EasyCID Documentation

Release 1.0

Wang Yue

Aug 12, 2022

CONTENTS

CHAPTER

ONE

QUICK EASYCID OVERVIEW

EasyCID is a tool for raman spectra identification. It provides user-friendly interfaces to manage the spectral library, train CNN model, predict the components in mixture, display the results intuitively and generate the analysis report. This documentation gives an detail introduction to EasyCID program, which describes the following:

- Install EasyCID
- Quickly use EasyCID for modeling and analysis
- Some Basic Concepts

CHAPTER

TWO

INSTALL EASYCID

2.1 Installation

The setup file of EasyCID is a single EXE created by Inno Setup, and it supports for all versions of Windows in use today. When the setup file are double-clicked, the setup wizard will take care of the whole installation procedure.

The EasyCID program consists of following entries:

File	Function
EasyCID.exe	main program
unins000.exe	Uninstall wizard for unistalling DeepCID_GUI from your computer.

2.2 System Requirements

If you want to run EasyCID on your PC, the recommanded system requirements are as follows:

Operating Systems

- Windows 7
- Windows 10

Recommanded Hardware

- GPU compute capability 3.5 or greater (visit NVIDIA DEVELOPER)
- CUDA and cuDNN
- Intel Core i5 or greater
- 4 GB RAM or more
- 3GB hard drive space
- monitor with 1024×768 pixels or higher

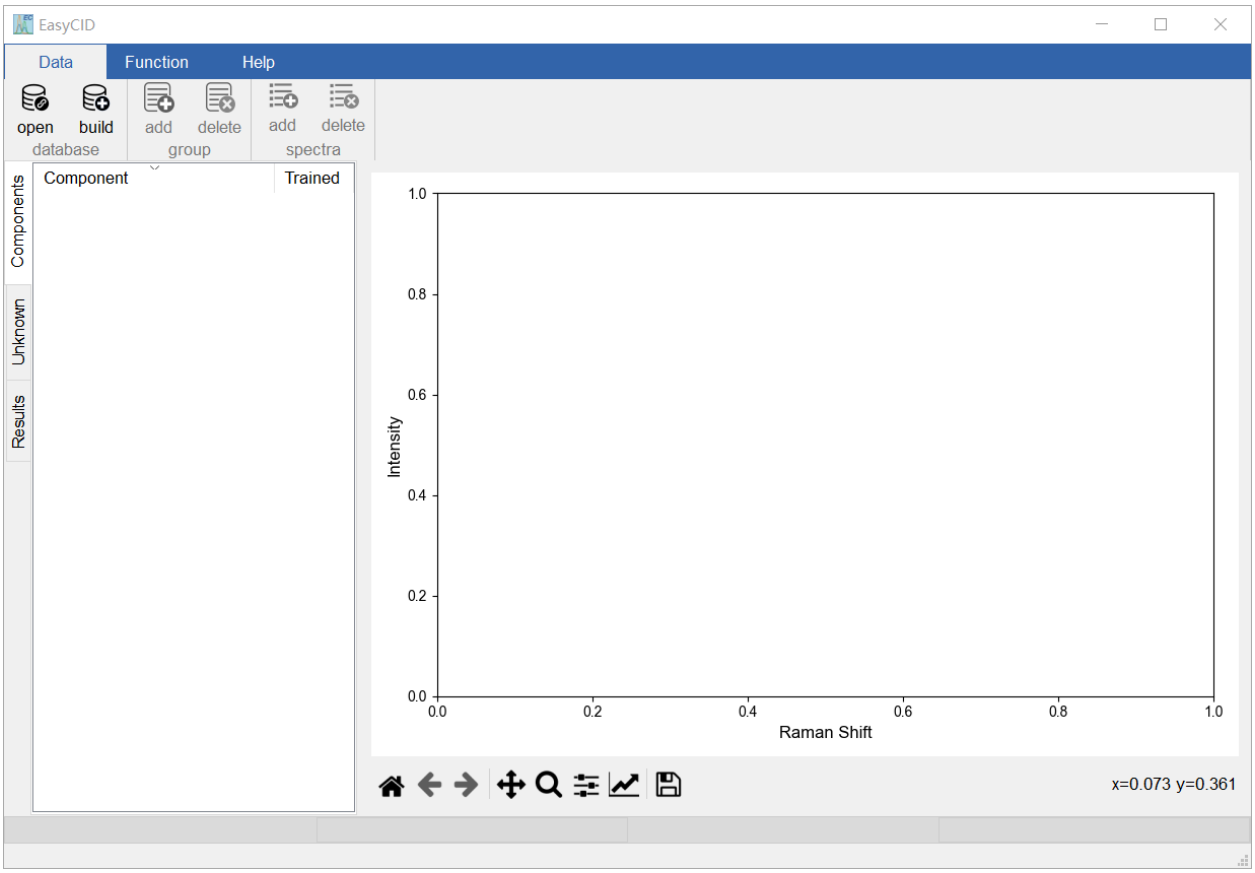
If your PC does not have the recommended configuration, you can run EasyCID, but it may take a little longer to train the CNN models.

CHAPTER

THREE

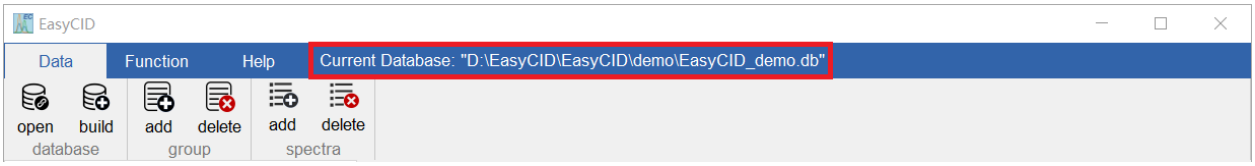
QUICKLY START EASYCID

This sub-section delivers you a brief introduction on how to use EasyCID. If EasyCID has been installed following the previous sub-section, it can be started by double clicking its icon on the desktop or single clicking from the start-menu of windows operation system. The Main Window of EasyCID is shown below:



3.1 build new database

EasyCID uses a local database based on sqlite to manage Raman spectral data. In order to set up a new database, please select **Data/build** in the menu of database. Then EasyCID will automatically link to the new database and give a tip message in the **Main Window**.



3.2 open database

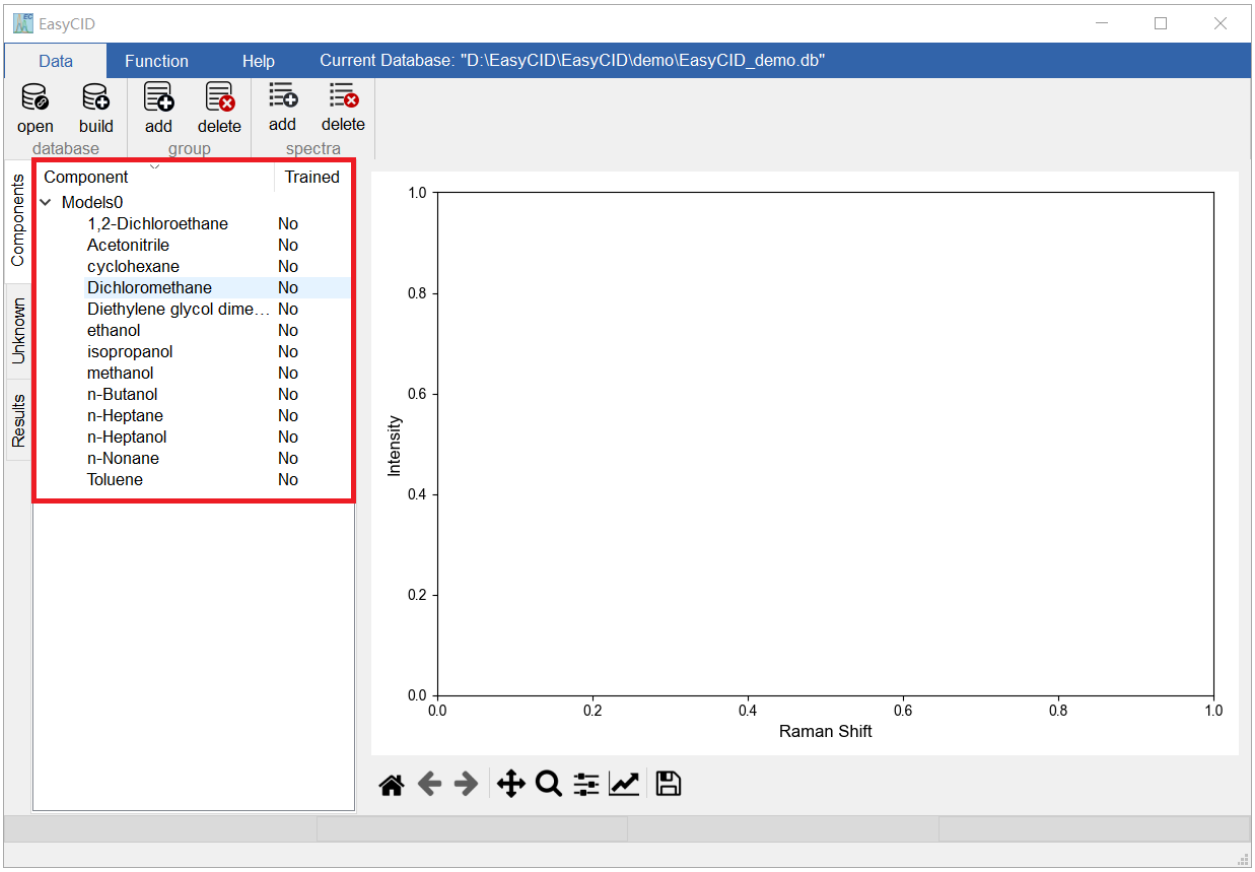
If you already have a built database, In order to link to this database, please select **Data/build** in the menu of database. *Note that only databases created through EasyCID can be successfully connected.*

3.3 Import Spectra of pure components

To carry out the next modeling process, the spectra file of pure components should be imported into Database firstly. In order to import spectra, please select **Data/add** in the menu of group. All the spectra file formats that EasyCID supports as follows:

- B&W Tek spectra files: pure text file which is proposed by B&W Tek Inc. for spectra storage and generated by BWSpec program.
- CSV/TXT files: Simple text format with comma-separated values or character-separated values (CSV), each row represents a spectrum.
- SPC files: SPC file format is a file format for storing and exchanging spectroscopic data, and it was invented by Galactic Industries.
- JDX files: JDX file is a JCAMP-DX Format Data. JCAMP is an acronym derived from Joint Committee on Atomic and Molecular Physical Data. It is a binary, chemical spectroscopy format.

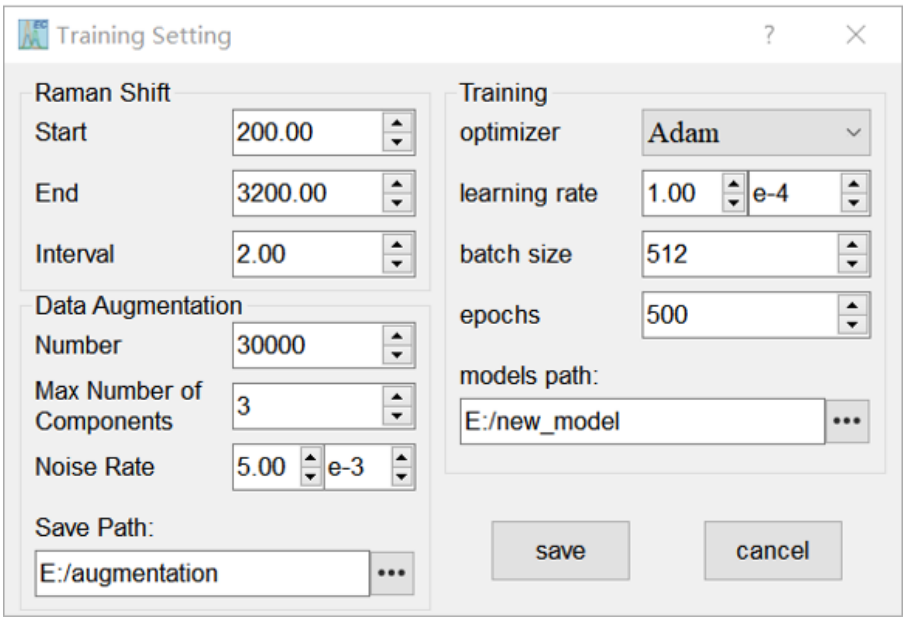
After the spectra file load into EasyCID, one can see the names of the spectra in the **Main Window**. The initial name of the corresponding group is similar as the format of “Models1”.



Double click on the name of component will plot its Raman spectrum on the right **Plot** Area.

3.4 Training the CNN models

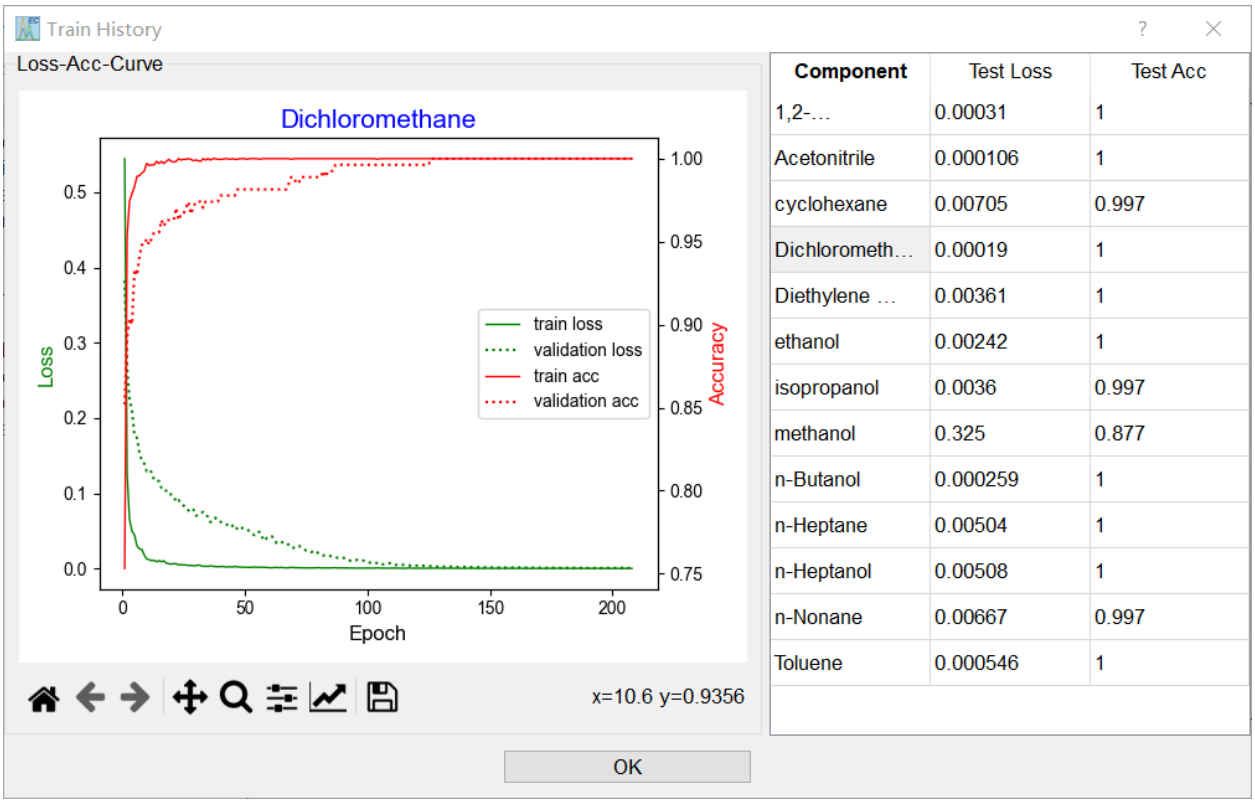
To build CNN models for a specific task, one need to adjust the parameters of the modeling process. In order to build CNN models, please select the name of a *Group* or a *spectrum*, then click **Function/start** in the menu of training. The **Training Parameters Window** will subsequently pop up for adjusting the parameters used in training process.



The description of all adjustable parameters and the corresponding recommend values are shown as follows:

	Name	Description	Recommend
Data Augmentation	number	the number of the generated simulated spectra	30000
	max number of components	the max number of components used to generate augmented spectra	3
	noise rate	control the noise level of the generated simulated spectra	0.5%
	save path	path for saving the augmented spectra. If empty, they are not saved	
Raman Shift	start	starting point of the analyzed Raman shift range	
	end	ending point of the analyzed Raman shift range	
	interval	sampling interval of the analyzed Raman shift	
Training Process	optimizer	minimize (or maximize) the loss function during the training procedure	Adam
	learning rate	control the adjust degree of the weights (and bias) of our network with respect the loss gradient.	0.00001
	batch size	the number of samples used for one weights (and bias) update step	512
	epochs	the number of times to train the entire sample set	500
	models path	save path of the weights (and bias) of CNN models	

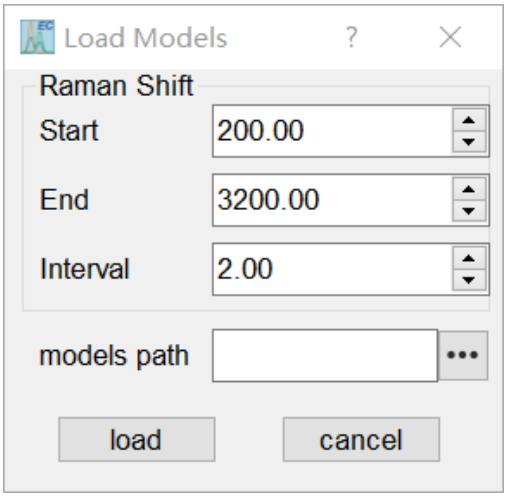
After setting all the parameters, the traning process will take place automatically and an progressBar of the training progress will appear in the **Main Window**. For better understand the concepts of parameters, please see *Hyperparameters of CNN models*. After the training process, a training report will given on **Training Report Window**.



Component	Test Loss	Test Acc
1,2-…	0.00031	1
Acetonitrile	0.000106	1
cyclohexane	0.00705	0.997
Dichlorometh…	0.00019	1
Diethylene …	0.00361	1
ethanol	0.00242	1
isopropanol	0.0036	0.997
methanol	0.325	0.877
n-Butanol	0.000259	1
n-Heptane	0.00504	1
n-Heptanol	0.00508	1
n-Nonane	0.00667	0.997
Toluene	0.000546	1

3.5 Load Models

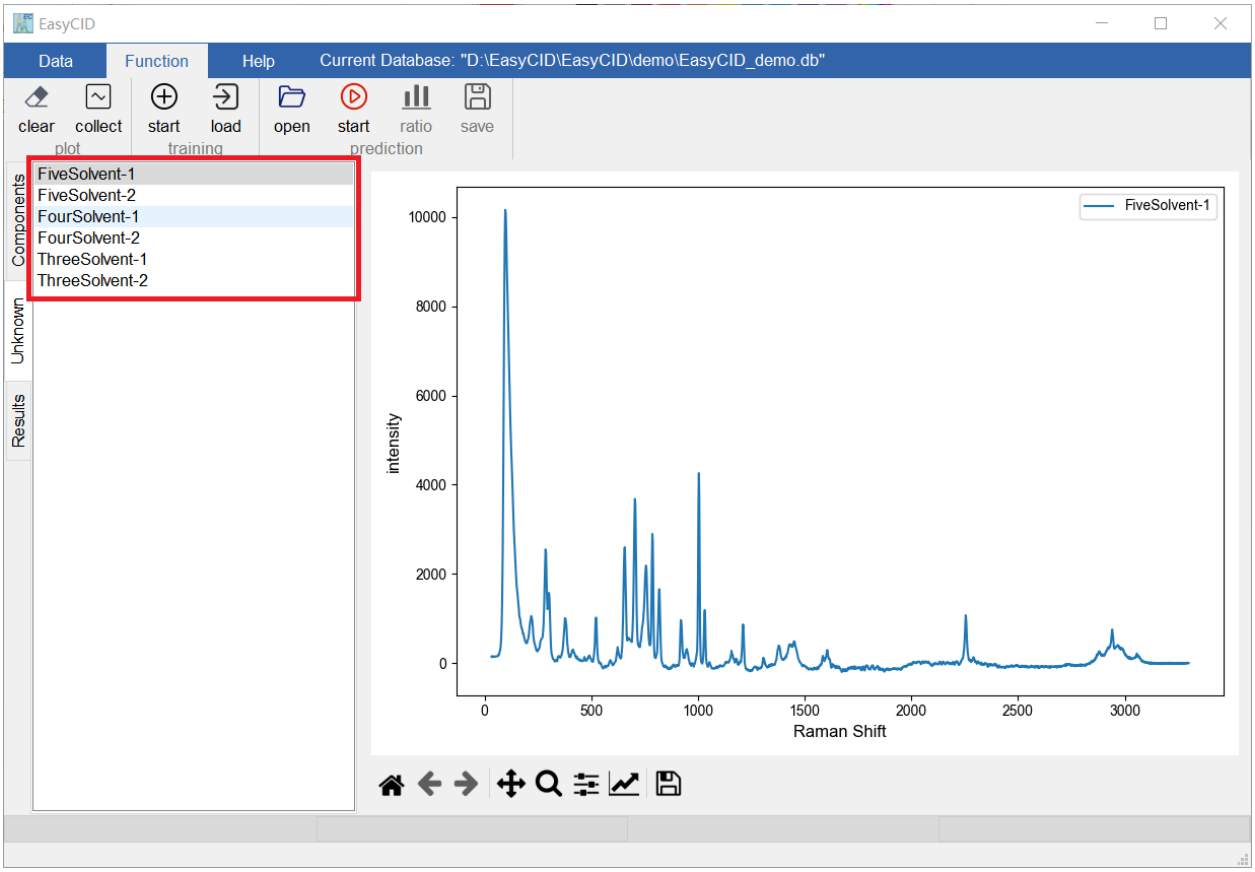
if you already have some models trained in EasyCID, in order to load those models, please select the group name and click on **Function/load** in the menu of training.



Then, **Link Models Window** will pop up. To successful link to the exist models, it is necessary to select the start, end and interval of the Raman shift corresponding to the models, and the storage folder for the models. *Note that* **If the models is built by EasyCID and the file named ModelsInfo.json is in the storage folder, the start, end and interval of the Raman shift will aotumatically set.**

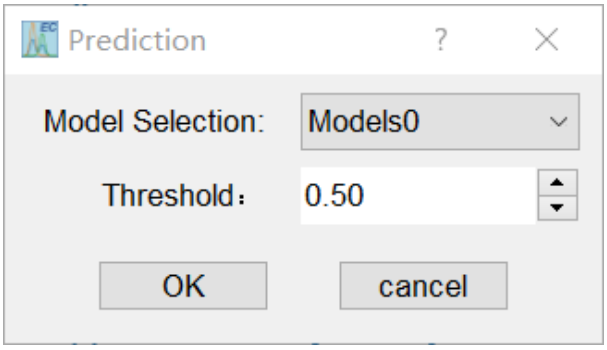
3.6 Perform analysis

To Perform analysis of unknown samples is simple in EasyCID. Since the CNN models have already been build in previous steps, it only takes the next two steps: Firstly, load spectra of samples for prediction by selecting the menu item **Function/open** in prediction, and select the spectra to be analyzed.

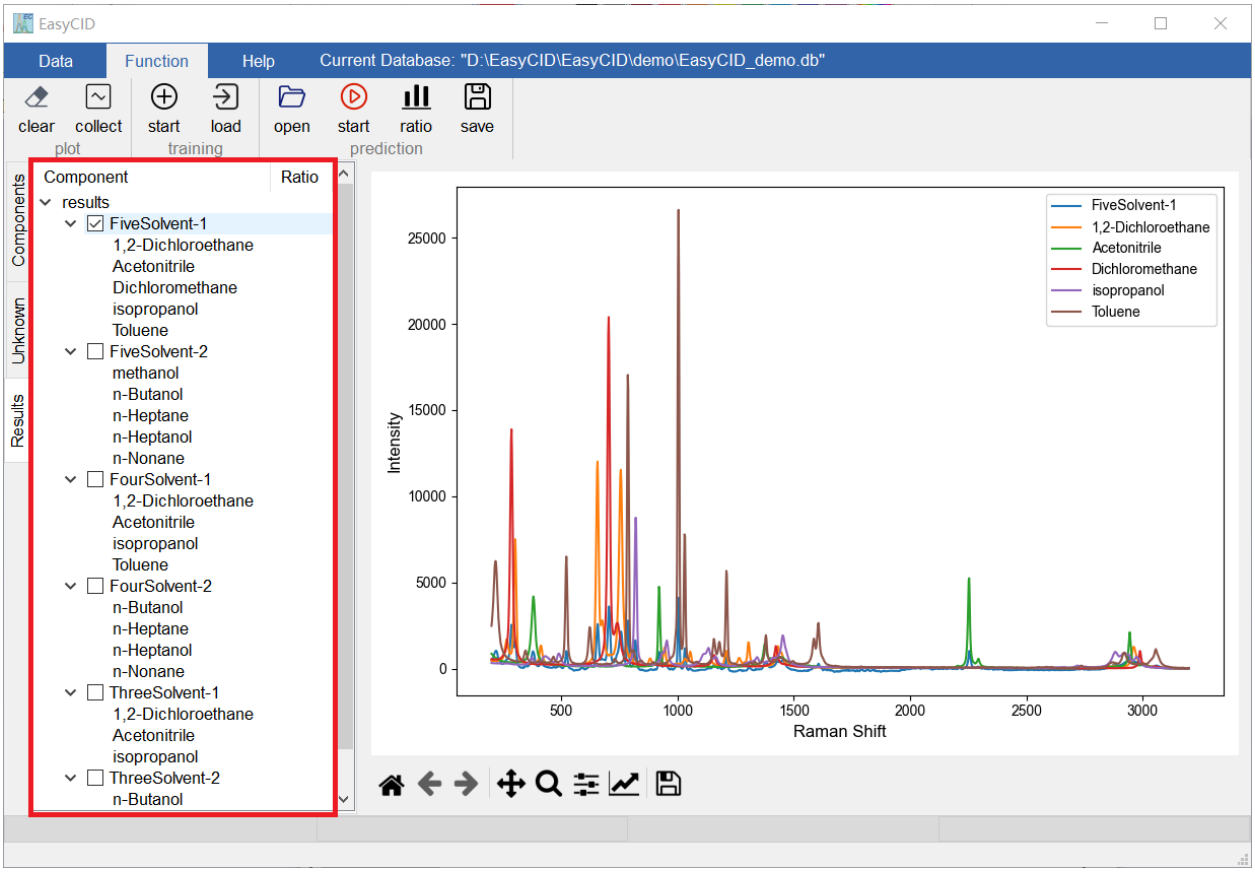


Double click on the name of mixture will plot its Raman spectrum on the right **Plot** Area.

Next, Clicking **Function/start** in the menu of prediction and then select the corrspound group and threshold in the pop-up **Prediction Window**.



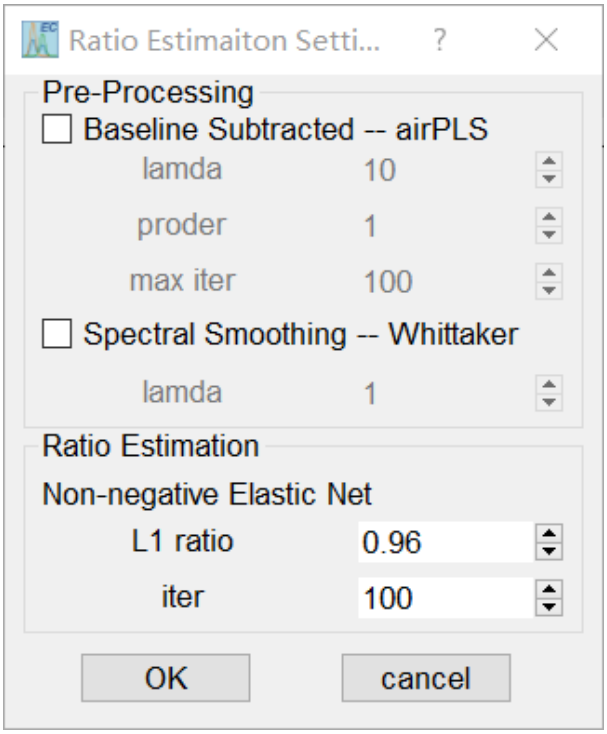
After prediction, the results will shown in **Main Window**.



Check the checkbox can plot the Raman spectra of the mixture and the corresponding components on the right **Plot** Area.

3.7 Ratio Estimation

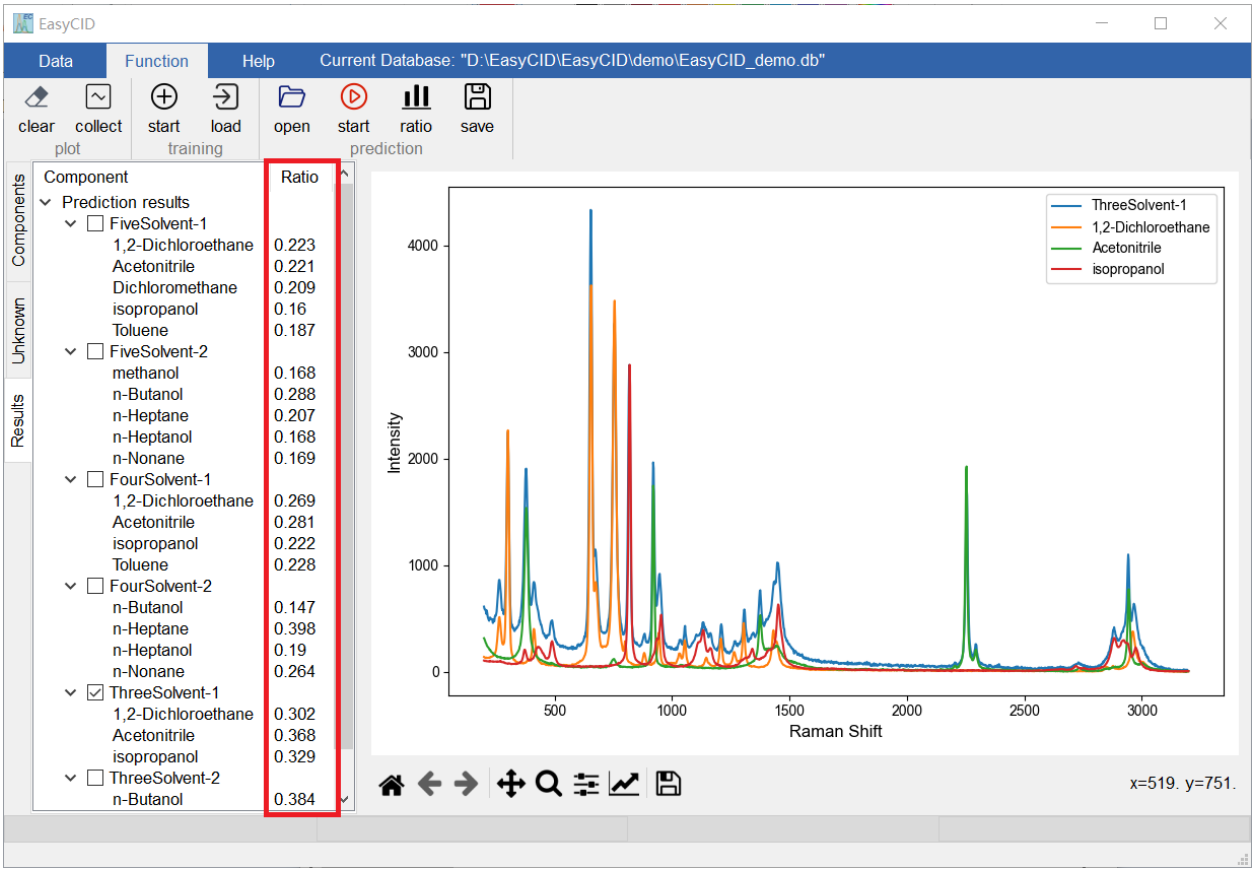
After obtaining the prediction results, the user can choose to perform further ratio estimation process. In order to perform ratio estimation process, please select **Function/ratio** in the menu of prediction. Then **Ratio Estimation Window** will pop up for adjusting the parameters of the used methods.



The description of all adjustable parameters and the corresponding recommend values are shown as follows:

	Name	Description	Recommend
AirPLS	lambda	the larger the lambda, the smoother the resulting background will be	10
	porder	adaptive iteratively reweighted penalized least squares for baseline fitting	1
	max iter	the maximum number of iterations	100
Whittaker smoother	lambda	starting point of the analyzed Raman shift range	1
Non-negative Elastic net	L1 ratio	adjust the proportion of L1 regularization in the overall penalty terms	0.96
	max iter	the maximum number of iterations	100

After setting all the parameters, the ratio estimation process will take place automatically and the ratio of each components will be shown in the **Main Window** at the end of the process. For better understand the concepts of parameters, please see *Basic Concepts*.



3.8 Save Results

You could save the prediction results and the corresponding ratios in CSV format. To save them, please select **Function/Save** in prediction.

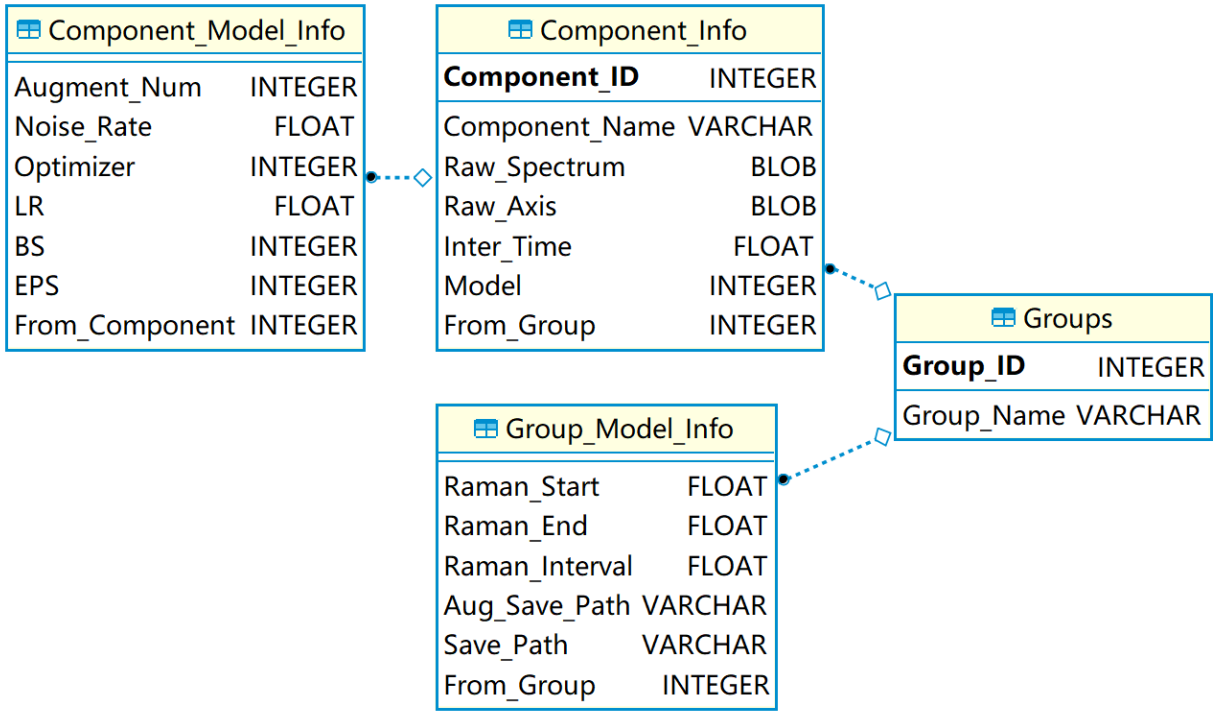
CHAPTER

FOUR

BASIC CONCEPTS

4.1 Database

The Raman spectra are stored in the SQLite database for better scaling, productivity and easy maintenance. The schema of the SQLite database is shown as follows:



EasyCID can extract the spectral information from raw Raman spectra in SPC, JCAMP-DX and TXT formats. Then, the extracted information, such intensity and Raman shift, is inserted into the database. Those Raman spectra are divided into groups for specific tasks. The database can also records the training information of the CNN model, which facilitates the tuning of the models. These information can be retrieved from the database for data augmentation, training and prediction.

4.2 Data Augmentation

Data augmentation can generate enough spectra to train deep neural networks and improve the generalization ability of models.

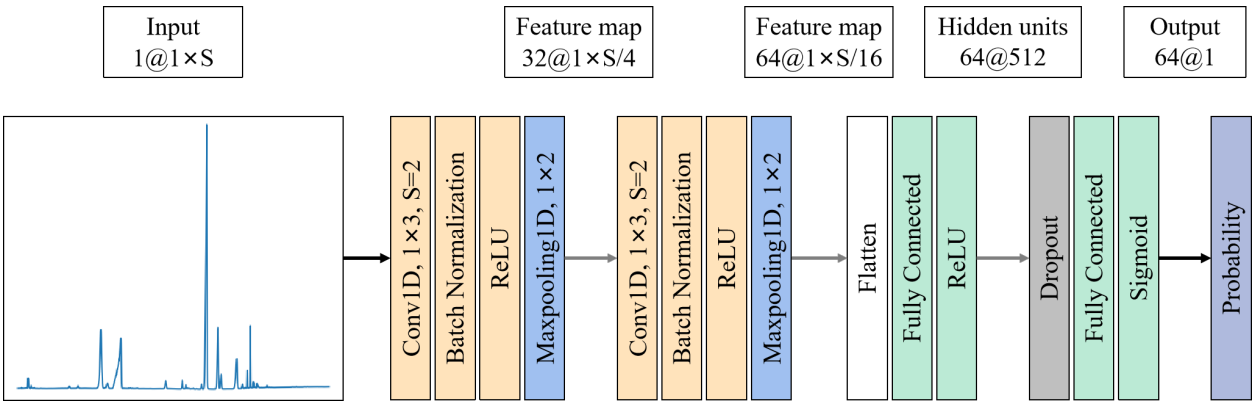
For the spectrum of each compound in a database, its augmented spectra were obtained by adding different pure spectra at random ratios with random Gaussian noise. There are two types of augmented spectra: positives and negatives. For positives, the spectra of the other interference compounds were randomly combined with the spectrum of the selected compound, and the ratio of the selected compound in each augmented spectra was set as not less than 10 percent. For negatives, spectra (without the selected component) were randomly sampled from the spectral database, and their ratios were also randomly generated.

Then, these spectra were summed according to the ratios. For each augmented spectra, the noise level (standard deviation of Gaussian noise) was equal to the product of the maximum value of the spectrum and the noise rate. The number of components for augmenting spectra ranged from 2 to the max number of components, which is a parameter that can be adjusted by users. The ratio of the augmented spectra of different max number of components was determined by their combinatorial numbers. For example, when the total number of components is 13 and the max components is 4. For positives, the ratio of the augmented spectra with different max number of components from 2 to 4 is equal to C(1,12):C(2,12):C(3,12). For negatives, the ratio is equal to C(2,12):C(3,12):C(4,12).

Finally, the augmented spectra were randomly divided into the training, validation, and test sets in the ratio of 8:1:1.

4.3 CNN models

CNN can extract the features of the spectra and learn to identify components of unknown spectra of mixtures.

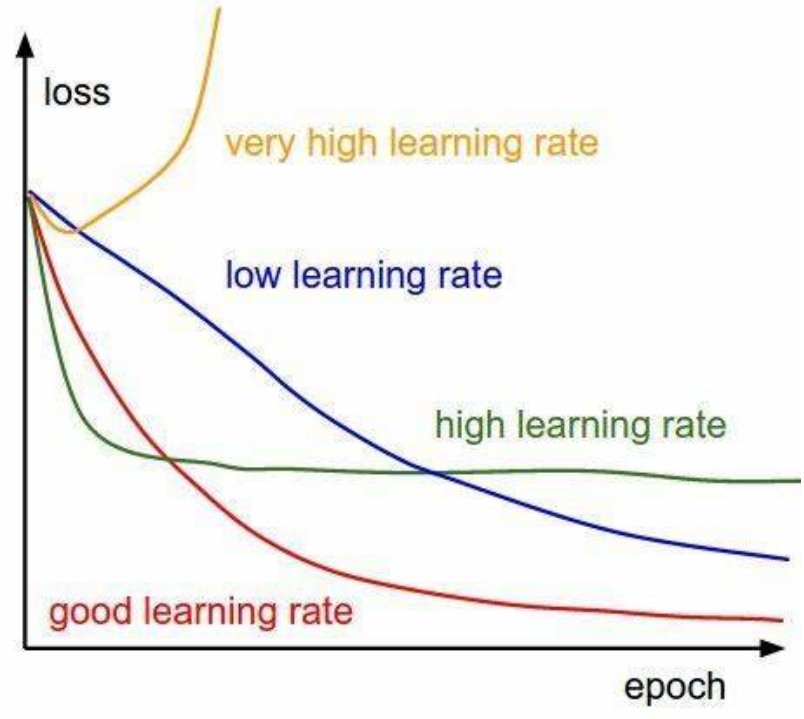


one-dimensional CNN with six hidden layers is used. The convolution layers are used to extract the features of spectra, which consists of a one-dimensional convolution, batch normalization and rectify linear unit (ReLU) function. Each convolution layer is followed by a pooling layer, which eliminates noisy information and reduce the amount of data. The fully connected layer are used as a classifier to determine the existence of the component based on the extracted features, where ReLU is used as the activation function in first layer while Sigmoid is used in second layer. Between two fully connected layers, Dropout is used to increase the generalization ability of the model. Binary cross entropy is used as the loss function.

4.4 Hyperparameters of CNN models

From all the adjustable hyperparameters for training CNN models, we selected the most important ones for the training step to reduce the learning cost for non-machine-learning experts.

- EasyCID offers four adaptive optimizers, and the default **Adam** optimizer typically requiring no tuning or little fine-tuning to accommodate a wide range of deep learning tasks.
- **Learning rate** is the most important hyperparameter for training CNN model. Its value can be adjusted by combining the training report given on **Training Report Window** and referring to the following figure:



It is recommended to set the initial learning rate from 0.0001 to 0.001.

- **Batch size** affects the optimization degree and speed of the algorithm, and it depends on the size of the GPU/CPU memory.
- It's better to set the **epochs** larger since the early stopping strategy was applied in training process. When the performance of the model does not improve within a certain period, this strategy will terminate the training process and save parameters with optimal performance of all time.

4.5 Baseline Subtracted

Adaptive iteratively reweighted penalized least squares (airPLS) is used as the baseline subtracted method that works by iteratively changing weights of sum squares errors between the fitted baseline and original signals.

The details for airPLS can be seen at Baseline correction using adaptive iteratively reweighted penalized least squares

4.6 Spectral Smoothing

Whittaker smoother is a spectral smoothing method that based on penalized least squares, which obtains smoothed data by adjusting the weight between the fidelity and roughness of data.

The details for Whittaker smoother can be seen at A Perfect Smoother

4.7 Regression Analysis

Non-negative elastic net is used to determine the relative ratio of each pure component in the mixture. The method can be treated as a hybrid of lasso[37] and ridge regression, which combines both the L1 and L2 regularization to improve the regularization of statistical models. Considering the concentration of the compound should be non-negative, the non-negative restrictions is imposed to make the results more reasonable.

The details for Non-negative elastic net can be seen at Mixture analysis using non-negative elastic net for Raman spectroscopy

CHAPTER

FIVE

INDICES AND TABLES

- genindex
- modindex
- search