Title

TaxaHFE: a machine learning approach to collapse microbiome datasets using taxonomic structure.

Journal

Bioinformatics Advances, 3(1)

Authors

Oliver, Andrew
Kay, Matthew
Lemay, Danielle

Publication Date

2023

DOI

10.1093/bioadv/vbad165

Peer reviewed

Bioinformatics Advances, 2023, **00**, vbad165
https://doi.org/10.1093/bioadv/vbad165
Advance Access Publication Date: 28 November 2023
Original Article

OXFORD

Data and text mining

TaxaHFE: a machine learning approach to collapse microbiome datasets using taxonomic structure

Andrew Oliver [1], **Matthew Kay** [2], **Danielle G. Lemay** [1,3,4,*]

[1]USDA-ARS Western Human Nutrition Research Center, Davis, CA 95616, United States
[2]Independent Researcher, Washington, DC 20002, United States
[3]Department of Nutrition, University of California, Davis, Davis, CA 95616, United States
[4]Genome Center, University of California, Davis, Davis, CA 95616, United States

*Corresponding author. Danielle G. Lemay, USDA-ARS Western Human Nutrition Research Center, 430 West Health Sciences Drive, Davis, CA 95616, United States. E-mail: danielle.lemay@usda.gov

Associate Editor: Sofia Forslund

Abstract

Motivation: Biologists increasingly turn to machine learning models not just to predict, but to explain. Feature reduction is a common approach to improve both the performance and interpretability of models. However, some biological datasets, such as microbiome data, are inherently organized in a taxonomy, but these hierarchical relationships are not leveraged during feature reduction. We sought to design a feature engineering algorithm to exploit relationships in hierarchically organized biological data.

Results: We designed an algorithm, called TaxaHFE, to collapse information-poor features into their higher taxonomic levels. We applied TaxaHFE to six previously published datasets and found, on average, a 90% reduction in the number of features (SD = 5.1%) compared to using the most complete taxonomy. Using machine learning to compare the most resolved taxonomic level (i.e. species) against TaxaHFE-preprocessed features, models based on TaxaHFE features achieved an average increase of 3.47% in receiver operator curve area under the curve. Compared to other hierarchical feature engineering implementations, TaxaHFE introduces the novel ability to consider both categorical and continuous response variables to inform the feature set collapse. Importantly, we find TaxaHFE's ability to reduce hierarchically organized features to a more information-rich subset increases the interpretability of models.

Availability and implementation: TaxaHFE is available as a Docker image and as R code at https://github.com/aoliver44/taxaHFE.

1 Introduction

With the cost of DNA sequencing continuing to drop faster than compute power increases (Wetterstrand), the analysis of large datasets remains a bottleneck in biological research. One method for analysis is machine learning (ML) (Choi *et al.* 2022), which is a blanket term that refers to computer algorithms designed to find patterns in data, iteratively optimizing performance without human input. While ML methods represent a suite of powerful and sensitive tools, they can suffer from a problem present in many humanomic studies: many features (i.e. microbial taxa) describing relatively few samples. Mathematician Richard Bellman referred to this problem as the "curse of dimensionality" (Bellman 2003). Practically speaking, too many features can result in "overfitting," leading to poor generalizability of the model. For this reason, implementing methods to reduce the size of data while retaining its important features can improve both the speed, generalizability, and interpretability of the data.

Feature engineering, or the set of preprocessing steps done to data prior to ML model evaluation, can help address problems imposed by high-dimensional data. While illustrating the totality of feature engineering is beyond the scope of this paper, some general examples include scaling or normalizing features, removing low variance features, collapsing highly correlated features, sophisticated methods for selecting subsets of features, and collapsing features into principal coordinate space. The goal of several of these methods is to reduce the feature space (dimensionality) and produce a highly discriminatory set of variables with respect to a response of interest. However, in biology models are not just used to make predictions, but to *explain*. This means that the reduced feature set needs to also be interpretable, rather than alternate ordinations, such as principal components.

Some biological data, such as microbiome and dietary data, can be represented using hierarchical structures (Jacobs and Steffen 2003; Johnson *et al.* 2019; Choi *et al.* 2022). Taxonomic assignments have long been used to identify microorganisms. This taxonomy is usually represented by a hierarchical classification scheme, whereby the ancestral level is the most general group, followed by increasingly specific grouping rules. More recently, researchers have begun to represent consumed foods in a similar taxonomic way (Johnson *et al.* 2019). More than merely identifying information, taxonomy represents relatedness, reflecting ecological patterns (for microorganisms) (Bevilacqua *et al.* 2021) or complex admixtures of similar chemicals and nutrients (for food) (Johnson *et al.* 2019). While data with a hierarchical structure presents many different levels by which to analyze a trait

Received: September 25, 2023; **Revised:** September 28, 2023; **Editorial Decision:** November 7, 2023; **Accepted:** November 27, 2023
Published by Oxford University Press 2023.

or response, researchers generally choose to collapse these data to a single level for ease of analysis (i.e. analyzing microbiome data at the family level) (Kleine Bardenhorst *et al.* 2021). This can be a useful strategy, especially if the trait/response of interest is known to be conserved at a certain phylogenetic depth which can be approximated by a taxonomic level (Martiny *et al.* 2015). Even without knowledge of a conserved phylogenetic depth *a priori*, tools exist to identify the average phylogenetic depth of a trait/response; however, many of these tools require a phylogenetic tree as input (Martiny *et al.* 2013). Often there is not *a priori* information available, and summarizing taxa to a specific level is weakly justified. Even more, if the data was summarized at a taxonomic level, but the response to a treatment is conserved at a different taxonomic level, it could lead to a false conclusion that, for example, the microbiome does not respond to a treatment of interest. A systematic review of current practices for the analysis of human microbiome data revealed a lack of consensus about which taxonomic level to study (Kleine Bardenhorst *et al.* 2021). Why not let information theory decide?

Here we introduce a method for hierarchical feature engineering (HFE) to dynamically collapse hierarchical data purely based on taxonomic relationships and information gain. Our algorithm, TaxaHFE, does not require the user to have knowledge *a priori* regarding the taxonomic level of conservation for a given trait. Rather, it seeks to maximize the information contained at various taxonomic levels while simultaneously reducing redundancy in the feature space. As a proof of concept, we apply TaxaHFE to microbiome data and compare it to an existing HFE algorithm (Oudah and Henschel 2018), assessing feature reduction and downstream ML performance. Additionally, we show that TaxaHFE's utility extends to other hierarchically organized data by applying our algorithm to hierarchical food data represented by taxonomic trees.

2 Methods

2.1 TaxaHFE algorithm

The algorithm (Fig. 1) can broadly be broken down into two main sections: (i) the creation of a taxonomic tree representing the hierarchical data and (ii) the competitions of each taxon in a post-order tree traversal. Within the competition section, four major steps occur: (i) a feature abundance and prevalence filter, (ii) a correlation competition between parent and child taxa, (iii) an ML step to determine the information content of the taxa from the previous step, and (iv) one additional ML step on all the "winning" features. These steps are graphically represented in Fig. 1 and algorithmically outlined below.

2.1.1 Build tree

1) Generate a node x in the tree T for each taxon.
2) Store the abundance values for the taxon on x, as well as calculating whether the abundances meet minimum mean abundance and prevalence thresholds (Fig. 1, Step 1)
3) To fill in any missing abundance data, traverse the tree $\mathbf{T}$ in post-order, generating a missing abundance vector a_n for a node x_n as the vector sum of all abundance vectors from the direct descendants $C_n = children(x_n)$, where a_c is the abundance vector of child node c.

$$a_n = \sum_{c \in C_n} a_c$$

2.1.2 Compete tree

Traverse the tree $\mathbf{T}$ in post-order, considering every subtree $\mathbf{T}_n$ using the following steps.

1) If $root(\mathrm{T}_n)$ has not met the minimum mean abundance and prevalence thresholds, this taxon will not be considered further (Fig. 1, Step 1)
2) If root($\mathbf{T}_\mathrm{n}$) is a leaf node, mark "winner" and proceed to the next subtree.
   a) Note: Being marked a "winner" is temporary, and any "winner" in a particular subtree T_n at level l will be reconsidered at level $l-1$ during the traversal.
3) Traverse $\mathrm{T_n}$, generating a set N_0 of nodes previously marked as "winner." If a node x is marked "winner," no descendant nodes are considered.
4) Generate a subset $\mathbf{N_{nc}} \in \mathbf{N}_0$, containing each node x in $\mathbf{N}_0$ whose abundance vector $\mathrm{a_x}$ is not correlated with the abundance vector $\mathrm{a_n}$ of root($\mathbf{T}_\mathrm{n}$), above the specified threshold t (Fig. 1, Step 2). For all nodes $\mathbf{N}_0 - \mathbf{N_{nc}}$, remove the "winner" designation:

$$\mathbf{N_{nc}} = \{x \mid x \in \mathbf{N}_0, corr(a_n, a_x) < t\}$$

5) Using the set of abundance vectors $\mathbf{A} = \{\mathrm{a_n}, \mathrm{a_{nc1}}, \mathrm{a_{nc2}}, \ldots\}$ (from root($\mathbf{T}_\mathrm{n}$) and the non-correlated nodes $\mathbf{N_{nc}}$) and response variables $\mathbf{M}$ (from the metadata input), fit a random forest (RF) model to determine the taxa importances to $\mathbf{M}$ calculated using Gini impurity-corrected scores (Nembrini *et al.* 2018) and generating a vector of scores $\mathbf{S}$ (Fig. 1, Step 3).

$$S = RFgi(\mathbf{A}, \mathbf{M})$$

Any node x with a score s_x greater than $root(\mathbf{T}_n)$ score s_n is marked "winner."

$$\mathbf{N}_w = \{x | x \in N_0, s_x > s_n\}$$

Otherwise, if the score of $\boldsymbol{root}(\mathbf{T}_n)$ is the highest value, only it is selected.

$$\mathbf{N}_w = \{\mathbf{root}(\mathbf{T}_n)\}$$

All other nodes $\mathbf{N_{nc}} - \mathbf{N}_w$ have the "winner" designation removed, as well as $\boldsymbol{root}(\mathbf{T}_n)$ if $\boldsymbol{root}(\mathbf{T}_n) \notin \mathbf{N}_w$.

Traversal will be stopped at a level $\boldsymbol{l}$, where $l \geq 2$, such that $\boldsymbol{level}(\boldsymbol{root}(\mathbf{T_n})) \geq \boldsymbol{l}$ for all competed subtrees $\mathbf{T_c}$. This prevents $\boldsymbol{root}(\mathbf{T}_0)$ (that contains the sum of all abundance vectors) from being included in the competition, and also allows for the preservation of taxonomic information at a definable level. Because they have not been traversed, any node $\boldsymbol{x}$ where $\boldsymbol{level}(\boldsymbol{x}) < \boldsymbol{l}$ is not considered in the algorithm and cannot be marked "winner."

The result of this is a set of distinct nodes $\mathbf{N}_w$, marked as "winner" across all competed subtrees $\mathbf{T_c}$ having $\boldsymbol{root}(\mathbf{T}_0)$ (the root of the full tree) as the only common ancestor. An

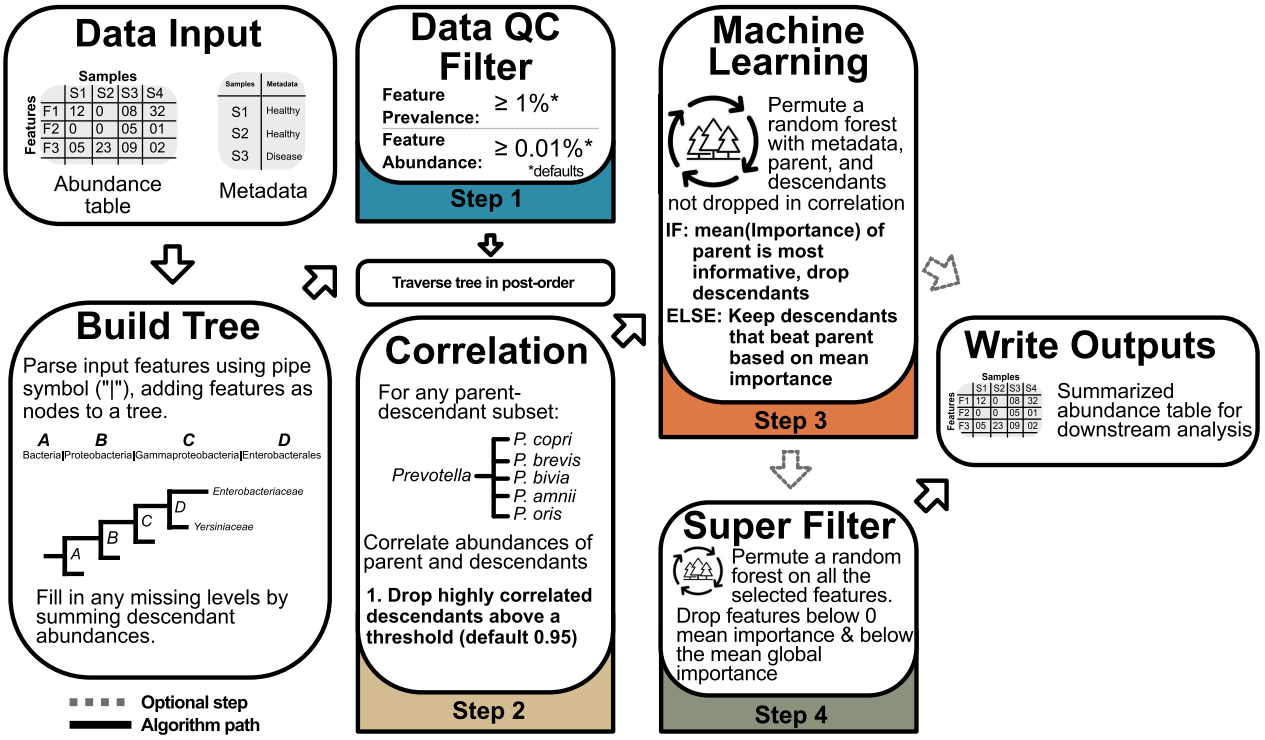

Figure 1. Overview of TaxaHFE.

optional final RF model (Fig. 1, Step 4) is then fit using the set of abundance vectors $\mathbf{A_W}$ from $\mathbf{N_W}$, and the response variables $\mathbf{M}$.

$$\mathbf{S_W} = RFgi(\mathbf{A_W}, \mathbf{M})$$

The final feature set $\mathbf{N_f}$ is the set of nodes x with scores s_x greater than zero and the average score Gini impurity-corrected scores of $\mathbf{N_f}$:

$$\mathbf{N_f} = \{\mathbf{x} | \mathbf{x} \in \boldsymbol{N_W}, s_x > 0, s_x > \mathbf{mean}(S_W)\}$$

2.2 Input data

The input data to TaxaHFE is (i) a flat file (comma- or tab-delimited) with a column labeled "clade_name" containing the taxonomic features with subsequent columns as samples containing numeric feature abundances and (ii) a response variable flat file (Fig. 1). The levels of the features within the input data should be delineated by a pipe symbol ("|"). For example, if the input data are microbiome features, the features should be Kingdom|Phylum|Class|…|last level. The taxonomic features can contain any number of levels present. The sample column names must be unique for each column in the input file and match a column in the response variable file, which also must contain a column encoding the response of interest (either categorical or continuous).

2.3 Implementation

A reference implementation of this algorithm was written in R as part of development. The implementation uses the data.tree package (v1.0.0) (Glur 2020) to represent the hierarchical structure of input taxa, as well as for the post-order traversal of the data used in several places. Data.tree parses the input file column by using the pipe symbol ("|") found in the input data, described above. Once imported into TaxaHFE, four broad steps occur, as described in the algorithm above: (i) feature abundance and prevalence filters, (ii) a correlation competition between parent taxon and child taxa, (iii) a RF competition selecting the most informative (relative to the response variable) features from the previous step, and (iv) and an optional final "super filter" (SF) RF of all the taxa that have survived from steps 1–3 (Fig. 1). TaxaHFE applies reasonable defaults to these steps. For instance, by default, features are only considered if they have a minimum mean feature abundance of 0.01% and a minimum feature prevalence of 1%. Both filters can be set to zero or any other number, depending on the scale of the input data. Additionally, if the user is predicting a rare class (i.e. class imbalance), it may be prudent to keep rare features by setting these filters to zero. For the correlation competition, a child taxon abundance must be correlated with the parent abundance at <0.95 Pearson correlation to pass to step 3. Finally, both step 3 and step 4 use RFs to report back Gini impurity-corrected scores per feature, indicating the feature's ability to predict the user-supplied response variable. These RF models are built and rebuilt 40 times by default, averaging out the Gini impurity scores per feature. The RF model fitting is done using the ranger package (v0.14.1) (Wright and Ziegler, 2017), including the ability to distinguish between numeric and factor response variable types in the input metadata. Randomness is seeded in the implementation using system time, but an allowance is made for defining a specific random seed. This randomness influences the outcome of the RF competitions run by the ranger package. For ease of use, the implementation has been released on GitHub and additionally packaged into a Docker container, allowing it to be run reliably in a variety of computing environments.

2.4 Output data

The outputs of TaxaHFE are individual files of summarized original abundances for each hierarchical level and two files containing either the TaxaHFE or the TaxaHFE + SF selected features alongside the response variable.

2.5 Downstream ML

The output of TaxaHFE is a reduced feature set, which may increase the performance of predictive analyses such as ML. To test this, we analyzed hierarchical datasets using a ML pipeline based around the Tidymodels package (v1.0.0) (Kuhn *et al.* 2020) in R. Input data was split using 70% data for training and hyperparameter tuning, and 30% was used for testing. Inside the training split, 10-fold repeated (3×) cross-validation was used for Bayesian-search hyperparameter tuning. A gentle correlation feature reduction step was also introduced within each cross-validation fold, yet at 0.95 Pearson (the same as TaxaHFE's internal correlation filter), it likely had minimal effect on ML performance. The search space was limited to 80 different hyperparameter combinations (optimizing mtry and minimum node size) or ten minutes of search time, whichever finished first. Bayesian hyperparameter search was allowed to preemptively end if the scoring metric was not improved after ten iterations. The metric optimized was balanced accuracy for classification or mean absolute error for regression. The best model was fit to the left-out test data and assessed using area under the receiver operator curve (ROC-AUC), balanced accuracy, and Cohen's kappa (Kuhn *et al.* 2023). For multi-class models, the Hand-Till ROC-AUC (Hand and Till 2001) and macro-averaged balanced accuracy were used. Each ML model was run using ten different random seeds to reduce stochasticity introduced from different train-test splits. To determine the importance of features, the package fastshap (v0.0.7) (Greenwell 2021; Štrumbelj and Kononenko 2014) was used. Briefly, the best model was fit to the entire input data, and Shapley values were calculated for each feature. The R package ShapViz (v0.4.1) (Mayer 2023) was used to plot these values.

2.6 Evaluation of TaxaHFE

Six previously published microbiome datasets (Lloyd-Price *et al.* 2019; Franzosa *et al.* 2019; Mars *et al.* 2020; Wang *et al.* 2020; Erawijantari *et al.* 2020; Muller *et al.* 2022; Oliver *et al.* 2022) and a dietary tree dataset (Kable *et al.* 2022) were used to assess TaxaHFE (Supplementary Table S1). For the microbiome datasets, TaxaHFE parameters were set to minimum mean feature abundance of 0.01% and minimum feature prevalence of 1%. The correlation threshold was set to 0.95, and nperm = 40 (defaults for TaxaHFE) and the random seed was set to 42. For the food dataset, the abundance filter was changed to 0. Each dataset was summarized at either the order, family, genus, or species level (using TaxaHFE, which applies the prevalence and abundance filters prior to writing the summary files). Additionally, the species summarized microbiome data was analyzed using a previously published algorithm (hfe_algorithm.py: Oudah and Henschel, 2018), using two correlation cutoffs: the program's default 0.7 and 0.95 (matching TaxaHFE's default correlation filter). These summarized and TaxaHFE-selected data were used in ML models to predict response variables associated with each study (Supplementary Table S1), as described above. To test whether ML performance was significantly different across different feature reduction methods, we used a linear mixed-effects model (from the R package nlme v3.1–157: Pinheiro *et al.* 2022), with study*level as a fixed effect and the random seed as a random effect. Quantile-quantile plots were investigated for normality of residuals. Study-specific models were also built like the above model, without the study interaction term. Finally, we performed estimated marginal means (EMMs) post hoc tests using the emmeans package (v1.8.8) (Lenth 2023) with Bonferroni adjusted *P*-values. We also analyzed the features compositionally by measuring the variance explained by the features selected using PERMANOVA models. To do so, we used the adonis(method = "bray," nperm = 999) function from the vegan (v2.6–4) package (Oksanen *et al.* 2022) in R. For comparative purposes, Lefse (Segata *et al.* 2011) was run on Galaxy (http://galaxy.biobakery.org/) using default parameters, and Boruta was run using the Boruta package (v8.0.0) (Kursa and Rudnicki, 2010) in R using default parameters.

2.7 Software and data availability

The code for TaxaHFE, along with installation instructions and example inputs, can be found on GitHub (https://github.com/aoliver44/taxaHFE). TaxaHFE version 2.0 was used for all analyses. The ML pipeline to assess the performance of TaxaHFE can also be found on GitHub (https://github.com/aoliver44/nutrition_tools). Location of datasets used for comparisons can be found in Supplementary Table S1. We downloaded five of the microbiome datasets from a microbiome-metabolome dataset collection (Muller *et al.* 2022).

3 Results

3.1 TaxaHFE improves feature reduction compared to alternative taxonomically informed methods

We initially investigated how well TaxaHFE performs compared to summarizing microbiome datasets to higher taxonomic levels (order (L4) to species (L7)), or when using a previously published HFE program (Oudah_70 and Oudah_95). In all six previously published studies, the dimensional reduction produced by TaxaHFE (±SF) selects features which, when used as input in ML models, result in the higher mean ROC-AUC scores compared to data summarized at specific taxonomic levels, or a previously published HFE algorithm (Fig. 2A, Supplementary Fig. S1 and Table S2). The mean ROC-AUC for TaxaHFE (+SF) across all studies assessed was 0.901 (SD = 0.071) followed by TaxaHFE (−SF) (0.897, SD = 0.071). Moreover, TaxaHFE (+SF) produces the best models for 4/6 studies when assessed using balanced accuracy (Fig. 2B, Supplementary Table S2) and 4/6 studies using Cohen's kappa (Fig. 2C, Supplementary Table S2). In the other two cases, TaxaHFE-preprocessed models were not significantly different from the best models produced (Supplementary Fig. S1).

In addition to performance improvements, TaxaHFE (+SF) utilizes less features than comparable methods (Fig. 2D). Models utilizing species-level features (after abundance and prevalence filters, see methods) utilized 469 (SD = 97) species on average, compared with the 45 (SD = 22) taxonomic features used by TaxaHFE (+SF) on average. Additionally, TaxaHFE (+SF) selects 95 less features on average compared to a previously published HFE algorithm by Oudah and Henschel (2018). Importantly, TaxaHFE's

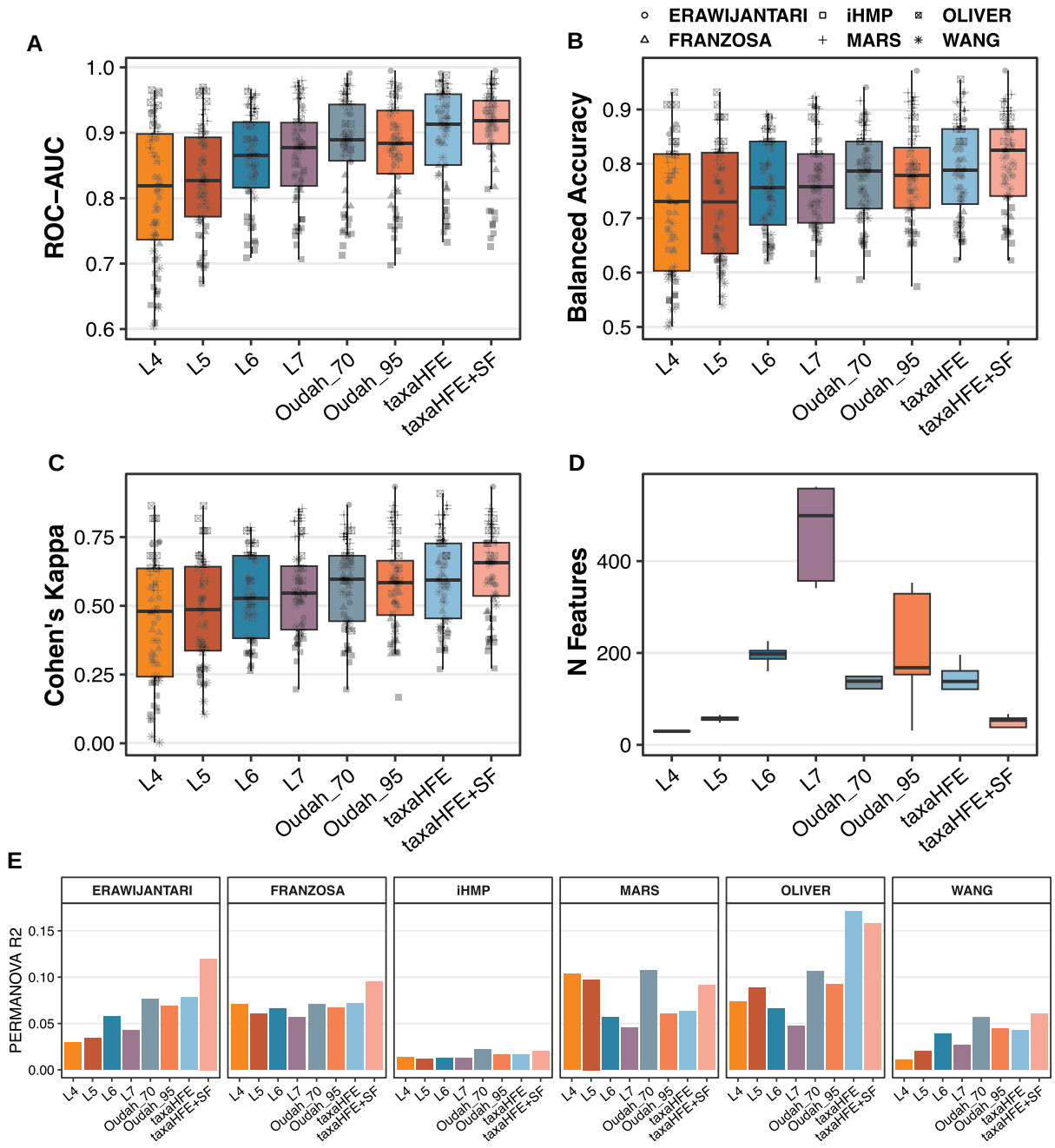

Figure 2. Comparisons made used summarized microbiome data (L4–L7, order through species), a previously published HFE program (Oudah) employing two different internal correlation cutoffs (Pearson 0.7 and 0.95), and TaxaHFE with and without the super-filter. ML performance metrics (A–C) and number of features used for model building (D). (E) Variance explained in PERMANOVA models by the composition of the features relative to the response variable used.

reduced feature set generally captures more variance in community composition compared to existing HFE methods or summarizing to a specific taxonomic level (Fig. 2E).

3.2 TaxaHFE can reduce features for both categorical and continuous predictions

A previous effort toward HFE produced an algorithm which maximized a performance metric with respect to a categorical variable (Oudah and Henschel 2018). We designed TaxaHFE to handle both categorical and continuous variables. To illustrate this, we used normalized antibiotic resistance gene abundance as a continuous or categorical response variable for selecting taxa using previously published data (Oliver *et al.* 2022). Specifically, we only analyzed samples in the highest or lowest ARG abundance quartile. Using categorical ARG abundance (low quartile versus high quartile), TaxaHFE, with or without the default super-filter, selected 4 and 14 taxon features respectively. Using these features as input to a RF classification model predicting categorical ARG abundance (high versus low) resulted in mean ROC-AUC values of 0.941 (+SF) and 0.952 (−SF) (Fig. 3). Features summarized at the L4 level performed slightly better in classification models than TaxaHFE+SF (L4 (order) ROC-AUC: 0.943). The previously published HFE algorithm by Oudah and

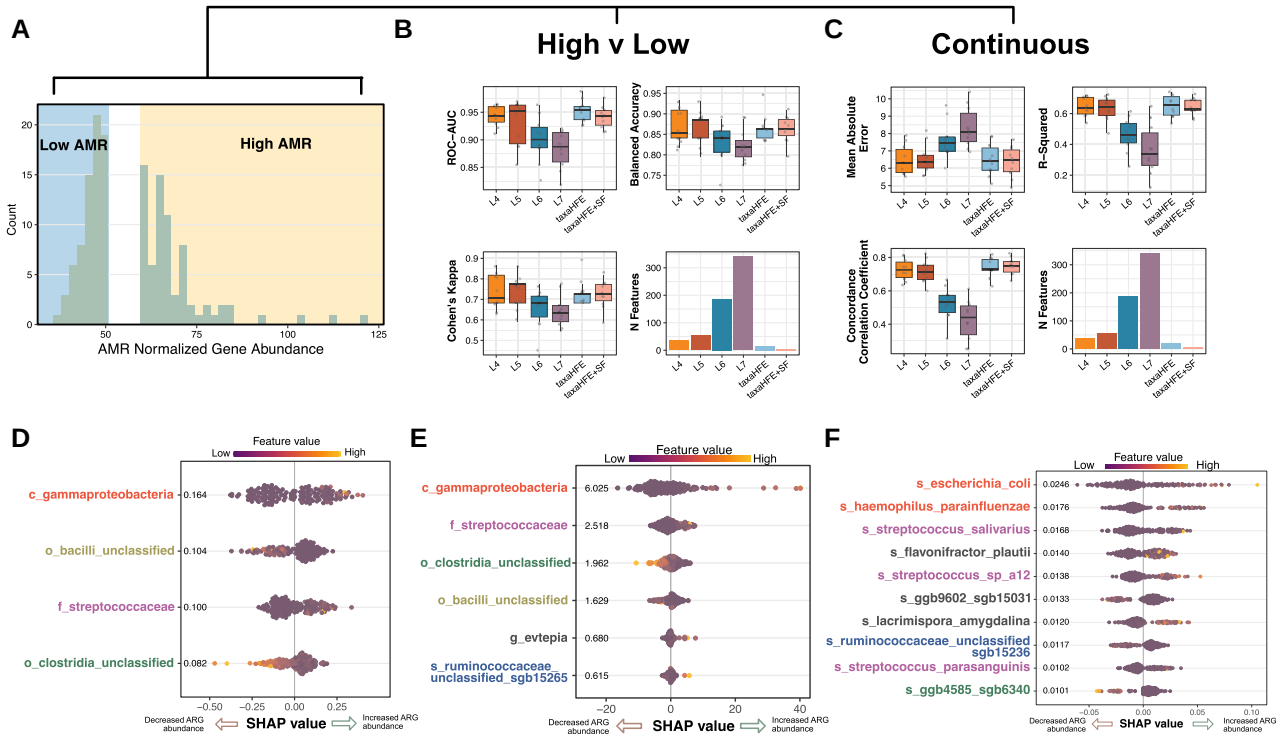

Figure 3. Analysis of TaxaHFE on categorical and continuous data. Histogram of antibiotic resistance gene abundance, showing just the top quartile and bottom quartile of the cohort (A). These data were analyzed as a categorical factor (low versus high ARG abundance) or continuous ARG abundance, and the (B, C) boxplots show the performance of ML models utilizing summarized taxonomic data or taxonomic HFE. (D–F) Shapley values of the most informative features from a model built using TaxaHFE+SF data as input for both categorical and continuous ARG abundance are shown, as well as a model built only using L7 level data against the categorical ARG outcome. Colored font levels share a common taxonomic group with TaxaHFE.

Henschel (Oudah and Henschel 2018) selected 134 features, a considerable increase from TaxaHFE (+SF) 4 features. Even with far fewer features, TaxaHFE (+SF) achieved better performance metrics compared to models built with Oudah engineered features (ROC-AUC: TaxaHFE+SF, 0.941; Oudah_70, 0.915).

When analyzing continuous ARG abundance, TaxaHFE, with or without the default super-filter, selected 6 and 18 taxon features, respectively. Using these features as input to an RF regression model predicting continuous ARG abundance resulted in mean R^2 (coefficient of determination) values of 0.644 (+SF) and 0.649 (−SF) (Fig. 3C). Like the categorical example, TaxaHFE+SF performed slightly behind the L4 model (by mean R^2 performance), which used L4 (order) summarized data (0.645), yet a post hoc test showed these differences were not significant, and the L4 model used far more features than TaxaHFE+SF (36 versus 6). No data is shown for the Oudah and Henschel algorithm as that algorithm does not support continuous outcome variables. In summary, TaxaHFE maximized performance while minimizing features for both categorical and continuous outcomes.

Both the categorical and continuous examples resulted in models utilizing similar features (Fig. 3D and E). Compared to a model built using Level 7 features (species), TaxaHFE identified higher taxonomic levels to discriminate categorical ARG abundance. These higher taxonomic levels, such as the class *Gammaproteobacteria* and family *Streptococcaceae*, resulted in nearly an order of magnitude higher SHAP values, suggesting TaxaHFE identified features that were more influential to model predictions than models built using species-level data.

3.3 TaxaHFE also works on hierarchically organized diet data

Since TaxaHFE works with hierarchical features, we sought to examine its utility beyond microbiome data. Dietary data represented as food trees could also be a useful feature set to apply HFE. To test this, we utilized a previously published dietary food tree and tested TaxaHFE's ability to select features that explain average fiber intake (Baldiviez *et al.* 2017). After prevalence and abundance filters, 456 foods were assessed. TaxaHFE+SF selected 4 features as the most informative for predicting average fiber intake, nearly a 99% reduction from the 456 features used as input. When assessed using the concordance correlation coefficient, a measure of both correlation and accuracy, Bonferroni corrected EMM post hoc test revealed that TaxaHFE+SF performed significantly better than summarized features in a RF ($P < 0.05$) (Fig. 4A). We next used Shapley values to determine which features were most important to a model built with TaxaHFE+SF (Fig. 4B) compared to a model built with Level 7 features (Fig. 4C). The most important features identified by TaxaHFE+SF were "other fruits" (Level 2), which include high fiber staples such as berries and avocados, and "dry beans peas other legumes nuts and seeds" (Level 1) (Fig. 4B). In contrast, the predictive model performance is lower for Level 7 (Fig. 4A) and the SHAP values, which indicate the magnitude of importance, are also lower (Fig. 4C) than those identified by

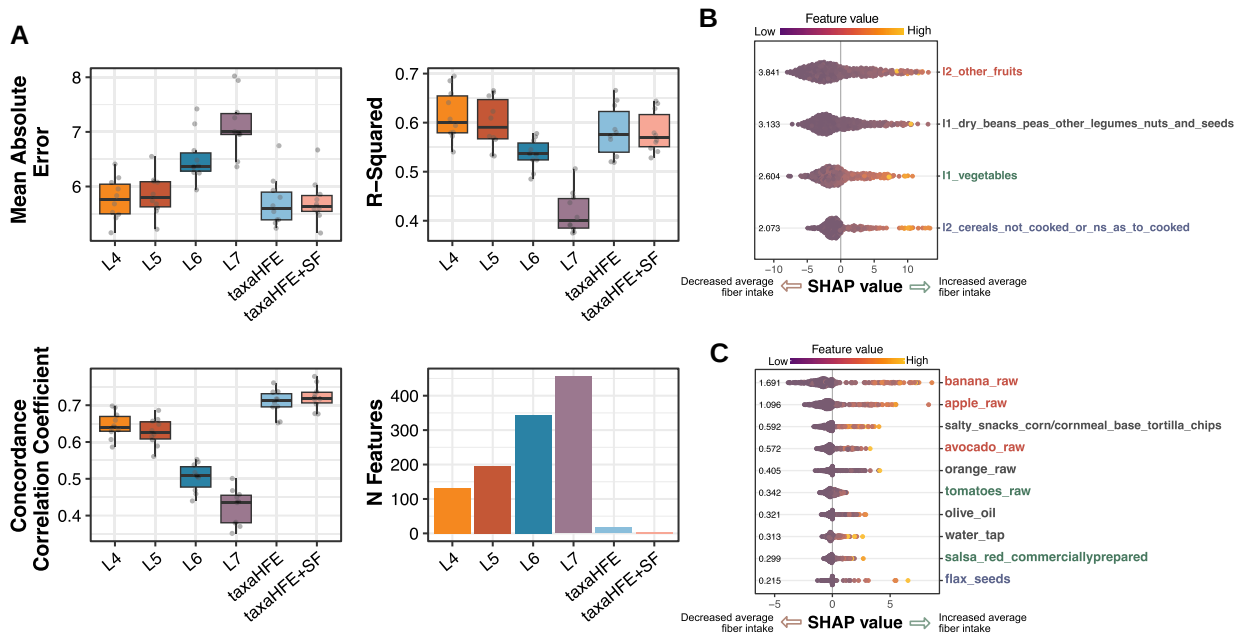

Figure 4. Testing TaxaHFE on hierarchically organized dietary data. ML performance metrics of models predicting average fiber intake using food tree data (A) and number of features used for model building. Shapley values (B, C) of the most informative features from a model built using TaxaHFE+SF data as input or L6 as input. Colored font levels are those shared with TaxaHFE.

TaxaHFE+SF (Fig. 4B). In summary, the use of TaxaHFE improved both model performance and feature importance when used with dietary data arranged in food trees.

4 Discussion

A common practice in microbiome analyses is the summarization of abundance data at a single taxonomic level. In one review examining 419 microbiome studies, the authors found that the genus level was the most commonly summarized taxonomic level, followed by analysis at the phylum level (Kleine Bardenhorst *et al.* 2021). However, the discernable variation in microbial composition with respect to a trait or response is strongly dependent on the taxonomic level analyzed (Martiny *et al.* 2015). Thus, an analysis at a single taxonomic level, especially at a level that is not optimized for the trait or response of interest, could lead investigators to incorrectly conclude the absence of a relationship between microbiome and response. Our results support this; to explain ARG abundance, a model built using species-level data would perform significantly worse than a model built using order level or TaxaHFE-preprocessed data (Fig. 3).

Here we propose an algorithm for the feature engineering of hierarchically organized data, particularly microbiome abundance data or dietary data represented using food trees. Currently, few algorithms exist to reduce the high dimensionality found in microbiome data with respect to the taxonomic structure inherent in microbial features. Oudah and Henschel (2018) provided an excellent implementation; however, the Oudah algorithm is unable to handle: (i) a continuous response variable, (ii) non-bacterial features and (iii) abundances that are not relative abundances. At the core of TaxaHFE is a RF, which is a particularly capable ML algorithm for handling both continuous and categorical response variables. Moreover, there is no need for specific feature names; TaxaHFE bases its understanding of hierarchical levels purely based on the use of a separator ("|") between levels. As such, any hierarchically organized data can be used as input.

One important shared logic of both the Oudah algorithm and our algorithm is the use of the parent taxon as the taxon to "beat" in the competitions. This decision was inspired by an early implementation of HFE (Ristoski and Paulheim 2014), for which the goal of the algorithm was to choose the most valuable features from the highest taxonomic levels possible. Indeed, when the goal is feature reduction, and the parent taxon and child taxa contain redundant information relative to a response of interest, choosing the parent taxon results in a feature that is likely representative of the child taxa in some way. The reverse is not necessarily true.

Other feature reduction programs such as Lefse (based on linear discriminant analysis) (Segata *et al.* 2011) and Boruta (based on RFs) (Kursa and Rudnicki 2010) are often used, particularly with microbiome data. However, when all taxonomic levels are supplied, these programs often choose features that carry redundant abundance information. For example, when using Lefse to select features based the Erawijantari study, the output contained *Lactobacillales*, *Streptococcaceae*, and *Streptococcus* (Supplementary Fig. S2A), which are all directly related features that carry nested but redundant feature abundance information (i.e., the abundance of *Streptococcus* is contained within the abundance of *Streptococcaceae*). And while Boruta performs slightly better than TaxaHFE in almost every case (Supplementary Fig. S2B), it exhibits similar behavior as Lefse, choosing features with redundant feature abundance information (Supplementary Fig. S2C). For some types of analyses this behavior is desirable. However, for explicit feature reduction, we suggest that TaxaHFE's method of removing overlapping features is usually preferable to aid interpretation of biological data.

Importantly, TaxaHFE-preprocessed data improves the performance of ML models, especially compared to models produced using the lowest taxonomic levels available (e.g. species). This is perhaps not altogether surprising; indeed, a common observation across many microbiome studies is the highly personalized nature of microbiomes (Oliver *et al.* 2021). It stands to reason then, that while utilizing the most resolved taxonomic data might best highlight differences from person to person, it will often mask more generalizable microbial responses within heterogeneous cohorts. HFE allows for the capture of these more generalizable responses, concomitantly increasing the accuracy of ML models in the process. We note, however, that if the goal is a generalizable model, using feature engineering inside of a cross-validation strategy is important to avoid data leakage (Aldehim and Wang 2017). We briefly examined the similarity of features selected across k = 3-fold partitions of the data and found a high amount of dissimilarity (83%, data not shown) among the features selected in each fold. One reason for this variability is the small number of samples in the published studies we used to assess TaxaHFE (mean 262 samples). Like other feature reduction tools, we expect the performance of TaxaHFE will suffer when samples are limited.

Overall, perhaps the most important aspect of TaxaHFE is the gains in interpretability. In the microbiome example of the status quo method of using the lowest level taxa (Fig. 3F), it is difficult to interpret the meaning of the ten species selected, each with a very low SHAP value. But with TaxaHFE (Fig. 3D and E), it is readily apparent that the class *Gammaproteobacteria* and family *Streptococcaceae* are top predictors of antimicrobial resistance in the human gut microbiome and their SHAP values are an order of magnitude higher. The relationship between *Gammaproteobacteria* and antibiotic resistance has been shown previously in a Hi-C study linking ARGs to their microbial hosts (Stalder *et al.* 2019). Moreover, when considered together TaxaHFE-selected features also appear to explain more compositional variance than using a single taxonomic level (Fig. 2E). In the diet example (Fig. 4C), the model based on the lowest taxonomic level is again difficult to interpret, with 456 individual food items (e.g. apples, bananas, avocados, oranges, olive oil, salty corn snacks, flax seeds, etc.) predictive of fiber intake. TaxaHFE reports low intake of other_fruits (non-citrus), legumes/nuts/seeds, uncooked cereals, and vegetables of low fiber intake. The results from TaxaHFE suggest, for example, that it is the entire class of legumes/nuts/seeds that is predictive, not just flax seeds, which is more reasonable. For this reason, dietary data is traditionally summarized and reported at the highest taxonomic level (e.g. fruits, vegetables, dairy, meat, etc.). However, nuances like the association of low intake of uncooked cereals with low fiber intake will be missed with such standard summary variables.

One shortcoming of TaxaHFE is its speed. For example, in a microbiome dataset with 4640 features, TaxaHFE took 2 min 44 s, whereas the Oudah algorithm took only 13 s on a 2.3 GHz Quad-Core Intel® Core i7 machine. Future iterations of TaxaHFE could utilize a more distributed parallelization implementation.

Another limitation of TaxaHFE is true of all ML algorithms in that large high-quality datasets are needed. Without sufficient information in the data (i.e., no relationship between the response variable and the features) all ML algorithms will suffer. However, what may not be obvious to users is that in the absence of information, ML algorithms will still report results, which often represent noise. Methods for avoiding pitfalls of ML have been described elsewhere (Whalen *et al.* 2021), but we would be remiss to not echo them here. Specifically, we implore users to not merely trust models based on their accuracy alone but to also investigate the features utilized for making predictions.

5 Conclusion

We demonstrate that TaxaHFE dramatically reduces the feature space of hierarchically organized data while generally increasing the performance of downstream ML models and improving interpretability. While our examples come from microbiology and nutrition research, TaxaHFE could be used with any dataset that has hierarchically related features. Moreover, TaxaHFE removes the prerequisite of choosing a taxonomic level that captures the most information relative to a response of interest, removing the need to ask, "At what taxonomic level should I analyze my data?" Overall, we suggest that HFE can lead to more accurate and interpretable models.

Acknowledgements

We would like to acknowledge Elizabeth Chin for early TaxaHFE discussions and Stephanie Wilson and Sarah Blecksmith for useful feedback provided while code testing. We would like to thank Rachel Waymack and Jules Larke for thoughtful comments and edits.

Author contributions

Andrew Oliver (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Methodology [equal], Software [equal], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Matthew Kay (Formal analysis [supporting], Methodology [equal], Software [equal], Writing—original draft [supporting]), and Danielle G. Lemay (Conceptualization [equal], Funding acquisition [lead], Project administration [lead], Resources [lead], Supervision [lead], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) [grant 2032–51530-026-00D]. A.O. was supported by an appointment to the Research Participation Program at USDA ARS, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U. S. Department of Energy and ARS. This research used resources provided by the SCINet project of the USDA ARS project number 0500–00093-001–00-D. USDA is an equal opportunity employer.

References

Aldehim G, Wang W. Determining appropriate approaches for using data in feature selection. *Int J Mach Learn Cyber* 2017;**8**:915–28.

Baldiviez LM, Keim NL, Laugero KD *et al.* Design and implementation of a cross-sectional nutritional phenotyping study in healthy US adults. *BMC Nutr* 2017;**3**:79–13.

Bellman RE. *Dynamic Programming*. Dover Publications, 2003.

Bevilacqua S, Anderson MJ, Ugland KI *et al.* The use of taxonomic relationships among species in applied ecological research: baseline, steps forward and future challenges. *Austral Ecol* 2021;**46**:950–64.

Choi Y, Hoops SL, Thoma CJ *et al.* A guide to dietary pattern–microbiome data integration. *J Nutr* 2022;**152**:1187–99.

Erawijantari PP, Mizutani S, Shiroma H *et al.* Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut* 2020;**69**:1404–15.

Franzosa EA, Sirota-Madi A, Avila-Pacheco J *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 2019;**4**:898.

Glur C. *data.tree: General Purpose Hierarchical Data Structure*. R package version 1.1.0, 2020. https://github.com/gluc/data.tree.

Greenwell B. *fastshap: Fast Approximate Shapley Value*. R package version 0.0.7, 2021. https://github.com/bgreenwell/fastshap.

Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn* 2001; **45**:171–86.

Jacobs DR, Steffen LM. Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr* 2003;**78**:508S–13S.

Johnson AJ, Vangay P, Al-Ghalith GA *et al.*; Personalized Microbiome Class Students. Daily sampling reveals personalized Diet-Microbiome associations in humans. *Cell Host Microbe* 2019;**25**: 789–802.e5.

Kable ME, Chin EL, Storms D *et al.* Tree-Based analysis of dietary diversity captures associations between fiber intake and gut microbiota composition in a healthy US adult cohort. *J Nutr* 2022; **152**:779–88.

Kleine Bardenhorst S, Berger T, Klawonn F *et al.* Data analysis strategies for microbiome studies in human populations—a systematic review of current practice. *mSystems* 2021;**6**:10.1128/msystems.01154-20.

Kuhn M *et al. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. 2020. https://www.tidymodels.org

Kuhn M, Vaughan D, Hvitfeldt E. *yardstick: Tidy Characterizations of Model Performance*. 2023. https://github.com/tidymodels/yardstick, https://yardstick.tidymodels.org.

Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Soft* 2010;**36**:1–13.

Lenth R. *Emmeans. R Packag*. R package version 1.8.9, 2023. https://github.com/rvlenth/emmeans.

Lloyd-Price J, Arze C, Ananthakrishnan AN *et al.*; IBDMDB Investigators. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62.

Mars RAT, Yang Y, Ward T *et al.* Longitudinal multi-omics reveals Subset-Specific mechanisms underlying irritable bowel syndrome. *Cell* 2020;**183**:1137–40.

Martiny AC, Treseder K, Pusch G . Phylogenetic conservatism of functional traits in microorganisms. *ISME J* 2013;**74**:830–8.

Martiny JBH, Jones SE, Lennon JT *et al.* Microbiomes in light of traits: a phylogenetic perspective. *Science* 2015;**350**:aac9323.

Mayer M. *shapviz: SHAP Visualizations*. R package version 0.4.1, 2023. https://github.com/mayer79/shapviz.

Muller E, Algavi YM, Borenstein E *et al.* The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *Npj Biofilms Microbiomes* 2022;**8**:79.

Nembrini S, König IR, Wright MN *et al.* The revival of the gini importance? *Bioinformatics* 2018;**34**:3711–8.

Oksanen J, Simpson G, Blanchet F *et al. vegan: Community Ecology Package*. R package version 2.6-4, 2022. https://github.com/vegandevs/vegan.

Oliver A, Xue Z, Villanueva YT *et al.* Association of diet and antimicrobial resistance in healthy U.S. Adults. *MBio* 2022;**13**:e0010122.

Oliver A, Chase AB, Weihe C *et al.* High-Fiber, Whole-Food dietary intervention alters the human gut microbiome but not fecal Short-Chain fatty acids. *mSystems* 2021;**6**

Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 2018;**19**:227–13.

Pinheiro J, Bates D, R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157, 2022. https://CRAN.R-project.org/package=nlme.

Ristoski P, Paulheim H. Feature selection in hierarchical feature spaces. In: Džeroski S, Panov P, Kocev D *et al.* (eds) *Discovery Science Lecture Notes in Computer Science*, Vol. 8777. Cham: Springer, 2014 288–300.

Segata N, Izard J, Waldron L *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;**12**:R60.

Stalder T, Press MO, Sullivan S *et al.* Linking the resistome and plasmidome to the microbiome. *Isme J* 2019;**13**:2437–46.

Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 2014;**41**:647–65.

Wang X, Yang S, Li S *et al.* Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. *Gut* 2020; **69**:2131–42.

Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. www.genome.gov/sequencingcostsdata (May 2023, date last accessed).

Whalen S *et al.* Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet* 2021;**2021**:169–81.

Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Soft* 2017;**77**:1–17.