MADE Plugin Tutorial

March 2023

Contents

1 Introduction

The MADE plugin represents a PyMOL plugin implementing the MADE (MAcromolecular DEnsity and Structure Analysis) approach. The plugin is capable of identifying the locations of metal ions in the binding sites of proteins. Besides metal ions, the plugin is applicable to any species in protein records, it can identify the important species in a particular structure through superposition with homologous proteins and subsequent clustering. The MADE plugin is the evolution of our previous toolset, the ProBiS H2O (MD) approach for the identification of conserved waters in protein structures.

The MADE approach consists of the following steps, depicted in Figure 1. Beginning with a query protein structure we first identify a set of protein chains homologous to the query protein. Subsequently, we apply a superposition algorithm to superimpose the set of homologous protein structures upon the query protein. Keeping track of the positions of all the atoms from the query and homologous structures we apply a clustering algorithm to locate areas where a particular species is present in a similar location among many of the homologous structures. Identified clusters are indicators of species that are likely structurally or mechanistically important to the query protein. The MADE plugin can be applied for numerous purposes. It can predict locations where important cofactors bind in a protein structure and distinguish them from species that are present as impurities or part of the experimental procedure. The plugin can also be applied to structures modeled by AlphaFold and predict the positions of cofactors in those structures. In addition, the plugin can highlight important pieces of a binding site, such as conserved water or other molecules, and help prepare structures for molecular docking or molecular dynamics simulations.

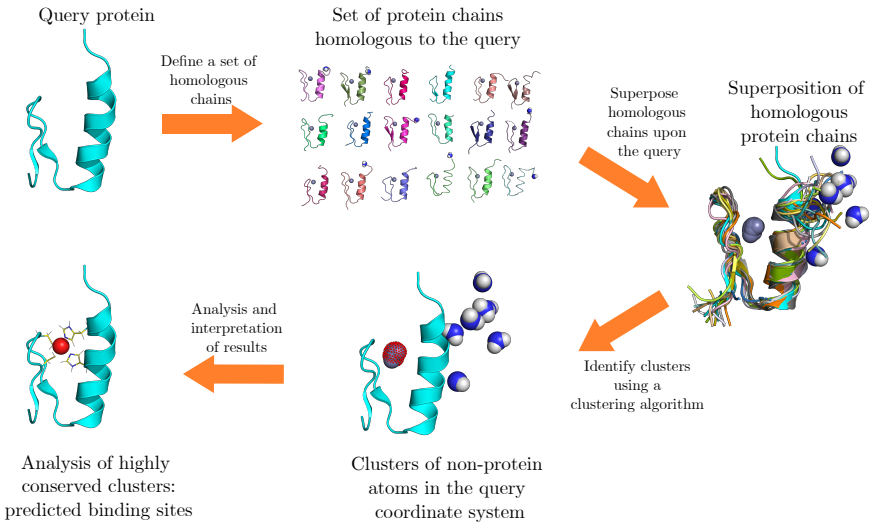

Figure 1: The MADE approach: first we identify a set of protein chains homologous to the query protein, these are subsequently superimposed onto the query protein, and dense areas of a particular species are located with a clustering algorithm. Clusters with many members likely represent species important to the query protein structure.

In the following sections we present instructions for installing and using the MADE plugin, which is a plugin for the PyMOL molecular visualisation software (v2.6, `https://pymol.org/`), it is written for python 3.

2 Installation

The MADE plugin is a PyMOL plugin, it requires PyMOL v2.x and python 3.x. It is compatible with the Windows and Linux operating systems. The MADE plugin also uses the scikit-learn python library, (`https://scikit-learn.org/`). You can install scikit-learn through pip by running:

```
pip install scikit-learn
```

Make sure you have downloaded the MADE plugin from the repository, this can be achieved with git by running:

```
git clone https://gitlab.com/Jukic/made_software.git
```

e Open PyMOL and in order to install a plugin go under **Plugin** → **Plugin Manager**, as shown in Figure 2A. This opens the **Plugin Manager**, navigate to the **Install New Plugin** tab, and select **Choose file...** as shown in Figure 2B. Navigate to where you have downloaded the plugin and select the **MADE_plugin.py** file.

You can now launch the MADE plugin from PyMOL by clicking on **Plugin** → **MADE plugin**. On the first startup, the plugin will ask for the location of the **.MADE_plugin** directory, which you have downloaded from the repository.

After inputting the location of the .MADE_plugin directory the MADE plugin will ask the user to setup the database. This is the folder in which the results of the calculations and other files will be

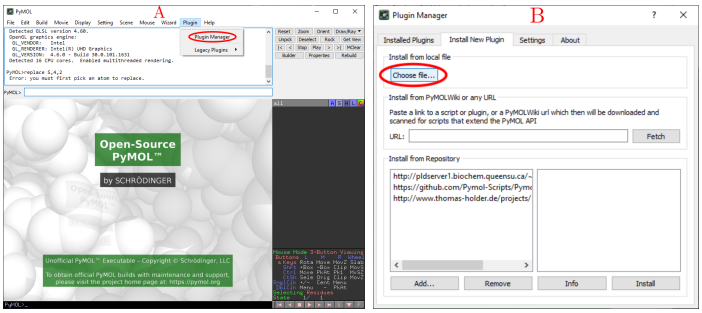

Figure 2: Installation of the MADE plugin, A) In PyMOL go under the Plugin tab and choose Plugin Manager, B) In the Plugin Manager go under Install New Plugin and select Choose file...

located, you can change its location with the **Local Database Directory** setting. After deciding on a location for the database directory (if you have changed it from the default location press the **SET** button to apply the change), click **SETUP DATABASE** (Figure 3). This will create the database folder and download the required sequence identity cluster files and superposition method executables.

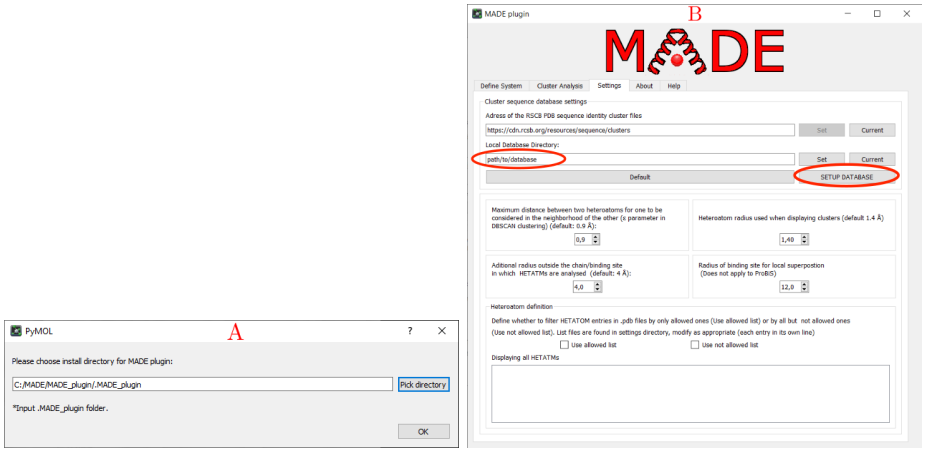

Figure 3: First Startup of the MADE plugin, A) Select the location of the .MADE_plugin directory you have downloaded from the repository, B) Click SETUP DATABASE to create the database folder and download the required sequence identity cluster files and superposition method executables.

3 Application

3.1 Typical workflow

3.1.1 Defining the Query and Homologous structures

After setting up the database, the MADE plugin in ready to use. The first step is defining the query protein structure and the set of homologous structures. Figure 4 presents the steps to define and

download the query and homologous protein structures:

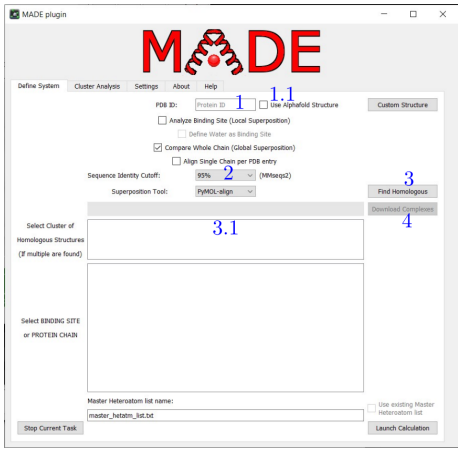

Figure 4: MADE Plugin typical workflow, input the **PDB ID** (**1**), select the **Sequence Identity Cutoff** (**2**) and press the **Find Homologous** (**3**) button. Subsequently press the **Download Complexes** (**4**) button.

- **1** The first step is defining the PDB ID of the query in the **PDB ID input field**. The plugin also supports computer-modeled structures (CMS) from AlphaFold2, to use enter the RCSB PDB CMS ID of the structure (e.g. AF_AFP00138F1, the AlphaFold naming scheme (AF-P00138-F1) is also supported (P00138 is the UniProt ID)) and check **Use Alphafold structure** (**1.1**)
- **2** Subsequently choose the **Sequence Identity Cutoff** which the homologous structures share with the query.
- **3** The **Find Homologous** button will identify complexes sharing at least the chosen sequence identity with the query structure. This is achieved using the pre- calculated sequence identity cluster files provided by the RCSB PDB. (`https://www.rcsb.org/docs/programmatic-access/file-download-services#sequence-clusters-data`). If the structure consists of multiple chains with different clusters of homologous structures, select the cluster to use in the **Select Cluster** field (**3.1**).
- **4** The **Download Complexes** button will download the PDB files of the query and homologous structures if they are not already present in the database directory.

3.1.2 Parameters of the calculation

After defining and downloading the query and homologous structures the subsequent step represents choosing the parameters of the calculation. This is achieved through the checkboxes under the PDB ID input field, as shown in Figure 5.

- **5** If **Analyse Binding Site (Local Superposition)** is checked the plugin will superpose the homologous structures around a binding site in the query (the binding sites are listed as eg. ZN.201.A representing the residue_name.residue_number.chain_name).

- **5.1** Checking **Define Water as Binding Site** will enable the selection of water molecules as binding sites.
- **6** If **Compare Whole Chain (Global Superposition)** is checked the plugin will superpose the homologous structures around a whole protein chain of the query.
- **7** If **Align Single Chain per PDB entry** is checked the plugin will use one homologous chain from each PDB structure (structures often have multiple identical chains).

A crucial decision is whether to perform the superposition globally, on a whole protein chain, or around a local binding site. Checking **Compare Whole Chain** will perform the superposition on a whole protein chain. On the other hand, local superposition, applied by ticking **Analyse Binding Site**, will superpose the structures around a binding site (locations of non-protein species in the structure), and locate only clusters near the binding site. Cluster conservation is often higher with local superposition and the results are easier to analyse, at the cost of requiring additional information about which binding site to choose. Thus it is often good practice to first run a whole chain superposition to identify an interesting binding site, followed by local superposition around the binding site.

After deciding on the parameters of the calculation with the checkboxes, the next steps involve deciding which binding site/protein chain we perform the superposition around and which method we use (Figure 5):

- **8** The MADE plugin will perform the superposition using the **Superposition Tool** selected from the dropdown menu.
- **9** When the query structure is downloaded, the protein chains or binding sites (depending on the **5** and **6** check-marks) will be listed in the **Select BINDING SITE/CHAIN** list. Clicking on a protein chains/binding site will select it to perform the superposition around.
- **10** The calculation will be launched by pressing the **Launch Calculation** button.

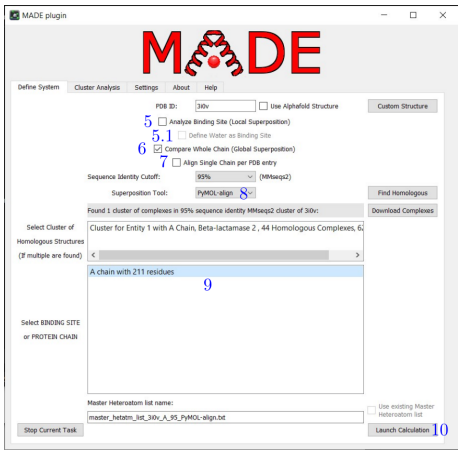

Figure 5: MADE Plugin typical workflow, Select the parameters of the analysis, check **Compare Whole Chain (Global Superposition)** (**6**) to run superposition on the whole protein chain, or check **Analyse Binding Site (Local Superposition)** (**5**) to focus the superposition around a binding site in the protein. When the query structure is downloaded select a protein chain or binding site from the **Select BINDING SITE or PROTEIN CHAIN** list (**9**). Select a **Superposition Tool** from the dropdown menu (**8**). Start the calculation with the **Launch Calculation** (**10**) button.

3.1.3 Results of the calculation

After all the parameters for the calculation have been selected, the calculation is launched with the **Launch Calculation** button. When the calculation is finished the user is moved to the **Cluster Analysis** tab. The MADE plugin lists identified clusters of heteroatoms (entries marked as HETATM in .pdb files, non-protein and other non standard species) on the Cluster Analysis tab, it ranks and colour codes the clusters according to their conservation. High values of cluster conservation, which the MADE plugin highlights with a red colour, represent clusters likely important the query protein, cluster conservation is calculated as:

$$\text{conservation} = \frac{\text{number of heteroatoms in a cluster}}{\text{total number of superimposed protein structures}} \quad (1)$$

The Cluster Analysis tab is presented in Figure 6.

- **1**: The **Info Panel**, contains information about the parameters of the calculation, as well as the identities and chemical formulae of heteroatoms found in the analysed system (taken from the HETNAM and FORMUL entries of the .pdf files).
- **2**: The **Heteroatom types** field, lists the total number and maximum conservation of clusters for all present heteroatom residue names. Results are sorted and colour coded by conservation. Selecting an entry filters the **Calculated cluster** field (**3**) by listing only clusters matching the selected residue name.
- **3**: The **Calculated cluster** field, lists the different clusters of heteroatoms identified. The clusters are listed by the number of clusters, the minimum number of atoms in the clusters, the minimum cluster conservation and the type of the heteroatom (RESIDUE_NAME-ATOM_NAME).

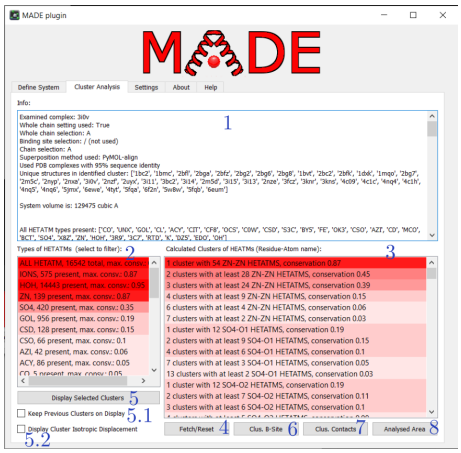

Figure 6: MADE Plugin typical workflow, the Cluster Analysis tab. The **Info Panel** (**1**), The **Heteroatom types** field (**2**), The **Calculated cluster** field (**3**), The **Fetch/Reset** button (**4**), The **Display Selected Clusters** button (**5**), The **Clus. B-Site** button (**6**), The **Clus. Contacts** button (**7**), The **Analysed Area** button (**8**).

- **4**: The **Fetch/Reset** button. Clears the PyMOL viewer and fetches the query structure. An example of its use is shown in Figure 7.
- **5**: The **Display Selected Clusters** button. Displays the cluster(s) selected from the **Calculated cluster** field (**4**) in the PyMOL viewer. Clusters are displayed as spheres with a certain radius (configurable in settings), colour coded matching the cluster conservation. The spheres are located at the average position of cluster members. An example of its use is shown in Figure 8. Pressing the **Display Selected Clusters** button will also produce a report file located in the Reports sub-directory of the database directory.
- **5.1**: The **Display Cluster Isotropic Displacement** check, if checked the **Display Selected Clusters** button (**6**) will also display an isotropic displacement of all the members of the displayed clusters. An example of its use is shown in Figure 9.
- **5.2**: The **Keep Previous Clusters on Display** check, if checked the **Display Selected Clusters** button (**6**) will keep all currently displayed clusters, otherwise it will display only the selected ones. An example of its use is shown in Figure 10.
- **6**: The **Clus. B-Site** button, will display the amino acid residues in the vicinity of displayed clusters in the PyMOL viewer. An example of its use is shown in Figure 11.
- **7**: The **Clus. Contacts** button, will display the distances between the displayed clusters and the closest amino acid residues in their vicinity. An example of its use is shown in Figure 12.
- **8**: The **Analysed Area** button. Displays a box around the area in which the plugin searched for clusters. The area can be increased in the Settings tab.

The **Calculated Clusters** field represents the main results of the MADE plugin, it lists detected clusters of different species. The clusters are presented with different minimum numbers of members, which the MADE plugin calculates with the iterative increase of the number of neighbours parameter

n of the clustering algorithm, 3D-DBSCAN. For example, the plugin may list **1 cluster with 70 ZN-ZN HETATMS, conservation 0.95** (Zinc ions, residue name ZN, atom name ZN in the PDB files) and **2 clusters with at least 62 ZN-ZN HETATMS, conservation 0.84**. This means the plugin located 1 cluster in which 95% of the superimposed protein structures exhibit a Zn ion in a similar position, and 2 clusters where at least 84% of the superimposed structures have a Zn ion in a similar position, including the previous 95% preserved cluster.

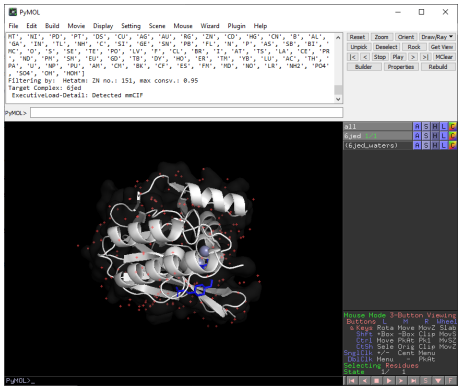

Figure 7: MADE Plugin typical workflow, an example of application of the **Fetch/Reset** Button.

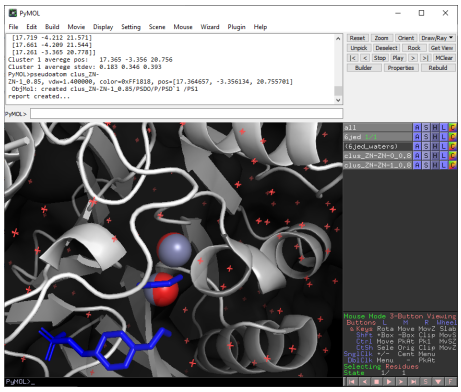

Figure 8: MADE Plugin typical workflow, an example of application of the **Display Selected Clusters** Button.

Figure 9: MADE Plugin typical workflow, an example of application of the **Display Selected Clusters** Button with **Display Cluster Isotropic Displacement** checked.

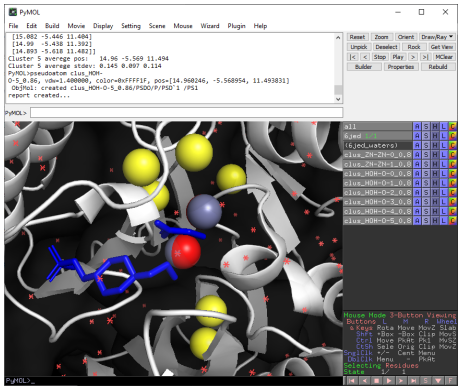

Figure 10: MADE Plugin typical workflow, an example of application of the **Display Selected Clusters** Button with **Keep Previous Clusters on Display** checked.

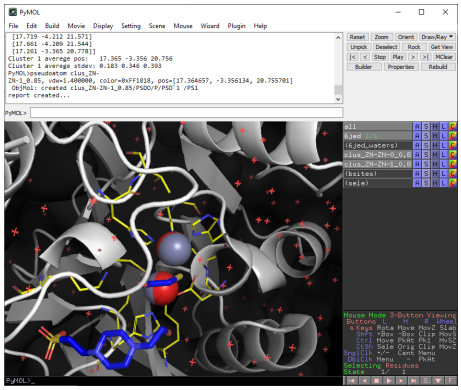

Figure 11: MADE Plugin typical workflow, an example of application of the **Clus. B-Site** Button.

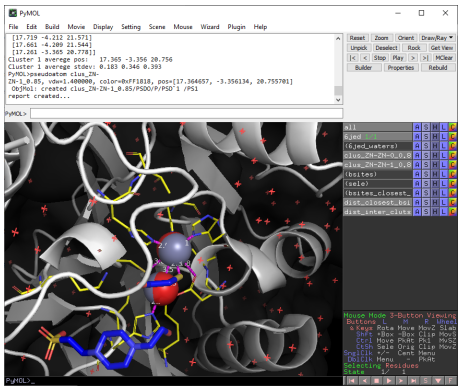

Figure 12: MADE Plugin typical workflow, an example of application of the **Clus. Contacts** Button.

3.2 Settings and advanced application

3.2.1 Settings

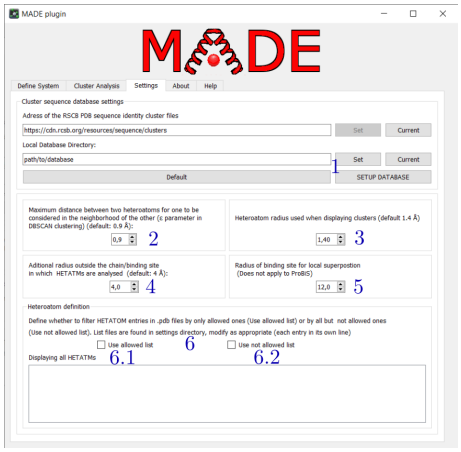

Figure 13: MADE Plugin settings tab, the tools for setting the database directory (**1**), the ϵ parameter of 3D-DBSCAN (**2**), displayed cluster heteroatom radius (**3**), the additional radius outside the selected protein chain/binding site in which the plugin analyses HETATMs (**4**), the radius of the binding site (**5**), tools for filtering unwanted heteroatom types (**6**).

- **1**: The tools of changing the database directory and setting up the database of cluster files and superposition algorithms.
- **2**: Setting for the ϵ parameter of 3D-DBSCAN.
- **3**: Setting for the heteroatom radius applied when displaying clusters.
- **4**: Setting for the additional radius outside the selected protein chain/binding site in which the plugin analyses HETATMs.

- **5**: Setting for the radius of the binding site when analysing a binding site, this applies to all superposition methods except ProBiS, which possesses an inbuilt local binding site superposition mode.
- **6**: Tools for filtering unwanted listed heteroatom types.
- **6.1**: The allowed list checkbox, if checked will only list heteroatoms matching an entry in the allowed list (.MADE_plugin/settings/List_allowed_hetams.txt). Can be modified to suit the users' needs.
- **6.2**: The not allowed list checkbox, if checked not display heteroatoms matching an entry in the not allowed list (.MADE_plugin/settings/List_not_allowed_hetams.txt). Can be modified to suit the users' needs.

3.2.2 Advanced application

Master Heteroatom lists

After protein superposition the MADE plugin writes the coordinates of all HETATM entries in the common coordinate system of the query into a single file, dubbed the Master Heteroatom List (MHL). These are located in the database directory, in the MHL sub-directory. The plugin supports the option to use an existing MHL to skip the superposition step and only run the clustering of heteroatoms. This can be achieved by entering the desired MHL name and checking the **Use existing Master Heteroatom list** in the **Define System** tab before running the calculation. This will skip the more time consuming process of protein superposition, make sure the settings are identical as when the MHL was first calculated, there are no checks for this in the code.

Report Files

When the **Display Selected Clusters** button is pressed a report file is produced in the Reports sub-directory of the database folder. The report file contains information about the parameters of the calculation as well as detailed information about the members of the displayed clusters.

Custom PDB file and homologous structures

To use a custom PDB file press the **Custom Structure** button on in the **Define System** tab. When using a custom structure the user is unable to use the sequence identity clusters from the PDB, a custom cluster must be provided. The custom cluster of complexes is provided in a file called clusters_custom_structure.txt. The file has to contain a cluster of structures per line, in the format of (Filename without .pdb)_(Protein chain to analyse) separated by white-space characters. The line must include the filename of the custom query file, and all the structures must be present in the database directory of the MADE plugin. The custom files feature is completely separate from RCSB PDB. The files for the query and homologous structures can be anything the user wants to analyse, e.g. snapshots from a molecular dynamics simulation.

To use a custom set of homologous structures (from the PDB database), select **Custom Cluster** in the **Sequence identify cutoff** dropdown menu. To define the custom cluster edit the clusters_custom.txt file in the database folder. Add a line containing the query PDB ID to the file, in the same format as the clusters-by-entity-.txt files from the RCSB PDB, PDB-ID_PDB-ENTITY entries separated by white-space characters. The line has to contain the PDB ID of the query.