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

Authors: Andrew Oliver[1], Danielle G. Lemay [1,2,3]

Affiliations:
[1]USDA-ARS Western Human Nutrition Research Center, 430 West Health Sciences Dr., Davis, CA, 95616
[2]Department of Nutrition, University of California, Davis, One Shields Ave., Davis, CA, 95616
[3]Genome Center, University of California, One Shields Ave., Davis, CA, 95616
*Corresponding author, 430 West Health Science Dr. Davis, CA 95616, danielle.lemay@usda.gov, (530) 752-4748

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 performance and interpretability of models. However, some biological data sets, 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, an 85% reduction in the number of features (s.d = 14.2%) 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 3.78% in receiver operator curve area under the curve (ROC-AUC). Compared to other hierarchical feature engineering implementations, TaxaHFE introduces the novel ability to consider both categorical and continuous dependent variables to inform the feature set collapse. TaxaHFE results in a dramatic decrease in the number of features, which in turn improves the performance and interpretability of machine learning models.

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

INTRODUCTION

With the cost of DNA sequencing continuing to drop faster than compute power increases (Wetterstrand.) the analysis of large data sets 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 human -omic 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, the 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. 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 hierarchical structure presents many different levels by which to analyze a trait 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.*, 2012). 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 hierarchical feature engineering 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.

METHODS

Input data

The input data to taxa HFE is (1) a flat file (comma- or tab-delimited) with the rows as taxonomic features and the columns as samples and (2) a dependent variable flat file. 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. For best functionality, data may be formatted in the style of Metaphlan-style output (Blanco-Míguez *et al.*, 2023) (referred to as taxon-parsed features throughout, see Supplemental Table 1) or formatted where each feature contains information about every hierarchical level (referred to as complete-features throughout, see Supplemental Table 2). The sample column names must be unique for each column in the input file and match a column in the dependent variable file, which also must contain a column encoding the response of interest (either categorical or continuous). The sum of each column must be equal, either through rarefaction or by converting counts to relative abundances.

TaxaHFE algorithm

TaxaHFE first reads in the input data and a dependent variable file. Next, if the input data is formatted as complete-features, TaxaHFE attempts to convert the data to taxon-parsed features. Two early filters are applied to the data prior to feature engineering, 1) a 1% prevalence filter and 2) an 0.0001 abundance filter (Figure 1). The prevalence filter removes any features that only occur in 1% or less of samples. The abundance filter removes features that occur at 0.0001 relative abundance or less. The sufficiently prevalent and abundant features then enter the main hierarchical competition section of the algorithm. Initially, all distinct child taxa for each parent taxa are tallied, and parent taxa with 2 or more child taxa are advanced in the competition. For parent taxon that only contain a single child, that child is removed from the final dataset; the child taxon likely contains redundant information with the parent. The first phase of the hierarchical competition is correlation-based. For child taxa that are Pearson correlated with their parent taxa at 0.95 or greater are removed from the final dataset. All child taxa are allowed to be removed at this step. If child taxa remain, a second phase of the hierarchical competition occurs. The remaining parent and child taxa are used as predictors in a permuted (ten times) random forest model (using the Ranger package (Wright and Ziegler, 2017)), predicting the dependent variable. The resulting impurity-corrected variable importance scores (Nembrini *et al.*, 2018) are averaged across permutations. If, on average, the parent taxon is more important than the remaining children, the children are removed from the final dataset. If there are children that are, on average, more important than the parent taxon, those children more important than the parent are kept. These competitions occur for taxon at every hierarchically adjacent level (e.g., genus vs species, family vs genus, etc.). Finally, an optional final filter step ("super filter (SF)") which permutes a random forest on the remaining features with the dependent variable being predicted, keeping the features which, on average, have variable importance greater than zero and greater than the mean variable importance for all features.

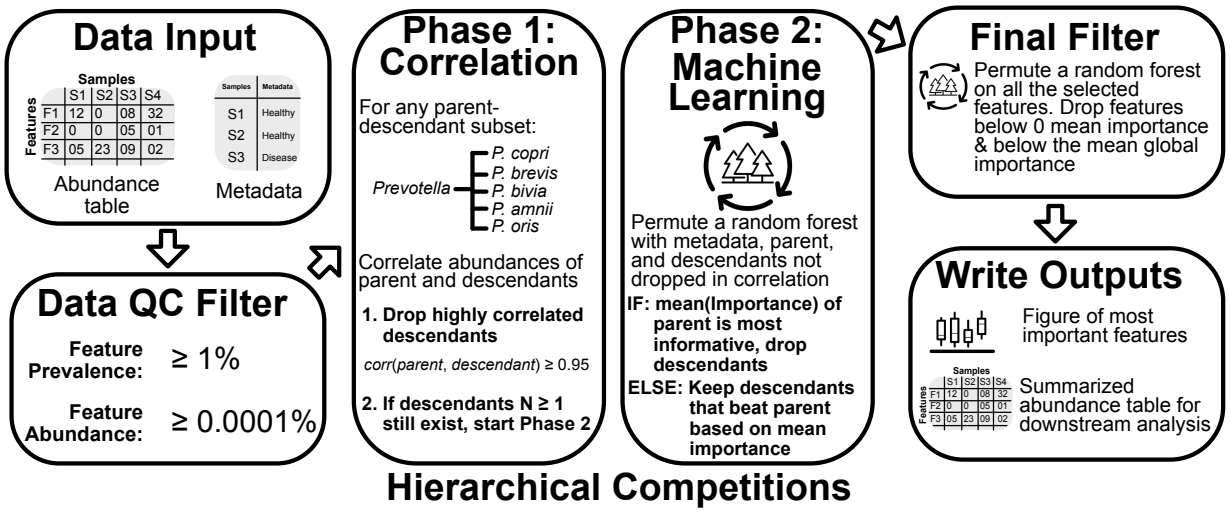

Figure 1: Outline of TaxaHFE algorithm.

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 dependent variable. An additional graph is output showing how the TaxaHFE selected taxa vary in relation to the dependent variable, plotted using ggPlot2 (Wickham, 2016).

Downstream machine learning

We analyzed hierarchical datasets using a machine learning pipeline based around the Tidymodels package (v1.0.0) (Kuhn and Wickham, 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 (3x) cross validation was used for Bayesian-search hyperparameter tuning. 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-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 was 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 and Stando, 2023) was used to plot these values.

Evaluation of TaxaHFE

Six previously published microbiome datasets (Lloyd-Price *et al.*, 2019; Wang *et al.*, 2020; Mars *et al.*, 2020; Erawijantari *et al.*, 2020; Franzosa *et al.*, 2019; Oliver *et al.*, 2022; Muller *et al.*,

2022) and a dietary tree dataset (Kable *et al.*, 2022) were used to assess TaxaHFE (Supplemental Table 3). 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_algroithm.py (Oudah and Henschel, 2018)), using two correlation cutoffs: the program's default 0.7 and 0.95 (matching TaxaHFE's initial correlation cutoff). These summarized and TaxaHFE-selected data were used in machine learning models to predict dependent variables associated with each study (Supplemental Table 3), as described above. Metrics of ML performance were evaluated within each study, using a one factor analysis of variance (ANOVA) and the post-hoc Tukey HSD test with Benjamini-Hochberg (BH) adjusted p-values.

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). 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 Supplemental Table 3. We downloaded five of the microbiome datasets from a microbiome-metabolome dataset collection (Muller *et al.*, 2022).

RESULTS

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 hierarchical feature engineering program (Oudah_70 and Oudah_95). In 5/6 previously published studies, the dimensional reduction produced by TaxaHFE (± super filter, SF) selects features which, when used as input in machine learning models, result in the higher mean ROC-AUC scores compared to data summarized at specific taxonomic levels, or a previously published HFE algorithm (Figure 2A, Supplemental Figure 1, Supplemental Table 4). The mean ROC-AUC for TaxaHFE (+SF) across all studies assessed was 0.906 (s.d. = 0.067) followed by TaxaHFE (-SF) (0.890, s.d. = 0.068). Moreover, TaxaHFE (+SF) produces the best models for 4/6 studies when assessed using balanced accuracy (Figure 2B, Supplemental Table 4) and 4/6 studies using mean Cohen's kappa (Figure 2C, Supplemental Table 4). In the other two cases, TaxaHFE preprocessed models were not significantly different from the best models produced (Supplemental Figure 1).

In addition to performance improvements, TaxaHFE (+SF) utilizes less features than comparable methods (Figure 2C). Models utilizing species-level features (after abundance and prevalence filters, see methods) utilized 645 (s.d. = 121) species on average, compared with the 97 (s.d. = 35) taxonomic features used by TaxaHFE (+SF) on average. Additionally, TaxaHFE (+SF) selects 41 less features on average compared to a previously published HFE algorithm by Oudah and Henschel (Oudah and Henschel, 2018).

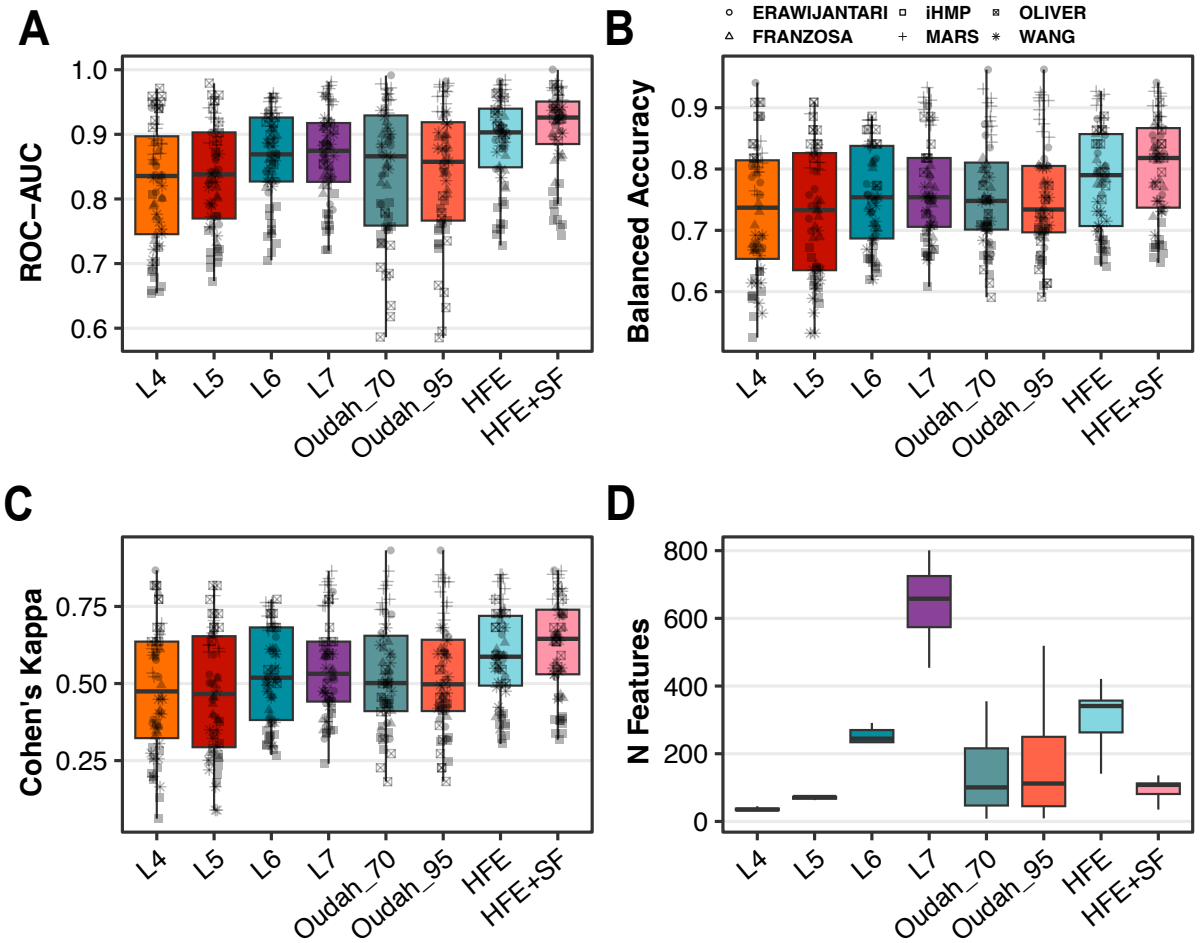

Figure 2: Comparisons made used summarized microbiome data (L4-L7, order through species), a previously published hierarchical feature engineering program (Oudah) employing two different internal correlation cutoffs (Pearson 0.7 and 0.95), and TaxaHFE with and without the super-filter. Machine learning performance metrics (A-C) and number of features used for model building (D).

TaxaHFE can reduce features for both categorical and continuous predictions.

A previous effort towards hierarchical feature engineering 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 dependent 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 vs high quartile), TaxaHFE, with or without the default super-filter, selected 35 and 141 taxon features respectively. Using these features as input to a random forest classification model predicting categorical ARG abundance (high vs low) resulted in mean ROC-AUC values of 0.939 (+SF) and 0.933 (-SF) (Figure 3). The best set of features to use in this example was L4 (order) summarized data (ROC-AUC = 0.945, Supplemental Figure 2A); however, differences in ML performance between models utilizing L4 summarized data and TaxaHFE+SF preprocessed data were not significant when assessed using a BH-corrected TukeyHSD post-hoc test. The previously published HFE algorithm by Oudah

and Henschel (Oudah and Henschel, 2018) selected 8 features (Supplemental Figure 2B), a considerable decrease from TaxaHFE (+SF) 35 features. However, this decrease came at a cost to the performance of the ML model; TaxaHFE (+SF) achieved substantially better performance metrics (balanced accuracy: TaxaHFE+SF, 0.836; Oudah_70, 0.691).

When analyzing continuous ARG abundance, TaxaHFE, with or without the default super-filter, selected 46 and 141 taxon features respectively. Using these features as input to a random forest regression model predicting continuous ARG abundance resulted in mean R^2 (coefficient of determination) values of 0.63 (+SF) and 0.615 (-SF) (Figure 3). Like the categorical example, TaxaHFE+SF performed slightly behind the best model (by mean R^2 performance), which used L4 (order) summarized data (0.641), yet BH-corrected TukeyHSD post-hoc test showed these differences were not significant, and both TaxaHFE+SF and the L4 model used nearly the same number of features (46 and 45 respectively). 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 example resulted in models utilizing similar features (Figure 3D, 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 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.

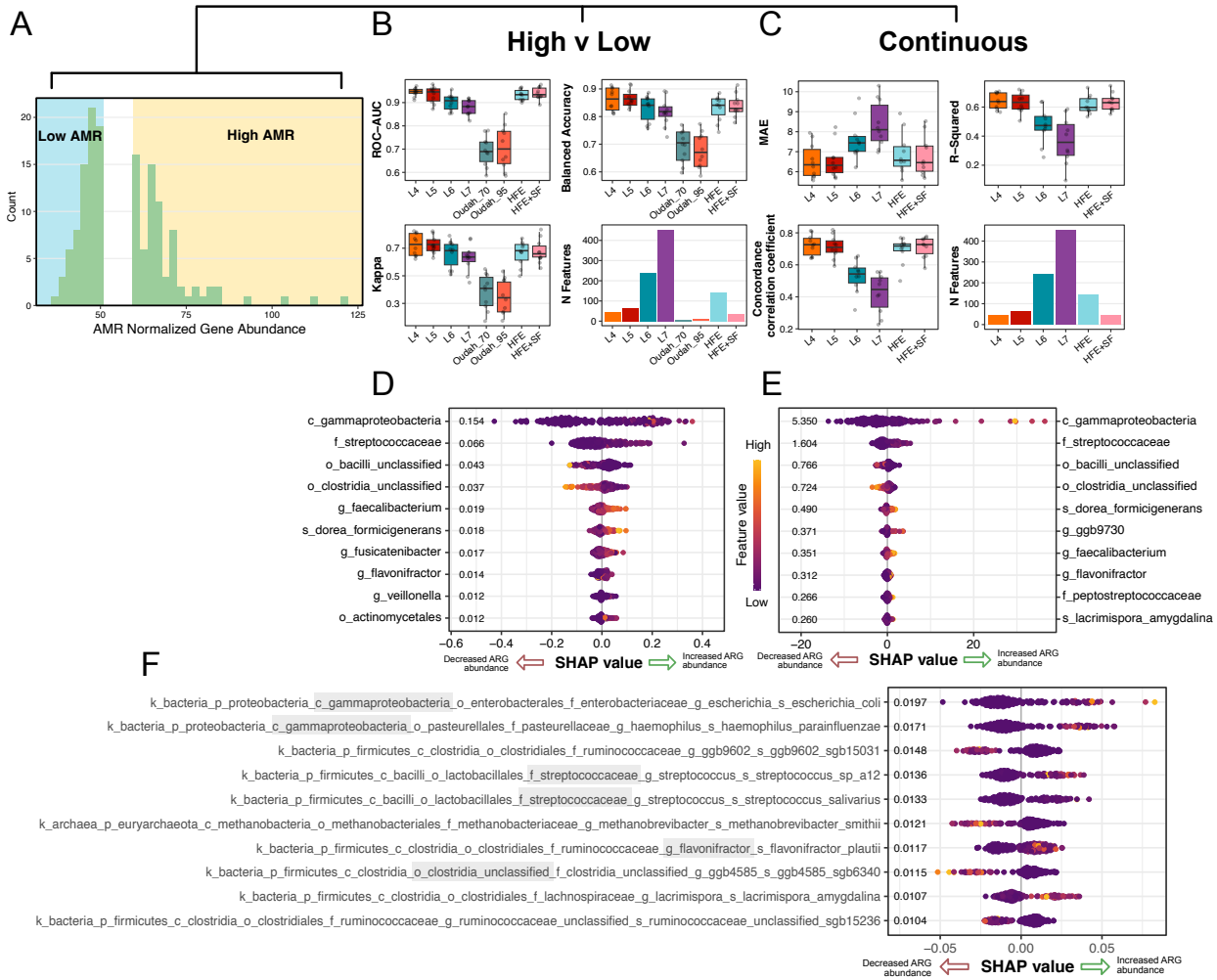

Figure 3: Analysis of TaxaHFE on categorical and continuous data. Histogram of antibiotic resistance gene abundance, showing just the top quartile and bottom quartile of the cohort (A). These data were analyzed as a categorical factor (low vs high ARG abundance) or continuous ARG abundance, and the (B,C) boxplots show the performance of machine learning models utilizing summarized taxonomic data or taxonomic hierarchical feature engineering. Formal statistical tests were run using the aov() function in R, and the model (for example) balanced_accuracy ~ taxonomic_level. (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. Grey highlighted levels share a common taxonomic group with TaxaHFE.

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 hierarchical feature engineering. To test this, we utilized a previously published dietary food tree and tested TaxaHFE's ability to select features which explain average fiber intake (Baldiviez *et al.*, 2017). After prevalence and abundance filters, 605 foods were assessed. TaxaHFE+SF selected 49 features as the most informative for predicting average fiber intake, nearly a 92% reduction from the 605 features used as input. When assessed using R^2, BH-corrected TukeyHSD post-hoc test revealed that TaxaHFE+SF performed significantly better

than summarized features in a random forest ($p < 0.05$) (Figure 4A). We next used Shapley values to determine which features were most important to a model built with TaxaHFE+SF. The most important features used in the model 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) (Figure 4B).

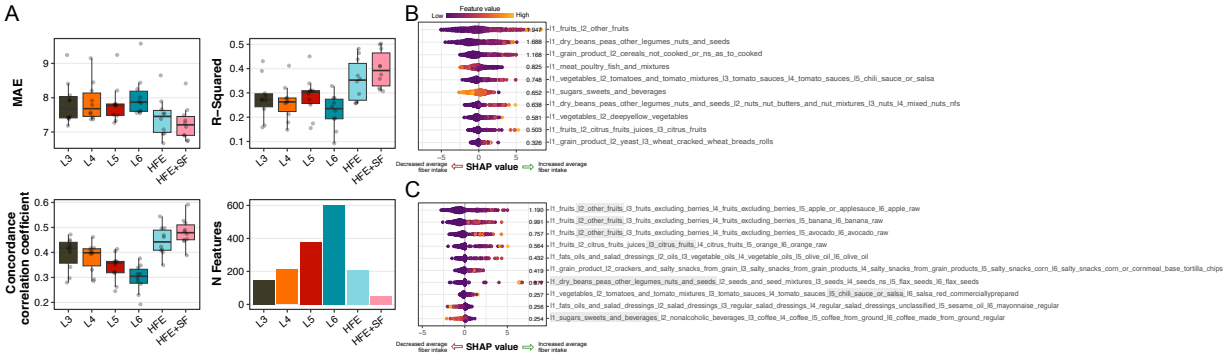

Figure 4: Testing TaxaHFE on hierarchically organized dietary data. Machine learning performance metrics of models predicting average fiber intake using food tree data (A) and number of features used for model building. Shapley values (B-C) of the most informative features from a model built using TaxaHFE+SF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

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 common 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 which 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 (Figure 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 (Oudah and Henschel, 2018) provided an excellent implementation; however, the Oudah algorithm is unable to handle a continuous dependent variable and non-microbial features. At the core of TaxaHFE is a random forest, which is a particularly capable ML algorithm for handling both continuous and categorical dependent 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.

Importantly, TaxaHFE preprocessed data improves machine learning 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 person to person, it will often mask more generalizable microbial responses within heterogeneous cohorts. Hierarchical feature engineering allows for the capture of these more generalizable responses, concomitantly increasing the accuracy of machine learning models in the process.

Consequently, 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 (Figure 3F), it is difficult to interpret the meaning of the ten species selected, each with a very low SHAP value. But with TaxaHFE (Figure 3D and 3E), 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. Moreover, 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). In the diet example (Figure 4C), the model based on the lowest taxonomic level is again difficult to interpret, with 605 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, and uncooked cereals and high intake of meat/poultry/fish as predictive 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. While it is not a memory intensive program, it does rely on relatively inefficient for-loops to iterate over each taxon in each taxonomic level. For example, in a microbiome dataset with 4640 features, TaxaHFE took 3 min 1 sec, whereas the Oudah algorithm took only 13 seconds 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 machine learning algorithms in that large high-quality data sets are needed. Without sufficient information in the data (i.e., no relationship between the dependent variable and the features) all ML algorithms will suffer. However, what may not be obvious to users is that in the absence of information, machine learning algorithms will still report results, which often represent noise. Methods for avoiding pitfalls of machine learning 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.

CONCLUSION

We demonstrate that TaxaHFE dramatically reduces the feature space of hierarchically organized data while generally increasing the performance of downstream machine learning models. While our examples come from microbiology and nutrition research, TaxaHFE could be used with any

data set 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 hierarchical feature engineering can lead to more accurate and interpretable models.

ACKNOWLEDGEMENTS

We would like to acknowledge Sarah Blecksmith for useful feedback provided while code testing. We would like to thank Rachel Waymack and Jules Larke for thoughtful comments and edits. A.O. was supported by an ORISE fellowship administered by Oak Ridge Associated Universities through an interagency agreement between the U.S. Department of Energy and the U.S. Department of Agriculture, Agricultural Research Service.

FUNDING INFORMATION

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.

SUPPLEMENTAL FIGURES

Supplemental Figure 1: Pairwise comparisons (Tukey HSD) within each study of ROC-AUC (A), balanced accuracy (B), and Cohen's Kappa (C). The green highlight represents the best model by averaging results from ten different random seeds.

Supplemental Figure 2: Shapley values from a RF model predicting categorical ARG abundance (high vs low) with the L4 features (A) the Oudah_95 (B) engineered features as input.

SUPPLEMENTAL TABLES

Supplemental Table 1: Example dataset of taxon-parsed features.

Supplemental Table 2: Example dataset of complete taxonomic features.

Supplemental Table 3: List of studies used in comparisons.

Supplemental Table 4: Performance of ML models using summarized data and hierarchical feature engineering algorithms. These data are visualized in Figure 2.

REFERENCES

Baldiviez,L.M. *et al.* (2017) Design and implementation of a cross-sectional nutritional

phenotyping study in healthy US adults. *BMC Nutr. 2017 31*, **3**, 1–13.
Bellman,R.E. (2003) Dynamic Programming Courier Corporation.
Bevilacqua,S. *et al.* (2021) The use of taxonomic relationships among species in applied ecological research: Baseline, steps forward and future challenges.
Blanco-Míguez,A. *et al.* (2023) Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol. 2023*, 1–12.
Choi,Y. *et al.* (2022) A Guide to Dietary Pattern–Microbiome Data Integration. *J. Nutr.*, **152**, 1187–1199.
Erawijantari,P.P. *et al.* (2020) Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut*, **69**, 1404–1415.
Franzosa,E.A. *et al.* (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.*
Greenwell,B. (2021) fastshap: Fast Approximate Shapley Value.
Hand,D.J. and Till,R.J. (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.*, **45**, 171–186.
Jacobs,D.R. and Steffen,L.M. (2003) Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am. J. Clin. Nutr.*, **78**, 508S-513S.
Johnson,A.J. *et al.* (2019) Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe*, **25**, 789-802.e5.
Kable,M.E. *et al.* (2022) Tree-Based Analysis of Dietary Diversity Captures Associations Between Fiber Intake and Gut Microbiota Composition in a Healthy US Adult Cohort. *J. Nutr.*, **152**, 779–788.
Kleine Bardenhorst,S. *et al.* (2021) Data Analysis Strategies for Microbiome Studies in Human Populations—a Systematic Review of Current Practice. *mSystems*.
Kuhn,M. *et al.* (2023) yardstick: Tidy Characterizations of Model Performance.
Kuhn,M. and Wickham,H. (2020) Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.
Lloyd-Price,J. *et al.* (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*.
Mars,R.A.T. *et al.* (2020) Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell*.
Martiny,A.C. *et al.* (2012) Phylogenetic conservatism of functional traits in microorganisms. *ISME J. 2013 74*, **7**, 830–838.
Martiny,J.B.H. *et al.* (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science (80-.).*, **350**.
Mayer,M. and Stando,A. (2023) shapviz: SHAP Visualizations.
Muller,E. *et al.* (2022) The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms Microbiomes*.
Nembrini,S. *et al.* (2018) The revival of the Gini importance? *Bioinformatics*, **34**, 3711–3718.
Oliver,A. *et al.* (2022) Association of Diet and Antimicrobial Resistance in Healthy U.S. Adults. *MBio*, **13**.
Oliver,A. *et al.* (2021) High-Fiber, Whole-Food Dietary Intervention Alters the Human Gut Microbiome but Not Fecal Short-Chain Fatty Acids. *mSystems*, **6**.
Oudah,M. and Henschel,A. (2018) Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics*, **19**, 1–13.
Stalder,T. *et al.* (2019) Linking the resistome and plasmidome to the microbiome. *ISME J. 2019*

1310, **13**, 2437–2446.
Štrumbelj,E. and Kononenko,I. (2014) Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*
Wang,X. *et al.* (2020) Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. *Gut*.
Wetterstrand.,K.A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
Whalen,S. *et al.* (2021) Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet. 2021 233*, **23**, 169–181.
Wickham,H. (2016) ggplot2 Elegant Graphics for Data Analysis.
Wright,M.N. and Ziegler,A. (2017) ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.*, **77**.