chem-bla-ics

The CDK/Metabolomics/ Chemometrics Unconference results

Egon Willighagen

Published April 7, 2008

Citation

Willighagen, E. (2008). The CDK/Metabolomics/Chemometrics Unconference results. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/f2ajz-tmr63

Keywords

Cdk, Defense, Phd, Metabolomics, Cheminf

Abstract

As announced earlier , Miguel, Velitchka, Christoph and I held a small CDK/Metabolomics/ Chemometrics unconference. We started late, and did not have an evening program, resulting in not overly much results. However, we did do molecular chemometrics.

Copyright

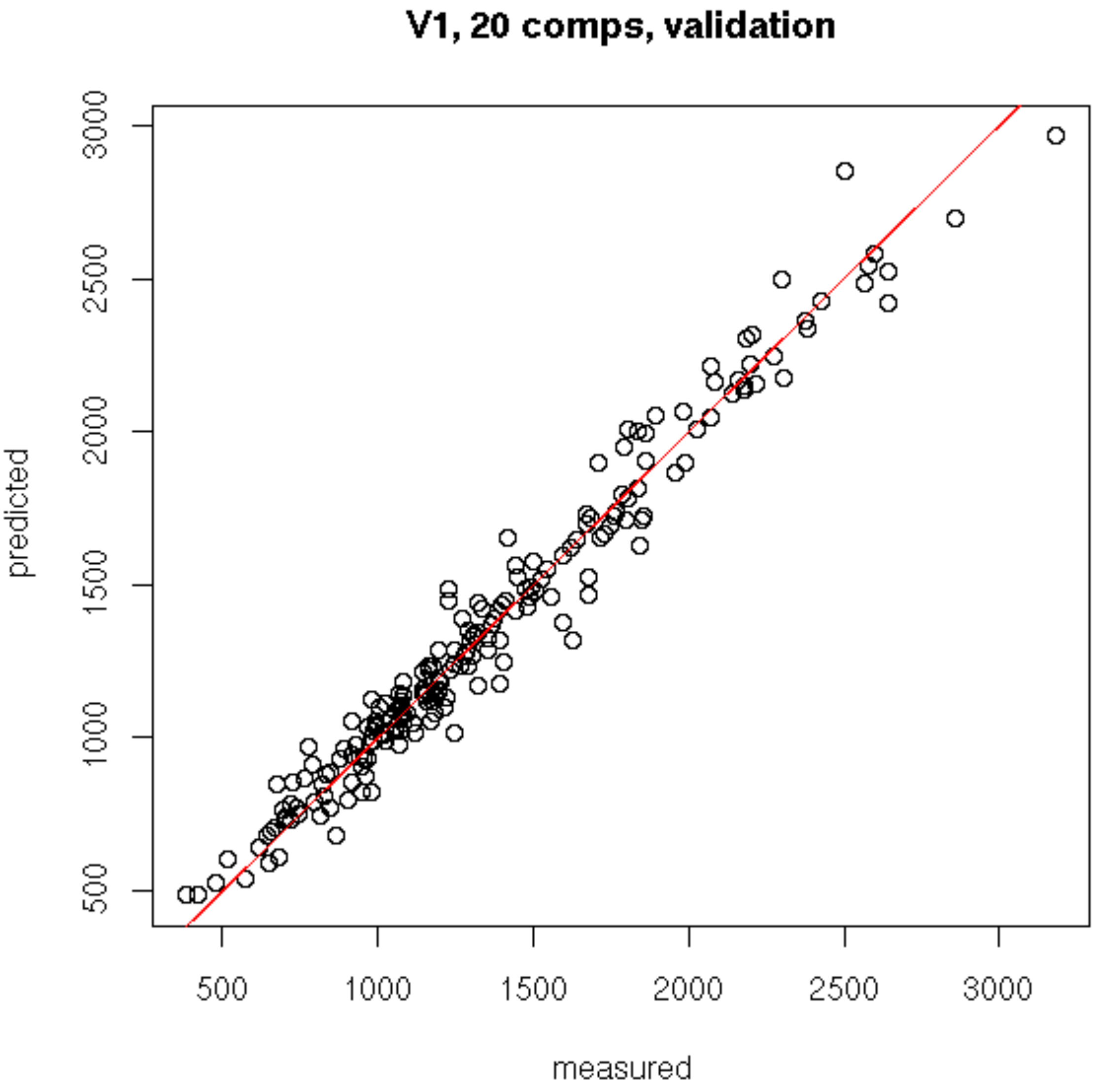
As announced earlier , Miguel, Velitchka, Christoph and I held a small CDK/Metabolomics/Chemometrics unconference. We started late, and did not have an evening program, resulting in not overly much results. However, we did do *molecular chemometrics*.

We used the R statistics software together with Rajarshi's rcdk package (an R wrapper around the CDK library) and Ron's (my PhD supervisor) PLS package (see this paper), to predict retention indices for a number of metabolites.

We ended up with this R script:

```
library("rJava")
library("rcdk")
library("pls")
mols = load.molecules("data_cdk.sdf")
selection = get.desc.names()
selection = selection[-
which(selection=="org.openscience.cdk.qsar.descriptors.molecular.AminoAcidCountDescriptor
x = eval.desc(mols, selection, verbose=TRUE)
x2 = x[,apply(x, 2, function(a) {all(!is.na(a))})]
y = read.table("data_cdk_RI")
input = data.frame(x2, y)
pls.model = plsr(V1 ~ ., 50, data=input, validation="CV")
summary(pls.model)
plot(RMSEP(pls.model))
plot(pls.model, ncomp=20)
abline(0,1, col="red")
plot(pls.model, "loadings", comps=1:2)
savehistory("finalHistory.R")
```

The `AminoAcidCountDescriptor` threw us a `NullPointerException` and there were a few NAs in the resulting matrix. The CV results were not so good as Velitchka's best models, but still a good start:



No variable selection; 200 objects, 190 variables.

Questions:

- Can we do this in Bioclipse2 too?
- Can we improve the default CDK descriptor parameters to maximize the column count?
- Rajarshi, what would be involved to write some wrapper code for atomic descriptors for rcdk?