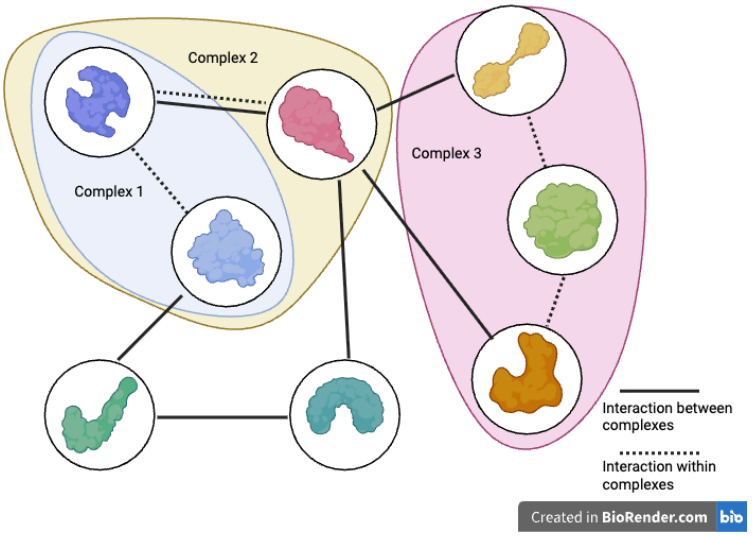
Attention: This project can have some collaboration with the project with *'DIGGER for intra-protein-complex interactions'* towards the end. If you are applying with a friend (with whom you work well together) and who wants to take the other project, please indicate this in your application.

Project Proposal: Investigate inter- vs. intra-protein complex interaction biases in existing PPI databases

Supervisor: Judith Bernett

Background and Motivation

Existing protein-protein interaction (PPI) databases (e.g., STRING, BioGRID, HIPPIE) annotate pairwise interactions, yet most proteins function as part of multi-subunit complexes. Importantly, the type of interaction captured depends heavily on the experimental method used to identify it. Yeast-2-Hybrid (Y2H) assays detect direct binary interactions between protein pairs and are particularly sensitive to weak and transient interactions. Affinity Purification coupled with Mass Spectrometry (AP-MS), on the other hand, preferentially captures stable interactions and can therefore also identify functional associations where proteins are not in direct contact but co-occur within the same complex or reaction.



We want to investigate what kind of interactions are stored in our PPI databases: do they mostly occur within complexes, between complexes, or is it somehow balanced? To this end, we want to build hypergraphs of protein complexes (as illustrated on the side) and analyze the topological properties. We want to compare these properties across databases and experimental methodologies to learn something about the biases.

Concrete steps:

1. Retrieve PPI data from multiple databases (e.g., STRING, BioGRID, HIPPIE), including experimental method annotations. Get protein complex data from resources like CORUM[1] or ComplexPortal[2].
2. Construct hypergraphs from the complex annotations for the different databases and experimental methodologies, and classify whether PPI edges are intra- or inter-complexes, or both.
3. Analyze graph properties like degree distributions (is the functional form of the degree distribution preserved, e.g., are both distributions power-law distributed), connectivity of individual proteins (are hub proteins preserved), and overall connectivity (diameter, number, and connected components), and compare them across databases and experimental methods.

1 Steinkamp et al. CORUM in 2024: protein complexes as drug targets, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D651–D657, https://doi.org/10.1093/nar/gkae1033, Website: https://mips.helmholtz-muenchen.de/corum/

2 Balu et al. Complex portal 2025: predicted human complexes and enhanced visualisation tools for the comparison of orthologous and paralogous complexes. Nucleic Acids Research 53, D644–D650. https://doi.org/10.1093/nar/gkae1085, Website: https://www.ebi.ac.uk/complexportal/home