Machine learning prediction of *C. difficile* colonization based on microbiota composition on day of challenge

Performance was measured by the area under the receiver-operator characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC).

- Performance on cross-validation folds of training data:
  - Mean AUROC 0.97 (s.d. 0.011).
- Performance on held-out test data:
  - Mean AUROC 0.95 (s.d. 0.05)
  - Mean AUPRC 0.84 (s.d. 0.065). Baseline AUPRC: 0.34

Feature importance: Top 20 OTUs

OTU	mean_decrease	stdev	% models	rank
Enterobacteriaceae (OTU 7)	-0.0078371	0.0121111	58.0	1
Erysipelotrichaceae (OTU 222)	-0.0022672	0.0047096	86.4	2
Porphyromonadaceae (OTU 14)	-0.0022489	0.0041229	69.4	3
Erysipelotrichaceae (OTU 10)	-0.0017732	0.0045827	78.4	4
Ruminococcaceae (OTU 56)	-0.0015730	0.0047184	60.6	5
Lachnospiraceae (OTU 54)	-0.0015468	0.0051692	65.8	6
Lachnospiraceae (OTU 27)	-0.0012169	0.0046644	59.6	7
Bifidobacterium (OTU 12)	-0.0010722	0.0034798	75.0	8
Lachnospiraceae (OTU 98)	-0.0009233	0.0031928	86.4	9
Firmicutes (OTU 69)	-0.0008873	0.0051307	77.6	10

OTU	mean_decrease	stdev	% models	rank
Clostridium IV (OTU 126)	-0.0007600	0.0042812	45.2	11
Lachnospiraceae (OTU 23)	-0.0007161	0.0036369	60.4	12
Oscillibacter (OTU 164)	-0.0006656	0.0034880	78.2	13
Porphyromonadaceae (OTU 29)	-0.0006512	0.0032917	65.4	14
Lactobacillus (OTU 26)	-0.0006454	0.0030530	85.2	15
Acetatifactor (OTU 172)	-0.0006135	0.0026414	87.8	16
Lactobacillus (OTU 15)	-0.0005947	0.0026614	85.4	17
Dorea (OTU 149)	-0.0005938	0.0025554	84.8	18
Lachnospiraceae (OTU 38)	-0.0005561	0.0030798	85.6	19
Helicobacter (OTU 21)	-0.0005545	0.0023031	93.6	20

Machine learning methods

Supervised machine learning was performed according to the best practices outlined by Topçuolu *et al.* (1) and implemented in the mikropml R package v1.2.1 (2). Models were trained on the abundance data on day 0 of the experiment to predict the presence of *C. difficile* on day 1 of the experiment. The data were first pre-processed by centering and scaling abundance counts, collapsing perfectly correlated OTUs, and removing OTUs with zero variance. For 100 random seeds, the data were randomly split into training and testing sets with 65% and 35% of the samples in each, respectively. Random forest models were trained on the training sets using 5-fold cross-validation to select the best hyper-parameter value (`mtry`: the number of OTUs included per tree), then the best models were evaluated with the held-out test sets by computing the AUROC and AUPRC.

The most important OTUs contributing to model performance were determined by

permutation feature importance tests. For each trained model, each OTU in the test dataset was randomly shuffled 100 times and the new permutation performance (AUROC) was measured. A given OTU was considered significantly important for a model at an alpha level of 0.05, where less than 5% of the permutation AUROC values were greater than the original test AUROC. The OTUs that decreased the AUROC the most when permuted were considered the most important for model performance.

The workflow used to perform the machine learning analysis is available at https://github.com/SchlossLab/Barron_IBD-CDI_2022

References

1. **Topçuolu BD**, **Lesniak NA**, **Ruffin MT**, **Wiens J**, **Schloss PD**. 2020. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio **11**. doi:10.1128/mBio.00434-20.

2. **Topçuolu BD**, **Lapp Z**, **Sovacool KL**, **Snitkin E**, **Wiens J**, **Schloss PD**. 2021. Mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. JOSS **6**:3073. doi:10.21105/joss.03073.

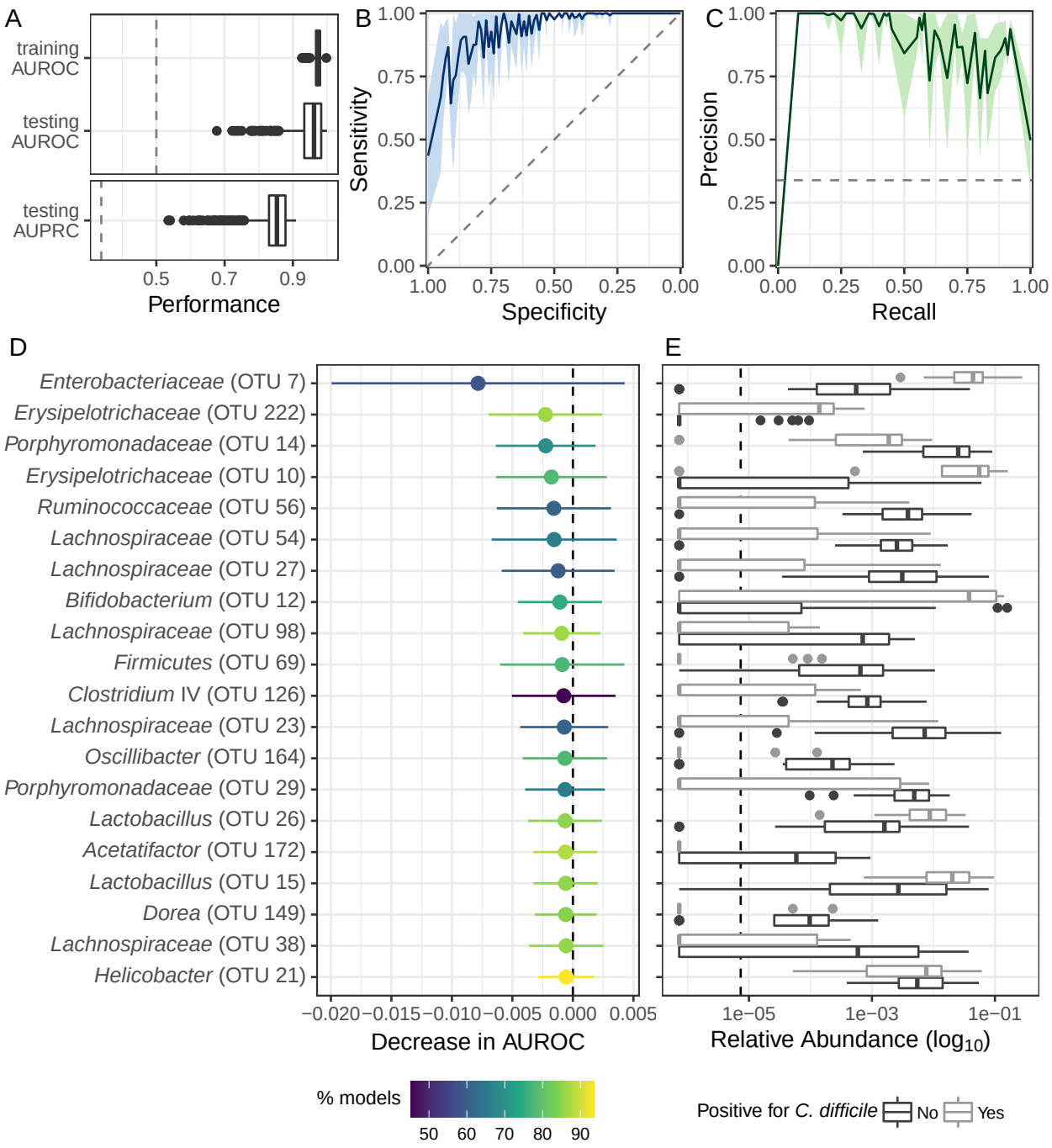

Figure 5. Machine learning analysis to predict *C. difficile* colonization. **A)** Mean area under the receiver-operator characteristic curve (AUROC) on the cross-validation folds during model training, mean AUROC on the held-out test data, and mean area under the precision-recall curve (AUPRC) on the held-out test data. The dashed grey lines represent the baseline AUROC (0.5) and AUPRC (0.34). **B)** Receiver-operator

characteristic curve for the test data, with mean specificity plotted against sensitivity. The light green shaded area shows the standard deviation. **C)** Precision-recall curve for the test data, with mean precision plotted against recall. The light blue shaded area shows the standard deviation. **D)** Top 20 most important OTUs as determined by permutation feature importance. OTUs with a greater decrease in AUROC when permuted are more important. The points represent the median decrease in AUROC with the tails as the standard deviation. Color represents the percentage of models for which an OTU's permutation AUROC was significantly different from the actual AUROC ($p < 0.05$). **E)** Log_{10} relative abundance for the top 20 most important OTUs on day 0 of the experiment, colored by *C. difficile* presence on day 1. The dashed line represents the limit of detection.