Trendline-based IM-MS Feature Filtering Software V1.0

User Manual

1. Software Introduction

Ion mobility (IM) is an emerging technology for separation of multiple compound ions based on their spatial sizes. Ion mobility coupled with mass spectrometry (IM-MS) integrates IM separation orthogonally with mass spectrometry (MS) detection, which could enable detection of more compounds. In addition to mass-to-charge ratios (*m/z*) measured from conventional MS detection, IM-MS could provide spatial size information of compound ions (e.g.: collision cross-section, CCS) for identification of their chemical structures. The IM-MS techniques have been increasingly applied to analysis of multiple compounds in complex matrix. Previous studies have demonstrated that CCS of compound ions with similar structures presented regular changes with their *m/z*. Performing *m/z*-CCS regressive prediction analysis on chemical analogues could explore their IM-MS feature trendline and construct their prediction intervals, which is conducive to search and characterize interested compounds in complex matrix. This software was developed to study the distribution trends of IM-MS features for compounds with specific structure types and realize the automatic and rapid filtering of potential chemical analogues from original IM-MS data, which improving the efficiency for multi-compound identification of complex matrix.

This software contains two functions. **The first function is analysis of IM-MS feature distribution trend for chemical analogues:** Users only need to import the *m/z* and CCS of chemical analogues used for model training. Different regression models could be automatically performed and compared on imported training *m/z*-CCS. The best regression model will be automatically selected as the IM-MS feature trendline of chemical analogues. **The second function is graded filtering of IM-MS features in complex matrix:** Users need to import the IM-MS experimental *m/z*-CCS data of complex matrix for filtering, as well as collected *m/z*-CCS of chemical analogues for model training. Firstly, the software automatically selects the best *m/z*-CCS regression model to fit IM-MS trendlines on the training *m/z*-CCS. Then, the estimation algorithms for upper/lower limits of CCS prediction and confidence intervals will be automatically established based on the *m/z*-CCS regression model type of training data. Finally, CCS prediction/confidence intervals for each experimental *m/z* could be estimated based on the established algorithms for upper/lower limits. By comparing the estimated CCS prediction/confidence intervals with experimental CCS associated to each experimental *m/z*, the IM-MS features in complex matrix will be automatically filtered/graded.

2. Software Installation

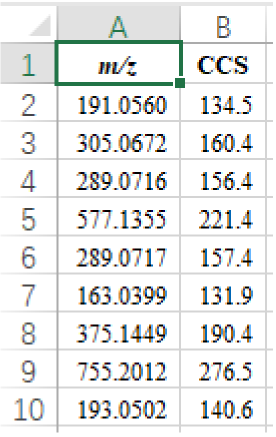
This software was developed by Python 3.9 programming and Qt framework, and packed into an executable file (.exe) that can be run directly on Windows 10 or later version system by Pyinstaller. Users just need to download the "**Trendline-based IM-MS Feature Filtering Software V1.0.exe**" to run directly on computers.

3. Software Operation Procedures (SOP)

3.1. Preparation of Import Files

3.1.1. Training m/z-CCS Data File of Chemical Analogues

Users need to collect the *m/z* and CCS of multiple chemical analogues from various sources, such as previous literatures, public databases or in-lab measurements. Appropriate candidate criteria of chemical analogues should be chosen or constructed by users based on recognized knowledges of chemistry. The *m/z* and CCS of chemical analogues should be input to an Excel file (.xlsx, xls. or .csv) as two adjacent columns with captions of “*m/z*” and “CCS” (**Fig. 1**). This data file is considered as “**Training *m/z*-CCS data**” and prepared for further profiling of IM-MS feature distribution trends and construction of IM-MS feature filtering intervals by *m/z*-CCS regressive prediction analysis.



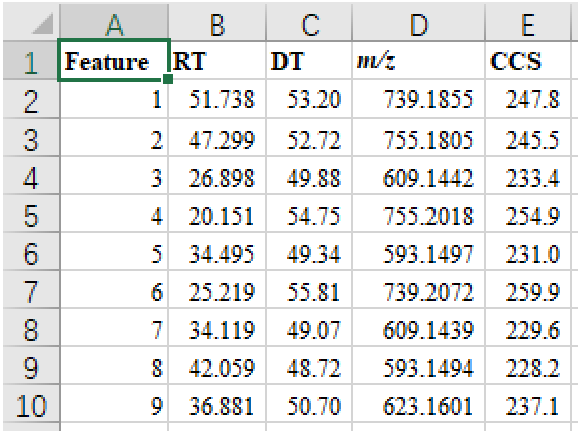
	A	B
1	*m/z*	CCS
2	191.0560	134.5
3	305.0672	160.4
4	289.0716	156.4
5	577.1355	221.4
6	289.0717	157.4
7	163.0399	131.9
8	375.1449	190.4
9	755.2012	276.5
10	193.0502	140.6

Fig. 1. Demonstration of Excel file for “Training *m/z*-CCS data” containing *m/z* and CCS of all collected chemical analogues.

3.1.2. Experimental IM-MS Data File of Complex Matrix

Users need to acquire the IM-MS data of complex matrix sample on self-owned IM-MS platform and export the *m/z* and CCS of all detected compounds by instrument-adapted data analysis software. **Please note that IM-MS data without available CCS is not applicable.** These experimental *m/z* and CCS should be input to another Excel file (.xlsx, xls. or .csv) also as two adjacent columns with captions of “*m/z*” and “CCS”. Combined inputting with other information, such as IM-MS feature number, retention time (RT) and drift time (DT), is acceptable (**Fig. 2**). This data file is considered as

"**Experimental IM-MS data**" and prepared for rapid filtering of IM-MS features for potential chemical analogues in complex matrix sample using previous constructed filtering interval based on *m/z*-CCS regressive prediction analysis on training data.

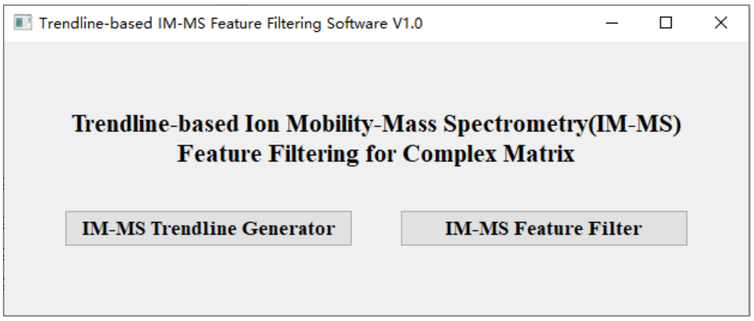


	A	B	C	D	E
1	**Feature**	**RT**	**DT**	***m/z***	**CCS**
2	1	51.738	53.20	739.1855	247.8
3	2	47.299	52.72	755.1805	245.5
4	3	26.898	49.88	609.1442	233.4
5	4	20.151	54.75	755.2018	254.9
6	5	34.495	49.34	593.1497	231.0
7	6	25.219	55.81	739.2072	259.9
8	7	34.119	49.07	609.1439	229.6
9	8	42.059	48.72	593.1494	228.2
10	9	36.881	50.70	623.1601	237.1

Fig. 2. Demonstration of Excel file for "Experimental IM-MS data" containing *m/z* and CCS (with or without other information) of all IM-MS detected compounds in complex matrix.

3.2. Main Interface

Double click "**Trendline-based IM-MS Feature Filtering Software V1.0.exe**" to open the software. Software title of "**Trendline-based Ion Mobility-Mass Spectrometry (IM-MS) Feature Filtering for Complex Matrix**" and two function buttons of "**IM-MS Trendline Generator**" and "**IM-MS Feature Filter**" could be observed on the main interface (**Fig. 3**).



Trendline-based IM-MS Feature Filtering Software V1.0

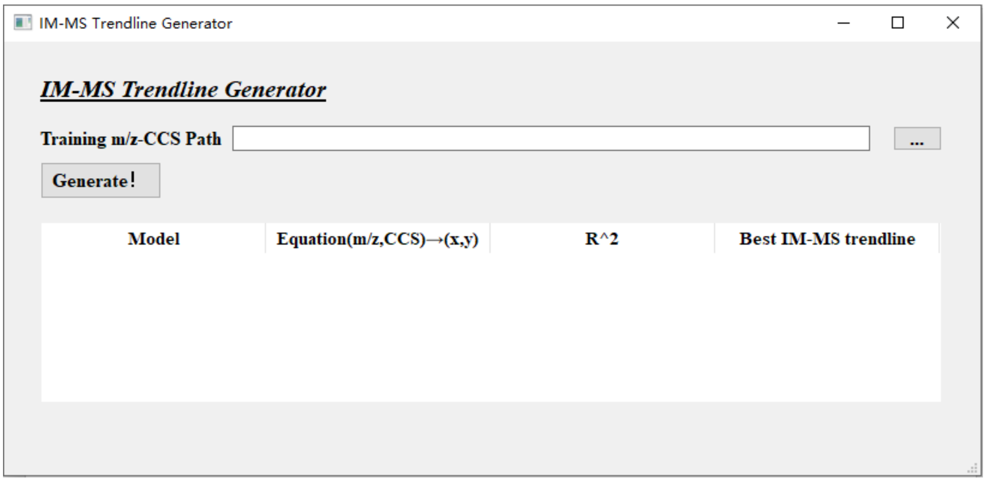
Trendline-based Ion Mobility-Mass Spectrometry(IM-MS)
Feature Filtering for Complex Matrix

Fig. 3. Main interface of "Trendline-based IM-MS Feature Filtering Software V1.0.exe"

3.3. Analysis of IM-MS Feature Distribution Trend for Chemical Analogues

3.3.1. Operation Procedures of "IM-MS Trendline Generator"

Users need to click the button "**IM-MS Trendline Generator**" on the main interface of the software to open the first functional interface for IM-MS feature distribution trend analysis.



Model	Equation(m/z,CCS)→(x,y)	R^2	Best IM-MS trendline

Fig. 4. Functional interface of "IM-MS Trendline Generator" in "Trendline-based IM-MS Feature Filtering Software V1.0.exe".

As shown in **Fig. 4**, users need to import previous prepared Excel file of "Training *m/z*-CCS data" (refer to *section 3.1.1*) at "**Training m/z-CCS Path**". Then clicking the button "**Generate!**", the software will automatically compare the fitting performances between different *m/z*-CCS regression models to select the best IM-MS feature trendline for chemical analogues. Detailed process on software is presented as follows:

I. Multi-model *m/z*-CCS regression

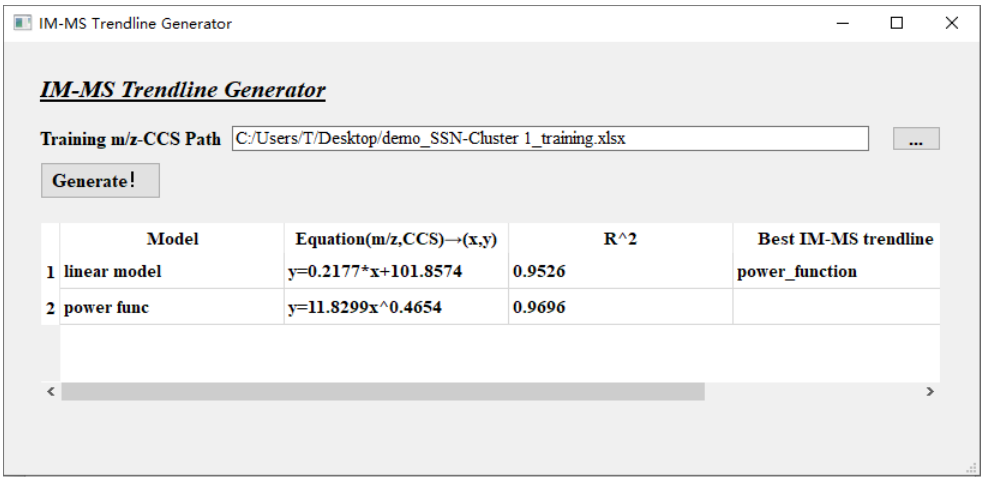
Original *m/z* and CCS in "Training *m/z*-CCS data" are defined as variables x_i and y_i, respectively. Linear model ($\hat{y}_i = ax_i + b$) and power function ($\hat{y}_i = ax_i^b$) are selected for *m/z*-CCS regression. Algorithms for estimation of *a* and *b* are different for two selected models.

i. Linear model ($\hat{y}_i = ax_i + b$)

For linear model regression, *a* and *b* could be calculated using least squares estimation directly with original values of *m/z* and CCS as independent variables (x_i) and dependent variables (y_i). Meanwhile, the coefficient of determination for *m/z*-CCS linear regression model (R^2) are calculated with the **Eq. 1**:

$$R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

The *n* in **Eq. 1** represents the count of training *m/z*-CCS data. Regression model type name "linear model", regressive equation, and R^2 value will be presented under "**Model**", "**Equation (m/z, CCS)→(x, y)**" and "**R^2**" on the result output area of "**IM-MS Trendline Generator**" as **Fig. 5**.



Training m/z-CCS Path: C:/Users/T/Desktop/demo_SSN-Cluster 1_training.xlsx

	Model	Equation(m/z,CCS)→(x,y)	R^2	Best IM-MS trendline
1	linear model	y=0.2177*x+101.8574	0.9526	power_function
2	power func	y=11.8299x^0.4654	0.9696	

Fig. 5. Demonstration of interface output for "IM-MS Trendline Generator" in "Trendline-based IM-MS Feature Filtering Software V1.0.exe".

ii. Power function ($\hat{y}_i = ax_i^b$)

For power function regression, least squares estimation could not be used directly. By taking the natural logarithm on both sides, original regressive equation of power function could transform to a pseudo-linear model ($\ln\hat{y}_i = b\ln x_i + \ln a$). The natural logarithms of *m/z* and CCS in "Training *m/z*-CCS data" are automatically calculated. Next, natural logarithm of *a* and original value *b* will be calculated by least squares estimation. Thus, original value of *a* could be calculated indirectly. The coefficient of determination for power function-transformed pseudo-linear regression model (R'^2) are calculated using the **Eq. 2**:

$$R'^2 = 1 - \sum_{i=1}^{n} (\ln y_i - \ln \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (\ln y_i - \overline{\ln y_i})^2$$

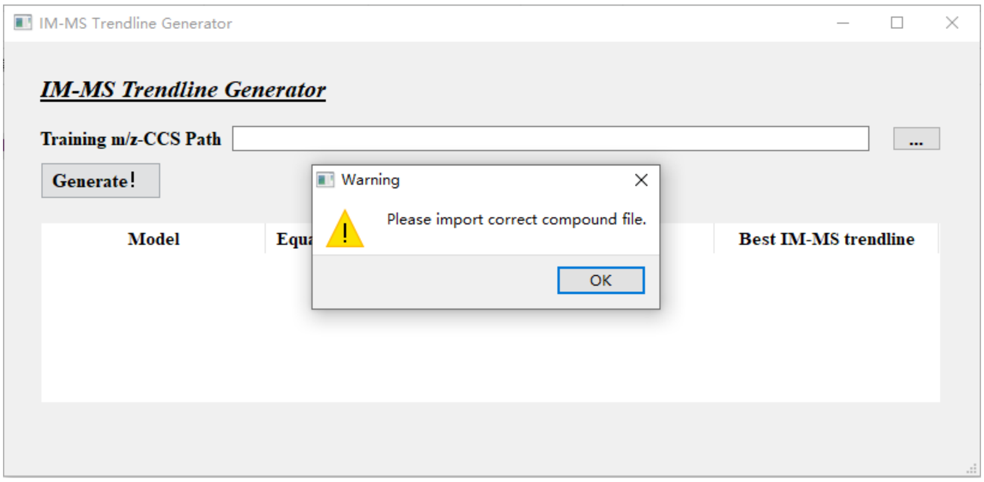
The *n* represents the count of training *m/z*-CCS data. Regression model type name of "**power func**", original regressive equation of power function, and R'^2 value of pseudo-linear regression model will be also presented under "**Model**", "**Equation (m/z, CCS)→(x, y)**" and "**R^2**" on the result output area of "**IM-MS Trendline Generator**" as **Fig. 5**.

II. Selection of the best IM-MS feature trendline

The best IM-MS feature trendline of training chemical analogues is selected by comparing the values under "**R^2**" on the result output area of "**IM-MS Trendline Generator**". If the maximum value under "**R^2**" is associated with "**power func**" under "**Model**", that is, R'^2 of power function is not smaller than R^2 of linear model during multi-model *m/z*-CCS regression, the software will automatically select power function model as the best IM-MS feature trendline. The result of best IM-MS feature trendline will be presented as "**power function**" under "**Best IM-MS trendline**" on the output area of "**IM-MS Trendline Generator**" as Fig. 5. Otherwise, the result of best IM-MS feature trendline will be presented as "**linear model**".

3.3.2. Common Errors on "IM-MS Trendline Generator"

If storage path at "**Training m/z-CCS Path**" was not selected with the Excel file that meet the format requirements of "**Training *m/z*-CCS data**" in *section 3.1.1*, the "IM-MS Trendline Generator" functional interface will prompt warnings as "Please import correct compound file" (Fig. 6).



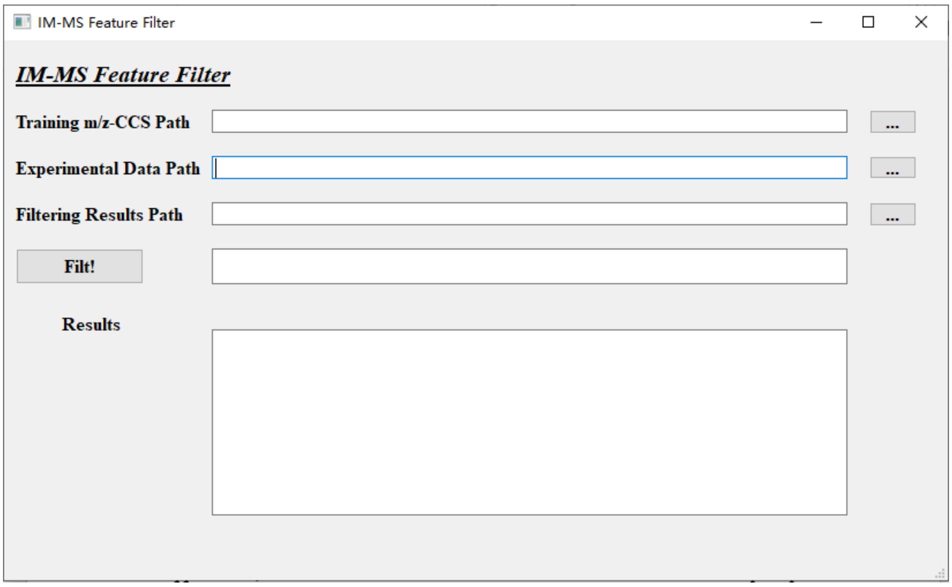
Please import correct compound file.

Fig. 6. Window warning message for not selecting the storage path of the Excel file met the format requirements of "Training m/z-CCS data" (refer to *section 3.1.1*) at "Training m/z-CCS Path" of "IM-MS Trendline Generator".

3.4. Graded Filtering of IM-MS Features for Potential Analogues in Complex Matrix

3.4.1. Operation Procedures of "IM-MS Feature Filter"

Users need to click the button "**IM-MS Feature Filter**" on the main interface of the software to open the second functional interface for IM-MS features filtering and grading.



IM-MS Feature Filter

IM-MS Feature Filter

Training m/z-CCS Path

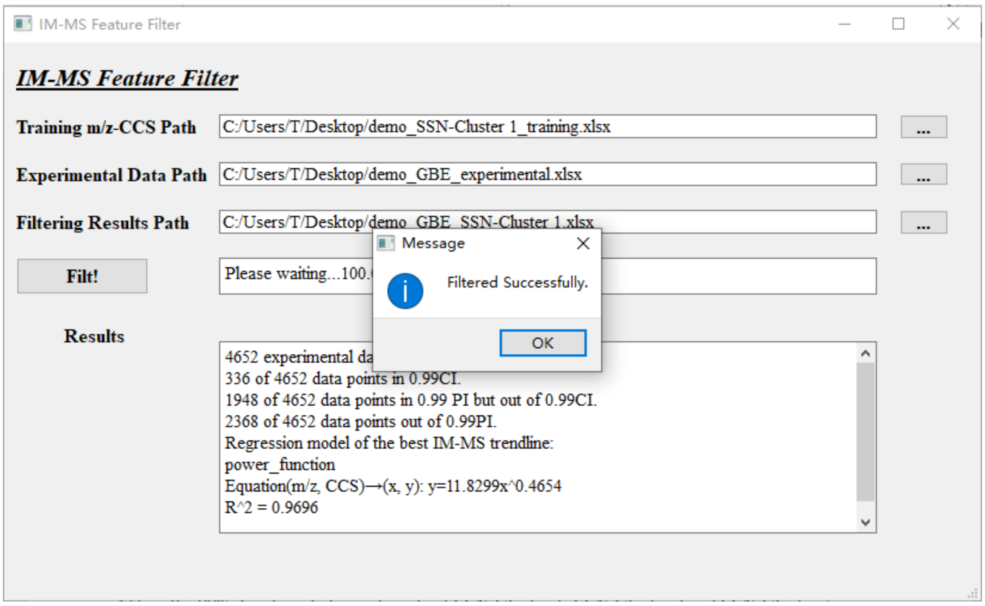
Experimental Data Path

Filtering Results Path

Results

Fig. 7. Functional interface of “IM-MS Feature Filter” in “Trendline-based IM-MS Feature Filtering Software V1.0.exe”.

As shown in Fig. 7, users need to import previous prepared Excel file of “**Training *m/z*-CCS data**” (refer to *section 3.1.1*) at “**Training m/z-CCS Path**”. Additionally, the Excel file of “**Experimental IM-MS data**” (refer to *section 3.1.2*) should be imported at “**Experimental Data Path**”. Moreover, the storage path for the results of graded IM-MS feature filtering should be selected at “**Filtering Results Path**”. Then clicking the button “**Filt!**”, the software will automatically filter and grade the IM-MS features in the experimental complex matrix data based on *m/z*-CCS regressive prediction analysis on training data of chemical analogues. Detailed processes on software is presented as follows:


IM-MS Feature Filter

IM-MS Feature Filter

Training m/z-CCS Path: C:/Users/T/Desktop/demo_SSN-Cluster 1_training.xlsx

Experimental Data Path: C:/Users/T/Desktop/demo_GBE_experimental.xlsx

Filtering Results Path: C:/Users/T/Desktop/demo_GBE_SSN-Cluster 1.xlsx

Please waiting...100.

Message

Filtered Successfully.

```
4652 experimental da
336 of 4652 data points in 0.99CI.
1948 of 4652 data points in 0.99 PI but out of 0.99CI.
2368 of 4652 data points out of 0.99PI.
Regression model of the best IM-MS trendline:
power_function
Equation(m/z, CCS)→(x, y): y=11.8299x^0.4654
R^2 = 0.9696
```

Fig. 8. Demonstration of interface output for "IM-MS Feature Filter" in "Trendline-based IM-MS Feature Filtering Software V1.0.exe".

I. Selection of the best IM-MS feature trendline for chemical analogues

Basing on the *m/z* (x_i) and CCS (y_i) in training data file of chemical analogue, the software will automatically select the best IM-MS feature trendline through comparing *m/z*-CCS regression performances between linear model and power function. Principles and detailed processes was highly similar with *section 3.3.1*, which is not detailedly described here. The regression model type, *m/z*-CCS regressive equation and coefficients of determination will be automatically output as results on the interface of "**IM-MS Feature Filter**" (**Fig. 8**).

II. Construction of graded filtering intervals for potential analogues

Graded filtering intervals for mining of potential analogues in complex matrix are constructed by the regressive prediction interval and confidence interval of training *m/z*-CCS data. The software will automatically estimate the *m/z*-CCS regressive prediction/confidence intervals using different algorithms according to the *m/z*-CCS regression model type of the best IM-MS feature trendline:

i. Linear model ($\hat{y}_i = ax_i + b$)

If the best IM-MS feature trendline of training chemical analogues fit the *m/z*-CCS linear regressive model, the residual (ε_i) for each training CCS will be directly calculated using the **Eq. 3**:

$$\varepsilon_i(Linear) = y_i - \hat{y}_i$$

Then, the standard deviation (SD_ε) and standard error (SE_ε) of ε_i are automatically calculated. In order to maximize the coverage of training data points for IM-MS feature filtering interval, 99%-level is selected for *m/z*-CCS regressive prediction/confidence statistics. The upper limit ($y_{upl\_0.99}$) and lower limit ($y_{lpl\_0.99}$) of prediction interval, as well as upper limit ($y_{ucl\_0.99}$) and lower limit ($y_{lcl\_0.99}$) of confidence interval, will be estimated using the following four equations (**Eq. 4-7**) for linear model *m/z*-CCS regression, respectively:

$$y_{upl\_0.99}(Linear) = \hat{y}_i + 2.58 \times SD_\varepsilon$$

$$y_{lpl\_0.99}(Linear) = \hat{y}_i - 2.58 \times SD_\varepsilon$$

$$y_{ucl\_0.99}(Linear) = \hat{y}_i + 2.58 \times SE_\varepsilon$$

$$y_{lcl\_0.99}(Linear) = \hat{y}_i + 2.58 \times SE_\varepsilon$$

ii. Power function ($\hat{y}_i = ax_i^b$)

If the best IM-MS feature trendline of training chemical analogues fit the *m/z*-CCS power regressive function, the regressive equation will transform to the pseudo-linear correlation ($\ln\hat{y}_i = b\ln x_i + \ln a$) between natural logarithms of *m/z* and CCS. The differences between natural logarithms of training CCS (y_i) and power function fitting CCS ($\hat{y}_i$) are automatically defined as pseudo-residuals (ε'_i) with calculating formula as follow (**Eq. 8**):

$$\varepsilon'_i(Power \rightarrow Pseudo\_linear) = \ln(y_i/\hat{y}_i)$$

Then, the standard deviation ($SD_{\varepsilon'}$) and standard error ($SE_{\varepsilon'}$) of ε'_i are also automatically calculated. Upper limits of 99%-level prediction interval, lower limits of 99%-level prediction interval, upper limits of 99%-level condidence interval, and lower limits of 99%-level condidence interval for natural logarithms of CCS are estimated by the following four equations (**Eq. 9-12**):

$$(\ln y)_{upl\_0.99} = \ln \hat{y}_i + 2.58 \times SD_{\varepsilon'}$$

$$(\ln y)_{lpl\_0.99} = \ln \hat{y}_i - 2.58 \times SD_{\varepsilon'}$$

$$(\ln y)_{ucl\_0.99} = \ln \hat{y}_i + 2.58 \times SD_{\varepsilon'}$$

$$(\ln y)_{lcl\_0.99} = \ln \hat{y}_i - 2.58 \times SD_{\varepsilon'}$$

On the basis of the increasing monotonicity of natural logarithm functions, the $y_{upl\_0.99}$, $y_{lpl\_0.99}$, $y_{ucl\_0.99}$ and $y_{lcl\_0.99}$ are estimated as below (*e* represents the natural constant) for power function *m/z*-CCS regression (**Eq. 13-16**), respectively:

$$y_{upl\_0.99}(Power) = \hat{y}_i \times exp(2.58 \times SD_{\varepsilon'})$$

$$y_{lpl\_0.99}(Power) = \hat{y}_i / exp(2.58 \times SD_{\varepsilon'})$$

$$y_{ucl\_0.99}(Power) = \hat{y}_i \times exp(2.58 \times SE_{\varepsilon'})$$

$$y_{ucl\_0.99}(Power) = \hat{y}_i / exp(2.58 \times SE_{\varepsilon'})$$

III. Filtering and grading of experimental IM-MS features for mining the potential chemical analogues in complex matrix

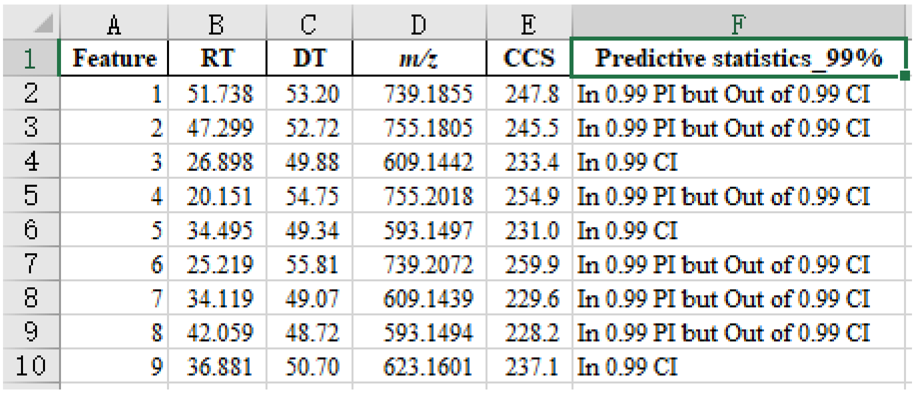
For each *m/z* (x_{exp}) in the experimental IM-MS data, their CCS upper limit ($y_{upl\_0.99}$) and CCS lower limit ($y_{lpl\_0.99}$) of 99%-level *m/z*-CCS regressive prediction interval, as well as CCS upper limit ($y_{ucl\_0.99}$) and CCS lower limit ($y_{lcl\_0.99}$) of 99%-level *m/z*-CCS regressive confidence interval, are estimated based on the best IM-MS feature trendline and corresponding alogrithms of training chemical analogues. Filtering and grading of experimental IM-MS features is operated by automatic comparing each experimental CCS (y_{exp}) with $y_{upl\_0.99}$, $y_{lpl\_0.99}$, $y_{ucl\_0.99}$ and $y_{lcl\_0.99}$ estimated from the corresponding experimental *m/z* (x_{exp}). Detailed assignment criteria and primary interpretation for filtering/grading results of experimental IM-MS features are described as below:

i. If CCS (y_{exp}) of an experimental IM-MS feature satisfies $y_{exp} > y_{upl\_0.99}$ or $y_{exp} < y_{lpl\_0.99}$, the filtering/grading result will be assigned as "Out of 0.99 PI", indicating that the *m/z*-CCS location of this experimental IM-MS feature is out of 99%-level prediction interval of the best IM-MS feature trendline for training chemical analogues. The *m/z*-CCS locations of these experimental IM-MS features are beyond the actual discrete range of training *m/z*-CCS data points. Therefore, they should be filtered out and not considered as chemical analogues of training compounds.

ii. If CCS (y_{exp}) of an experimental IM-MS feature satisfies $y_{ucl\_0.99} < y_{exp} \leq y_{upl\_0.99}$ or $y_{upl\_0.99} \leq y_{exp} < y_{lcl\_0.99}$, the filtering/grading result will be assigned as "In 0.99 PI but Out of 0.99 CI", indicating that the *m/z*-CCS location of this experimental IM-MS feature is in 99%-level prediction interval but out of 99%-level confidence interval of the best IM-MS feature trendline for training chemical analogues. Even though the *m/z*-CCS locations of these experimental IM-MS features are away from the distribution trendline, they are still within the actual discrete range of training *m/z*-CCS data points. Therefore, they should be retained for mining the potential analogues of training compounds from complex matrix IM-MS data.

iii. If CCS (y_{exp}) of an experimental IM-MS feature satisfies $y_{lcl\_0.99} \le y_{exp} \le y_{ucl\_0.99}$, the filtering/grading result will be assigned as "In 0.99 CI", indicating that the *m/z*-CCS location of this experimental IM-MS feature is in 99%-level confidence interval of the best IM-MS feature trendline for training chemical analogues. The *m/z*-CCS locations of these experimental IM-MS features are not only within the actual discrete range, but also close to the distribution trendline of training *m/z*-CCS data points. When mining the potential analogues of training compounds in complex matrix, they should be not only retained but considered with higher priorities than features assigned as "In 0.99 PI but Out of 0.99 CI".

Results of experimental IM-MS feature filtering/grading will be generated as an extra column with the caption "**Predictive statistics_99%**" added to the original Excel file of "**Experimental IM-MS data**" (Fig. 9).



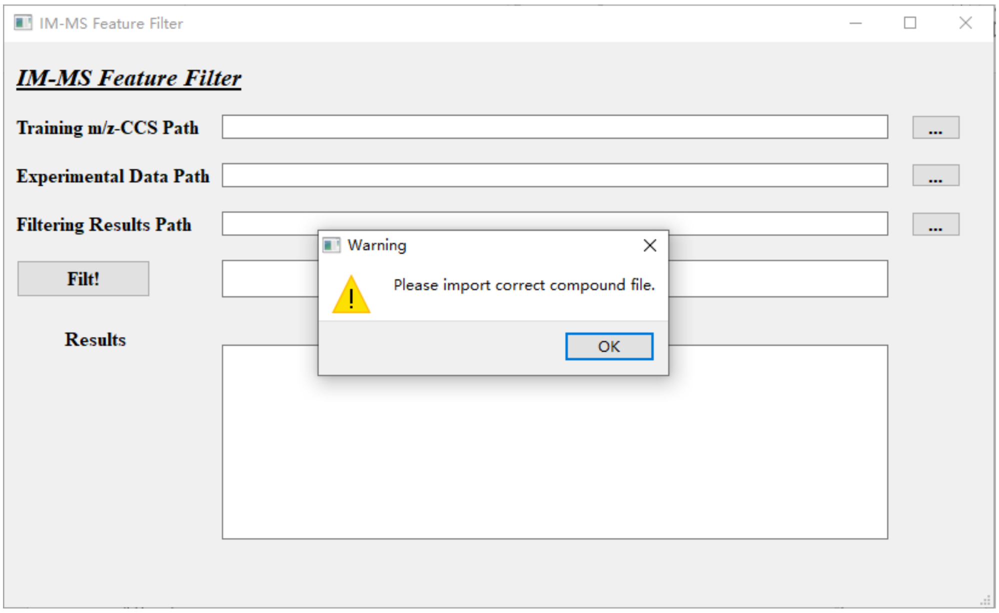
	A	B	C	D	E	F
1	Feature	RT	DT	*m/z*	CCS	Predictive statistics_99%
2	1	51.738	53.20	739.1855	247.8	In 0.99 PI but Out of 0.99 CI
3	2	47.299	52.72	755.1805	245.5	In 0.99 PI but Out of 0.99 CI
4	3	26.898	49.88	609.1442	233.4	In 0.99 CI
5	4	20.151	54.75	755.2018	254.9	In 0.99 PI but Out of 0.99 CI
6	5	34.495	49.34	593.1497	231.0	In 0.99 CI
7	6	25.219	55.81	739.2072	259.9	In 0.99 PI but Out of 0.99 CI
8	7	34.119	49.07	609.1439	229.6	In 0.99 PI but Out of 0.99 CI
9	8	42.059	48.72	593.1494	228.2	In 0.99 PI but Out of 0.99 CI
10	9	36.881	50.70	623.1601	237.1	In 0.99 CI

Fig. 9. Demonstration of Excel file for experimental IM-MS feature filtering/grading results of complex matrix sample data.

Meanwhile, the counts of experimental IM-MS features with filtering/grading results assigned as "**Out of 0.99 PI**", "**In 0.99 PI but Out of 0.99 CI**" and "**In 0.99 CI**", as well as total count of experimental IM-MS features, are also output on the interface of "**IM-MS Feature Filter**" (**Fig. 8**).

3.4.2. Common Errors on "IM-MS Feature Filter"

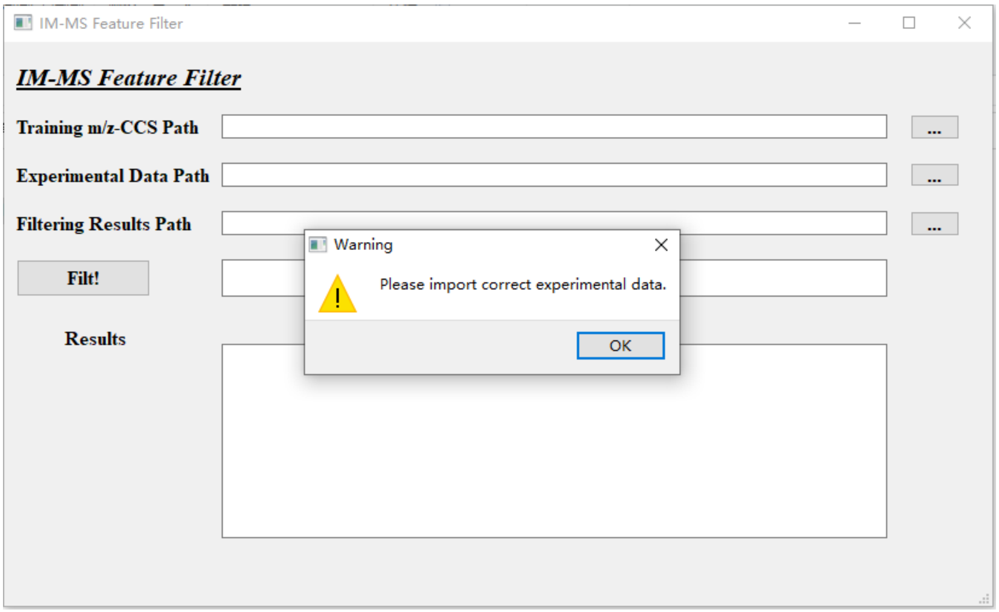
If the Excel file with the format requirements of "**Training *m/z*-CCS data**" in *section 3.1.1* is not selected at "**Training m/z-CCS Path**", the "IM-MS Feature Filter" functional interface will prompt warnings as "Please import correct compound file" (**Fig. 10**).



Please import correct compound file.

Fig. 10. Window warning message for not selecting the storage path of the Excel file met the format requirements of "Training *m/z*-CCS data" (refer to *section 3.1.1*) at "Training m/z-CCS Path" of "IM-MS Feature Filter".

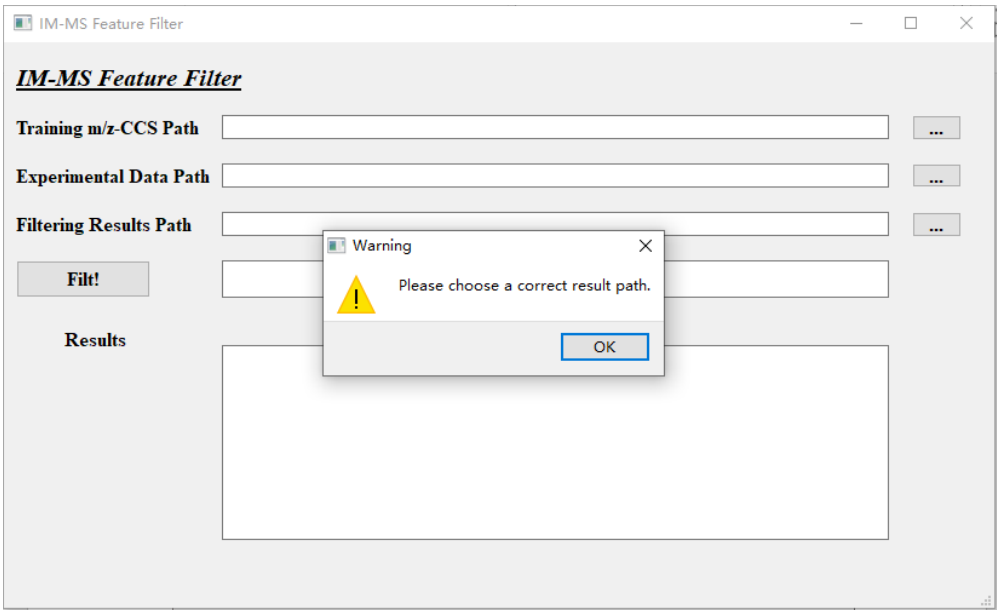
If the Excel file with the format requirements of "**Experimental IM-MS data**" in *section 3.1.2* is not selected at "**Experimental Data Path**", the "IM-MS Feature Filter" functional interface will prompt warnings as “Please import correct experimental data” (**Fig. 11**).



Please import correct experimental data.

Fig. 11. Window warning message for not selecting the storage path of the Excel file met the format requirements of “Experimental IM-MS data” (refer to *section 3.1.2*) at “Experimental Data Path” of “IM-MS Feature Filter”.

If the storage path of IM-MS feature filtering results is not selected at “**Filtering Results Path**”, the “IM-MS Feature Filter” functional interface will prompt warnings as “Please choose a correct result path” (**Fig. 12**).



Please choose a correct result path.

Fig. 12. Window warning message for not selecting the storage path of IM-MS feature filtering results at “Filtering Results Path” of “IM-MS Feature Filter”.