Decision on target clinical outcome variable (continuous/discrete) or contrast (group membership case/control)
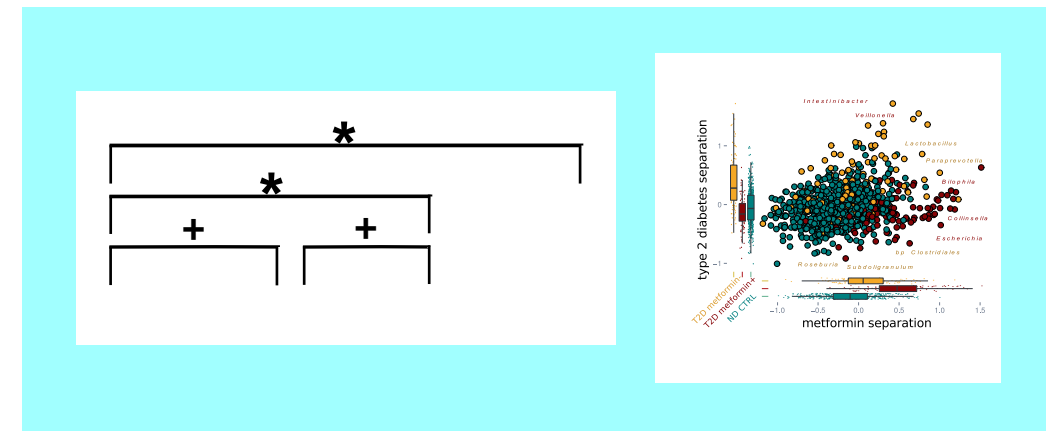
Preparatory (pre-omics) analysis and confounder detection

Clustering/correlation analysis for health/lifestyle/diet metadata patterns/dependencies

MD clusters correlating with clinical outcome are modelled as confounders in subsequent analysis
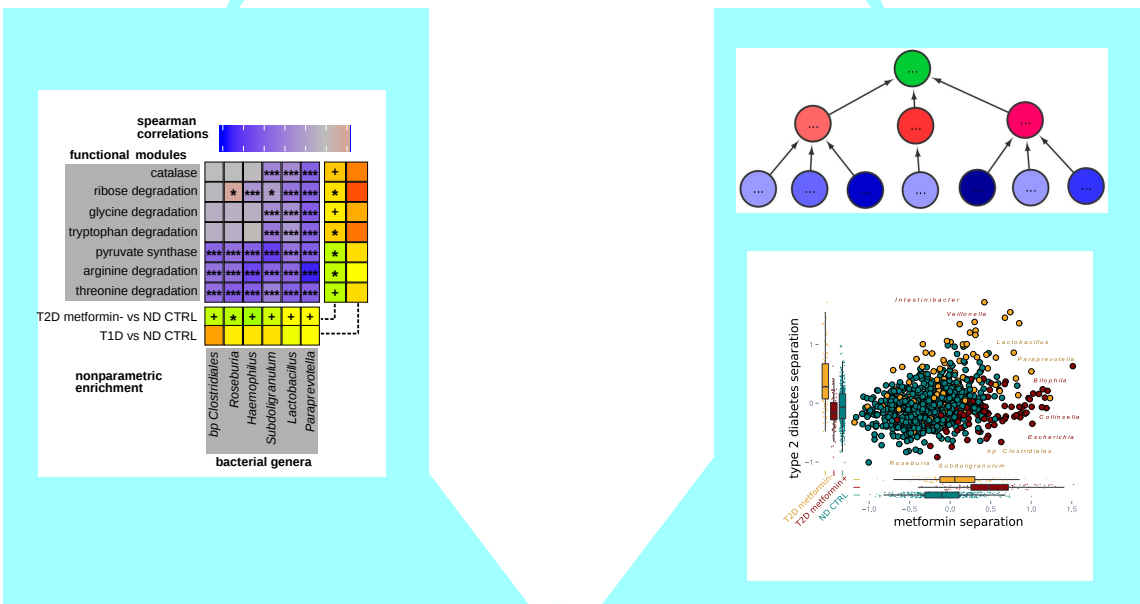
Univariate step

Multivariate step

Omics measurement spaces (mOTUs, MGSs, functional modules, metabolites, gene expressions) correlated against health/lifestyle metadata (with particular interest in findings from Task 5.3)

Confounders from previous steps are modelled in either by cohort subsetting, conditional/blocked tests, or nested model comparisons

Automated feature selection/machine intelligence/feature extraction

Multivariate contrast/correlation testing, ordination, visualizing by clustering

NB: Need to control for confounders from previous step, either by subsetting or by incorporating into methods (blocked ordination, confounder explicitly placed in data-organizing network...)

Data merge - accept as high-reliability those features selected both through univariate and multivariate tests

Omics vs metadata analysis
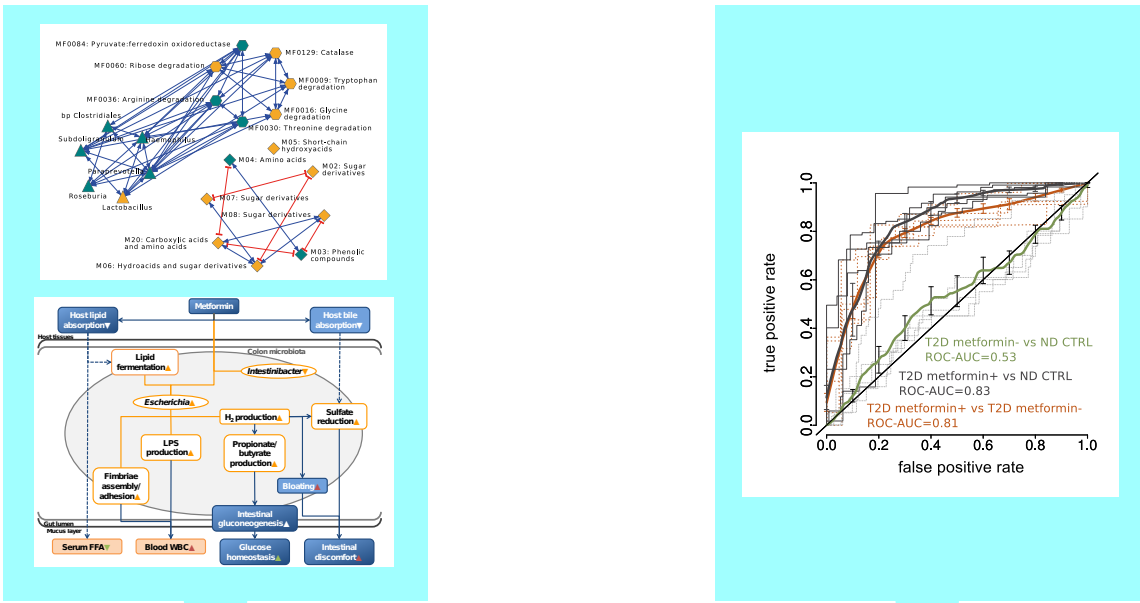
Stage II

Union of features

Elucidation step

Prediction step

Correlation between features that are significant predictors from both univariate and multivariate analysis (high-confidence set)

Linking bacteria to their functional repertoires to health/lifestyle factors, confounders and outcome variables - mechanistic network view of target pathology

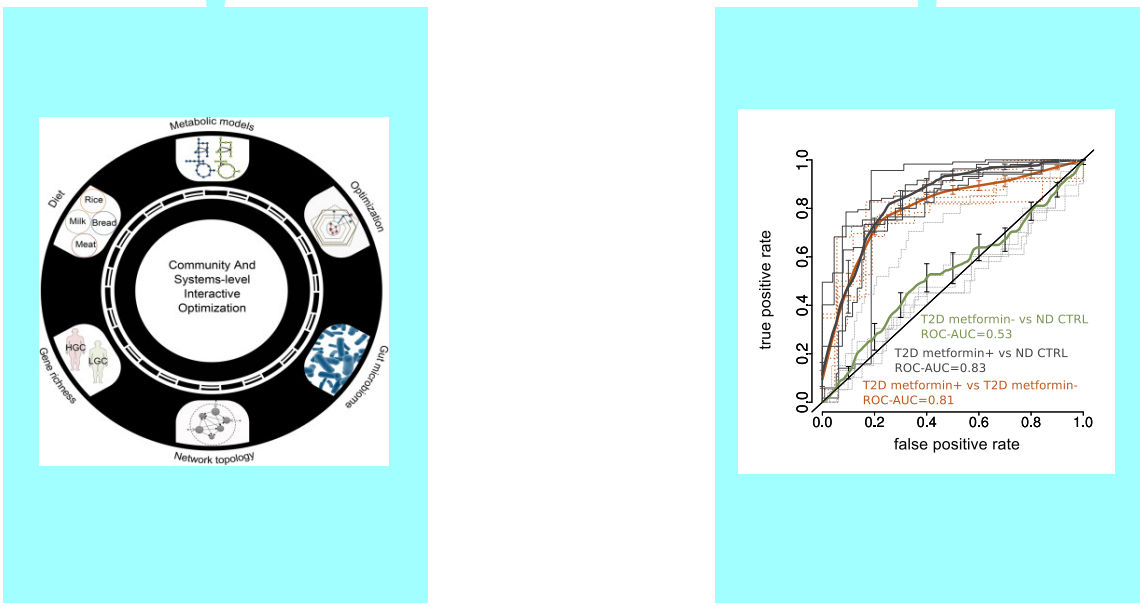Visualization by heatmap, barcode plot, (multipartite) network, egg diagram...

Building predictors from selected features (NBNs, SVMs, other more complex network structures) drawing on features in high-confidence sparse set, and incorporating confounders explicitly

Optimizing using holdout and crossvalidation

Visualization by ROC plots etc.

Omics vs omics vs metadata analysis

Stage III

Reconstruction step

Validation step

Build CASINO GEM ecosystem and test behaviour of the constructed model under perturbations

Design and test experimental interventions

Validate predictors on fully independent cohorts including legacy and holdout sets

Prune and pare down predictors to minimal signature of biomarkers for exploitation

Expanding beyond the cohort

Stage IV

Reporting and exploitation