User Manual

1 Introduction

This manual will guide users to deploy and operate AutoMolDesigner V1.1 with graphical user interface (GUI). Briefly, it implements automated molecular design through focused library generation by deep molecular generation and machine learning-based virtual screening. It is based on a chemical language model (ACS Cent. Sci. 2018, 4, 120-131) and an open-source automated machine learning (AutoML) framework named AutoGluon (https://auto.gluon.ai/) developed by Amazon Co. Ltd.,.

2 Instructions

This software has been tested on both Windows and MacOS operating systems. There are central processing unit (CPU) version or graphical processing unit (GPU) version for Windows platform while only CPU version for MacOS platform. There is no functional difference among these versions. Basic requirement on CPU is Intel(R) Core(TM) i5-8400 @ 2.80GHz or AMD Ryzen(TM) 3 3300X @ 3.80GHz or Apple M1 @ 3.20GHz; basic requirement on GPU is NVIDIA GeForce GTX 1070 8GB. The RAM not less than 8 GB is recommended.

If a Windows-based machine is equipped with NVIDIA GPU, the GPU version is highly recommended. This software can be used to detect whether GPU is available

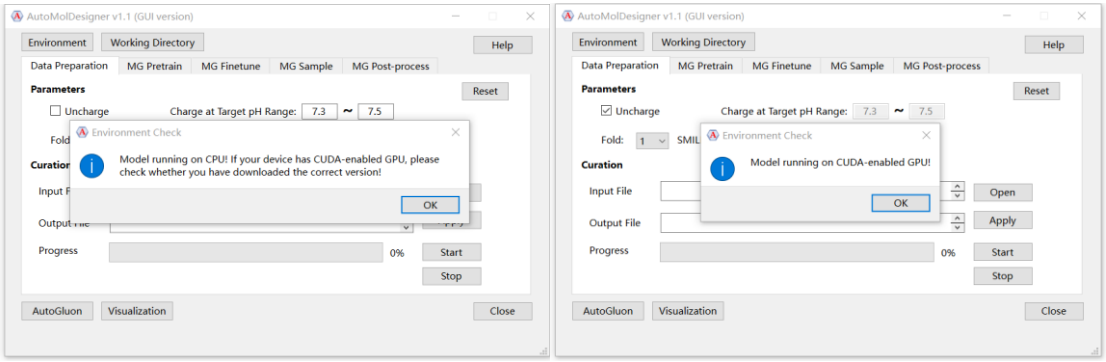


Figure 1. Detect software running environment.

(Figure 1, click the "Environment" button). Please note that the CPU version will only run on CPU while the GPU version can use GPU for acceleration only if it is CUDA-enabled and the latest driver has been installed, otherwise it will run on CPU instead.

Since operating logics are exactly the same for CPU and GPU version, this manual will take Windows-based CPU version for illustration.

2.1 Environmental Settings

(1) Double click "AutoMolDesigner-GUI.exe" under the directory of "\program" to launch the software (it may take seconds depending on the machine). For MacOS

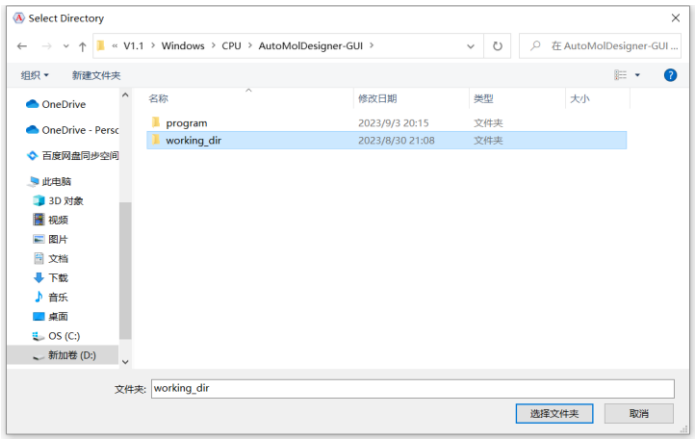


Figure 2. Set working directory.

users, a line of command ("chmod -R 755 program/") needs executing in the shell and apps downloaded from anywhere should be allowed in the system. It may take minutes to launch the program on MacOS.

(2) Click the "Environment" button to detect whether GPU is enabled (Figure 1).

(3) Click the "Working Directory" button to set current working directory (Figure 2). Please note that the program will try to set the directory "\working_dir" as the default working directory. Users can click this button to reset the working directory.

2.2 Functional Modules

This software is comprised of two main functional modules. The first module is deep molecular generation for focused library generation. The main workflow includes "MG Pretrain", "MG Finetune" and "MG Sample" (Nat. Commun. 2023, 14, 114). Moreover, another two supplementary modules named "Data Preparation" and "MG Post-process" are also provided. The second main module is AutoML-based molecular property prediction. This module takes molecular descriptors to characterize molecules and train AutoGluon models for binary classification or regression task. It consists of two sub-modules, i.e., "Model Training" and "Model Prediction". Related data curation can share the "Data Preparation" module.

Herein, this manual will take the molecular design of novel candidate antibiotics for *Escherichia coli* (*E. coli*) as a case study to demonstrate the usage of each module. All related files are available under the directory of "\working_dir\datasets".

2.2.1 Deep Molecular Generation

2.2.1.1 Pretrain

Model pretraining aims at learning general features of drug-like molecules. This module is **NOT** recommended for personal computers due to the high computational cost. A pretrained model is provided under the directory of "\working_dir\projects\ChEMBL32-pretrain-linux" that can be used for model finetuning or sampling. If users plan to pretrain their own models, they should refer to the following instructions:

(1) Use "Data Preparation" tab to prepare raw data. Preparations that will be performed automatically includes：de-duplication, salts removal, molecular standardization and stereochemical type removal. The "Uncharge" box is checked by default to perform charge neutralization or it can be manually unchecked to implement protonation at a certain range of pH. In detail, the lower limit value and upper limit value can be typed into the two adjacent text field (Figure 3). Data augmentation with SMIELS enumeration can be adopted by selecting the specific fold from the dropdown combo box (value 1 denoting no augmentation). Click "Open" button to choose raw molecules input for preparation (herein, the file "pretrain-chembl-mini.smi"), and enter the **COMPLETE** directory and the output file name (herein, named "pretrain-chembl-mini_preped.smi") in the text field below. Click "Apply" button to check the typed directory (Figure 4). Now, click "Start" button to start

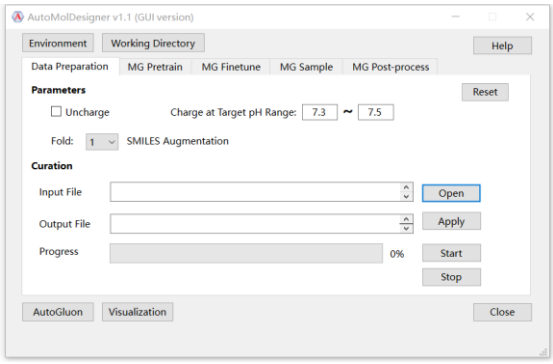


Figure 3. Set "Data Preparation" tab.

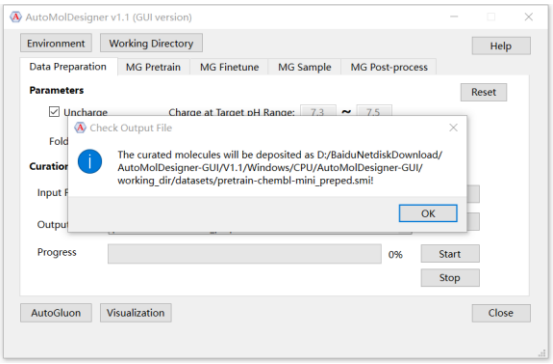


Figure 4. Start preparing.

preparation. When task is running, the "Stop" button can be clicked to manually terminate the process. The "Reset" button locating at the upper right corner can be used to reset this tab.

(2) Use "MG Pretrain" tab to implement model pretraining. Firstly, type project name in the "Project Name" text field (herein, named "Ecoli-Pretrain-Demo") to set "\working_dir\projects\{project name}" as the storage path for all generated data during pretraining. Click "Open" button to load prepared molecules (herein, the file "pretrain-chembl-mini_preped.smi"). Users who desire more flexible net architecture can use the below options for customization while most users are recommended to leave them as default. Now, click "Start" button to initiate model

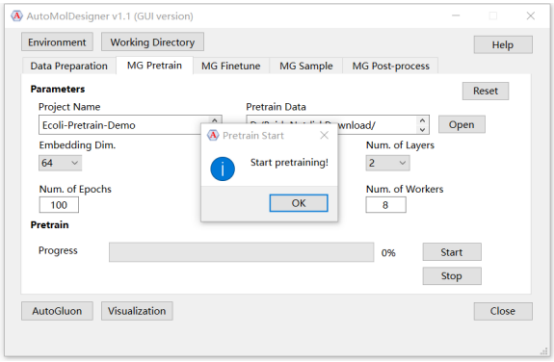


Figure 5. Start pretraining.

pretraining (Figure 5). The best mode parameters will be named as "CLM-pretrain_best-{epoch}-{validation loss}.pkl". Since early stopping has been included, the task may stop before reaching the last epoch. When task is running, the "Stop" button can be clicked to manually terminate the process while it will not stop immediately until the current iteration is completed. The "Reset" button locating at the upper right corner can be used to reset this tab.

2.2.1.2 Finetune

Model finetuning (transfer learning) is an effective strategy to steer the direction of molecular generation towards the chemical space that medicinal chemists are interested in. To perform finetuning, users are required to provide pretrained model and curated dataset consisting of tens to thousands of molecules that have shared features [for example, with bioactivity against a specific target, protein family or even phenotype, and herein, the file "finetune-Ecoli-32ugmL-preped.smi" contains 570 diverse molecules with minimum inhibitory concentration (MIC) ≤ 32μg/mL against *E. coli*]. The instructions are as followed:

(1) Use "Data Preparation" to prepare raw data, which can be referred to section "2.2.1.1 (1)". The dataset for finetuning provided in this case study has been prepared before thus no more preparations will be performed.

(2) Use "MG Finetune" tab to implement model finetuning. Firstly, type project name in the "Project Name" text field (herein, named "Ecoli-Finetune-Demo") to set

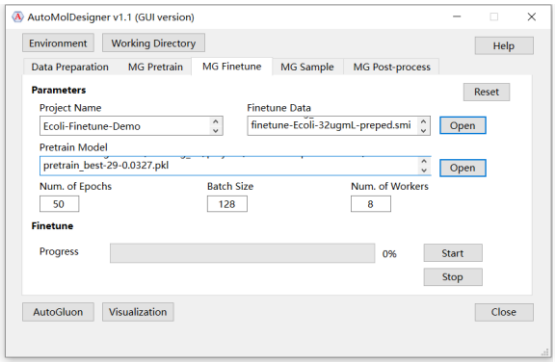


Figure 6. Set "MG Finetune" tab

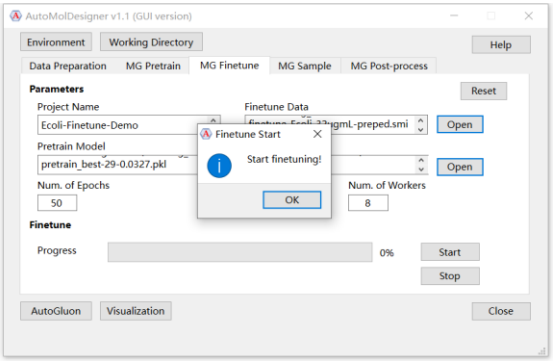


Figure 7. Start finetuning.

"\working_dir\projects\{project name}" as the storage path for all generated data during finetuning. Click "Open" button to load prepared molecules (herein, the file "finetune-Ecoli-32ugmL-preped.smi"). Click another "Open" button below to load pretrained models (*.pkl file, see Figure 6). Users can customize remaining arguments or leave them as default. Now, click "Start" button to initiate model finetuning (Figure 7). The best mode parameters will be named as "CLM-finetune_best-{epoch}-{validation loss}.pkl". Since early stopping has been included, the task may stop before reaching the last epoch. When task is running, the "Stop" button can be clicked to manually terminate the process while it will not stop immediately until the current iteration is completed. The "Reset" button locating at the upper right corner can be used to reset this tab.

2.2.1.3 Sample

Molecular sampling can be used to generate a certain number of molecules. To perform sampling, users are required to provide trained models (*.pkl file, pretrained or finetuned model). The instructions are as followed:

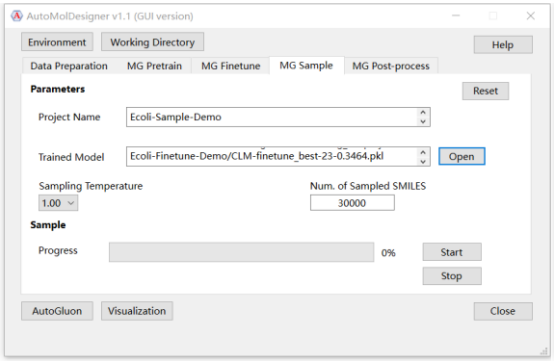


Figure 8. Set "MG sample" tab.

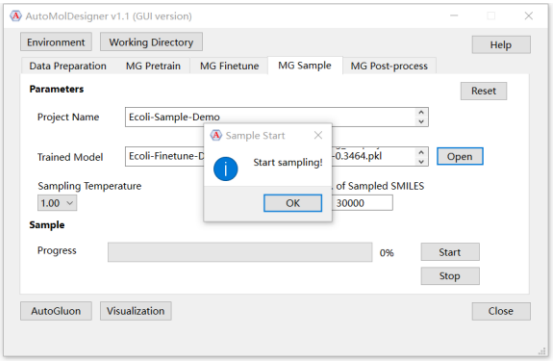


Figure 9. Start molecular sampling.

(1) Use "MG Sample" tab to implement molecular sampling. Firstly, type project name in the "Project Name" text field (herein, named "Ecoli-Sample-Demo") to set "\working_dir\projects\{project name}" as the storage path for all generated data during finetuning. Click "Open" button to load the trained model. In this module, "Sampling Temperature" is introduced to control the scaffold novelty of sampled molecules (Figure 8), which can be referred to the work reported by Gupta et al. (Mol. Inform. 2018, 37, 1700111). Users are recommended to sample hundreds of molecules for testing.

(2) Now, click "Start" button to initiate molecular sampling. The sampled molecules will be saved as the file named "sampled_SMILES.smi" (Figure 9). When task is running, the "Stop" button can be clicked to manually terminate the process. The "Reset" button locating at the upper right corner can be used to reset this tab.

2.2.1.4 Post-process

This module is utilized to post-process generated SMILES to remove invalid and duplicate molecules, with three metrics including "Validity", "Uniqueness" and

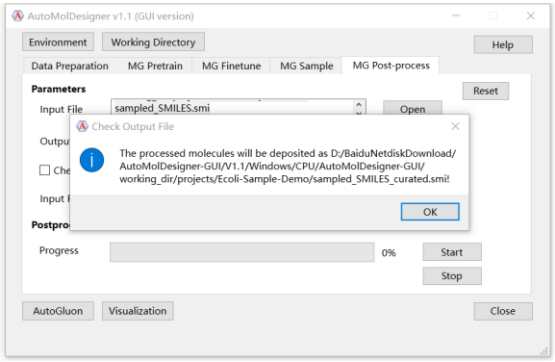


Figure 10. Set "MG Post-process" tab.

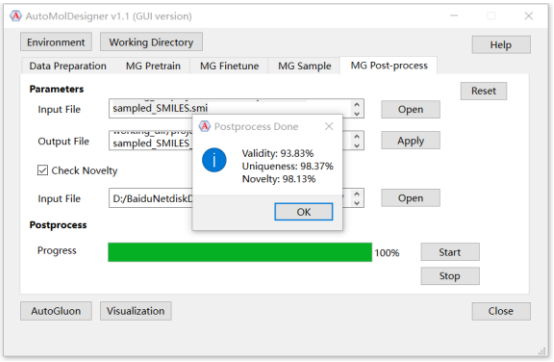


Figure 11. Validity, Uniqueness and Novelty.

"Novelty" reported (Front. Pharmacol. 2020, 11, 565644). The instructions are as followed:

(1) Set the files for input and output. Firstly, click "Open" button to load raw SMILES (herein, the file "sampled_SMILES.smi"), and enter the output file name (herein, named "sampled_SMILES_curated.smi") in the text field below. Click "Apply" button to check the typed directory (Figure 10). Now, click "Start" to start post-processing. A message box showing the values of "Validity" and "Uniqueness" will pop up upon task finishes.

(2) If metric "Novelty" is required, the "Check Novelty" check box should be firstly checked which will enable the "Open" button that can be used to load molecules for finetuning (herein, the file "finetune-Ecoli-32ugmL-preped.smi"). Now, click "Start" to start post-processing. A message box showing the values of "Validity" and "Uniqueness" alongside "Novelty" will pop up upon task finishes (Figure 11).

2.2.2 AutoML-based Molecular Property Prediction

2.2.2.1 Model Training

Click the "AutoGluon" button locating at the lower left corner of main interface to launch the "AutoGluon" tab. This module can be used to train machine learning-based molecular property prediction models for benchmarking or new data prediction. AutoGluon is able to complete all modeling work automatically by implementing feature engineering, model architecture selection, hyperparameter optimization and ensemble learning. In this case study, we will walk through the construction of a binary classification model for phenotypic antibacterial activity prediction. The input file for model training should be tabular data, of which the format can be Comma-Separated Values (*.csv) or Microsoft Excel (*.xlsx). Herein, we made a maximal unbiased benchmarking dataset (MUBD) based on known antibiotics with high activity (MIC ≤ 1μg/mL) against *E. coli* using MUBD-DecoyMaker2.0 (Mol. Inform. 2020, 39, 1000151) which is available under directory of "\working_dir\datasets". It was split into the training set ("MIC_Ecoli_train.xlsx") and test set ("MIC_Ecoli_test.xlsx") to

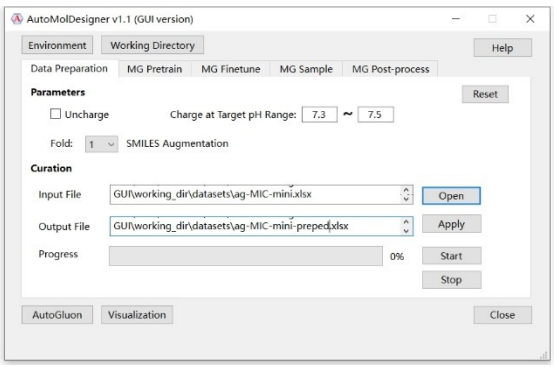


Figure 12. Prepare training data.

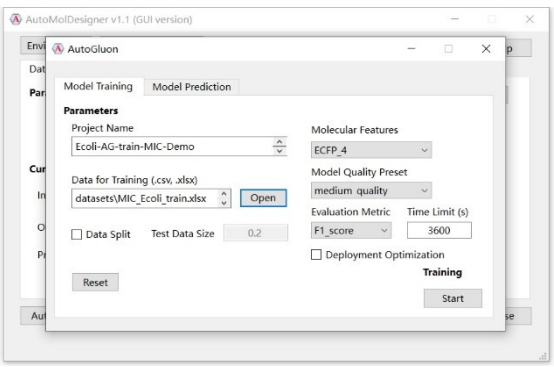


Figure 13. Set model training.

demonstrate two modes – retrospective training denoting that trained model will be used for benchmarking and prospective training denoting that trained model will be used for new data prediction. The instructions are as followed:

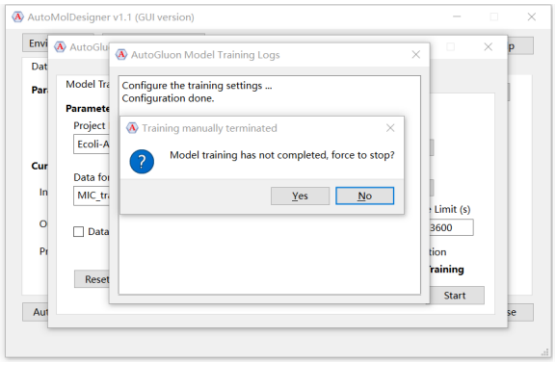


Figure 14. Terminate model training.

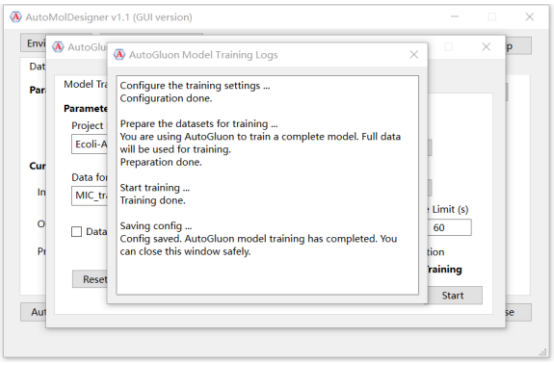


Figure 15. Model training completed.

(1) Use “Data Preparation” tab to prepare raw molecules, which can be referred to section “2.2.1.1 (1)”. This module also supports csv file and Microsoft Excel file (herein, the file “ag-MIC-mini.xlsx” is provided as a test case for preparation, Figure 12). In this case study, the supplied datasets were all prepared before.

(2) Use “Model Training” tab in “AutoGluon” interface to train machine learning models. Firstly, type project name in the “Project Name” text field (herein, named “Ecoli-AG-Train-MIC-Demo”) to set “\working_dir\projects\{project name}” as the storage path for all generated data. Click “Open” button to select dataset for model training (herein, the file “MIC_Ecoli_train.xlsx”, Figure 13). Users can exploit dropdown combo boxes, text field and checkboxes to customize the training configuration. Table 1 lists all available arguments with their descriptions.

Table 1. Optional arguments in AutoGluon model training with brief descriptions.

Argument	Description
Molecular Features	Computational descriptors used to characterize molecules. Five classes are provided - RDKit_2D and its normalized version, ECFP_4 (Morgan2), FCFP_6 (featured Morgan3) and MACCS structural keys. Their definitions can be referred to RDKit Documentation.
Model Quality Preset	The preset quality of AutoGluon models. It has four levels ranging from low to high, i.e., "medium_quality", "good_quality", "high_quality" and "best_quality". Detailed descriptions can be referred to AutoGluon Documentation.
Evaluation Metric	The metric function used for training. For binary classification, F1 score and ROC_AUC are provided; For regression, MAE and RMSE are provided.
Time Limit (seconds)	Maximum amount of time for model training. The training process may stop early before the amount of time is met.
Deployment Optimization	Whether to perform optimization for model deployment. Detailed descriptions can be referred to AutoGluon Documentation.
Data Split & Test Data Size	Whether to hold out test set from input data and the size of test set. Being check denoting retrospective training, and three files – "ag_binary.csv", "ag_train.csv" and "ag_test.csv" will be generated under the directory of "\working_dir\projects\{project

name}", namely only the file "ag_train.csv" will be used for training. Default size of test set is 20% of the input data. Being uncheck denoting prospective training, and only the file "ag_binary.csv" will be generated，namely all input data will be used for training.(If the user has prepared the training and testing sets before, leave it unchecked)

(3) Click "Start" button to initiate model training. Herein, retrospective training will be conducted. Considering that training and testing sets were prepared and split before, the "Data Split" box is left unchecked. A logging dialogue will pop up during model training. If the user tries to close this dialogue before the task is completed, a message box will pop up as a reminder and the task will be manually terminated provided that the further confirmation is made (Figure 14). The user can close this dialogue when it reminds the task has finished (Figure 15). All trained models will be under the directory of "\working_dir\projects\{project name}\ag_models". The "Reset" button locating at the lower left corner can be

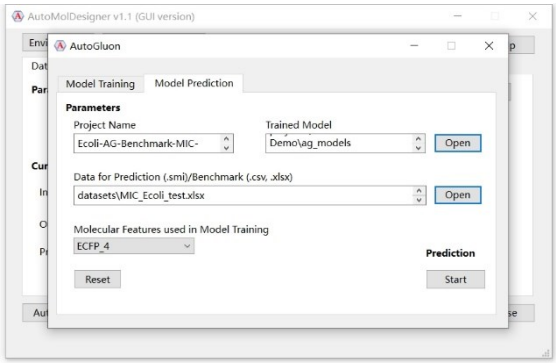


Figure 16. Set "Model Prediction" tab for retrospective prediction.

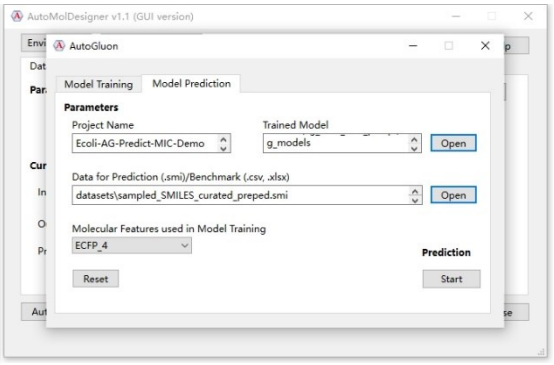


Figure 17. Set "Model Prediction" tab for prospective prediction.

used to reset this tab.

2.2.2.2 Model Prediction

Switch to “Model Prediction” tab in the “AutoGluon” interface. This module aims at assessing model performance (retrospective prediction) or predicting the properties of new molecules (prospective prediction). Users should provide the directory of trained AutoGluon models and datasets for prediction. The instructions are as followed:

(1) Use “Data Preparation” tab to prepare raw molecules, which can be referred to section “2.2.1.1 (1)”. If the model benchmarking is conducted on the test set that is automatically split during model training, this step can be ignored. Otherwise, the data supplied by users themselves should be prepared. Note that the file “MIC_Ecoli_test.xlsx” used here was prepared before.

(2) Use “Model Prediction” tab to predict data. If retrospective prediction is adopted, type project name in the “Project Name” text field (herein, named “Ecoli-AG-Benchmark-MIC-Demo”). Click “Open” button to load the directory of trained models (for example, the directory of “\working_dir\projects\Ecoli-AG-Train-Demo\ag_models”, Figure 16). Next, click “Open” button below to load the test data (can be the file “ag_test.csv” generated in section “2.2.2.1”, herein, the test data is “MIC_test.xlsx”) and choose the descriptor used for model training with the dropdown combo box. If prospective prediction is adopted, type project name in the “Project Name” text field (herein, named “Ecoli-AG-Predict-MIC-Demo”),

and load the model that is trained on full dataset. The new data for prediction can be the molecules generated at the section "2.2.1.4" (herein, the file "sampled_SMILES_curated.smi" was further prepared in "Data preparation" tab to give the file "sampled_SMILES_curated_preped.smi", Figure17).

(3) Click "Start" button to initiate model prediction. A logging dialogue will pop up during prediction. If the user tries to close this dialogue before the task is completed, a message box will pop up as a reminder and the task will be manually terminated provided that the further confirmation is made (Figure 18). The user can close this dialogue when it reminds the task has finished (Figure 19). The prediction results can be different according to the mode. If retrospective prediction is adopted, a file named "metrics.csv" will generated under the directory of "\working_dir\projects\{project name}". For binary classification tasks, it records the model performance measured by five metrics including confusion matrix, accuracy, area under receiver operating characteristics (AUROC), Matthews

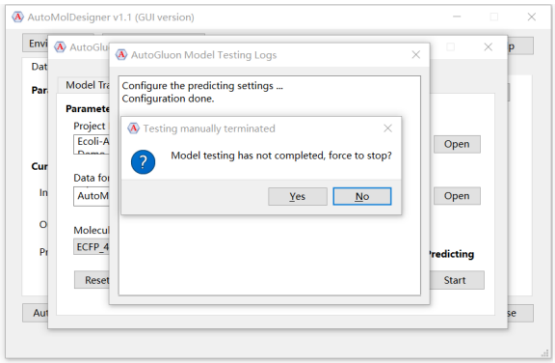


Figure 18. Terminate model prediction.

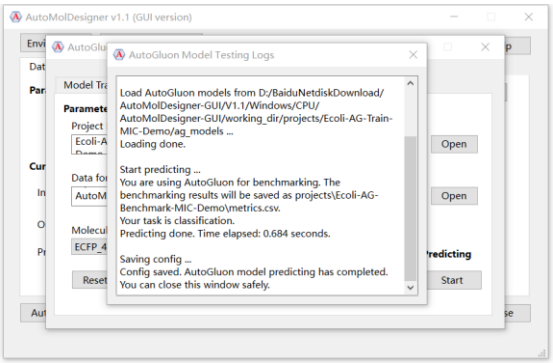


Figure 19. Model prediction completed.

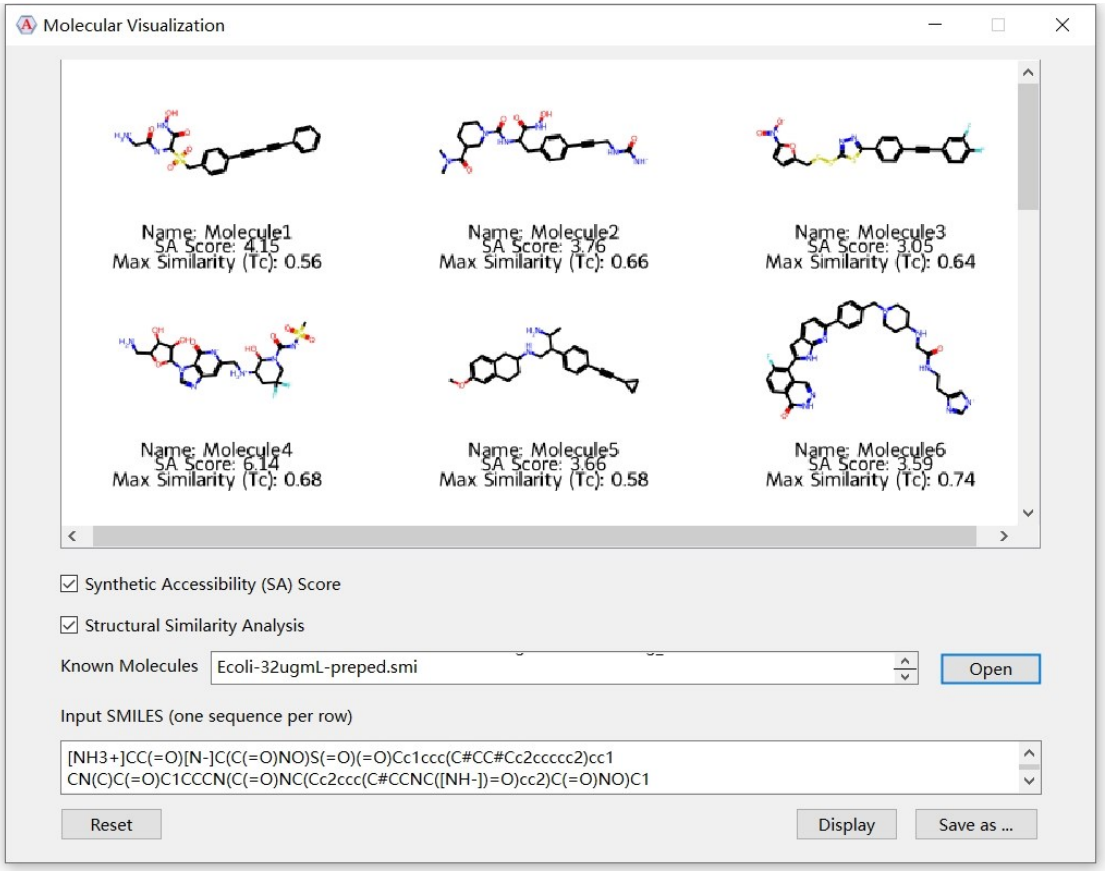


Figure 20. Molecular visualization.

Correlation Coefficient (MCC) and F1 Score. For regression tasks, the metrics are mean absolute error (MAE), root mean squared error (RMSE), *R* square (R^2) and median absolute error (MedianAE). If prospective prediction is adopted, two files including "pred_results.csv" and "pred_results.sdf" will be dumped under the same directory. The csv file records the predicted molecules represented by SMILES and their scores (predicted positive probability or numerical regressive value) while the additional Structural Data File (*.sdf) can be further visualized by molecular informatics tool such as Discovery Studio The "Reset" button locating at the lower left corner can be used to reset this tab.

2.3 Molecular Visualization

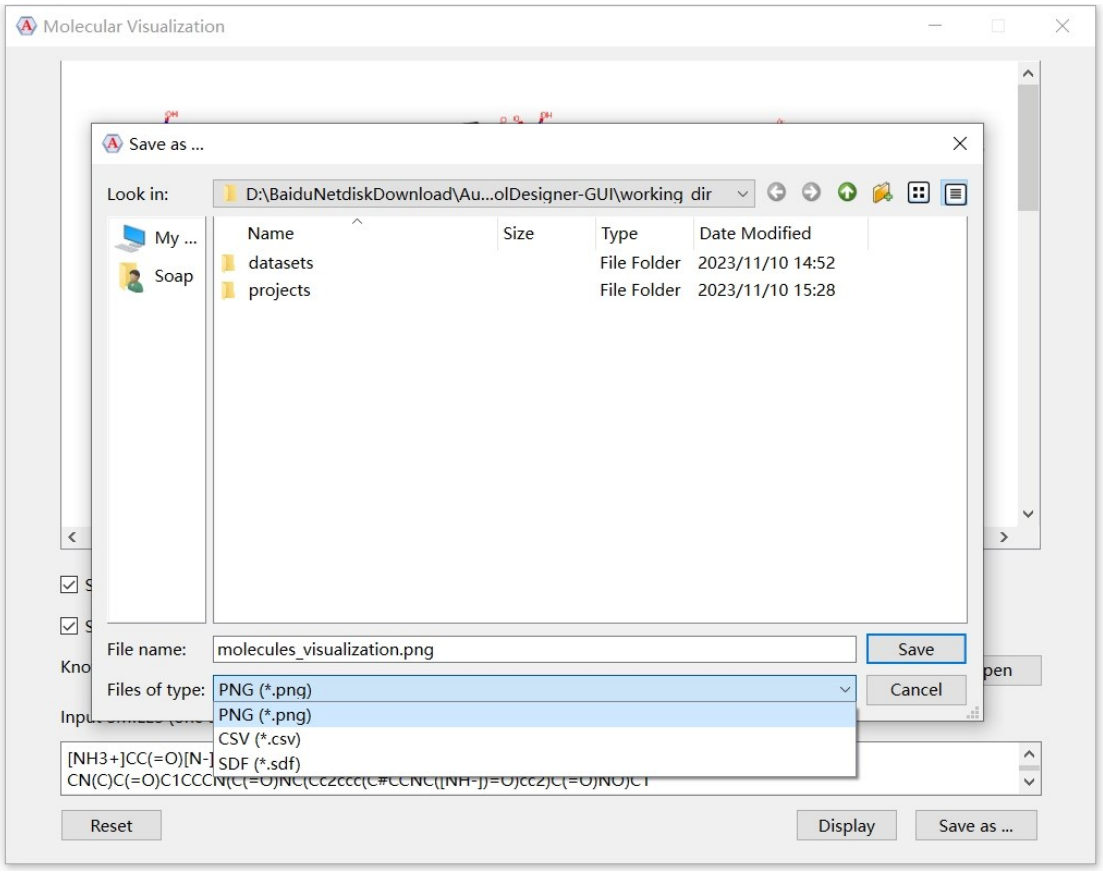


Figure 21. Save the visualization result.

Users can visualize the generated molecules through the "Molecular Visualization" tab and perform extra calculations. The instructions are as followed:

(1) Click the "Visualization" tab to pop up this sub-tab. Users can enter molecules in the text field below (one sequence per row, and click "Display" to view the corresponding molecules.

(2) Users can calculate the synthetic accessibility score (J. Cheminform. 2009, 1, 8) of the input molecules by checking the corresponding box. Moreover, users can perform structural similarity analysis by comparing with the known molecules. The

application will record the closest molecule to the query molecule and the maximum Tanimoto coefficient (Figure 20).

(3) The quality of bitmap displayed in this tab is limited due to the restricted size of graphical box. However, by clicking "Save as …" button, users can obtain the high-resolution bitmap through saving locally. Users can also save the visualized molecules as other formats (.csv, .sdf). The "Reset" button locating at the lower left corner can be used to reset this tab (Figure 21).

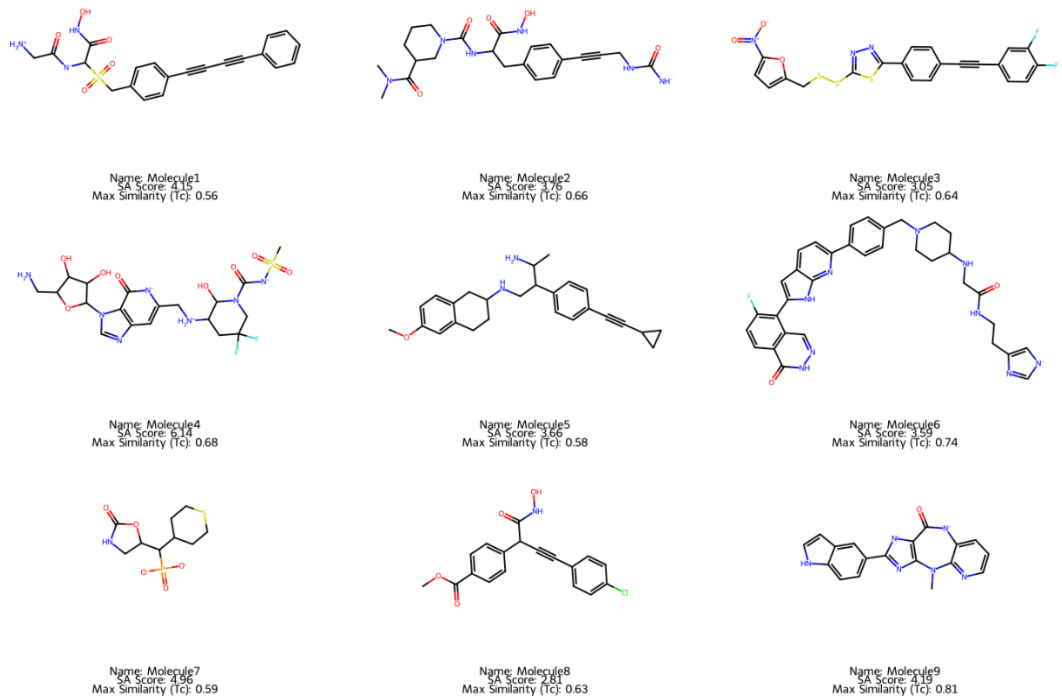


Figure 22. 9 molecules designed in this case study.

Figure 22 displays 9 molecules with high predicted antibacterial activity against *E. coli* by means of this module.

3 References

1. Segler, M. et al. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* 4, 120-131 (2018).

2. Gupta, A. et al. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* 37, 1700111 (2018).

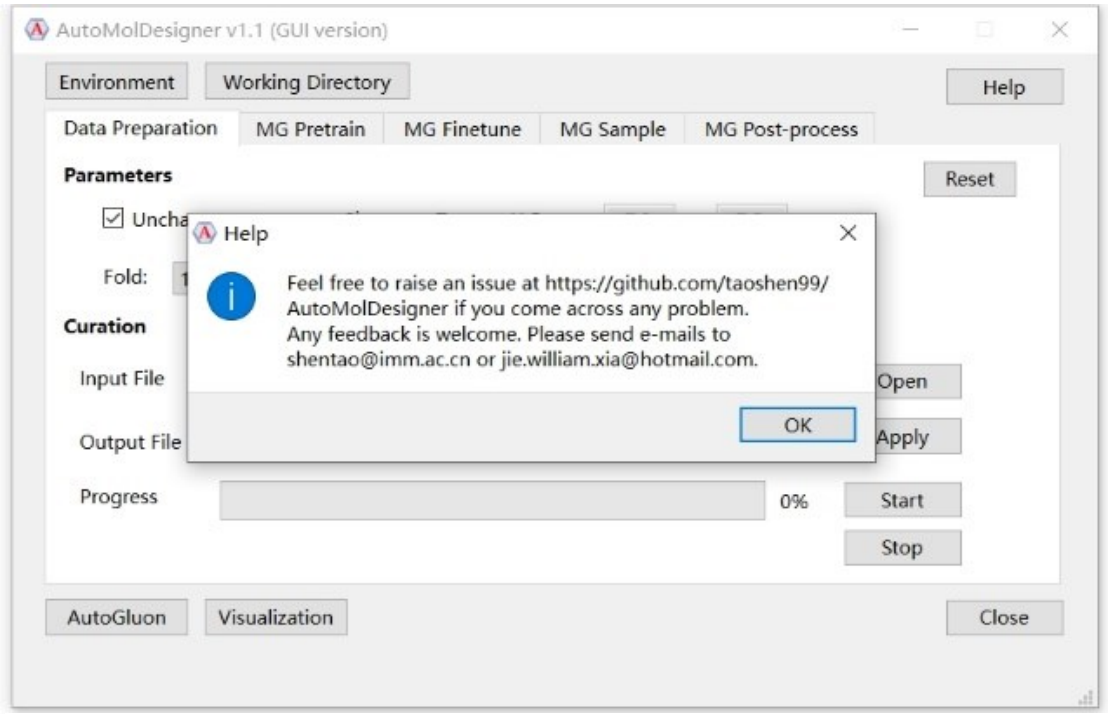
3. Polykovskiy, D. et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* 11, 565644 (2020).

4. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* 14, 114 (2023).

5. Xia, J. et al. MUBD-DecoyMaker 2.0: A Python GUI Application to Generate Maximal Unbiased Benchmarking Data Sets for Virtual Drug Screening. *Mol. Inform.* 39, 1900151 (2020).

6. Ertl, P. et al. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* 1, 8 (2009).

PS: Users can click “Help” button to report any problem or provided feedbacks to the software developers. Licenses of the open-source packages used in this software can be found at “\program\LICENSE”.



Appended figure: contact information of software developers.