2025-04-17

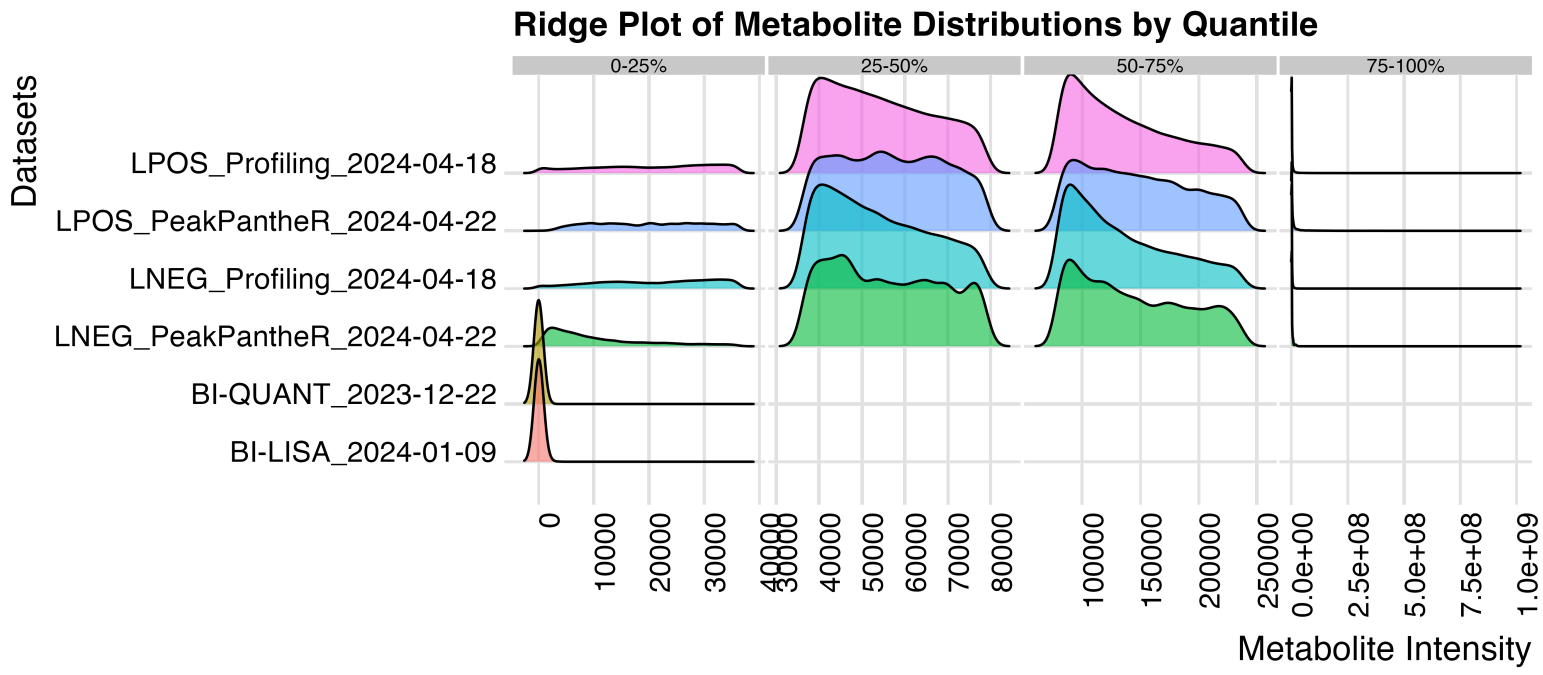
Data

Data Summary

	sample number	feature number	classification
LNEG_PeakPantheR_2024-04-22	101	125	Named Lipids
LNEG_Profiling_2024-04-18	101	1640	Untargeted mass spectrometry
LPOS_PeakPantheR_2024-04-22	100	302	Named Lipids
LPOS_Profiling_2024-04-18	100	7067	Untargeted mass spectrometry
BI-LISA_2024-01-09	98	112	Targeted metabolites from kits
BI-QUANT_2023-12-22	98	41	Targeted metabolites from kits

1. LPOS and LNEG datasets have similar distribution of intensity values
2. BI-QUANT and BI-LISA have similar distribution of intensity values

Different distribution of intensity between datasets



Results

Using both (un)annotated datasets to build PLS model

In first try, we used all 6 datasets to filter for important variables by statistical tests such as ANOVA or T-tests and build PLS-models accordingly based on the filtered metabolites/lipids. We then built ROC curves based on the PLS-models to test whether the model is predictive of the patient treatment response.

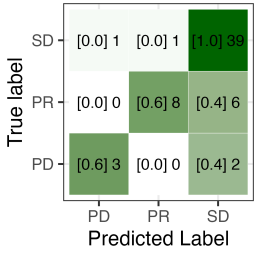
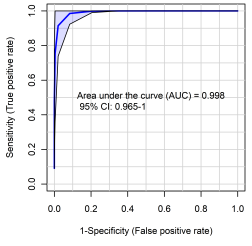
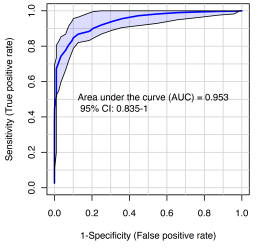
Explaining the Metrics

1. R^2 is the coefficient of Determination
    a. It is the proportion of variance in the **response variable** (e.g. the treatment response) explained by the PLS component.

   b. A high R^2 would therefore mean that the model captures most of the variance in your data.
   c. Inflation of R^2 may occur by overfitting, especially when there are too many metabolites (features) and too less samples.
2. Q^2 is the cross-validated predictive ability
   a. It is the proportion of variance predicted during **cross-validation** (e.g. leave one out)
   b. Q^2 measures how well your model generalises to new data, a Q^2 over 0.5 is considered good, while $Q^2 < 0$ indicates that the model is even worse than random guesses
   c. When $Q^2 < R^2$ it's usually a sign of overfitting.

Result Summary and Interpretation

	3 Groups	PD vs PR	SD vs PD	SD vs PR
Metabolites used in PLS-model ––– selection of metabolites were performed by ANOVA analysis (for 3 groups) and T-test for comparison of two groups.				
LNEG PeakPantheR	0	0	0	0
LNEG Profiling	0	0	0	0
LPOS PeakPantheR	2	0	1	1
LPOS Profiling	28	4	9	5
BI-QUANT	0	0	0	0
BI-LISA	0	0	0	0
Total Number of Metabolites/Lipids used in PLS-model	30	4	10	6
Interpretation of PLS-model results				
n-component used	2	1	1	2
model accuracy	0.78	1	0.91	0.80

R2	0.41	0.87	0.43	0.24
Interpretation	Moderate accuracy but **low Q2/R2**: The model distinguishes groups but has poor predictive power. Likely due to heterogeneity in PD/SD/PR groups.	**Perfect accuracy, high Q2/R2**: Strong separation between PD (progressive disease) and PR (partial response). Risk of overfitting (sample size = 19).	High accuracy but **low Q2**: Model fits training data but generalizes poorly. SD (stable disease) and PD share overlapping metabolic features.	**Poor Q2/R2**: Weak predictive power. SD and PR may lack distinct metabolic drivers.
Result from ROC-curves				
Models Used	Random Forest (because there are more than 2 responses)	Partial Least Squares	Partial Least Squares	Partial Least Squares
AUC/ROC	NA, confusion matrix used instead 	AUC = 0.998 CI = 0.965-1 	AUC = 0.953 CI = 0.835-1 	AUC = 0.708 CI = 0.423-0.938