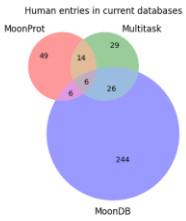
Moonlighting proteins: a steep climb on the tip of an iceberg

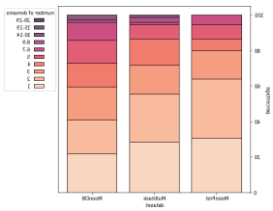
Supplementary Material

Databases of moonlighting proteins (MLPs)

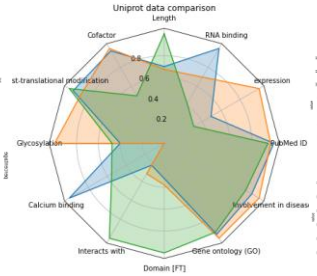
Manually curated: MoonProt; predicted: MoonDB



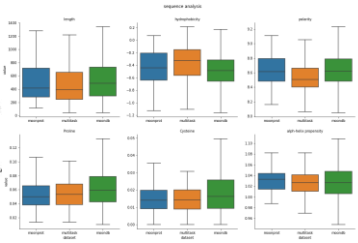
Human Swissprot entries in different databases of moonlighting proteins: Little overlap witnesses MLPs are largely unexplored



Interpro domain number: Barplot representing the fraction of proteins with a specific number of domains



Uniprot annotations: MoonProt (blue), MultitaskProtDB (orange), MoonDB (green).

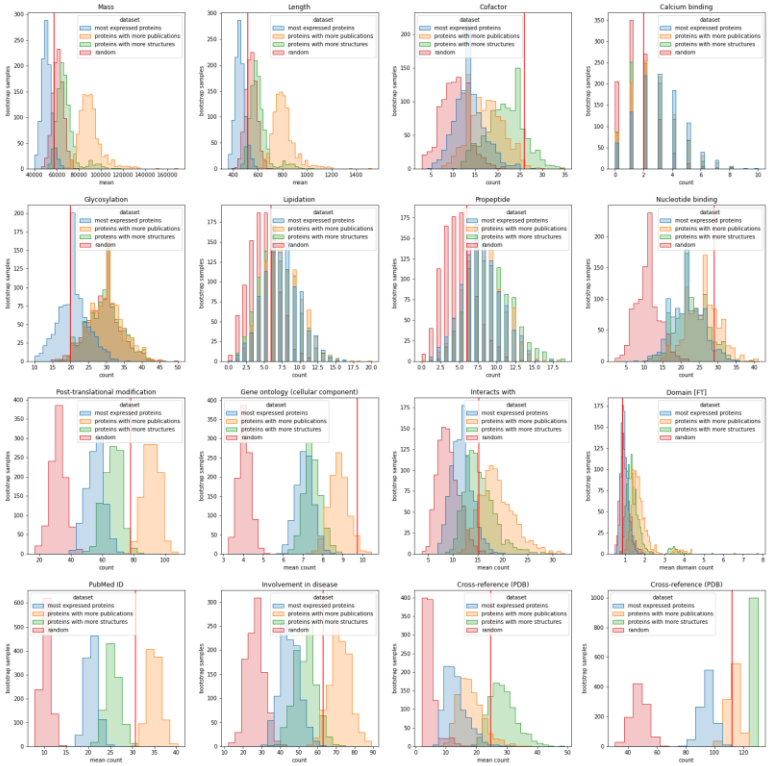


Primary sequence characterisation: MoonProt (blue), MultitaskProtDB (orange), MoonDB (green).

Proteins from MoonDB are longer and with more domains. Therefore, I chose to exclude them from my database of MLPs.

Bootstrap analysis of Uniprot annotations

1. Associate moonlighting database with a statistical parameter of a feature (e.g. mean length).
2. Extract random samples of the same size from the human proteome (or a restricted dataset) associate each to the same parameter.
3. Calculate probability of finding the value associated to the moonlighting database by chance

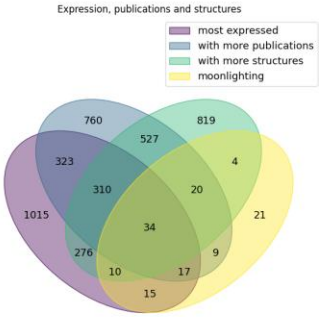


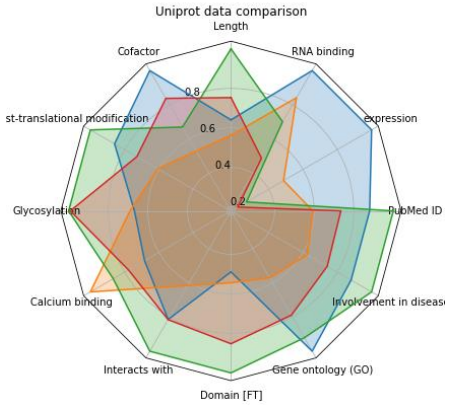
Bootstrap analysis of following datasets:
2000 most expressed proteins,
2000 proteins with more publications,
2000 proteins with more PDB structures.
Comparison with Human MLPs, red line

MLPs are longer than highly expressed proteins, but less than most non-MLPs; are enriched in cofactors; are not particularly glycosylated (intracellular at least for one function), often bind nucleotides; enriched in PTMs, subcellular localisations; they are often involved in disease and are well studied: average of 30 publications and 25 structure per MLP.

The choice of such dataset aims at reducing bias due to lack of annotations. Expression, number of PubMed IDs and PDB structures can be used as proxy of good characterisation.

Comparison between MLPs and different datasets of human proteins

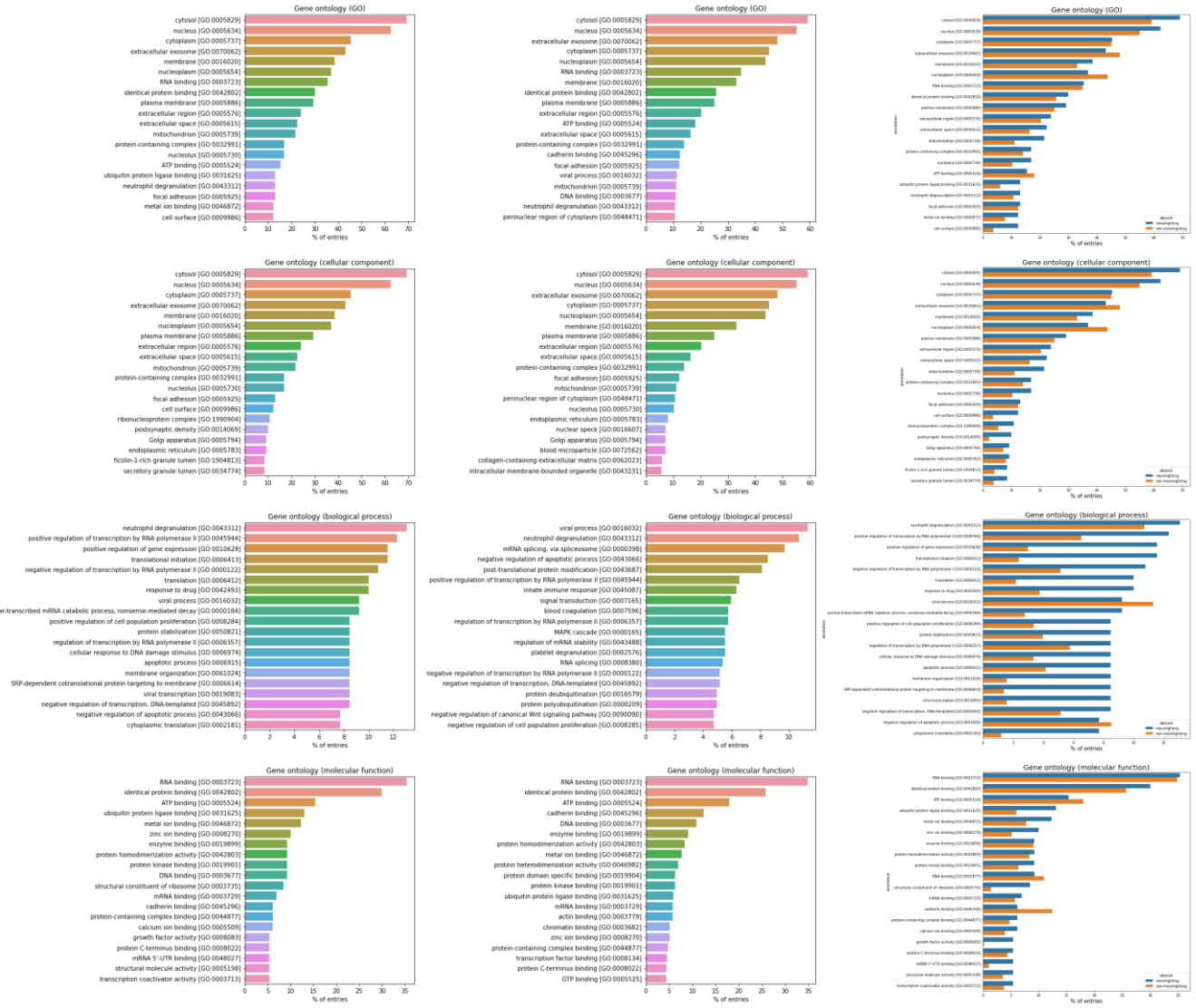




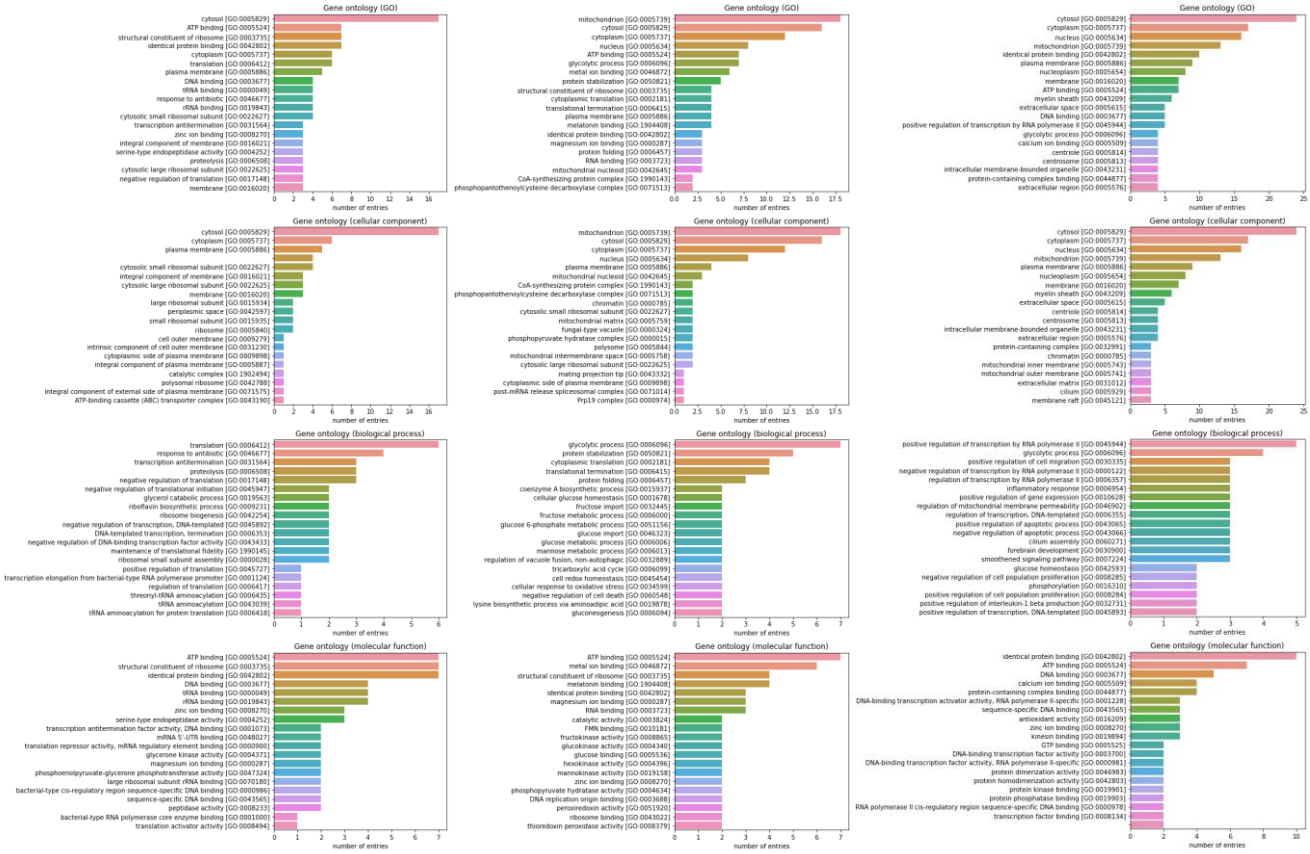
Right: Venn diagram of following datasets:
Human MLPs,
2000 most expressed proteins,
2000 proteins with more publications,
2000 proteins with more PDB structures.

Left: radarplot of Uniprot annotations
moonlighting (blue),
2000 most expressed proteins (orange),
2000 proteins with more publications (green),
2000 proteins with more PDB structures (red).

Gene ontology (GO) enrichment

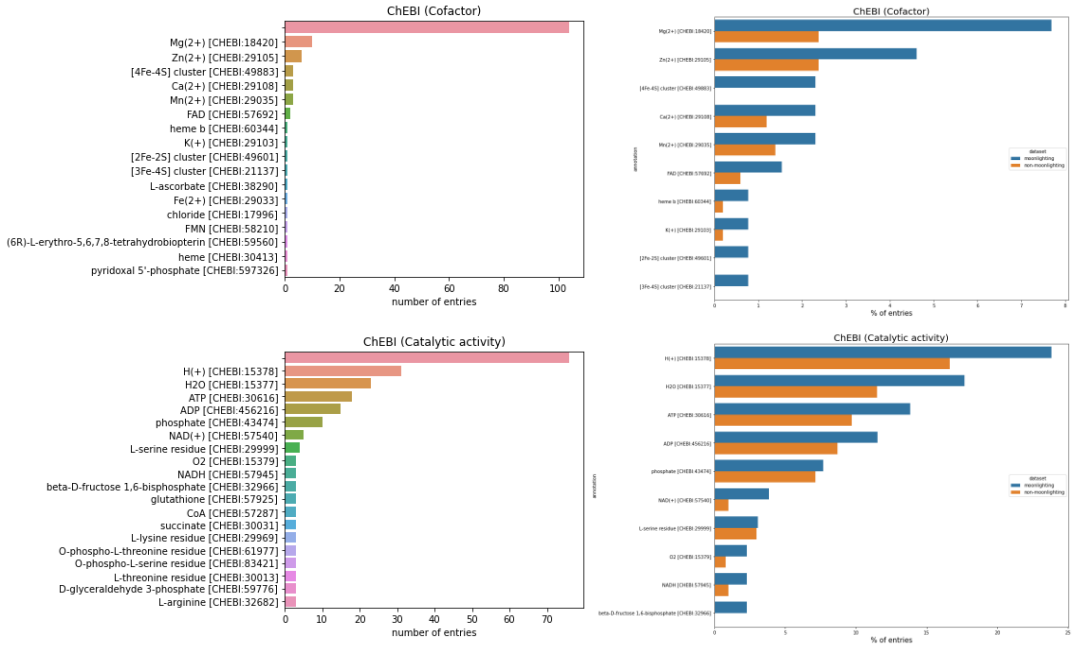


Right: human MLPs; middle: non-MLP; left: 20 most enriched annotations from MLPs compared with non-MLPs



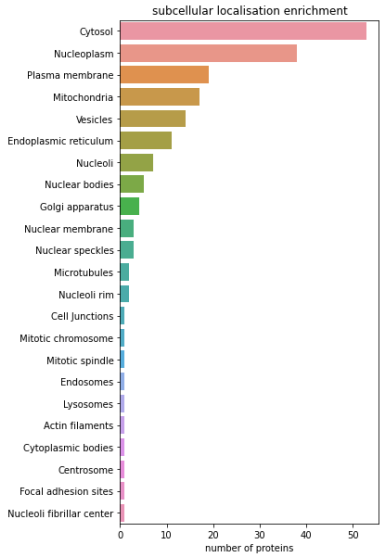
GO enrichment of MLPs in other organisms: right: E. coli; middle: yeast; left: mouse.

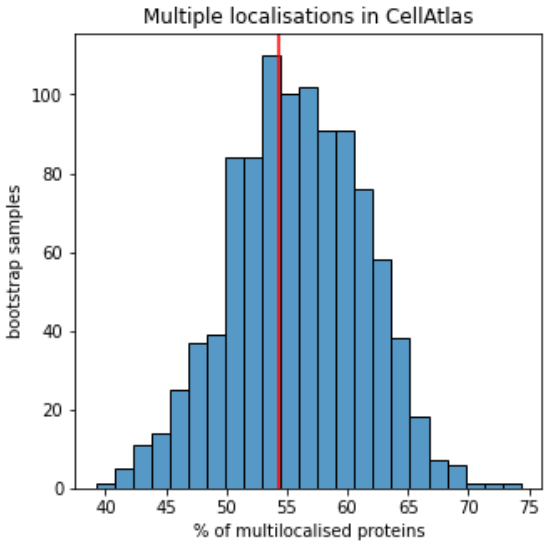
ChEBI enrichment



Right: enrichment in MLPs; enriched annotations in MLPs compared to non-MLPs

Subcellular Localisation in Cell Atlas



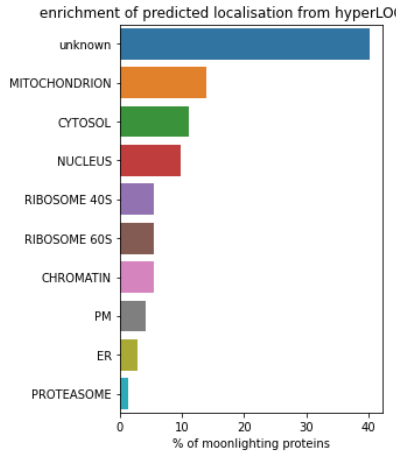


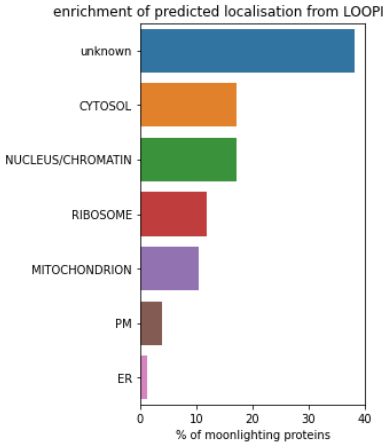
Right: enrichment of subcellular localisations from Cell Atlas.

Left: Bootstrapping analysis of multilocalised proteins in Cell Atlas.

MLPs do not have any unexpected localisations and are not particularly multilocalised. The red line represents the number of multilocalised MLPs.

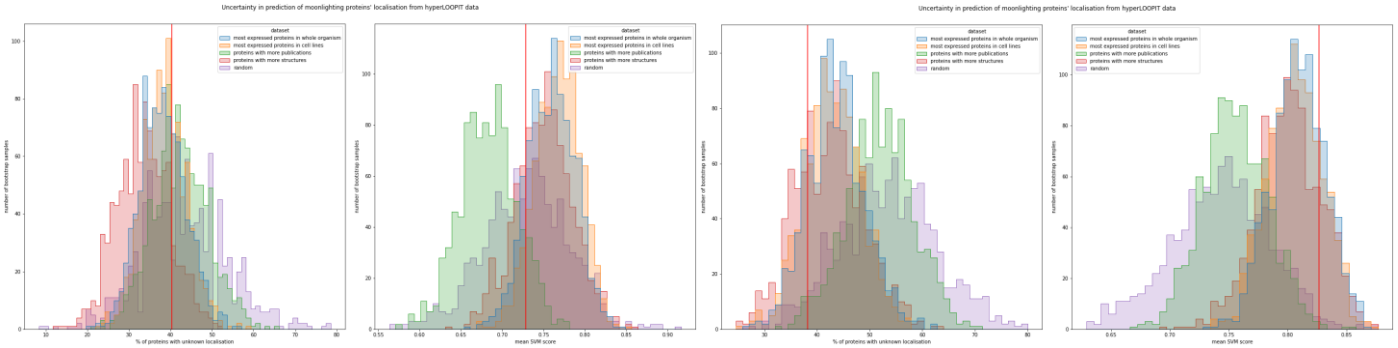
Subcellular Localisation in LOPIT databases





Right: enrichment of subcellular localisations SVM-predicted annotations in hyperLOPIT database.

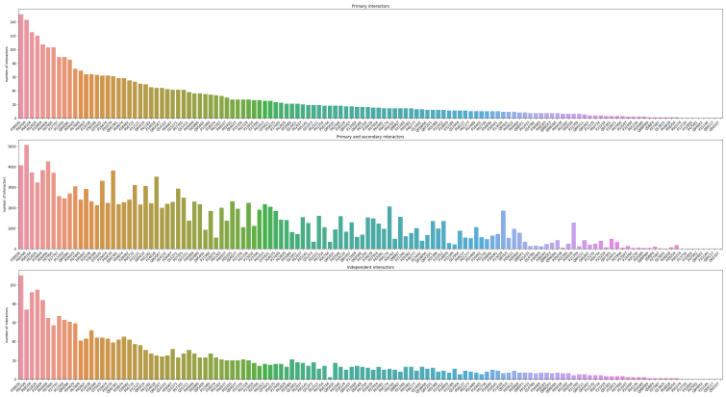
Left: enrichment of subcellular localisations SVM-predicted annotations in LOPIT-DC database.



Bootstrapping analysis for multilocalised proteins in LOPIT datasets: right: hyperLOPIT; left: LOPIT-DC. The red line represents the mean value for MLPs.

Multilocalised proteins should be annotated as “unknown” localisation and have low SVM prediction scores, but it is not the case for MLPs.

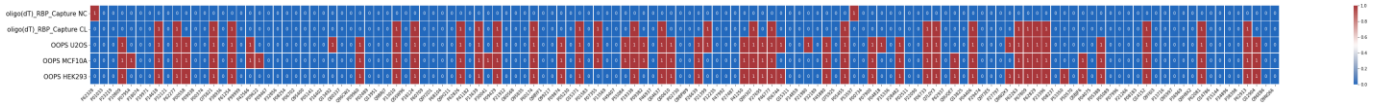
Protein-protein interactions: IntAct



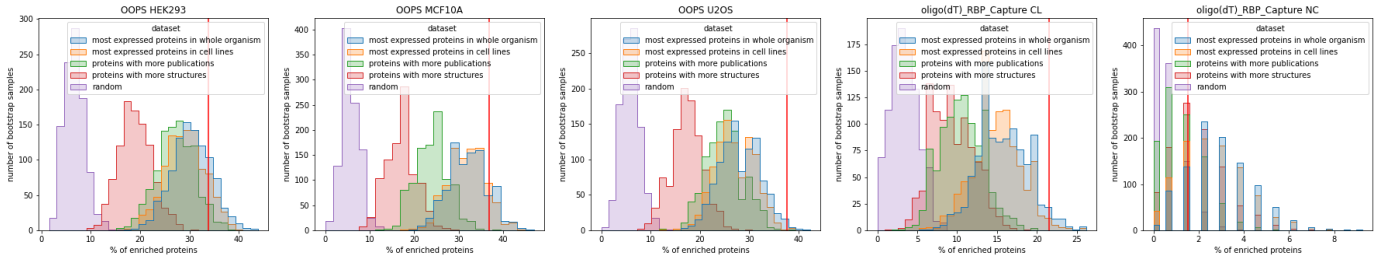
Number of interactors of MLPs:
Top: primary interactors,
Middle: primary and secondary interactors,
Bottom: independent interactor clusters.

RNA-binding in OOPS database

RNA-binding was discovered in glycolytic enzymes supporting their moonlighting activity. However, it has never been systematically assessed across other moonlighting proteins. Moonlighting proteins are significantly enriched in manually curated annotations for RNA-binding. Impressively this is confirmed in the unbiased database OOPS with more than 40% of moonlighting proteins found at the interface between RNA and protein enriched phases (indicative of an RNA-binding activity) in at least one of the 3 examined human cell lines. Bootstrapping analysis of different datasets confirmed the statistical relevance of this result.



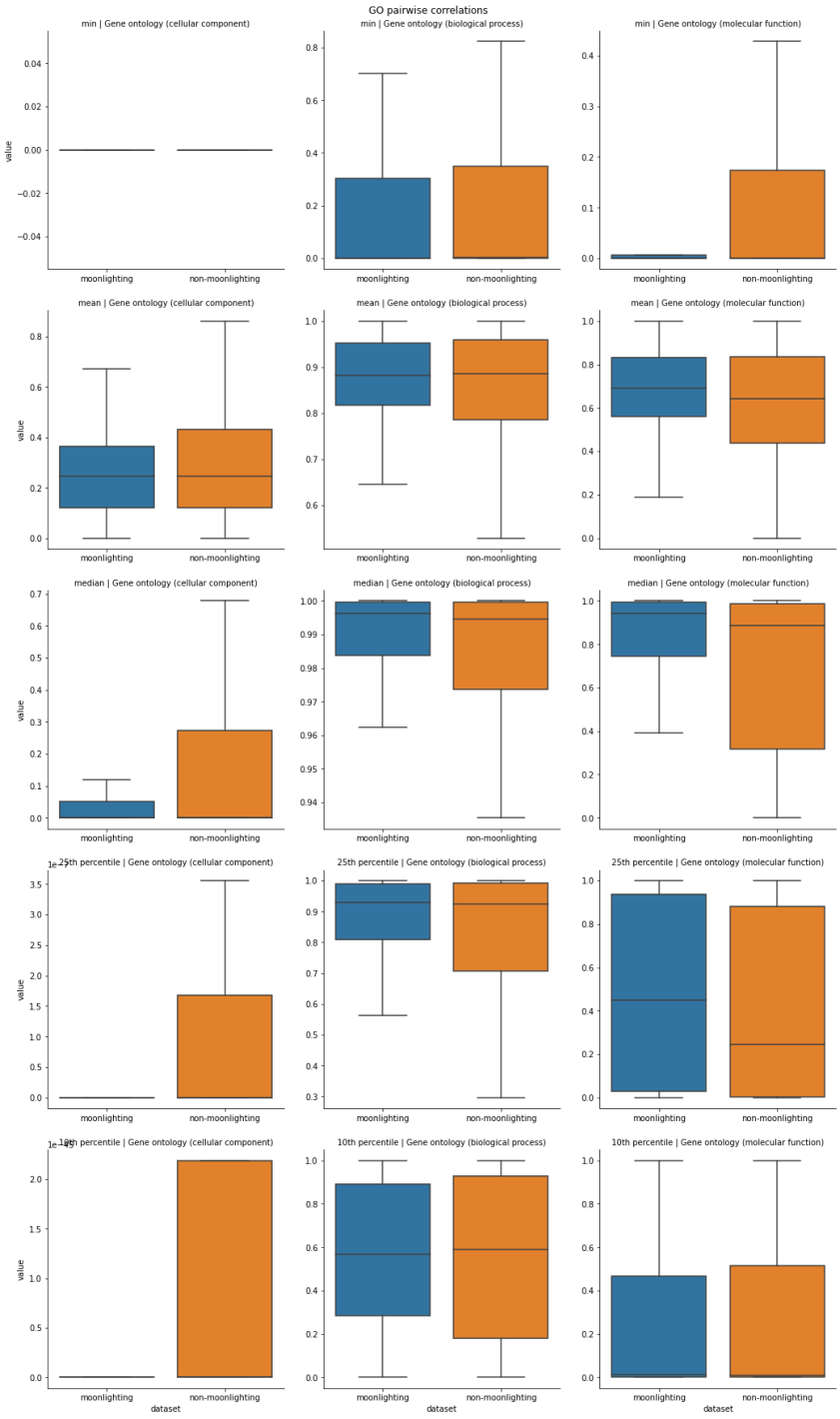
Heatmap of RNA-binding proteins in OOPS (U2OS, MCF10A and HEK293 human cell lines) and Oligo(dT) RBP capture datasets (with or without crosslinking). RNA-binding MLPs are usually enriched OOPS datasets for every cell-line and in CL Oligo(dT) RBP capture, but not in NC Oligo(dT) RBP capture (negative control), witnessing a real RNA-binding capacity.



Bootstrapping analysis for RBP in different OOPS and Oligo(dT) capture datasets. The red line represents the number of RNA-binding MLPs.

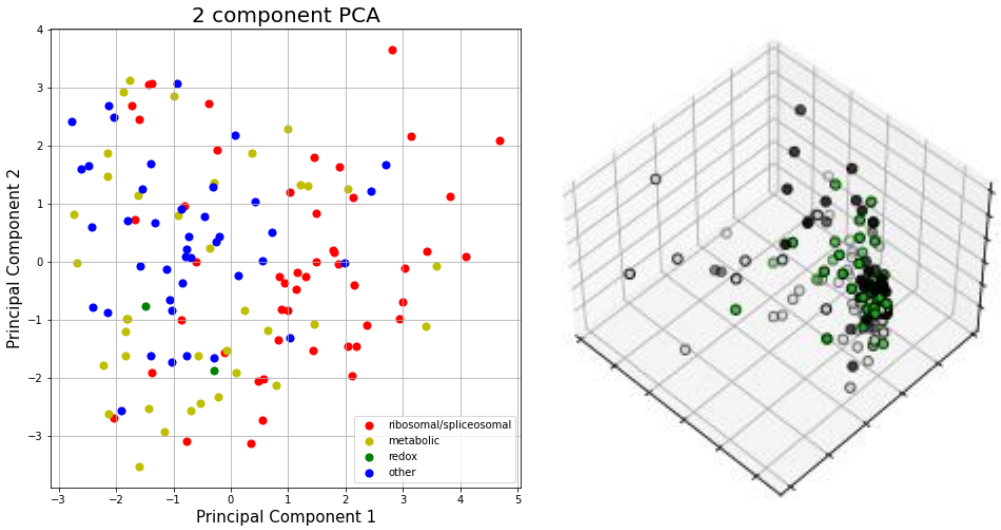
GO correlations

1. Quantify how often two GO annotations are found in the same protein or the same cluster of interacting proteins across the human genome.
2. Count how often each GO and each GO pair is found in the proteome.
3. Assume a hypergeometric distribution of independent GO annotations as in https://www.frontiersin.org/articles/10.3389/fgene.2015.00200/full
4. Make a correlation matrix
5. Check if moonlighting proteins (and their primary interactors) have GO annotations that are less correlated than a control database.



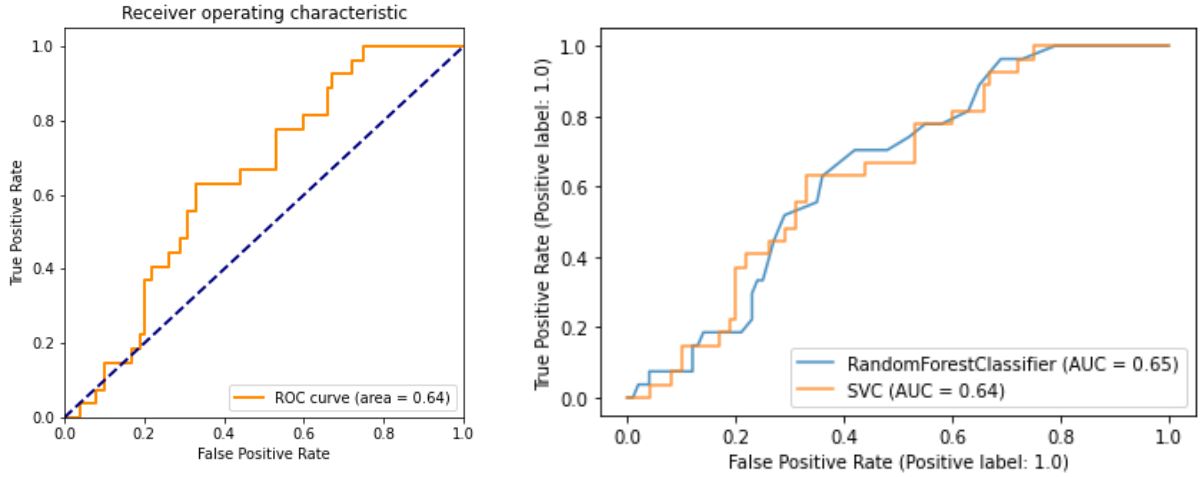
Note low minimum correlation in MLPs for all GO domains and very low mean, median and percentiles for cellular components.

Machine learning and MLPs



Failure of PCA to classify moonlighting proteins both in2d (right) and 3D (left)

MLPs are significantly enriched in specific classes, including ribosomal and spliceosomal proteins, glycolytic enzymes and other metabolic enzymes, peroxidases, thioredoxins and other oxidoreductases, and chaperones. However Principal component analysis (PCA) fails in clustering them even when these annotations are used as inputs, probably because of the low correlation between these recurrent canonical functions and very variable moonlighting function that are different in each single case.



Left: failure of SVM to predict moonlighting proteins, right: failure of SVM and Random Forest classifiers.

The features identified in my analysis are not sufficient to distinguish MLPs from non-MLPs.