Analysis of Binary Classification Results Using Imputed Drug Signatures

Rachel Hodos

Aug 3, 2017

Load the results

```
load('../results/classification/2017-07-30-03-47-00/results_ROC_counts_params.RData')
load('../results/classification/2017-07-30-03-47-00/RC.RData')
```

Do some processing...

```
# Melt AUC results
R = melt(ROC)
names(R) = c('AUC', 'eval', 'model', 'subset', 'feature', 'outcome')

# Split outcome into outcome and category
R = SplitOutcome(R)

# Subset to top three represented ATCs
R = subset(R, outcome %in% c('L','C','D') | category == 'Target')

# Subset to evaluations on measured signatures and reformat data
Rmeas = RemoveDfColumns(subset(R, eval == 'eval_meas'), 'eval')
Rmeas$AUC[is.na(Rmeas$AUC)] = 0.5
Rmeas = dcast(Rmeas, model + feature + outcome + category ~ subset, value.var='AUC')
Rmeas = suppressWarnings(ChangeColumnName(Rmeas, from=c('full','obs'), to=c('AUC.full', 'AUC.obs')))
Rmeas$diff = Rmeas$AUC.full - Rmeas$AUC.obs
idx = Rmeas$category == 'ATC'
Rmeas$outcome[idx] = paste0('ATC ' , Rmeas$outcome[idx])

# Results where AUCs were both below this in the two comparisons are thrown out
threshold = 0.5
Rmeas = FilterByAUC(Rmeas, threshold)
```

Analysis of overall results

Look at the overall difference between the two groups:

```
t.test(x=Rmeas$AUC.full, y=Rmeas$AUC.obs, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  Rmeas$AUC.full and Rmeas$AUC.obs
## t = 6.1389, df = 338, p-value = 2.329e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02200544 0.04275600
## sample estimates:
## mean of the differences
##              0.03238072
```

Then count the number of cases where there is a positive change in AUC:

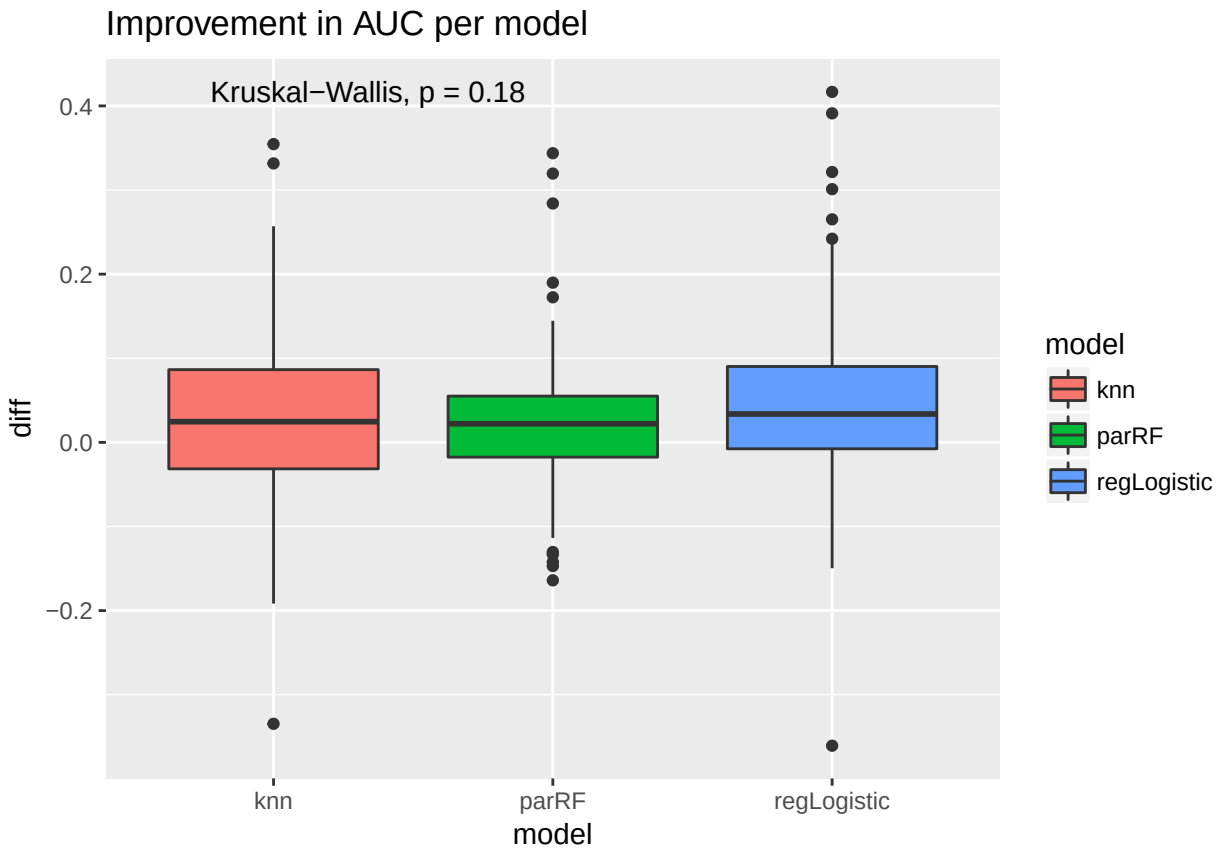
```
nPos = length(which(Rmeas$diff > 0))
nTot = nrow(Rmeas)
print(sprintf('%d out of %d increased AUC', nPos, nTot))
```

```
## [1] "223 out of 339 increased AUC"
```

Analysis of results per model

Plot distributions.

```
print(ggplot(Rmeas, aes(x=model, y=diff, group=model, fill=model)) + geom_boxplot() + ggtitle('Improveme
```



Check significance of deltas per model.

```
print(lapply(split(Rmeas, Rmeas$model), function(x) t.test(x$AUC.full, x$AUC.obs, paired=TRUE)))
```

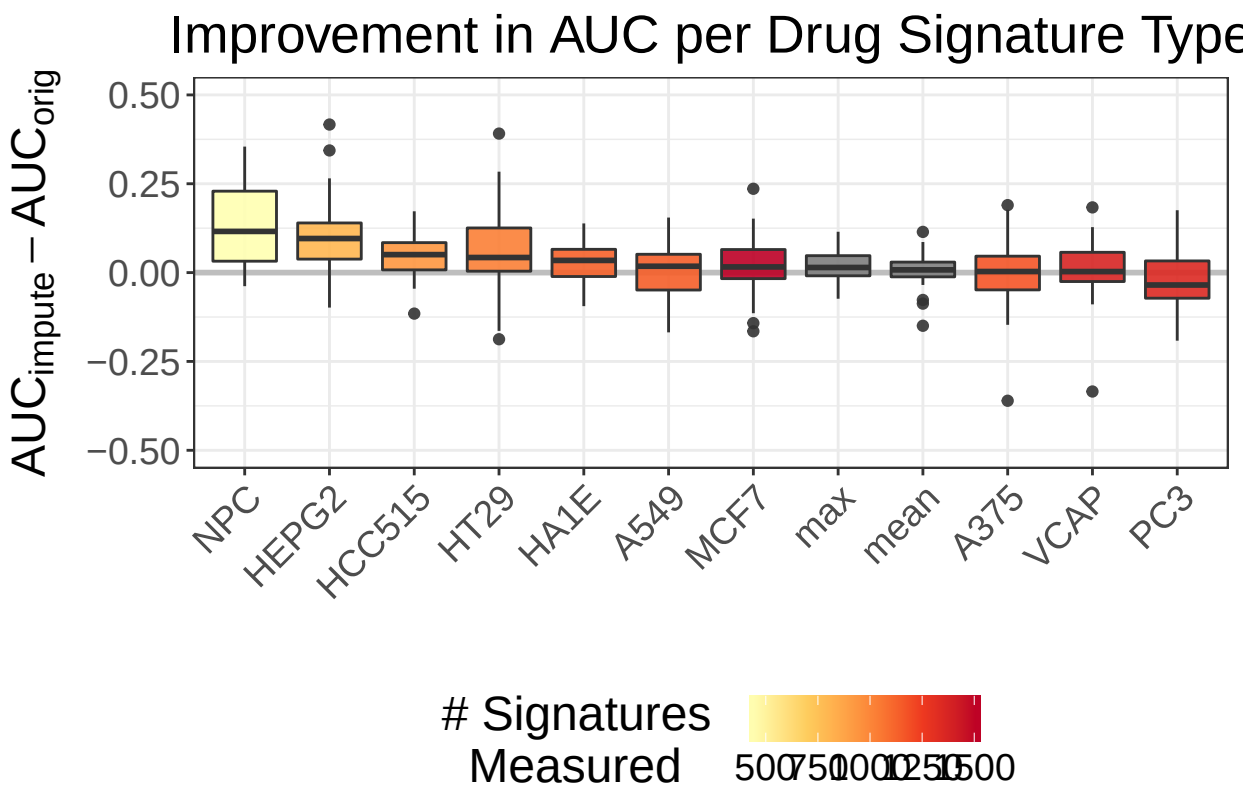
```
## $knn
##
##  Paired t-test
##
## data:  x$AUC.full and x$AUC.obs
## t = 2.6128, df = 117, p-value = 0.01016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.006023552 0.043751360
## sample estimates:
## mean of the differences
```

```
##              0.02488746
##
##
## $parRF
##
##  Paired t-test
##
## data:  x$AUC.full and x$AUC.obs
## t = 3.0247, df = 108, p-value = 0.00311
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.008209386 0.039425494
## sample estimates:
## mean of the differences
##              0.02381744
##
##
## $regLogistic
##
##  Paired t-test
##
## data:  x$AUC.full and x$AUC.obs
## t = 5.0294, df = 111, p-value = 1.907e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02945744 0.06776118
## sample estimates:
## mean of the differences
##              0.04860931
```

Analysis of results per signature type

```
numSigs = c(MCF7=1505, VCAP=1368, PC3=1340, A375=1168, A549=1139,
            HA1E=1127, HT29=1022, HCC515=934, HEPG2=798, NPC=441)
Rmeas$num_sigs = numSigs[Rmeas$feature]

ggplot(Rmeas, aes(x=reorder(feature, -diff, FUN=median), y=diff, group=feature, fill=num_sigs)) +  geom
  geom_boxplot(alpha=0.9) +
  ggtitle(sprintf('Deltas per feature, threshold = %0.1f', threshold)) +
  theme_bw() + ylim(c(-0.5,0.5)) + scale_fill_gradientn(colours=brewer.pal(5,'YlOrRd')) +
  theme(text = element_text(size=18), axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size=20, hjust=0.5),
        legend.justification='bottom', legend.position='bottom') +
  labs(fill='# Signatures  \n  Measured\n', x='', y=expression(AUC [impute] - AUC [orig]),
       title='Improvement in AUC per Drug Signature Type')
```

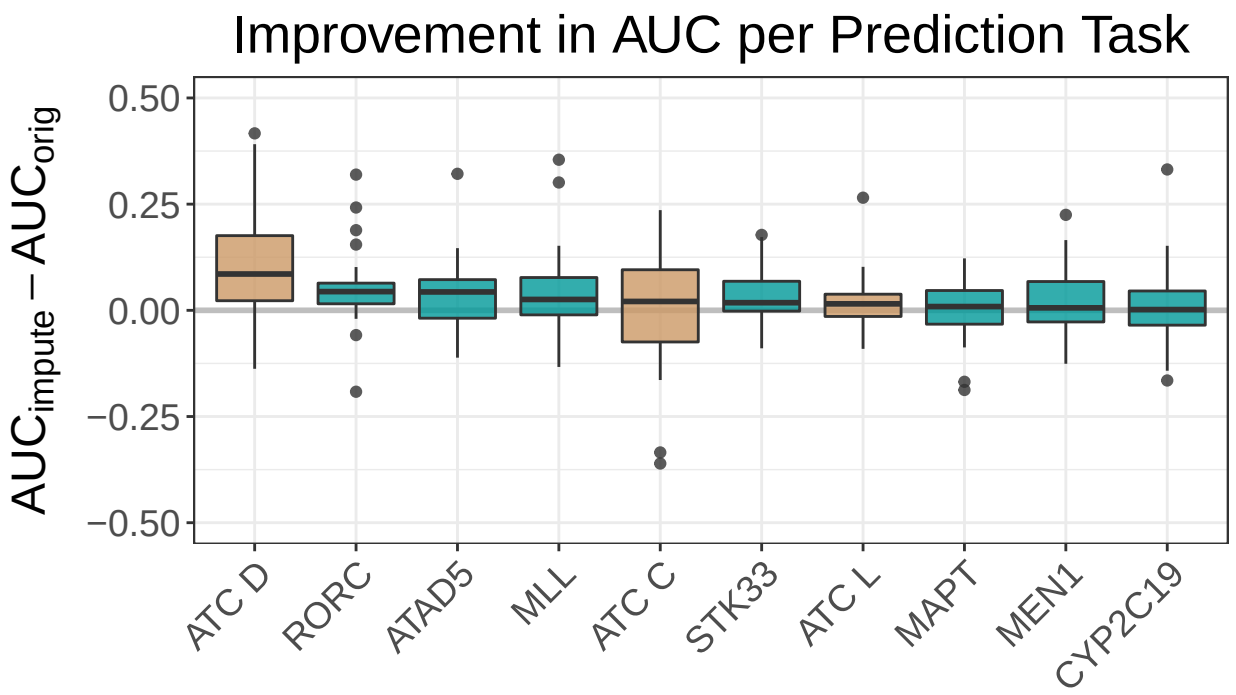



```
# ggsave(PlotDir('DeltaAUC_per_Feature.svg'), height=7, width=8)
```

```
p_feature = sapply(split(Rmeas, Rmeas$feature), function(x) t.test(x$AUC.full, x$AUC.obs, paired=TRUE)$p
print(adjp_feature <- sort(p.adjust(p_feature, method='BH')))
```

```
##          NPC       HCC515        HEPG2         HA1E         HT29
## 8.423282e-05 7.981616e-04 7.981616e-04 1.212124e-02 5.966839e-02
##          max         MCF7          PC3         mean         VCAP
## 5.966839e-02 2.747964e-01 5.020732e-01 7.338712e-01 8.067996e-01
##         A375         A549
## 8.545544e-01 8.545544e-01
```

```
ggplot(Rmeas, aes(x=reorder(outcome, -diff, FUN=median), y=diff, group=outcome, fill=category)) + ylim(
  geom_hline(yintercept = 0, color='grey', lwd=1) + geom_boxplot(alpha=0.8) + scale_fill_manual(values=
  ggtitle(sprintf('Deltas per outcome, threshold = %0.1f', threshold)) + theme_bw() +
  labs(x='', y=expression(AUC [impute] - AUC [orig]), title='Improvement in AUC per Prediction Task',
       fill='Category') +
  theme(text = element_text(size=18), axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size=20, hjust=0.5),
        legend.justification='bottom', legend.position='bottom')
```

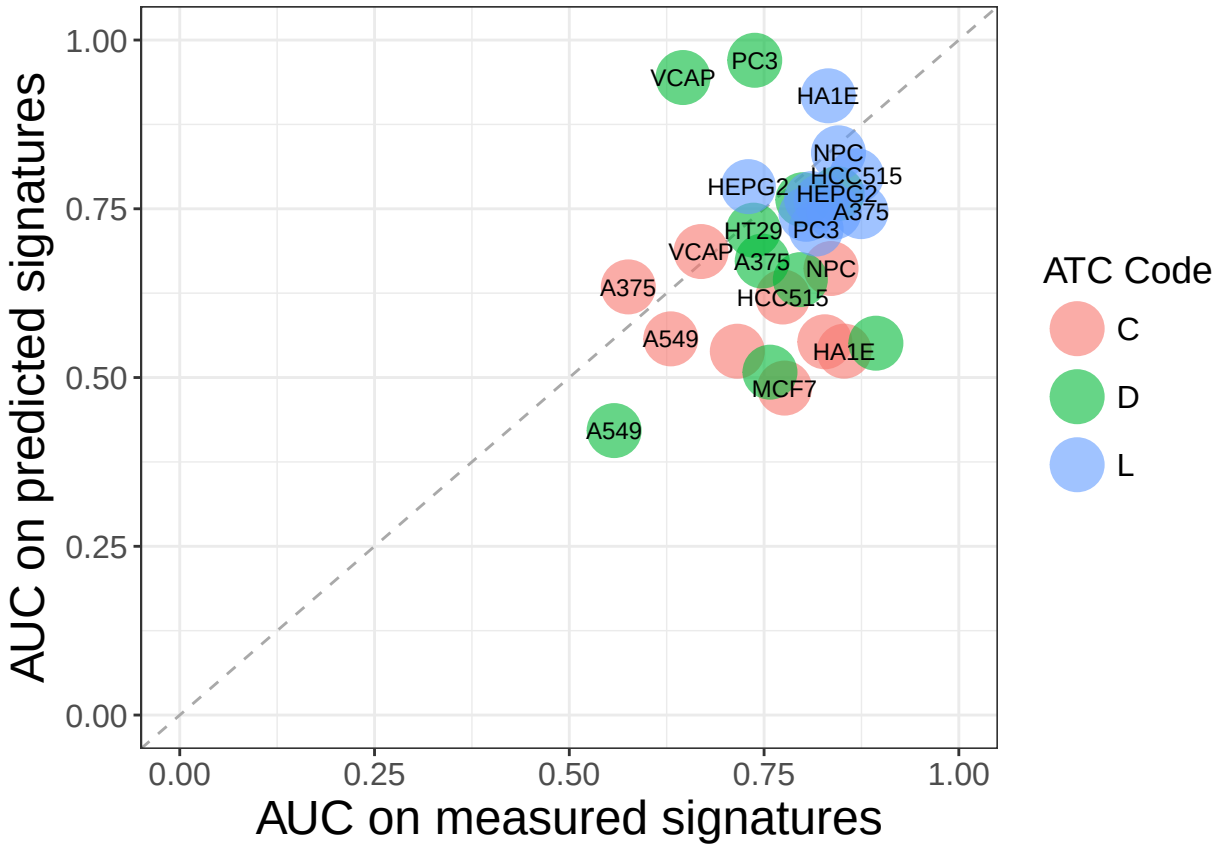



```
# ggsave(PlotDir('DeltaAUC_per_Outcome.svg'), height=7, width=8)
```

```
p_outcome = lapply(split(Rmeas, Rmeas$outcome), function(x) t.test(x$AUC.full, x$AUC.obs, paired=TRUE)$p
print(adjp_outcome <- sort(p.adjust(p_outcome, method='BH')))
```

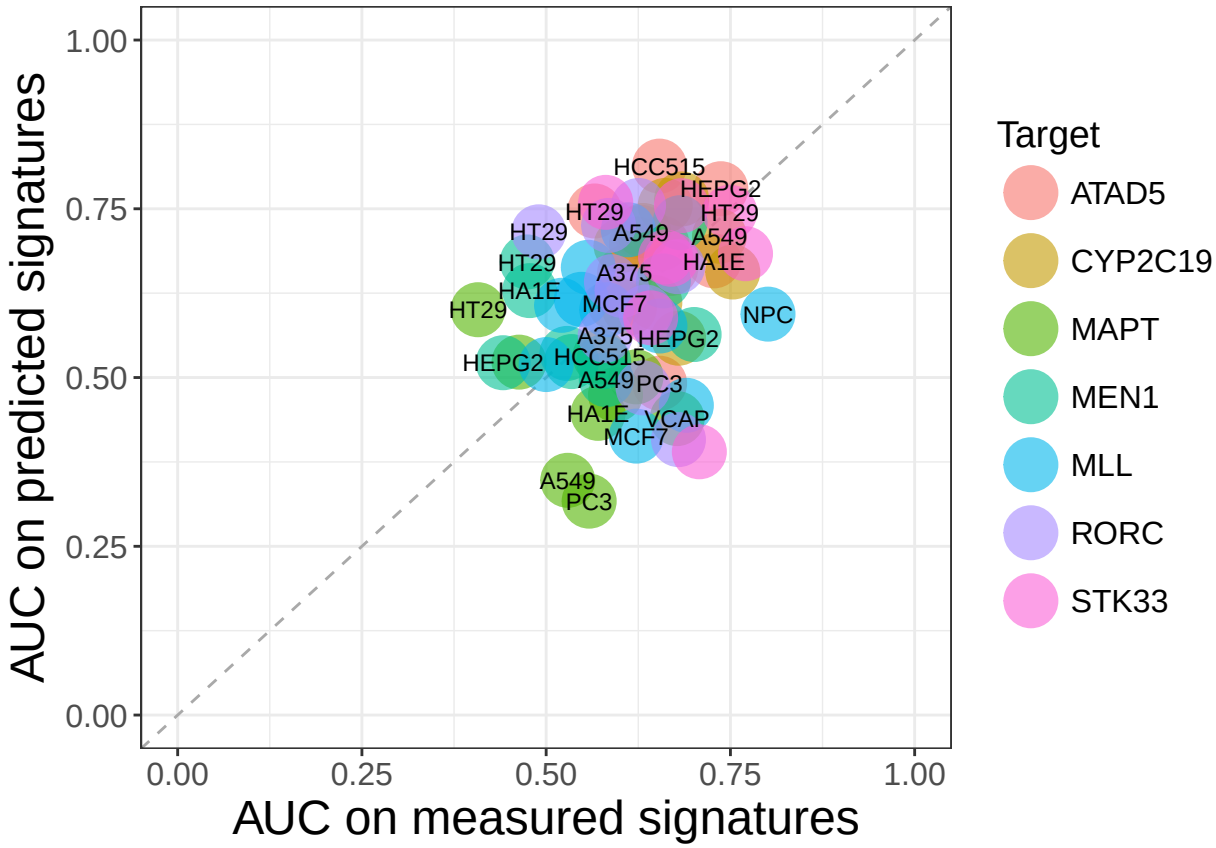
```
##        ATC D         RORC        ATAD5        STK33          MLL
## 0.0001256311 0.0082305315 0.0100504714 0.0100504714 0.0513355468
##        ATC L         MEN1        ATC C      CYP2C19         MAPT
## 0.1608094447 0.3687120271 0.8957620459 0.8957620459 0.8957620459
```

```
ggplot(ATC, aes(x=AUC_meas, y=AUC_imp, color=outcome, label=feature)) +
  geom_abline(slope=1, intercept=0, color='DarkGrey', lty='dashed') + geom_point(size=9, alpha=0.6) +
  xlim(c(0,1)) + ylim(c(0,1)) +
  geom_text(size=3, color='black', check_overlap=TRUE) + theme_bw() +
  guides(color=guide_legend(title='ATC Code')) +
  xlab('AUC on measured signatures') + ylab('AUC on predicted signatures') +
  theme(axis.text=element_text(size=12), axis.title=element_text(size=18),
        legend.text=element_text(size=12), legend.title=element_text(size=14))
```




```
# ggsave(PlotDir('ATC_code_imputed_vs_measured.svg'), width=8, height=7)
```

```
ggplot(Targets, aes(x=AUC_meas, y=AUC_imp, color=outcome, label=feature)) +
  geom_abline(slope=1, intercept=0, color='DarkGrey', lty='dashed') + geom_point(size=9, alpha=0.6) +
  xlim(c(0,1)) + ylim(c(0,1)) +
  geom_text(size=3, color='black', check_overlap=TRUE) + theme_bw() + guides(color=guide_legend(title='
  xlab('AUC on measured signatures') + ylab('AUC on predicted signatures') +
  theme(axis.text=element_text(size=12), axis.title=element_text(size=18),
        legend.text=element_text(size=12), legend.title=element_text(size=14))
```




```
#ggsave(PlotDir('Target_imputed_vs_measured.svg'), width=8.5, height=7)
```