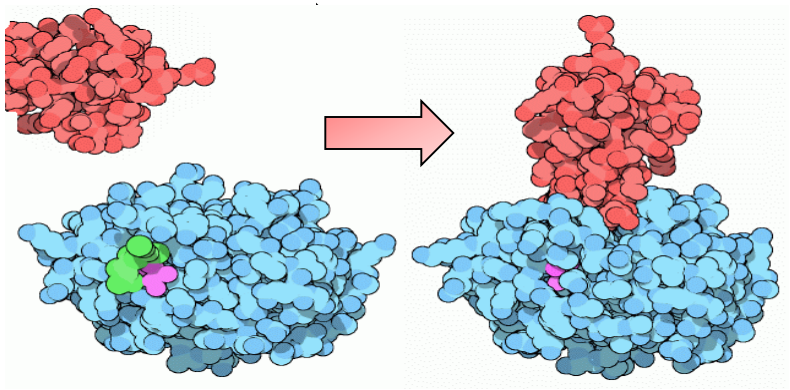
Protein complex prediction by date-hub removal

Iana Pyrogova

Limsoon Wong

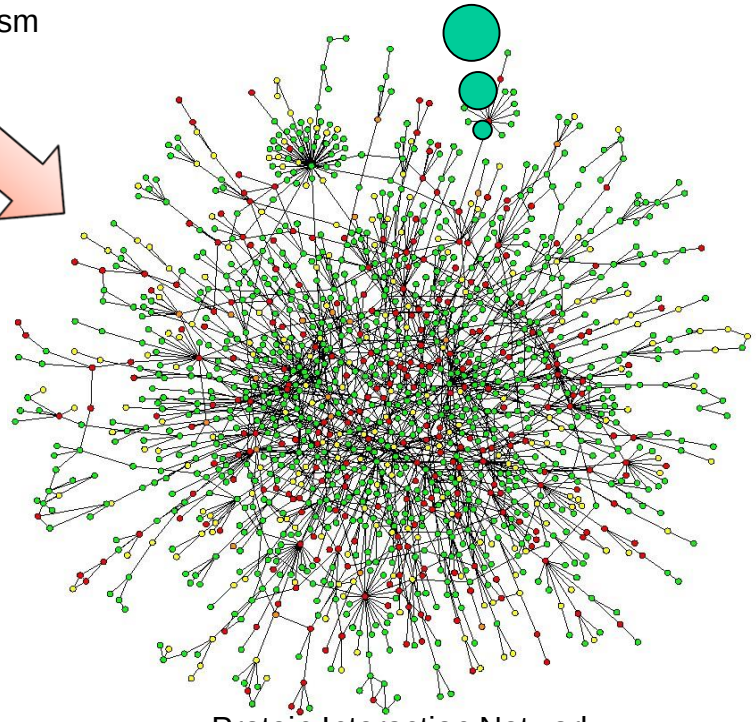
Protein interaction networks



Individual proteins come together and interact

PPIN
Valuable source of knowledge

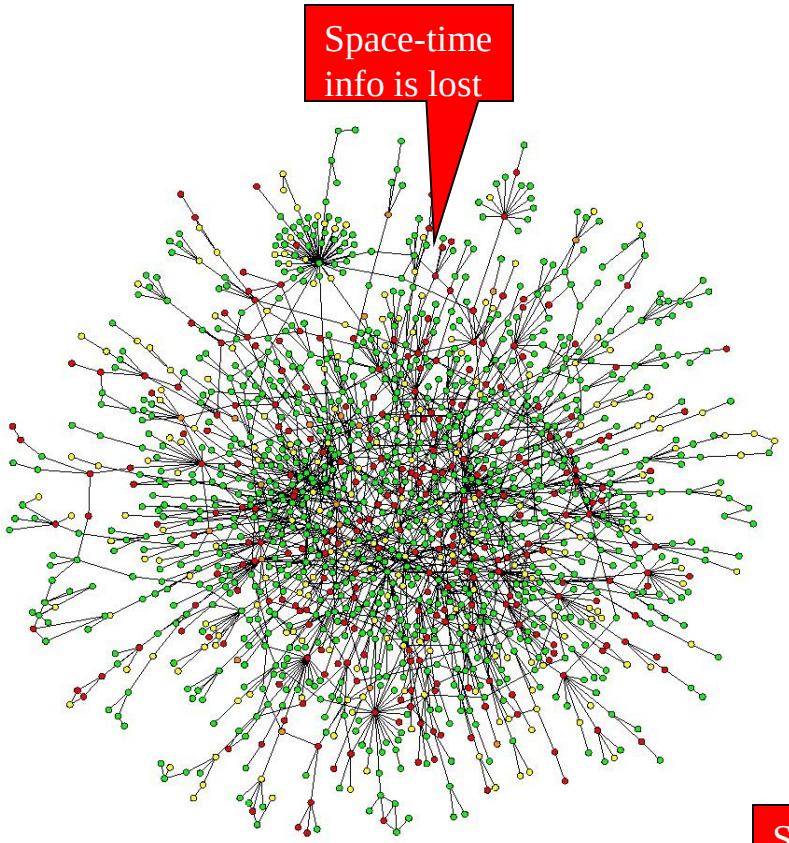
Collection of such interactions in an organism



Protein Interaction Network

- **Proteins come together & interact**
- **The collection of these interactions forms a Protein-Protein Interaction Network or PPIN**

Detection & analysis of protein complexes in PPIN



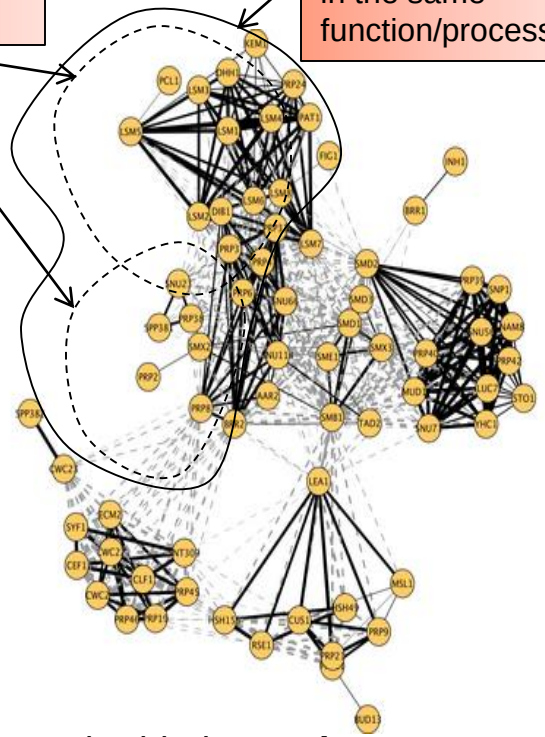
PPIN derived from several high-throughput experiments

Identifying embedded complexes

Space-time info is "recovered"

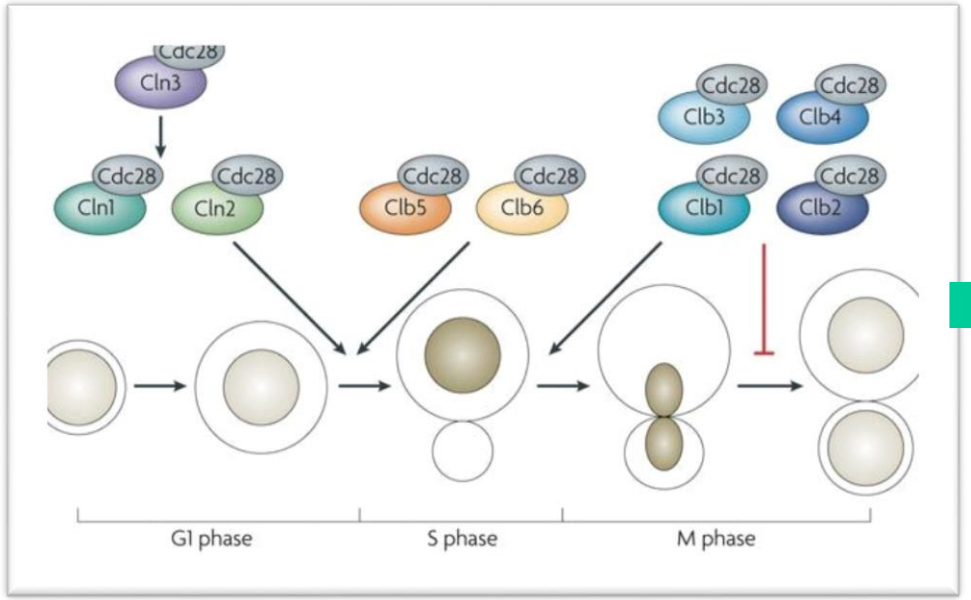
Individual complexes (Some might share proteins)

Entire module might be involved in the same function/process

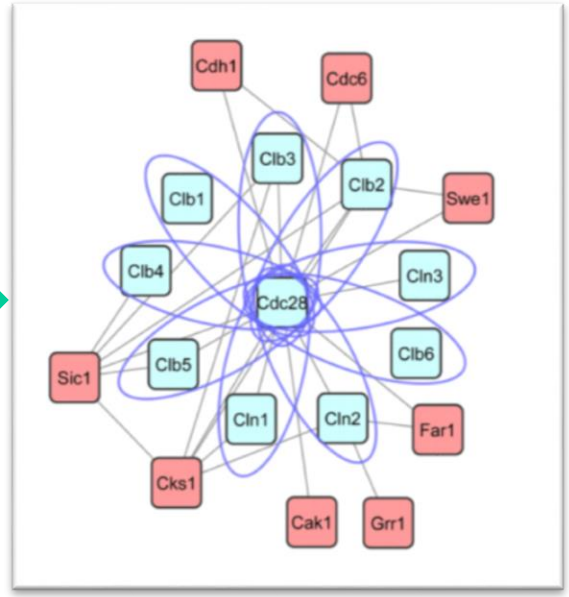


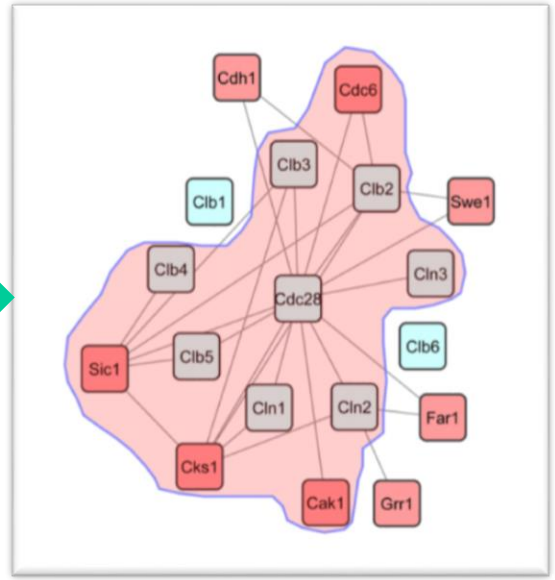
Embedded **complexes** identified from PPIN

A challenge in protein complex prediction



J. Bloom et al. (2007)



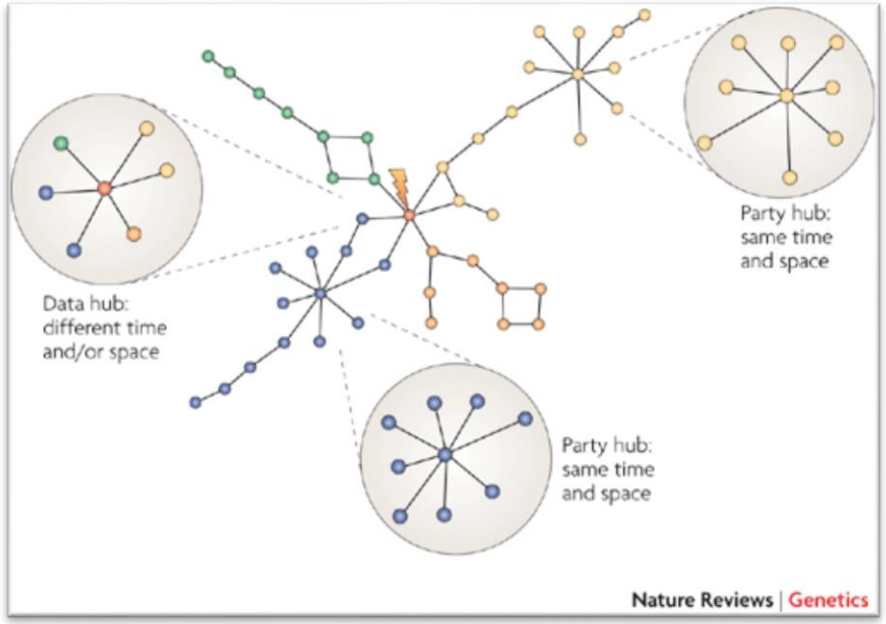


C.H. Yong et al. (2015)

- **It is difficult for protein complex prediction algorithms to identify the overlapping complexes' boundaries**

Date and party hubs

- **Party hub**
  - Interacts with its partners at the same time
- **Date hub**
  - Participates in different complexes at different times or at different locations



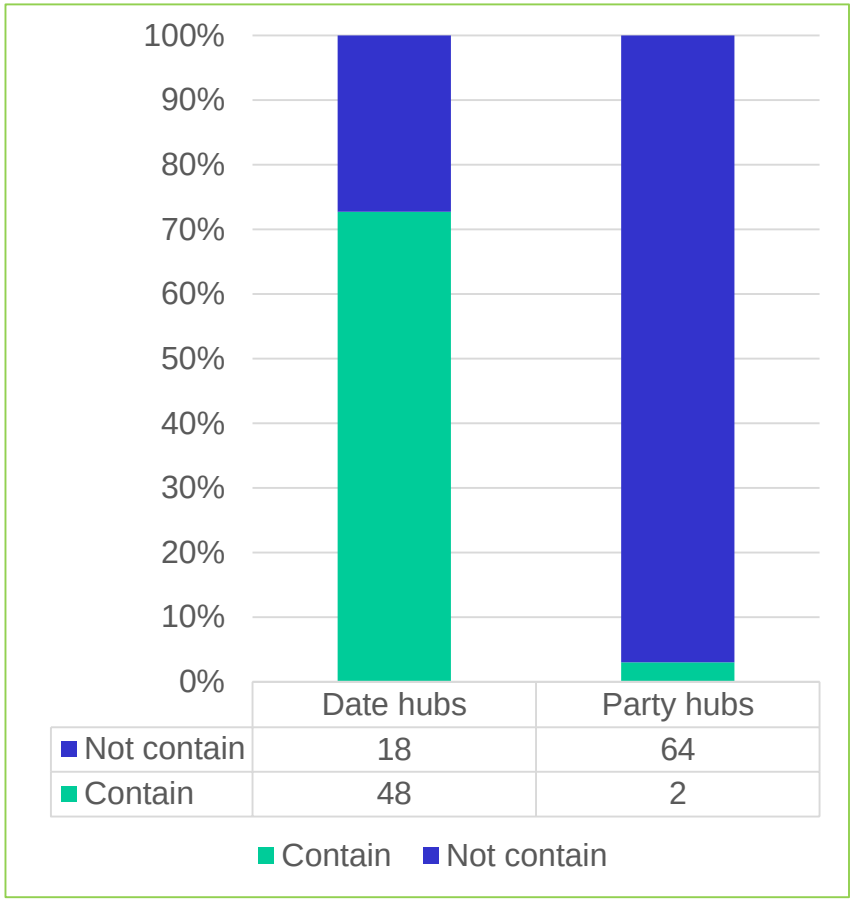
Han et al. (2004)

What roles do date hubs play in complexes? Can we use them to deconvolute overlapping complexes?

Date hubs and overlapping complexes

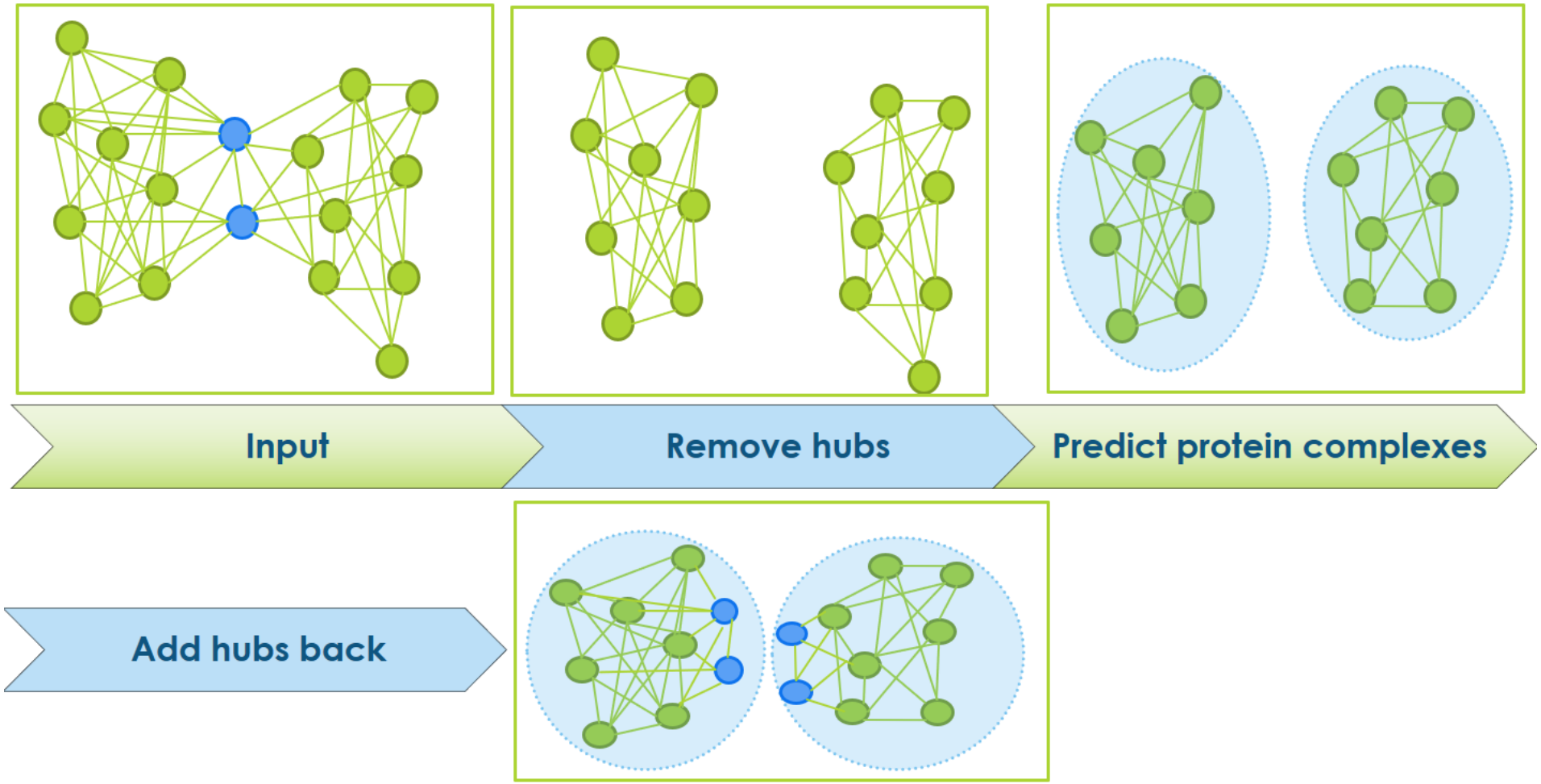
If two reference complexes overlap, then the proteins within their intersection should correspond to the date hubs

Let's test this on 66 overlapping yeast protein complex pairs…



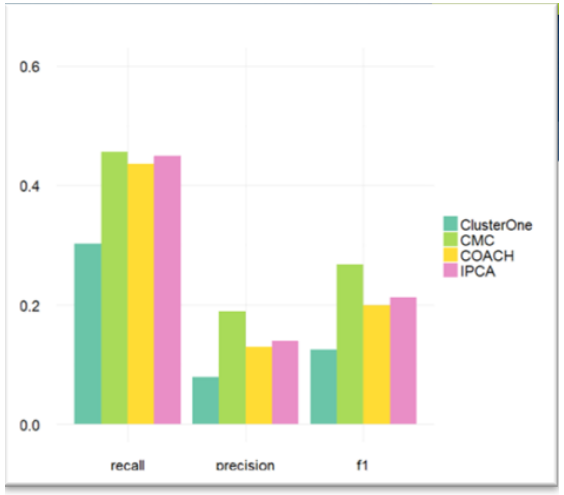
Date & party hubs from Pritykin et al. 2013

Network decomposition by (date) hub removal

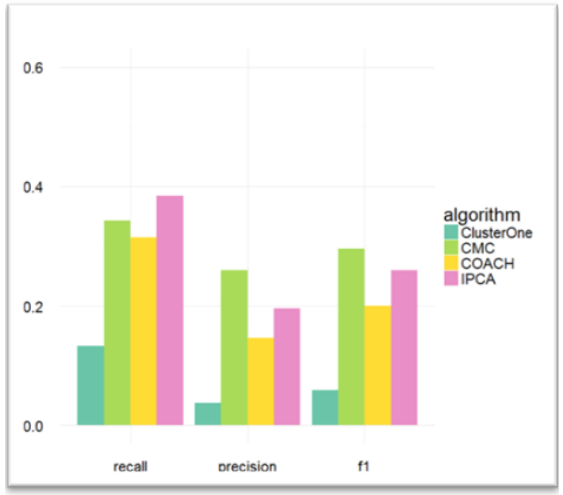


Experiments setup

- **Data sources**
  - PPI networks: yeast, human
  - Reference Complexes
    - **Yeast: 149 big complexes (size≥ 4) from CYC2008**
      - *68 reference complexes are overlapping*
    - **Human: 659 big complexes CORUM (2013)**
      - *90% reference complexes are overlapping*
- **Protein complex prediction approaches**
  - CMC
  - COACH
  - ClusterONE
  - IPCA



Yeast (match_thresh = 0.75)



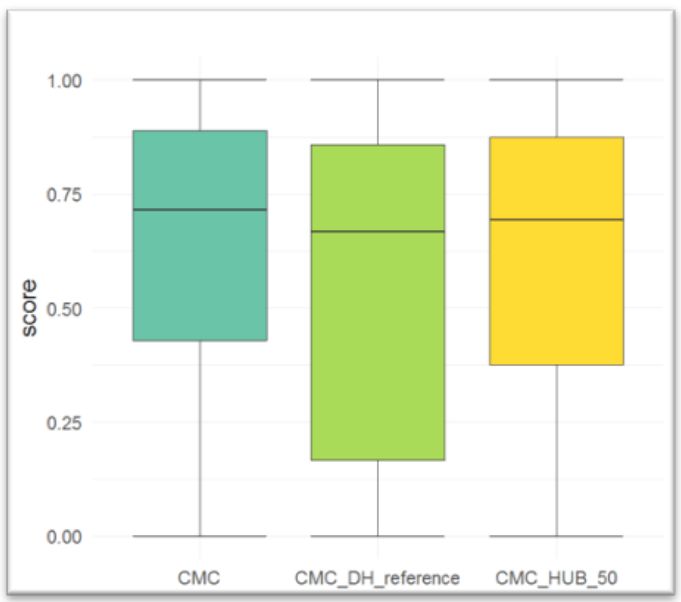
Human match_thresh = 0.5

How do we evaluate predictions?

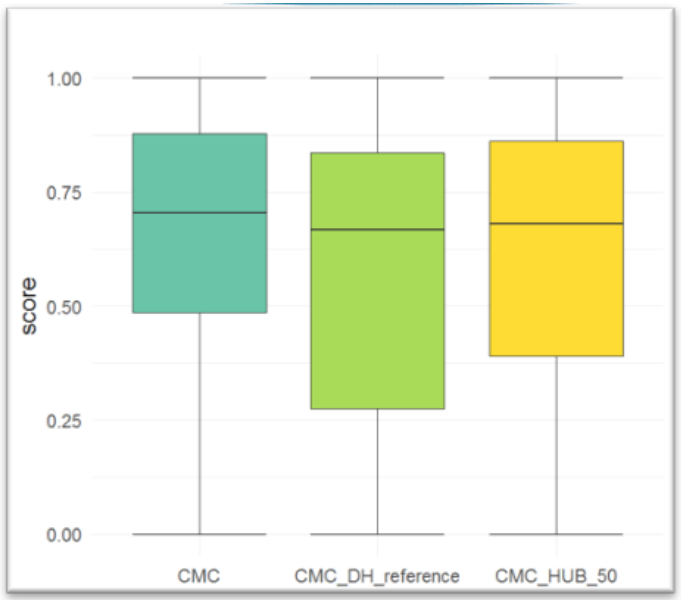
- **A predicted complex is said to match a reference complex when their jaccard coefficient exceeds a threshold**
  - 0.75 for yeast complexes
  - 0.5 for human complexes

- **An effective approach would be characterized by:**
  - High recall and precision values
    - **Precision = matched predictions / total predictions**
    - **Recall = matched complexes / total reference complexes**
  - High a best-match cluster score distribution

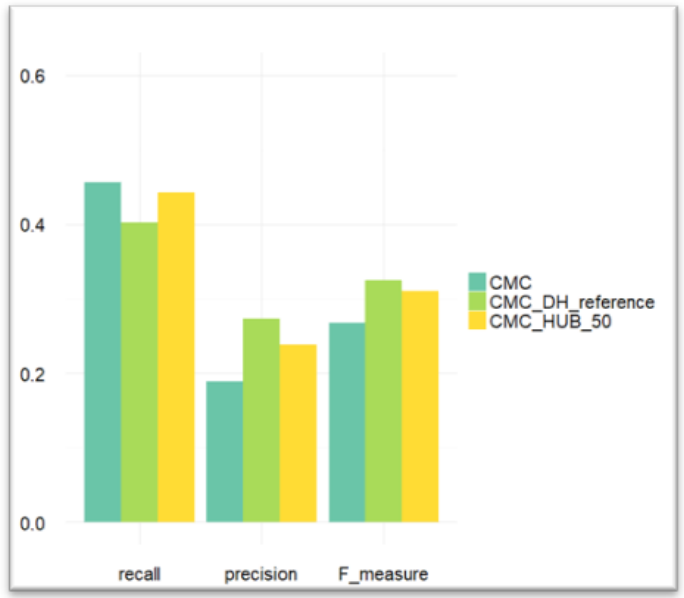
Network decomposition by date-hub removal, in yeast



All real complexes



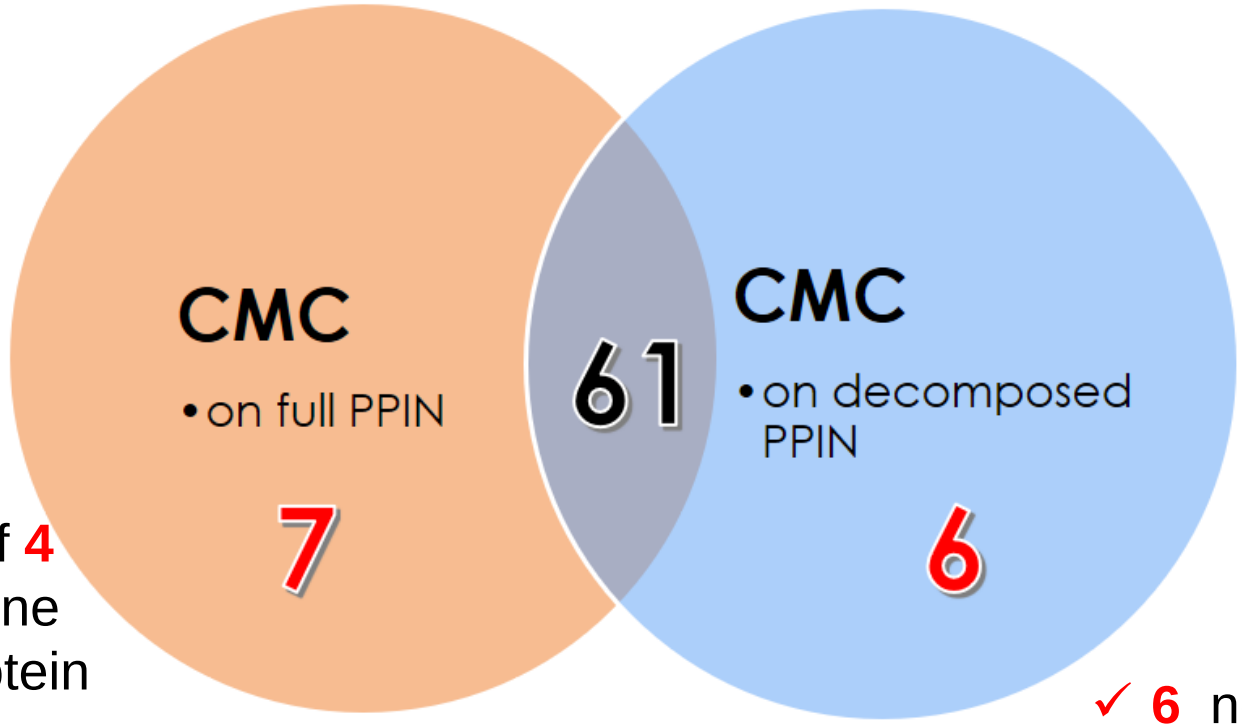
Overlapping real complexes



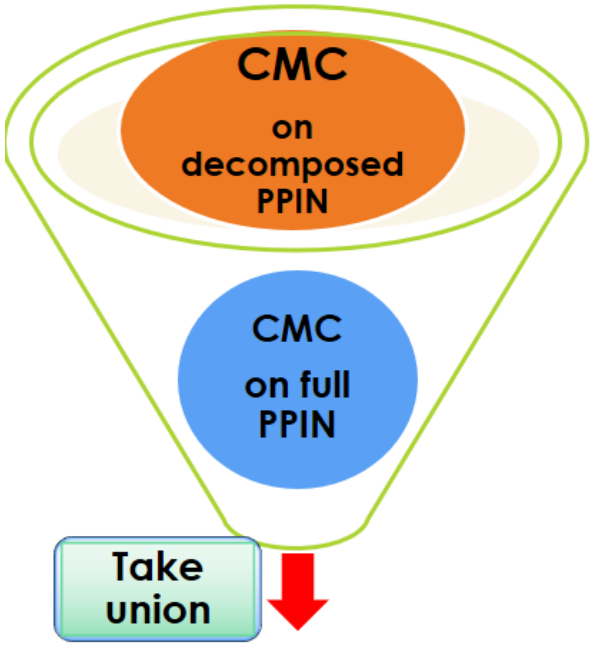
match_thresh = 0.75

- **Observations**
  - Higher precision and F-measure
  - Why do we get lower median best-match cluster score?
    - **CMC is not able to recover some reference complexes after date-hub removal**

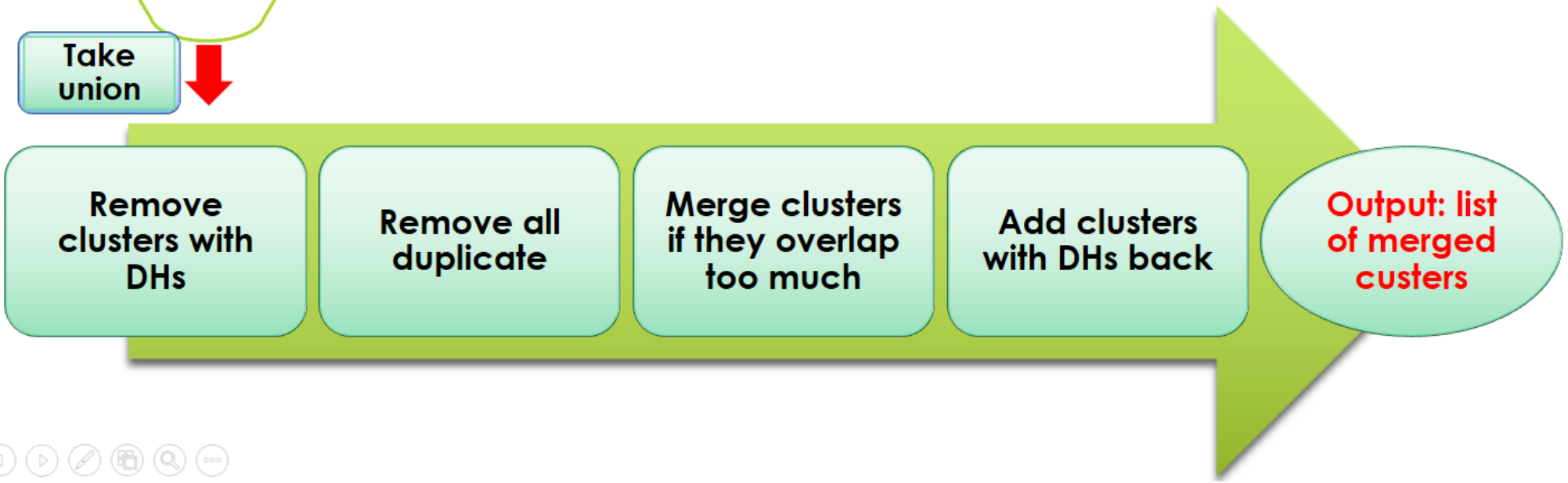
Further investigation



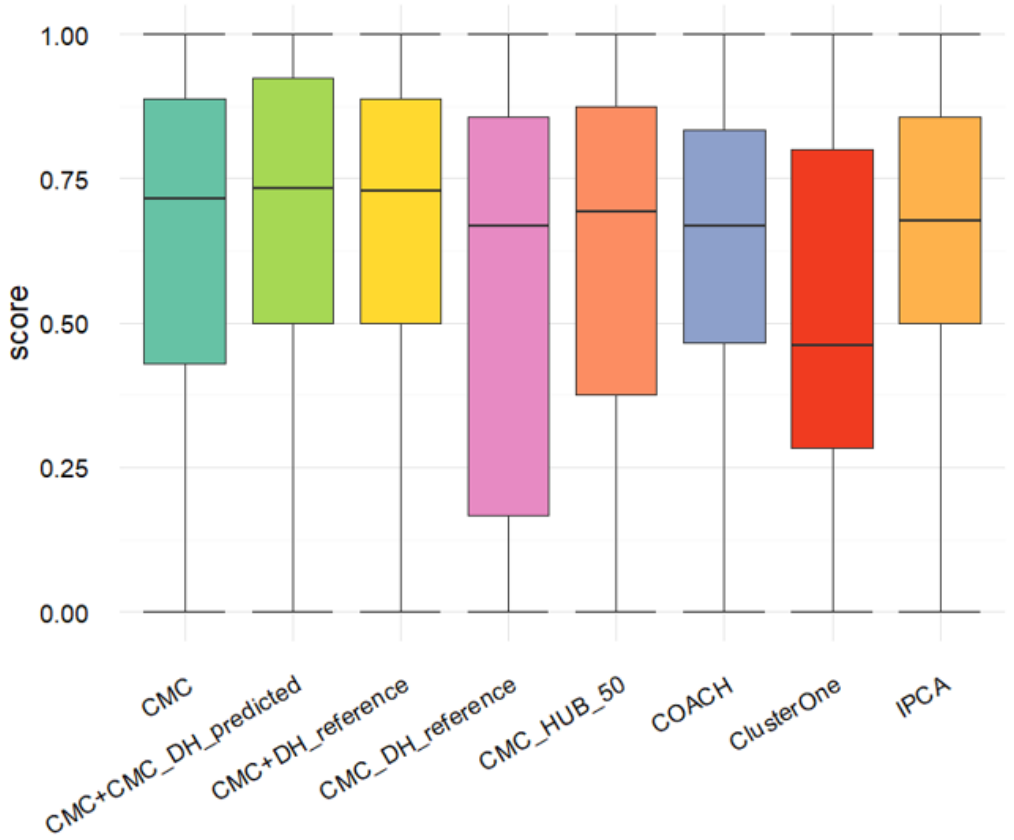
- ✓ **7** complexes consist of **4** proteins with at least one predicted date hub protein
- ✓ Once we decompose the network, CMC is not able to generate clusters of size **3** to predict those complexes
- ✓ **6** new complexes were predicted after network decomposition



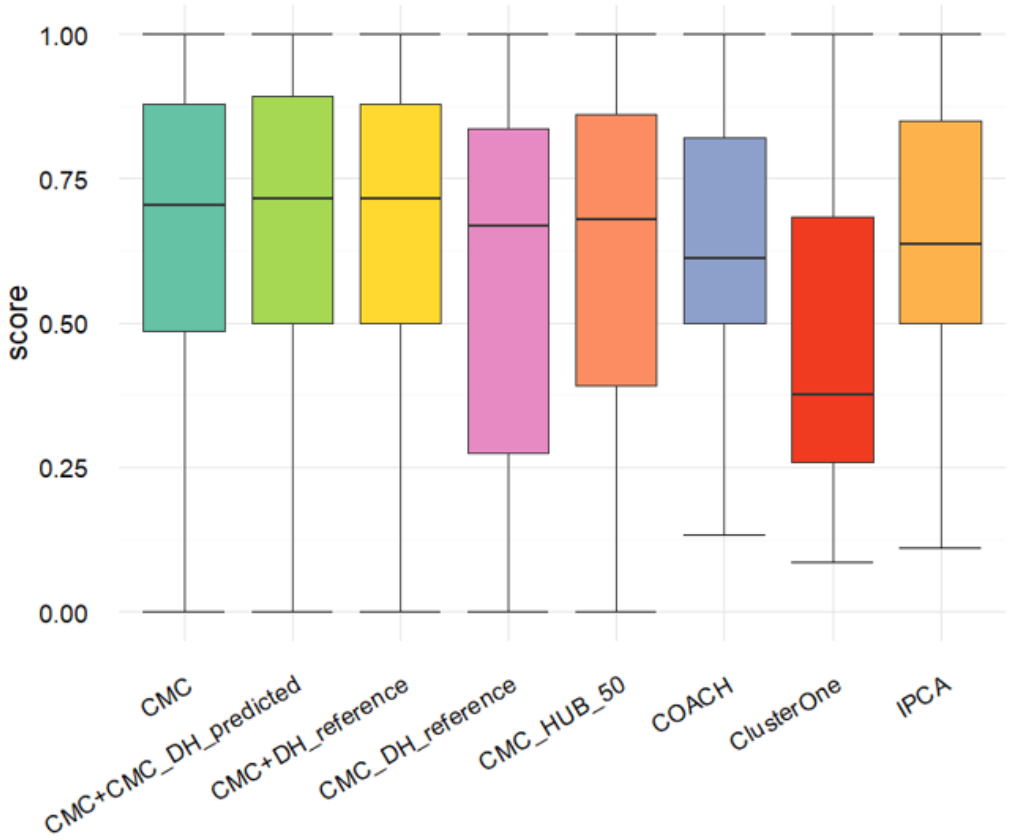
Possible solution: Combine the two sets of predicted clusters to improve the performance of CMC



Results of this “double-barrel” approach, in yeast

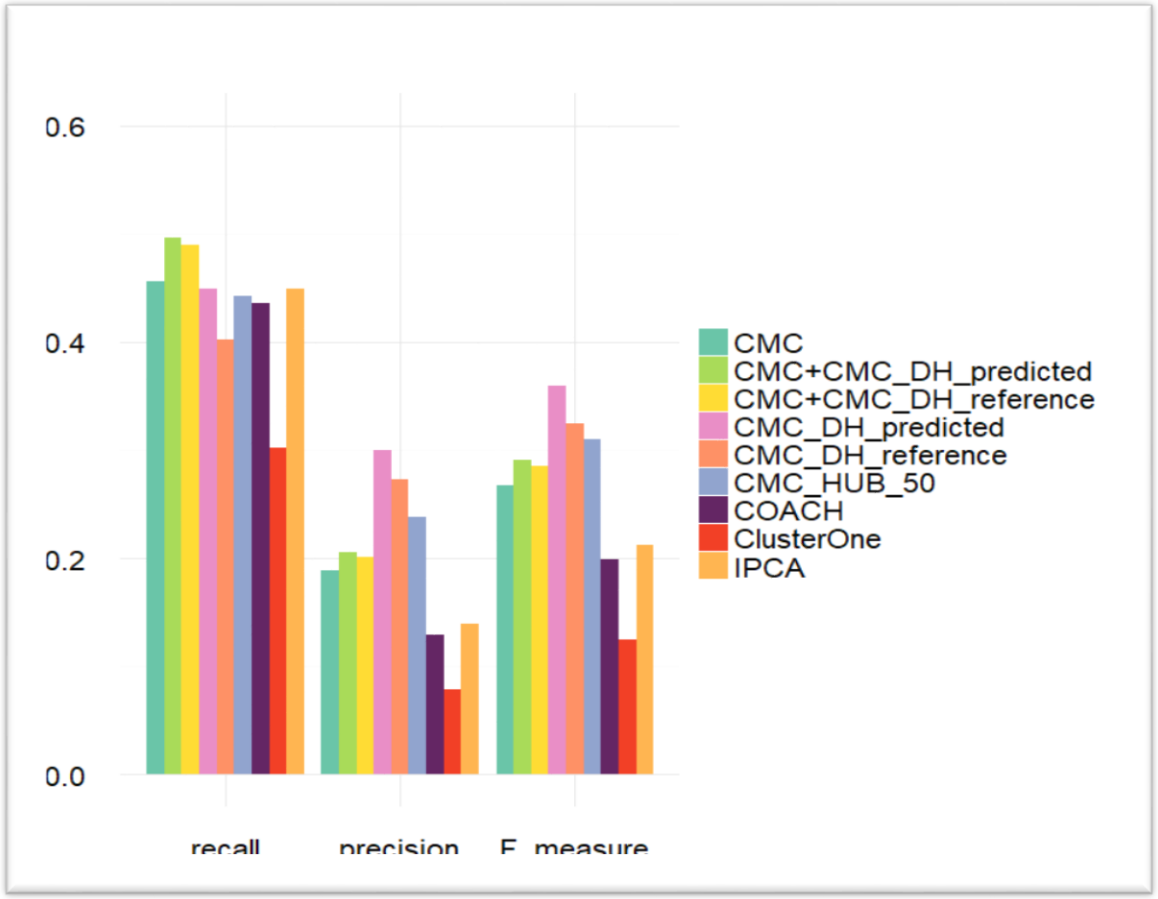


All real complexes



Overlapping real complexes

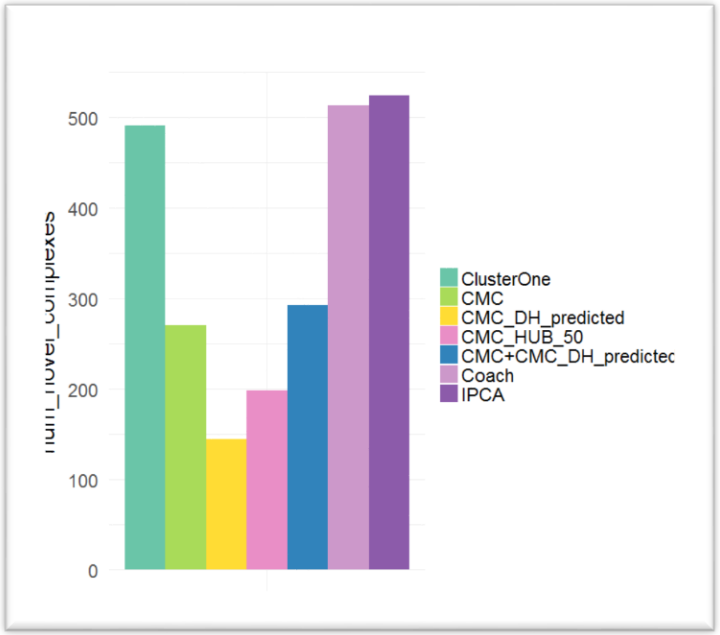
Observation, in yeast

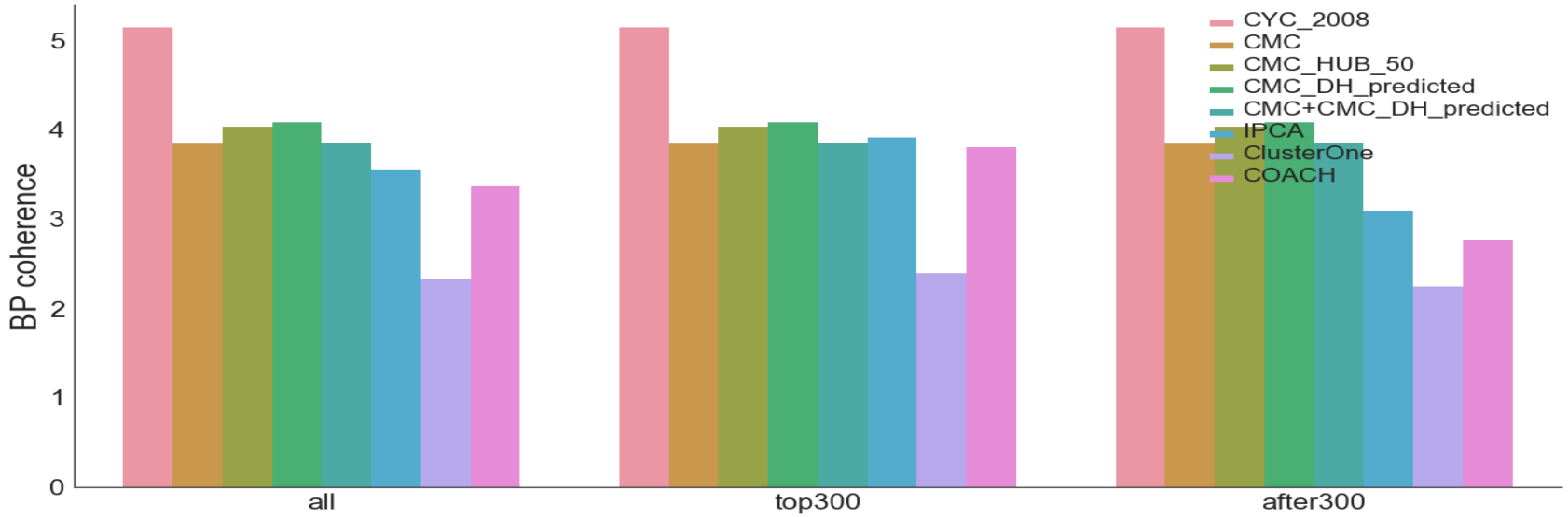


- **Taking the union increases the recall substantially: CMC+CMC_DH_predicted, CMC+CMC_DH_reference**
- **Many predicted clusters may correspond to novel complexes, because the set of reference complexes is incomplete**

Quality of novel complexes predicted

- **Novel yeast complexes are predicted complexes which do not match any reference complex at match-thresh = 0.75**





Summary

1

We confirmed that the date hubs from reference dataset tend to occur within the intersection of real overlapping protein complexes.

2

We observed that CMC benefits much from date hub removal.

3

The distribution of the best match cluster score has the lowest median score

4

We proposed a simple strategy to combine the clusters predicted by CMC before and after we remove date hubs to improve the overall CMC performance.