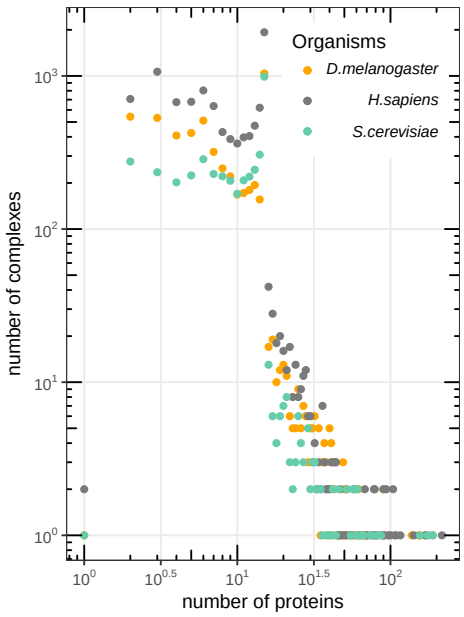
# Appendix: COMPLEAT database

The COMPLEAT database (Vinayagam et al. 2013) provides both protein complex data and a platform for annotation and enrichment of RNAi and other data. To our knoweledge it's the most complete database for protein complexes of *D.melanogaster*, *S.cerevisiae* and *H.sapiens* yet. While analysing the complex data we discovered an irregularity in the complex size distridution. The authors didn't mention this irregularity which is apperent in Figure .1a.There is a gap in the distribution between 15 and 16 number of proteins of complexes size for all organisms (Figure .1a). This gap disappears in the reverse distribution which is the only distribution that was published by the authors, i.e the number of complexes that each protein participates in (Figure .1b). This pattern looks like a phase transition which if it was true then it would have huge biological meaning. But with a more thorough look we discovered the source of this irregularity (Figures .1d and .1c).

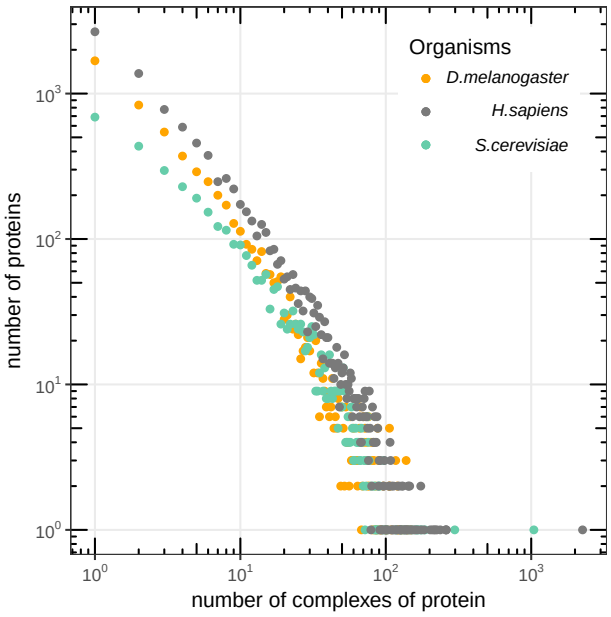
Table .1: Summary of COMPLEAT database for *D.melanogaster*

| Source | Complexes | Proteins |
|---|---|---|
| Literature (326 distinct experiments) | 2045 | 4501 |
| NetworkBlast | 2893 | 3525 |
| CFinder | 389 | 1362 |
| Total | 5327 | 5786 |

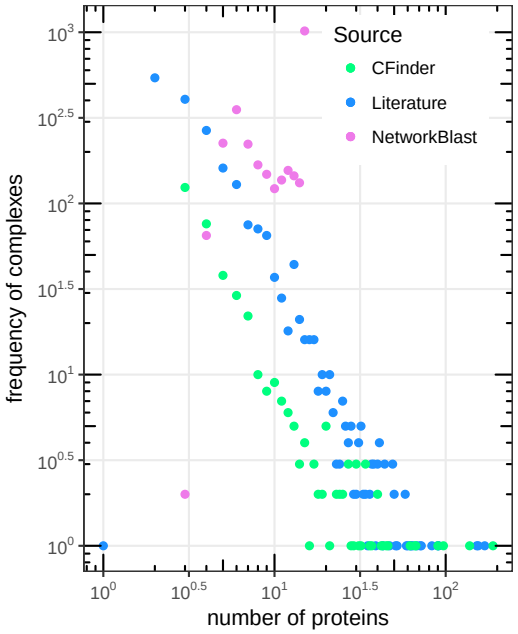
The complexes are provided by 3 different approaches, literature from specific experiments both small scale and highthrouput and computationally inferred from CFinder and NetworkBlast algorithms (Kalaev et al. 2008). In table .1 we see that half of the complexes are provided from the NetworkBlast algoritm. This tool has a plateau of 16 proteins as maximum complex size (Figure .2) although the other methods show a heavy-tailed distribution to complexes size. This creates a bias towards medium sized complexes that is reflected to other analysis like the modular essentiality discussed here (Figure **??**). Further investigation is needed to determine if this bias of NetworkBlast is due to authors' implementation of NetworkBlast or the algorithm has an inherent bias towards medium sized protein complexes.



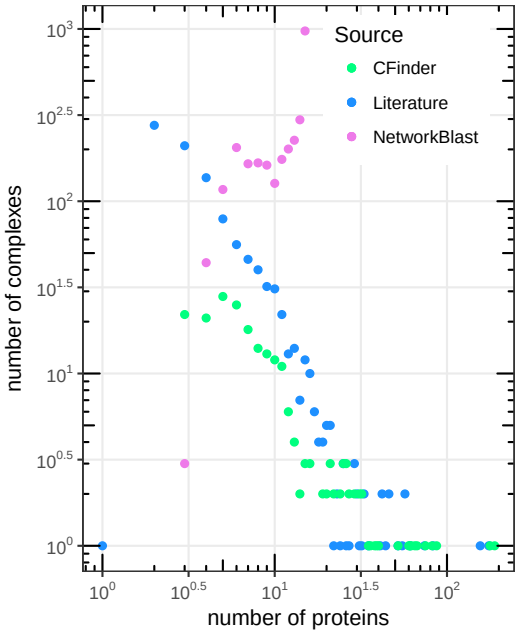

(a) Complexes size distribution.



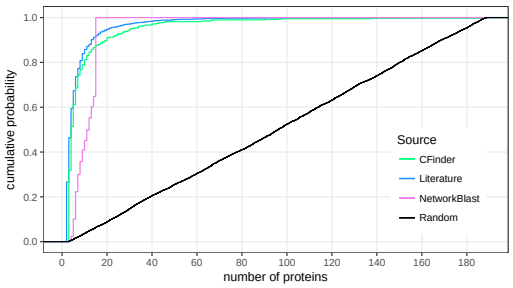

(b) Proteins participation in complexes distribution.




(c) *D.melanogaster* complexes size distribution with different methods.




(d) *S.cerevisiae* complexes size distribution with different methods.




(e) COMPLEAT database proteins cumulative distribution of *D.melanogaster* based on inference methods.

Figure .1: COMPLEAT2 database distributions.

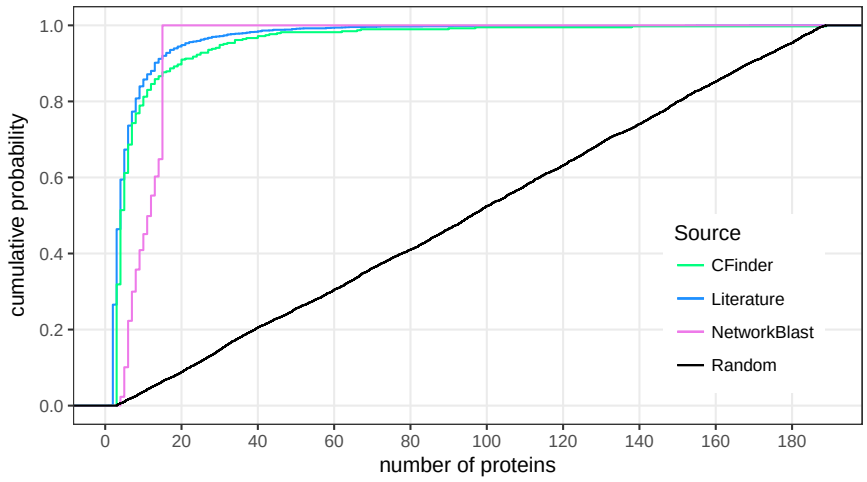



Figure .2: COMPLEAT database proteins cumulative distribution of *D.melanogaster* based on inference methods.

# Appendix: Network contraction with complexes

## Complexes in the signed network

Which of these protein complexes are present in our data set? To answer this question we annotated the signed network proteins with complexes data. Most complexes asre missing proteins in the interval [0.10] (Figure .3) which is expected since most complexes are small (Figure .1a). We found that 585 complexes were complete (Figure .4).

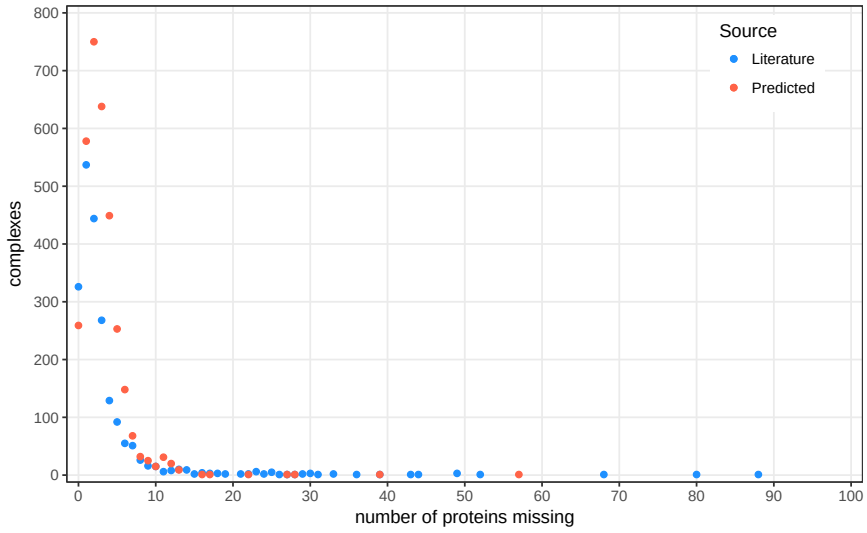



Figure .3: Histogram of the missing proteins of complexes when comparered to the signed network.

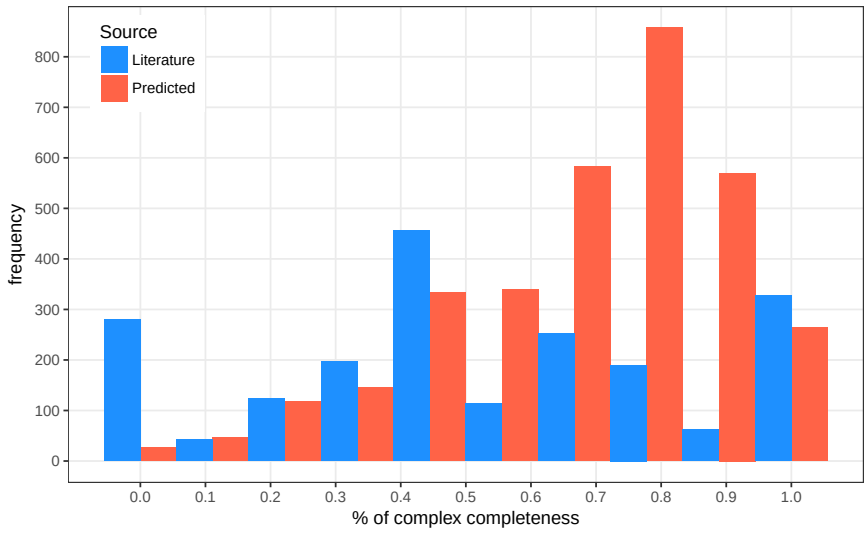



Figure .4: Histogram of the percentage of proteins that appear in the signed PPI network per complex.

## Network contraction with complexes

Since the complexes are molecules that facilitate the processes of the organisms and not the individual proteins it is very important to construct networks with complexes interactions. This requires experimental procedures and computational tools that can change the resolution to complexes scale. Scalability is one of the main goals for network science in the following years. To contract network with complexes from the protein - protein interaction network it is necessery to determine which complexes to use. The rule we applied in this instance is to use only the complexes that all of their proteins are present in the network. This resulted in 585 complexes. Others can use a different threshold, like to use complexes that have >80% of their proteins present. Or take a completely different approach, like using GO annotation in the original network for the selection of complexes or applying clustering methods in protein networks like linked communities (Ahn, Bagrow, and Lehmann 2010; Kalinka and Tomancak 2011).

These 585 complexes contain 1063 proteins which have 2123 interactions in the signed network. between these. So the 1/3 of the signed network is used. After we created the complexes network, two complexes are interacting if their proteins interact in the signed network. We got a network that contained duplicated edges and self loops which we deleted. There were multiple edges between complexes, we kept those that were distinct in the signed network.

In order to keep as much information as possible so treated positive and negative edges independently. More specifically, from all the redundant edges with the same direction, we kept 2, one positive and one negative. The weight of the positive edge and negative edge will be the normalized weight from all the positive and negetive edges, respectively. Finally we normalized all the weights with the absolute value of the maximun weight, in order to have all the edge weights in the [-1,1]. This methods resulted in a very dense network (table .2).

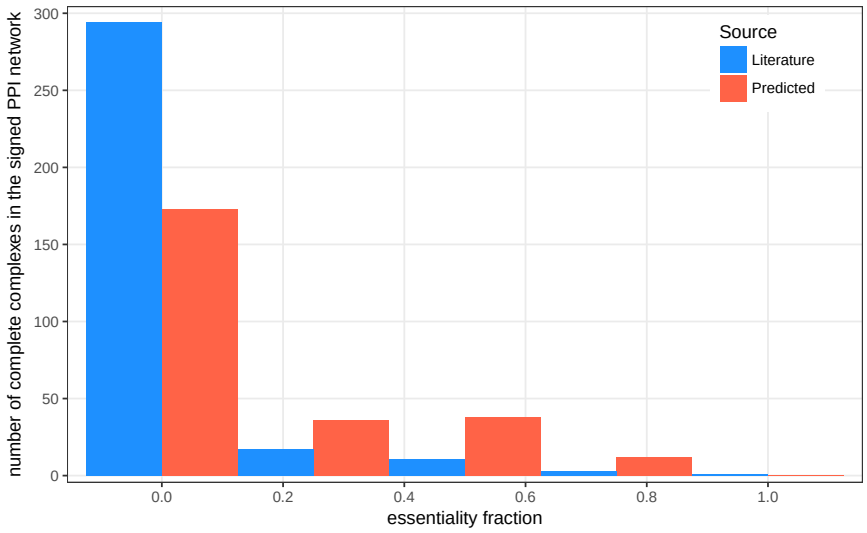



Figure .5: Histogram of the essentiality fraction of the complexes that have all of their proteins in the signed PPI network. Forty nine protein complexes, from the 585 complexes that are complete in the signed network of drosophila, consist of 50% or more essential proteins.

Table .2: This is a summary of the network between complexes based on the signed PPI network. Unique edges are those between complexes that

| Type | Total |
|---|---|
| Positive edges | 14269 |
| Negative edges | 6081 |
| Total | 20350 |

Ahn, Yong-Yeol, James P Bagrow, and Sune Lehmann. 2010. "Link communities reveal multiscale complexity in networks." *Nature* 466 (7307). Nature Publishing Group: 761–64. doi:10.1038/nature09182.

Kalaev, Maxim, Mike Smoot, Trey Ideker, and Roded Sharan. 2008. "NetworkBLAST: Comparative analysis of protein networks." *Bioinformatics* 24 (4): 594–96. doi:10.1093/bioinformatics/btm630.

Kalinka, Alex T., and Pavel Tomancak. 2011. "linkcomm: An R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type." *Bioinformatics* 27 (14): 2011–2. doi:10.1093/bioinformatics/btr311.

Vinayagam, A, Y Hu, M Kulkarni, C Roesel, R Sopko, S E Mohr, and N Perrimon. 2013. "Protein complex-based analysis framework for high-throughput data sets." *Sci Signal* 6 (264): rs5. doi:10.1126/scisignal.2003629.