promor: An R package for Label-free Proteomics Data Analysis and Building Machine Learning Models with Candidate Proteins

Chathurani Ranathunge[1], Sagar S. Patel[1], Lubna Pinky[1, 2], Vanessa L. Correll[1], Shimin Chen[1], O. John Semmes[1],
Robert K. Armstrong[1], C. Donald Combs[1], Julius O. Nyalwidhe[1]

[1]Eastern Virginia Medical School, Norfolk, VA, [2]Meharry Medical College, Nashville, TN

Motivation

- **Label-free quantification (LFQ)** approaches are commonly used in proteomics research.
- There is still a great need for specialized tools that can make the downstream statistical analysis of LFQ data **widely accessible and reproducible.**
- **promor** is a user-friendly, comprehensive **R package** that streamlines differential expression analysis of LFQ data and building machine learning (ML) - based predictive models with top protein candidates.

promor includes specialized tools for:

- Quality control
- Visualization
- Missing data imputation
- Data normalization
- Differential expression analysis
- Feature selection
- Building ML-based models
- Model evaluation

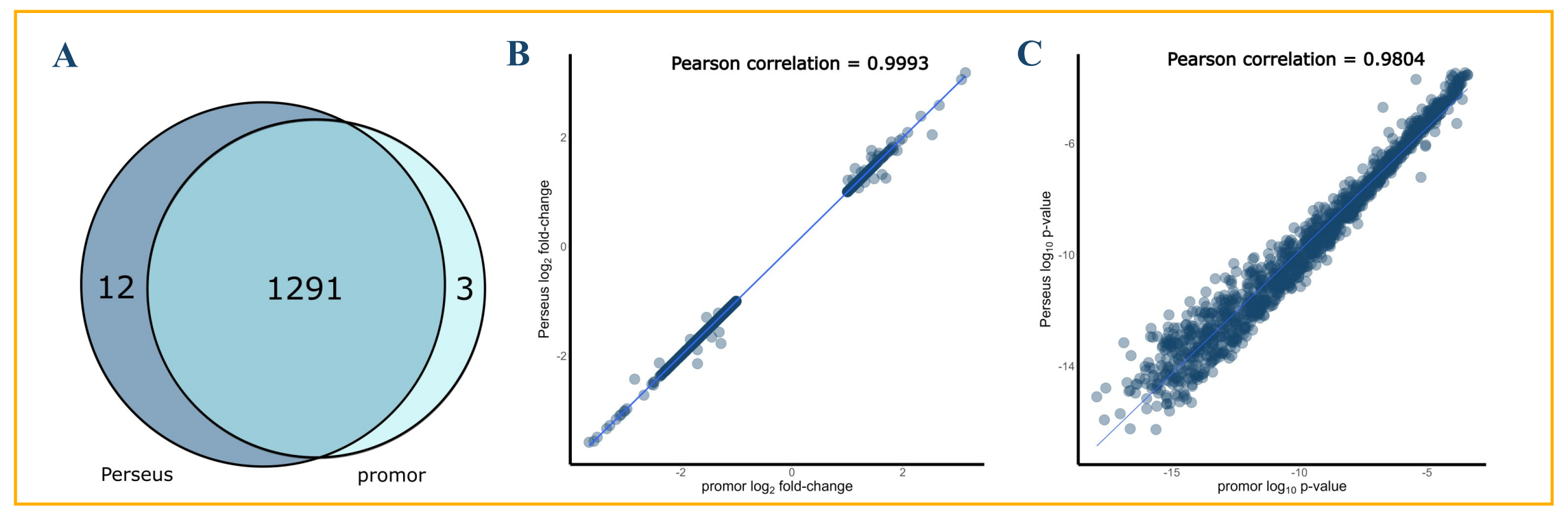
Input files for promor are:

- A proteinGroups.txt file or a quantitative matrix of protein intensities
- A text file containing the experimental design

Benchmarking

- promor was benchmarked on a publicly available dataset **(PRIDE ID: PXD000279)**.
- The results from the differential expression analysis were compared against those from **Perseus[1]**.

promor and Perseus share 98.85% DE proteins in common (A) and show strong correlation between log-fold changes and P-values (B and C) calculated by both programs.

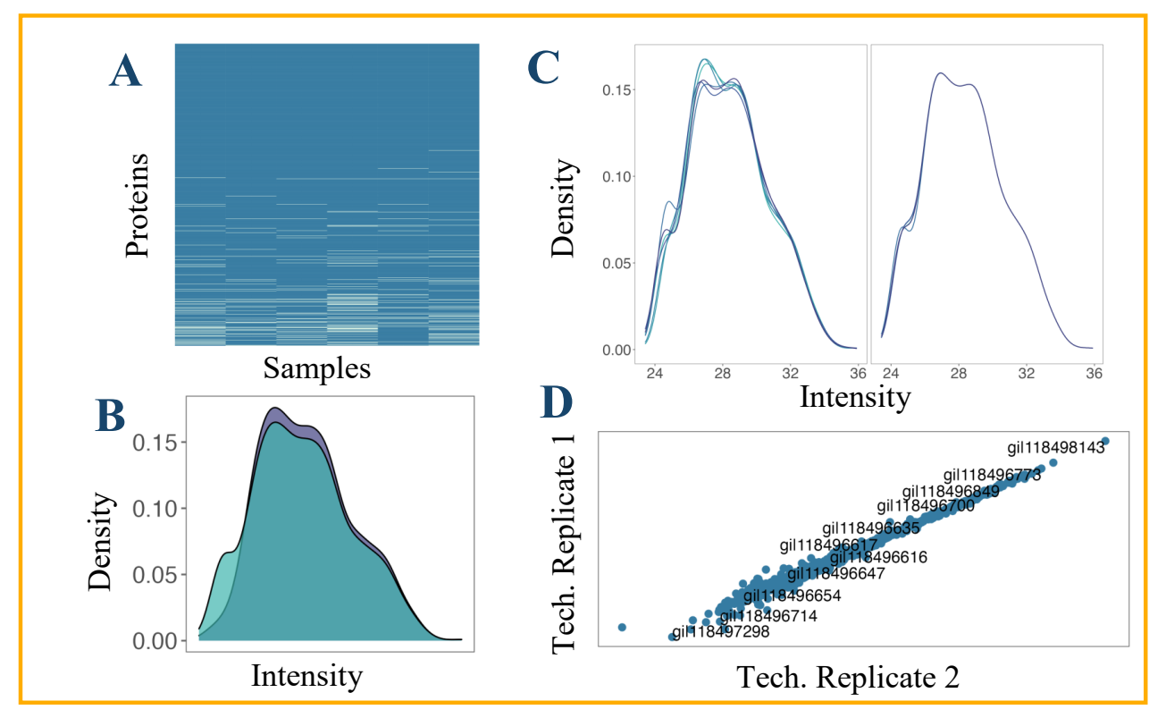


Quality control & Visualization

promor provides tools for:

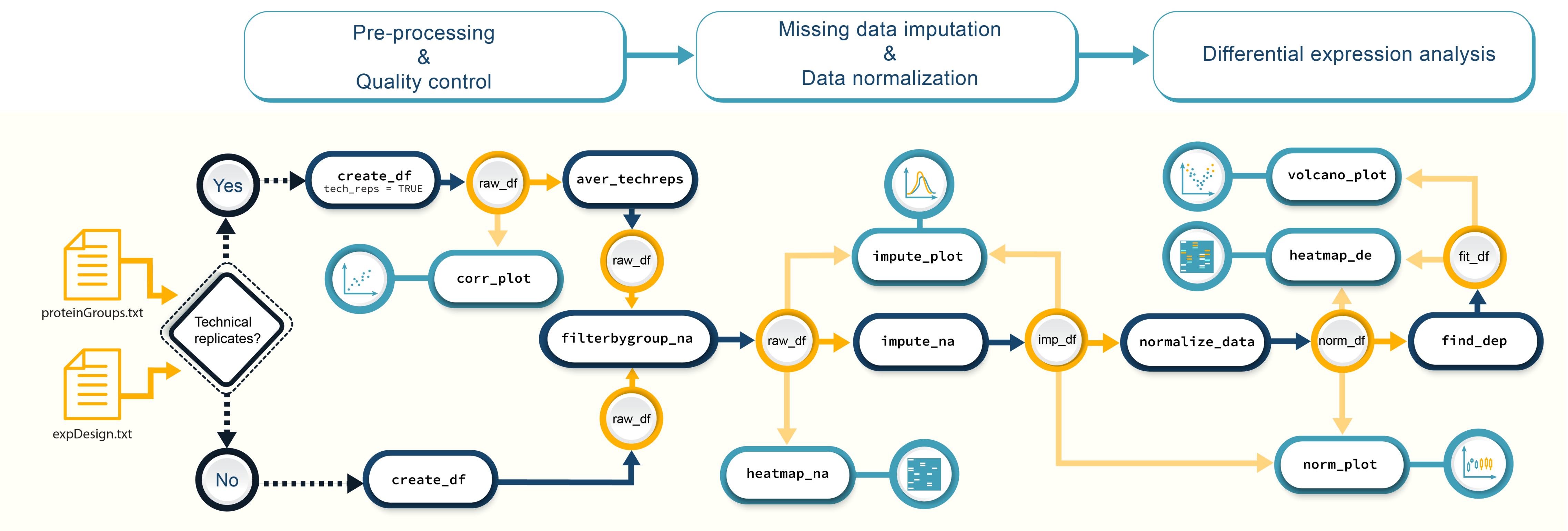
- **Filtering proteins** based on different criteria
- **Missing data imputation** with a variety of methods
- **Data normalization** with multiple methods
- A variety of **data visualization** options

Visualize (A) missing data distribution, (B) missing data imputation, (C) data normalization, and (D) correlation between technical replicates

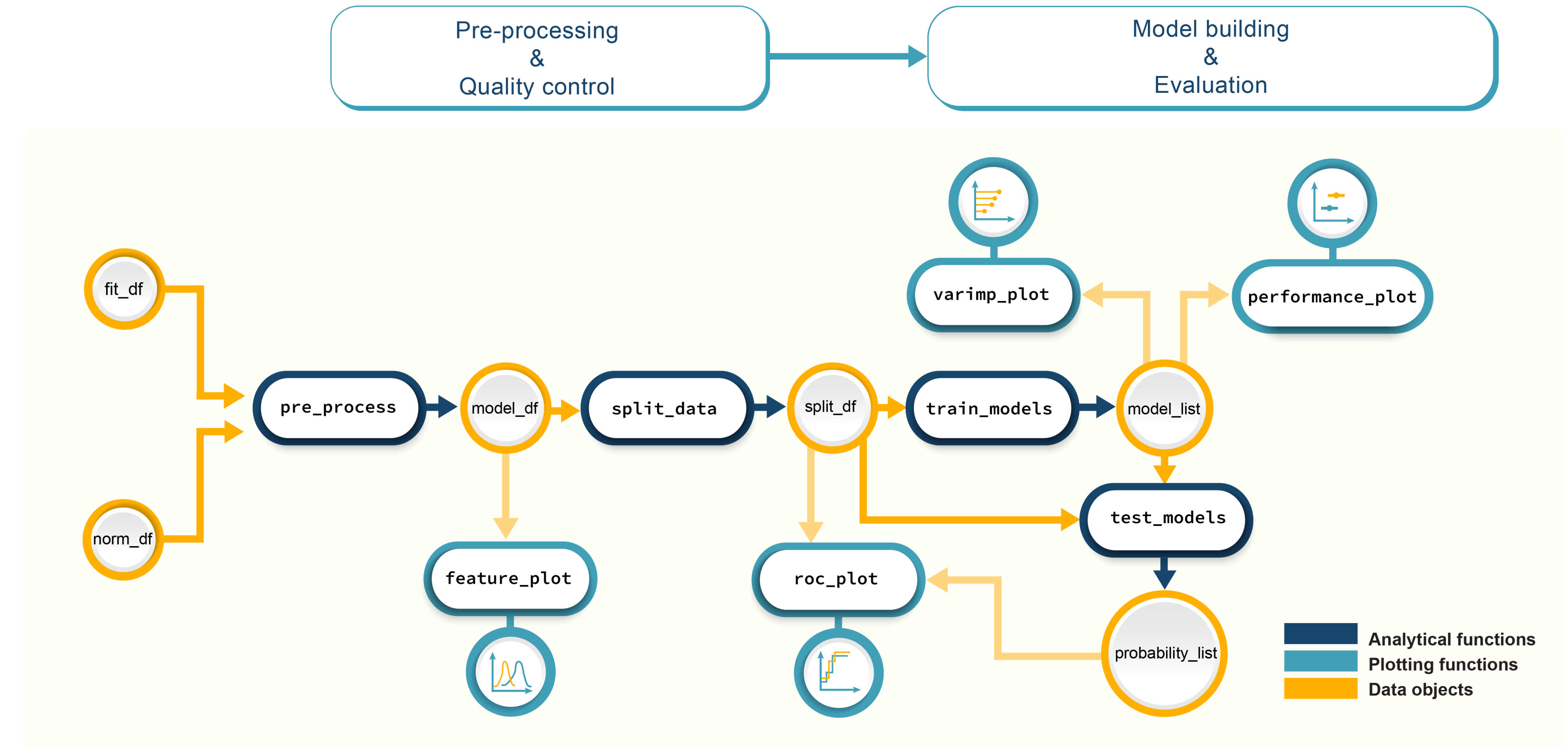


Workflows

Suggested promor workflows for analyzing LFQ proteomics data



Suggested promor workflows for building predictive models with protein candidates



Data & code availability

Data and code used for conducting analyses and making plots shown on this poster are available at: **https://github.com/caranathunge/asms2023**

Give promor a try!

References

1. Tyanova,S. et al. The Perseus computational platform for comprehensive analysis of proteomics data. Nat. Methods, 13 (2016): 731–740 .

2. Ritchie, Matthew E., et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic acids research 43.7 (2015): e47-e47.

Funding

This research was supported by **The Hampton Roads Biomedical Research Consortium.**
The authors declare no competing financial interest.

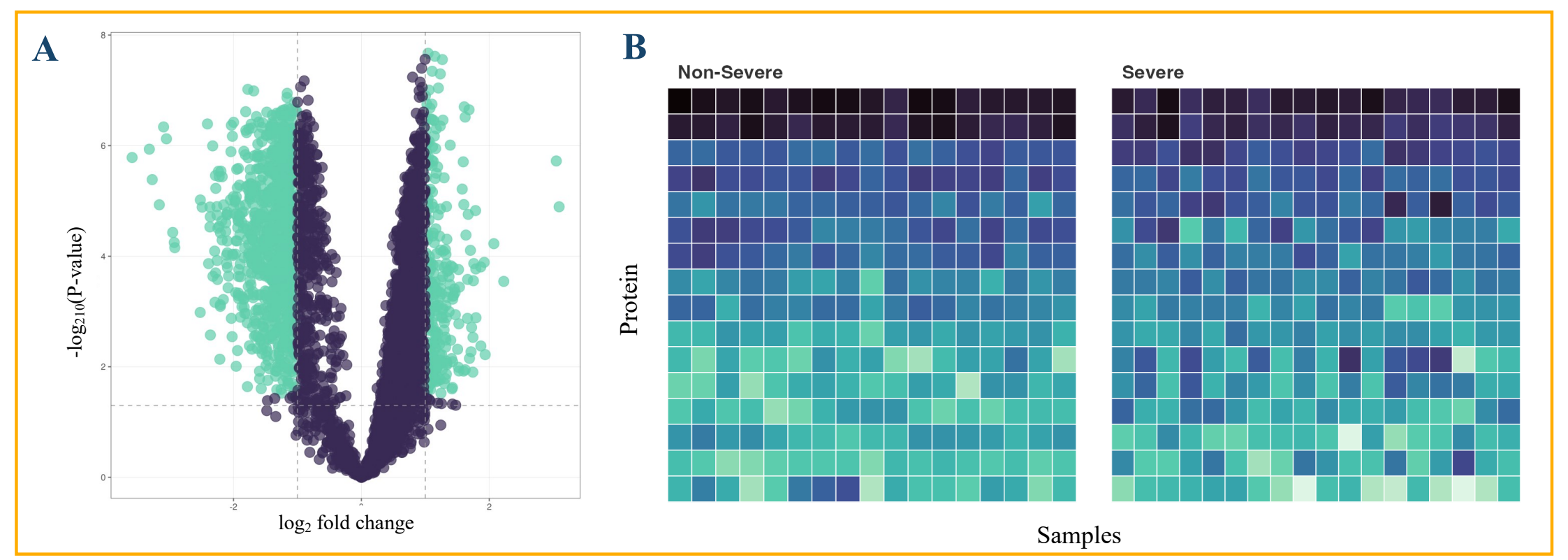
HAMPTON ROADS BIOMEDICAL RESEARCH CONSORTIUM.

EVMS Eastern Virginia Medical School

Differential expression analysis

- **promor** uses the empirical Bayes method implemented in the R package **limma[2]**, for differential expression analysis.

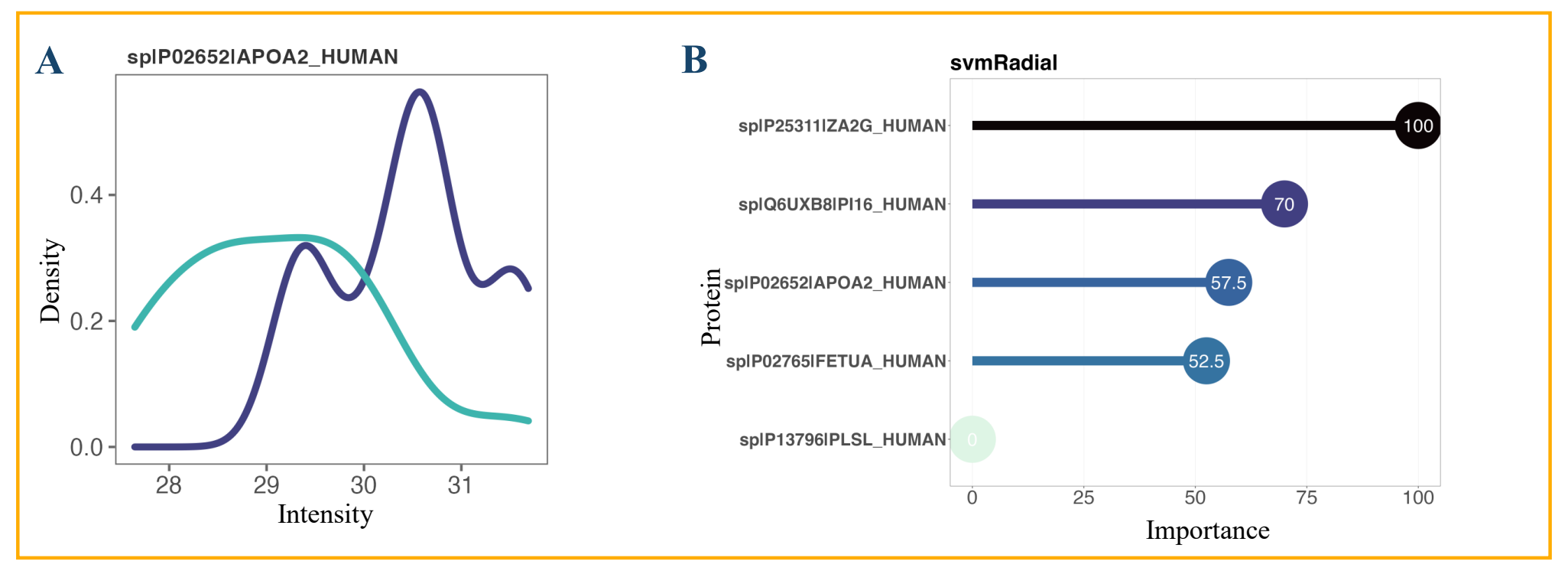
Visualize differentially expressed proteins using (A) volcano plots and (B) heatmaps



Feature selection

- **Feature plots:** Visualize protein (feature) variation among groups prior to building ML-based models.
- **Variable importance plots:** Assess the importance of different proteins (features) in the models built using different machine learning algorithms.

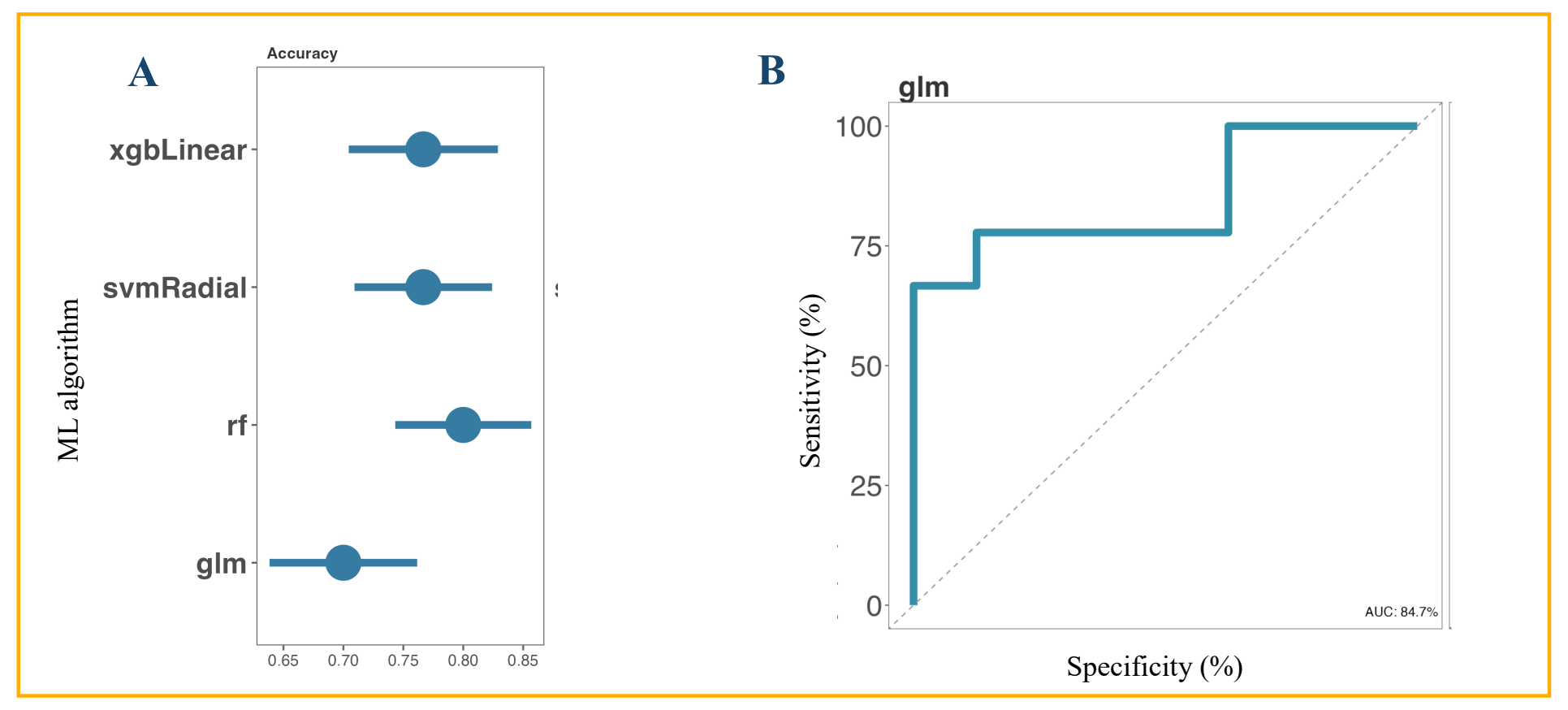
Visualize feature (protein) variation among groups with (A) feature plots, and the importance of different features (proteins) in the models built with (B) variable importance plots



Model building & Evaluation

- **promor** provides a default list of five widely used machine learning algorithms (out of about 200 accessible) to build predictive models.
- **Performance plots:** Use multiple metrics to assess the performance of models trained on training data.
- **ROC plots:** Build Receiver Operating Characteristic (ROC) curves to evaluate the diagnostic ability of models built using different algorithms.

Visualize model performance with (A) performance plots, and the diagnostic ability of models with (B) ROC plots


A
Accuracy
xgbLinear
svmRadial
rf
glm
ML algorithm
B
glm
Sensitivity (%)
Specificity (%)
AUC: 84.7%