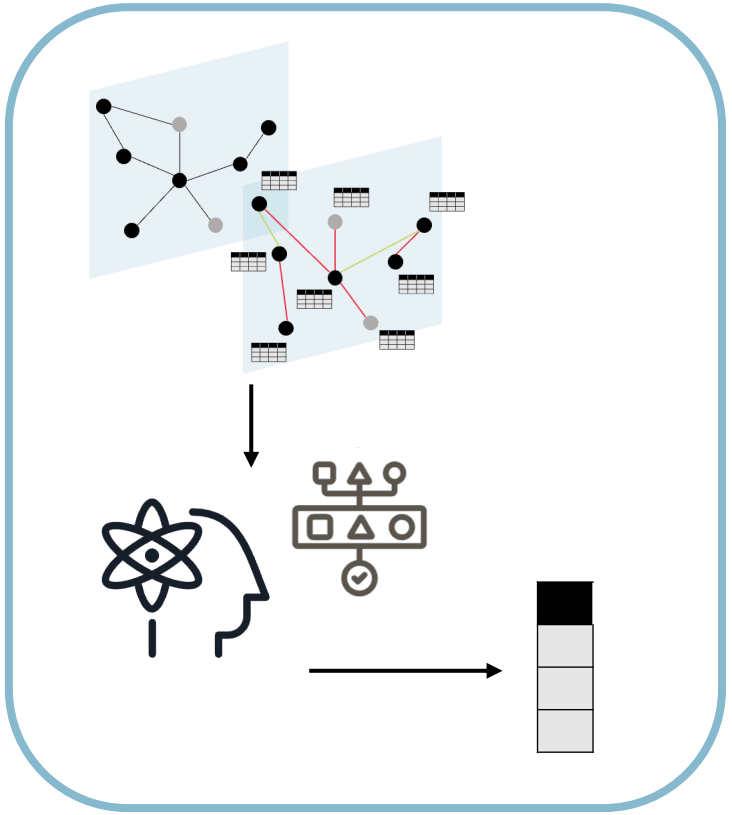
Machine learning-based prediction of functionally similar protein pairs

Liv Toft



Supervisor: Prof. Yu (Brandon) Xia

April 11, 2024

Outline

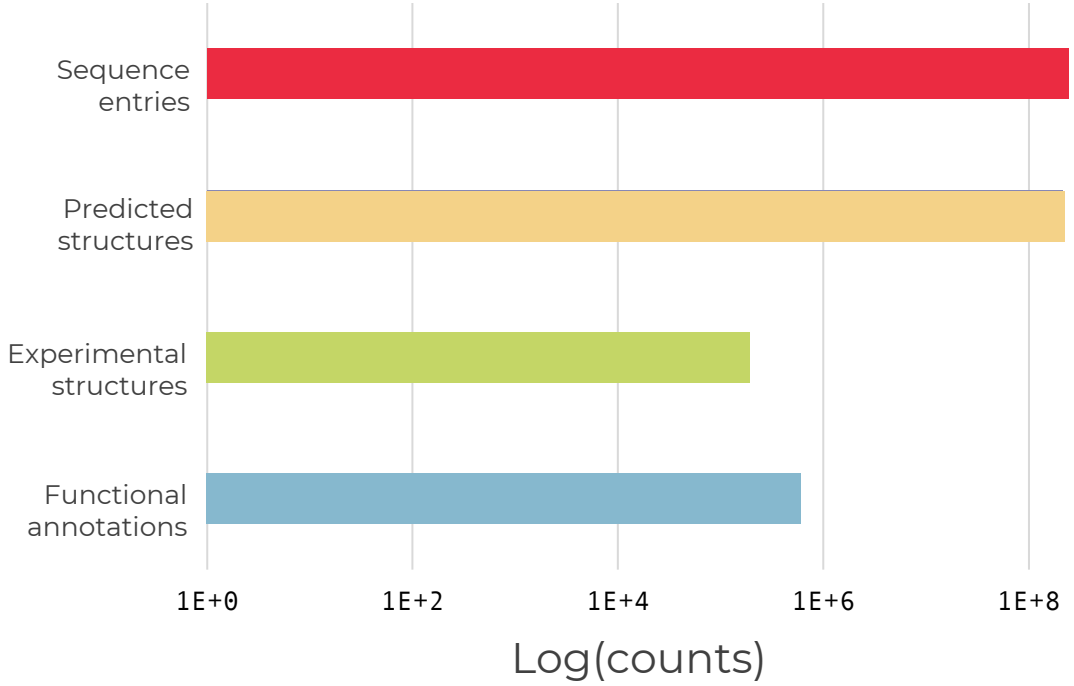
1. Background & Motivation
2. Data Integration
3. Machine Learning Models
4. Results
5. Future Work

1

Background & Motivation

The Protein Function Annotation Problem

- 230 million protein sequences exist
- Only 0.25% have functional annotations
- Experimental functional annotation is slow and costly



(Bugnon *et al., Patterns*, 2023;
Varadi *et al., GigaScience*, 2022;
Varadi *et al.; Nucleic Acids Res.*, 2023)

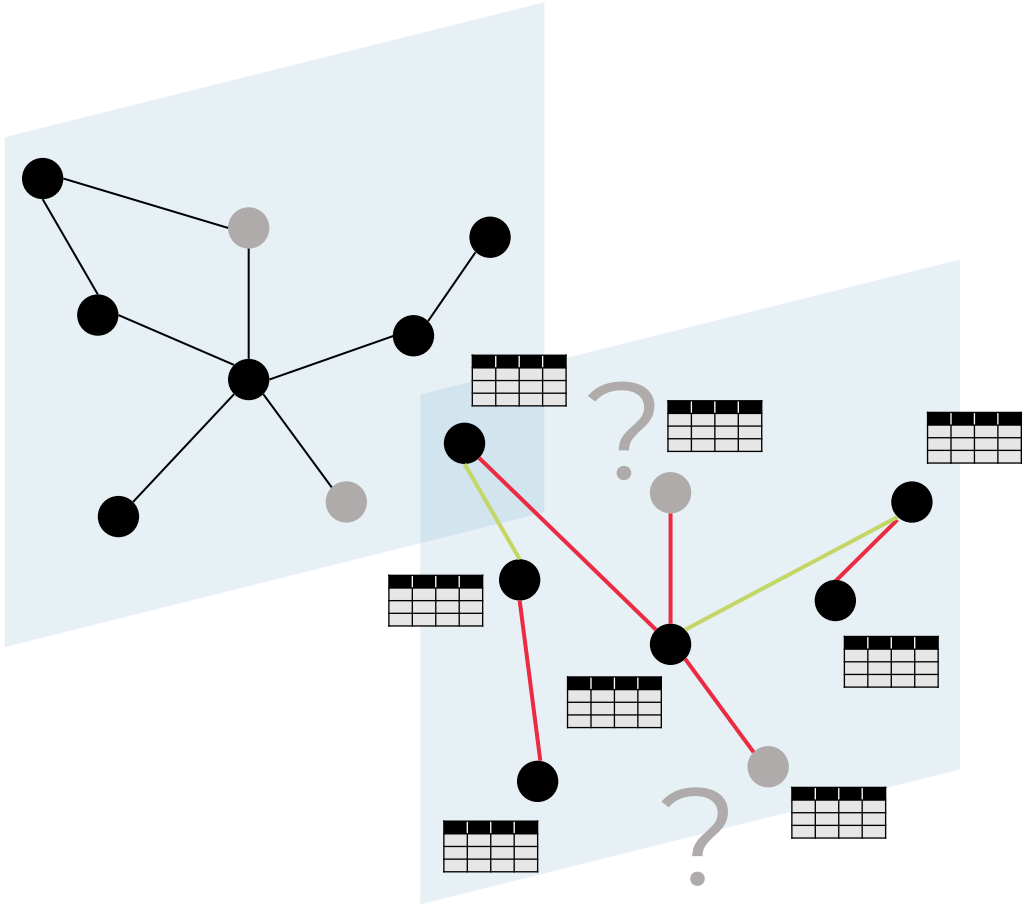
Motivations

1. Advance our knowledge of biological systems and processes
2. Discover novel disease genes
3. Identify new drug targets

(Xia *et al.*, *Data Mining for Systems Bio*logy, 2013; Derry and Altman, *bioRxiv*, 2023)

Filling In The Gaps With Machine Learning

- A lot of data already exists → use different genomic data sources to infer protein function from related proteins

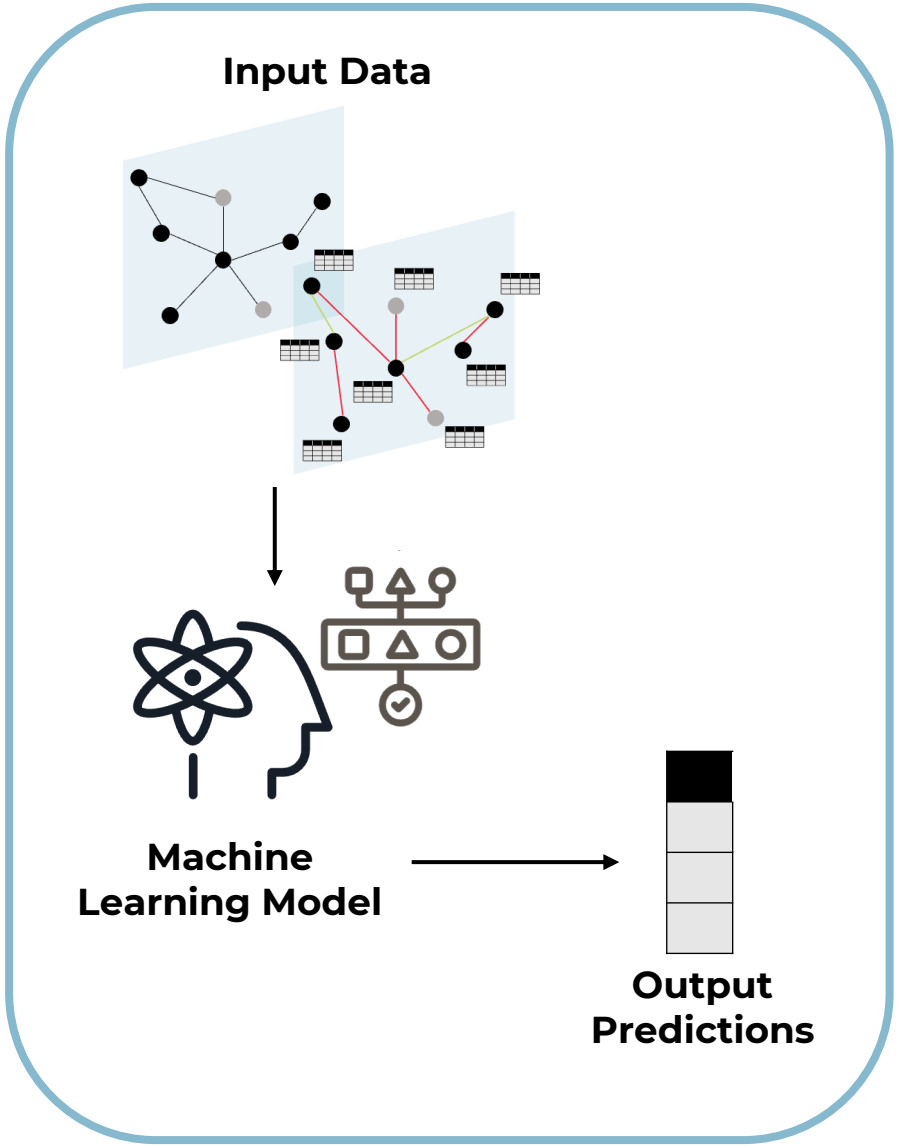


- Rigorous solution to heterogenous data integration
- Fast
- Inexpensive
- Accurate

(Xia *et al.*, *Data Mining for Systems Biology*, 2013)

Project Objective

Create a pipeline for combining heterogeneous genomic data sources for yeast and use it to train a machine learning model that can predict functionally similar protein pairs

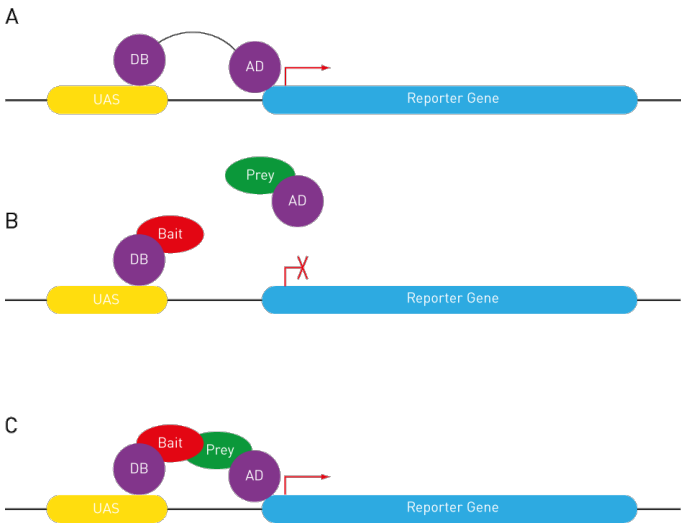


2

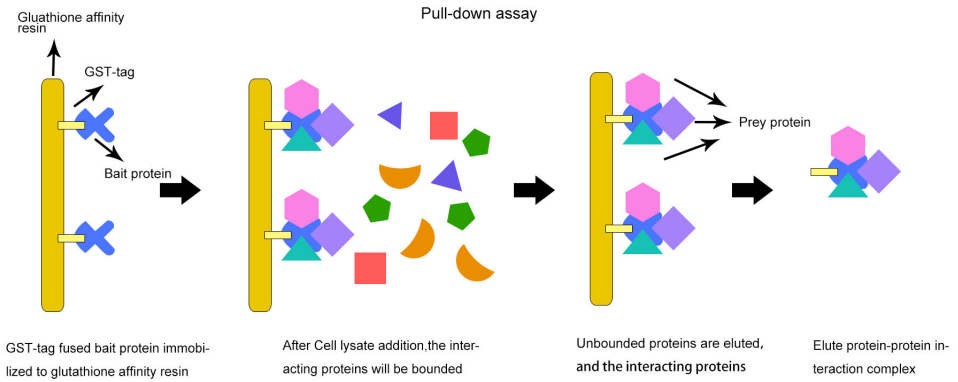
Data Integration

Experimental Protein-Protein Interactions (PPI)

Experimental Method(s)	Description	Rationale
Yeast-two hybrid, affinity capture	Physical protein-protein interactions experimentally validated by two or more independent sources	Proteins that physically interact have been found to be more likely to share a similar function



Yeast-two hybrid

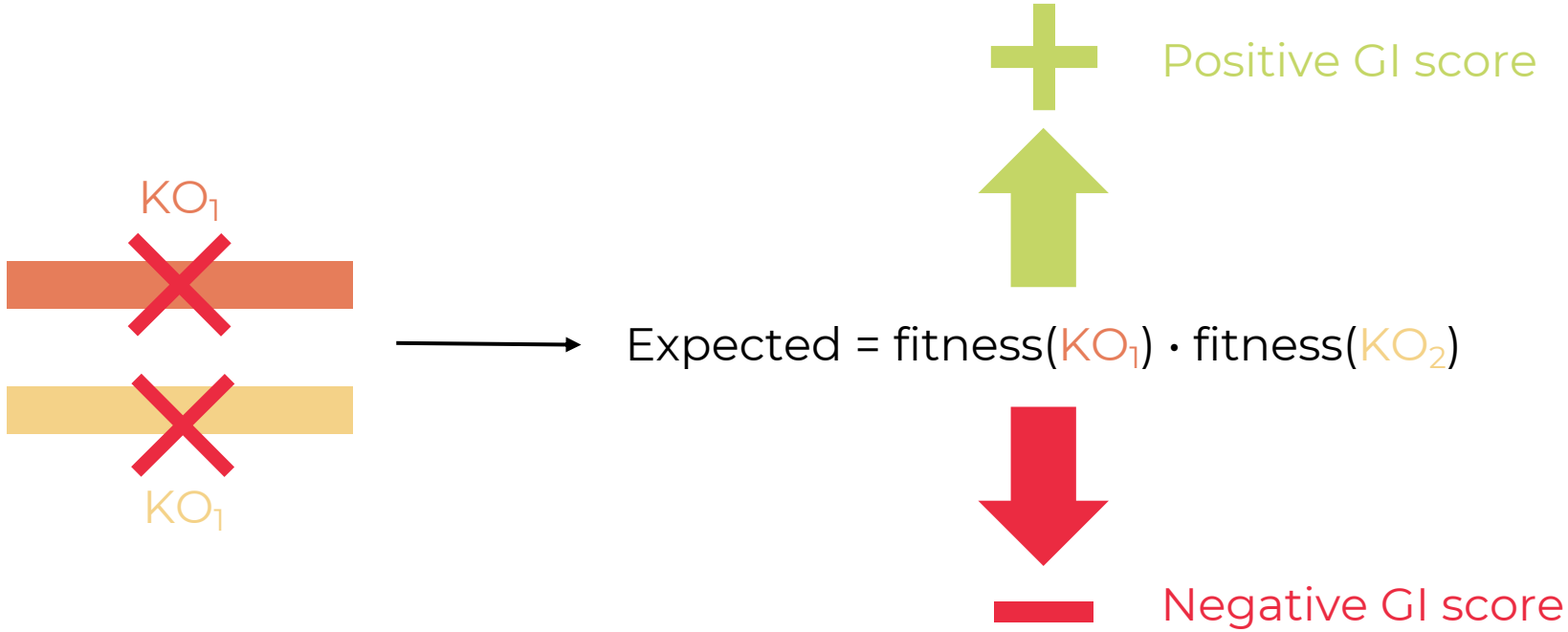


Affinity capture

(Stark *et al.*, *Nucleic Acids Research*, 2006)

Genetic Interaction (GI) Score

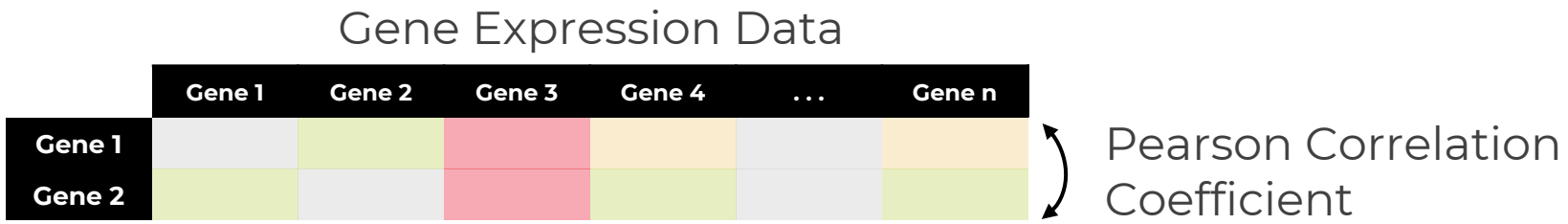
Experimental Method(s)	Description	Rationale
Synthetic genetic array	GI score for a pair of genes found by comparing the expected versus true fitness of double-mutant knockouts	Protein pairs that have a greater genetic interaction are shown to have similar functions



(Usai *et al.*, *G3*, 2017)

GI Profile Similarity

Experimental Method(s)	Description	Rationale
Synthetic genetic array	Pearson correlation coefficient for the GI profiles (GI scores for all proteins) of a pair of proteins	Proteins with similar GI profiles are more likely to be involved in the same biological processes



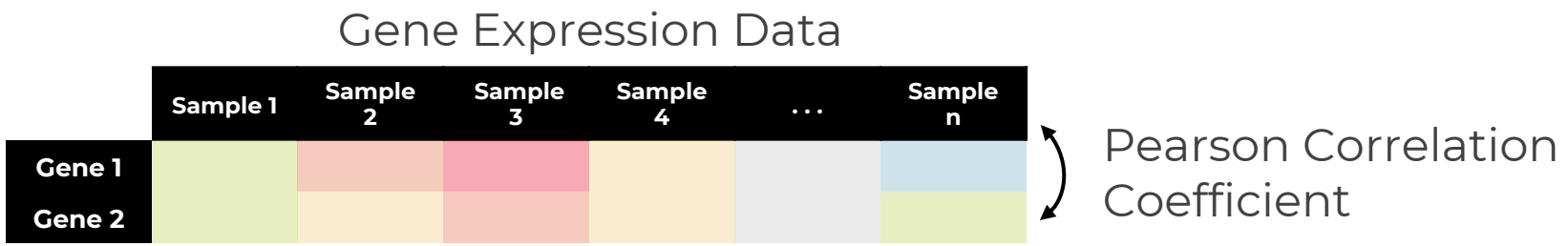
(Usai *et al.*, *G3*, 2017)

PPI Structural Information

Experimental Method(s)	Description	Reasoning
Experimental co-crystal PPI structures and PPI homology models	PPI structural information derived from three-dimensional structures from the Protein Data Bank (PDB) and PPI homology models	Proteins that physically interact are more likely to share a similar function

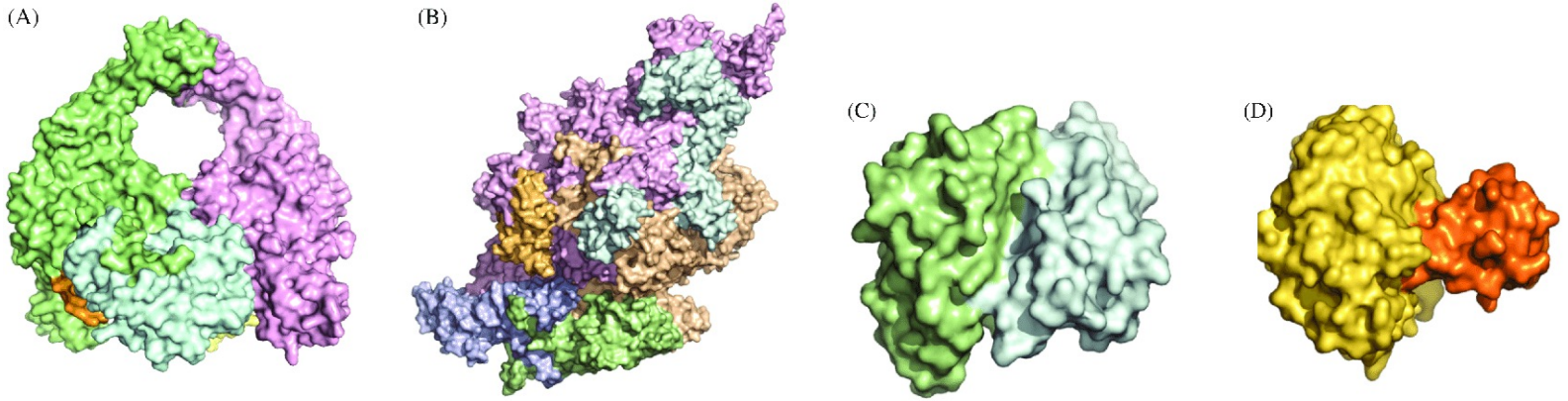
Co-Expression

Experimental Method(s)	Description	Reasoning
cDNA microarray hybridization assay	Pearson correlation coefficient for the pair of gene expression profiles corresponding to expression under 300 different mutations and chemical treatments	Pairs of proteins with similar functions have similar expression profiles



Co-Complex

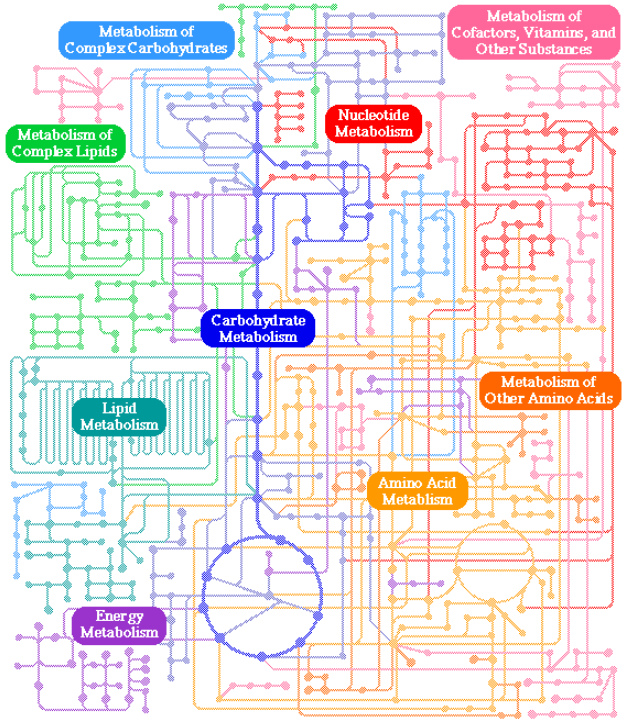
Experimental Method(s)	Description	Reasoning
Various (E.g. affinity capture + mass spectrometry)	Whether a pair of proteins are found in the same protein complex and are demonstrated to have the same molecular function	Proteins in the same complex share similar functions. This may capture more interacting proteins than PPI data



(Meldal *et al.*, *Nucleic Acids Res.*, 2018)

Co-Pathway

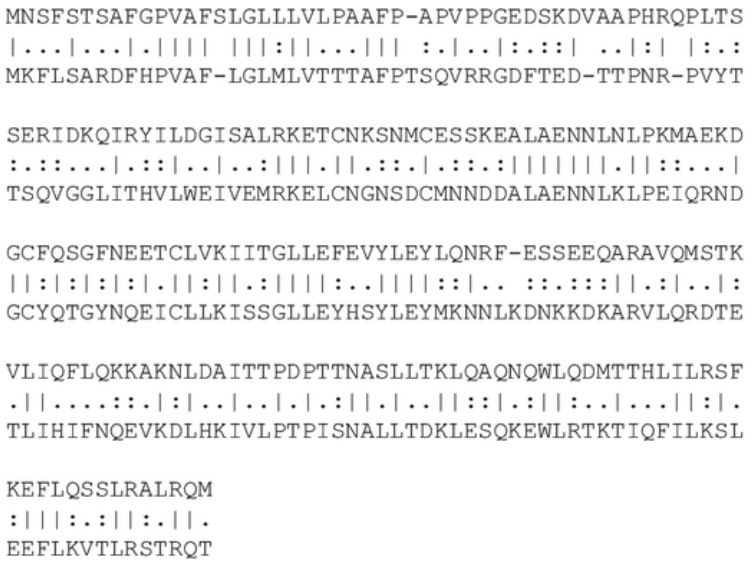
Experimental Method(s)	Description	Reasoning
Various (E.g. mass spectrometry + omics data analysis)	Whether a pair of proteins belong to the same pathway	Proteins in the same pathway share similar functions



(Wishart *et al., Nucleic Acids Res.*, 2020)

Sequence Similarity

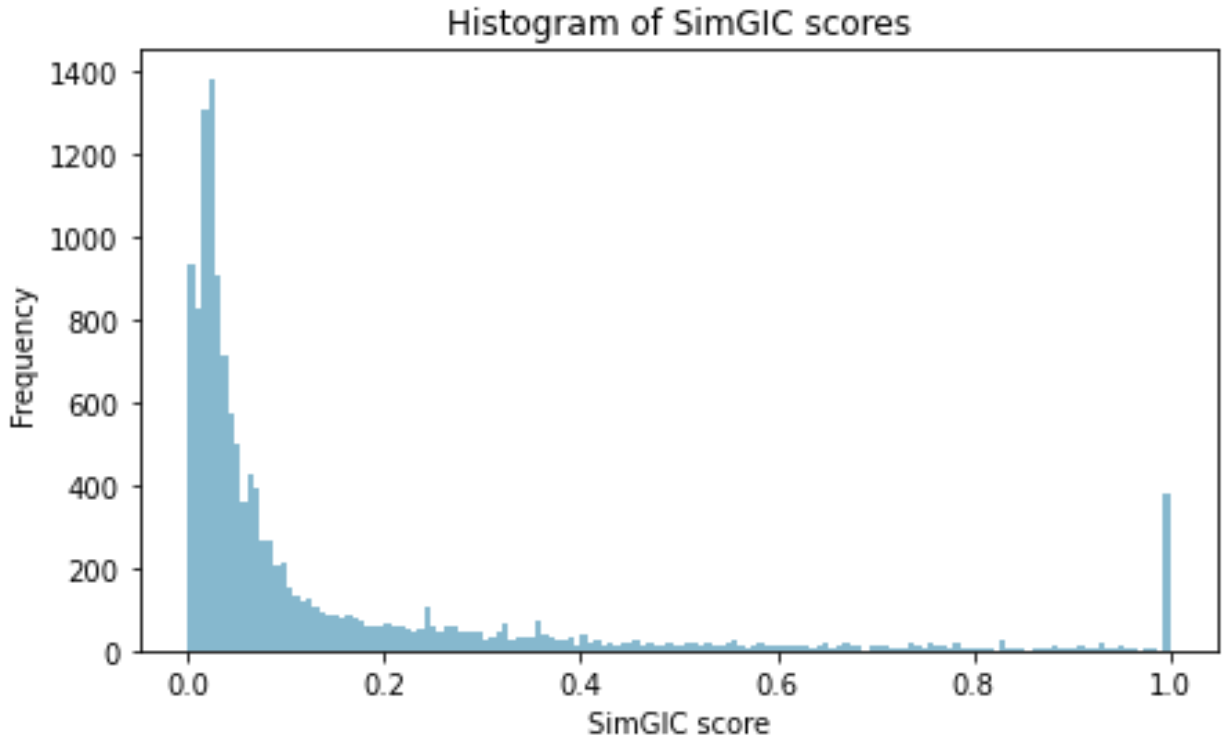
Experimental Method(s)	Description	Reasoning
BLASTp pairwise protein sequence alignment	Log(E-value) for the BLAST alignment of each protein pairs' sequences	Sequence similarity may be an indicator of structural-functional similarity



(Altschul *et al.*, *J. Mol. Bio.*, 1990)

Target: Protein Similarity

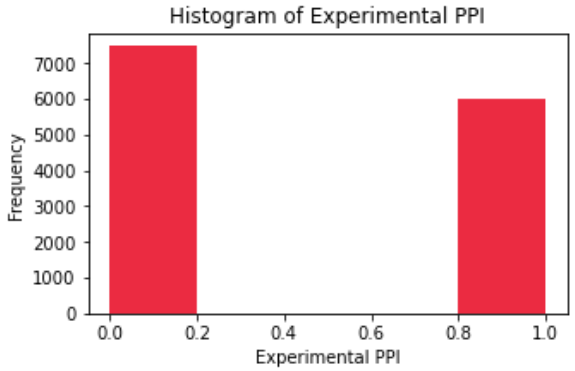
- SimGIC is a quantitative measurement of protein pair functional similarity
  - Based on Gene Ontology (GO) annotation and information entropy
- Set a threshold to **classify** similar and dissimilar protein pairs
  - More on this later

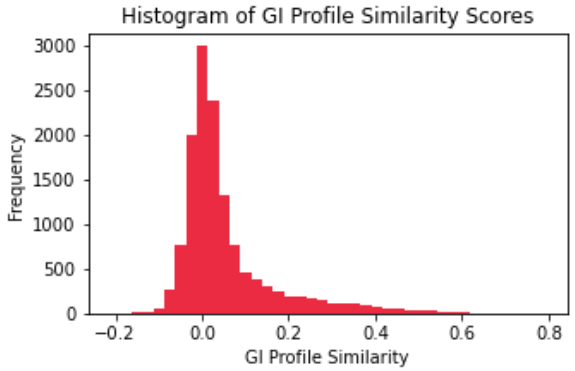


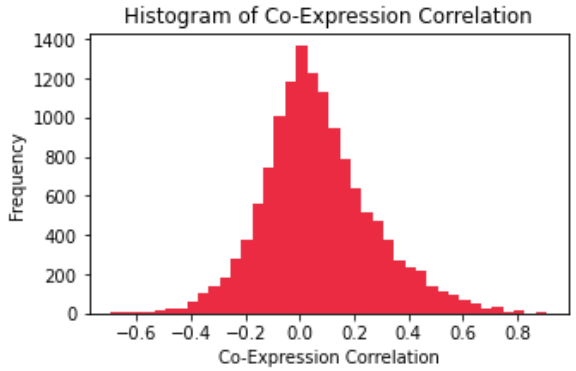
(Pesquita *et al.*, *BMC Bioinformatics*, 2008)

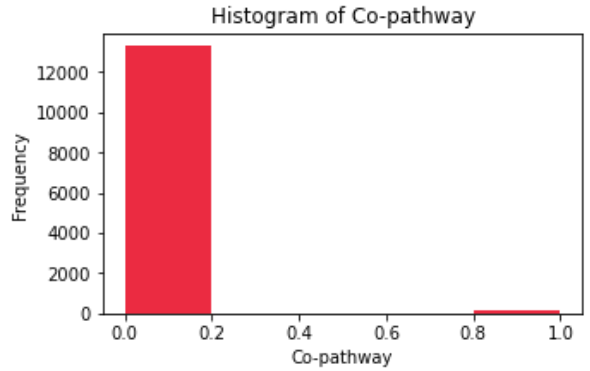
Integrated Dataset

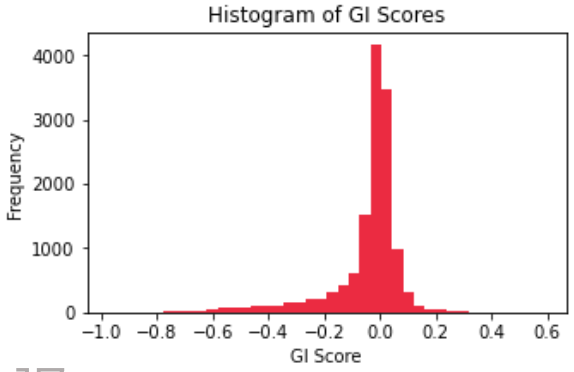
Protein pair	Experimental PPI	GI score	GI profile similarity	Structural information	Co-expression correlation	Co-complex	Co-pathway	Log(E-value)	Similar
P1_ID, P2_ID	[0, 1]	(-1, 1)	(-1, 1)	[0, 1]	(-1, 1)	[0, 1]	[0, 1]	[-460, 4.6]	[0, 1]

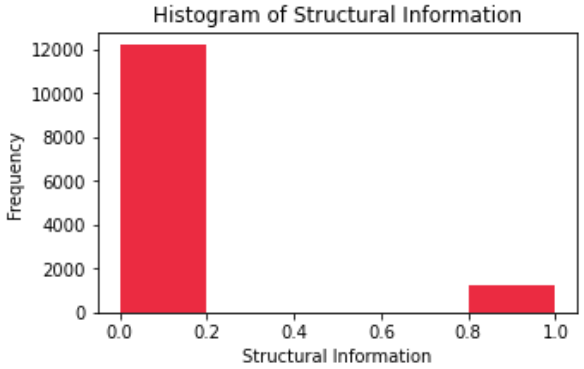


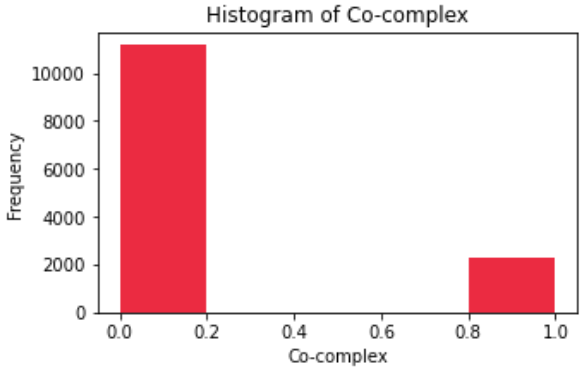


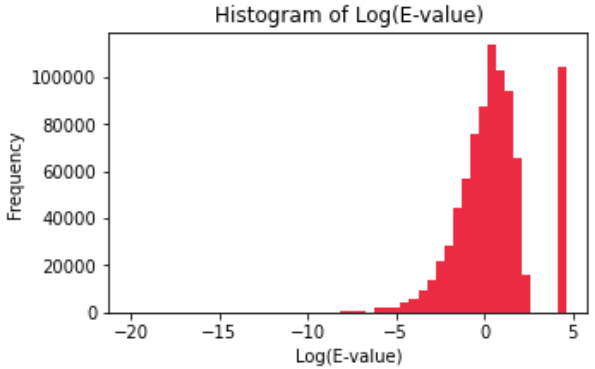






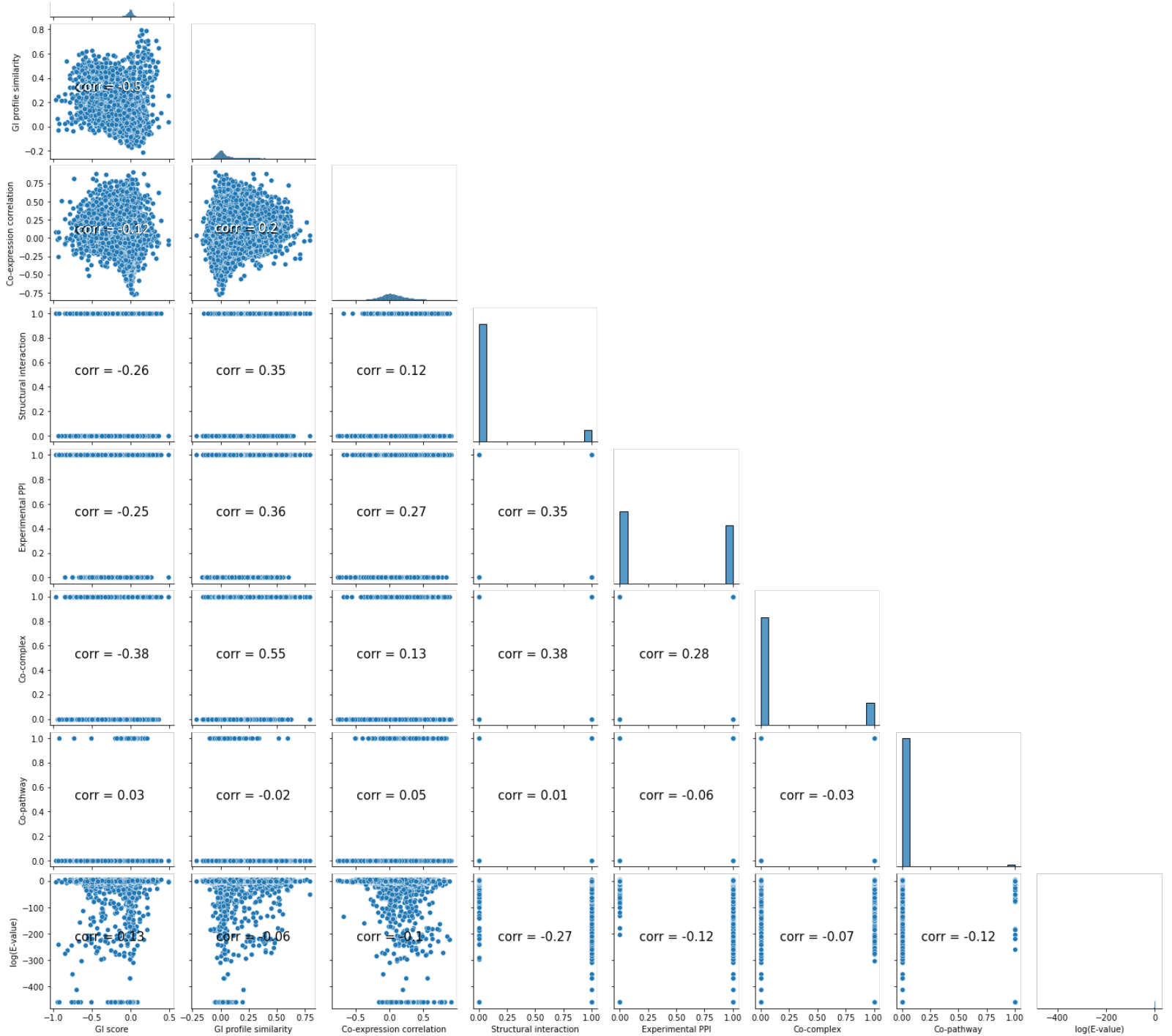






Integrated Dataset

- 13,458 entries
- Pairwise correlation plots show no feature pairs have a correlation > 0.55



3

Machine Learning Models

Naïve Bayes

- For benchmarking
- Utilizes Bayes' Theorem
- Simple, fast running times
- Conditional independence assumption can lead to low performance

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}}\left[\underbrace{\log\left(p(y=c)\right)}_{\text{Log prior}} + \underbrace{\sum_{i=0}^{k} \log\left(p(x_i|y=c)\right)}_{\text{Log likelihood}}\right]$$

C = [similar, not similar]

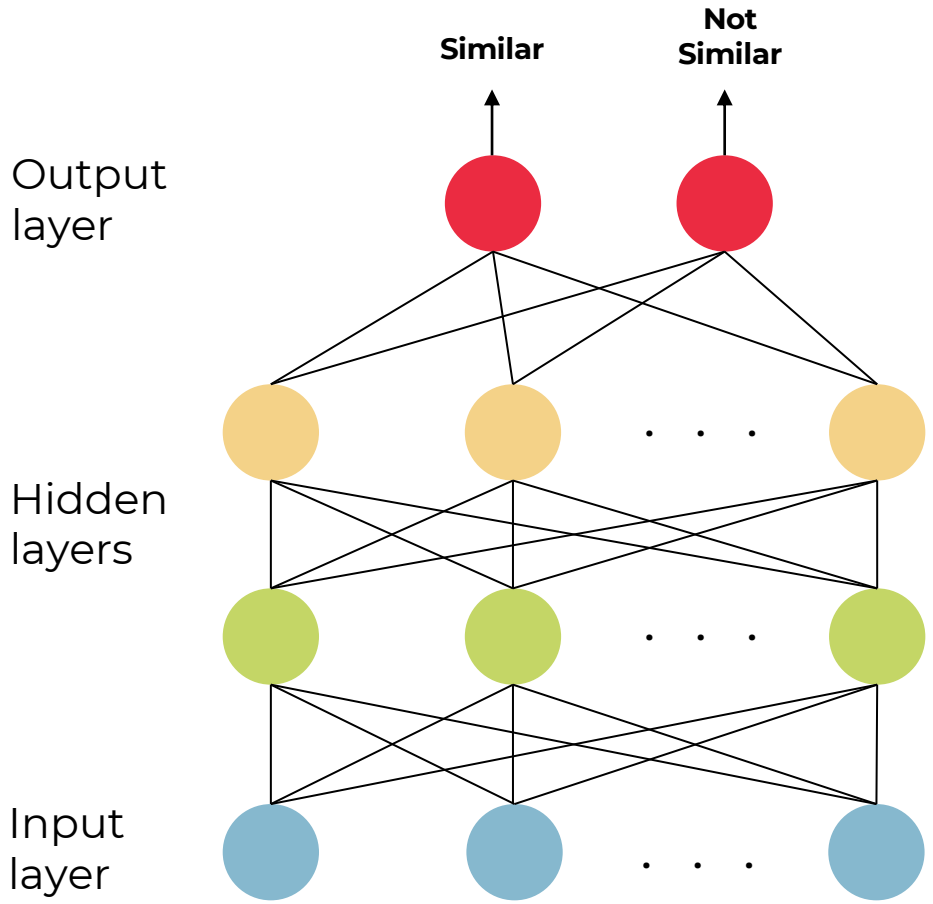
x_i = observed feature value

y = target

$\hat{y}$ = prediction

(Ranganthan *et al.*, *ABC of Bioinformatics*, 2019)

Multi-Layer Perceptron (MLP)



- Neural network-based model
- Input data goes through multiple layers of abstraction that maps it to the outputs
- Architecture that is more suitable for tabular data
  - Does not assume a spatial relationship between features

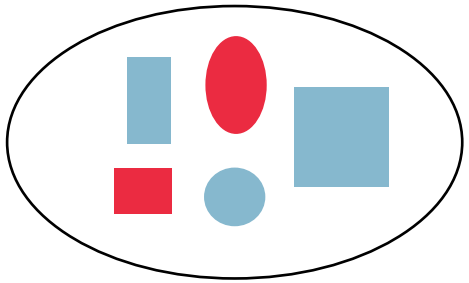
Key Parameters

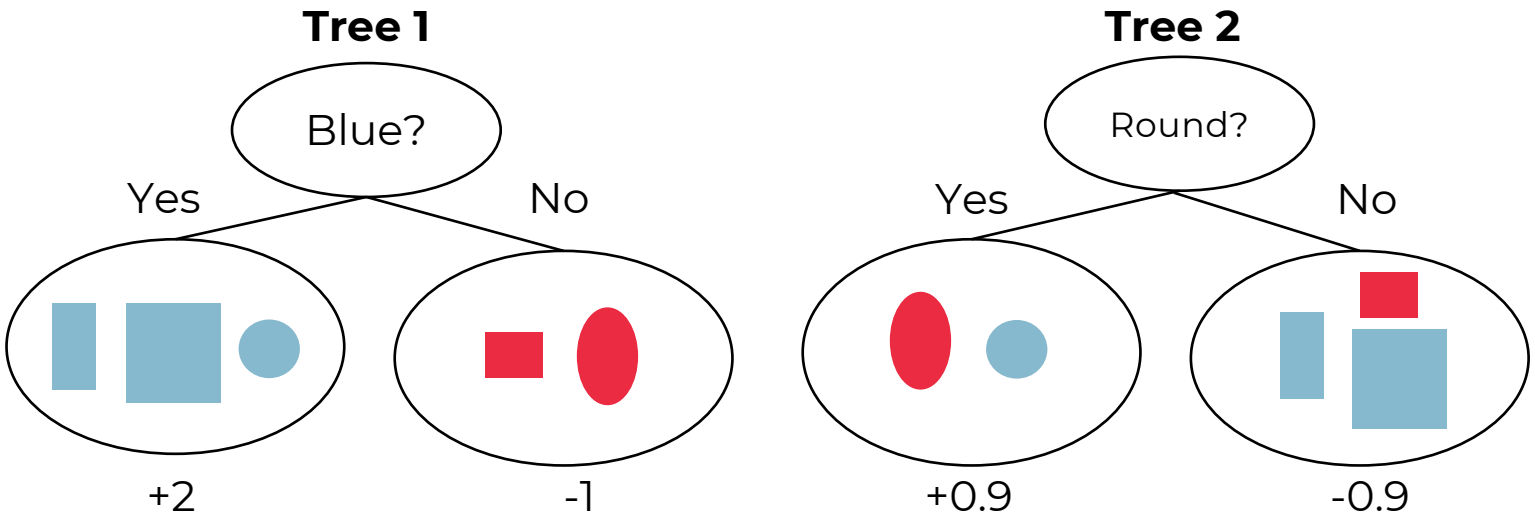
- Hidden layers
- Nodes/layer
- Activation function
- Alpha

(Brownlee, *Machine Learning Mastery*, 2022; Borisov *et al.*, *IEEE Transactions on Neural Networks and Learning Systems*, 2022)

Extreme Gradient Boosted Trees (XGBoost)

Input:
Color, shape, aspect ratio,...





f( ) = (+2) + (+0.9) = +2.9

- Decision tree ensemble model
  - Individual classification and regression trees are combined to partition the data
- Shown to outperform neural network models on tabular data

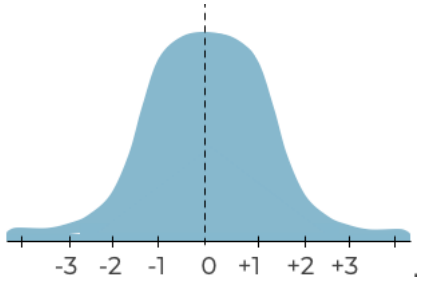
Key Parameters

- Number of trees
- Max depth
- Learning rate
- Regularization

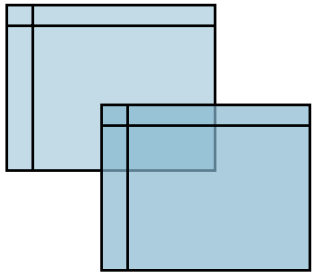
(XGBoost Developers, *XGBoost Tutorials,* 2022; Grinsztajn, *36th Conference on Neural Information Processing Systems*, 2022)

Model Training

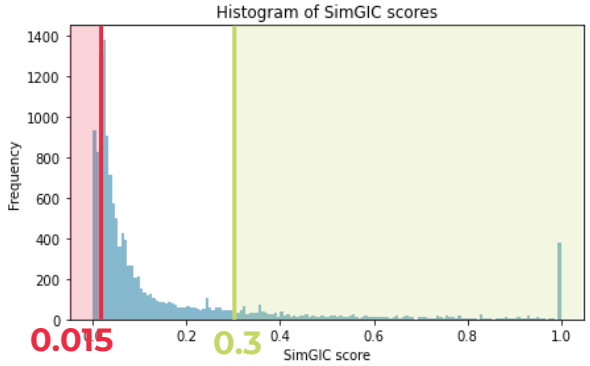
Feature standardization



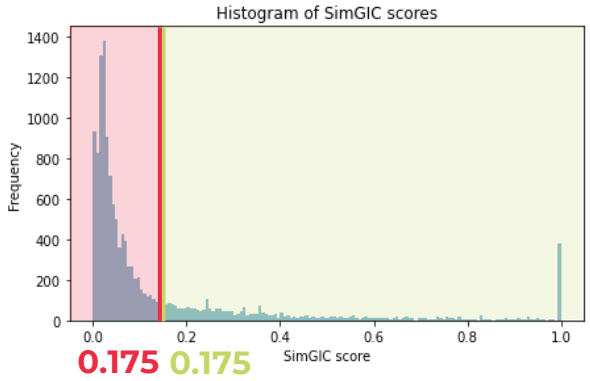
80-20 train-test data split



Remove 'ambiguous' pairs for training



Keep all pairs for testing



Hyperparameter tuning with 10-fold cross-validation

MLP

Hidden layers: 2
Nodes/layer: 20, 60
Activation func.: relu
Alpha: 5e-6

XGBoost

 # trees: 20
Max depth: 4
LR: 0.22
Regularization: 4

4

Results

Confusion Matrices

Threshold = 0.175

Naïve Bayes

		Actual	
		S	Not S
Predicted	S	334	347
	Not S	240	1771

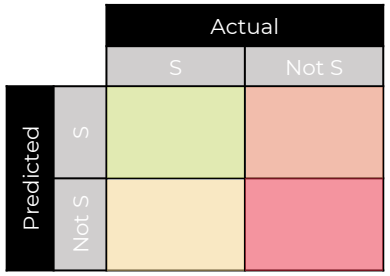
MLP

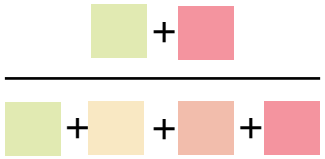
		Actual	
		S	Not S
Predicted	S	399	555
	Not S	175	1563

XGBoost

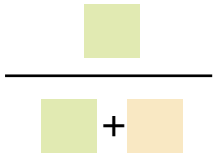
		Actual	
		S	Not S
Predicted	S	388	523
	Not S	186	1595

Performance Comparison

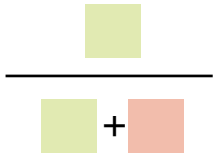




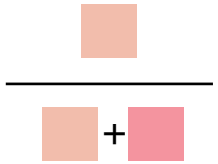
How often the model makes the right prediction



How often similar protein pairs are predicted as similar



How often a positive prediction is right

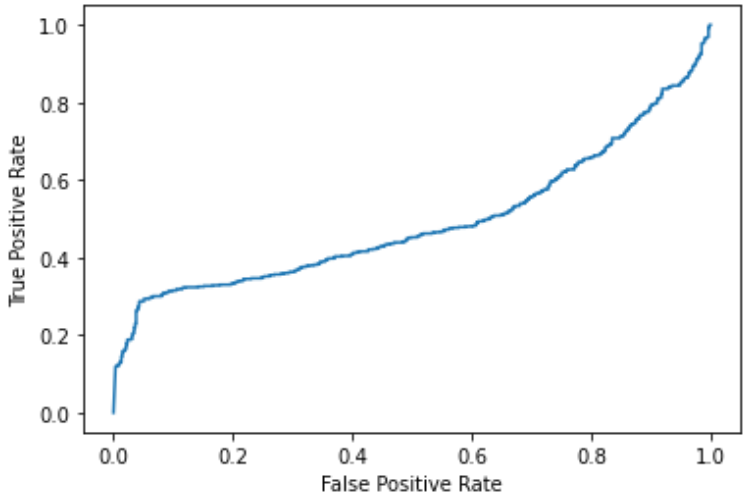


How often dissimilar protein pairs are predicted as similar

Model	Accuracy	Sensitivity	Positive Predictive Value	False Positive Rate
Naïve Bayes	0.766	0.526	0.756	**0.097**
MLP	**0.795**	0.632	**0.763**	0.113
XGBoost	0.789	**0.702**	0.712	0.162

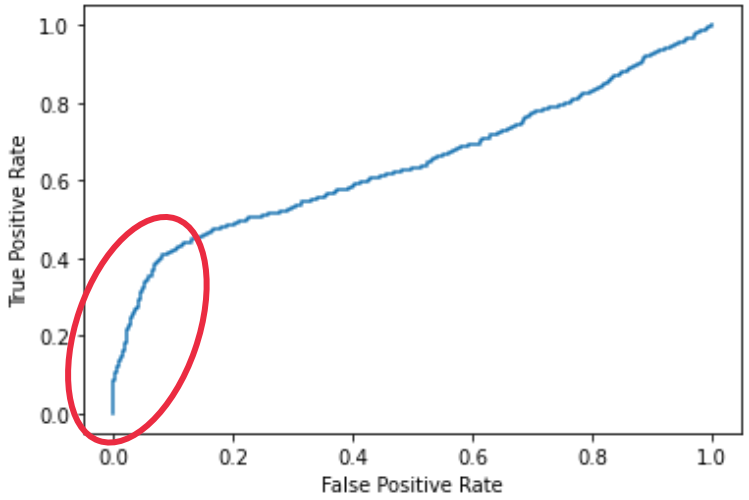
Receiver Operating Characteristic (ROC) Curves

Naïve Bayes



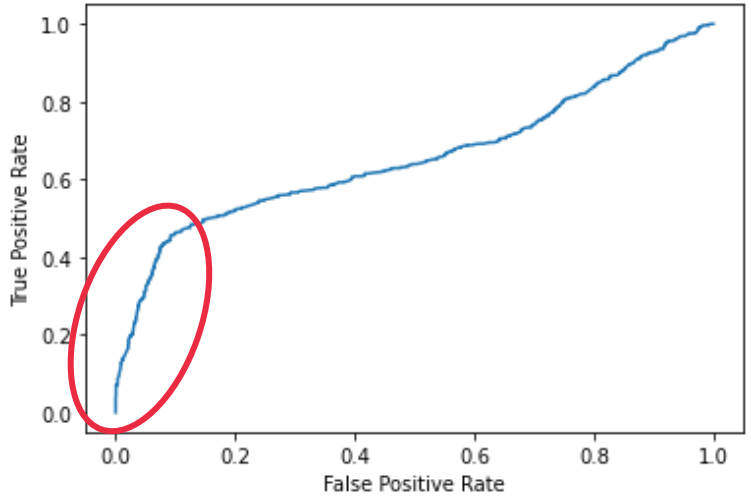
AUC-ROC = 0.714

MLP



AUC-ROC = 0.760

XGBoost



AUC-ROC = 0.770

Performance improvement comes from the models' abilities to distinguish similar protein pairs at higher cutoff thresholds

5 Future Work

Future Work

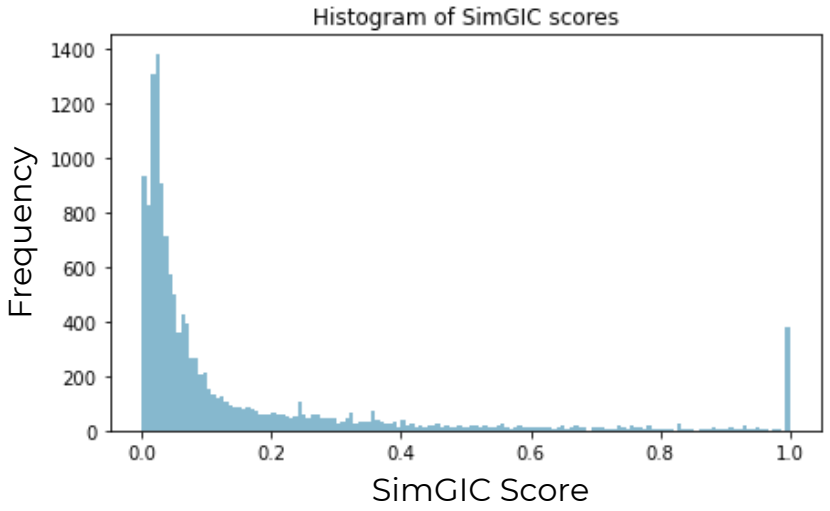
- There is always more data to integrate
- Regression models to directly predict SimGIC score
- Building a functional linkage network

Thank you.

Appendix

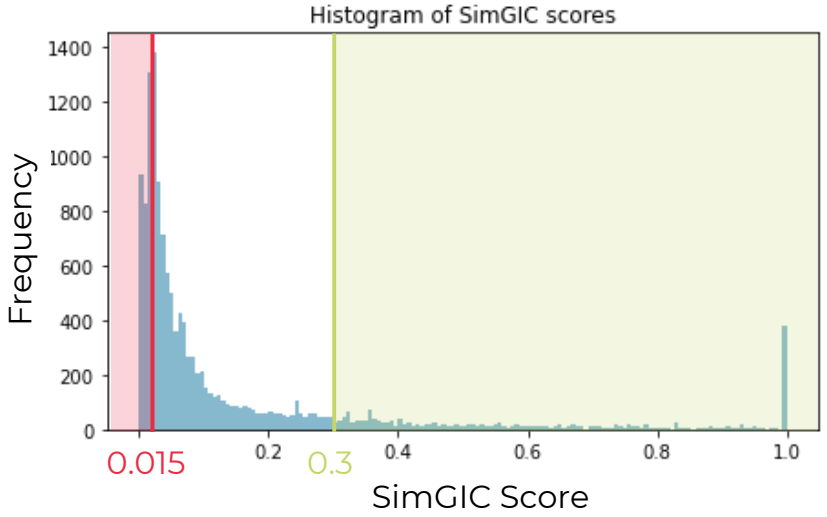
Train + Test Thresholds

A)

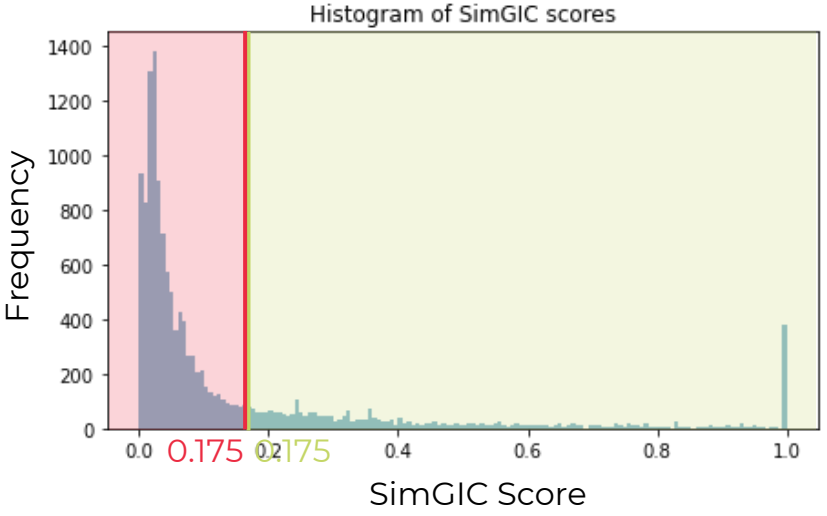


Thresholding

B) Train



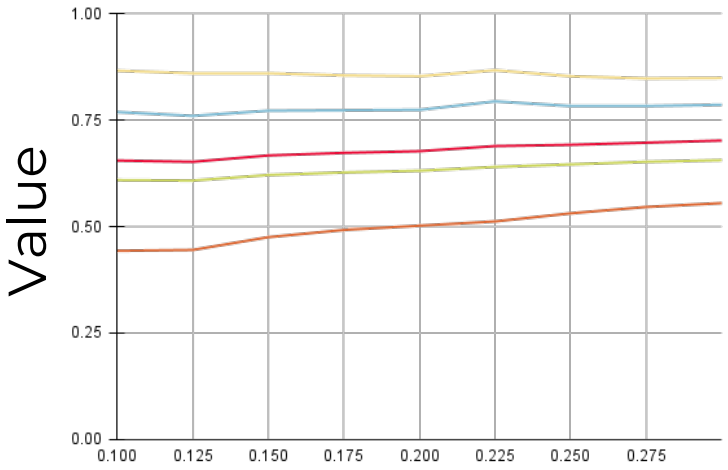
C) Test



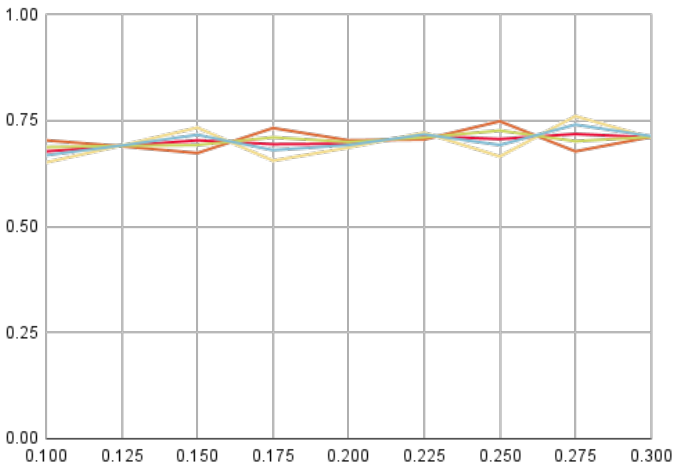
Appendix

Test Threshold Selection

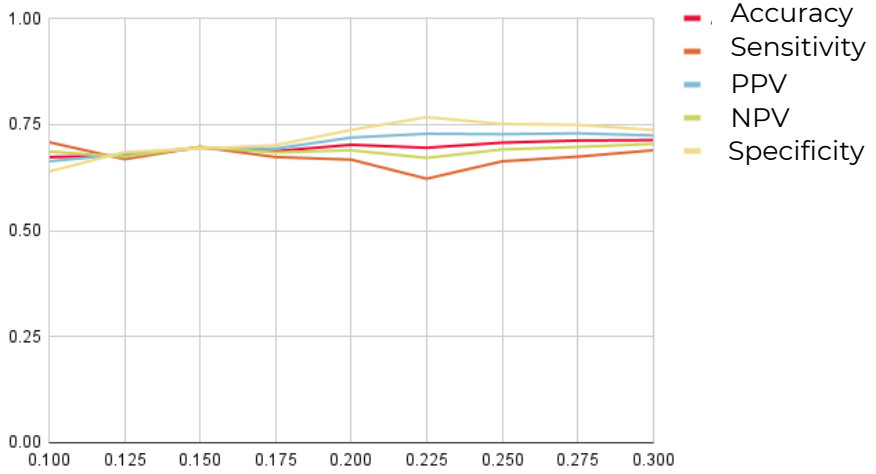
Naïve Bayes



MLP



XGBoost



Value

Threshold

Accuracy
Sensitivity
PPV
NPV
Specificity

References

1. L. A. Bugnon, E. Fenoy, A. A. Edera, J. Raad, G. Stegmayer, and D. H. Milone, "Transfer learning: The key to functionally annotate the protein universe," *Patterns,* vol. 4, no. 2, 2023, doi: 10.1016/j.patter.2023.100691.

2. M. Varadi *et al.*, "3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources," *GigaScience,* vol. 11, 2022, doi: https://doi.org/10.1093/gigascience/giac118.

3. M. Varadi *et al.*, "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences," *Nucleic Acids Research,* vol. 52, no. D1, pp. D368-D375, 2023, doi: https://doi.org/10.1093/nar/gkad1011.

4. B. Linghu, E. A. Franzosa, and Y. Xia, "Construction of functional linkage gene networks by data integration," *Data Mining for Systems Biology,* vol. 939, pp. 215-232, 2013, doi: https://doi.org/10.1007/978-1-62703-107-3_14.

5. A. Derry and R. B. Altman, "Explainable protein function annotation using local structure embeddings," *bioRxiv,* 2023, doi: https://doi.org/10.1101/2023.10.13.562298.

6. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research,* vol. 34, no. 1, pp. D535-D539, 2006, doi: 10.1093/nar/gkj109.

7. M. Usaj *et al.*, "TheCellMap.org: A web-accessible database for visualizing and mining the global yeast genetic interaction network," *G3: Genes, Genomes, Genetics,* vol. 7, no. 5, pp. 1539-1549, 2017, doi: 10.1534/g3.117.040220.

References

8. T. R. Hughes *et al.*, "Functional discovery via a compendium of expression profiles," *The Cell,* vol. 102, no. 1, pp. 109-126, 2000, doi: https://doi.org/10.1016/S0092-8674(00)00015-5.

9. Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horácková A, Melicher F, Perfetto L, Pokorný D, Lopez MR, Türková A, Wong ED, Xie Z, Casanova EB, Del-Toro N, Koch M, Porras P, Hermjakob H, Orchard S (2019). Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res, 47(d1):D550-D558, 01 Jan 2019, PMID: 30357405

10. Wishart DS, Li C, Marcu A, et al. PathBank: A Comprehensive Pathway Database for Model Organisms. *Nucleic Acids Res.* 2020 Jan 8;48(D1):D470-D478.

11. C. Pesquita, D. Faria, H. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinformatics,* vol. 9, 2008, doi: https://doi.org/10.1186/1471-2105-9-S5-S4.

12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) “Basic local alignment search tool.” J. Mol. Biol. 215:403-410.

13. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. (2008) “BLAST+: architecture and applications.” BMC Bioinformatics 10:421.

14. S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, "Bayes’ theorem and naive Bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*: Elsevier, 2019.

15. V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Transactions on Neural Networks and Learning Systems,* 2022, doi: 10.1109/TNNLS.2022.3229161.

16. J. Brownlee. "Crash course on multi-layer perceptron neural networks." Machine Learning Mastery. https://machinelearningmastery.com/neural-networks-crash-course/ (accessed January 28, 2025).

References

17. XGBoost developers. "Introduction to Boosted Trees." XGBoost Tutorials. https://xgboost.readthedocs.io/en/stable/tutorials/model.html (accessed January 28, 2024).

18. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," presented at the 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022.