

CaGe App

Computer-aided approach for Genetic
prediction

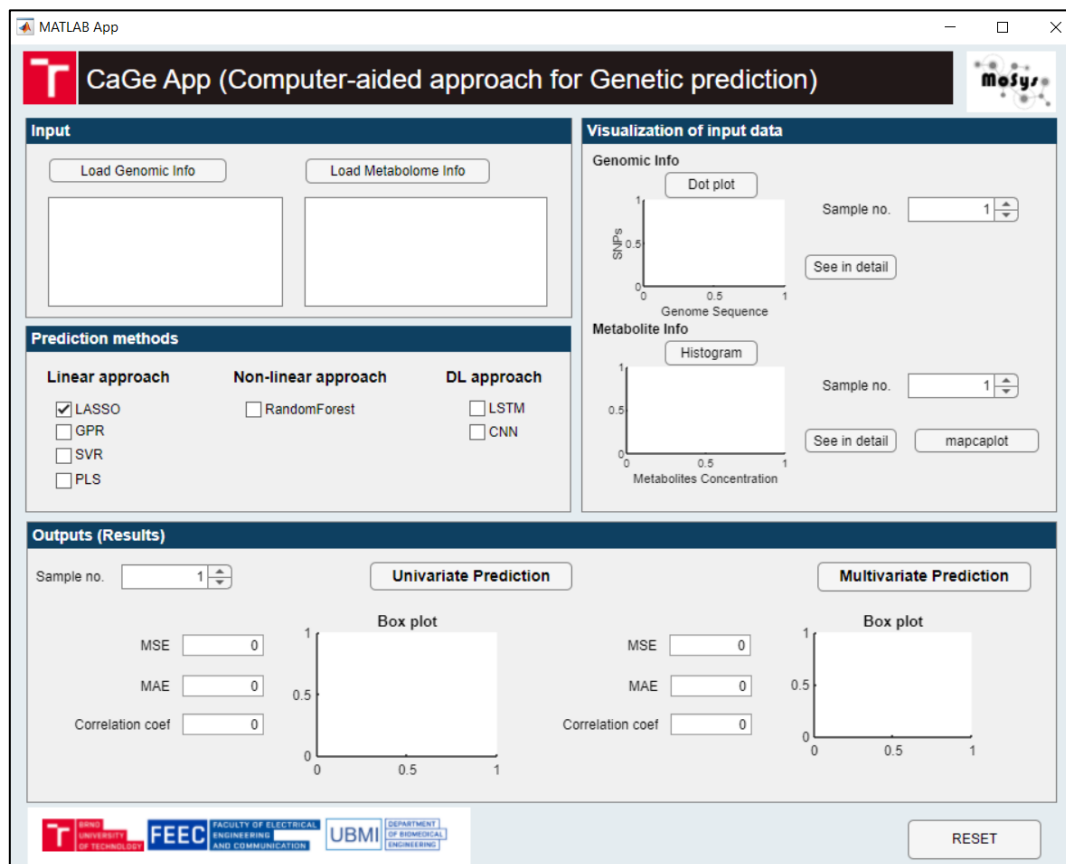


Department of Biomedical Engineering, FEEC, BUT
[Molecular Systems Biology University of Vienna](#)
31.1.2023

..

CaGe App

Computer-aided approach for Genetic prediction



Content

Introduction	2
1. Input	3
2. Visualization	5
3. Prediction methods	6
4. Outputs (Results)	9

Introduction

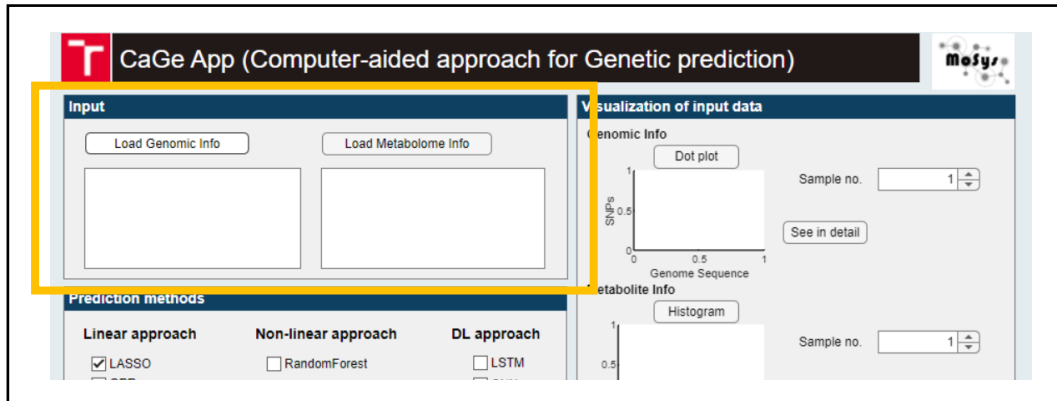
Currently advanced sequencing methods have upgraded molecular biology research into so-called post-genomic era. Thus, the research is focused on understanding functional relationships between individual genes causing the final phenotype. This completely new software implemented in Matlab2022b App Designer view has a great potential for opening new paths in understanding predictions tools in genomic prediction challenges.

We developed user-friendly app for uploading genomic (in form single nucleotide polymorphisms (SNPs)) and metabolomic data, their visualization and subsequently prediction analysis using different methods such as linear, non-linear and deep learning which is implemented in Matlab2022b.

This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, which is financed from the OP RDE.

1. Input

The first section in app is to upload data. App expects load genomic dataset in form SNPs where 0 represents undetected SNP and 1 represents detected SNP. Load Metabolome Info needs to upload metabolite information relating to individual samples associated with SNPs data.



Example data:

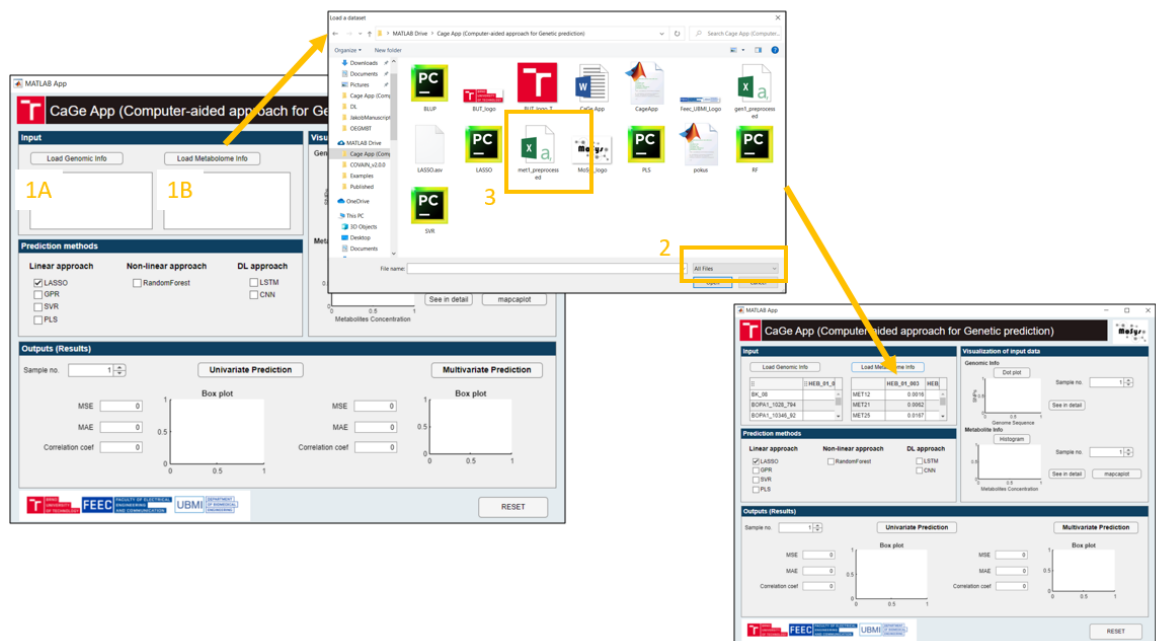
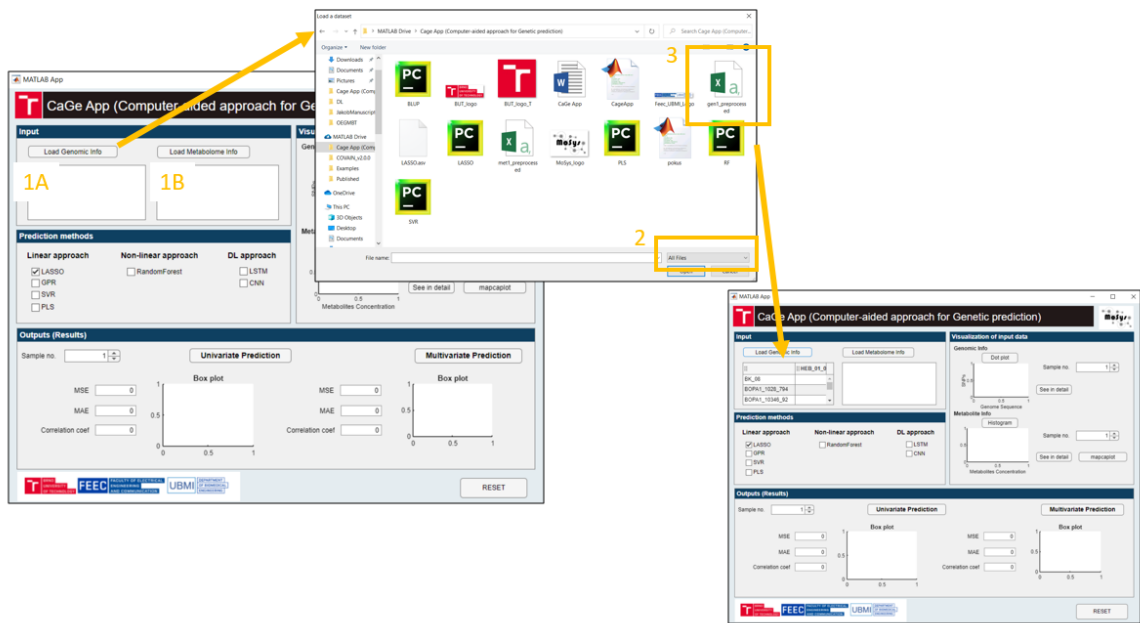
The data on which the software was built is taken from the study by Gemmer et al. [1].

Genomic Info

Genomic information was obtained using a 50k Illumina Infinium iSelect 9k SNP chip. Genotyping resulted in 1429 SNP lines on 37 chromosomes. Static information has a value of 0 if there were no changes compared to the reference Barke. A value of 1 represents heterozygous changes and a value of 2 represents homozygous changes.

Metabolome Info

Metabolomic information refers to the above mentioned HEB-25 population. Metabolomic information was obtained using GC-MS and processed using MassHunter Qualitative Analysis software. This data includes 158 metabolite concentrations of 1419 lines. The NAM method was used to place exotic alleles of 25 barley plants into the genetic background of the German spring barley variety Barke. This presents in the NAM population HEB-25 (which represents our tested dataset).



2. Visualization

Section called visualization of input data is used to better understand the individual input samples. Genomic information is possible to shown as dot plot where the user can see the changing distribution of individual SNPs in a particular sample. The sample number can be changed by spinner button which is in right site of this section. Detail preview is also available.

Regarding metabolite information, the user can see distribution of metabolite concentration. Also, for metabolite information the user can switch between different samples using the spinner button and see the detail preview. Moreover, we incorporated button called '*mapcaplot*' (<https://www.mathworks.com/help/bioinfo/ref/mapcaplot.html>) for metabolite information. This button call `mapcaplot()` [2] matlab function which:

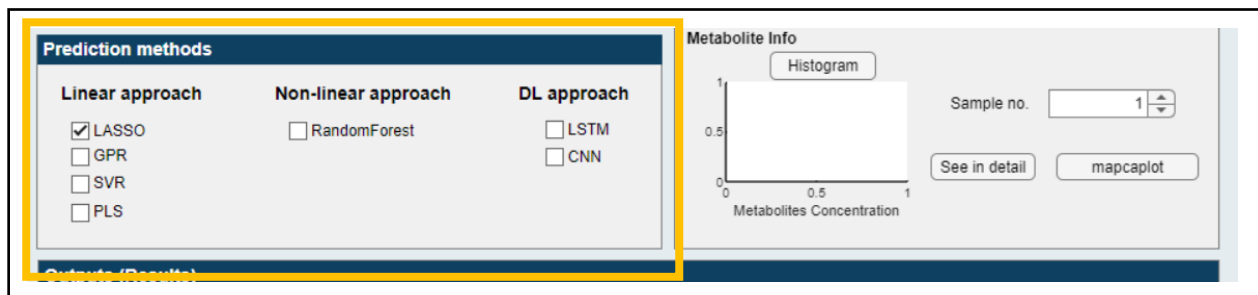
creates 2-D scatter plots of principal components of data. Once you plot the principal components, you can:

- Select principal components for the x and y axes from the drop-down list below each scatter plot.
- Click a data point to display its label.
- Select a subset of data points by dragging a box around them. Points in the selected region and the corresponding points in the other axes are then highlighted. The labels of the selected data points appear in the list box.
- Select a label in the list box to highlight the corresponding data point in the plot. Press and hold **Ctrl** or **Shift** to select multiple data points.
- Export the gene labels and indices to the MATLAB® workspace.

The screenshot displays the CaGe App interface, titled "CaGe App (Computer-aided approach for Genetic prediction)". The interface is divided into several sections:

- Input:** Contains two buttons, "Load Genomic Info" and "Load Metabolome Info". Below these are two data tables. The first table, "HEB_01_0", lists genomic information with columns for gene names (BK_08, BOPA1_1028_794, BOPA1_10346_92) and a value column. The second table, "HEB_01_003", lists metabolite concentrations with columns for metabolite names (MET12, MET21, MET25) and their respective values (0.0016, 0.0062, 0.0167).
- Prediction methods:** This section is divided into three columns: "Linear approach" (with checkboxes for LASSO, GPR, SVR, and PLS), "Non-linear approach" (with a checkbox for RandomForest), and "DL approach" (with checkboxes for LSTM and CNN).
- Visualization of input data:** This section is highlighted with a yellow border and contains two sub-sections:
 - Genomic Info:** Features a "Dot plot" button, a "Sample no." spinner set to 1, and a "See in detail" button. The plot area shows a graph of "Genome Sequence" on the x-axis (0 to 1) and "SNP" on the y-axis (0 to 1).
 - Metabolite Info:** Features a "Histogram" button, a "Sample no." spinner set to 1, a "See in detail" button, and a "mapcaplot" button. The plot area shows a graph of "Metabolites Concentration" on the x-axis (0 to 1) and "SNP" on the y-axis (0 to 1).

3. Prediction methods



The major part of this App is section called Prediction methods. This section contains a total of seven different prediction methods that can be used for genomic prediction. The section is very user friendly as the user only has to choose the method they want to use. However, hidden in the background of this section are settings that appear to be optimal for genomic prediction.

Linear approaches:

- linear methods include 5-fold cross validation to train models.

LASSO

LASSO prediction uses predict students' exam scores using lasso and the elastic net method. Thus, predict exam scores for the test data. Compare the predicted values to the actual exam grades using a reference line:

```
app.Pred_Val = XTest*coef + coef0;
```

GPR

Fit a Gaussian process regression (GPR) model

`gprMdl = fitrgp(Tbl, ResponseVarName)` returns a Gaussian process regression (GPR) model trained using the sample data in Tbl, where ResponseVarName is the name of the response variable in Tbl.

(<https://www.mathworks.com/help/stats/fitrgp.html>)

SVR

Fit a support vector machine regression (SVR) model

`fitrsvm` trains or cross-validates a support vector machine (SVM) regression model on a low-through moderate-dimensional predictor data set. `fitrsvm` supports mapping the predictor data using kernel functions, and supports SMO, ISDA, or $L1$ soft-margin minimization via quadratic programming for objective-function minimization.

To train a linear SVM regression model on a high-dimensional data set, that is, data sets that include many predictor variables, use `fitrlinear` instead.

To train an SVM model for binary classification, see `fitcsvm` for low- through moderate-dimensional predictor data sets, or `fitclinear` for high-dimensional data sets.

(<https://www.mathworks.com/help/stats/fitrsvm.html>)

PLS

Partial least-squares (PLS) regression

`plsregress(X,Y,ncomp)` returns the predictor and response loadings XL and YL, respectively, for a partial least-squares (PLS) regression of the responses in matrix Y on the predictors in matrix X, using ncomp PLS components.

(<https://www.mathworks.com/help/stats/plsregress.html>)

Non-linear approaches:

RF

Random forest implemented by **oobQuantilePredict** - Quantile predictions for out-of-bag observations from bag of regression trees

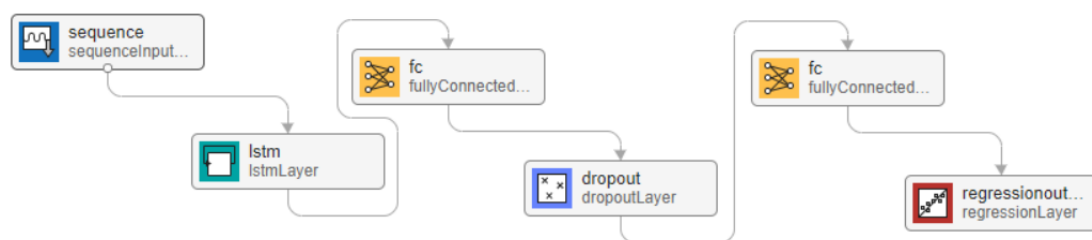
(<https://www.mathworks.com/help/stats/treebagger.oobquantilepredict.html>)

Deep Learning approaches:

LSTM

Long Short-Term Memory (LSTM) is an artificial neural network used in artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points but also entire sequences of data (in our case – genome represented by SNPs sequence). This property makes LSTM networks ideal for data processing and prediction.

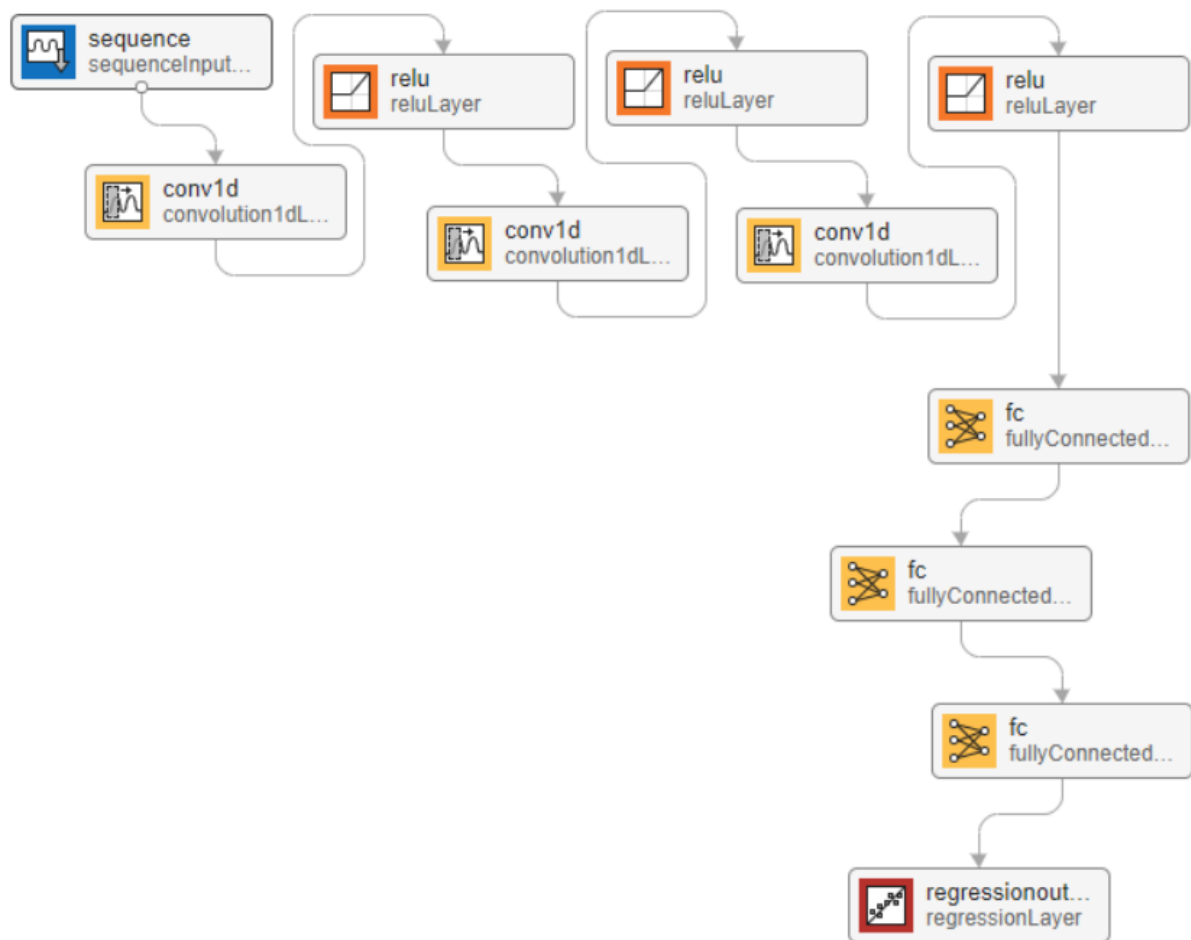
In CaGe software we implemented LSTM using architecture:



CNN

Convolutional Neural Network (CNN) is a class of artificial neural network most commonly used to prediction analysis. CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN), based on a shared weight architecture of convolutional kernels or filters that shift along input features to provide translation-equivalent responses known as feature maps.

In CaGe software we implemented CNN using architecture:

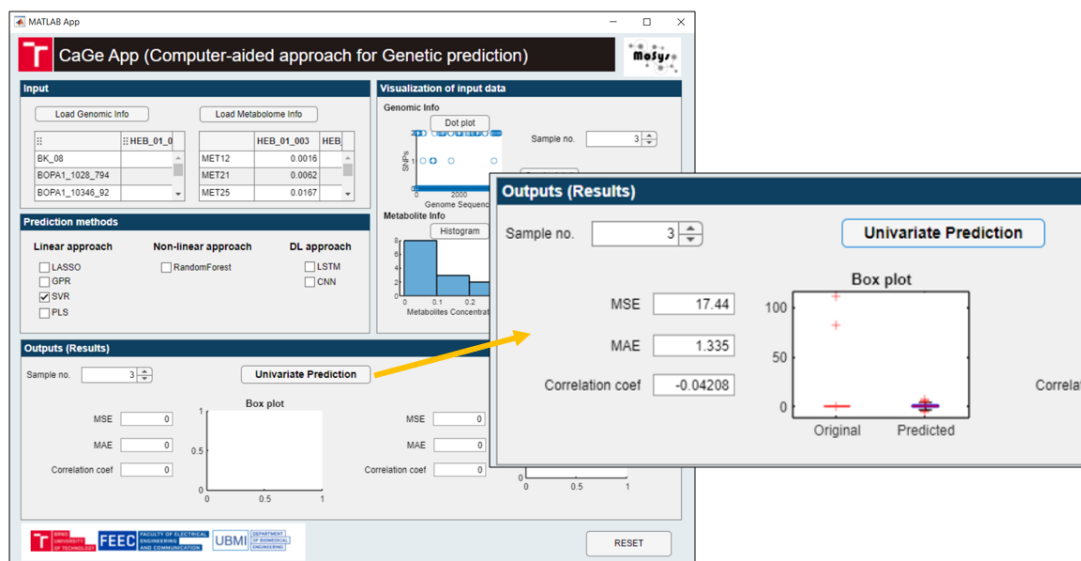


4. Outputs (Results)

The results are divided into two parts rely on genomic prediction analysis. In general, the prediction analysis is evaluated using mean square error (MSE), mean absolute error (MAE) and correlation coefficient. However, since all three of these evaluation parameters can be very misleading, the CaGe software also allows to visualize a box plot of predicted and original values.

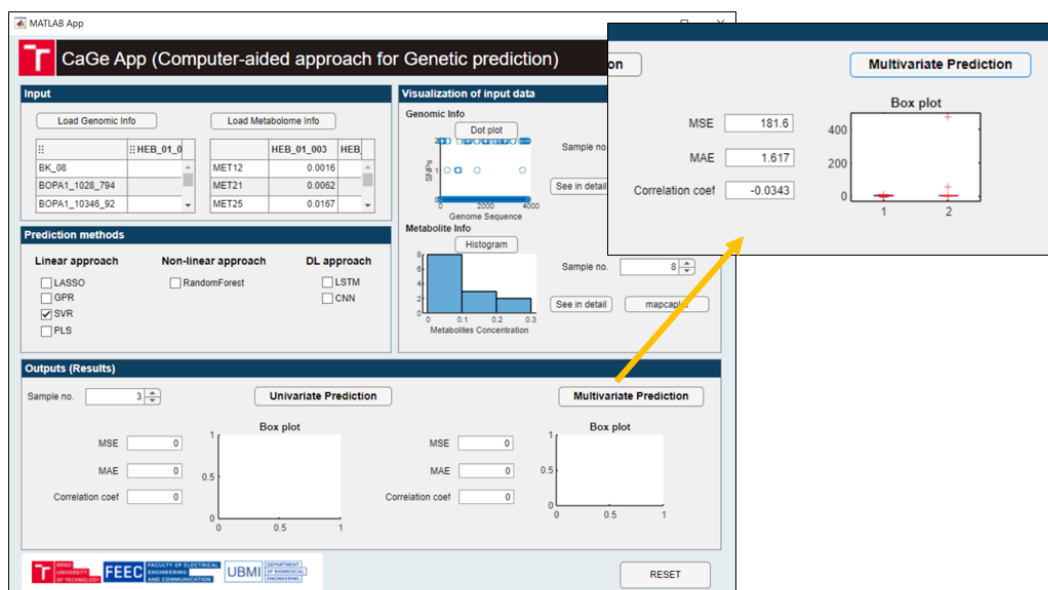
Univariate prediction

Univariate prediction analysis focuses on single metabolites. The selected metabolite we want to observe, we can select using the slider button which is on the left. Then the user just wait for the results of our prediction analysis:



Multivariate prediction

Multivariate prediction analysis includes all metabolites. The use is analogous as in univariate prediction.



At the end, the user has two options, he can switch to another method in section three of the *Prediction methods* and again by pressing the buttons in section four *Outputs/(Results)* to activate the new prediction. Or he can use the RESET button to restore the application to its default settings.

LITERATURE

- [1] M. R. Gemmer et al.: Can metabolic prediction be an alternative to genomic prediction in barley? PloS one, 2020, DOI: 10.1371/journal.pone.0234052
- [2] DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686s.