EXPLANA: A user-friendly workflow for EXPLoratory ANAlysis and feature selection in cross-sectional and longitudinal microbiome studies

Jennifer Fouquier, Maggie Stanislawski, John O'Connor,
Ashley Scadden, Catherine Lozupone

Department of Biomedical Informatics, School of Medicine,
University of Colorado, Anschutz Medical Campus, Aurora, CO

Abstract

The potential for disease treatment through gut microbiome modification has contributed to an increase in longitudinal microbiome studies (LMS). Gut microbiome modification can occur through factors such as diet, probiotics, or fecal transplants. Scientific data often motivates researchers to perform exploratory analyses to identify features that relate to a response. However, LMS are challenging to analyze, often leading to lost information and research barriers. LMS analytic challenges include data integration, compositionality, dimensionality reduction, and the need for mixed-effects models for non-independent data. Additionally, LMS can be observational or interventional, and relevant comparisons of interest might differ for these two study types. For example, in an observational study, measurements are made over time and show natural fluctuations in symptoms/measurements, so the baseline measurement might not be a reference point of primary interest; whereas, in an interventional study, the baseline value often coincides with the start of treatment and is a key reference point. Thus, the optimal way to calculate feature changes for each subject over time is dependent on different reference values. To address these challenges, we developed EXPLANA, a data-driven feature-selection workflow that supports numerical and categorical data. We implemented machine-learning models for repeated measures, feature-selection methods, and visualizers explaining how selected features relate to the response. With one script, analysts can build models to select and evaluate important features and obtain an analytic report that textually and graphically summarizes results. EXPLANA had good performance using twenty simulated data models yielding an average area under the curve (AUC) of 0.91 (range: 0.79-1.0; SD = 0.05) and better performance compared to an existing tool (AUC: 0.95 and 0.56; precision: 0.82, and 0.14, respectively). EXPLANA is a flexible, data-driven tool that simplifies LMS analyses and can identify unique features that are predictive of outcomes of interest through a straightforward workflow.

Background/Introduction

In our current era, scientific studies often include the collection of complex multiomic data[1], such as microbiome[2], transcriptome[3], or metabolome[4], where it is of interest to determine whether any novel features, or collections of features, may be related to a response in an exploratory manner. Adding to the complexity, studies often collect other data from individuals that may impact an outcome, such as demographic and health data, or surveys on diet or medications. The growing quantity of available data complicates statistical decisions regarding variable inclusion, which is often based on hypotheses that motivated the initial study design. Additionally, studies can include both categorical and numerical variables and can often contain longitudinal, non-independent data, posing greater statistical challenges. As research advancements are made, collaborative efforts with different research laboratories produce more data per study, and human biases are often introduced during study design and analytics.

These challenges have ultimately stimulated a growing interest in data-driven methods. Therefore, we developed a data-driven feature selection workflow that streamlines exploratory analyses for hypothesis generation, accommodating longitudinal data and both numerical and categorical variables.

One field particularly impacted by an abundance of data is microbiome research, which focuses on characterizing the community of viruses, fungi and bacteria and their genes found in different environments. Characterization of the microbiome is often performed by 16S ribosomal RNA (rRNA) gene sequencing, which identifies the microorganisms in an environment. One well-studied microbial environment is the gut microbiome because of the metabolic potential of the bacterial community and its association with numerous human diseases, including obesity[5], depression[6] , autism[7], cancer[8,9], HIV[10] and cardiovascular disease[11]. The relationship between the gut microbiome and human disease suggests that gut microbiome modification through interventions like dietary changes, probiotics, or fecal microbial transplants may provide treatment options for many diseases.

To understand changes in health response and address the impact of individual variation, longitudinal studies that collect data from multiple individuals, at different timepoints, are essential. These studies involve repeated measurements on individuals, requiring special statistical considerations to identify relationships between features within non-independent data[12]. These studies often include diverse subject data, such as demographics, nutrition, or health symptom questionnaire data, often with both numerical and categorical features. Random Forest (RF)[13] based machine learning (ML) approaches are powerful for combining different data types to predict outcomes and identify important features. RFs work well with high-dimensional data (more features than samples/instances)[14], find non-linear relationships, work with non-normal data distributions, and are more interpretable than many other ML models because they are based on simple decision trees. Additionally, mixed-effects RF (MERF)[15] models can be used for longitudinal study designs. However, numerous challenges can hinder effective application of these methods.

MERFs can be run on original (raw) data from longitudinal studies or by using changes (Δs) between different reference timepoints, which can reveal unique insights in some studies[16–20]. However, the research question of interest can affect decisions regarding optimal calculation of Δs. In some designs, such as interventions, or other observational studies with an expected trend over time (e.g., gut microbiome changes over the first years of a baby's life[16]), changes are expected compared to a baseline reference value, so Δs can be calculated using baseline as a reference[17,18]. However, some observational studies have no meaningful baseline, and it might instead be of interest to relate an outcome variable to changes in predictors between adjacent timepoints or all pairs of timepoints[21,22]. For instance, in an observational longitudinal study of children with autism spectrum disorder (ASD) that we conducted[22], we evaluated children with ASD over time to identify relationships between diet, gastrointestinal distress, or the microbiome and ASD-associated behaviors. Because of high interpersonal variation in gut microbiome, this longitudinal study design revealed relationship between the gut microbiome and ASD behaviors as a correlation between the degree of microbiome change and ASD behavior change between timepoints. However, because we studied more than two timepoints, and because the first timepoint was not a meaningful baseline, we performed pairwise analyses. Pairwise analysis is useful for identification of effects that are time-delayed (i.e., a change from time 2 to time 4), order-dependent, or reference dependent. Different longitudinal study designs highlight the importance of understanding changes (Δs) in features, for each subject over time, and that feature changes differ depending on their reference values. Statistical methods do differ regarding how and when to apply change analysis and can even lead to different conclusions[23]. Thus, our method compares results from original and Δ datasets for a more complete picture of a longitudinal study.

Another analytic challenge encountered in the application of RFs to complex microbiome studies is the integration of microbiome data as a predictor value together with other data types. Data collected in surveys/questionnaires or from clinical reports can be numerical and categorical. To the best of our knowledge, there are no software tools that create and select order-dependent categorical feature changes that impact a response. For example, the drugs amiodarone and quinidine for heart arrhythmia treatment have an interaction that could lead to a dangerously rapid heartbeat[24], but an interaction risk is higher if amiodarone precedes quinidine since amiodarone has a much longer elimination half-life (days[25] vs hours[26]). This example highlights how calculating categorical Δs in an order-dependent way might uncover relationships that have differential impact if introduced in opposite order, such as in crossover study designs (AB/BA designs). This led us to the hypothesis that we could identify unique features dependent on different contexts of change, including novel order-dependent categorical features by tracking text changes as an engineered feature value (e.g., "amiodarone__quinidine").

Finally, another key challenge is the analytic complexity of performing longitudinal microbiome analysis in a reproducible way that facilitates communication about results. These workflows can involve inputs of diverse data types, calculation of Δs with different reference points, feature selection using mixed-effects ML methods, and methods for explaining why each feature was selected, in addition to their importance ranks. Although there are tools for feature selection in microbiome data[16,27–31], none provide the combination of methods we describe. For example, timeOmics[31] is useful for multi-omic integration with an emphasis on time as the response, while we wanted to find features related to different response variables over time. QIIME 2 longitudinal feature-volatility[16] allowed for looking at different responses, but did not incorporate importance metrics that explain the selected feature's impact on the response. Both tools, although useful for feature selection in longitudinal data, did not incorporate categorical deltas like we needed. It is cumbersome for scientists to research and implement this complex array of tools.

For these reasons, we developed a software workflow, called EXPLANA (EXPLoratory ANAlysis) to streamline identification of important features in both cross-sectional and longitudinal microbiome studies and reduce analytic barriers to data-driven hypothesis generation. Our tool identifies unique features important in different contexts of change, including important changes in categorical features related to a response. EXPLANA combines novel methods and popular tools to address various analytic challenges and provide broad applicability for scientific research.

Results

Overview

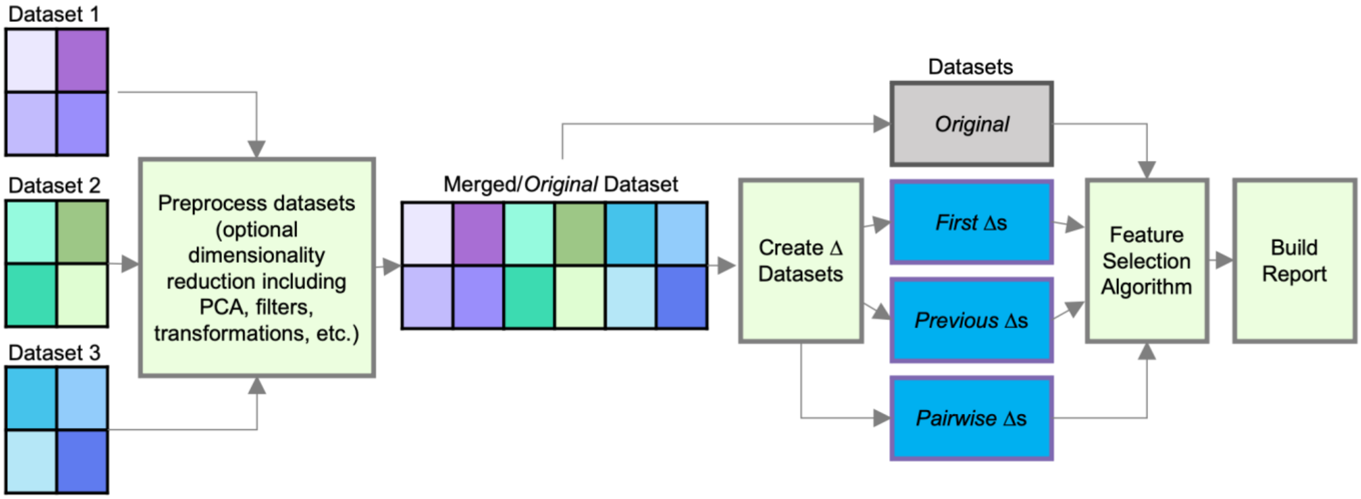

Dataset 1
Dataset 2
Dataset 3
Preprocess datasets (optional dimensionality reduction including PCA, filters, transformations, etc.)
Merged/*Original* Dataset
Create Δ Datasets
Datasets
Original
First Δs
Previous Δs
Pairwise Δs
Feature Selection Algorithm
Build Report

Figure 1. Feature selection workflow diagram. Individual datasets can be preprocessed individually to reduce dimensionality using a variety of methods, including principal components analysis (PCA), center-log-ratio (CLR) transformation or through data filtering. Datasets are merged to form the *Original* dataset prior to creation of Δ datasets for longitudinal studies. *First*, *Previous*, *Pairwise* Δ datasets, as shown in blue, are created as explained in Fig 2. Feature selection is performed for up to four models built from each dataset (*Original, First*, *Previous*, and *Pairwise*). An .html report is created which summarizes features selected per model.

EXPLANA was developed to create a comprehensive feature selection report. The workflow is guided by directions from a configuration file where the user provides paths to the different input datasets, defines whether to perform optional preprocessing steps per dataset, and defines input variables and the response variable of interest (Fig 1). The input datasets are merged to form the comprehensive *Original* dataset. If more than one timepoint is sampled per subject, feature changes are calculated using different reference points to uncover important features in different contexts of change. Thus, the *Original* dataset is used to compute three Δ datasets, *First*, *Previous*, and *Pairwise* (Fig 2). For *First*, differences are calculated compared to baseline/first measures. For *Previous*, differences are calculated compared to the immediately previous timepoint only. For *Pairwise*, all pairwise comparisons between timepoints are computed. Notation throughout is as follows: Timepoints 1, 2, 3, etc. are referred to as T1, T2, T3, etc., respectively. Accordingly, differences/changes (Δs) such as the difference between T1 and T2 is T1_T2, which is labeled in chronological order for clarity. For T1_T2, T1 is the reference point and is subtracted from T2 (Fig 2).

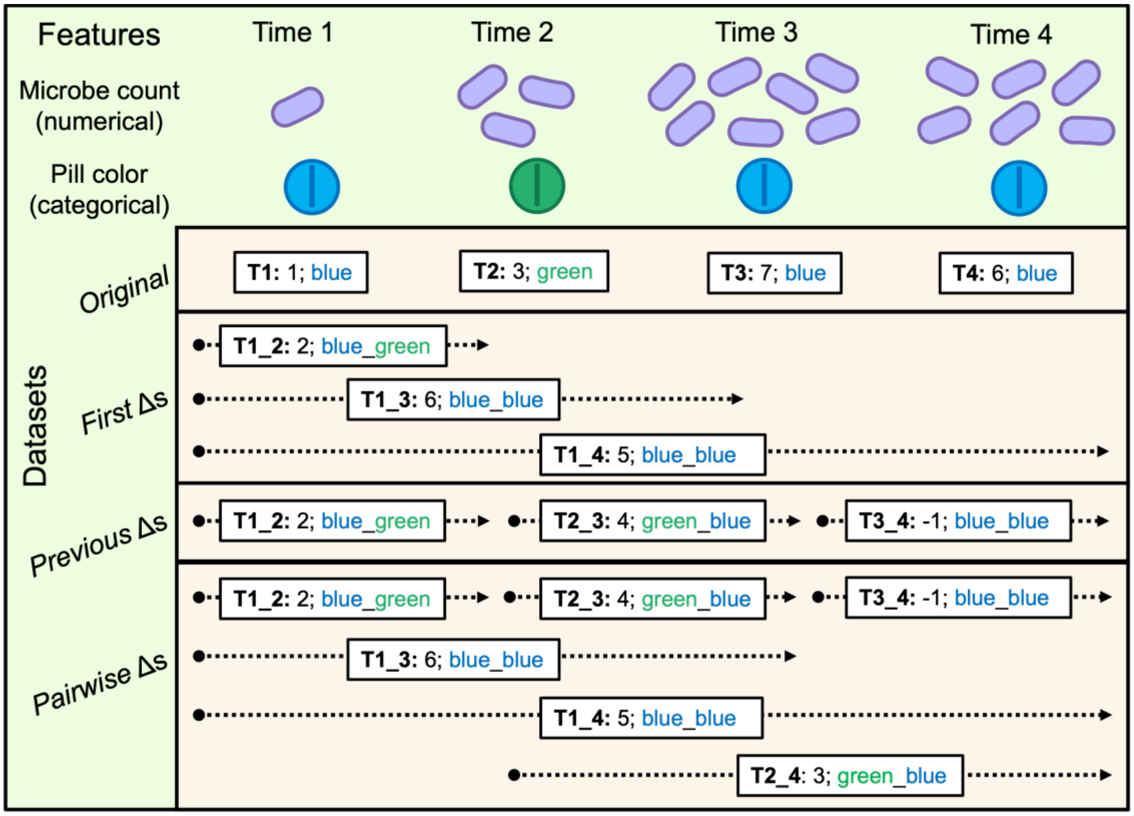
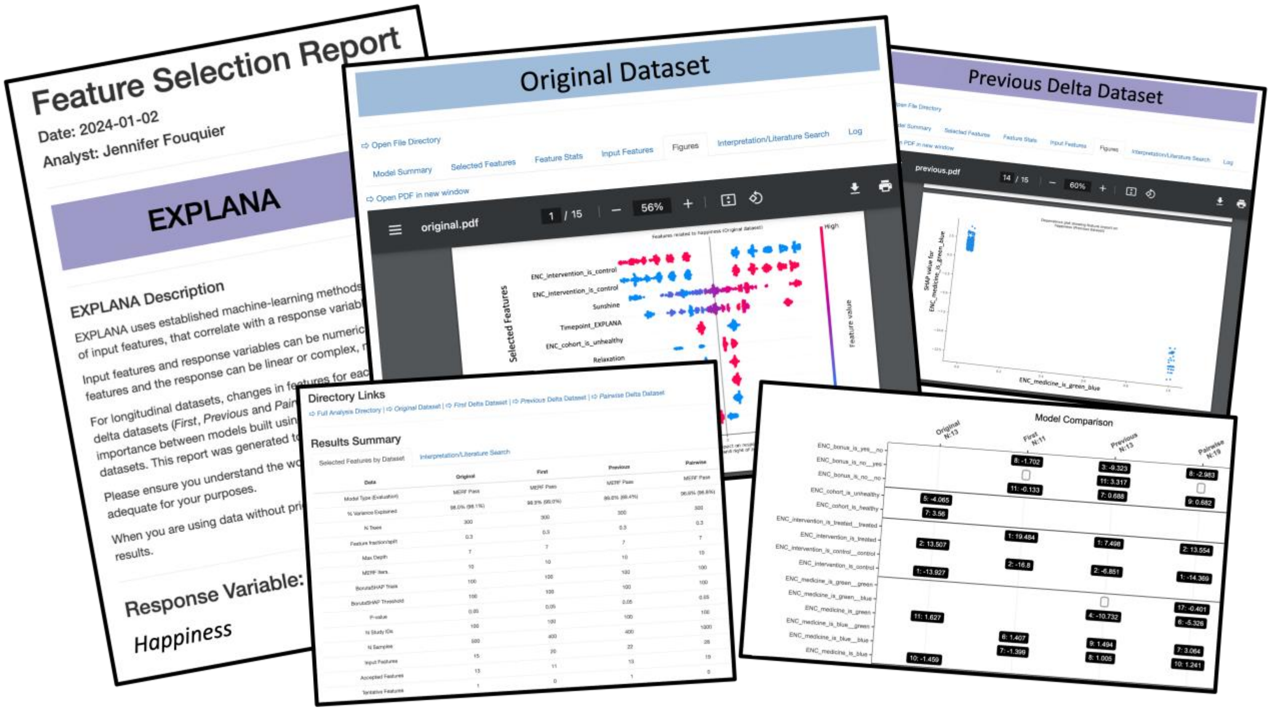


Figure 2. Example calculations of *First*, *Previous* and *Pairwise* ∆ datasets for numerical and categorical features in a four timepoint study. The *Original* dataset includes original feature values (without change analysis) and ∆ datasets contain differences/changes in features, per subject, between timepoints. Different reference points are used for each ∆ dataset. For categorical features (e.g. pill color), text is used to track the order of categorical changes and for numerical features (e.g. microbe count), the reference is subtracted from the later timepoint. The comparison between two timepoints is indicated before each colon (e.g., a ∆ between timepoint 1 and 2 is indicated as T1_2). For a four timepoint study: *First* ∆s: 1_2, 1_3, 1_4; *Previous* ∆s: 1_2, 2_3, and 3_4; and *Pairwise* ∆s: 1_2, 2_3, 3_4, 1_3, 1_4, and 2_4.

Support for numerical and categorical predictors and response variables is implemented, including novel functionality to track categorical feature changes over time. Microbiome-specific challenges addressed include the option to use a center-log-ratio (CLR) transformation for compositional data[32] as well as accommodating distance matrix incorporation during ∆ calculations, which allows users to evaluate differences between microbiome samples, such as those calculated with UniFrac or other beta diversity measures[33]. MERFs are used as the ML method for feature selection in non-independent, repeated measures data, while RFs are used when repeated measures are excluded. The Boruta[34] method combined with SHAP[35] (BorutaSHAP[36]) is used to identify which important features identified with MERFs or RFs contribute to a more accurate prediction of the outcome better than expected by random chance. SHAP not only provides importance scores that rank features by their importance for model performance, but also produces plots that allow one to assess whether features have a positive or negative impact on a response, thereby improving results interpretation. Upon workflow completion, a report is generated that contains a description of the analysis, as well as tables and figures that explain why features were selected (Fig 3).



Feature Selection Report

Date: 2024-01-02

Analyst: Jennifer Fouquier

EXPLANA

EXPLANA Description

EXPLANA uses established machine-learning methods
of input features, that correlate with a response variab

Input features and response variables can be numeric
features and the response can be linear or complex, r

For longitudinal datasets, changes in features for eac
delta datasets (*First*, *Previous* and *Pair*
importance between models built usin
datasets. This report was generated t

Please ensure you understand the wo
adequate for your purposes.

When you are using data without pri
results.

Response Variable:

Happiness

Directory Links

⇨ Full Analysis Directory | ⇨ Original Dataset | ⇨ First Delta Dataset | ⇨ Previous Delta Dataset | ⇨ Pairwise Delta Dataset

Results Summary

Data	Original	First	Previous	Pairwise
Model Type (Evaluation)	MERF Pass	MERF Pass	MERF Pass	MERF Pass
% Variance Explained	98.0% (98.1%)	99.9% (99.0%)	99.0% (99.4%)	96.8% (96.8%)
N Trees	300	300	300	300
Feature fraction/split	0.3	0.3	0.3	0.3
Max Depth	7	7	7	7
MERF Iters	10	10	10	10
BorutaSHAP Trials	100	100	100	100
BorutaSHAP Threshold	100	100	100	100
P-value	0.05	0.05	0.05	0.05
N Study IDs	100	100	100	100
N Samples	500	400	400	1000
Input Features	15	20	22	28
Accepted Features	13	11	13	19
Tentative Features	1	0	1	0

Figure 3. Example screenshots of a feature selection report. The report is in an interactive .html format to facilitate interpretation of complex machine-learning method results through figures and written explanations. The report includes a written summary of methods containing arguments and information from the configuration file. The dynamic methods can be copy/pasted into manuscripts for efficiency; tables and figures summarize feature selection results for models built from *Original* and, if longitudinal, *First*, *Previous* and *Pairwise* Δ datasets. Figures include SHAP summary plots and SHAP dependence plots. Links are provided to directories containing files used for report generation to facilitate exploration of data and results.

Simulated Data Design

The workflow was evaluated using both simulated and published datasets (both detailed in *Methods*). A simulated longitudinal happiness study, *SimFeatures*, was created for performance evaluation and modeled as an intervention where 100 individuals were treated with one of two therapies to improve happiness over five timepoints. Happiness is based on a numerical score where higher values indicate better mental health. For interpretability, features are recognizable as factors that could affect real-life happiness such as relaxation, sunshine, salary, medication, etc. Categorical and numerical features were included with and without relationships to the response, thus representing predictive and not predictive engineered features (Table S1). Some features were designed to be important only in some of the four models (Fig 2; *Original, First, Previous* or *Pairwise*) to validate whether the tool could select unique features dependent on different contexts of change. To show how order-dependent categorical changes are found, a pill color/medicine variable was created to mimic a drug interaction between blue and green pills resulting in lower happiness. Specifically, green *before* blue (green_blue) negatively impacted happiness due to green having a longer half-life than blue.

A simulated compositional, microbiome feature table was also created using MicrobiomeDASim[37] with 25 differentially abundant microbes linearly correlated to happiness changes over time and 175 that are not related. The dataset with the response and simulated microbes was called *SimMicrobiome*. The dataset with *SimFeatures* and *SimMicrobiome* combined was called *SimFeaturesMicrobiome0.* To evaluate the effects of including many features with no relation to the response in an analysis, we added an increasing number of random variables (number indicated in name) from a variety of not predictive data distributions to the *SimFeaturesMicrobiome* dataset. Thus, the five simulated datasets are *SimFeatures*, *SimMicrobiome*, *SimFeaturesMicrobiome0*, *SimFeaturesMicrobiome500*, and *SimFeaturesMicrobiome1000*. (Table S1).

For features in the five simulated datasets, workflow evaluation was performed by appropriate selection or rejection of engineered features. Accordingly, true positives (*selected* predictive features; TPs) and true negatives (*rejected* not-predictive features; TNs) were considered correctly classified, while false positives (*selected* not-predictive features; FPs) and false negatives (*rejected* predictive features; FNs) were considered incorrectly classified (Fig S1). These datasets allowed us to: 1) evaluate workflow performance from classification accuracy of engineered features and 2) to test our hypothesis that we could identify unique features dependent on different contexts of change, including novel order-dependent categorical changes related to a response.

Simulated Data Results

We used EXPLANA to select features related to happiness for the five simulated datasets using the four models (*Original, First, Previous* or *Pairwise*). Features were ranked and performance was determined for each of the four models for all five datasets (Fig 4, Table 1, Fig S2). AUC and F1-score (a metric that accounts for both precision and recall) respective ranges for *Original, First, Previous* and *Pairwise* were: 0.79-1.00 and 0.83-1.00 *SimFeatures*; 0.80-0.96, and 0.73-0.87 for *SimMicrobiome*; 0.88-0.94 and 0.78-0.91 for *SimFeaturesMicrobiome0*; 0.87-0.95 and 0.69-0.91 for *SimFeaturesMicrobiome500*; and 0.90-0.95 and 0.66-0.92 for *SimFeaturesMicrobiome1000.* Thus. *Original* yielded the highest F1-score and AUCs. The average workflow ability to recall predictive features was good/excellent (average 0.87, SD = 0.09) and good for precision (avg 0.82, SD = 0.14), with some ∆ datasets having lower precision or recall. Of the four models analyzed with EXPLANA for *SimMicrobiome*, *Previous* had the lowest percent variation explained (65.4%), a low recall (0.60), and failed to correctly classify some predictive features. The lowest F1-score was for *SimFeaturesMicrobiome1000* using *Pairwise* (0.66), which had 30 FPs that affected precision (0.55). Still for *SimFeaturesMicrobiome1000*, the proportion of selected predictive features (recall) was 0.81, AUC was 0.91, and out of 1000 random variables, only 30 were FPs (see confusion matrix in Table 1). For simulated datasets with not-predictive microbes or random variables, *Pairwise* had the poorest precision.

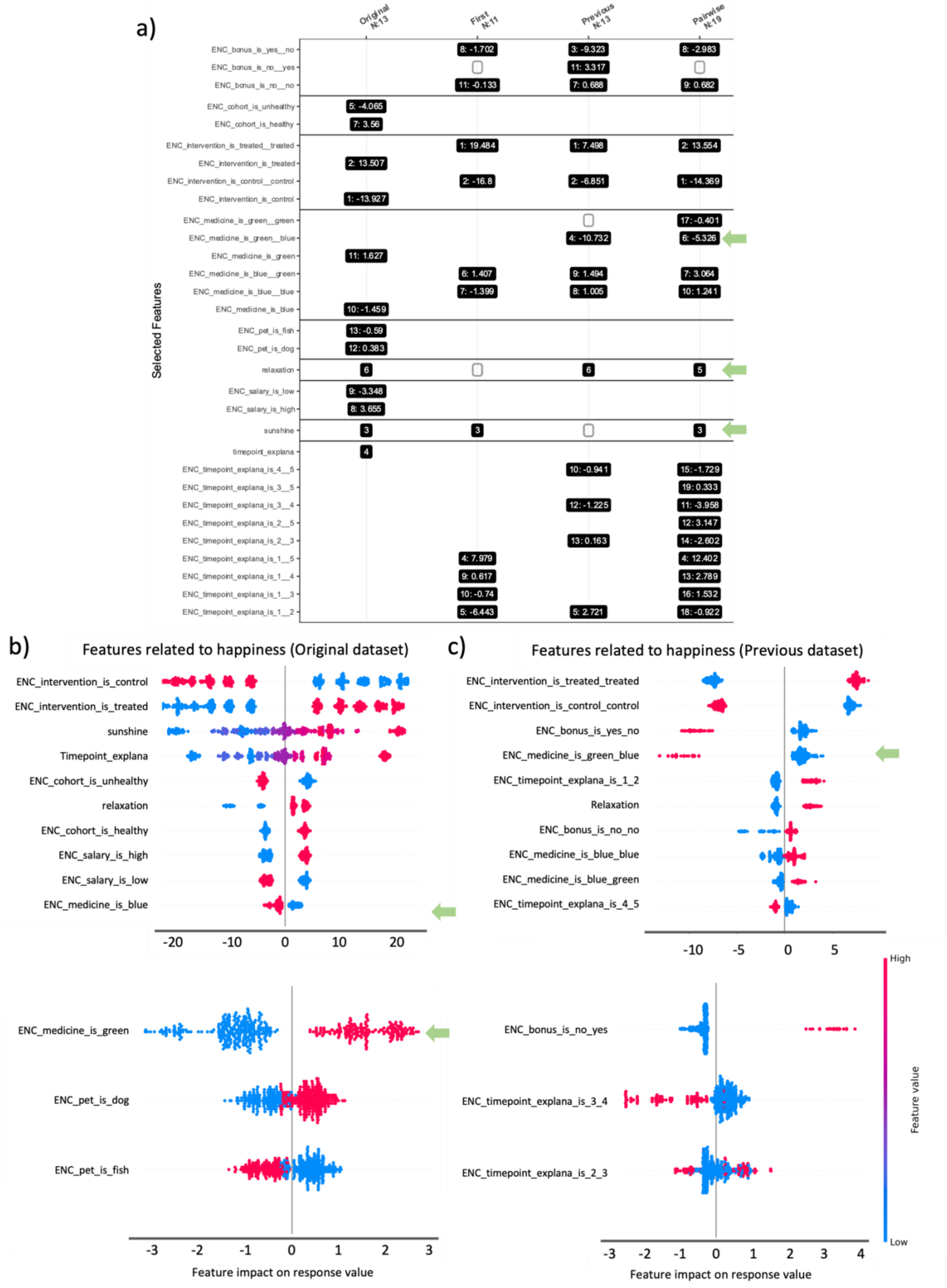
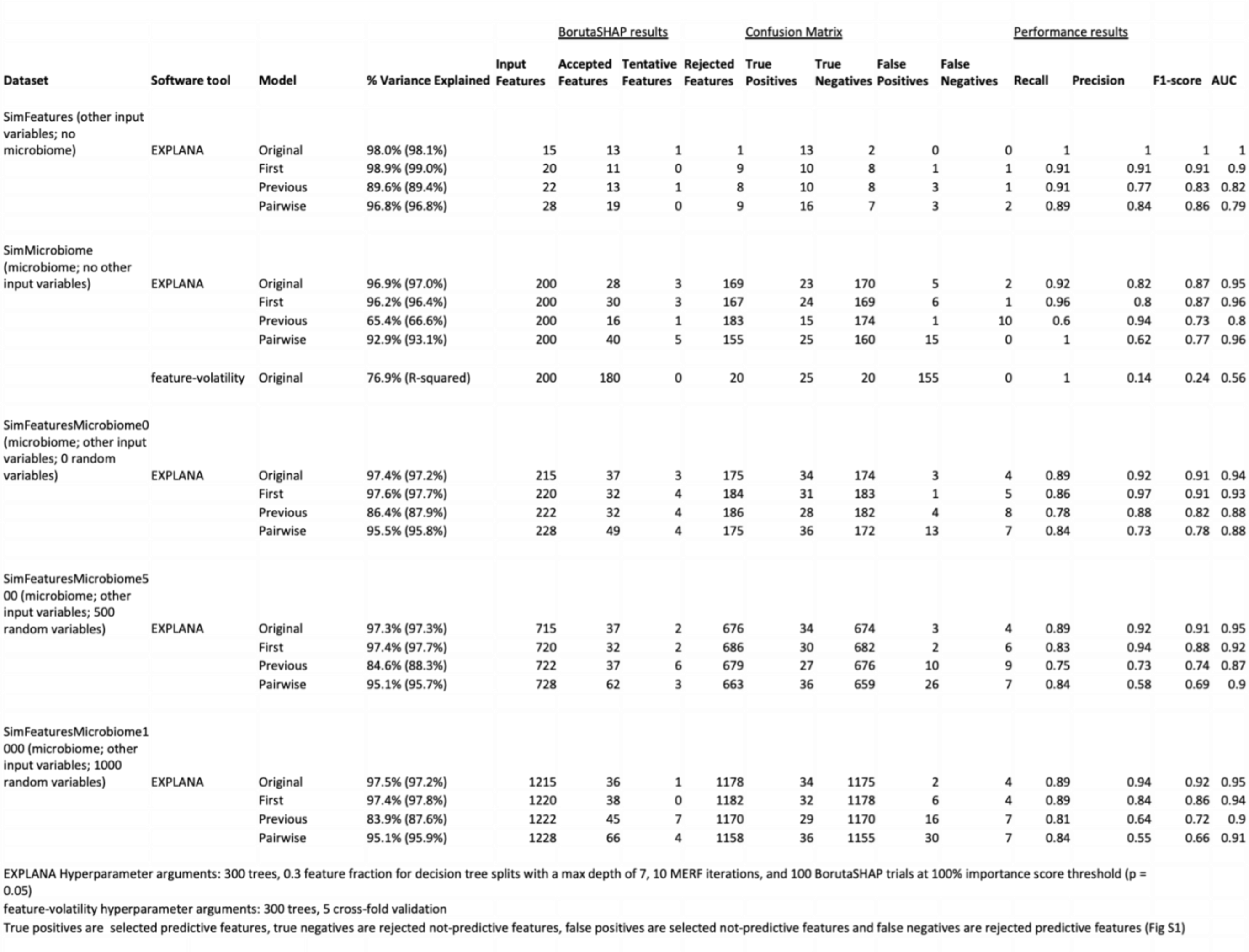

a)
Original N:13
First N:11
Previous N:13
Pairwise N:19
Selected Features
ENC_bonus_is_yes__no: 8: -1.702; 3: -9.323; 8: -2.983
ENC_bonus_is_no__yes: □; 11: 3.317; □
ENC_bonus_is_no__no: 11: -0.133; 7: 0.688; 9: 0.682
ENC_cohort_is_unhealthy: 5: -4.065
ENC_cohort_is_healthy: 7: 3.56
ENC_intervention_is_treated__treated: 1: 19.484; 1: 7.498; 2: 13.554
ENC_intervention_is_treated: 2: 13.507
ENC_intervention_is_control__control: 2: -16.8; 2: -6.851; 1: -14.369
ENC_intervention_is_control: 1: -13.927
ENC_medicine_is_green__green: □; 17: -0.401
ENC_medicine_is_green__blue: 4: -10.732; 6: -5.326
ENC_medicine_is_green: 11: 1.627
ENC_medicine_is_blue__green: 6: 1.407; 9: 1.494; 7: 3.064
ENC_medicine_is_blue__blue: 7: -1.399; 8: 1.005; 10: 1.241
ENC_medicine_is_blue: 10: -1.459
ENC_pet_is_fish: 13: -0.59
ENC_pet_is_dog: 12: 0.383
relaxation: 6; □; 6; 5
ENC_salary_is_low: 9: -3.348
ENC_salary_is_high: 8: 3.655
sunshine: 3; 3; □; 3
timepoint_explana: 4
ENC_timepoint_explana_is_4__5: 10: -0.941; 15: -1.729
ENC_timepoint_explana_is_3__5: 19: 0.333
ENC_timepoint_explana_is_3__4: 12: -1.225; 11: -3.958
ENC_timepoint_explana_is_2__5: 12: 3.147
ENC_timepoint_explana_is_2__3: 13: 0.163; 14: -2.602
ENC_timepoint_explana_is_1__5: 4: 7.979; 4: 12.402
ENC_timepoint_explana_is_1__4: 9: 0.617; 13: 2.789
ENC_timepoint_explana_is_1__3: 10: -0.74; 16: 1.532
ENC_timepoint_explana_is_1__2: 5: -6.443; 5: 2.721; 18: -0.922
b)
Features related to happiness (Original dataset)
ENC_intervention_is_control
ENC_intervention_is_treated
sunshine
Timepoint_explana
ENC_cohort_is_unhealthy
relaxation
ENC_cohort_is_healthy
ENC_salary_is_high
ENC_salary_is_low
ENC_medicine_is_blue
-20
-10
0
10
20
ENC_medicine_is_green
ENC_pet_is_dog
ENC_pet_is_fish
-3
-2
-1
0
1
2
3
Feature impact on response value
c)
Features related to happiness (Previous dataset)
ENC_intervention_is_treated_treated
ENC_intervention_is_control_control
ENC_bonus_is_yes_no
ENC_medicine_is_green_blue
ENC_timepoint_explana_is_1_2
Relaxation
ENC_bonus_is_no_no
ENC_medicine_is_blue_blue
ENC_medicine_is_blue_green
ENC_timepoint_explana_is_4_5
-10
-5
0
5
ENC_bonus_is_no_yes
ENC_timepoint_explana_is_3_4
ENC_timepoint_explana_is_2_3
-3
-2
-1
0
1
2
3
4
Feature impact on response value
High
Feature value
Low

Figure 4. Feature selection results for one analysis using *SimFeatures* dataset. The *SimFeatures* dataset is a simulated longitudinal intervention with 100 individuals sampled over five timepoints (see *Methods* for detailed description). A) Feature occurrence figure with selected features organized by model (*Original, First, Previous* and *Pairwise)* and ranked with one being the highest importance. For presence (yes/1) of a categorical variable, SHAP values are indicated following the colon. SHAP summary beeswarm plots for ranked selected features are shown for B) *Original* model and C) *Previous* model. Each point represents one sample, and the horizontal position indicates impact on the response as indicated on the x-axis. Points to the left indicate a negative impact, and points to the right indicate a positive impact. The colors represent the selected feature values, where red is larger, and blue is smaller. For binary encoded features ('ENC') red is yes/1 and blue is no/0. Note that scales differ between the top and bottom SHAP plots, as they are grouped by a maximum of ten features per SHAP plot. Some features were designed to be identified in only certain models. Interesting results include: ENC_medicine_is_green_blue (pill color), a categorical feature important in *Previous* and *Pairwise* models; "relaxation", a numerical feature important in *Previous* and *Pairwise* and undetectable in *First* as described in the discussion; and "sunshine" which was unable to be detected in Previous but a high rank in other models. Multiple analyses create a more comprehensive picture for longitudinal studies. 300 trees were used, with a feature fraction of 0.3, max depth of 7, with 10 iterations of mixed-effects Random Forests (MERFs), and 100 BorutaSHAP trials (100% importance threshold; p = 0.05).

Table 1. Performance Results from Classification Accuracy of Engineered Predictive and Not Predictive Features in Five Variations of Simulated Data



					BorutaSHAP results			Confusion Matrix				Performance results			
Dataset	Software tool	Model	% Variance Explained	Input Features	Accepted Features	Tentative Features	Rejected Features	True Positives	True Negatives	False Positives	False Negatives	Recall	Precision	F1-score	AUC
SimFeatures (other input variables; no microbiome)	EXPLANA	Original	98.0% (98.1%)	15	13	1	1	13	2	0	0	1	1	1	1
		First	98.9% (99.0%)	20	11	0	9	10	8	1	1	0.91	0.91	0.91	0.9
		Previous	89.6% (89.4%)	22	13	1	8	10	8	3	1	0.91	0.77	0.83	0.82
		Pairwise	96.8% (96.8%)	28	19	0	9	16	7	3	2	0.89	0.84	0.86	0.79
SimMicrobiome (microbiome; no other input variables)	EXPLANA	Original	96.9% (97.0%)	200	28	3	169	23	170	5	2	0.92	0.82	0.87	0.95
		First	96.2% (96.4%)	200	30	3	167	24	169	6	1	0.96	0.8	0.87	0.96
		Previous	65.4% (66.6%)	200	16	1	183	15	174	1	10	0.6	0.94	0.73	0.8
		Pairwise	92.9% (93.1%)	200	40	5	155	25	160	15	0	1	0.62	0.77	0.96
	feature-volatility	Original	76.9% (R-squared)	200	180	0	20	25	20	155	0	1	0.14	0.24	0.56
SimFeaturesMicrobiome0 (microbiome; other input variables; 0 random variables)	EXPLANA	Original	97.4% (97.2%)	215	37	3	175	34	174	3	4	0.89	0.92	0.91	0.94
		First	97.6% (97.7%)	220	32	4	184	31	183	1	5	0.86	0.97	0.91	0.93
		Previous	86.4% (87.9%)	222	32	4	186	28	182	4	8	0.78	0.88	0.82	0.88
		Pairwise	95.5% (95.8%)	228	49	4	175	36	172	13	7	0.84	0.73	0.78	0.88
SimFeaturesMicrobiome500 (microbiome; other input variables; 500 random variables)	EXPLANA	Original	97.3% (97.3%)	715	37	2	676	34	674	3	4	0.89	0.92	0.91	0.95
		First	97.4% (97.7%)	720	32	2	686	30	682	2	6	0.83	0.94	0.88	0.92
		Previous	84.6% (88.3%)	722	37	6	679	27	676	10	9	0.75	0.73	0.74	0.87
		Pairwise	95.1% (95.7%)	728	62	3	663	36	659	26	7	0.84	0.58	0.69	0.9
SimFeaturesMicrobiome1000 (microbiome; other input variables; 1000 random variables)	EXPLANA	Original	97.5% (97.2%)	1215	36	1	1178	34	1175	2	4	0.89	0.94	0.92	0.95
		First	97.4% (97.8%)	1220	38	0	1182	32	1178	6	4	0.89	0.84	0.86	0.94
		Previous	83.9% (87.6%)	1222	45	7	1170	29	1170	16	7	0.81	0.64	0.72	0.9
		Pairwise	95.1% (95.9%)	1228	66	4	1158	36	1155	30	7	0.84	0.55	0.66	0.91

EXPLANA Hyperparameter arguments: 300 trees, 0.3 feature fraction for decision tree splits with a max depth of 7, 10 MERF iterations, and 100 BorutaSHAP trials at 100% importance score threshold (p = 0.05)

feature-volatility hyperparameter arguments: 300 trees, 5 cross-fold validation

True positives are selected predictive features, true negatives are rejected not-predictive features, false positives are selected not-predictive features and false negatives are rejected predictive features (Fig S1)

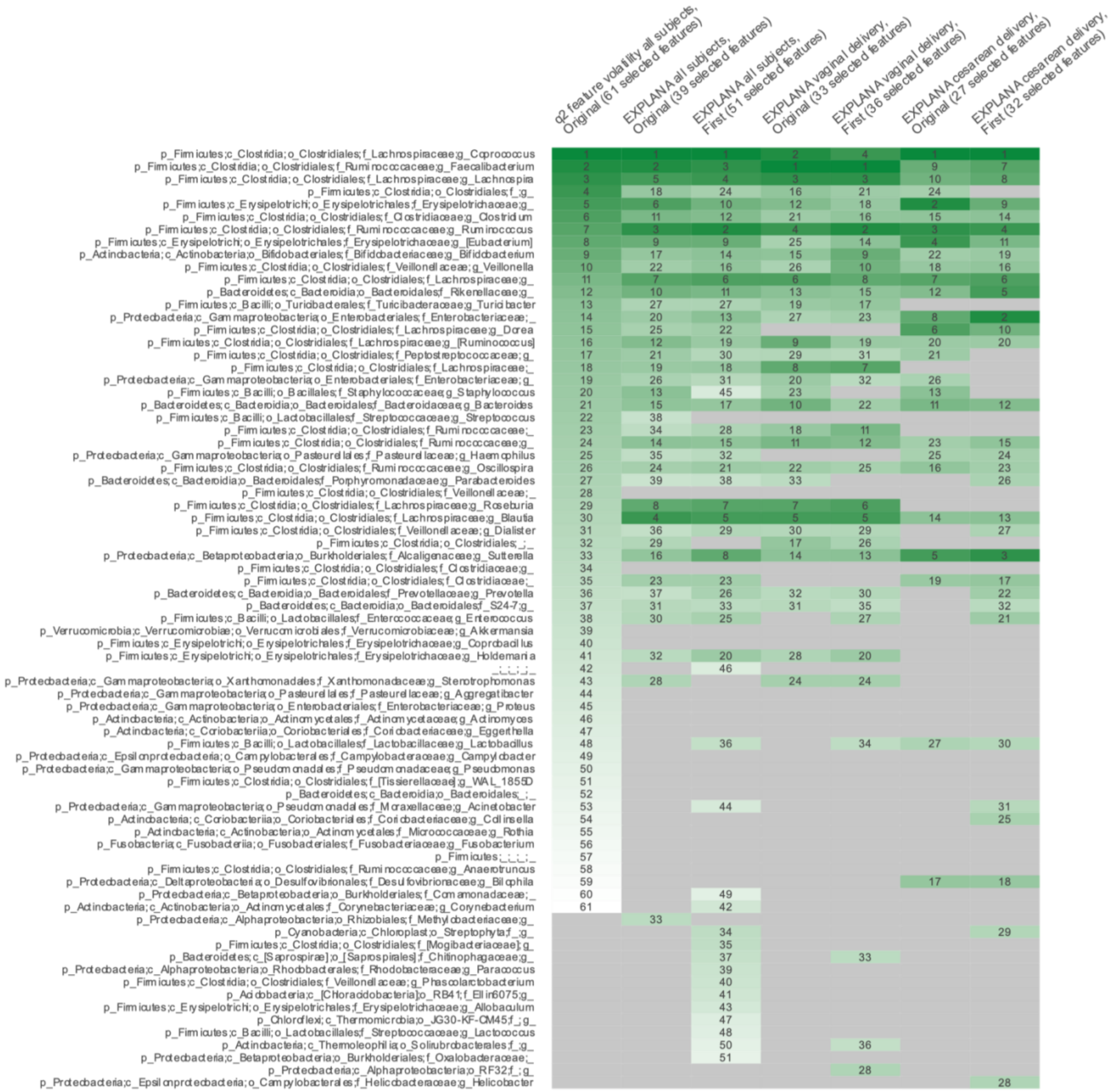
The features selected from the smallest dataset containing a variety of input feature types, *SimFeatures*, are shown in Fig 4a. Details about their motivation for workflow demonstration are explained in Table S1. One feature that emphasizes the advantage of calculating ∆s using the previous values or pairwise values rather than only first/baseline values is pill color, a categorical feature engineered to have a negative impact on happiness when blue pills were taken *after* green ("green_blue"), and not conversely. The feature green_blue had relatively high ranks and a negative impact on the response using *Pairwise* and *Previous* (respective rank:impact: *Pairwise* = 6/19:-5.3, and *Previous* 4/13: -10.7; Fig 4a and Fig 4c). Green was not consumed by individuals at baseline/T1, therefore green_blue could not be identified with *First*. Green alone was selected with *Original* at a lower importance rank and a small positive impact (11/13: +1.6) and blue was selected with a small negative impact (10/13: -1.5). Green consumption occurred more at later timepoints, while happiness was also increasing, so selection of green with a small positive impact on happiness was appropriate in the *Original* despite being designed without independent effects. However, without the additional information provided by ∆s, an assumption about positive impact on happiness when consuming green pills could have been made.

A numerical example that emphasizes the benefit of using different methods for calculating ∆s in longitudinal analysis is "relaxation," which was selected in *Original*, *Previous* and *Pairwise* models but not *First*. Relaxation had one value for baseline/T1 and a different equivalent value at all later timepoints, resulting in T1 comparisons to T2, T3, T4 and T5 having identical values (e.g., T1=1, T2-T5 = 5 differences compared to baseline would be 4 for comparisons in the *First* dataset). The lack of change in *First* makes it ineffective for discrimination and pattern recognition for "relaxation" despite its relationship to happiness. Another numerical feature only important in some contexts of change is "sunshine," which was selected using *Original, First,* and *Pairwise*, but not *Previous*. Sunshine has a linear relationship to happiness over time which can sometimes be less impactful upon calculating ∆s using differences between adjacent timepoints only.

Although all analyses shown thus far have demonstrated the application of EXPLANA to longitudinal study designs, we note that EXPLANA can also be used to analyze cross-sectional data, which excludes ∆s. To illustrate this, we applied EXPLANA to *SimFeatures* using only timepoint 1 data. Perfect precision and recall were obtained, with six features selected: high and low salary, fish and dog as pets, healthy and unhealthy individuals (Fig S4).

EXPLANA performance was then compared to an existing feature selection tool for longitudinal microbiome studies, specifically the QIIME 2[38] longitudinal[16] feature-volatility tool. Because the feature-volatility tool itself does not create ∆ datasets (other tools in q2-longitudinal can create ∆ datasets) and does not include categorical variables like EXPLANA, *SimMicrobiome* (the simulated microbiome, with no other study input variables) was used with *Original* only (Fig S3). All performance measures were substantially better for EXPLANA compared to feature-volatility, except for recall, which was 0.92 and 1.00, respectively (Table 1). Of the 25 predictive microbes, EXPLANA selected 23 feature-volatility identified all 25. Of the 175 not-predictive microbes, EXPLANA selected five FPs and feature-volatility selected 155 FPs.

EXPLANA was next applied to identify bacteria related to month-of-life in babies from the early childhood and microbiome (ECAM) study[39] which was also used to compare results with the QIIME2[38] longitudinal feature-volatility feature selection[16]. The ECAM study produced 16S rRNA targeted sequencing data from fecal samples collected monthly from 43 babies, over the first 2 years of life, and was used with EXPLANA and feature-volatility to identify bacterial genera related to month-of-life among 455 input genera (Fig 5). Of 455 input genera, the selected genera for all individuals in the *Original* dataset was 61 (13% of the input features) for QIIME 2 longitudinal feature-volatility and 39 (8.6% of the input features) for EXPLANA. *First* ∆s using EXPLANA selected 51 features and uncovered 11 unique features related to month-of-life when making comparisons to values at baseline/birth.



	q2 feature volatility all subjects, Original (61 selected features)	EXPLANA all subjects, Original (39 selected features)	EXPLANA all subjects, First (51 selected features)	EXPLANA vaginal delivery, Original (33 selected features)	EXPLANA vaginal delivery, First (36 selected features)	EXPLANA cesarean delivery, Original (27 selected features)	EXPLANA cesarean delivery, First (32 selected features)
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus	1	1	1	2	4	1	1
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium	2	2	3	1	1	9	7
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Lachnospira	3	5	4	3	3	10	8
p_Firmicutes;c_Clostridia;o_Clostridiales;f_;g_	4	18	24	16	21	24	
p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_	5	6	10	12	18	2	9
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium	6	11	12	21	16	15	14
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus	7	3	2	4	2	3	4
p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_[Eubacterium]	8	9	9	25	14	4	11
p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium	9	17	14	15	9	22	19
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Veillonella	10	22	16	26	10	18	16
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_	11	7	6	6	8	7	6
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;g_	12	10	11	13	15	12	5
p_Firmicutes;c_Bacilli;o_Turicibacterales;f_Turicibacteraceae;g_Turicibacter	13	27	27	19	17		
p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;_	14	20	13	27	23	8	2
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea	15	25	22			6	10
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus]	16	12	19	9	19	20	20
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Peptostreptococcaceae;g_	17	21	30	29	31	21	
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;_	18	19	18	8	7		
p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_	19	26	31	20	32	26	
p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Staphylococcus	20	13	45	23		13	
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides	21	15	17	10	22	11	12
p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus	22	38					
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;_	23	34	28	18	11		
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_	24	14	15	11	12	23	15
p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus	25	35	32			25	24
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira	26	24	21	22	25	16	23
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides	27	39	38	33			26
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;_	28						
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Roseburia	29	8	7	7	6		
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia	30	4	5	5	5	14	13
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister	31	36	29	30	29		27
p_Firmicutes;c_Clostridia;o_Clostridiales;_;_	32	29		17	26		
p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella	33	16	8	14	13	5	3
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_	34						
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;_	35	23	23			19	17
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella	36	37	26	32	30		22
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_S24-7;g_	37	31	33	31	35		32
p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus	38	30	25		27		21
p_Verrucomicrobia;c_Verrucomicrobiae;o_Verrucomicrobiales;f_Verrucomicrobiaceae;g_Akkermansia	39						
p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Coprobacillus	40						
p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Holdemania	41	32	20	28	20		
_;_;_;_;_	42		46				
p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;g_Stenotrophomonas	43	28		24	24		
p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Aggregatibacter	44						
p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Proteus	45						
p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces	46						
p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Eggerthella	47						
p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactobacillus	48		36		34	27	30
p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Campylobacteraceae;g_Campylobacter	49						
p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Pseudomonadaceae;g_Pseudomonas	50						
p_Firmicutes;c_Clostridia;o_Clostridiales;f_[Tissierellaceae];g_WAL_1855D	51						
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;_;_	52						
p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Moraxellaceae;g_Acinetobacter	53		44				31
p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella	54						25
p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia	55						
p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;g_Fusobacterium	56						
p_Firmicutes;_;_;_;_	57						
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Anaerotruncus	58						
p_Proteobacteria;c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;g_Bilophila	59					17	18
p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Comamonadaceae;_	60		49				
p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Corynebacteriaceae;g_Corynebacterium	61		42				
p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;f_Methylobacteriaceae;g_		33					
p_Cyanobacteria;c_Chloroplast;o_Streptophyta;f_;g_			34				29
p_Firmicutes;c_Clostridia;o_Clostridiales;f_[Mogibacteriaceae];g_			35				
p_Bacteroidetes;c_[Saprospirae];o_[Saprospirales];f_Chitinophagaceae;g_			37		33		
p_Proteobacteria;c_Alphaproteobacteria;o_Rhodobacterales;f_Rhodobacteraceae;g_Paracoccus			39				
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Phascolarctobacterium			40				
p_Acidobacteria;c_[Chloracidobacteria];o_RB41;f_Ellin6075;g_			41				
p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Allobaculum			43				
p_Chloroflexi;c_Thermomicrobia;o_JG30-KF-CM45;f_;g_			47				
p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Lactococcus			48				
p_Actinobacteria;c_Thermoleophilia;o_Solirubrobacterales;f_;g_			50		36		
p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Oxalobacteraceae;_			51				
p_Proteobacteria;c_Alphaproteobacteria;o_RF32;f_;g_					28		
p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Helicobacteraceae;g_Helicobacter							28

Figure 5. Heatmap of bacterial genera predictive of month-of-life in newborns selected by EXPLANA and QIIME 2 longitudinal feature-volatility using the early childhood and microbiome (ECAM) dataset[39]. Results from EXPLANA were compared to QIIME2 longitudinal feature-volatility results. Using both tools 500 trees were used. For EXPLANA, arguments included a feature fraction of 0.2, max depth of 7, with 10 iterations of mixed-effects Random Forests (MERFs), and 100 BorutaSHAP trials (100% threshold; p = 0.05). For feature-volatility, a feature fraction of 1.0 was used with Random Forest (RF). Feature-volatility results are shown in the first column (*Original*; all individuals), followed by EXPLANA analysis by stratifying newborns by vaginal or cesarean delivery mode and using *Original* and *First* models. Figure is sorted by QIIME2 longitudinal feature-volatility ranks, where a darker green and lower number indicates a more important rank.

The ease of running different datasets through the EXPLANA workflow enables easier exploratory analyses of datasets, such as conducting stratified analyses. For instance, to do stratified analyses in EXPLANA the user simply needs to use small R scripts inside the configuration file which will be documented in the report. EXPLANA was additionally used to select genera after stratifying by vaginal or cesarean delivery mode, due to their impact on the gut microbiome[40], and using *Original* and *First* ∆s. Some genera were only selected as important in vaginal delivery (*Roseburia*, *Stenotrophomonas, Turicibacter* and *Holdemania*), while others only in cesarean delivery (*Dorea*, *Bilophila*, and *Haemophilus)*. There were 14 genera unique to EXPLANA overall compared to QIIME 2 feaure-volatility,13 of which came from using the *First* model, including *Paracoccus, Allobaculum, Helicobacter* and *Lactococcus*.

To demonstrate the versatility for EXPLANA to work with categorical variables using the ECAM dataset we identified categorical features that were related to month-of-life while excluding the microbiome data. Variables included were delivery type (cesarean/vaginal), predominant diet during first three months (breast/formula milk), sex (male/female), and antibiotic exposure (yes/no). Change in antibiotic use from no to yes (n_y) at later timepoints was positively related to month-of-life, indicating that babies are more likely to have an antibiotic treatment event as they age. Similarly, antibiotic use “no to no” showing a negative impact on month-of-life (Fig S5).

Discussion

To address challenges with longitudinal microbiome analytics, we developed a feature selection tool for longitudinal data to expedite discovery. We implemented supervised ML methods to identify features that relate to response values. For meaningful results, it was essential to implement methods that provide rationale and explanations about why features were selected and how their occurrence impacts values of dependent variables.

When using simulated longitudinal datasets, EXPLANA had good performance for most analyses, as determined by selection or rejection of engineered features. The ∆ datasets highlighted that performing change analyses can produce unique insights compared to using *Original* longitudinal data without ∆ calculations. Several studies have applied MERFs for feature selection[41,42], however, they did not use ∆s, which could lead to a possible loss of valuable insights.

The four models can have strengths and limitations with feature selection capabilities for a variety of reasons or in different situations. For example, the *Previous* model’s failure to select sunshine with *SimFeatures*, which had an engineered linear relationship to happiness over time, is conceptually similar to why *Previous* had a low percent variation explained and recall for *SimMicrobiome.* This is because the simulated predictive microbes and sunshine had a similar linear trend over time. Other temporal trends can include quadratic, hockey stick, etc[37]. *Previous* can miss predictive features in cases where changes are minimized such as when the reference time is closer and predictor variables linearly relate to the response. This contrasts with *First*, where changes would be emphasized compared to baseline for this study. A limitation of *Pairwise*, is that more comparisons are made, which increases the compute time, and the number of comparisons from overlapping time spans, so it is important to consider a higher chance of FPs due to more feature comparisons, leading to more values. A more stringent p-value cutoff should be considered for *Pairwise*. As expected, *Pairwise* had the most FPs for our simulations. For all simulations, *First* and *Original* had the best overall percent variation explained as expected for this study design consisting of engineered features with known changes from a baseline value. Despite model limitations in particular instances, different engineered features emphasized the importance in building models using *Original* and Δ datasets. Pill color "green_blue" was only able to be found in *Previous* and *Pairwise*, and demonstrated the ability for EXPLANA to find an order-dependent categorical variable that impacted a response. "Relaxation" is impossible to select in *First,* and was not selected, because it lacked variation compared to baseline values, and it was selected by all other models. "Sunshine" which is positively linearly related to the response, was not selected with *Previous*, due to smaller differences between closer points in time, but was selected with a high importance rank with the three other models. These features provide examples that demonstrate how different models, for different contexts of change, are needed for a more comprehensive exploratory analysis in longitudinal datasets. The Δs allow for identification of order-dependent categorical features (also seen with "no_yes" for antibiotics with the ECAM data), which are impossible to detect using *Original*.

Dissimilarities between data included in each of the four models can create complications with interpreting feature selection results, such as dropping samples from Δ datasets due to missing timepoints. Another challenge with interpreting results from multiple models arises from including distance matrices, which can only be included in Δ datasets because they represent changes between samples. Thus, care should be taken with interpreting results obtained by different models within one report.

When using the ECAM dataset, the higher percentage of important genera identified using feature-volatility is likely attributed to false positives due to a lack of automated statistical testing as performed with BorutaSHAP in EXPLANA. Many false positives were indeed identified using feature-volatility with a simulated microbiome dataset (*SimMicrobiome*) which had a known amount of predictive and not predictive microbes. Importance ranks differ for many genera between EXPLANA and feature-volatility leading to different conclusions about the degree of importance regarding developmental microbiome changes. When genera were selected using EXPLANA by delivery mode, which influences the gut microbiome[39,40], four were important only in vaginal delivery while three were important only in cesarean delivery (Fig 5). Of these seven genera, a review on gut microbiota by birth type[43], noted two studies corroborating *Haemophilus* as important in cesarean delivery mode despite a small sample size of 19 babies.[44,45] The small number of individuals per delivery mode (24 for vaginal delivery and 19 for cesarean) might limit the power needed to detect effects. However, other studies have demonstrated RFs to be useful with small sample sizes, including 26[46], 30[47] and 35[48] samples. Because RFs can perform internal validation using subsets of data and out-of-bag (OOB) scores, there is less concern associated with small datasets.

The tools first-differences and first-distances (for distance matrices) from q2-longitudinal[1] can create Δs from continuous data, which can be used with feature-volatility. However, creation of Δs is not part of the feature-volatility feature-selection process, and comparing results from different models is cumbersome without comparative figures, such as the feature occurrences figure provided with EXPLANA (e.g., Fig 4a). We have also not observed the incorporation of categorical Δs in any feature selection tool. The ML model used in feature-volatility is RF, which is not designed for repeated measures analyses like MERF, which is used in EXPLANA. The importance score used in feature-volatility is Gini, which is biased when categorical and numerical variables are combined[49], while EXPLANA uses SHAP, which works well for this combination of feature types. Additionally, SHAP provides feature impact on response, in addition to rank, as well as statistical testing from BorutaSHAP.

There is no one-size-fits all model and it is not possible to know which parameter adjustments will lead to optimal results. However, tuning of the algorithm can address some issues, especially the number of features available per decision tree split, which is affected by the proportion of meaningful input variables. This workflow provides a means for finding a better model because it provides a relatively simple approach for testing. Interpretation of exploratory analysis should be done with care, and *post-hoc* testing and further investigation should be considered.

The barrier to performing data-driven feature selection for cross-sectional and longitudinal microbiome studies has been lessened by EXPLANA. Different applications are more attainable including analyzing less timepoints or segments of time, such as during plateau or active time periods. Additionally, stratifying by factors such as sex, geography, disease symptom, or a combination of factors could be worthwhile. EXPLANA also provides the opportunity to investigate different variables (including responses) from prior hypotheses or from results of another exploratory analysis.

Overall, we addressed many challenges and removed various barriers to performing feature selection by combining existing tools and novel methods to generate research-motivating results.

Methods

Workflow Overview

EXPLANA was developed using Snakemake[50] to facilitate piping inputs and outputs from different scripts written in different software languages, primarily R and Python (Fig 1). The workflow is executed from user-input arguments from a configuration file which pipes files to different scripts concluding with an .html report. The configuration file includes a list of datasets (microbiome, surveys, etc.) in long format (rows are samples; columns are features). First, individual datasets can be preprocessed through filters, dimensionality reduction, or transformation. If multiple files exist, they are merged to create the *Original* dataset. For longitudinal data, Δ datasets are computed (Fig 2). Finally, a feature selection algorithm is implemented by building a model from each of the four datasets (*Original, First Previous* and *Pairwise*): First, RFs[13] or MERFs[15] (if multiple samples exist per subject), are trained; Next, BorutaSHAP[36] is used to rank features by importance if they perform better than expected by random chance, and determine feature impact on response. The final report includes figures, tables, and a written analytic summary.

Analyses were completed locally to ensure reasonable compute time for typical academic microbiome studies or those without server access. For 1000 features, 5 timepoints and 100 individuals, run time is less than 30 minutes using a MacBook Pro (Memory: 32 GB 2400 MHz DDR4; Processor: 2.9 GHz; 6-Core Intel Core i9).

Software and Data Availability

Description of workflow implementation and user documentation can be found at www.explana.io and software, dataset and licensing at https://github.com/JTFouquier/explana.

Configuration file

Each configuration file is associated with one analysis. Users must modify a configuration file that specifies dataset files, a response/outcome variable, sample identifier column, timepoint column, distance matrices (if applicable), optional dimensionality reduction steps prior to feature selection, and ML method decisions/arguments. Feature values as well as feature columns can be kept or dropped for individual datasets or for the merged *Original* dataset using small scripts within the configuration file.

Preprocessing datasets

For each feature selection analysis, one or more dataset files can be used depending on project and data organization. Each dataset can be preprocessed, as needed, which may include dropping entire features or specific feature values, or other filters on a per dataset basis. Dimensionality reduction can be performed prior to feature selection using principal components analysis (PCA), transformation or filters. PCA is used on a set of related variables to capture the maximum variance using fewer variables. Short R scripts can be added to the configuration file to modify each dataset or the complete dataset after merging individual datasets.

Dataset Integration

After preprocessing, individual datasets are merged using the sample identifier column to create the combined "*Original”* dataset. The *Original* dataset is named accordingly because it contains original values of features that may have been sampled over time. The *Original* dataset does not include any intra-individual changes/differences between timepoints like the Δ datasets. Data integration is performed through a left merge, where samples in the top/first dataset listed are prioritized. This means that additional samples in other datasets will not be included. For some analyses, merging data prior to implementation may be simpler.

Delta (Δ) dataset creation

For longitudinal analyses, the *Original* dataset is used to compute three Δ datasets, *First*, *Previous*, and *Pairwise* by calculating feature changes over time, per subject, using different reference points (Fig 2). For *First*, differences are calculated compared to baseline/first measures. For *Previous*, differences are calculated compared to the previous timepoint. For *Pairwise*, all pairwise comparisons between timepoints are computed. For longitudinal studies with only two timepoints, only *Original* and *First* are built.

For categorical variables, the order of categorical values for each subject at both timepoints per comparison is tracked (e.g., for the pill color feature, if T2 is green and T3 is blue, the T2_T3 Δ is green_blue). For numerical variables, reference values are subtracted from the later timepoint (e.g., for subject 1, if T2 has 3 microbes and T3 has 7 microbes, the Δ is 4).

Timepoint is numerical for *Original* to provide information about order of events, and as a categorical feature for Δ datasets due to overlap in timepoints (i.e., T1_2 and T1_3 overlap each other at T1_2). This overlap can be thought of as though time were a generic categorical feature rather than an abstract concept. In other words, if T1, T2 and T3 were recoded as A, B, and C, respectively, the comparisons A_B, A_C and B_C are potentially interesting.

Feature Selection Algorithm

Feature selection is performed using all four models, as needed. For categorical features, all unique values/classes per feature are converted to new binary features, where feature presence in a sample is 1 and absence is 0. This enables selection of uncommon feature values that influence a response.

Next, regression is performed to select features related to the response. When more than one measurement per subject exists, MERF is used instead of RFs. Both use Scikit-Learn[51] RandomForestRegressor as the fixed effects forest. Boruta[34] is a method that uses shuffled versions of input features to assess importance score comparison to shuffled versions of the same features. This process is repeated to identify features that more important than expected by random chance. Features are categorized as accepted, tentative or rejected. BorutaSHAP[36] is implemented because it works with the unique properties of SHAP (SHapley Additive exPlanations)[35], which provides feature ranks and explains feature impact on the response.

Rejected features are dropped, RF or MERF are rerun, and visualizations are generated without irrelevant features that might hide true signal from important features. Percent variation explained, using OOB scores, is provided using the complete-feature and reduced-feature forest.

Report Details

The result of each analysis is an interactive .html report that includes figures, tables, links to directories for data exploration, links to PubMed for researching interesting findings, and written descriptions of processes leading to the selected feature set (Fig 3). An analytic methods section is dynamically created based on user inputs, other variables, and defaults that can be directly included in a manuscript. A feature occurrence figure summarizes selected feature ranks by all four models, followed by model-specific details. Impact on response is provided for categorical features because positive instances per sample are clear, while numerical feature relationships are more complex (i.e., a hockey curve pattern). Figures include SHAP summary and dependence plots with detailed feature impact on the response.

Simulated dataset design

A simulated happiness dataset was designed to facilitate testing performance using variables that are predictive and not predictive. A summary of the motivation for including variables, and their effects on the response is detailed in Table 1.

All individuals started with the same happiness score and effects from all features were used to update each subject's happiness score at each timepoint. Predictive features had effect values stored in a different column labeled with the suffix "_effect". For example, the "salary" column contained values "high" or "low" and had a corresponding "salary_effect" column with numerical values reflecting positive or negative effects on happiness, corresponding to high salary and low salary, respectively. All effect column values were added to the original happiness scores and columns labeled "effect" were removed prior to feature selection. This way, the engineered effects were contained in the happiness value, and predictive features can be identified if they corresponded to the effect.

We used R package faux[52] to add subject random effects, five timepoints, and a control and test group. The test group was simulated to linearly increase with a positive slope of 30 (correlation coefficient of 0.7 and standard deviation of 5) to simulate a treatment effect that improved happiness.

For longitudinal microbiome simulations containing differentially abundant and not differentially abundant microbes we used microbiomeDASim[37]. A first-order autoregressive correlation structure was used that linearly increased with a slope of 30 to correlate with happiness (Correlation coefficient = 0.7; standard deviation = 5).

Eight data distributions were used with random, not predictive variables. Normal, Bernoulli, Binomial, Poisson, Exponential, Gamma, Weibull[53] and Dirichlet. The number of random variables is indicated in the dataset name. Accordingly, for *SimFeaturesMicrobiome0*, *SimFeaturesMicrobiome500* and *SimFeaturesMicrobiome1000*, the number of random variables is 0, 500, and 1000, respectively and all include *SimFeatures* and *SimMicrobiome*.

When using EXPLANA arguments were set based on recommendations of the underlying tools or from previous studies on hyperparameter tuning[54,55], which included for MERF, 300 trees, 0.3 feature fraction for decision tree splits with a max depth of 7 and 10 MERF iterations and 100 BorutaSHAP trials were run (100% importance threshold; p = 0.05).

Performance Testing

Performance was assessed from simulated microbiome datasets using F1-scores and AUC metrics. Recall, also called the true positive rate, measures the proportion of the predictive features correctly selected Recall = TP/(TP+FN). Precision describes the proportion of all positive predictions that are predictive features (Precision = TP/(TP+FP) (Fig S1). An F1-score is calculated using precision and recall 2*(Precision * Recall) / (Precision + Recall), as well as the AUC is the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate.

Early childhood and microbiome dataset analysis

To explore feature selection results using EXPLANA with previously published data, the ECAM dataset was used because it is also used to demonstrate feature selection with QIIME2 longitudinal feature-volatility (feature-volatility). Metadata was filtered to remove duplicate months to facilitate Δ calculations performed by EXPLANA. 500 trees were used for both tools. For EXPLANA, arguments included a feature fraction of 0.2, max depth of 7, with 10 iterations of MERF and 100 BorutaSHAP trials (100% threshold; p = 0.05). Defaults for Q2 feature-volatility included 1.0 for feature fraction and 5 k-fold cross-validations.

Acknowledgements:

We would like to acknowledge the following individuals from The University of Colorado, Anschutz Medical Campus: Jennifer's dissertation committee, which consists of Laura Saba, Michael Strong and Katerina Kechris, and co-authors Catherine Lozupone and Maggie Stanislawski; Nichole Nusbacher (Researcher and Lozupone Lab Manager) and Emily Mastej (Doctoral Candidate, Computational Bioscience Graduate Program); and students who provided feedback on software testing and installation which include Brook Santangelo, John Sterrett, Angela Sofia Burkhart Colorado, and Erik Serrano. We would also like to thank Scott Kelley (Professor and Program Director of Bioinformatics and Medical Informatics at San Diego State University), Sergio Serrano-Villar and Alba Talavera-Rodríguez for feedback and suggestions.

Funding:

Jennifer Fouquier was funded by T15 LM009451(NIH U.S. National Library of Medicine, Colorado Biomedical Informatics Training Program) and R01 DK108366 (NIH Diet/gut Microbiome Interaction Influence Inflammatory Disease in HIV Patients) and is currently funded by U01 AI150589 (NIH Dietary and synbiotic strategies to protect against *Clostridioides difficile* infection). Maggie Stanislawski is funded by K01 HL157658.

Conflict of interest statement:

Jennifer Fouquier would receive compensation from future licensing agreements and analytic services performed by Jennifer Fouquier, LLC. Catherine Lozupone would receive compensation from future licensing agreements. Other authors declare no conflict of interest.

1. Santiago-Rodriguez, T. M. & Hollister, E. B. Multi ‘omic data integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology* **45**, 151456 (2021).
2. Ursell, L. K., Metcalf, J. L., Parfrey, L. W. & Knight, R. Defining the human microbiome. *Nutrition Reviews* **70**, S38–S44 (2012).
3. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *WIREs RNA* **8**, e1364 (2017).
4. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Molecular Cell* **58**, 699–706 (2015).
5. Maruvada, P., Leone, V., Kaplan, L. M. & Chang, E. B. The Human Microbiome and Obesity: Moving beyond Associations. *Cell Host & Microbe* **22**, 589–599 (2017).
6. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature microbiology* **4**, 623–632 (2019).
7. Krajmalnik-Brown, R., Lozupone, C., Kang, D.-W. & Adams, J. B. Gut bacteria in children with autism spectrum disorders: challenges and promise of studying how a complex community influences a complex disease. *Microbial Ecology in Health and Disease* **26**, 26914 (2015).
8. Rebersek, M. Gut microbiome and its role in colorectal cancer. *BMC cancer* **21**, 1–13 (2021).
9. Zhuang, H. *et al.* Dysbiosis of the gut microbiome in lung cancer. *Frontiers in Cellular and Infection Microbiology* **9**, 112 (2019).
10. Williams, B., Landay, A. & Presti, R. M. Microbiome alterations in HIV infection a review. *Cellular microbiology* **18**, 645–651 (2016).
11. Witkowski, M., Weeks, T. L. & Hazen, S. L. Gut microbiota and cardiovascular disease. *Circulation research* **127**, 553–570 (2020).
12. Linear Mixed-Effects Models: Basic Concepts and Examples. in *Mixed-Effects Models in S and S-PLUS* (eds. Pinheiro, J. C. & Bates, D. M.) 3–56 (Springer, New York, NY, 2000). doi:10.1007/0-387-22747-4_1.
13. Breiman, L. Random Forests -- Random Features.
14. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).
15. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* **84**, 1313–1328 (2014).
16. Bokulich, N. A. *et al.* q2-longitudinal: Longitudinal and Paired-Sample Analyses of Microbiome Data. *mSystems* **3**, 10.1128/msystems.00219-18 (2018).
17. Frey, D. L. *et al.* Changes in Microbiome Dominance Are Associated With Declining Lung Function and Fluctuating Inflammation in People With Cystic Fibrosis. *Front. Microbiol.* **13**, (2022).
18. Ferrocino, I. *et al.* Changes in the gut microbiota composition during pregnancy in patients with gestational diabetes mellitus (GDM). *Sci Rep* **8**, 12216 (2018).
19. Meslier, V. *et al.* Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. *Gut* **69**, 1258–1268 (2020).

20. Rodenes-Gavidia, A. *et al.* An insight into the functional alterations in the gut microbiome of healthy adults in response to a multi-strain probiotic intake: a single arm open label trial. *Front Cell Infect Microbiol* **13**, 1240267 (2023).
21. Zhang, L., Luo, H. & Kang, G. Longitudinal study of physical activity with various methods in maintenance hemodialysis patients. *Hemodialysis International* **25**, 249–256 (2021).
22. Fouquier, J. *et al.* The Gut Microbiome in Autism: Study-Site Effects and Longitudinal Analysis of Behavior Change. *mSystems* **6**, e00848-20 (2021).
23. Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. (Cambridge University Press, Cambridge, 2013). doi:10.1017/CBO9781139342834.
24. Tartini, R., Steinbrunn, W., Kappenberger, L. & Meyer, U. A. Dangerous interaction between amiodarone and quinidine. *The Lancet* **319**, 1327–1329 (1982).
25. Vassallo, P. & Trohman, R. G. Prescribing Amiodarone: An Evidence-Based Review of Clinical Indications. *JAMA* **298**, 1312–1322 (2007).
26. Mullen, S. A. *et al.* Precision therapy for epilepsy due to KCNT1 mutations: A randomized trial of oral quinidine. *Neurology* **90**, e67–e72 (2018).
27. Queen, O. & Emrich, S. J. LASSO-based feature selection for improved microbial and microbiome classification. in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2301–2308 (2021). doi:10.1109/BIBM52615.2021.9669485.
28. Wang, C., Hu, J., Blaser, M. J. & Li, H. Microbial trend analysis for common dynamic trend, group comparison, and classification in longitudinal microbiome study. *BMC Genomics* **22**, 667 (2021).
29. Calle, M. L., Pujolassos, M. & Susin, A. coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics* **24**, 82 (2023).
30. Lee, K. H., Coull, B. A., Moscicki, A.-B., Paster, B. J. & Starr, J. R. Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. *Biostatistics* **21**, 499–517 (2018).
31. Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A. & Droit, A. timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics* **38**, 577–579 (2022).
32. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8**, (2017).
33. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**, 169–172 (2011).
34. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **36**, 1–13 (2010).
35. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
36. Keany, E. BorutaSHAP. (2021).
37. Williams, J., Bravo, H. C., Tom, J. & Paulson, J. N. microbiomeDASim: Simulating longitudinal differential abundance for microbiome data. *F1000Res* **8**, 1769 (2020).
38. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852–857 (2019).
39. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* **8**, 343ra82-343ra82 (2016).

40. Dogra, S. *et al.* Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *mBio* **6**, e02419-14 (2015).
41. Zeamer, A. L. *et al.* Association between microbiome and the development of adverse posttraumatic neuropsychiatric sequelae after traumatic stress exposure. *Transl Psychiatry* **13**, 1–14 (2023).
42. Haran, J. P. *et al.* The high prevalence of Clostridioides difficile among nursing home elders associates with a dysbiotic microbiome. *Gut Microbes* **13**, 1897209 (2021).
43. Coelho, G. D. P. *et al.* Acquisition of microbiota according to the type of birth: an integrative review. *Rev Lat Am Enfermagem* **29**, e3446.
44. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
45. Hesla, H. M. *et al.* Impact of lifestyle on the gut microbiota of healthy infants and their mothers—the ALADDIN birth cohort. *FEMS Microbiol Ecol* **90**, 791–801 (2014).
46. Ward, D., Miller, R. & Nikolaev, A. Evaluating three stuttering assessments through network analysis, random forests and cluster analysis. *Journal of Fluency Disorders* **67**, 105823 (2021).
47. Luan, J., Zhang, C., Xu, B., Xue, Y. & Ren, Y. The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research* **227**, 105534 (2020).
48. Hassan, S. S., Farhan, M., Mangayil, R., Huttunen, H. & Aho, T. Bioprocess data mining using regularized regression and random forests. *BMC Syst Biol* **7**, S5 (2013).
49. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).
50. Snakemake—a scalable bioinformatics workflow engine | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/28/19/2520/290322.
51. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
52. DeBruine, L. faux: Simulation for Factorial Designs. Zenodo https://doi.org/10.5281/ZENODO.2669586 (2023).
53. K, J. 7 Statistical Distributions that every Data Scientist should know— with intuitive explanations. *Medium* https://towardsdatascience.com/7-statistical-distributions-that-every-data-scientist-should-know-with-intuitive-explanations-bf967db81f0b (2020).
54. Weerts, H. J. P., Mueller, A. C. & Vanschoren, J. Importance of Tuning Hyperparameters of Machine Learning Algorithms. Preprint at http://arxiv.org/abs/2007.07588 (2020).
55. Probst, P., Wright, M. N. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* **9**, e1301 (2019).