



HAL
open science

Multiblock omics data fusion using the Consensus OPLS R package

Céline Bougel, van Du T. Tran, Julien Boccard, Marie Tremblay-Franco,
Florence Mehl

► To cite this version:

Céline Bougel, van Du T. Tran, Julien Boccard, Marie Tremblay-Franco, Florence Mehl. Multiblock omics data fusion using the Consensus OPLS R package. Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM), Jun 2024, Toulouse (FRANCE), France. hal-04623919

HAL Id: hal-04623919

<https://hal.science/hal-04623919>

Submitted on 25 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multiblock omics data fusion using the Consensus OPLS R package



UNIVERSITÉ DE GENÈVE

Toxalim

INRAE

Genotoul Metatoul



Céline Bougel^{1,2}, Van Du Tran³, Julien Boccard⁴, Marie Tremblay-Franco^{1,2}, Florence Mehl³

¹ Toxalim (Research Center in Food Toxicology), Toulouse University, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France ³ Vital-IT Group, SIB Swiss Institute for Bioinformatics, 1015 Lausanne, Switzerland
² MetaboHUB-Metatoul, National Infrastructure of Metabolomics and Fluxomics, Toulouse, France ⁴ School of Pharmaceutical Sciences, University of Geneva, 1211, Geneva 4, Switzerland



Introduction

Omic approaches have proven their value in providing a broad monitoring of biological systems. However, despite the wealth of data generated by modern analytical platforms, the analysis of a single dataset is still limited and insufficient to reveal the full biochemical complexity of biological samples. The fusion of information from several data sources constitutes therefore a relevant approach to assessing biochemical events more comprehensively. However, inherent problems encountered when analysing single tables are amplified with the

generation of multi-block datasets. Finding the relationships between data layers of increasing complexity constitutes a challenging task. Here we propose an extension to the versatile methodology combining the strength of established data analysis strategies, multiblock approaches with Orthogonal Partial Least Squares Discriminant analysis (OPLS-DA) framework [1], to offer an efficient tool as an R package for the fusion of Omics data obtained from multiple sources.

Multi-Omics Data Integration

	ConsensusOPLS [1]	KPCA [2]	DIABLO [3]	MOFA [4]
R Package	ConsensusOPLS	kernlab	mixOmics	MOFA2
Aim	Multi-block and multivariate relationship modeling	Dimensionality reduction	Multi-block and multivariate relationship modeling	Identification of common factors
Approach	Consensus Kernel-based OPLS	Kernel-based PCA	Latent projections for integrating multiple omics datasets	Bayesian model to discover shared latent factors
Model type	Supervised (regression or classification)	Unsupervised	Supervised (classification)	Unsupervised
Non-linearity	Yes in kernel	Yes in kernel	No	No
Parameters	Kernel, Cross-validation type	Kernel	Design matrix	Bayesian inference



Check here MATLAB in-house scripts available at [GitHub Unige](#) with demo_data [5]

R package available in [CRAN](#) New features



Upgrades

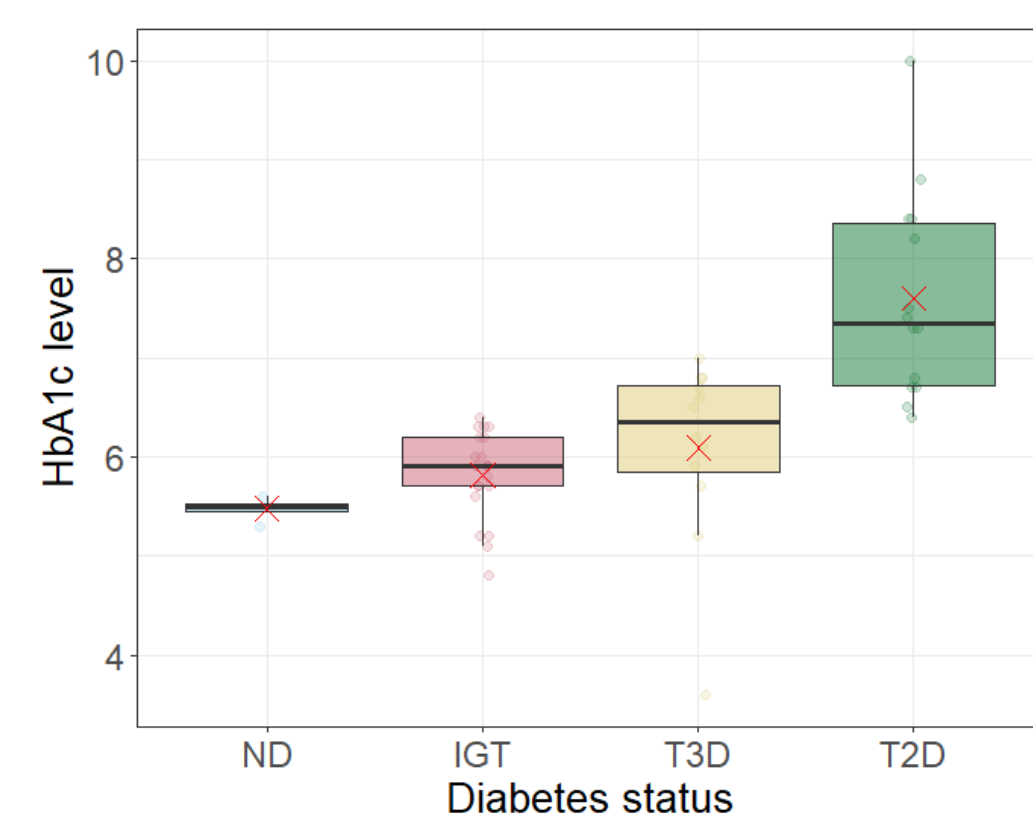
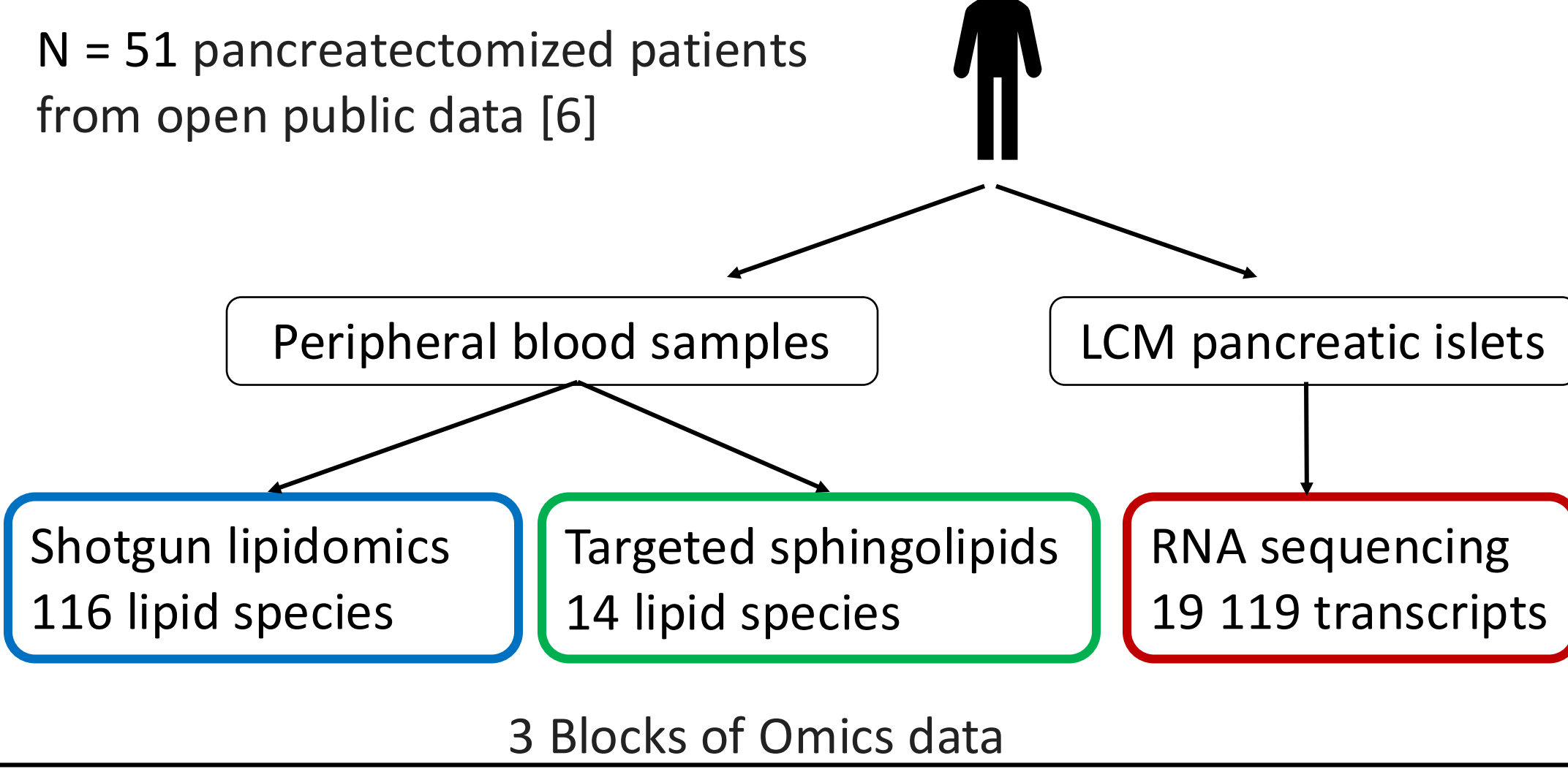
Kernels non-linear
Multiple predictive components

Variable Importance in Projection (VIP)
Prediction of new samples

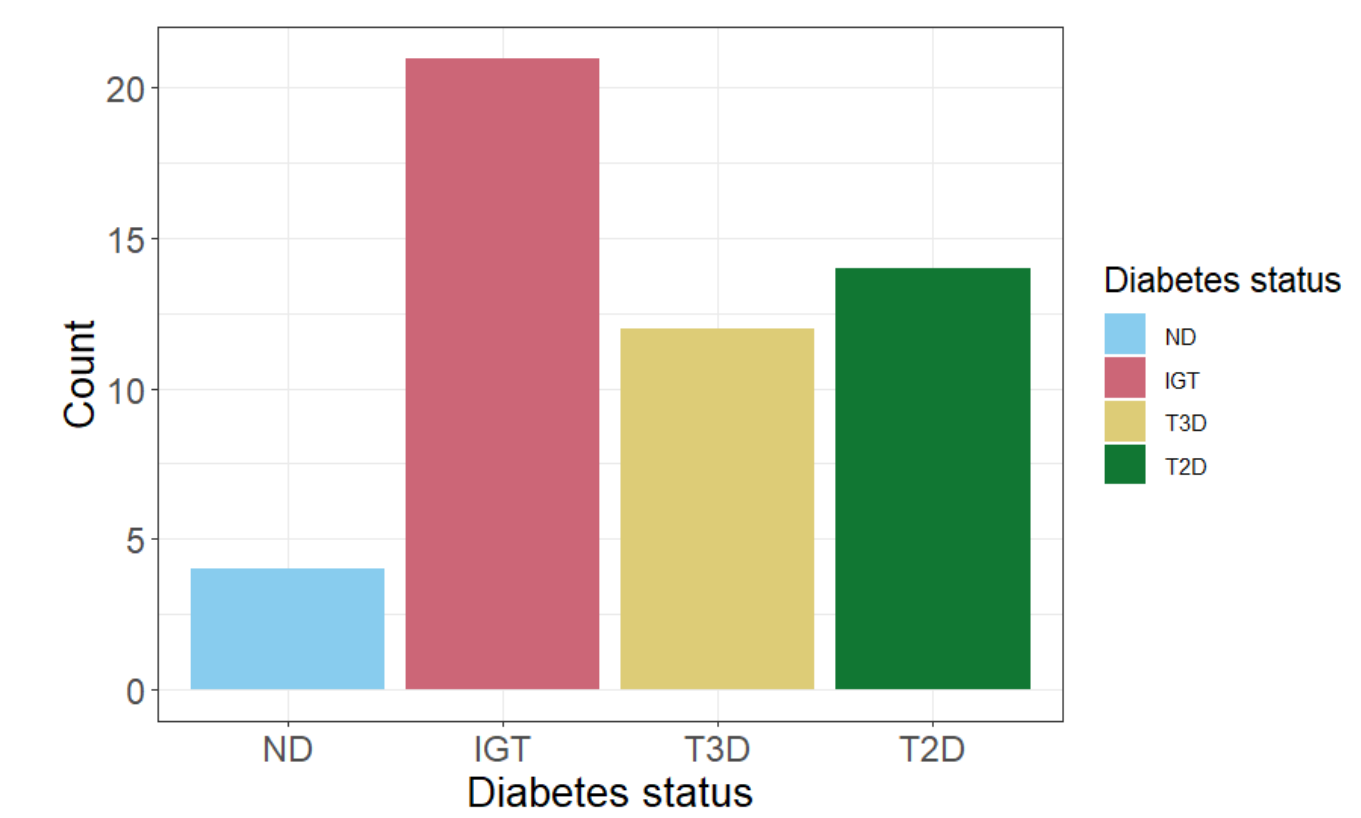
Parallel computing

Use case

4 groups of patients based on glycemic values at fasting and at the 2-hour timepoint of an oral glucose tolerance test, using the thresholds defined in the guidelines of the American Diabetes Association :



Non-diabetic (ND); impaired glucose tolerance (IGT); type 2 diabetes (T2D); type 3c diabetes (T3D).



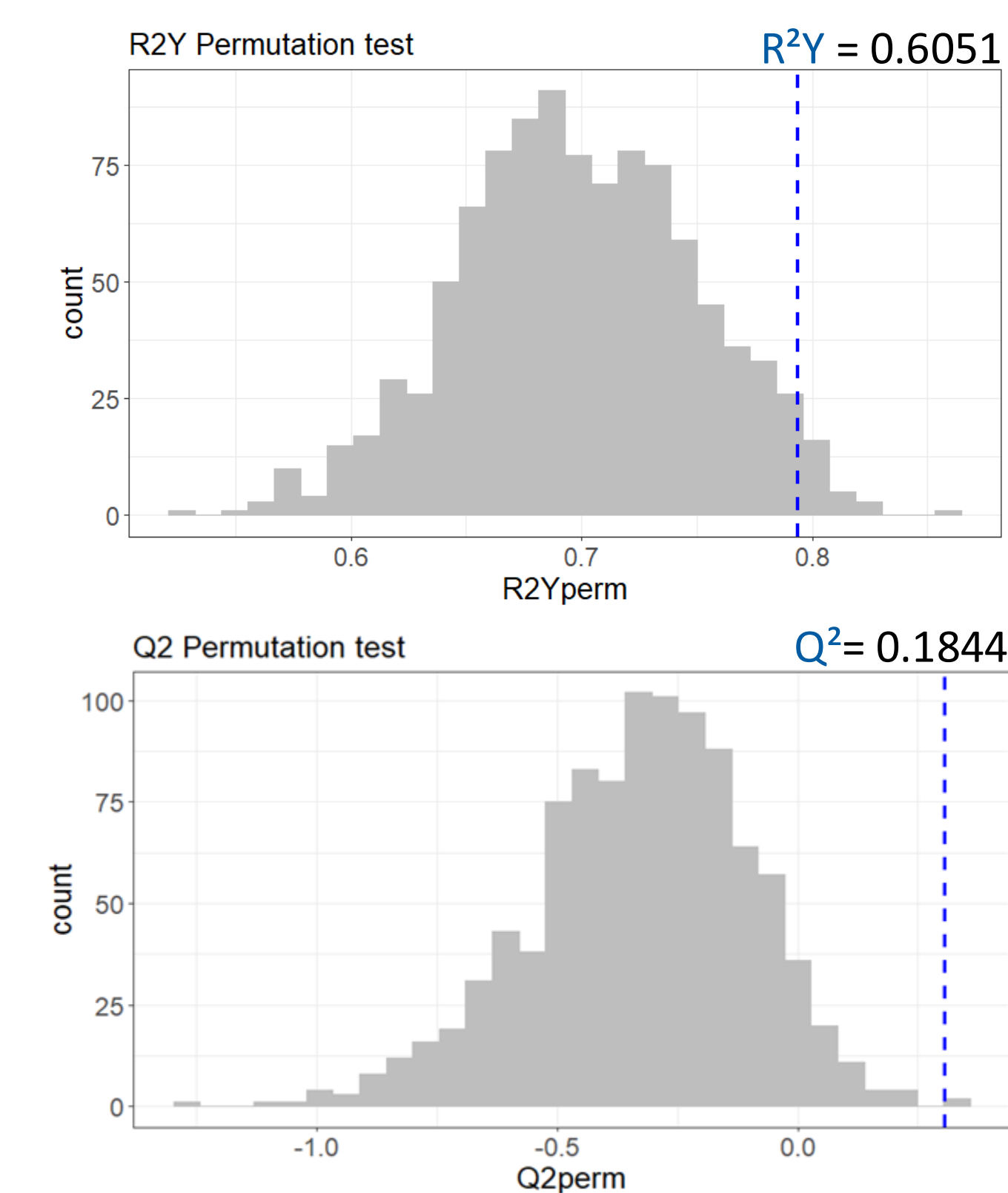
Results and discussion

Model computation

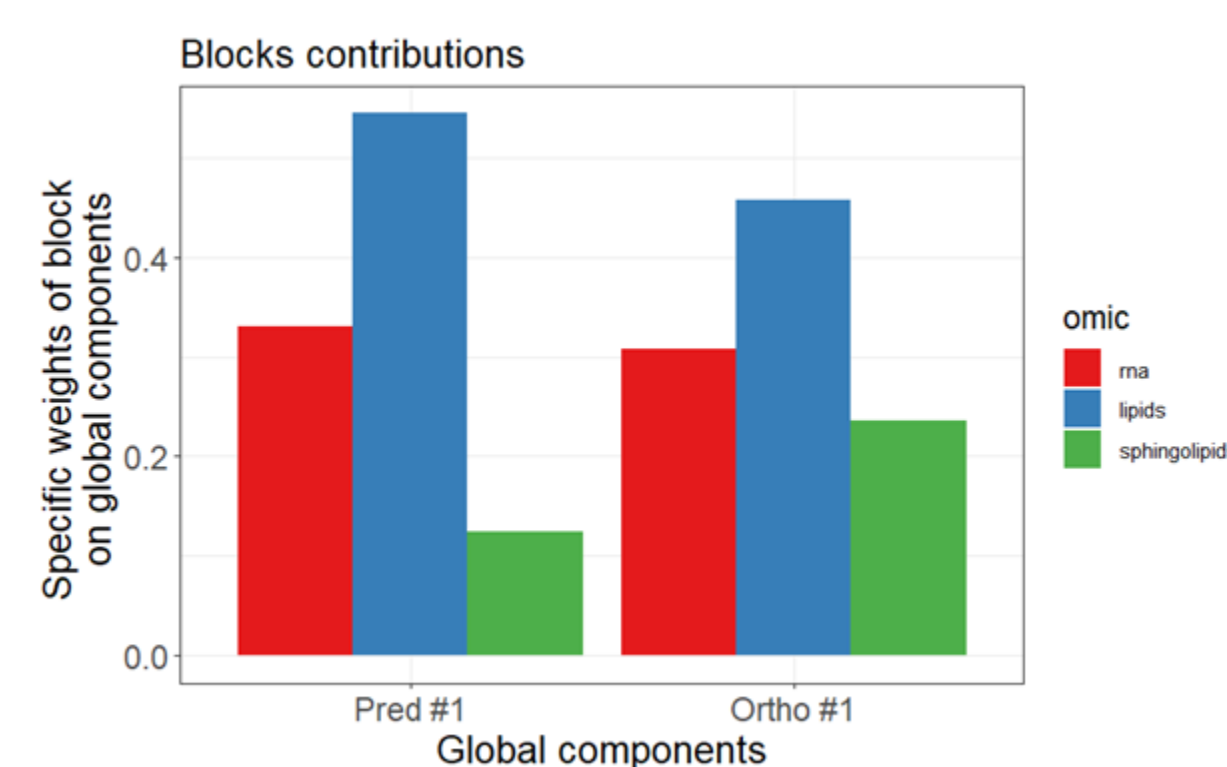
```
COPLS_results <- ConsensusOPLS(
  data = my_data_with_3_blocks_of_omics_data,
  y = response_HbA1c_variable,
  maxComp = 1, #one predictive component
  maxOrthoComp = 1, #one orthogonal component
  modelType = "reg", #regression model
  nperm = 1000, #number of permutations
  cvType = "nfold", #type of cross-validation
  nfold = 5, #number of subjects = leave-one-out
  kernelParams = list(type = "p",
    params = c(order = 1)),
  mc.cores = 1 #how many cores for parallelization
)
```

Quality assessment

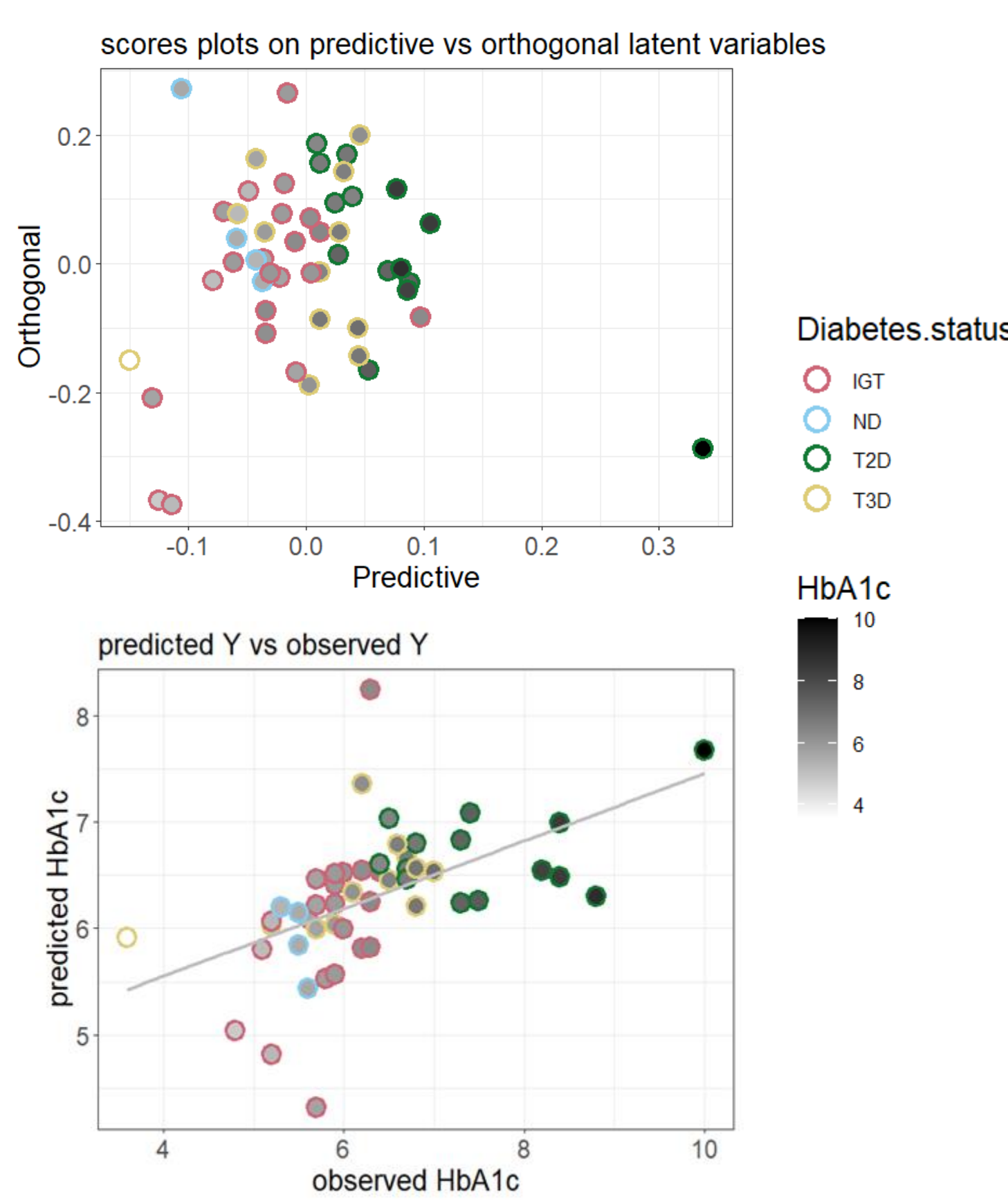
1000 models are computed with permuted Y values. The results of the optimal model are significantly different from the results of the permuted models.



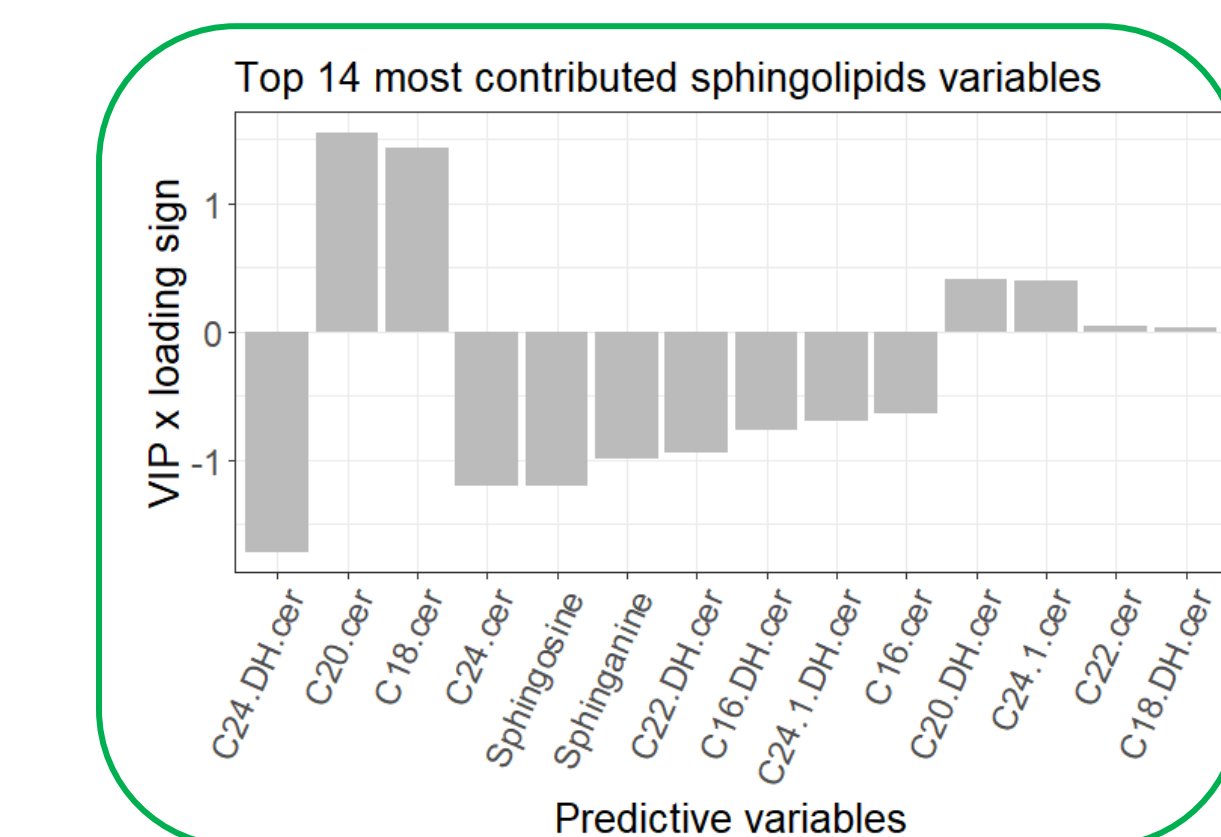
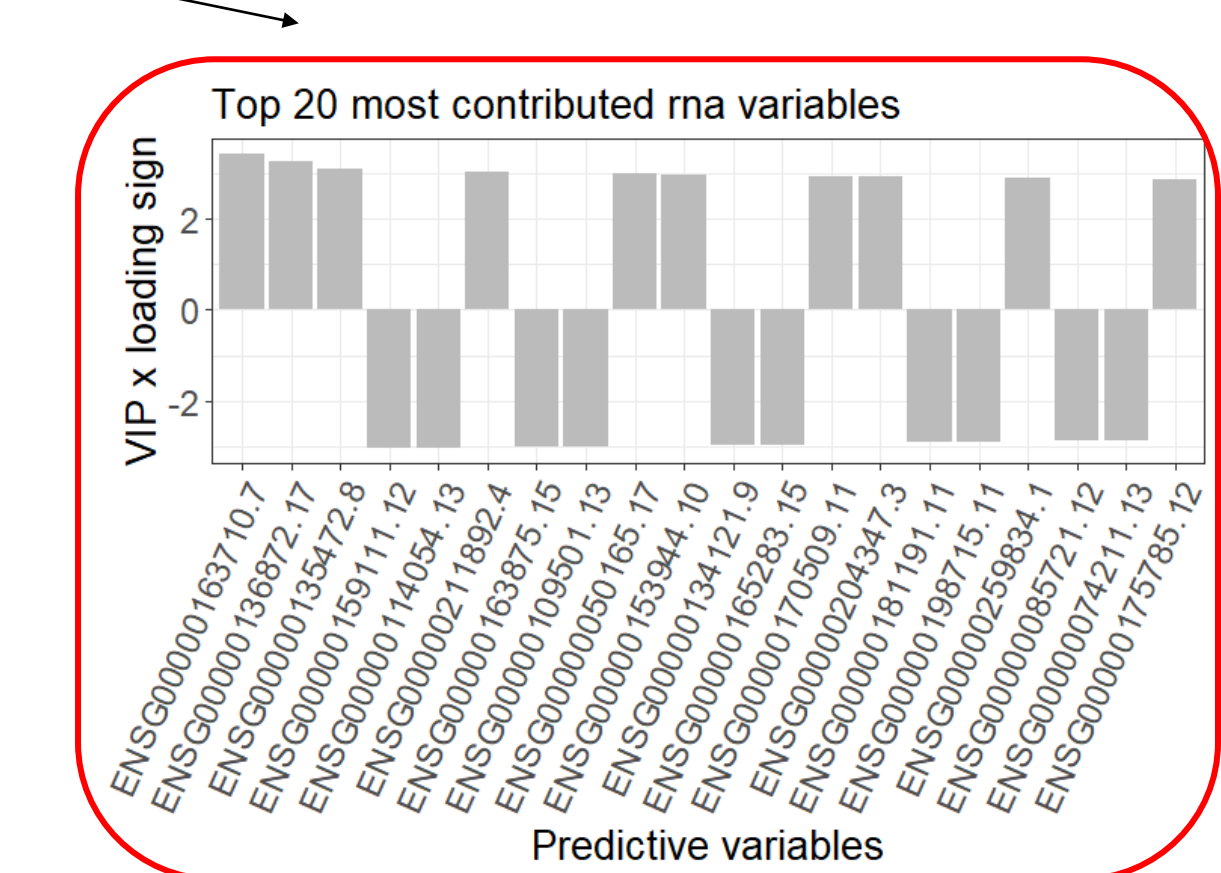
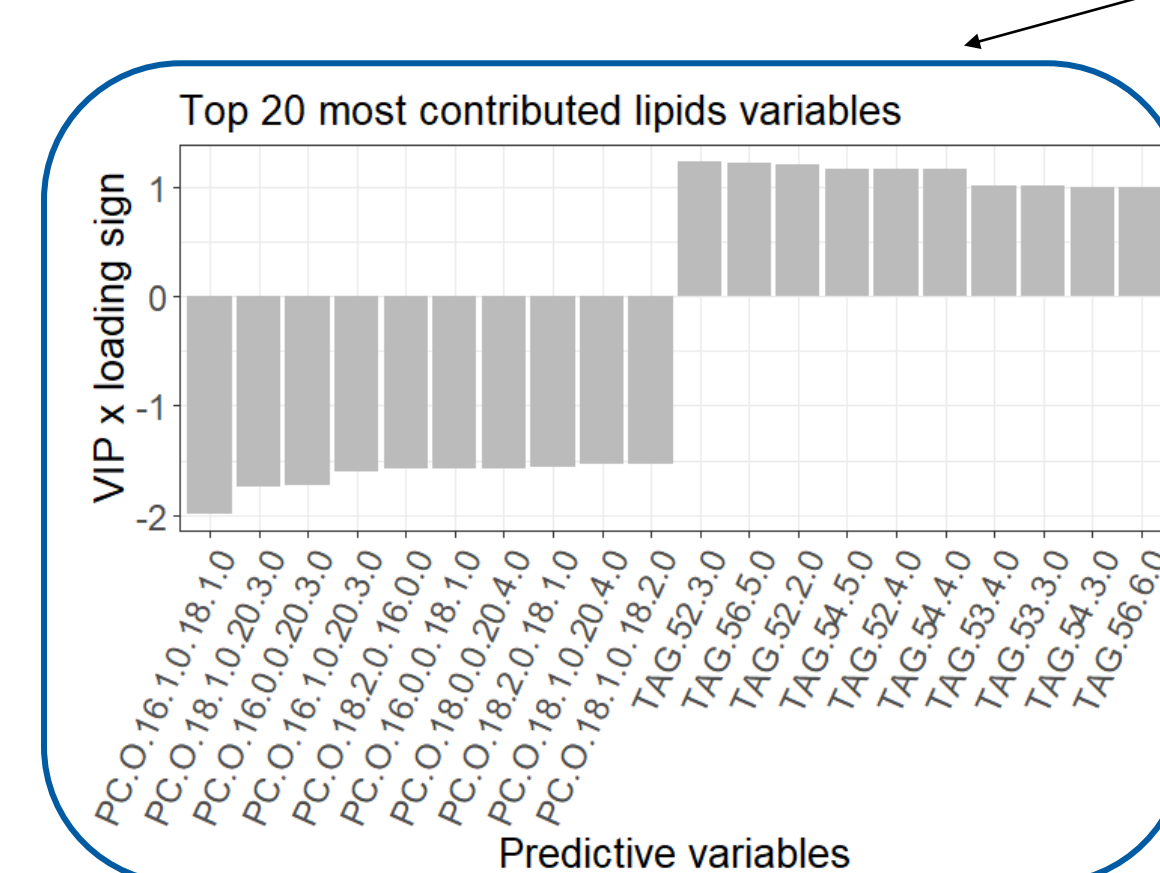
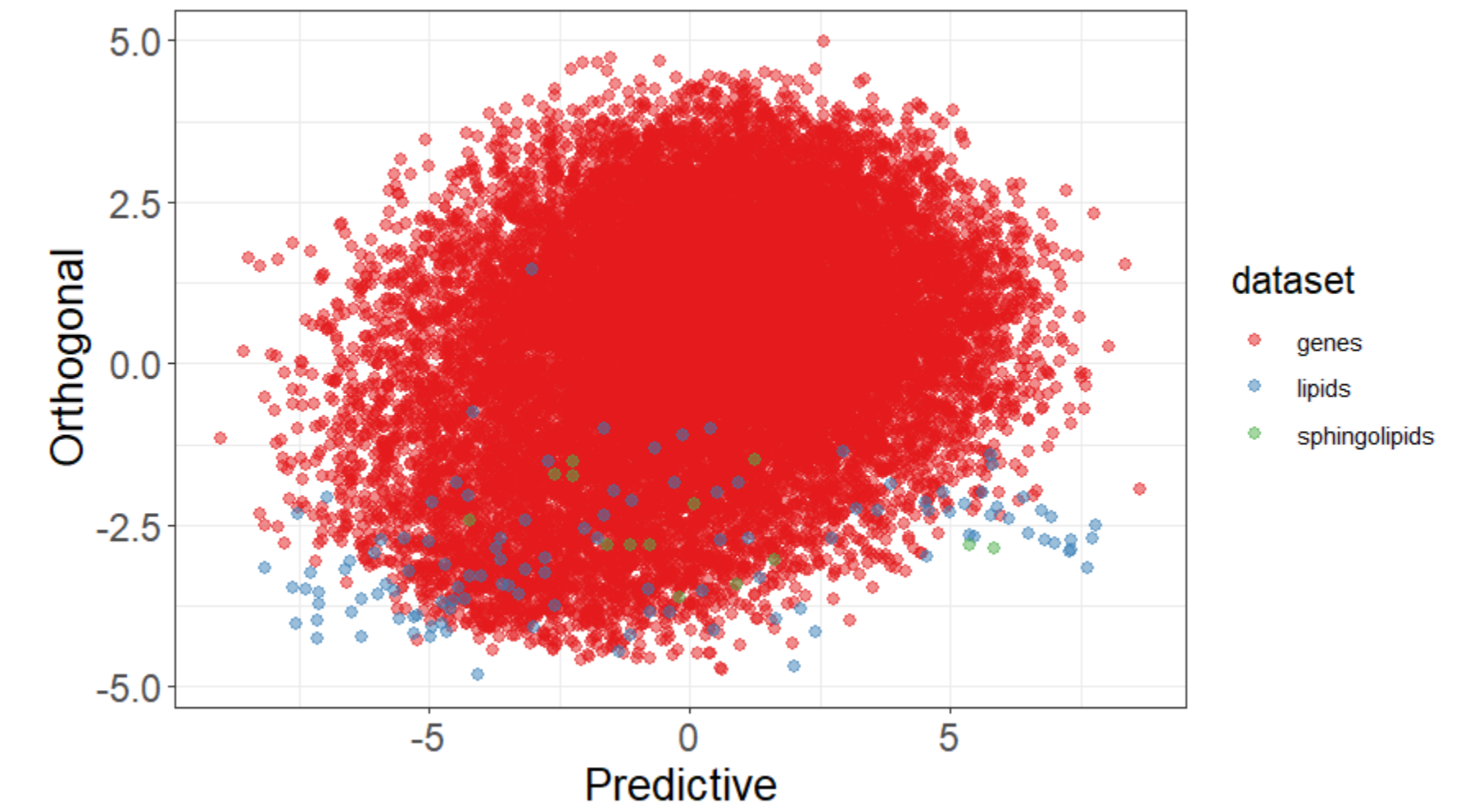
Regression Analysis (ConsensusOPLS-reg) of HbA1c levels



The 3 omics contribute to the modelization of HbA1c levels. The untargeted lipidomics dataset brings the largest part of the information.



Loadings plots on predictive vs orthogonal latent variables



Conclusions

The **ConsensusOPLS R package** proposes a **free, open-source, and easy-to-use software** for the Consensus OPLS method, which was demonstrated as a **relevant and widely applicable method** for the **horizontal integration of omics data**, with multiple functionality upgrades.

References

The **ConsensusOPLS R package** is available at <https://CRAN.R-project.org/package=ConsensusOPLS>.
 [1] Boccard J, Rutledge DN. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Analytica Chimica Acta* [Internet] 2013;769:30–9. Available from: <https://doi.org/10.1016/j.aca.2013.01.022>.
 [2] Schölkopf B, Smola A, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* [Internet] 1998;10:1299–319. Available from: <https://doi.org/10.1162/089976698300017467>.
 [3] Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* [Internet] 2019;35:3055–62. Available from: <https://doi.org/10.1093/bioinformatics/bty1054>.
 [4] Argelaguet R, Velten B, Arno D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* [Internet] 2018;14:e8124. Available from: <https://doi.org/10.15252/msb.20178124>.
 [5] Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* [Internet] 2006;6:813–23. Available from: <https://doi.org/10.1038/nrc1951>.
 [6] Wigger L, Barovic M, Brunner AD, Marzetta F, Schöniger E, Mehl F, et al. Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories towards type 2 diabetes. *Nat Metab* [Internet] 2021;3:1017–31. Available from: <https://doi.org/10.1038/s42255-021-00420-9>.

Acknowledgments

➤ This research was funded by the French National Infrastructure for metabolomics and fluxomics MetaboHUB ANR-11-INBS-0010.
 ➤ Vital-IT Group, SIB Swiss Institute for Bioinformatics.

Contacts

➤ marie.tremblay-franco@inrae.fr
 ➤ florence.mehl@sib.swiss