|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Metabolite ID** | **MSV dataset** | **Title** | **Description** | **Organisms** | **Cosine score** | **Matched Peaks** | **All matches** |
| 4724 | MSV000083446 | GNPS - Antibiotic Mice All Tissue | These are raw and mzXML data from mice tissue obtained from the Raffatellu lab. This excludes the fecal data collected from the time series. | Animalia | 0.93 | 22 | https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=78cd1c95b76b4655a0ce3c0fecfebcea&view=view\_all\_datasets\_matched |
| 4000 | MSV000080655 | GNPS Cross Sectional CF Sputum Samples | A collection of cross sectional sputum samples and some longitudinal from adult patients at the UCSD adult CF center. These samples were extracted in both ethyl actetate/methanol and in 80% ethanol. | *Homo sapiens* (NCBITaxon:9606) | 0.97 | 19 | https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=e0d9f4d16f71467e87ece31fa11c39ab&view=view\_all\_datasets\_matched |
| 184 | MSV000080016 | GNPS - Sugarcane Microbe interaction | LC-MSMS based 3D spatial metabolomics of two sugarcane plants and their interaction with a Xanthomonas isolate. | Sugarcane | 0.85 | 14 | https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=d56022e5ac924fd498ba917a9597df79&view=view\_all\_datasets\_matched |
| 7453 | MSV000080050 | GNPS - Global Stool Dataset - C18-LC-MS/MS Positive polarity | Human and rat stool sample Data was acquired using a Bruker Daltonics maXis Impact and C18 RP-UHPLC. Positive polarity acquisition of LC-MS/MS. | *Homo sapiens* (NCBITaxon:9606);*Rattus norvegicus* (NCBITaxon:10116) | 0.84 | 14 | https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=4bd6c51c36e341cb9d86c2ecf8024739&view=view\_all\_datasets\_matched |

Supplementary Table 3. Further details of MASST matches. Full results including mirrored peaks can be seen via the link.



**Supplementary Figure 1.** Cross validation of DAPC analysis, retaining different numbers of principal components (PCs) to determine the number of PCs resulting in the smallest root mean squared error. Default values for train-test split were used with the *xvalDapc* function in adegenet package implemented in R.



**Supplementary Figure 2.** Cumulative variance explained by principal component analysis.



**Supplementary Figure 3.** Confusion matrix for each of the three machine learning models used showing the proportions of true positives, true negatives, false positives and false negatives.

**Random Forest**



**Gradient Boosting**

**Logistic Regression**

****

**Supplementary Figure 4.** Example explanations from each model for sample 118. These explanations show the proportion that each feature pushes the model from the base value. The base value is determined by averaging the model output over the entire training set. Features pushing the model output higher are shown in red, while those pushing the model output lower are shown in blue. All code can be found at https://github.com/ntobias-85/Rprolixus\_metabolites.



**Supplementary Figure 5.** Metabolite abundance of top nine machine learning features, broken down by gut compartment (AM: anterior midgut, PM: posterior midgut, H: hindgut). Box plots are separated based on annotated metadata (Supplementary Table 1). Box plots show interquartile range, with median, minimum, maximum and outliers shown. All plots were generated with the GNPS Feature Based Molecular Networking dashboard (v0.1).