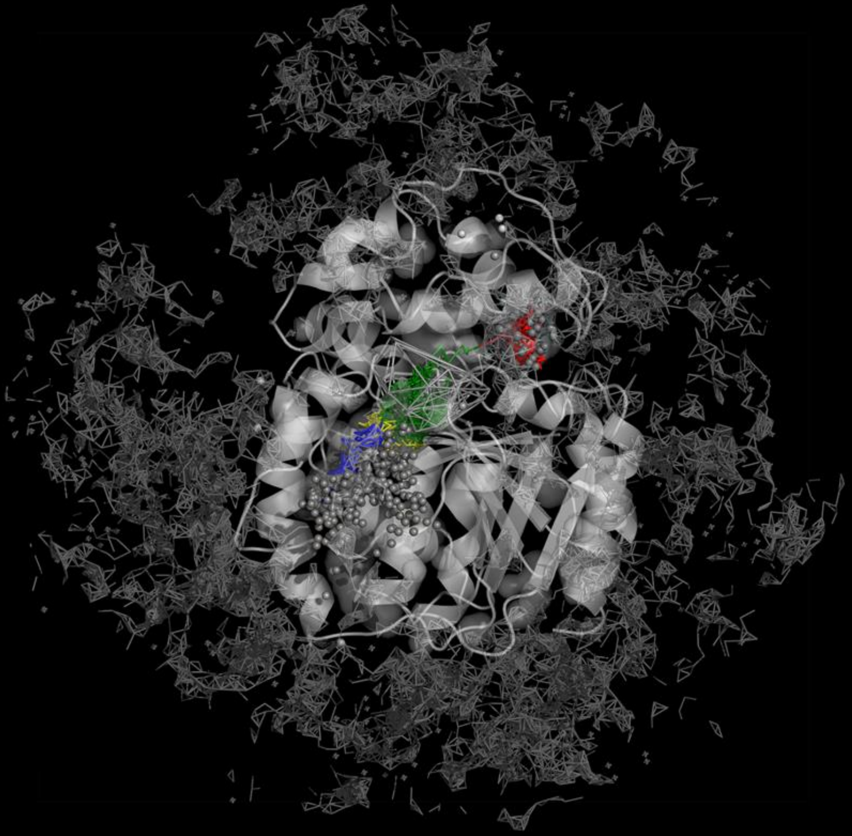
AQUA DUCT by TG



TIPS & TRICKS - clustering

TIPS & TRICKS - clustering

Adjusting method of clustering

This section will show you how your results can vary from chosen method of clustering and it also contains tips about adjusting method of clustering to acquire desired results. A 10 ns simulation of murine soluble epoxide hydrolase structure (PDB ID: 1CQZ) was used to run sample calculations and test different clustering methods. Method of clustering in AQUA-DUCT can be easily changed through **config.txt** file. The inspection of MD simulations suggests that optimal clustering method should group inlets into one main cluster (C1), and two smaller clusters (C2 and C3). Defined clusters with AQUA-DUCT correspond to the hollows in the protein surface (Figure 1b) and groups of entry/exits of smoothed trajectories of water molecules (Figure 1c). Other inlets should be classified as outliers, since transport of solvent molecules through the rest of the tunnels is limited in comparison to the main clusters.

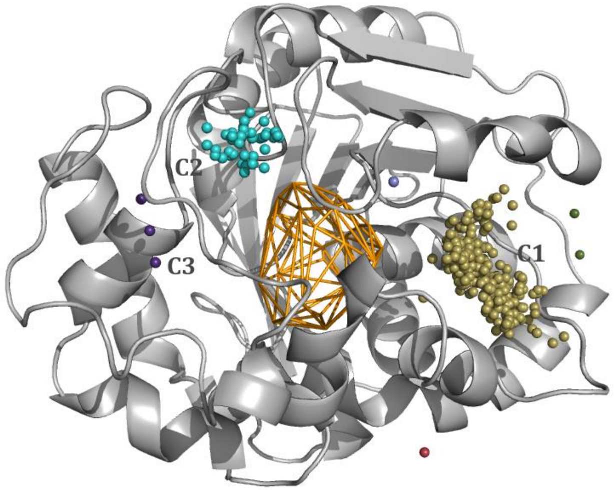


Fig. 1a. Optimal results of clustering of AQUA-DUCT results with `meanshift` method,
`bandwidth = Auto`
`cluster_all = True`
of 1CQZ with three main clusters provided:
C1 (gold), C2 (light blue), C3 (purple) and other minor 1- or 2-inlet clusters

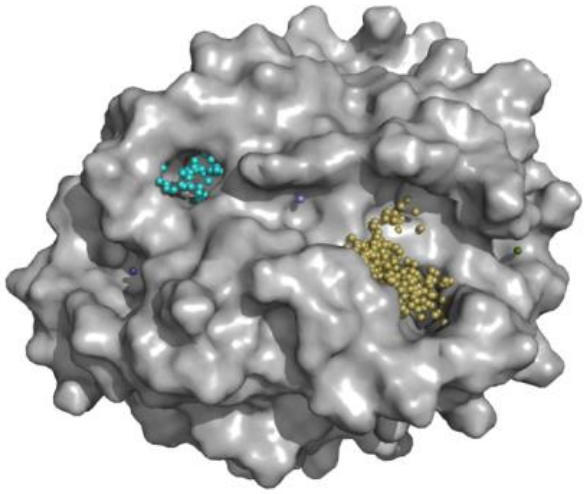


Fig. 1b. Surface of the protein with clusters located in the proximity of the hollows

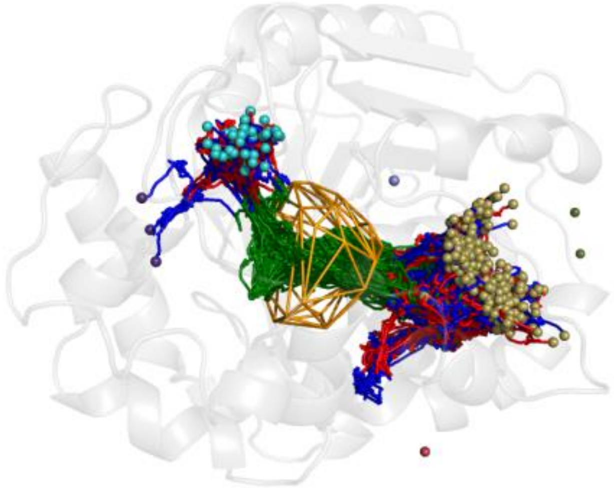


Fig. 1c. Smoothed trajectories of water molecules with entry/exit detected in one of the main clusters

TIPS & TRICKS - clustering

Barber

Barber method is set as a default one. Barber works by constructing a collection of spheres according to Auto Barber separate_paths stage settings or according to parameters given in clustering section, and then coherent clouds of inlets of mutually intersecting spheres are grouped into separate clusters. Results acquired with this method correspond to protein surface hollows. This method frequently is an optimal one and works well on defining clusters. The barber clustering method results for the sample system show three major clusters: C1 (gold), C2 (light blue) and C3 (purple). Outliers were classified as separate clusters (Figure 2).

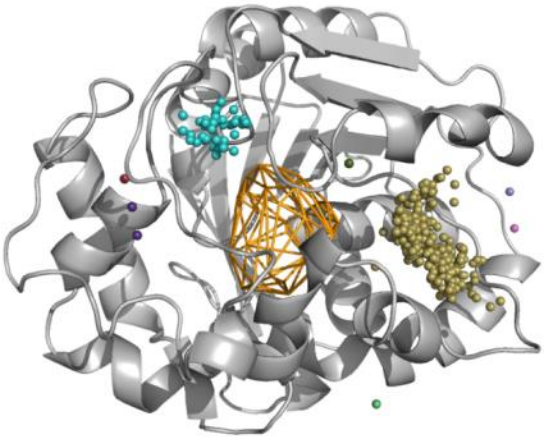


Fig 2. Results of clustering with barber method on default parameters*, i.e.

```
auto_barber = protein
auto_barber_maxcut = None
auto_barber_mincut = None
auto_barber_tovdw = True
```

* More details about barber method parameters in Tips&Trics - Trimming paths.

TIPS & TRICKS - clustering

Meanshift

Another clustering method which frequently gives consistent results is `meanshift` (Figure 3a). For this method two options are of crucial importance: `bandwidth` and `cluster_all`. `Bandwidth` can either be set to `Auto` (Figure 3a), then its value is generated in an automatic way, or a particular value can be chosen by the user (Figure 3b, c). Increasing this value caused unification of clusters C2 and C3, whereas C1 remained isolated (Figure 3c). `Cluster_all` is a flag that is by default set to `True` (Figure 3a). Setting it to `False` may result in classifying of some inlets on the edges of clusters as outliers (Figure 3d).

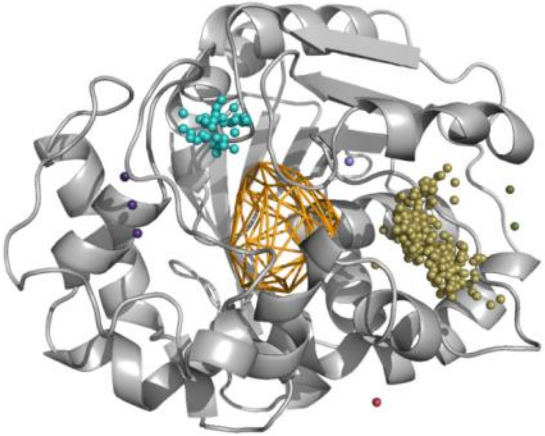


Fig. 3a. Results of clustering with `meanshift` method on default parameters, i.e.
`bandwidth = Auto`
`cluster_all = True`

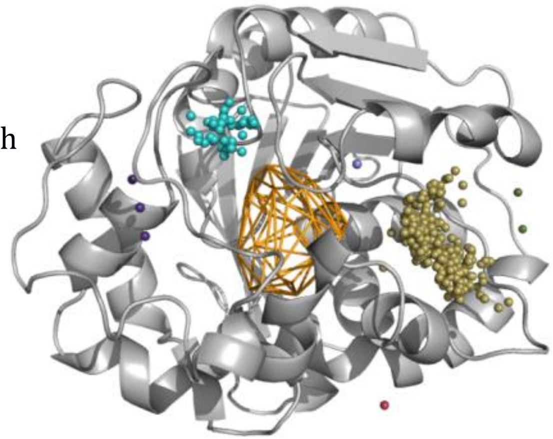


Fig. 3b. Results of clustering with `meanshift` method
`bandwidth = 5`
`cluster_all = True`

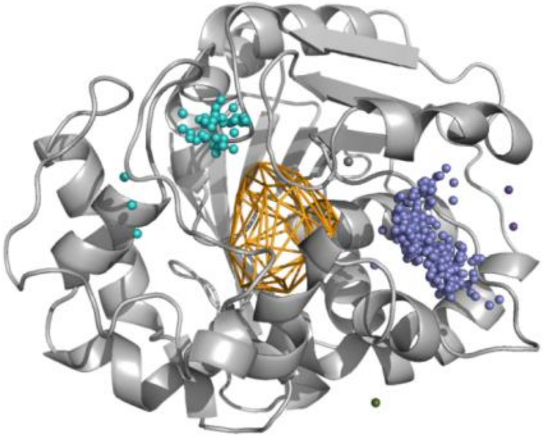


Fig. 3c. Results of clustering with `meanshift` method
`bandwidth = 7.5`
`cluster_all = True`

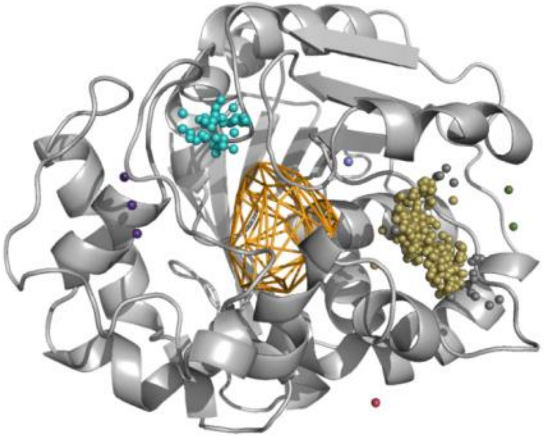


Fig. 3d. Results of clustering with `meanshift` method
`bandwidth = Auto`
`cluster_all = False`

TIPS & TRICKS - clustering

Birch

Clustering method which can be efficiently used to separate large clusters is `birch`. Running this method on default settings does not always give consistent results of inlets clustering (Figure 4a). This problem, however, can be easily solved by setting the `n_clusters` value to the desired number of clusters. If the clustering of inlets is still not satisfying, recursive clustering has to be conducted. For this structure, it was not possible to set `n_clusters = 3` because clusters C2 and C3 were not separated and cluster C1 was divided into two sections (Figure 4a). Option `n_clusters` was first needed to be set to 2 so that all solvent molecules could be assigned to two clusters (Figure 4b). Then the smaller cluster (cyan) was divided by setting `recursive_threshold` and adjusting it to its size. Option `max_level` had to be set to 1 and the method for recursive clustering was set to `birch` with `n_clusters = 2`. The results of the next calculations showed three separate clusters: C1, C2 and C3 (Figure 4c).

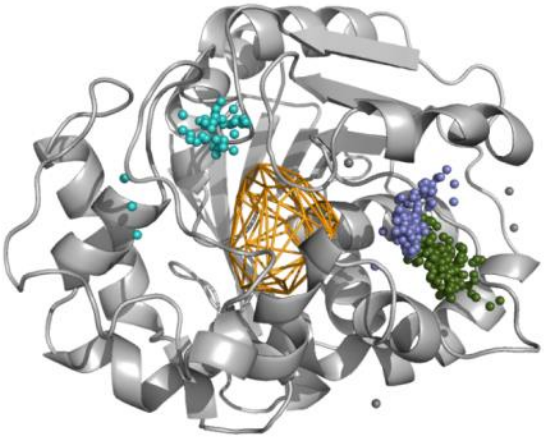


Fig. 4a. Results of clustering with `birch` method on default parameters, i.e.
`n_clusters = 3`

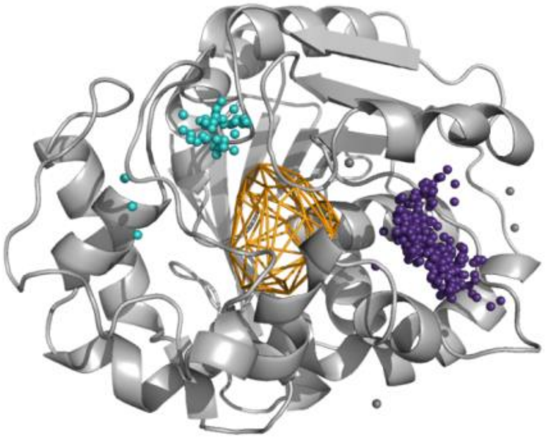


Fig. 4b. Results of clustering with `birch` method
`n_clusters = 2`

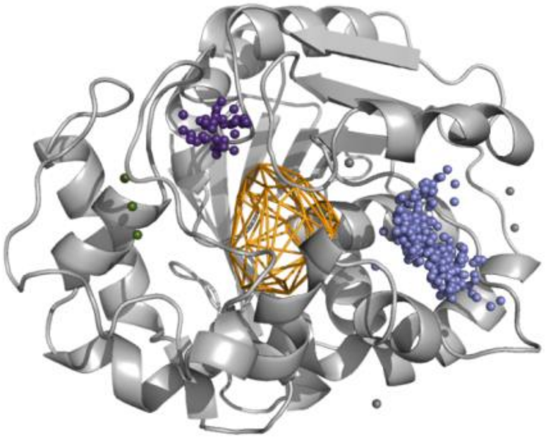


Fig. 4c. Results of clustering with `birch` method
`n_clusters = 2`
recursive clustering by `birch` with
`n_clusters = 2`

TIPS & TRICKS - clustering

Kmeans

Similar results to `birch` clustering method can be provided by `kmeans` method. Default settings yielded results with traced residues classified to an excessive number of clusters with cluster C1 divided into four sections (Figure 5a). Same as with `birch`, the `n_clusters` option had to be set to 2 (Figure 5b) and then recursive clustering set to `kmeans` with `n_clusters = 2` was used to separate clusters C2 and C3 (Figure 5c).

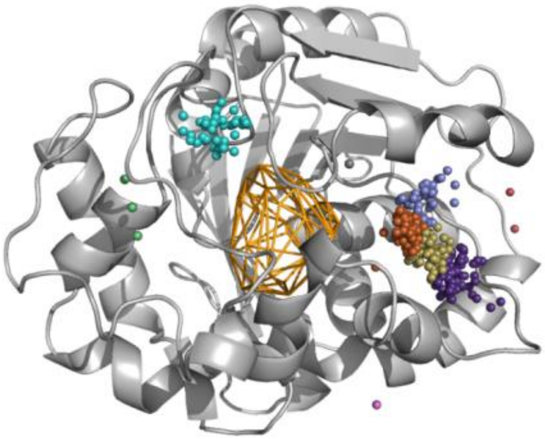


Fig. 5a. Results of clustering with `kmeans` method on default parameters i.e.
`n_clusters = 8`

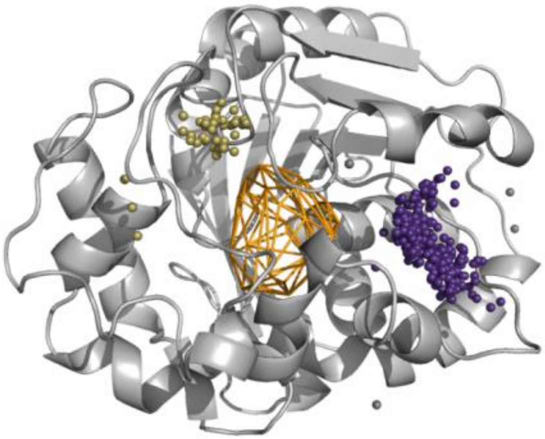


Fig. 5b. Results of clustering with `kmeans` method
`n_clusters = 2`

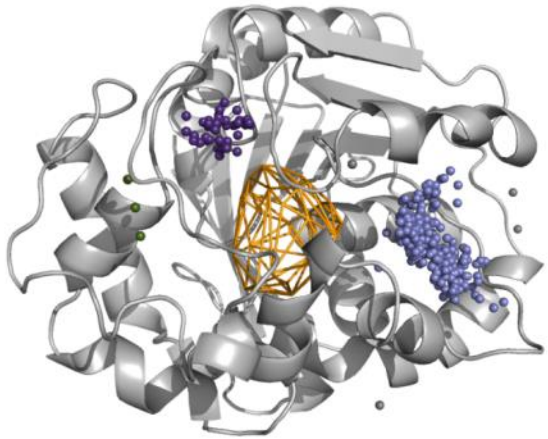


Fig. 5c. Results of clustering with `kmeans` method
`n_clusters = 2`
recursive clustering by `kmeans` with
`n_clusters = 2`

TIPS & TRICKS - clustering

Dbscan

Clustering method `dbscan` on default settings did not provide satisfying results. Most traced molecules were classified as outliers (gray) (Figure 6a). In `dbscan` method it is important to set appropriate value of `eps` option. Alongside with it, `metric` option can be adjusted to optimize results of the clustering. The default `metric` is set to `euclidean`. The suitable value of `eps` can vary depending on `metric` (Figure 6e versus 6h). Increasing the value of `eps` leads to reduction of number of inlets classified as outliers at the edges of the clusters. In spite of augmentation of `eps`, C3 remained classified as outliers and it was not possible to cluster it.

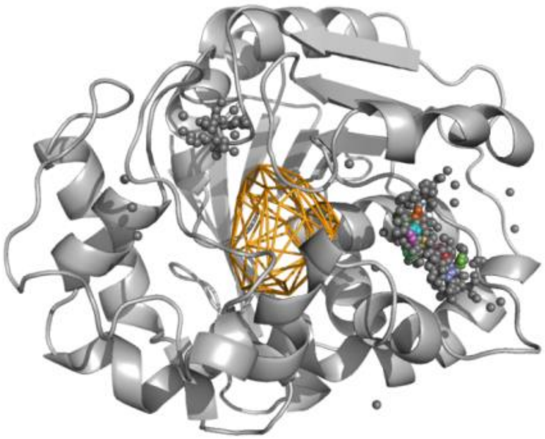


Fig. 6a. Results of clustering with `dbscan` method on default parameters i.e.
`metric = cityblock`
`eps = 0.5`

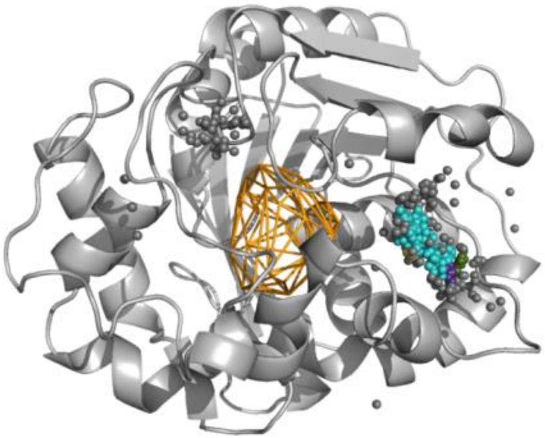


Fig. 6b. Results of clustering with `dbscan` method
`metric = cityblock`
`eps = 1.0`

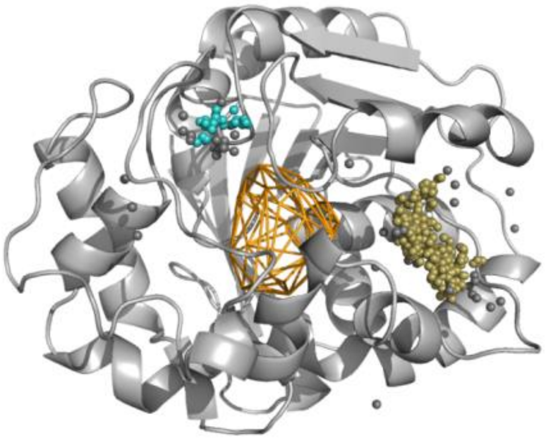


Fig. 6c. Results of clustering with `dbscan` method
`metric = cityblock`
`eps = 1.5`

TIPS & TRICKS - clustering

Dbscan cd.

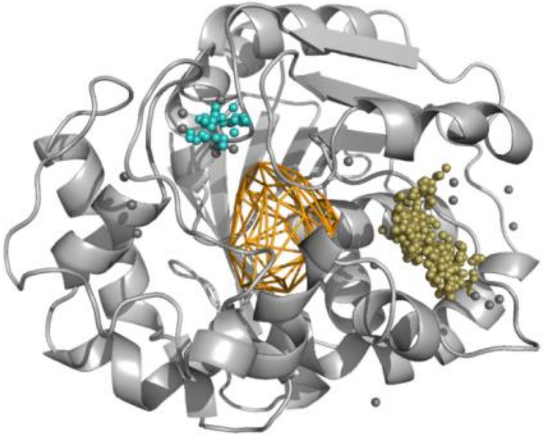


Fig. 6d. Results of clustering with `dbscan` method
`metric = cityblock`
`eps = 2.0`

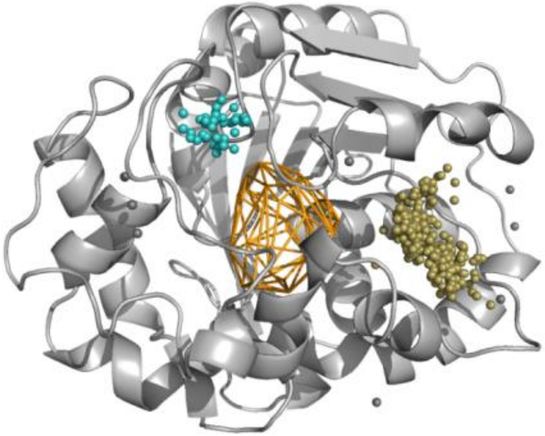


Fig. 6e. Results of clustering with `dbscan` method
`metric = cityblock`
`eps = 4.0`

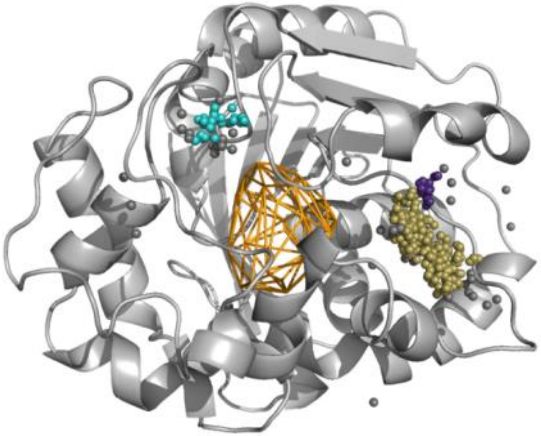


Fig. 6f. Results of clustering with `dbscan` method
`metric = euclidean`
`eps = 1.0`

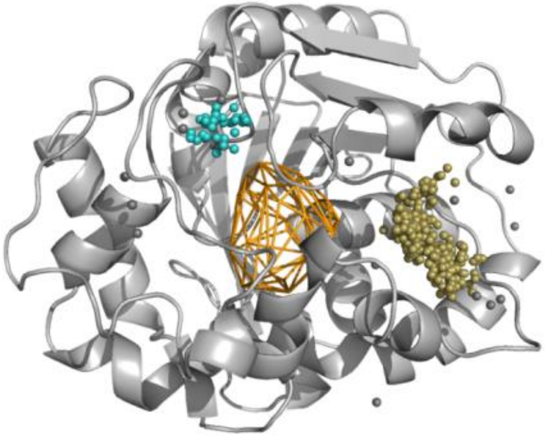


Fig. 6g. Results of clustering with `dbscan` method
`metric = euclidean`
`eps = 1.5`

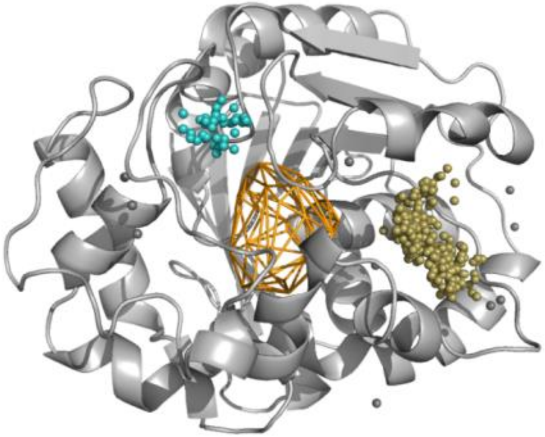


Fig. 6h. Results of clustering with `dbscan` method
`metric = euclidean`
`eps = 2.0`

AQUA-DUCT by TTG

TIPS & TRICKS - clustering

Affprop

Running `affprop` clustering method on default settings gave an excessive number of clusters with C1 and C2 divided into sections. For `affprop` option `damping` can be adjusted. Setting `damping = 0.9` gave results with cluster C1 divided into two sections, whereas C2 and C3 were clustered as one (Figure 7b). When `damping = 0.99` only two clusters were obtained (Figure 7c) with C1 divided into two sections and bigger cluster (cyan) including clusters C2, C3 and one part of C1.

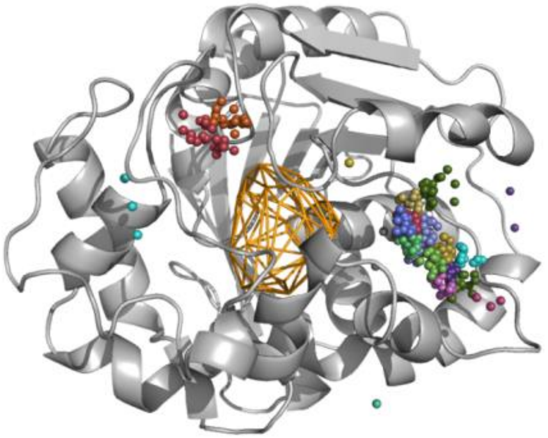


Fig. 7a. Results of clustering with `affprop` method on default parameters, i.e. `damping = 0.5`

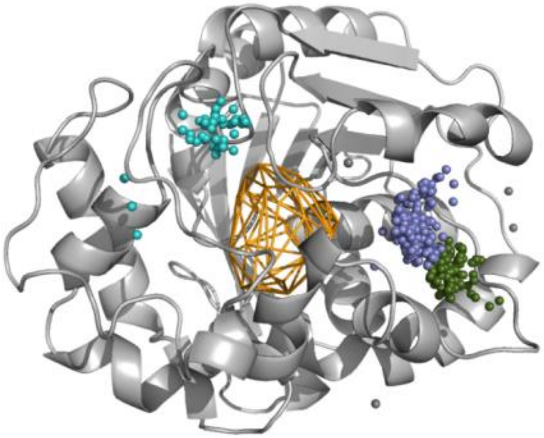


Fig. 7b. Results of clustering with `affprop` method `damping = 0.9`

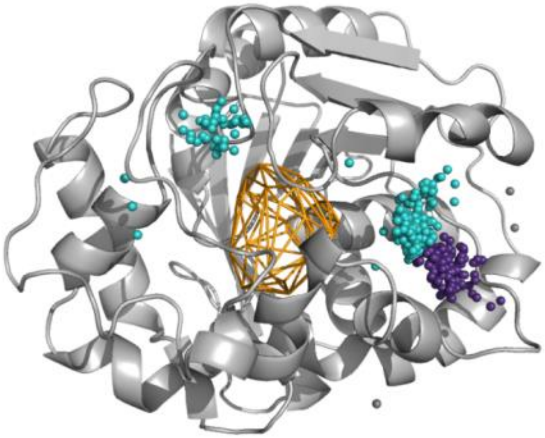


Fig. 7c. Results of clustering with `affprop` method `damping = 0.99`

TIPS & TRICKS - clustering

Our tips for clustering

- `barber` clustering method provides results consistent with protein surface hollows and is recommended as a first option to define the number of clusters,

- `meanshift` clustering method can provide similar results to `barber` clustering method and can be applied interchangeably with it,

- `birch` and `kmeans` methods are useful for division of large clusters into smaller well defined ones which can be useful for asymmetrical tunnels exits fragmentation, both methods can be forced to merge smaller clusters into larger one (`n_clusters` option) and these methods are especially recommended when the needed number of clusters is known,

- `affprop` clustering method can also be used to separate large clusters, as well as to merge smaller clusters into larger ones, however, it works differently to `birch` and `kmeans` methods, e.g. the asymmetric tunnels are divided in different ways,

- `dbscan` clustering method with optimal parameters provides results consistent with protein surface hollows and this method can be used to define the number of inlets marked as outliers at the edges of the clusters by setting the value of `eps`,

- Outliers are defined differently depending on the method of clustering, in `dbscan` number of inlets classified as outliers can be changed by setting `eps` option, in `birch` and `kmeans` all inlets that are not part of the major clusters, which number is set by user, are defined as outliers, and in `barber` and `meanshift` separate inlets are marked as 1- or 2-inlet clusters, however, in `meanshift` inlets at the edges of clusters can be marked as outliers by setting `cluster_all` parameter to `False`.

TIPS & TRICKS – clustering

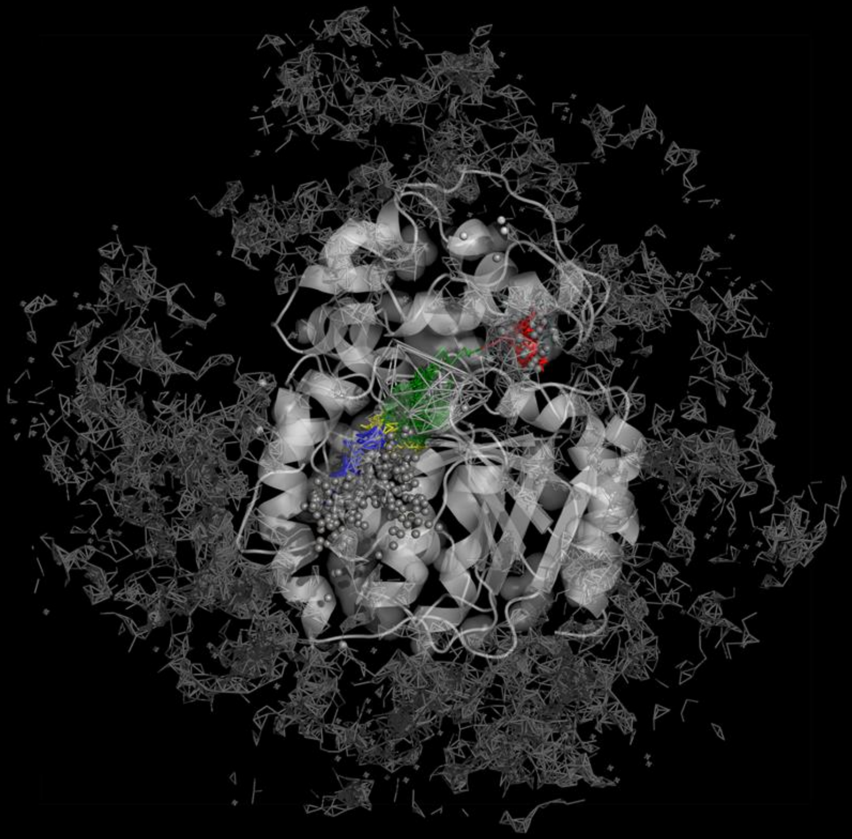
AQUA-DUCT

Instalation and guide:

http://www.aquaduct.pl

More info:

info@aquaduct.pl



TG

Tunneling Group
Biotechnology Centre
Silesian University of Technology
ul. Krzywoustego 8
44-100 Gliwice
POLAND