Supporting Information

C-Graphs Tool with Graphical User Interface to Dissect Conserved Hydrogen-Bond Networks: Applications to Visual Rhodopsins

Éva Bertalan,[1] Elena Lesca,[2,3] Gebhard FX Schertler,[2,3]

and Ana-Nicoleta Bondar[1,4,5]

[1]Freie Universität Berlin, Department of Physics, Theoretical Molecular Biophysics, Arnimallee 14, D-14195 Berlin, Germany

[2]ETH Zürich, Paul Scherrer Institut, Department of Biology and Chemistry, Laboratory of Biomolecular Research, 5303 Villigen-PSI, Switzerland

[3]ETH Zürich, Department of Biology, 8093 Zürich, Switzerland

[4]University of Bucharest, Faculty of Physics, Str. Atomiştilor nr. 405, Măgurele 077125, Romania

[5]Forschungszentrum Jülich, Institute for Neuroscience and Medicine and Institute for Advanced Simulations, Computational Biomedicine, IAS-5/INM-9, 52425 Jülich, Germany

Correspondence to nbondar@fizica.unibuc.ro

`

This Supporting Information File contains the C-Graphs User's Manual with additional Supporting Information Figures S1-S26.

This C-Graphs User's Manual was written by Éva Bertalan and Ana-Nicoleta Bondar.

License:
C-Graphs is made open access on an as-it-is basis with GNU General Public License v3.0.

How to cite:
When using C-Graphs, please cite the following two references:

1. Bertalan E, Lesnik S, Bren U, Bondar A-N. Protein-water hydrogen-bond networks of G Protein-Coupled Receptors: Graph-based analyses of static structures and molecular dynamics. *Journal of Structural Biology* 212, 107634 (2020)
2. Bertalan E, Lesca E, Deupi X, Schertler GFX, Bondar A-N. C-Graphs tool with graphical user interface to dissect conserved hydrogen-bond networks: Applications to visual rhodopsins.

How to install C-Graphs
Instructions for how to install C-Graphs are given in the Readme file included with the zipped directory of the source code.

C-Graphs as standalone Phyton scripts
The source code is available in the github repository evabertalan/cgraph.

Content of User's Manual

1. Description of main analysis window for performing conserved H-bond network analyses of protein crystal structures
2. Figures illustrating windows of the C-Graphs interface
3. References

Note
Figures included in the User's manual have been generated using as test systems a set of crystal structures of squid rhodopsin, and a short simulation trajectory of Jumping Spider Rhodopsin-1, JSR-1. All numerical values of various parameters listed in the C-Graphs interface are for illustration purposes. The set of computations being discussed with each figure is highlighted with an orange frame.

1. Main window for conserved network analysis of crystal structures

The C-Graphs interface opens with the main window illustrated in Figure 1. Here, the User can choose between analyses on Crystal structures, MD trajectories or comparison of two structures.

1.1 Input locations contains the tab *Select PDB folder*, with which the User selects the folder containing protein coordinate files in standard Protein Data Bank (PDB) format (Figure 2), and the tab *Select reference file* (Figure 3) to select one coordinate file as a reference.
The reference structure selected here will be used as a reference for all conserved graphs computations, as basis to compute relative sequence identities for all other structures, for all structural overlaps, and for preparing the plots of conserved H-bond graphs.
We recommend that structures of membrane proteins are oriented along the membrane normal, as this will facilitate interpretation of PCA-projected H-bond graphs.
The *Minimum sequence identity (%)* tab (Figure 1) expects an integer number that indicates the value of the minimum sequence identity between any structure read under the tab Select PDB folder, and the reference structure. Structures whose sequence identity relative to the reference is below the minimum value set here will be discarded from analyses.

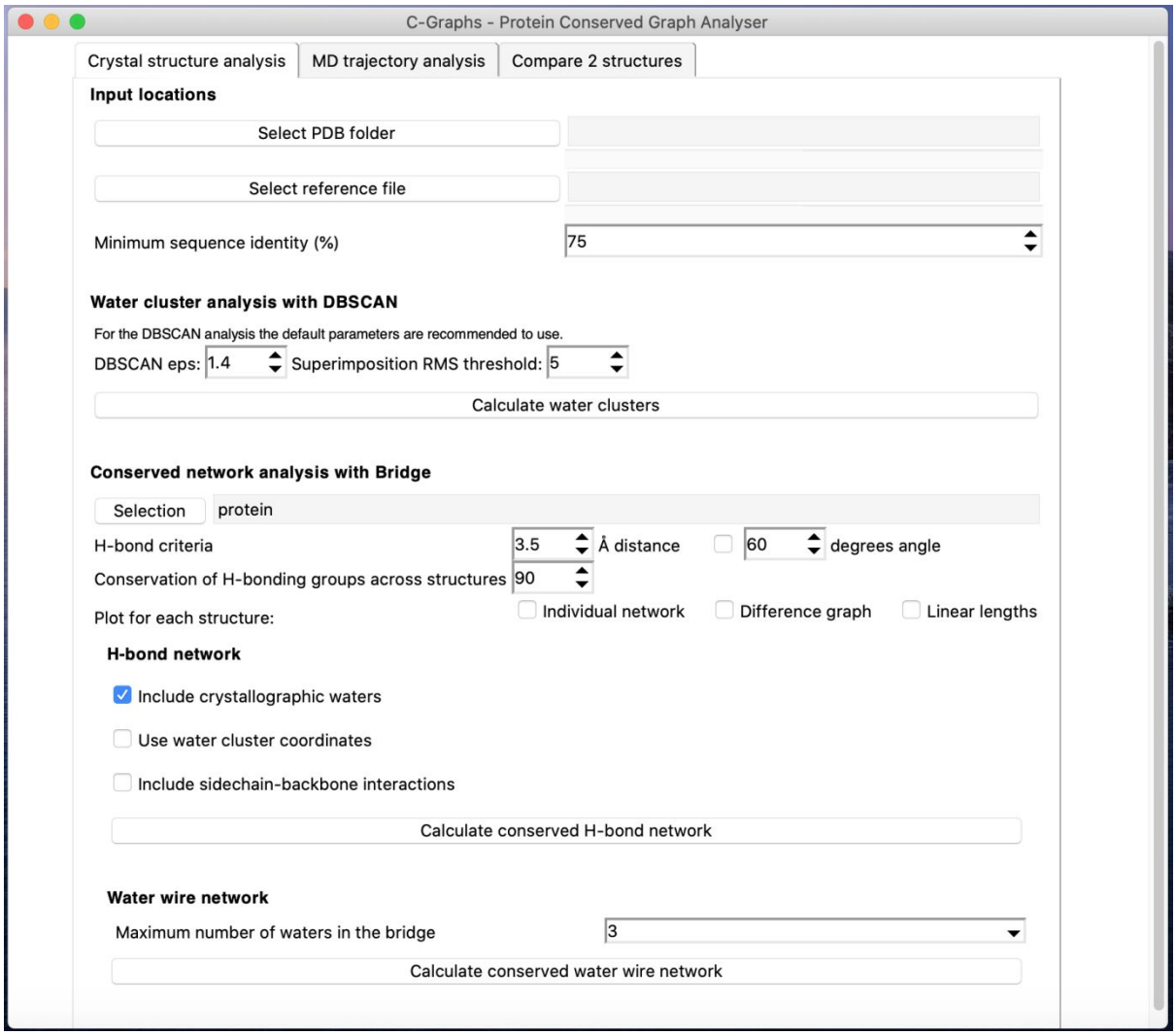

Figure S1. The main window of C-Graphs. The first two tabs, 'Crystal structure analysis' and 'MD trajectory analysis', are used to choose between analyses of static structures, or of MD simulation trajectories. The third tab, 'Compare two structures', can be used for both static structures and MD simulations.

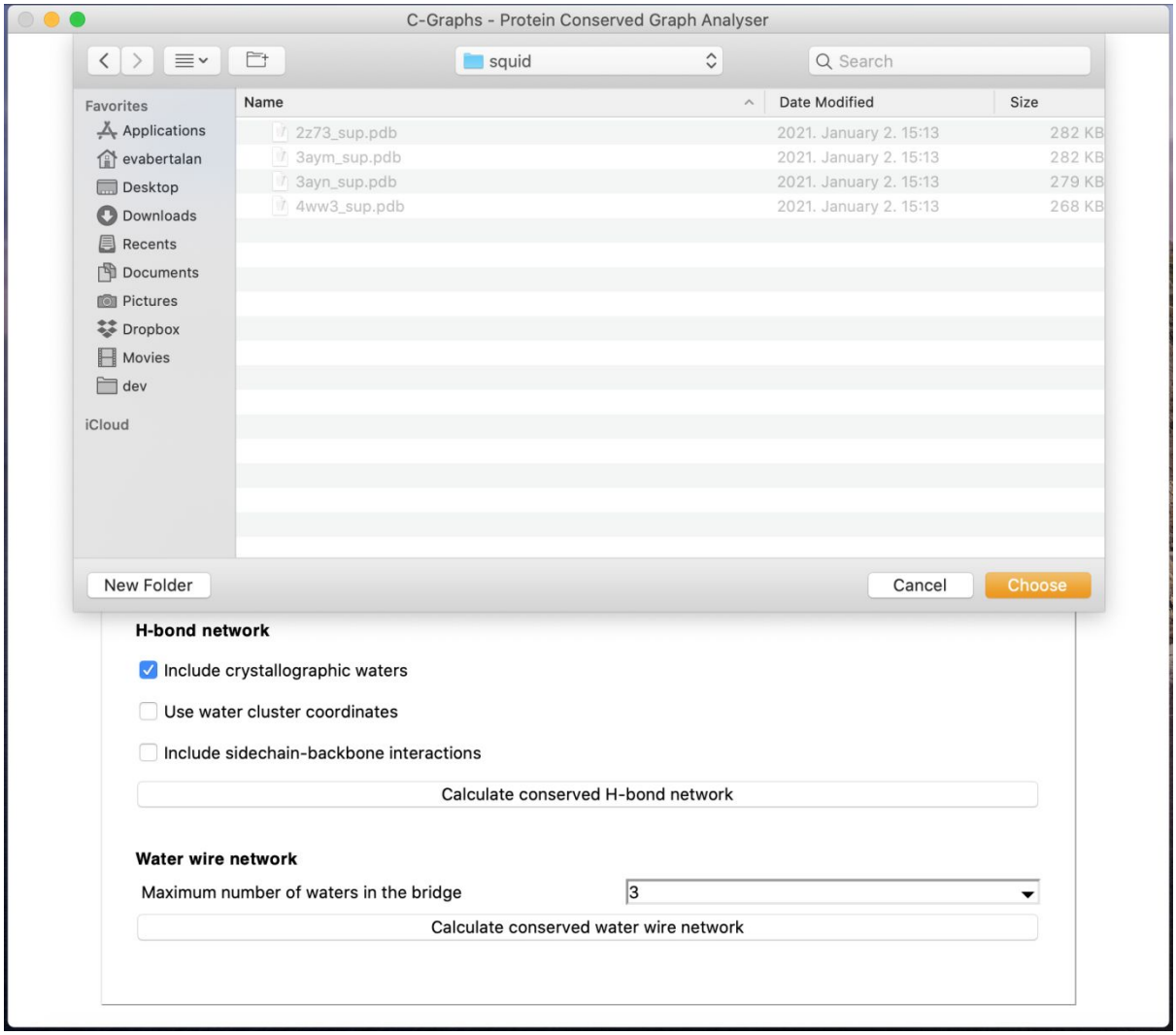

Figure S2. Illustration of the selection of protein coordinate files in PDB format. All coordinate files listed in this directory will be included in analyses. Any number of coordinate files can be read here. The order in which files are listed in the directory is irrelevant, as graphs are computed for each individual structure before the computation of the conserved graph. The root-mean-squared difference (RMSD) between the structures and the reference structure is reported in the bash terminal of the interface, and in the log file of the computation. The work folder that contains results of all data analyses will be generated automatically as a subdirectory of the folder with the PDB files to be analyzed.

The structures will be automatically uploaded by C-Graphs once a specific analysis type has been selected in a subsequent step. All results of the various analyses that can be performed with C-Graphs are stored in the directory workfolder that is generated automatically as a subdirectory of the directory containing the input PDB coordinate files.

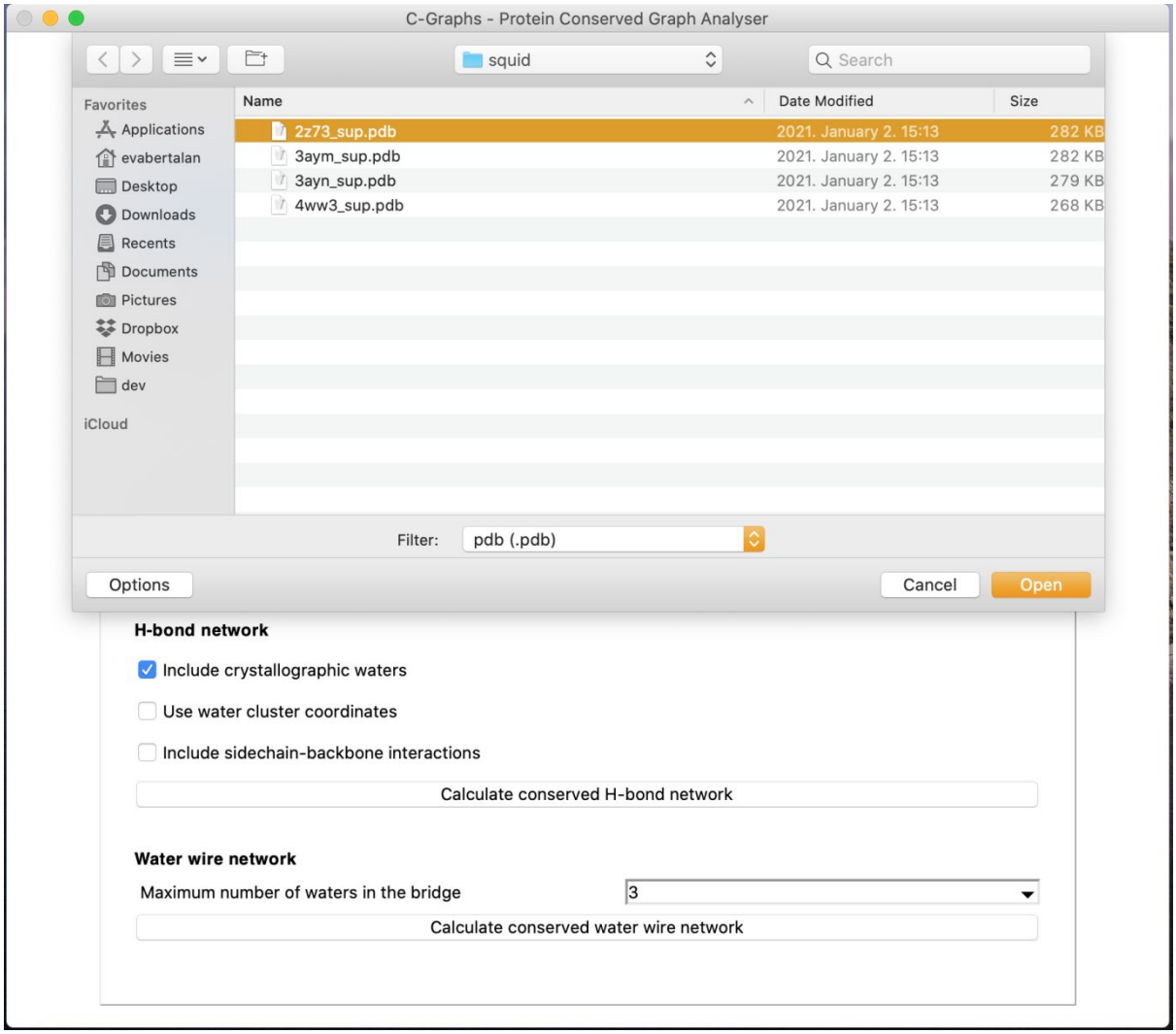

Figure S3. Selection of a protein structure to be used as reference for all computations. The coordinate file for the reference structure must be in standard PDB format. We recommend that structures of membrane proteins are oriented along the membrane normal.

1.2 Water cluster analysis with DBSCAN

C-Graphs uses the DBSCAN algorithm [1] to cluster waters that are located close to each other in protein structures included in analyses (Figures 4-6). The search for water molecules located close to each other is made using protein structures superimposed on the reference structure. Based on test computations [2], the recommended eps parameter is 1.4. Both the RMS value and eps parameter can be modified, however we recommend using the default parameters. Changing the RMS and eps values will impact results of the water clustering algorithm.

Results of the DBSCAN water cluster analysis are collected in the water_clusters directory that is generated automatically as a subfolder of the workfolder. The water_clusters directory contains the following:

- For each PDB file read as input, C-Graphs generats a .txt file with the *xyz* coordinates of all water molecules found in the PDB (Figure 6)
- The file denoted as DBSCAN_water_cc_chimera.bild can be read in Chimera [3] to generate an intuitive molecular graphics of the volume in which waters that cluster together are found in structures included in analyses (Figure 5).

- Files DBSCAN_water_cc_chimera.txt and DBSCAN_water_cc_chimera.xyz contain the same results as the .bild file above. These .txt and .xyz files could be read, for example, in Visual Molecular Dynamics, VMD [4], or PyMol [5].
- The file water_clusters.png contains an image of the PCA projection of all water molecules found in the superimposed structures. Scatter points with the same color indicate water molecules identified as members of the same cluster, whereas outlier water molecules are denoted as black points.

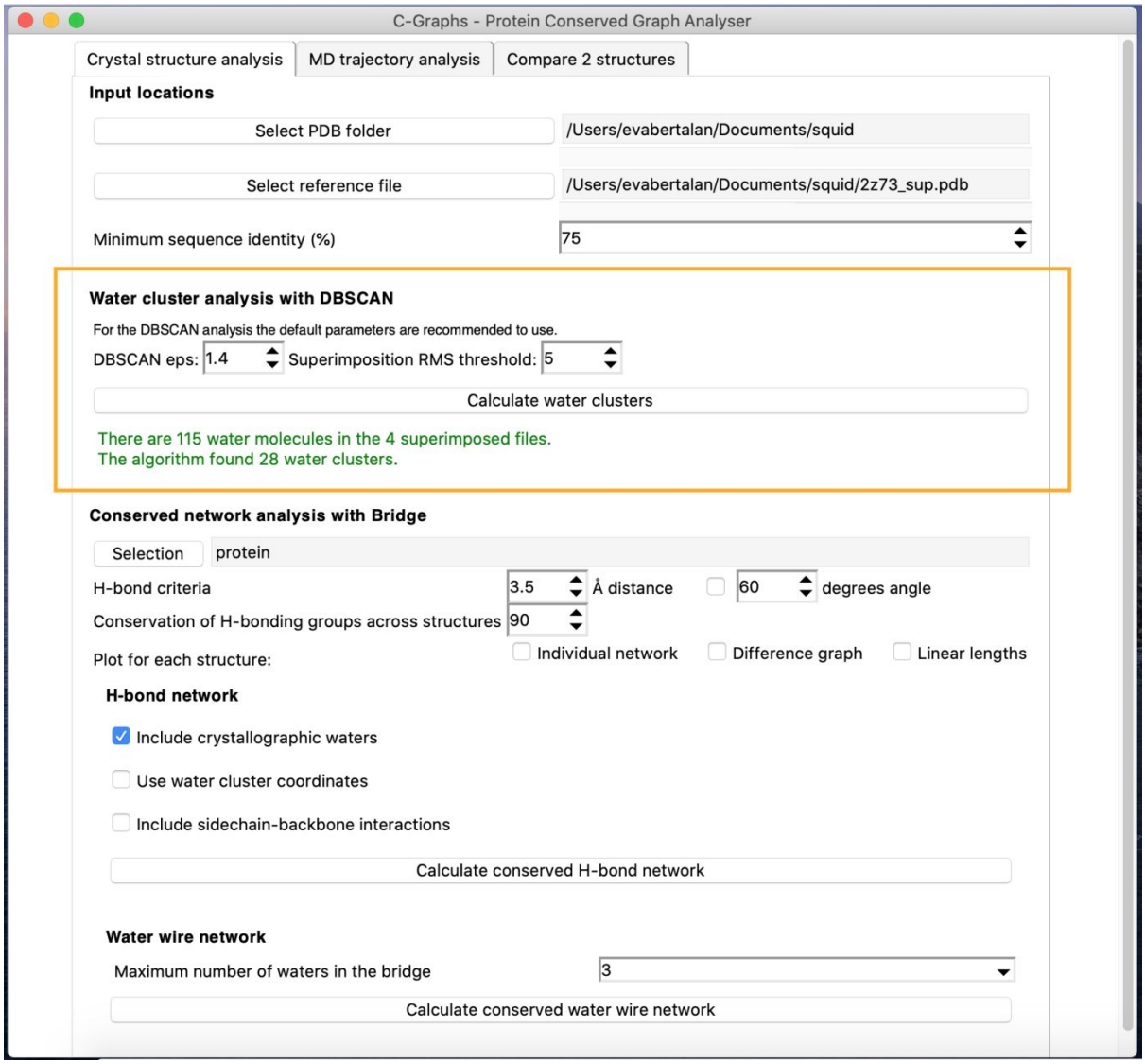

Figure S4. The DBSCAN water clustering is initiated by selecting the button 'Calculate water clusters'. Once the calculation has been completed, the interface reports the total number of water molecules present in the ensemble of protein structures included in the analysis, and the total number of clusters identified for the water molecules.

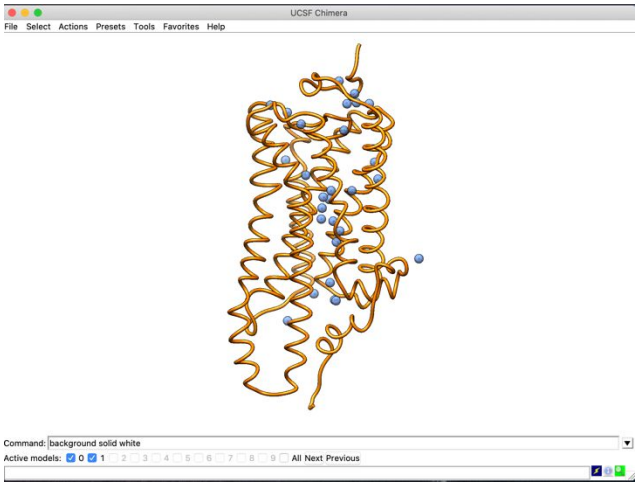

Figure S5. Molecular graphics of a protein structure (PDB ID: 2Z73) with water clusters shown as blue spheres. The image was generated using Chimera [3].

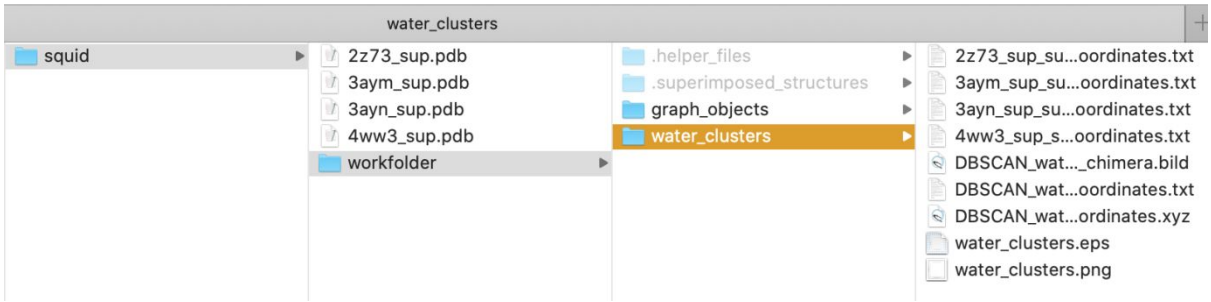

Figure S6. Screen snapshot of the directory that contains results of the water clustering computation. Files with the water coordinates are labeled according to the original PDB file.

1.3 Conserved H-bond network analysis with Bridge

C-Graphs relies on Bridge [6] to compute H-bond networks. All H-bond computations are performed according to the default geometric criteria of H-bonding. When the input protein structures are experimental structures that lack H atoms, H-bonds are computed with a single distance-based criterion of 3.5Å between the heavy atoms of the H-bond. When protein structures are read from an MD simulation trajectory, H-bonds are computed with the default 3.5Å distance criterion, and an additional H-bond angle criterion of 60º can be turned on. Both the distance and the angle criteria can be modified in the interface.

By default, C-Graphs computes the H-bond network for protein groups and water molecules. The selection of atoms for which H-bond graphs shall be computed can be adjusted from the *Selection* button. When selected, this button opens a window (Figure 9) in which a selection string may be used that supports the MDAnalysis[7, 8] atom selection language.

To include them in H-bond network calculations, non-protein donor and acceptor atoms must be declared in the Selection window. For example, to include a sodium ion H-bond graoh computations, the selection string will read “protein or resname NA”, where *NA* is the residue name of the sodium ion in the PDB file; additionally, the name of sodium ion, e.g., Na, needs to be added to the list of acceptor and donor atoms.

The MDAnalysis atom selection language is described under
https://userguide.mdanalysis.org/stable/selections.html

Conservation of H-bonding groups across structures expects an integer value indicating the percentage of structures that must include the graph nodes (H-bonding groups) to be included in the computation of conserved H-bond graphs. In the example illustrated in Figure

7, plots of the conserved H-bond network analysis will display only H-bonding groups present in 90% of the structures included in analyses.
The User can then choose to have one, two, or all three sets of data plots generated in corresponding subfolders of the work directory.
The *Use water clusters* option includes in the conserved H-bond network calculation the water cluster centers obtained with DBSCAN. In the case where more than one of the structures contain H-bonding between protein group and a water molecule, and the number of structures meets the threshold value set in *Conservation of H-bonding group,* the water molecule will be included in the conserved H-bond plot, marked as conserved water site, and labeled as W1, W2, …, Wn, where *n* is the number of water clusters.

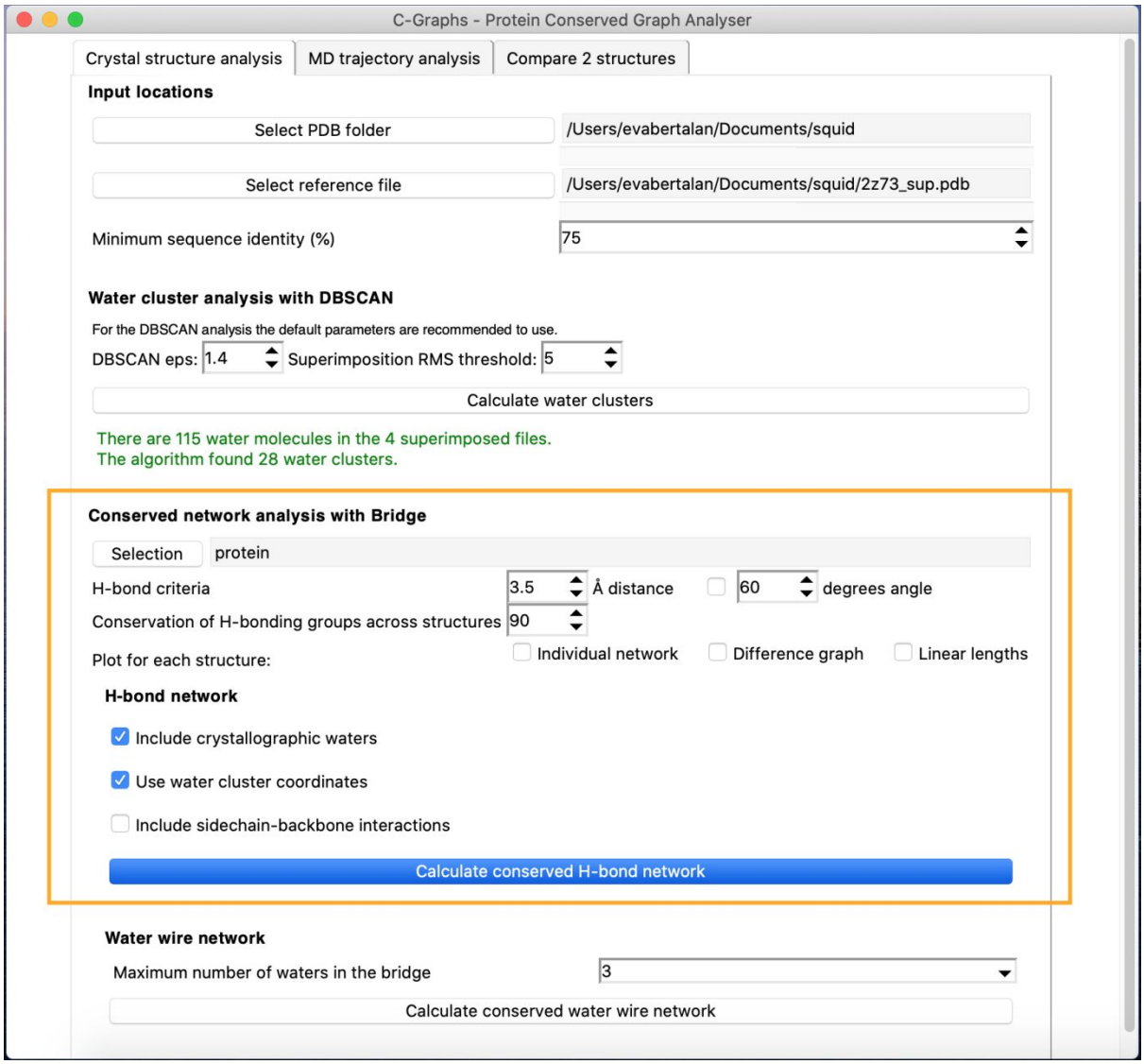

Figure S7. Computation of conserved protein-water H-bond networks. The User can choose which plots will be generated as result of the computation. Plots of conserved H-bond graphs are saved as .png and .eps formats in the subdirectory H-bond_graphs of workfolder. Each plot is provided twice, with and without without labels for the amino acid residues (Figure 8). The tab *Calculate conserved H-bond network* will be colored blue as long as the computation is ongoing, and return to a white color once the computation is completed.

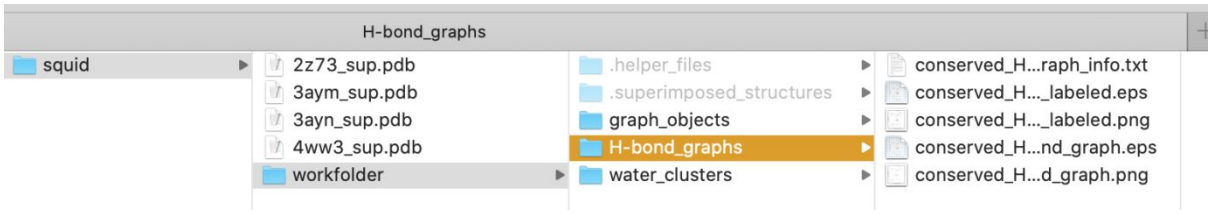

Figure S8. The H-bond_graphs subdirectory of workfolder includes the data plots in .png and .eps formats. Files denoted as '_labeled' contain the labels of the amino acid residues that are part of the network. For clarity, and to facilitate preparation of figures for publications, C-Graphs also generates a version of the plots in which nodes representing amino acid residues are not labeled, and the User may add manually labels only for nodes of particular interest.

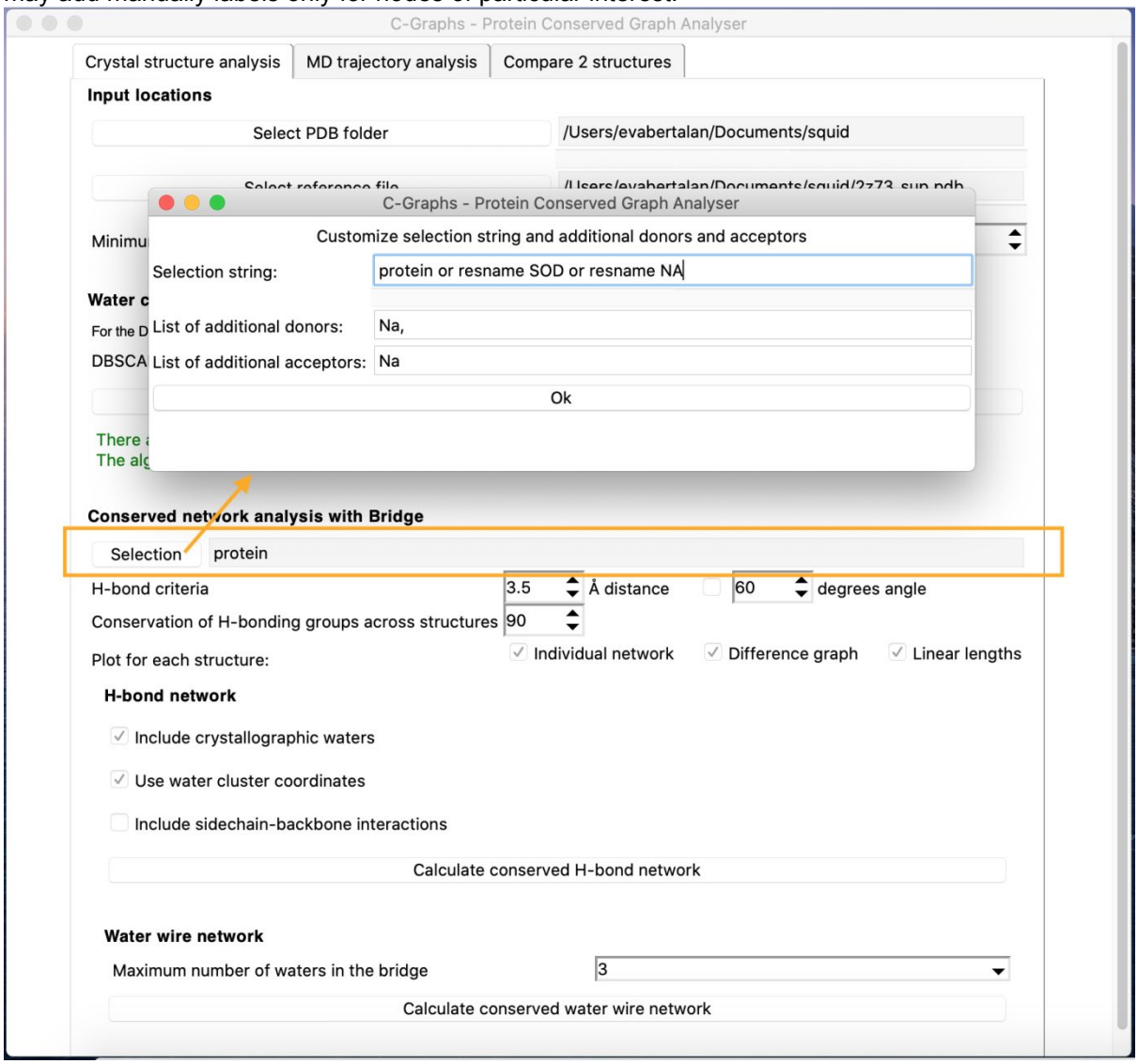

Figure S9. Custom atom selection. The Selection button opens a window in which any atom selection can be included according to the MDAnalysis[7, 8] atom selection language. To be included in H-bond graph calculations, non-protein atoms must be declared in the list of additional donors and acceptors.

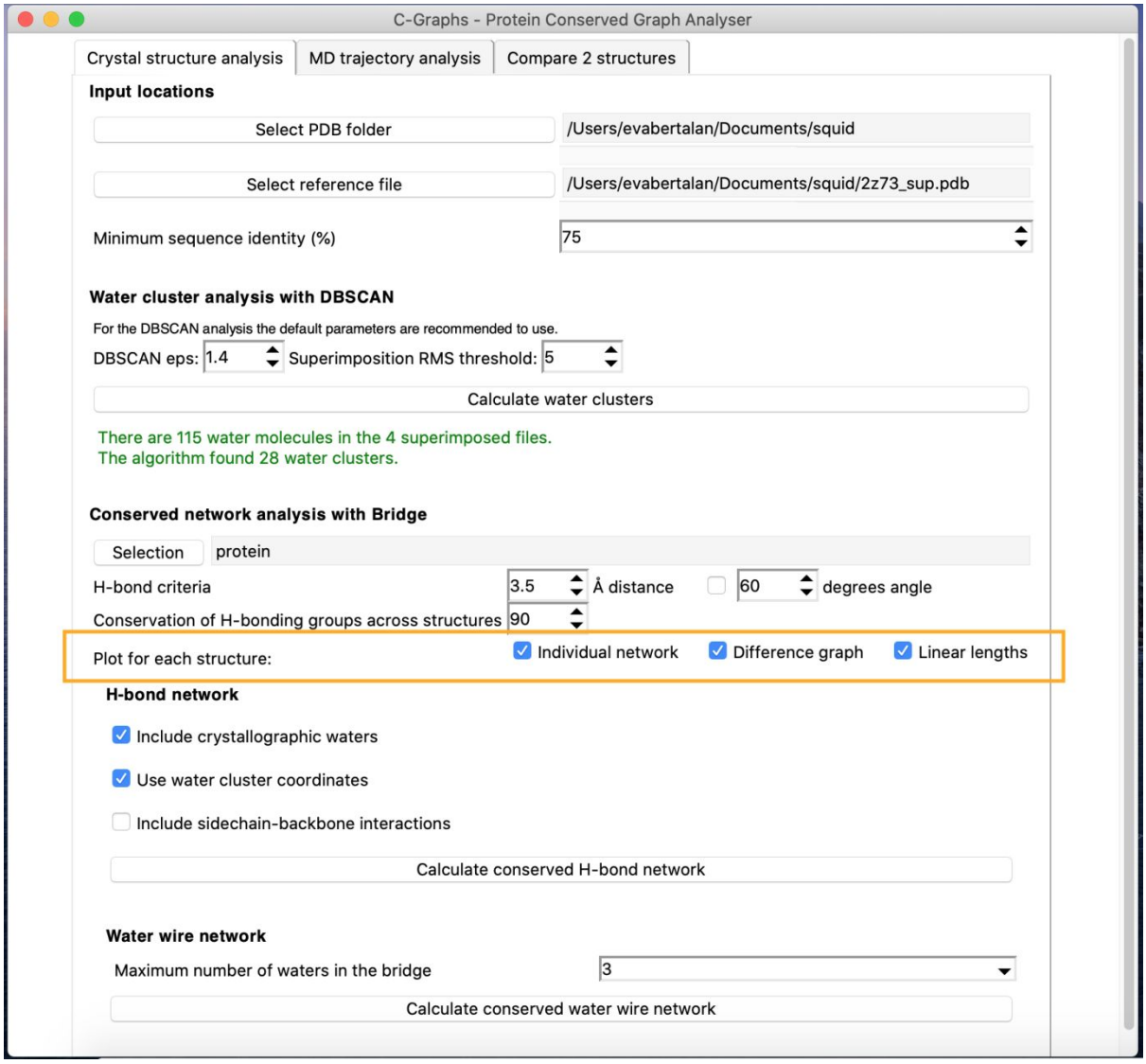

Figure S10. Selections for generating data plots of the conserved water analyses. *Plot network for each structure* generates a Bridge H-bond graph for each protein structure included in analyses. *Plot difference graphs for each structure* generates, for each protein structure, the difference H-bond graph relative to the conserved graph. The difference graph shows the unique H-bonds that are not conserved across the structures. Nodes and edges are colored gray when conserved in all structures, and blue when present only in the structure for which the plot is generated. *Plot linear lengths of continuous networks for each structure* generates, for each structure, a two-dimensional plot with the linear length of H-bond networks as vertical axis, and the linear projection of the 3rd, Cartesian *Z* coordinate of the amino acid residue node, as horizontal axis. All plots are collected in the subdirectory H-bond graphs (Figure 11).

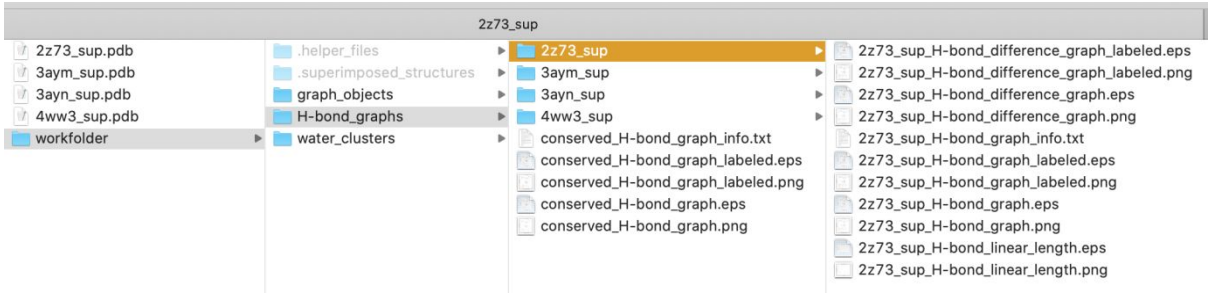

Figure S11. Content of the subdirectory H-bond_graphs when all three options for data plots are active (see Figure 10). A separate folder is generated for each protein structure included in analysis.

This folder includes data plots in .png and .eps formats, and each plot is generated twice, with and without labels for the nodes representing amino acid residues that are part of the H-bond graph.

Conserved network analysis with Bridge has two sub-sets of computations as follows:

1.3.1 H-bond network computes the H-bond network for protein groups. By default the H-bond network is computed only for protein sidechains. By selecting the corresponding boxes, the User may include in the computation water clusters identified previously with the DBSCAN computation (Figure 4), and protein backbone groups.

1.3.2 Water wire network includes in the H-bond graph computation water wires, which are computed with Bridge [6]. The C-Graphs interface requests here an integer number that sets the maximum number of H-bonding waters in the chain (Figure 12); the maximum length of water chains between two protein groups is five waters. The user can select here water bridges with 1, 2, 3, 4, or 5 waters. Results of computations of conserved water wire networks are saved under corresponding subfolders of workfolder. For example, when chains with up to three waters are computed, results are saved in a directory labeled 3_water_wires (Figure 13).

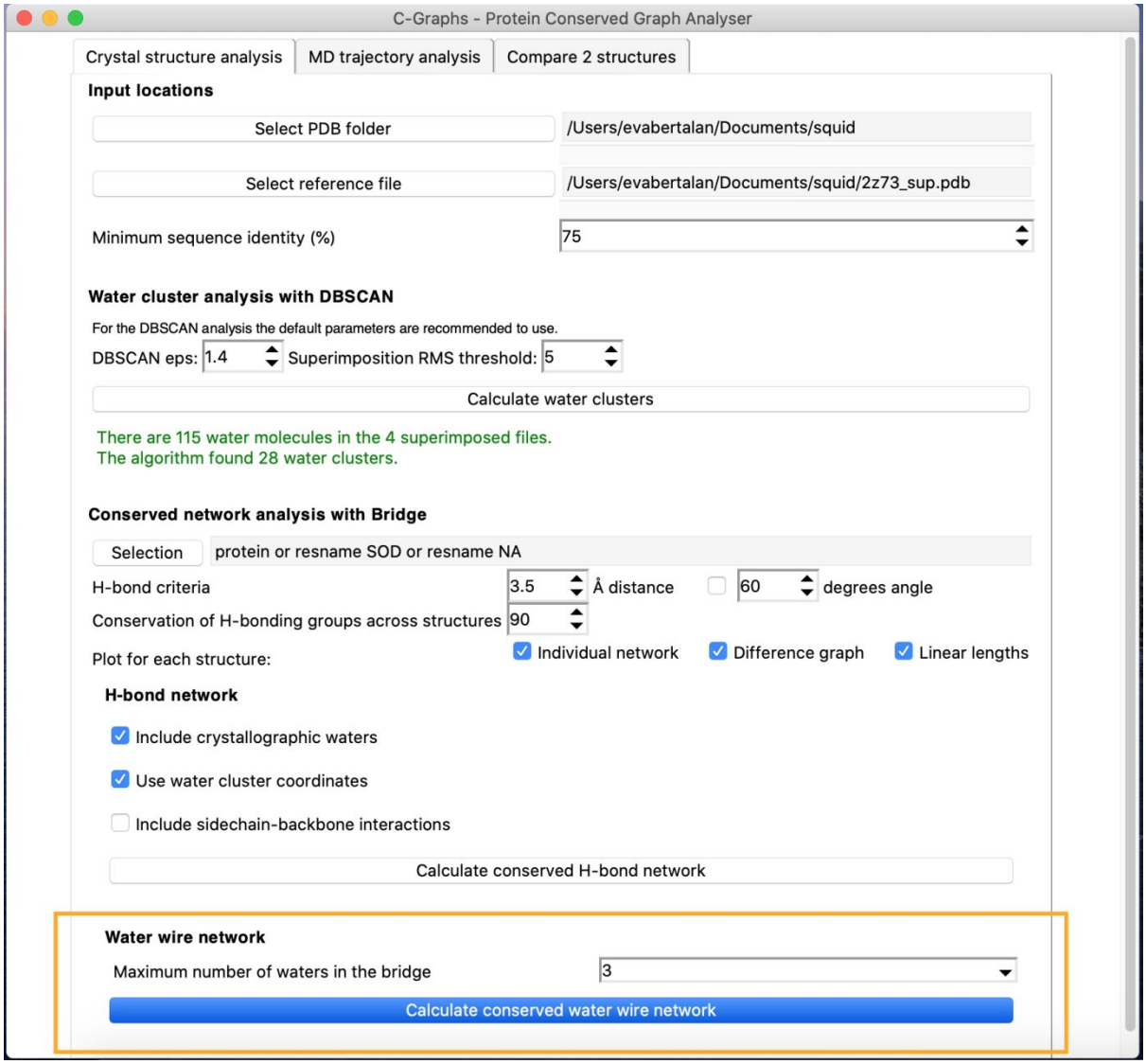

Figure S12. The Water wire network analysis with the C-graphs interface. The computation is initiated by setting an integer number ≤ 5, and pressing the tab highlighted blue.

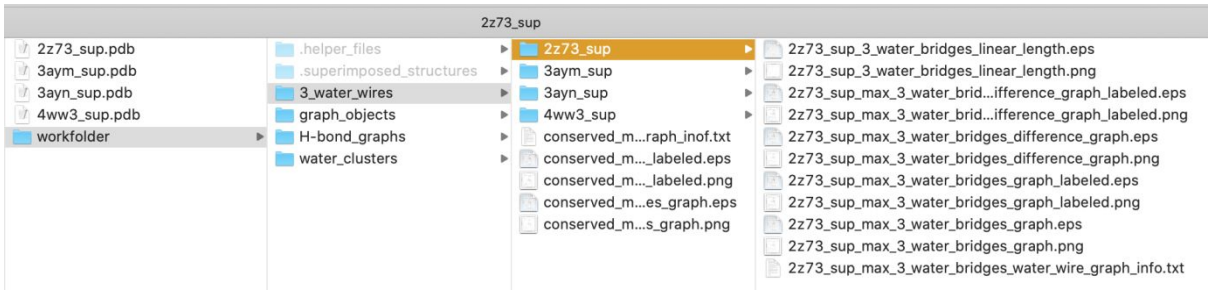

Figure S13. Content of folders with results of conserved H-bond network analyses. H-bond graphs and water wires are saved as separate sub-folders workfolder. For each protein structure included in analyses, results are saved as .png and .eps files, with and without labels for the amino acid residues.

2. Conserved network analysis of MD trajectories

MD trajectory analysis (Figure 14) expects as input files a protein structure file and one or more trajectory files. Once a set of .psf and .dcd files has been read, it must be given a name using the *Name as* tab (Figure 14). The name is important when H-bond graphs computed from at least two independent simulation sets must be compared with each other.

All results of C-Graphs analyses will be saved in the workfolder whose location must be declared before selecting the MD simulation file(s). When a name is assigned to a simulation set used for analyses, a sub-folder with the same name will be generated automatically for storing the corresponding results. Assigning the same name to two different simulation sets will result in sub-folders with analysis results being overwritten.

The minimum and maximum numbers of water molecules allowed in a wire are 1 and 5, respectively. Analyses of MD trajectories have been tested with dcd trajectory files in standard NAMD format.

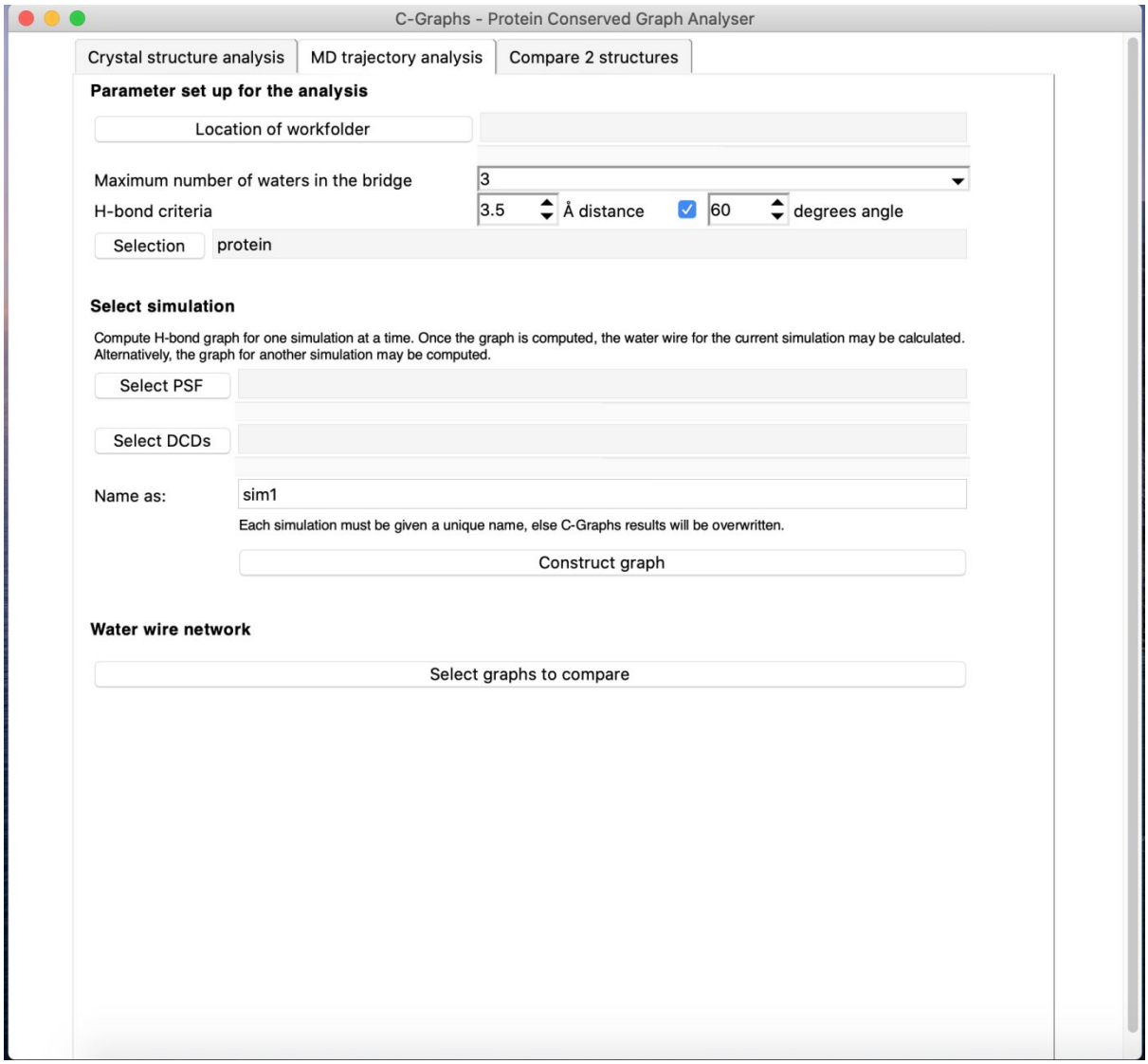

Figure S14. Main window for C-Graphs analyses of MD simulation trajectories. The interface expects that the location of the workfolder and the maximum number of waters in a wire are declared prior to reading input files for data analyses.

Once the input .psf and .dcd files have been selected (Figure 15), C-Graphs computations are initiated by pressing the button *Construct graph* (Figures 14,16). Depending on the size of the protein and on the number of coordinate sets read, the construction of H-bond graphs might require some time. C-Graphs will report when the calculation is completed (Figure 17).

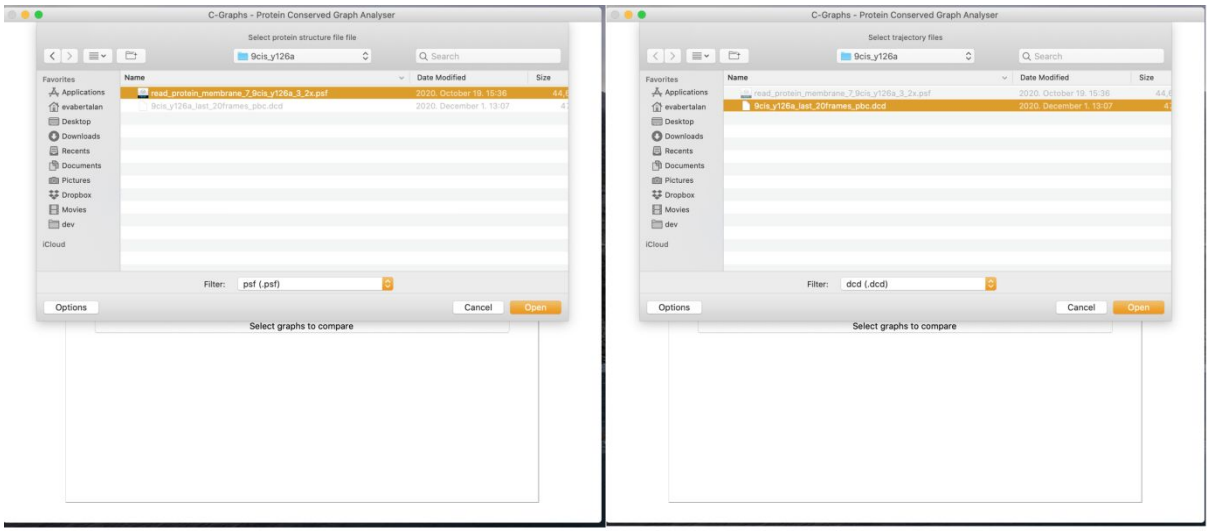

Figure S15. Selection of .psf and .dcd files for analyses with C-Graphs. Left: Selection of the .psf file. Right: Selection of one or more .dcd files for data analyses.

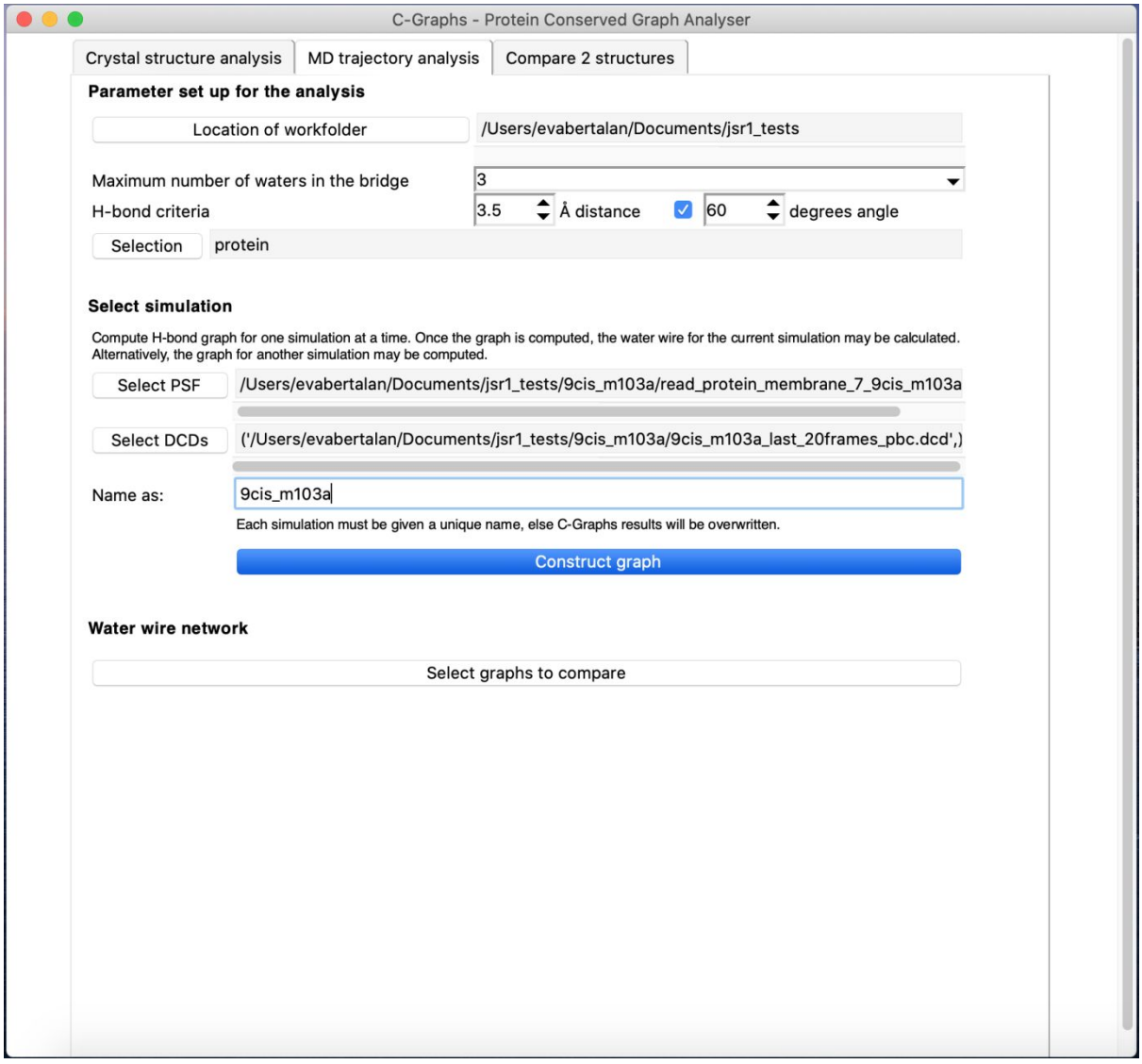

Figure S16. Initiation of a C-Graphs computation for an MD simulation trajectory. Once the input .psf and .dcd files have been read, and the simulation set has been assigned a unique name, the computation is initiated by selecting the button *Construct graph*.

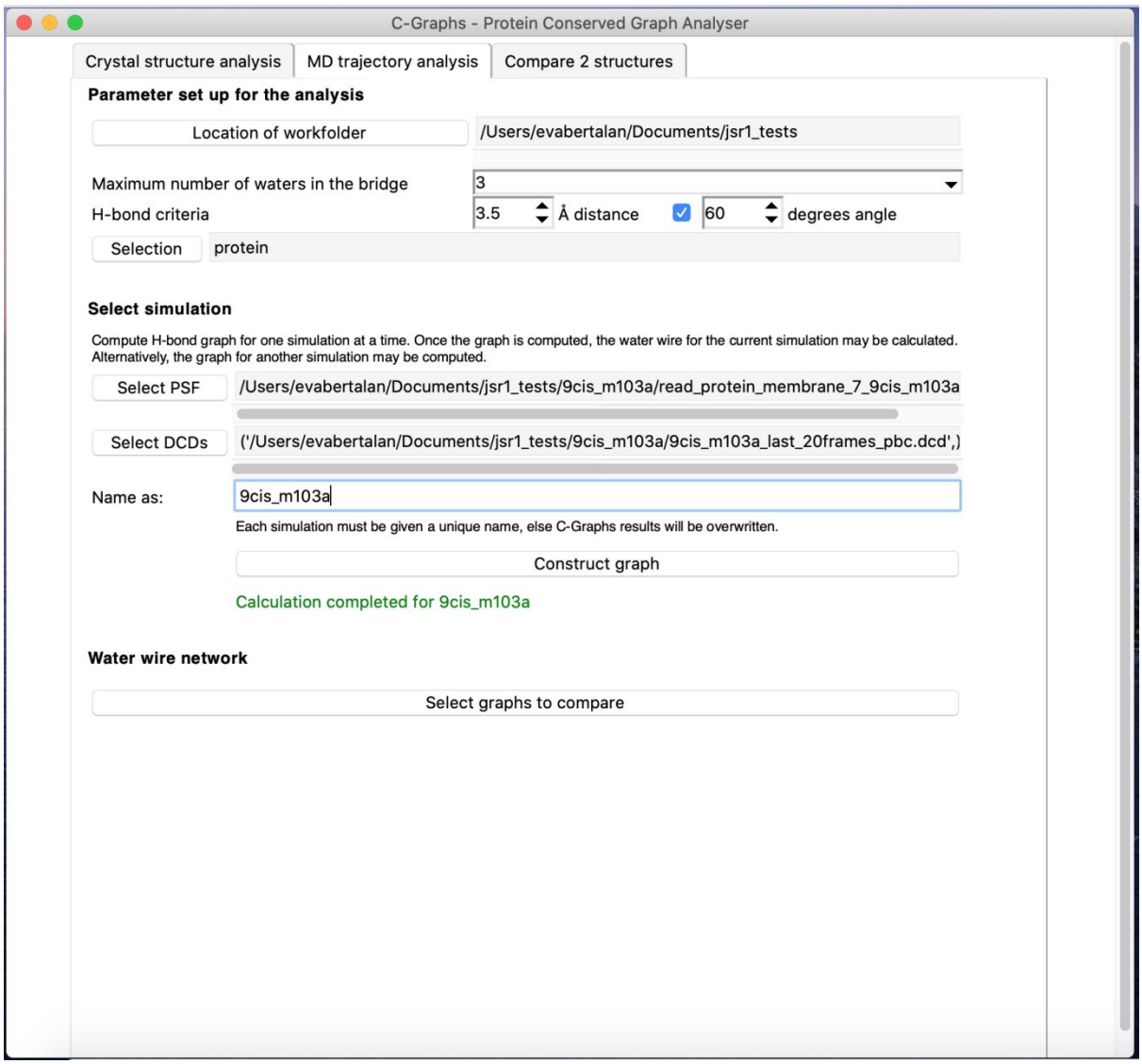

Figure S17. Once the H-bond graph for a given set of MD simulations has been completed, the C-Graphs interface indicates Calculation completed for that set. At this time, when H-bond graphs are to be compared between two or more simulation sets, the subsequent simulation set may be read, and the calculation of H-bond graphs for this subsequent set may be initiated.

Water wire network uses as input the H-bond graphs computed for the MD simulation trajectories. Once all H-bond graphs of interest have been computed and stored in sub-folders with unique names, comparisons among the H-bond graphs may be performed.
The tab Select graphs for conserved network calculation opens a dialogue to select the location of pre-computed H-bond graphs. By default, all H-bond graphs are stored in a subfolder of theworkfolder, denoted as graph objects (Figure 18). The example in Figure 18 illustrates the content of graph objects from a computation of H-bond networks with up to three waters in H-bonded water wires.
Once the User has selected the H-bond graphs (Figure 18), and activated the computation of conserved H-bond graphs (Figure 19), C-Graphs uses these graphs as basis to compute conserved networks and to generate corresponding data plots according to the options illustrated in Figure 20. C-Graphs expects here settings for the minimum conservation of H-bonding groups (graph nodes) among different graphs, and of the minimum occupancy of H-

bonds (Figure 20). Data plots are saved, with names corresponding to the simulation sets, in a subfolder of the workfolder (Figures 20, 21).
The computation of conserved graphs might be time intensive, and it requires that all input .psf and .dcd files used for the computation of H-bond graphs have remained in the original location.

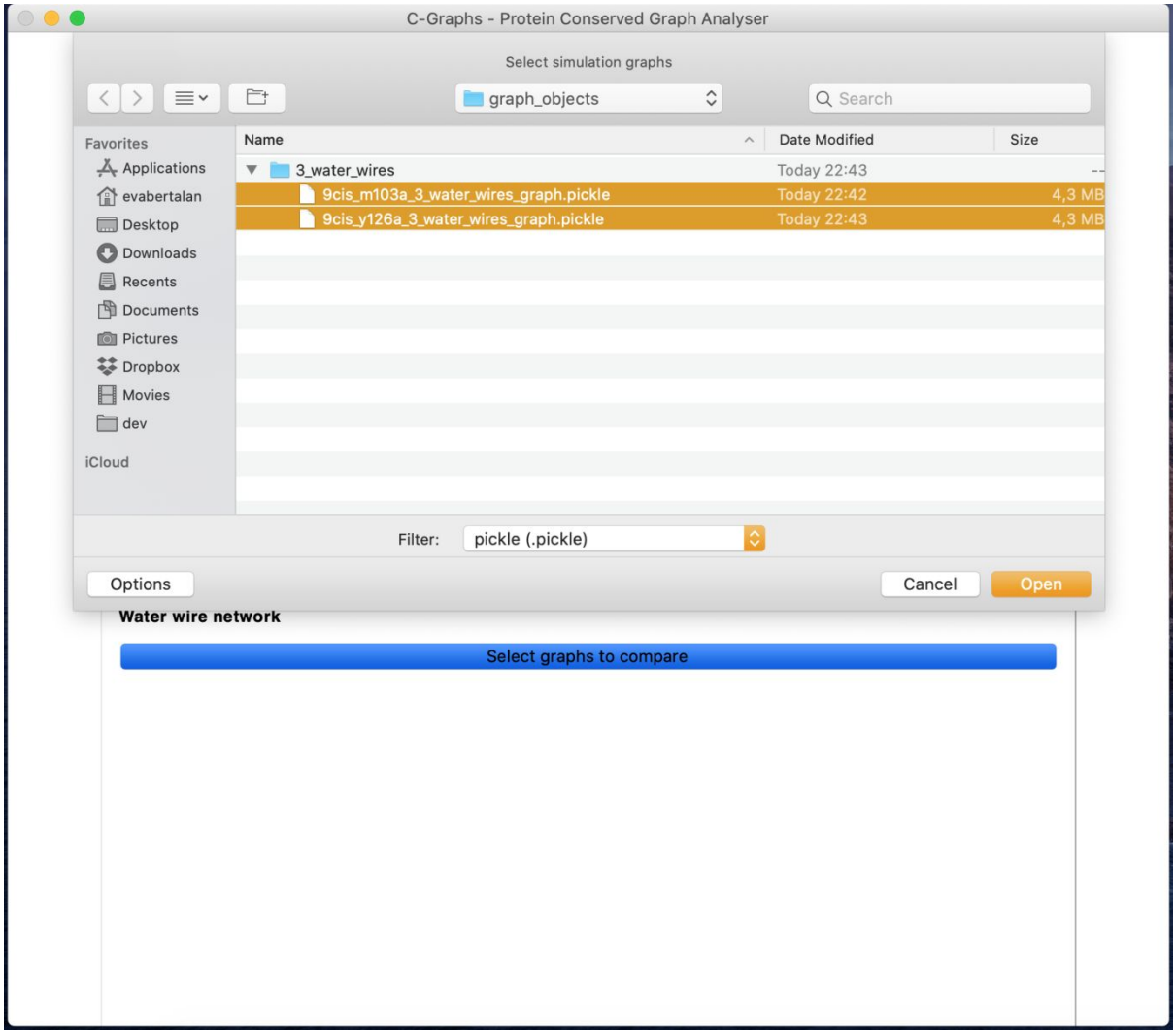

Figure S18. Selection of H-bond graphs computed from independent MD simulations. The computation can also be performed for only for one simulation, in which case only one subfolder will be generated for the individual graphs, and the conserved water wire network will be the same as the water wire network computed for that simulation.

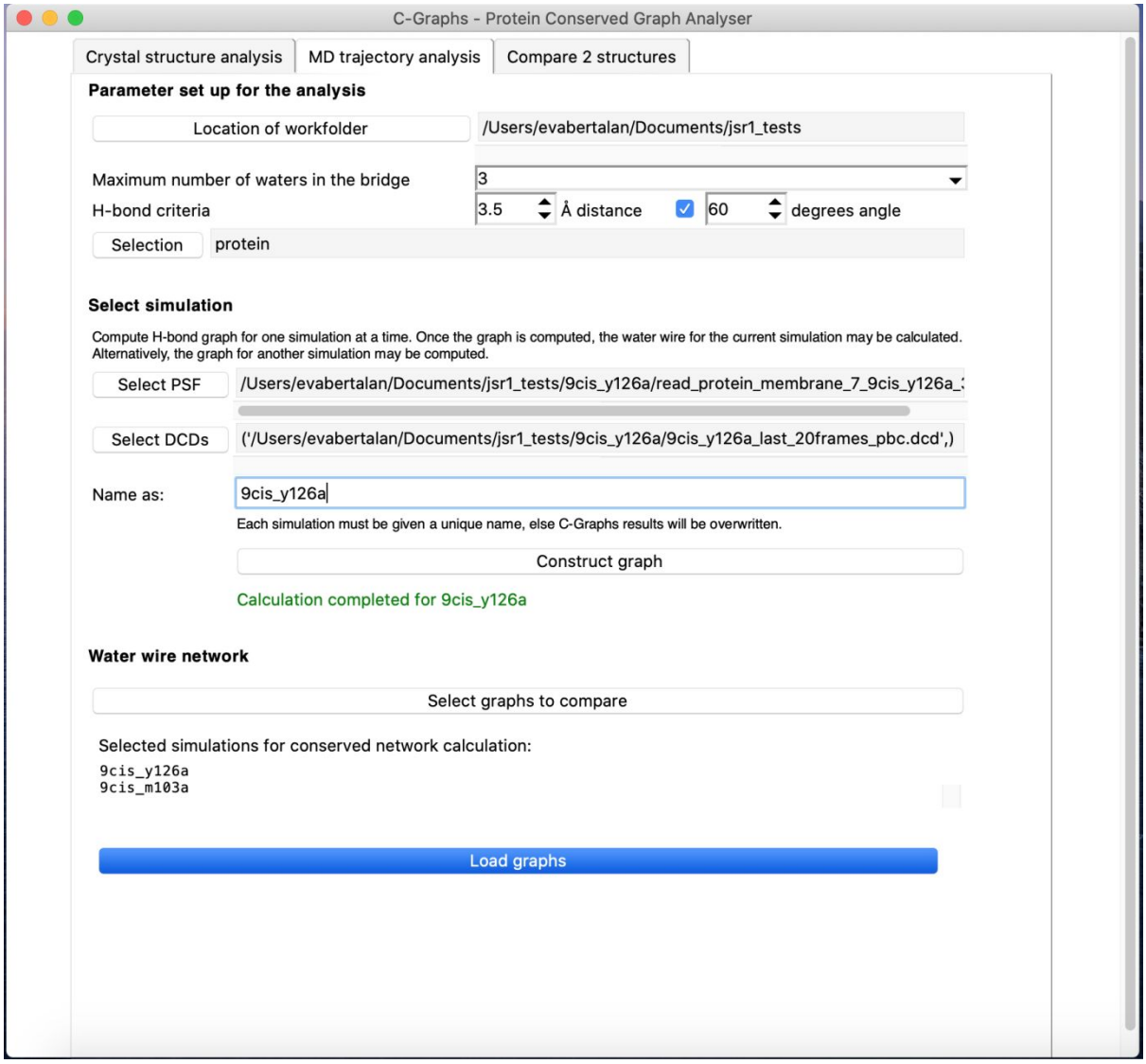

Figure S19. H-bond graphs selected for comparisons are listed in the C-Graphs interface, labeled according to the names assigned to the simulation sets. The computation of H-bond graphs is initiated by pressing the button *Load graphs*, when the pre-calculated graphs of the individual simulations are loaded into the program, to enable further analysis.

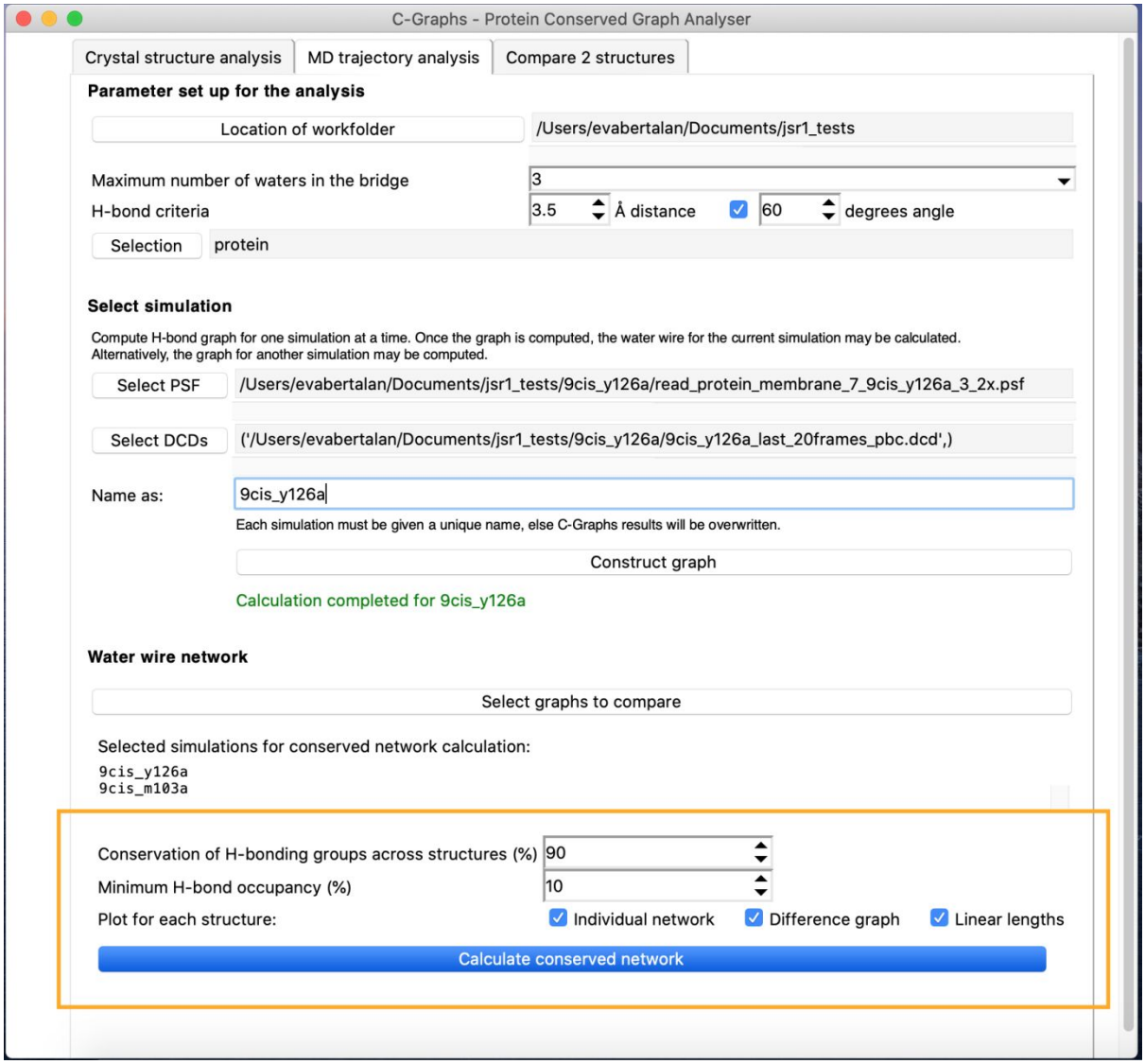

Figure S20. Options for generating data plots for conserved H-bond graphs computed from one or more simulation trajectories. When only one MD simulation is provided here as input, C-Graphs returns the Bridge H-bond graph computed according to the criteria set by the user.
Plot network for each structure generates data plots for the H-bond network computed for each MD simulation set. With Plot difference graph for each structure, C-Graphs computes the difference between the H-bond network of a given structure and the conserved graph. Graph nodes and edges present only in the given structure of interest for analysis are colored blue, conserved amino acid residues and their connections, gray. Same as in the case of crystal structure analysis, the difference graphs show the unique H-bonds present in the given structure, but not conserved across all structures included in the analysis.
Plot linear length of continuous networks for each structure generates plots of the H-bond graphs projected along the *z* axis of the coordinate system.

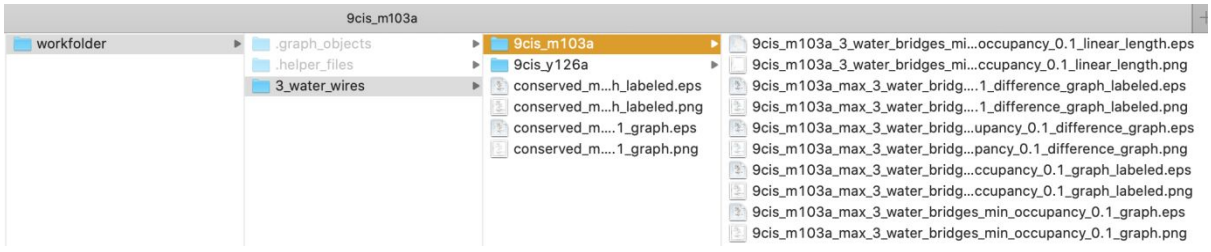

Figure S21. Content of the sub-folder in which results of the conserved networks are saved. All data plots are generated as .png and .eps files, with and without labels for the amino acid residues.

Compare 2 structures

This tab of C-Graph enables direct comparison between the H-bond networks of two structures. The comparison can be performed between two PDB structures (Figure 22), with results stored in a dedicated subdirectory of the workfolder (Figure 23), or between two separate sets of MD simulations of proteins of the same family (Figure 24).

To ease comparison of H-bond graphs, the user can select colors for the nodes and edges that are present only in the graph of that protein (Figure 22). Nodes (amino acid residues) and edges (H-bonds) present in both structures will be colored gray.

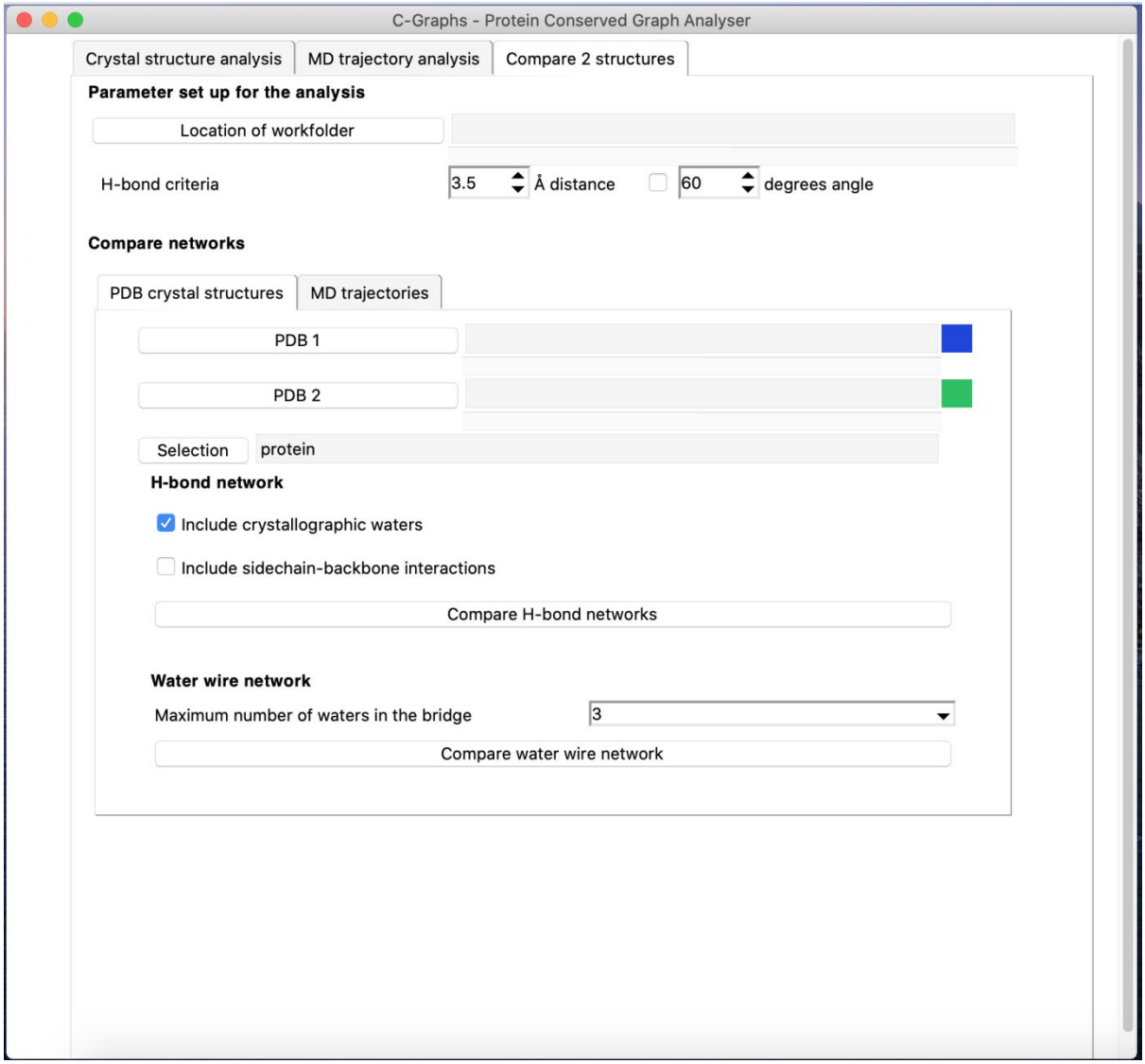

Figure S22. The C-Graph tab to compare two PDB structures. For static PDB structures computation of H-bond graphs and water wires are supported.

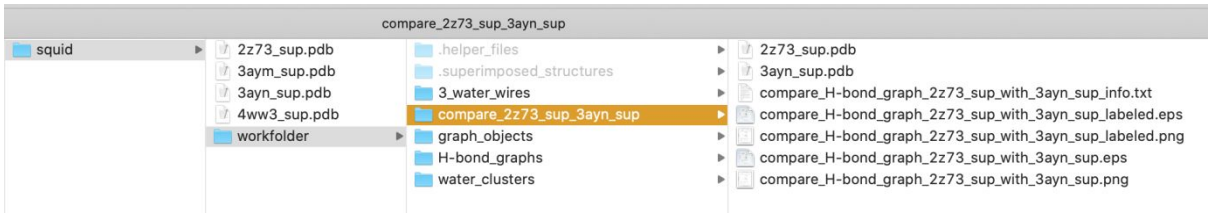

Figure S23. Content of the subfolder, where the results of the structure comparison are saved. The folder is named “compare_pdb1_pdb2” indicating the name of the two selected structures.

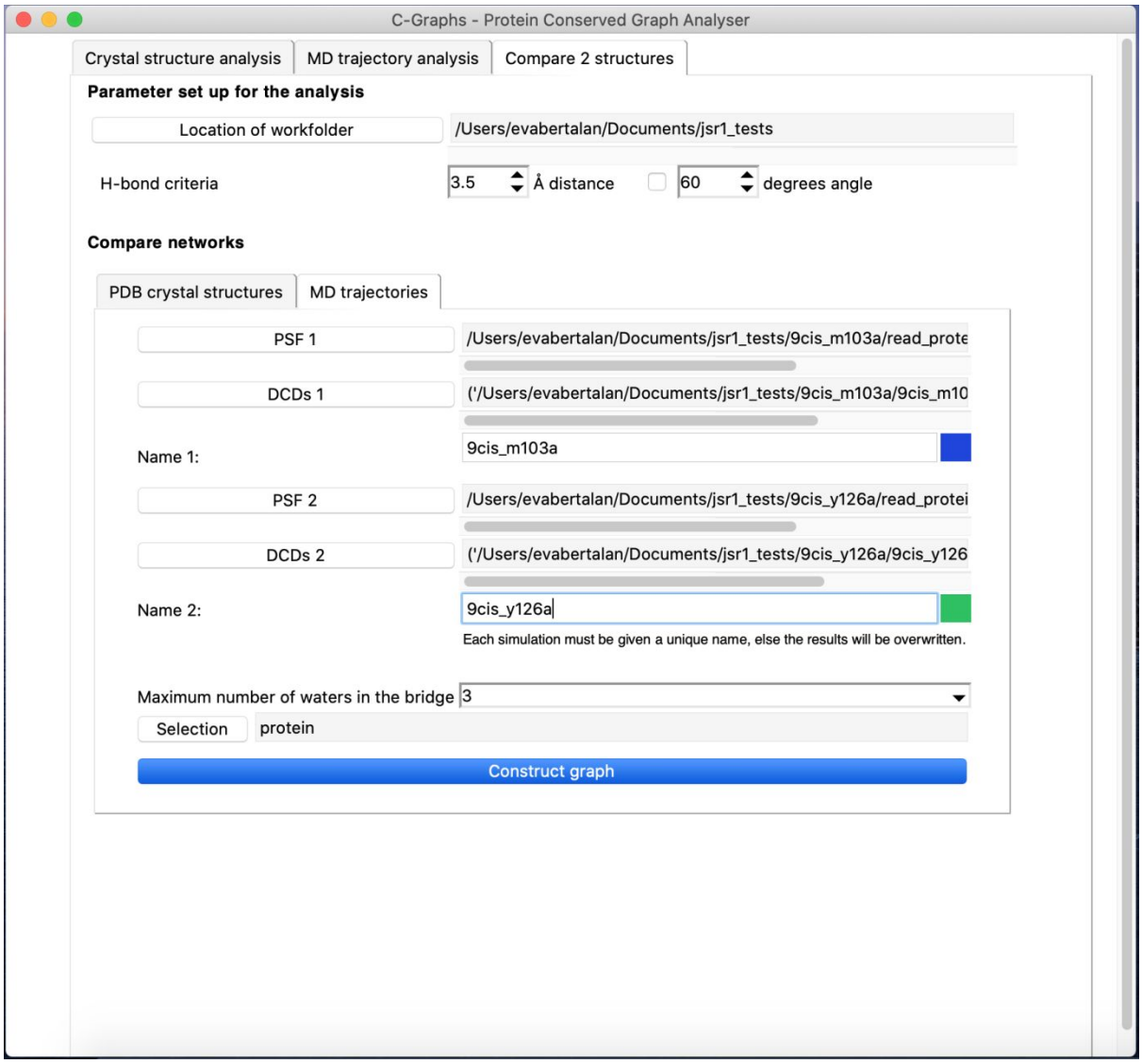

Figure S24. Comparison of H-bond graphs computed from two distinct MD trajectories. C-Graphs will first construct the H-bond graph from each simulation, and it expects here as input the criteria to be used for the H-bond graph.

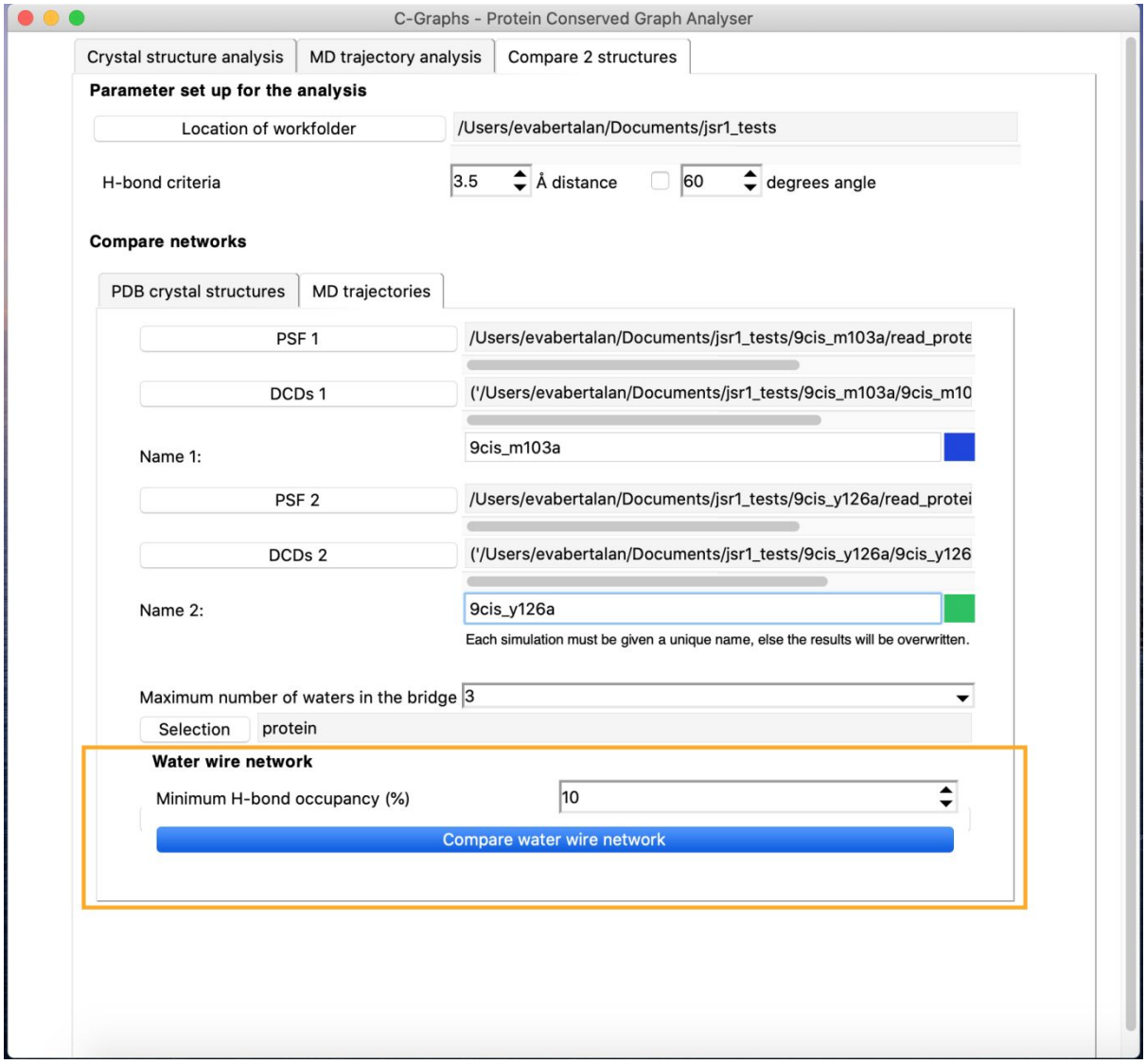

Figure S25. Comparison of the water wires computed for two distinct sets of MD simulations. Water wires are computed separately for the two simulations, and the criteria for water wires need not be the same in the two graphs. C-Graphs will generate a separate set of data plots for each simulation.

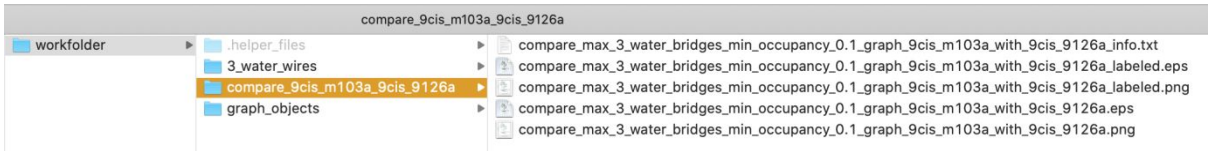

Figure S26. Subfolder for C-Graphs results of MD trajectory comparisons. In the example shown here, the maximum number of water molecules in the bridge was 3. Folder "3_water_wires" > "compare_sim1_sim2" contains plots that compare H-bond graphs computed from sim1 and sim2. Each plot is provided with and without labels for the amino acid residues, and in both .png and .eps formats.

Additional files generated by C-Graphs

Log file. For each C-Graph run a log file is generated and saved in the hidden folder *workfolder/.helper_files* under the name of *cgraph_logs.log*. This log file contains all ending logs of the latest C-Graph calculations.

Graph info file. In the individual folders of each structure, there is a text file created e.g.: *workfolder/H-bond_graphs/3ayn/ 3ayn_H-bond_graph_info.txt.* This file contains information about the number of nodes and edges in the graph and the list of the nodes and edges.

Conserved graph info file. Similarly to the individual graph info file, this file contains information about the number of conserved nodes and edges, and lists the nodes and edges that are conserved. The file is located under
workfolder/H-bond_graphs/conserved_H-bond_graph_info.txt

Edge info file. This is a json object located under
workfolder/.helper_files/name_of_the_structure_water_wire_graph_edge_info.json.
and contains information about the number of waters in each edge of the graph, and the occupancy values of these edges.

References

1. Gaonkar, M. N.; Sawant, K., AutoEpsDBSCAN: DBSCAN with Eps Automatic for Large Dataset. *Intl. J. Adv. Computer Theory & Eng.* **2013**, *2*, 11-16.
2. Bertalan, É.; Lešnik, S.; Bren, U.; Bondar, A.-N., Protein-Water Hydrogen-Bond Networks of G Protein-Coupled Receptors: Graph-Based Analyses of Static Structures and Molecular Dynamics. *J. Struct. Biol.* **2020**, *212*, 107634.
3. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera - a Vizualization System for Exploratory Eesearch and Analysis. *J. Comput. Chem* **2004**, *25*, 1605-1612.
4. Humphrey, W.; Dalke, W.; Schulten, K., VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33-38.
5. Schrödinger, L., The PyMol Molecular Graphics System, Version 1.8. **2015**.
6. Siemers, M.; Lazaratos, M.; Karathanou, K.; Guerra, F.; Brown, L. S.; Bondar, A.-N., Bridge: A Graph-Based Algorithm to Analyze Dynamic H-Bond Networks in Membrane Proteins. *J. Chem. Theory Comput.* **2019**, *15*, 6781-6798.
7. Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Dotson, D. L.; Domanski, J.; Buchoux, S.; Kenney, I. M.; Beckstein, O., MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *S. Benthall and S. Rostrup, Editors, Proceedings of the 15th Phyton in Science Conference, Austin, TX, 2016 SciPy* **2016**, 102-109.
8. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B., MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem* **2011**, *32*, 2319-2327.