to the large amount of data these files can hold, they are only recorded if requested.
- ***Conf_Anal_B_NAME_1-2-3.txt*** Conformational analysis results. “B” (or “S”) refers to results obtained for aggregates with more (or less) elements than the threshold defined by the user in the question section, followed by the molecule name (“NAME” in the example) and the atoms (in this case, 1-2-3 is the angle between atoms 1, 2 and 3 was evaluated).
- ***2DConf_Anal_B_NAME_X.txt*** Combined distribution functions (CDF) computed between two molecular conformational analyses (e.g., distance between atoms and an angle). “B” (or “S”) refers to results obtained for aggregates with more (or less) elements than the threshold defined by the user in the questions section. "NAME" is the molecule name.
- ***Conf_Anal_time_NAME_A-B-C-D.txt*** Conformational analysis results as a function of simulation time. "NAME" is the molecule name, followed by the position of the atoms forming a bond, angle, or dihedral.

7. ADDITIONAL NOTES

- The volume calculation choice only considers the atoms selected in the aggregation analysis. In other words, if the interaction is computed only between two atoms, only these atoms will be used to compute the aggregate volume or its apparent volume.

- The normalization of CDFs should be used with caution. Considering that these functions are computed only for aggregates of a specific size, neglecting all the remaining molecules in the simulation box may lead to abnormal results.

- The use of PDB files requires knowledge of the simulation box dimensions. For this purpose, two different approaches can be used: *i*) it is possible to give the length of a cubic box and assume an NVT simulation; *ii*) provide in each frame a line with the box parameters e.g.:

```
CRYST1   60.385  60.385  60.385  90.00  90.00  90.00
```

In the earlier example, 60.385 refers to x, y, and z dimensions (in Å), and 90.00 to the α, β, γ angles in degrees). In this case, an NPT simulation is considered. Note, however, that if some PDB files do not have this line, the last value read by the program will be considered.

- The arguments **–dcddir**, **–hstdir**, **-lmpsdir**, and **–xtcdir**, may be used to read all trajectory files contained in a folder. This folder should only contain trajectory files. Each file can have a different number of frames.

8. RELEASE NOTES

** VERSION 3.3.1 **

Released on 11 January 2024

News

- New sub-menu for the analysis of computed data from the main module (e.g., outputted frames). This includes the possibility to find dimeric molecular units and the computation of their interaction energy. The output is a PDB file with the dimers and corresponding energy contributions (Coulomb and VDW), and a file with data to plot the energy variation as a function of the distance between the centers of mass of the molecules.

- Solved a problem related to the compilation of the *libxdrf* library, due to a support change in the more recent OSs (e.g., Debian-based operating systems).

** VERSION 3.2.0 **

Released on 22 January 2021

New Features

- A new aggregation type of calculation, which is based on a type 2 procedure, is now available. In this case, after the initial analyses, it allows to exclude molecules depending on their number of neighbors.
- Center large molecules (e.g., polymer chains) at the center of the simulation box [2].
- Computation of aggregates apparent volume using a variable number of vectors and for a single molecule.
- New advanced menu section, which allows: (i) control reconstruction of broken molecules; (ii) trajectory read control; (iii) apparent volume calculation options.
- Improved output report, which includes glitch corrections and printing additional information about the calculation and system topology.

Bug Fixes

- Memory leak problem solved.
- Improved compatibility with GROMACS topology files.
- Corrected problems during ETA calculation.

** VERSION 3.1.6 **

Released on 21 August 2019

A memory leak occurs when the first shell analysis is performed. The origin of this problem is under investigation. It only happens when the program is compiled using GFORTRAN 7.0 or higher. The program memory usage was improved to minimize this issue, but a permanent fix is expected to be released only soon.

Bug Fixes

- Corrected problem while handling periodic boundary conditions with trajectories of triclinic simulation boxes produced with LAMMPS.

** VERSION 3.1.5 **

Released on 09 August 2018

New Features

- Compatibility improvement with GROMACS *top files.
- SDFs are now computed with units in nm^{-3} and normalized considering a uniform particle density when type 5 calculation is used.
- Other small tweaks!

Bug Fixes

- Corrected the problem that prevented the calculation of CDFs with normalization.
- Corrected problem when reading *PDB files in a folder.

**** VERSION 3.1.4 ****

Released on 8 March 2018

New Features

- It is now possible to compute the bonds, angles, and dihedrals for each molecule in the simulation box, for each step.
- Extended support for LAMMPS trajectory files.

Bug Fixes

- Fixed an issue that could prevent the program from reading from GROMACS topology files.
- Fixed an issue that could lead to a wrong assignment of molecules during molecular conformation analysis.

**** VERSION 3.1.3 ****

Released on 19 September 2017

New Features

- New user interface.

Bug Fixes

- Corrected problem reading GROMACS topology files.
- Corrected memory allocation problem using calculation options 3 and 4.

**** VERSION 3.1.2 ****

Released on 28 July 2017

New Features

- Allows restraining 1st shell calculation according to VDW radii.
- Perform 1st shell analysis between different molecules.
- Revision of the normalization procedure of CDFs.
- Ability to write conformational CDFs as matrices.

Bug Fixes

- Corrected problems related to the computation of conformational combined distribution functions.

**** VERSION 3.1.1 ****

Released on 17 February 2017

New Features

- It is now possible to read multiple trajectory files.
- A few tweaks of the code and a questionnaire.

Bug Fixes

- Solved issues related to the setup of the packing coefficient calculation.
- Fixed atoms identification problem while using PSF files and when the program is compiled with Intel® Fortran Compiler.
- Solved issue that prevents the correct reading of LAMMPS trajectory files.

**** VERSION 3.1 ****

Released on 17 January 2017

New Features

- Computation of combined conformational analysis (CCA).
- Restrain the computation of CDFs, SDFs, FSAs, CAs, and CCA according to the number of neighbors of the molecules in the aggregates.
- Record a PDB trajectory file with the molecules that compose the aggregates.
- Other small tweaks and optimizations.
- Ability to read GROMACS topology files.

Bug Fixes

- Corrected bugs while performing the simultaneous calculation of several conformational analyses.
- Corrected bug while reading PSF files, when the program was compiled with Intel® Fortran Compiler.

** VERSION 3.0 **

Released on 01 September 2016

New Features

- Ability to read trajectory files from GROMACS and LAMMPS, and also to use PDB files.
- Computation of CDFs, SDFs, FSA, and molecular conformational analysis.
- The ability to restrict the configurations used in the analysis.
- The questionnaire was redesigned, and new features were implemented to avoid problems in the calculations.
- Calculation of the average number of aggregates during the simulation and distribution of the number of neighbors.
- Record the Aggregates in the simulation box in PDB files, colored according to the number of neighbors of the molecules.
- Several other tweaks and optimizations.

** VERSION 2.3.1 **

Released on 11 December 2015

New Features

- The connectivity criteria can now be based on the Van der Waals radii, R_{vdw}, of the atoms involved in the evaluation. Giving atoms A and B the criteria, D, is defined as $D = R_{vdw}(A) + R_{vdw}(B) + C^{te}$, where C^{te} is a constant value, that can be defined by the user. The atoms names are automatically retrieved from the initial two letters of the atoms name and, the following atomic species are recognized: H, He, Li, Be, B, C, N, O, F, Ne, Na, Mg, Al, Si, P, S, Cl, Ar, K, Ca, Se, Br, Kr, I and Xe. For an unidentified item, a value of 1.75A is assumed. The program is case-sensitive. Thus "CA" and "Ca" can be used in the same run to define an aromatic carbon or calcium atom, respectively.
- For the aggregate volume calculation using VdW radii, a constant value can also be added as described above.
- In this version all calculation reports are recorded in different files, avoiding overwriting previous results. For this purpose, an incremental index is added to the output files that now take the form

AGGOUT_xxx.dat. The output files also contain a description of the calculation details.

- The program is now able to record an input file (INPUT_xx) that can be used to run the program with the same conditions in a future calculation or for sequential calculations.

**** VERSION 2.3 ****

Released on 24 August 2015

New Features

- Ability to read CHARMM output files (*.dcd and *.psf);
- Calculation of aggregate area and volume using atomic VDW radii. The new option relies on the GEPOLE (*J. Comput. Chem.* **1991**, *12*, 1077-1088) and ARVO (*Comput. Phys. Comm.* **2005**, *165*, 59–96) codes and, by default, uses Van der Waals radii recommended by Bondi (*J. Phys. Chem.* **1964**, *68*, 441-451; *Chem. Eng. News.* **2009**, *87*, 19).
- Analysis of aggregates composed of different centers (in the same or different molecules). For example, between OH groups in different molecules, that form continuous chains of hydrogen bonds (alternate between the centers is not required, as in the case of the polar networks of ionic liquids).

Bug Fixes

- Major memory optimization to handle large simulation boxes and trajectory files.
- A problem about the computation of an average number of neighbors in large simulations.

9. ACKNOWLEDGEMENTS

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal through projects:

UIDP/00100/2020, with DOI 10.54499/UIDP/00100/2020 (https://doi.org/10.54499/UIDP/00100/2020).

UIDB/00100/2020, with DOI 10.54499/UIDB/00100/2020 (https://doi.org/10.54499/UIDB/00100/2020).

LA/P/0056/2020, with DOI 10.54499/LA/P/0056/2020 (https://doi.org/10.54499/LA/P/0056/2020).

2021.03239.CEECIND/CP1650/CT0003, with DOI 10.54499/2021.03239.CEECIND/CP1650/CT0003 (https://doi.org/10.54499/2021.03239.CEECIND/CP1650/CT0003).

10. REFERENCES

[1] Bernardes, CES; *J. Comput. Chem.* **2017**, 38, 753-765. http://dx.doi.org/10.1002/jcc.24735

[2] Zanatta, M; Lopes, M; Cabrita, EJ; Bernardes, CES; Corvo, MC; *J. Co2 Util.* **2020**, 41, 101225. http://dx.doi.org/10.1016/j.jcou.2020.101225