# ABSTRACT

The structure and the function of the cell arise from interactions between molecules inside and outside it. Investigation of those interactions is, therefore, necessary (but not sufficient) to obtain the understanding of the living systems. Gathering interaction data and depositing it in a central database for open and easy access by the research community can and does accelerate progress in the field. An attempt to unify interaction data from multiple sources meets many challenges, such as standardisation of annotation, interoperability with other resources and time it takes to manually gather the data. As new network analysis methods become available and research group start to use molecular interaction data to make inferences, generate new hypotheses and explain the results of transcriptomics and proteomics the coverage of the real interactome by our knowledge and the bias present in the databases begin to influence research results. This motivates the need to identify which proteins have no interactions are available and understand biases in our interaction data. We focus our analysis on the data deposited to the IMEx consortium of primary databases, which includes the IntAct database, the resource supported by our group.

The best coverage we observe is for yeast, *E.coli* and human. Isoform coverage is limited, but still significant for human. We have investigated if IntAct database coverage is biased towards physicochemical properties of the protein and how well described the protein is in the literature. Proteins with no interactions in IntAct are on average smaller, less well-studied overall, have a lower fraction of charged residues and higher mean hydropathy. Next, we have investigated if two high-throughput interaction detection methods (which use distinct strategies) may be biased towards physicochemical properties of the protein (such as mass): AP-MS and two-hybrid. Affinity purification followed by mass spectrometry (AP-MS) seems to capture a higher proportion of larger proteins as compared to two-hybrid methods. By performing enrichment analysis of the molecular function (Gene Ontology) we have found that databases and datasets which contain experimentally derived data are enriched and depleted in the same functional categories. STRING database, which includes computational prediction data, is the least biased. These results can inform literature curation by IMEx consortium teams and data integration efforts.