Spatio-temporal expression analysis of genes involved in the electron transport chain of *Caenorhabditis elegans* using machine learning

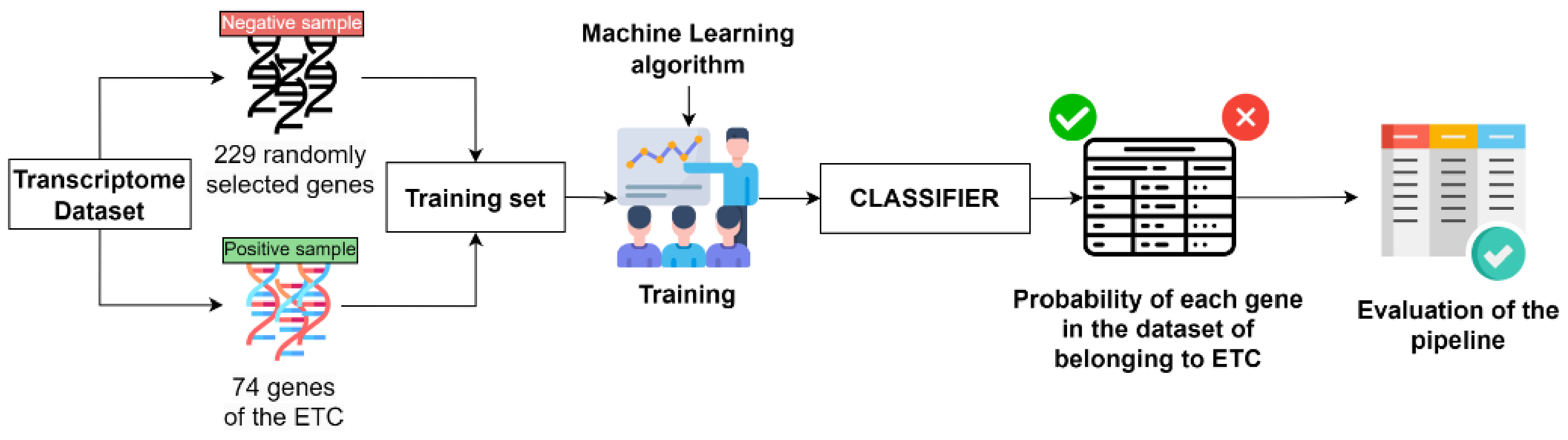
Sofía Zeballos[1,2]; Flavio Pazon-Obregón[1,3]; Gustavo Salinas[2]

1 - Department of Neurodevelopmental Biology, Instituto de Investigación Biológica Clemente Estable
2 - Worm Biology Laboratory, Institut Pasteur Montevideo
3 - Analytical Biochemistry and Proteomics Unit, Institut Pasteur Montevideo

INTRODUCTION

The electron transport chain (ETC) is a key process in cellular energy production and is vital for the proper functioning for most organisms. In the nematode *C. elegans*, the genes involved in this process have been well characterized, but there is still much to learn about genes associated with this process. In this study, we used a supervised machine learning approach to predict new genes associated with the ETC in *C. elegans* using an RNA-seq dataset that includes single cell expression during different points of embryo development and whole animal samples at each stage of the life cycle. We found that our approach was able to accurately identify several new genes that were previously unknown to be involved in this process. These discoveries are within the framework of a master's thesis that seeks to contribute to the annotation of gene functions in C. elegans through machine learning techniques.

VISUAL ABSTRACT


Negative sample
229 randomly selected genes
Transcriptome Dataset
Positive sample
74 genes of the ETC
Training set
Machine Learning algorithm
Training
CLASSIFIER
Probability of each gene in the dataset of belonging to ETC
Evaluation of the pipeline

METHODS

RESEARCH

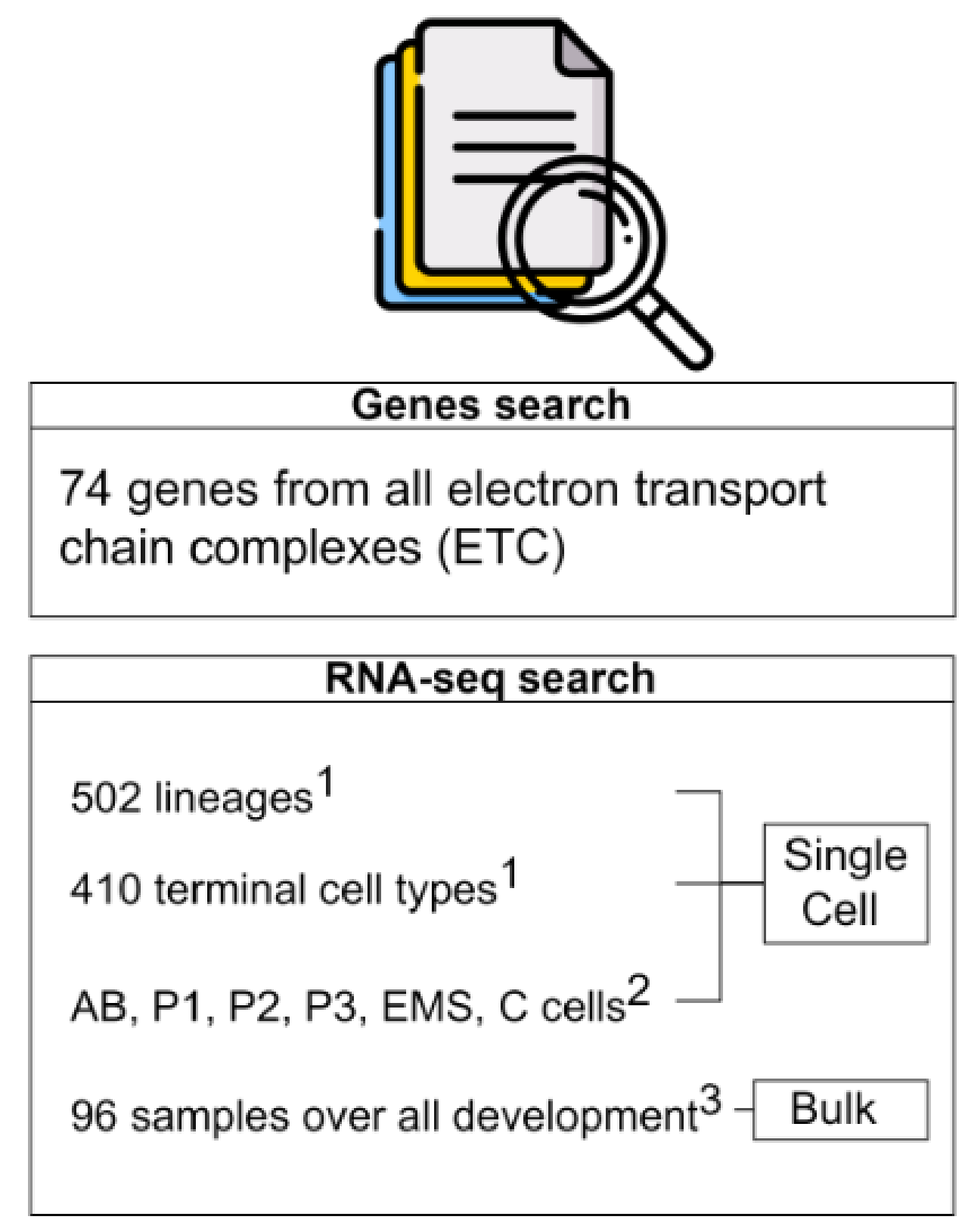

Genes search
74 genes from all electron transport chain complexes (ETC)
RNA-seq search
502 lineages[1]
410 terminal cell types[1]
AB, P1, P2, P3, EMS, C cells[2]
Single Cell
96 samples over all development[3]
Bulk

Fig. 1 - Bibliographic search to define the dataset and training sample was conducted. Three studies with freely accessible gene expression data were found, two of which performed single-cell assays and one on whole organisms. Additionally, the genes that form the five complexes of the electron transport chain (ETC) were defined by mapping human genes to the *C. elegans* genome. A list of 74 genes was identified, including 36 from complex I, 5 from complex II, 10 from complex III, 8 from complex IV, and 15 from complex V.

TRAINING

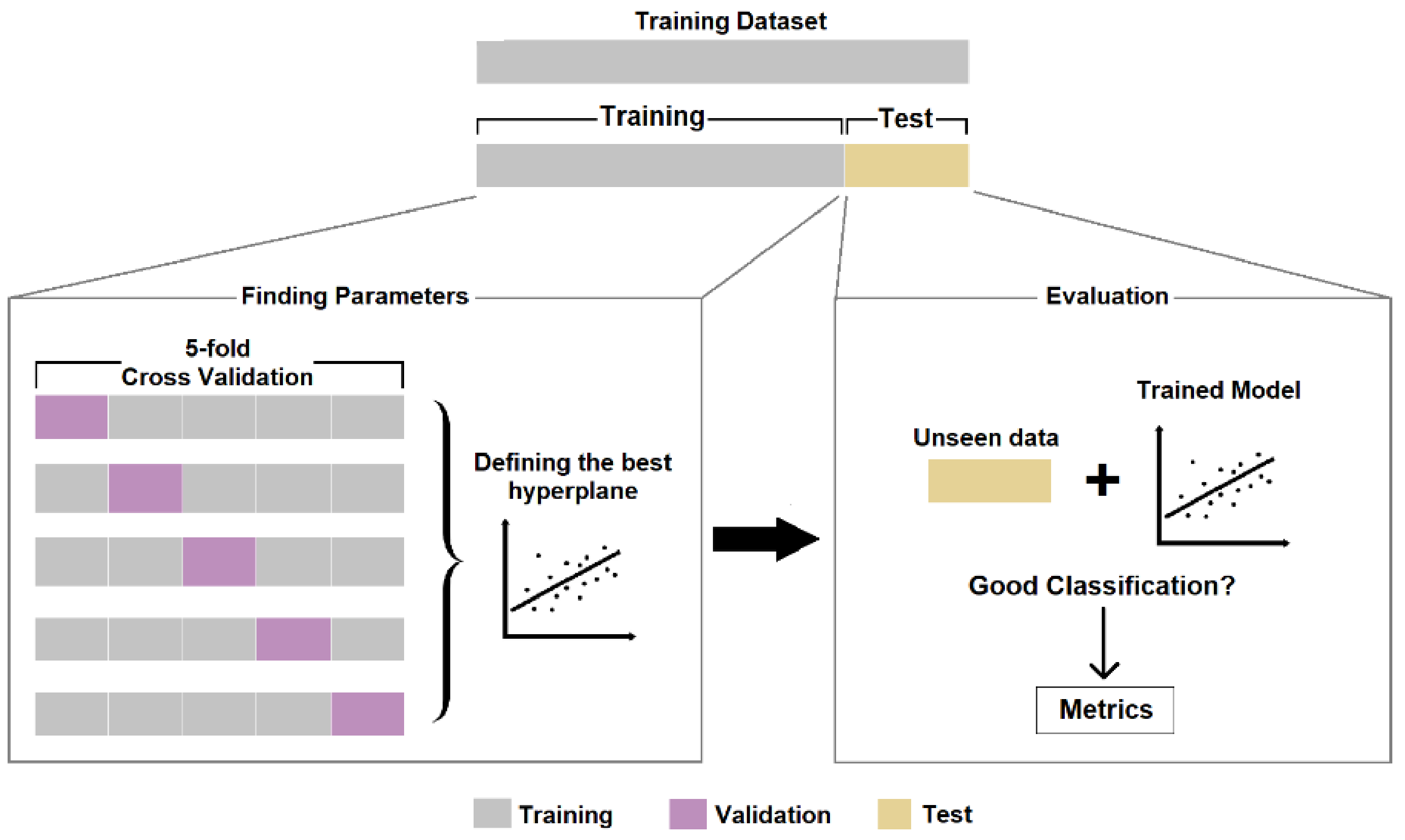
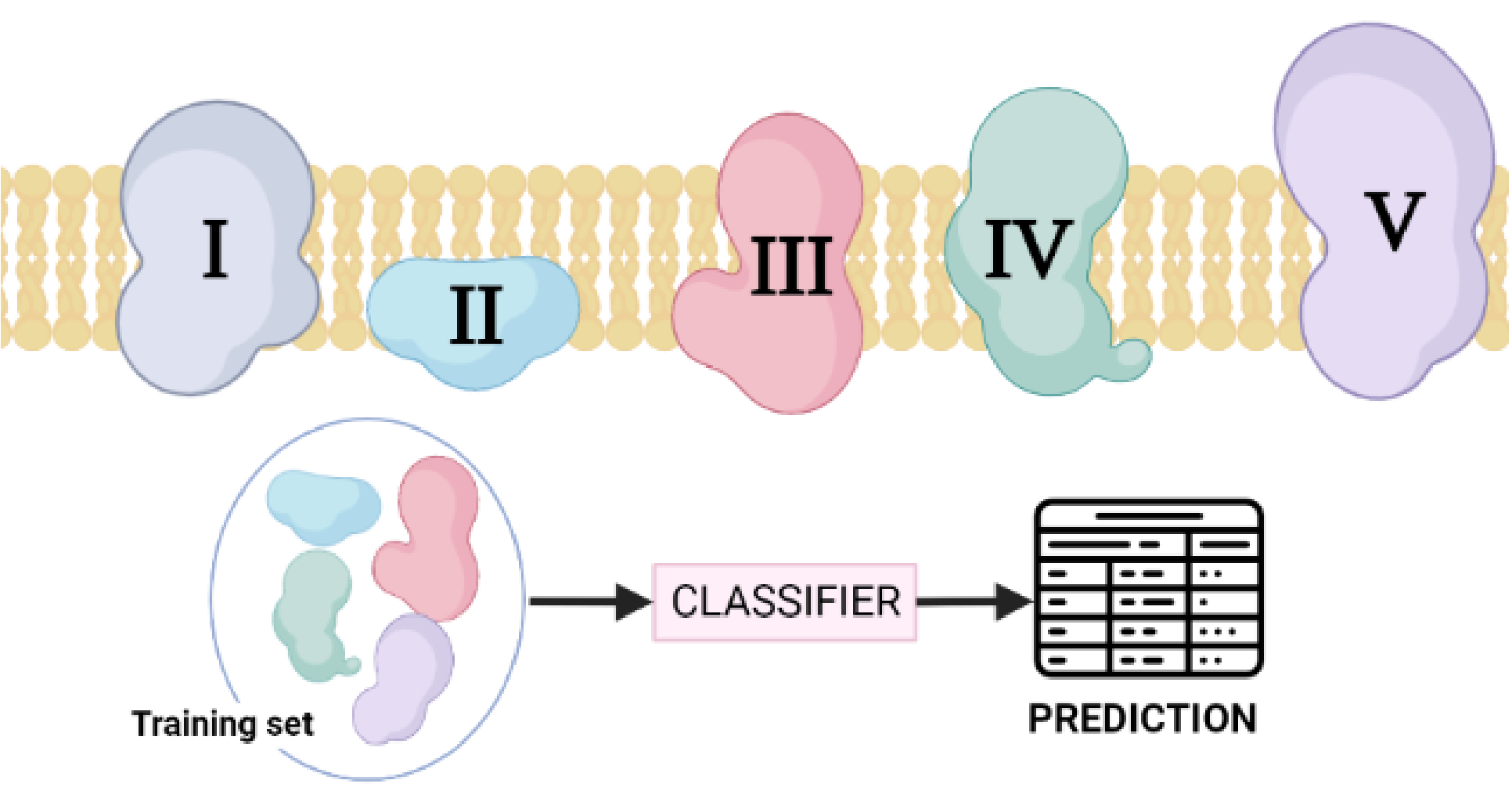


Fig. 2 - Workflow for training a Support Vector Machine (SVM) algorithm. The 74 genes were used as a positive sample, while 229 randomly selected genes were used as a negative sample for training. Once the model was trained, the probability of belonging to the ETC was calculated for the remaining genes in the dataset. A list of 1839 genes was obtained.

OPTIMIZATION



Steps:

1.Train the algorithm with 4 of the 5 complexes

2. Predict which genes of the rest of the RNA dataset may be similar to the training set

3. Repete with all complexes

Fig. 3 - Optimization of the list of candidate genes belonging to the ETC. In this step, five new training samples were created, in which each of the five complexes was removed, and the algorithm was trained accordingly. In total, six lists of candidate genes were obtained.

EVALUATION

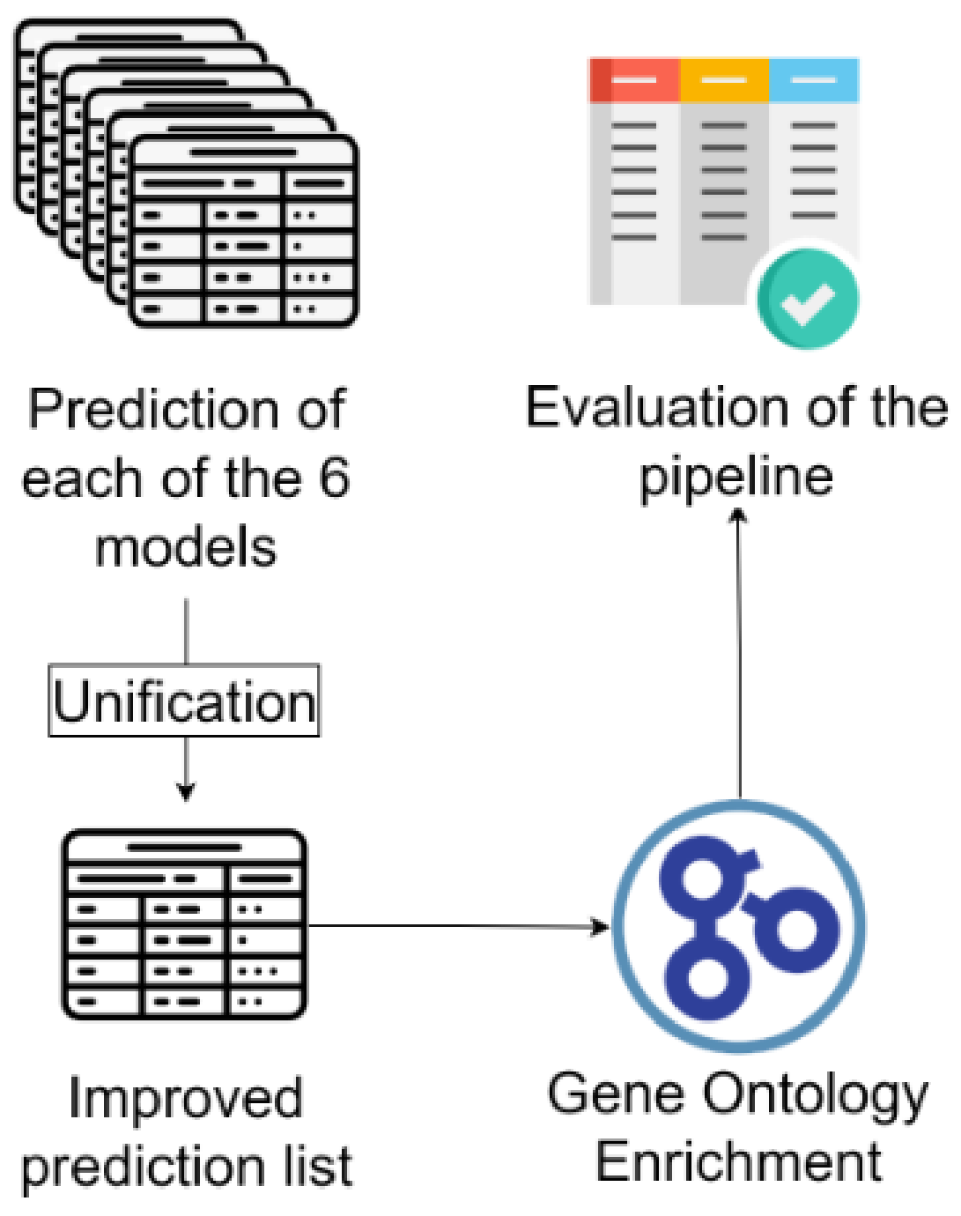
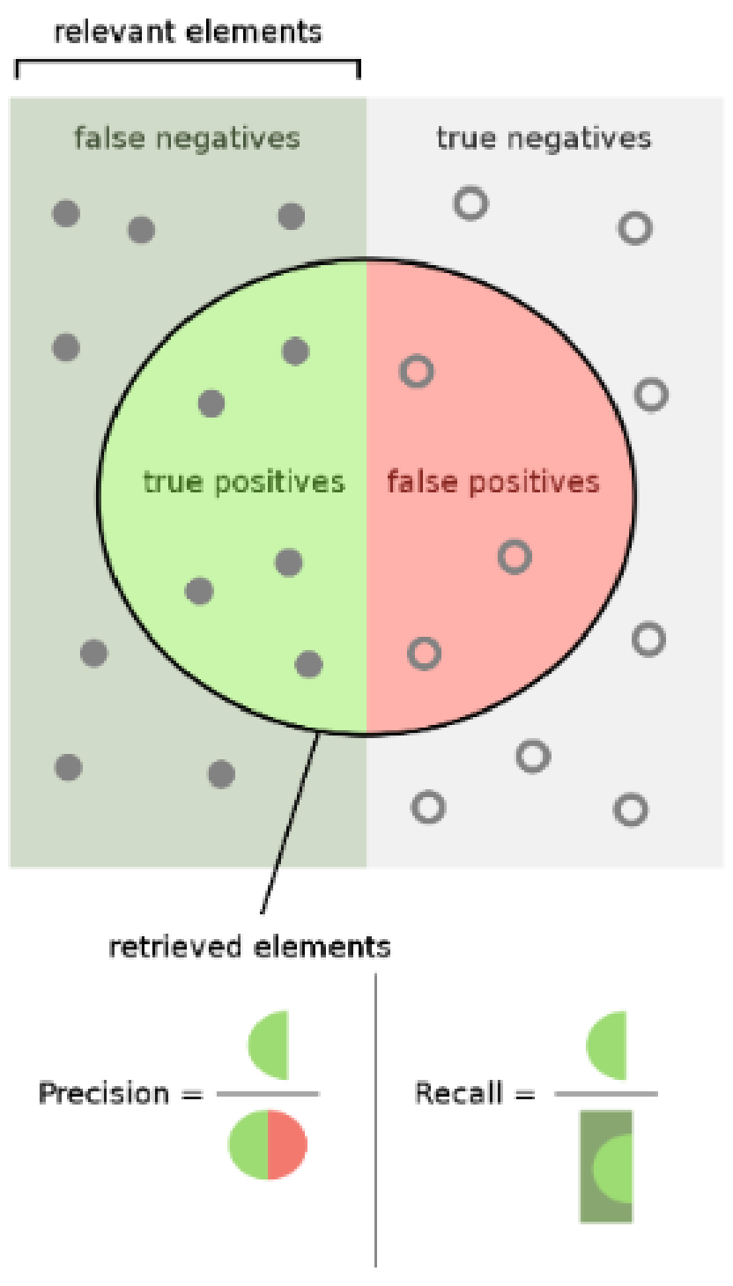

Prediction of each of the 6 models
Evaluation of the pipeline
Unification
Improved prediction list
Gene Ontology Enrichment

Fig. 4 - Model evaluation. A single list of candidate genes was created by taking the intersection of the six lists. A Gene Ontology (GO) term enrichment analysis was performed using this new list of genes, and it was evaluated whether it was enriched in terms associated with cellular respiration.

RESULTS



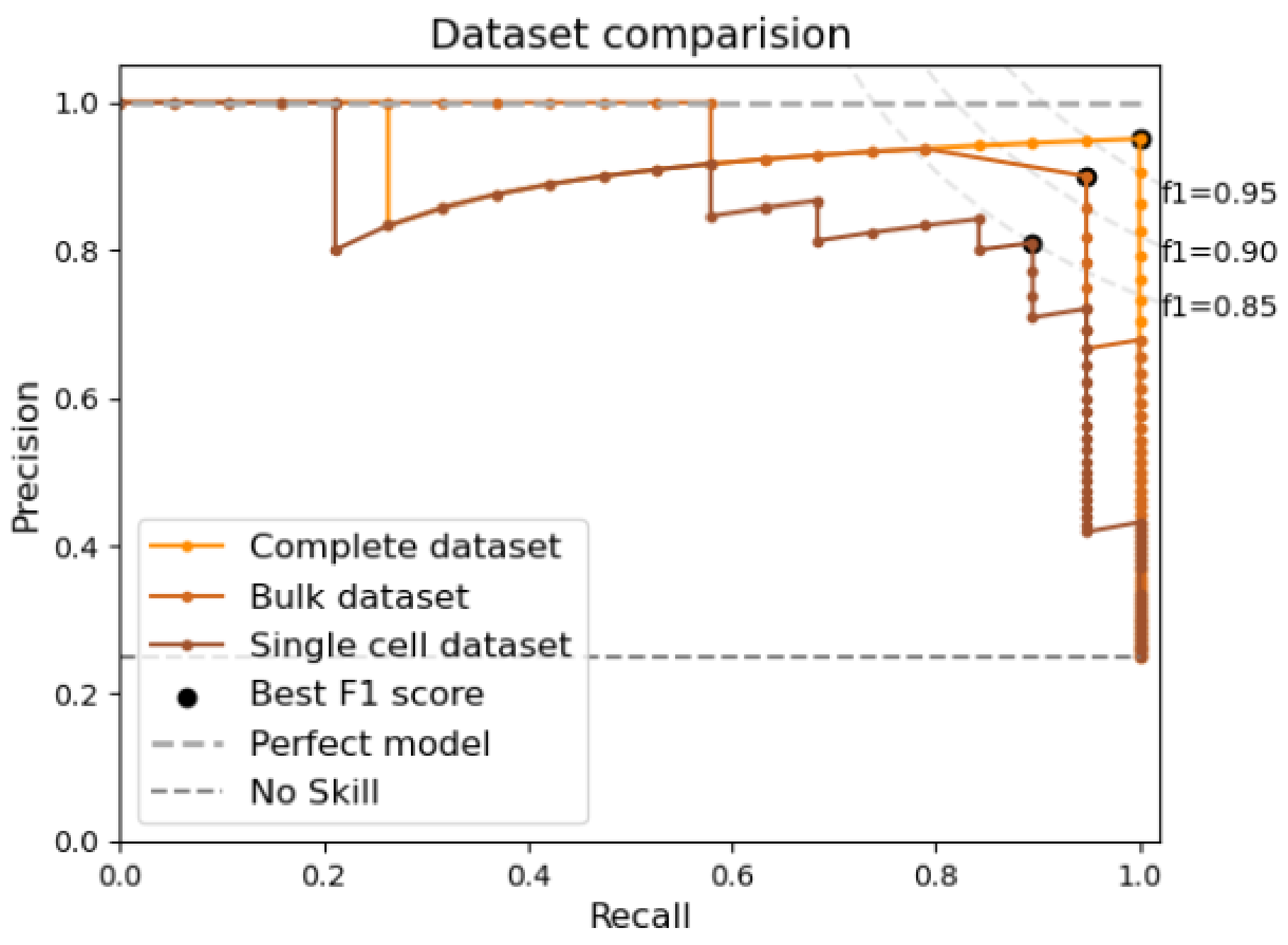


Fig. 6 - Comparison of different gene expression subsets. The performance of the model was evaluated using the complete training set (303 genes) and different subsets of expression data. The figure shows the performance based on the precision-recall curve for all the data (light orange), using only whole organism data (orange) and single-cell data (brown). It can be observed that although all three models perform well, the best metrics are obtained using the complete dataset (F1=0.97).

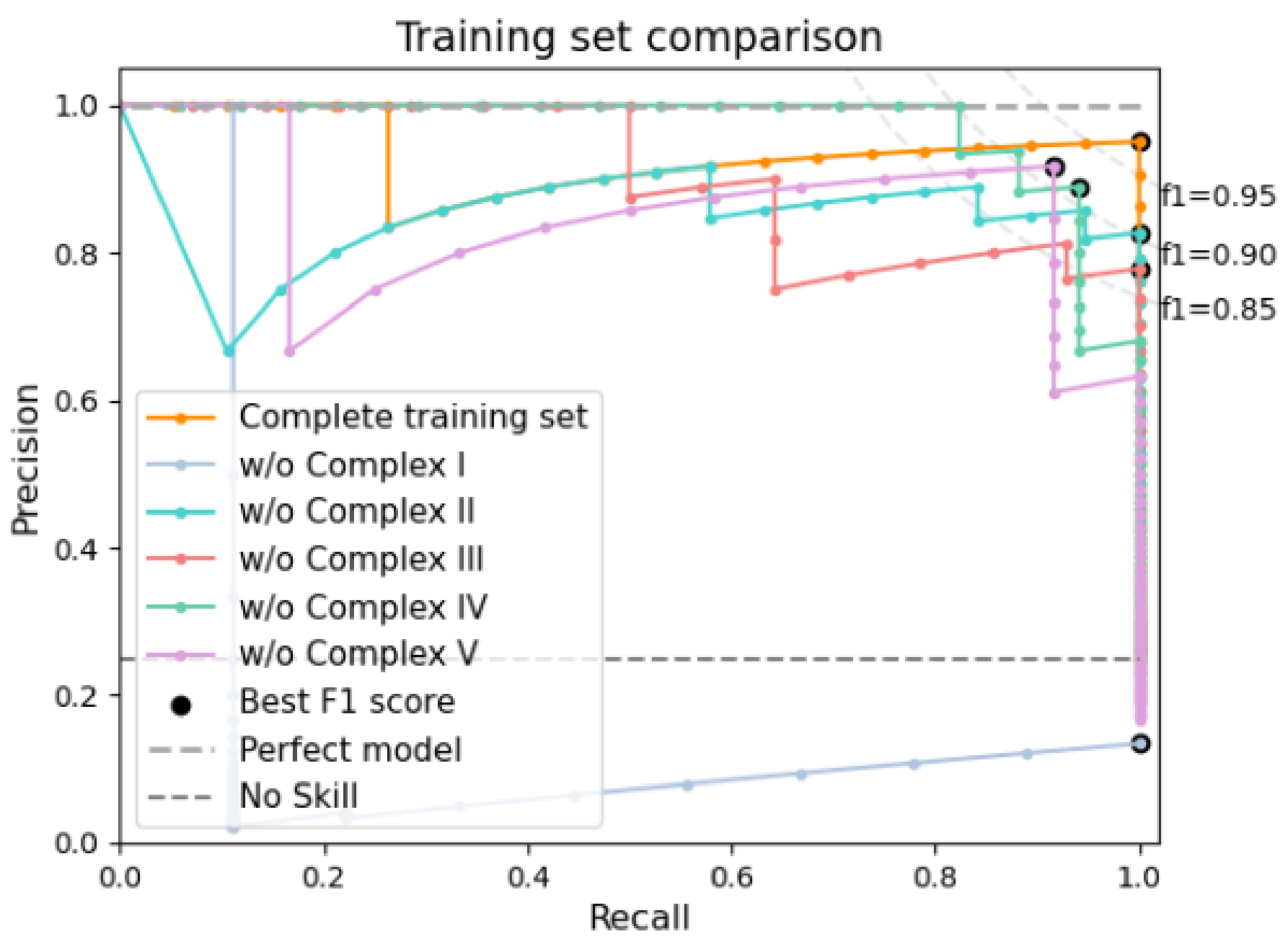


Fig. 7 - Comparison of different training sets. Five new training s were created by removing the genes from each of the complexes. The performance of the five models was evaluated using the precision-recall curve. Although they generally have good performance, none of the new models outperforms the use of all the genes (F1=0.97).

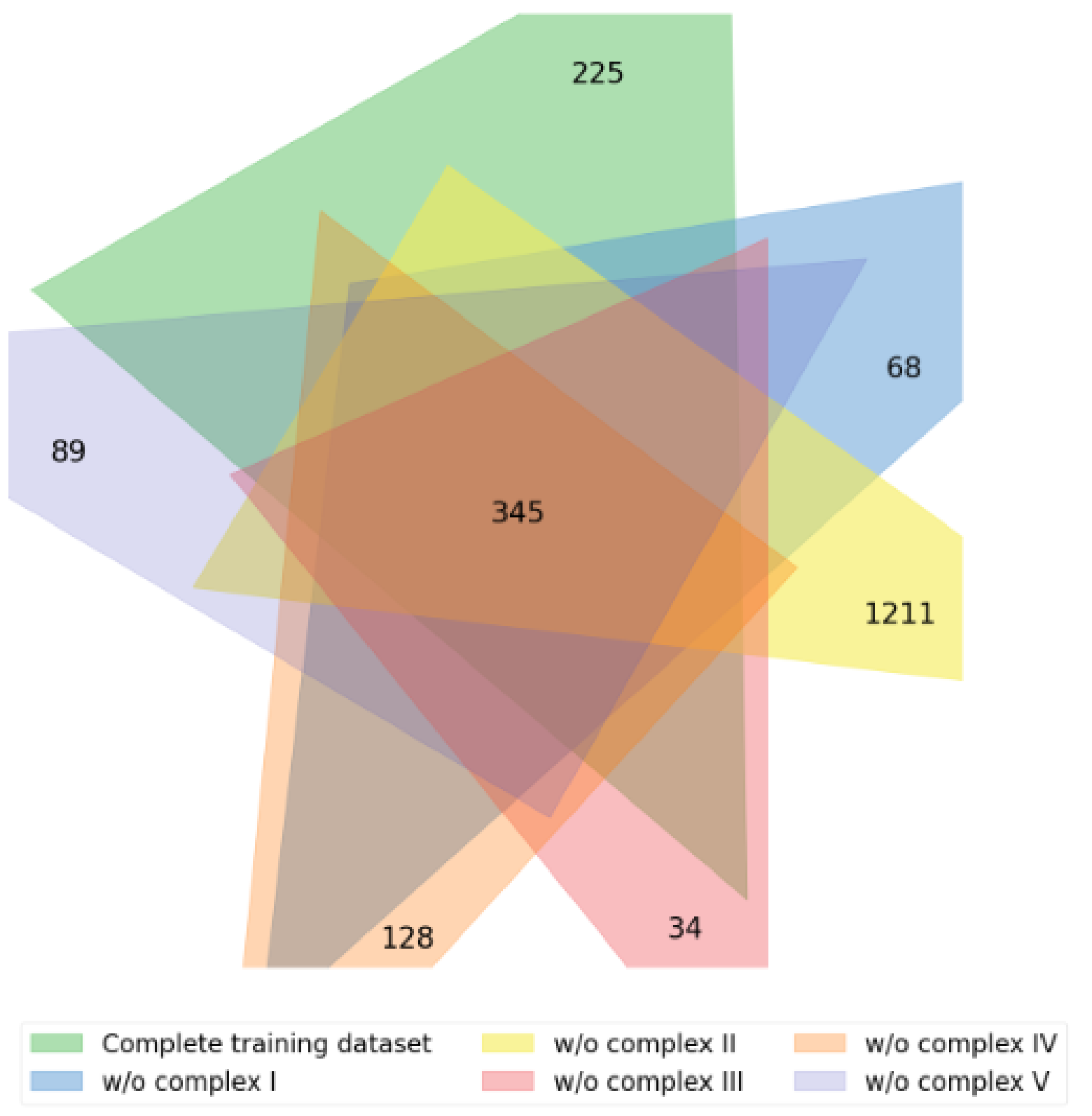


Fig. 8: Venn diagram of the predicted gene lists with the 6 models. It can be observed that the 6 models recover 382 genes in common.

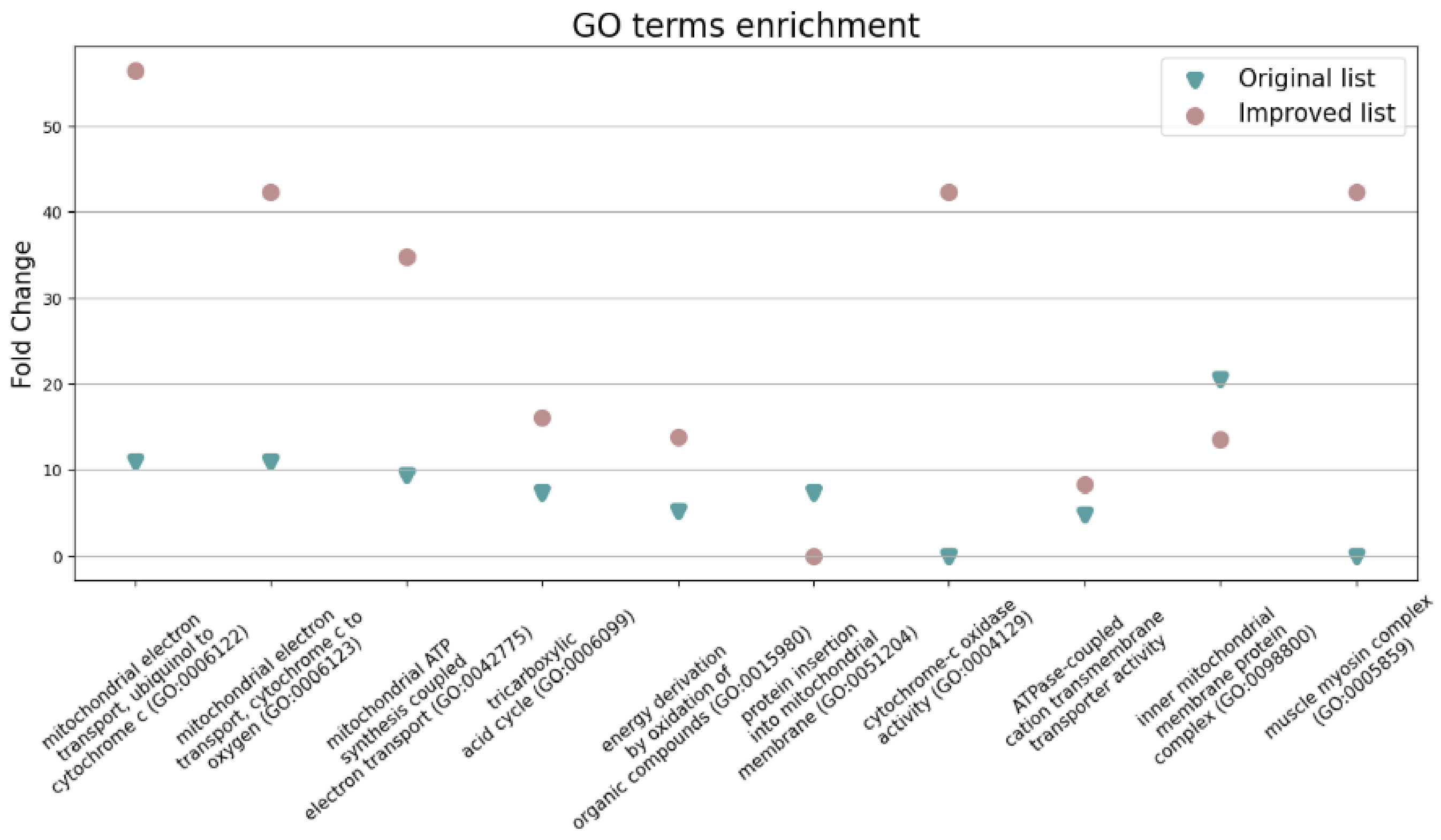


Fig. 9: GO term enrichment. Enrichment was evaluated for the gene list obtained by predicting with all ETC genes compared to the gene list optimized with the intersection of the 6 lists.

CONCLUSION

Single cell and bulk datasets together are the best combination to obtain a model with high precision.

Although there are complexes within the ETC that are associated with other biological processes that could distort the prediction, the model accuracy is better when using the complete chain.

There are 345 genes in common in the prediction lists obtained from the use of the six models

Using the intersection between the six models, we obtain a larger fold change enrichment of processes associated with aerobic respiration, Krebs cycle, and muscle components, which are associated with the ETC.

94 of the 345 genes are not anotated in Gene Ontology

The use of artificial intelligence techniques is a promising strategy for analyzing massive biological data. This work presents a possible approach to identifying candidate genes related to a specific biological process, which can be taken as input for experimental validation.

1 - Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., Waterston, R. H., & Murray, J. I. (2020). A lineage-resolved molecular atlas of C. elegans embryogenesis at single cell resolution. e*nce, 3*65(6459). https://doi.org/doi:10.1126/science.aax1971
2- r, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., Waterston, R. H., & Murray, J. I. (2020). A lineage-resolved molecular atlas of C. elegans embryogenesis at single cell resolution. Scien*ce, 365*(6459). https://doi.org/doi:10.1126/science.aax1971
3- Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. 2017. "Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism." *Science* 357(6352):661–67. doi: 10.1126/science.aam8940.

Created with BioRender Poster Builder