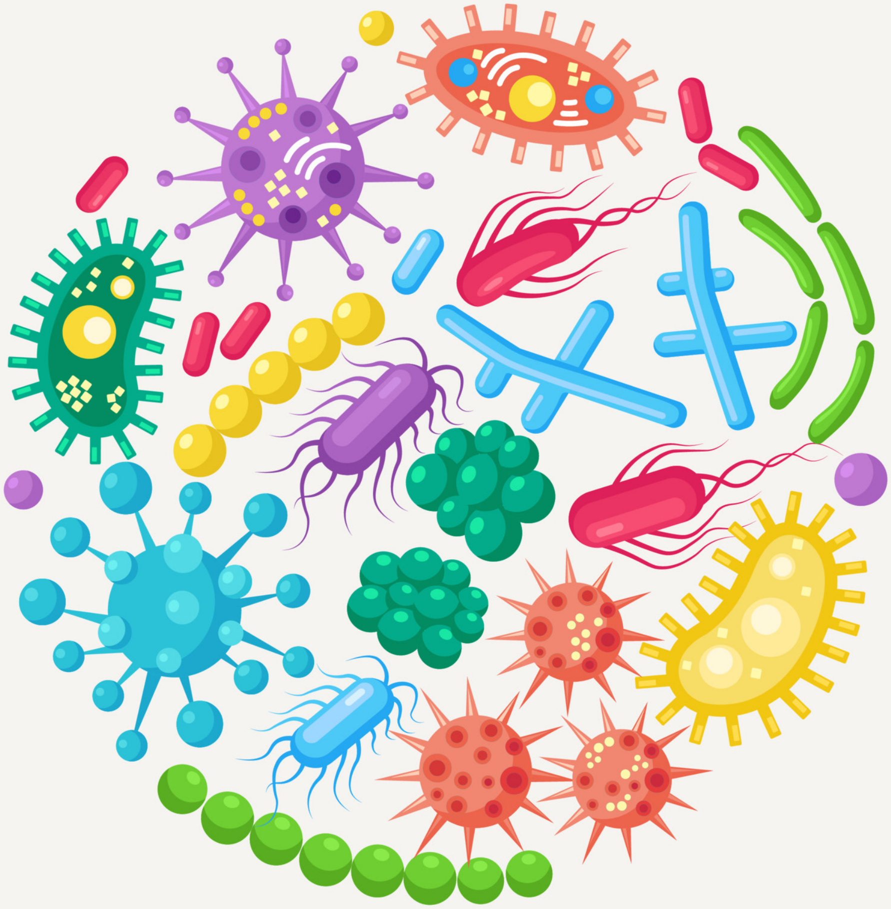
The artificially generated microbiome

A study on the generation and potential use cases of predicted meta-omics data



Bianca-Maria Cosma

TU Delft

The artificially generated microbiome

A study on the generation and potential use cases of predicted meta-omics data

by Bianca-Maria Cosma

To obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on Monday, 24th of June, 2024, at 09:00 AM.

Student number: 5019710
Project duration: November 13th, 2023 – June 24th, 2024

Thesis committee:
Dr. T.E.P.M.F. Abeel, TU Delft, Supervisor, Committee Chair
Dr. M.A. Migut, TU Delft
S. Pillay, TU Delft, Daily supervisor
Dr. D. Calderón-Franco, Hologenomix Life Sciences, Daily supervisor

Cover image credits: buravleva_stock on Vecteezy

Acknowledgements

I would like to thank Thomas Abeel, who was not only a great mentor throughout my academic journey, but also a very patient and considerate supervisor. My sincere gratitude also goes towards my daily supervisors, Stephanie Pillay and David Calderón-Franco, who offered me valuable advice and insight, and encouraged me to persevere in my work. I am very lucky that I had the chance to learn from such wonderful scientists.

Abstract

Motivation: Imbalances in the human gut microbiome have been linked to various conditions, including inflammatory bowel disease (IBD), diabetes, and mental health disorders. While metagenomics and amplicon sequencing are the most commonly used technologies to characterize microbial communities, they do not capture all layers of functional activity of the microbiome. Unfortunately, data from other meta-omics modalities is generally difficult to obtain, due to high costs and error-prone technologies, among other issues. The growing availability of paired meta-omics data offers an opportunity to develop machine learning models that can infer connections between metagenomics data and other forms of meta-omics data. The aim is to enable the prediction of these other forms of meta-omics data from metagenomics data. To that end, we evaluated several machine learning model architectures on the task of predicting meta-omics features from various meta-omics inputs, and analyzed the robustness of these models, as well as potential use cases of artificially generated microbiome data.

Results: Machine learning models, in particular simpler architectures such as elastic net regression models and random forests, generated reliable predictions of transcript and metabolite abundances, with correlations of up to 0.77 and 0.74, respectively, but predicting protein profiles proved more difficult, with correlations of at most 0.42. We also identified a core set of well-predicted features for each meta-omics output type, and showed that multi-output regression neural networks performed similarly when trained using fewer output features. Lastly, our experiments demonstrated that predicted features can be used for the downstream task of inflammatory bowel disease prediction. For instance, accuracy obtained using predicted metabolite abundances was 77%, compared to the 80% accuracy achieved using real metabolomics data.

Code availability: The scripts used for model design, training and benchmarking are available on GitHub.

Contents

1. Introduction

The human microbiome directly and indirectly engages with various physiological subsystems, including the nervous, gastrointestinal, cardiovascular, and immune systems. Research has proven that imbalances in the human microbiome, commonly referred to as dysbiosis, are associated with the onset and progression of various health conditions. For instance, the composition of the gut microbiome, along with its associated metabolites, was found to be significantly different between healthy individuals and those suffering from inflammatory bowel disease (IBD), type I and II diabetes, cardiovascular disease, as well as mental health disorders such as depression and anxiety [1]–[3]. Dysbiosis in the vaginal microbiome has been linked to cervical cancer, as it can affect the development and advancement of HPV (human papillomavirus) infection [4]. Beyond the human microbiome, recent literature also highlights the importance of investigating microbial communities in non-clinical sectors, with applications ranging from surveillance of antibiotic resistance genes [5] to the study of greenhouse gases [6].

Microbial communities can be characterized across various layers of functional activity, with state-of-the-art meta-omics technologies such as metagenomics, metatranscriptomics, metaproteomics, and metabolomics. To obtain a complete picture of the microbiome, each sample should be characterized using all of these meta-omics modalities. Although the metagenome encodes the functional potential of a microbial community, the presence of genes is not synonymous with that of messenger RNA (mRNA), and the latter is not always translated into active proteins. Additionally, even though some associations between microbes and metabolites are known, these can be ambiguous and inconclusive [7], due to the fact that two microbes may produce the same metabolite, or that some metabolites may only be produced under certain conditions. The importance of data accessibility across diverse meta-omics layers is not only emphasized in experimental research [8]–[11], but also in the development of machine learning models capable of performing a wide range of predictive tasks, including disease detection [12], [13].

However, measuring meta-omics data comes with many challenges, including significant costs and reliability issues [14]. DNA sequencing data, whether in the form of amplicon or shotgun metagenome sequencing, is currently the most accessible option, due to its lower cost and the higher reliability provided by next-generation sequencing. At the same time, paired multi-meta-omics data is becoming increasingly available, through initiatives such as the Integrative Human Microbiome Project [15]. This availability of microbiome data across multiple meta-omics layers presents an opportunity to develop machine learning models capable of inferring connections between metagenomics data and other forms of meta-omics data, with the eventual goal of predicting the latter from the former.

Several studies have already described the use of machine learning to predict metabolite abundances in microbiome samples, starting with features derived from metagenomics data. Some model architectures that have been proposed include MelonnPan (elastic net regression) [16], SparseNED (sparse neural encoder-decoder) [17], MiMeNet (multilayer perceptron) [18], mNODE (neural ordinary differential equations) [19], MMINP (two-way orthogonal partial least squares (O2-PLS)) [20] and LOCATE (neural network) [21]. However, the prediction of other meta-omics modalities, in addition to metabolomics, has not been investigated.

Consequently, this study addresses three main research questions:

1. To what extent can model architectures such as the ones previously mentioned be applied to the prediction of other meta-omics data, in addition to metabolomics? We benchmark these cross-omics prediction models on several input-output combinations of single- and multi-omics data modalities. Due to the nature of compositional microbiome data, which is shared across meta-omics modalities, we hypothesize that generalization should be achievable.

2. How can predictions of meta-omics features be improved? We hypothesize that it is possible to optimize model training for a small set of features (transcripts, proteins, or metabolites).

3. What is a potential use case of predicted meta-omics data? Our hypothesis is that predicted data can be useful in phenotype classification. We evaluate models on the task of IBD classification using predicted features.

This thesis is structured as follows. In Section 2, we begin with a brief overview of the biological context of this problem, outlining the main characteristics of meta-omics analysis. We describe our experimental pipeline in detail in Section 3, and present and discuss our results in Section 4. This report ends with a conclusion (Section 5), a short glossary of terms and a list of acronyms.

2. Meta-omics and the microbiome

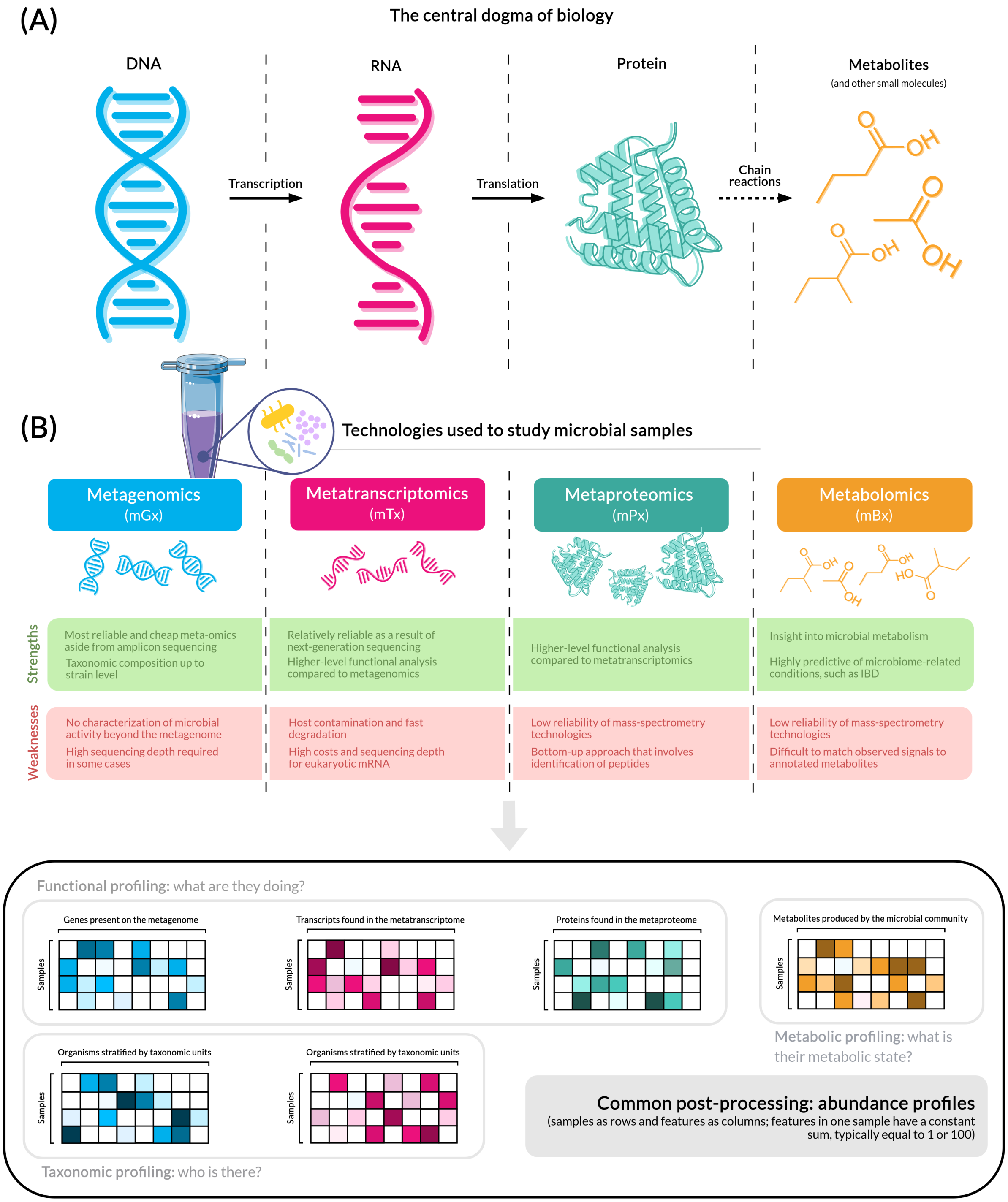

(A)
The central dogma of biology
DNA
RNA
Protein
Metabolites
(and other small molecules)
Transcription
Translation
Chain reactions
OH
HO
(B)
Technologies used to study microbial samples
Metagenomics
(mGx)
Metatranscriptomics
(mTx)
Metaproteomics
(mPx)
Metabolomics
(mBx)
Strengths
Most reliable and cheap meta-omics aside from amplicon sequencing
Taxonomic composition up to strain level
Relatively reliable as a result of next-generation sequencing
Higher-level functional analysis compared to metagenomics
Higher-level functional analysis compared to metatranscriptomics
Insight into microbial metabolism
Highly predictive of microbiome-related conditions, such as IBD
Weaknesses
No characterization of microbial activity beyond the metagenome
High sequencing depth required in some cases
Host contamination and fast degradation
High costs and sequencing depth for eukaryotic mRNA
Low reliability of mass-spectrometry technologies
Bottom-up approach that involves identification of peptides
Low reliability of mass-spectrometry technologies
Difficult to match observed signals to annotated metabolites
Functional profiling: what are they doing?
Genes present on the metagenome
Transcripts found in the metatranscriptome
Proteins found in the metaproteome
Metabolites produced by the microbial community
Samples
Metabolic profiling: what is their metabolic state?
Organisms stratified by taxonomic units
Organisms stratified by taxonomic units
Taxonomic profiling: who is there?
Common post-processing: abundance profiles
(samples as rows and features as columns; features in one sample have a constant sum, typically equal to 1 or 100)

Figure 1: Visual summary of the four main types of meta-omics technologies used to study the microbiome. **(A)** The central dogma of biology, modified to include metabolomics. **(B)** Characterization of meta-omics technologies in parallel to the central dogma of biology, including their strengths and drawbacks, as well as common post-processing methods used for downstream analysis.
The following images were modified and integrated in this figure: DNA and RNA icons (original on Vecteezy), protein icon (original on Wikimedia Commons).

The central dogma of biology (Figure 1A), introduced by Francis Crick during his famous lecture of September 1957 [22], states that genetic information encoded by the genes on our DNA is first transcribed into messenger RNA, which ultimately gets translated into protein. For the purposes of this study, this diagram can be slightly modified to also account for the existence of small molecules, such as metabolites. Upon synthesis, proteins may act as enzymes or structural components that drive metabolic reactions, therefore leading to the production of metabolites.

This simplified view of molecular biology makes up the foundation for our understanding of the microbiome, mainly driven by four meta-omics technologies: metagenomics, metatranscriptomics, metaproteomics, and metabolomics (Figure 1B). In this section, we give a brief account of these four technologies, their strengths and weaknesses, as well as some post-processing data formats relevant to this thesis. For an even more condensed explanation of these biological concepts, please consult the Glossary at the end of this thesis.

2.1. Amplicon sequencing and metagenomics

Sequencing techniques for microbial DNA belong to one of two broad categories: amplicon sequencing and whole metagenome sequencing (WMS), also known as shotgun metagenomics. Amplicon sequencing involves a polymerase chain reaction (PCR) that amplifies the production of selected marker genes within a microbial community. Commonly used marker genes include the 16S ribosomal DNA (rDNA) for prokaryotes, and 18S rDNA or internal transcribed spacers (ITS) for eukaryotic genomes [23]. Such marker genes are chosen because they are highly conserved between bacterial species, while still retaining a degree of variation that enables the identification of bacterial taxa [24]. To that end, the sequenced DNA fragments are queried against a reference database, generating the taxonomic profile of a microbial community.

The advantage of WMS over amplicon sequencing is that it generates sequencing data from the full bacterial genomes present in a sample, rather than just isolated genes. This allows for more fine-grained taxonomic profiling, at the species- or strain-level [23], [25], and it can also be used to determine the functional potential of the microbial community, by mapping reads to references that are functionally annotated, or through ORF (open-reading frame) mining [26]. It is also possible to assemble the genomes of the organisms present in a sample, either through reference-based methods or with *de novo* metagenomic assemblers [5], [27]. However, due to their higher cost and relatively time-consuming analysis [23], [28], WMS technologies are not always the better alternative to amplicon sequencing. As a consequence of the higher volume of data, in addition to the challenges posed by host DNA contamination, shotgun metagenomics requires greater sequencing depth in order to capture the presence of less abundant taxa [29].

2.2. Metatranscriptomics

A shortcoming of standalone metagenomics analysis is that genes encoded by microbial DNA are not always expressed, or may originate from dead and inactive organisms within the microbiome [8]. The microbial metatranscriptome describes the genes from the metagenome that are actively transcribed, under varying conditions, providing insight beyond functional potential and advancing our understanding of host-microbiome interactions. In addition, while several studies confirm that the microbiome metagenome varies significantly between individuals, metatranscriptomic data may also be useful in longitudinal studies of disease progression, because it highlights higher levels of intraindividual variation compared to its metagenomic counterpart [30].

State-of-the-art metatranscriptomics research is now dominated by RNA sequencing (RNA-seq) technologies. Similar to metagenomic samples, RNA-seq is usually performed through NGS (next-generation sequencing) technologies [14]. These approaches involve the isolation of messenger RNA (mRNA) during artificially induced reverse transcription, also known as complementary DNA (cDNA) synthesis. The sequenced cDNA reads can then be used for downstream analysis, which often includes alignments to reference genomes to identify the active functional profile of a community.

Selection of mRNA is more easily achievable for eukaryotes than for prokaryotes, due to the presence of a poly-A tail on eukaryotic mRNA [8]. As a result, it is challenging to separate microbial mRNA from other, non-protein-coding RNA types, and methods that have been devised for this purpose require much greater sequencing depth and are therefore not cost-effective [31]. This is only further complicated by the high amounts of host contamination present in many human microbiome samples [32]–[34], as well as the difficulties involved in preserving microbial transcriptome samples, which tend to suffer from fast degradation [9].

2.3. Metaproteomics

Because the presence of mRNA does not always imply the presence of translated active protein, a quantification of the metaproteome is necessary to provide a more fine-grained view of the functional activity in a microbial community [8]. Metaproteomics can be used to construct a taxonomic profile of the active members of a microbial community, and it also enables the identification of interaction networks in the microbiome [35]. As a consequence, several studies rely on metaproteomics data to find relevant pathways that characterize specific microbial systems, in disease progression [36], [37], but also beyond the human microbiome, for instance, in aerobic communities where it is important to assess the involvement of bacterial communities in the production of harmful greenhouse gases [6].

Unlike metagenomics and metatranscriptomics analysis, which is made more accessible by recent developments in next-generation sequencing, the presence, quantification, and annotation of multiple proteins in a sample are generally studied using a bottom-up approach, because direct sequencing of intact proteins in a microbiome setting is not yet technically feasible [35]. This entails that proteins are first digested into peptides (short protein fragments), which are then separated using liquid chromatography (LC). Lastly, peptide masses are analyzed through high-resolution mass spectrometry (MS), producing experimental mass spectra that are then matched against theoretical mass spectra from a protein sequence database, such as UniProt [11]. Among current challenges in metaproteomics, we point out the lack of standardized practices across research facilities [35], difficulties in proper sample preservation to avoid protein degradation [11], reduced protein yields, MS signal mapping, noise removal and database gaps, particularly for the identification of peptides that originate from homologous protein sequences in different organisms [14].

2.4. Metabolomics

As a complementary approach to the previously described meta-omics techniques, metabolomics does not directly measure the (active) composition of the microbiome, focusing instead on the products of its metabolism. In the human gut microbiome, the presence and quantity of various microbial metabolites have been linked to chronic diseases [1]. It has also been shown that the microbial metabolome is closely linked to the host's metabolism and immune system, and some host metabolites are only produced when certain gut microbes are present [38]. Metabolomics can therefore reveal the pathways involved in such correlations, for instance through the analysis of metabolic networks [39], especially as the conservation of metabolic pathways across species opens up many possibilities for research on microbial communities beyond the human body [40].

The metabolome can be studied in a global setting, through the use of untargeted approaches, but individual metabolites can also be measured, as in the case of targeted metabolomics. The latter approach is generally preferred when the goal is to isolate metabolites known to be associated with specific conditions [7]. As in the case of metaproteomics, the metabolites produced by the microbiome can be identified through the use of mass spectrometry, although another common approach is nuclear magnetic resonance spectroscopy (NMR) [14]. MS-based metabolomics shares many of the shortcomings of metaproteomics, including the difficulties involved in matching observed signals of metabolic mass spectra to annotated metabolites, and post-processing steps such as data denoising [41]. Another active area of research in metabolomics is finding associations between microbes and metabolites. Even in the presence of other omics data, it remains challenging to determine which microbes produce specific metabolites, particularly since it is possible for two microorganisms to produce the same metabolite [7].

3. Materials and methods

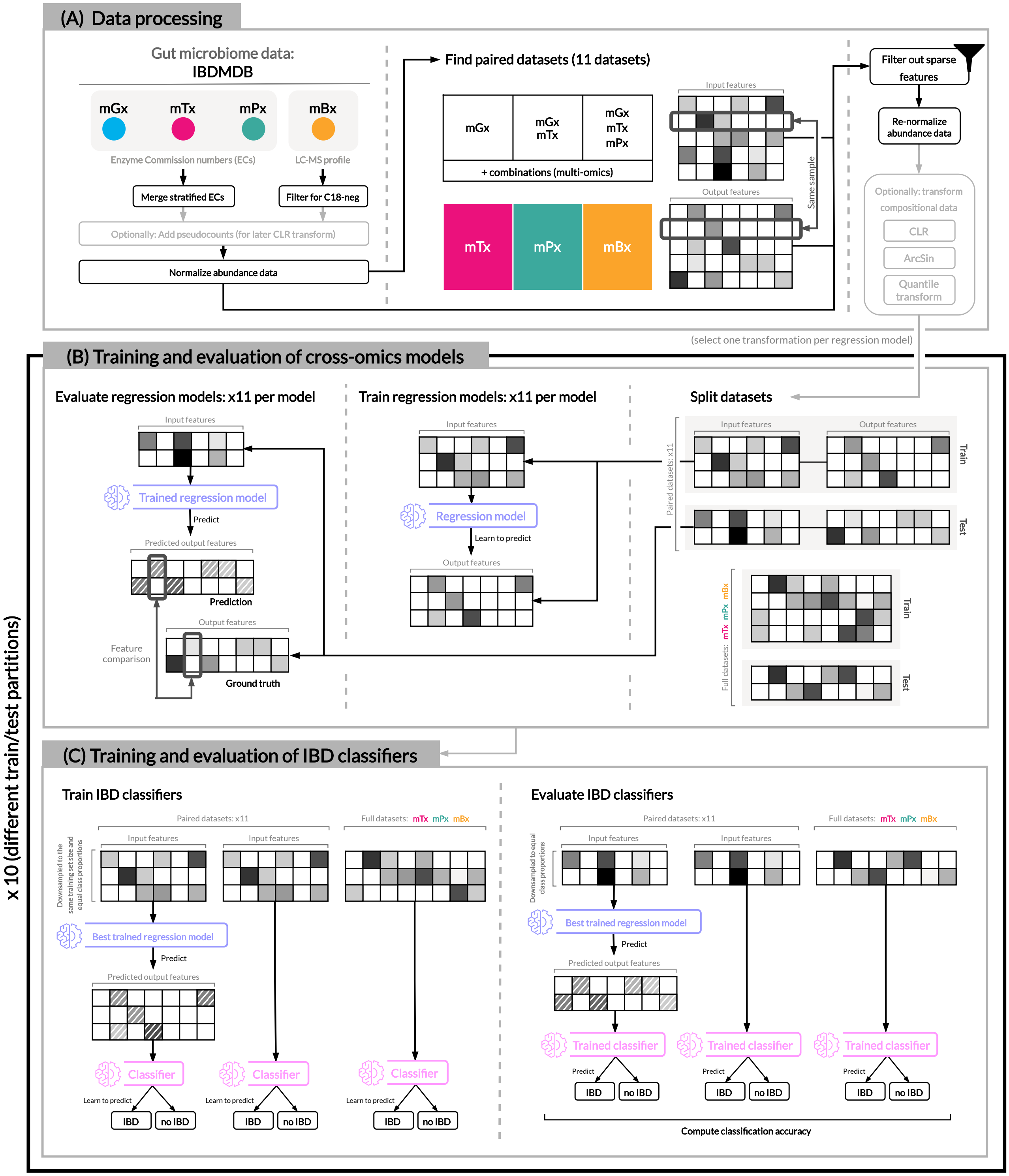

(A) Data processing
Gut microbiome data:
IBDMDB
mGx
mTx
mPx
mBx
Enzyme Commission numbers (ECs)
LC-MS profile
Merge stratified ECs
Filter for C18-neg
Optionally: Add pseudocounts (for later CLR transform)
Normalize abundance data
Find paired datasets (11 datasets)
mGx
mGx mTx
mGx mTx mPx
+ combinations (multi-omics)
mTx
mPx
mBx
Input features
Output features
Same sample
Filter out sparse features
Re-normalize abundance data
Optionally: transform compositional data
CLR
ArcSin
Quantile transform
(select one transformation per regression model)
(B) Training and evaluation of cross-omics models
Evaluate regression models: x11 per model
Input features
Trained regression model
Predict
Predicted output features
Prediction
Feature comparison
Output features
Ground truth
Train regression models: x11 per model
Input features
Regression model
Learn to predict
Output features
Split datasets
Paired datasets: x11
Input features
Output features
Train
Test
Full datasets: mTx mPx mBx
Train
Test
x 10 (different train/test partitions)
(C) Training and evaluation of IBD classifiers
Train IBD classifiers
Paired datasets: x11
Full datasets: mTx mPx mBx
Downsampled to the same training set size and equal class proportions
Input features
Input features
Best trained regression model
Predict
Predicted output features
Classifier
Learn to predict
IBD
no IBD
Evaluate IBD classifiers
Paired datasets: x11
Full datasets: mTx mPx mBx
Downsampled to equal class proportions
Input features
Input features
Best trained regression model
Predict
Predicted output features
Trained classifier
Predict
IBD
no IBD
Compute classification accuracy

Figure 2: Overview of our experimental set-up. We use pre-processed gut microbiome data (A) to train regression models as cross-omics predictors (B), and subsequently evaluated these predictions for the downstream task of IBD prediction (C). Abbreviations: The Inflammatory Bowel Disease Multi'omics Database (IBDMDB), genes derived from metagenomics (mGx), metatranscriptomics (mTx), metaproteomics (mPx), metabolomics (mBx), enzyme commission numbers (ECs), liquid chromatography-mass spectrometry (LC-MS), centered log-ratio (CLR), inflammatory bowel disease (IBD).

3.1. Preliminary experiments

3.1.1. Datasets

We performed preliminary benchmarking of our experimental approach on three datasets containing paired metagenomics and metabolomics data (Supplementary Table S1): Franzosa, Sirota-Madi, Avila-Pacheco, *et al.* [12] (inflammatory bowel disease), Wang, Yang, Li, *et al.* [42] (end-stage renal disease) and Yachida, Mizutani, Shiroma, *et al.* [43] (colorectal cancer). The latter two datasets were downloaded from The Curated Gut Microbiome Metabolome Data Resource [44], release v2.1.0, while the IBD dataset was downloaded from the paper's supplement. We ran MelonnPan [16] on all three datasets, with default settings, to predict metabolite abundances from metagenomics data. The IBD dataset contains metagenomics data in the form of gene families, while the other two contain taxonomic profiles at the species level. We note that MelonnPan [16] was originally trained and tested on the same IBD dataset.

3.1.2. Feature filtering

We tested two approaches for filtering out low-abundance features (species, genes, transcripts, proteins and metabolites):

- **strict filtering:** similarly to Mallick, Franzosa, McIver, *et al.* [16], we retained only those features with at least 0.01% abundance in more than 10% of samples. For all datasets, with the exception of taxonomic profiles and metaproteomics, we additionally filtered out features with less than 0.0001% abundance in more than 10% of samples. Features with more than 95% zeros were also filtered out across all feature types.
- **lenient filtering:** we retained only features with at least 0.005% abundance in more than 10% of samples. Features with more than 95% zeros were filtered out.

3.1.3. Prediction of inflammatory bowel disease

The data collected by Franzosa, Sirota-Madi, Avila-Pacheco, *et al.* [12] included an independently sampled validation cohort, which we used as a test set. We split the two remaining datasets into a training set and a test set, with a ratio of 80% to 20%. We subsequently trained 10 random forest classifiers, initialized with different random seeds (the same ones shown in Supplementary Table S3) to predict phenotypes specific to each dataset. We used `scikit-learn`'s RandomForestClassifier (v1.4.1.post1), with default parameters.

3.2. Main experimental pipeline

Our experimental pipeline and setup are summarized in Figure 2. We divided our workflow into three main steps. We first processed gut microbiome data, created paired datasets for cross-omics prediction, and applied feature filtering (Figure 2(A)). We then trained several regression models to predict one type of meta-omics data from another, and subsequently evaluated them on independent test samples (Figure 2(B)). Lastly, we used predicted data for the downstream task of IBD prediction (Figure 2(C)).

3.2.1. Data processing

Gut microbiome meta-omics data

We downloaded pre-processed metagenomics, metatranscriptomics, metaproteomics and metabolomics data from IBDMDB (Inflammatory Bowel Disease Multi'omics Database) [45], which was assembled as part of the Integrative Human Microbiome Project. The dataset contains longitudinal samples from 132 subjects, including a control group, as well as patients diagnosed with ulcerative colitis (UC) or Crohn's disease (CD). Download links and dates are recorded in Supplementary Table S2. Before feature filtering, all meta-omics abundance profiles were normalized, such that feature values per sample sum up to 1.

Metagenomics (mGx), metatranscriptomics (mTx) and metaproteomics (mPx). We used gene, transcript and protein abundance profiles annotated using Enzyme Commission numbers (ECs). As this data was originally stratified, we summed up ECs across taxonomic groupings to reduce dimensionality and sparsity. Additional experiments supporting all of our main results were performed using pathways and species abundances derived from mGx data, as shown in some of the supplementary results (see Supplementary Tables S3 and S7).

Metabolomics (mBx). We retained mBx data for one LC-MS technology, namely C18 negative (C18-neg).

Imputation of zeros

To enable log transformation of features at a later stage in our pipeline, we also generated versions of these datasets with imputed zeros. For mGx and mTx data, we added $\epsilon = 1\text{e-}6$ to all abundances, which is less than all other non-zero values in the matrices, while for mPx and mBx, which were available as count data, we added a pseudocount.

Paired meta-omics datasets

We generated paired meta-omics datasets for multiple input-output combinations of meta-omics modalities (Figure 2A). Experiments were set up as follows:

- predicting transcripts (mTx) from genes (mGx);
- predicting proteins (mPx) from genes (mGx) and transcripts (mTx);
- and, lastly, predicting metabolites (mBx) from genes (mGx), transcripts (mTx) and proteins (mPx).

In addition, we also predicted mPx and mBx data from multi-omics input, obtained as combinations of single-omics input, constructed using standard feature concatenation. In total, we analyzed results from 11 different input-output combinations of meta-omics modalities. Supplementary experiments were performed for a total of 32 models, including input data types represented as taxonomic profiles and pathways extracted from metagenomics data. A full overview of all paired datasets is provided in Supplementary Table S3.

Feature filtering

Across all datasets, we applied the lenient feature filtering approach previously described in Section 3.1.2.

Data transformations

Following feature filtering, each sample was normalized and the data was transformed using common practices to handle compositionality, sparsity, and feature scaling (Figure 2A). To that end, we compared two standard transformations for compositional data, namely the centered log-ratio (CLR) transformation, which requires the imputation of zeros, and the arcsin square root transformation, which also works on non-imputed data. The CLR transformation of a sample $x \in \mathbb{R}^D$, with sum of elements $\sum_{i=1}^{D} x_i = 1$, $x_i > 0, \forall i \in 1, ..., D$, and $g(x)$ defined as the geometric mean of x, is given by:

$$clr(x) = \left[\log \frac{x_i}{g(x)}\right]_{i \in 1,...,D}. \tag{1}$$

For $x \in \mathbb{R}^D$, with sum of elements $\sum_{i=1}^{D} x_i = 1$ and $0 \leq x_i \leq 1, \forall i \in 1, ..., D$, the arcsin transformation is as follows:

$$a(x) = \left[\arcsin \sqrt{x_i}\right]_{i \in 1,...,D}. \tag{2}$$

As initial experiments showed that MelonnPan [16] performed best among all benchmarked models, we also tested the quantile transformation, which maps normalized features to the quantiles of a normal distribution. This transformation was shown to improve the predictive power of standard regression models and neural networks [46], [47]. Although Mallick, Franzosa, McIver, *et al.* [16] only apply this transformation to the input features, we transformed the output features as well, to preserve consistency with other transformations that we benchmarked. This was done using `scikit-learn`'s QuantileTransformer (v1.4.1post1), with the output distribution set to "normal".

3.2.2. Training and evaluation of cross-omics regression models

Training and testing partitions

To make up for the lack of an independently sampled test set and provide a fair evaluation, we generated 10 train/test splits of each paired dataset, based on a fixed set of random seeds, with a ratio of 80% to 20% (Supplementary Table S3). To reduce overfitting, we performed each split on patients instead of samples, such that samples belonging to the same patient would not be present in both the training and test sets. Each partition was stratified, preserving the proportion of classes (UC, CD, and healthy control (HC)) between the training and test samples. Aside from the paired datasets, we applied the same procedure for the full datasets containing metatranscriptomics (mTx), metaproteomics (mPx) and metabolomics (mBx) data (Figure 2B, Supplementary Table S4), which are later used in benchmarking IBD classifiers (see Section 3.2.3).

Benchmarking of cross-omics regression models

We benchmarked four models and a baseline on several cross-omics prediction tasks. From the literature, we selected MelonnPan [16] (elastic net regression), SparseNED [17] (sparse neural encoder-decoder) and MiMeNet [18] (feed-forward neural network). These architectures were all originally designed to predict metabolite abundances from metagenomics. All models were run with default parameters, except for MiMeNet, where some parameters were changed to reduce runtime (see Supplementary Listing 1). Each model was trained and tested using different data transformations (see Supplementary Table S5). We also trained a deep neural network (Deep NN), with data augmentation (Supplementary Section A.1.1), and a RandomForestRegressor baseline (`scikit-learn` v1.4.1.post1), initialized with default parameters and a random seed equal to 42. For more details regarding the network architecture, as well as the loss used for training, see Section A.1 of the Supplementary Methods, particularly Subsections A.1.2 and A.1.3. Hyperparameter tuning for the neural network is also recorded in Supplementary Table S9.

Model evaluation

We evaluated all models on each independent test set by comparing predicted features (transcripts, proteins, and metabolites) with the ground truth data. Consistent with methods reported in the literature [16], [18], [19], [21], we used Spearman's rank correlation coefficient to compare a predicted feature vector $\hat{y} \in \mathbb{R}^N$ with the ground truth $y \in \mathbb{R}^N$, transformed to ranks $R(\hat{y})$ and $R(y)$:

$$r(\hat{y}, y) = \frac{cov(R(\hat{y}), R(y))}{\sigma_{R(\hat{y})}\sigma_{R(y)}}, \tag{3}$$

where $cov(R(\hat{y}), R(y))$ is the covariance of the rank variables, and $\sigma_{R(\hat{y})}$, $\sigma_{R(y)}$ are the standard deviations.

To compute scores across the 10 test partitions, we first computed the mean Spearman's rank correlation coefficient per individual feature. We then reported the average for the top predicted features. The error was calculated as the mean standard deviation across features.

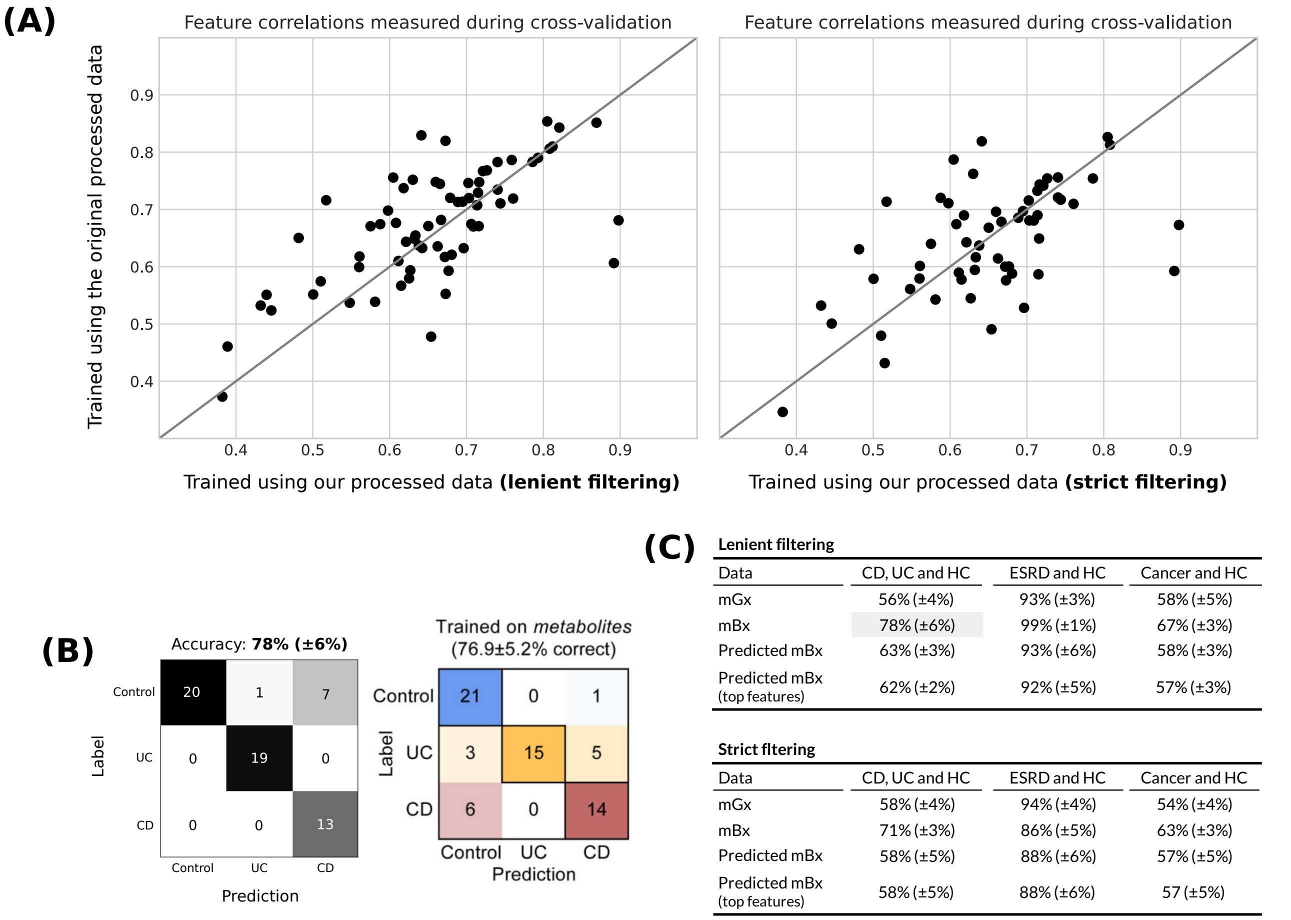
3.2.3. Predicting inflammatory bowel disease

To evaluate the applicability of cross-omics regression models, we used the predicted features for the downstream task of inflammatory bowel disease (IBD) prediction (Figure 2C). All classification tasks were performed using `scikit-learn`'s RandomForestClassifier (v1.4.1.post1), trained using random search cross-validation (Supplementary Table S6) with 50 iterations and a random state equal to the seed corresponding to each train/test partition (all random seeds are listed in Supplementary Table S3). We used 5 stratified cross-validation folds, divided based on study participants.

For each paired dataset, we compared the performance of IBD classifiers trained on the predicted data to that of classifiers trained on the input data used to generate the corresponding predictions. We additionally benchmarked these results against classifiers trained on ground-truth datasets of metatranscriptomics (mTx), metaproteomics (mPx), and metabolomics (mBx) data. The training and test partitions were kept as the same ones used to train the cross-omics regression models (Figure 2B). To provide a fair comparison, for each meta-omics output type (mTx, mPx, mBx), we downsampled the classifier training sets to the size of the smallest paired meta-omics dataset. In addition, each train and test set was downsampled to equal class proportions (IBD and healthy control).

4. Results and discussion

4.1. Preliminary experiments confirm the accuracy and robustness of our pipeline, in accordance with the established literature



Lenient filtering

Data	CD, UC and HC	ESRD and HC	Cancer and HC
mGx	56% (±4%)	93% (±3%)	58% (±5%)
mBx	78% (±6%)	99% (±1%)	67% (±3%)
Predicted mBx	63% (±3%)	93% (±6%)	58% (±3%)
Predicted mBx (top features)	62% (±2%)	92% (±5%)	57% (±3%)

Strict fltering

Data	CD, UC and HC	ESRD and HC	Cancer and HC
mGx	58% (±4%)	94% (±4%)	54% (±4%)
mBx	71% (±3%)	86% (±5%)	63% (±3%)
Predicted mBx	58% (±5%)	88% (±6%)	57% (±5%)
Predicted mBx (top features)	58% (±5%)	88% (±6%)	57 (±5%)

Figure 3: **(A)** Spearman's rank correlations obtained by training MelonnPan [16] using our processed data (x-axis) and the data processed by the authors (y-axis). Correlations were computed during training across 10 folds of cross-validation. We include two feature filtering alternatives: less restrictive (left) and more restrictive (right). The original dataset was published by Franzosa, Sirota-Madi, Avila-Pacheco, *et al.* [12] (see also Supplementary Table S1). **(B)** On the left, a confusion matrix for the result highlighted in sub-figure (C). On the right, a confusion matrix taken from Figure 6 in the study published by Franzosa, Sirota-Madi, Avila-Pacheco, *et al.* [12]. **(C)** Performance of random forest classifiers for three different classification tasks, corresponding to the datasets [12], [42], [43] in Supplementary Table S1. Top mBx features were determined by MelonnPan during cross-validation, with a correlation cut-off equal to 0.3. Abbreviations: Crohn's disease (CD), ulcerative colitis (UC), healthy control (HC), end-stage renal disease (ESRD), genes derived from metagenomics (mGx), metabolomics (mBx).

To validate the steps in our experimental pipeline, we first reproduced a selection of results from the literature (Figure 3(A) and (B)). To that end, we trained MelonnPan [16] on the dataset provided on GitHub as part of the model's package, as well as our processed version of the dataset, originally published by Franzosa, Sirota-Madi, Avila-Pacheco, *et al.* [12]. The processed and predicted data were subsequently used to classify IBD sub-types and healthy patients.

We found that MelonnPan's performance in predicting metabolite abundances from metagenomics data was robust to the data processing and filtering approach that was used (Figure 3(A)). Metabolite correlations measured during cross-validation were mostly consistent for each of the two experiments, with some variations likely occurring as a result of the randomness involved in MelonnPan's training procedure, which includes cross-validation.

However, the filtering procedure did have an impact on classification performance. Figure 3(B) shows that our lenient feature filtering approach resulted in classification accuracy comparable to the one obtained by the authors of the study containing the original dataset [12]. This result is also highlighted in Figure 3(C), which includes results for two other datasets (see Supplementary Table S1), and two filtering approaches. Across all three datasets, we noticed a significant drop in performance for classifiers trained on real metabolomics data, with the strict filtering approach. This suggests that features important for distinguishing phenotype were discarded during filtering, so we opted for the more lenient filtering approach for our main experimental pipeline.

4.2. Machine learning models can reliably predict a subset of meta-omics features

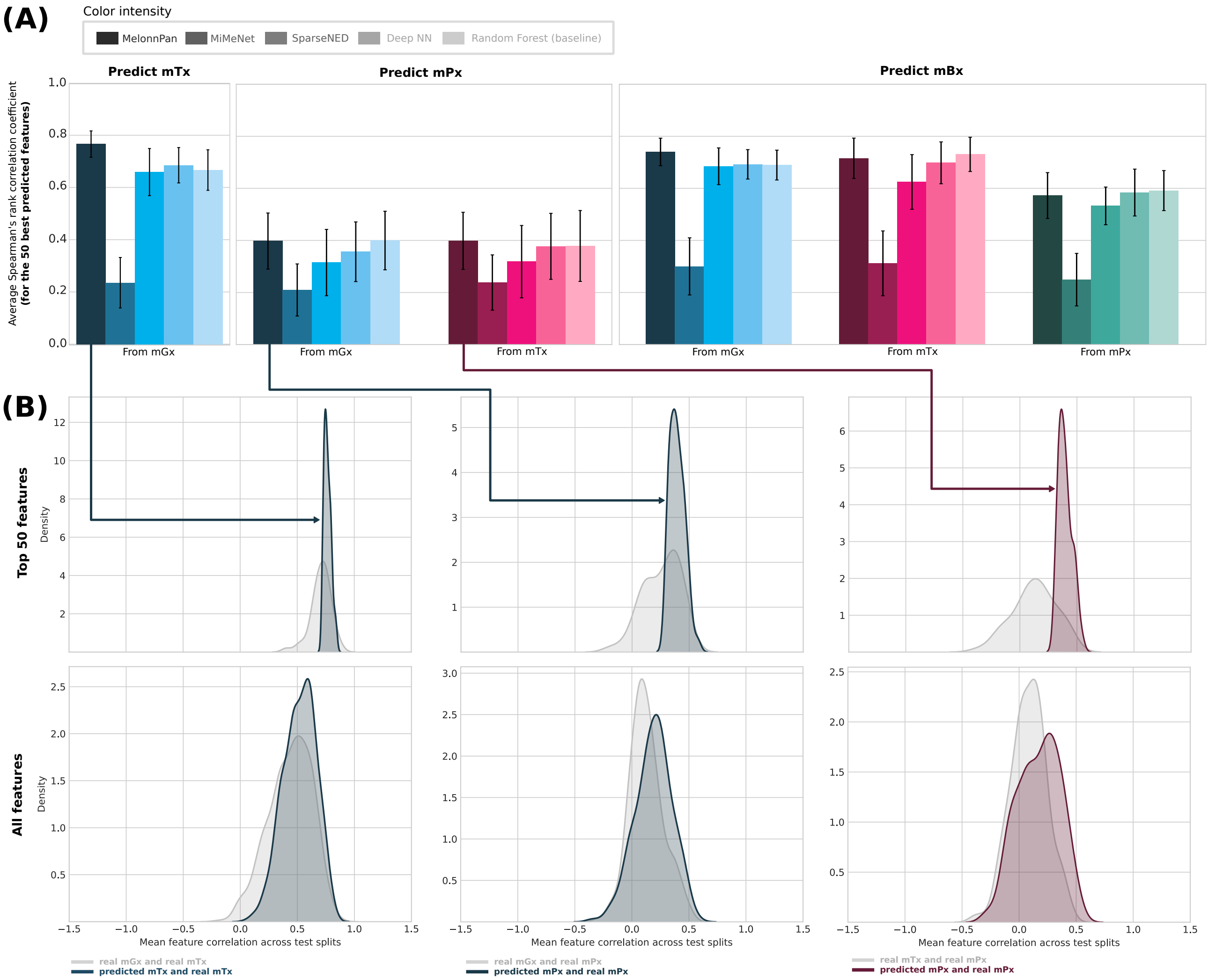


Figure 4: (A) Mean test performance results of cross-omics regression models on several prediction tasks, calculated across 10 different dataset partitions. The average Spearman's rank correlation coefficient was calculated for the 50 best predicted features for each output type. **(B)** For metatranscriptomics (mTx) and metaproteomics (mPx) predictions generated by MelonnPan [16], we also plot kernel density estimates comparing correlations between the input data and the ground-truth mTx/mPx data with those computed between the predicted data and the ground truth data. We perform this analysis on the 50 best predicted features for each output type, as well as all predicted features. Input types are represented through different colors, while cross-omics models are represented using different color intensities. Abbreviations: neural network (NN), genes derived from metagenomics (mGx), metatranscriptomics (mTx), metaproteomics (mPx), metabolomics (mBx).

To assess the generalization performance of machine learning pipelines designed for metabolite prediction, we first trained and evaluated several models from the literature on the task of predicting transcript and protein abundances, in addition to metabolite abundances. Average scores across 10 different train/test partitions were computed for 6 single-omics input-output combinations and 3 models from the literature (MelonnPan [16], MiMeNet [17], SparseNED [18]), along with a deep neural network (Deep NN) and a random forest regressor baseline (Figure 4). Following a pattern established in the literature [16], [18], [19], [21], we plotted the performance of cross-omics regression models for the 50 best predicted features (Figure 4A).

Considering the top predictions, cross-omics regression models performed similarly in predicting metatranscriptomics and metabolomics (Figure 4A). Protein abundances (mPx) were the most challenging to predict. This is expected, given the fact that metaproteomics was characterized by the highest sparsity and that paired data available to train models that predict metaproteomics was most scarce (Supplementary Table S3). Good performance for metatranscriptomics prediction is also not surprising, given the similarities between metagenomics and metatranscriptomics as next-generation sequencing technologies. Lastly, we note that architectures like elastic nets [16] and random forests were more robust across input-output combinations and generally performed best.

To investigate whether machine learning models provide more accurate estimations of transcript and protein abundances compared to the "gene-to-transcript-to-protein" assumption, we generated density plots comparing the distributions of correlations between different types of meta-omics data (Figure 4B). Top predicted transcript and protein abundances were significantly more highly correlated with the ground-truth data, when compared to the distribution of correlations between the input data and the ground-truth. This suggests that a subset of features (transcripts or proteins) can be more reliably predicted using machine learning approaches, rather than relying on the assumption that genes encoded in the metagenome will be transcribed into mRNA and subsequently translated into protein. When plotting correlations for all features, this was still the case, but to a much lesser extent, ultimately indicating that only a subset of features can be reliably predicted. One explanation for this result is that machine learning becomes challenging when the number of features is high relative to the number of samples. This phenomenon, known as the curse of dimensionality, is particularly problematic for multi-output regression tasks such as cross-omics prediction.

4.3. Multi-omics integration does not improve the prediction accuracy of machine learning models

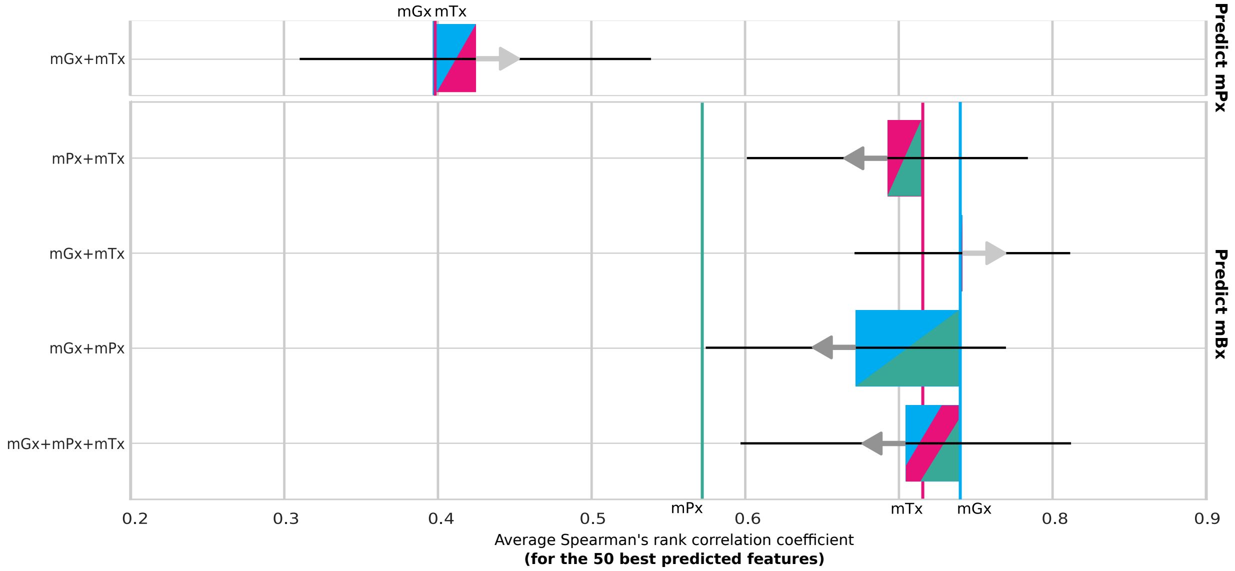


Figure 5: Performance comparison of MelonnPan [16] using multi-omics and single-omics input data. Single-omics input types are shown as lines, using different colors, while model performance on multi-omics data is indicated with the corresponding two- and three-color combination, with diagonally spliced bars. Improvements or downgrades in performance are indicated with arrows. Abbreviations: genes derived from metagenomics (mGx), metatranscriptomics (mTx), metaproteomics (mPx), metabolomics (mBx).

To determine whether using a combination of different types of input features leads to better predictions of protein and metabolite abundances, we additionally trained MelonnPan, the overall best performing model identified in the previous section, on multi-omics input. Figure 5 shows a comparison between single-omics and multi-omics input in predicting metaproteomics and metabolomics. Results for other input types, such as taxonomic profiles and pathways derived from metagenomics, are recorded in Supplementary Table S7.

While metaproteomics predictions slightly improved when combining metagenomics and metatranscriptomics, combining single-omics modalities did not lead to more accurate predictions of metabolite abundances. High variances in model performance also suggest that the increased input dimensionality made learning more difficult and less robust across training sets. We also note that the number of samples available for training decreased with the amount of meta-omics modalities involved (Supplementary Table S3), and that likely also had an influence on these results. Comparable performance was obtained when using metagenomics data processed in the form of pathways or species-level taxonomic profiles (Supplementary Table S7).

We also designed a more elaborate multi-omics integration scheme, using an autoencoder trained with a joint reconstruction and regression loss (Supplementary Section A.2 and Supplementary Figure S3), but preliminary experiments did not show promising results. Therefore, we did not pursue this line of research further. Supplementary Table S8 shows the performance of MelonnPan [16] trained on concatenated multi-omics, compared to the embeddings learned by the autoencoder architecture. Although we observed a decline in prediction accuracy, the models trained on latent features were more robust, as suggested by the very low variation in model performance across test partitions. These results imply that such a dimensionality reduction approach leads to more stable predictions, and that more extensive research and benchmarking may lead to a more reliable multi-meta-omics integration architecture.

4.4. There is a core set of well-predicted features that are robust to changes in input types and dataset partitions

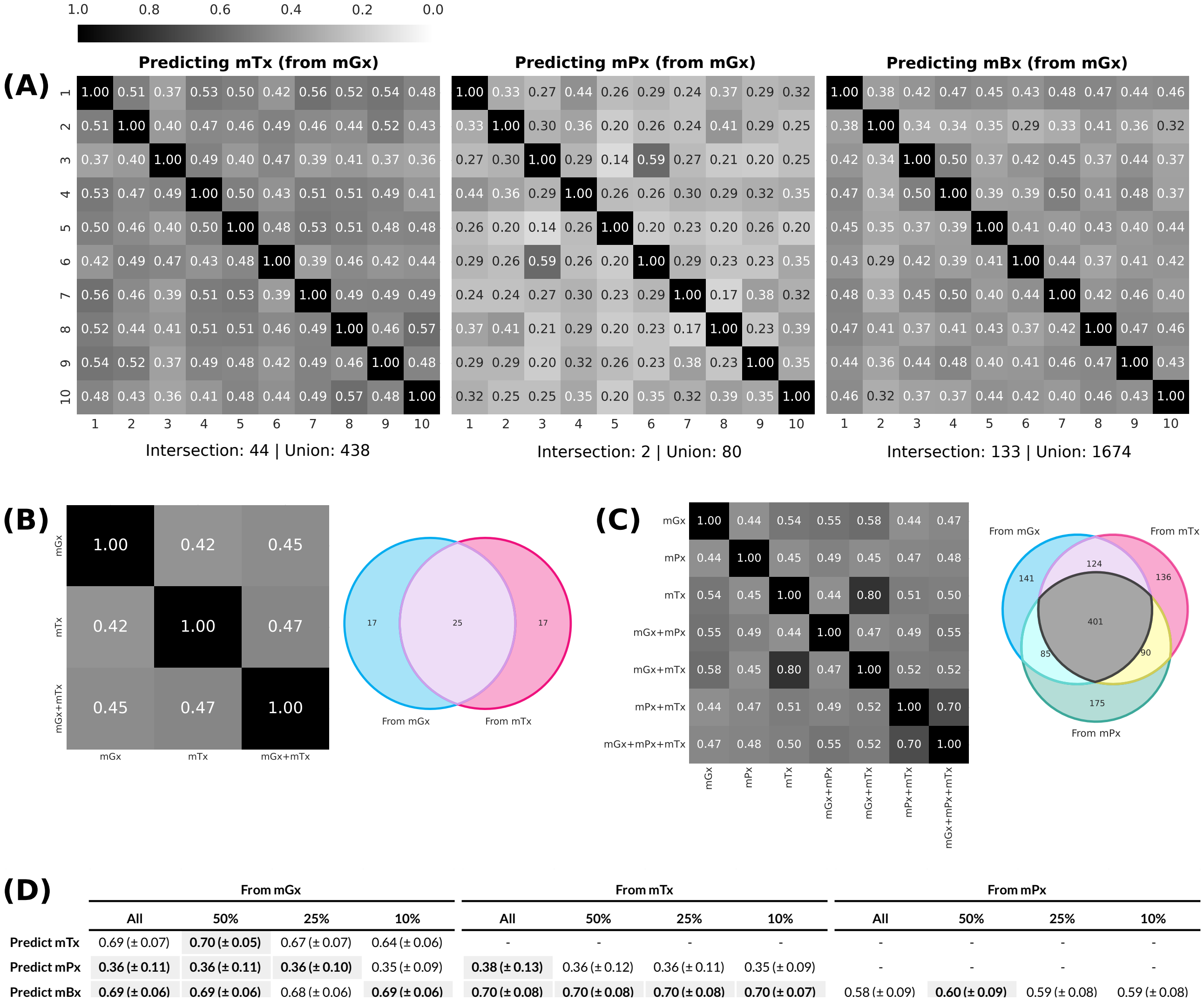
We evaluated the robustness of cross-omics models through an analysis of well-predicted features across dataset partitions and input types (Figure 6(A), (B) and (C)). We limited this analysis to results produced by MelonnPan [16], as we found this model performed best overall for the task of cross-omics prediction (Section 4.2). However, additional experiments were performed with a deep neural network model (Supplementary Section A.1), to study the effect of feature selection on model performance (Figure 6(D)).

A pairwise comparison of the top 25% well-predicted features across train/test partitions showed that these subsets share a selection of features, with some features being well-predicted across all test splits. While Figure 6(A) only shows predictions generated from mGx data, we found this to be the case for all single-omics input types (see Supplementary Figure S1). For each output type, a small set of features was found to be consistently well predicted across dataset partitions. For metaproteomics, this number was especially low (2.5% of the feature union). Glutamate dehydrogenase (1.4.1.3) and DNA-directed RNA polymerase (2.7.7.6) were the two enzymes included in this subset. Low abundance levels for these two enzymes have been associated with IBD diagnosis [48]–[50], with glutamate dehydrogenase being linked to *Clostridium difficile* infections in IBD patients.

Some well-predicted features were also shared across single-omics and multi-omics input types (Figure 6(B) and (C)). In total, 25 proteins were well-predicted from both metagenomics and metatranscriptomics data, while 401 metabolites were well-predicted from metagenomics, metatranscriptomics and metaproteomics data. We also note that we did not find significant correlations between feature variance and prediction quality (Supplementary Figure S2). The characterization of well-predicted features remains a largely unanswered question, requiring more in-depth analysis.

However, regardless of what makes features easy to predict, these results imply that there is a core subset of features, for each output type, that can be reliably predicted. Consequently, we hypothesized that training a model on just a subset of features would lead to better predictions, as the trade-off between data dimensionality and the number of samples would be more balanced in that case. To that end, we ran a pre-training iteration which consisted of training 10 random forest models on different cross-validation splits, averaging feature correlations and retaining only a proportion of the top features (Supplementary Section A.3 and Supplementary Figure S4). We then trained a deep neural network (Supplementary Section A.1), restricting the output to subsets containing 50%, 25% and 10%

of features based on individual correlations obtained during pre-training (Figure 6(D)). As opposed to elastic net architectures such as MelonnPan [16], which combine multiple single-output regression models [51], neural networks can exploit dependencies between output variables in multi-output regression. In addition, deep architectures were shown to be better at bypassing the curse of dimensionality, particularly when modelling compositional functions [52]. The results included in Figure 6D do not fully support our hypothesis, as we did not find that training on a smaller set of features improved network performance. However, in many cases, performance also did not decline when using fewer output features to train the network. One possible cause could be that important dependencies between features are preserved, to an extent, during feature selection, and that is worth further investigation.



Predicting mTx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.51	0.37	0.53	0.50	0.42	0.56	0.52	0.54	0.48
2	0.51	1.00	0.40	0.47	0.46	0.49	0.46	0.44	0.52	0.43
3	0.37	0.40	1.00	0.49	0.40	0.47	0.39	0.41	0.37	0.36
4	0.53	0.47	0.49	1.00	0.50	0.43	0.51	0.51	0.49	0.41
5	0.50	0.46	0.40	0.50	1.00	0.48	0.53	0.51	0.48	0.48
6	0.42	0.49	0.47	0.43	0.48	1.00	0.39	0.46	0.42	0.44
7	0.56	0.46	0.39	0.51	0.53	0.39	1.00	0.49	0.49	0.49
8	0.52	0.44	0.41	0.51	0.51	0.46	0.49	1.00	0.46	0.57
9	0.54	0.52	0.37	0.49	0.48	0.42	0.49	0.46	1.00	0.48
10	0.48	0.43	0.36	0.41	0.48	0.44	0.49	0.57	0.48	1.00

Intersection: 44 | Union: 438

Predicting mPx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.33	0.27	0.44	0.26	0.29	0.24	0.37	0.29	0.32
2	0.33	1.00	0.30	0.36	0.20	0.26	0.24	0.41	0.29	0.25
3	0.27	0.30	1.00	0.29	0.14	0.59	0.27	0.21	0.20	0.25
4	0.44	0.36	0.29	1.00	0.26	0.26	0.30	0.29	0.32	0.35
5	0.26	0.20	0.14	0.26	1.00	0.20	0.23	0.20	0.26	0.20
6	0.29	0.26	0.59	0.26	0.20	1.00	0.29	0.23	0.23	0.35
7	0.24	0.24	0.27	0.30	0.23	0.29	1.00	0.17	0.38	0.32
8	0.37	0.41	0.21	0.29	0.20	0.23	0.17	1.00	0.23	0.39
9	0.29	0.29	0.20	0.32	0.26	0.23	0.38	0.23	1.00	0.35
10	0.32	0.25	0.25	0.35	0.20	0.35	0.32	0.39	0.35	1.00

Intersection: 2 | Union: 80

Predicting mBx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.38	0.42	0.47	0.45	0.43	0.48	0.47	0.44	0.46
2	0.38	1.00	0.34	0.34	0.35	0.29	0.33	0.41	0.36	0.32
3	0.42	0.34	1.00	0.50	0.37	0.42	0.45	0.37	0.44	0.37
4	0.47	0.34	0.50	1.00	0.39	0.39	0.50	0.41	0.48	0.37
5	0.45	0.35	0.37	0.39	1.00	0.41	0.40	0.43	0.40	0.44
6	0.43	0.29	0.42	0.39	0.41	1.00	0.44	0.37	0.41	0.42
7	0.48	0.33	0.45	0.50	0.40	0.44	1.00	0.42	0.46	0.40
8	0.47	0.41	0.37	0.41	0.43	0.37	0.42	1.00	0.47	0.46
9	0.44	0.36	0.44	0.48	0.40	0.41	0.46	0.47	1.00	0.43
10	0.46	0.32	0.37	0.37	0.44	0.42	0.40	0.46	0.43	1.00

Intersection: 133 | Union: 1674

	mGx	mTx	mGx+mTx
mGx	1.00	0.42	0.45
mTx	0.42	1.00	0.47
mGx+mTx	0.45	0.47	1.00

	mGx	mPx	mTx	mGx+mPx	mGx+mTx	mPx+mTx	mGx+mPx+mTx
mGx	1.00	0.44	0.54	0.55	0.58	0.44	0.47
mPx	0.44	1.00	0.45	0.49	0.45	0.47	0.48
mTx	0.54	0.45	1.00	0.44	0.80	0.51	0.50
mGx+mPx	0.55	0.49	0.44	1.00	0.47	0.49	0.55
mGx+mTx	0.58	0.45	0.80	0.47	1.00	0.52	0.52
mPx+mTx	0.44	0.47	0.51	0.49	0.52	1.00	0.70
mGx+mPx+mTx	0.47	0.48	0.50	0.55	0.52	0.70	1.00

	From mGx				From mTx				From mPx			
	All	50%	25%	10%	All	50%	25%	10%	All	50%	25%	10%
Predict mTx	0.69 (± 0.07)	**0.70 (± 0.05)**	0.67 (± 0.07)	0.64 (± 0.06)	-	-	-	-	-	-	-	-
Predict mPx	**0.36 (± 0.11)**	**0.36 (± 0.11)**	**0.36 (± 0.10)**	0.35 (± 0.09)	**0.38 (± 0.13)**	0.36 (± 0.12)	0.36 (± 0.11)	0.35 (± 0.09)	-	-	-	-
Predict mBx	**0.69 (± 0.06)**	**0.69 (± 0.06)**	0.68 (± 0.06)	**0.69 (± 0.06)**	**0.70 (± 0.08)**	**0.70 (± 0.08)**	**0.70 (± 0.08)**	**0.70 (± 0.07)**	0.58 (± 0.09)	**0.60 (± 0.09)**	0.59 (± 0.08)	0.59 (± 0.08)

Figure 6: **(A)** Jaccard similarities between the sets of the 25% best predicted features by MelonnPan [16] for each output type (mTx, mPx, mBx), compared across 10 different train/test partitions. All predictions were generated from mGx data. **(B)** Jaccard similarities and Venn diagram of the sets of the 25% best predicted proteins, compared across input types. **(C)** Jaccard similarities and Venn diagram of the sets of the 25% best predicted metabolites, compared across input types. **(D)** Performance of a deep neural network model (Supplementary Section A.1) trained on different feature subsets, based on a pre-training step for feature selection (Supplementary Section A.3). The best results for each input-output combination are highlighted. Abbreviations: genes derived from metagenomics (mGx), metatranscriptomics (mTx), metaproteomics (mPx), metabolomics (mBx).

4.5. Predicted meta-omics data can be used to classify phenotypes, with performance comparable to that of real data

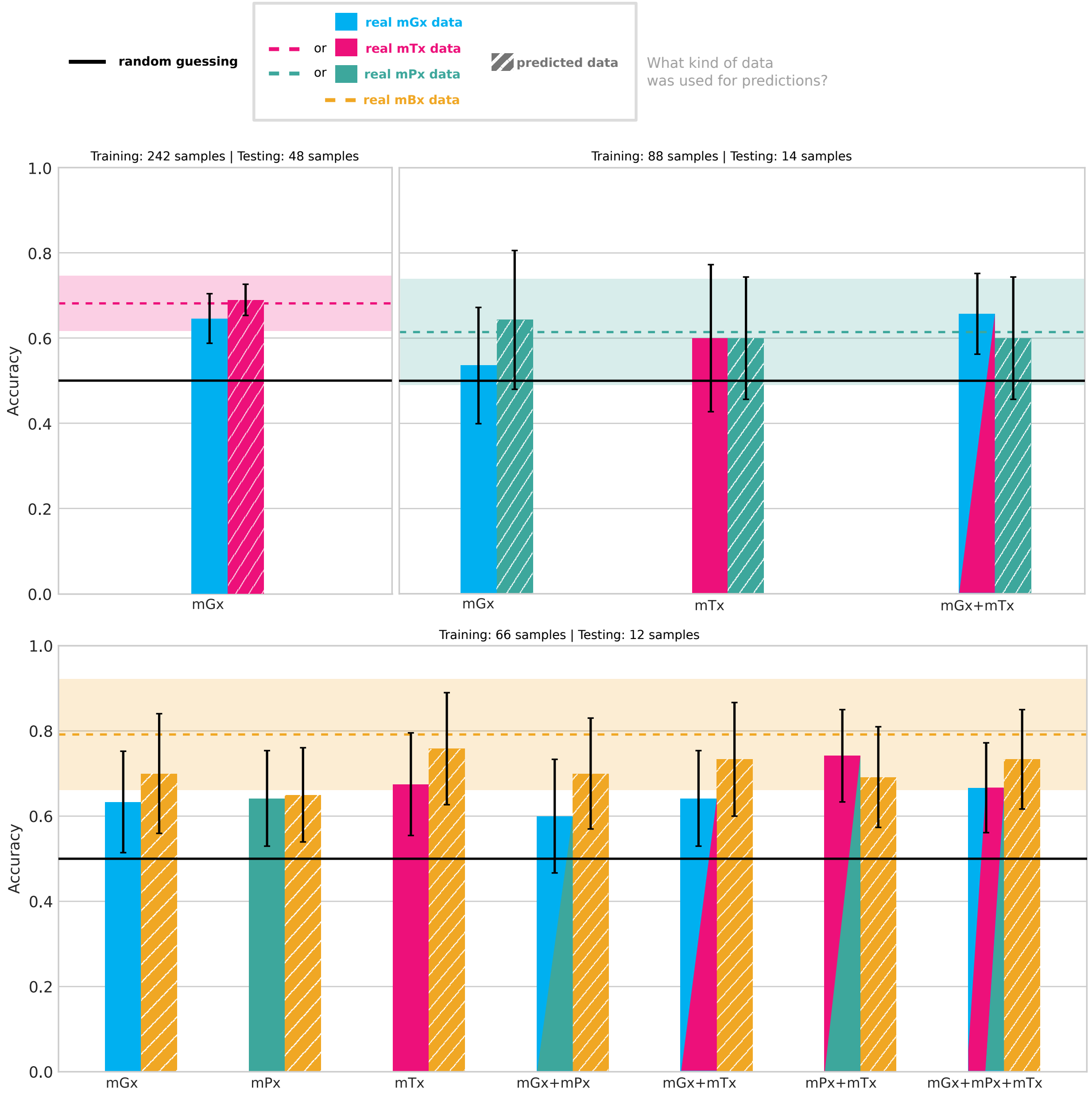


Figure 7: Accuracy of random forest classifiers on the binary task of inflammatory bowel disease prediction, averaged across 10 test partitions. Left bars indicate input types, while right, striped bars, indicate the output predicted from these inputs. Predicted data was generated with MelonnPan [16]. Abbreviations: genes derived from metagenomics (mGx), metatranscriptomics (mTx), metaproteomics (mPx), metabolomics (mBx).

Lastly, to demonstrate an application of cross-omics regression models, we tested whether predictions generated by cross-omics prediction models could be used for the downstream task of IBD prediction. We compared the classification performance of random forest classifiers trained on input and predicted data, as well as real data of the same modality as the predictions (Figure 7).

Overall, predicted features could be used to distinguish between IBD and the healthy control, with accuracies comparable to those obtained with real data. Metabolomics was identified as the best predictor of IBD, with a mean accuracy of almost 80%. Metabolite abundances predicted from metatranscriptomics almost matched this performance. In the majority of cases, predicted features were better predictors of inflammatory bowel disease compared to the input data from which they were generated. Using input data as opposed to predicted data was only preferred in multi-omics settings.

Notably, classification performance was quite low across all datasets, but this outcome was not surprising. We did not expect IBD classifiers to perform particularly well on this dataset, especially due to the limited number of training samples. This is a result of downsampling to equal class proportions, on top of downsampling to the size of the smallest paired dataset for each output type. Additionally, the IBD samples in IBDMDB [45] were not all collected from patients with active disease, making it harder to distinguish between the two classes of samples. Our initial experiments on other datasets (Section 4.1) also provided evidence that this issue is dataset-related, as random forest classifiers with no hyperparameter tuning were able to achieve good performance with real metabolomics data, even for more challenging classification tasks. Ultimately, more extensive benchmarking on other datasets is required to gain a deeper understanding on the uses of predicted meta-omics data for phenotype identification.

5. Conclusion

As an answer to our first research question, regarding the generalization performance of microbe-to-metabolite models on other meta-omics types, we conclude that regression models for cross-omics prediction are able to accurately predict a subset of features, whether those features are transcripts, proteins or metabolites. This confirms our initial hypothesis, to a certain extent. Although metaproteomics prediction was challenging, our experiments showed that metatranscriptomics and metabolomics features were reliably predicted. Therefore, we were able to validate similar results from the literature on metabolomics prediction [16]–[18].

In addition, we showed that machine learning models provide reliable insights into metatranscriptomics and metaproteomics. Feature correlations between ground-truth and predicted data were generally higher than those obtained between genes and transcripts, or transcripts and proteins.

In relation to our second research objective, we hypothesized that a smaller set of meta-omics features should be easier to predict. Although our results did not fully support this hypothesis, we found that it is possible to achieve similar prediction accuracy when using fewer output features to train a model. This suggests that essential dependencies between predicted features were preserved during the feature selection process.

Finally, we proposed that predicted meta-omics features could be used to distinguish between different phenotypes. For the binary task of IBD classification, this hypothesis was confirmed. Some predicted datasets lead to classification performance similar to that obtained using real data, and generally better than that obtained using the input data at the basis of those predictions.

5.1. Future work

First of all, we propose that future research should focus on the creation of curated datasets of paired meta-omics data. Although the Integrative Human Microbiome Project achieves this to a much appreciated extent, this database is still characterized by heterogeneous data, generally available in large volumes and unprocessed. One improvement to such initiatives is the availability of homogeneous, pre-processed data, which would further streamline and improve the performance of machine learning models and deep learning architectures for cross-omics prediction.

Secondly, it is worthwhile to perform more extensive benchmarking on multi-omics integration approaches. In addition to naive feature concatenation, we only analyzed results from one autoencoder architecture, and those results were not very encouraging, at least not performance-wise. While out-of-scope for the timeline of this project, we believe that further hyperparameter tuning, particularly focused on the loss function, could lead to higher prediction accuracy.

Lastly, there is considerable potential in improving cross-omics prediction models. Although our experiments on feature selection were not fully successful, our results show that trying to learn fewer output features leads to similar performance as using the full set of output features. This implies that an essential set of dependencies between output features was captured with just those selected features. To that end, further investigation should focus on a more in-depth characterization of well-predicted transcripts, proteins or metabolites, and, consequently, strategies to improve the robustness of prediction models.

Glossary

cross-omics We use this term to refer to the prediction of one type of meta-omics from another, for instance, predicting protein abundances (metaproteomics) from transcript abundances (metatranscriptomics). This is done through the use of machine learning architectures that can perform regression tasks.

meta-omics Umbrella term for omics technologies used to study the microbiome, including metagenomics, metatranscriptomics, metaproteomics and metabolomics. Note the use of the prefix "meta", which differentiates meta-omics from standard omics technologies. While "omics" refers to the study of a single organism, "meta-omics" implies the study of multiple microorganisms (the microbiome). One exception to the rule is "metabolomics", for which the double use of the prefix would have made an awkward noun.

metabolomics Meta-omics technologies used to study the metabolites produced as part of microbial metabolism. As metaproteomics, metabolomics relies on mass spectrometry approaches.

metagenomics Meta-omics technologies used to study the DNA of microbial communities. Metagenomics is nowadays usually done through next-generation sequencing, and can be used to identify the microorganisms present in a microbial sample, and quantify their functional potential based on the genes found on the metagenome.

metaproteomics Meta-omics technologies used to study the proteins produced by a microbial community. Unlike in the case of metagenomics and metatranscriptomics, it is not currently possible to sequence the metaproteome, and we rely instead on mass spectrometry to identify the proteins in a microbial sample. Protein identification enables functional profiling at a higher level compared to metatranscriptomics and metagenomics.

metatranscriptomics Meta-omics technologies used to study the transcripts (RNA fragments) of microbial communities. Metatranscriptomics is nowadays achieved using next-generation sequencing, and it is generally used to quantify the functional potential of microbial communities, through the analysis of gene expression levels.

microbiome A community of microorganisms, including bacteria, viruses, fungi and archaea, that are found in a particular environment. For example, the gut microbiome generally refers to the microorganisms that populate the human gut.

Acronyms

CD Crohn's disease.

cDNA complementary DNA.

CLR centered log-ratio.

ECs enzyme commission numbers.

ESRD end-stage renal disease.

HC healthy control.

IBD inflammatory bowel disease.

IBDMDB The Inflammatory Bowel Disease Multi'omics Database.

ITS internal transcribed spacers.

LC liquid chromatography.

LC-MS liquid chromatography-mass spectrometry.

mBx metabolomics.

mGx genes derived from metagenomics.

mGx_pa pathways derived from metagenomics.

mGx_taxa taxonomic profile derived from metagenomics.

mPx metaproteomics.

mRNA messenger RNA.

MS mass spectrometry.

mTx metatranscriptomics.

NGS next-generation sequencing.

NMR nuclear magnetic resonance.

NN neural network.

ORF open-reading frame.

PCR polymerase chain reaction.

rDNA ribosomal DNA.

UC ulcerative colitis.

WMS whole-metagenome sequencing.

6. List of publications

Aysun Urhan, **Bianca-Maria Cosma**, Ashlee M Earl, Abigail L Manson, Thomas Abeel, SAFPred: synteny-aware gene function prediction for bacteria using protein embeddings, *Bioinformatics*, Volume 40, Issue 6, June 2024, btae328, https://doi.org/10.1093/bioinformatics/btae328

Bianca-Maria Cosma, Ramin Shirali Hossein Zade, Erin Noel Jordan, Paul van Lent, Chengyao Peng, Stephanie Pillay, Thomas Abeel, Evaluating long-read de novo assembly tools for eukaryotic genomes: insights and considerations, *GigaScience*, Volume 12, 2023, giad100, https://doi.org/10.1093/gigascience/giad100

7. List of presentations

Oral presentations

The artificially generated microbiome: a study on the generation and potential use cases of predicted meta-omics data, *BioSB*, June 2024, Egmond aan Zee, Netherlands

The artificially generated microbiome: a study on the generation and potential use cases of predicted meta-omics data, presented by Thomas Abeel, *AI4b.io Annual Symposium*, April 2024, Delft, Netherlands

SAP: Synteny-aware gene function prediction for bacteria using protein embeddings, presented by Aysun Urhan, *ISMB/ECCB*, July 2023, Lyon, France, https://www.youtube.com/watch?v=3c8BvjQY0Lo

Poster presentations

The artificially generated microbiome: a study on the generation and potential use cases of predicted meta-omics data, *BioDay 2024*, April 2024, Delft, Netherlands

References

[1] A. Vijay and A. M. Valdes, "Role of the gut microbiome in chronic diseases: A narrative review," *European Journal of Clinical Nutrition*, vol. 76, no. 4, pp. 489–501, Apr. 2022, Number: 4 Publisher: Nature Publishing Group, ISSN: 1476-5640. DOI: 10.1038/s41430-021-00991-6. [Online]. Available: https://www.nature.com/articles/s41430-021-00991-6.

[2] D. Radjabzadeh, J. A. Bosch, A. G. Uitterlinden, *et al.*, "Gut microbiome-wide association study of depressive symptoms," *Nature Communications*, vol. 13, no. 1, p. 7128, Dec. 6, 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-022-34502-3. [Online]. Available: https://www.nature.com/articles/s41467-022-34502-3.

[3] H. Kaur, Y. Singh, S. Singh, and R. B. Singh, "Gut microbiome-mediated epigenetic regulation of brain disorder and application of machine learning for multi-omics data analysis," *Genome*, vol. 64, no. 4, pp. 355–371, Apr. 2021, ISSN: 1480-3321. DOI: 10.1139/gen-2020-0136.

[4] M. Kyrgiou and A.-B. Moscicki, "Vaginal microbiome and cervical cancer," *Seminars in Cancer Biology*, vol. 86, pp. 189–198, Nov. 1, 2022, ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2022.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1044579X2200061X.

[5] S. Pillay, D. Calderón-Franco, A. Urhan, and T. Abeel, "Metagenomic-based surveillance systems for antibiotic resistance in non-clinical settings," *Frontiers in Microbiology*, vol. 13, 2022, ISSN: 1664-302X. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1066995.

[6] N. Roothans, M. Gabriëls, M. Pabst, M. C. M. v. Loosdrecht, and M. Laureni, *Aerobic denitrification as n2o source in microbial communities*, Pages: 2023.06.14.544945 Section: New Results, Jun. 14, 2023. DOI: 10.1101/2023.06.14.544945. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.06.14.544945v1.

[7] E. B.-M. Daliri, F. K. Ofosu, R. Chelliah, B. H. Lee, and D.-H. Oh, "Challenges and perspective in integrated multi-omics in gut microbiota studies," *Biomolecules*, vol. 11, no. 2, p. 300, Feb. 2021, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2218-273X. DOI: 10.3390/biom11020300. [Online]. Available: https://www.mdpi.com/2218-273X/11/2/300.

[8] S. Bashiardes, G. Zilberman-Schapira, and E. Elinav, "Use of metatranscriptomics in microbiome research," *Bioinformatics and Biology Insights*, vol. 10, BBI.S34610, Jan. 1, 2016, Publisher: SAGE Publications Ltd STM, ISSN: 1177-9322. DOI: 10.4137/BBI.S34610. [Online]. Available: https://doi.org/10.4137/BBI.S34610.

[9] T. Ojala, E. Kankuri, and M. Kankainen, "Understanding human health through metatranscriptomics," *Trends in Molecular Medicine*, vol. 29, no. 5, pp. 376–389, May 1, 2023, Publisher: Elsevier, ISSN: 1471-4914, 1471-499X. DOI: 10.1016/j.molmed.2023.02.002. [Online]. Available: https://www.cell.com/trends/molecular-medicine/abstract/S1471-4914(23)00034-5.

[10] M. Storr, H. J. Vogel, and R. Schicho, "Metabolomics: Is it useful for IBD?" *Current opinion in gastroenterology*, vol. 29, no. 4, pp. 378–383, Jul. 2013, ISSN: 0267-1379. DOI: 10.1097/MOG.0b013e328361f488. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693066/.

[11] F. Salvato, R. L. Hettich, and M. Kleiner, "Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes," *PLoS Pathogens*, vol. 17, no. 2, e1009245, Feb. 25, 2021, ISSN: 1553-7366. DOI: 10.1371/journal.ppat.1009245. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7906368/.

[12] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, *et al.*, "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, no. 2, pp. 293–305, Feb. 2019, Number: 2 Publisher: Nature Publishing Group, ISSN: 2058-5276. DOI: 10.1038/s41564-018-0306-4. [Online]. Available: https://www.nature.com/articles/s41564-018-0306-4.

[13] N. A. Bokulich, P. Łaniewski, A. Adamov, D. M. Chase, J. G. Caporaso, and M. M. Herbst-Kralovetz, "Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment," *PLoS computational biology*, vol. 18, no. 2, e1009876, Feb. 2022, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009876.

[14] Y. Wang, Y. Zhou, X. Xiao, J. Zheng, and H. Zhou, "Metaproteomics: A strategy to study the taxonomy and functionality of the gut microbiota," *Journal of Proteomics*, vol. 219, p. 103 737, May 15, 2020, ISSN: 1876-7737. DOI: 10.1016/j.jprot.2020.103737.

[15] L. M. Proctor, H. H. Creasy, J. M. Fettweis, *et al.*, "The integrative human microbiome project," *Nature*, vol. 569, no. 7758, pp. 641–648, May 2019, Number: 7758 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1238-8. [Online]. Available: https://www.nature.com/articles/s41586-019-1238-8.

[16] H. Mallick, E. A. Franzosa, L. J. Mclver, *et al.*, "Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences," *Nature Communications*, vol. 10, no. 1, p. 3136, Jul. 17, 2019, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-019-10927-1. [Online]. Available: https://www.nature.com/articles/s41467-019-10927-1.

[17] V. Le, T. P. Quinn, T. Tran, and S. Venkatesh, "Deep in the bowel: Highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome," *BMC Genomics*, vol. 21, no. 4, p. 256, Jul. 20, 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-020-6652-7. [Online]. Available: https://doi.org/10.1186/s12864-020-6652-7.

[18] D. Reiman, B. T. Layden, and Y. Dai, "MiMeNet: Exploring microbiome-metabolome relationships using neural networks," *PLOS Computational Biology*, vol. 17, no. 5, e1009021, May 17, 2021, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009021. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009021 (visited on 03/02/2024).

[19] T. Wang, X.-W. Wang, K. A. Lee-Sarwar, *et al.*, "Predicting metabolomic profiles from microbial composition through neural ordinary differential equations," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 284–293, Mar. 2023, Number: 3 Publisher: Nature Publishing Group, ISSN: 2522-5839. DOI: 10.1038/s42256-023-00627-3. [Online]. Available: https://www.nature.com/articles/s42256-023-00627-3 (visited on 03/02/2024).

[20] W. Tang, H. Zheng, S. Xu, *et al.*, "MMINP: A computational framework of microbe-metabolite interactions-based metabolic profiles predictor based on the o2-PLS algorithm," *Gut Microbes*, vol. 15, no. 1, p. 2 223 349, 2023, ISSN: 1949-0976. DOI: 10.1080/19490976.2023.2223349. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10262779/ (visited on 03/02/2024).

[21] O. Shtossel, O. Koren, I. Shai, E. Rinott, and Y. Louzoun, "Gut microbiome-metabolome interactions predict host condition," *Microbiome*, vol. 12, no. 1, p. 24, Feb. 10, 2024, ISSN: 2049-2618. DOI: 10.1186/s40168-023-01737-1. [Online]. Available: https://doi.org/10.1186/s40168-023-01737-1 (visited on 03/02/2024).

[22] M. Cobb, "60 years ago, francis crick changed the logic of biology," *PLoS Biology*, vol. 15, no. 9, e2003243, Sep. 18, 2017, ISSN: 1544-9173. DOI: 10.1371/journal.pbio.2003243. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5602739/ (visited on 05/29/2024).

[23] Y.-X. Liu, Y. Qin, T. Chen, *et al.*, "A practical guide to amplicon and metagenomic analysis of microbiome data," *Protein & Cell*, vol. 12, no. 5, pp. 315–330, May 1, 2021, ISSN: 1674-800X. DOI: 10.1007/s13238-020-00724-8. [Online]. Available: https://doi.org/10.1007/s13238-020-00724-8.

[24] C. P. Kolbert and D. H. Persing, "Ribosomal DNA sequencing as a tool for identification of bacterial pathogens," *Current Opinion in Microbiology*, vol. 2, no. 3, pp. 299–305, Jun. 1, 1999, ISSN: 1369-5274. DOI: 10.1016/S1369-5274(99)80052-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369527499800526.

[25] P. Rausch, M. Rühlemann, B. M. Hermes, *et al.*, "Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms," *Microbiome*, vol. 7, no. 1, p. 133, Sep. 14, 2019, ISSN: 2049-2618. DOI: 10.1186/s40168-019-0743-1. [Online]. Available: https://doi.org/10.1186/s40168-019-0743-1.

[26] B. J. Woodcroft, J. A. Boyd, and G. W. Tyson, "OrfM: A fast open reading frame predictor for metagenomic data," *Bioinformatics*, vol. 32, no. 17, pp. 2702–2703, Sep. 1, 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw241. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw241.

[27] A. L. Lapidus and A. I. Korobeynikov, "Metagenomic data assembly – the way of decoding unknown microorganisms," *Frontiers in Microbiology*, vol. 12, 2021, ISSN: 1664-302X. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.613791.

[28] Y. Xu and F. Zhao, "Single-cell metagenomics: Challenges and applications," *Protein & Cell*, vol. 9, no. 5, pp. 501–510, May 1, 2018, ISSN: 1674-800X. DOI: 10.1007/s13238-018-0544-5. [Online]. Available: https://doi.org/10.1007/s13238-018-0544-5.

[29] G. M. Douglas, R. Hansen, C. M. A. Jones, *et al.*, "Multi-omics differentially classify disease state and treatment outcome in pediatric crohn's disease," *Microbiome*, vol. 6, no. 1, p. 13, Jan. 15, 2018, ISSN: 2049-2618. DOI: 10.1186/s40168-018-0398-3. [Online]. Available: https://doi.org/10.1186/s40168-018-0398-3.

[30] R. S. Mehta, G. S. Abu-Ali, D. A. Drew, *et al.*, "Stability of the human faecal microbiome in a cohort of adult men," *Nature Microbiology*, vol. 3, no. 3, pp. 347–355, Mar. 2018, Number: 3 Publisher: Nature Publishing Group, ISSN: 2058-5276. DOI: 10.1038/s41564-017-0096-0. [Online]. Available: https://www.nature.com/articles/s41564-017-0096-0.

[31] Y. Zhang, K. N. Thompson, T. Branck, *et al.*, "Metatranscriptomics for the human microbiome and microbial community functional profiling," *Annual Review of Biomedical Data Science*, vol. 4, no. 1, pp. 279–311, Jul. 20, 2021, ISSN: 2574-3414, 2574-3414. DOI: 10.1146/annurev-biodatasci-031121-103035. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-031121-103035.

[32] B. Ditz, J. Boekhoudt, N. Couto, *et al.*, "The microbiome in bronchial biopsies from smokers and ex-smokers with stable COPD - a metatranscriptomic approach," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 19, no. 1, pp. 81–87, Dec. 31, 2022, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15412555.2022.2033193, ISSN: 1541-2555. DOI: 10.1080/15412555.2022.2033193. [Online]. Available: https://doi.org/10.1080/15412555.2022.2033193.

[33] T. L. Masters, C. A. Hilker, P. R. Jeraldo, *et al.*, "Comparative evaluation of cDNA library construction approaches for RNA-seq analysis from low RNA-content human specimens," *Journal of Microbiological Methods*, vol. 154, pp. 55–62, Nov. 1, 2018, ISSN: 0167-7012. DOI: 10.1016/j.mimet.2018.10.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167701218303749.

[34] A. J. Westermann, L. Barquist, and J. Vogel, "Resolving host–pathogen interactions by dual RNA-seq," *PLOS Pathogens*, vol. 13, no. 2, e1006033, Feb. 16, 2017, Publisher: Public Library of Science, ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1006033. [Online]. Available: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006033.

[35] M. Kleiner, "Metaproteomics: Much more than measuring gene expression in microbial communities," *mSystems*, May 21, 2019, Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC. DOI: 10.1128/msystems.00115-19. [Online]. Available: https://journals.asm.org/doi/10.1128/msystems.00115-19.

[36] R. H. Mills, Y. Vázquez-Baeza, Q. Zhu, *et al.*, "Evaluating metagenomic prediction of the metaproteome in a 4.5-year study of a patient with crohn's disease," *mSystems*, vol. 4, no. 1, e00337–18, 2019, ISSN: 2379-5077. DOI: 10.1128/mSystems.00337-18.

[37] H. Zhong, H. Ren, Y. Lu, *et al.*, "Distinct gut metagenomics and metaproteomics signatures in prediabetics and treatment-naïve type 2 diabetics," *eBioMedicine*, vol. 47, pp. 373–383, Sep. 1, 2019, Publisher: Elsevier, ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2019.08.048. [Online]. Available: https://www.thelancet.com/article/S2352-3964(19)30572-9/fulltext.

[38] N. E. Diether and B. P. Willing, "Microbial fermentation of dietary protein: An important factor in diet–microbe–host interaction," *Microorganisms*, vol. 7, no. 1, p. 19, Jan. 2019, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-2607. DOI: 10.3390/microorganisms7010019. [Online]. Available: https://www.mdpi.com/2076-2607/7/1/19.

[39] Z. Liu, A. Ma, E. Mathé, M. Merling, Q. Ma, and B. Liu, "Network analyses in microbiome based on high-throughput multi-omics data," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1639–1655, Mar. 1, 2021, ISSN: 1477-4054. DOI: 10.1093/bib/bbaa005. [Online]. Available: https://doi.org/10.1093/bib/bbaa005.

[40] J. Chong and J. Xia, "Computational approaches for integrative analysis of the metabolome and microbiome," *Metabolites*, vol. 7, no. 4, p. 62, Dec. 2017, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2218-1989. DOI: 10.3390/metabo7040062. [Online]. Available: https://www.mdpi.com/2218-1989/7/4/62.

[41] A. Bauermeister, H. Mannochio-Russo, L. V. Costa-Lotufo, A. K. Jarmusch, and P. C. Dorrestein, "Mass spectrometry-based metabolomics in microbiome investigations," *Nature reviews. Microbiology*, vol. 20, no. 3, pp. 143–160, Mar. 2022, ISSN: 1740-1526. DOI: 10.1038/s41579-021-00621-9. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9578303/.

[42] X. Wang, S. Yang, S. Li, *et al.*, "Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents," *Gut*, vol. 69, no. 12, pp. 2131–2142, Dec. 2020, ISSN: 1468-3288. DOI: 10.1136/gutjnl-2019-319766.

[43] S. Yachida, S. Mizutani, H. Shiroma, *et al.*, "Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer," *Nature Medicine*, vol. 25, no. 6, pp. 968–976, Jun. 2019, Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: 10.1038/s41591-019-0458-7. [Online]. Available: https://www.nature.com/articles/s41591-019-0458-7 (visited on 06/01/2024).

[44] E. Muller, Y. M. Algavi, and E. Borenstein, "The gut microbiome-metabolome dataset collection: A curated resource for integrative meta-analysis," *npj Biofilms and Microbiomes*, vol. 8, no. 1, pp. 1–7, Oct. 15, 2022, Publisher: Nature Publishing Group, ISSN: 2055-5008. DOI: 10.1038/s41522-022-00345-5. [Online]. Available: https://www.nature.com/articles/s41522-022-00345-5 (visited on 05/18/2024).

[45] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, *et al.*, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, May 2019, Number: 7758 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1237-9. [Online]. Available: https://www.nature.com/articles/s41586-019-1237-9.

[46] X. Zhou and M. Stephens, "Efficient multivariate linear mixed model algorithms for genome-wide association studies," *Nature Methods*, vol. 11, no. 4, pp. 407–409, Apr. 2014, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/nmeth.2848. [Online]. Available: https://www.nature.com/articles/nmeth.2848 (visited on 06/06/2024).

[47] I. Zwiener, B. Frisch, and H. Binder, "Transforming RNA-seq data to improve the performance of prognostic gene signatures," *PLOS ONE*, vol. 9, no. 1, e85150, Jan. 8, 2014, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0085150. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085150 (visited on 06/06/2024).

[48] T. Lehmann, K. Schallert, R. Vilchez-Vargas, *et al.*, "Metaproteomics of fecal samples of crohn's disease and ulcerative colitis," *Journal of Proteomics*, vol. 201, pp. 93–103, Jun. 15, 2019, ISSN: 1874-3919. DOI: 10.1016/j.jprot.2019.04.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874391919301137 (visited on 06/06/2024).

[49] H. Sokol, V. Lalande, C. Landman, *et al.*, "*Clostridium difficile* infection in acute flares of inflammatory bowel disease: A prospective study," *Digestive and Liver Disease*, vol. 49, no. 6, pp. 643–646, Jun. 1, 2017, ISSN: 1590-8658. DOI: 10.1016/j.dld.2017.01.162. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1590865817301895 (visited on 06/06/2024).

[50] M. Desai, K. Knight, J. M. Gray, V. Nguyen, J. Boone, and D. Sorrentino, "Low glutamate dehydrogenase levels are associated with colonization in clostridium difficile PCR-only positive patients with inflammatory bowel disease," *European Journal of Gastroenterology & Hepatology*, vol. 32, no. 9, p. 1099, Sep. 2020, ISSN: 0954-691X. DOI: 10.1097/MEG.0000000000001762. [Online]. Available: https://journals.lww.com/eurojgh/abstract/2020/09000/low_glutamate_dehydrogenase_levels_are_associated.4.aspx (visited on 06/06/2024).

[51] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *WIREs Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, Sep. 2015, ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1157. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1157 (visited on 05/24/2024).

[52] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review," *International Journal of Automation and Computing*, vol. 14, no. 5, pp. 503–519, Oct. 2017, ISSN: 1476-8186, 1751-8520. DOI: 10.1007/s11633-017-1054-2. [Online]. Available: http://link.springer.com/10.1007/s11633-017-1054-2 (visited on 05/24/2024).

[53] E. Gordon-Rodriguez, T. Quinn, and J. P. Cunningham, "Data augmentation for compositional data: Advancing predictive models of the microbiome," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 551–20 565, 2022.

[54] M. T. Hira, M. A. Razzaque, C. Angione, J. Scrivens, S. Sawan, and M. Sarker, "Integrated multi-omics analysis of ovarian cancer using variational autoencoders," *Scientific Reports*, vol. 11, no. 1, p. 6265, Mar. 18, 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-021-85285-4. [Online]. Available: https://www.nature.com/articles/s41598-021-85285-4 (visited on 02/16/2024).

A. Supplementary methods

A.1. Deep fully-connected neural network

We trained 36 deep neural network (Deep NN) architectures, with various numbers of hidden layers, different data augmentation factors and loss functions. In the end, the architecture that was used throughout our experiments included 3 fully connected hidden layers and a loss function including equal proportions of Pearson's correlation and the mean-squared error (MSE). Additionally, training was performed without any data augmentation. For a comprehensive benchmark of different architectures and augmentation factors, see Supplementary Table S9.

A.1.1. Data augmentation

We augmented the paired cross-omics datasets using an approach inspired by the Aitchison mixup described by Gordon-Rodriguez, Quinn, and Cunningham [53]. However, one important distinction is that the authors describe augmentation of compositional data in the simplex, before transformations are applied, but we apply augmentation on the transformed data. Data was transformed using the quantile transformation, as described in Section 3.2.1.

Let x_i and $x_j \in \mathbb{R}^D$ be two input samples, with corresponding output samples y_i and $y_j \in \mathbb{R}^M$. We construct augmented data points x' and y' using a linear combination:

$$x' = \lambda \cdot x_i + (1 - \lambda) \cdot x_j,$$
$$y' = \lambda \cdot y_i + (1 - \lambda) \cdot y_j,$$

where $\lambda \in [0, 1]$ and $i, j \in \{1, ..., N\}$. To generate multiple data points, i and j are chosen randomly, and λ is sampled from a uniform distribution.

A.1.2. Architecture

Layer dimensions were chosen based on input size. To that end, we constructed architectures of the form: `input_size` − ⌊`1.25` · `input_size`⌋ − ... − ⌊`1.25` · `input_size`⌋ − ⌊`2.5` · `input_size`⌋ − `output_size`. Layer norm and ReLU were applied after each layer, excluding the output layer.

A.1.3. Loss function

We defined a loss function based on a combination of Pearson's correlation and the mean squared error, between the ground truth $Y \in \mathbb{R}^{N \times M}$ and the prediction $\hat{Y} \in \mathbb{R}^{N \times M}$:

$$L(Y, \hat{Y}) = \alpha_{MSE} \cdot MSE(Y, \hat{Y}) + \alpha_{corr} \cdot (1 - \rho(Y, \hat{Y})), \tag{4}$$

where $\rho(Y, \hat{Y})$ represents the average Pearson correlation coefficient between ground-truth and predicted features, $MSE(Y, \hat{Y})$ is the mean squared error between the ground truth and the prediction, and α_{MSE} and $\alpha_{corr} \in [0, 1]$.

To compute the mean squared error, we used `torch`'s (2.1.2.post300) MSE loss, from the `nn.functional` module, with "mean" reduction. To determine Pearson's correlation coefficient, we applied the CosineEmbeddingLoss from the same module. Prior to this, each feature was centered around its mean, and the data batch was transposed.

A.1.4. Training procedure

All network models (Supplementary Table S9) were constructed using the Pytorch Lightning API, version 2.2.1, with a random seed equal to 42. Training and validation sets were split based on study participants, as described in Section 3, using a shuffled split, with a random seed equal to 42. We used a batch size of 16 and the Adam optimizer, with a learning rate equal to `1e-4`, a patience of 3 for early stopping, and a maximum of 35 epochs.

A.2. Multi-omics autoencoder

As an alternative to naive feature concatenation, we trained an autoencoder model for multi-omics integration. A diagram of the architecture and loss function is included in Supplementary Figure S3. After training the network, we used the latent features to train the best-performing model in our benchmark, MelonnPan [16]. Below we provide details on the network architecture and the loss function used during training.

A.2.1. Architecture

As shown in Supplementary Figure S3, we divided the model architecture into two main parts: the autoencoder and a multi-layer perceptron, which takes as input the latent features of the autoencoder, and then predicts a meta-omics output. The autoencoder was organized using a symmetric architecture, with a hidden layer of dimension $\lfloor 0.75 \cdot \texttt{input\_size} \rfloor$, and a latent space of dimension $\lfloor 0.5 \cdot \texttt{input\_size} \rfloor$. The multi-layer perceptron included one hidden layer, of dimension $\lfloor 0.25 \cdot \texttt{input\_size} \rfloor$. Following a similar approach as described previously (Supplementary Section A.1.3), we applied layer norm and a ReLU activation after each layer, excluding the output layers.

A.2.2. Loss function

We constructed an architecture to enable learning of embeddings for the task of cross-omics prediction. To that end, we trained an autoencoder wth a combined loss, integrating the reconstruction loss with a regression loss. This was inspired by the approach described by Hira, Razzaque, Angione, *et al.* [54], who jointly trained a variational autoencoder and classifier for ovarian cancer, using a combined loss.

Let $X \in \mathbb{R}^{N \times D}$ be the input multi-meta-omics feature matrix, and let $Y \in \mathbb{R}^{N \times M}$ be the output meta-omics feature matrix. In addition, let $\hat{X} \in \mathbb{R}^{N \times D}$ be the prediction produced by the autoencoder, and let $\hat{Y} \in \mathbb{R}^{N \times M}$ be the prediction produced by the multi-layer perceptron. We computed the following loss:

$$L(X, \hat{X}, Y, \hat{Y}) = L(X, \hat{X}) + L(Y, \hat{Y}), \tag{5}$$

where $L(X, \hat{X})$ and $L(Y, \hat{Y})$ are defined as in equation 4.

A.2.3. Training procedure

We followed the same procedure as described in Supplementary Section A.1.4, with the exception that the maximum number of epochs was set to 50.

A.3. Feature selection

We designed a pre-training step to select a small set features to be used later during model training (Supplementary Figure S4). To that end, we split each training set into 10 training/validation partitions, and trained a random forest regressors on each partition. Feature correlations were subsequently calculated on the validation set, and each feature was assigned a score, equal to the mean across validation sets. Based on these scores, we retained a fraction of features to be later used for cross-omics model training.

B. Supplementary tables

Table S1: Description of paired metagenomics and metabolomics datasets used for experimental validation and reproducibility testing.

				Lenient filtering		Strict Filtering		No filtering	
Dataset name	**Study**	**# train samples**	**# test samples**	**Input dim.**	**Output dim.**	**Input dim.**	**Output dim.**	**Input dim.**	**Output dim.**
CD, UC and HC	Franzosa et al. [12]	102	60	1231	4541	722	1415	2113	8848
ESRD and HC	Wang et al. [42]	106	28	1401	274	795	94	56961	276
Cancer and HC	Yachida et al. [43]	202	52	1462	269	698	76	57702	450

Table S2: Datasets downloaded from the Inflammatory Bowel Disease Multi'omics Database (IBDMB), used for training and testing machine learning models in our experiments.

Data description	Download link	Download date
Metaproteomics ECs	https://www.ibdmdb.org/downloads/products/HMP2/MPX/2017-03-20/HMP2_proteomics_ecs.tsv.gz	16.11.2023
Metagenomics ECs	https://www.ibdmdb.org/downloads/products/HMP2/MGX/2018-05-04/ecs_relab.tsv.gz	16.11.2023
Metabolomics	https://www.ibdmdb.org/downloads/products/HMP2/MBX/HMP2_metabolomics.biom	16.11.2023
Metatranscriptomics ECs	https://www.ibdmdb.org/downloads/products/HMP2/MTX/2017-12-14/ecs_relab.tsv.gz	16.11.2023
Taxonomic profile	https://www.ibdmdb.org/downloads/products/HMP2/MGX/2018-05-04/taxonomic_profiles.tsv.gz	16.11.2023
Patient metadata	https://ibdmdb.org/downloads/metadata/hmp2_metadata_2018-08-20.csv	17.01.2024

Table S3: Metadata for the paired-omics datasets used in our experiments. Input-output pairs used for the results of the main manuscript are highlighted. Results for the other datasets are only reported as part of the supplementary material.

			Input/output dim. after filtering						Train/test # samples per seed																			
Output	Input	# samples	Non-imputed data		Imputed data		Train/test mean # samples		2		3		5		7		11		13		17		23		29		31	
mTx	**mGx**	**733**	**983**	**851**	**992**	**851**	**584,5**	**148,5**	**577**	**156**	**587**	**146**	**619**	**114**	**600**	**133**	**569**	**164**	**589**	**144**	**570**	**163**	**582**	**151**	**597**	**136**	**555**	**178**
mTx	mGx_taxa	731	130	850	130	850	582,8	148,2	575	156	586	145	618	113	599	132	567	164	587	144	568	163	580	151	595	136	553	178
mTx	mGx_pa	732	323	850	323	851	583,6	148,4	576	156	587	145	618	114	599	133	568	164	588	144	569	163	581	151	596	136	554	178
mTx	mGx+mGx_taxa	731	1113	850	1117	850	582,8	148,2	575	156	586	145	618	113	599	132	567	164	587	144	568	163	580	151	595	136	553	178
mTx	mGx+mGx_pa	732	1306	850	1315	851	583,6	148,4	576	156	587	145	618	114	599	133	568	164	588	144	569	163	581	151	596	136	554	178
mTx	mGx_pa+mGx_taxa	731	453	850	453	850	582,8	148,2	575	156	586	145	618	113	599	132	567	164	587	144	568	163	580	151	595	136	553	178
mPx	**mGx**	**280**	**1000**	**281**	**1004**	**909**	**222,2**	**57,8**	**217**	**63**	**223**	**57**	**218**	**62**	**221**	**59**	**223**	**57**	**223**	**57**	**222**	**58**	**226**	**54**	**223**	**57**	**226**	**54**
mPx	mGx_taxa	278	135	281	135	909	220,4	57,6	215	63	222	56	216	62	220	58	221	57	221	57	220	58	224	54	221	57	224	54
mPx	mGx_taxa+mTx	185	1009	274	1010	909	146,7	38,3	150	35	146	39	150	35	136	49	145	40	147	38	145	40	153	32	150	35	145	40
mPx	**mTx**	**186**	**874**	**274**	**875**	**909**	**147,6**	**38,4**	**151**	**35**	**146**	**40**	**151**	**35**	**137**	**49**	**146**	**40**	**148**	**38**	**146**	**40**	**154**	**32**	**151**	**35**	**146**	**40**
mPx	mGx+mGx_pa	278	1339	281	1329	909	220,4	57,6	215	63	222	56	216	62	220	58	221	57	221	57	220	58	224	54	221	57	224	54
mPx	mGx_pa	278	323	281	325	909	220,4	57,6	215	63	222	56	216	62	220	58	221	57	221	57	220	58	224	54	221	57	224	54
mPx	mGx+mGx_taxa	278	1151	281	1154	909	220,4	57,6	215	63	222	56	216	62	220	58	221	57	221	57	220	58	224	54	221	57	224	54
mPx	**mGx+mTx**	**186**	**1902**	**274**	**1903**	**909**	**147,6**	**38,4**	**151**	**35**	**146**	**40**	**151**	**35**	**137**	**49**	**146**	**40**	**148**	**38**	**146**	**40**	**154**	**32**	**151**	**35**	**146**	**40**
mPx	mGx_pa+mGx_taxa	278	458	281	458	909	220,4	57,6	215	63	222	56	216	62	220	58	221	57	221	57	220	58	224	54	221	57	224	54
mPx	mGx_pa+mTx	185	1194	274	1201	909	146,7	38,3	150	35	146	39	150	35	136	49	145	40	147	38	145	40	153	32	150	35	145	40
mBx	**mGx**	**388**	**1038**	**3247**	**1040**	**3247**	**311,7**	**76,3**	**314**	**74**	**315**	**73**	**311**	**77**	**316**	**72**	**309**	**79**	**316**	**72**	**321**	**67**	**307**	**81**	**307**	**81**	**301**	**87**
mBx	**mGx+mPx**	**203**	**1277**	**3345**	**1912**	**3345**	**161,3**	**41,7**	**161**	**42**	**163**	**40**	**160**	**43**	**158**	**45**	**162**	**41**	**164**	**39**	**167**	**36**	**157**	**46**	**162**	**41**	**159**	**44**
mBx	**mGx+mPx+mTx**	**154**	**2196**	**3373**	**2830**	**3373**	**121,2**	**32,8**	**124**	**30**	**118**	**36**	**116**	**38**	**118**	**36**	**121**	**33**	**123**	**31**	**116**	**38**	**128**	**26**	**125**	**29**	**123**	**31**
mBx	mGx_taxa	386	129	3240	129	3240	310,1	75,9	313	73	315	71	309	77	314	72	307	79	314	72	319	67	305	81	306	80	299	87
mBx	mGx_taxa+mTx	298	1002	3246	1002	3246	238,2	59,8	240	58	239	59	243	55	239	59	243	55	240	58	248	50	235	63	231	67	224	74
mBx	**mTx**	**299**	**870**	**3252**	**871**	**3252**	**239,1**	**59,9**	**241**	**58**	**239**	**60**	**244**	**55**	**240**	**59**	**244**	**55**	**241**	**58**	**249**	**50**	**236**	**63**	**232**	**67**	**225**	**74**
mBx	mGx+mGx_pa	386	1360	3240	1363	3247	310,1	75,9	313	73	315	71	309	77	314	72	307	79	314	72	319	67	305	81	306	80	299	87
mBx	mGx_pa	386	322	3240	323	3247	310,1	75,9	313	73	315	71	309	77	314	72	307	79	314	72	319	67	305	81	306	80	299	87
mBx	mGx+mGx_taxa	386	1167	3240	1169	3240	310,1	75,9	313	73	315	71	309	77	314	72	307	79	314	72	319	67	305	81	306	80	299	87
mBx	**mPx+mTx**	**154**	**1158**	**3373**	**1792**	**3373**	**121,2**	**32,8**	**124**	**30**	**118**	**36**	**116**	**38**	**118**	**36**	**121**	**33**	**123**	**31**	**116**	**38**	**128**	**26**	**125**	**29**	**123**	**31**
mBx	**mGx+mTx**	**299**	**1922**	**3252**	**1925**	**3252**	**239,1**	**59,9**	**241**	**58**	**239**	**60**	**244**	**55**	**240**	**59**	**244**	**55**	**241**	**58**	**249**	**50**	**236**	**63**	**232**	**67**	**225**	**74**
mBx	mGx_pa+mPx	202	602	3339	1233	3345	160,4	41,6	160	42	163	39	159	43	157	45	161	41	163	39	166	36	156	46	161	41	158	44
mBx	mGx_pa+mGx_taxa	386	451	3240	451	3240	310,1	75,9	313	73	315	71	309	77	314	72	307	79	314	72	319	67	305	81	306	80	299	87
mBx	**mPx**	**307**	**282**	**3298**	**909**	**3298**	**242,4**	**64,6**	**242**	**65**	**241**	**66**	**249**	**58**	**241**	**66**	**239**	**68**	**247**	**60**	**238**	**69**	**243**	**64**	**242**	**65**	**242**	**65**
mBx	mGx_pa+mTx	298	1190	3246	1194	3252	238,2	59,8	240	58	239	59	243	55	239	59	243	55	240	58	248	50	235	63	231	67	224	74
mBx	mGx_taxa+mPx	202	410	3339	1041	3339	160,4	41,6	160	42	163	39	159	43	157	45	161	41	163	39	166	36	156	46	161	41	158	44

Table S4: Metadata for the full meta-omics datasets used for IBD classification, before downsampling.

		Dimensionality after filtering				Train/test # samples per seed																			
Data type	**# samples**	**Non-imputed data**	**Imputed data**	**Train/test mean # samples**		**2**		**3**		**5**		**7**		**11**		**13**		**17**		**23**		**29**		**31**	
mTx	735	851	851	586,1	148,9	578	157	589	146	620	115	602	133	570	165	590	145	572	163	584	151	599	136	557	178
mPx	449	283	909	358	91	350	99	356	93	359	90	360	89	356	93	358	91	361	88	364	85	353	96	363	86
mBx	546	3183	3183	433,5	112,5	439	107	430	116	434	112	433	113	438	108	430	116	435	111	430	116	434	112	432	114

Table S5: Comparison of data processing methods for three "metagenomics-to-metabolomics" models: MelonnPan [16], MiMeNet [18] and SparseNED [17]. We use the word "default" to refer to the data processing approach applied internally by the model, on normalized data. Some experiments were not performed. For example, MelonnPan already uses an arcsine and quantile transformation in its default pipeline, so we omitted that comparison. For each model, we highlight the data processing approach selected to report results for the model. Performance was measured using the average Spearman's rank correlation coefficient for the 50 best predicted features.

	Predict mTx	**Predict mPx**		**Predict mBx**		
	From mGx	**From mGx**	**From mTx**	**From mGx**	**From mPx**	**From mTx**
MelonnPan						
Normalized	0.72 (± 0.07)	0.35 (± 0.08)	0.29 (± 0.06)	0.64 (± 0.07)	0.49 (± 0.06)	0.61 (± 0.10)
ArcSin	-	-	-	-	-	-
CLR	0.76 (± 0.05)	**0.42 (± 0.09)**	**0.47 (± 0.07)**	0.72 (± 0.06)	**0.58 (± 0.08)**	**0.72 (± 0.08)**
Quantile transform	-	-	-	-	-	-
Model default	**0.77 (± 0.05)**	0.40 (± 0.11)	0.40 (± 0.11)	**0.74 (± 0.05)**	0.57 (± 0.09)	**0.72 (± 0.08)**
SparseNED						
Normalized	0.53 (± 0.12)	0.30 (± 0.12)	0.26 (± 0.13)	0.52 (± 0.13)	0.40 (± 0.13)	0.50 (± 0.13)
ArcSin	0.64 (± 0.09)	**0.34 (± 0.12)**	0.30 (± 0.13)	0.65 (± 0.07)	0.51 (± 0.09)	0.59 (± 0.11)
CLR	0.64 (± 0.10)	0.32 (± 0.12)	0.31 (± 0.12)	0.54 (± 0.12)	0.45 (± 0.07)	0.61 (± 0.10)
Quantile transform	**0.66 (± 0.09)**	0.31 (± 0.13)	**0.32 (± 0.14)**	**0.68 (± 0.07)**	**0.53 (± 0.07)**	**0.62 (± 0.11)**
Model default	-	-	-	-	-	-
MiMeNet						
Normalized	**0.24 (± 0.11)**	0.20 (± 0.09)	0.23 (± 0.08)	0.29 (± 0.09)	0.27 (± 0.08)	0.28 (± 0.09)
ArcSin	**0.24 (± 0.10)**	0.21 (± 0.10)	0.24 (± 0.11)	**0.30 (± 0.11)**	0.25 (± 0.10)	**0.31 (± 0.12)**
CLR	-	-	-	-	-	-
Quantile transform	0.23 (± 0.10)	0.19 (± 0.10)	0.22 (± 0.10)	0.28 (± 0.10)	**0.32 (± 0.11)**	**0.31 (± 0.10)**
Model default	0.23 (± 0.08)	**0.22 (± 0.12)**	**0.26 (± 0.10)**	0.28 (± 0.09)	0.31 (± 0.11)	0.30 (± 0.11)

Table S6: Grid values for hyperparameter tuning of random forest classifiers for IBD prediction.

Parameter name	**Parameter values**
n_estimators	64, 128, 256, 512, 1024
max_depth	16, 32, 64, 128, None
min_samples_split	2, 4, 8
min_samples_leaf	1, 2, 4
random_state	[equal to train/test partition seed]

Table S7: Average Spearman's rank correlation coefficient of MelonnPan [16] predictions (top 50) for multiple single-omics and multi-omics input data types, including pathways (mGx_pa) and taxonomic profiles (mGx_taxa). The best results for each output type are highlighted.

	Output type		
Input type	mTx	mPx	mBx
mGx	0.77 (± 0.05)	0.40 (± 0.11)	0.74 (± 0.05)
mGx+mGx_pa	0.77 (± 0.05)	0.41 (± 0.10)	0.74 (± 0.05)
mGx+mGx_taxa	**0.78 (± 0.05)**	**0.42 (± 0.10)**	**0.75 (± 0.05)**
mGx+mPx	-	-	0.67 (± 0.10)
mGx+mPx+mTx	-	-	0.70 (± 0.11)
mGx+mTx	-	**0.42 (± 0.12)**	0.74 (± 0.07)
mGx_pa	0.72 (± 0.06)	0.40 (± 0.10)	0.71 (± 0.05)
mGx_pa+mGx_taxa	0.76 (± 0.06)	0.41 (± 0.11)	0.74 (± 0.05)
mGx_pa+mPx	-	-	0.66 (± 0.10)
mGx_pa+mTx	-	0.41 (± 0.12)	0.74 (± 0.07)
mGx_taxa	0.75 (± 0.06)	0.38 (± 0.11)	0.72 (± 0.06)
mGx_taxa+mPx	-	-	0.66 (± 0.09)
mGx_taxa+mTx	-	0.41 (± 0.11)	**0.75 (± 0.07)**
mPx	-	-	0.57 (± 0.09)
mPx+mTx	-	-	0.69 (± 0.09)
mTx	-	0.40 (± 0.11)	0.72 (± 0.08)

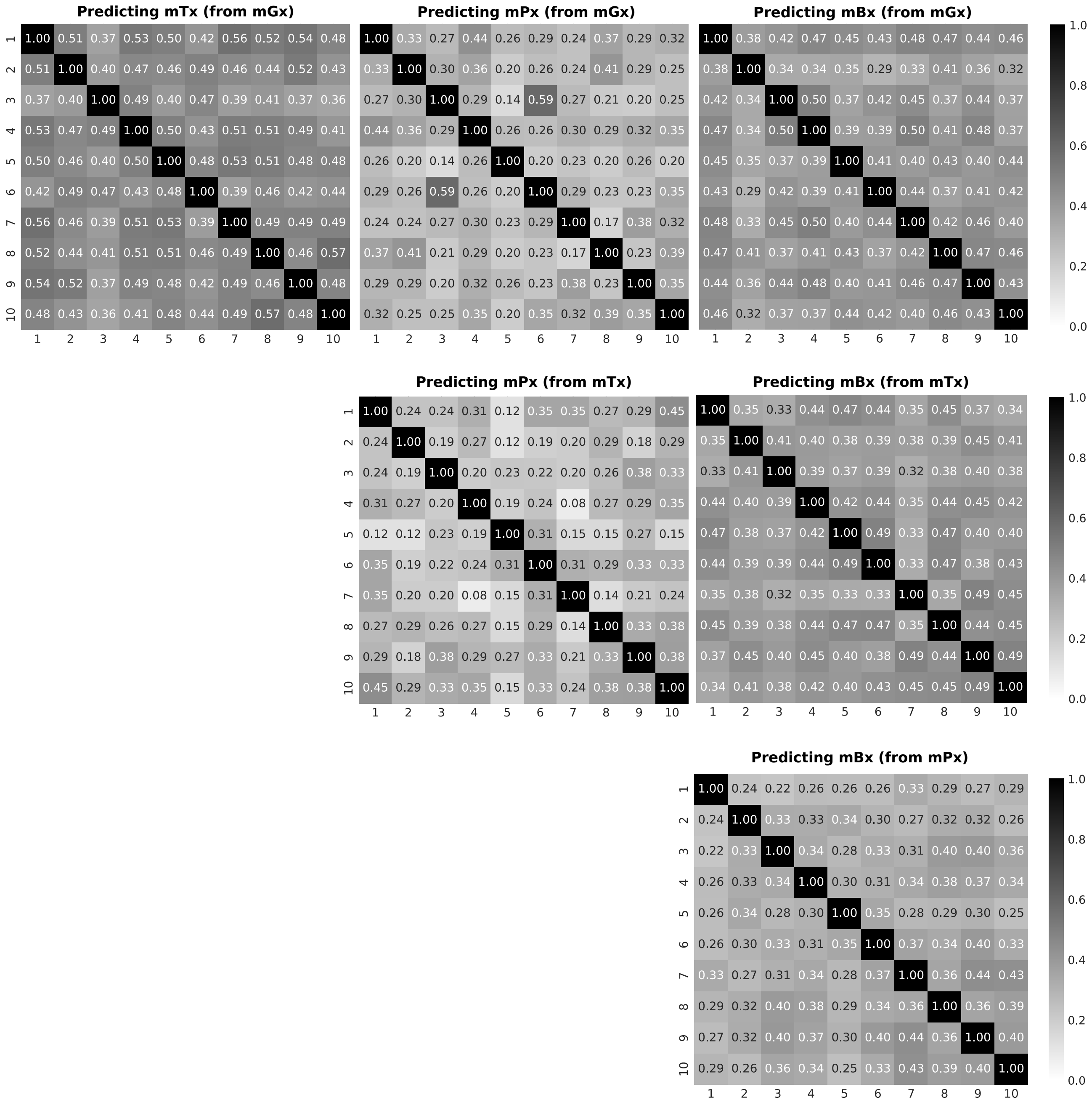
Table S8: Average Spearman's rank correlation coefficient of MelonnPan [16] predictions (top 50) for multi-omics input, comparing the model trained on a latent space, using the autoencoder in section A.2, to the model trained on naively concatenated multi-omics.

Prediction task	MelonnPan trained on latent features	MelonnPan trained on concatenated multi-omics
mGx+mTx -> mPx	0.16 (± 0.01)	0.42 (± 0.12)
mGx+mPx -> mBx	0.35 (± 0.01)	0.67 (± 0.10)
mGx+mPx+mTx -> mBx	0.40 (± 0.00)	0.70 (± 0.11)
mGx+mTx -> mBx	0.34 (± 0.01)	0.74 (± 0.07)
mPx+mTx -> mBx	0.40 (± 0.00)	0.69 (± 0.09)

Table S9: Spearman's rank correlation coefficient of the 50 best predicted features, for multiple combinations of network hyperparameters and input-output combinations, for a deep neural network model. Performance was computed on a validation set, separate from the test sets described in Section 3.2. An augmentation factor equal to 1 indicates that no augmentation was applied, while an augmentation factor equal to $n > 1$ indicates that the final number of data points is equal to the initial size of the dataset multiplied by n. The best result for each input-output combination is highlighted.

	Predict mTx	**Predict mPx**		**Predict mBx**		
Model architecture	**From mGx**	**From mGx**	**From mTx**	**From mGx**	**From mPx**	**From mTx**
Num. layers: 3 Aug. factor: 1 Loss: corr	0.62 (± 0.07)	0.32 (± 0.12)	0.32 (± 0.15)	0.61 (± 0.08)	0.46 (± 0.11)	0.59 (± 0.10)
Num. layers: 3 Aug. factor: 4 Loss: corr	0.63 (± 0.07)	0.30 (± 0.13)	0.32 (± 0.14)	0.61 (± 0.08)	0.47 (± 0.12)	0.59 (± 0.09)
Num. layers: 3 Aug. factor: 16 Loss: corr	0.62 (± 0.07)	0.32 (± 0.12)	0.35 (± 0.14)	0.59 (± 0.08)	0.45 (± 0.13)	0.56 (± 0.10)
Num. layers: 3 Aug. factor: 32 Loss: corr	0.64 (± 0.07)	0.32 (± 0.12)	0.36 (± 0.14)	0.61 (± 0.09)	0.40 (± 0.11)	0.59 (± 0.09)
Num. layers: 3 Aug. factor: 1 Loss: mse + corr	**0.67 (± 0.06)**	**0.40 (± 0.11)**	**0.43 (± 0.15)**	**0.68 (± 0.07)**	0.53 (± 0.10)	0.65 (± 0.06)
Num. layers: 3 Aug. factor: 4 Loss: mse + corr	**0.67 (± 0.06)**	0.35 (± 0.12)	0.40 (± 0.15)	0.66 (± 0.07)	0.50 (± 0.11)	0.65 (± 0.07)
Num. layers: 3 Aug. factor: 16 Loss: mse + corr	0.64 (± 0.06)	0.32 (± 0.12)	0.37 (± 0.14)	0.62 (± 0.08)	0.48 (± 0.10)	0.60 (± 0.09)
Num. layers: 3 Aug. factor: 32 Loss: mse + corr	0.62 (± 0.07)	0.32 (± 0.13)	0.37 (± 0.14)	0.62 (± 0.07)	0.44 (± 0.10)	0.61 (± 0.09)
Num. layers: 3 Aug. factor: 1 Loss: mse	**0.67 (± 0.06)**	**0.40 (± 0.12)**	0.42 (± 0.15)	**0.68 (± 0.07)**	0.53 (± 0.10)	0.67 (± 0.08)
Num. layers: 3 Aug. factor: 4 Loss: mse	**0.67 (± 0.06)**	0.36 (± 0.12)	0.40 (± 0.15)	0.67 (± 0.07)	0.53 (± 0.10)	0.66 (± 0.07)
Num. layers: 3 Aug. factor: 16 Loss: mse	0.64 (± 0.06)	0.33 (± 0.12)	0.37 (± 0.14)	0.61 (± 0.08)	0.48 (± 0.12)	0.60 (± 0.08)
Num. layers: 3 Aug. factor: 32 Loss: mse	0.63 (± 0.06)	0.33 (± 0.12)	0.37 (± 0.14)	0.61 (± 0.09)	0.44 (± 0.11)	0.61 (± 0.09)
Num. layers: 4 Aug. factor: 1 Loss: corr	0.62 (± 0.07)	0.31 (± 0.12)	0.31 (± 0.14)	0.62 (± 0.08)	0.44 (± 0.12)	0.59 (± 0.09)
Num. layers: 4 Aug. factor: 4 Loss: corr	0.64 (± 0.07)	0.30 (± 0.12)	0.33 (± 0.14)	0.64 (± 0.08)	0.49 (± 0.11)	0.60 (± 0.08)
Num. layers: 4 Aug. factor: 16 Loss: corr	0.63 (± 0.07)	0.32 (± 0.12)	0.35 (± 0.14)	0.62 (± 0.08)	0.42 (± 0.11)	0.58 (± 0.09)
Num. layers: 4 Aug. factor: 32 Loss: corr	0.66 (± 0.06)	0.34 (± 0.12)	0.36 (± 0.14)	0.62 (± 0.08)	0.42 (± 0.11)	0.59 (± 0.09)
Num. layers: 4 Aug. factor: 1 Loss: mse + corr	0.66 (± 0.07)	0.36 (± 0.12)	0.40 (± 0.15)	0.65 (± 0.07)	0.53 (± 0.12)	**0.70 (± 0.06)**
Num. layers: 4 Aug. factor: 4 Loss: mse + corr	**0.67 (± 0.06)**	0.35 (± 0.11)	0.39 (± 0.16)	0.66 (± 0.07)	0.50 (± 0.11)	0.66 (± 0.07)
Num. layers: 4 Aug. factor: 16 Loss: mse + corr	0.65 (± 0.06)	0.34 (± 0.12)	0.39 (± 0.14)	0.63 (± 0.09)	0.47 (± 0.12)	0.60 (± 0.08)
Num. layers: 4 Aug. factor: 32 Loss: mse + corr	0.65 (± 0.06)	0.34 (± 0.12)	0.36 (± 0.14)	0.65 (± 0.07)	0.44 (± 0.11)	0.61 (± 0.08)
Num. layers: 4 Aug. factor: 1 Loss: mse	0.66 (± 0.07)	0.37 (± 0.12)	0.39 (± 0.15)	0.65 (± 0.09)	0.53 (± 0.10)	0.67 (± 0.07)
Num. layers: 4 Aug. factor: 4 Loss: mse	**0.67 (± 0.06)**	0.37 (± 0.13)	0.40 (± 0.14)	0.67 (± 0.07)	0.52 (± 0.12)	0.67 (± 0.06)
Num. layers: 4 Aug. factor: 16 Loss: mse	0.65 (± 0.06)	0.35 (± 0.12)	0.39 (± 0.14)	0.63 (± 0.08)	0.48 (± 0.11)	0.62 (± 0.08)
Num. layers: 4 Aug. factor: 32 Loss: mse	0.65 (± 0.06)	0.35 (± 0.12)	0.38 (± 0.14)	0.64 (± 0.08)	0.43 (± 0.11)	0.63 (± 0.08)
Num. layers: 5 Aug. factor: 1 Loss: corr	0.62 (± 0.08)	0.29 (± 0.13)	0.33 (± 0.15)	0.64 (± 0.07)	0.45 (± 0.12)	0.59 (± 0.08)
Num. layers: 5 Aug. factor: 4 Loss: corr	0.64 (± 0.07)	0.30 (± 0.13)	0.33 (± 0.15)	0.63 (± 0.08)	0.48 (± 0.12)	0.61 (± 0.08)
Num. layers: 5 Aug. factor: 16 Loss: corr	0.64 (± 0.07)	0.32 (± 0.12)	0.36 (± 0.14)	0.62 (± 0.08)	0.43 (± 0.11)	0.60 (± 0.09)
Num. layers: 5 Aug. factor: 32 Loss: corr	**0.67 (± 0.06)**	0.35 (± 0.13)	0.35 (± 0.14)	0.63 (± 0.09)	0.42 (± 0.12)	0.59 (± 0.09)
Num. layers: 5 Aug. factor: 1 Loss: mse + corr	0.66 (± 0.07)	0.36 (± 0.12)	0.42 (± 0.15)	**0.68 (± 0.08)**	0.53 (± 0.10)	0.65 (± 0.08)
Num. layers: 5 Aug. factor: 4 Loss: mse + corr	**0.67 (± 0.06)**	0.37 (± 0.13)	0.40 (± 0.15)	0.65 (± 0.07)	0.51 (± 0.10)	0.66 (± 0.06)
Num. layers: 5 Aug. factor: 16 Loss: mse + corr	0.66 (± 0.06)	0.34 (± 0.12)	0.37 (± 0.14)	0.63 (± 0.08)	0.47 (± 0.11)	0.63 (± 0.08)
Num. layers: 5 Aug. factor: 32 Loss: mse + corr	0.66 (± 0.06)	0.35 (± 0.12)	0.38 (± 0.14)	0.64 (± 0.08)	0.45 (± 0.11)	0.64 (± 0.08)
Num. layers: 5 Aug. factor: 1 Loss: mse	0.66 (± 0.07)	0.37 (± 0.12)	0.41 (± 0.15)	0.66 (± 0.07)	**0.55 (± 0.11)**	0.68 (± 0.08)
Num. layers: 5 Aug. factor: 4 Loss: mse	**0.67 (± 0.06)**	0.38 (± 0.11)	0.41 (± 0.14)	0.67 (± 0.07)	0.52 (± 0.11)	0.67 (± 0.07)
Num. layers: 5 Aug. factor: 16 Loss: mse	0.66 (± 0.06)	0.36 (± 0.12)	0.39 (± 0.15)	0.65 (± 0.08)	0.48 (± 0.11)	0.64 (± 0.07)
Num. layers: 5 Aug. factor: 32 Loss: mse	0.66 (± 0.06)	0.36 (± 0.12)	0.38 (± 0.14)	0.65 (± 0.07)	0.45 (± 0.11)	0.64 (± 0.07)

C. Supplementary figures



Predicting mTx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.51	0.37	0.53	0.50	0.42	0.56	0.52	0.54	0.48
2	0.51	1.00	0.40	0.47	0.46	0.49	0.46	0.44	0.52	0.43
3	0.37	0.40	1.00	0.49	0.40	0.47	0.39	0.41	0.37	0.36
4	0.53	0.47	0.49	1.00	0.50	0.43	0.51	0.51	0.49	0.41
5	0.50	0.46	0.40	0.50	1.00	0.48	0.53	0.51	0.48	0.48
6	0.42	0.49	0.47	0.43	0.48	1.00	0.39	0.46	0.42	0.44
7	0.56	0.46	0.39	0.51	0.53	0.39	1.00	0.49	0.49	0.49
8	0.52	0.44	0.41	0.51	0.51	0.46	0.49	1.00	0.46	0.57
9	0.54	0.52	0.37	0.49	0.48	0.42	0.49	0.46	1.00	0.48
10	0.48	0.43	0.36	0.41	0.48	0.44	0.49	0.57	0.48	1.00

Predicting mPx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.33	0.27	0.44	0.26	0.29	0.24	0.37	0.29	0.32
2	0.33	1.00	0.30	0.36	0.20	0.26	0.24	0.41	0.29	0.25
3	0.27	0.30	1.00	0.29	0.14	0.59	0.27	0.21	0.20	0.25
4	0.44	0.36	0.29	1.00	0.26	0.26	0.30	0.29	0.32	0.35
5	0.26	0.20	0.14	0.26	1.00	0.20	0.23	0.20	0.26	0.20
6	0.29	0.26	0.59	0.26	0.20	1.00	0.29	0.23	0.23	0.35
7	0.24	0.24	0.27	0.30	0.23	0.29	1.00	0.17	0.38	0.32
8	0.37	0.41	0.21	0.29	0.20	0.23	0.17	1.00	0.23	0.39
9	0.29	0.29	0.20	0.32	0.26	0.23	0.38	0.23	1.00	0.35
10	0.32	0.25	0.25	0.35	0.20	0.35	0.32	0.39	0.35	1.00

Predicting mBx (from mGx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.38	0.42	0.47	0.45	0.43	0.48	0.47	0.44	0.46
2	0.38	1.00	0.34	0.34	0.35	0.29	0.33	0.41	0.36	0.32
3	0.42	0.34	1.00	0.50	0.37	0.42	0.45	0.37	0.44	0.37
4	0.47	0.34	0.50	1.00	0.39	0.39	0.50	0.41	0.48	0.37
5	0.45	0.35	0.37	0.39	1.00	0.41	0.40	0.43	0.40	0.44
6	0.43	0.29	0.42	0.39	0.41	1.00	0.44	0.37	0.41	0.42
7	0.48	0.33	0.45	0.50	0.40	0.44	1.00	0.42	0.46	0.40
8	0.47	0.41	0.37	0.41	0.43	0.37	0.42	1.00	0.47	0.46
9	0.44	0.36	0.44	0.48	0.40	0.41	0.46	0.47	1.00	0.43
10	0.46	0.32	0.37	0.37	0.44	0.42	0.40	0.46	0.43	1.00

Predicting mPx (from mTx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.24	0.24	0.31	0.12	0.35	0.35	0.27	0.29	0.45
2	0.24	1.00	0.19	0.27	0.12	0.19	0.20	0.29	0.18	0.29
3	0.24	0.19	1.00	0.20	0.23	0.22	0.20	0.26	0.38	0.33
4	0.31	0.27	0.20	1.00	0.19	0.24	0.08	0.27	0.29	0.35
5	0.12	0.12	0.23	0.19	1.00	0.31	0.15	0.15	0.27	0.15
6	0.35	0.19	0.22	0.24	0.31	1.00	0.31	0.29	0.33	0.33
7	0.35	0.20	0.20	0.08	0.15	0.31	1.00	0.14	0.21	0.24
8	0.27	0.29	0.26	0.27	0.15	0.29	0.14	1.00	0.33	0.38
9	0.29	0.18	0.38	0.29	0.27	0.33	0.21	0.33	1.00	0.38
10	0.45	0.29	0.33	0.35	0.15	0.33	0.24	0.38	0.38	1.00

Predicting mBx (from mTx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.35	0.33	0.44	0.47	0.44	0.35	0.45	0.37	0.34
2	0.35	1.00	0.41	0.40	0.38	0.39	0.38	0.39	0.45	0.41
3	0.33	0.41	1.00	0.39	0.37	0.39	0.32	0.38	0.40	0.38
4	0.44	0.40	0.39	1.00	0.42	0.44	0.35	0.44	0.45	0.42
5	0.47	0.38	0.37	0.42	1.00	0.49	0.33	0.47	0.40	0.40
6	0.44	0.39	0.39	0.44	0.49	1.00	0.33	0.47	0.38	0.43
7	0.35	0.38	0.32	0.35	0.33	0.33	1.00	0.35	0.49	0.45
8	0.45	0.39	0.38	0.44	0.47	0.47	0.35	1.00	0.44	0.45
9	0.37	0.45	0.40	0.45	0.40	0.38	0.49	0.44	1.00	0.49
10	0.34	0.41	0.38	0.42	0.40	0.43	0.45	0.45	0.49	1.00

Predicting mBx (from mPx)

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.24	0.22	0.26	0.26	0.26	0.33	0.29	0.27	0.29
2	0.24	1.00	0.33	0.33	0.34	0.30	0.27	0.32	0.32	0.26
3	0.22	0.33	1.00	0.34	0.28	0.33	0.31	0.40	0.40	0.36
4	0.26	0.33	0.34	1.00	0.30	0.31	0.34	0.38	0.37	0.34
5	0.26	0.34	0.28	0.30	1.00	0.35	0.28	0.29	0.30	0.25
6	0.26	0.30	0.33	0.31	0.35	1.00	0.37	0.34	0.40	0.33
7	0.33	0.27	0.31	0.34	0.28	0.37	1.00	0.36	0.44	0.43
8	0.29	0.32	0.40	0.38	0.29	0.34	0.36	1.00	0.36	0.39
9	0.27	0.32	0.40	0.37	0.30	0.40	0.44	0.36	1.00	0.40
10	0.29	0.26	0.36	0.34	0.25	0.33	0.43	0.39	0.40	1.00

Figure S1: Jaccard similarities computed between sets of the top 25% well-predicted features across 10 train/test partitions. Predictions were generated with MelonnPan [16].

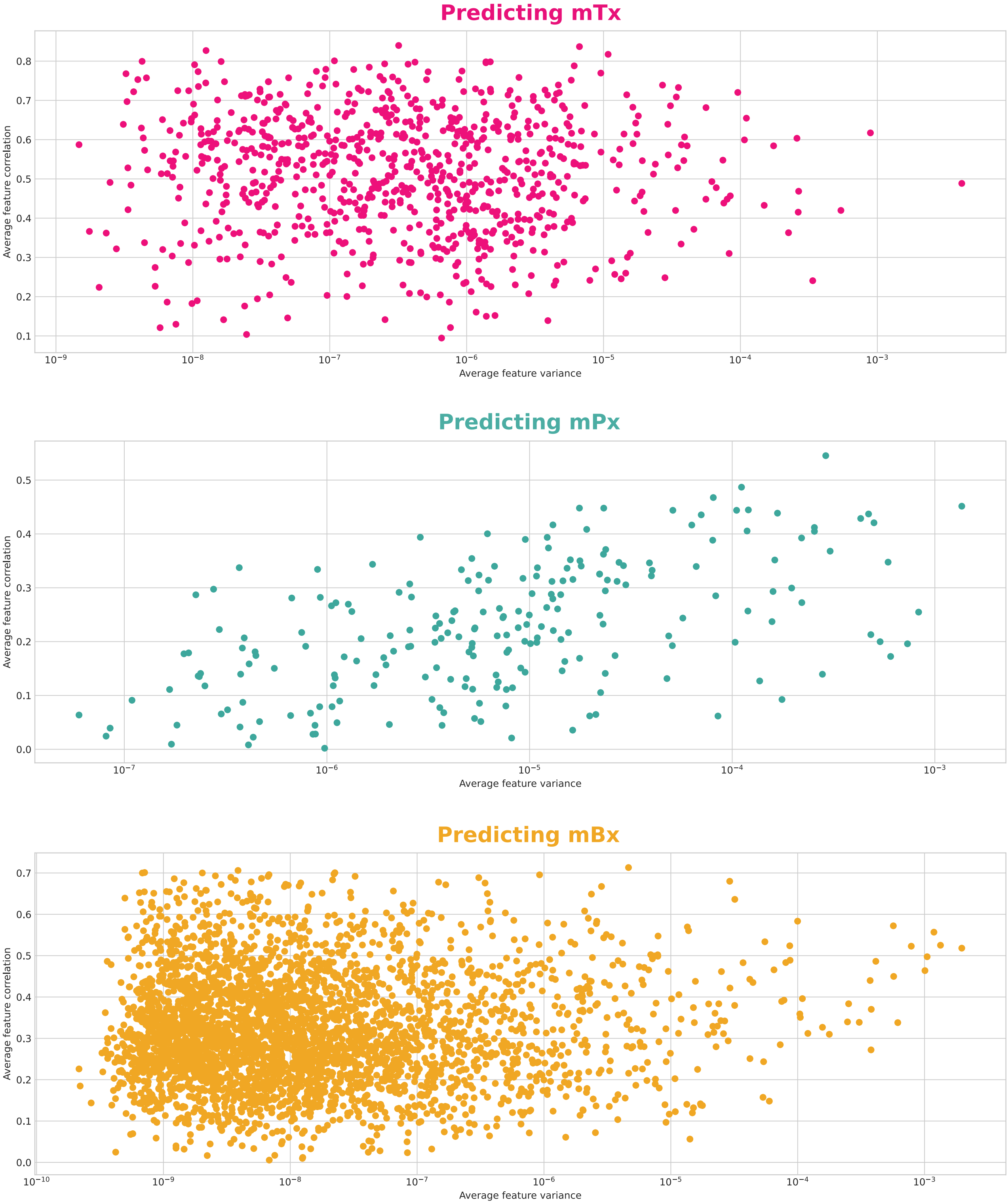


Figure S2: Feature variance plotted against feature correlation, for the three output data types. Correlation was computed between predicted features and the ground-truth. Variance and correlation were both computed on test sets, and averaged across dataset partitions and single- and multi-omics input types.

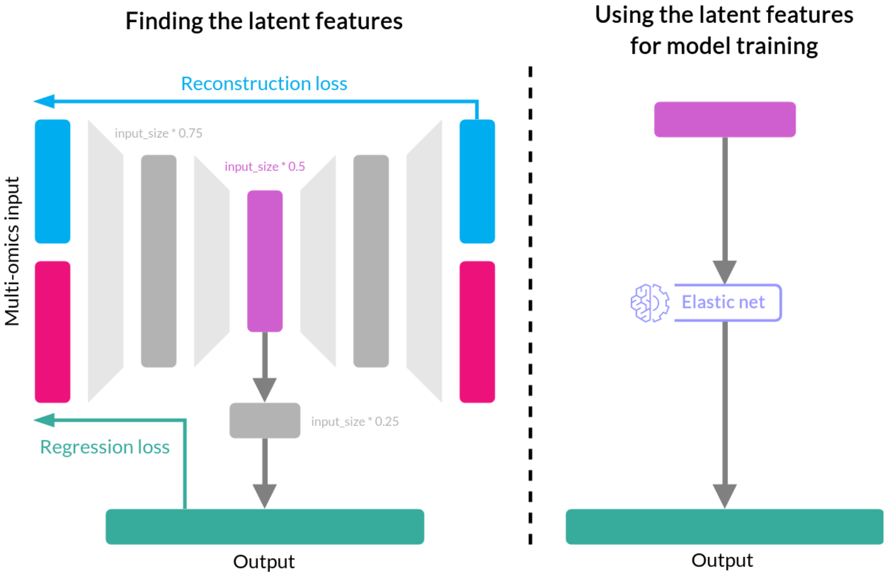


Figure S3: Training a multi-omics autoencoder (Supplementary Section A.2) with a combined loss, followed by training an elastic net model (MelonnPan [16]) on the latent features.

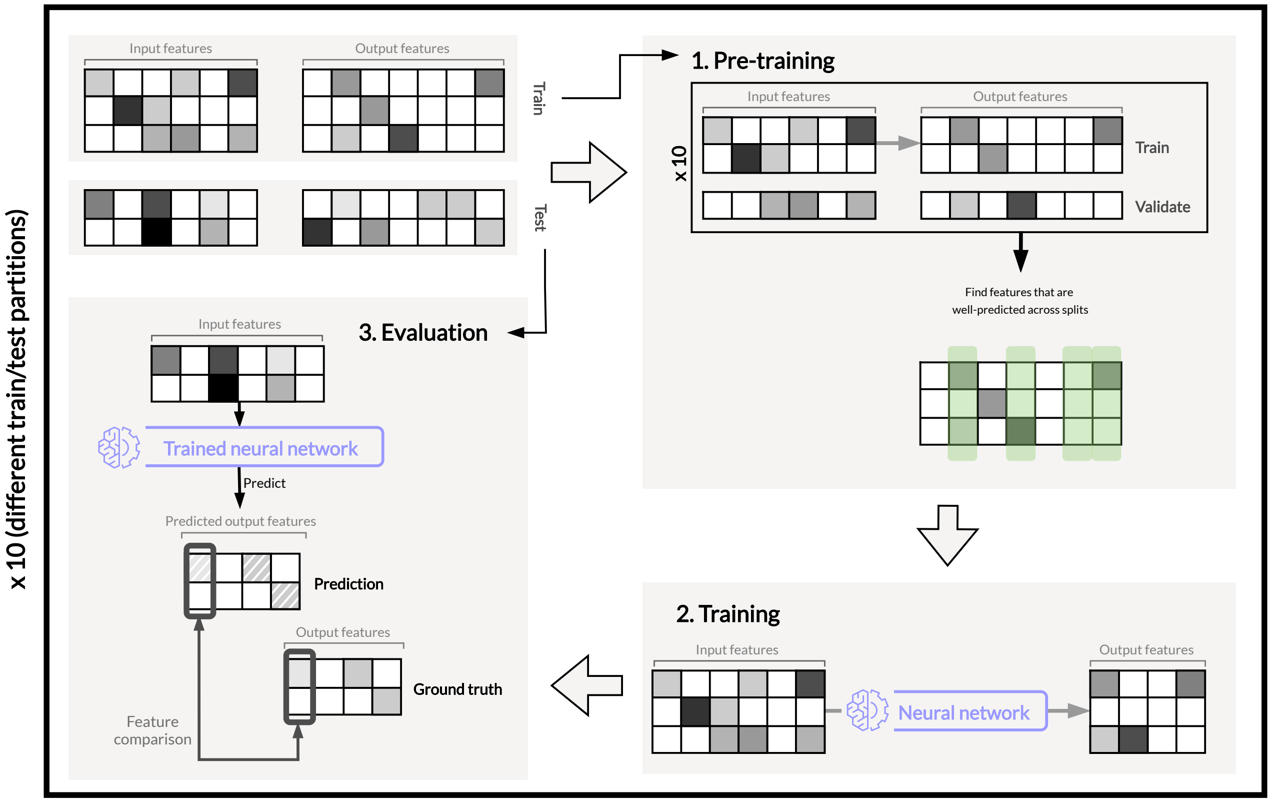


Figure S4: Pre-training for feature selection, performed on cross-validation folds. Selected features are subsequently used to train a neural network, as the one described in SectionA.1.

D. Code listings

```
# MelonnPan
# Training
Rscript train_metabolites.R -i [path_to_input_data] -g [path_to_output_data] -p 20 -o [
    path_to_output_folder]
# Making predictions with the re-trained model
Rscript predict_metabolites.R -w [path_to_trained_weights] -i [path_to_input_test_data] -o [
    path_to_output_folder]


# SparseNED
python3 main_cv_1dir.py --model BiomeAESnip --sparse 0.06 --learning_rate 0.01 --batch_size 20
     --latent_size 70 --activation "tanh_tanh" --data_type [dataset_name] --data_root [
    path_to_data_folder] --nonneg_weight --normalize_input

# MiMeNet
python3 MiMeNet_train.py -micro [path_to_input_data] -metab [path_to_output_data] -micro_norm
    None -metab_norm None -net_params None -external_micro + [path_to_input_test_data] -
    external_metab [path_to_output_test_data] -num_background 10 -num_run 5 -num_cv 5
```

Listing 1: Commands used to run “metagenomics-to-metabolomics” tools: MelonnPan [16], SparseNED[17] and MiMeNet [18].