

# UV RNA FAIMS S100

## Load required packages

```
## corrrplot 0.84 loaded
## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.0    v purrr  0.3.4
## v tidyr  1.0.3     v dplyr  0.8.5
## v readr  1.3.1     v forcats 0.5.0
## -- 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:XL 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 UV 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 (UV, DEB).

### Setting up the dataset

Harmonize the number of trues and falses and dropping all categorical columns.

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

m1_UV_RNA<-data1 %>%
  filter(XLinker == "UV", Curated == "0", Nucleotides == "RNA") %>%
  sample_n(838) %>%
  rbind(RNA_UV)
rows_UV_RNA <- sample(nrow(m1_UV_RNA))
m1_UV_RNA <- m1_UV_RNA[rows_UV_RNA, ] %>%
  dplyr::select_if(is.numeric)
m1_UV_RNA<-m1_UV_RNA[, !(colnames(m1_UV_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_UV_RNA$Curated<-as.factor(ml_UV_RNA$Curated)
ml_UV_RNA$Curated<-factor(ml_UV_RNA$Curated, levels = c("0", "1"), labels = c("False", "True"))
ml_UV_RNA_train <- ml_UV_RNA[1:1340, ]
ml_UV_RNA_test <- ml_UV_RNA[1341:1676, ]
```

## 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_UV_RNA_test_knn<-ml_UV_RNA_test[,!(names(ml_UV_RNA_test) %in% knn_drop)]
ml_UV_RNA_test_knn<-as.data.frame(lapply(ml_UV_RNA_test_knn[1:39], normalize))
ml_UV_RNA_test_knn_labels<-ml_UV_RNA_test[1:336, 5]

ml_UV_RNA_train_knn<-ml_UV_RNA_train[,!(names(ml_UV_RNA_train) %in% knn_drop)]
ml_UV_RNA_train_knn<-as.data.frame(lapply(ml_UV_RNA_train_knn[1:39], normalize))
ml_UV_RNA_train_knn_labels<-as.data.frame(ml_UV_RNA_train[1:1340, 5])
```

## Evaluate Model Performance UV RNA

Here, I chose 5 neighbors:

```
UV_RNA_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = ml_UV_RNA_test_knn,
  cl = ml_UV_RNA_train_knn_labels$Curated, k =5 )
CrossTable(x = ml_UV_RNA_test_knn_labels$Curated, y = UV_RNA_knn_pred, prop.chisq = F)
```

```
##
##
##   Cell Contents
## |-----|
## |                N |
## |          N / Row Total |
## |          N / Col Total |
## |          N / Table Total |
## |-----|
##
##
## Total Observations in Table:  336
##
##
##                                | UV_RNA_knn_pred
## ml_UV_RNA_test_knn_labels$Curated |      False |      True | Row Total |
## -----|-----|-----|-----|
```

```

##           False |         147 |         11 |         158 |
##           |         0.930 |         0.070 |         0.470 |
##           |         0.948 |         0.061 |         |
##           |         0.438 |         0.033 |         |
## -----|-----|-----|-----|
##           True |          8 |         170 |         178 |
##           |         0.045 |         0.955 |         0.530 |
##           |         0.052 |         0.939 |         |
##           |         0.024 |         0.506 |         |
## -----|-----|-----|-----|
##           Column Total |         155 |         181 |         336 |
##           |         0.461 |         0.539 |         |
## -----|-----|-----|-----|
##
##

```

## Decision trees and classification rules

### C5.0

```

library(C50)

C5_drop <- c("isotope_error", "RNPx1_a_ion_score", "precursor_purity")

ml_UV_RNA_train_C5<-ml_UV_RNA_train[,!(names(ml_UV_RNA_train) %in% C5_drop)]
ml_UV_RNA_test_C5<-ml_UV_RNA_test[,!(names(ml_UV_RNA_test) %in% C5_drop)]

```

### UV 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_UV_RNA_boost_pen<-C5.0(x=ml_UV_RNA_train_C5[-5], y=ml_UV_RNA_train_C5$Curated,
                                costs = error_cost, trials = 10)
summary(C5_model_UV_RNA_boost_pen)

```

```

##
## Call:
## C5.0.default(x = ml_UV_RNA_train_C5[-5], y = ml_UV_RNA_train_C5$Curated,
## trials = 10, costs = error_cost)
##
##
## C5.0 [Release 2.07 GPL Edition]      Wed Jun 23 11:03:15 2021
## -----
##
## Class specified by attribute `outcome'
##
## Read 1340 cases (40 attributes) from undefined.data
## Read misclassification costs from undefined.costs

```

```

##
## ----- Trial 0: -----
##
## Decision tree:
##
## score > 0.061008:
## :...RNPxl_pl_pc_MIC <= 0.007726523:
## :   :...RNPxl_pl_MIC > 0.04139248:
## :   :   :...precursor <= 753.8025: True (11)
## :   :   :   precursor > 753.8025: False (16/3)
## :   :   RNPxl_pl_MIC <= 0.04139248:
## :   :   :...RNPxl_pl_pc_MIC <= 0.0001836105: False (561/11)
## :   :   :   RNPxl_pl_pc_MIC > 0.0001836105:
## :   :   :   :...RNPxl_pl_modds <= 0.730354: False (72/9)
## :   :   :   :   RNPxl_pl_modds > 0.730354:
## :   :   :   :   :...precursor_intensity <= 1838890: False (8/3)
## :   :   :   :   :   precursor_intensity > 1838890: True (11)
## :   RNPxl_pl_pc_MIC > 0.007726523:
## :   :...aa_class_polar > 5: False (5/1)
## :   :   aa_class_polar <= 5:
## :   :   :...RNPxl_total_MIC > 0.1157202: True (64)
## :   :   :   RNPxl_total_MIC <= 0.1157202:
## :   :   :   :...RNPxl_RNA_MASS_z0 > 404.0022:
## :   :   :   :   :...Hydro <= -0.49: False (6/3)
## :   :   :   :   :   Hydro > -0.49: True (22)
## :   :   :   :   RNPxl_RNA_MASS_z0 <= 404.0022:
## :   :   :   :   :...score > 0.419009: False (5)
## :   :   :   :   :   score <= 0.419009:
## :   :   :   :   :   :...RNPxl_marker_ions_score <= 0.06687339: True (6)
## :   :   :   :   :   :   RNPxl_marker_ions_score > 0.06687339: False (2)
## score <= 0.061008:
## :...RNPxl_modds <= 0.2422339:
## :   :...aIndex <= 90.69: False (21/7)
## :   :   aIndex > 90.69: True (21)
## :   RNPxl_modds > 0.2422339:
## :   :...pI > 6.16:
## :   :   :...Hydro > -0.39: True (263)
## :   :   :   Hydro <= -0.39:
## :   :   :   :...RT <= 1995.088: True (26)
## :   :   :   :   RT > 1995.088: False (9/6)
## :   pI <= 6.16:
## :   :...RNPxl_RNA_MASS_z0 <= 244.0695: False (5/2)
## :   :   RNPxl_RNA_MASS_z0 > 244.0695:
## :   :   :...aa_class_charged <= 2:
## :   :   :   :...RNPxl_best_localization_score <= 0.0497132: False (3)
## :   :   :   :   RNPxl_best_localization_score > 0.0497132: True (5)
## :   :   :   aa_class_charged > 2:
## :   :   :   :...aa_class_polar <= 3: True (79/1)
## :   :   :   :   aa_class_polar > 3:
## :   :   :   :   :...RNPxl_RNA_MASS_z0 > 635.0778: True (49/1)
## :   :   :   :   :   RNPxl_RNA_MASS_z0 <= 635.0778:
## :   :   :   :   :   :...RNPxl_pl_MIC > 0.02491653: True (26)
## :   :   :   :   :   :   RNPxl_pl_MIC <= 0.02491653:
## :   :   :   :   :   :   :...RNPxl_marker_ions_score <= 0.09317322: False (32/22)

```

```

##                                     RNPxl_marker_ions_score > 0.09317322: True (12)
##
## ----- Trial 1: -----
##
## Decision tree:
##
## score <= 0.0726375:
## :...RNPxl_partial_loss_score <= 0.7555764: False (28/19.6)
## :   RNPxl_partial_loss_score > 0.7555764: True (869/27.5)
## score > 0.0726375:
## :...RNPxl_pl_pc_MIC <= 0.002406057:
##   :...RNPxl_marker_ions_score > 0.03210174: False (101.9/77.7)
##   :   RNPxl_marker_ions_score <= 0.03210174:
##   :     :...RNPxl_pl_modds <= 0.33299:
##   :     :       :...aa_class_polar <= 6: False (272)
##   :     :       :       aa_class_polar > 6:
##   :     :       :         :...RNPxl_MIC <= 0.4404297: False (73.2/12.7)
##   :     :       :         :         RNPxl_MIC > 0.4404297: True (13.5/0.8)
##   :     :       :         RNPxl_pl_modds > 0.33299:
##   :     :       :         :...score > 0.759875: True (59.1/8.3)
##   :     :       :         :         score <= 0.759875:
##   :     :       :         :         :...RNPxl_MIC <= 0.4907319: False (134.1/46.4)
##   :     :       :         :         :         RNPxl_MIC > 0.4907319: True (16.5/1.5)
##   RNPxl_pl_pc_MIC > 0.002406057:
##   :...precursor > 1085.438: False (7.6/1.5)
##   precursor <= 1085.438:
##   :...RNPxl_RNA_MASS_z0 <= 244.0695: False (9.8/4.5)
##   RNPxl_RNA_MASS_z0 > 244.0695:
##   :...precursor_error_ppm <= 0.2314823: False (15.1/9.1)
##   precursor_error_ppm > 0.2314823:
##   :...RNPxl_pl_Morph <= 2.00818: False (21.2/11.3)
##   RNPxl_pl_Morph > 2.00818: True (198.5/2.3)
##
## ----- Trial 2: -----
##
## Decision tree:
##
## RNPxl_pl_pc_MIC > 0.001713034: True (952.1/52.3)
## RNPxl_pl_pc_MIC <= 0.001713034:
## :...RNPxl_precursor_score > 0.002855255: True (94.3/1.2)
##   RNPxl_precursor_score <= 0.002855255:
##   :...RNPxl_total_loss_score > 13.47351:
##   :   :...RNPxl_pl_pc_MIC > 0.001647775: False (7.3)
##   :   :   RNPxl_pl_pc_MIC <= 0.001647775:
##   :   :     :...RNPxl_pl_im_MIC <= 0.007103287: True (153.1/2.1)
##   :   :     :       RNPxl_pl_im_MIC > 0.007103287: False (33.9/25.7)
##   RNPxl_total_loss_score <= 13.47351:
##   :...RNPxl_pl_MIC > 0.04215598: True (124.6/9.4)
##   RNPxl_pl_MIC <= 0.04215598:
##   :...RNPxl_pl_modds <= 0.33299: False (317.8/30.2)
##   RNPxl_pl_modds > 0.33299:
##   :...precursor > 1005.797: False (33.2)
##   precursor <= 1005.797:
##   :...RNPxl_score <= -2.285719: False (24.1)

```

```

##           RNPxl_score > -2.285719:
##           :...RNPxl_total_MIC <= 0.04661144: True (91.7/2.4)
##           RNPxl_total_MIC > 0.04661144: False (130.2/85.5)
##
## ----- Trial 3: -----
##
## Decision tree:
##
## score > 0.279249:
## :...RNPxl_pl_pc_MIC > 0.021649: True (131/7.8)
## :   RNPxl_pl_pc_MIC <= 0.021649:
## :     :...RNPxl_total_MIC > 0.4968228: True (59.5/7.4)
## :     RNPxl_total_MIC <= 0.4968228:
## :       :...RNPxl_marker_ions_score > 0.07374955: True (56.5/8.1)
## :       RNPxl_marker_ions_score <= 0.07374955:
## :         :...RNPxl_pl_modds <= 0.07581517: False (132)
## :         RNPxl_pl_modds > 0.07581517:
## :           :...RNPxl_partial_loss_score <= 5.0293: False (194.9/52.1)
## :           RNPxl_partial_loss_score > 5.0293:
## :             :...precursor > 1005.471: False (60)
## :             precursor <= 1005.471:
## :               :...RNPxl_best_localization_score <= 0.2996045: True (134.5/13.6)
## :               RNPxl_best_localization_score > 0.2996045: False (7.2)
## score <= 0.279249:
## :...RNPxl_precursor_score > 0.01703667: True (127.6)
##   RNPxl_precursor_score <= 0.01703667:
##     :...RNPxl_total_loss_score <= 1.654075:
##     :...RNPxl_partial_loss_score > 4.360192: True (36.3/0.5)
##     :   RNPxl_partial_loss_score <= 4.360192:
##     :     :...RT <= 2036.58: True (31.8/1.4)
##     :     RT > 2036.58: False (43.1/5.6)
##     RNPxl_total_loss_score > 1.654075:
##     :...RNPxl_pl_pc_MIC > 0.01650992: True (135.5)
##     RNPxl_pl_pc_MIC <= 0.01650992:
##     :...RNPxl_pl_MIC > 0.04691937: True (81.4)
##     RNPxl_pl_MIC <= 0.04691937:
##     :...RNPxl_score <= 0.1529219: False (53.1/33.2)
##     RNPxl_score > 0.1529219:
##     :...Hydro > -0.2: True (312.3/9.5)
##     Hydro <= -0.2:
##     :...RNPxl_Da_difference <= -0.0007993339: True (87.8)
##     RNPxl_Da_difference > -0.0007993339:
##     :...aa_class_charged > 5: True (30.3)
##     aa_class_charged <= 5:
##     :...RNPxl_total_MIC > 0.5777186: True (25.5)
##     RNPxl_total_MIC <= 0.5777186:
##     :...RNPxl_modds > 11.55659: True (16.4)
##     RNPxl_modds <= 11.55659: [S1]
##
## SubTree [S1]
##
## RNPxl_immonium_score <= 0.03291839: True (54.3/5.5)
## RNPxl_immonium_score > 0.03291839:
## :...RNPxl_pl_MIC <= 0.006622917: True (13.8/0.9)

```



```

##      RNPxl_pl_MIC > 0.006622917: False (64.3/21.6)
##
## ----- Trial 4: -----
##
## Decision tree:
##
## score <= 0.279249:
## :...RNPxl_pl_modds <= 0.06587414: False (73.6/44.2)
## :   RNPxl_pl_modds > 0.06587414:
## :     :...RNPxl_precursor_score > 0.01703667: True (98.8)
## :       RNPxl_precursor_score <= 0.01703667:
## :         :...RT <= 1707.911: True (112.7/2.2)
## :           RT > 1707.911:
## :             :...Peplength <= 9: False (72.9/46.2)
## :               Peplength > 9:
## :                 :...Hydro <= -0.9: False (30.2/20.6)
## :                   Hydro > -0.9:
## :                     :...aIndex > 126.96: False (27.9/19.5)
## :                       aIndex <= 126.96:
## :                         :...RNPxl_pl_MIC > 0.01001978: True (562.5/20.6)
## :                           RNPxl_pl_MIC <= 0.01001978:
## :                             :...RNPxl_pl_modds <= 0.1546841: True (83.5/0.7)
## :                               RNPxl_pl_modds > 0.1546841:
## :                                 :...RNPxl_precursor_score <= 0.002451922: False (79.4/52)
## :                                   RNPxl_precursor_score > 0.002451922: True (23.6)
## score > 0.279249:
## :...RNPxl_pl_pc_MIC > 0.007726523:
## :   :...RNPxl_RNA_MASS_z0 <= 324.0359: False (60/36.7)
## :     :   RNPxl_RNA_MASS_z0 > 324.0359: True (149.6/4.9)
## :   RNPxl_pl_pc_MIC <= 0.007726523:
## :     :...RNPxl_pl_Morph <= 3.006252: False (144.9)
## :       RNPxl_pl_Morph > 3.006252:
## :         :...RT > 3562.469: False (67.8)
## :           RT <= 3562.469:
## :             :...precursor > 1058.797: False (35.4)
## :               precursor <= 1058.797:
## :                 :...aa_class_hydrophobic > 12: False (11.9)
## :                   aa_class_hydrophobic <= 12:
## :                     :...RNPxl_modds <= 1.15274e-005: True (47.4/0.7)
## :                       RNPxl_modds > 1.15274e-005:
## :                         :...RNPxl_pl_modds <= 0.2060084: False (21.3)
## :                           RNPxl_pl_modds > 0.2060084:
## :                             :...Peplength > 18: False (11.9)
## :                               Peplength <= 18:
## :                                 :...RNPxl_RNA_MASS_z0 <= 244.0695: False (6.9)
## :                                   RNPxl_RNA_MASS_z0 > 244.0695: [S1]
##
## SubTree [S1]
##
## RNPxl_best_localization_score > 0.2416947: True (29.5)
## RNPxl_best_localization_score <= 0.2416947:
## :...RNPxl_pl_pc_MIC <= 0.0004818717: False (79.6/49.9)
## :   RNPxl_pl_pc_MIC > 0.0004818717: True (88.3/3.1)
##

```

```

## ----- Trial 5: -----
##
## Decision tree:
##
## score <= 0.0792426:
## :...RT <= 2026.161: True (267.9)
## :   RT > 2026.161:
## :     :...RNPxl_total_MIC <= 0.1033425: False (26.9/15.5)
## :       RNPxl_total_MIC > 0.1033425:
## :         :...RNPxl_score <= -0.6372114: False (24.8/15.3)
## :           RNPxl_score > -0.6372114:
## :             :...pI > 6.5: True (202/2.2)
## :               pI <= 6.5:
## :                 :...RNPxl_best_localization_score > 0.5767938: True (71.9)
## :                   RNPxl_best_localization_score <= 0.5767938:
## :                     :...RNPxl_RNA_MASS_z0 > 635.0778: True (138.3/2.8)
## :                       RNPxl_RNA_MASS_z0 <= 635.0778:
## :                         :...score > 0.00985416: True (49.4)
## :                           score <= 0.00985416:
## :                             :...RNPxl_immonium_score > 0.1319419: False (6.3/0.3)
## :                               RNPxl_immonium_score <= 0.1319419:
## :                                 :...RNPxl_immonium_score > 0.1050521: True (35)
## :                                   RNPxl_immonium_score <= 0.1050521:
## :                                     :...RNPxl_immonium_score > 0.0983247: False (5.5/0.3)
## :                                       RNPxl_immonium_score <= 0.0983247:
## :                                         :...precursor > 1184.961: True (44.3)
## :                                           precursor <= 1184.961: [S1]
## score > 0.0792426:
## :...RNPxl_pl_pc_MIC > 0.007726523:
## :   :...RNPxl_total_MIC > 0.1157202: True (233.4/1.3)
## :     :   RNPxl_total_MIC <= 0.1157202:
## :       :     :...aa_class_charged <= 1: True (29.5)
## :         :       aa_class_charged > 1: False (97.6/71.4)
## :       RNPxl_pl_pc_MIC <= 0.007726523:
## :         :...RNPxl_pl_Morph <= 4.002007: False (171.8/18)
## :           RNPxl_pl_Morph > 4.002007:
## :             :...RT > 3562.469: False (53.9)
## :               RT <= 3562.469:
## :                 :...precursor > 1062.842: False (29.1)
## :                   precursor <= 1062.842:
## :                     :...aa_class_hydrophobic > 12: False (9)
## :                       aa_class_hydrophobic <= 12:
## :                         :...RNPxl_pl_pc_MIC > 0.00714745: False (7.6)
## :                           RNPxl_pl_pc_MIC <= 0.00714745:
## :                             :...RNPxl_pl_MIC > 0.04691937: True (57.8)
## :                               RNPxl_pl_MIC <= 0.04691937:
## :                                 :...RNPxl_precursor_score > 0: True (21.8)
## :                                   RNPxl_precursor_score <= 0:
## :                                     :...RNPxl_Morph > 4.07617: False (19.8/4.7)
## :                                       RNPxl_Morph <= 4.07617:
## :                                         :...RNPxl_Morph <= 2.374591: False (14.1)
## :                                           RNPxl_Morph > 2.374591: [S2]
##
## SubTree [S1]

```

```

##
## RNPxl_immonium_score > 0.08947683: True (34)
## RNPxl_immonium_score <= 0.08947683:
## :...RNPxl_pl_im_MIC <= 0.001106155: False (69.2/45.5)
##   RNPxl_pl_im_MIC > 0.001106155:
##     :...RNPxl_Da_difference <= -0.002918677: False (3.3)
##       RNPxl_Da_difference > -0.002918677: True (155.7/10.8)
##
## SubTree [S2]
##
## OMS_precursor_mz_error_ppm <= -1.892518: False (12.1/4.8)
## OMS_precursor_mz_error_ppm > -1.892518:
## :...RNPxl_total_MIC > 0.4968228: True (74.4/0.3)
##   RNPxl_total_MIC <= 0.4968228:
##     :...RNPxl_pl_im_MIC > 0.001746027: False (8.6)
##       RNPxl_pl_im_MIC <= 0.001746027:
##         :...aa_class_charged <= 1: False (2.9)
##           aa_class_charged > 1:
##             :...aa_class_polar <= 7: True (155.2/7.2)
##               aa_class_polar > 7: False (3.1)
##
## ----- Trial 6: -----
##
## Decision tree:
##
## RNPxl_pl_pc_MIC <= 0.001713034:
## :...score <= 0.01433:
## :   :...RNPxl_Morph > 22.34067: False (17.1/8.9)
## :   :   RNPxl_Morph <= 22.34067:
## :   :     :...RNPxl_pl_im_MIC <= 0.01459797: True (379.7/20.5)
## :   :     :       RNPxl_pl_im_MIC > 0.01459797: False (12.5/6.7)
## :   :   score > 0.01433:
## :   :     :...RNPxl_pl_MIC > 0.03875625: True (88.6/14.3)
## :   :     :       RNPxl_pl_MIC <= 0.03875625:
## :   :     :         :...RNPxl_pl_modds <= 0.33299: False (172.8/18.7)
## :   :     :         :       RNPxl_pl_modds > 0.33299:
## :   :     :         :         :...RNPxl_score <= 0.2081771: False (162.8/82.1)
## :   :     :         :         :       RNPxl_score > 0.2081771: True (42.1/4.3)
## RNPxl_pl_pc_MIC > 0.001713034:
## :...RNPxl_pl_pc_MIC > 0.01692257: True (565.7/9.9)
##   RNPxl_pl_pc_MIC <= 0.01692257:
##     :...score <= 0.358071:
##       :...RNPxl_modds <= 8.219478: True (584/14.1)
##       :   RNPxl_modds > 8.219478: False (68.7/54.5)
##       score > 0.358071:
##         :...RNPxl_precursor_score > 0.009979509: True (24.5)
##         :       RNPxl_precursor_score <= 0.009979509:
##         :         :...aIndex <= 20: True (29.7/1)
##         :         :       aIndex > 20:
##         :         :         :...OMS_precursor_mz_error_ppm <= 0.6800386: False (58.8/14.3)
##         :         :         :       OMS_precursor_mz_error_ppm > 0.6800386: True (49.4/3)
##
## ----- Trial 7: -----
##

```

```

## Decision tree:
##
## aa_class_hydrophobic > 12:
## :...RNPxl_total_loss_score <= 11.3473: False (49.6/0.2)
## :   RNPxl_total_loss_score > 11.3473: True (14)
## aa_class_hydrophobic <= 12:
## :...RNPxl_pl_pc_MIC > 0.003447095:
##   :...aa_class_polar > 5: False (31.7/18.1)
##   :   aa_class_polar <= 5:
##   :     :...RNPxl_score <= -1.828889:
##   :     :       :...RNPxl_marker_ions_score > 0.04883788: False (10.7/2.7)
##   :     :       :       RNPxl_marker_ions_score <= 0.04883788:
##   :     :       :         :...RNPxl_score > -1.871791: False (10.9/4.4)
##   :     :       :         :       RNPxl_score <= -1.871791:
##   :     :       :         :         :...precursor_intensity <= 7494384: True (171.6/4.6)
##   :     :       :         :         :       precursor_intensity > 7494384: False (7.9/3.5)
##   :     :       :         :       RNPxl_score > -1.828889:
##   :     :       :         :         :...precursor <= 999.3894: True (577.1/0.8)
##   :     :       :         :         :       precursor > 999.3894:
##   :     :       :         :         :         :...precursor <= 1032.44: False (29.2/18.3)
##   :     :       :         :         :         :       precursor > 1032.44:
##   :     :       :         :         :         :         :...precursor_error_ppm <= 0.2342775: False (29.9/23.6)
##   :     :       :         :         :         :         :       precursor_error_ppm > 0.2342775: True (254.5/1.6)
##   RNPxl_pl_pc_MIC <= 0.003447095:
##   :...RNPxl_partial_loss_score <= 0.04691507: False (52.7)
##   :   RNPxl_partial_loss_score > 0.04691507:
##   :     :...score <= 0.0121387:
##   :     :       :...RNPxl_score > 20.26862: True (84.5)
##   :     :       :       RNPxl_score <= 20.26862:
##   :     :       :         :...aa_class_hydrophobic <= 5: True (39.6)
##   :     :       :         :       aa_class_hydrophobic > 5:
##   :     :       :         :         :...RNPxl_score > 19.49398: False (9.4/2.9)
##   :     :       :         :         :       RNPxl_score <= 19.49398:
##   :     :       :         :         :         :...score > 0.00985416: True (40.5)
##   :     :       :         :         :         :       score <= 0.00985416:
##   :     :       :         :         :         :         :...score > 0.00165906: False (20.4/11.2)
##   :     :       :         :         :         :         :       score <= 0.00165906:
##   :     :       :         :         :         :         :         :...RNPxl_RNA_MASS_z0 > 635.0778: True (72.3)
##   :     :       :         :         :         :         :         :       RNPxl_RNA_MASS_z0 <= 635.0778:
##   :     :       :         :         :         :         :         :         :...RNPxl_total_MIC <= 0.1422275: False (5)
##   :     :       :         :         :         :         :         :         :       RNPxl_total_MIC > 0.1422275: [S1]
##   :     :       :         :         :         :         :         :         :         :...score > 0.0121387:
##   :     :       :         :         :         :         :         :         :         :         :...RNPxl_precursor_score > 0.001085989: True (59.3/3)
##   :     :       :         :         :         :         :         :         :         :         :       RNPxl_precursor_score <= 0.001085989:
##   :     :       :         :         :         :         :         :         :         :         :         :...RNPxl_pl_Morph <= 3.113975: False (63.1/2.9)
##   :     :       :         :         :         :         :         :         :         :         :         :       RNPxl_pl_Morph > 3.113975:
##   :     :       :         :         :         :         :         :         :         :         :         :         :...RNPxl_pl_im_MIC > 0.009181726: False (23.9)
##   :     :       :         :         :         :         :         :         :         :         :         :         :       RNPxl_pl_im_MIC <= 0.009181726:
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :...precursor > 1043.75: False (17.9)
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :       precursor <= 1043.75:
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :         :...RNPxl_score <= -2.23855: False (13.8)
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :         :       RNPxl_score > -2.23855:
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :         :         :...RNPxl_pl_MIC > 0.03875625: True (48.4)
##   :     :       :         :         :         :         :         :         :         :         :         :         :         :         :         :       RNPxl_pl_MIC <= 0.03875625:

```

```

##                                     :...RNPxl_immonium_score > 0.01925475: False (11.5)
##                                     RNPxl_immonium_score <= 0.01925475:
##                                     :...aIndex <= 56.67: False (15/2.9)
##                                     aIndex > 56.67: [S2]
##
## SubTree [S1]
##
## aa_class_hydrophobic > 9: True (63.2)
## aa_class_hydrophobic <= 9:
## :...Hydro > -0.16: True (67.2/1.4)
##   Hydro <= -0.16:
##     :...RNPxl_modds <= 0.7470319: False (6.1)
##     RNPxl_modds > 0.7470319:
##       :...RNPxl_pl_err <= 0.003015395: False (75/55.2)
##       RNPxl_pl_err > 0.003015395: True (47.9)
##
## SubTree [S2]
##
## precursor_error_ppm <= 0.4417974: False (17/7.6)
## precursor_error_ppm > 0.4417974:
## :...precursor_error_ppm > 2.94749: False (4)
##   precursor_error_ppm <= 2.94749:
##     :...RT > 3023.513: False (30/20.9)
##     RT <= 3023.513:
##       :...RNPxl_total_loss_score <= 2.523285: True (156.3/2.2)
##       RNPxl_total_loss_score > 2.523285: False (16.9/13.3)
##
## ----- Trial 8: -----
##
## Decision tree:
##
## score <= 0.061008:
## :...RNPxl_pl_MIC <= 0.0083432:
## :   :...RNPxl_modds <= 0.5975084: False (35.7/14.8)
## :   :   RNPxl_modds > 0.5975084: True (155.2/7)
## :   RNPxl_pl_MIC > 0.0083432:
## :   :...RNPxl_pl_im_MIC <= 0.006229385:
## :   :   :...RNPxl_pl_err > 0.001676484: True (871.6/7.3)
## :   :   :   RNPxl_pl_err <= 0.001676484:
## :   :   :   :...RNPxl_pl_err <= 0.001613237: True (167.3/4)
## :   :   :   :   RNPxl_pl_err > 0.001613237: False (2.5/0.3)
## :   :   RNPxl_pl_im_MIC > 0.006229385:
## :   :   :...RNPxl_RNA_MASS_z0 <= 244.0695: False (5.6)
## :   :   :   RNPxl_RNA_MASS_z0 > 244.0695:
## :   :   :   :...RNPxl_mass_error_p <= 1.028149: False (5.3)
## :   :   :   :   RNPxl_mass_error_p > 1.028149: True (182/9.9)
## score > 0.061008:
## :...RNPxl_pl_pc_MIC > 0.021649: True (255.6/8.4)
##   RNPxl_pl_pc_MIC <= 0.021649:
##     :...precursor > 1057.878: False (57.2)
##     precursor <= 1057.878:
##       :...RNPxl_pl_Morph <= 3.113975:
##       :   :...RNPxl_precursor_score <= 0.0103176: False (134.7/30.4)
##       :   :   RNPxl_precursor_score > 0.0103176: True (29.1)

```

```

##          RNPxl_pl_Morph > 3.113975:
##          :...aa_class_hydrophobic > 12: False (13.8)
##          aa_class_hydrophobic <= 12:
##          :...RNPxl_pl_im_MIC > 0.009181726: False (11.2)
##          RNPxl_pl_im_MIC <= 0.009181726:
##          :...RT > 3783.566: False (10.1)
##          RT <= 3783.566:
##          :...RNPxl_pl_pc_MIC > 0.001696311:
##          :...RNPxl_immonium_score > 0.02689577: False (4.2/1.5)
##          :   RNPxl_immonium_score <= 0.02689577:
##          :   :...precursor_error_ppm <= 0.06626505: False (2.5)
##          :   precursor_error_ppm > 0.06626505: True (179.2/1)
##          RNPxl_pl_pc_MIC <= 0.001696311:
##          :...Peplength > 23: True (30.4/0.1)
##          Peplength <= 23:
##          :...RNPxl_pl_MIC > 0.04691937: True (25.8)
##          RNPxl_pl_MIC <= 0.04691937:
##          :...precursor > 946.3495: False (16.5)
##          precursor <= 946.3495: [S1]
##
## SubTree [S1]
##
## RNPxl_partial_loss_score <= 3.192354: False (40.8/15.8)
## RNPxl_partial_loss_score > 3.192354: True (149.5/15.6)
##
## ----- Trial 9: -----
##
## Decision tree:
##
## RNPxl_pl_pc_MIC <= 0.0001836105:
## :...RNPxl_marker_ions_score > 0.1432697: True (66.3/0.1)
## :   RNPxl_marker_ions_score <= 0.1432697:
## :   :...RNPxl_pl_MIC > 0.05793984: True (59.3/2.6)
## :   RNPxl_pl_MIC <= 0.05793984:
## :   :...RNPxl_precursor_score > 0.002900583: True (54.5/3.6)
## :   RNPxl_precursor_score <= 0.002900583:
## :   :...RNPxl_total_MIC <= 0.4894923: False (403.3/152.6)
## :   RNPxl_total_MIC > 0.4894923:
## :   :...RNPxl_RNA_MASS_z0 <= 324.0359: True (118.8/2.9)
## :   RNPxl_RNA_MASS_z0 > 324.0359: False (28.4/14.9)
## RNPxl_pl_pc_MIC > 0.0001836105:
## :...RNPxl_RNA_MASS_z0 <= 244.0695: False (48.7/26.4)
## RNPxl_RNA_MASS_z0 > 244.0695:
## :...RNPxl_partial_loss_score <= 0.03149006: False (53.5/34.4)
## RNPxl_partial_loss_score > 0.03149006:
## :...RNPxl_pl_modds > 1.364098: True (311)
## RNPxl_pl_modds <= 1.364098:
## :...RNPxl_Da_difference <= -0.0004365474:
## :...RNPxl_pl_MIC <= 0.02963713: True (502.5/4.1)
## :   RNPxl_pl_MIC > 0.02963713: False (17.3/10.6)
## RNPxl_Da_difference > -0.0004365474:
## :...RNPxl_Da_difference <= -0.0003878434: False (10.4/0.4)
## RNPxl_Da_difference > -0.0003878434:
## :...RNPxl_pl_pc_MIC > 0.01306356: True (169.6)

```

```

##          RNPxl_pl_pc_MIC <= 0.01306356:
##          :...score > 0.463182: False (38.7/17.6)
##          score <= 0.463182:
##          :...precursor > 1176.198: False (14.3/5.1)
##          precursor <= 1176.198:
##          :...RNPxl_pl_im_MIC > 0.009081018: False (7.5/1.9)
##          RNPxl_pl_im_MIC <= 0.009081018:
##          :...RNPxl_Morph <= 2.544801: False (5.3/1)
##          RNPxl_Morph > 2.544801: True (307.4/9.6)
##
##
## Evaluation on training data (1340 cases):
##
## Trial          Decision Tree
## -----
##      Size      Errors  Cost
##
##      0      26   69( 5.1%)  0.06
##      1      14  129( 9.6%)  0.20
##      2      11  106( 7.9%)  0.24
##      3      23  121( 9.0%)  0.23
##      4      23  160(11.9%)  0.18
##      5      33  105( 7.8%)  0.14
##      6      14  130( 9.7%)  0.22
##      7      37   77( 5.7%)  0.10
##      8      23   67( 5.0%)  0.16
##      9      18  111( 8.3%)  0.13
## boost          7( 0.5%)  0.01  <<
##
##
##      (a)  (b)  <-classified as
##      ----  ----
##      676   4   (a): class False
##      3   657  (b): class True
##
##
## Attribute usage:
##
## 100.00% score
## 100.00% RNPxl_pl_pc_MIC
## 100.00% aa_class_hydrophobic
## 99.10% RNPxl_pl_modds
## 95.15% RNPxl_precursor_score
## 94.33% RNPxl_partial_loss_score
## 94.25% RNPxl_pl_MIC
## 91.04% RNPxl_total_MIC
## 90.15% precursor
## 83.58% RNPxl_total_loss_score
## 74.93% aa_class_polar
## 73.21% RT
## 67.54% RNPxl_score
## 66.72% RNPxl_marker_ions_score
## 60.90% RNPxl_RNA_MASS_z0
## 57.99% RNPxl_pl_Morph

```

```

## 56.49% RNPxl_pl_im_MIC
## 53.58% RNPxl_modds
## 44.40% Peplength
## 44.25% Hydro
## 40.00% pI
## 37.69% aIndex
## 37.31% RNPxl_Da_difference
## 29.63% RNPxl_best_localization_score
## 28.81% RNPxl_pl_err
## 26.87% RNPxl_Morph
## 22.99% aa_class_charged
## 22.69% RNPxl_immonium_score
## 21.42% precursor_error_ppm
## 16.34% RNPxl_MIC
## 8.28% OMS_precursor_mz_error_ppm
## 5.67% RNPxl_mass_error_p
## 4.78% precursor_intensity
##
##
## Time: 0.3 secs

```

```

C5_UV_RNA_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_test_C5)
CrossTable(ml_UV_RNA_test_C5$Curated, C5_UV_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: 336
##
##
##      | predicted TRUE
## actual TRUE |      False |      True | Row Total |
## -----|-----|-----|-----|
##      False |      143 |      15 |      158 |
##      |      0.426 |      0.045 |      |
## -----|-----|-----|-----|
##      True |      7 |      171 |      178 |
##      |      0.021 |      0.509 |      |
## -----|-----|-----|-----|
## Column Total |      150 |      186 |      336 |
## -----|-----|-----|-----|
##
##

```



# Classifiers

## UV RNA

### RIPPER

```
library(rJava)
library(RWeka)
```

```
RIPPER_model_UV_RNA <- JRip(Curated ~ ., data = ml_UV_RNA_train_C5)
RIPPER_model_UV_RNA
```

```
## JRIP rules:
```

```
## =====
```

```
##
```

```
## (score <= 0.079243) and (RNPxl_pl_pc_MIC >= 0.003078) => Curated=True (342.0/5.0)
```

```
## (score <= 0.098059) and (RNPxl_pl_Morph >= 3.007485) and (RNPxl_precursor_score >= 0.000504) => Cura
```

```
## (RNPxl_pl_pc_MIC >= 0.01727) => Curated=True (78.0/4.0)
```

```
## (score <= 0.015775) and (RNPxl_pl_MIC >= 0.010216) => Curated=True (72.0/6.0)
```

```
## (score <= 0.29027) and (RNPxl_score >= 0.319302) and (Hydro >= 0.26) => Curated=True (27.0/4.0)
```

```
## (RNPxl_pl_pc_MIC >= 0.00049) and (aIndex <= 78) and (score <= 0.535324) => Curated=True (27.0/2.0)
```

```
## (RNPxl_pl_MIC >= 0.038842) and (precursor <= 738.084215) => Curated=True (11.0/1.0)
```

```
## => Curated=False (688.0/34.0)
```

```
##
```

```
## Number of Rules : 8
```

```
summary(RIPPER_model_UV_RNA)
```

```
##
```

```
## === Summary ===
```

```
##
```

```
## Correctly Classified Instances      1280          95.5224 %
```

```
## Incorrectly Classified Instances     60           4.4776 %
```

```
## Kappa statistic                     0.9104
```

```
## Mean absolute error                 0.0844
```

```
## Root mean squared error             0.2054
```

```
## Relative absolute error             16.8818 %
```

```
## Root relative squared error         41.0875 %
```

```
## Total Number of Instances          1340
```

```
##
```

```
## === Confusion Matrix ===
```

```
##
```

```
##   a   b   <-- classified as
```

```
## 654 26 |   a = False
```

```
## 34 626 |   b = True
```

```
RIPPER_UV_RNA_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_test_C5)
```

```
CrossTable(ml_UV_RNA_test_C5$Curated, RIPPER_UV_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: 336
##
##
##          | predicted TRUE
## actual TRUE |   False |   True | Row Total |
## -----|-----|-----|-----|
##          False |      145 |      13 |      158 |
##          |      0.432 |      0.039 |          |
## -----|-----|-----|-----|
##          True |       14 |     164 |      178 |
##          |      0.042 |      0.488 |          |
## -----|-----|-----|-----|
## Column Total |      159 |      177 |      336 |
## -----|-----|-----|-----|
##
##
```

## UV RNA FAIMS runs S100

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

### Data setup UV RNA CV 35\_45

```
UV_RNA_FAIMS_35_45_KNN <- read_delim("AWulf_030919_Ecoli_UV_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(UV_RNA_FAIMS_35_45_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_35_45_KNN))
names(UV_RNA_FAIMS_35_45_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_35_45_KNN))
names(UV_RNA_FAIMS_35_45_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_35_45_KNN))
names(UV_RNA_FAIMS_35_45_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_35_45_KNN))
UV_RNA_FAIMS_35_45_KNN$XLinker <- "UV"
UV_RNA_FAIMS_35_45_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_35_45_KNN$Sample <- "S100"
UV_RNA_FAIMS_35_45_KNN <- UV_RNA_FAIMS_35_45_KNN[UV_RNA_FAIMS_35_45_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
```

```

UV_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_35_45_KNN$sequence))
UV_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_35_45_KNN$prepPep))
UV_RNA_FAIMS_35_45_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_35_45_KNN$prepPep))
UV_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_35_45_KNN$prepPep)
UV_RNA_FAIMS_35_45_KNN$prepPep <- gsub("\\\\.\\.*", "", UV_RNA_FAIMS_35_45_KNN$prepPep)
UV_RNA_FAIMS_35_45_KNN$Pepseq <- UV_RNA_FAIMS_35_45_KNN$prepPep
UV_RNA_FAIMS_35_45_KNN$Peplength <- nchar(UV_RNA_FAIMS_35_45_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_35_45_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_35_45_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_35_45_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_35_45_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_35_45_KNN$pI <- round(pI(UV_RNA_FAIMS_35_45_KNN$Pepseq), 2)
UV_RNA_FAIMS_35_45_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_35_45_KNN$Pepseq), 2)
UV_RNA_FAIMS_35_45_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_35_45_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_35_45_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_35_45_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_35_45_KNN$mw <- round(mw(UV_RNA_FAIMS_35_45_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_35_45_KNN$cv <- c("35_45")

names(UV_RNA_FAIMS_35_45_KNN)[names(UV_RNA_FAIMS_35_45_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_35_45_KNN)[names(UV_RNA_FAIMS_35_45_KNN) == "precursor_error_(ppm)"] <- "precursor"
UV_RNA_FAIMS_35_45_KNN <- UV_RNA_FAIMS_35_45_KNN %>% filter(RNPxl_isPhospho == "0")

UV_RNA_FAIMS_35_45_KNN_model <- UV_RNA_FAIMS_35_45_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_35_45_KNN_model_norm <- UV_RNA_FAIMS_35_45_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_35_45_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_35_45_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_35_45_KNN_model_norm[is.na(UV_RNA_FAIMS_35_45_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_35_45_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_35_45_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_35_45_KNN_model$Prediction <- UV_RNA_FAIMS_35_45_knn_pred
table(UV_RNA_FAIMS_35_45_KNN_model$Prediction)

```

```

##
## False True
## 13464 320

```

```

UV_RNA_FAIMS_35_45_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_35_45_test_C5<-UV_RNA_FAIMS_35_45_KNN %>%
  select(colnames(ml_UV_RNA_train))

C5_UV_RNA_FAIMS_35_45_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_35_45_test_C5)

ml_UV_RNA_FAIMS_35_45_test_C5$Prediction<-C5_UV_RNA_FAIMS_35_45_predict_boost_pen
table(ml_UV_RNA_FAIMS_35_45_test_C5$Prediction)

##
## False True
## 11143 2641

UV_RNA_FAIMS_35_45_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_35_45_predict_boost_pen

```

## RIPPER

```

RIPPER_UV_RNA_FAIMS_35_45_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_35_45_test_C5)
ml_UV_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_predict
table(ml_UV_RNA_FAIMS_35_45_test_C5$Prediction_RIPPER)

##
## False True
## 10042 3742

UV_RNA_FAIMS_35_45_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_predict

```

## Data setup UV RNA CV 40\_50

```

UV_RNA_FAIMS_40_50_KNN <- read_delim("AWulf_030919_Ecoli_UV_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(UV_RNA_FAIMS_40_50_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_40_50_KNN))
names(UV_RNA_FAIMS_40_50_KNN) <- gsub(":", "_", names(UV_RNA_FAIMS_40_50_KNN))
names(UV_RNA_FAIMS_40_50_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_40_50_KNN))
names(UV_RNA_FAIMS_40_50_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_40_50_KNN))

```

```

UV_RNA_FAIMS_40_50_KNN$XLinker <- "UV"
UV_RNA_FAIMS_40_50_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_40_50_KNN$Sample <- "S100"
UV_RNA_FAIMS_40_50_KNN <- UV_RNA_FAIMS_40_50_KNN[UV_RNA_FAIMS_40_50_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
UV_RNA_FAIMS_40_50_KNN$prepPep <- gsub("Carbamyl", "", as.character(UV_RNA_FAIMS_40_50_KNN$sequence))
UV_RNA_FAIMS_40_50_KNN$prepPep <- gsub("Oxidation", "", as.character(UV_RNA_FAIMS_40_50_KNN$prepPep))
UV_RNA_FAIMS_40_50_KNN$prepPep <- gsub("Phospho", "", as.character(UV_RNA_FAIMS_40_50_KNN$prepPep))
UV_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_40_50_KNN$prepPep)
UV_RNA_FAIMS_40_50_KNN$prepPep <- gsub("\\\\.\\.", "", UV_RNA_FAIMS_40_50_KNN$prepPep)
UV_RNA_FAIMS_40_50_KNN$Pepseq <- UV_RNA_FAIMS_40_50_KNN$prepPep
UV_RNA_FAIMS_40_50_KNN$Peplength <- nchar(UV_RNA_FAIMS_40_50_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_40_50_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_40_50_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_40_50_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_40_50_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_40_50_KNN$pI <- round(pI(UV_RNA_FAIMS_40_50_KNN$Pepseq), 2)
UV_RNA_FAIMS_40_50_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_40_50_KNN$Pepseq), 2)

## Warning: Sequence 9973 has unrecognized amino acid types. Output value might be
## wrong calculated
UV_RNA_FAIMS_40_50_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_40_50_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_40_50_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_40_50_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)

## Warning: Sequence 9973 has unrecognized amino acid types. Output value might be
## wrong calculated
UV_RNA_FAIMS_40_50_KNN$mw <- round(mw(UV_RNA_FAIMS_40_50_KNN$Pepseq, monoisotopic = F),
  2)

## Warning: Sequence 9973 has unrecognized amino acid types. Output value might be
## wrong calculated
UV_RNA_FAIMS_40_50_KNN$cv <- c("40_50")

names(UV_RNA_FAIMS_40_50_KNN)[names(UV_RNA_FAIMS_40_50_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_40_50_KNN)[names(UV_RNA_FAIMS_40_50_KNN) == "precursor_error_(ppm)"] <- "precursor"
UV_RNA_FAIMS_40_50_KNN <- UV_RNA_FAIMS_40_50_KNN %>% filter(RNPxl_isPhospho == "0")

UV_RNA_FAIMS_40_50_KNN_model <- UV_RNA_FAIMS_40_50_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_40_50_KNN_model_norm <- UV_RNA_FAIMS_40_50_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_40_50_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_40_50_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_40_50_KNN_model_norm[is.na(UV_RNA_FAIMS_40_50_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```
UV_RNA_FAIMS_40_50_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_40_50_KNN_model_norm
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_40_50_KNN_model$Prediction <- UV_RNA_FAIMS_40_50_knn_pred
table(UV_RNA_FAIMS_40_50_KNN_model$Prediction)
```

```
##
## False True
## 10829 502
```

```
UV_RNA_FAIMS_40_50_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_40_50_test_C5<-UV_RNA_FAIMS_40_50_KNN %>%
  select(colnames(ml_UV_RNA_train))
```

```
C5_UV_RNA_FAIMS_40_50_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_40_50_test_C5)
```

```
ml_UV_RNA_FAIMS_40_50_test_C5$Prediction<-C5_UV_RNA_FAIMS_40_50_predict_boost_pen
table(ml_UV_RNA_FAIMS_40_50_test_C5$Prediction)
```

```
##
## False True
## 9631 1700
```

```
UV_RNA_FAIMS_40_50_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_40_50_predict_boost_pen
```

### RIPPER

```
RIPPER_UV_RNA_FAIMS_40_50_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_40_50_test_C5)
ml_UV_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_40_50_predict
table(ml_UV_RNA_FAIMS_40_50_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 8616 2715
```

```
UV_RNA_FAIMS_40_50_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_40_50_predict
```

## Data setup UV RNA CV 55\_65

```
UV_RNA_FAIMS_55_65_KNN <- read_delim("AWulf_030919_Ecoli_UV_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(UV_RNA_FAIMS_55_65_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_55_65_KNN))
names(UV_RNA_FAIMS_55_65_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_55_65_KNN))
names(UV_RNA_FAIMS_55_65_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_55_65_KNN))
names(UV_RNA_FAIMS_55_65_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_55_65_KNN))
UV_RNA_FAIMS_55_65_KNN$XLinker <- "UV"
UV_RNA_FAIMS_55_65_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_55_65_KNN$Sample <- "S100"
UV_RNA_FAIMS_55_65_KNN <- UV_RNA_FAIMS_55_65_KNN[UV_RNA_FAIMS_55_65_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
UV_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_55_65_KNN$sequence))
UV_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_55_65_KNN$prepPep))
UV_RNA_FAIMS_55_65_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_55_65_KNN$prepPep))
UV_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_55_65_KNN$prepPep)
UV_RNA_FAIMS_55_65_KNN$prepPep <- gsub("\\\\.*", "", UV_RNA_FAIMS_55_65_KNN$prepPep)
UV_RNA_FAIMS_55_65_KNN$Pepseq <- UV_RNA_FAIMS_55_65_KNN$prepPep
UV_RNA_FAIMS_55_65_KNN$Peplength <- nchar(UV_RNA_FAIMS_55_65_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_55_65_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_55_65_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_55_65_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_55_65_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_55_65_KNN$pI <- round(pI(UV_RNA_FAIMS_55_65_KNN$Pepseq), 2)
UV_RNA_FAIMS_55_65_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_55_65_KNN$Pepseq), 2)
UV_RNA_FAIMS_55_65_KNN$XL_aa <- gsub("[A-Z:]", "", UV_RNA_FAIMS_55_65_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_55_65_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_55_65_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_55_65_KNN$mw <- round(mw(UV_RNA_FAIMS_55_65_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_55_65_KNN$cv <- c("55_65")

names(UV_RNA_FAIMS_55_65_KNN)[names(UV_RNA_FAIMS_55_65_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_55_65_KNN)[names(UV_RNA_FAIMS_55_65_KNN) == "precursor_error_(ppm)"] <- "precursor"
UV_RNA_FAIMS_55_65_KNN <- UV_RNA_FAIMS_55_65_KNN %>% filter(RNPxl_isPhospho == "0")

UV_RNA_FAIMS_55_65_KNN_model <- UV_RNA_FAIMS_55_65_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_55_65_KNN_model_norm <- UV_RNA_FAIMS_55_65_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_55_65_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_55_65_KNN_model_norm,

```

```
normalize))
UV_RNA_FAIMS_55_65_KNN_model_norm[is.na(UV_RNA_FAIMS_55_65_KNN_model_norm)] <- 0
```

## Classifying UV RNA FAIMS data

### KNN5

```
UV_RNA_FAIMS_55_65_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_55_65_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_55_65_KNN_model$Prediction <- UV_RNA_FAIMS_55_65_knn_pred
table(UV_RNA_FAIMS_55_65_KNN_model$Prediction)
```

```
##
## False True
## 16622 283
```

```
UV_RNA_FAIMS_55_65_KNN$Prediction_KNN <- UV_RNA_FAIMS_55_65_knn_pred
```

```
d <- UV_RNA_FAIMS_55_65_KNN %>% filter(Prediction_KNN == "True") %>% select(accessions) %>%
  distinct()
nrow(d)
```

```
## [1] 245
```

### 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_UV_RNA_FAIMS_55_65_test_C5<-UV_RNA_FAIMS_55_65_KNN %>%
  select(colnames(ml_UV_RNA_train))
```

```
C5_UV_RNA_FAIMS_55_65_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_55_65_test_C5)
```

```
ml_UV_RNA_FAIMS_55_65_test_C5$Prediction<-C5_UV_RNA_FAIMS_55_65_predict_boost_pen
table(ml_UV_RNA_FAIMS_55_65_test_C5$Prediction)
```

```
##
## False True
## 13551 3354
```

```
UV_RNA_FAIMS_55_65_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_55_65_predict_boost_pen
```

### RIPPER

```
RIPPER_UV_RNA_FAIMS_55_65_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_55_65_test_C5)
ml_UV_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_55_65_predict
table(ml_UV_RNA_FAIMS_55_65_test_C5$Prediction_RIPPER)
```

```
##
## False True
## 8184 8721
```



```
UV_RNA_FAIMS_55_65_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_55_65_predict
```

## Data setup UV RNA CV 60\_70

```
UV_RNA_FAIMS_60_70_KNN <- read_delim("AWulf_050919_Ecoli_UV_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(UV_RNA_FAIMS_60_70_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_60_70_KNN))  
names(UV_RNA_FAIMS_60_70_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_60_70_KNN))  
names(UV_RNA_FAIMS_60_70_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_60_70_KNN))  
names(UV_RNA_FAIMS_60_70_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_60_70_KNN))  
UV_RNA_FAIMS_60_70_KNN$XLinker <- "UV"  
UV_RNA_FAIMS_60_70_KNN$Nucleotides <- "RNA"  
UV_RNA_FAIMS_60_70_KNN$Sample <- "S100"  
UV_RNA_FAIMS_60_70_KNN <- UV_RNA_FAIMS_60_70_KNN[UV_RNA_FAIMS_60_70_KNN$RNPxl_RNA !=  
  "none", ] %>% filter(target_decoy == "target")  
UV_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_60_70_KNN$sequence))  
UV_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_60_70_KNN$prepPep))  
UV_RNA_FAIMS_60_70_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_60_70_KNN$prepPep))  
UV_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\(|\\|)", "", UV_RNA_FAIMS_60_70_KNN$prepPep)  
UV_RNA_FAIMS_60_70_KNN$prepPep <- gsub("\\.\\.*", "", UV_RNA_FAIMS_60_70_KNN$prepPep)  
UV_RNA_FAIMS_60_70_KNN$Pepseq <- UV_RNA_FAIMS_60_70_KNN$prepPep  
UV_RNA_FAIMS_60_70_KNN$Peplength <- nchar(UV_RNA_FAIMS_60_70_KNN$Pepseq, type = "chars")  
UV_RNA_FAIMS_60_70_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_60_70_KNN$Pepseq,  
  paste("R|K|D|E", collapse = "|"))  
UV_RNA_FAIMS_60_70_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_60_70_KNN$Pepseq,  
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))  
UV_RNA_FAIMS_60_70_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_60_70_KNN$Pepseq,  
  paste("A|I|L|M|F|V|P|G", collapse = "|"))  
UV_RNA_FAIMS_60_70_KNN$pI <- round(pI(UV_RNA_FAIMS_60_70_KNN$Pepseq), 2)  
UV_RNA_FAIMS_60_70_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_60_70_KNN$Pepseq), 2)  
UV_RNA_FAIMS_60_70_KNN$XL_aa <- gsub("[A-Z:]", "", UV_RNA_FAIMS_60_70_KNN$RNPxl_best_localization)  
UV_RNA_FAIMS_60_70_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_60_70_KNN$Pepseq,  
  scale = "KyteDoolittle"), 2)  
UV_RNA_FAIMS_60_70_KNN$mw <- round(mw(UV_RNA_FAIMS_60_70_KNN$Pepseq, monoisotopic = F),  
  2)  
UV_RNA_FAIMS_60_70_KNN$cv <- c("60_70")
```

```

names(UV_RNA_FAIMS_60_70_KNN)[names(UV_RNA_FAIMS_60_70_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_60_70_KNN)[names(UV_RNA_FAIMS_60_70_KNN) == "precursor_error_(ppm)"] <- "precursor_error_ppm"
UV_RNA_FAIMS_60_70_KNN <- UV_RNA_FAIMS_60_70_KNN %>% filter(RNPxl_isPhospho == "0")

UV_RNA_FAIMS_60_70_KNN_model <- UV_RNA_FAIMS_60_70_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_60_70_KNN_model_norm <- UV_RNA_FAIMS_60_70_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_60_70_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_60_70_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_60_70_KNN_model_norm[is.na(UV_RNA_FAIMS_60_70_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_60_70_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_60_70_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_60_70_KNN_model$Prediction <- UV_RNA_FAIMS_60_70_knn_pred
table(UV_RNA_FAIMS_60_70_KNN_model$Prediction)

```

```

##
## False True
## 12450 345

```

```
UV_RNA_FAIMS_60_70_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_60_70_test_C5 <- UV_RNA_FAIMS_60_70_KNN %>%
  select(colnames(ml_UV_RNA_train))

```

```
C5_UV_RNA_FAIMS_60_70_predict_boost_pen <- predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_60_70_test_C5)
```

```

ml_UV_RNA_FAIMS_60_70_test_C5$Prediction <- C5_UV_RNA_FAIMS_60_70_predict_boost_pen
table(ml_UV_RNA_FAIMS_60_70_test_C5$Prediction)

```

```

##
## False True
## 10810 1985

```

```
UV_RNA_FAIMS_60_70_KNN$Prediction_C5 <- C5_UV_RNA_FAIMS_60_70_predict_boost_pen
```

### RIPPER

```

RIPPER_UV_RNA_FAIMS_60_70_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_60_70_test_C5)
ml_UV_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_60_70_predict

```

```
table(ml_UV_RNA_FAIMS_60_70_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 6968 5827
```

```
UV_RNA_FAIMS_60_70_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_60_70_predict
```

## collapsed Proteins

### S100 FAIMS 35\_45 KNN

```
nrow(UV_RNA_FAIMS_35_45_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 241
```

### S100 FAIMS 35\_45 C5.0

```
nrow(UV_RNA_FAIMS_35_45_KNN %>%  
  filter(Prediction_C5 == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1648
```

### S100 FAIMS 35\_45 RIPPER

```
nrow(UV_RNA_FAIMS_35_45_KNN %>%  
  filter(Prediction_RIPPER == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 2042
```

### S100 FAIMS 40\_50 KNN

```
nrow(UV_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 198
```

### S100 FAIMS 40\_50 C5.0

```
nrow(UV_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_C5 == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 924
```

### S100 FAIMS 40\_50 RIPPER

```
nrow(UV_RNA_FAIMS_40_50_KNN %>%  
  filter(Prediction_RIPPER == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1446
```

### S100 FAIMS 55\_65 KNN

```
nrow(UV_RNA_FAIMS_55_65_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 245
```

### S100 FAIMS 55\_65 C5.0

```
nrow(UV_RNA_FAIMS_55_65_KNN %>%  
  filter(Prediction_C5 == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1954
```

### S100 FAIMS 55\_65 RIPPER

```
nrow(UV_RNA_FAIMS_55_65_KNN %>%  
  filter(Prediction_RIPPER == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 3025
```

### S100 FAIMS 60\_70 KNN

```
nrow(UV_RNA_FAIMS_60_70_KNN %>%  
  filter(Prediction_KNN == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 201
```

### S100 FAIMS 60\_70 C5.0

```
nrow(UV_RNA_FAIMS_60_70_KNN %>%  
  filter(Prediction_C5 == "True") %>%  
  select(accessions) %>%  
  distinct())
```

```
## [1] 1233
```

## S100 FAIMS 60\_70 RIPPER

```
nrow(UV_RNA_FAIMS_60_70_KNN %>%
  filter(Prediction_RIPPER == "True") %>%
  select(accessions) %>%
  distinct())
```

```
## [1] 2521
```

## Data setup UV RNA CV 35\_45\_55

```
UV_RNA_FAIMS_35_45_55_KNN <- read_delim("AWulf_030919_Ecoli_UV_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(UV_RNA_FAIMS_35_45_55_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_35_45_55_KNN))
names(UV_RNA_FAIMS_35_45_55_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_35_45_55_KNN))
names(UV_RNA_FAIMS_35_45_55_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_35_45_55_KNN))
names(UV_RNA_FAIMS_35_45_55_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_35_45_55_KNN))
UV_RNA_FAIMS_35_45_55_KNN$XLinker <- "UV"
UV_RNA_FAIMS_35_45_55_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_35_45_55_KNN$Sample <- "S100"
UV_RNA_FAIMS_35_45_55_KNN <- UV_RNA_FAIMS_35_45_55_KNN[UV_RNA_FAIMS_35_45_55_KNN$RNPxl_RNA !=
  "none", ] %>% filter(target_decoy == "target")
UV_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_35_45_55_KNN$sequence))
UV_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_35_45_55_KNN$prepPep))
UV_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_35_45_55_KNN$prepPep))
UV_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("\\(|\\|)", "", UV_RNA_FAIMS_35_45_55_KNN$prepPep)
UV_RNA_FAIMS_35_45_55_KNN$prepPep <- gsub("\\\\.\\.*", "", UV_RNA_FAIMS_35_45_55_KNN$prepPep)
UV_RNA_FAIMS_35_45_55_KNN$Pepseq <- UV_RNA_FAIMS_35_45_55_KNN$prepPep
UV_RNA_FAIMS_35_45_55_KNN$Peplength <- nchar(UV_RNA_FAIMS_35_45_55_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_35_45_55_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_35_45_55_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_35_45_55_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_35_45_55_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_35_45_55_KNN$I <- round(pI(UV_RNA_FAIMS_35_45_55_KNN$Pepseq), 2)
UV_RNA_FAIMS_35_45_55_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_35_45_55_KNN$Pepseq),
  2)
```

```

UV_RNA_FAIMS_35_45_55_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_35_45_55_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_35_45_55_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_35_45_55_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_35_45_55_KNN$mw <- round(mw(UV_RNA_FAIMS_35_45_55_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_35_45_55_KNN$cv <- c("35_45")

names(UV_RNA_FAIMS_35_45_55_KNN)[names(UV_RNA_FAIMS_35_45_55_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_35_45_55_KNN)[names(UV_RNA_FAIMS_35_45_55_KNN) == "precursor_error_(ppm)"] <- "precursor_error_ppm"
UV_RNA_FAIMS_35_45_55_KNN <- UV_RNA_FAIMS_35_45_55_KNN %>% filter(RNPxl_isPhospho ==
  "0")

UV_RNA_FAIMS_35_45_55_KNN_model <- UV_RNA_FAIMS_35_45_55_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_35_45_55_KNN_model_norm <- UV_RNA_FAIMS_35_45_55_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_35_45_55_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_35_45_55_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_35_45_55_KNN_model_norm[is.na(UV_RNA_FAIMS_35_45_55_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_35_45_55_knn_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_35_45_55_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_35_45_55_KNN_model$Prediction <- UV_RNA_FAIMS_35_45_55_knn_pred
table(UV_RNA_FAIMS_35_45_55_KNN_model$Prediction)

```

```

##
## False True
## 13464 320

```

```

UV_RNA_FAIMS_35_45_55_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_35_45_55_test_C5 <- UV_RNA_FAIMS_35_45_55_KNN %>%
  select(colnames(ml_UV_RNA_train))

```

```

C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen <- predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_35_45_55_test_C5)

```

```

ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction <- C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction)

```

```

##
## False True

```

```
## 11143 2641
```

```
UV_RNA_FAIMS_35_45_55_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
```

## RIPPER

```
RIPPER_UV_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_35_45_55_test_C5)
ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 10042 3742
```

```
UV_RNA_FAIMS_35_45_55_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
```

## Data setup UV RNA CV 40\_50\_60

```
UV_RNA_FAIMS_40_50_60_KNN <- read_delim("AWulf_060919_Ecoli_UV_S100_FAIMS_40_50_60_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(UV_RNA_FAIMS_40_50_60_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_40_50_60_KNN))
```

```
names(UV_RNA_FAIMS_40_50_60_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_40_50_60_KNN))
```

```
names(UV_RNA_FAIMS_40_50_60_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_40_50_60_KNN))
```

```
names(UV_RNA_FAIMS_40_50_60_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_40_50_60_KNN))
```

```
UV_RNA_FAIMS_40_50_60_KNN$XLinker <- "UV"
```

```
UV_RNA_FAIMS_40_50_60_KNN$Nucleotides <- "RNA"
```

```
UV_RNA_FAIMS_40_50_60_KNN$Sample <- "S100"
```

```
UV_RNA_FAIMS_40_50_60_KNN <- UV_RNA_FAIMS_40_50_60_KNN[UV_RNA_FAIMS_40_50_60_KNN$RNPx1_RNA !=
"none", ] %>% filter(target_decoy == "target")
```

```
UV_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_40_50_60_KNN$sequence))
```

```
UV_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_40_50_60_KNN$sequence))
```

```
UV_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_40_50_60_KNN$sequence))
```

```
UV_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_40_50_60_KNN$prepPep)
```

```
UV_RNA_FAIMS_40_50_60_KNN$prepPep <- gsub("\\\\.\\.*", "", UV_RNA_FAIMS_40_50_60_KNN$prepPep)
```

```
UV_RNA_FAIMS_40_50_60_KNN$Pepseq <- UV_RNA_FAIMS_40_50_60_KNN$prepPep
```

```
UV_RNA_FAIMS_40_50_60_KNN$Peplength <- nchar(UV_RNA_FAIMS_40_50_60_KNN$Pepseq, type = "chars")
```

```
UV_RNA_FAIMS_40_50_60_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_40_50_60_KNN$Pepseq,
paste("R|K|D|E", collapse = "|"))
```

```

UV_RNA_FAIMS_40_50_60_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_40_50_60_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_40_50_60_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_40_50_60_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_40_50_60_KNN$pI <- round(pI(UV_RNA_FAIMS_40_50_60_KNN$Pepseq), 2)
UV_RNA_FAIMS_40_50_60_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_40_50_60_KNN$Pepseq),
  2)
UV_RNA_FAIMS_40_50_60_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_40_50_60_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_40_50_60_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_40_50_60_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_40_50_60_KNN$mw <- round(mw(UV_RNA_FAIMS_40_50_60_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_40_50_60_KNN$cv <- c("35_45")

names(UV_RNA_FAIMS_40_50_60_KNN)[names(UV_RNA_FAIMS_40_50_60_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_40_50_60_KNN)[names(UV_RNA_FAIMS_40_50_60_KNN) == "precursor_error_(ppm)"] <- "precursor_error_ppm"
UV_RNA_FAIMS_40_50_60_KNN <- UV_RNA_FAIMS_40_50_60_KNN %>% filter(RNPxl_isPhospho ==
  "0")

UV_RNA_FAIMS_40_50_60_KNN_model <- UV_RNA_FAIMS_40_50_60_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_40_50_60_KNN_model_norm <- UV_RNA_FAIMS_40_50_60_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_40_50_60_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_40_50_60_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_40_50_60_KNN_model_norm[is.na(UV_RNA_FAIMS_40_50_60_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_40_50_60_KNN_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_40_50_60_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_40_50_60_KNN_model$Prediction <- UV_RNA_FAIMS_40_50_60_KNN_pred
table(UV_RNA_FAIMS_40_50_60_KNN_model$Prediction)

##
## False True
## 17100 565

UV_RNA_FAIMS_40_50_60_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_35_45_55_test_C5 <- UV_RNA_FAIMS_40_50_60_KNN %>%
  select(colnames(ml_UV_RNA_train))

```



```

C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_35_45_55)

ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction)

##
## False True
## 13581 4084
UV_RNA_FAIMS_40_50_60_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen

```

## RIPPER

```

RIPPER_UV_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_35_45_55_test_C5)
ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)

##
## False True
## 9891 7774
UV_RNA_FAIMS_40_50_60_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict

```

## Data setup UV RNA CV 30\_50\_70

```

UV_RNA_FAIMS_30_50_70_KNN <- read_delim("AWulf_050919_Ecoli_UV_S100_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(UV_RNA_FAIMS_30_50_70_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_30_50_70_KNN))
names(UV_RNA_FAIMS_30_50_70_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_30_50_70_KNN))
names(UV_RNA_FAIMS_30_50_70_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_30_50_70_KNN))
names(UV_RNA_FAIMS_30_50_70_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_30_50_70_KNN))
UV_RNA_FAIMS_30_50_70_KNN$XLinker <- "UV"
UV_RNA_FAIMS_30_50_70_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_30_50_70_KNN$Sample <- "S100"
UV_RNA_FAIMS_30_50_70_KNN <- UV_RNA_FAIMS_30_50_70_KNN[UV_RNA_FAIMS_30_50_70_KNN$RNPxl_RNA !=
"none", ] %>% filter(target_decoy == "target")
UV_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Carbamyl)", "", as.character(UV_RNA_FAIMS_30_50_70_KNN$sequenc
UV_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_30_50_70_KNN$prepe

```

```

UV_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_30_50_70_KNN$prepPep))
UV_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_30_50_70_KNN$prepPep)
UV_RNA_FAIMS_30_50_70_KNN$prepPep <- gsub("\\.\\.*", "", UV_RNA_FAIMS_30_50_70_KNN$prepPep)
UV_RNA_FAIMS_30_50_70_KNN$Pepseq <- UV_RNA_FAIMS_30_50_70_KNN$prepPep
UV_RNA_FAIMS_30_50_70_KNN$Peplength <- nchar(UV_RNA_FAIMS_30_50_70_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_30_50_70_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_30_50_70_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_30_50_70_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_30_50_70_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_30_50_70_KNN$pI <- round(pI(UV_RNA_FAIMS_30_50_70_KNN$Pepseq), 2)
UV_RNA_FAIMS_30_50_70_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_30_50_70_KNN$Pepseq),
  2)
UV_RNA_FAIMS_30_50_70_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_30_50_70_KNN$RNPxl_best_localization)
UV_RNA_FAIMS_30_50_70_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_30_50_70_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_30_50_70_KNN$mw <- round(mw(UV_RNA_FAIMS_30_50_70_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_30_50_70_KNN$cv <- c("35_45")

names(UV_RNA_FAIMS_30_50_70_KNN)[names(UV_RNA_FAIMS_30_50_70_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_30_50_70_KNN)[names(UV_RNA_FAIMS_30_50_70_KNN) == "precursor_error_(ppm)"] <- "precursor_error_ppm"
UV_RNA_FAIMS_30_50_70_KNN <- UV_RNA_FAIMS_30_50_70_KNN %>% filter(RNPxl_isPhospho ==
  "0")

UV_RNA_FAIMS_30_50_70_KNN_model <- UV_RNA_FAIMS_30_50_70_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_30_50_70_KNN_model_norm <- UV_RNA_FAIMS_30_50_70_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_30_50_70_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_30_50_70_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_30_50_70_KNN_model_norm[is.na(UV_RNA_FAIMS_30_50_70_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_30_50_70_KNN_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_30_50_70_KNN_model_norm,
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_30_50_70_KNN_model$Prediction <- UV_RNA_FAIMS_30_50_70_KNN_pred
table(UV_RNA_FAIMS_30_50_70_KNN_model$Prediction)

##
## False True
## 15698 377

UV_RNA_FAIMS_30_50_70_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_35_45_55_test_C5<-UV_RNA_FAIMS_30_50_70_KNN %>%
  select(colnames(ml_UV_RNA_train))

C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_35_45_55_test_C5)

ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction)

##
## False True
## 13068 3007

UV_RNA_FAIMS_30_50_70_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen

```

## RIPPER

```

RIPPER_UV_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_35_45_55_test_C5)
ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)

##
## False True
## 9757 6318

UV_RNA_FAIMS_30_50_70_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict

```

## Data setup UV RNA CV 45\_55\_65

```

UV_RNA_FAIMS_45_55_65_KNN <- read_delim("AWulf_060919_Ecoli_UV_S100_FAIMS_45_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(UV_RNA_FAIMS_45_55_65_KNN) <- gsub("\\.", "_", names(UV_RNA_FAIMS_45_55_65_KNN))
names(UV_RNA_FAIMS_45_55_65_KNN) <- gsub("\\:", "_", names(UV_RNA_FAIMS_45_55_65_KNN))
names(UV_RNA_FAIMS_45_55_65_KNN) <- gsub("\\ ", "_", names(UV_RNA_FAIMS_45_55_65_KNN))
names(UV_RNA_FAIMS_45_55_65_KNN) <- gsub("\\|", "_", names(UV_RNA_FAIMS_45_55_65_KNN))

```

```

UV_RNA_FAIMS_45_55_65_KNN$XLinker <- "UV"
UV_RNA_FAIMS_45_55_65_KNN$Nucleotides <- "RNA"
UV_RNA_FAIMS_45_55_65_KNN$Sample <- "S100"
UV_RNA_FAIMS_45_55_65_KNN <- UV_RNA_FAIMS_45_55_65_KNN[UV_RNA_FAIMS_45_55_65_KNN$RNPx1_RNA !=
  "none", ] %>% filter(target_decoy == "target")
UV_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Carbamy)", "", as.character(UV_RNA_FAIMS_45_55_65_KNN$sequ
UV_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Oxidation)", "", as.character(UV_RNA_FAIMS_45_55_65_KNN$prep
UV_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("(Phospho)", "", as.character(UV_RNA_FAIMS_45_55_65_KNN$prepP
UV_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("\\(|\\)", "", UV_RNA_FAIMS_45_55_65_KNN$prepPep)
UV_RNA_FAIMS_45_55_65_KNN$prepPep <- gsub("\\.|\\.", "", UV_RNA_FAIMS_45_55_65_KNN$prepPep)
UV_RNA_FAIMS_45_55_65_KNN$Pepseq <- UV_RNA_FAIMS_45_55_65_KNN$prepPep
UV_RNA_FAIMS_45_55_65_KNN$Peplength <- nchar(UV_RNA_FAIMS_45_55_65_KNN$Pepseq, type = "chars")
UV_RNA_FAIMS_45_55_65_KNN$aa_class_charged <- str_count(UV_RNA_FAIMS_45_55_65_KNN$Pepseq,
  paste("R|K|D|E", collapse = "|"))
UV_RNA_FAIMS_45_55_65_KNN$aa_class_polar <- str_count(UV_RNA_FAIMS_45_55_65_KNN$Pepseq,
  paste("Q|N|H|S|T|Y|C|W", collapse = "|"))
UV_RNA_FAIMS_45_55_65_KNN$aa_class_hydrophobic <- str_count(UV_RNA_FAIMS_45_55_65_KNN$Pepseq,
  paste("A|I|L|M|F|V|P|G", collapse = "|"))
UV_RNA_FAIMS_45_55_65_KNN$pI <- round(pI(UV_RNA_FAIMS_45_55_65_KNN$Pepseq), 2)
UV_RNA_FAIMS_45_55_65_KNN$aIndex <- round(aIndex(UV_RNA_FAIMS_45_55_65_KNN$Pepseq),
  2)
UV_RNA_FAIMS_45_55_65_KNN$XL_aa <- gsub("[:A-Z:]", "", UV_RNA_FAIMS_45_55_65_KNN$RNPx1_best_localization
UV_RNA_FAIMS_45_55_65_KNN$Hydro <- round(hydrophobicity(UV_RNA_FAIMS_45_55_65_KNN$Pepseq,
  scale = "KyteDoolittle"), 2)
UV_RNA_FAIMS_45_55_65_KNN$mw <- round(mw(UV_RNA_FAIMS_45_55_65_KNN$Pepseq, monoisotopic = F),
  2)
UV_RNA_FAIMS_45_55_65_KNN$cv <- c("35_45")

names(UV_RNA_FAIMS_45_55_65_KNN)[names(UV_RNA_FAIMS_45_55_65_KNN) == "precursor_m/z"] <- "precursor"
names(UV_RNA_FAIMS_45_55_65_KNN)[names(UV_RNA_FAIMS_45_55_65_KNN) == "precursor_error_(ppm)"] <- "pre
UV_RNA_FAIMS_45_55_65_KNN <- UV_RNA_FAIMS_45_55_65_KNN %>% filter(RNPx1_isPhospho ==
  "0")

UV_RNA_FAIMS_45_55_65_KNN_model <- UV_RNA_FAIMS_45_55_65_KNN %>% select(colnames(ml_UV_RNA_train_knn))

```

## Normalizing

```

UV_RNA_FAIMS_45_55_65_KNN_model_norm <- UV_RNA_FAIMS_45_55_65_KNN_model %>% select_if(is.numeric)
UV_RNA_FAIMS_45_55_65_KNN_model_norm <- as.data.frame(lapply(UV_RNA_FAIMS_45_55_65_KNN_model_norm,
  normalize))
UV_RNA_FAIMS_45_55_65_KNN_model_norm[is.na(UV_RNA_FAIMS_45_55_65_KNN_model_norm)] <- 0

```

## Classifying UV RNA FAIMS data

### KNN5

```

UV_RNA_FAIMS_45_55_65_KNN_pred <- knn(train = ml_UV_RNA_train_knn, test = UV_RNA_FAIMS_45_55_65_KNN_mod
  cl = ml_UV_RNA_train_knn_labels$Curated, k = 5)
UV_RNA_FAIMS_45_55_65_KNN_model$Prediction <- UV_RNA_FAIMS_45_55_65_KNN_pred
table(UV_RNA_FAIMS_45_55_65_KNN_model$Prediction)

```

```

##
## False True

```

```
## 17251 510
```

```
UV_RNA_FAIMS_45_55_65_KNN$Prediction_KNN <- UV_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_UV_RNA_FAIMS_35_45_55_test_C5<-UV_RNA_FAIMS_45_55_65_KNN %>%
  select(colnames(ml_UV_RNA_train))
```

```
C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen<-predict(C5_model_UV_RNA_boost_pen, ml_UV_RNA_FAIMS_35_45_55_test_C5)
```

```
ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction)
```

```
##
```

```
## False True
```

```
## 13470 4291
```

```
UV_RNA_FAIMS_45_55_65_KNN$Prediction_C5<-C5_UV_RNA_FAIMS_35_45_55_predict_boost_pen
```

## RIPPER

```
RIPPER_UV_RNA_FAIMS_35_45_55_predict <- predict(RIPPER_model_UV_RNA, ml_UV_RNA_FAIMS_35_45_55_test_C5)
```

```
ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
table(ml_UV_RNA_FAIMS_35_45_55_test_C5$Prediction_RIPPER)
```

```
##
```

```
## False True
```

```
## 8894 8867
```

```
UV_RNA_FAIMS_45_55_65_KNN$Prediction_RIPPER <- RIPPER_UV_RNA_FAIMS_35_45_55_predict
```

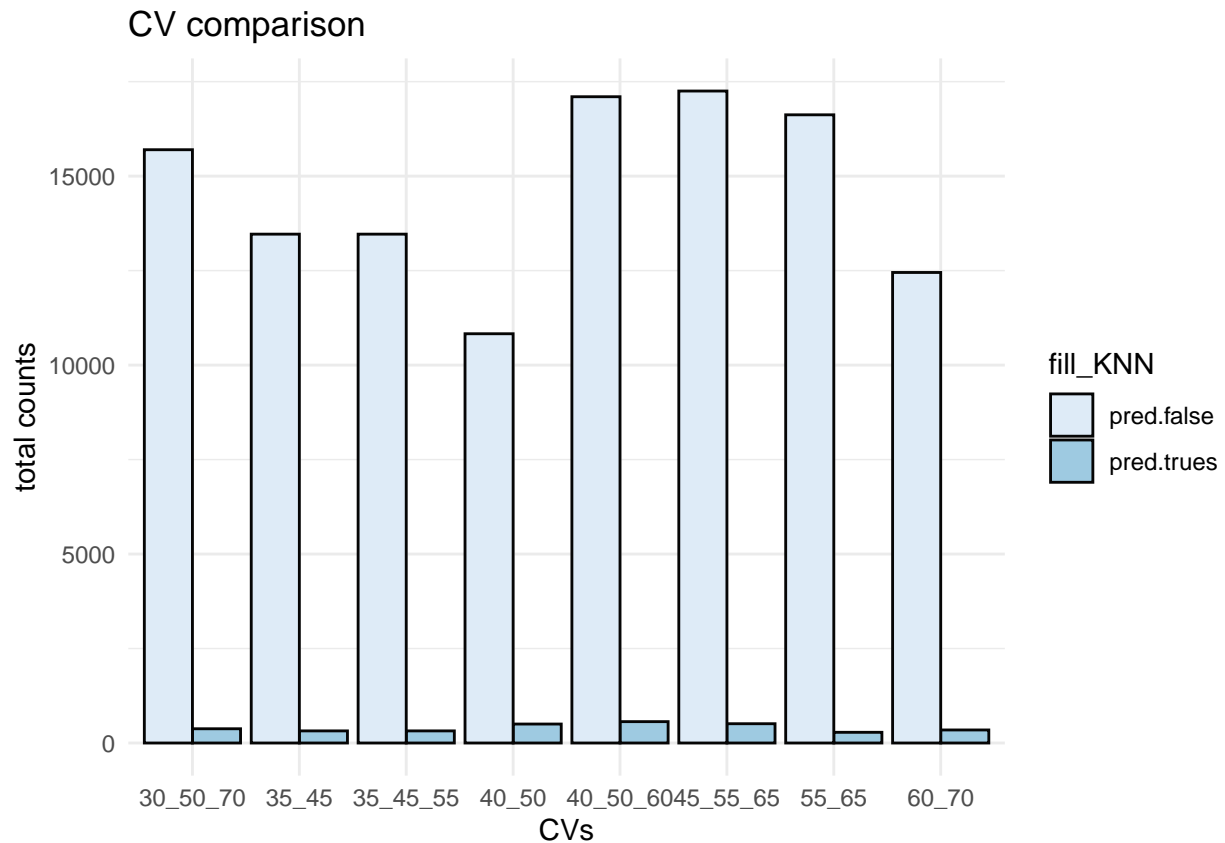
## Comparing the most conservative model KNN

```
a<-as.data.frame(table(UV_RNA_FAIMS_35_45_KNN$Prediction_KNN))
b<-as.data.frame(table(UV_RNA_FAIMS_40_50_KNN$Prediction_KNN))
c<-as.data.frame(table(UV_RNA_FAIMS_55_65_KNN$Prediction_KNN))
d<-as.data.frame(table(UV_RNA_FAIMS_60_70_KNN$Prediction_KNN))
e<-as.data.frame(table(UV_RNA_FAIMS_35_45_55_KNN$Prediction_KNN))
f<-as.data.frame(table(UV_RNA_FAIMS_30_50_70_KNN$Prediction_KNN))
g<-as.data.frame(table(UV_RNA_FAIMS_40_50_60_KNN$Prediction_KNN))
h<-as.data.frame(table(UV_RNA_FAIMS_45_55_65_KNN$Prediction_KNN))
```

```
f_KNN<-c("double", "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")
```

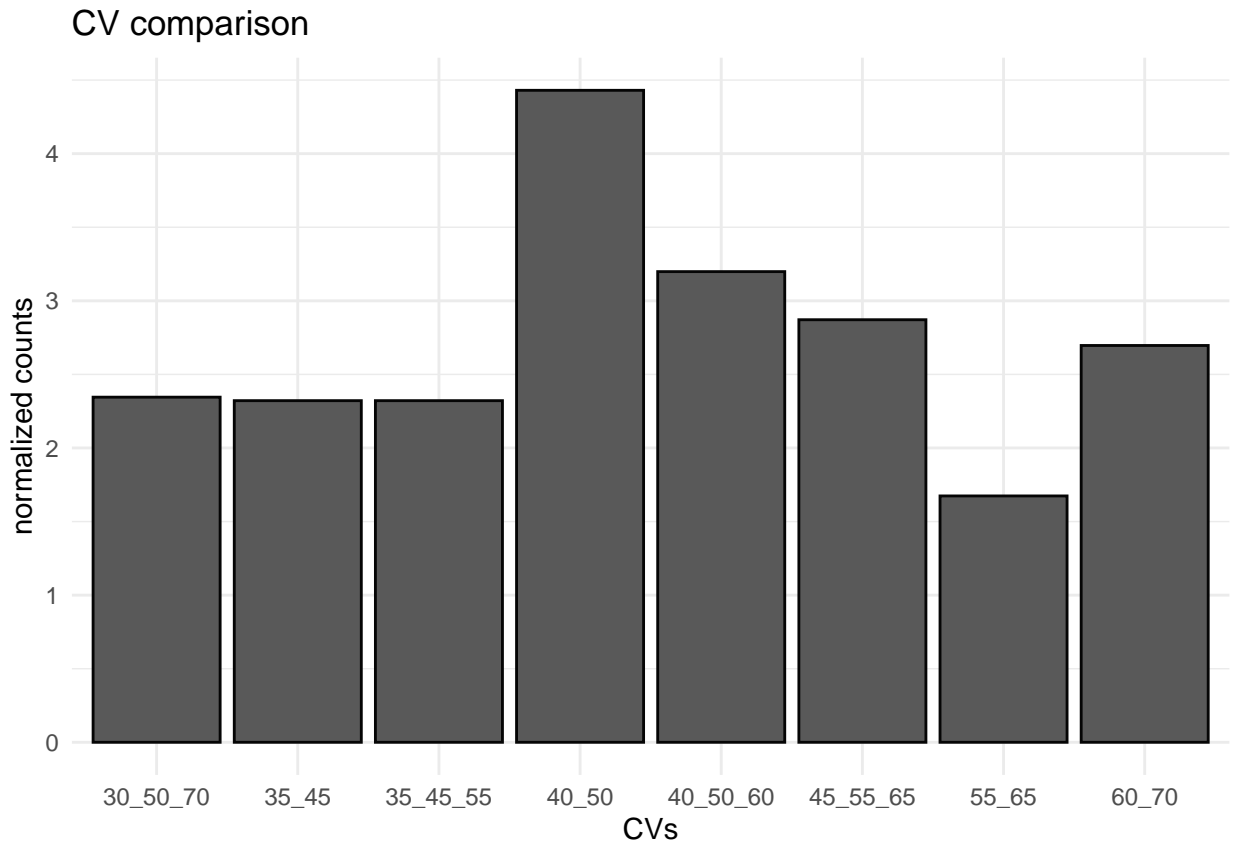
```
df_KNN_comparison2<-filter(df_KNN_comparison, fill_KNN!="pred.norm")
ggplot(df_KNN_comparison1, aes(x=x_KNN, y=y_KNN, fill = fill_KNN))+
  geom_bar(stat = "identity", position=position_dodge(), bandwidth =1, bins = 3, color="gray2")+
  scale_fill_brewer(palette = "Blues")+
  ggtitle("CV comparison")+
  theme_minimal()+
  ylab("total counts")+
  xlab("CVs")
```

## Warning: Ignoring unknown parameters: bandwidth, bins



```
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()
```

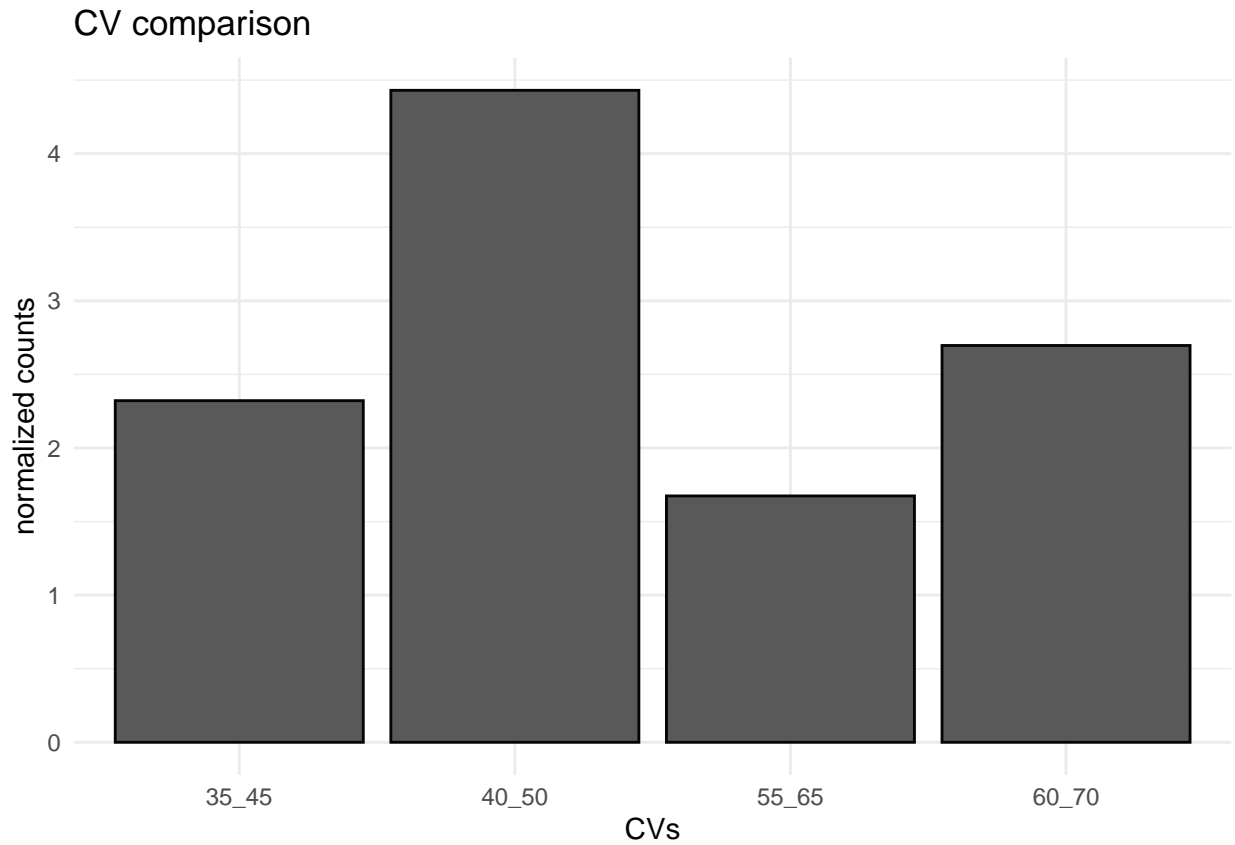
## Warning: Ignoring unknown parameters: bandwidth, bins



```
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
```

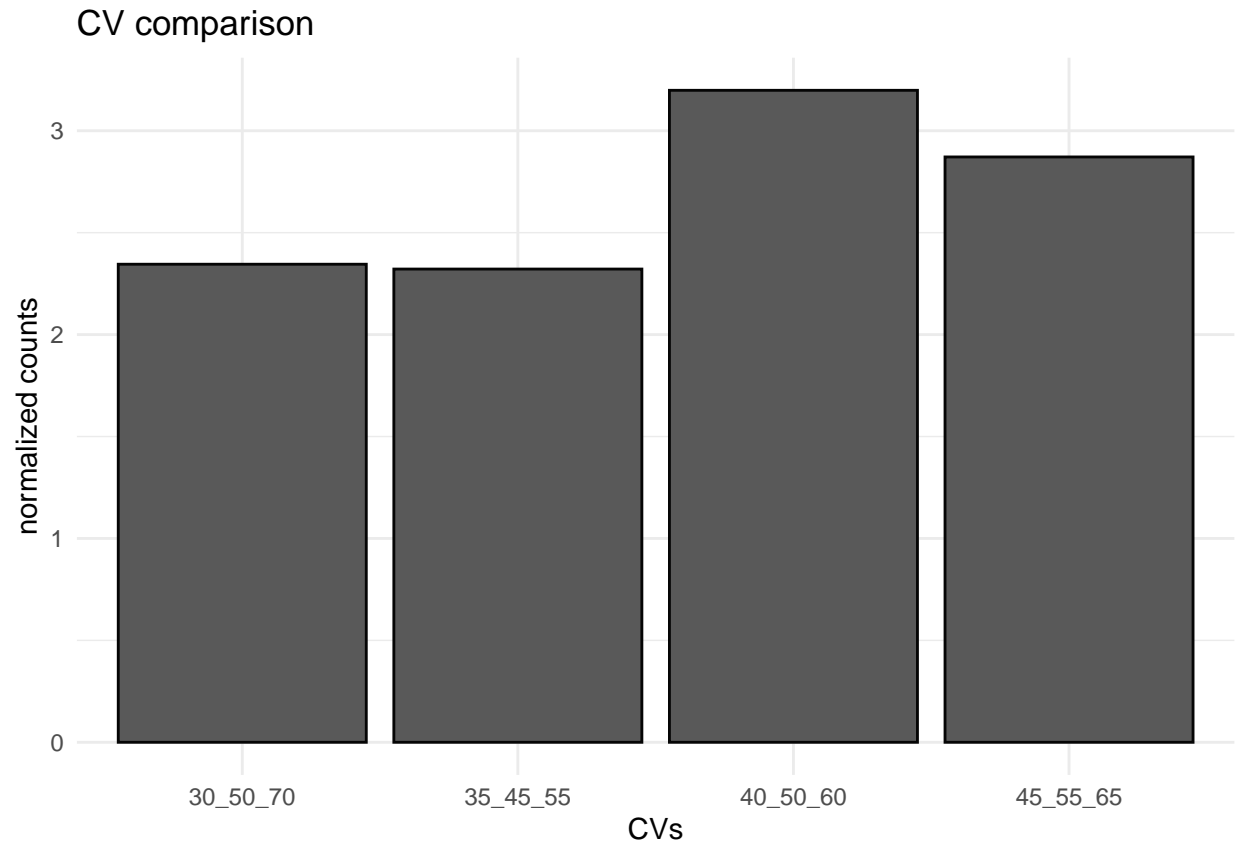


```
df_KNN_comparison4<-filter(df_KNN_comparison2, f_KNN=="triple")

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_UV_RNA_FAIMS_total_predicted_trues<- rbind(UV_RNA_FAIMS_35_45_KNN, UV_RNA_FAIMS_40_50_KNN, UV_RNA_1
  filter(Prediction_KNN == "True")
write_excel_csv(S100_UV_RNA_FAIMS_total_predicted_trues, "S30_UV_RNA_FAIMS_total_KNN_prediction.csv")
```