

DEB RNA FAIMS S100

Load required packages

```
## corrplot 0.84 loaded
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.0    v purrr  0.3.4
## v tidyr   1.0.3    v forcats 0.5.0
## v readr   1.3.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'kableExtra'
## The following object is masked from 'package:dplyr':
##
##   group_rows
##
## Attaching package: 'magrittr'
## The following object is masked from 'package:purrr':
##
##   set_names
## The following object is masked from 'package:tidyr':
##
##   extract
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##   combine
##
## *****
```

```

## Note: As of version 1.0.0, cowplot does not change the
## default ggplot2 theme anymore. To recover the previous
## behavior, execute:
## theme_set(theme_cowplot())
## *****
##
## Attaching package: 'cowplot'
## The following object is masked from 'package:ggpubr':
##
## get_legend
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
## src, summarize
## The following objects are masked from 'package:base':
##
## format.pval, units
##
## Attaching package: 'caret'
## The following object is masked from 'package:survival':
##
## cluster
## The following object is masked from 'package:purrr':
##
## lift
## Loading required package: usethis

```

Loading RNP:XI data into the R environment

I included the manually evaluated datasets for my ASMS talk to see if machine learning can help me with identifying a possible score cutoff or other metric to help reduce the numbers of entries I have to go through after RNP:xl output. The datasets include a merged version of both the S100 and S100 fraction that were initially considered individually:

*Ecoli DEB RNA

Data wrangling

First, we need to add an extra column to each dataset, indicating which cross-linker was used in the experiment.

Then, we need to combine the dataframes.

First filtering step

Remove any entry that does not contain an RNA adduct and fits to a target better than a decoy.

Compute additional metrics (columns)

Peplength

classification of amino acids

Create new column with peptide length and classify identified amino acids in peptides as follows:

1. charged amino acids: R, K, D, E
2. polar amino acids: Q, N, H, S, T, Y, C, W
3. hydrophobic amino acids: A, I, L, M, V, P, G

Peptide pI (isoelectric point)

Compute the isoelectric point for each peptide in each entry.

aI (aliphatic index)

Compute the aliphatic Index for each peptide in each entry. See the following paper for more info:

Thermostability and aliphatic index of globular proteins. J Biochem. 1980 Dec;88(6):1895-8.

“A statistical analysis shows that the aliphatic index, which is defined as the relative volume of a protein occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine), of proteins of thermophilic bacteria is significantly higher than that of ordinary proteins. The index may be regarded as a positive factor for the increase of thermostability of globular proteins.”

In general, an aI of greater 92.6 is particularly stable thermally as found in thermophiles. Now, since we are investigating tryptic peptides, I could only find one paper on that matter.

GETTING INTIMATE WITH TRYPSIN, THE LEADING PROTEASE IN PROTEOMICS Published online 15 June 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/mas.21376

According to this paper, human proteins result in 61 tryptic peptides on average.

proposed cross-linking amino acid

Hydrophobicity index

Molecular weight of peptide

Note: returns the average masses, not the monoisotopic masses!

Machine learning

Split dataset

Separate different species of nucleotides (RNA, DNA) and cross-linker (DEB, DEB).

Setting up the dataset

Harmonize the number of trues and falses and dropping all categorical columns.50% T and %0 F

```
set.seed(42)
drops_general<-c("precursor_purity", "RNPxl_a_ion_score")
```

```

ml_DEB_RNA<-data1 %>%
  filter(XLinker == "DEB", Curated == "0", Nucleotides == "RNA") %>%
  sample_n(459) %>%
  rbind(RNA_DEB)
rows_DEB_RNA <- sample(nrow(ml_DEB_RNA))
ml_DEB_RNA <- ml_DEB_RNA[rows_DEB_RNA, ] %>%
  dplyr::select_if(is.numeric)
ml_DEB_RNA<-ml_DEB_RNA[, !(colnames(ml_DEB_RNA) %in% drops_general)]

```

Splitting into train and test data

I will use 80% of the data to train, and 20% to test the model

```

ml_DEB_RNA$Curated<-as.factor(ml_DEB_RNA$Curated)
ml_DEB_RNA$Curated<-factor(ml_DEB_RNA$Curated, levels = c("0", "1"), labels = c("False", "True"))
ml_DEB_RNA_train <- ml_DEB_RNA[1:734, ]
ml_DEB_RNA_test <- ml_DEB_RNA[735:918, ]

```

KNN

K nearest neighbor classification based on numeric variables, min-max normalized.

```

library(class)
library(gmodels)

knn_drop <- c("Curated", "isotope_error", "RNPxl_a_ion_score", "RNPxl_precursor_purity")

normalize <- function(x) {
  return((x-min(x))/(max(x)-min(x)))
}

ml_DEB_RNA_test_knn<-ml_DEB_RNA_test[,!(names(ml_DEB_RNA_test) %in% knn_drop)]
ml_DEB_RNA_test_knn<-as.data.frame(lapply(ml_DEB_RNA_test_knn[1:39], normalize))
ml_DEB_RNA_test_knn_labels<-ml_DEB_RNA_test[1:184, 5]

ml_DEB_RNA_train_knn<-ml_DEB_RNA_train[,!(names(ml_DEB_RNA_train) %in% knn_drop)]
ml_DEB_RNA_train_knn<-as.data.frame(lapply(ml_DEB_RNA_train_knn[1:39], normalize))
ml_DEB_RNA_train_knn_labels<-as.data.frame(ml_DEB_RNA_train[1:734, 5])

```

Evaluate Model Performance DEB RNA

Here, I chose 5 neighbors:

```

DEB_RNA_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = ml_DEB_RNA_test_knn,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5 )
CrossTable(x = ml_DEB_RNA_test_knn_labels$Curated, y = DEB_RNA_knn_pred, prop.chisq = F)

```

```

##
##
##   Cell Contents
## |-----|
## |                N |
## |      N / Row Total |
## |      N / Col Total |

```

```
## |          N / Table Total |
## |-----|
##
##
## Total Observations in Table:  184
##
##
##                | DEB_RNA_knn_pred
## ml_DEB_RNA_test_knn_labels$Curated |      False |      True | Row Total |
## -----|-----|-----|-----|
##                False |          79 |          17 |          96 |
##                |      0.823 |      0.177 |      0.522 |
##                |      0.859 |      0.185 |           |
##                |      0.429 |      0.092 |           |
## -----|-----|-----|-----|
##                True |          13 |          75 |          88 |
##                |      0.148 |      0.852 |      0.478 |
##                |      0.141 |      0.815 |           |
##                |      0.071 |      0.408 |           |
## -----|-----|-----|-----|
##                Column Total |          92 |          92 |          184 |
##                |      0.500 |      0.500 |           |
## -----|-----|-----|-----|
##
##
```

Decision trees and classification rules

C5.0

```
library(C50)

C5_drop <- c("isotope_error", "RNPxl_a_ion_score", "precursor_purity")

ml_DEB_RNA_train_C5<-ml_DEB_RNA_train[!(names(ml_DEB_RNA_train) %in% C5_drop)]
ml_DEB_RNA_test_C5<-ml_DEB_RNA_test[!(names(ml_DEB_RNA_test) %in% C5_drop)]
```

DEB RNA

with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

C5_model_DEB_RNA_boost_pen<-C5.0(x=ml_DEB_RNA_train_C5[-5], y=ml_DEB_RNA_train_C5$Curated,
                                costs = error_cost, trials = 10)
summary(C5_model_DEB_RNA_boost_pen)

##
## Call:
## C5.0.default(x = ml_DEB_RNA_train_C5[-5], y =
```

```

## ml_DEB_RNA_train_C5$Curated, trials = 10, costs = error_cost)
##
##
## C5.0 [Release 2.07 GPL Edition]      Wed Jun 23 10:46:29 2021
## -----
##
## Class specified by attribute `outcome'
##
## Read 734 cases (40 attributes) from undefined.data
## Read misclassification costs from undefined.costs
##
## ----- Trial 0: -----
##
## Decision tree:
##
## RNPxl_total_MIC > 0.2365065:
## :...aa_class_polar > 4:
## :   :...RNPxl_pl_im_MIC <= 0.01192007: True (5)
## :   :   RNPxl_pl_im_MIC > 0.01192007: False (4/1)
## :   aa_class_polar <= 4:
## :     :...Hydro <= 0.6: True (205/3)
## :     :   Hydro > 0.6:
## :       :...pI <= 9.45: False (6/3)
## :       :   pI > 9.45: True (8)
## RNPxl_total_MIC <= 0.2365065:
## :...aa_class_hydrophobic > 3:
## :   :...RNPxl_pl_pc_MIC <= 0: False (322/26)
## :   :   RNPxl_pl_pc_MIC > 0:
## :     :...RNPxl_pl_pc_MIC > 0.04382946: True (8)
## :     :   RNPxl_pl_pc_MIC <= 0.04382946:
## :       :...RNPxl_modds <= 0.3741593: False (25/3)
## :       :   RNPxl_modds > 0.3741593:
## :         :...RNPxl_RNA_MASS_z0 <= 449.0948: False (5)
## :         :   RNPxl_RNA_MASS_z0 > 449.0948: True (14)
## aa_class_hydrophobic <= 3:
## :...aa_class_charged <= 2:
## :   :...RNPxl_pl_modds <= 0.4927427: False (14/9)
## :   :   RNPxl_pl_modds > 0.4927427:
## :     :...RNPxl_precursor_score <= 0.001193983: True (60)
## :     :   RNPxl_precursor_score > 0.001193983: False (1)
## aa_class_charged > 2:
## :...RT <= 818.5449: True (7)
## :   RT > 818.5449:
## :     :...RNPxl_pl_modds <= 0.6882493: False (19/2)
## :     :   RNPxl_pl_modds > 0.6882493:
## :       :...pI <= 4.43: True (10)
## :       :   pI > 4.43:
## :         :...RNPxl_pl_err <= 0.006021813: False (8/1)
## :         :   RNPxl_pl_err > 0.006021813:
## :           :...RNPxl_immonium_score <= 0.02881235: True (12)
## :           :   RNPxl_immonium_score > 0.02881235: False (1)
##
## ----- Trial 1: -----
##

```

```

## Decision tree:
##
## mw > 1856:
## :...RNPxl_pl_modds <= 0.8764216: False (128.9/3)
## :   RNPxl_pl_modds > 0.8764216:
## :     :...RNPxl_err <= 0.002779919: True (40.4/0.8)
## :     :   RNPxl_err > 0.002779919: False (45.6/22.1)
## mw <= 1856:
## :...RNPxl_pl_modds <= 0.04243729: False (18.2/2.3)
## :   RNPxl_pl_modds > 0.04243729:
## :     :...RNPxl_err > 0.009156602: True (73.1/0.8)
## :     :   RNPxl_err <= 0.009156602:
## :     :     :...precursor > 973.7704: False (19/3)
## :     :     :   precursor <= 973.7704:
## :     :     :     :...RNPxl_total_MIC > 0.3725607: True (79.6)
## :     :     :     :   RNPxl_total_MIC <= 0.3725607:
## :     :     :     :     :...aa_class_hydrophobic > 7: False (31.1/14.4)
## :     :     :     :     :   aa_class_hydrophobic <= 7:
## :     :     :     :     :     :...pI > 11.13: False (16.7/5.3)
## :     :     :     :     :     :   pI <= 11.13:
## :     :     :     :     :     :     :...RNPxl_immonium_score > 0.008700988:
## :     :     :     :     :     :     :   :...Peplength <= 7: True (51.7/3)
## :     :     :     :     :     :     :   :   Peplength > 7:
## :     :     :     :     :     :     :   :     :...RNPxl_marker_ions_score > 0.0006902145: True (22.8/1.5)
## :     :     :     :     :     :     :   :     :   RNPxl_marker_ions_score <= 0.0006902145: [S1]
## :     :     :     :     :     :     :   RNPxl_immonium_score <= 0.008700988:
## :     :     :     :     :     :     :     :...RNPxl_partial_loss_score > 6.786638: True (100.5)
## :     :     :     :     :     :     :     :   RNPxl_partial_loss_score <= 6.786638:
## :     :     :     :     :     :     :     :     :...Peplength > 10: False (12.1/1.5)
## :     :     :     :     :     :     :     :     :   Peplength <= 10:
## :     :     :     :     :     :     :     :     :     :...aIndex > 128.12: False (5.3/0.8)
## :     :     :     :     :     :     :     :     :     :   aIndex <= 128.12:
## :     :     :     :     :     :     :     :     :     :     :...Peplength > 9: True (69.2)
## :     :     :     :     :     :     :     :     :     :     :   Peplength <= 9:
## :     :     :     :     :     :     :     :     :     :     :     :...pI <= 4.18: False (8.3/3.8)
## :     :     :     :     :     :     :     :     :     :     :     :   pI > 4.18: [S2]
##
## SubTree [S1]
##
## RNPxl_partial_loss_score > 13.77283: True (9.9)
## RNPxl_partial_loss_score <= 13.77283:
## :...mw <= 861.97: True (9.9)
## :   mw > 861.97:
## :     :...RNPxl_pl_MIC <= 0.05950294: False (49.7/13.6)
## :     :   RNPxl_pl_MIC > 0.05950294: True (15.2/1.5)
##
## SubTree [S2]
##
## RNPxl_Da_difference <= -0.0001694992: False (34.4/26.6)
## RNPxl_Da_difference > -0.0001694992: True (133.8/2.3)
##
## ----- Trial 2: -----
##
## Decision tree:

```

```

##
## mw > 2184.58: False (106.7/20.2)
## mw <= 2184.58:
##   ...RNPxl_total_MIC > 0.2095862: True (384.2/11.5)
##     RNPxl_total_MIC <= 0.2095862:
##       ...score <= 0.0594036: True (70.1/0.6)
##         score > 0.0594036:
##           ...RNPxl_pl_modds <= 0.8817645:
##             ...RNPxl_precursor_score > 0.0002412407: True (30.2)
##             : RNPxl_precursor_score <= 0.0002412407:
##             :   ...Peplength > 12: False (64.6/5)
##             :   Peplength <= 12:
##             :     ...RNPxl_Da_difference <= -0.0001225632: False (21.8/2.9)
##             :     RNPxl_Da_difference > -0.0001225632:
##             :       ...RNPxl_err <= 0.004044351: False (62.2/43.4)
##             :       RNPxl_err > 0.004044351: True (46/0.6)
##           RNPxl_pl_modds > 0.8817645:
##             ...Hydro > 0.28: False (8.2/0.6)
##             Hydro <= 0.28:
##               ...RNPxl_immonium_score <= 0.008185891:
##               : ...RNPxl_total_MIC <= 0.06030513: False (19.3/15.2)
##               : RNPxl_total_MIC > 0.06030513: True (199/2.3)
##               RNPxl_immonium_score > 0.008185891:
##               : ...RNPxl_pl_err <= 0.003974453: False (10.1)
##               RNPxl_pl_err > 0.003974453:
##               : ...aa_class_polar <= 5: True (95.1/8.2)
##               aa_class_polar > 5: False (4.7)
##
## ----- Trial 3: -----
##
## Decision tree:
##
## aa_class_hydrophobic <= 3:
##   ...RNPxl_RNA_MASS_z0 <= 391.0781: False (31.7/20)
##   : RNPxl_RNA_MASS_z0 > 391.0781:
##   :   ...Hydro <= -2.13: False (32.3/24.7)
##   :   Hydro > -2.13: True (395.6/9.2)
## aa_class_hydrophobic > 3:
##   ...RNPxl_total_MIC > 0.3001701:
##   : ...RNPxl_mass_error_p <= 1.941169: True (126.3)
##   : RNPxl_mass_error_p > 1.941169: False (7.7/3.2)
##   RNPxl_total_MIC <= 0.3001701:
##   : ...mw > 2405.7: False (50.9)
##   : mw <= 2405.7:
##   :   ...RNPxl_pl_modds <= 0.07640828: False (21.3)
##   :   RNPxl_pl_modds > 0.07640828:
##   :     ...RNPxl_immonium_score > 0.008149372:
##   :     : ...RNPxl_mass_error_p <= 1.912772: False (89.1/21)
##   :     : RNPxl_mass_error_p > 1.912772: True (81.1/12.7)
##   :     RNPxl_immonium_score <= 0.008149372:
##   :     : ...RNPxl_err <= 0.001154643: False (19.5/6.8)
##   :     RNPxl_err > 0.001154643:
##   :     : ...score <= 0.813925: True (207.1/8.2)
##   :     score > 0.813925:

```



```

##           :...RNPxl_pl_Morph <= 8.029901: False (18)
##           RNPxl_pl_Morph > 8.029901: True (49.3/5.9)
##
## ----- Trial 4: -----
##
## Decision tree:
##
## RNPxl_immonium_score > 0.02881235:
## :...RNPxl_pl_pc_MIC <= 0.01601188: False (45/1.1)
## :   RNPxl_pl_pc_MIC > 0.01601188: True (9.6)
## RNPxl_immonium_score <= 0.02881235:
## :...RNPxl_total_MIC > 0.2148431:
##   :...aa_class_polar <= 6: True (325/12.5)
##   :   aa_class_polar > 6: False (8/6.1)
##   RNPxl_total_MIC <= 0.2148431:
##   :...aa_class_hydrophobic <= 3:
##     :...precursor <= 524.2557: False (8.2/3.7)
##     :   precursor > 524.2557:
##       :   :...RNPxl_immonium_score > 0.009534214: False (64.4/50)
##       :     RNPxl_immonium_score <= 0.009534214:
##         :     :...RNPxl_score <= 0.8732423: True (253.2/6.8)
##         :       RNPxl_score > 0.8732423: False (3.8)
##     aa_class_hydrophobic > 3:
##       :...aa_class_hydrophobic > 12: False (22.1)
##       aa_class_hydrophobic <= 12:
##         :...RNPxl_pl_pc_MIC > 0.04382946: True (35.6)
##         RNPxl_pl_pc_MIC <= 0.04382946:
##           :...RNPxl_pl_Morph > 12.05118: True (50.6/3.8)
##           RNPxl_pl_Morph <= 12.05118:
##             :...aa_class_charged > 5: False (24.5)
##             aa_class_charged <= 5:
##               :...aa_class_polar > 5: False (18.3)
##               aa_class_polar <= 5:
##                 :...precursor <= 727.8291: False (28.6/9.3)
##                 precursor > 727.8291:
##                   :...aa_class_charged <= 1: True (25.1/0.7)
##                   aa_class_charged > 1:
##                     :...pI > 6.45: False (55.9/18.6)
##                     pI <= 6.45:
##                       :...RNPxl_pl_Morph <= 4.00529: False (8.3)
##                       RNPxl_pl_Morph > 4.00529: [S1]
##
## SubTree [S1]
##
## RNPxl_pl_im_MIC <= 0.01466738: True (120.9/8.8)
## RNPxl_pl_im_MIC > 0.01466738: False (6.1/0.4)
##
## ----- Trial 5: -----
##
## Decision tree:
##
## aa_class_hydrophobic > 5:
## :...RNPxl_pl_pc_MIC > 0.01601188: True (40.4/1.3)
## :   RNPxl_pl_pc_MIC <= 0.01601188:

```

```

## :   :...score <= 0.145311: True (31.8/3.1)
## :       score > 0.145311:
## :       :...RNPxl_pl_MIC > 0.1462661: True (30.6/1.8)
## :           RNPxl_pl_MIC <= 0.1462661:
## :           :...RNPxl_pl_modds <= 1.37235: False (144/33.1)
## :               RNPxl_pl_modds > 1.37235:
## :               :...RT <= 1838.091: True (32.7)
## :                   RT > 1838.091: False (65.9/30)
## aa_class_hydrophobic <= 5:
## :...RNPxl_total_MIC <= 0.04410198: False (54/33.5)
##     RNPxl_total_MIC > 0.04410198:
##     :...aIndex > 78.5:
##         :...RNPxl_total_MIC > 0.3017294: True (34.4)
##         :     RNPxl_total_MIC <= 0.3017294:
##         :         :...RNPxl_marker_ions_score <= 0.001819908: False (55.7/31.2)
##         :             RNPxl_marker_ions_score > 0.001819908: True (10.1)
##         aIndex <= 78.5:
##         :...Peplength <= 6: True (92)
##             Peplength > 6:
##             :...aa_class_hydrophobic <= 1: False (26.6/18)
##                 aa_class_hydrophobic > 1:
##                 :...RNPxl_Morph <= 4.025711: True (206.3/1.1)
##                     RNPxl_Morph > 4.025711:
##                     :...precursor_intensity <= 1252038: False (70.7/51)
##                         precursor_intensity > 1252038:
##                         :...RNPxl_err <= 0.001018666: False (12.5/6.4)
##                             RNPxl_err > 0.001018666: True (204.2/3.4)
##
## ----- Trial 6: -----
##
## Decision tree:
##
## Peplength > 10:
## :...RNPxl_total_MIC > 0.2480816: True (68.5/2.2)
## :     RNPxl_total_MIC <= 0.2480816:
## :     :...aIndex <= 20: True (38.7/3.1)
## :         aIndex > 20:
## :         :...RNPxl_pl_modds <= 0.8905877: False (110.2/15.7)
## :             RNPxl_pl_modds > 0.8905877:
## :             :...precursor_error_ppm > 4.391985: True (29.9)
## :                 precursor_error_ppm <= 4.391985:
## :                 :...RNPxl_pl_pc_MIC > 0.01709038: True (17)
## :                     RNPxl_pl_pc_MIC <= 0.01709038:
## :                     :...RNPxl_immonium_score > 0.008237569: False (29.4/2.9)
## :                         RNPxl_immonium_score <= 0.008237569:
## :                         :...RNPxl_mass_error_p <= 1.497245: False (8.7)
## :                             RNPxl_mass_error_p > 1.497245: True (81.4/5.7)
## Peplength <= 10:
## :...score <= 0.0752907: True (87.8/0.2)
##     score > 0.0752907:
##     :...aIndex > 86.25:
##         :...RNPxl_pl_Morph <= 2.227052: False (9.9)
##             RNPxl_pl_Morph > 2.227052:
##             :     :...RNPxl_Morph <= 4.530347: True (102.8/11)

```

```

##      :      RNPxl_Morph > 4.530347: False (9.7/2.1)
##      aIndex <= 86.25:
##      :...RNPxl_RNA_MASS_z0 > 758.1211: False (26.4/18.5)
##      RNPxl_RNA_MASS_z0 <= 758.1211:
##      :...aa_class_charged <= 2: True (335.1/9.3)
##      aa_class_charged > 2:
##      :...precursor_error_ppm <= 0.6376348: True (82.6)
##      precursor_error_ppm > 0.6376348:
##      :...RNPxl_precursor_score > 0.0007802289: False (3.9/0.4)
##      RNPxl_precursor_score <= 0.0007802289:
##      :...RNPxl_pl_err > 0.01726561: False (3.1)
##      RNPxl_pl_err <= 0.01726561:
##      :...RNPxl_marker_ions_score > 0.001386256: False (2.6)
##      RNPxl_marker_ions_score <= 0.001386256:
##      :...RNPxl_pl_modds > 3.219504: False (3.4)
##      RNPxl_pl_modds <= 3.219504:
##      :...mw <= 1219.46: True (144.4/6.9)
##      mw > 1219.46: False (2.5)
##
## ----- Trial 7: -----
##
## Decision tree:
##
## aa_class_hydrophobic > 6:
## :...RNPxl_pl_pc_MIC > 0.01601188: True (30.1/1.7)
## :   RNPxl_pl_pc_MIC <= 0.01601188:
## :     :...OMS_precursor_mz_error_ppm <= 3.760296: False (188.2/80.2)
## :     :   OMS_precursor_mz_error_ppm > 3.760296: True (27.7/1.6)
## aa_class_hydrophobic <= 6:
## :...RNPxl_pl_MIC > 0.1492823: True (76.8)
##   RNPxl_pl_MIC <= 0.1492823:
##     :...RNPxl_RNA_MASS_z0 <= 335.123:
##     :...score > 0.929751: True (22.6)
##     :   score <= 0.929751:
##     :     :...precursor_error_ppm <= 0.5690612: True (17.5/0.2)
##     :     precursor_error_ppm > 0.5690612:
##     :       :...RNPxl_Da_difference > 0.00196853: True (14.2/0.3)
##     :       RNPxl_Da_difference <= 0.00196853:
##     :         :...RNPxl_peptide_mass_z0 <= 827.5593: True (12.2/0.2)
##     :         RNPxl_peptide_mass_z0 > 827.5593: False (37.5/0.3)
##     RNPxl_RNA_MASS_z0 > 335.123:
##     :...RNPxl_total_MIC <= 0.037681: False (41.6/21.6)
##     RNPxl_total_MIC > 0.037681:
##     :...aa_class_hydrophobic > 5: False (27.9/15.8)
##     aa_class_hydrophobic <= 5:
##     :...Peplength > 12: False (25.4/16.1)
##     Peplength <= 12:
##     :...OMS_precursor_mz_error_ppm > 2.227397: True (115/0.2)
##     OMS_precursor_mz_error_ppm <= 2.227397:
##     :...RNPxl_Da_difference <= -0.0007595894:
##     :...precursor_error_ppm <= 3.304309: False (55.9/33.1)
##     :   precursor_error_ppm > 3.304309: True (11.9)
##     RNPxl_Da_difference > -0.0007595894:
##     :...RNPxl_modds > 1.71208: False (17.2/12.3)

```

```

##                               RNPxl_modds <= 1.71208:
##                               :...RT <= 4286.273: True (391.1/13.1)
##                               RT > 4286.273: False (5.1/1.5)
##
## ----- Trial 8: -----
##
## Decision tree:
##
## RNPxl_MIC > 0.3118303: True (111.2)
## RNPxl_MIC <= 0.3118303:
## :...mw > 2184.58: False (73.9/29.7)
##   mw <= 2184.58:
##     :...RT <= 1262.902: True (274.2/14.8)
##     RT > 1262.902:
##       :...RNPxl_pl_modds <= 0.1330409: False (22)
##       RNPxl_pl_modds > 0.1330409:
##         :...RNPxl_RNA_MASS_z0 <= 351.1179:
##         :...RNPxl_pl_MIC > 0.08438432: True (39.5/1.2)
##         :   RNPxl_pl_MIC <= 0.08438432:
##         :     :...RNPxl_pl_modds <= 3.201042: False (68.3/24.2)
##         :     RNPxl_pl_modds > 3.201042: True (10.7)
##         RNPxl_RNA_MASS_z0 > 351.1179:
##         :...pI > 9.3:
##         :...precursor_intensity <= 953246.9: True (66.8/4)
##         :   precursor_intensity > 953246.9:
##         :     :...RNPxl_MIC <= 0.2795199: False (84.3/40.6)
##         :     RNPxl_MIC > 0.2795199: True (18.6)
##         pI <= 9.3:
##         :...Hydro <= -1.89: False (7.5)
##         Hydro > -1.89:
##         :...aa_class_polar <= 1: False (29.3/21.2)
##         aa_class_polar > 1:
##         :...Peplength <= 9: True (205.9/0.1)
##         Peplength > 9:
##         :...precursor_error_ppm > 2.518739: False (5.9/0.4)
##         precursor_error_ppm <= 2.518739:
##         :...pI <= 3.88: False (3.1)
##         pI > 3.88: True (170.5/12)
##
## ----- Trial 9: -----
##
## Decision tree:
##
## RNPxl_total_MIC <= 0.2148431:
## :...aa_class_hydrophobic > 3:
## :   :...Peplength > 20: False (27.5)
## :   :   Peplength <= 20:
## :   :     :...RNPxl_score <= -0.640421: False (145.5/61.1)
## :   :     RNPxl_score > -0.640421:
## :   :       :...RNPxl_pl_pc_MIC > 0.0002567632: True (93.8/4.2)
## :   :       RNPxl_pl_pc_MIC <= 0.0002567632:
## :   :         :...RNPxl_immonium_score <= 0.008149372: True (112.5/9.9)
## :   :         RNPxl_immonium_score > 0.008149372: False (51.5/12.5)
## :   aa_class_hydrophobic <= 3:

```

```

## :   ...aa_class_charged <= 2: True (179.1/6)
## :       aa_class_charged > 2:
## :       ...RNPxl_pl_modds > 2.211792: True (27.6)
## :           RNPxl_pl_modds <= 2.211792:
## :           ...precursor_error_ppm <= 0.51688: True (48.3/3.4)
## :               precursor_error_ppm > 0.51688:
## :               ...RNPxl_peptide_mass_z0 <= 678.4429: True (7.7)
## :                   RNPxl_peptide_mass_z0 > 678.4429:
## :                   ...RNPxl_Morph <= 2.469005: True (11.3/0.6)
## :                       RNPxl_Morph > 2.469005: False (37.1/8.2)
## RNPxl_total_MIC > 0.2148431:
## ...aa_class_polar > 6: False (13.2/9.6)
##     aa_class_polar <= 6:
##     ...RNPxl_total_MIC > 0.3725607: True (150.8)
##         RNPxl_total_MIC <= 0.3725607:
##         ...RNPxl_score <= -0.9358985: True (99.4)
##             RNPxl_score > -0.9358985:
##             ...Hydro <= -2.04: False (2.8)
##                 Hydro > -2.04:
##                 ...RT > 2551.2: False (20/12.4)
##                     RT <= 2551.2:
##                     ...RNPxl_pl_pc_MIC <= 0.1841059: True (119.7/2.3)
##                         RNPxl_pl_pc_MIC > 0.1841059: False (3.1)
##
##
## Evaluation on training data (734 cases):
##
## Trial          Decision Tree
## -----
##      Size      Errors  Cost
##
##      0      19  48( 6.5%)  0.08
##      1      22  92(12.5%)  0.18
##      2      14  62( 8.4%)  0.21
##      3      13 107(14.6%)  0.39
##      4      19  74(10.1%)  0.26
##      5      16  87(11.9%)  0.20
##      6      21  63( 8.6%)  0.27
##      7      18  90(12.3%)  0.22
##      8      16  86(11.7%)  0.32
##      9      18  58( 7.9%)  0.21
## boost          2( 0.3%)  0.00  <<
##
##
##      (a)  (b)  <-classified as
##      ----  ----
##      363      (a): class False
##      2  369  (b): class True
##
##
## Attribute usage:
##
## 100.00% RNPxl_MIC
## 100.00% RNPxl_immonium_score

```

```

## 100.00% RNPxl_pl_modds
## 100.00% RNPxl_total_MIC
## 100.00% Peplength
## 100.00% aa_class_hydrophobic
## 100.00% mw
## 98.23% score
## 97.14% RNPxl_pl_MIC
## 96.73% aIndex
## 76.29% RNPxl_err
## 76.16% OMS_precursor_mz_error_ppm
## 74.52% RT
## 72.21% RNPxl_RNA_MASS_z0
## 70.98% precursor
## 70.30% aa_class_polar
## 65.94% aa_class_charged
## 64.31% RNPxl_pl_pc_MIC
## 61.72% Hydro
## 61.44% RNPxl_score
## 59.54% pI
## 43.32% RNPxl_Da_difference
## 41.14% RNPxl_precursor_score
## 40.87% RNPxl_Morph
## 40.19% precursor_error_ppm
## 38.69% RNPxl_pl_Morph
## 36.51% RNPxl_partial_loss_score
## 35.56% RNPxl_mass_error_p
## 28.47% RNPxl_modds
## 27.52% precursor_intensity
## 24.39% RNPxl_marker_ions_score
## 19.21% RNPxl_pl_err
## 8.31% RNPxl_peptide_mass_z0
## 7.08% RNPxl_pl_im_MIC
##
##
## Time: 0.2 secs

```

```

C5_DEB_RNA_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, m1_DEB_RNA_test_C5)
CrossTable(m1_DEB_RNA_test_C5$Curated, C5_DEB_RNA_predict_boost_pen, prop.chisq = FALSE,
           prop.c = FALSE, prop.r = FALSE, dnn = c("actual TRUE", "predicted TRUE"))

```

```

##
##
## Cell Contents
## |-----|
## |                N |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table: 184
##
##
##          | predicted TRUE
## actual TRUE | False | True | Row Total |
## -----|-----|-----|-----|

```

```

##      False |      86 |      10 |      96 |
##          |    0.467 |    0.054 |          |
## -----|-----|-----|-----|
##      True  |       5 |      83 |      88 |
##          |    0.027 |    0.451 |          |
## -----|-----|-----|-----|
## Column Total |      91 |      93 |     184 |
## -----|-----|-----|-----|
##
##

```

Classifiers

DEB RNA

One Rule (1R)

RIPPER

```

library(rJava)
library(RWeka)

```

```

RIPPER_model_DEB_RNA <- JRip(Curated ~ ., data = ml_DEB_RNA_train_C5)
RIPPER_model_DEB_RNA

```

```

## JRIP rules:

```

```

## =====

```

```

##

```

```

## (mw >= 1164.34) and (RNPxl_total_MIC <= 0.148872) and (RNPxl_pl_modds <= 0.848058) and (score >= 0.4
## (RNPxl_total_MIC <= 0.236507) and (Hydro >= -0.67) and (RNPxl_immonium_score >= 0.008186) => Curated
## (RNPxl_peptide_mass_z0 >= 1130.473637) and (RNPxl_total_MIC <= 0.193156) and (RNPxl_RNA_MASS_z0 <= 6
## (RNPxl_total_MIC <= 0.213968) and (precursor_intensity <= 1138125.875) and (RNPxl_pl_modds <= 0.6873
## (RNPxl_total_MIC <= 0.28572) and (aa_class_hydrophobic >= 4) and (OMS_precursor_mz_error_ppm >= -0.1
## (aIndex >= 70) and (RNPxl_total_MIC <= 0.213968) and (RNPxl_pl_MIC <= 0.01229) => Curated=False (12.
## => Curated=True (364.0/21.0)

```

```

##

```

```

## Number of Rules : 7

```

```

summary(RIPPER_model_DEB_RNA)

```

```

##

```

```

## === Summary ===

```

```

##

```

```

## Correctly Classified Instances      685          93.3243 %
## Incorrectly Classified Instances    49           6.6757 %
## Kappa statistic                     0.8665
## Mean absolute error                 0.1161
## Root mean squared error            0.2409
## Relative absolute error             23.2226 %
## Root relative squared error        48.1898 %
## Total Number of Instances          734

```

```

##

```

```

## === Confusion Matrix ===

```

```

##

```

```

##   a   b   <-- classified as

```

```
## 342 21 | a = False
## 28 343 | b = True

RIPPER_DEB_RNA_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_test_C5)
CrossTable(ml_DEB_RNA_test_C5$Curated, RIPPER_DEB_RNA_predict, prop.chisq = FALSE,
  prop.c = FALSE, prop.r = FALSE, dnn = c("actual TRUE", "predicted TRUE"))

##
##
## Cell Contents
## |-----|
## | N |
## | N / Table Total |
## |-----|
##
##
## Total Observations in Table: 184
##
##
##      | predicted TRUE
## actual TRUE | False | True | Row Total |
## -----|-----|-----|-----|
##      False | 81 | 15 | 96 |
##      | 0.440 | 0.082 | |
## -----|-----|-----|-----|
##      True | 15 | 73 | 88 |
##      | 0.082 | 0.397 | |
## -----|-----|-----|-----|
## Column Total | 96 | 88 | 184 |
## -----|-----|-----|-----|
##
##
```

DEB RNA FAIMS runs S100

Now, let us test KNN, C5, and RIPPER models on unknown data: Ecoli DEB FAIMS runs

Data setup DEB RNA CV 35_45

```
DEB_RNA_FAIMS_35_45_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_35_45_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),
##   `RNPxl:RNA` = col_character(),
##   `RNPxl:best_localization` = col_character(),
##   `RNPxl:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )
```



```
## )
```

```
## See spec(...) for full column specifications.
```

```
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_35_45_KNN))
DEB_RNA_FAIMS_35_45_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_35_45_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_35_45_KNN$Sample <- "S100"
DEB_RNA_FAIMS_35_45_KNN <- DEB_RNA_FAIMS_35_45_KNN[DEB_RNA_FAIMS_35_45_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$sequence))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$prepPep))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$prepPep))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_35_45_KNN$prepPep)
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_35_45_KNN$prepPep)
DEB_RNA_FAIMS_35_45_KNN$Pepseq <- DEB_RNA_FAIMS_35_45_KNN$prepPep
DEB_RNA_FAIMS_35_45_KNN$Peplength <- nchar(DEB_RNA_FAIMS_35_45_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_35_45_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$pI <- round(pI(DEB_RNA_FAIMS_35_45_KNN$Pepseq), 2)
DEB_RNA_FAIMS_35_45_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_35_45_KNN$Pepseq), 2)
DEB_RNA_FAIMS_35_45_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_35_45_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_35_45_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_35_45_KNN$mw <- round(mw(DEB_RNA_FAIMS_35_45_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_35_45_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_35_45_KNN)[names(DEB_RNA_FAIMS_35_45_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_35_45_KNN)[names(DEB_RNA_FAIMS_35_45_KNN) == "precursor_error_(ppm)"] <- "precursor_error"
DEB_RNA_FAIMS_35_45_KNN <- DEB_RNA_FAIMS_35_45_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_35_45_KNN_model <- DEB_RNA_FAIMS_35_45_KNN %>% select(colnames(ml_DEB_RNA_train_knn))
```

Normalizing

```
DEB_RNA_FAIMS_35_45_KNN_model_norm <- DEB_RNA_FAIMS_35_45_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_35_45_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_35_45_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_35_45_KNN_model_norm[is.na(DEB_RNA_FAIMS_35_45_KNN_model_norm)] <- 0
```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_35_45_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_35_45_KNN_model_1,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_35_45_KNN_model$Prediction <- DEB_RNA_FAIMS_35_45_knn_pred
table(DEB_RNA_FAIMS_35_45_KNN_model$Prediction)

##
## False True
## 13505 2524
DEB_RNA_FAIMS_35_45_KNN$Prediction_KNN <- DEB_RNA_FAIMS_35_45_knn_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

ml_DEB_RNA_FAIMS_35_45_test_C5<-DEB_RNA_FAIMS_35_45_KNN %>%
  select(colnames(ml_DEB_RNA_train))

C5_DEB_RNA_FAIMS_35_45_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_test_C5)

ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction<-C5_DEB_RNA_FAIMS_35_45_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction)

##
## False True
## 11339 4690
DEB_RNA_FAIMS_35_45_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_predict_boost_pen

```

RIPPER

```

RIPPER_DEB_RNA_FAIMS_35_45_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_test_C5)
ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_predict
table(ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER)

##
## False True
## 10035 5994
DEB_RNA_FAIMS_35_45_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_predict

```

Data setup DEB RNA CV 40_50

```

DEB_RNA_FAIMS_40_50_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_40_50_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),

```

```

## `RNPxl:RNA` = col_character(),
## `RNPxl:best_localization` = col_character(),
## `RNPxl:localization_scores` = col_character(),
## protein_references = col_character(),
## target_decoy = col_character(),
## `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_40_50_KNN))
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_40_50_KNN))
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_40_50_KNN))
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_40_50_KNN))
DEB_RNA_FAIMS_40_50_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_40_50_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_40_50_KNN$Sample <- "S100"
DEB_RNA_FAIMS_40_50_KNN <- DEB_RNA_FAIMS_40_50_KNN[DEB_RNA_FAIMS_40_50_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$sequence))
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$prepPep))
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$prepPep))
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_40_50_KNN$prepPep)
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_40_50_KNN$prepPep)
DEB_RNA_FAIMS_40_50_KNN$Pepseq <- DEB_RNA_FAIMS_40_50_KNN$prepPep
DEB_RNA_FAIMS_40_50_KNN$Peplength <- nchar(DEB_RNA_FAIMS_40_50_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_40_50_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_40_50_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_40_50_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_40_50_KNN$pI <- round(pI(DEB_RNA_FAIMS_40_50_KNN$Pepseq), 2)
DEB_RNA_FAIMS_40_50_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_40_50_KNN$Pepseq), 2)
DEB_RNA_FAIMS_40_50_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_40_50_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_40_50_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_40_50_KNN$mw <- round(mw(DEB_RNA_FAIMS_40_50_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_40_50_KNN$cv <- c("40_50")

names(DEB_RNA_FAIMS_40_50_KNN)[names(DEB_RNA_FAIMS_40_50_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_40_50_KNN)[names(DEB_RNA_FAIMS_40_50_KNN) == "precursor_error_(ppm)"] <- "precursor"
DEB_RNA_FAIMS_40_50_KNN <- DEB_RNA_FAIMS_40_50_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_40_50_KNN_model <- DEB_RNA_FAIMS_40_50_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_40_50_KNN_model_norm <- DEB_RNA_FAIMS_40_50_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_40_50_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_40_50_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_40_50_KNN_model_norm[is.na(DEB_RNA_FAIMS_40_50_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```
DEB_RNA_FAIMS_40_50_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_40_50_KNN_model_1,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_40_50_KNN_model$Prediction <- DEB_RNA_FAIMS_40_50_knn_pred
table(DEB_RNA_FAIMS_40_50_KNN_model$Prediction)
```

```
##
## False True
## 12115 2101
```

```
DEB_RNA_FAIMS_40_50_KNN$Prediction_KNN <- DEB_RNA_FAIMS_40_50_knn_pred
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)
```

```
ml_DEB_RNA_FAIMS_40_50_test_C5<-DEB_RNA_FAIMS_40_50_KNN %>%
  select(colnames(ml_DEB_RNA_train))
```

```
C5_DEB_RNA_FAIMS_40_50_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_40_50_test_C5)
```

```
ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction<-C5_DEB_RNA_FAIMS_40_50_predict_boost_pen
table(ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction)
```

```
##
## False True
## 10326 3890
```

```
DEB_RNA_FAIMS_40_50_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_40_50_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_40_50_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_40_50_test_C5)
ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_40_50_predict
table(ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 8625 5591
```

```
DEB_RNA_FAIMS_40_50_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_40_50_predict
```

Data setup DEB RNA CV 55_65

```
DEB_RNA_FAIMS_55_65_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_55_65_table.csv",
  delim = "\t")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
```

```

## sequence = col_character(),
## accessions = col_character(),
## `RNPxl:NT` = col_character(),
## `RNPxl:RNA` = col_character(),
## `RNPxl:best_localization` = col_character(),
## `RNPxl:localization_scores` = col_character(),
## protein_references = col_character(),
## target_decoy = col_character(),
## `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_55_65_KNN))
DEB_RNA_FAIMS_55_65_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_55_65_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_55_65_KNN$Sample <- "S100"
DEB_RNA_FAIMS_55_65_KNN <- DEB_RNA_FAIMS_55_65_KNN[DEB_RNA_FAIMS_55_65_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$sequence))
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$prepPep))
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$prepPep))
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_55_65_KNN$prepPep)
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_55_65_KNN$prepPep)
DEB_RNA_FAIMS_55_65_KNN$Pepseq <- DEB_RNA_FAIMS_55_65_KNN$prepPep
DEB_RNA_FAIMS_55_65_KNN$Peplength <- nchar(DEB_RNA_FAIMS_55_65_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_55_65_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$pI <- round(pI(DEB_RNA_FAIMS_55_65_KNN$Pepseq), 2)
DEB_RNA_FAIMS_55_65_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_55_65_KNN$Pepseq), 2)
DEB_RNA_FAIMS_55_65_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_55_65_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_55_65_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_55_65_KNN$mw <- round(mw(DEB_RNA_FAIMS_55_65_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_55_65_KNN$cv <- c("55_65")

names(DEB_RNA_FAIMS_55_65_KNN)[names(DEB_RNA_FAIMS_55_65_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_55_65_KNN)[names(DEB_RNA_FAIMS_55_65_KNN) == "precursor_error_(ppm)"] <- "precursor"
DEB_RNA_FAIMS_55_65_KNN <- DEB_RNA_FAIMS_55_65_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_55_65_KNN_model <- DEB_RNA_FAIMS_55_65_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_55_65_KNN_model_norm <- DEB_RNA_FAIMS_55_65_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_55_65_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_55_65_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_55_65_KNN_model_norm[is.na(DEB_RNA_FAIMS_55_65_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_55_65_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_55_65_KNN_model_norm,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_55_65_KNN_model$Prediction <- DEB_RNA_FAIMS_55_65_knn_pred
table(DEB_RNA_FAIMS_55_65_KNN_model$Prediction)

```

```

##
## False True
## 18600 3222

```

```

DEB_RNA_FAIMS_55_65_KNN$Prediction_KNN <- DEB_RNA_FAIMS_55_65_knn_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

```

```

ml_DEB_RNA_FAIMS_55_65_test_C5<-DEB_RNA_FAIMS_55_65_KNN %>%
  select(colnames(ml_DEB_RNA_train))

```

```

C5_DEB_RNA_FAIMS_55_65_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_55_65_test_C5)

```

```

ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction<-C5_DEB_RNA_FAIMS_55_65_predict_boost_pen
table(ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction)

```

```

##
## False True
## 14960 6862

```

```

DEB_RNA_FAIMS_55_65_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_55_65_predict_boost_pen

```

RIPPER

```

RIPPER_DEB_RNA_FAIMS_55_65_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_55_65_test_C5)
ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_55_65_predict
table(ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER)

```

```

##
## False True
## 12322 9500

```

```

DEB_RNA_FAIMS_55_65_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_55_65_predict

```

Data setup DEB RNA CV 60_70

```

DEB_RNA_FAIMS_60_70_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_60_70_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),
##   `RNPxl:RNA` = col_character(),
##   `RNPxl:best_localization` = col_character(),
##   `RNPxl:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.

names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\\\|", "_", names(DEB_RNA_FAIMS_60_70_KNN))
DEB_RNA_FAIMS_60_70_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_60_70_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_60_70_KNN$Sample <- "S100"
DEB_RNA_FAIMS_60_70_KNN <- DEB_RNA_FAIMS_60_70_KNN[DEB_RNA_FAIMS_60_70_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$sequence))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$prepPep))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$prepPep))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_60_70_KNN$prepPep)
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_60_70_KNN$prepPep)
DEB_RNA_FAIMS_60_70_KNN$Pepseq <- DEB_RNA_FAIMS_60_70_KNN$prepPep
DEB_RNA_FAIMS_60_70_KNN$Peplength <- nchar(DEB_RNA_FAIMS_60_70_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_60_70_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$pI <- round(pI(DEB_RNA_FAIMS_60_70_KNN$Pepseq), 2)
DEB_RNA_FAIMS_60_70_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_60_70_KNN$Pepseq), 2)
DEB_RNA_FAIMS_60_70_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_60_70_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_60_70_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_60_70_KNN$mw <- round(mw(DEB_RNA_FAIMS_60_70_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_60_70_KNN$cv <- c("60_70")

names(DEB_RNA_FAIMS_60_70_KNN)[names(DEB_RNA_FAIMS_60_70_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_60_70_KNN)[names(DEB_RNA_FAIMS_60_70_KNN) == "precursor_error_(ppm)"] <- "precursor"
DEB_RNA_FAIMS_60_70_KNN <- DEB_RNA_FAIMS_60_70_KNN %>% filter(RNPxl_isPhospho ==

```

```
"0")
```

```
DEB_RNA_FAIMS_60_70_KNN_model <- DEB_RNA_FAIMS_60_70_KNN %>% select(colnames(ml_DEB_RNA_train_knn))
```

Normalizing

```
DEB_RNA_FAIMS_60_70_KNN_model_norm <- DEB_RNA_FAIMS_60_70_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_60_70_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_60_70_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_60_70_KNN_model_norm[is.na(DEB_RNA_FAIMS_60_70_KNN_model_norm)] <- 0
```

Classifying DEB RNA FAIMS data

KNN5

```
DEB_RNA_FAIMS_60_70_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_60_70_KNN_model_norm,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_60_70_KNN_model$Prediction <- DEB_RNA_FAIMS_60_70_knn_pred
table(DEB_RNA_FAIMS_60_70_KNN_model$Prediction)
```

```
##
## False True
## 33120 4143
```

```
DEB_RNA_FAIMS_60_70_KNN$Prediction_KNN <- DEB_RNA_FAIMS_60_70_knn_pred
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions <- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)
```

```
ml_DEB_RNA_FAIMS_60_70_test_C5 <- DEB_RNA_FAIMS_60_70_KNN %>%
  select(colnames(ml_DEB_RNA_train))
```

```
C5_DEB_RNA_FAIMS_60_70_predict_boost_pen <- predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_60_70_test_C5)
```

```
ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction <- C5_DEB_RNA_FAIMS_60_70_predict_boost_pen
table(ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction)
```

```
##
## False True
## 22354 14909
```

```
DEB_RNA_FAIMS_60_70_KNN$Prediction_C5 <- C5_DEB_RNA_FAIMS_60_70_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_60_70_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_60_70_test_C5)
ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_60_70_predict
table(ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER)
```

```
##
```



```
## False True
## 18746 18517
DEB_RNA_FAIMS_60_70_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_60_70_predict
```

DEB RNA FAIMS runs S100

Now, let us test KNN, C5, and RIPPER models on unknown data: Ecoli DEB FAIMS runs

Data setup DEB RNA CV 35_45

```
DEB_RNA_FAIMS_35_45_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_35_45_table.csv",
  delim = "\t")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),
##   `RNPxl:RNA` = col_character(),
##   `RNPxl:best_localization` = col_character(),
##   `RNPxl:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_35_45_KNN))
names(DEB_RNA_FAIMS_35_45_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_35_45_KNN))
DEB_RNA_FAIMS_35_45_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_35_45_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_35_45_KNN$Sample <- "30"
DEB_RNA_FAIMS_35_45_KNN <- DEB_RNA_FAIMS_35_45_KNN[DEB_RNA_FAIMS_35_45_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$sequence))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$prepPep))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_35_45_KNN$prepPep))
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_35_45_KNN$prepPep)
DEB_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_35_45_KNN$prepPep)
DEB_RNA_FAIMS_35_45_KNN$Pepseq <- DEB_RNA_FAIMS_35_45_KNN$prepPep
DEB_RNA_FAIMS_35_45_KNN$Peplength <- nchar(DEB_RNA_FAIMS_35_45_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_35_45_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_35_45_KNN$pI <- round(pI(DEB_RNA_FAIMS_35_45_KNN$Pepseq), 2)
DEB_RNA_FAIMS_35_45_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_35_45_KNN$Pepseq), 2)
DEB_RNA_FAIMS_35_45_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_35_45_KNN$RNPxl_best_localization)
```

```

DEB_RNA_FAIMS_35_45_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_35_45_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_35_45_KNN$mw <- round(mw(DEB_RNA_FAIMS_35_45_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_35_45_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_35_45_KNN)[names(DEB_RNA_FAIMS_35_45_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_35_45_KNN)[names(DEB_RNA_FAIMS_35_45_KNN) == "precursor_error_(ppm)"] <- "precursor_error_(ppm)"
DEB_RNA_FAIMS_35_45_KNN <- DEB_RNA_FAIMS_35_45_KNN %>% filter(RNPs1_isPhospho ==
  "0")

DEB_RNA_FAIMS_35_45_KNN_model <- DEB_RNA_FAIMS_35_45_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_35_45_KNN_model_norm <- DEB_RNA_FAIMS_35_45_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_35_45_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_35_45_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_35_45_KNN_model_norm[is.na(DEB_RNA_FAIMS_35_45_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_35_45_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_35_45_KNN_model_norm,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_35_45_KNN_model$Prediction <- DEB_RNA_FAIMS_35_45_knn_pred
table(DEB_RNA_FAIMS_35_45_KNN_model$Prediction)

```

```

##
## False True
## 13506 2523

```

```

DEB_RNA_FAIMS_35_45_KNN$Prediction_KNN <- DEB_RNA_FAIMS_35_45_knn_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions <- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

```

```

ml_DEB_RNA_FAIMS_35_45_test_C5 <- DEB_RNA_FAIMS_35_45_KNN %>%
  select(colnames(ml_DEB_RNA_train))

```

```

C5_DEB_RNA_FAIMS_35_45_predict_boost_pen <- predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_test_C5)

```

```

ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction <- C5_DEB_RNA_FAIMS_35_45_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction)

```

```

##
## False True
## 11339 4690

```

```
DEB_RNA_FAIMS_35_45_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_35_45_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_test_C5)
ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_predict
table(ml_DEB_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 10035 5994
```

```
DEB_RNA_FAIMS_35_45_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_predict
```

Data setup DEB RNA CV 40_50

```
DEB_RNA_FAIMS_40_50_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_40_50_table.csv",
  delim = "\t")
```

```
## Parsed with column specification:
```

```
## cols(
```

```
##   .default = col_double(),
```

```
##   sequence = col_character(),
```

```
##   accessions = col_character(),
```

```
##   `RNPxl:NT` = col_character(),
```

```
##   `RNPxl:RNA` = col_character(),
```

```
##   `RNPxl:best_localization` = col_character(),
```

```
##   `RNPxl:localization_scores` = col_character(),
```

```
##   protein_references = col_character(),
```

```
##   target_decoy = col_character(),
```

```
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
```

```
## )
```

```
## See spec(...) for full column specifications.
```

```
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_40_50_KNN))
```

```
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub(":", "_", names(DEB_RNA_FAIMS_40_50_KNN))
```

```
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_40_50_KNN))
```

```
names(DEB_RNA_FAIMS_40_50_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_40_50_KNN))
```

```
DEB_RNA_FAIMS_40_50_KNN$XLinker <- "DEB"
```

```
DEB_RNA_FAIMS_40_50_KNN$Nucleotides <- "RNA"
```

```
DEB_RNA_FAIMS_40_50_KNN$Sample <- "S100"
```

```
DEB_RNA_FAIMS_40_50_KNN <- DEB_RNA_FAIMS_40_50_KNN[DEB_RNA_FAIMS_40_50_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
```

```
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$sequence))
```

```
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$prepPep))
```

```
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_40_50_KNN$prepPep))
```

```
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_40_50_KNN$prepPep)
```

```
DEB_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_40_50_KNN$prepPep)
```

```
DEB_RNA_FAIMS_40_50_KNN$Pepseq <- DEB_RNA_FAIMS_40_50_KNN$prepPep
```

```
DEB_RNA_FAIMS_40_50_KNN$Peplength <- nchar(DEB_RNA_FAIMS_40_50_KNN$Pepseq, type = "chars")
```

```
DEB_RNA_FAIMS_40_50_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
```

```
DEB_RNA_FAIMS_40_50_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
```

```

paste("Q|N|H|S|T|Y|C|W", collapse = "|")
DEB_RNA_FAIMS_40_50_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_40_50_KNN$pI <- round(pI(DEB_RNA_FAIMS_40_50_KNN$Pepseq), 2)
DEB_RNA_FAIMS_40_50_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_40_50_KNN$Pepseq), 2)
DEB_RNA_FAIMS_40_50_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_40_50_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_40_50_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_40_50_KNN$Pepseq,
scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_40_50_KNN$mw <- round(mw(DEB_RNA_FAIMS_40_50_KNN$Pepseq, monoisotopic = F),
2)
DEB_RNA_FAIMS_40_50_KNN$cv <- c("40_50")

names(DEB_RNA_FAIMS_40_50_KNN)[names(DEB_RNA_FAIMS_40_50_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_40_50_KNN)[names(DEB_RNA_FAIMS_40_50_KNN) == "precursor_error_(ppm)"] <- "precursor_error_ppm"
DEB_RNA_FAIMS_40_50_KNN <- DEB_RNA_FAIMS_40_50_KNN %>% filter(RNPxl_isPhospho ==
"0")

DEB_RNA_FAIMS_40_50_KNN_model <- DEB_RNA_FAIMS_40_50_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_40_50_KNN_model_norm <- DEB_RNA_FAIMS_40_50_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_40_50_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_40_50_KNN_model_norm,
normalize))
DEB_RNA_FAIMS_40_50_KNN_model_norm[is.na(DEB_RNA_FAIMS_40_50_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_40_50_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_40_50_KNN_model_norm,
cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_40_50_KNN_model$Prediction <- DEB_RNA_FAIMS_40_50_knn_pred
table(DEB_RNA_FAIMS_40_50_KNN_model$Prediction)

##
## False True
## 12114 2102

DEB_RNA_FAIMS_40_50_KNN$Prediction_KNN <- DEB_RNA_FAIMS_40_50_knn_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions <- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

ml_DEB_RNA_FAIMS_40_50_test_C5 <- DEB_RNA_FAIMS_40_50_KNN %>%
select(colnames(ml_DEB_RNA_train))

C5_DEB_RNA_FAIMS_40_50_predict_boost_pen <- predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_40_50_test_C5)

```

```
ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction<-C5_DEB_RNA_FAIMS_40_50_predict_boost_pen
table(ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction)
```

```
##
## False True
## 10326 3890
```

```
DEB_RNA_FAIMS_40_50_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_40_50_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_40_50_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_40_50_test_C5)
ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_40_50_predict
table(ml_DEB_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 8625 5591
```

```
DEB_RNA_FAIMS_40_50_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_40_50_predict
```

Data setup DEB RNA CV 55_65

```
DEB_RNA_FAIMS_55_65_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_55_65_table.csv",
  delim = "\t")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),
##   `RNPxl:RNA` = col_character(),
##   `RNPxl:best_localization` = col_character(),
##   `RNPxl:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_55_65_KNN))
names(DEB_RNA_FAIMS_55_65_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_55_65_KNN))
```

```
DEB_RNA_FAIMS_55_65_KNN$XLinker <- "DEB"
```

```
DEB_RNA_FAIMS_55_65_KNN$Nucleotides <- "RNA"
```

```
DEB_RNA_FAIMS_55_65_KNN$Sample <- "S100"
```

```
DEB_RNA_FAIMS_55_65_KNN <- DEB_RNA_FAIMS_55_65_KNN[DEB_RNA_FAIMS_55_65_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
```

```
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$sequence))
```

```
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$prepPep))
```

```
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_55_65_KNN$prepPep))
```

```
DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_55_65_KNN$prepPep)
```

```

DEB_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\\\.*", "", DEB_RNA_FAIMS_55_65_KNN$prepPep)
DEB_RNA_FAIMS_55_65_KNN$Pepseq <- DEB_RNA_FAIMS_55_65_KNN$prepPep
DEB_RNA_FAIMS_55_65_KNN$Peplength <- nchar(DEB_RNA_FAIMS_55_65_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_55_65_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_55_65_KNN$pI <- round(pI(DEB_RNA_FAIMS_55_65_KNN$Pepseq), 2)
DEB_RNA_FAIMS_55_65_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_55_65_KNN$Pepseq), 2)
DEB_RNA_FAIMS_55_65_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_55_65_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_55_65_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_55_65_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_55_65_KNN$mw <- round(mw(DEB_RNA_FAIMS_55_65_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_55_65_KNN$cv <- c("55_65")

names(DEB_RNA_FAIMS_55_65_KNN)[names(DEB_RNA_FAIMS_55_65_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_55_65_KNN)[names(DEB_RNA_FAIMS_55_65_KNN) == "precursor_error_(ppm_)"] <- "precursor"
DEB_RNA_FAIMS_55_65_KNN <- DEB_RNA_FAIMS_55_65_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_55_65_KNN_model <- DEB_RNA_FAIMS_55_65_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_55_65_KNN_model_norm <- DEB_RNA_FAIMS_55_65_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_55_65_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_55_65_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_55_65_KNN_model_norm[is.na(DEB_RNA_FAIMS_55_65_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_55_65_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_55_65_KNN_model_norm,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_55_65_KNN_model$Prediction <- DEB_RNA_FAIMS_55_65_knn_pred
table(DEB_RNA_FAIMS_55_65_KNN_model$Prediction)

```

```

##
## False True
## 18603 3219

```

```

DEB_RNA_FAIMS_55_65_KNN$Prediction_KNN <- DEB_RNA_FAIMS_55_65_knn_pred

```

```

d <- DEB_RNA_FAIMS_55_65_KNN %>% filter(Prediction_KNN == "True") %>% select(accessions) %>%
  distinct()
nrow(d)

```

```

## [1] 1726

```

```
write_excel_csv(as.data.frame(d), "S100_DEB_RNA_FAIMS_55_65_accessions.csv")
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

ml_DEB_RNA_FAIMS_55_65_test_C5<-DEB_RNA_FAIMS_55_65_KNN %>%
  select(colnames(ml_DEB_RNA_train))

C5_DEB_RNA_FAIMS_55_65_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_55_65_test_C5)

ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction<-C5_DEB_RNA_FAIMS_55_65_predict_boost_pen
table(ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction)

##
## False True
## 14960 6862

DEB_RNA_FAIMS_55_65_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_55_65_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_55_65_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_55_65_test_C5)
ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_55_65_predict
table(ml_DEB_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER)

##
## False True
## 12322 9500

DEB_RNA_FAIMS_55_65_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_55_65_predict
```

Data setup DEB RNA CV 60_70

```
DEB_RNA_FAIMS_60_70_KNN <- read_delim("AWulf_080221_Ecoli_DEB_S100_FAIMS_60_70_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPx1:NT` = col_character(),
##   `RNPx1:RNA` = col_character(),
##   `RNPx1:best_localization` = col_character(),
##   `RNPx1:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )
```

```

## See spec(...) for full column specifications.
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_60_70_KNN))
names(DEB_RNA_FAIMS_60_70_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_60_70_KNN))
DEB_RNA_FAIMS_60_70_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_60_70_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_60_70_KNN$Sample <- "S100"
DEB_RNA_FAIMS_60_70_KNN <- DEB_RNA_FAIMS_60_70_KNN[DEB_RNA_FAIMS_60_70_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$sequence))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$prepPep))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_60_70_KNN$prepPep))
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_60_70_KNN$prepPep)
DEB_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_60_70_KNN$prepPep)
DEB_RNA_FAIMS_60_70_KNN$Pepseq <- DEB_RNA_FAIMS_60_70_KNN$prepPep
DEB_RNA_FAIMS_60_70_KNN$Peplength <- nchar(DEB_RNA_FAIMS_60_70_KNN$Pepseq, type = "chars")
DEB_RNA_FAIMS_60_70_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_60_70_KNN$pI <- round(pI(DEB_RNA_FAIMS_60_70_KNN$Pepseq), 2)
DEB_RNA_FAIMS_60_70_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_60_70_KNN$Pepseq), 2)
DEB_RNA_FAIMS_60_70_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_60_70_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_60_70_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_60_70_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_60_70_KNN$mw <- round(mw(DEB_RNA_FAIMS_60_70_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_60_70_KNN$cv <- c("60_70")

names(DEB_RNA_FAIMS_60_70_KNN)[names(DEB_RNA_FAIMS_60_70_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_60_70_KNN)[names(DEB_RNA_FAIMS_60_70_KNN) == "precursor_error_(ppm)"] <- "precursor"
DEB_RNA_FAIMS_60_70_KNN <- DEB_RNA_FAIMS_60_70_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_60_70_KNN_model <- DEB_RNA_FAIMS_60_70_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_60_70_KNN_model_norm <- DEB_RNA_FAIMS_60_70_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_60_70_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_60_70_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_60_70_KNN_model_norm[is.na(DEB_RNA_FAIMS_60_70_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_60_70_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_60_70_KNN_model_norm,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_60_70_KNN_model$Prediction <- DEB_RNA_FAIMS_60_70_knn_pred

```



```
DEB_RNA_FAIMS_60_70_KNN$Prediction_KNN <- DEB_RNA_FAIMS_60_70_knn_pred
table(DEB_RNA_FAIMS_60_70_KNN_model$Prediction)
```

```
##
## False True
## 33116 4147
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)
```

```
ml_DEB_RNA_FAIMS_60_70_test_C5<-DEB_RNA_FAIMS_60_70_KNN %>%
  select(colnames(ml_DEB_RNA_train))
```

```
C5_DEB_RNA_FAIMS_60_70_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_60_70_test_C5)
ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction<-C5_DEB_RNA_FAIMS_60_70_predict_boost_pen
DEB_RNA_FAIMS_60_70_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_60_70_predict_boost_pen
table(ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction)
```

```
##
## False True
## 22354 14909
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_60_70_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_60_70_test_C5)
ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_60_70_predict
table(ml_DEB_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 18746 18517
```

```
DEB_RNA_FAIMS_60_70_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_60_70_predict
```

collapsed Proteins

S100 FAIMS 35_45 KNN

```
nrow(DEB_RNA_FAIMS_35_45_KNN %>%
  filter(Prediction_KNN == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 1382
```

S100 FAIMS 35_45 C5.0

```
nrow(DEB_RNA_FAIMS_35_45_KNN %>%
  filter(Prediction_C5 == "True") %>%
```

```
select(accessions) %>%  
distinct()
```

```
## [1] 2153
```

S100 FAIMS 35_45 RIPPER

```
nrow(DEB_RNA_FAIMS_35_45_KNN %>%  
  filter(Prediction_RIPPER == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 2461
```

S100 FAIMS 40_50 KNN

```
nrow(DEB_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1285
```

S100 FAIMS 40_50 C5.0

```
nrow(DEB_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_C5 == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1930
```

S100 FAIMS 40_50 RIPPER

```
nrow(DEB_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_RIPPER == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 2354
```

S100 FAIMS 55_65 KNN

```
nrow(DEB_RNA_FAIMS_55_65_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1726
```

S100 FAIMS 55_65 C5.0

```
nrow(DEB_RNA_FAIMS_55_65_KNN %>%  
  filter(Prediction_C5 == "True") %>%
```

```
select(accessions) %>%
distinct()
```

```
## [1] 2684
```

S100 FAIMS 55_65 RIPPER

```
nrow(DEB_RNA_FAIMS_55_65_KNN %>%
  filter(Prediction_RIPPER == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 3048
```

S100 FAIMS 60_70 KNN

```
nrow(DEB_RNA_FAIMS_60_70_KNN %>%
  filter(Prediction_KNN == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 2133
```

S100 FAIMS 60_70 C5.0

```
nrow(DEB_RNA_FAIMS_60_70_KNN %>%
  filter(Prediction_C5 == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 3505
```

S100 FAIMS 60_70 RIPPER

```
nrow(DEB_RNA_FAIMS_60_70_KNN %>%
  filter(Prediction_RIPPER == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 3692
```

Data setup DEB RNA CV 35_45_55

```
DEB_RNA_FAIMS_35_45_55_KNN <- read_delim("AWulf_050919_Ecoli_DEB_S30_FAIMS_35_45_55_table.csv",
  delim = "\t")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPx1:NT` = col_character(),
##   `RNPx1:RNA` = col_character(),
##   `RNPx1:best_localization` = col_character(),
```

```

## `RNPxl:localization_scores` = col_character(),
## protein_references = col_character(),
## target_decoy = col_character(),
## `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.
names(DEB_RNA_FAIMS_35_45_55_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_35_45_55_KNN))
names(DEB_RNA_FAIMS_35_45_55_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_35_45_55_KNN))
names(DEB_RNA_FAIMS_35_45_55_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_35_45_55_KNN))
names(DEB_RNA_FAIMS_35_45_55_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_35_45_55_KNN))
DEB_RNA_FAIMS_35_45_55_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_35_45_55_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_35_45_55_KNN$Sample <- "S30"
DEB_RNA_FAIMS_35_45_55_KNN <- DEB_RNA_FAIMS_35_45_55_KNN[DEB_RNA_FAIMS_35_45_55_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_35_45_55_KNN$seq))
DEB_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_35_45_55_KNN$prepPep))
DEB_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_35_45_55_KNN$prepPep))
DEB_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("\\(|\\|)", "", DEB_RNA_FAIMS_35_45_55_KNN$prepPep)
DEB_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("\\..*", "", DEB_RNA_FAIMS_35_45_55_KNN$prepPep)
DEB_RNA_FAIMS_35_45_55_KNN$Pepseq <- DEB_RNA_FAIMS_35_45_55_KNN$prepPep
DEB_RNA_FAIMS_35_45_55_KNN$Peplength <- nchar(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq,
  type = "chars")
DEB_RNA_FAIMS_35_45_55_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_35_45_55_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_35_45_55_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_35_45_55_KNN$pI <- round(pI(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq), 2)
DEB_RNA_FAIMS_35_45_55_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq),
  2)
DEB_RNA_FAIMS_35_45_55_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_35_45_55_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_35_45_55_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_35_45_55_KNN$mw <- round(mw(DEB_RNA_FAIMS_35_45_55_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_35_45_55_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_35_45_55_KNN)[names(DEB_RNA_FAIMS_35_45_55_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_35_45_55_KNN)[names(DEB_RNA_FAIMS_35_45_55_KNN) == "precursor_error_(ppm)"] <- "precursor_error"
DEB_RNA_FAIMS_35_45_55_KNN <- DEB_RNA_FAIMS_35_45_55_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_35_45_55_KNN_model <- DEB_RNA_FAIMS_35_45_55_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```

DEB_RNA_FAIMS_35_45_55_KNN_model_norm <- DEB_RNA_FAIMS_35_45_55_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_35_45_55_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_35_45_55_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_35_45_55_KNN_model_norm[is.na(DEB_RNA_FAIMS_35_45_55_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```
DEB_RNA_FAIMS_35_45_55_knn_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_35_45_55_KNN_1
  c1 = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_35_45_55_KNN_model$Prediction <- DEB_RNA_FAIMS_35_45_55_knn_pred
table(DEB_RNA_FAIMS_35_45_55_KNN_model$Prediction)

##
## False True
## 31918 5021

DEB_RNA_FAIMS_35_45_55_KNN$Prediction_KNN <- DEB_RNA_FAIMS_35_45_55_knn_pred
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

ml_DEB_RNA_FAIMS_35_45_55_test_C5<-DEB_RNA_FAIMS_35_45_55_KNN %>%
  select(colnames(ml_DEB_RNA_train))

C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_55_test_C5)

ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction)

##
## False True
## 23372 13567

DEB_RNA_FAIMS_35_45_55_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_55_test_C5)
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)

##
## False True
## 20432 16507

DEB_RNA_FAIMS_35_45_55_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
```

Data setup DEB RNA CV 40_50_60

```
DEB_RNA_FAIMS_40_50_60_KNN <- read_delim("AWulf_060919_Ecoli_DEB_S30_FAIMS_40_50_60_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
```

```

## sequence = col_character(),
## accessions = col_character(),
## `RNPxl:NT` = col_character(),
## `RNPxl:RNA` = col_character(),
## `RNPxl:best_localization` = col_character(),
## `RNPxl:localization_scores` = col_character(),
## protein_references = col_character(),
## target_decoy = col_character(),
## `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.
names(DEB_RNA_FAIMS_40_50_60_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_40_50_60_KNN))
names(DEB_RNA_FAIMS_40_50_60_KNN) <- gsub("\\:", "_", names(DEB_RNA_FAIMS_40_50_60_KNN))
names(DEB_RNA_FAIMS_40_50_60_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_40_50_60_KNN))
names(DEB_RNA_FAIMS_40_50_60_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_40_50_60_KNN))
DEB_RNA_FAIMS_40_50_60_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_40_50_60_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_40_50_60_KNN$Sample <- "S30"
DEB_RNA_FAIMS_40_50_60_KNN <- DEB_RNA_FAIMS_40_50_60_KNN[DEB_RNA_FAIMS_40_50_60_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_40_50_60_KNN$sequence))
DEB_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_40_50_60_KNN$sequence))
DEB_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_40_50_60_KNN$sequence))
DEB_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_40_50_60_KNN$prepPep)
DEB_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("\\\\.\\.*", "", DEB_RNA_FAIMS_40_50_60_KNN$prepPep)
DEB_RNA_FAIMS_40_50_60_KNN$Pepseq <- DEB_RNA_FAIMS_40_50_60_KNN$prepPep
DEB_RNA_FAIMS_40_50_60_KNN$Peplength <- nchar(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq,
  type = "chars")
DEB_RNA_FAIMS_40_50_60_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_40_50_60_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_40_50_60_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_40_50_60_KNN$pI <- round(pI(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq), 2)
DEB_RNA_FAIMS_40_50_60_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq),
  2)
DEB_RNA_FAIMS_40_50_60_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_40_50_60_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_40_50_60_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_40_50_60_KNN$mw <- round(mw(DEB_RNA_FAIMS_40_50_60_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_40_50_60_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_40_50_60_KNN)[names(DEB_RNA_FAIMS_40_50_60_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_40_50_60_KNN)[names(DEB_RNA_FAIMS_40_50_60_KNN) == "precursor_error_(ppm)"] <- "precursor_error"
DEB_RNA_FAIMS_40_50_60_KNN <- DEB_RNA_FAIMS_40_50_60_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_40_50_60_KNN_model <- DEB_RNA_FAIMS_40_50_60_KNN %>% select(colnames(ml_DEB_RNA_train_knn))

```

Normalizing

```
DEB_RNA_FAIMS_40_50_60_KNN_model_norm <- DEB_RNA_FAIMS_40_50_60_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_40_50_60_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_40_50_60_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_40_50_60_KNN_model_norm[is.na(DEB_RNA_FAIMS_40_50_60_KNN_model_norm)] <- 0
```

Classifying DEB RNA FAIMS data

KNN5

```
DEB_RNA_FAIMS_40_50_60_KNN_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_40_50_60_KNN,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_40_50_60_KNN_model$Prediction <- DEB_RNA_FAIMS_40_50_60_KNN_pred
table(DEB_RNA_FAIMS_40_50_60_KNN_model$Prediction)
```

```
##
## False True
## 30945 5226
```

```
DEB_RNA_FAIMS_40_50_60_KNN$Prediction_KNN <- DEB_RNA_FAIMS_40_50_60_KNN_pred
```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```
matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)
```

```
ml_DEB_RNA_FAIMS_35_45_55_test_C5<-DEB_RNA_FAIMS_40_50_60_KNN %>%
  select(colnames(ml_DEB_RNA_train))
```

```
C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_55_test_C5)
```

```
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction)
```

```
##
## False True
## 20805 15366
```

```
DEB_RNA_FAIMS_40_50_60_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_55_test_C5)
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 17649 18522
```

```
DEB_RNA_FAIMS_40_50_60_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
```

Data setup DEB RNA CV 30_50_70

```
DEB_RNA_FAIMS_30_50_70_KNN <- read_delim("AWulf_050919_Ecoli_DEB_S30_FAIMS_30_50_70_table.csv",
  delim = "\t")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sequence = col_character(),
##   accessions = col_character(),
##   `RNPxl:NT` = col_character(),
##   `RNPxl:RNA` = col_character(),
##   `RNPxl:best_localization` = col_character(),
##   `RNPxl:localization_scores` = col_character(),
##   protein_references = col_character(),
##   target_decoy = col_character(),
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
## )

## See spec(...) for full column specifications.

names(DEB_RNA_FAIMS_30_50_70_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_30_50_70_KNN))
names(DEB_RNA_FAIMS_30_50_70_KNN) <- gsub(":", "_", names(DEB_RNA_FAIMS_30_50_70_KNN))
names(DEB_RNA_FAIMS_30_50_70_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_30_50_70_KNN))
names(DEB_RNA_FAIMS_30_50_70_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_30_50_70_KNN))
DEB_RNA_FAIMS_30_50_70_KNN$XLinker <- "DEB"
DEB_RNA_FAIMS_30_50_70_KNN$Nucleotides <- "RNA"
DEB_RNA_FAIMS_30_50_70_KNN$Sample <- "S30"
DEB_RNA_FAIMS_30_50_70_KNN <- DEB_RNA_FAIMS_30_50_70_KNN[DEB_RNA_FAIMS_30_50_70_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
DEB_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_30_50_70_KNN$sequence))
DEB_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_30_50_70_KNN$sequence))
DEB_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_30_50_70_KNN$sequence))
DEB_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_30_50_70_KNN$prepPep)
DEB_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("\\\\.\\*", "", DEB_RNA_FAIMS_30_50_70_KNN$prepPep)
DEB_RNA_FAIMS_30_50_70_KNN$Pepseq <- DEB_RNA_FAIMS_30_50_70_KNN$prepPep
DEB_RNA_FAIMS_30_50_70_KNN$Peplength <- nchar(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq,
  type = "chars")
DEB_RNA_FAIMS_30_50_70_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
DEB_RNA_FAIMS_30_50_70_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
DEB_RNA_FAIMS_30_50_70_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
DEB_RNA_FAIMS_30_50_70_KNN$pI <- round(pI(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq), 2)
DEB_RNA_FAIMS_30_50_70_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq),
  2)
DEB_RNA_FAIMS_30_50_70_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_30_50_70_KNN$RNPxl_best_localization)
DEB_RNA_FAIMS_30_50_70_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_30_50_70_KNN$mw <- round(mw(DEB_RNA_FAIMS_30_50_70_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_30_50_70_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_30_50_70_KNN)[names(DEB_RNA_FAIMS_30_50_70_KNN) == "precursor_m/z"] <- "precursor"
```



```

names(DEB_RNA_FAIMS_30_50_70_KNN)[names(DEB_RNA_FAIMS_30_50_70_KNN) == "precursor_error_(ppm)"] <- "p
DEB_RNA_FAIMS_30_50_70_KNN <- DEB_RNA_FAIMS_30_50_70_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_30_50_70_KNN_model <- DEB_RNA_FAIMS_30_50_70_KNN %>% select(colnames(ml_DEB_RNA_train_knn)

```

Normalizing

```

DEB_RNA_FAIMS_30_50_70_KNN_model_norm <- DEB_RNA_FAIMS_30_50_70_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_30_50_70_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_30_50_70_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_30_50_70_KNN_model_norm[is.na(DEB_RNA_FAIMS_30_50_70_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_30_50_70_KNN_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_30_50_70_KNN,
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_30_50_70_KNN_model$Prediction <- DEB_RNA_FAIMS_30_50_70_KNN_pred
table(DEB_RNA_FAIMS_30_50_70_KNN_model$Prediction)

```

```

##
## False True
## 27680 3776

```

```

DEB_RNA_FAIMS_30_50_70_KNN$Prediction_KNN <- DEB_RNA_FAIMS_30_50_70_KNN_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

```

```

ml_DEB_RNA_FAIMS_35_45_55_test_C5<-DEB_RNA_FAIMS_30_50_70_KNN %>%
  select(colnames(ml_DEB_RNA_train))

```

```

C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_55_test_C5)

```

```

ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction)

```

```

##
## False True
## 18977 12479

```

```

DEB_RNA_FAIMS_30_50_70_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen

```

RIPPER

```

RIPPER_DEB_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_55_test_C5)
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict

```

```
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 16185 15271
```

```
DEB_RNA_FAIMS_30_50_70_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
```

Data setup DEB RNA CV 45_55_65

```
DEB_RNA_FAIMS_45_55_65_KNN <- read_delim("AWulf_050919_Ecoli_DEB_S30_FAIMS_35_45_55_table.csv",  
  delim = "\t")
```

```
## Parsed with column specification:
```

```
## cols(
```

```
##   .default = col_double(),
```

```
##   sequence = col_character(),
```

```
##   accessions = col_character(),
```

```
##   `RNPxl:NT` = col_character(),
```

```
##   `RNPxl:RNA` = col_character(),
```

```
##   `RNPxl:best_localization` = col_character(),
```

```
##   `RNPxl:localization_scores` = col_character(),
```

```
##   protein_references = col_character(),
```

```
##   target_decoy = col_character(),
```

```
##   `PeakAnnotations(mz|intensity|charge|annotation;)` = col_character()
```

```
## )
```

```
## See spec(...) for full column specifications.
```

```
names(DEB_RNA_FAIMS_45_55_65_KNN) <- gsub("\\.", "_", names(DEB_RNA_FAIMS_45_55_65_KNN))
```

```
names(DEB_RNA_FAIMS_45_55_65_KNN) <- gsub(":", "_", names(DEB_RNA_FAIMS_45_55_65_KNN))
```

```
names(DEB_RNA_FAIMS_45_55_65_KNN) <- gsub("\\ ", "_", names(DEB_RNA_FAIMS_45_55_65_KNN))
```

```
names(DEB_RNA_FAIMS_45_55_65_KNN) <- gsub("\\|", "_", names(DEB_RNA_FAIMS_45_55_65_KNN))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$XLinker <- "DEB"
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Nucleotides <- "RNA"
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Sample <- "S30"
```

```
DEB_RNA_FAIMS_45_55_65_KNN <- DEB_RNA_FAIMS_45_55_65_KNN[DEB_RNA_FAIMS_45_55_65_KNN$RNPxl_RNA !=  
  "none", ] %>% filter(target_decoy == "target")
```

```
DEB_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(DEB_RNA_FAIMS_45_55_65_KNN$sequence))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Oxidation)", "", as.character(DEB_RNA_FAIMS_45_55_65_KNN$sequence))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Phospho)", "", as.character(DEB_RNA_FAIMS_45_55_65_KNN$sequence))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("\\(|\\)", "", DEB_RNA_FAIMS_45_55_65_KNN$prepPep)
```

```
DEB_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("\\..*", "", DEB_RNA_FAIMS_45_55_65_KNN$prepPep)
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Pepseq <- DEB_RNA_FAIMS_45_55_65_KNN$prepPep
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Peplength <- nchar(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq,  
  type = "chars")
```

```
DEB_RNA_FAIMS_45_55_65_KNN$aa_class_charged <- str_count(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq,  
  paste("R|K|D|E", collapse = "|"))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$aa_class_polar <- str_count(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq,  
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$aa_class_hydrophobic <- str_count(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq,  
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
```

```
DEB_RNA_FAIMS_45_55_65_KNN$pI <- round(pI(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq), 2)
```

```
DEB_RNA_FAIMS_45_55_65_KNN$aIndex <- round(aIndex(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq),  
  2)
```

```

DEB_RNA_FAIMS_45_55_65_KNN$XL_aa <- gsub("[:A-Z:]", "", DEB_RNA_FAIMS_45_55_65_KNN$RNPxl_best_localizat
DEB_RNA_FAIMS_45_55_65_KNN$Hydro <- round(hydrophobicity(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
DEB_RNA_FAIMS_45_55_65_KNN$mw <- round(mw(DEB_RNA_FAIMS_45_55_65_KNN$Pepseq, monoisotopic = F),
  2)
DEB_RNA_FAIMS_45_55_65_KNN$cv <- c("35_45")

names(DEB_RNA_FAIMS_45_55_65_KNN)[names(DEB_RNA_FAIMS_45_55_65_KNN) == "precursor_m/z"] <- "precursor"
names(DEB_RNA_FAIMS_45_55_65_KNN)[names(DEB_RNA_FAIMS_45_55_65_KNN) == "precursor_error_(ppm_)"] <- "p
DEB_RNA_FAIMS_45_55_65_KNN <- DEB_RNA_FAIMS_45_55_65_KNN %>% filter(RNPxl_isPhospho ==
  "0")

DEB_RNA_FAIMS_45_55_65_KNN_model <- DEB_RNA_FAIMS_45_55_65_KNN %>% select(colnames(ml_DEB_RNA_train_knn

```

Normalizing

```

DEB_RNA_FAIMS_45_55_65_KNN_model_norm <- DEB_RNA_FAIMS_45_55_65_KNN_model %>% select_if(is.numeric)
DEB_RNA_FAIMS_45_55_65_KNN_model_norm <- as.data.frame(lapply(DEB_RNA_FAIMS_45_55_65_KNN_model_norm,
  normalize))
DEB_RNA_FAIMS_45_55_65_KNN_model_norm[is.na(DEB_RNA_FAIMS_45_55_65_KNN_model_norm)] <- 0

```

Classifying DEB RNA FAIMS data

KNN5

```

DEB_RNA_FAIMS_45_55_65_KNN_pred <- knn(train = ml_DEB_RNA_train_knn, test = DEB_RNA_FAIMS_45_55_65_KNN_
  cl = ml_DEB_RNA_train_knn_labels$Curated, k = 5)
DEB_RNA_FAIMS_45_55_65_KNN_model$Prediction <- DEB_RNA_FAIMS_45_55_65_KNN_pred
table(DEB_RNA_FAIMS_45_55_65_KNN_model$Prediction)

```

```

##
## False True
## 31911 5028

```

```

DEB_RNA_FAIMS_45_55_65_KNN$Prediction_KNN <- DEB_RNA_FAIMS_45_55_65_KNN_pred

```

C5.0 with boosting and penalty for false positives

penalize False Positives 4 times more than false negatives.

```

matrix_dimensions<- list(c("True", "False"), c("True", "False"))
names(matrix_dimensions) <- c("predicted", "actual")
error_cost <- matrix(c(0,1,4,0), nrow = 2, dimnames = matrix_dimensions)

ml_DEB_RNA_FAIMS_35_45_55_test_C5<-DEB_RNA_FAIMS_45_55_65_KNN %>%
  select(colnames(ml_DEB_RNA_train))

C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_DEB_RNA_boost_pen, ml_DEB_RNA_FAIMS_35_45_
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction)

```

```

##
## False True

```

```
## 23372 13567
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Prediction_C5<-C5_DEB_RNA_FAIMS_35_45_55_predict_boost_pen
```

RIPPER

```
RIPPER_DEB_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_DEB_RNA, ml_DEB_RNA_FAIMS_35_45_55_test_C5)
ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
table(ml_DEB_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 20432 16507
```

```
DEB_RNA_FAIMS_45_55_65_KNN$Prediction_RIPPER <- RIPPER_DEB_RNA_FAIMS_35_45_55_predict
```

Comparing the most conservative model KNN

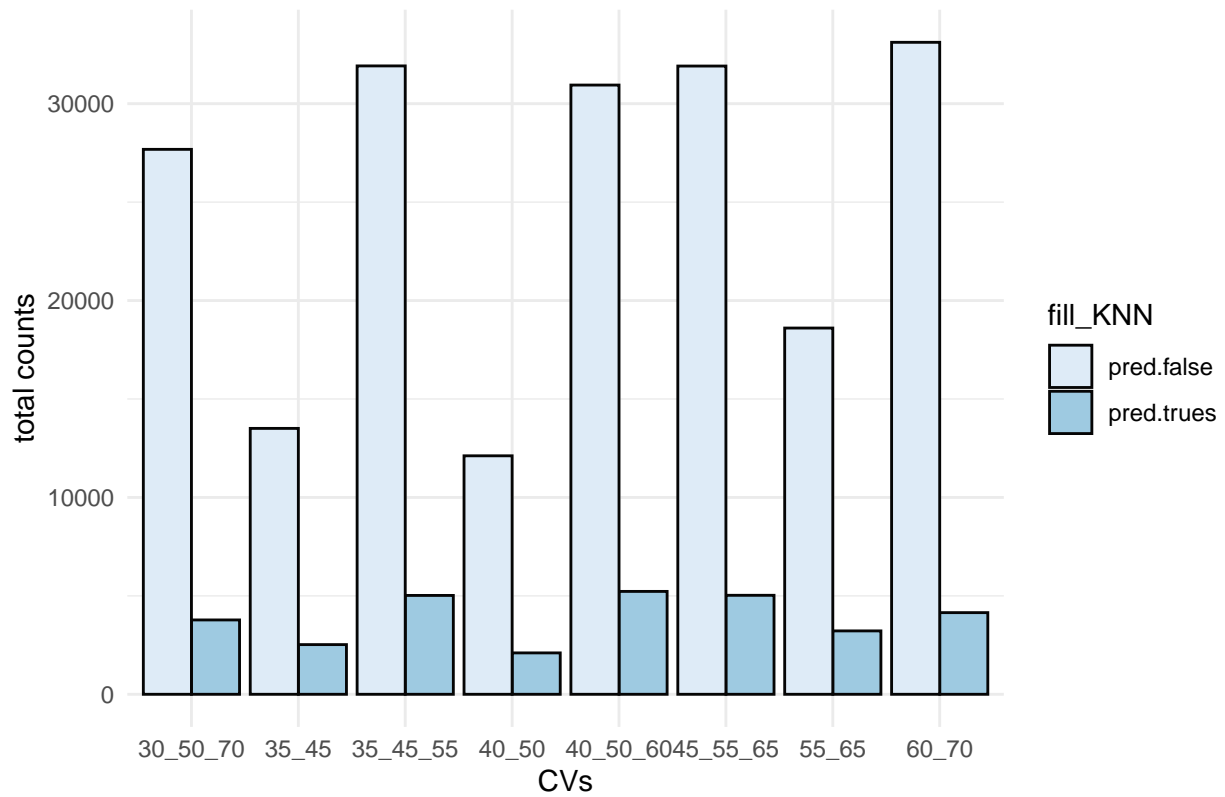
```
a<-as.data.frame(table(DEB_RNA_FAIMS_35_45_KNN$Prediction_KNN))
b<-as.data.frame(table(DEB_RNA_FAIMS_40_50_KNN$Prediction_KNN))
c<-as.data.frame(table(DEB_RNA_FAIMS_55_65_KNN$Prediction_KNN))
d<-as.data.frame(table(DEB_RNA_FAIMS_60_70_KNN$Prediction_KNN))
e<-as.data.frame(table(DEB_RNA_FAIMS_35_45_55_KNN$Prediction_KNN))
f<-as.data.frame(table(DEB_RNA_FAIMS_30_50_70_KNN$Prediction_KNN))
g<-as.data.frame(table(DEB_RNA_FAIMS_40_50_60_KNN$Prediction_KNN))
h<-as.data.frame(table(DEB_RNA_FAIMS_45_55_65_KNN$Prediction_KNN))
```

```
f_KNN<-c("double", "double", "double", "double", "double", "double", "double", "double", "double", "double")
x_KNN<-c("35_45", "35_45", "35_45", "40_50", "40_50", "40_50", "55_65", "55_65", "55_65", "60_70", "60_70")
y_KNN<-c(a[2,2], a[1,2], a[2,2]/(a[2,2]+a[1,2])*100, b[2,2], b[1,2], b[2,2]/(b[2,2]+b[1,2])*100, c[2,2], c[1,2], c[2,2]/(c[2,2]+c[1,2])*100, d[2,2], d[1,2], d[2,2]/(d[2,2]+d[1,2])*100, e[2,2], e[1,2], e[2,2]/(e[2,2]+e[1,2])*100, f[2,2], f[1,2], f[2,2]/(f[2,2]+f[1,2])*100, g[2,2], g[1,2], g[2,2]/(g[2,2]+g[1,2])*100, h[2,2], h[1,2], h[2,2]/(h[2,2]+h[1,2])*100)
fill_KNN<-c("pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm", "pred.trues", "pred.false", "pred.norm")
df_KNN_comparison<-data.frame(x_KNN, y_KNN, fill_KNN, f_KNN)
df_KNN_comparison1<-filter(df_KNN_comparison, fill_KNN!="pred.norm")
df_KNN_comparison2<-filter(df_KNN_comparison, fill_KNN=="pred.norm")
```

```
ggplot(df_KNN_comparison1, aes(x=x_KNN, y=y_KNN, fill = fill_KNN))+
  geom_bar(stat = "identity", position=position_dodge(), bandwidth =1, bins = 3, color="gray2")+
  scale_fill_brewer(palette = "Blues")+
  ggtitle("CV comparison")+
  theme_minimal()+
  ylab("total counts")+
  xlab("CVs")
```

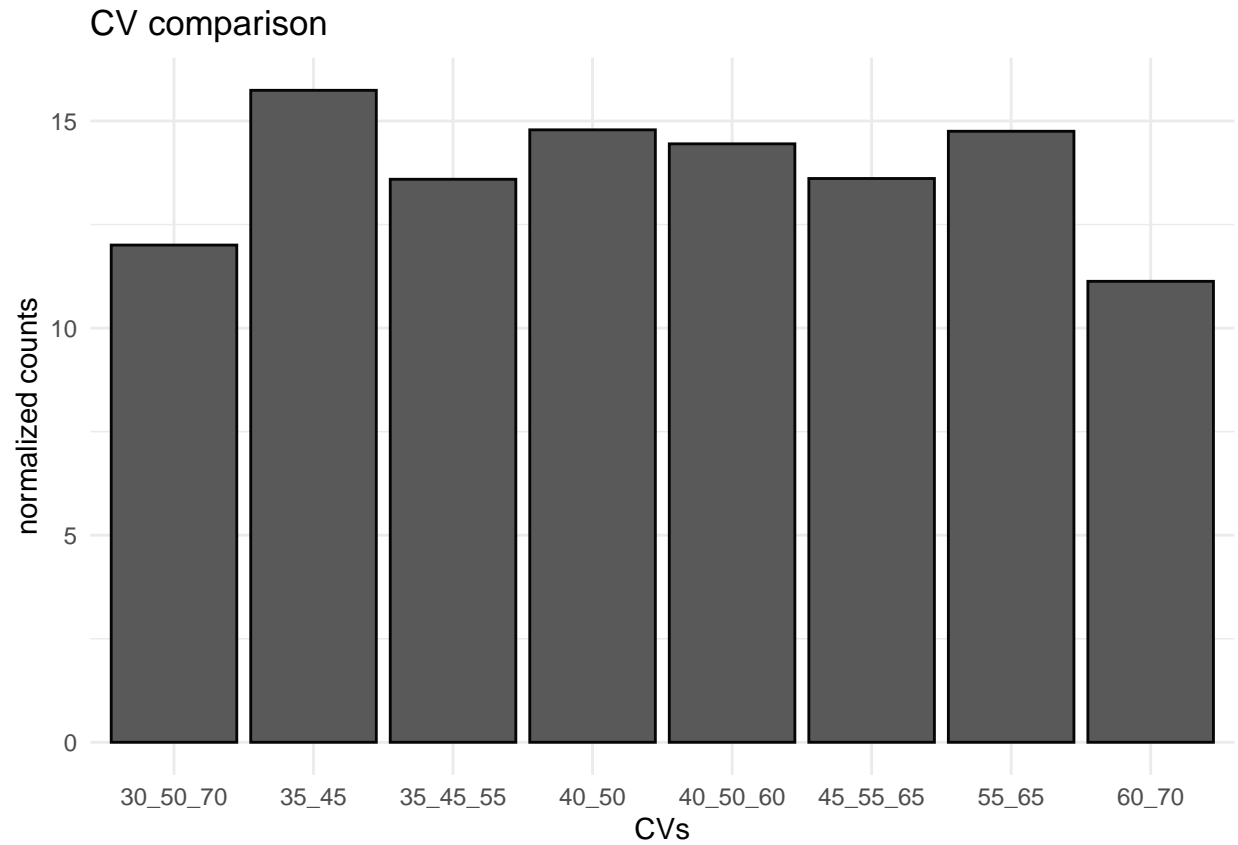
```
## Warning: Ignoring unknown parameters: bandwidth, bins
```

CV comparison



```
ggplot(df_KNN_comparison2, aes(x=x_KNN, y=y_KNN))+  
  geom_bar(stat="identity", bandwidth =1, bins = 3, color="gray2")+  
  ggtitle("CV comparison")+  
  ylab("normalized counts")+  
  xlab("CVs")+  
  theme_minimal()
```

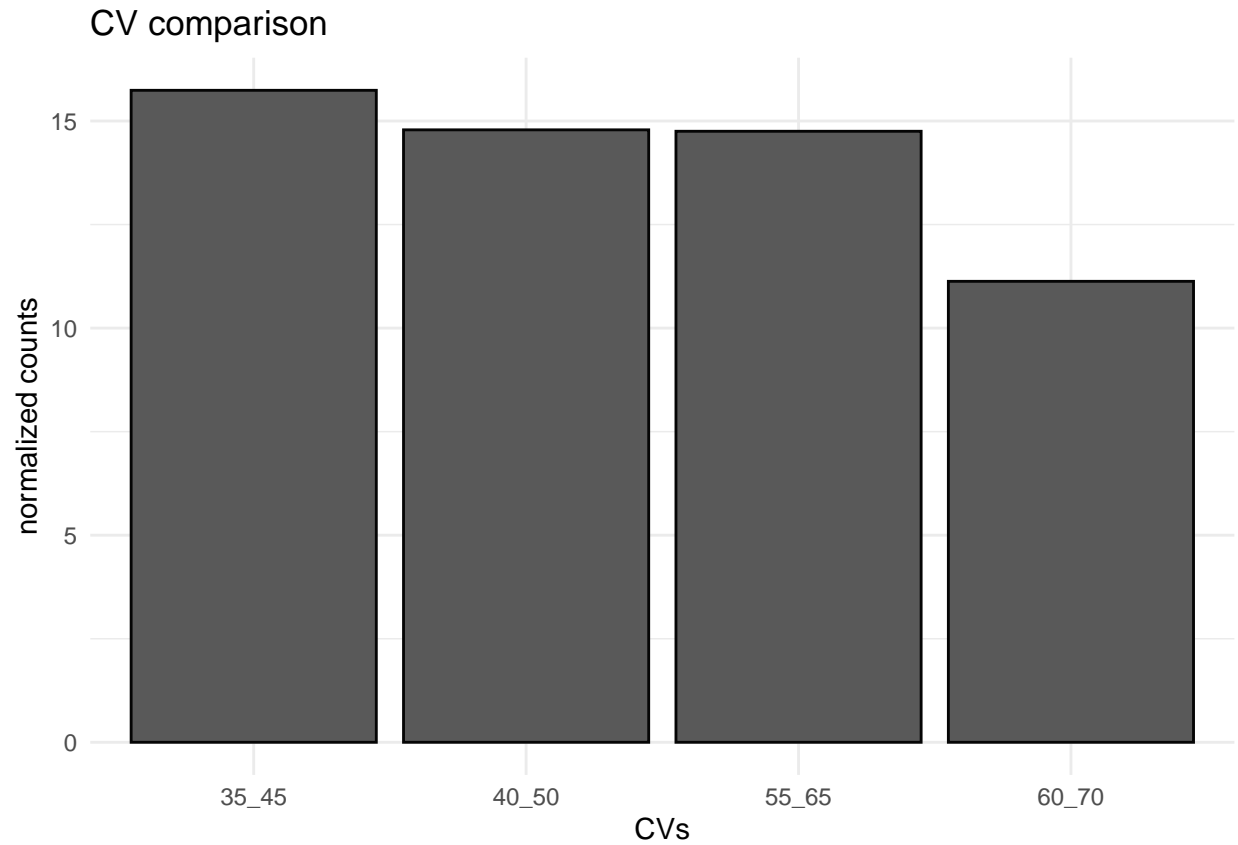
Warning: Ignoring unknown parameters: bandwidth, bins



```
df_KNN_comparison3<-filter(df_KNN_comparison2, f_KNN=="double")

ggplot(df_KNN_comparison3, aes(x=x_KNN, y=y_KNN))+
  geom_bar(stat="identity", bandwidth =1, bins = 3, color="gray2")+
  ggtitle("CV comparison")+
  ylab("normalized counts")+
  xlab("CVs")+
  theme_minimal()
```

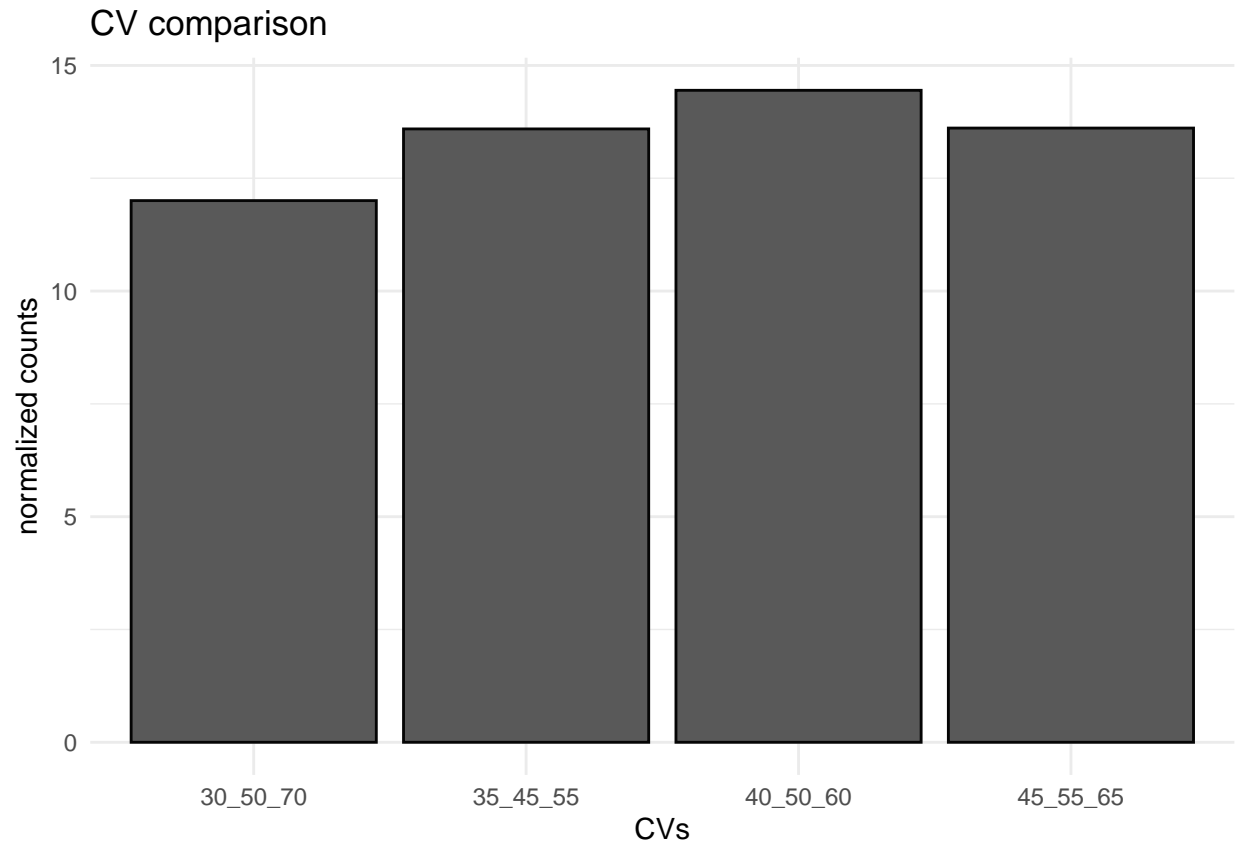
```
## Warning: Ignoring unknown parameters: bandwidth, bins
```



```
df_KNN_comparison4<-filter(df_KNN_comparison2, f_KNN=="triple")

ggplot(df_KNN_comparison4, aes(x=x_KNN, y=y_KNN))+
  geom_bar(stat="identity", bandwidth =1, bins = 3, color="gray2")+
  ggtitle("CV comparison")+
  ylab("normalized counts")+
  xlab("CVs")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: bandwidth, bins
```



Export lists

```
S100_DEB_RNA_FAIMS_total_predicted_trues<- rbind(DEB_RNA_FAIMS_35_45_KNN, DEB_RNA_FAIMS_40_50_KNN, DEB_RNA_FAIMS_45_55_KNN, DEB_RNA_FAIMS_50_60_KNN, DEB_RNA_FAIMS_55_65_KNN)
filter(Prediction_KNN == "True")

write_excel_csv(S100_DEB_RNA_FAIMS_total_predicted_trues, "S100_DEB_RNA_FAIMS_total_KNN_prediction.csv")
```