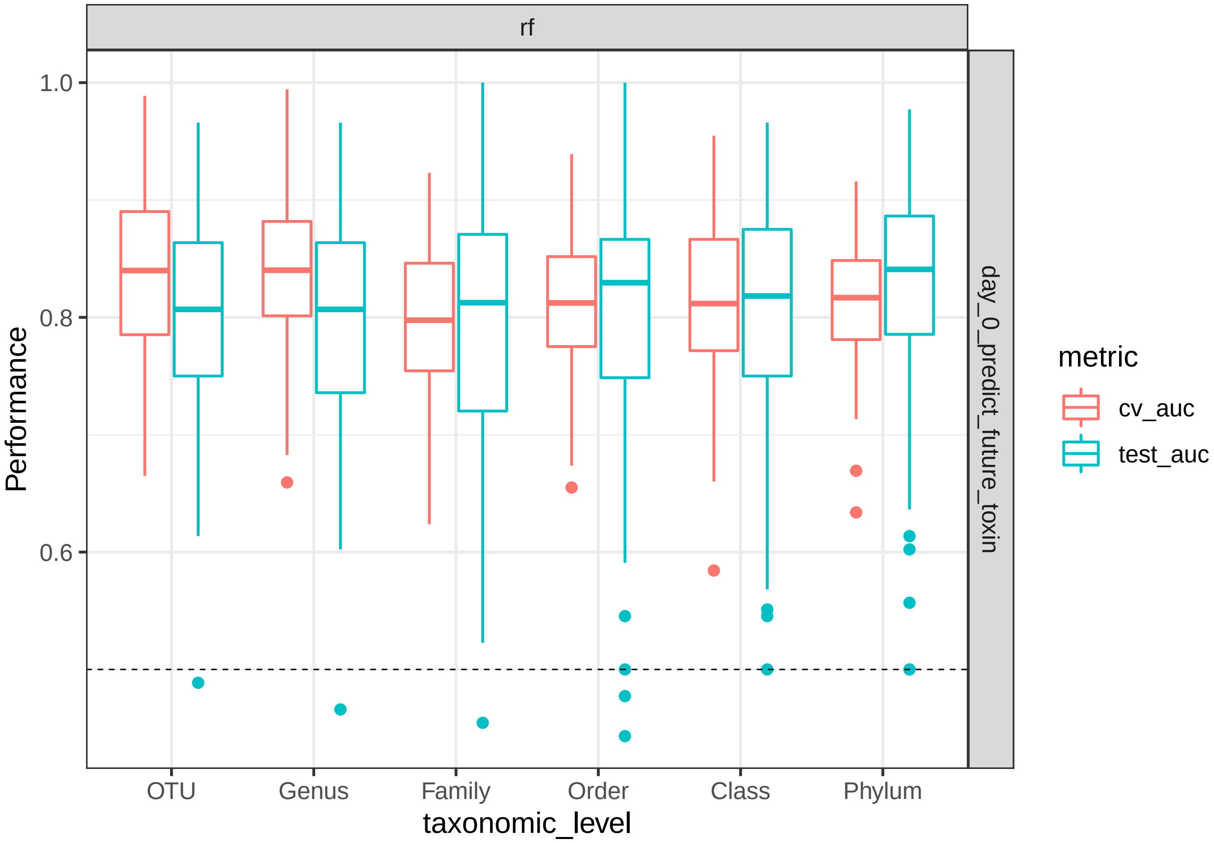
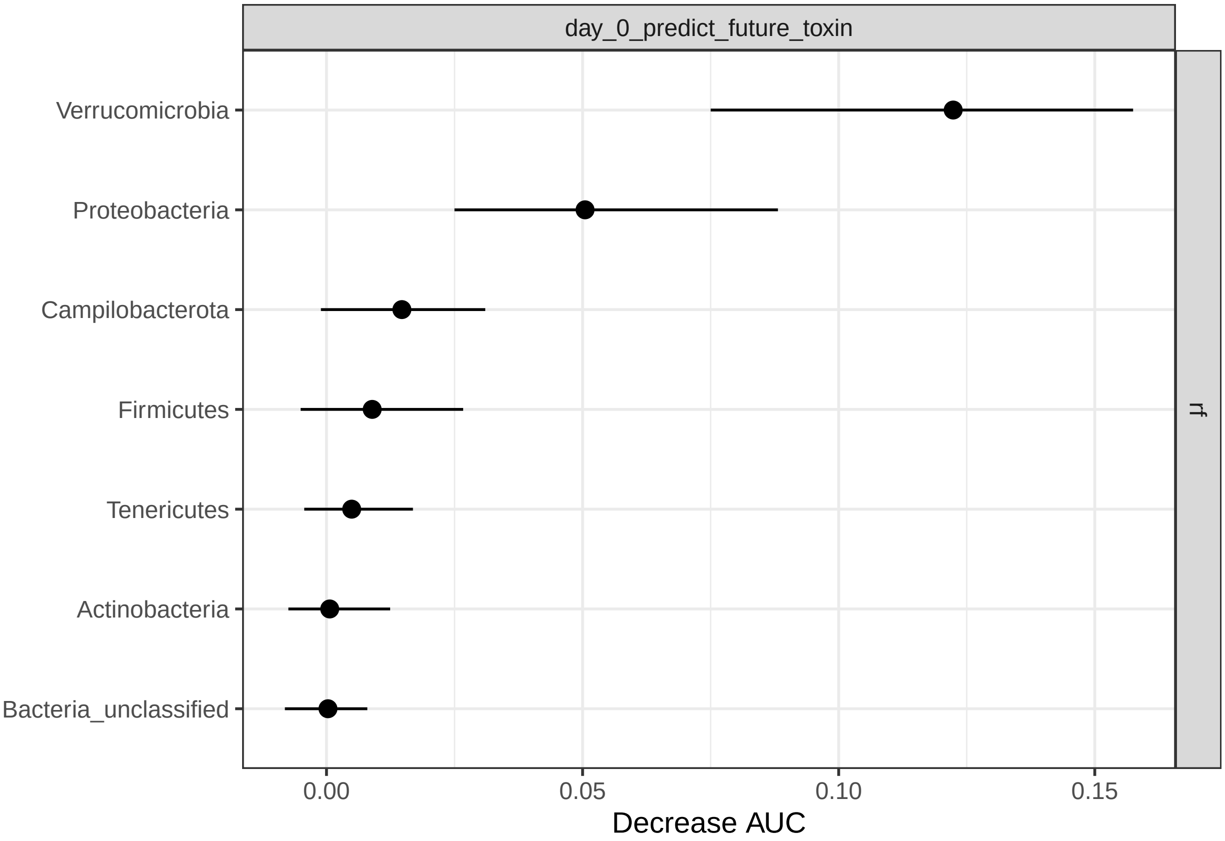
feature_correlation_notebook

Nick Lesniak

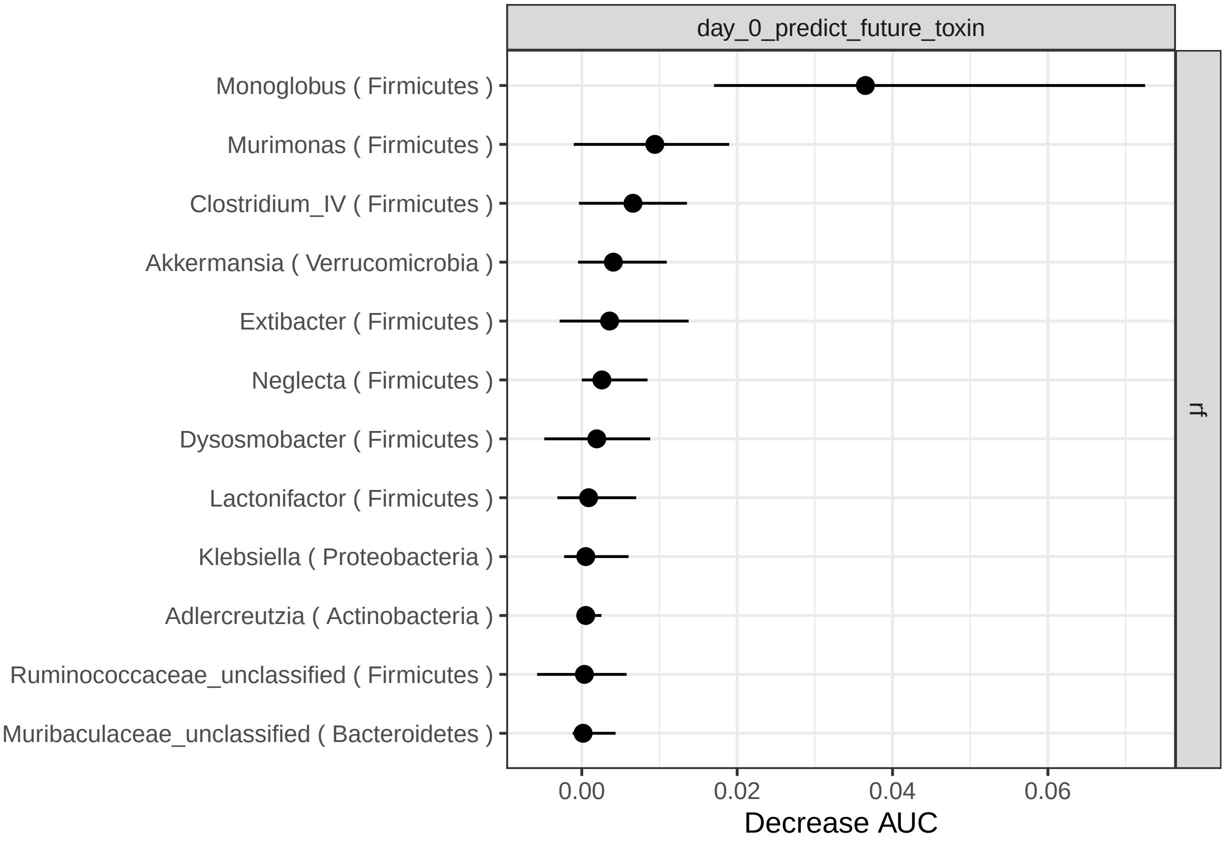
6/16/2021



Using RF at the Phylum level the features with median differences greater than 0 are:



Verrucomicrobia (Akkermansia) has a median decrease in AUC of 0.12 Using RF at the Genus level the features with median differences greater than 0 are:



Akkermansia now has a median decrease in AUC of less than 0.005 runnning my model at all taxa levels, Verrucomicrobia (akkermansia) is top feature with median decrease in AUC of 0.12 for phylum but at the genus level it drops to 0.005 but both models overall perform the same. only one genus correlates close to akkermansia (0.8) but has low effect on auc. so i was trying to think if its possible another set of features could be replace the information lost by akkermansia, but unsure how to show or evaluate that possibility

- run ml to predict akkermansia
- add multiple genera together and check for correlation between akkermansia and combined genera sets
- compare distibution of abundance relative to outcome

Are other genera or combinations of genera able to replace signal from akkermansia?

Correlation with Akkermansia

Feature	P value	Rho
Akkermansia	0.00e00	1.00
Rubneribacter	1.50e-12	0.812

Try combining features and then testing correlations w/ Akkermansia

Correlation to Akkermansia (only running up to the combination of Ileibacterium_Rikenellaceae_unclassified)

Feature	P value	Rho
Alloiococcus	1.412915e-03	0.443538316
Enterobacterales_unclassified	1.027612e-01	0.235874618

Feature	P value	Rho
Alloiococcus_Enterobacterales_unclassified	0.0002539063	0.5000263

Could the effect in decrease AUC be due to the number of features being much larger than the number of samples?

This model has 49 samples
- the phylum level model has 11 features
- the genus level model has 111 features

The literature suggests as few as 5 samples/feature but Ploeg et al says 20-50 samples/feature for LR and >200 was insufficient in some cases for RF, but LR can use penalization to reduce the number of features