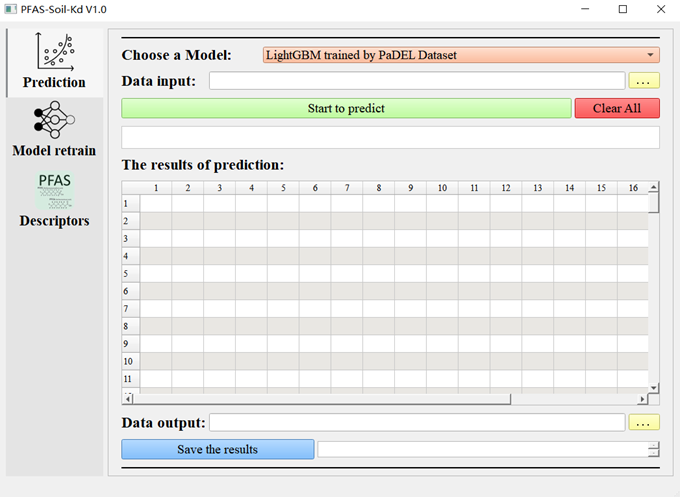
**PFAS-Soil-Kd V1.0**

**Software User Manual**

Running this software does not require installation. Simply find the executable file "PFAS-Soil-Kd V1.0.exe" in the folde r, double-click to run it, and the main interface of the software will pop up, as shown in Figure 1. (Due to file size limitations, we are unable to upload PFAS-Soil-Kd V1.0.exe. We have uploaded the code file for PFAS-Soil-Kd V1.0.py, which can be run in Python to achieve the same interface and functionality. If you need an exe file, please contact the author for assistance.)



**Figure 1.** Software model prediction interface.

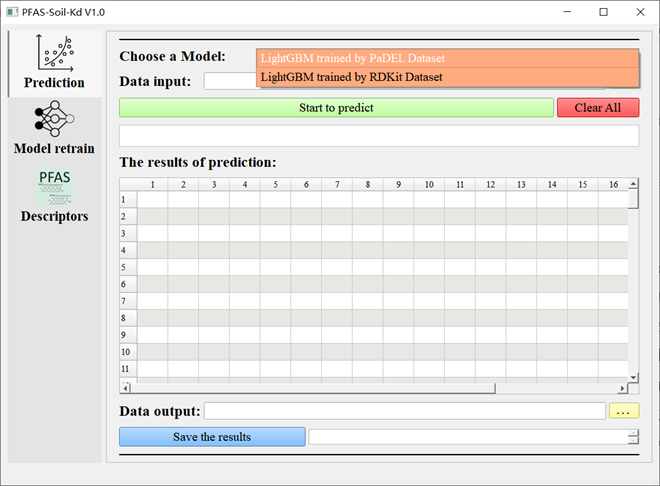
# 1 Instructions for the use of PFAS soil adsorption capacity (logKd) prediction function

## 1.1 Prediction interface

Click on "Prediction" to jump to the PFAS soil adsorption capacity prediction interface, as shown in Figure 1.

## 1.2 Selecting a predictive model

Select a model for prediction in the checkbox corresponding to Choose a Model, as shown in Figure 2. This software integrates two LightGBM algorithms trained on different PFAS molecular descriptor datasets (RDKit, PaDEL), with similar predictive performance. Users can choose according to their own needs.



**Figure 2.** Selecting a Prediction Model.

## 1.3 Data Preparation

In the ‘Data’ folder, there are two data import templates "Data template for PaDEL training model prediction.xlsx" and "Data template for RDKit training model prediction.xlsx". According to the model selected in step (1), open the corresponding data import template and fill in the corresponding data under the corresponding header, as shown in Figure 3. In the header, SOC represents soil organic carbon content (%); CEC represents the cation exchange capacity of soil (cmol/kg). Sand, Silt, and Clay respectively represent the sand content (%), silt content (%), and clay content (%) in the soil, pH represents the environmental pH value during adsorption, Rws represents the water soil ratio (mL/g), Ca2+ represents the calcium ion content in the environment (mM), and logCe represents the PFAS content in the aqueous solution (μg/L). The other features are RDKit or PaDEL molecular descriptors of PFAS, and their calculations can be found in the operating instructions of the subsequent Descriptors interface.

This software currently supports importing data in “xlsx” and “csv” formats. Please ensure that the data is filled in and saved according to the template format, otherwise it will affect the normal operation of the software.



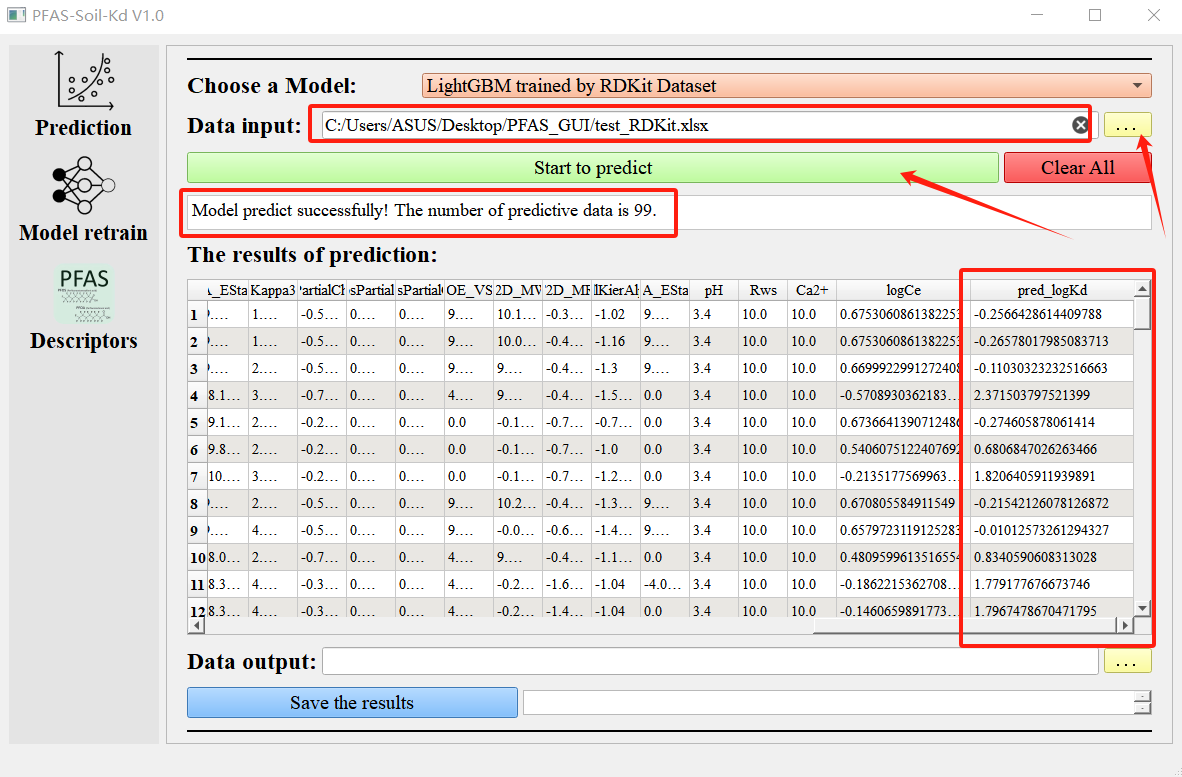


**Figure 3.** Prediction Data Import Template.

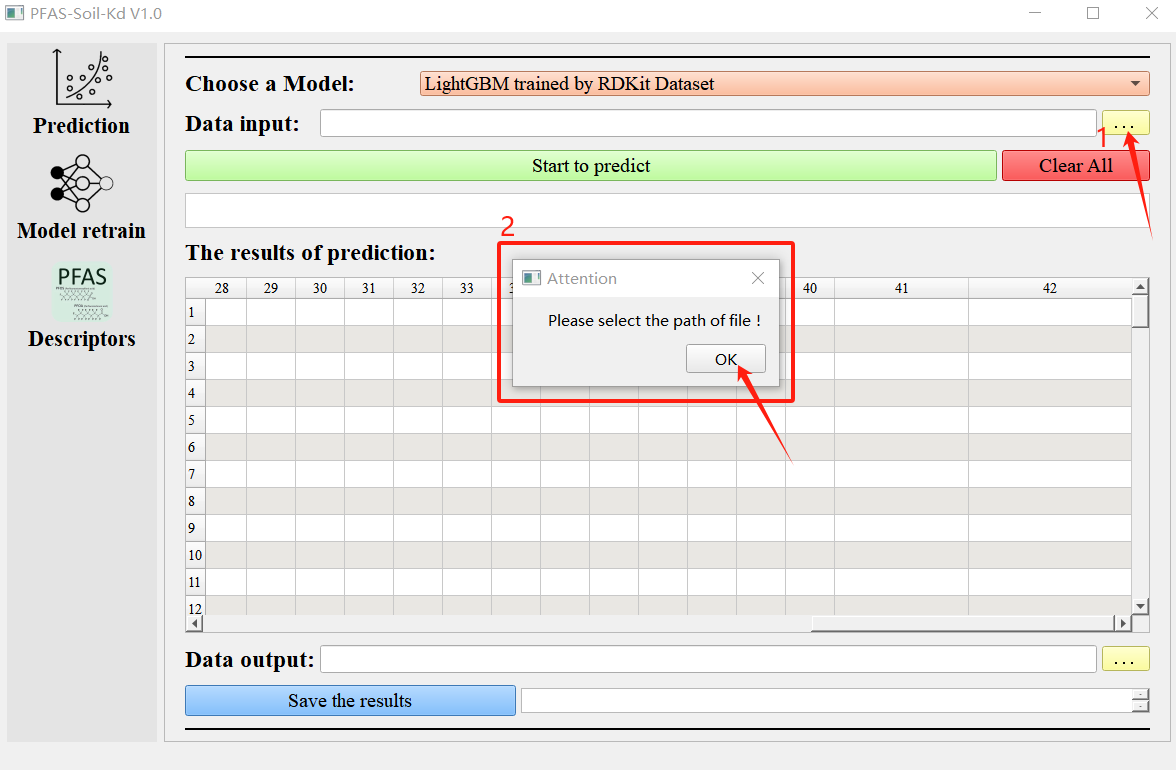
## 1.4 Data Import and Prediction

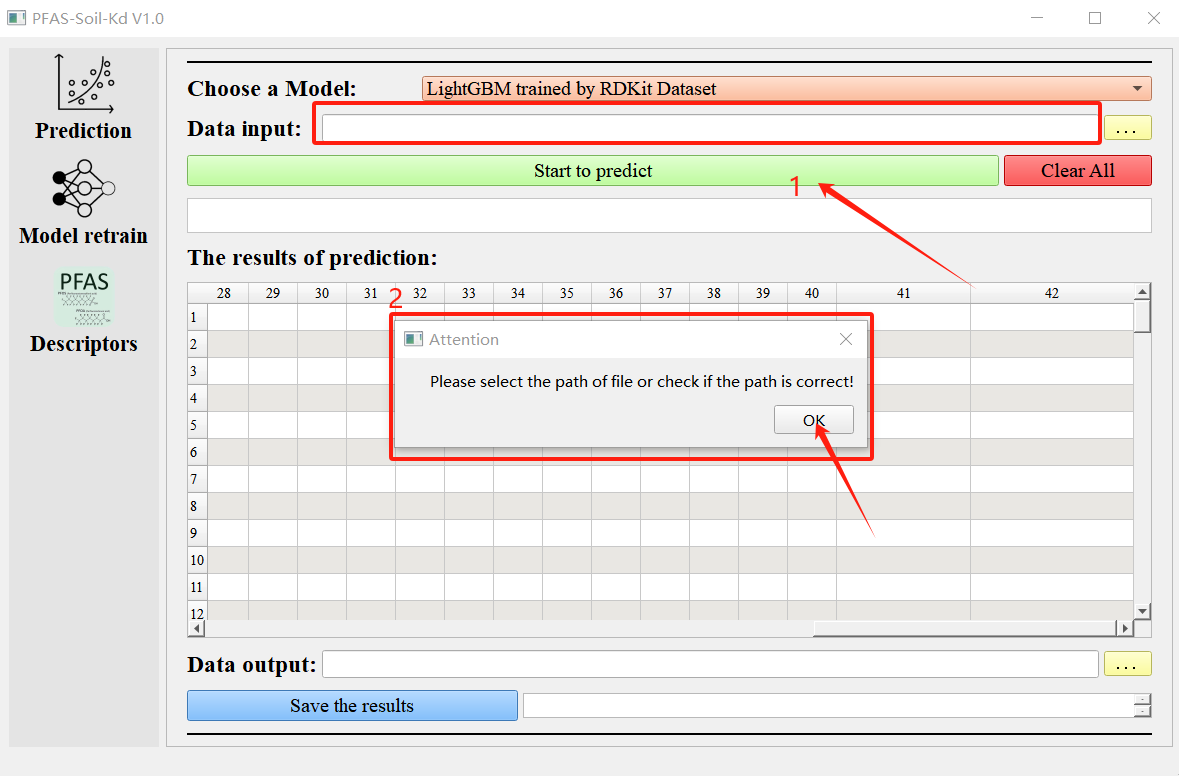
Click the yellow button with the "Data input" to enter the data import interface. Select the data file prepared in step 1.2 that corresponds to the model selection and import it. After the import is complete, the file path will be displayed in the text box after "Data input". In addition to importing data through the yellow button, you can also manually enter the location of the file in this text box. With the correct input of the file path and file name, data import can also be achieved. Then click the "Start to predict" button to make a prediction. After the prediction is successful, the text box below the button will display a message indicating the success of the prediction and inform you of the number of data involved in the prediction. The specific prediction results will be displayed in the table corresponding to "The results of prediction". In addition to displaying the logKd value of the adsorption distribution coefficient of PFAS in soil predicted for each data (with the header as pred-logKd), the entire data involved in the prediction will also be displayed, which corresponds one-to-one with pred-logKd. As shown in Figure 4.

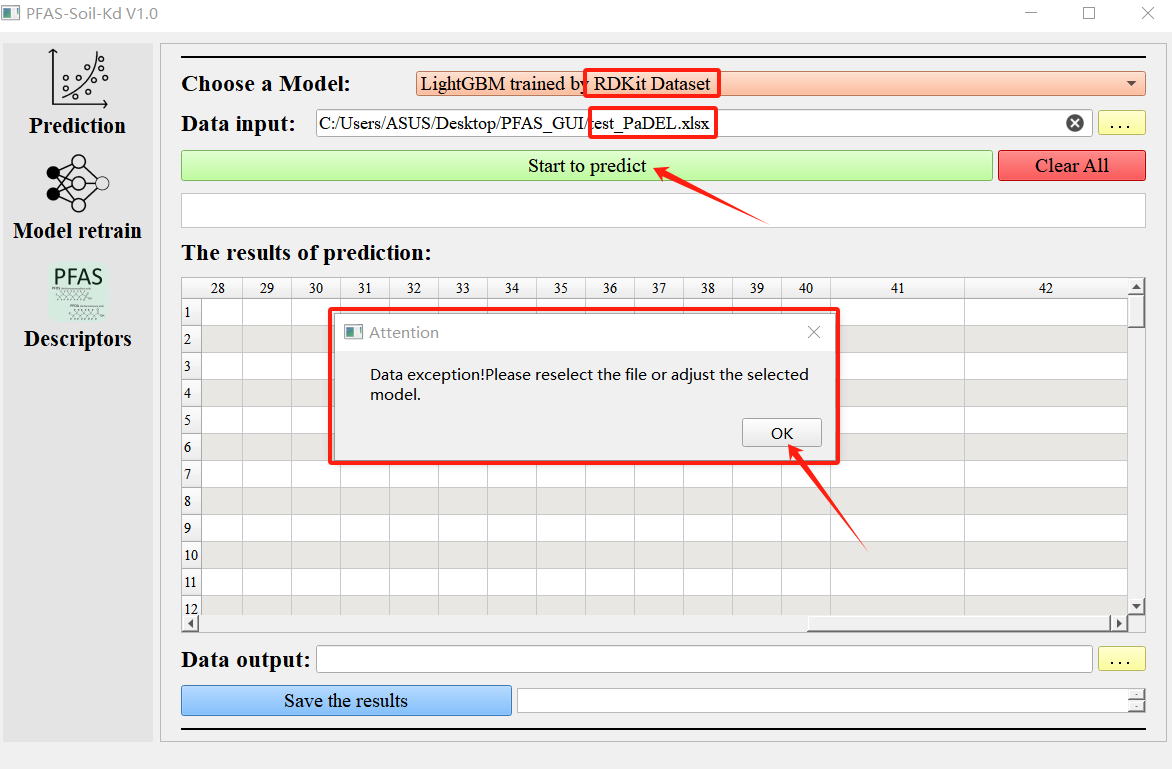
If the yellow button is clicked for data import operation but no file is selected, the software will pop up a message prompt window "Please select the path of file!". If the "Start to predict" button is clicked without importing data or inputting the wrong data path, the software will pop up an information prompt window "Please select the path of file or check if the path is correct. If the selected model does not match the imported data, the software will pop up an information prompt window "Data exception! Please reselect the file or adjust the selected model.”. All information prompt boxes can be closed by clicking "OK". Users can adjust according to the prompt information before continuing with the prediction, as shown in Figure 5.



**Figure 4.** Data import and prediction results.





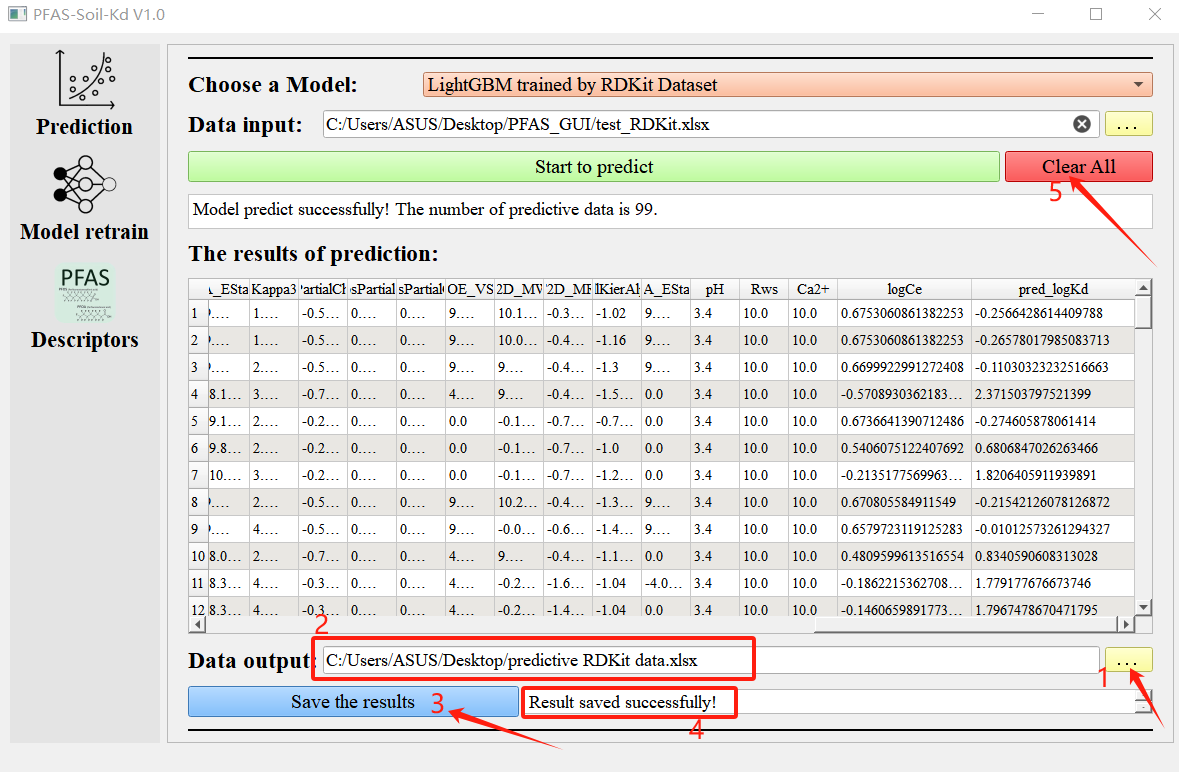


**Figure 5.** Prompt window for data import and prediction process.

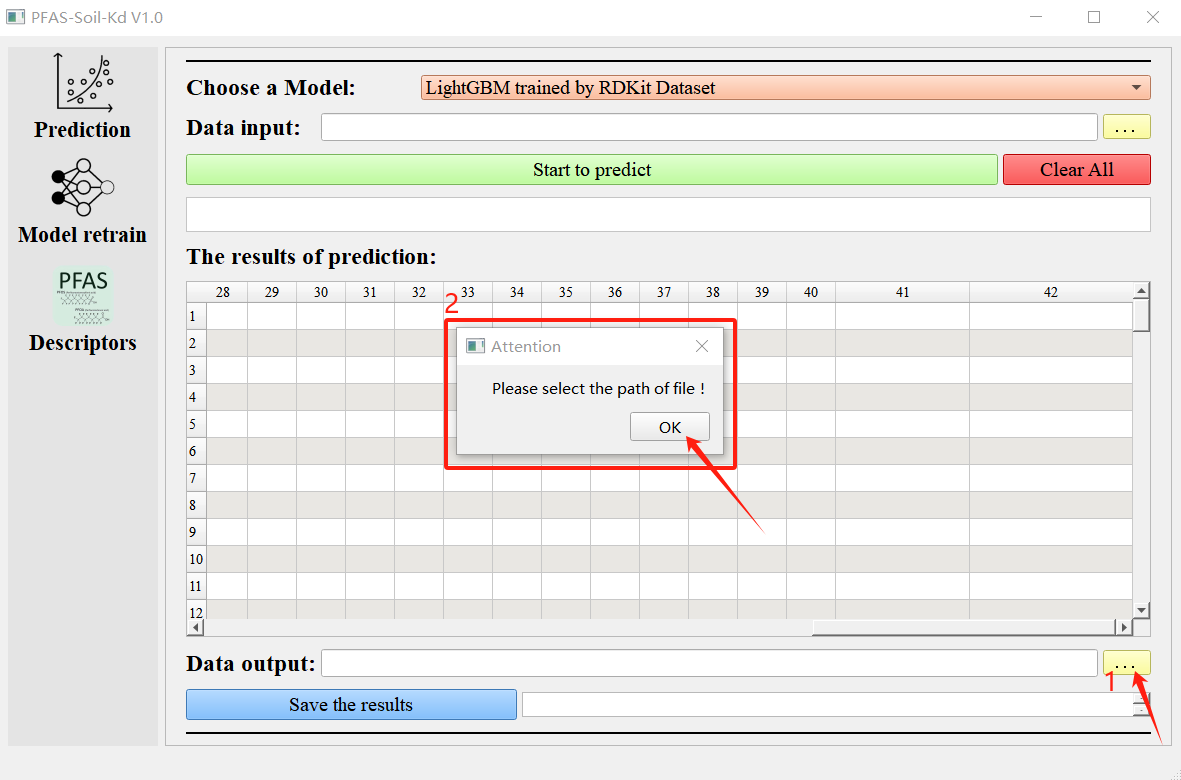
## 1.5 Data export and clearing

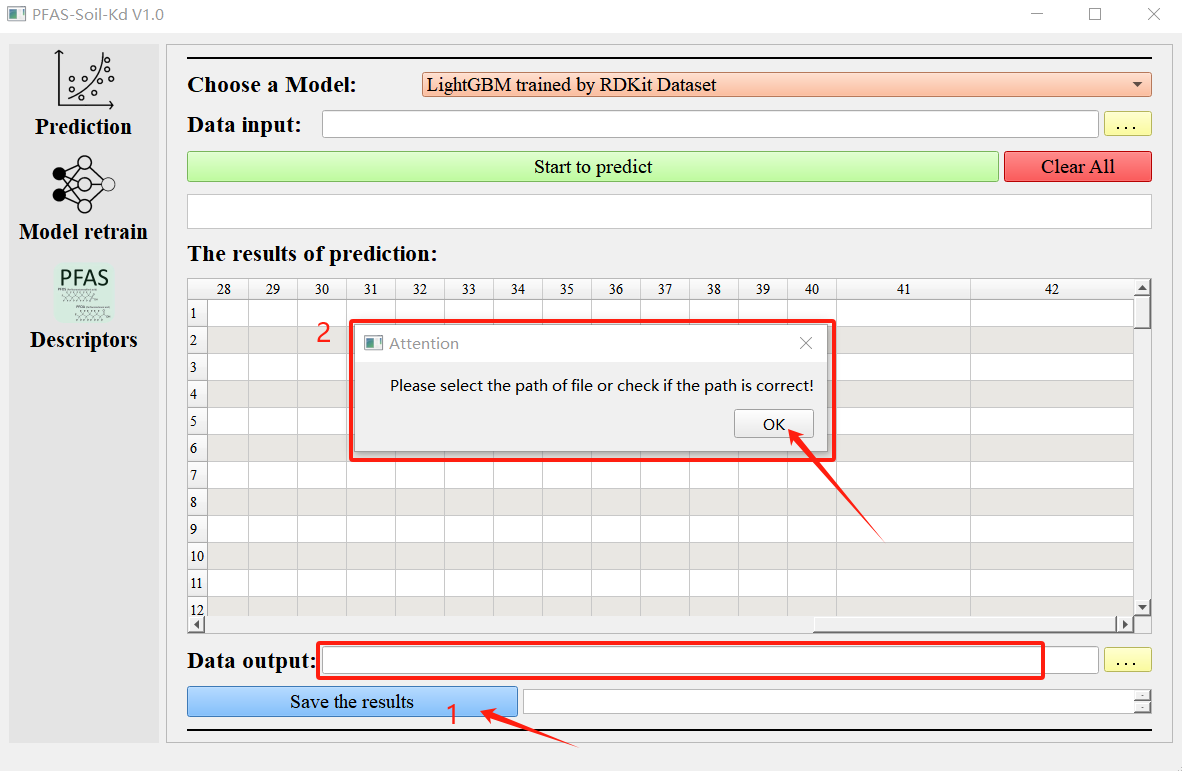
Click the yellow button behind the "Data output" column, select the path to save the data, the name of the saved file, and the format of the file. Data export supports exporting files in three formats: xlsx, csv, and txt. Users can choose according to their needs. After selecting the path, name, and format, it will be displayed in the text box after "Data output". The same text box also supports manual input, but it is important to ensure that the file format is one of the three mentioned above. Then click the "Save the results" button to save the data, and a prompt will appear in the text box after the button once saved successfully. After saving the data or preparing to predict data from another file, you can click the "Clear All" button to easily clear the data on the interface for subsequent operations. The cleared interface is shown in Figure 1.

Similar to step 2.4, if the yellow button is clicked to export data and select a file, but no file path is selected, the software will pop up an information prompt window "Please select the path of file!". If the "Save the results" button is clicked when the export file path, name, and format are not selected properly or the path input is incorrect, the software will pop up an information prompt window "Please select the path of file or check if the path is correct!". Click "OK" to close all information prompt boxes, as shown in Figure 7.



**Figure 6.** Export and clearing of data.





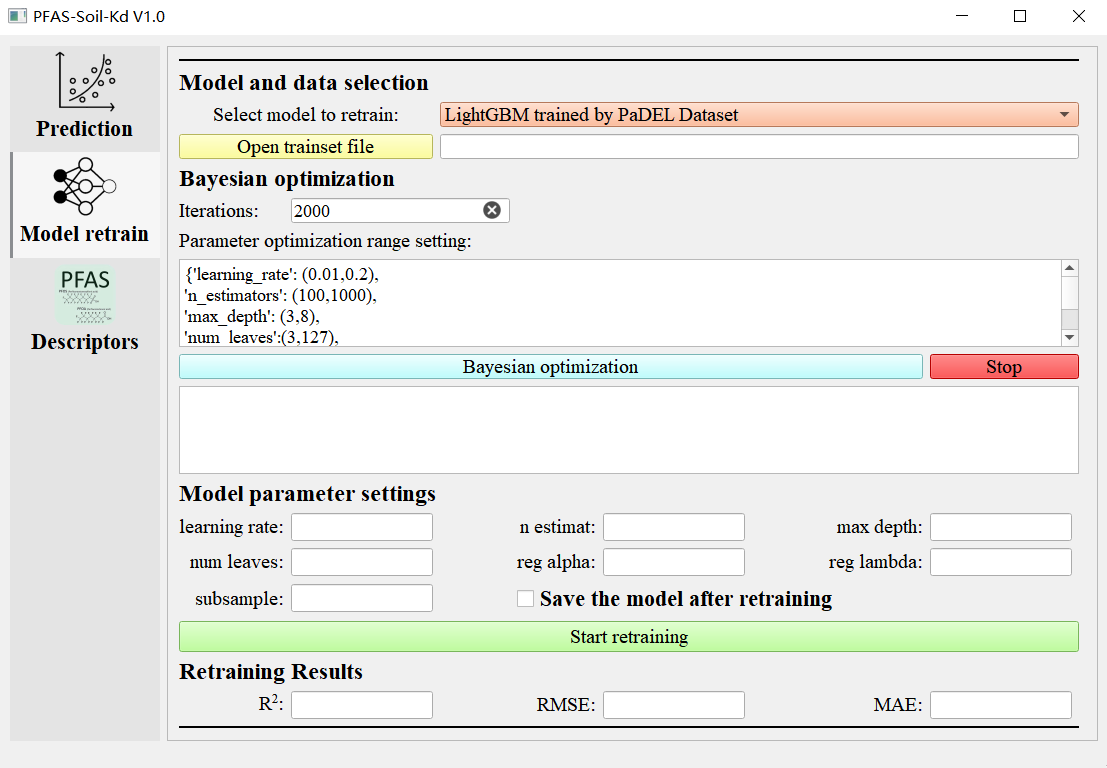
**Figure 7.** Prompt window for data export and clearing.

# 2 Instructions for using the model retraining function

When users have data outside the original dataset and want to incorporate this part of the data into the model, they can use this interface to retrain the model.

## 2.1 Model retraining interface

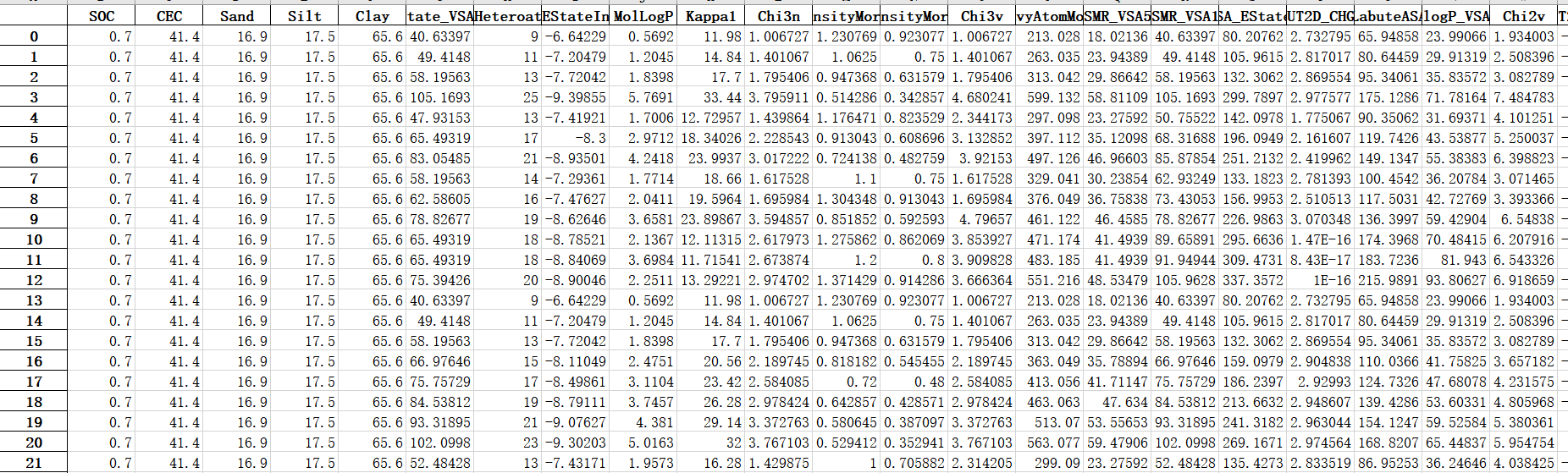
Click on "Model Retrain" on the left to jump to the model retraining interface, as shown in Figure 8.

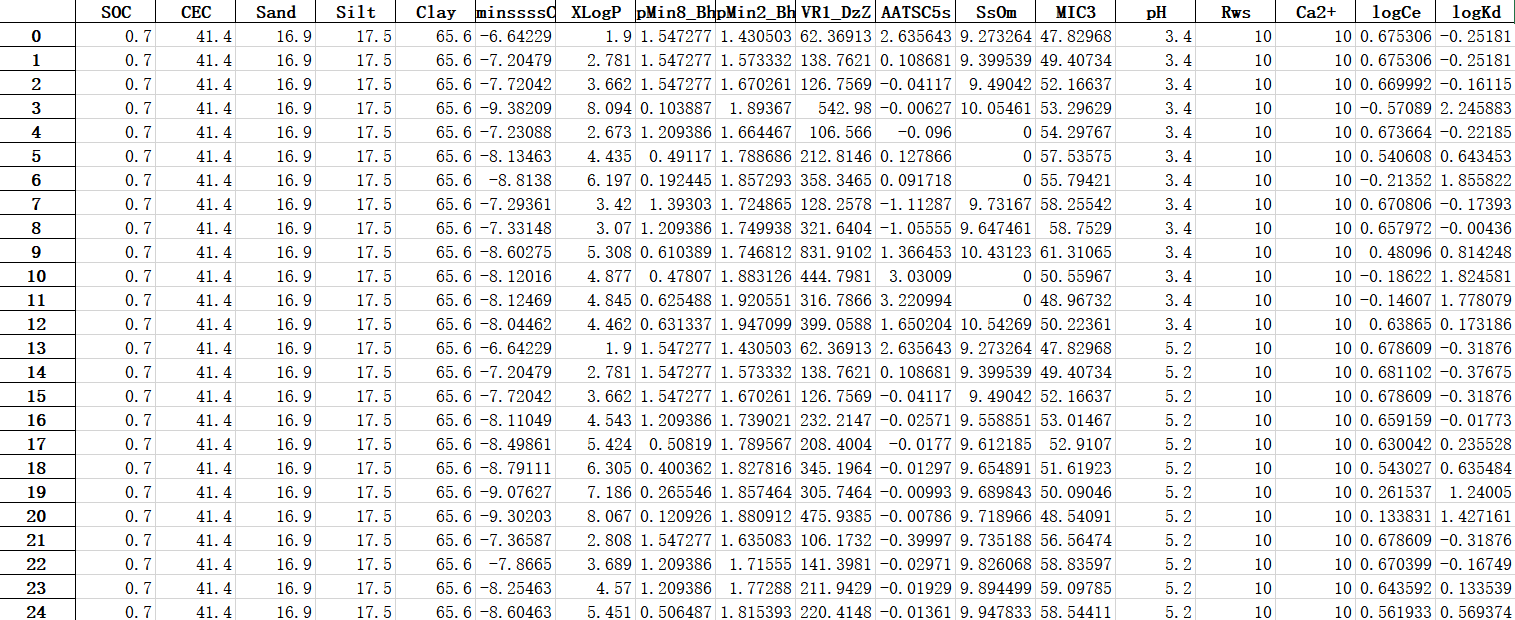


**Figure 8.** Model retraining interface.

## 2.2 Preparation of retraining data

For the convenience of users, this software folder ‘Data’ provides data for training two original models, namely "PaDEL dataset for model training.xlsx" and "RDKit dataset for model training.xlsx". Users only need to correspond to the header of the original data and fill in their own data after the original data. The raw data of the two models are shown in Figure 9.



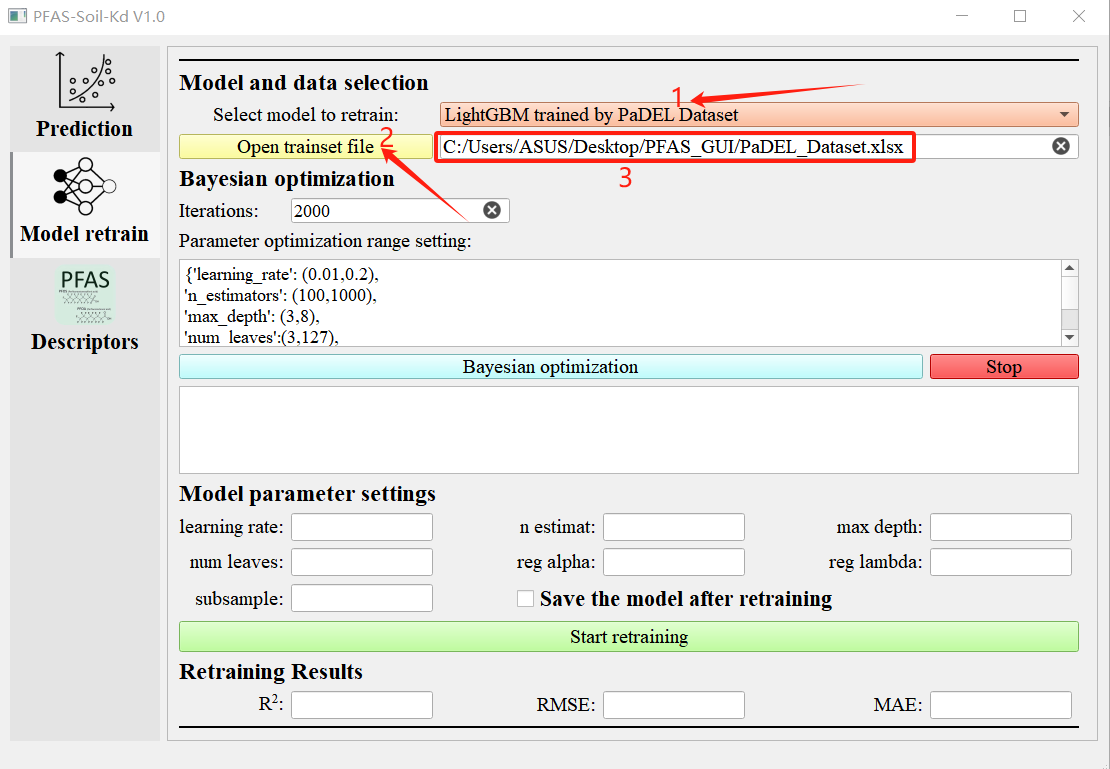
**Figure 9**. Raw data of RDKit dataset and PaDEL dataset.

## 2.3 Select models that need to be retrained and import training data

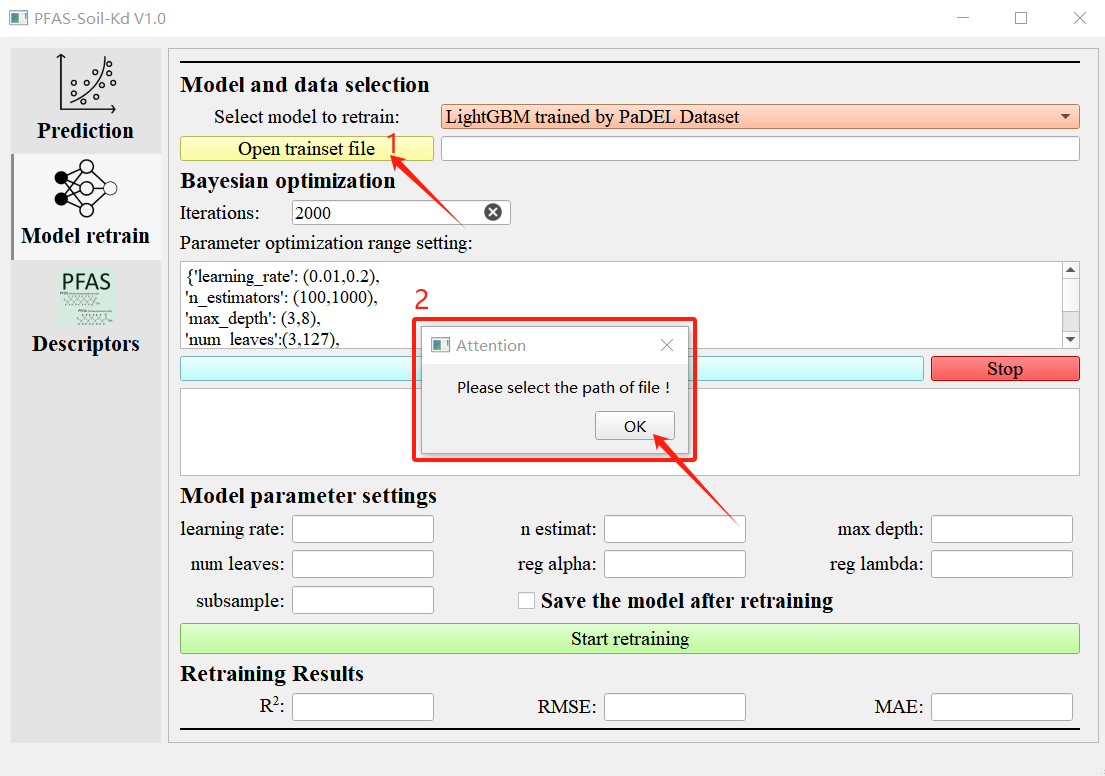
Select a model that needs to be retrained in the checkbox corresponding to “Select model to retrain”. Users can choose the model to be retrained based on the type of data they have, for example, if they have data representing PFAS properties using RDKit, they can choose the model trained using RDKit Dataset.

Click the “Open trainset file” button and select the dataset prepared in step 2.2 for model retraining that corresponds to the model selection result. The software supports selecting data in both xlsx and csv formats. After selection, the path of the selected file will be displayed in the text box behind the button. The specific process is shown in Figure 10.

If the “Open trainset file" button is clicked to import data and select a file, but no file path is selected, the software will pop up an information prompt window “Please select the path of file!”, as shown in Figure 11.



**Figure 10.** Model selection and data import for retraining.

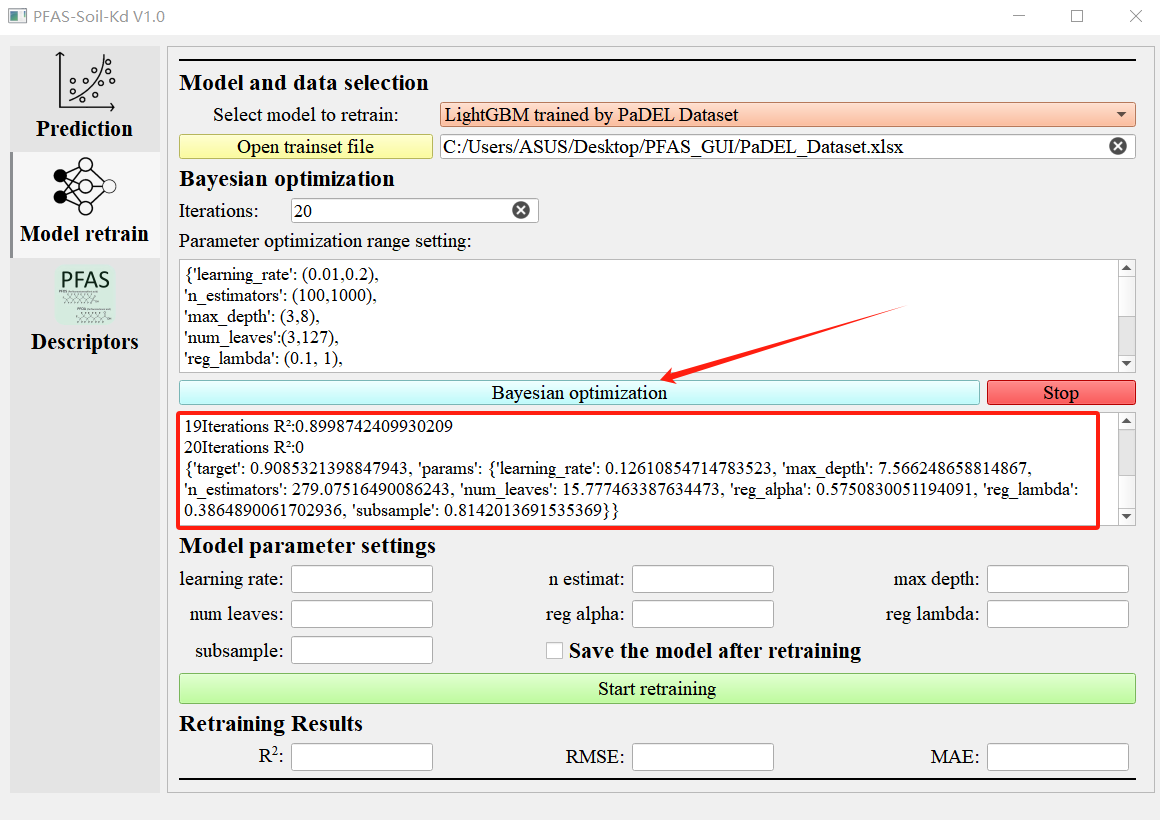


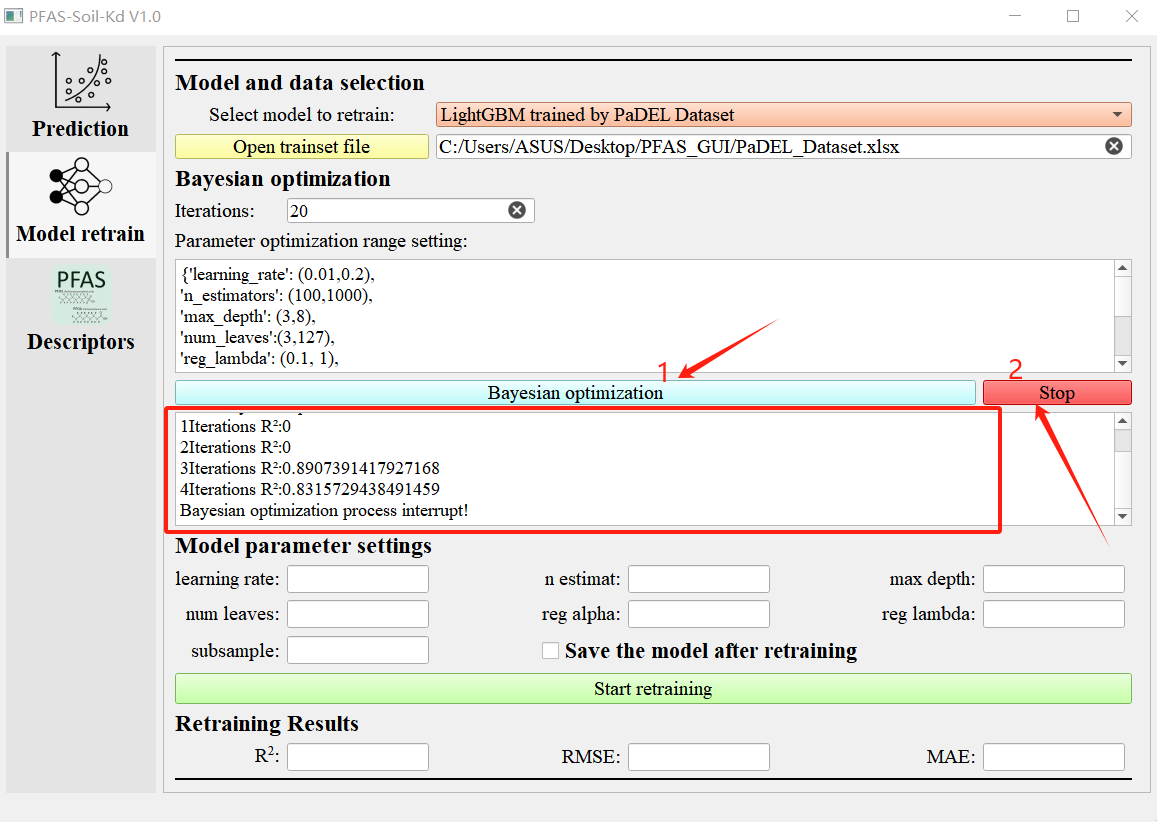
**Figure 11.** Prompt message for data import process during retraining.

## 2.4 Bayesian Optimization

In order to facilitate users to retrain the model and obtain the optimal model prediction performance, this software has implanted Bayesian optimization algorithm and set the default iteration number (Iterations) of Bayesian optimization to 2000 times, as well as set the hyperparameter search range of machine learning algorithm LightGBM in the text box under “Parameter optimization range setting”. These parameters are all default values, and users can adjust the values according to their own needs. Please do not change the format of hyperparameters, otherwise it will affect the software operation.

The prerequisite for performing Bayesian optimization is to complete the content in step 2.3, and then click the “Bayesian optimization” button directly. The Bayesian optimization process will be displayed in the text box below the button, where you can see the current number of iterations and the R2 value of the model performance at that iteration. After reaching the set number of iterations, Bayesian optimization will automatically stop and return the best prediction performance obtained among all iterations, and output its parameters at the end of the text box. Users can obtain the results of Bayesian optimization by copying the content. In the process of Bayesian optimization, you can click the “Stop” button to terminate the process. After termination, the message “Bayesian optimization process interrupt!” will be output in the text box below. As shown in Figure 12.

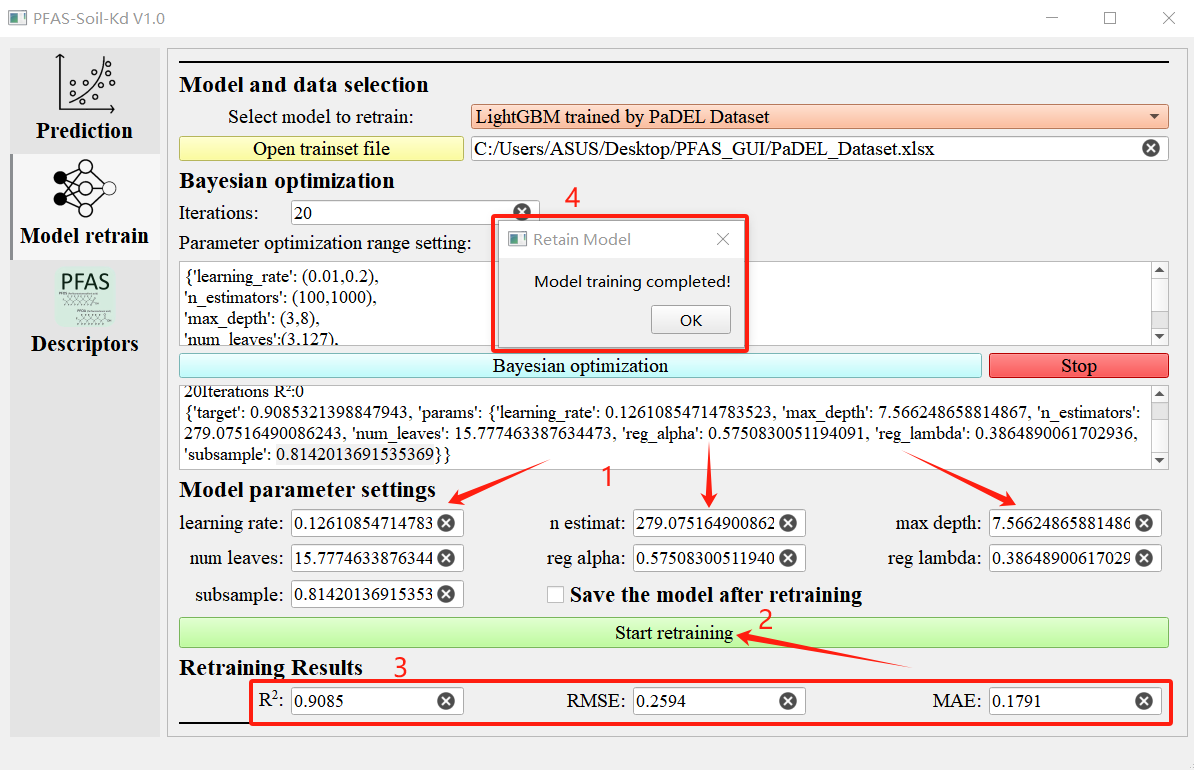


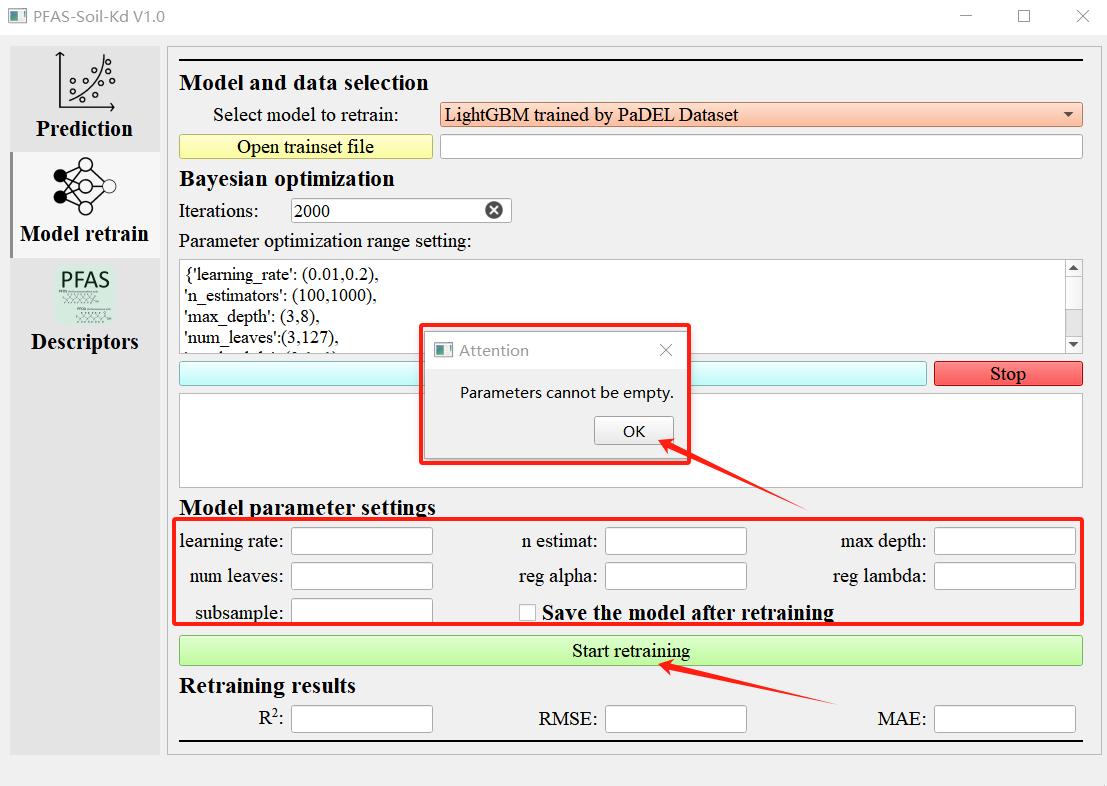


**Figure 12.** Beginning and termination of Bayesian optimization process.

## 2.5 Model Retraining and Saving

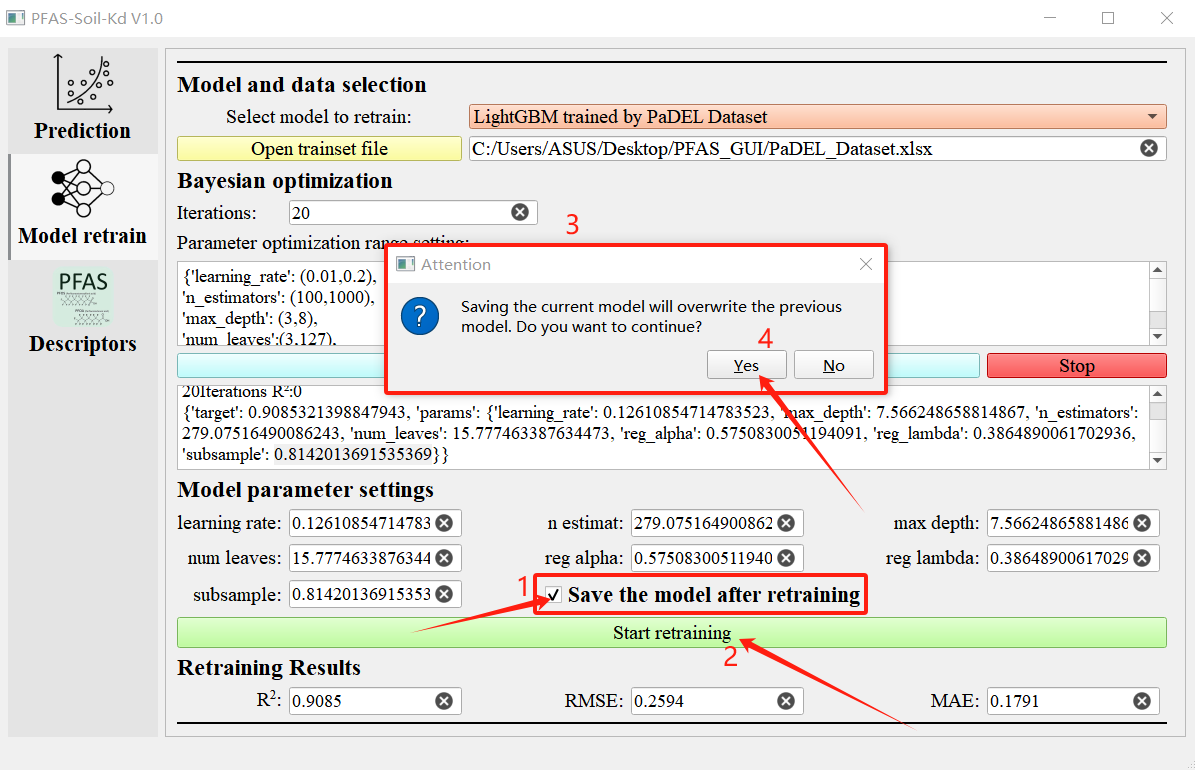
Based on the results of Bayesian optimization, the optimization results of hyperparameters can be copied into the corresponding text boxes, and then click “Start retraining” to retrain the model. The performance of the retrained model is represented by the mean of R2, RMSE, and MAE in its five fold group cross validation on the training set. These results are displayed in their respective text boxes in the “Retraining results” column. After successful training, the software will prompt a window "Model training completed!". Users can also freely design hyperparameters and fill them in the "Model parameter settings" column without performing Bayesian optimization. Based on the trained performance, they can select the best performing model to save. If the hyperparameters are not fully set when clicking "Start retreating", the software will display a message saying "Parameters cannot be empty.", as shown in Figure 13.

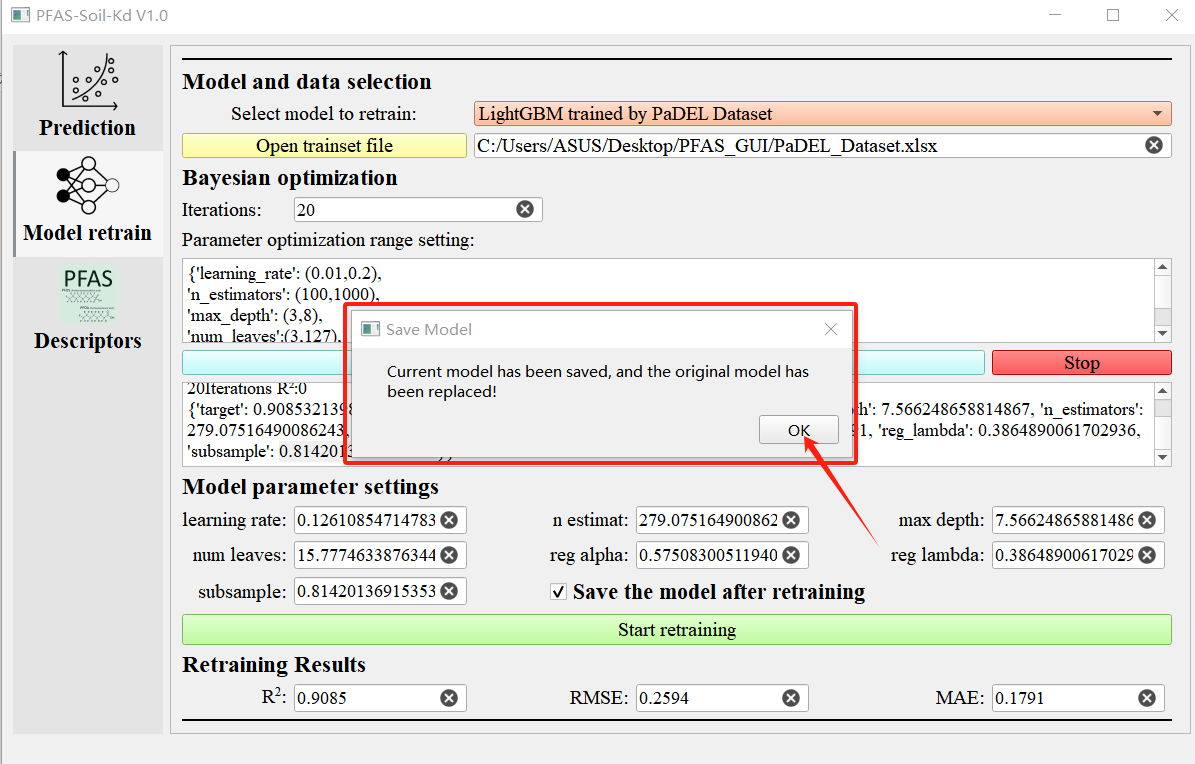


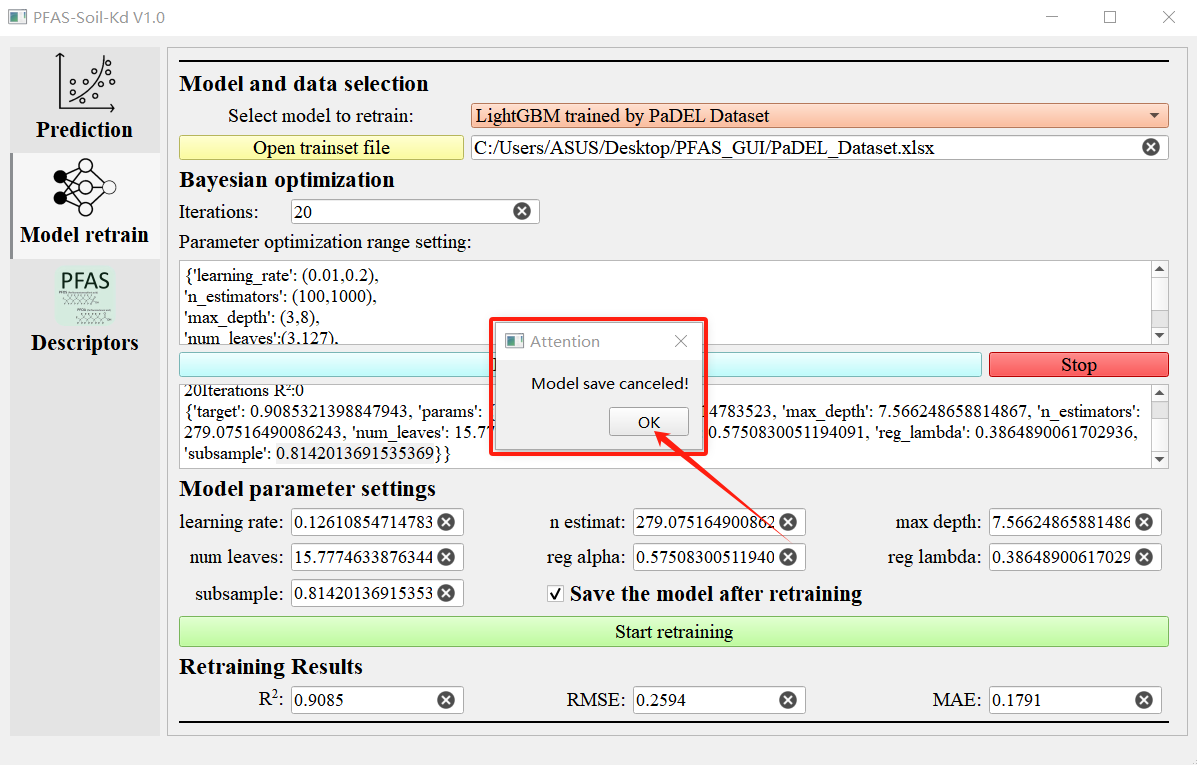


**Figure 13.** Model retraining process.

To save the model, simply check the box in front of "Save the model after retirement" after confirming all hyperparameters, and then click "Start retirement". The software will prompt the user to save the model and replace the original model. If the user clicks "Yes", the model will continue to be saved. After successful saving, the software will display a prompt window saying "Current model has been saved, and the original model has been replaced; Clicking 'No' will cancel the saving of the model, and the software will prompt a window saying 'Model save cancelled!'. As shown in Figure 14.





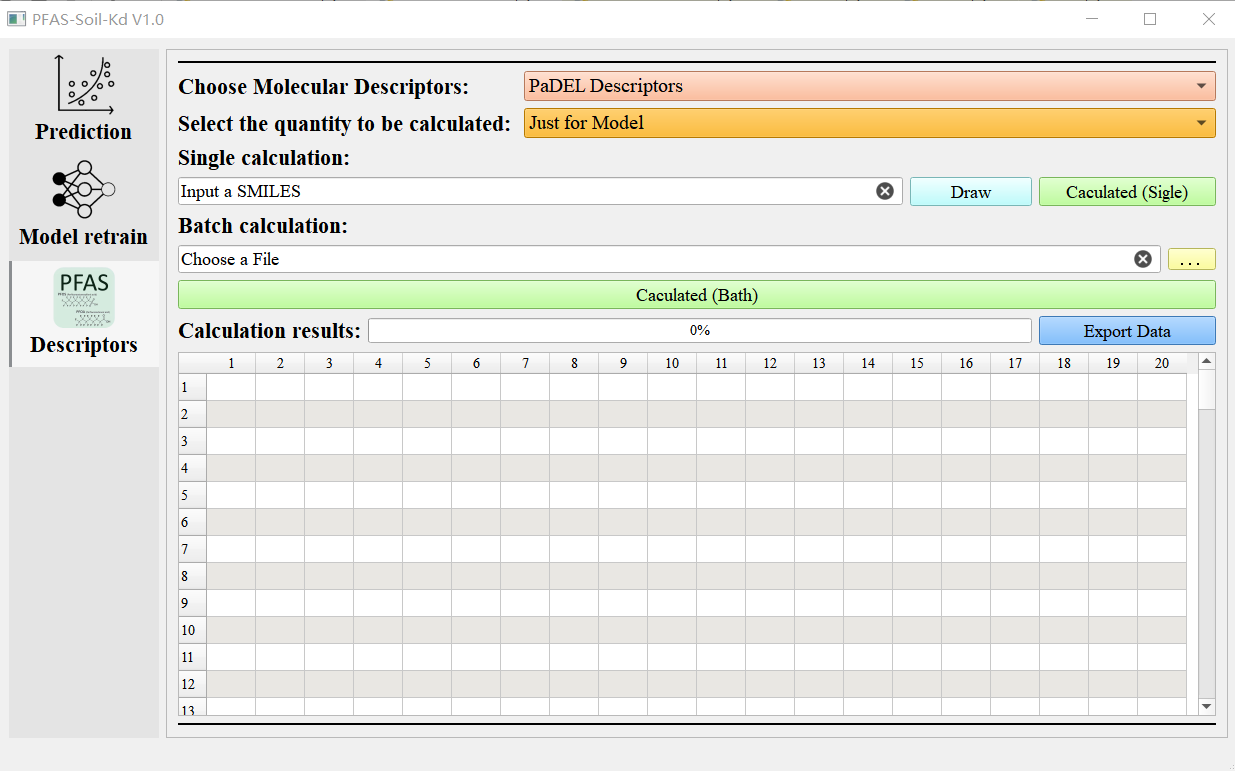


**Figure 14.** Saving the retrained model.

# 3 Instructions for using the calculation function of molecular descriptors

## 3.1 Molecular descriptor calculation interface

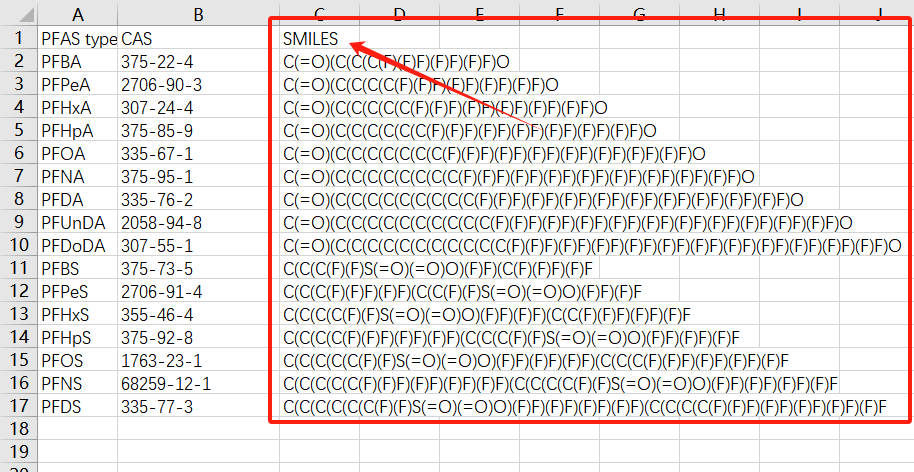
Click on "Descriptors" on the left to jump to the molecular descriptor calculation interface, as shown in Figure 15. The calculation molecular descriptor interface of this software is mainly designed to provide convenience for predicting the adsorption performance of PFAS in soil. Currently, it only supports calculating two types of molecular descriptors, namely RDKit and PaDEL molecular descriptors. However, the calculation of compounds is not limited to PFAS, and other compounds can also use the functionality of this interface to calculate their RDKit and PaDEL molecular descriptors.



**Figure 15.** Molecular descriptor calculation interface.

## 3.2 Data Preparation

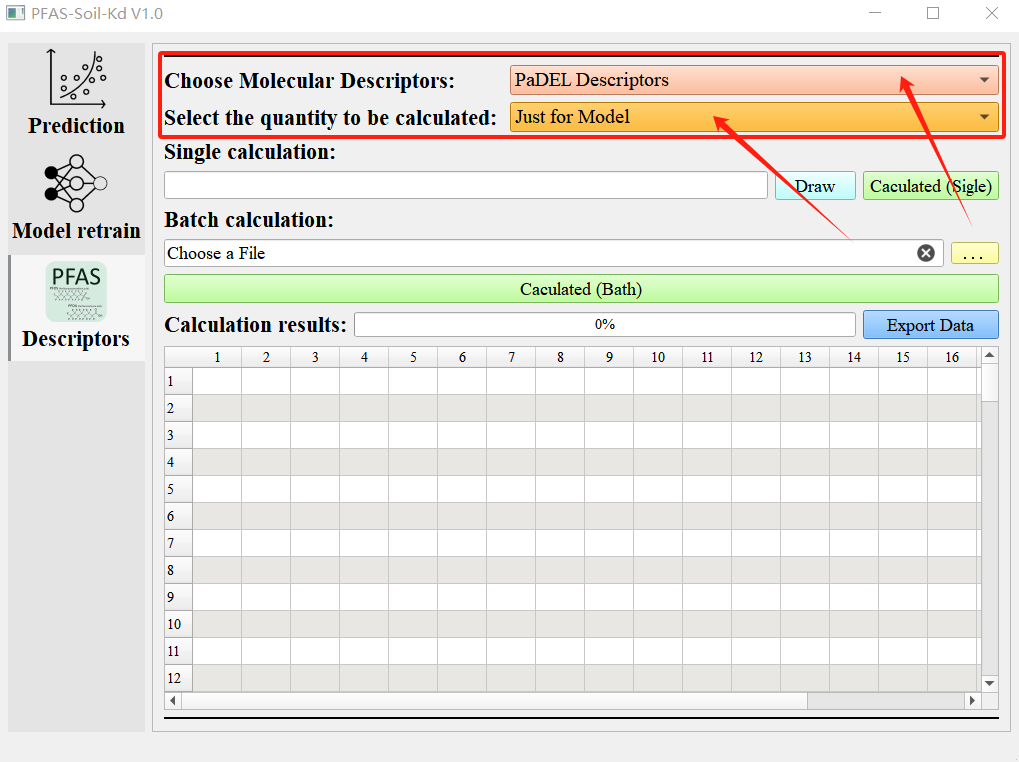
The calculation of molecular descriptors for a single compound only requires the corresponding SMILES of the compound. For a large number of compounds that require batch calculation, their SMILES need to be stored in an Excel file in xlsx or csv format for data import. Note that the only requirement for storing SMILES files is that the header of the column where the compound SMILES is located should be "SMILES". There is no requirement for data in other columns, and users can decide on their own. As shown in Figure 16.



**Figure 16.** Data preparation for molecular descriptor calculation.

## 3.3 Selection of molecular descriptor types and quantities

In the checkbox after the "Choose Molecular Descriptors" column, you can select the type of molecular descriptor you want to calculate, which are "RDKit Descriptors" and "PaDEL Descriptors". In the checkbox after the "Select the quantity to be calculated" column, you can select the number of molecular descriptors you want to calculate, which is divided into two options: "Just for Model" and "All". Selecting "Just for Model" will only calculate the molecular descriptors used by each prediction model. RDKit has 32 descriptors, and PaDEL has 8 descriptors. Selecting "All" will calculate all molecular descriptors contained in RDKit or PaDEL. As shown in Figure 17.

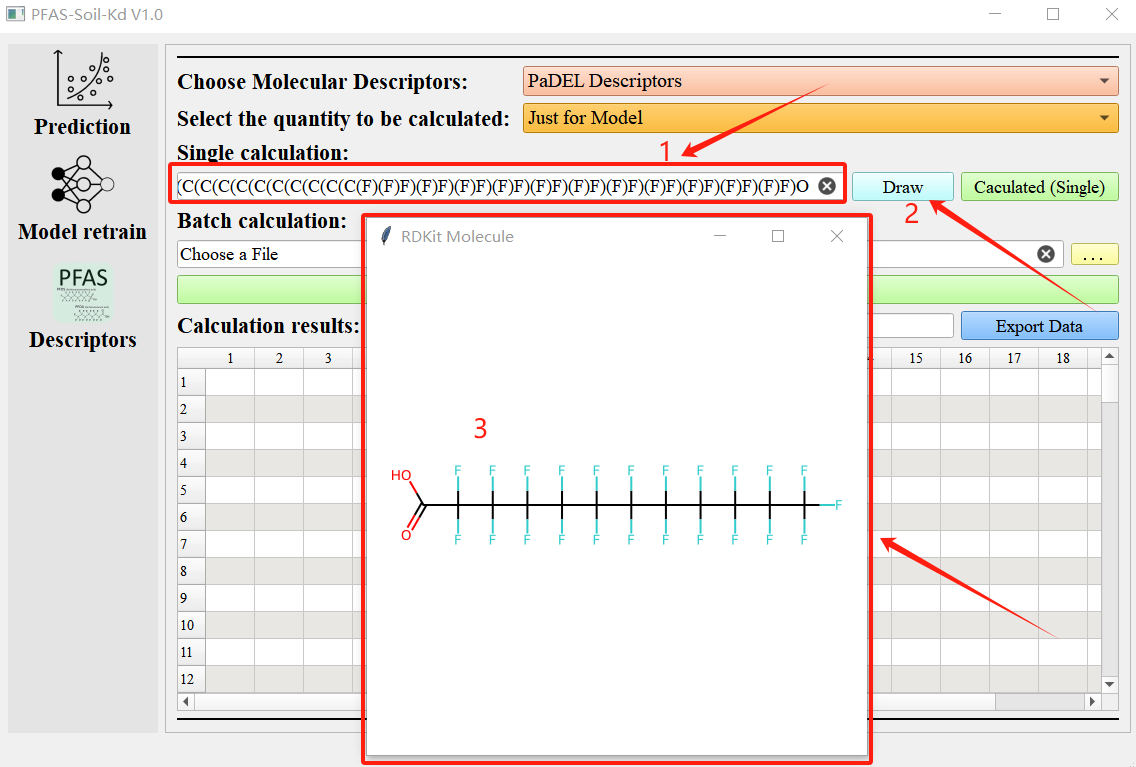


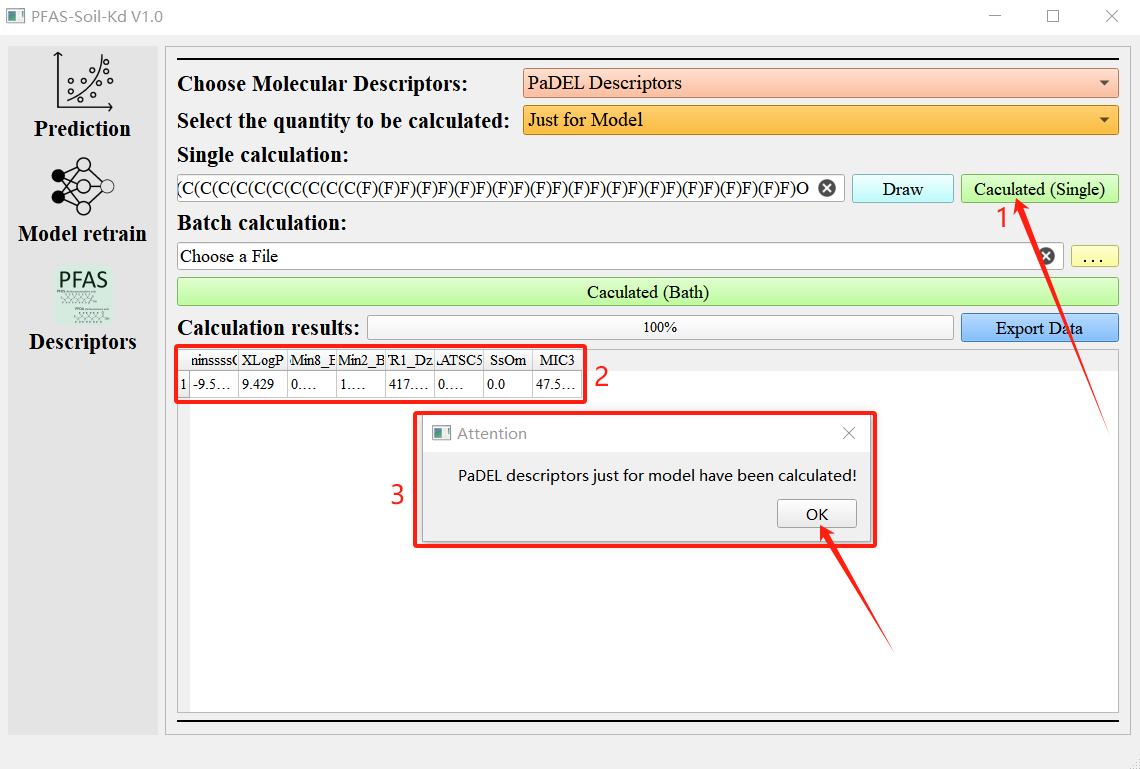
**Figure 17.** Selection of molecular descriptor types and quantities.

## 3.4 Calculate the molecular descriptor of a single compound

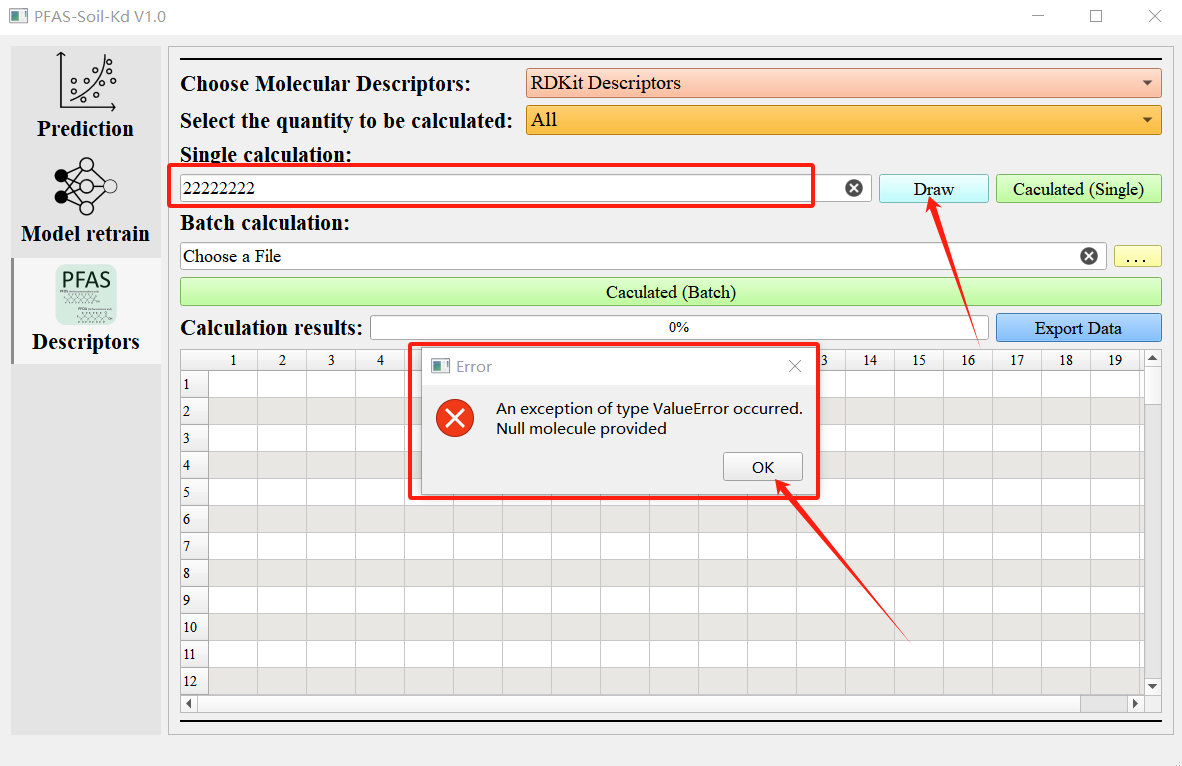
Calculate the molecular descriptor of a single compound and paste its SMILES into the text box under 'Single calculation'. As the software integrates the function of drawing molecular structures based on SMILES in the RDKit package, users can click the "Draw" button behind the text box after entering SMILES to view the structure of the corresponding compound, making it easier for users to check whether the SMILES of the compound is entered correctly. After closing the compound structure window, you can click "Calculated (Single)" to calculate the molecular descriptors. After the calculation is completed, the results will be displayed in the table below "Calculation results", and an information prompt window will be generated based on the type and quantity of selected molecular descriptors. If "PaDEL Descriptors" and "Just for Model" are selected here, it will prompt "PaDEL descriptors just for model have been calculated!", as shown in Figure 18.

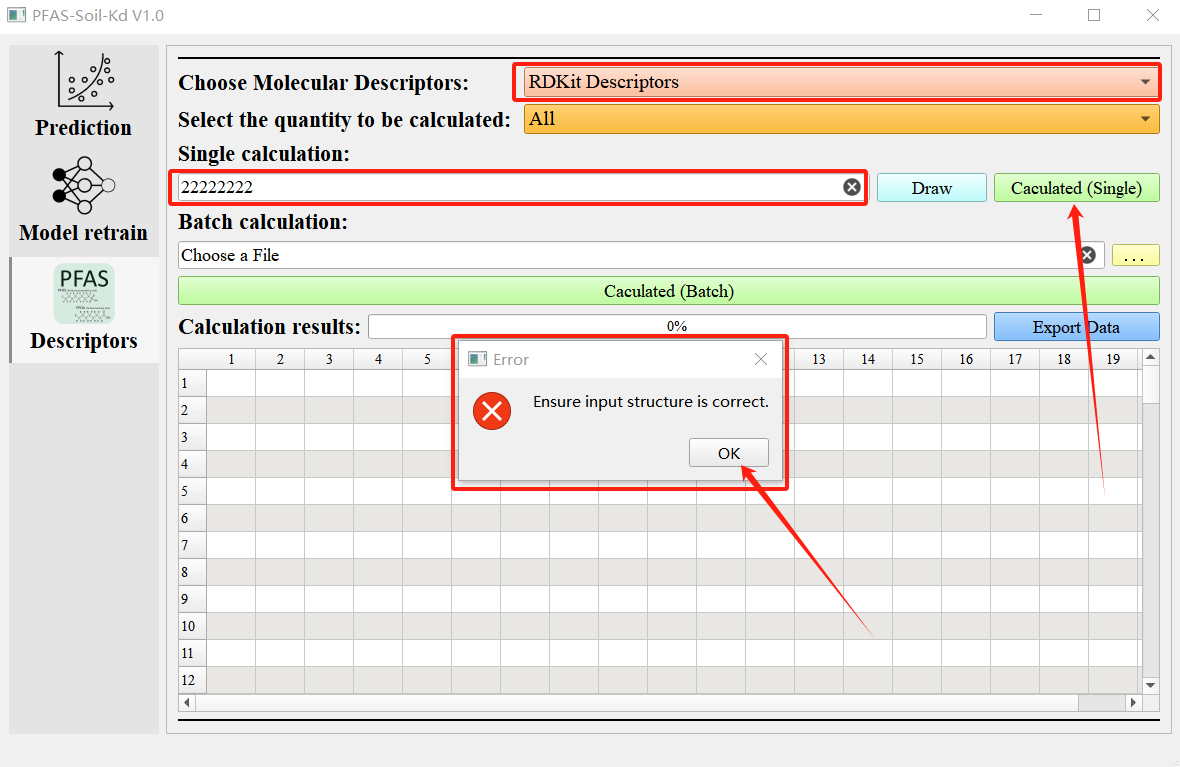
When the text box for SMILES is empty, clicking the "Draw" or "Caculated (Single)" button will not have any response; However, if there is an error in the SMILES of the input compound, clicking the "Draw" or "Caculated (Single)" button will cause the software to display an error message, as shown in Figure 19.

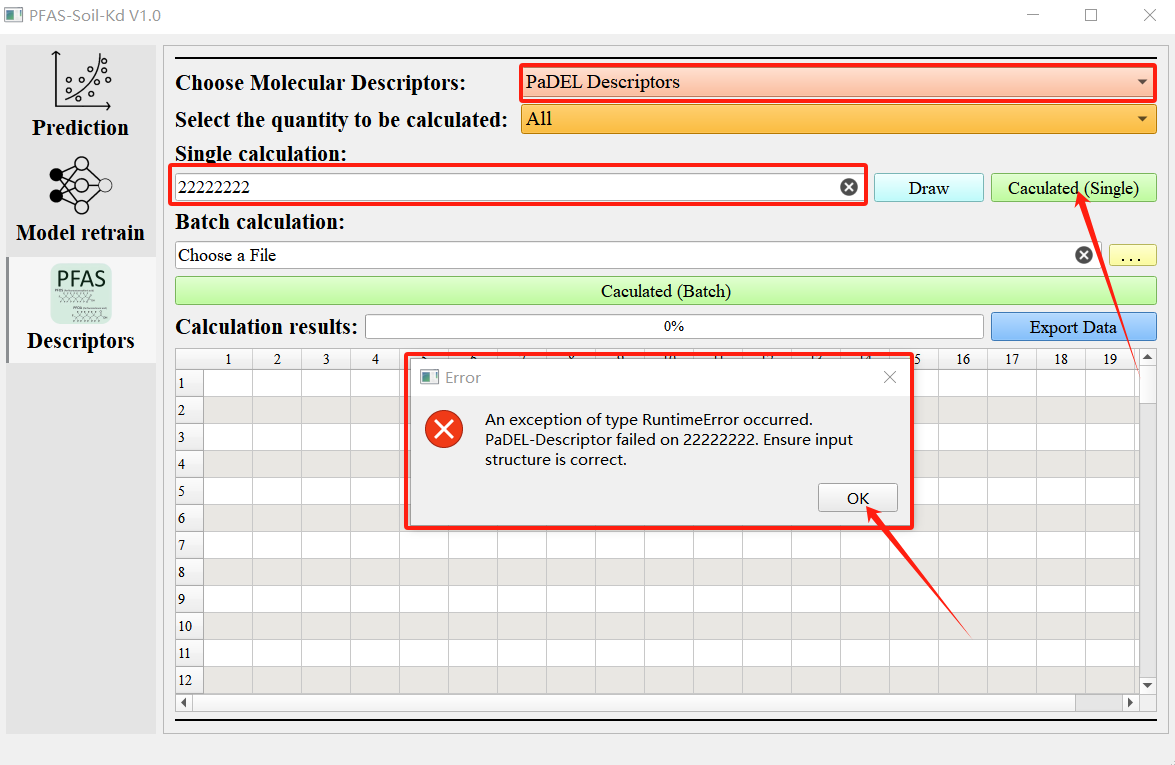




**Figure 18.** Calculation of single compound molecular descriptors.





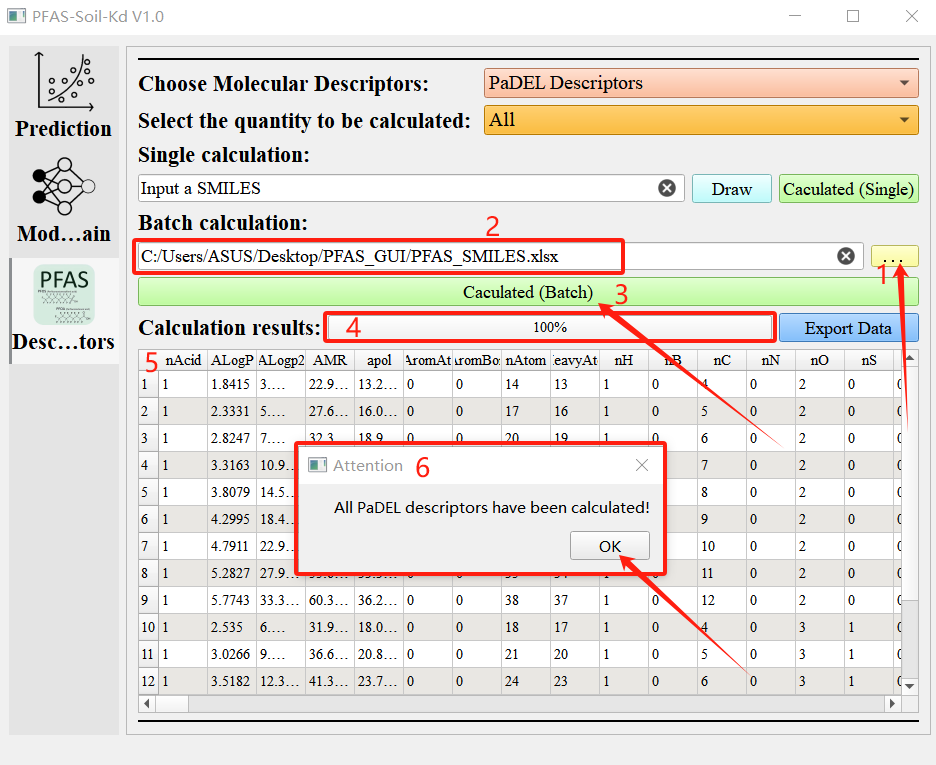


**Figure 19.** Error message indicating incorrect input of compound structure.

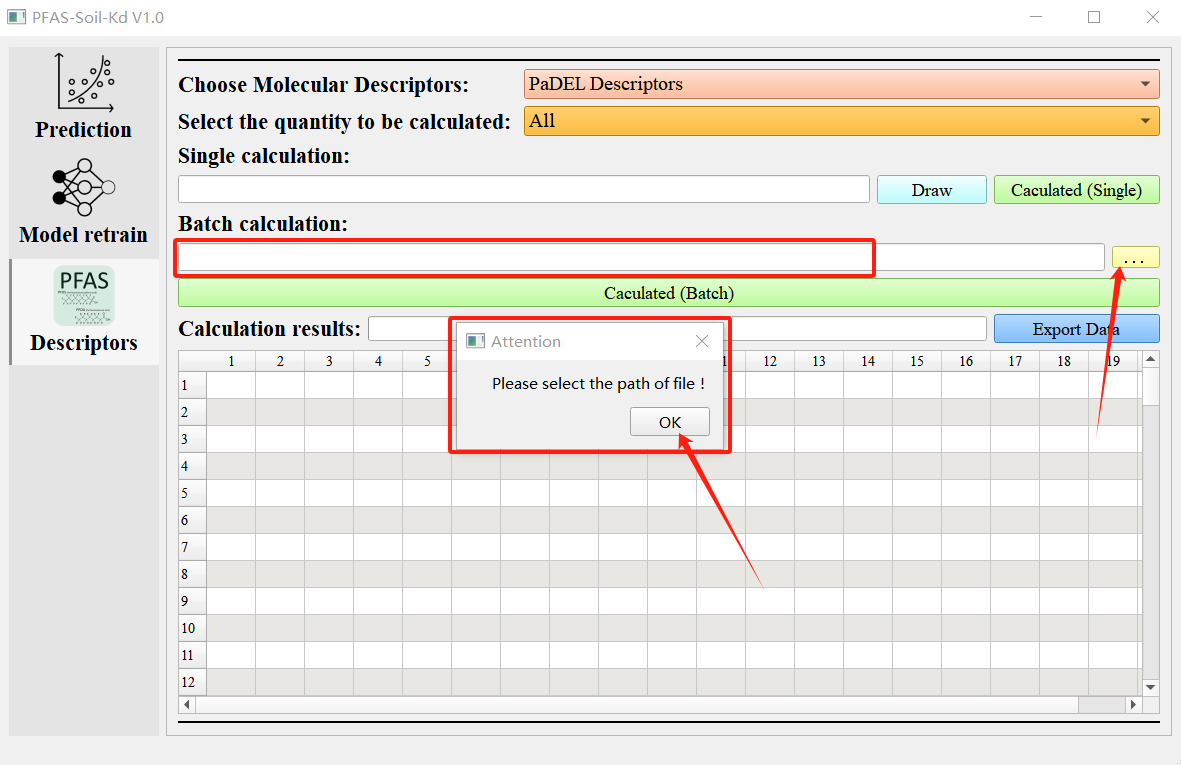
## 3.5 Batch calculation of molecular descriptors for compounds

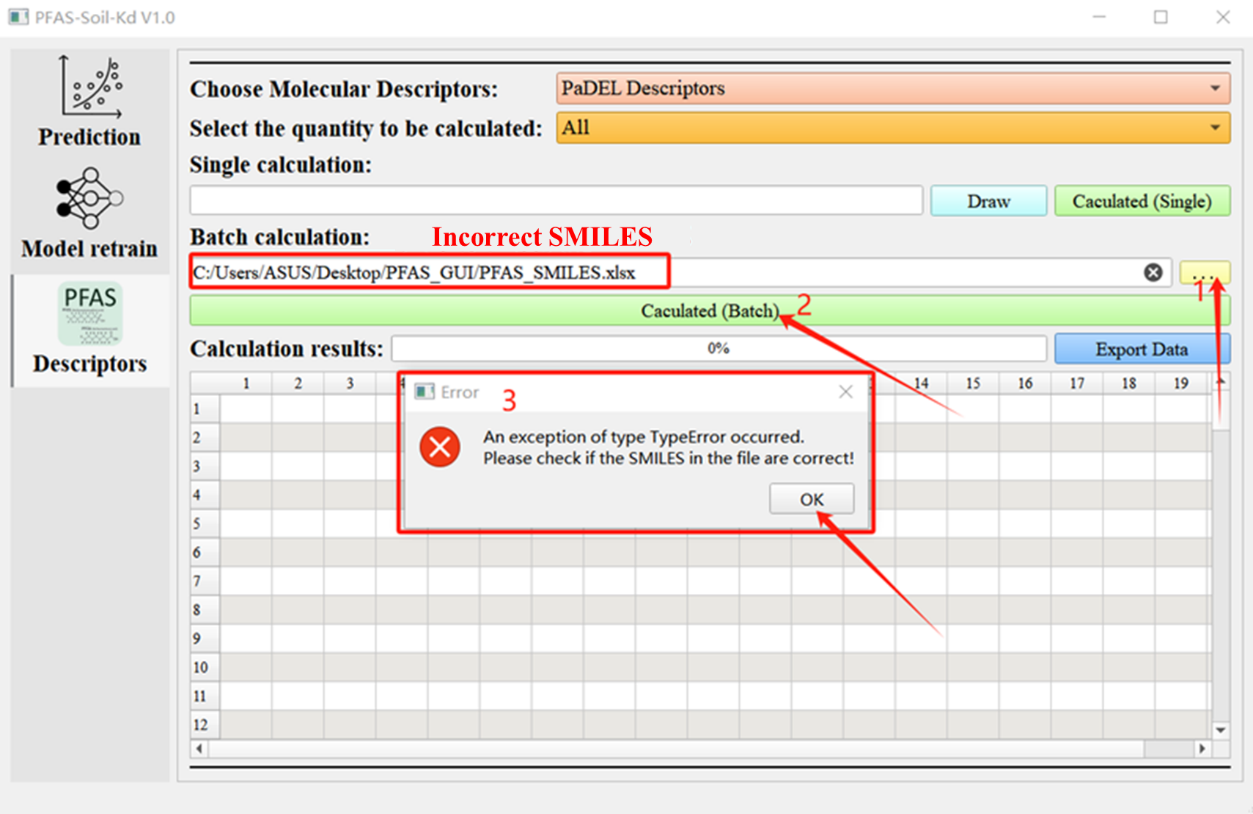
After preparing the SMILES data of the compound for batch calculation as shown in step 3.2, click the yellow button behind the text box under "Batch calculation" to select the data file. After successful selection, the path of the data file will be displayed in the text box in front of the button. Then click on "Calculated (Batch)" to start batch calculation. After the calculation starts, the progress bar below will display the calculation progress. If the data volume is large, it may take a long time. Please be patient and wait. After the calculation is completed, the calculation results will be displayed in the table below according to the SMILES order in the data file. The software will generate an information prompt window based on the selected molecular descriptor type and quantity, such as the prompt message "All PaDEL descriptors have been calculated!". The specific process is shown in Figure 20.

If the yellow button is clicked but no data file is selected, the software will prompt "Please select the path of file!". If the selected data file path is incorrect or empty, clicking the "Caculated (Batch)" button will not produce any response. Please carefully check if the file path is correct. If there are incorrect SMILES in the selected SMILES data file, the software will report an error message "Please check if the SMILES in the file are correct!" after clicking the "Caculated (Batch)" button. At this time, please check to ensure that the SMILES used for batch calculation in the file are correct. The specific instructions are shown in Figure 21.



**Figure 20.** Batch calculation of molecular descriptors for compounds.





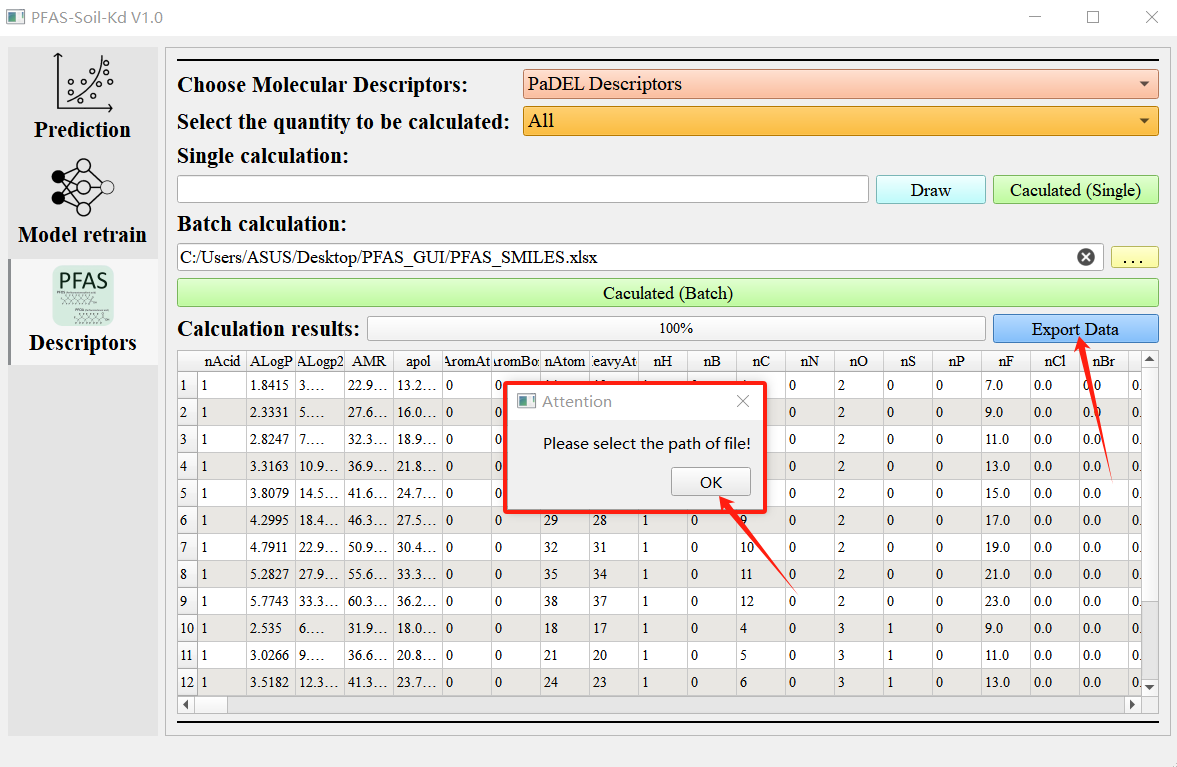
**Figure 21.** Error message in batch calculation.

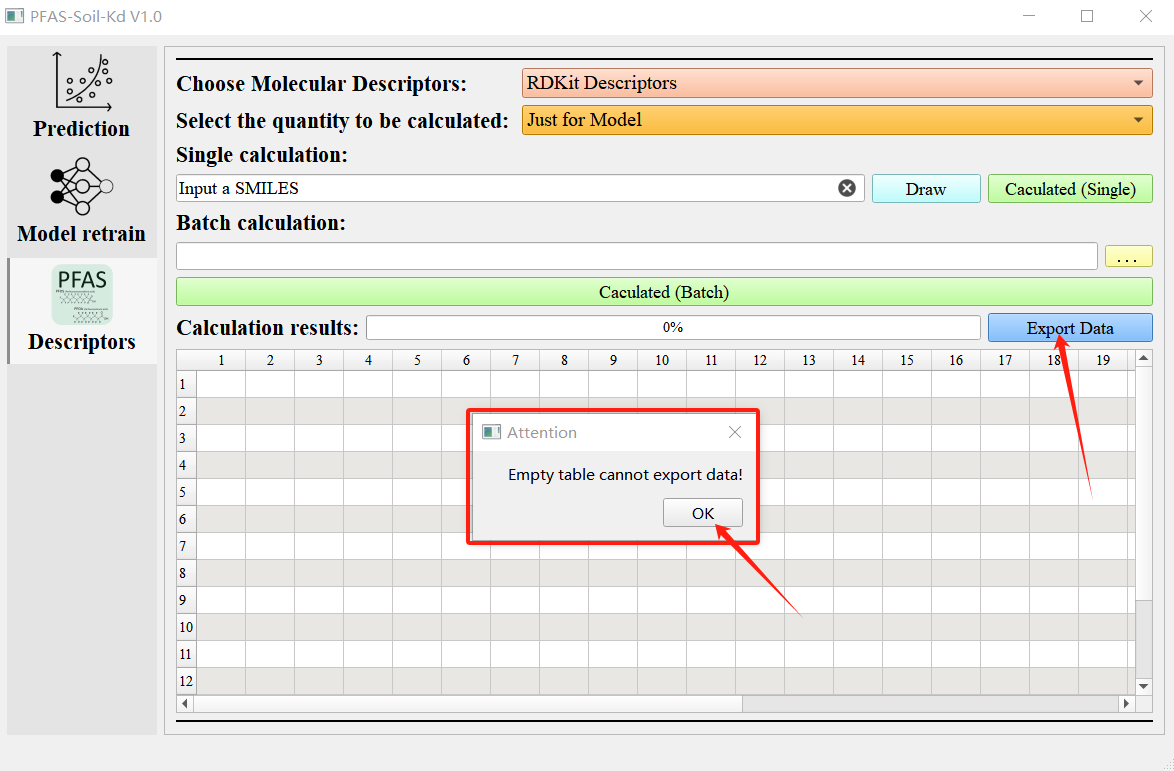
## 3.6 Molecular descriptor data export

Whether it is individual calculation data or batch calculation data, as long as it is displayed in the table below, it can be exported through the "Export Data" button. Click the "Export Data" button, select the path to save the data, the name of the saved file, and the file format. Data export supports exporting files in three formats: xlsx, csv, and txt. Users can choose according to their needs. After successful data export, there will be a message prompt "Export data successfully!". As shown in Figure 22.

If you click the "Export Data" button and return without selecting the save path and name of the data, the software will prompt you with the message "Please select the path of file!". Due to the fact that data export is based on the data in the table, when the data in the table is empty, performing the data export operation will prompt the software with the message "Empty table cannot export data!". The specific instructions are shown in Figure 23.

## D:\微信\文档\WeChat Files\wxid_uhn16qzpg3gc22\FileStorage\Temp\1725254245499.png

**Figure 22.** Molecular descriptor calculation data export.



**Figure 23.** Prompt information for the export process of molecular descriptor results.