PCN User Guide

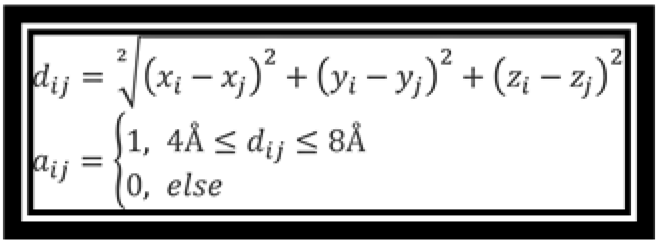
Protein Contact Networks Miner: A tool for the Analysis of Protein Contact Networks

1 - Description

Protein Contact Networks Miner is a tool with a Graphic User Interface designed to annotate allosteric domains of a protein based on its representation through an unweighted graph.. This graph is also known as Protein Contact Network (PCN).

A PCN is an unweighted graph where: (i) nodes are the amino acids of the protein and (ii) there exists an edge that connects two nodes *i* and *j* only if the euclidean distance between them has a value included in the interval: 4 Angstrom (threshold for only non covalent interactions) and 8 Angstrom (threshold for only significant interactions). The distance between two aminoacids *i* and *j* is approximated with the distance between the amino acids alpha carbons.

Given the distance between two nodes *i* and *j,* the user can only modify covalent (min) and significant (max) threshold distance for PCN construction.



$$d_{ij} = \sqrt[2]{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

$$a_{ij} = \begin{cases} 1, & 4\text{Å} \le d_{ij} \le 8\text{Å} \\ 0, & else \end{cases}$$

Where x,y and z represent the coordinates in the 3D space of amino acids and aij represents the considered interaction among two aminoacids depending on their distance as reported above.

PCN global descriptors (like graph diameter) or local descriptors (like node centrality measures) are useful to model and analyze protein functions.

2 - PCN Miner tool short description

PCN Miner allows to:

(i) identify the putative allosteric paths and regions in protein structures (thus to help the design of allosteric drugs);

(ii) define hypothesis and make tests about the functional effects of mutations (example variants of SARS CoV-2 Spike Protein);

(iii) recognise functional domains (communities or clusters) in proteins.

It expones three groups of algorithms:

1. Centrality Analysis: it implements main centrality algorithms
2. spectral clustering: it extracts clusters from a graph with a clustering approach based on the Laplacian matrix eigenvectors;
3. embedding+clustering: it uses one of the embedding algorithm in the GEM library and then apply clustering;
4. community detection: it uses one of the community detection algorithms in the cdlib library.

Also, PCN-Miner includes the following implemented algorithms:

1. Spectral Clustering: Both Hard (K-Means) and Soft (Fuzzy C-Means) clustering approach used on the eigenvectors of the Laplacian matrix (both normalized or unnormalized form). The Shi Malik spectral clustering approach is also supported, to resolve the generalized eigenvalues problem;
2. Embedding + Clustering: Node2vec, HOPE, Laplacian-eigenmaps embedding followed by a supported spectral clustering algorithm;
3. Community Detection: Louvain, Leiden, Walktrap, Infomap, Asyn FluidC, Greedy Modularity, Spinglass;
4. Centrality Measures: Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, Degree Centrality.

3 - PCN Miner installation guide

It can be installed and used by using the setup files on github and using a command line in the operating system.

-Microsoft Windows O.S. :

```
git clone https://github.com/hguzzi/ProteinContactNetworks.git
setupWindows.bat
```

-Linux-MACOSX O.S. :

```
git clone https://github.com/hguzzi/ProteinContactNetworks.git
source setupLinux-MacOSX.sh
```

It is possible to install the tool by using Python Package Index (the pip).

Requirements : pymol and Gem library

This can be performed by using the anaconda prompt and typing the following command:

```
conda create -n PCN python=3.8.3
conda activate PCN
conda install -c schrodinger pymol-bundle
```

GEM library can be installed by using pip+git and this library using TESTPYPI as:

```
pip install git+https://github.com/palash1992/GEM.git
pip install --extra-index-url https://pypi.org/simple -i
https://test.pypi.org/simple/PCN-Miner
```

After installing these libraries, you may install PCN-Miner:

```
pip install git+https://github.com/hguzzi/ProteinContactNetworks.git
```

4 - Examples, use case and how to use of PCN Miner

The command line tool version can be used as follows:

```
conda activate PCN

cd pcn

cd tools

python pcn_main.py
```

The GUI tool version can be used by running is as:

```
conda activate PCN

cd pcn

cd tools

python pcn_gui_main.py
```

PCN-Miner user GUI application and use cases

GUI interface instances are reported in the following case study examples.

Case Study 1.: Centrality Analysis.

After launching the GUI of PCN-Miner, the user may insert the PDB identifier into the textbar on the upper part of the GUI. For instance, let us insert the code 6nxc that stands for 6nxc.pdb protein file.

For the calculation of the PCN, the user has to set a lower and upper bound. The lower bound is used to discard trivial connections due to closeness along the backbone, while the upper bound is defined by considering the peptide bonds geometry.

Description: Calculation of PCN from PDB file

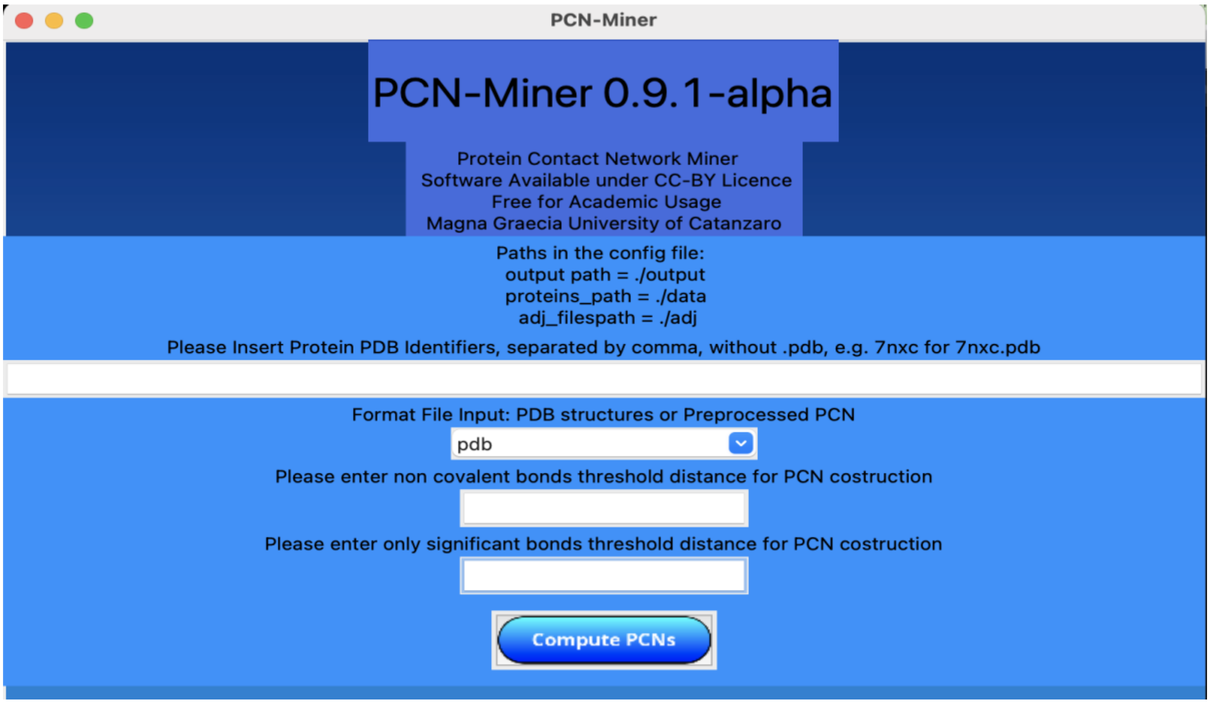
Use the entry PDB code: *6nxc*

Analysis of PCN: *Method: Centrality Analysis*

Algorithm: *Betweenness Centrality Analysis*

Visualization: *Pymol Embedded*

Step 1: Loading the PDB File into PCN-Miner



PCN-Miner

PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

Paths in the config file:
output path = ./output
proteins_path = ./data
adj_filespath = ./adj

Please Insert Protein PDB Identifiers, separated by comma, without .pdb, e.g. 7nxc for 7nxc.pdb

Format File Input: PDB structures or Preprocessed PCN

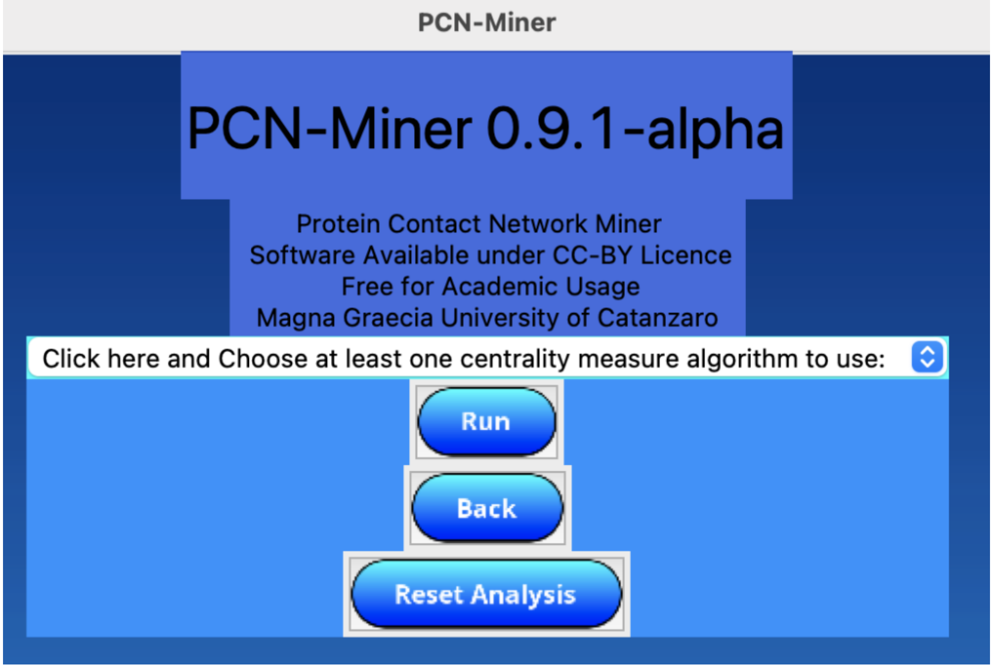
pdb

Please enter non covalent bonds threshold distance for PCN costruction

Please enter only significant bonds threshold distance for PCN costruction

Step 2 Choice of the desired centrality measure.

We inserted following centrality measures: Betweenness Centrality, Closeness Centrality, Eigenvector Centrality,



PCN-Miner

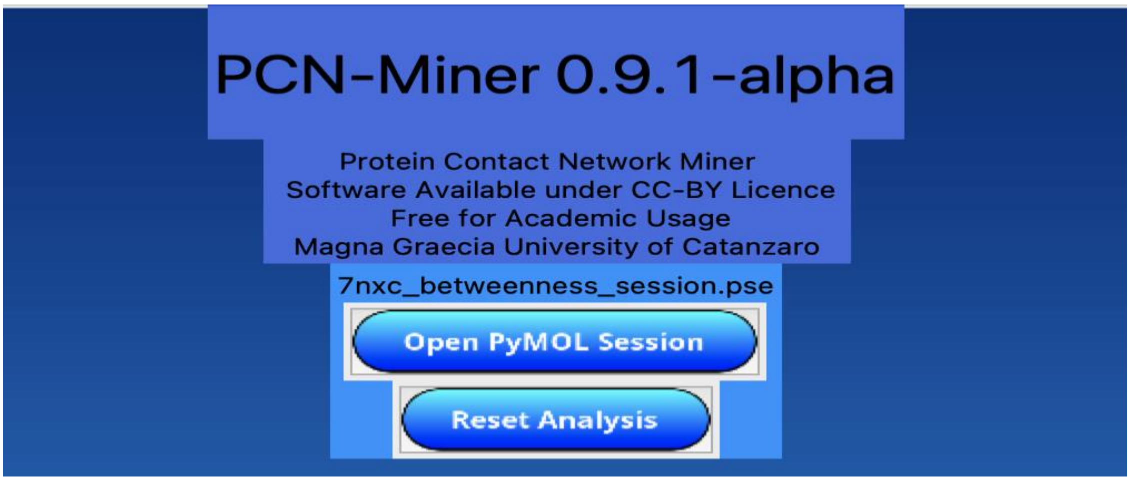
PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

Click here and Choose at least one centrality measure algorithm to use:
Run
Back
Reset Analysis

Step 3 Loading and Visualization of the results through Pymol.

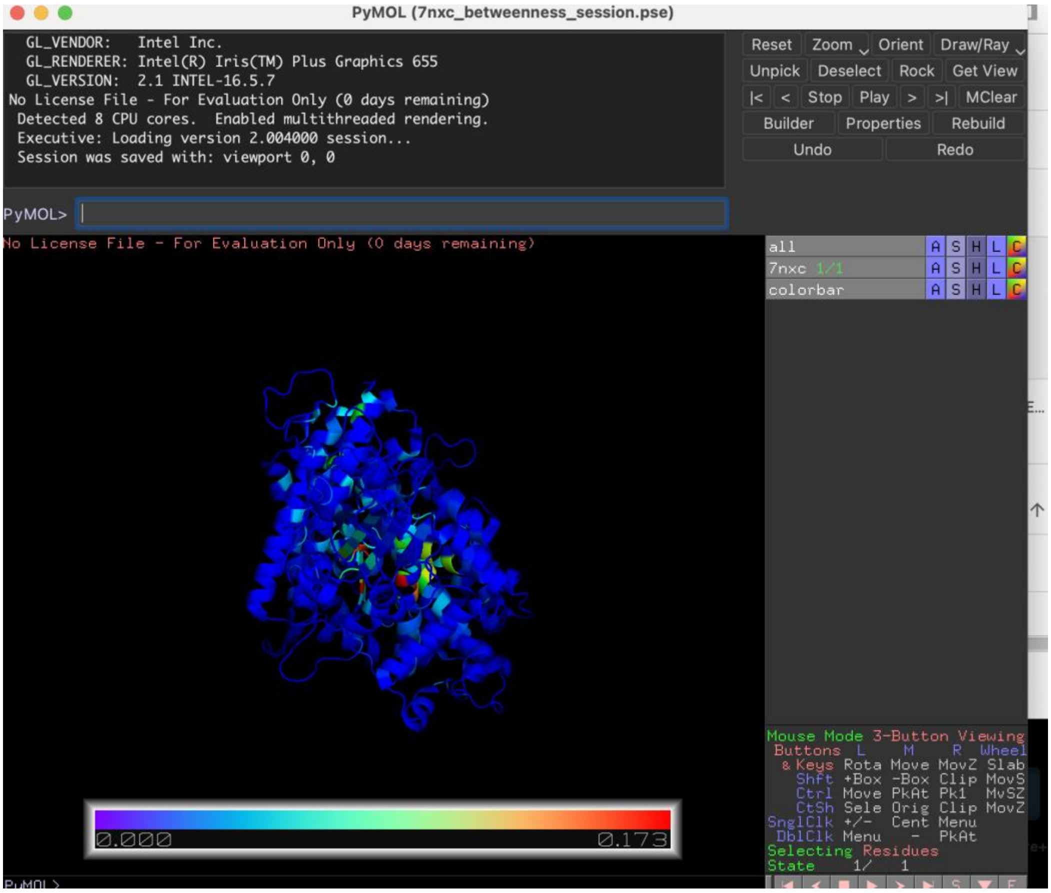
After the calculation of the centrality values, the user may automatically visualize these coefficients in Pymol.



PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

7nxc_betweenness_session.pse



PyMOL (7nxc_betweenness_session.pse)

```
  GL_VENDOR:   Intel Inc.
  GL_RENDERER: Intel(R) Iris(TM) Plus Graphics 655
  GL_VERSION:  2.1 INTEL-16.5.7
No License File - For Evaluation Only (0 days remaining)
 Detected 8 CPU cores.  Enabled multithreaded rendering.
 Executive: Loading version 2.004000 session...
 Session was saved with: viewport 0, 0
PyMOL>
No License File - For Evaluation Only (0 days remaining)
```

Case Study 2: Spectral Clustering.

Description: Analysis of Modularity through Spectral Clustering

Use the entry PDB code: *6VXX (Coronavirus Spike Protein)*

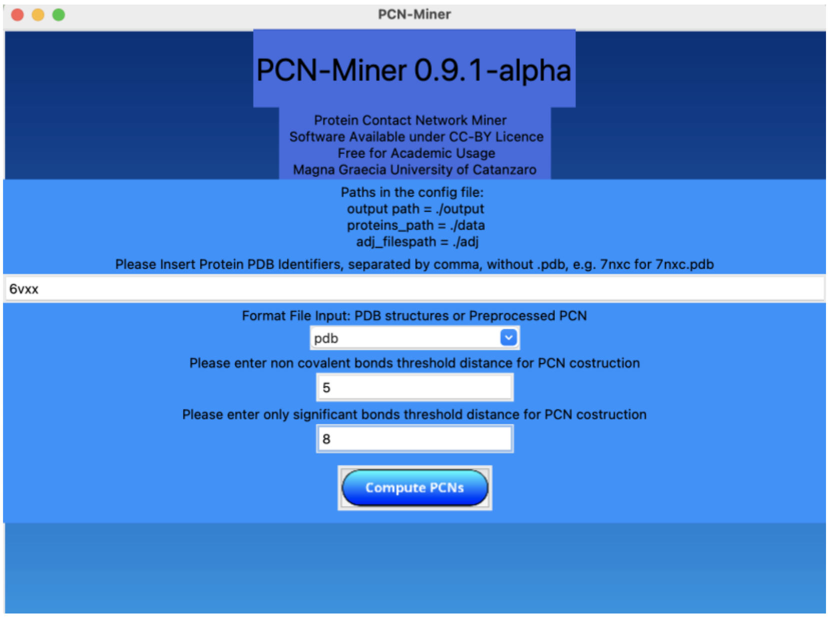
Analysis of PCN: *Method: Spectral Clustering*

Parameters of Clustering: Number of Clusters 5

Algorithm:

Visualization: *Pymol Embedded*

Step 1: Loading of 6Vxx



PCN-Miner

PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

Paths in the config file:
output path = ./output
proteins_path = ./data
adj_filespath = ./adj

Please Insert Protein PDB Identifiers, separated by comma, without .pdb, e.g. 7nxc for 7nxc.pdb

6vxx

Format File Input: PDB structures or Preprocessed PCN

pdb

Please enter non covalent bonds threshold distance for PCN costruction

5

Please enter only significant bonds threshold distance for PCN costruction

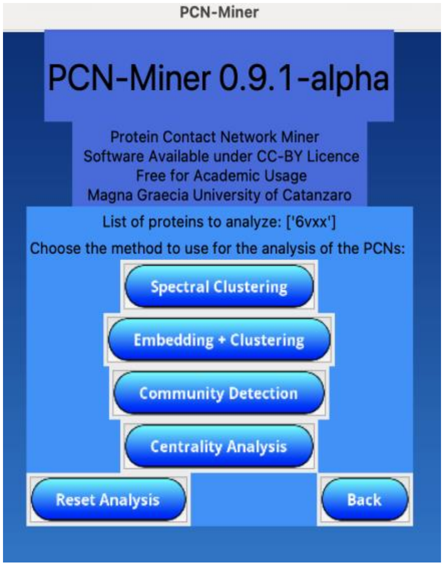
8

Step 2: Selection of Spectral Clustering Algorithm

Number of Clusters 5

Algorithm*: Un-normalised Hard Spectral Clustering*

Clustering has only one parameter: the number of desired clusters. User may select this number equal to the number of modules that should be evidenced. More details may be found on Di Paola et al 2020.

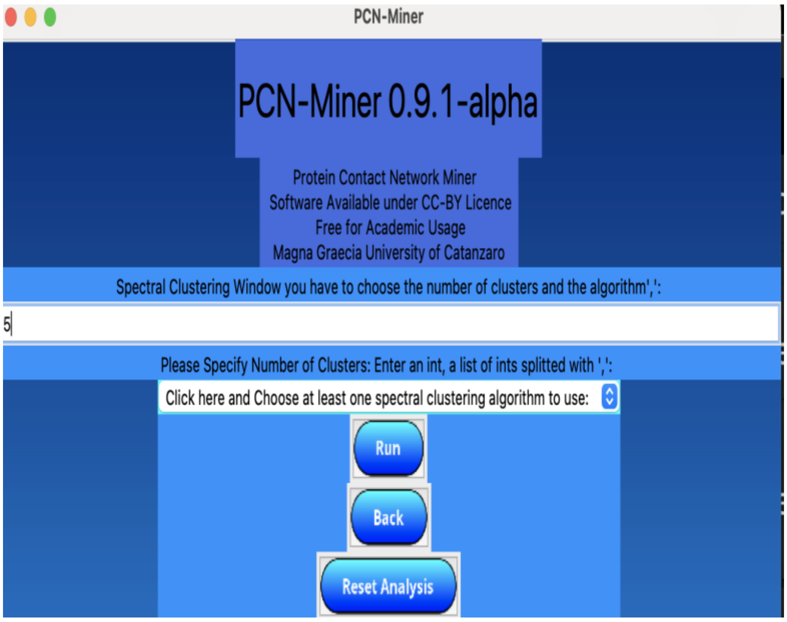


PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

List of proteins to analyze: ['6vxx']

Choose the method to use for the analysis of the PCNs:



PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

Spectral Clustering Window you have to choose the number of clusters and the algorithm',':

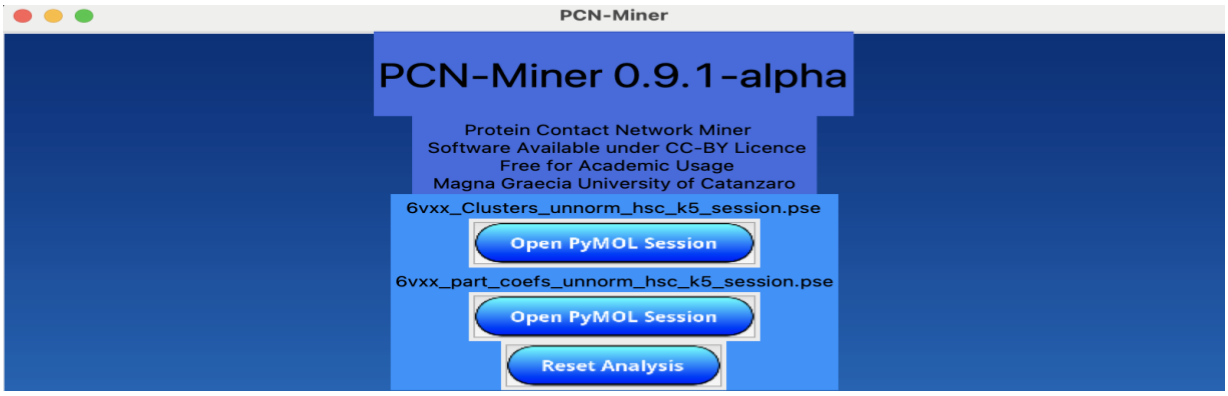
5

Please Specify Number of Clusters: Enter an int, a list of ints splitted with ',':

Click here and Choose at least one spectral clustering algorithm to use:

Step 4: Selection of Visualization:

After that clusters have been obtained, user may visualize clusters mapped into the initial structure of the protein. Alternatively he/she may also visualize the coefficient of participation.



PCN-Miner

PCN-Miner 0.9.1-alpha

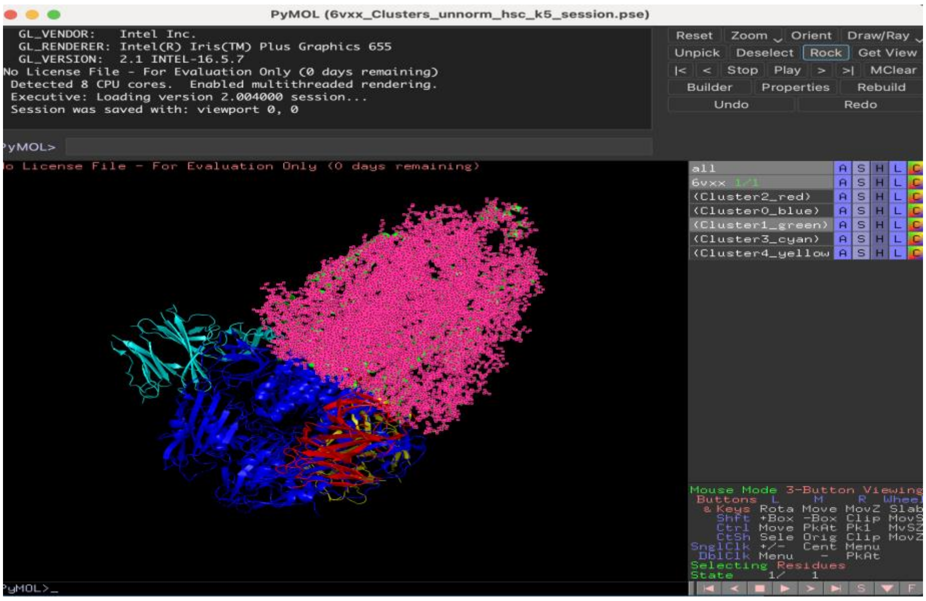
Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

6vxx_Clusters_unnorm_hsc_k5_session.pse

6vxx_part_coefs_unnorm_hsc_k5_session.pse

Step 5 Visualization

Visualization of Clusters



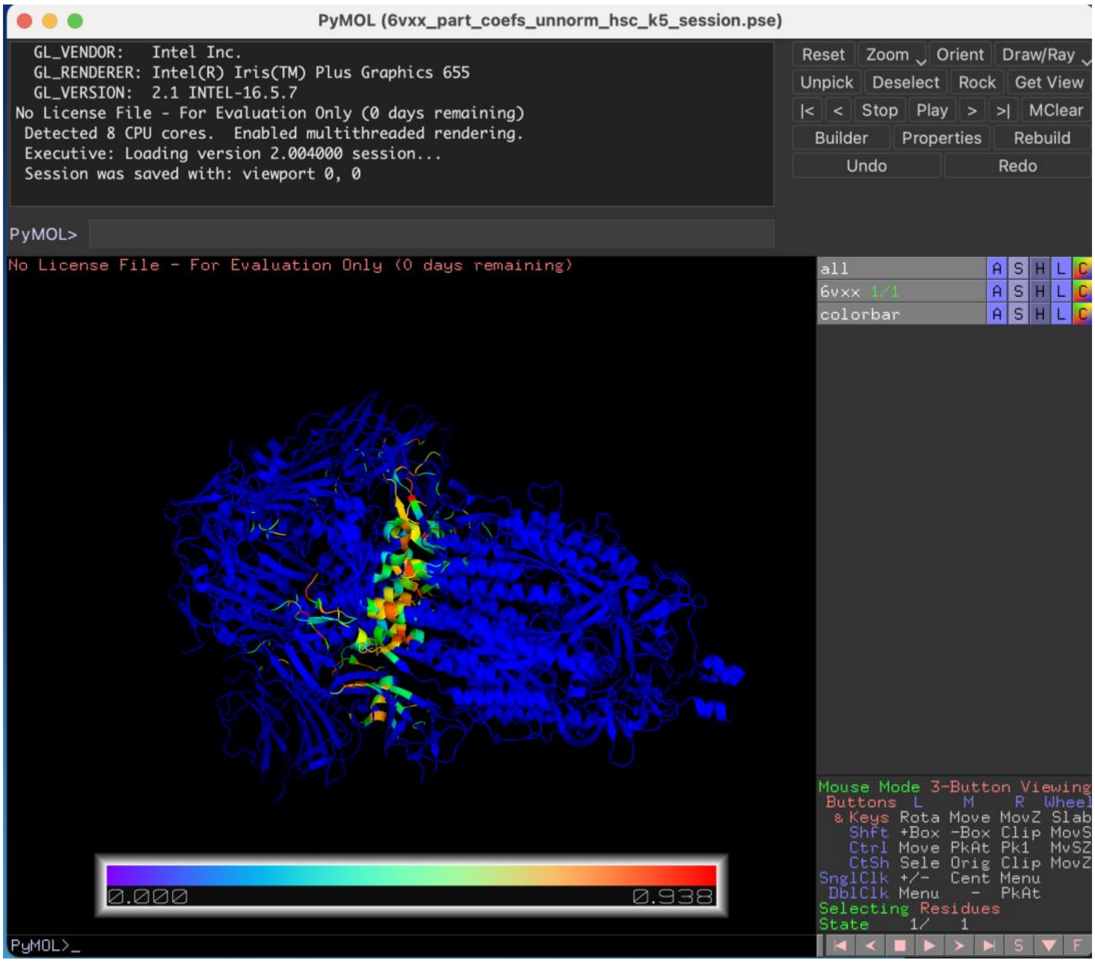
PyMOL (6vxx_Clusters_unnorm_hsc_k5_session.pse)

```
  GL_VENDOR:   Intel Inc.
  GL_RENDERER: Intel(R) Iris(TM) Plus Graphics 655
  GL_VERSION:  2.1 INTEL-16.5.7
No License File - For Evaluation Only (0 days remaining)
 Detected 8 CPU cores.  Enabled multithreaded rendering.
 Executive: Loading version 2.004000 session...
 Session was saved with: viewport 0, 0
PyMOL>
No License File - For Evaluation Only (0 days remaining)
```

Clusters are evidenced on the left menu of Pymol. In this case we have 5 clusters.

Visualization of Participation Coefficients.

Alternatively user may plot the participation coeefficients.



PyMOL (6vxx_part_coefs_unnorm_hsc_k5_session.pse)

```
  GL_VENDOR:   Intel Inc.
  GL_RENDERER: Intel(R) Iris(TM) Plus Graphics 655
  GL_VERSION:  2.1 INTEL-16.5.7
No License File - For Evaluation Only (0 days remaining)
 Detected 8 CPU cores.  Enabled multithreaded rendering.
 Executive: Loading version 2.004000 session...
 Session was saved with: viewport 0, 0
PyMOL>
```

```
No License File - For Evaluation Only (0 days remaining)
```

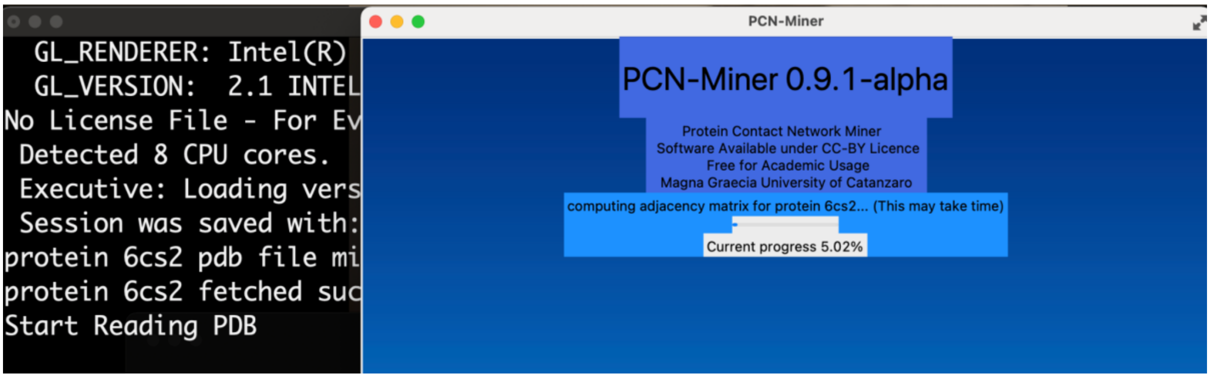
```
PyMOL>_
```

Case Study 3: Discovery of a Putative Allosteric Site

We show how to discover a putative allosteric site, as shown in Di Paola et al 2020.

Step 1

Load the SARS-CoV S/ACE2 complex (PDB code 6cs2) Protein into PCN-Miner



```
  GL_RENDERER: Intel(R)
  GL_VERSION:  2.1 INTEL
No License File - For Ev
 Detected 8 CPU cores.
 Executive: Loading vers
 Session was saved with:
protein 6cs2 pdb file mi
protein 6cs2 fetched suc
Start Reading PDB
```

PCN-Miner

PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

computing adjacency matrix for protein 6cs2... (This may take time)

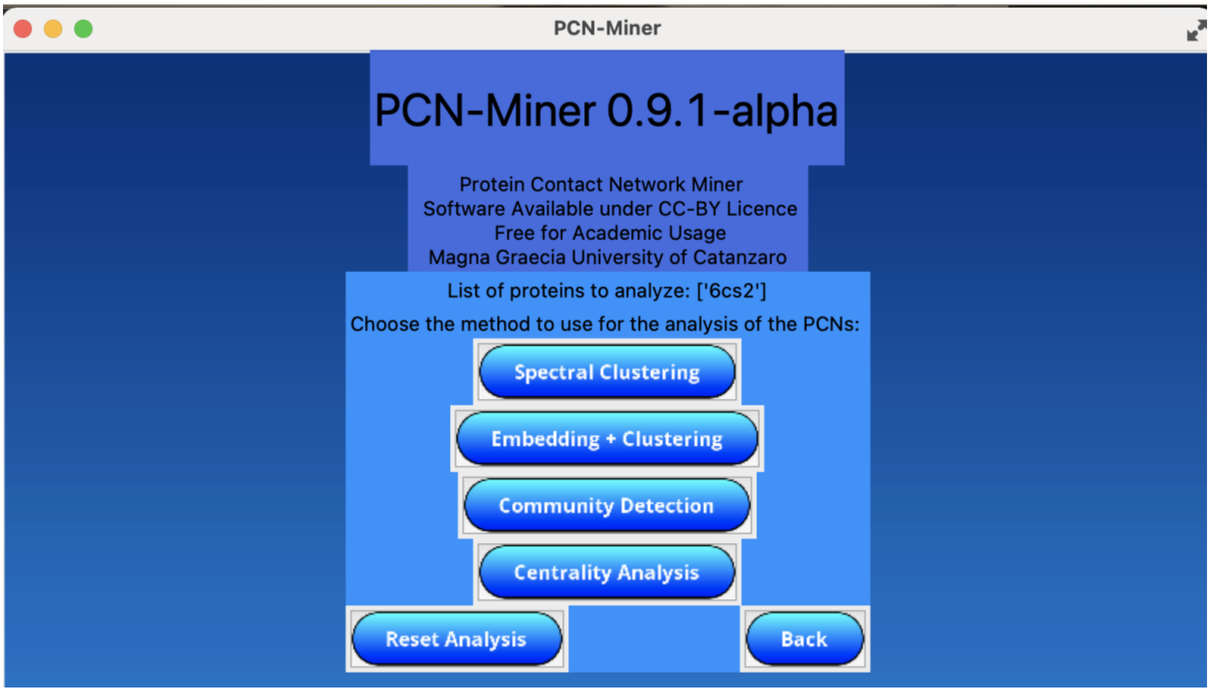
Current progress 5.02%

PCN-Miner is used to gather the structural information of the PDB files into a PCN. Network nodes are the amino acid residues represented by α-carbons, while links between nodes (residues) exist if the mutual distance of the residues (centered on α-carbons) were in the range between 5 and 8 Å.

To detect allosteric sites and functional regions activating upon binding we follow the approach described in Di Paola et al J Prot Res 2020 that demonstrated that network clusters (group of residues) correspond to functional regions in the protein. This method is based on network spectral clustering of the PCN.

Step 2: Spectral Clustering of the PCN

Parameters: Cluster =2
Algorithm SKL Spectral Clustering

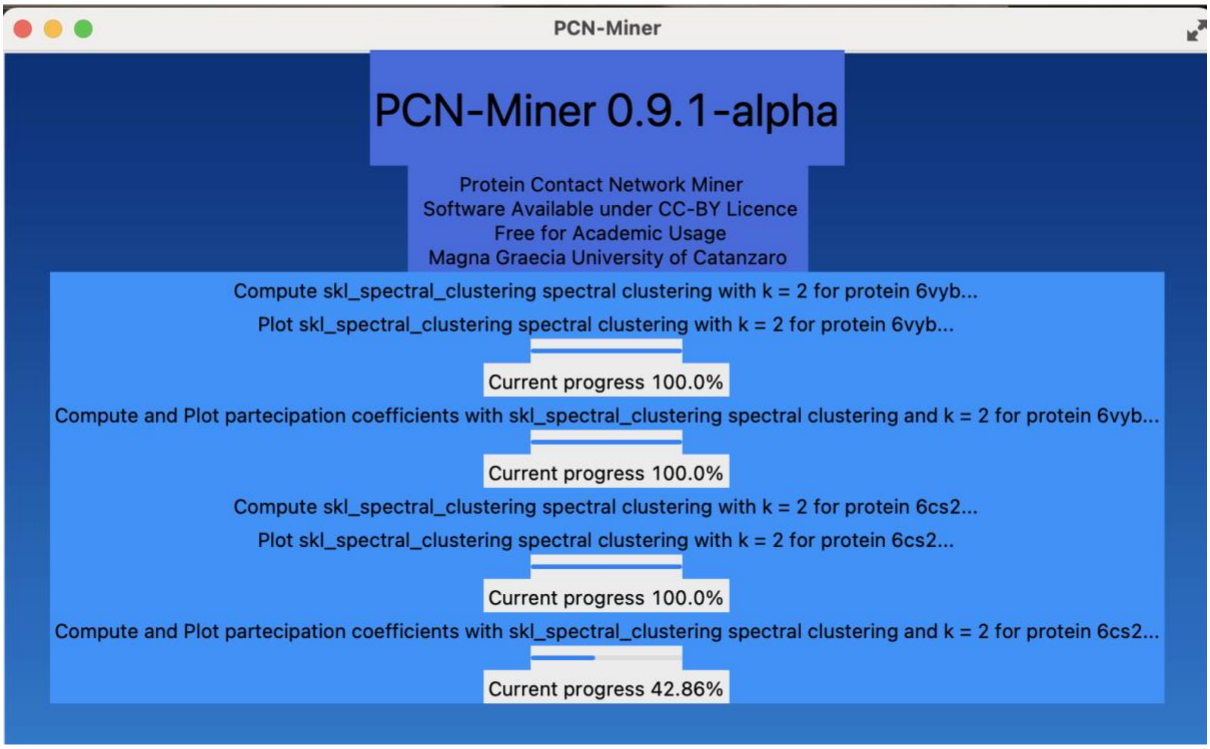


PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

List of proteins to analyze: ['6cs2']

Choose the method to use for the analysis of the PCNs:



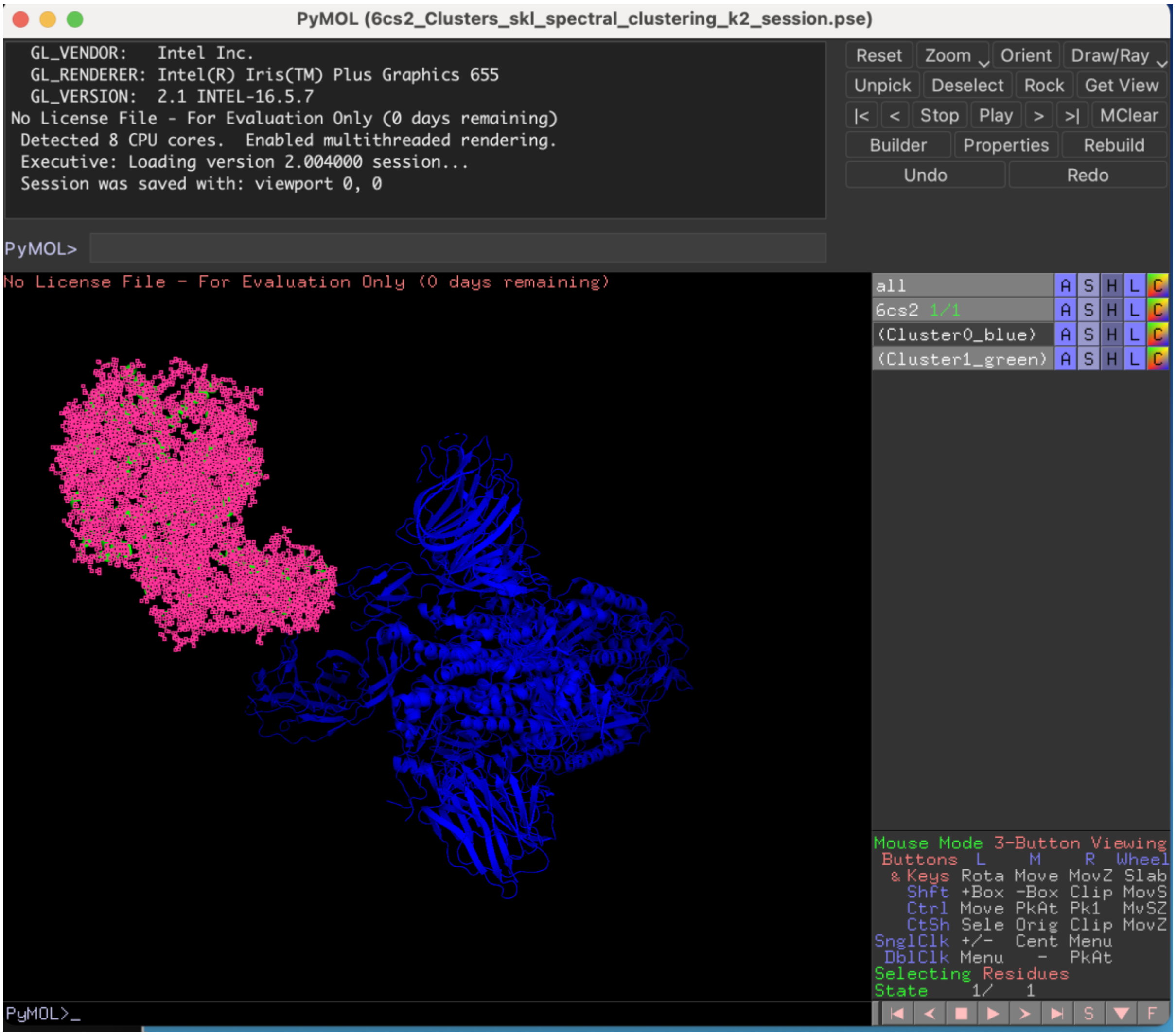
PCN-Miner 0.9.1-alpha

Protein Contact Network Miner
Software Available under CC-BY Licence
Free for Academic Usage
Magna Graecia University of Catanzaro

```
Compute skl_spectral_clustering spectral clustering with k = 2 for protein 6vyb...
Plot skl_spectral_clustering spectral clustering with k = 2 for protein 6vyb...
Current progress 100.0%
Compute and Plot partecipation coefficients with skl_spectral_clustering spectral clustering and k = 2 for protein 6vyb...
Current progress 100.0%
Compute skl_spectral_clustering spectral clustering with k = 2 for protein 6cs2...
Plot skl_spectral_clustering spectral clustering with k = 2 for protein 6cs2...
Current progress 100.0%
Compute and Plot partecipation coefficients with skl_spectral_clustering spectral clustering and k = 2 for protein 6cs2...
Current progress 42.86%
```

We plot both the detected clusters and the map of the participation coefficient projected onto the ribbon structure of the SARS-CoV/ACE2 complex.

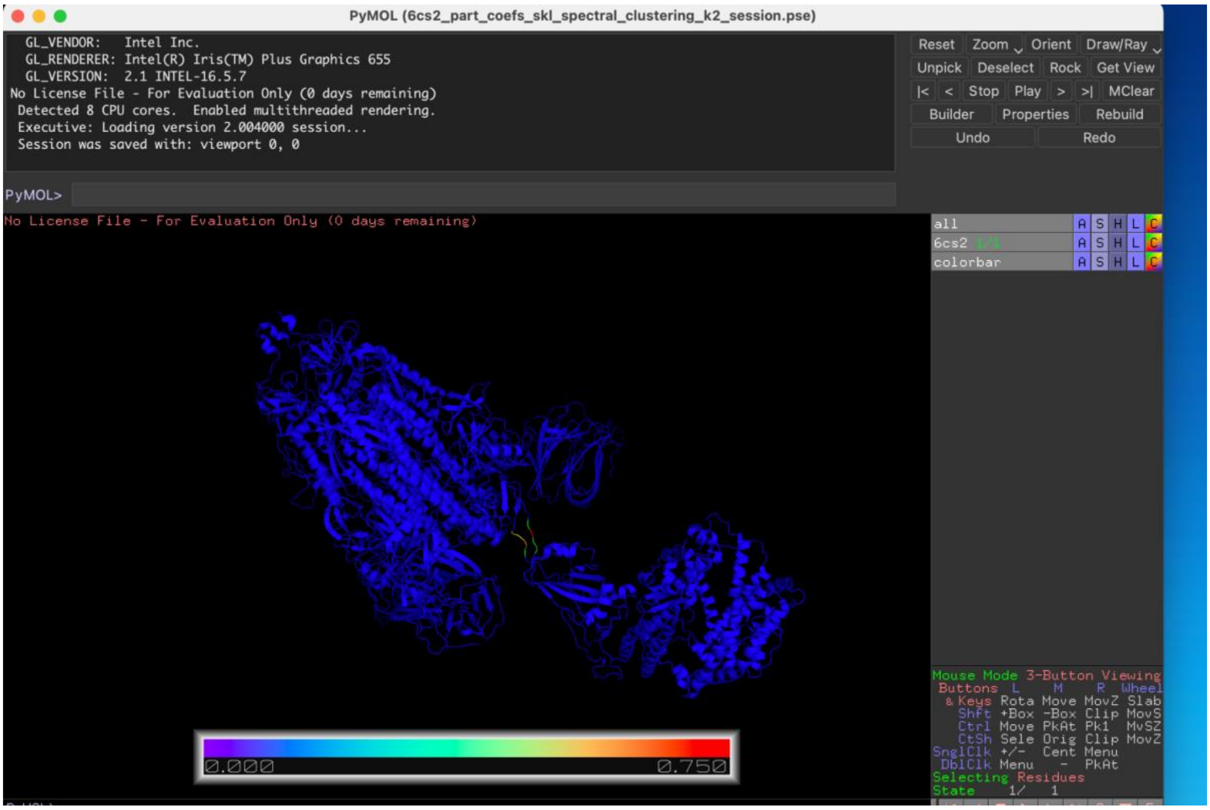
The following figure shows that the complex does not have a modular structure,



```
  GL_VENDOR:   Intel Inc.
  GL_RENDERER: Intel(R) Iris(TM) Plus Graphics 655
  GL_VERSION:  2.1 INTEL-16.5.7
No License File - For Evaluation Only (0 days remaining)
 Detected 8 CPU cores.  Enabled multithreaded rendering.
 Executive: Loading version 2.004000 session...
 Session was saved with: viewport 0, 0

PyMOL>
No License File - For Evaluation Only (0 days remaining)
PyMOL>_
```

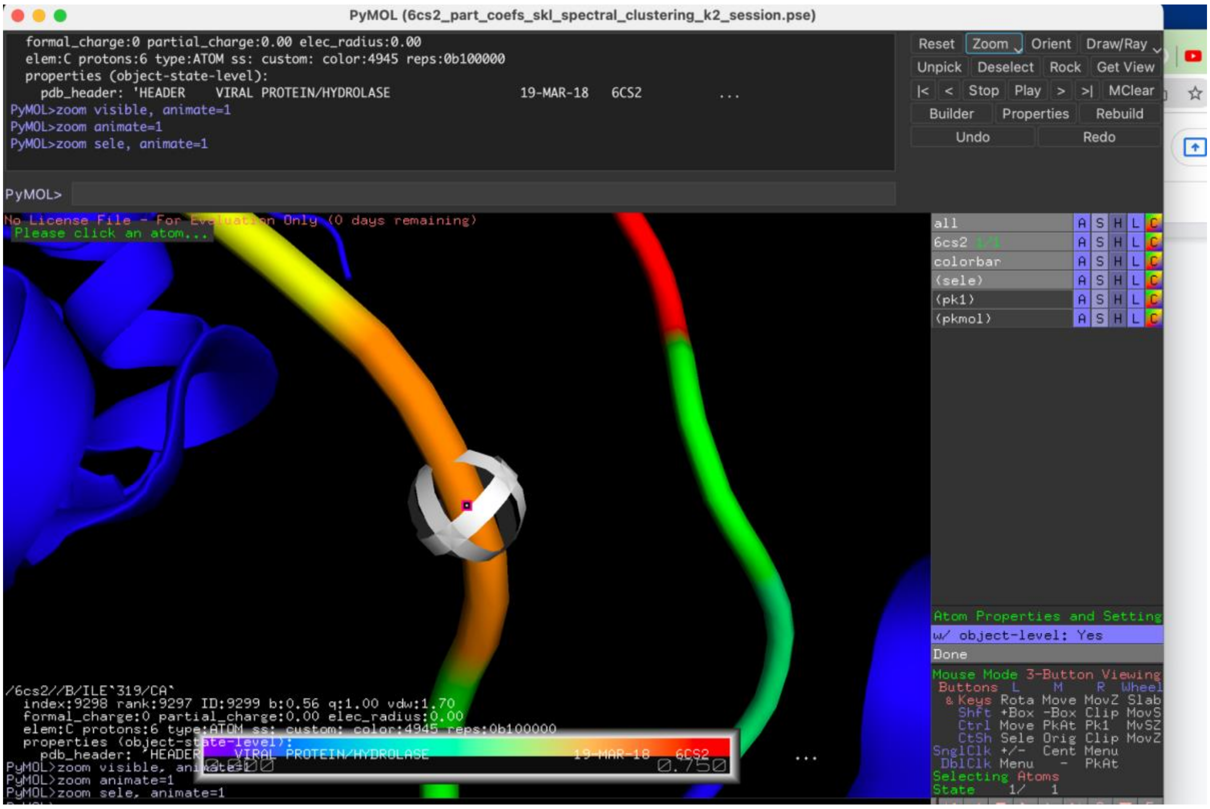
Conversely, the following figure shows an active region (P > 0) in the junction between the fusion peptide and the trimeric bulk phase of the spike protein. This region is a good candidate to intervene in the allosteric regulation of complex formation.



PyMOL (6cs2_part_coefs_skl_spectral_clustering_k2_session.pse)

```
  GL_VENDOR:   Intel Inc.
  GL_RENDERER: Intel(R) Iris(TM) Plus Graphics 655
  GL_VERSION:  2.1 INTEL-16.5.7
No License File - For Evaluation Only (0 days remaining)
 Detected 8 CPU cores.  Enabled multithreaded rendering.
 Executive: Loading version 2.004000 session...
 Session was saved with: viewport 0, 0
PyMOL>
No License File - For Evaluation Only (0 days remaining)
```

The potential allosteric regions is more clear when zooming on the image closest to the interesting region obtaining the following.



PyMOL (6cs2_part_coefs_skl_spectral_clustering_k2_session.pse)

```
  formal_charge:0 partial_charge:0.00 elec_radius:0.00
  elem:C protons:6 type:ATOM ss: custom: color:4945 reps:0b100000
  properties (object-state-level):
    pdb_header: 'HEADER    VIRAL PROTEIN/HYDROLASE                 19-MAR-18   6CS2          ...
PyMOL>zoom visible, animate=1
PyMOL>zoom animate=1
PyMOL>zoom sele, animate=1
PyMOL>
No License File - For Evaluation Only (0 days remaining)
 Please click an atom...
/6cs2//B/ILE`319/CA`
  index:9298 rank:9297 ID:9299 b:0.56 q:1.00 vdw:1.70
  formal_charge:0 partial_charge:0.00 elec_radius:0.00
  elem:C protons:6 type:ATOM ss: custom: color:4945 reps:0b100000
  properties (object-state-level):
    pdb_header: 'HEADER    VIRAL PROTEIN/HYDROLASE                 19-MAR-18   6CS2          ...
PyMOL>zoom visible, animate=1
PyMOL>zoom animate=1
PyMOL>zoom sele, animate=1
```

5 - Algorithms Parameters

We here report the parameters that can be used in the algorithms.

Spectral Clustering Parameters.

When selecting spectral clustering, the software requires the number of clusters as input selection. User may: (i) select a single value as the number of cluster, (ii) a set of values; or he can leave to the software the choice of the best number of clusters through an optimisation scheme.

This number depends on the number of the domain the user wants to explore (as in Di Paola et Al Allostery 2020 https://link.springer.com/protocol/10.1007/978-1-0716-1154-8_2-).

The current version of the software implements following algorithms wrapping the Scikit library:

Unnorm SSC : The unnormalised soft spectral clustering

Norm SSC : it computes a soft clustering using a normalised laplacian

Norm HSC : it calculates a hard clustering on a normalised Laplacian;

Unnorm HSC : it calculates a hard clustering on a unnormalised Laplacian;

HSC ShiMalik it calculates a hard clustering on a normalised Laplacian by Shi-Malik

SSC ShiMalik it calculates a soft clustering on a normalised Laplacian by Shi-Malik

Spectral clustering algorithms have been implemented following the algorithm proposed in *Von Luxmburg U. et al.* Hard clustering is obtained by using k-means, while soft clustering uses fuzzy c-means.

Community Detection Parameters.

The current version of PCN-Miner supports the following algorithms for community detection:

Louvain
Leiden
Walktrap
asyn_fluid
greedy_modularity
infomap
spinglass.

Through the gui the user has only to select the number of desired communities for asyn fluid, while other algorithms proceed without user-defined parameters. We run each algorithm with best parameters set by the literature. User can launch the CLI version of PCN-Miner to control all the parameters.

6 - Additional Information

The PCN-Miner tool is an open source one. Users are invited to use and give additional contributions both in terms of code enrichment as well as use case and data.

***References**:*

Di Paola, L., Hadi-Alijanvand, H., Song, X., Hu, G., & Giuliani, A. (2020). The Discovery of a Putative Allosteric Site in the SARS-CoV-2 Spike Protein Using an Integrated Structural/Dynamic Approach. Journal of proteome research, 19(11), 4576–4586. https://doi.org/10.1021/acs.jproteome.0c00273

Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.