AutoGP Help document V2

brief introduction

In order to effectively promote the continuous development of the field of genomic selection (GS), we have developed ——AutoGP, a platform that integrates data storage, data pre-processing, data analysis, genome selection, and GS forum.

The AutoGP design aims to provide a highly integrated environment that reduces the workload of users in programming, and provides an intuitive, easy-to-use operating interface that allows users to easily perform a range of complex tasks. Specifically speaking, the AutoGP has the following characteristics:

- **Multi-mode database management platform: support multi-mode database management, convenient for users to conduct efficient and flexible processing of data.**
- **Simple and convenient data pre-processing tool: to extract high-quality SNP data from gene data, and to easily extract the phenotype through video.**
- **Convenient and intuitive data analysis tools: provide a variety of convenient and intuitive data analysis tools to help users quickly understand and process data.**
- **Genome prediction methods for multiple model selection: integrating a variety of traditional statistical methods, machine learning (ML) and deep learning (DL) technology, providing a powerful and flexible tool for genome selection.**

This document provides two versions of the operating guidelines:

1. **Start fast: It is suitable for users who have rich experience in the GS field and have a certain knowledge of AutoGP. This guide allows you to quickly complete your personal requirements tasks.**
2. **Detailed guide: Suitable for users unfamiliar with the platform operation steps. This guide specifies the various input data format requirements to help you use the platform smoothly.**

Start fast

1. Login

Please use the right side of the navigation bar to log in or register for your personal account.

Please use your personal user name and password to log on to the platform.

2. Data management

The platform provides users with three types of databases:

1. **Personal database: The user has full permission to the data in this database. Such data is the private data of the account, and can only be uploaded, downloaded and deleted by the account itself.**

2. **Shared database: Users can upload and download the shared data. Such data are spontaneously uploaded by breeding researchers to facilitate the field of GS and allow other users to conduct relevant research. After the platform certification, part of the shared data will be upgraded to quality data.**

3. **Quality database: users can only download such data. The data in the high-quality database is strictly screened and confirmed by the platform to ensure that it is high-quality data.**

Using personal databases as an example, users can upload the desired gene vcf files, phenotype csv files, and genotype sequence data txt files of the offspring to be predicted. Shared database, high-quality database operation is the same as above, only file upload, deletion and other permission differences.

In addition, for the additional "phenotype database", the "phenotype extraction" function module can store crop video data and point cloud data.

3. Pre-data processing

High-quality SNP extraction

To upload files:

- Gene vcf file (genome or with large-scale SNP data)

After completing the above content, click the "Submit" button to submit the task. When the Download Now button is illuminated, it indicates that the high-quality SNP data corresponding through the specific gene regulatory network has been extracted. Click "Download Now" to download the corresponding vcf file.

Phenotype extraction

Please enter the phenotype extraction interface through the navigation bar.

Data upload and selection:

- **Upload video: Click the "Upload video" button, users can scan the QR code and upload the video data using the wechat applet.**
- **Video selection: After the data is uploaded, it will be added to the website database, and users can select the video through the drop-down box. After the selection is complete, click Submit, and the background will begin the 3 D reconstruction. The reconstruction process usually takes more than 45 minutes. After completion, the user can slide the page to view the 3 D model and perform**

phenotype extraction, and can drag the mouse to view.

Automatic Phenotype Extraction:

- After clicking the Auto Extraction button, the reference phenotype based on the deep learning algorithm will be displayed.

Interactive phenotype Extraction:

- After clicking the "Interactive Extraction" button, the interactive phenotype extraction interface is enter.
- First, the mouse selects the length of the "marker" and inputs the real length in the upper left corner to form the reference length, which facilitates the scale transformation.
- There are two interactive modes: measurement and counting. The user can click to make the length measurement or obtain the number of counting points.

4. Data analysis

GWAS analyse

To upload files:

- Genes and the vcf files
- The target csv file

After completing the above content, click the "Submit" button to submit the task. When the Download Now button is lit, the relevant GWAS analysis is completed.

Population division

To upload files:

- Genes and the vcf files

After completing the above content, click the "Submit" button to submit the task. When the Download Now button is lit, the relevant group analysis is completed.

Analysis of phenotypic data

To upload files:

- The target csv file

After completing the above content, click the "Submit" button to submit the task. When the Download Now button is lit, it indicates that a descriptive analysis of the relevant csv file is completed.

5. Genomic selection

This platform provides users with four algorithm functions for different tasks, namely model training, phenotype prediction, training prediction integration and optimal parent selection.

1. **Model training: This function allows users to train a trait prediction model with certain prediction ability through machine learning (ML) or deep learning (DL) by using a batch of crops (VCF format files) and corresponding phenotypic trait data.**

2. **Phenotype prediction: This function allows users to predict the genotype data of the predicted material using the model file.**

3. **Training and prediction integration: This function integrates the two major algorithm functions of "model training" and "phenotype prediction", providing a more convenient operation interface.**

4. **Selection of optimal parents: This function allows the user to hybridize seeds of known genotypes to a batch of homozygous data through a trained model file, predicting which combination of the offspring is most expected.**

Next, introduce the type file requirements required for each algorithm model in turn:

Model training

To upload files:

- Genes and the vcf files
- Single-phenotype csv file

To select the model:

- One of the 10 ML / DL methods was randomly selected

Select and fill in content:

Note information (note: do not appear space, only for simple marking)

After completing the above content, click the "Start Training" button to submit the task. When the Download Now button is on, the model training is complete. Click "Download Now" to download the corresponding model weight file.

Time reference:

Model training using 5000 SNP of 1000 accessions for approximately 1 min.

Model training (including environmental information)

To upload files:

- Genes and the vcf files
- The environment csv file
- Single-phenotype csv files corresponding to each environment

Select and fill in content:

Note information (note: do not appear space, only for simple marking)

After completing the above content, click the "Start Training" button to submit the task. When the Download Now button is on, the model training is complete. Click "Download Now" to download the corresponding model weight file.

Phenotype prediction

To upload files:

- Genes and the vcf files
- Model weight files of the "model training" output

Select and fill in content:

If the model file you upload is a DNNGP model, then you need to upload an additional PCA model weight file.

After completing the above content, click the "Start forecast" button to submit the task. When the Download Now button is lit, the model training is complete. Click

"Download Now" to download the corresponding document.

Time reference:

Use 6,000 copies for about half a minute.

Phenotype prediction (including environmental information)

To upload files:

- Genes and the vcf files
- The environment csv file
- The model weight file produced by the "model training (including environmental information)"

Select and fill in content:

If you upload the model file for the DNNGP model, then you need to upload an additional PCA model weight file.

After completing the above content, click the "Start forecast" button to submit the task. When the Download Now button is lit, the model training is complete. Click "Download Now" to download the corresponding document.

Training and prediction integration

To upload files:

- Genes and the vcf files
- The vcf file corresponding to the single-phenotype csv file
- Gene vcf files of the material to be predicted

To select the model:

- One of the 10 ML / DL methods was randomly selected

Select and fill in content:

Note information (note: do not appear space, only for simple annotation)

After completing the above content, click the "Start Training" button to submit the task. When the Download Now button is on, the model training is complete. Click "Download Now" to download the corresponding document.

Time reference:

5000 SNP for model training + 6000 accessions using 1000 accessions

were predicted for approximately 2 minutes.

Training and prediction integration (including environmental information)

To upload files:

- Genes and the vcf files
- The environment csv file
- The vcf file corresponding to the single-phenotype csv file
- Gene vcf files of the material to be predicted

Select and fill in content:

Note information (note: do not appear space, only for simple marking)

After completing the above content, click the "Start Training" button to submit the task. When the Download Now button is on, the model training is complete. Click "Download Now" to download the corresponding document.

Selection of optimal parents

To upload files:

- Genes and the vcf files
- Model weight files of the "model training" output
- Gene sequence txt files of the material to be evaluated
- Expected phenotype value (Max / Min / arbitrary value)

Select and fill in content:

If the model file you upload is a DNNGP model, then you need to upload an additional PCA model weight file.

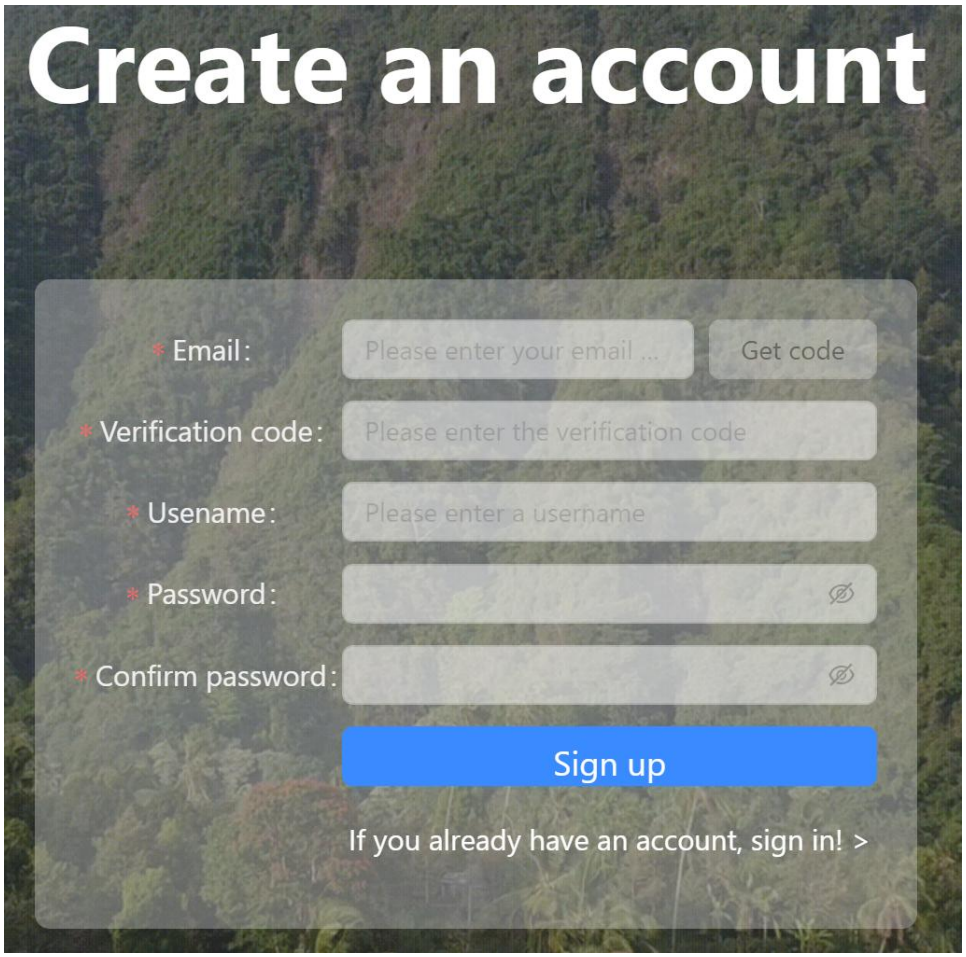
After completing the above content, click the "Start forecast" button to submit the task. When the Download Now button is lit, the model training is complete. Click "Download Now" to download the corresponding document.

Detailed guide

The detailed guide expands from the sample demonstration

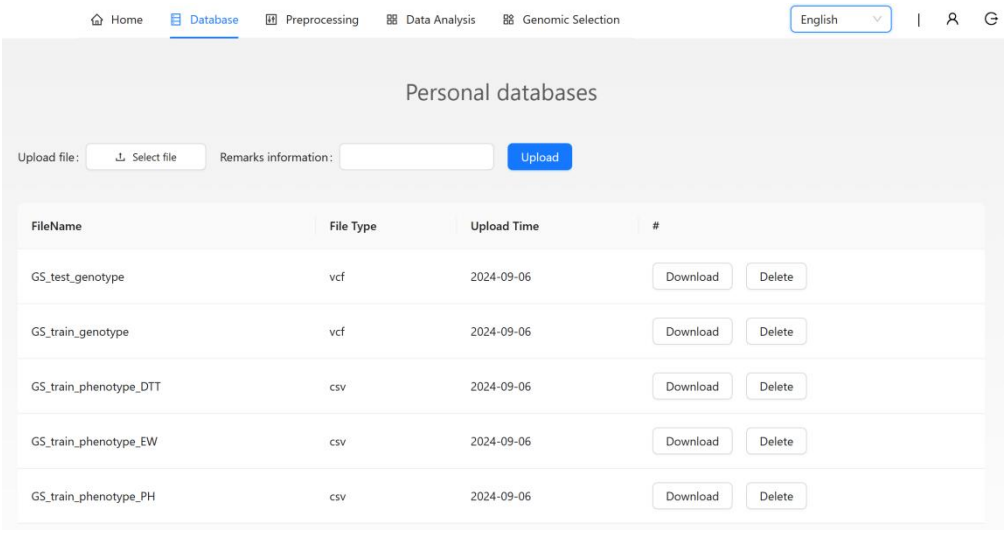
1. Login

For the new account, users can experience logging in through the two accounts provided on the login page, or register their exclusive personal account through their personal email.

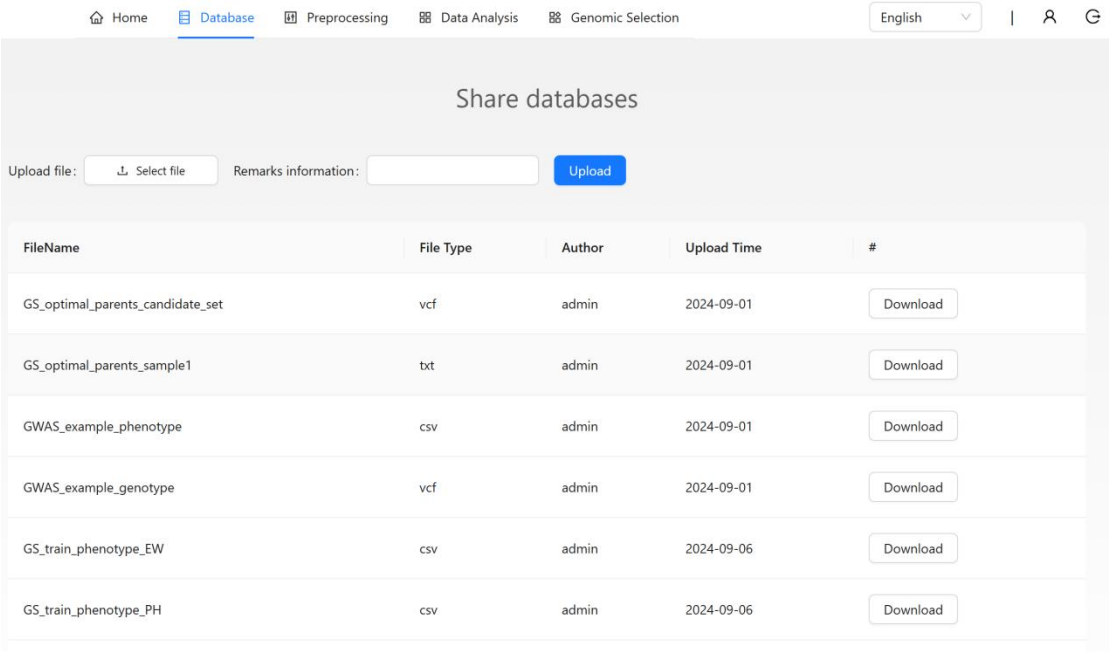


2. Data management

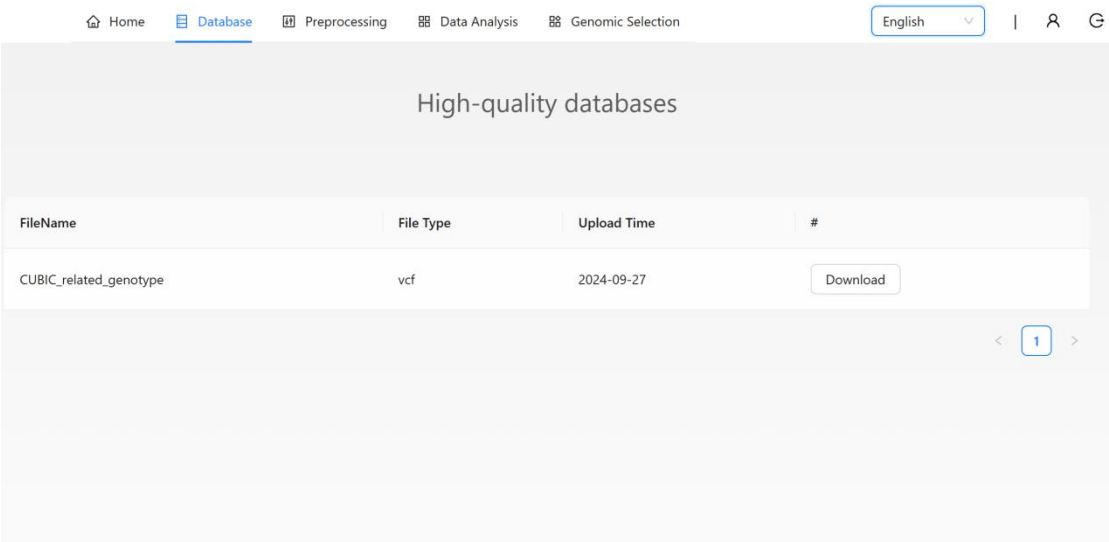
Personal database: the user manages himself and has the permission to upload, download and delete.



Shared database: Users can upload their personal data to the shared platform and allow other users to use it. You can also use other people data through this database.



Selected database: the database is collected, organized and managed by the platform itself, and users can use it.



3. Pre-data processing

High-quality SNP extraction

The platform provides sample data:

- High-confidence_original_data.vcf

By extracting the original gene VCF file (not screening the SNP through the regulatory network), the user can obtain the gene VCF file of the "corresponding gene regulatory network of DTT".

Phenotype extraction

Video shooting requirements

1. The shooting should place the marker beside the target corn, and the square object with a fixed height should not be too small, if the sign inserted into the ground should reach 1 / 3 of the height of the corn field. If the seedling is small and the sign is large, the sign can be cut short to facilitate shooting.

2. Pay attention to the target corn for not shaking too much when shooting, and pay attention to the wind or collision. The shaking corn will affect the final reconstruction effect.

3. Try to shoot a 1280 * 720, or 720p HD video format.

Shooting skills

(1) Tip 1: First, take a distant shot and scan around the object at 360° . Then take close-up shots, also around 360° . The shooting speed should be moderate, with a total length of about 25 seconds, to avoid too fast to prevent motion blur. If the object is large, such as corn, you can slow down appropriately.

(2) Tip 2: scan from top to bottom. Ensure moderate speed up and down and scan up and down from about six different directions.

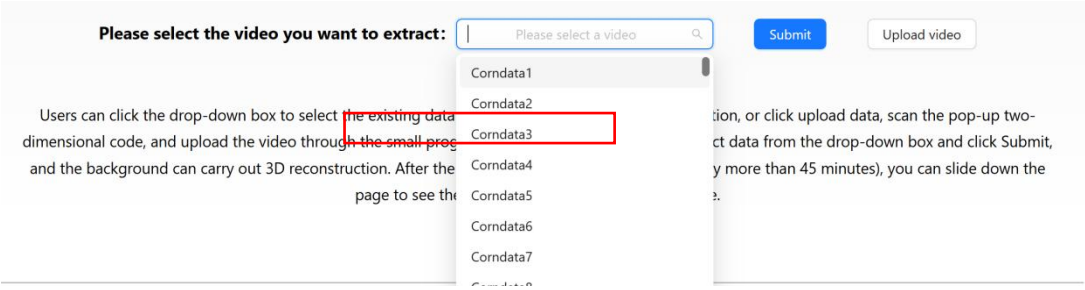
We used Corndata1 as an example to demonstrate the phenotype extraction process.

Data upload and selection:

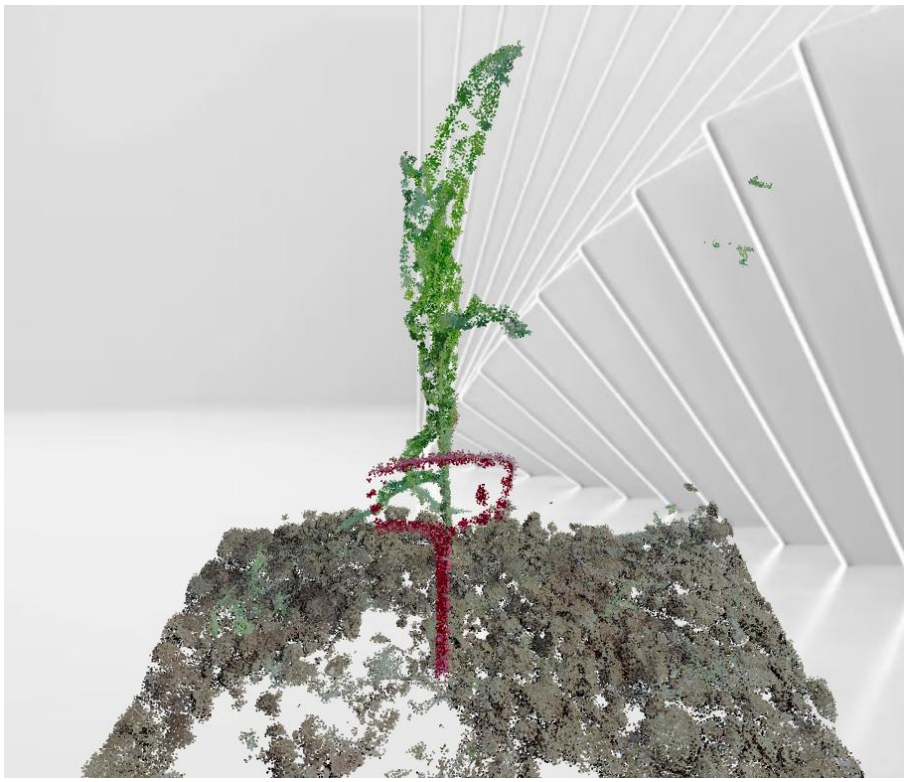
- **Upload video: Click the "Upload video" button, scan the QR code, upload the Corndata1 video data, and fill in the name, region, variety and other information of the uploader in the small program.**

- **Video selection: After uploading, the data will be added to the**

website database. Click the drop-down box, select Corndata1 data, and the video will automatically start 3 D reconstruction, and the reconstruction can be completed after about 45min.



After the 3 D reconstruction is completed, pull down the web page and see the 3 D model of Corndata1.



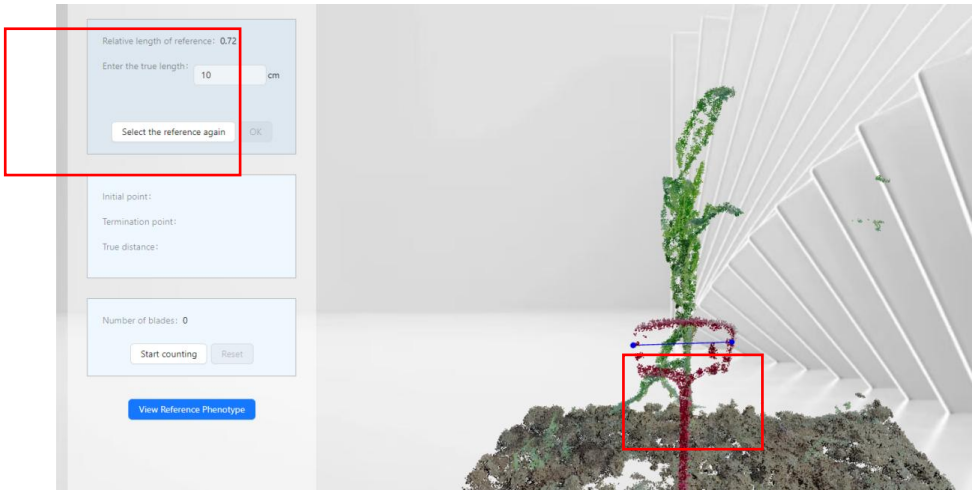
Automatic Phenotype Extraction:

- After clicking the "Automatic Extraction" button, the reference phenotype will be displayed at the left end of the page (this function is being upgraded and maintained).

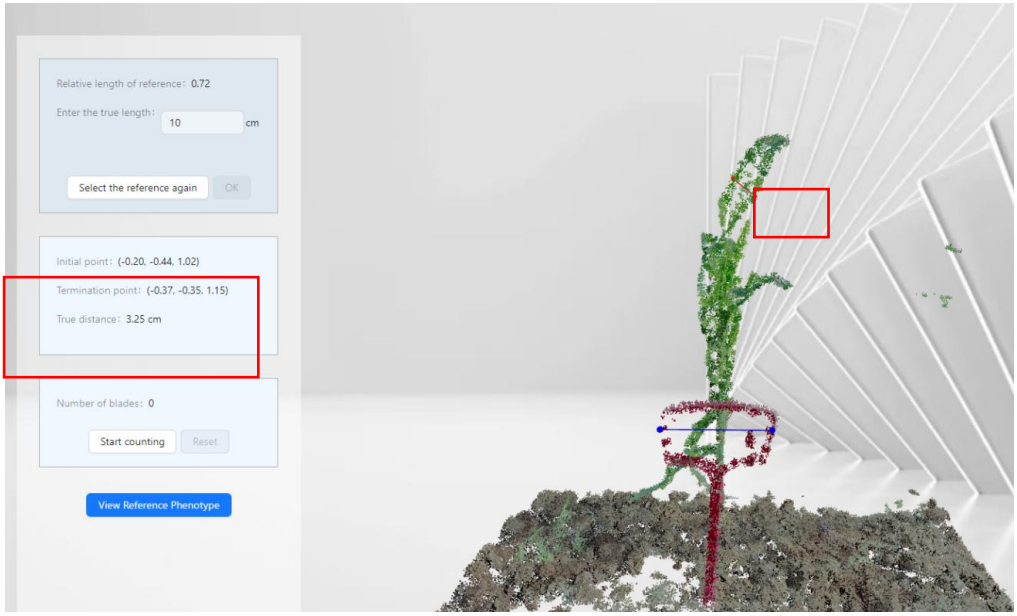
Automated phenotype extraction robustness still needs to be improved, and users can choose to use interactive phenotype extraction for further analysis.

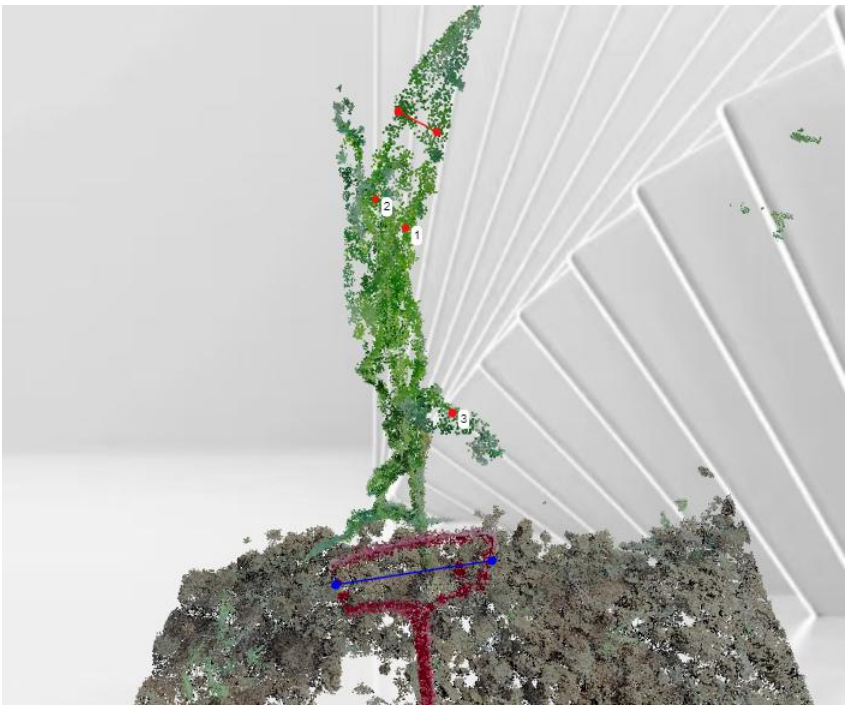
Interactive phenotype Extraction:

- After clicking the "Interactive Extraction" button on the left end of the page, the user can select the key points in the three-dimensional point cloud for interaction through the mouse.
- First, the mouse clicks on both sides of the Corndata1 "marker" to obtain the relative length of the marker, and enters the real length of 10cm in the upper left corner. From clicking OK, the conversion ratio can be generated.



- Subsequently, we can interactively click the leaf width and plant higher phenotype, as shown in the red line segment below. We can also click the "Start counting" button and click the leaf to count the leaves. In counting mode, the points will be displayed on the plant point cloud.





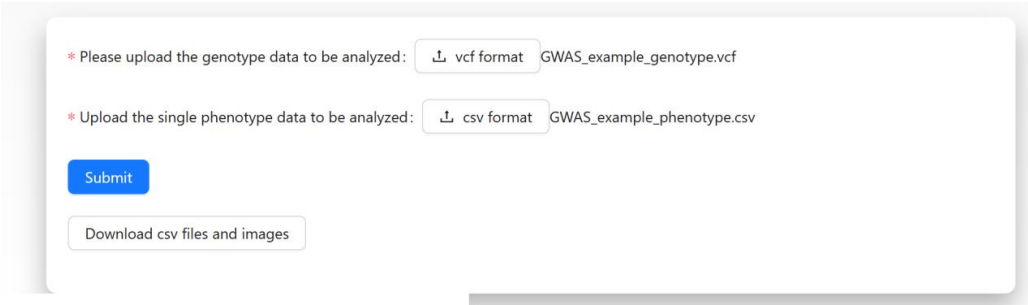
4. Data analysis

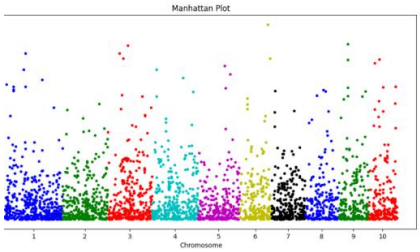
GWAS analyse

The platform provides sample data:

- GWAS_example_genotype.vcf
- GWAS_example_genotype. csv

By extracting the VCF files and CSV files of genes and clicking Submit, users can obtain the P-value of the SNP and the corresponding Manhattan diagram.[P-value csv files corresponding to Manhattan plot and each SNP]



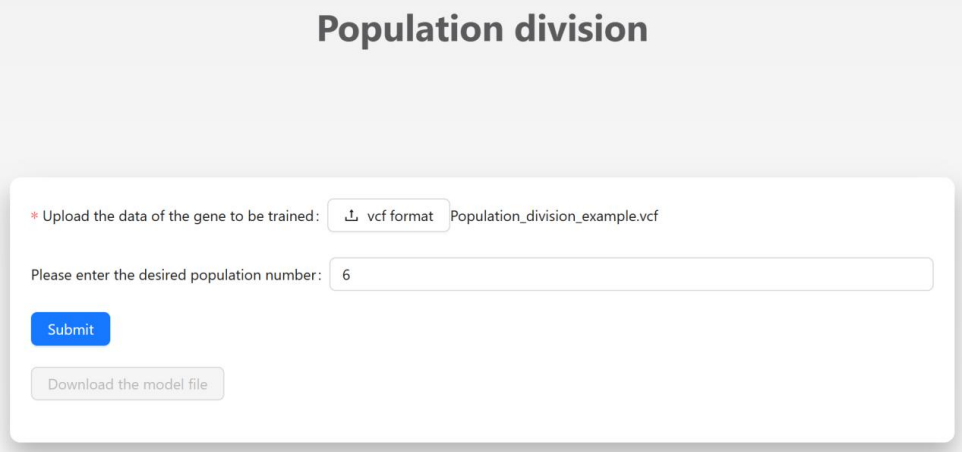


Group division

The platform provides the sample data:

- Population_division_example.vcf

By extracting the VCF file of the gene and the "expected number of group division scale", click the corresponding group division distribution map, and the CSV file of the corresponding group details.[Download the CSV file with the distribution map and group details]

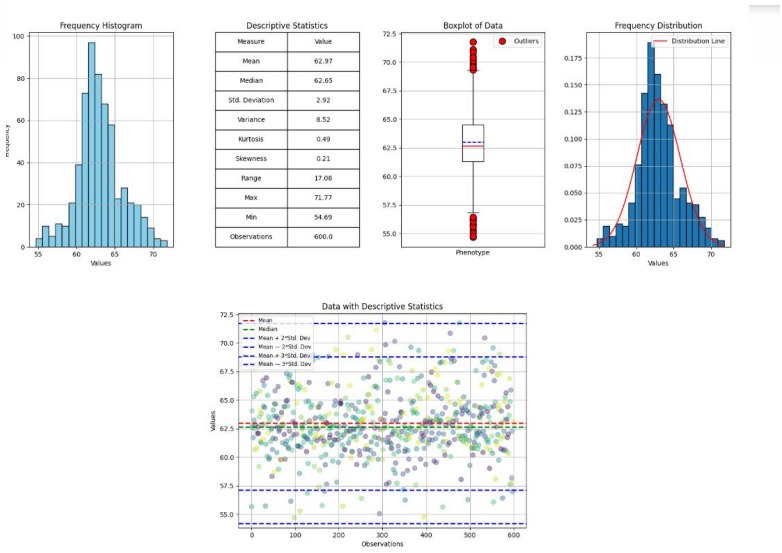


Analysis of phenotypic data

The platform provides the sample data:

- Data_analysis_example.csv [The csv file of any GS has rows]

Through the CSV file, users can obtain the corresponding descriptive analysis map, so as to intuitively view the basic information of the data.



5. Genomic selection

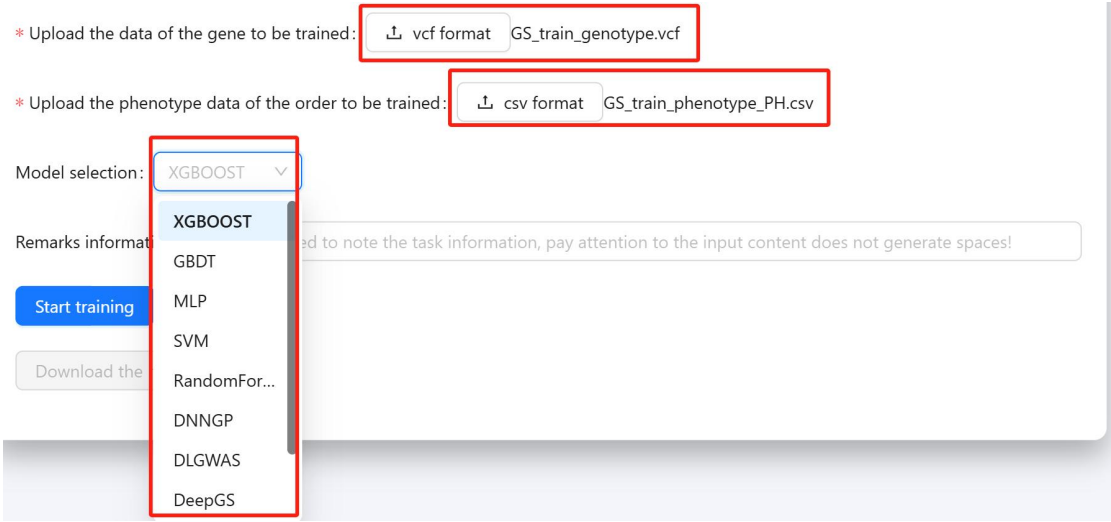
Model training

The platform provides sample data: [in order]

- GS_train_genotype.vcf
- GS_train_phenotype_PH.csv

Users can obtain the corresponding model file by submitting the file (and completing other deployments) and clicking on the submission.

[Note: The output is of. pth.zip, No decompression operation, can be directly used as a weight file for subsequent input operation]

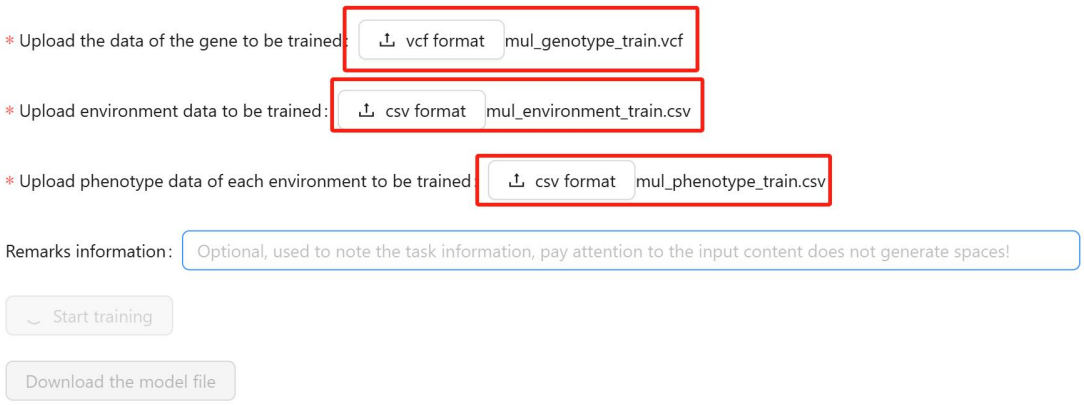


Model training (including environmental information)

The platform provides sample data: [in order]

- mul_genotype_train.vcf
- mul_environment_train.csv [Difference point from "model training"]
- mul_phenotype_train.csv [Including environmental information, so you can fill in the phenotypes corresponding to several environments]

Users can obtain the corresponding model file by submitting the file (and completing other deployments) and clicking on the submission.

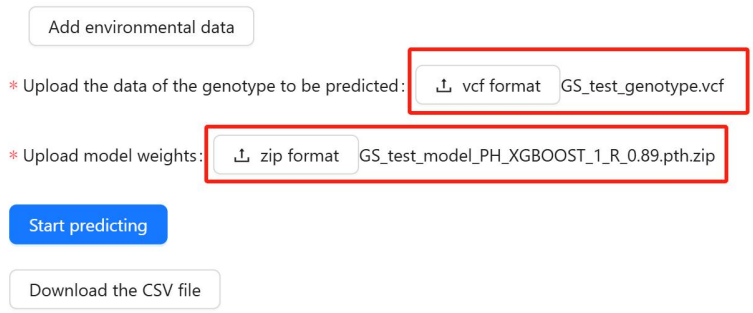


Phenotype prediction

The platform provides sample data: [in order]

- GS_test_genotype.vcf
- GS_test_model.pth.zip

[Use the platform to output weight files and avoid the file name (1) (2)]



Users can submit the file (and complete the other deployment), and click submit to get the corresponding predicted csv file.

[Note: The output csv file can be visualized in "Phenotype Data Analysis"]

id	predict
MG_115_X_MG_1528	238.24701
MG_991_X_MG_1524	257.93466
MG_162_X_MG_1540	274.8416
MG_204_X_MG_1520	270.39908
MG_68_X_MG_1545	262.42966
MG_621_X_MG_1532	241.12639
F349_X_MG_1535	244.07361
MG_204_X_MG_1536	242.68143
MG_923_X_MG_1541	361.16336
MG_1303_X_MG_1527	263.64697
MG_556_X_MG_1535	239.821
MG_447_X_MG_1518	270.0877

Phenotype prediction (including environmental information)

The platform provides sample data: [in order]

- mul_test_genotype.vcf
- mul_environment_test.csv [Difference point from "phenotype prediction"]
- mul_test_model.pth.zip [The model source must come from the "model training (including environmental information)"]

Users can obtain the corresponding results by submitting the document (and completing other deployments) and clicking Submit.

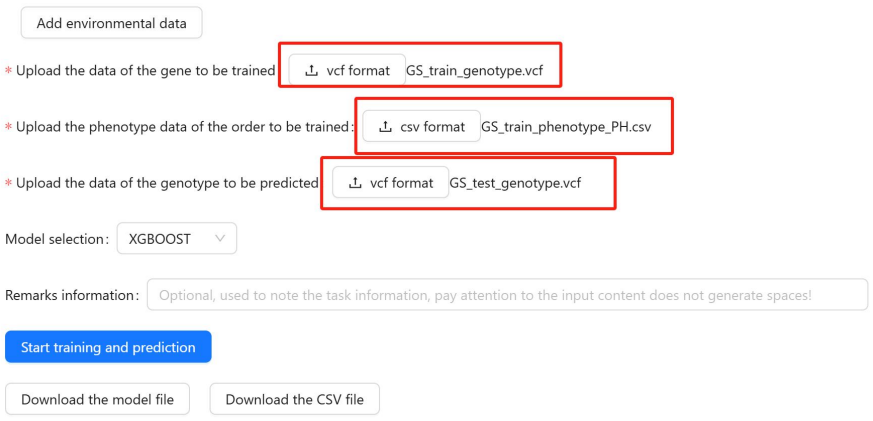
ID	BJ	HeB	JL	LN	HN
MG_115_X_MG_1528	17.143845	17.152086	17.18957	17.165375	17.111414
MG_991_X_MG_1524	17.140156	17.147812	17.185732	17.160892	17.109158
MG_162_X_MG_1540	17.136444	17.144053	17.18276	17.157333	17.105593
MG_204_X_MG_1520	17.13905	17.146381	17.184267	17.159435	17.10772
MG_68_X_MG_1545	17.136644	17.144276	17.183714	17.158031	17.106506
MG_621_X_MG_1532	17.13686	17.144392	17.180227	17.156788	17.107155
F349_X_MG_1535	17.14204	17.149588	17.18623	17.162268	17.111845
MG_204_X_MG_1536	17.138123	17.145449	17.180069	17.157507	17.109976
MG_923_X_MG_1541	17.138506	17.145596	17.180561	17.15708	17.107817
MG_1303_X_MG_1527	17.136435	17.14428	17.183271	17.157352	17.105667
MG_556_X_MG_1535	17.14391	17.152077	17.189358	17.165434	17.112888
MG_447_X_MG_1518	17.138042	17.145672	17.182762	17.158417	17.10751
MG_631_X_MG_1539	17.143354	17.15051	17.186666	17.162378	17.112217
MG_1236_X_MG_1526	17.13969	17.148052	17.192438	17.163284	17.108932
MG_1242_X_MG_1544	17.137096	17.14543	17.187044	17.159653	17.105282

Training and prediction integration

The platform provides sample data: [in order]

- GS_train_genotype.vcf
- GS_train_phenotype_PH.csv
- GS_test_genotype.vcf

By submitting documents (and completing other deployments), users can get the corresponding weight file and the prediction phenotype file.



Training and prediction integration (including environmental information)

The platform provides sample data: [in order]

- mul_genotype_train.vcf
- mul_environment_train.csv [Difference point from "training and prediction integration"]
- mul_phenotype_train.csv
- mul_genotype_test.vcf

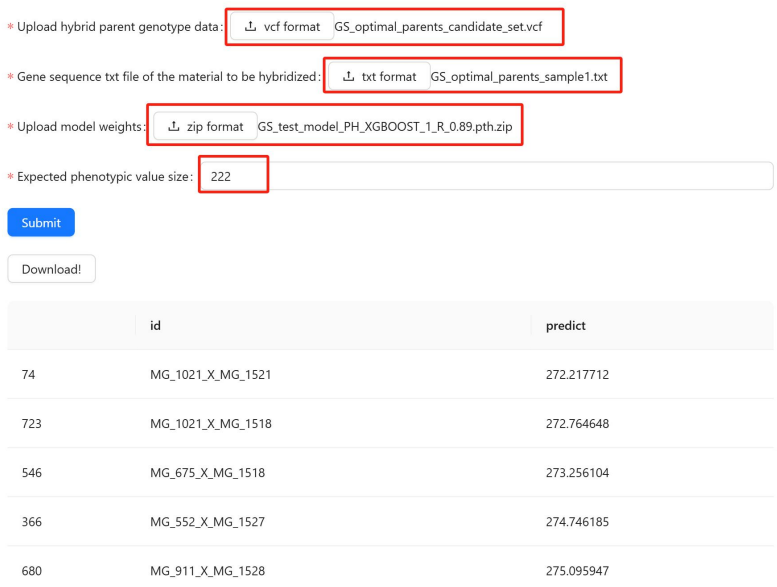
Users can obtain the corresponding results by submitting the document (and completing other deployments) and clicking Submit.

Selection of optimal parents

The platform provides sample data: [in order]

- GS_optimal_parents_candidate.vcf
- GS_optimal_parents_sample1.txt
- GS_test_model.pth.zip
- Prediction correlation value: 55

Users can obtain the corresponding results by submitting the document (and completing other deployments) and clicking Submit.



	id	predict
74	MG_1021_X_MG_1521	272.217712
723	MG_1021_X_MG_1518	272.764648
546	MG_675_X_MG_1518	273.256104
366	MG_552_X_MG_1527	274.746185
680	MG_911_X_MG_1528	275.095947

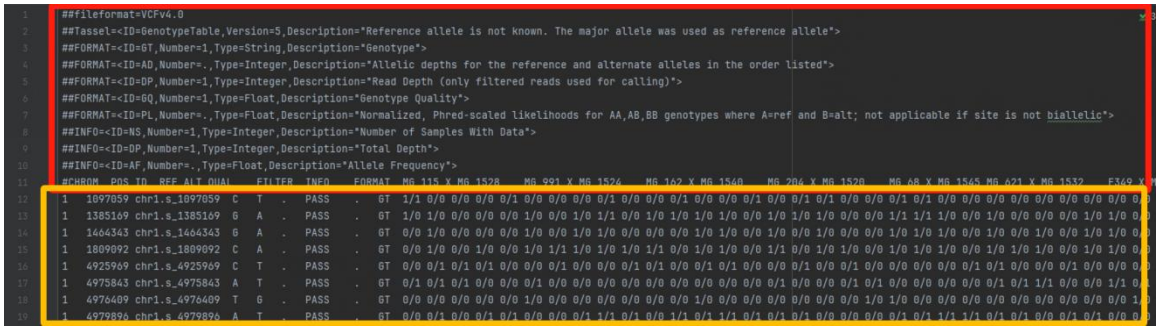
supplementary specification

1. Data format requirements

In addition, the following strict requirements for the format of vcf files, csv files, and txt files for the AutoGP platform:

VCF document

Corresponding to the vcf files: rows are represented as each SNP and columns as each sample.



The vcf file top 11 acts remarks in conventional vcf format, starting from line 12 for the read information, where the first 9 columns are remarks about the SNP data, column 10 starts each column corresponds to the details of each material on each SNP, where "0 / 0", "1 / 0", "0 / 1", "1 / 1" indicates no allele mutation, single mutation, double mutation, respectively.

CSV document

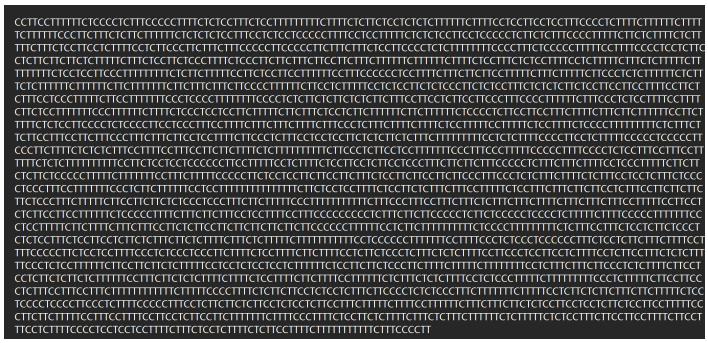
Corresponding to the csv file: rows are represented as each sample and columns as the phenotype.

The csv file, each row in the first column is represented as the sample name (this sample order should be consistent with the sample order in line 10 in the VCF file), and each row in the second column is represented as the phenotype value corresponding to the sample.

Note: Currently, users only need to provide csv format files in the above two-column formats.

Lineid	DTT
MG_115_X_MG_1528	55
MG_991_X_MG_1524	64
MG_162_X_MG_1540	63
MG_204_X_MG_1520	64
MG_68_X_MG_1545	63
MG_621_X_MG_1532	65
F349_X_MG_1535	63
MG_204_X_MG_1536	60
MG_923_X_MG_1541	68
MG_1303_X_MG_1527	63
MG_556_X_MG_1535	63
MG_447_X_MG_1518	66
MG_631_X_MG_1539	
MG_1236_X_MG_1526	60
MG_1242_X_MG_1544	63
MG_891_X_MG_1533	62
MG_552_X_MG_1546	62
MG_161_X_MG_1536	59
MG_522_X_MG_1520	63
MG_220_X_MG_1525	66
MG_447_X_MG_1523	64

TXT document



The txt file, representing the gene sequence formed from the corresponding SNP splicing of a single sample to be predicted. For example, if 20000 SNP are involved in vcf, the txt file corresponding to a single sample represents a gene sequence of 20000 in length with 20000 SNP splicing.

2. Details of the algorithm design

In order to give GS researchers more details about the algorithm architecture design, the platform details the algorithm architecture design:

SVM

In this platform, if the support vector machine (SVM) is selected as the model configuration, the system will automatically use SVR (support vector regression) for training and prediction. This article did not do much operation, only use model = SVR (verbose=0) model.fit (train_x, train_y) for related training.

XGBOOST

In this platform, if XGBoost is selected as the model configuration, the system will be trained with XGBRegressor and perform parameter tuning by grid search.first,

Define the initial parameters of the model, Where the learning rate (learning_rate) is set to 0.05, The initial number of decision trees (n_estimators) is 500, Maximum depth (max _ tap) is 5, The minimum child node weight (min_child_weight) is 1, Random seed (seed) is 0, The subsample proportion (subsample) was 0.8, The column sampling ratio (colsample_bytree) is 0.8, The minimum loss function drop value (gamma) required to split the nodes of the tree is 0, The L1 regularization coefficient (reg _ alpha) is 0, The L2 regularization coefficient (reg _ lambda) is 1. Meanwhile, the parameter range of the grid search is set, and the number of candidate decision trees (n_estimators) is 500,600 and 700. Then, a XGBRegressor model object is initialized using these parameters. Next, the model underwent parameter tuning by GridSearchCV and the evaluation index was R^2 and 5-fold cross-validation. Finally, the tuned model is trained using the provided training data, and the details are output during the training process. Through the above steps, the XGBoost model completes the initialization, parameter tuning and training, which can be used for subsequent data prediction tasks.

GBDT

In this platform, if the gradient lifting decision tree (GBDT) is selected as the model configuration, the system will be trained with GradientBoostingRegressor. First, the system initializes a GradientBoostingRegressor model object where the random seed (random_state) is set to 123 to ensure the repeatability of the results and set verbose to 0 to turn off the detail output during the training process. The system then trains the model using the provided training data (train_x and train_y). Through these steps, the GBDT model completes the initialization and training, which can be used for subsequent data prediction tasks.

MLP

In this platform, if the multi-layer perceptron (MLP) is selected as the model configuration, the system will be trained with the MLPRegressor. First, the system will initialize a MLPRegressor model object, where the structure of the hidden layer (hidden_layer_sizes) is set to two layers, the first layer has 3060 neurons, the second layer has 64 neurons, activation function (activation) selects ReLU, optimization

algorithm (solver) selects Adam, L2, the parameter (alpha) of regularization term is set to 0.01 and the maximum number of iterations (max_iter) is set to 200. With these steps, the MLP model completes the initialization and can be used for subsequent data training and prediction tasks.

RF

In this platform, if the random forest (RandomForest) is selected as the model configuration, the system will be trained with the RandomForestRegressor. First, the system initializes an RandomForestRegressor model object, where the number of decision trees (n_estimators) is set to 20. The system then trains the model using the provided training data (train_x and train_y). Through these steps, the random forest model completes the initialization and training and can be used for subsequent data prediction tasks. The design of this platform ensures the flexibility and ease of use of the model configuration, allowing researchers to focus on the research itself without paying much attention to the complexity of the underlying implementation.

DeepGS、 DLGWAS、 DNNGP、 SoyDNGP

This platform reproduced all the DL models following the source paper architecture, and designed them to ensure that its architecture can be reused for different size matrices.

The following is the configuration situation for the other parameters:

The batch size during training is set by the train_batch parameter and the default value is 64; the batch size for validation is set by the variable _ batch parameter and the default value is 64. The total number of rounds for training is specified by the epochs parameter, with a default value of 28. To ensure repeatability of the results, random seeds can be set using the seed parameter with a default value of 6. The learning rate is controlled by the lr parameter with a default value of 0.001, while the weight decay coefficient was used for L2 regularization, set by the weight_decay parameter with a default value of 1e-5.

In the model training, a smooth L1 loss function was used to calculate the loss. The optimizer selects Adam, whose learning rate and weight decay coefficient are specified by the lr and weight_decay parameters, respectively. To adjust the learning

rate, a stepped learning rate scheduler was used, which reduces the learning rate to 10% per 10 epoch. With these configurations and settings, the training process of the model can be flexibly adjusted to demand to achieve better performance and results.

frequently asked questions

If you have any suggestions or questions about AutoGP, please contact us via the feedback provided by the platform or via email.

This is currently the V2 version, and you want this version to meet your requirements.

contact us

If you have any suggestions or questions about AutoGP, please contact us by:

- mail box:**wuh_hzau@163.com**
- Address: School of Information, Huazhong Agricultural University, Wuhan City, Hubei Province
- telephone:

Before contacting us, you can visit our frequently Asked Questions (FAQ) page to see if you questions any answered.