# EXPLAINING PROTEIN-PROTEIN INTERACTION PREDICTIONS WITH GENETIC PROGRAMMING

evo 2022

Rita T. Sousa, Sara Silva, Catia Pesquita
LASIGE, Faculdade de Ciências da Universidade de Lisboa

U LISBOA  Ciências ULisboa  LASIGE  FCT Fundação para a Ciência e a Tecnologia

# WHY DO WE NEED
# EXPLAINABLE
# ARTIFICIAL INTELLIGENCE?

# WHY DO WE NEED EXPLAINABLE ARTIFICIAL INTELLIGENCE?

Users' trust

# WHY DO WE NEED EXPLAINABLE ARTIFICIAL INTELLIGENCE?

Users' trust

Ensure algorithmic fairness

# WHY DO WE NEED EXPLAINABLE ARTIFICIAL INTELLIGENCE?

Users' trust

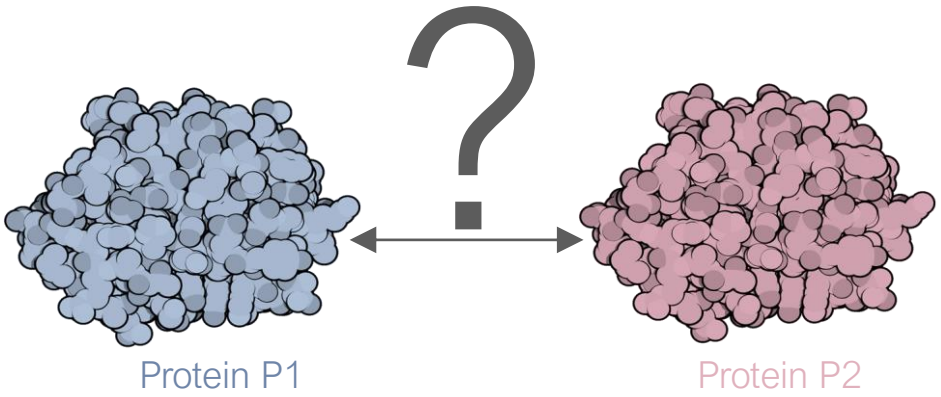
Ensure algorithmic fairness

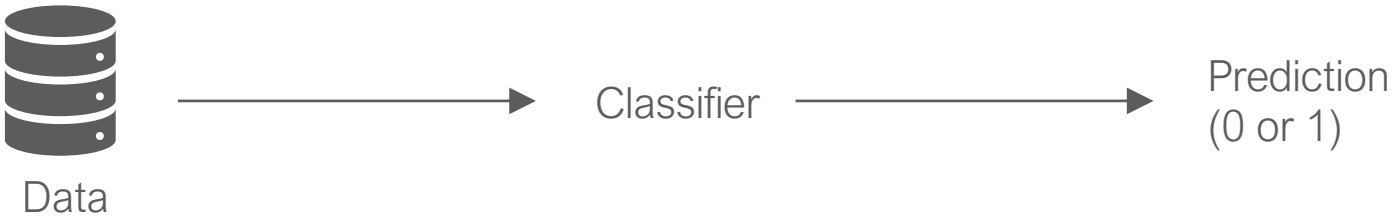
Identification of potential bias

# WHY DO WE NEED EXPLAINABLE ARTIFICIAL INTELLIGENCE?

- Users' trust
- Ensure algorithmic fairness
- Identification of potential bias
- Discovery of new knowledge

# PROTEIN-PROTEIN INTERACTION (PPI) PREDICTION

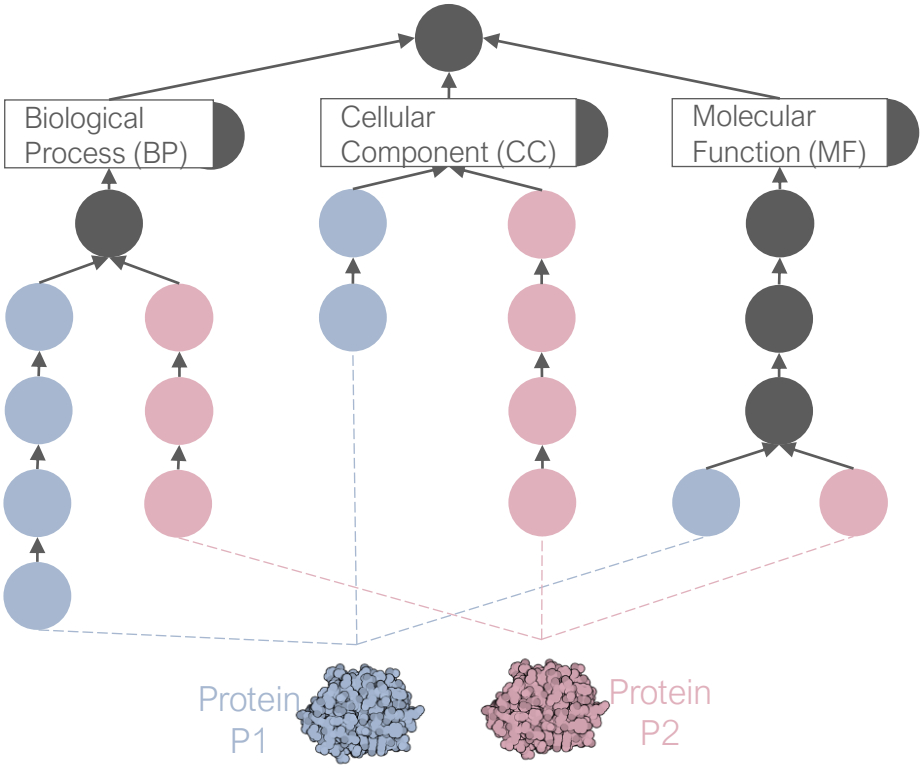



However, experts want to understand the biological mechanisms that underlie the natural phenomena they are predicting.

# ONTOLOGIES AND KNOWLEDGE GRAPHS ARE A UNIQUE OPPORTUNITY FOR EXPLAINABILITY

Ontologies and Knowledge Graphs (KGs) provide semantics (i.e., meaning) to the entities they represent through different semantic aspects.




GO KG = GO + GO annotations

# KNOWLEDGE GRAPH EMBEDDINGS ARE NOT EXPLAINABLE BY DEFAULT

An embedding is a vector representation that maps each node to a lower-dimensional space in which its graph position and the structure of its local graph neighborhood are preserved.

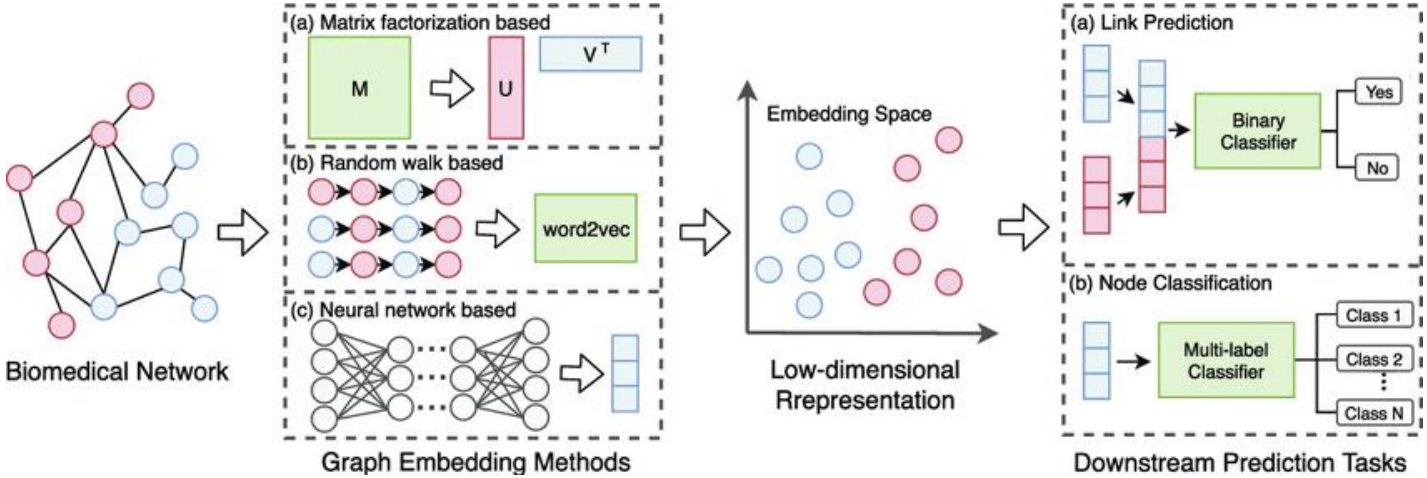
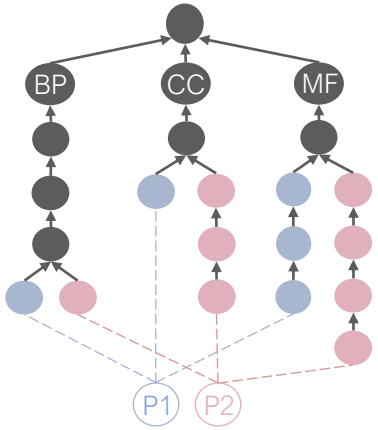



Image credit: Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., ... & Sun, H. (2020). Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics, 36(4), 1241-1251.

# USING SEMANTIC SIMILARITY AS AN ALTERNATIVE EXPLANATORY STRATEGY

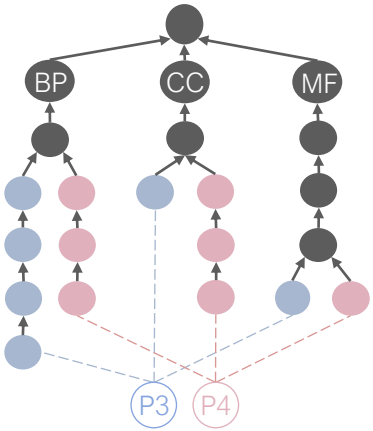
- KGs provide the scaffolding for comparing entities at a higher level of complexity by comparing the ontology classes with which they are annotated.
- Semantic similarity computed using different portions of the KG to reflect different semantic aspects (SA) can provide more granular explanations with higher information content.




P1 is vacuolar protein-sorting-associated protein 36.
P2 is vacuolar-sorting protein SNF8.

P1 and P2 interact.

$$SS_{BP} = 0.793$$
$$SS_{CC} = 0.326$$
$$SS_{MF} = 0.061$$



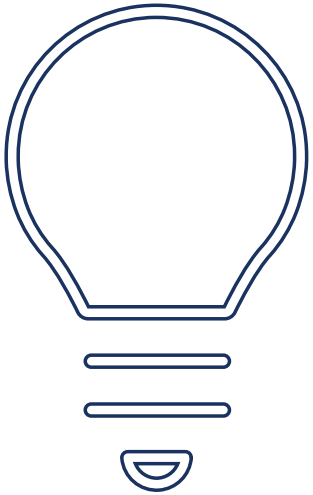

P3 is eukaryotic translation initiation factor 5B, isoform B.
P4 is Mig-2-like GTPase Mtl.

P3 and P4 do not interact.

$$SS_{BP} = 0.006$$
$$SS_{CC} = 0.202$$
$$SS_{MF} = 0.713$$

# HOW CAN WE GENERATE GLOBAL AND INTERPRETABLE EXPLANATION FOR PPI PREDICTION USING SEMANTIC SIMILARITY AS REPRESENTATION?

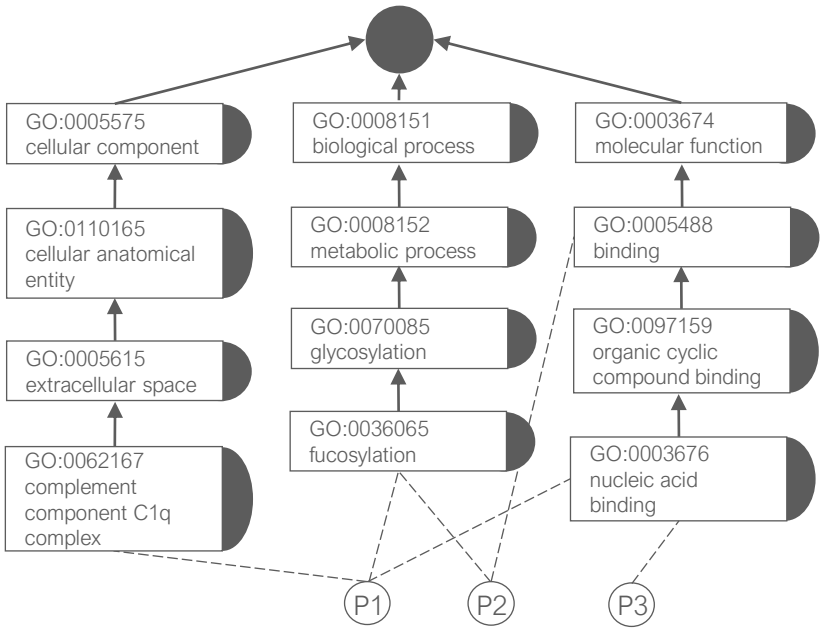
# HOW CAN WE GENERATE GLOBAL AND INTERPRETABLE EXPLANATION FOR PPI PREDICTION USING SEMANTIC SIMILARITY AS REPRESENTATION?



Using Genetic Programming (GP) over a set of semantic similarity values, each describing a semantic aspect represented in the knowledge graph.

# OUR APPROACH

Gene Ontology KG




Computing KG-based semantic similarity between entity pairs for each semantic aspect


PPI from STRING Database

P1 P2 1

P1 P3 1

+

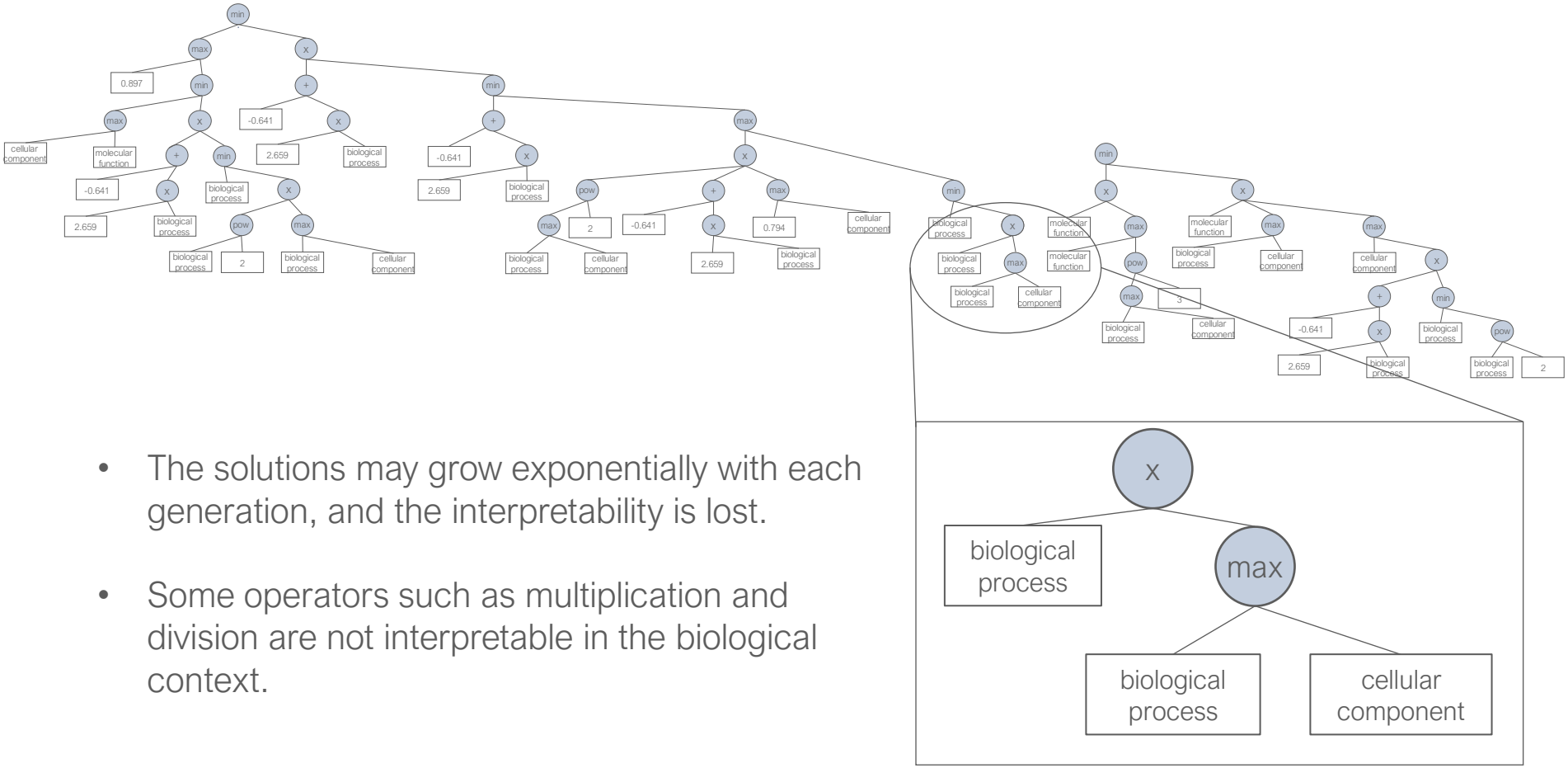
Random Negative Sampling

P2 P3 0

Evolving a GP model

Predicting on unseen data using the GP model

# GP IS NOT ALWAYS INTERPRETABLE




- The solutions may grow exponentially with each generation, and the interpretability is lost.
- Some operators such as multiplication and division are not interpretable in the biological context.

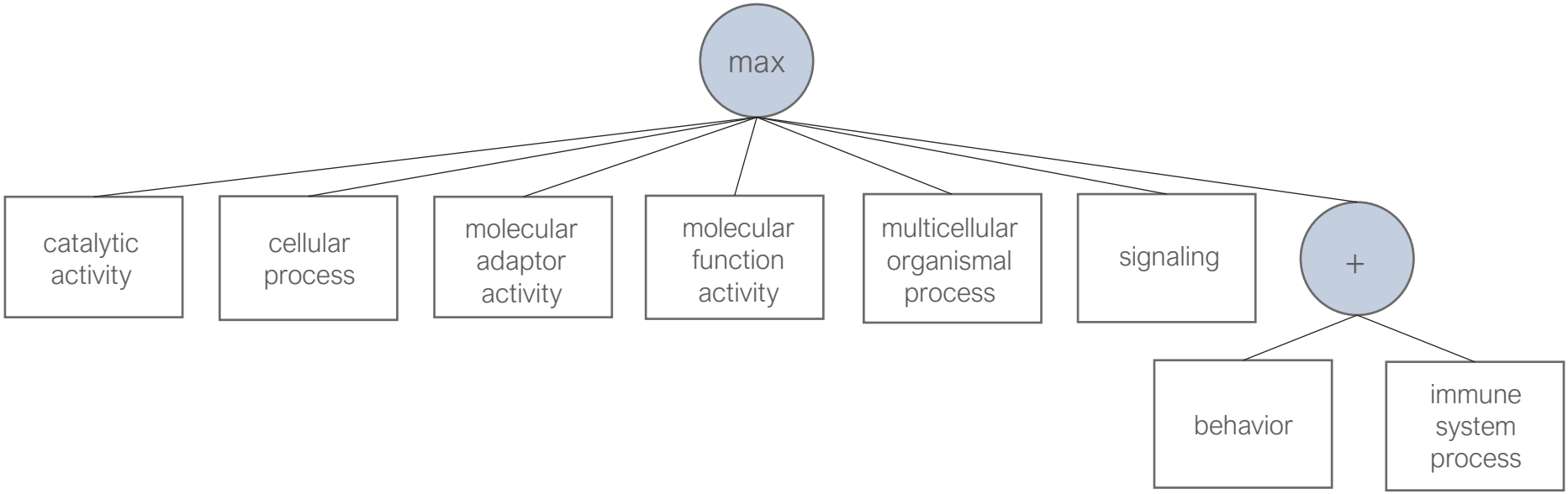
# GP VERSUS GP6X

| GP | GP6x |
|---|---|
| • No depth penalization<br>• 6 operators: multiplication, division, maximum, minimum, addition and subtraction | • Penalize solutions with a depth greater than 6<br>• Interpretable operators, namely maximum, minimum, addition and subtraction |

Results for 10-fold cross-validation :

| | Median weighted average of F-measures (WAF) | Median number of nodes |
|---|---|---|
| GP | 0.875 | 49 |
| GP6x | 0.866 | 17 |

# GP6X MODEL ANALYSIS




$$\max(SS_{catalytic\ activity}, SS_{cellular\ process}, SS_{molecular\ adaptor\ activity}, SS_{molecular\ function\ regulator}, SS_{multicellular\ organismal\ process}, SS_{signaling}, SS_{behavior} + SS_{immunne\ system\ process})$$

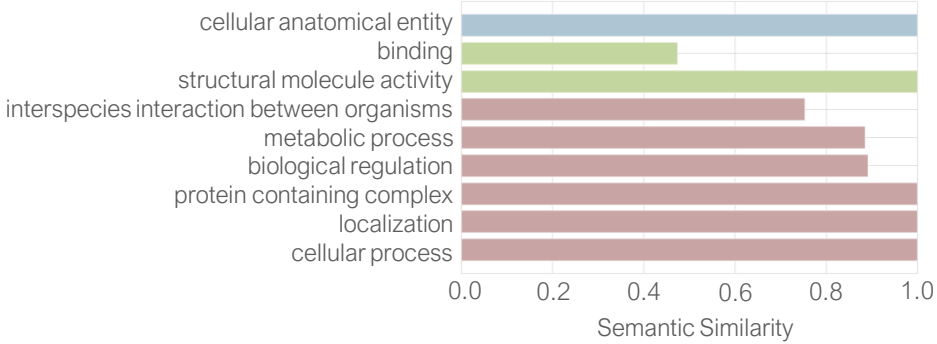
Two proteins that interact usually participate in the same biological processes.

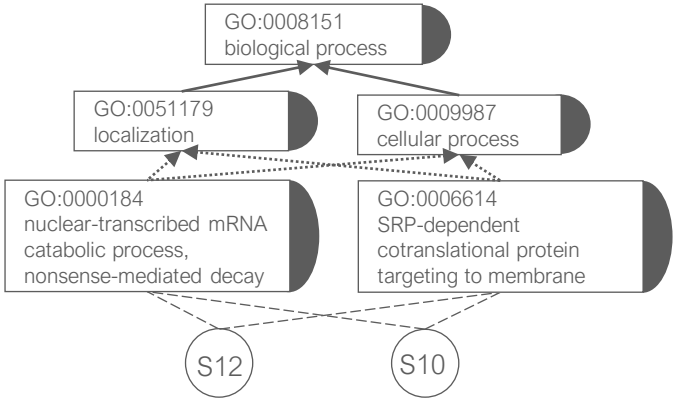
# WHEN GP CORRECTLY CLASSIFIES

$$\max(SS_{catalytic\ activity}, SS_{cellular\ process}, SS_{molecular\ adaptor\ activity}, SS_{molecular\ function\ regulator}, SS_{multicellular\ organismal\ process}, SS_{signaling}, SS_{behavior} + SS_{immunne\ system\ process})$$

## True Positive (+/+)

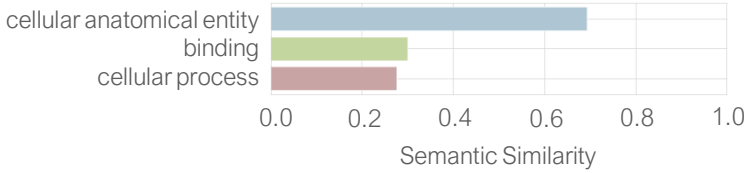
40S ribosomal protein S12 and
40S ribosomal protein S10

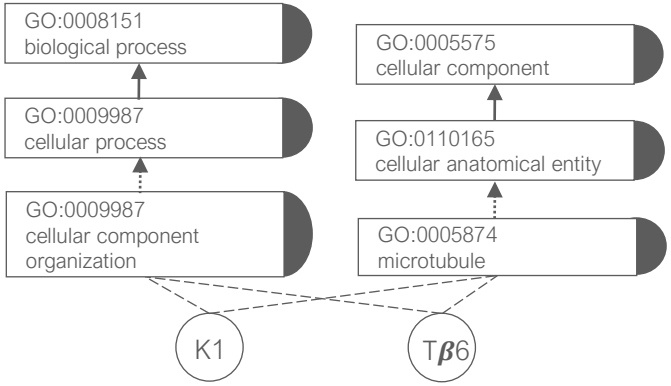



## True Negative (-/-)

Kinetochore-associated protein 1 and
Tubulin $\beta$-6 chain

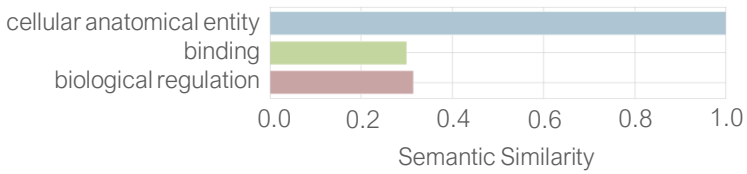


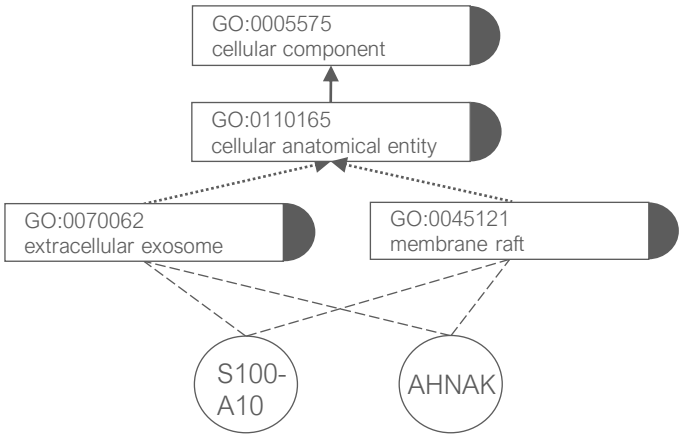



# WHEN GP FAILS

False Negative (+/-)

S100-A10 protein and neuroblast differentiation-associated protein AHNAK




Protein S100-A10 works together with neuroblast differentiation-associated protein AHNAK in the development of the intracellular membrane.

JCB Journal of Cell Biology

This article at JCB.org | Editors | Contact | Instructions for Authors

J Cell Biol. 2004 Jan 5; 164(1): 133–144. PMCID: PMC2171952

doi: 10.1083/jcb.200307098 PMID: 14699089

AHNAK interaction with the annexin 2/S100A10 complex regulates cell membrane cytoarchitecture

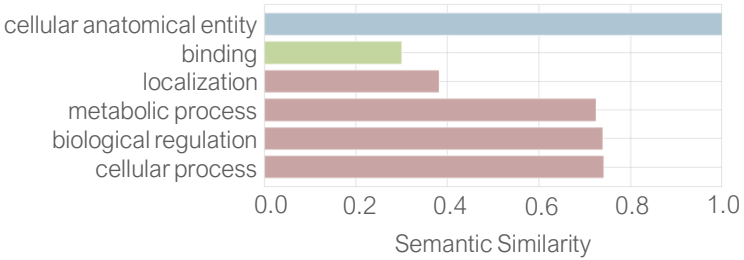
Christelle Benaud,[1] Benoît J. Gentil,[1] Nicole Assard,[1] Magalie Court,[2] Jerome Garin,[2] Christian Delphin,[1] and Jacques Baudier[1]

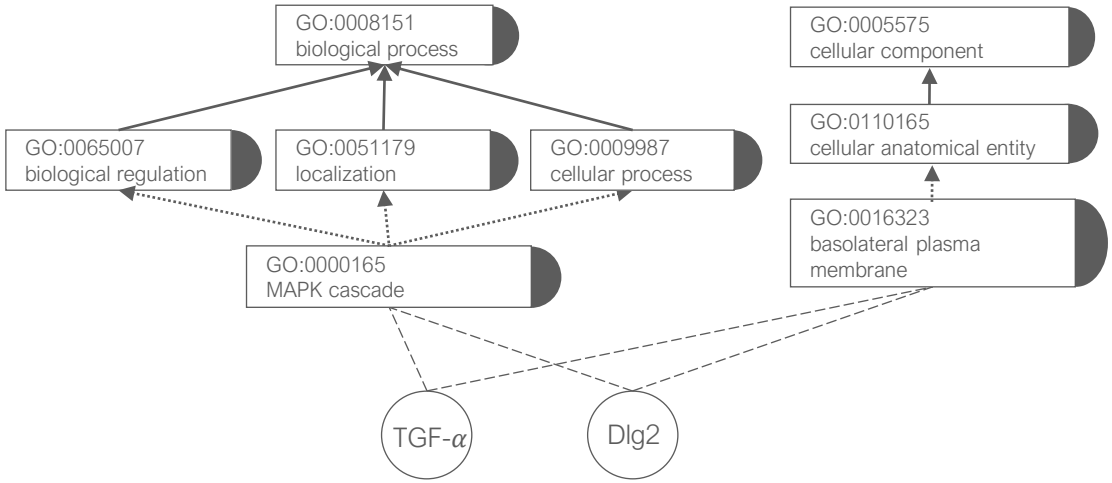
J Cell Biol

# WHEN MISCLASSIFICATIONS ARE NOT MISTAKES

False Positive (-/+)

Protransforming growth factor $\alpha$ and Disks large homolog 2




The literature describes interactions between proteins of the same family of the pair, indicating that this is likely a true but still unknown interaction.

# CLOSING REMARKS

- Explainability can be key to **uncover issues with the underlying data** and even **pose new hypothesis**.
- When we use GP, the explanation is the model itself, avoiding the need for local explanations or post-hoc techniques. Moreover, operators can be tailored to the domain.
- The performance of the more interpretable GP methods is not substantially lower, but what little they sacrifice in performance is more than gained in explainability.

# ACKNOWLEDGEMENTS

Catia Pesquita

Sara Silva

risousa@ciencias.ulisboa.pt

@RitaTorresSousa

Read the short paper: https://github.com/liseda-lab/ExplainingPPIpredictions/blob/main/Explaining_PPI_predictions_with_GP.pdf

This work was funded by FCT through LASIGE Research Unit (UIDB/00408/2020, UIDP/00408/2020); projects GADgET (DSAIPA/DS/0022/2018) and BINDER (PTDC/CCI-INF/29168/2017); PhD grant SFRH/BD/145377/2019.