ToBvalid version 0.9.7

ToBvalid is a Python library and a program for the statistical analysis and validation of ADPs. It comes under MPL-2.0 license and supports Python3.

Installation

In order to install ToBvalid from source:

```
pip install tobvalid
```

Github repository (*https://github.com/ToBvalid/)* also may be used:

```
pip install git+https://github.com/ ToBvalid
```

or clone the project (or download a zip file) and from the top-level directory do:

```
pip install .
```

Dependencies

ToBvalid has the next dependencies: pandas, fire, matplotlib, numpy, scipy, gemmi>=0.3.8, seaborn, statsmodels. All dependencies will be automatically checked and installed if it is necessary during ToBvalid installation.

ToBvalid functionalities

This tool is designed for modelling of ADP distribution and their validation on both global and local levels. Internal functionalities of ToBvalid include:

- Overall statistical analysis of ADP distribution.
- Parametrisation of ADP distribution (mixture) and validation of parameters .
- Search for potential lighter and heavier atoms which may have been modelled wrongly
- Comparison of the relative occupancy of two atoms. (only in the library, not implemented in the program yet.)
- Validation of ligands.

Usage

ToBvalid reads pdb files and implements validation of ADPs. It is strongly advised to re-redine the structure with any refinemet software before the validation.
The input file is pdb and output files are html reports, graphs and text files with outliers' lists of both local and global analysis.

```
tobvalid [-h] [-i <pdb file>] [-o <output file directory>]
           [-m <number of modes | auto>] [-p <json parameter file>][-g]
           [-l] [-c]
```

Defaults values:
o - current directory
m – 1
p - None

Sometimes ADP distribution may be multimodal (as a mixture of Shifted Inverse Gamma Distributions (SIGD)). The default number of modes is one. If the number of modes is needed to be defined use -*auto*.

If only local/global/chain analysis is needed –*l*, -*g* or -*c* respectively should be used.

Example of json config file:

```
{
    "gmm":{
        "maxiteration": 200,
        "tolerance":1e-04,

    },
    "igmm":{
        "maxiteration": 100,
        "tolerance":1e-04,
        "ext":"stochastic"
    },
    "plot":{
        "dpi":150
    },
    "local":{
        "r_main": 4.2,
        "r_water": 3.2,
        "olowmin": 0.7,
        "olowq3":0.99,
        "ohighmax":1.2,
        "ohighq1":1.01
    },
    "ligand":{
        "r_main": 4.2,
        "olowmin": 0.7,
        "ohighmax":1.2
    }
}
```

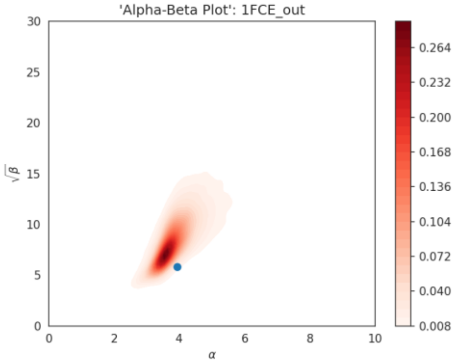
The details of the input may be printed with *help:*

tobvalid -h

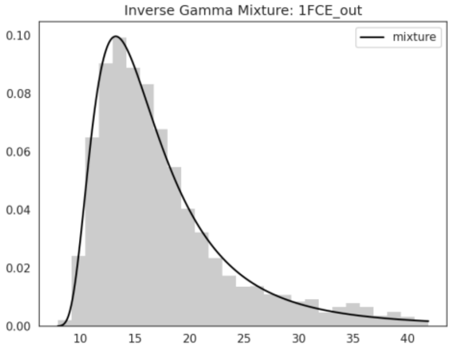
Outputs:

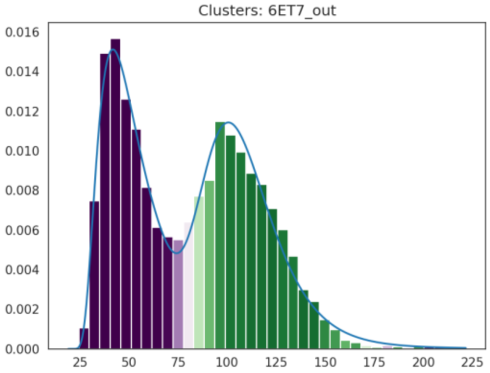
Outputs of the global analysis are listed below:

1. The results of the global analysis and overall statistics reports are given as html files.
2. alpha-beta plot. SIGD parameters (α, β) are placed on contour plot. If these parameters are out of the contour this structure should be considered for rebuilding and re-refinement.

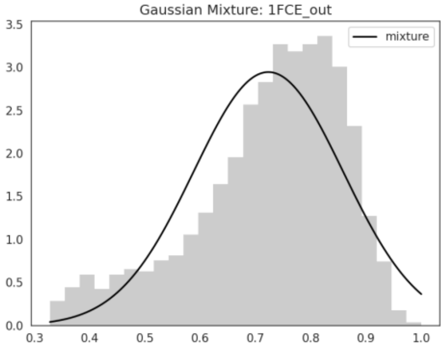


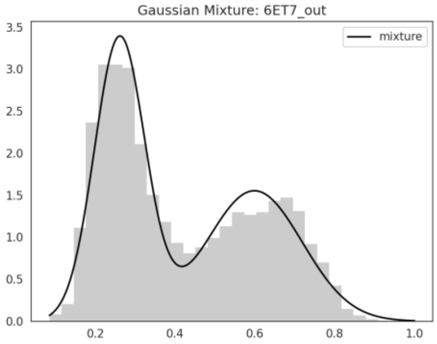
3. SIGD (mixture) plot. If there is one mode in the ADP distribution the is just one plot, but for the multimodal cases the plot of clusters are also given.



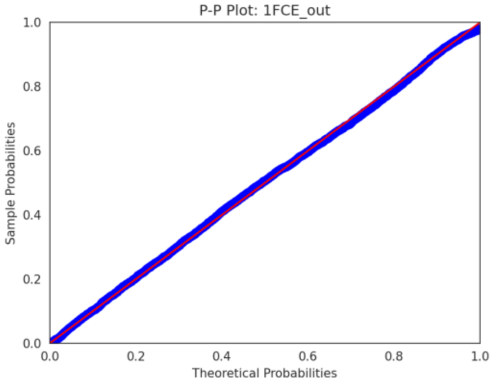


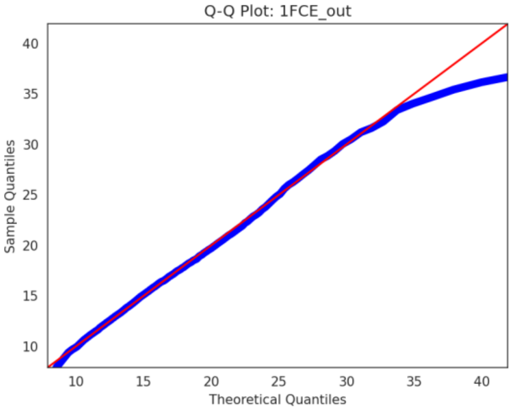
4. Peak height distribution plot. Here is the relation between B values and the resolution is demostrated. In this plot several atoms in the very left/right may be the sigh of light and heavy atoms respectively.





5. qq and pp plots illustrate the agreement the ADP distribution of the given structure to its theoretical probabilities.





6. List of Interquartile outliers. Very small/large values of ADPs are not only considered susbicious, but also affect the parametrization. We make our data out of these values in the beginning of global analysis and provide users with the list.

Outputs of the global analysis are listed below:

1. File named {filename}_local.txt lists the report of the overall local analysis. In this validation each atom is checked with consideration of its environment (neighbouring atoms). In this file, potential lighter/heavier atoms are listed with the calculated optimal occupancy and also with the list of neighbouring atoms and the basic statistics for the environment.
2. File named {filename}_water.txt lists water molecules with the number of neighbouring atom 6 and more.
3. File named {filename}_ligand.txt lists the results of the local analysis for each ligand.
4. ligands_validation.txt provides the list of wrong ligand.