How to generate/replicate Suppl. Figure 5:

Suppl. Figure 5. Comparing the performance of CH and AFM in predicting random 5%GSP+5%GSN protein interactions of the Yeast DIP network.

1. Running the code

To run the code, execute the function *run_suppl_fig5* with one of these options:

- Option 1: 1 to generate item with existing results. Usage: *run_suppl_fig5(1).*
- Option 2: 2 to recreate item from original data, involving all required computations. Usage: *run_suppl_fig5(2).*

Here is an overview of the execution of each option. The execution times reported below are measured executing the code in Windows 10 Pro with 256 GB RAM, and AMD Ryzen Threadripper PRO 3995WX 64-Cores CPU with 2.70 GHz. The software environment is MATLAB 2019a.

Option 1: Total execution time **~1 min**

Runs *plot_suppl_fig5* to generate item with existing results located in data folder. In data folder, there is:

- original_data: contains the raw data downloaded from its source.
- matrix: contains the adjacency matrix of the Yeast DIP network and a cell array whose first column contains list of positive protein pairs and second column list of negative protein pairs to test for each simulation (10).
- sparsified_matrices: contains the Yeast DIP sparsified network with 5% of the links removed at random within the GSP set.
- network_similarities: contains CH similarity scores matrices of the sparsified network.
- INTS_outputs: contains the AFM-IS (interface score) for the protein pairs in GSP and GSN sets.
- PITMS_outputs: contains the AFM-piTMS for the protein pairs in GSP and GSN sets.

- table: contains an excel sheet with the Uniprot IDs of mapped proteins in Yeast DIP network.
- Yeast_AFM_output: contains the AFM results of AFM-IS and AFM-piTMS and the timing file for each AFM-modelled protein pair for the best AFM setting we have selected, meaning AlphaFold2-multimer-v2 unpaired+paired.
- script: contains *plot_suppl_fig5* function and its called function *prediction_evaluation*.

Option 2: Total execution time ~**16 min**.

All the results of the following scripts are stored in the directory data_replicated. Below are the different steps to implement the computation:

- In order to download the data, follow the instructions reported in the file *instructions to download the data* located inside the folder "data".
- *create_Scere_DIP_net* : creates the adjacency matrix for Yeast DIP network. Output in "matrix". Total execution time: ~**13 s**.
- *protein_pairs_ints_data_processing* and *protein_pairs_pitms_data_processing*: For each model preset/pair mode combination of AFM, they calculate the average interface score (INTS) and piTM score values respectively for protein pairs in positive and negative sets, and saves the results to text files. Outputs in "INTS_outputs" and "PITMS_outputs". Total execution time: **~1 min.**
- *create_list_GSP_GSN_and_perturbed_net_5perc*: Creates the GSP and GSN sets and perturbed the network by removing 5% of links. Outputs in "matrix". Total execution time: ~**1 min**
- *run_CHA_linkpred5_monopartite_Yeast_net_perturbed*: Computes the CH similarity scores of the Yeast DIP perturbed network. Outputs in "network_similaritites/CH_L2_L3/results". Total execution time: ~**13 min**
- *plot_suppl_fig5*: Plot Suppl. Figure 5. Total execution time: **~1 min**.

2. Required modules for Python codes

The Python version used is 3.9.13 and the list of required modules is:

- pathlib
- itertools
- numpy
- glob
- csv
- re
- sys
- json
- pickle
- itertools