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

- We now see that microbiota are sufficient for colonization susceptibility/resistance
- Some taxa are suggestive of being protective vs unprotective (they have cropped up multiple times; think Lachno and Porphyro vs Entero and Lacto)
- Goal for this section: Generate a model through which to predict susceptibility based on microbiota
- Samples:
  - 16S sequences from all experiments.
  - Determine whether susceptible based on who was colonized at any point throughout experiment
    * Random Forest
    * Taxa that were predictive
- This is a hypothesis generating step to computationally identify relevant taxa to advance future biological/mechanistic investigations.

performance 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)

TODO feature importance

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.

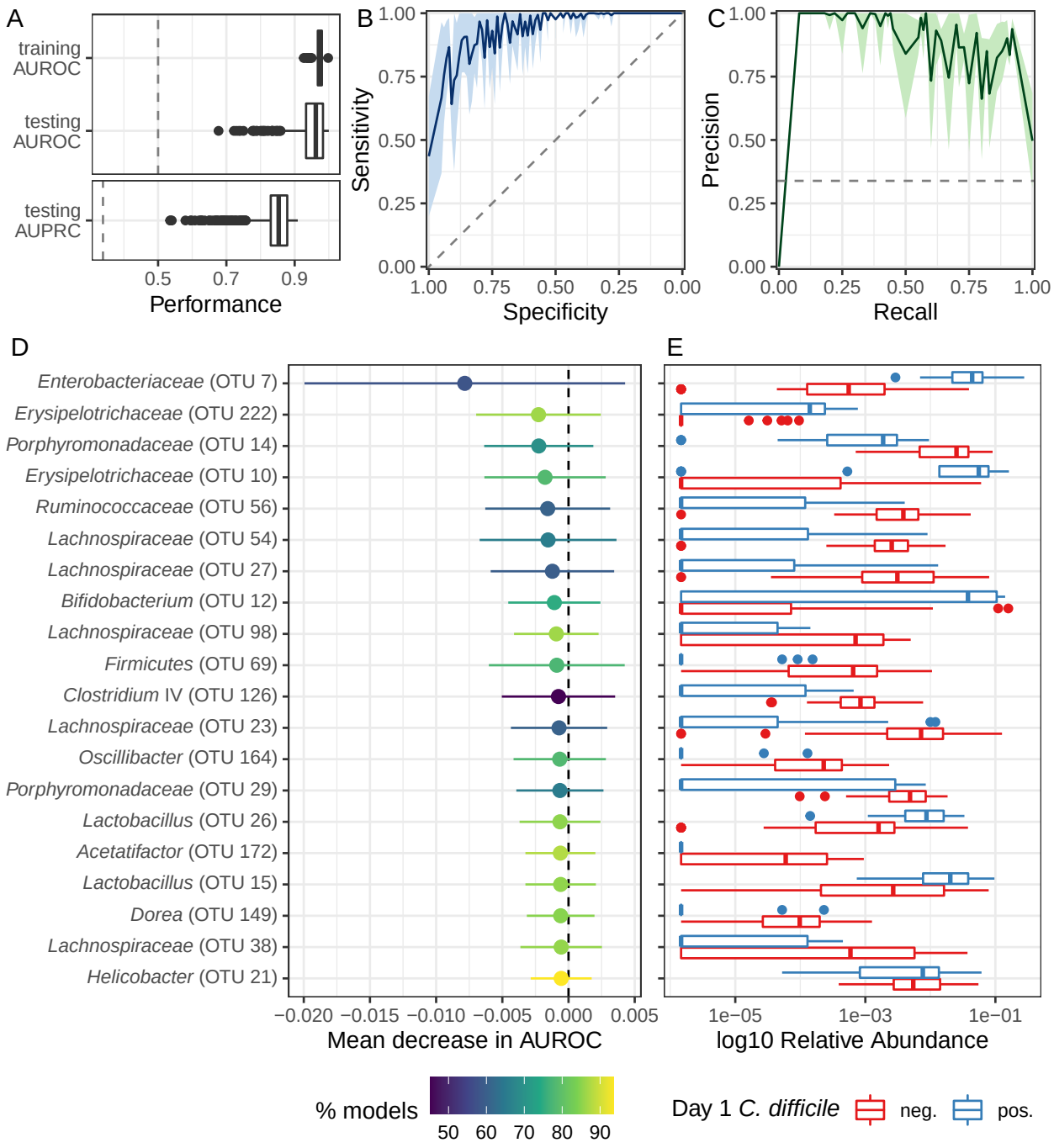

Figure 5. Machine learning analysis to predict *C. difficile* colonization. **A)** Mean area under the receiver-operator characteristic curve (AUROC) on the cross-validation folds during model training, mean AUROC on the held-out test data, and mean area under the precision-recall curve (AUPRC) on the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