MoSys

CaGeOMICs App

CaGeOMICs App

BRNO UNIVERSITY OF TECHNOLOGY FEEC FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION UBMI DEPARTMENT OF BIOMEDICAL ENGINEERING

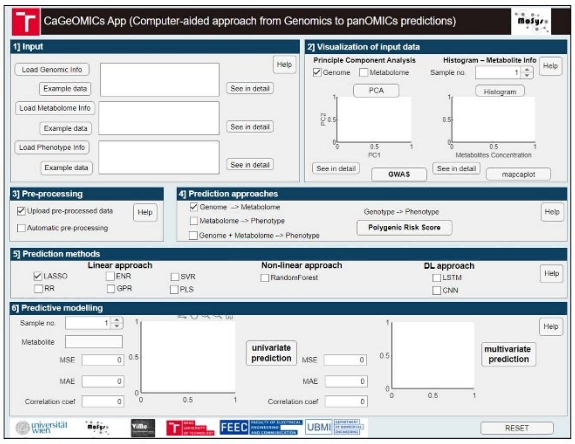
Department of Biomedical Engineering, FEEC, BUT
Molecular Systems Biology University of Vienna

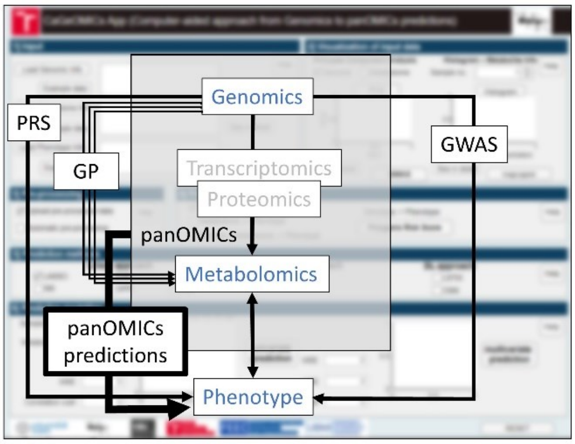
10.8.2024

..

CaGeOMICs App

Computer-aided approach from Genomics to panOMICs predictions





Content

Introduction

Advancements in sequencing technologies have propelled molecular biology research into the post-genomic era. Research now focuses on understanding functional relationships between individual genes and their impact on the final phenotype.

The CaGeOMICs app, implemented using Matlab2023b App Designer, provides a user-friendly tool for uploading genomic and metabolomic data, visualizing it, and performing prediction analysis using various methods, including linear, non-linear, and deep learning techniques.

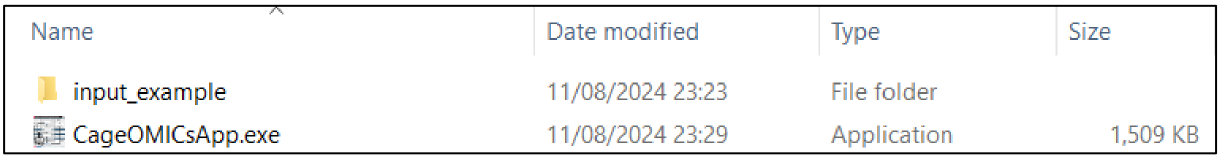
This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, which is financed from the OP RDE.

1. Installation

To install the CaGeOMICs App, follow these instructions based on your Matlab setup:

- If you have Matlab installed on your PC: You can run the app directly using the executable file provided. Navigate to the folder CageApp1/for_testing/ and execute CaGeOMICs.exe.

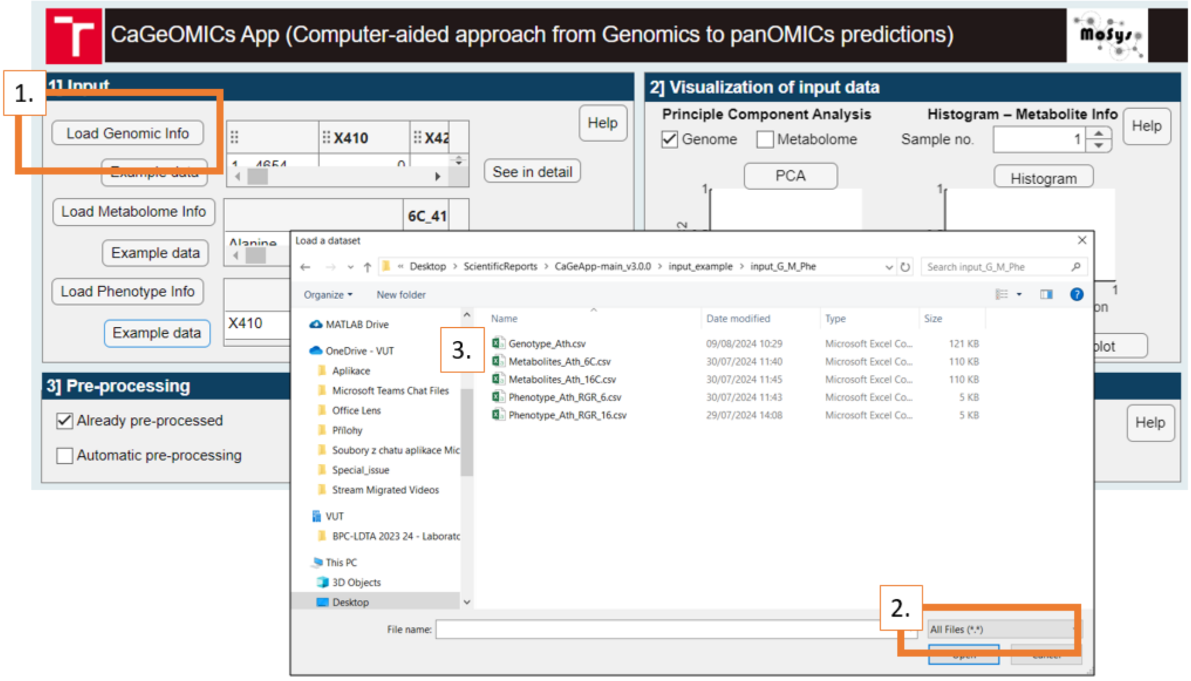
- If you do not have Matlab or a valid Matlab license: You can install the app using the standalone installer. Download and run MyAppInstaller_web.exe from the folder CageApp_1/for_redistribution/ to install the application on your computer. After installation, to the folder where you have the ".exe" application installed, download the folder with the input example data ""



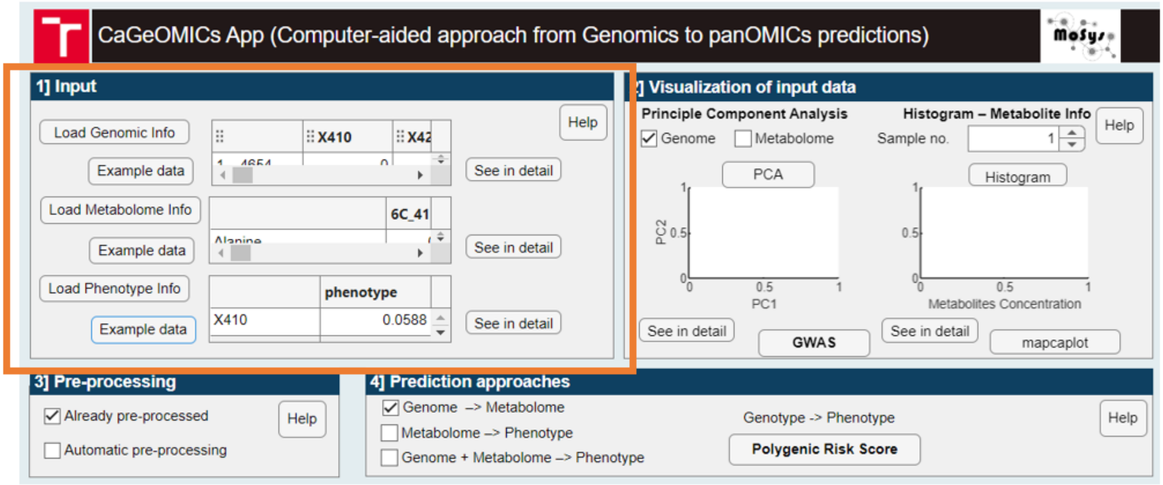
Name	Date modified	Type	Size
input_example	11/08/2024 23:23	File folder	
CageOMICsApp.exe	11/08/2024 23:29	Application	1,509 KB

2. Input

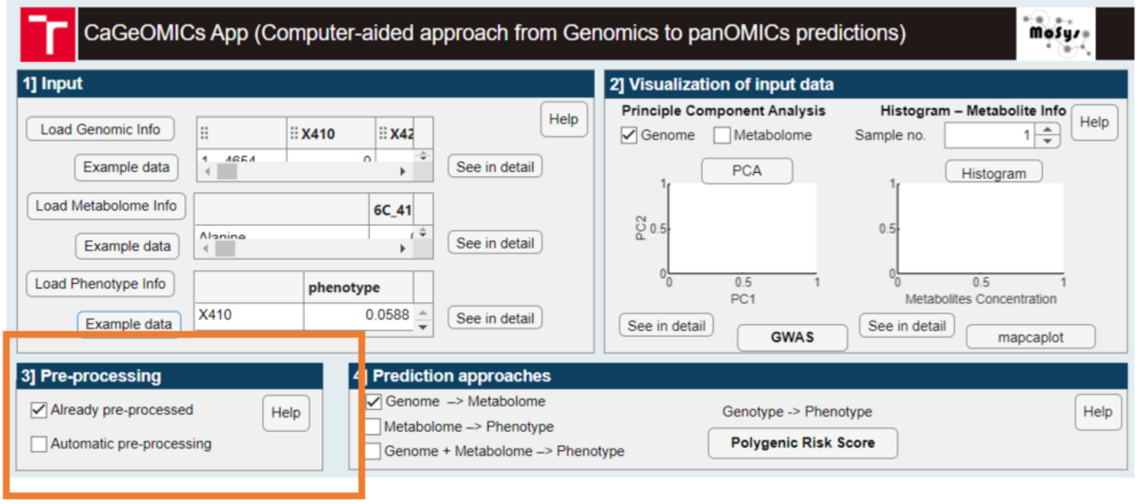
The first section of the app is designed for uploading data. The app expects genomic data in the form of SNP matrices, where '0' indicates an undetected SNP and '1' indicates a detected SNP. Users can also upload metabolomic data related to individual samples associated with the SNP data. Additionally, phenotypic data can be uploaded for calculating Polygenic Risk Scores (PRS) and panOMICs predictions.



If the data has already been pre-processed, users can proceed directly to analysis.



If pre-processing is needed, the app offers an automated pre-processing option. This feature fills missing values in genomic and phenotypic datasets with zeros, and for metabolomics data, it replaces missing values with half the minimum value and performs a log10 transformation on the absolute values.



CaGeOMICs App (Computer-aided approach from Genomics to panOMICs predictions)

1] Input

2] Visualization of input data

Principle Component Analysis: ☑ Genome □ Metabolome

Histogram – Metabolite Info: Sample no. 1

3] Pre-processing

☑ Already pre-processed

□ Automatic pre-processing

4] Prediction approaches

☑ Genome –> Metabolome

□ Metabolome –> Phenotype

□ Genome + Metabolome –> Phenotype

Genotype -> Phenotype

Tested data during app development:

Arabidopsis thaliana (Ath) datasets

For testing, we utilized datasets from the Weiszmann study [1] which include measurements from two temperature conditions: 6°C and 16°C. These datasets feature 37 metabolites representing the core metabolome of 241 different Arabidopsis ecotypes. The core genotype data comprises 16,544 SNPs. Phenotype information is provided as relative growth rates at both temperatures.

Barley datasets

The barley datasets [2] include genotype matrices for 1,363 lines from the wild barley nested association mapping (NAM) population HEB-25, genotyped with a 50k Illumina SNP Array. The SNP dataset includes 33,005 SNPs that meet the quality criteria; after excluding missing values, 3,878 SNPs remain. Metabolite data for barley includes 158 metabolites. The overlap between barley and Arabidopsis metabolite sets comprises 25 metabolites: alanine, aspartic acid, butanoic acid, fructose, fumarate, galactinol, galactose, glucose, glutamate, glycine, isoleucine, lactic acid, leucine, malic acid, myo-inositol, oxoglutaric acid, proline, putrescine, pyruvic acid, serine, succinic acid, sucrose, threonic acid, threonine, and valine.

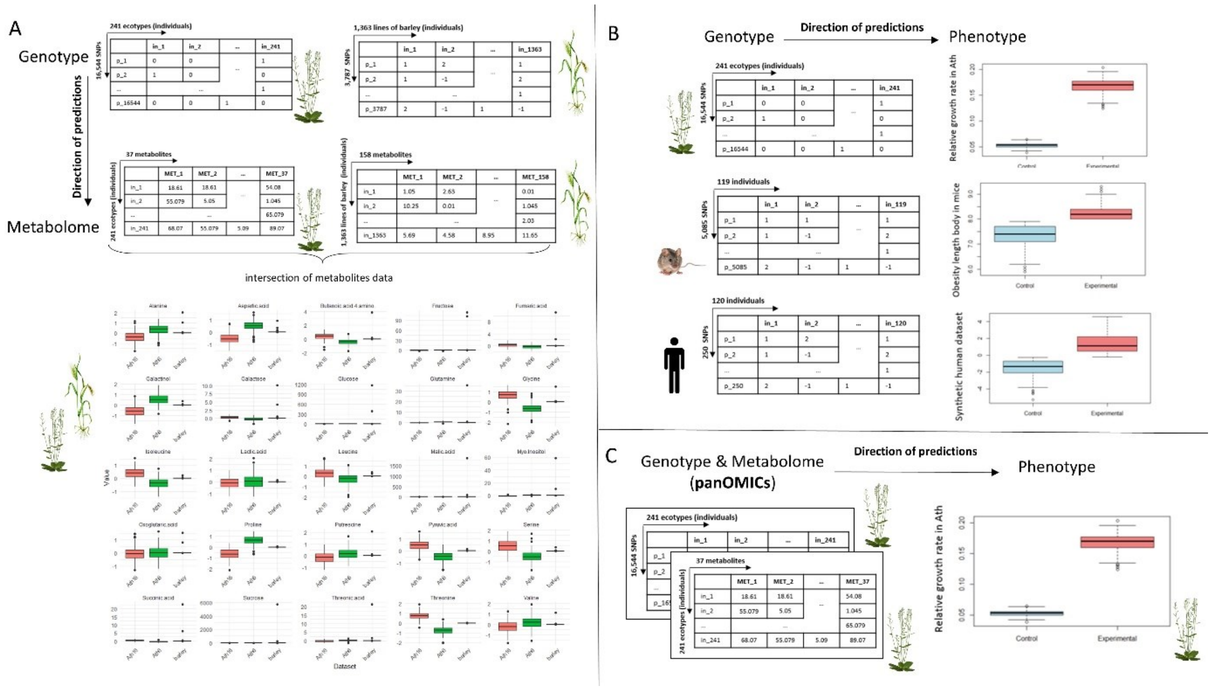
Mice Datasets

Mice datasets were analyzed using R/BGLR (Bayesian Generalized Linear Regression) [3]. This dataset includes genomic and phenotypic data focused on obesity-related traits and other complex characteristics. Body length measurements among individuals with obesity range from 5.9 to 9.3, with most values clustered in the mid to upper 7s. The distribution of phenotype data is relatively normal with fewer extreme values, indicating a concentration around a central mean. The dataset comprises 89 control (healthy) subjects and 30 obese (experimental) subjects.

Human Datasets

For human data, we used datasets from Kaggle [4], specifically the 'SNP dataset for GWAS'. This dataset includes SNP information crucial for genome-wide association studies (GWAS), which analyze genetic predispositions to various diseases and traits. The synthetic dataset was divided evenly with 60 control and 60 experimental samples.

Each dataset presents unique challenges and opportunities for testing and validating the predictive models in the CaGeOMICs app, ensuring its versatility and robustness across different biological domains. Comprehensive overview of the datasets used, is shown here:



	in_1	in_2	...	in_241
p_1	0	0	...	1
p_2	1	0		0
...	...			1
p_16544	0	0	1	0

	in_1	in_2	...	in_1363
p_1	1	2	...	1
p_2	1	-1		2
...	...			1
p_3787	2	-1	1	-1

	MET_1	MET_2	...	MET_37
in_1	18.61	18.61	...	54.08
in_2	55.079	5.05		1.045
...	...			65.079
in_241	68.07	55.079	5.09	89.07

	MET_1	MET_2	...	MET_158
in_1	1.05	2.63	...	0.01
in_2	10.25	0.01		1.045
...	...			2.03
in_1363	5.69	4.58	8.95	11.65

	in_1	in_2	...	in_241
p_1	0	0	...	1
p_2	1	0		0
...	...			1
p_16544	0	0	1	0

	in_1	in_2	...	in_119
p_1	1	1	...	1
p_2	1	-1		2
...	...			1
p_5085	2	-1	1	-1

	in_1	in_2	...	in_120
p_1	1	2	...	1
p_2	1	-1		2
...	...			1
p_250	2	-1	1	-1

	in_1	in_2	...	in_241
p_1				
p_2				
...				
p_165				

	MET_1	MET_2	...	MET_37
in_1	18.61	18.61	...	54.08
in_2	55.079	5.05		1.045
...	...			65.079
in_241	68.07	55.079	5.09	89.07

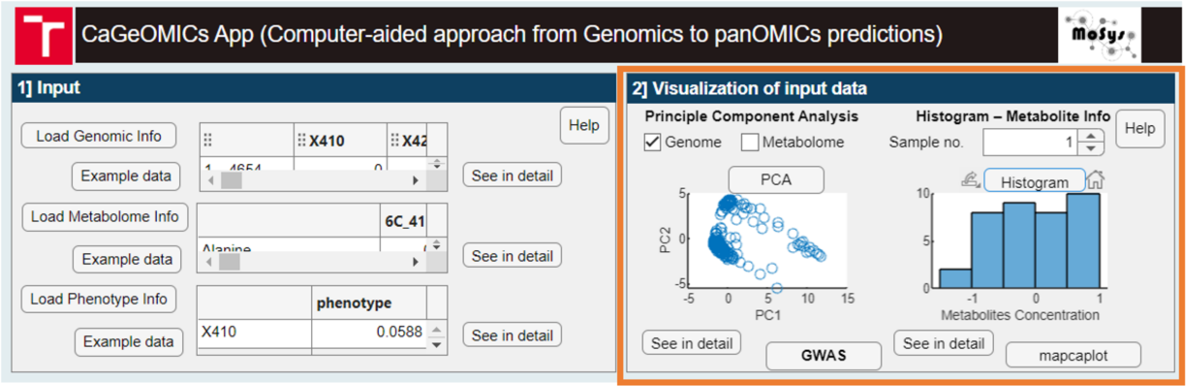
3. Visualization

The 'Visualization' section enhances the understanding of individual input samples. Genomic information can be visualized using Principal Component Analysis (PCA), allowing users to observe the distribution of individual SNPs or metabolites. The app also includes GWAS (Genome-Wide Association Study) calculations for genetic information. For metabolite data, users can view the distribution of metabolite concentrations using histogram plots.

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()`matlab function which:

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

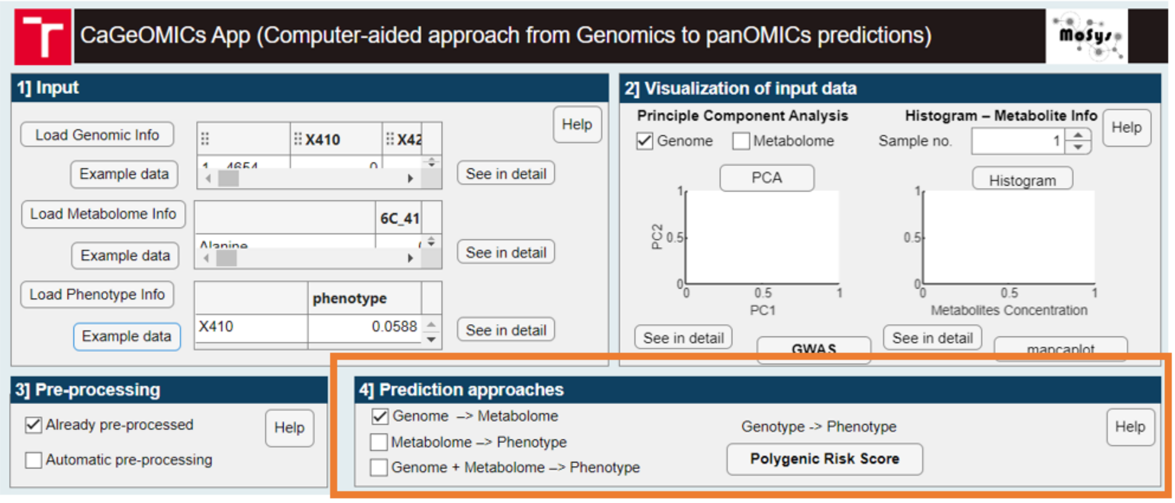
- Select principal components for the *x* and *y* axes from the drop-down list below each scatter plot.
- Click a data point to display its label.
- Select a subset of data points by dragging a box around them. Points in the selected region and the corresponding points in the other axes are then highlighted. 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.



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

4. Prediction approaches

The "Prediction Approaches" section is a core component of the app, offering users a range of options for making predictions. In this section, users can choose their desired prediction types and specify the methods to be used. The app supports four prediction approach applicable to genomic predictions, metabolome predictions, OMICs predictions and polygenic risk score (PRS) calculations.


CaGeOMICs App (Computer-aided approach from Genomics to panOMICs predictions)

1] Input

2] Visualization of input data

Principle Component Analysis: ☑ Genome □ Metabolome

Histogram – Metabolite Info: Sample no. 1

3] Pre-processing

☑ Already pre-processed

□ Automatic pre-processing

4] Prediction approaches

☑ Genome –> Metabolome

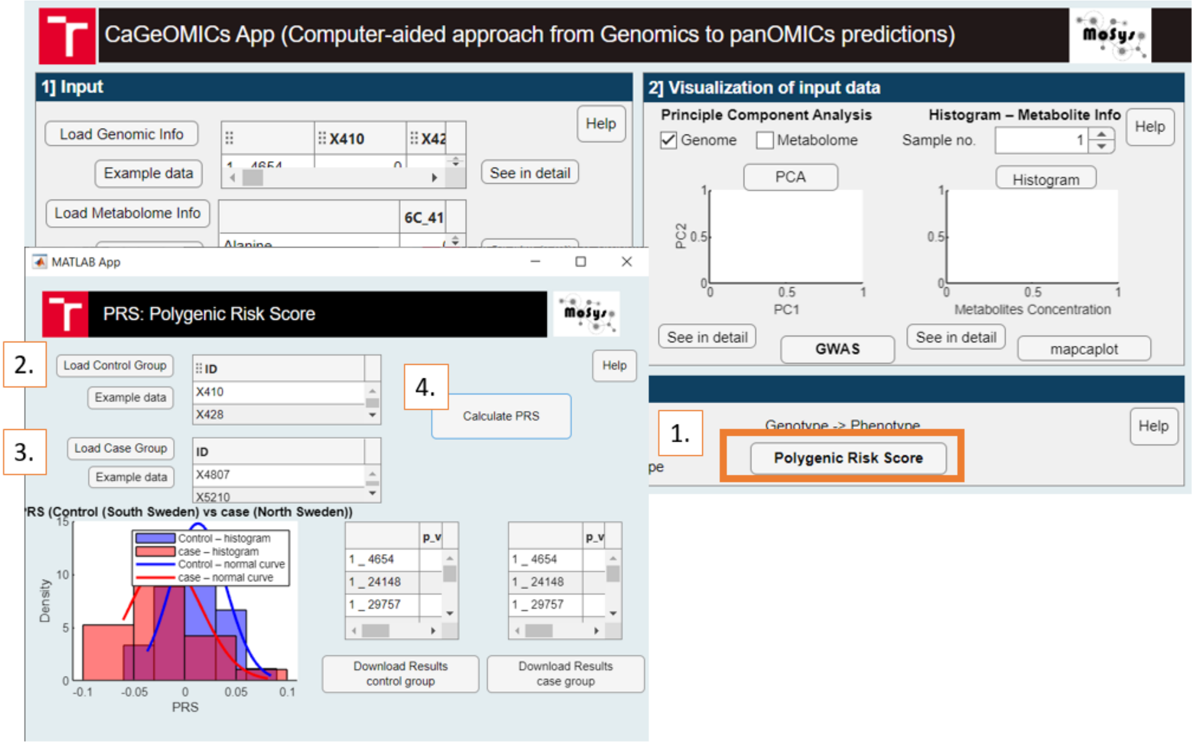
□ Metabolome –> Phenotype

□ Genome + Metabolome –> Phenotype

Genotype -> Phenotype

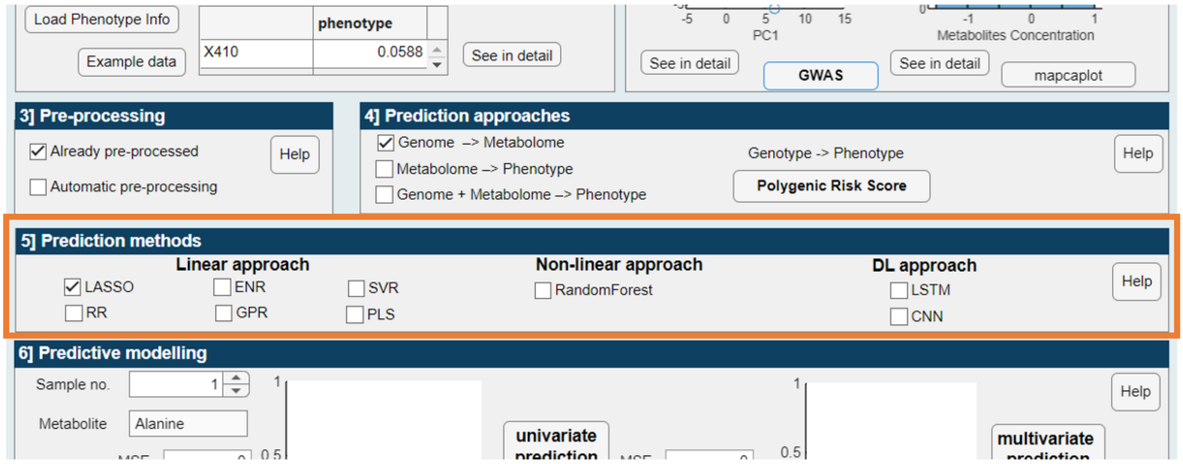
Polygenic risk score

The Polygenic Risk Score (PRS) feature in the app calculates an aggregate measure of genetic risk based on multiple SNPs associated with a particular phenotype. By summing the effects of individual SNPs, weighted by their respective risk scores, the PRS provides a comprehensive estimate of an individual's genetic predisposition to specific traits or diseases. This calculation requires SNP location information and outputs a CSV file with the PRS results, which can be further analyzed or visualized. The PRS functionality enables users to assess genetic risk in both control and experimental groups, offering insights into the genetic factors influencing phenotypic outcomes.


CaGeOMICs App (Computer-aided approach from Genomics to panOMICs predictions)
MoSys
1] Input
Load Genomic Info
Example data
X410
X42
Help
See in detail
Load Metabolome Info
6C_41
Alanine
2] Visualization of input data
Principle Component Analysis
☑ Genome
□ Metabolome
PCA
PC1
PC2
See in detail
GWAS
Histogram – Metabolite Info
Sample no.
1
Histogram
Metabolites Concentration
See in detail
mapcaplot
Help
Genotype -> Phenotype
Polygenic Risk Score
Help
1.
MATLAB App
PRS: Polygenic Risk Score
2.
Load Control Group
Example data
ID
X410
X428
3.
Load Case Group
Example data
ID
X4807
X5210
4.
Calculate PRS
Help
PRS (Control (South Sweden) vs case (North Sweden))
Control – histogram
case – histogram
Control – normal curve
case – normal curve
Density
PRS
15
10
5
0
-0.1
-0.05
0
0.05
0.1
p_v
1 _ 4654
1 _ 24148
1 _ 29757
p_v
1 _ 4654
1 _ 24148
1 _ 29757
Download Results control group
Download Results case group

5. Prediction methods

The core components of the app are 'Prediction Methods' sections. Users can select their desired predictions and specify the methods to be used. The app offers nine different prediction methods.



	phenotype
X410	0.0588

3] Pre-processing

☑ Already pre-processed

☐ Automatic pre-processing

4] Prediction approaches

☑ Genome –> Metabolome

☐ Metabolome –> Phenotype

☐ Genome + Metabolome –> Phenotype

Genotype -> Phenotype

5] Prediction methods

Linear approach: ☑ LASSO ☐ RR ☐ ENR ☐ GPR ☐ SVR ☐ PLS

Non-linear approach: ☐ RandomForest

DL approach: ☐ LSTM ☐ CNN

6] Predictive modelling

Sample no. 1

Metabolite Alanine

Linear approaches:

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

LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) regression applies L1 regularization to enhance model performance by performing both variable selection and regularization. In this method, the model is trained to predict exam scores using the LASSO and elastic net methods. Predictions are compared with actual exam grades using the formula:

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

RR

RR (Ridge Regression) is a linear model incorporating L2 regularization to penalize large coefficients and mitigate multicollinearity. In the CaGeOMICs application, RR is implemented with a very small alpha parameter (e.g., 0.000001), approximating ridge regression. The model is trained using 5-fold cross-validation to find the optimal lambda value, enhancing stability and prediction accuracy.

ENR

ENR (Elastic Net Regularization) combines L1 and L2 penalties to leverage the strengths of both LASSO and Ridge Regression. In this approach, the alpha parameter is set to 0.75, balancing LASSO and Ridge contributions. The ENR model undergoes 5-

fold cross-validation to determine the best regularization parameters, which are then used for predicting outcomes.

GPR

Fit a Gaussian process regression (GPR) model; GPR involves fitting a Gaussian Process Regression model using the fitrgp function. This model is trained on sample data, accommodating both univariate and multivariate outputs, and provides flexible prediction capabilities

`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; SVR fits a support vector machine regression model to predictor data, utilizing kernel functions and soft-margin minimization techniques. The ***fitrsvm*** function supports various methods, including SMO and quadratic programming. For high-dimensional data, fitrlinear is used. For binary classification, fitcsvm and fitclinear are employed based on data dimensions

`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 – PLS regression calculates predictor and response loadings using the ***plsregress*** function, which identifies relationships between predictors and responses by specifying the number of components.

`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

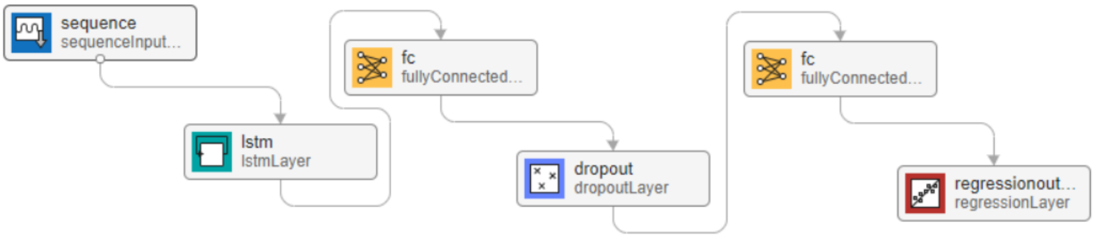
(https://www.mathworks.com/help/stats/treebagger.oobquantilepredict.html)

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.

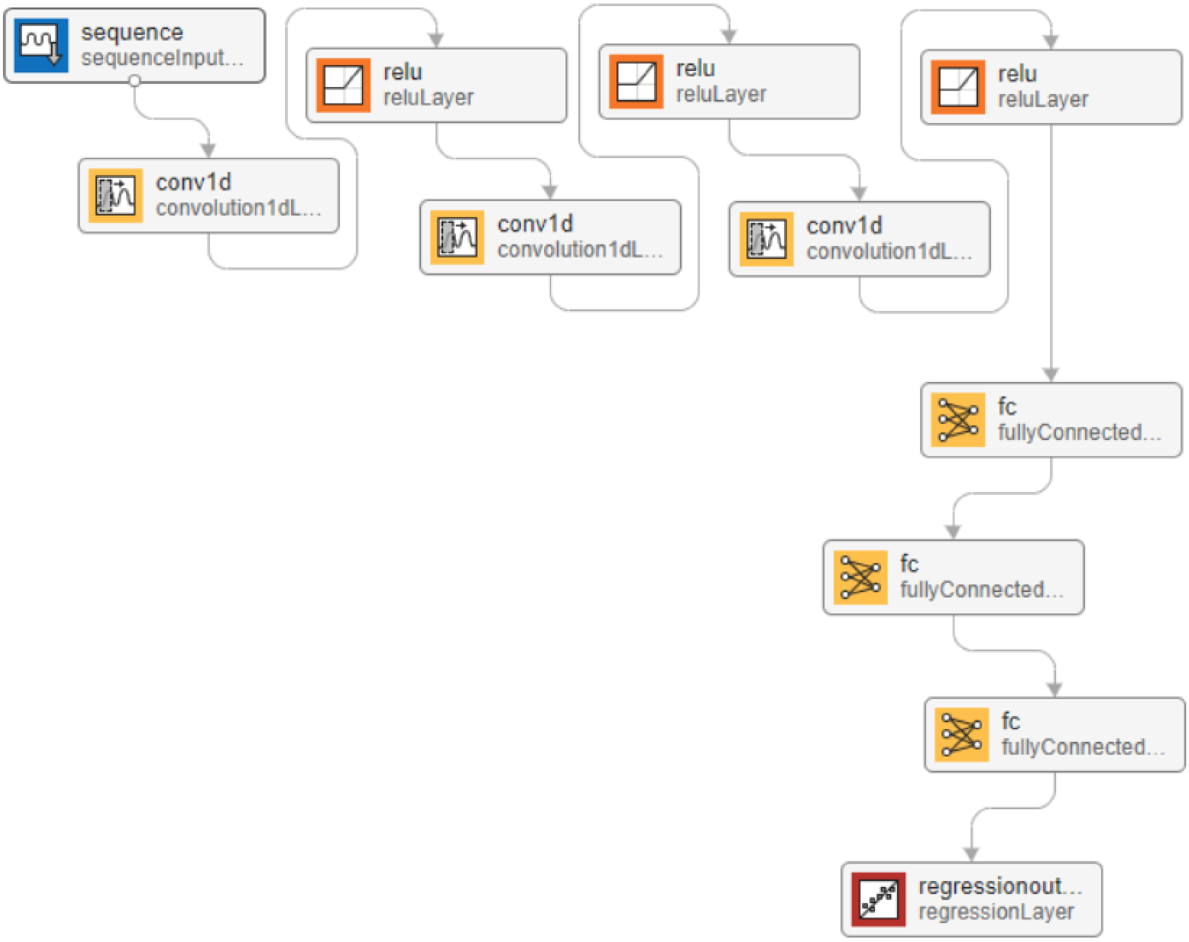
In CaGeOMICs software we implemented LSTM using architecture:



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 CaGeOMICs software we implemented CNN using architecture:

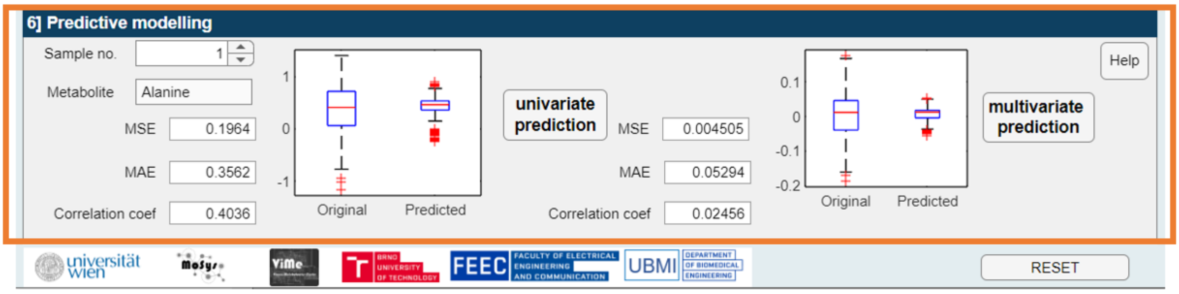


6. Outputs (Results)

The results are divided into two parts rely on genomics/metabolomics/OMICs prediction analysis. In general, the prediction analysis is evaluated using mean square error (MSE), mean absolute error (MAE) and correlation coefficient (cc). However, since all three of these evaluation parameters can be very misleading, the CaGeOMICs 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:



6] Predictive modelling

Sample no. 1

Metabolite Alanine

MSE 0.1964

MAE 0.3562

Correlation coef 0.4036

MSE 0.004505

MAE 0.05294

Correlation coef 0.02456

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] Weiszmann, Jakob, et al. "Metabolome plasticity in 241 Arabidopsis thaliana accessions reveals evolutionary cold adaptation processes." *Plant physiology* 193.2 (2023): 980-1000.

[2] Gemmer, M.R., Richter, C., Jiang, Y., Schmutzer, T., Raorane, M.L., Junker, B., Pillen, K. and Maurer, A., 2020. Can metabolic prediction be an alternative to genomic prediction in barley?. *PLoS One*, *15*(6), p.e0234052.

[3] Villar-Hernández, B.D.J., Dreisigacker, S., Crespo, L., Pérez-Rodríguez, P., Pérez-Elizalde, S., Toledo, F. and Crossa, J., 2024. A Bayesian optimization R package for multitrait parental selection. *The Plant Genome*, p.e20433.

[4] Bojer, C.S. and Meldgaard, J.P., 2021. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), pp.587-603.