

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).

Mean AUROC 0.95 (s.d. 0.029)

Mean AUPRC 0.85 (s.d. 0.039)

TODO feature importance

Figure 5

TODO caption

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. **Topçuoğlu 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. **Topçuoğlu 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.