

## 5. Methodology

### 5.1 Metabolomics of the METACARDIS Study and Serum Drug Identification

METACARDIS was carried out between 2013-2016 in three centres across the EU: France, Germany and Denmark, to evaluate the role of the metabolome and microbiome in the cardiovascular disease spectrum (Figure 2). Work carried out by Forslund et al. (42), Vieira-Silva et al (40) and Fromentin et al. (41) evaluated three aspects of the disease spectrum: the impact of drugs on the metabolome and microbiome in cardiometabolic diseases, how statin intake relates to the microbiome and metabolome in dyslipidaemia and what metabolomic and microbiome features are associated with the development of ischaemic heart disease.

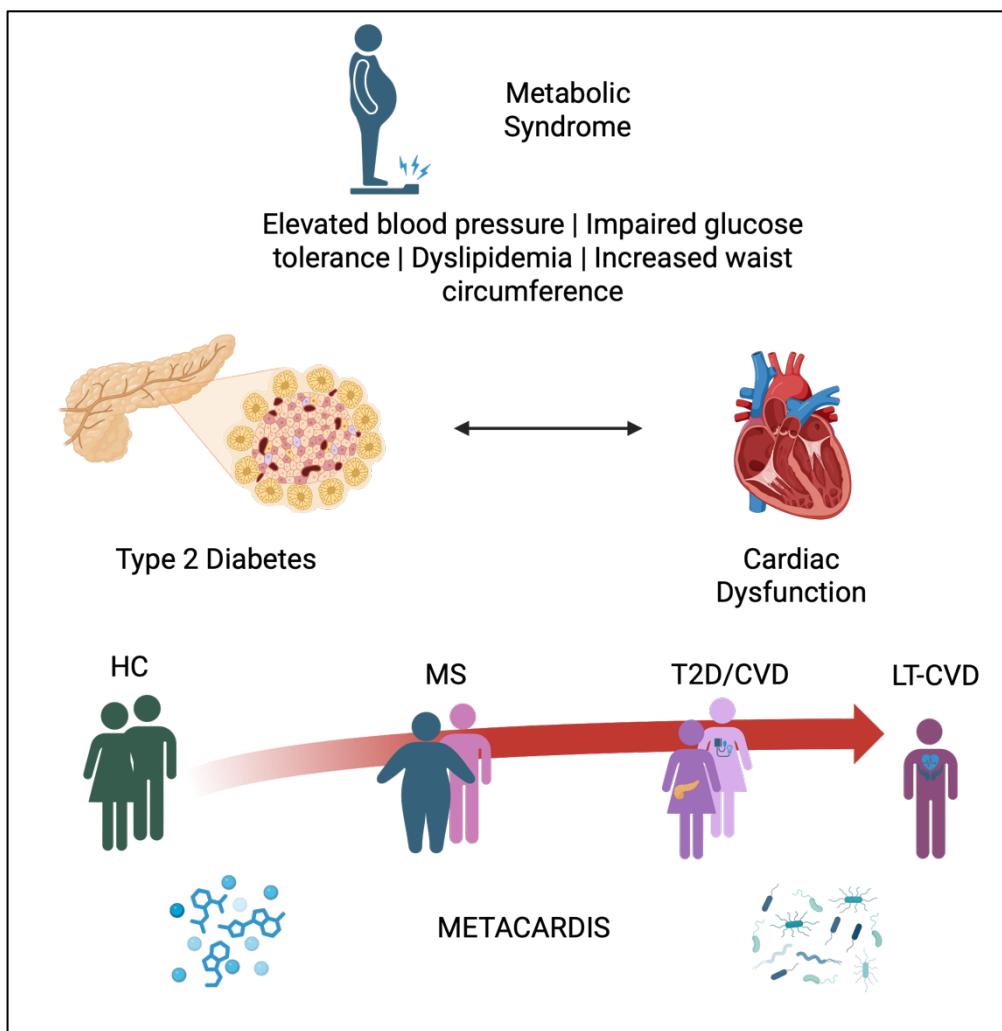


Figure 2: Flowchart linking metabolic syndromes to the METACARDIS study; Abbreviations used: HC: Healthy Control, MS: Metabolic Syndrome, T2D: Type 2 Diabetes, CVD: Cardiovascular Disease, LT-CVD: Long Term Cardiovascular Disease, Figure made in Biorender

Untargeted metabolomics of the metabolon in the cohort were carried out using a UPLC/MS methodology (44) by the Metabolon Inc. Annotated metabolites and unknown features were identified by comparing sample features with previously tested standards and unknowns, followed by quality control and visual inspection (45). The bulk metabolome dataset was normalised and imputed. Out of the bulk metabolome, metformin's (metformin is not metabolised in the body (14)) signal was selected and denoted as serum metformin value for further analyses

## 5.2 Data Selection and Group Level Analysis

To evaluate the microbiome, the following datasets, processed as per protocols defined by Forslund et al. (42) and Fromentin et al. (41), were selected for the study: microbial abundance (taxonomic data) for evaluating the microbiome composition , and KEGG KOs, KEGG modules, KEGG pathway, and gut metabolic modules (GMMs) level data for studying the functional aspects. To evaluate the broad impact of the drug on the microbiome, microbial load index, enterotype (46), carbohydrate, protein and fat degradation potential of the microbiome, microbial richness at both genus and species level and Shannon Diversity were evaluated as well.

From the cohort, individuals with a type 2 diabetes diagnosis excluding cardiovascular disease were evaluated in this study. To evaluate the role of the drug on clinical parameters, liver function parameters [aspartate aminotransferase (AST), alanine aminotransferase (ALAT), gamma-glutamyl transferase (GGT)], kidney function parameters [estimated glomerular filtration rate (eGFR)] and parameters of metabolic health and inflammation [Body Mass Index (BMI), C-reactive protein (CRP)] were evaluated at a group level. Demographics (age, sex, country of recruitment) were also extracted to be used as covariates in further analyses. Only patients with both metabolomic and metagenomic data were evaluated in this study.

The selected individuals were then split into two cohorts: true drug intake and true drug non-intake. True drug intake represented the individuals where there was a non-zero metformin dose reported and a non-zero metformin serum level detected (LC-MS/MS metformin signal). True drug non-intake had zero values for both dose and serum levels. From the true drug non-intake group, a further subset was created, "no drug from class", which included patients who had a zero-dose value for all hypoglycaemics given to patients in the study: GLP-1 agonists, insulin(bolus), PPAR $\gamma$  agonists, sulfonylureas, DPP-IV inhibitors and SGLT-2 inhibitors.

Using all clinical and metagenomic variables, a one-way ANOVA, followed by Tukey's Post Hoc Testing, was performed to evaluate group-level differences. All p-values were FDR adjusted using the Benjamini-Hochberg (BH) method. Significance was defined for p-value (adjusted) < 0.05.

### 5.3 Association Testing with Drug Intake

To evaluate what metabolites and metagenomic variables were associated with metformin intake, the following approach was adopted (summarised in Figure 3):

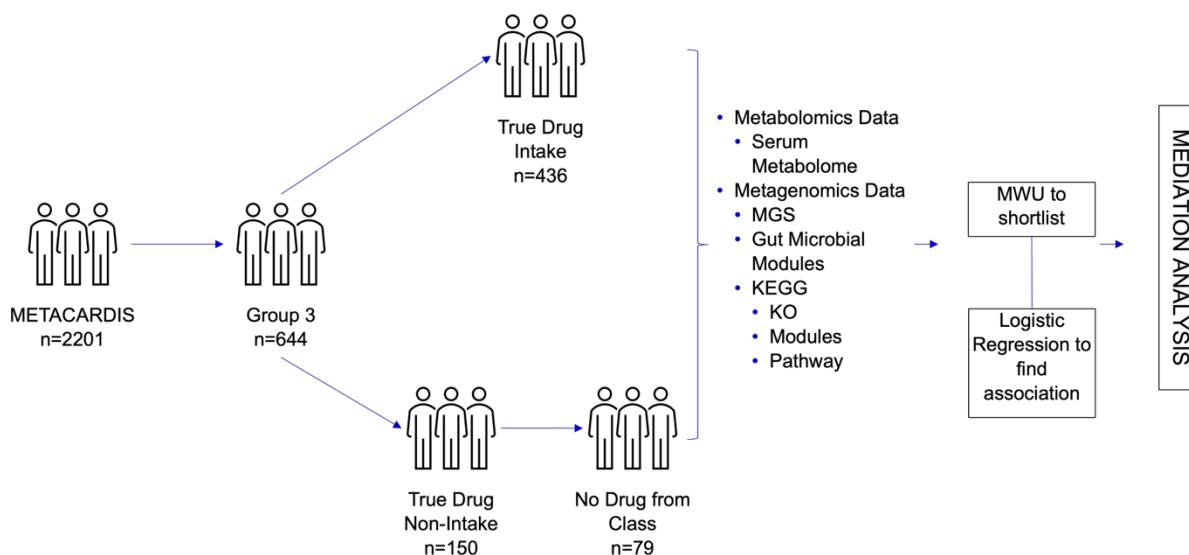
Shortlisting of significant variables was done by the Mann-Whitney U Test ( $p$  (adjusted)  $< 0.05$ ) (adjustment by the BH Method). Shortlisted variables were tested in a logistic regression model to see if a given variable was predictive of group (true drug intake, true drug non-intake, no drug from class) association. The logistic regression model was adjusted for age, sex and country of recruitment. Significance was defined as  $p$  (adjusted)  $< 0.05$  (BH).

The logistic regression formula used was:

$$\text{Group} \sim \text{Variable} + \text{Age} + \text{Sex} + \text{Country of Recruitment}$$

using the `glm` function (family binomial) in the stats package. Results were plotted using the `ggplot2` package (47). The results were plotted post-transformation with  $\log_{10}$  Odds Ratio vs  $-(\log_{10}p)$  being plotted. The threshold for labelling a point on the plot was defined as those belonging to the top 10th percentile on either axis (threshold was dropped to 1st percentile for KEGG KO, due to the density of the graph).

The above pipeline was used for two comparisons: true drug intake vs true drug non-intake and true drug intake vs no drug from class. Variables found significant in both these comparisons were selected for further analyses.



*Figure 3: Workflow (until Mediation Analyses): The METACARDIS cohort was subset to evaluate individuals who were diabetics without cardiovascular disease (Group 3). This group was further subset on the basis of reported dose and serum values for metformin into true drug intake and true drug non-intake. A further sub-setting of true drug non-intake was carried out based on dose of other hypoglycaemics (if zero across all: individual placed in no drug from class). Bulk metabolon and metagenomic data were evaluated using a two-step Mann-Whitney and Logistic Regression (Adjusted for demographics) testing pipeline to find associations with drug intake.*

To reduce dimensionality, a sparse canonical correlational analysis (CCA)-based approach was performed (48). A sparse-CCA was used to select those variables that maximised the correlation between the associated variables and a matrix of dose drug, serum drug (both were selected as the correlation between dose and serum levels was  $< 0.01$  (Pearson Correlation)) and 1,5-anhydroglucitol (AG, correlation between 1,5-AG and HbA1c  $\cong 0.6$ ). The methodology of the sCCA is as follows:

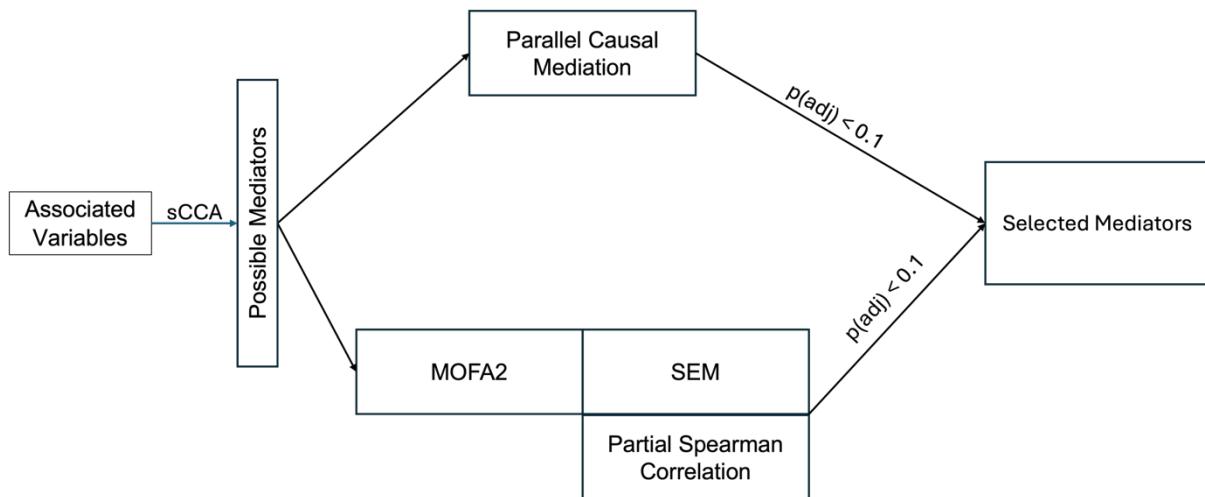
L1 penalty was optimised using a range of penalty values and running a 10-fold cross-validation at each penalty 3 times. The L1 penalty yielding the highest average canonical correlation was selected. The sparse CCA model was run 1000 times in a bootstrap, and the frequency of the mediator being selected was recorded. A threshold of 30% was defined for selection, meaning mediators that appeared in more than 30% of the bootstrap runs were selected. If less than 10% of the input features were above the threshold, the top 25% (top quartile) in terms of selection frequency were selected.

A causal parallel mediation framework using the mediation package (49,50) in R was used for four combinations of output-input: with the inputs being dose metformin and serum metformin, and outputs being 1,5-AG and HbA1c. All mediators were individually tested across all four combinations, and testing was bootstrapped 1000 times for robustness. The mediation p-values were adjusted by using the BH method (for each input-output combination), and p-values (adjusted)  $< 0.1$  were selected as statistically significant mediators.

To account for the fact that causal mediation assumes that mediators do not have any confounding within themselves, which is generally not seen in multiomics datasets, a MOFA-SEM (in lavaan) (51–53) based mediation was done. This approach is a latent space approach (54), and a similar approach has been used with neural networks for response prediction in breast cancer (55). In latent space approaches, multiple mediators are evaluated together using latent vectors (54).

MOFA2 (MultiOmic Factor Analysis) was carried out across all the separated omic blocks. Weights, variance explained, and the model itself were extracted, with model performance parameters reported. The factors obtained were put in structural equation modelling-based mediation, bootstrapped 1000 times, adjusted for demographics (covariates with collinearity  $> 0.8$  were dropped). The p-values were adjusted by using the BH method for each combination, and p-values  $< 0.1$  were considered significant. A partial Spearman correlation (adjusted for demographics) (56) was also carried out for all factors vs 1,5-AG and HbA1c. P-values were adjusted using the BH method, with statistical significance defined for p-values (adjusted)  $< 0.1$ . A factor which was seen as a mediator and was significantly correlated with the respective marker variable was selected. The features making up

the top quartile of absolute weights in the selected factor were selected for further analysis. (Figure 4)



*Figure 4: Mediation Framework: Associated variables were first selected by an sCCA approach, maximising correlation between dose/serum and 1,5-AG. Selected features were tested individually in a 2 x 2 parallel causal mediation framework (Dose/Serum x HbA1c/1,5-AG) and via latent factors constructed via MOFA2 and evaluated via SEM/Partial Correlation simultaneously. Significant mediators ( $p < 0.1$ ) were selected for co-occurrence analyses.*

## 5.4 Co-Occurrence and Metabolome-Microbiome Interaction Analyses

The relationship between the microbiome and its associated metabolites was evaluated by constructing a co-occurrence network, followed by a regression/mediation approach that linked the metabolome and microbiome.

A partial correlation matrix (adjusting for demographics) using Spearman correlation was built (56). A correlation threshold of 0.4 was selected to keep moderate associations. The two matrices used were associated metabolites and associated microbes. Intra-matrix and intermatrix relationships were both built to evaluate microbiome/microbiome, microbiome/metabolome and metabolite/metabolite relationships.

The network was constructed using the correlation matrix using the igraph package(57). Edge length was defined by the inverse of the correlation strength from the previous step (the more strongly correlated nodes are closer). Edge weight was defined by the absolute value of the correlation. Each node(metabolite/microbe) was annotated to its sub-pathway/genus to evaluate relationships in an interpretable way. To identify non-biological groups, Louvain clustering was carried out within the network (58). Identification of nodes of interest was carried out by identifying microbes-metabolite connections. The Louvain cluster, which a metabolite-microbe interaction belonged to, was extracted for evaluation. All microbes found in the clusters were correlated with the dose of metformin using a partial Spearman correlation, adjusting for demographics(56). Microbes found significant were regressed against the KEGG variables found significantly associated with drug

intake (using the lm function, adjusting for demographics) to see what proportion of the KEGG results could be explained by the microbes of interest. As multiple Co-Abundance Gene groups (CAG) IDs can be annotated to the same microbe, a union of explanation was also seen, where if CAGx and CAGy were classified to the same microbe A, the proportion explained by microbe A was given by (CAGx  $\cup$  CAGy).

To evaluate the interplay of the microbiome and the metabolome post-metformin intake, the following regression/mediation pipeline was carried out:

All microbes were regressed linearly and in LASSO (glmnet)(59) against all metabolites (adjusted for demographics), in a cross-validated bootstrapped model (10-fold, 1000 bootstraps), to see univariate and multivariate implications on the metabolon. LASSO is a multivariate linear regression approach that uses L1 penalisation to minimise the coefficients that do not explain the outcome variable, thereby giving a multivariate linear relationship between multiple independent variables and a dependent variable. Microbes significant in the univariate approach and those appearing more than 75% (60) of the time in the LASSO approach were linked back to microbes found correlated with the dose of the drug in the previous step using the distance() function in igraph (47). Those found linked were then put in a mediation using lavaan (51), where the metabolite-linked microbe was taken as the mediator, the dose-linked microbe was taken as the exposure, and the metabolite was taken as the outcome. All models were FDR adjusted (BH), and significant results ( $p(\text{adjusted}) < 0.05$ , pairwise adjustment) were reported. The results were graphed using igraph (57), and these metabolites were then removed from the next set of testing, being defined as microbiome-associated metabolites.

The remaining metabolites were evaluated as host actions of metformin via AMPK( $\beta$ -subunit(PRKAB1)) (61). To validate this, MetaboSignal (62) was used to link PRKAB1 to metabolites, wherein the metabolites were mapped to human pathways on KEGG (63) using the KEGGREST (64) package using KEGG IDs. Molecules which had generic equivalents used in pathways were annotated to their class compound ID (Supplementary Table S3). The shortest path network was built and visualised using ggraph (65).

The targets of any metabolite/metabolite class of interest were predicted using SwissTarget Prediction (66).