CaGe App

CaGe App

Department of Biomedical Engineering, FEEC, BUT

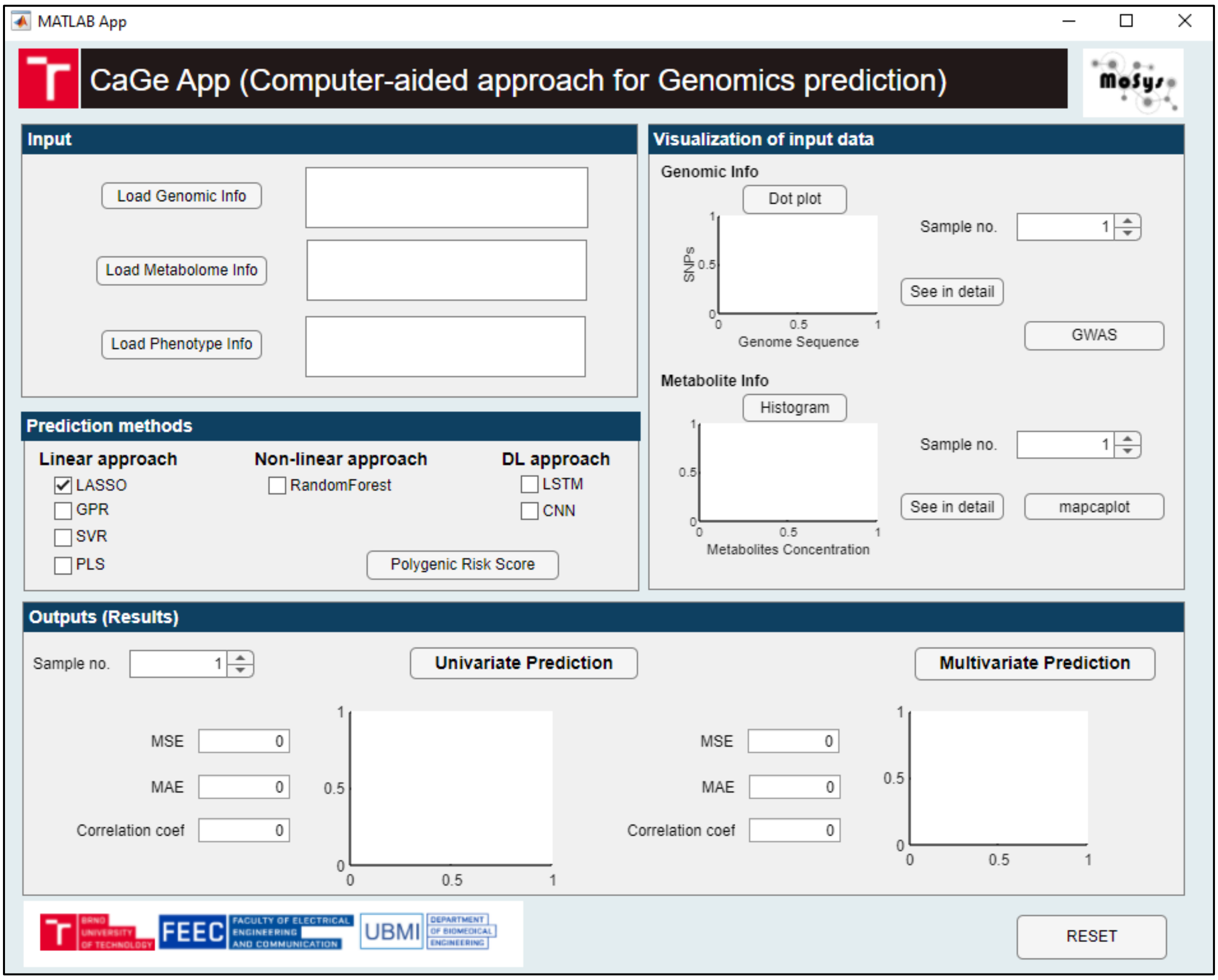
Molecular Systems Biology University of Vienna

3.5.2024

..

CaGe App

Computer-aided approach for genomics prediction



Linear approach: ☑ LASSO □ GPR □ SVR □ PLS

Non-linear approach: □ RandomForest

DL approach: □ LSTM □ CNN

Sample no. 1

MSE 0 MAE 0 Correlation coef 0

MSE 0 MAE 0 Correlation coef 0

Content

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.

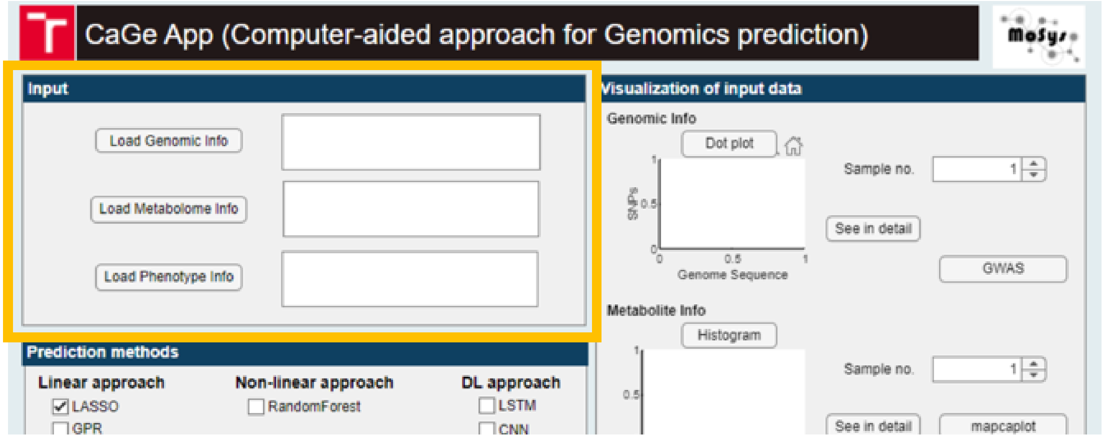
This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, which is financed from the OP RDE.

1. Input

The first section of the app is for uploading data. The app expects genomic datasets in the form of SNPs, where '0' represents an undetected SNP and '1' represents a detected SNP.

The 'Load Metabolome Info' section is used to upload metabolite information related to individual samples associated with the SNP data.

Last but not least, there is the possibility to upload Phenotype Info. For plants, this could include relative growth rate or yield. For mice or humans, it could be a specific attribute of the disease or the BMI index, etc.



☑ LASSO
□ GPR
□ RandomForest
□ LSTM
□ CNN

Example data:

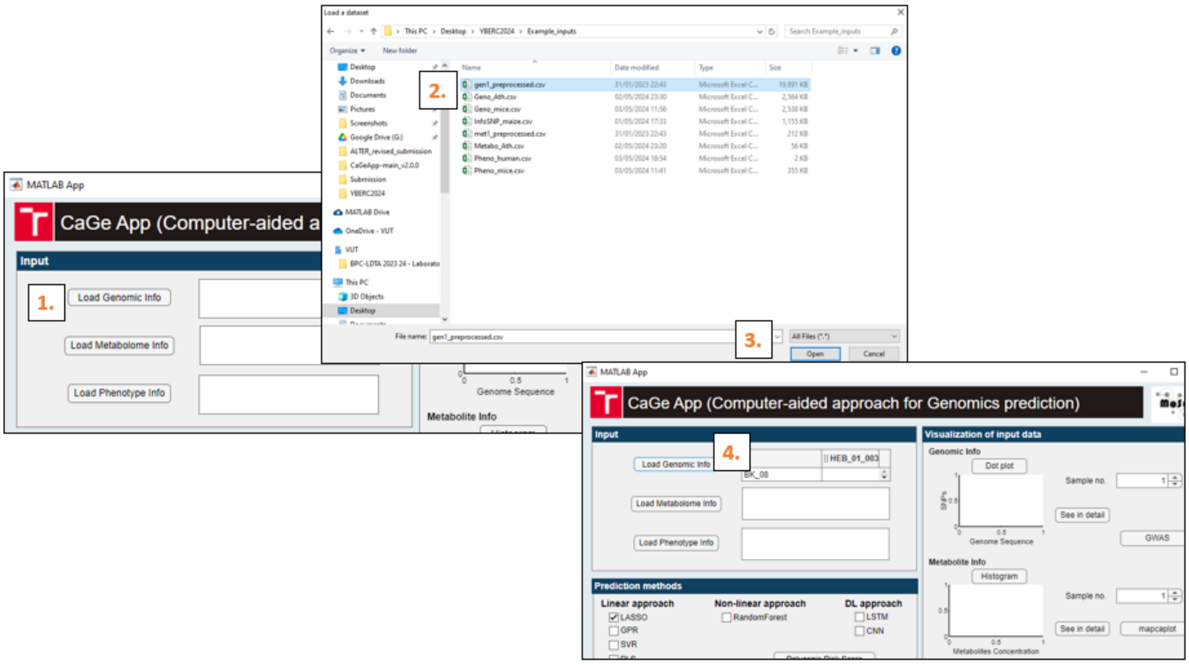
The data upon which the software is built was extracted from the study conducted by Gemmer et al. [1]. Additional test elements encompass maize [1][2][3], *Arabidopsis thaliana* [4], as well as available mouse datasets or human data – the mouse data were utilized similarly to the BGLR testing [5]. Regarding human data, we leveraged datasets sourced from Kaggle (https://www.kaggle.com/), specifically the 'SNP dataset for GWAS'.

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).



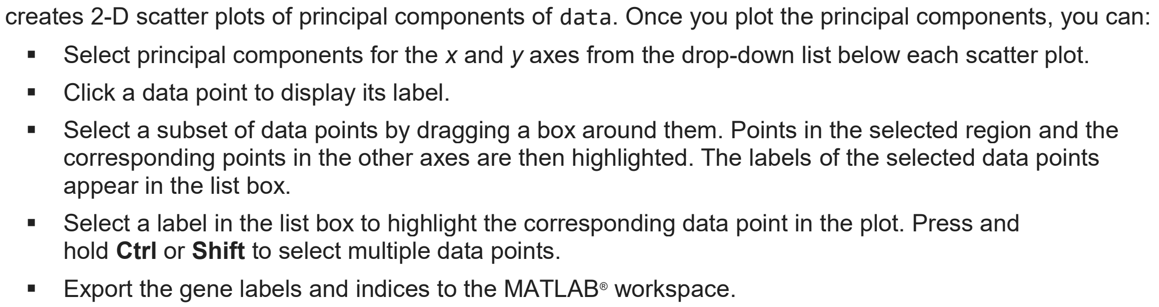
Name	Date modified	Type	Size
gen1_preprocessed.csv	31/01/2023 22:43	Microsoft Excel C...	19,891 KB
Geno_Ath.csv	02/05/2024 23:30	Microsoft Excel C...	2,564 KB
Geno_mice.csv	03/05/2024 11:56	Microsoft Excel C...	2,538 KB
InfoSNP_maize.csv	01/05/2024 17:33	Microsoft Excel C...	1,155 KB
met1_preprocessed.csv	31/01/2023 22:43	Microsoft Excel C...	212 KB
Metabo_Ath.csv	02/05/2024 23:20	Microsoft Excel C...	56 KB
Pheno_human.csv	03/05/2024 18:54	Microsoft Excel C...	2 KB
Pheno_mice.csv	03/05/2024 11:41	Microsoft Excel C...	355 KB

File name: gen1_preprocessed.csv

2. Visualization

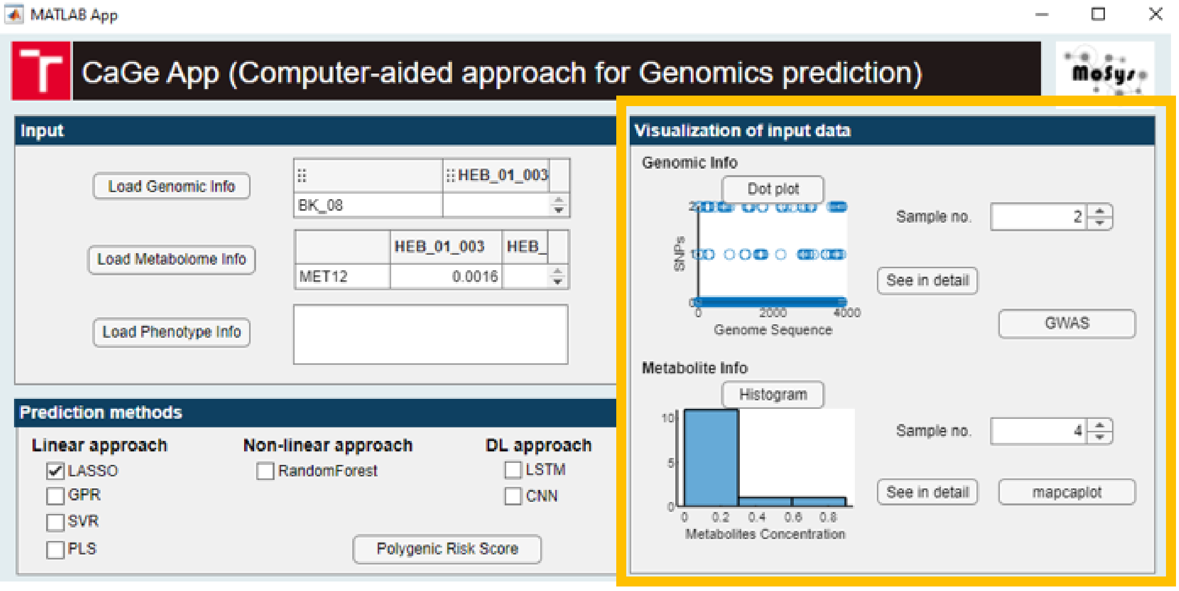
Section called visualization of input data is used to better understand the individual input samples. Genomic information is possible to shown as dot plot where the user can see the changing distribution of individual SNPs in a particular sample. The sample number can be changed by spinner button which 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()` [6] 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.



MATLAB App

CaGe App (Computer-aided approach for Genomics prediction)

Input

⠿	⠿HEB_01_003
BK_08	

	HEB_01_003	HEB_
MET12	0.0016	

Prediction methods

Linear approach: ☑ LASSO, □ GPR, □ SVR, □ PLS

Non-linear approach: □ RandomForest

DL approach: □ LSTM, □ CNN

Visualization of input data

Genomic Info

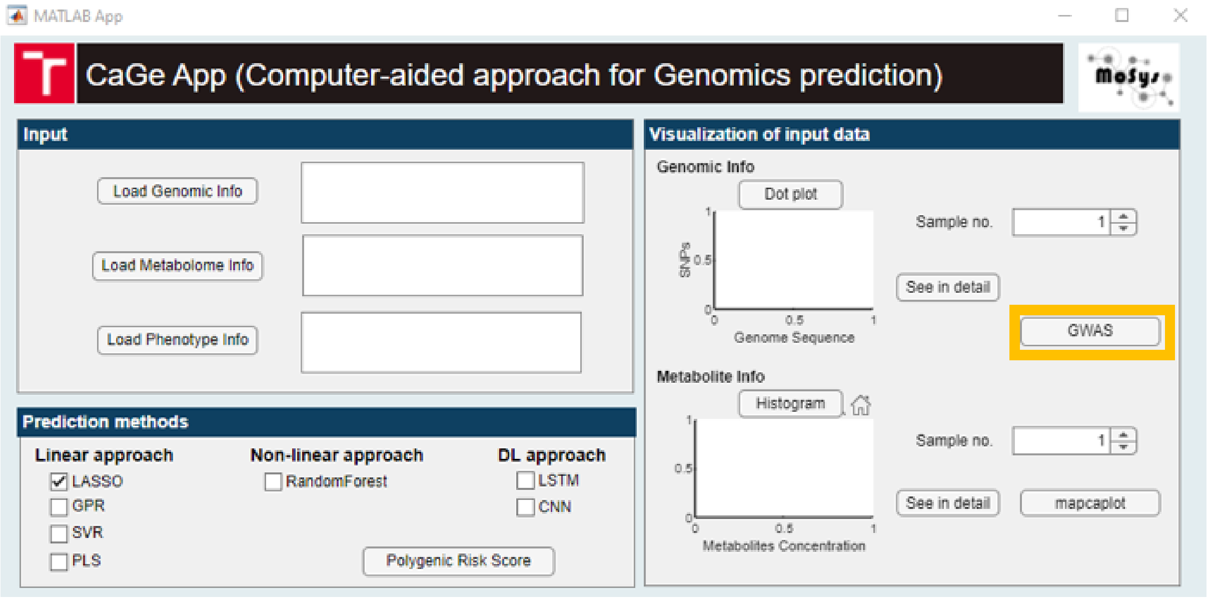
Sample no. 2

Metabolite Info

Sample no. 4

Figure 1: Example for inputs data – gen1_preprocessed.csv & met1_preprocessed.csv

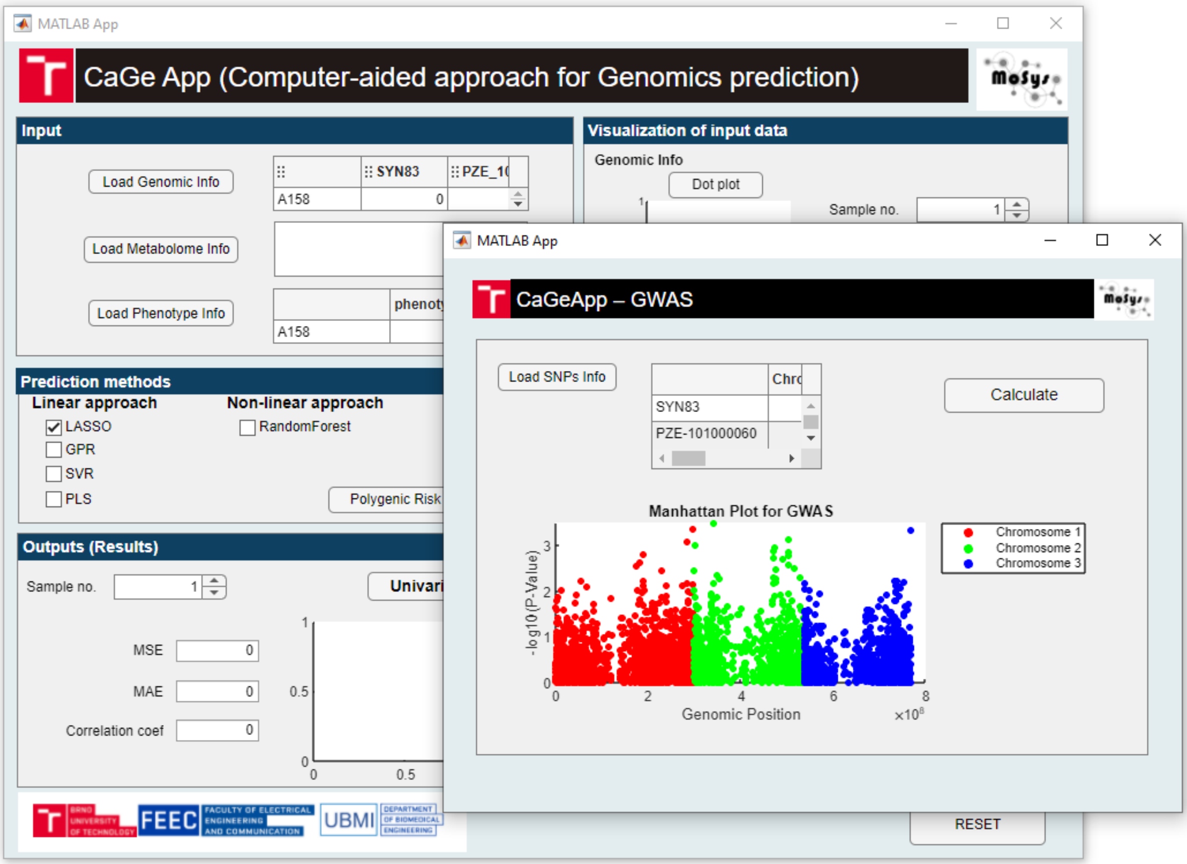
Newly, CaGe includes a GWAS button - which prompts the user to add information about SNPs, for correct rendering in the Manhattan plot.



☑ LASSO
□ GPR
□ SVR
□ PLS
□ RandomForest
□ LSTM
□ CNN

Sample no. 1

Sample no. 1



::	:: SYN83	:: PZE_10
A158	0	

	phenot
A158	

☑ LASSO
□ GPR
□ SVR
□ PLS
□ RandomForest

Sample no. 1

Sample no. 1

MSE 0

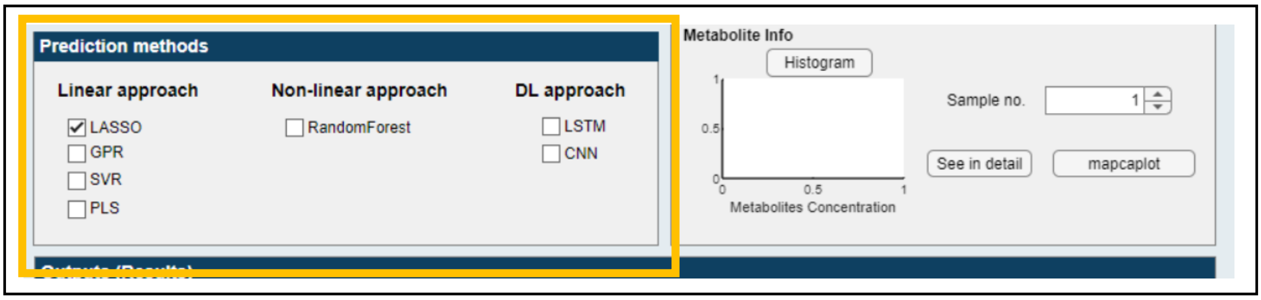
MAE 0

Correlation coef 0

	Chrc
SYN83	
PZE-101000060	

Figure 2: Example for inputs data – Geno_maize.csv & Pheno_maize.csv & InfoSNP_maize.csv

3. Prediction methods



Prediction methods

Linear approach	Non-linear approach	DL approach
☑ LASSO	□ RandomForest	□ LSTM
□ GPR		□ CNN
□ SVR		
□ PLS		

Metabolite Info

Sample no. 1

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 *L*1 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

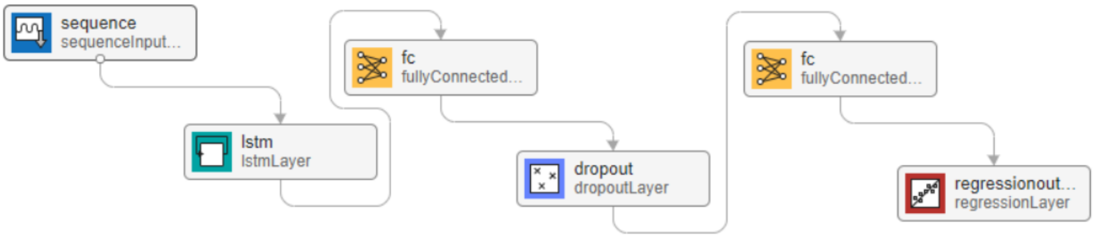
(https://www.mathworks.com/help/stats/treebagger.oobquantilepredict.html)

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.

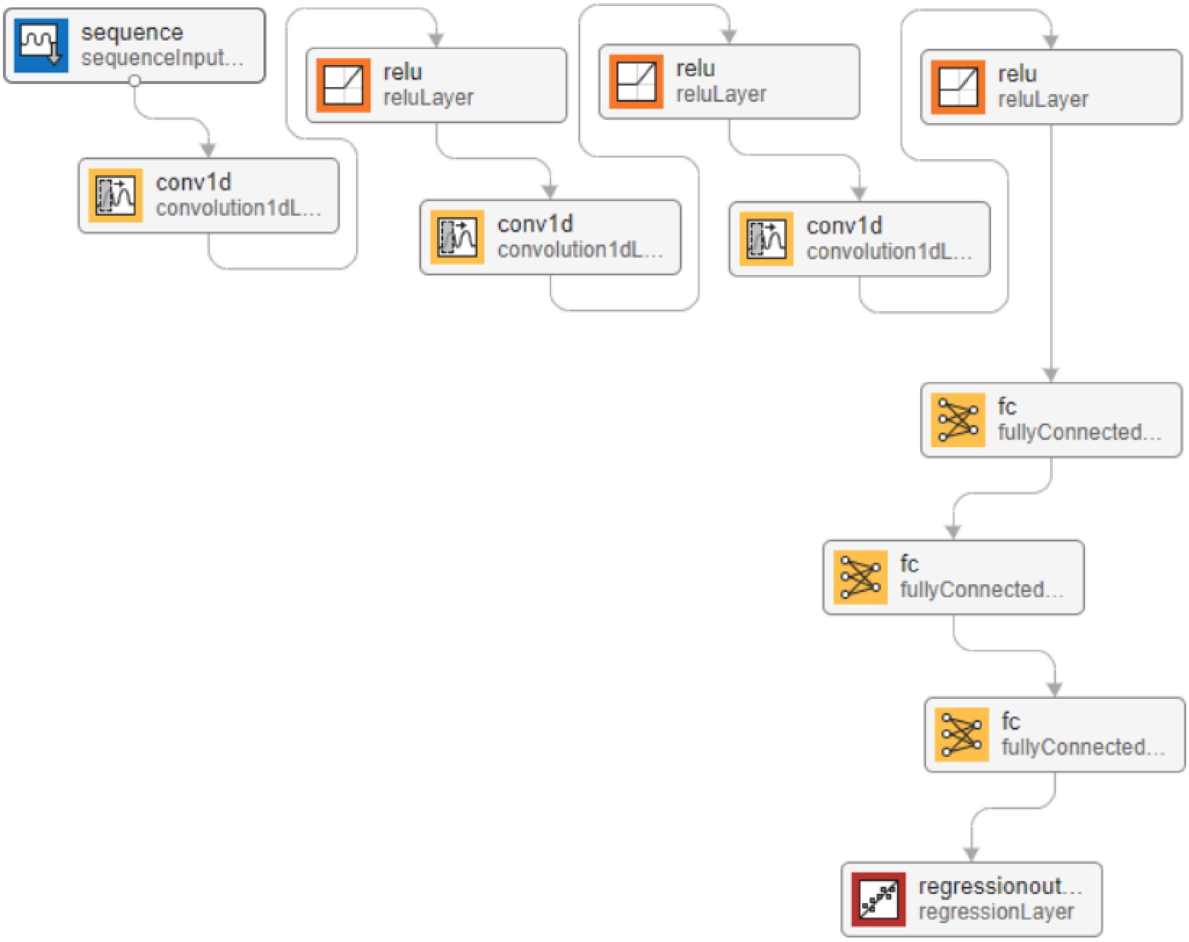
In CaGe software we implemented LSTM using architecture:



CNN

Convolutional Neural Network (CNN) is a class of artificial neural networkmost 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:

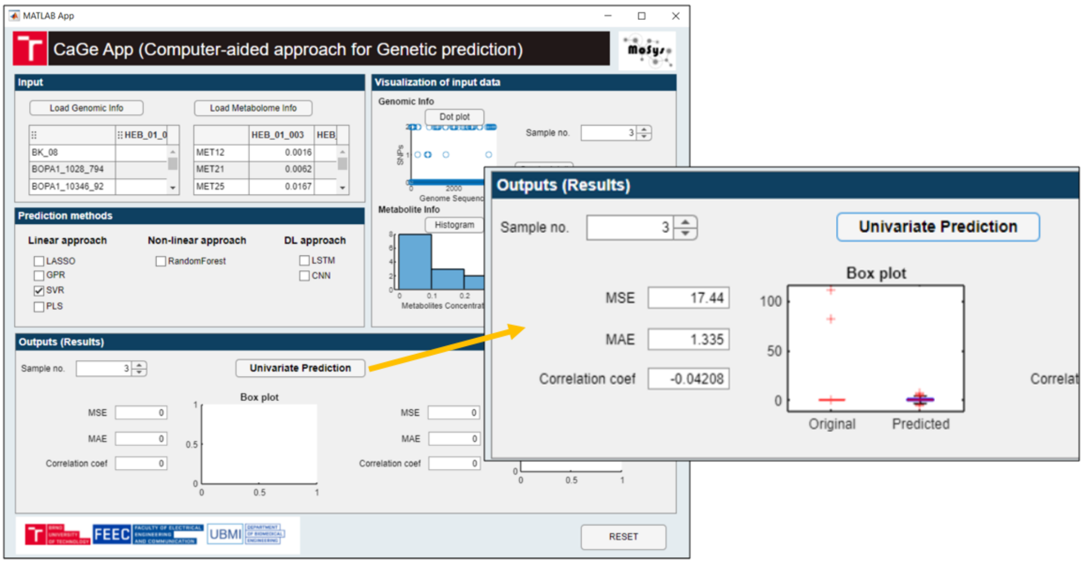


4. Outputs (Results)

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:



	HEB_01_0
BK_08	
BOPA1_1028_794	
BOPA1_10346_92	

	HEB_01_003	HEB
MET12	0.0016	
MET21	0.0062	
MET25	0.0167	

Linear approach: □ LASSO, □ GPR, ☑ SVR, □ PLS

Non-linear approach: □ RandomForest

DL approach: □ LSTM, □ CNN

Sample no. 3

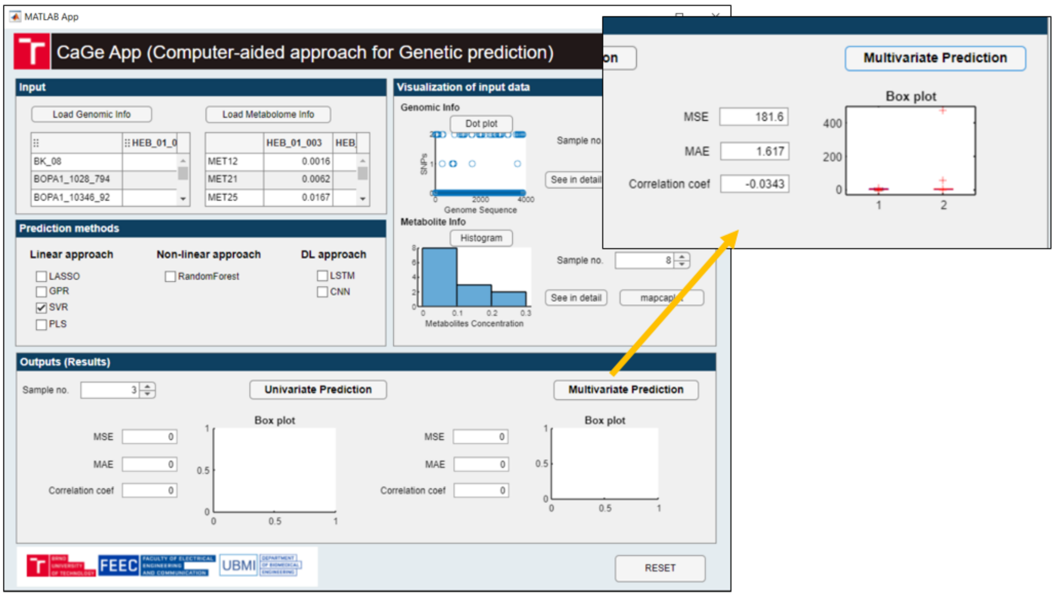
MSE 17.44

MAE 1.335

Correlation coef -0.04208

Multivariate prediction

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



	HEB_01_0
BK_08	
BOPA1_1028_794	
BOPA1_10346_92	

	HEB_01_003	HEB
MET12	0.0016	
MET21	0.0062	
MET25	0.0167	

Linear approach: □ LASSO, □ GPR, ☑ SVR, □ PLS

Non-linear approach: □ RandomForest

DL approach: □ LSTM, □ CNN

Sample no. 3

MSE 181.6

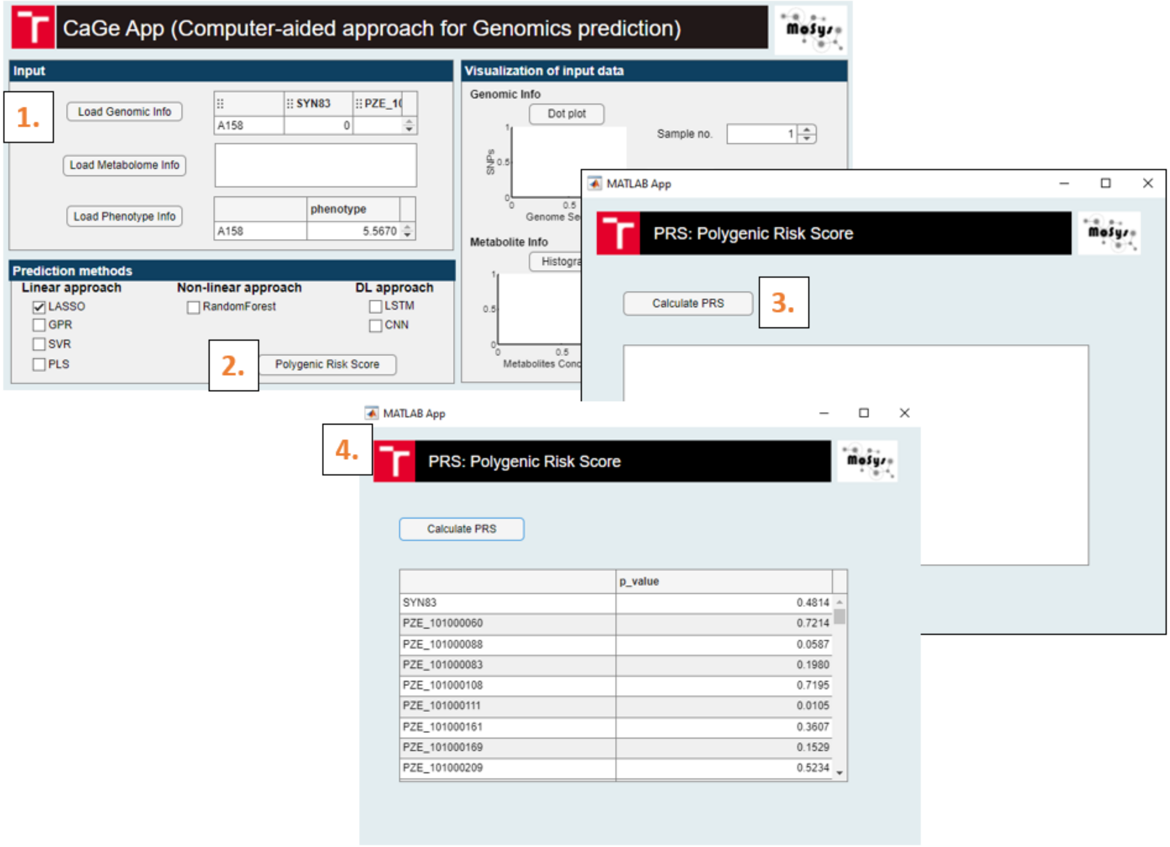
MAE 1.617

Correlation coef -0.0343

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.

Polygenic Risk Score

As part of genomic prediction, the tool also provides the calculation of a polygenic score, which represents a p-value. When interpreted correctly, this score can be scaled and understood as a comprehensive indicator affecting the genetic information for a specific phenotype.



	p_value
SYN83	0.4814
PZE_101000060	0.7214
PZE_101000088	0.0587
PZE_101000083	0.1980
PZE_101000108	0.7195
PZE_101000111	0.0105
PZE_101000161	0.3607
PZE_101000169	0.1529
PZE_101000209	0.5234

Figure 3: Example for inputs data – Geno_maize.csv & Pheno_maize.csv

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] Millet, E.J., et al., 2016. Genome-wide Bojer, C.S. and Meldgaard, J.P., 2021. Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting, 37(2), pp.587-603

[3] Hurta, Martin, et al. "Utilizing Genetic Programming to Enhance Polygenic Risk Score Calculation." 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023

[4] Weiszmann, J., et al., I., 2023. Metabolome plasticity in 241 Arabidopsis thaliana accessions reveals evolutionary cold adaptation processes. Plant Physiology, 193(2), pp.980-1000. analysis of yield in Europe: allelic effects vary with drought and heat scenarios. Plant Physiology, 172(2), pp.749-764

[5] Pérez, P. and de Los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics, 198(2), pp.483-495.

[6] 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.