ChemModLab_1-17_bugfixes

Jeremy Ash

January 17, 2017

There were several bugs that needed to be addressed so that Jackie could start using the package. Many of these have resulted from alterations to the code that I had made so that the package has the structure recommended by Wickham's R Packages. Some of these bugs went undetected because I was not installing the package in a new R environment (some older versions of chemmodlab functions were still in my environment).

I have nearly finished reading R Packages now, so the package should now be structured in a way that will be acceptable to CRAN (though more documentation still needs to be written). I am following Wickham's suggestions precisely even though they can be a bit fastidious at times. This is because I want to minimize the need for manual intervention from the people at CRAN. This should increase the chances of the package being accepted.

Bugs - recently fixed

- Not all of the functions that should be accessible to users (eg. CombineSplits) were being exported. I needed to read up on NAMESPACES. Users can now access all the functions we want them to use. All other functions have been kept internal.
- The chemmodlab object constructor now assigns the class "chemmodlab" not "ChemModLab". plot.chemmodlab expects the former.
- Some functions used for fitting Pls models did not have the package namespace specified explicitly (eg. `pls::lda`). Those packages were not imported by the functions that used them and so those functions could not be found. This has been fixed.
- Each time a plot is created in ChemModLab, the user's graphical parameters need to be changed. Wickham advises minimizing changes to a user's global environment. I am saving the user's grpahical parameters and resetting them. I have found a way to suppress the large number of warnings that resulted from trying to reset graphical parameters that can't be set.
- The legends for the plot.chemmodlab plots are now being placed correctly. They were at times being placed in the middle of the plot.

After these bug fixes, the package should at least be usable.

Bugs that I am currently working on

- I had not tested ChemModLab thoroughly on a dataset with a continuous response. Now, when I try to run `ModelTrain` with the `USArrests` dataset, PLS is not run due to this error: `replacement has length zero`. I have not tracked down this bug yet.
- Anytime I plot a chemmodlab object, there are some concerning issues with the plots:

- There are some lines that are plotting points corresponding to 0 selected compounds, or 1 larger than the max number of selected compounds (Figure 1). I have noticed that this was also happening in the plots that I was originally provided. This does not make sense to me, as everywhere in the plotting code the x values are specified as the sequence: `1:x.max`. Also, the accumulation curves for the continuous models are no longer extending all the way to the max number of selected compounds.

- We are no longer generating all of the descriptor set accumulation curve plots (the plots in which the accumulation curves for each descriptor set are being compared for a particular model). I only see plots for the LAR and PLS models now. This is a recent development. (See below)

```
yfilein <- read.csv("C:/Users/Vestige/Dropbox/ChemModLab/example_run/AID_364.csv")
xfilein <- read.csv("C:/Users/Vestige/Dropbox/ChemModLab/example_run/BurdenNumbers.csv")
data <- cbind(yfilein, xfilein[,-1])
head(data[1:6])
```

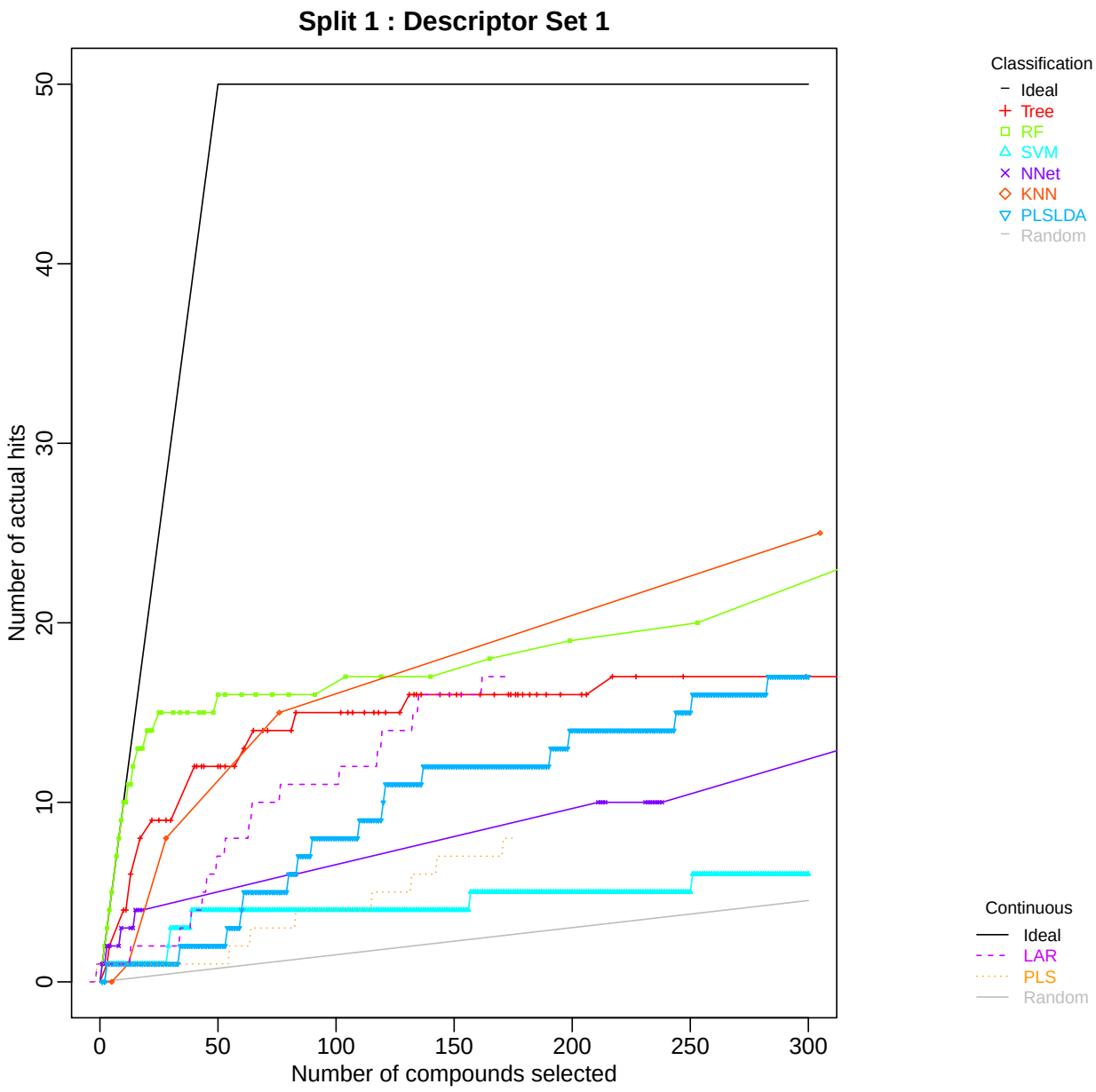
```
##        CID Outcome WBN_GC_L_0.25 WBN_GC_H_0.25 WBN_GC_L_0.50 WBN_GC_H_0.50
## 1  5388992       1      -2.40010       1.98339      -2.52864       2.50835
## 2  5388983       1      -2.40010       1.98240      -2.52868       2.50398
## 3   663143       1      -2.41650       1.32890      -2.53910       2.05778
## 4    10607       1      -2.38337       2.17677      -2.52643       2.33232
## 5  5388972       1      -2.29039       1.97468      -2.41743       2.46177
## 6 11970251       1      -2.29039       2.22488      -2.41748       2.56161
```

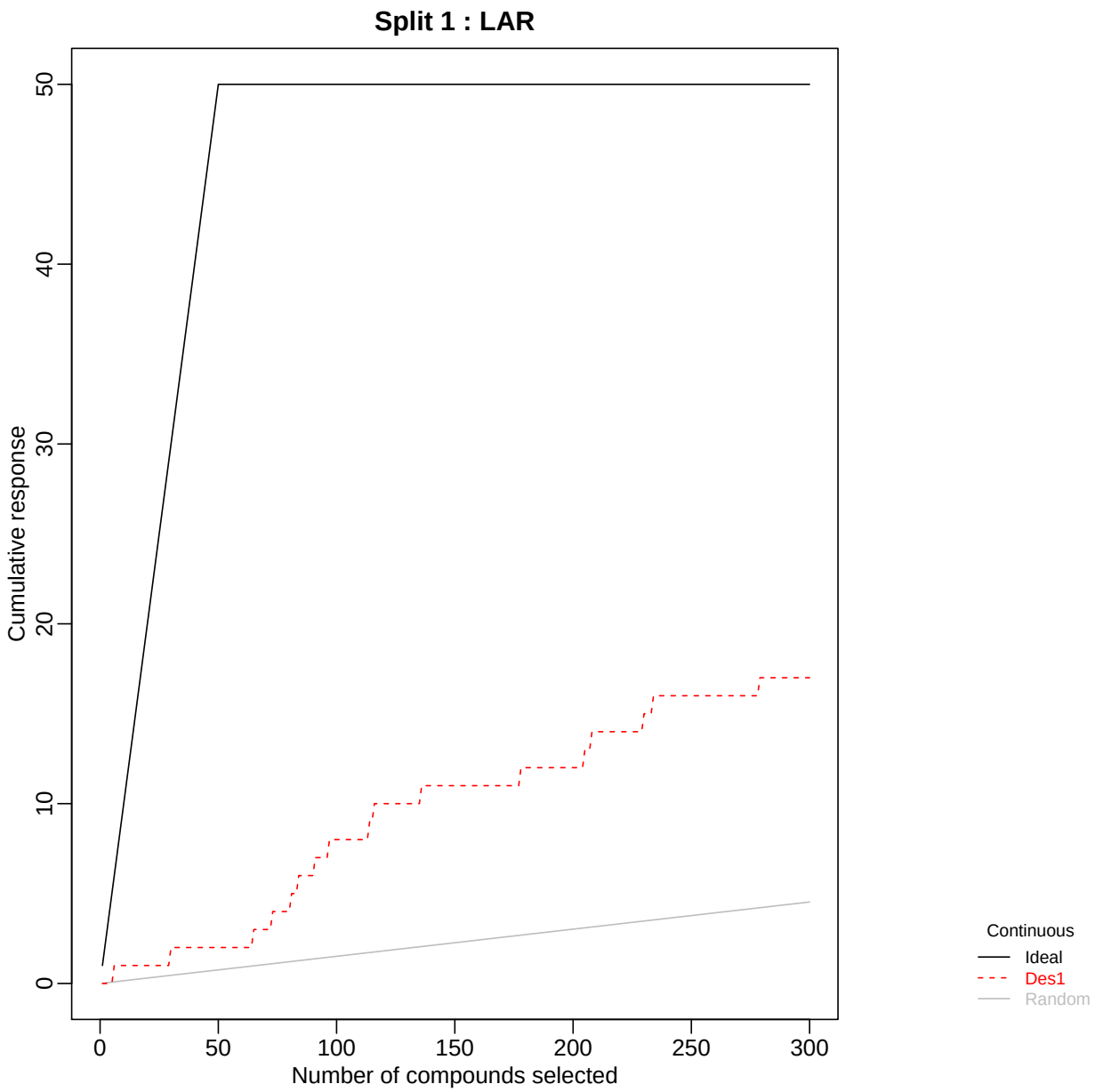
```
cml.big <- ModelTrain(data, idcol = 1, ycol = 2,
                  nsplits = 1)
```

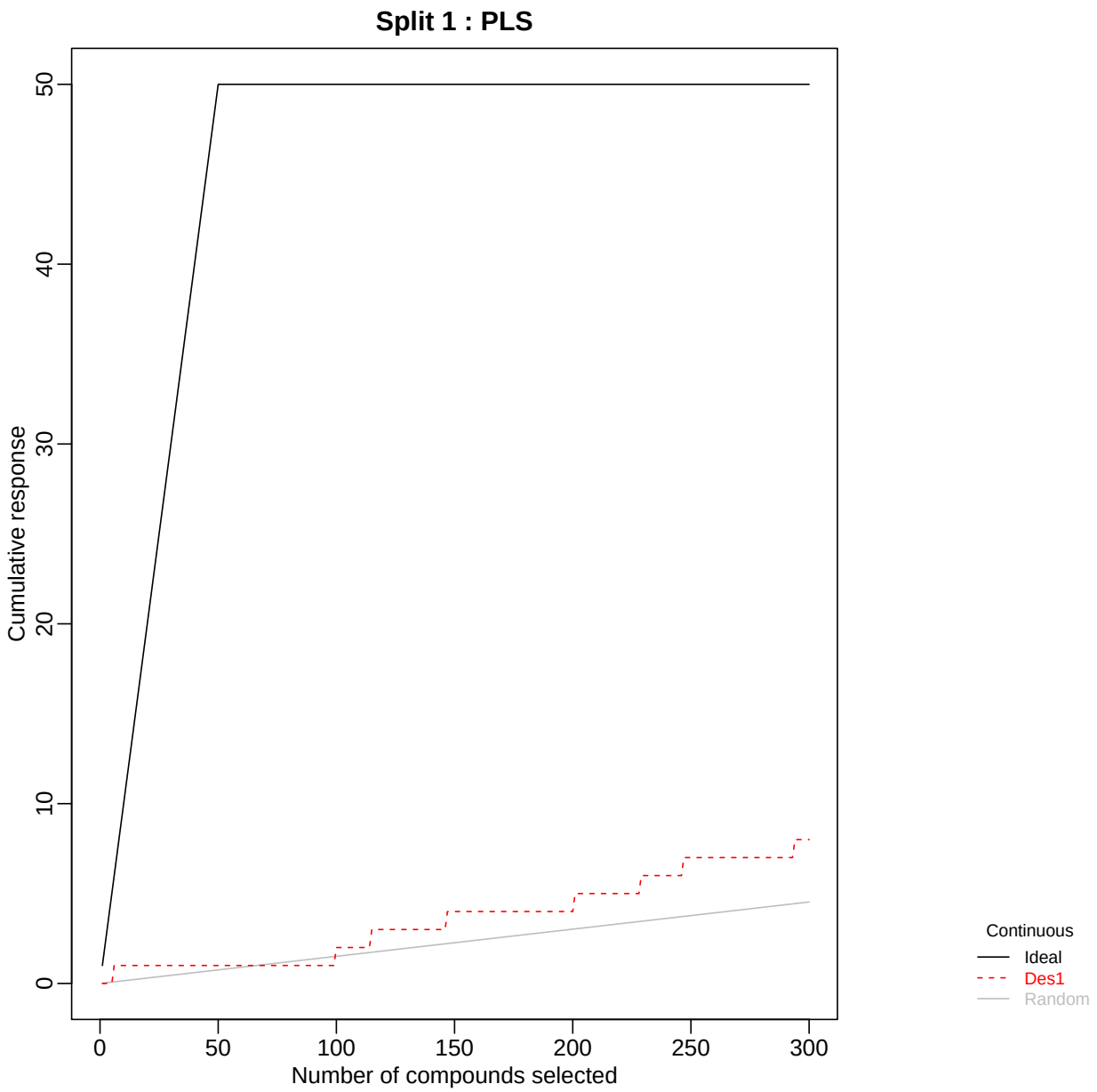
```
## Begining Analysis for Split:  1 and Descriptor Set:  Descriptor Set 1
## Number of Descriptors:  24
## Responses:  3311
## Starting Seed:  11111
## Number of CV folds:  10
##
## Tree ---------------------------------------
## Real time used:  3.08
##  CPU time used:  3.03
##
##
## Forest -------------------------------------
## Real time used:  8.09
##  CPU time used:  8.06
##
##
## SVM ----------------------------------------
## Real time used:  103.13
##  CPU time used:  103.1
##
##
## NNet ---------------------------------------
## Real time used:  2.72
##  CPU time used:  2.68
##
##
## KNN ----------------------------------------
## Real time used:  0.57
##  CPU time used:  0.53
##
##
## PLSLDA -------------------------------------
## Real time used:  2.73
```

```
##  CPU time used:  2.67
##
##
## LARs ----------------------------------------
## Real time used:  0.38
##  CPU time used:  0.33
##
##
## PLS ----------------------------------------
## Real time used:  1.08
##  CPU time used:  1.02
##
##
## Ending Analysis for Split:  1 and Descriptor Set:  Descriptor Set 1
```

```r
plot(cml.big)
```


Split 1 : Descriptor Set 1
Number of actual hits
Number of compounds selected
0
10
20
30
40
50
0
50
100
150
200
250
300
Classification
Ideal
Tree
RF
SVM
NNet
KNN
PLSLDA
Random
Continuous
Ideal
LAR
PLS
Random


Split 1 : LAR
Cumulative response
0
10
20
30
40
50
0
50
100
150
200
250
300
Number of compounds selected
Continuous
Ideal
Des1
Random



- The MCS plot is not formatted correctly when there are only few models being plotted. It is small, hard to read, and not positioned correctly. Some of these formatting issues seems to be a recent developments. I believe that the graphical parameters are not being set and reset properly. I also may need to make changes to the plot margins.

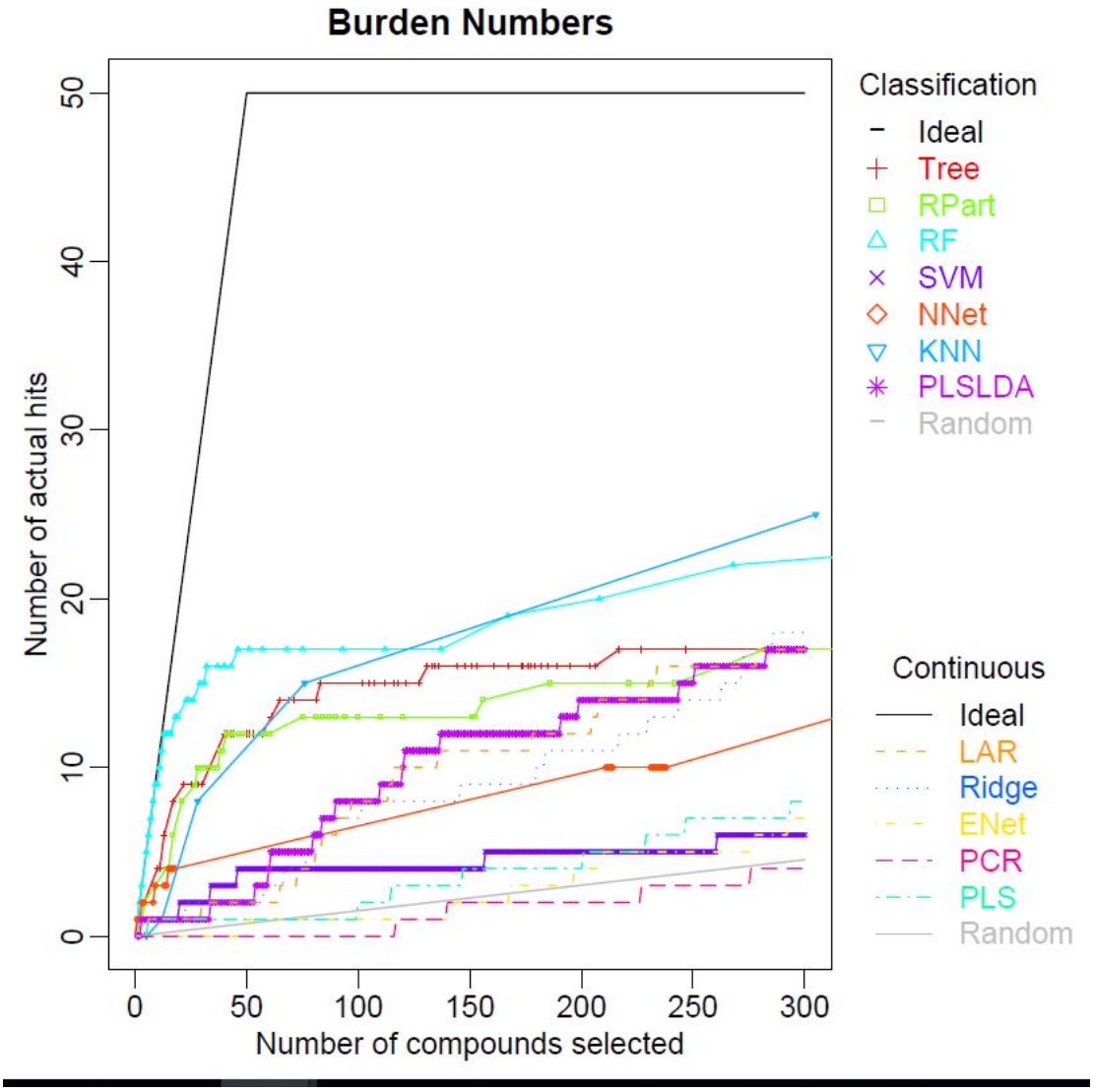


Figure 1: Original accumulation plot

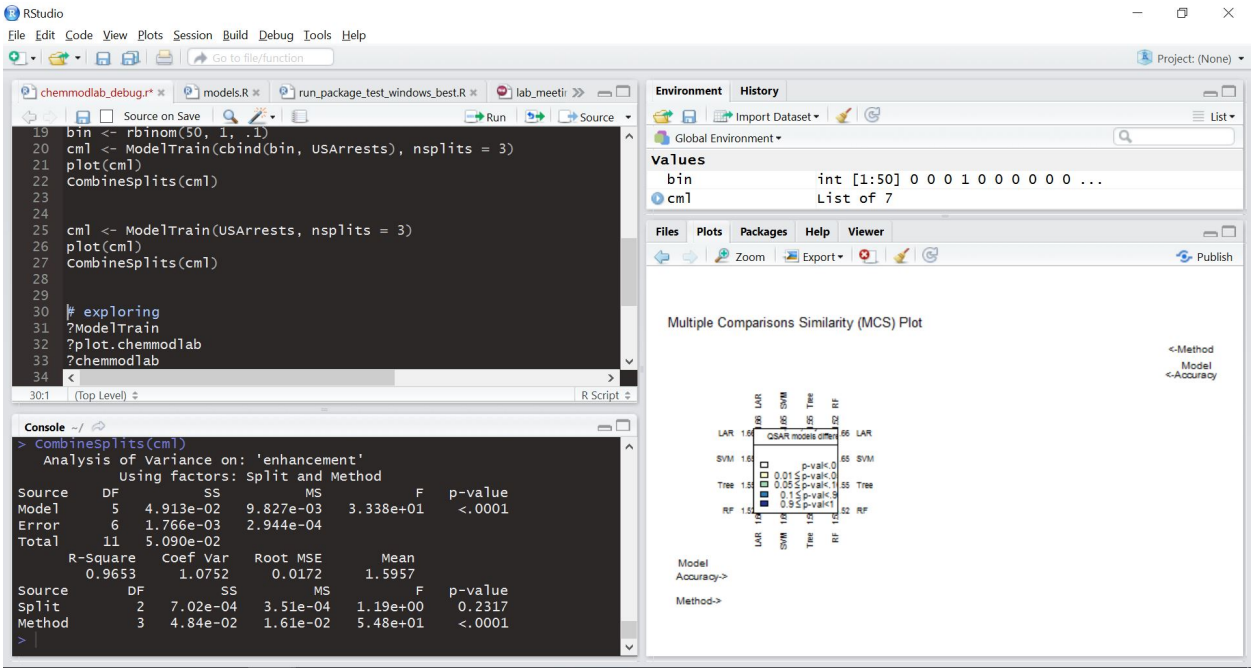


Figure 2: With a small number of models the plot is not visible

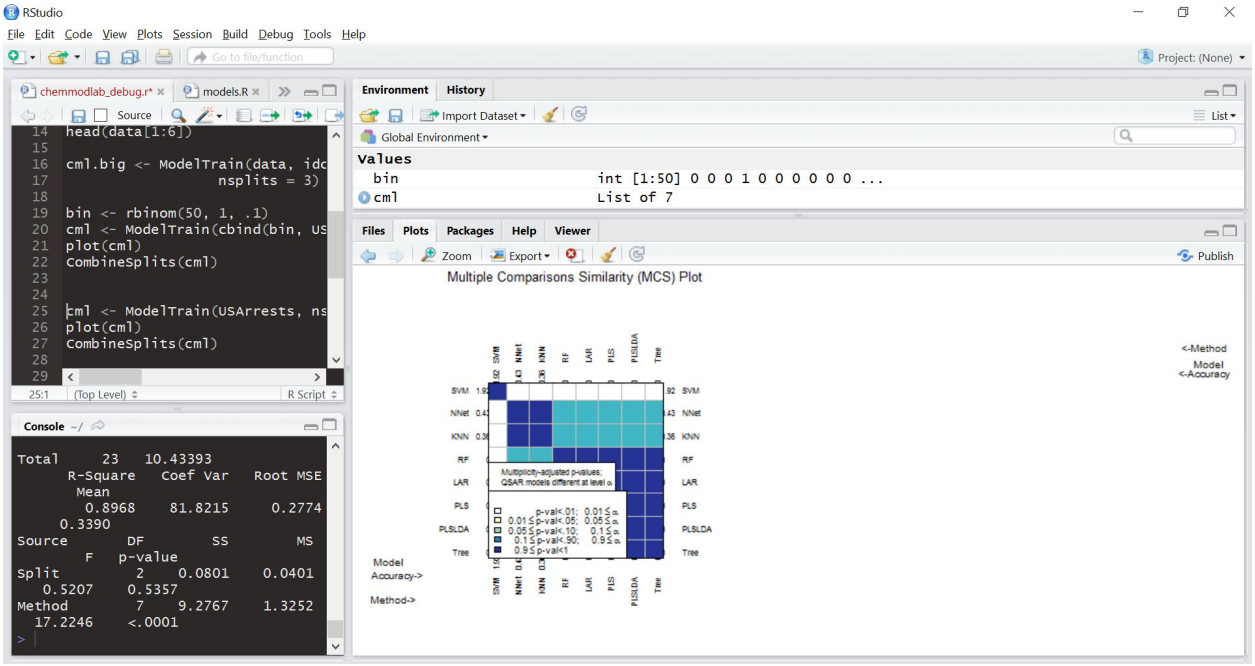


Figure 3: Making the plotting window larger improves the visibility