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)

Feature importance: top 20 OTUs

OTU	stdev	% models
Helicobacter (OTU 21)	0.0023031	93.6
Lachnospiraceae (OTU 38)	0.0030798	85.6
Dorea (OTU 149)	0.0025554	84.8
Lactobacillus (OTU 15)	0.0026614	85.4
Acetatifactor (OTU 172)	0.0026414	87.8
Lactobacillus (OTU 26)	0.0030530	85.2
Porphyromonadaceae (OTU 29)	0.0032917	65.4
Oscillibacter (OTU 164)	0.0034880	78.2
Lachnospiraceae (OTU 23)	0.0036369	60.4
Clostridium IV (OTU 126)	0.0042812	45.2

OTU	stdev	% models
Firmicutes (OTU 69)	0.0051307	77.6
Lachnospiraceae (OTU 98)	0.0031928	86.4
Bifidobacterium (OTU 12)	0.0034798	75.0
Lachnospiraceae (OTU 27)	0.0046644	59.6
Lachnospiraceae (OTU 54)	0.0051692	65.8
Ruminococcaceae (OTU 56)	0.0047184	60.6
Erysipelotrichaceae (OTU 10)	0.0045827	78.4
Porphyromonadaceae (OTU 14)	0.0041229	69.4
Erysipelotrichaceae (OTU 222)	0.0047096	86.4
Enterobacteriaceae (OTU 7)	0.0121111	58.0

Machine learning methods

TODO describe pipeline (1)

mikropml version 1.2.1 (2)

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.

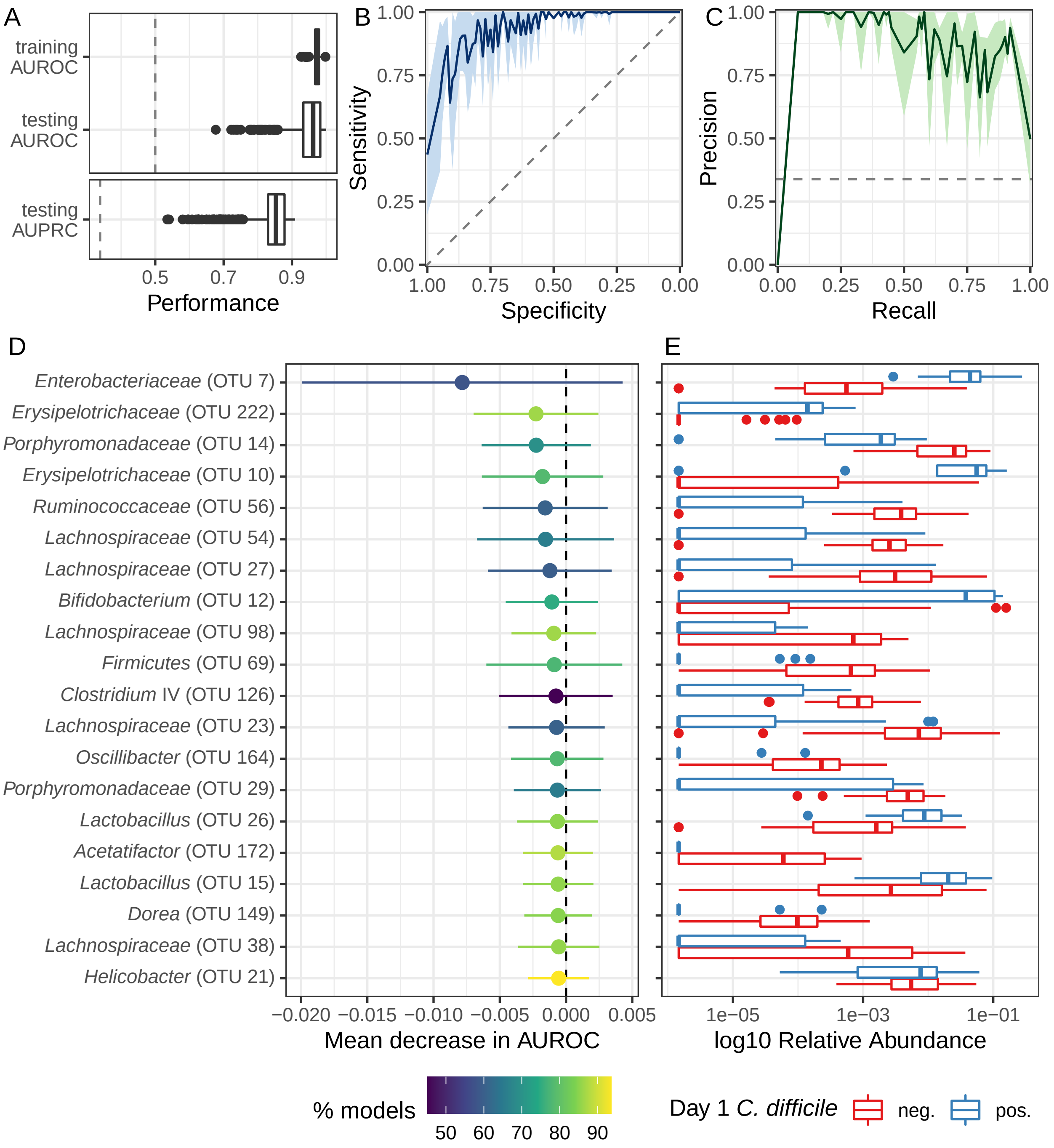

Figure 5. Machine learning analysis to predict *C. difficile* colonization. **A)** Mean area under the receiver-operator characteristic curve (AUROC) on the cross-validation folds during model training, mean AUROC on the held-out test data, and mean area under the precision-recall curve (AUPRC) on the 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. Mean specificity is plotted against sensitivity. The light green shaded area shows the standard deviation. **C)** Precision-recall curve for the test data. Mean precision is plotted against recall. The light blue shaded area shows the standard deviation. **D)** Top 20 most important OTUs as determined by permutation tests. Features with a larger decrease in AUROC when permuted are more important. The points are the median decrease in AUROC while the tails are the standard deviation. Color represents the percentage of models for which the feature's permutation AUROC was significantly different from the actual AUROC. **E)** Log$_{10}$ relative abundance for the top 20 most important OTUs on day 0 of the experiment.