

AllCCS Tutorial V1.00

Zhiwei Zhou

2019-11-01

Contents

About AllCCS	9
Citation	11
News	13
1 Quick Start Guide	15
1.1 Create a AllCCS Account	15
1.2 Browser CCS database	15
1.3 Perform CCS prediction/Annotation	16
2 CCS Database	21
2.1 Compound Browser	21
2.2 Compound Card	24
2.3 Advanced Search	28
3 CCS Prediction	31
3.1 Data preparation	31
3.2 Result	33
4 Metabolite Annotation	37
4.1 Feature to candidates (Feature match)	37
4.2 Complement to MS/MS result (Candidate rank)	38

List of Tables

2.1	Basic statistics of Unified CCS database	22
2.2	Definition of confidence level	24

List of Figures

1.1	Register AllCCS account	16
1.2	Browser CCS database	17
1.3	Details of compounds	18
1.4	CCS prediction function	19
1.5	Metabolite match and candidate rank	19
2.1	Browser conditions	23
2.2	Download interested compounds	25
2.3	Download interested compounds	26
2.4	Unified CCS	27
2.5	Experimental CCS records	27
2.6	Predicted CCS records	28
2.7	Database link	28
2.8	Advanced search – single mode	29
2.9	Advanced search – batch mode	30
3.1	CCS prediction	32
3.2	CCS prediction results	33
3.3	Preview results	35
3.4	Download results	36
4.1	Feature annotation	38
4.2	Feature match results	39
4.3	Candidate rank function	41
4.4	CCS scoring	43
4.5	CCS filtering result of candidates	44
4.6	CCS scoring result of candidates	45

About AllCCS

Copyright (c) 2019 AllCCS Development Team

AllCCS is a powerful platform to support various applications in Ion Mobility – Mass Spectrometry (IM-MS). It is designed to contain three major parts: **1) Unified CCS database, 2) Machine learning based CCS prediction, and 3) small molecule annotation.** The unified CCS database is one of the most comprehensive CCS databases, covering ~1,700,000 small molecules. It provides a universal platform to contain both experimental CCS values (3,539) and predicted CCS values (over 10,000,000). Machine learning based CCS prediction function supports convenient prediction from SMILES structure to CCS values. This function utilizes the second generation CCS prediction algorithm to generate CCS values and RSS score for novel structures. Small molecule annotation provides an easy-to-use annotation function for various features or compounds. It is supported to search database with measured m/z and CCS for annotation, or in conjunct with any other annotation tools, such as MetFrag, CFM-ID, MS-Finder, and SIRUS etc.

Zhiwei Zhou (zhouzw@sioc.ac.cn) Zheng-Jiang Zhu (jiangzhu@sioc.ac.cn)
Laboratory for Mass Spectrometry and Metabolomics IRCBC, Shanghai
Institute of Organic Chemistry Chinese Academy of Sciences, Shanghai,
China

Citation

If AllCCS is useful in your project, please cite our articles.

- Z. Zhou, Z.-J. Zhu* etc. Advancing CCS database towards metabolite annotation, In preparing

News

November 25, 2019

- Update tutorial
- Fix some known problems

November 6, 2019

- Test in web server
- Fix bugs in admin system
- Fix some known bugs
- Update formal database and install AllCCS package V0.1.61
- Correct compound name of drugbank compound

September 22, 2019

- Demo webserver online

Chapter 1

Quick Start Guide

1.1 Create a AllCCS Account

Users need to register AllCCS account to use the webserver. This process is very sample, and **completely private**. To register an account, navigate to the home page. The “Sign up” button is in the upper right corner of the page. Then, you need to fill in your email and verification code which would be sent to your email (Figure 1.1). Finally, input some basic information to complete the registration. You could log in and enjoy all functions in the web server.

1.2 Browser CCS database

You could search you interested compound (name, formula, smiles, inchi, InChIKey etc.) in the search box in navigation bar, or directly browser the all database records in the “browser” page (Figure 1.2).

Then, you can click the link in the column of AllCCS ID to browser detail information of this compound (Figure 1.3). It includes basic meta information, unified CCS values, experimental CCS records, predicted CCS records and other database links etc. Please see section 2 for more details.



Figure 1.1: Register AllCCS account

1.3 Perform CCS prediction/Annotation

AllCCS also provides CCS prediction function (Section ??) and metabolite annotation functions (Section ??). You could click the link or corresponding item in navigation bar. For CCS prediction function, please input the SMILES list of your compounds in the input panel. The result would be returned on the project panel within several seconds (Figure 1.4).

For annotation function, you could search experimental feature to search the database with your settings, or filter/rerank candidates to conjunct with MS/MS annotation tools (Figure 1.5).

Figure 1.2: Browser CCS database

Figure 1.3: Details of compounds

Figure 1.4: CCS prediction function

Figure 1.5: Metabolite match and candidate rank

Chapter 2

CCS Database

Unified CCS database aims to be a **unified platform** to host both literature-reported CCS values and in-silico predicted CCS values for ion mobility - mass spectrometry (IM-MS). It is **open-access** and **downloadable**. It contains 3,539 unified CCS values which are summarized from 5,119 experimental CCS records. These experimental CCS values are acquired with variable platform including DTIMS, TWIMS and TIMS etc., and have definitive confidence level. In addition, **~10,000,000 predicted CCS** values are provided for **~1,700,000** small molecules from multiple public database to support widespread applications, including metabolomics, lipidomics, drug screening, pesticide screening etc (Table 2.1). For each compound, its compound card contains meta information, complete records and links to other database. Finally, users can search interested compounds' CCS values with the function of "Browser" and/or "Advance search" in this part.

2.1 Compound Browser

Compound Browser Function provides a simple and straightforward way to browser the database. There are several browser conditions set in "Browser" part (Figure 2.1).

- **Type:** It provides the choice of CCS values generated from experiments or prediction.
- **Database:** this option includes variable databases that cover all com-

Table 2.1: Basic statistics of Unified CCS database

No.	Database	Compounds	Coverage
1	[KEGG](https://www.genome.jp/kegg/)	16085	Metabolites & lipids
2	[HMDB](http://www.hmdb.ca/)	113989	Metabolites & lipids
3	[LMSD](https://www.lipidmaps.org/)	40532	Metabolites & lipids
4	[MINE](https://minedatabase.mcs.anl.gov/)	592175	Metabolites & lipids
5	[DrugBank](https://www.drugbank.ca/)	9546	Drugs & xenobiotics
6	[DSSTox](https://comptox.epa.gov/dashboard)	856919	Drugs & xenobiotics
7	[UNPD](NA)	213188	Natural products
8	[ZhuLab](NA)	1417	Metabolites & lipids

pounds in our unified database (Table 2.1). And users can choose specific database(s) for further execution.

- **Level:** it includes confidence level (See Section 2.2) of compounds in the unified database. It helps to choose compounds in the clearly defined level.

With browser conditions, users can screen out a series of interested compounds. Compounds entries would be displayed according to defined condition. In below text, it contains brief information for each compound.

- **AllCCS ID:** As described in the section 1.2, users can click the link in the column of AllCCS ID to browse the corresponding compound card (Figure 1.2).
- **Name:** compound name
- **Structure:** the image of compound structure
- **Formula:** chemical formula
- **Experimental CCS:** The unified CCS value reported in literature (See Section 2.2.3)
- **Predicted CCS:** The predicted CCS values using machine-learning algorithm. (See Section 2.2.4)
- **Highest level:** The highest confidence level of CCS values (See Section 2.2.2)

Users can check the interested compounds in the last column. Click the download option, you could download a CSV table containing the information

Figure 2.1: Browser conditions

Table 2.2: Definition of confidence level

Confidence level	Platform	Reported labs (N)	Maximum relative error (%)
Level 1	DTIMS	N 2	1%
Level 2	DTIMS/TWIMS/TIMS	N 2	3%
Level 3	DTIMS/TWIMS/TIMS	N=1	—
Level 4	Predicted CCS	—	—
Conflict	DTIMS/TWIMS/TIMS	N 2	>3%

of you checked compounds (Figure 2.2).

Note:

- Download function supports up to **100** items for one time.

2.2 Compound Card

In the compound card, it contains detail information of the compound. Next, we would like to explain each parts in the compound card.

2.2.1 Compound information

It contains the basic information of the compound, including ALLCCS ID, name, formula, exact mass, SMILES, InChI, InChIKey, classification and structure in the right panel (Figure 2.3). Here, [ClassyFire](#) is used for compounds' classification [ref8](#).

2.2.2 Unified CCS

This part contains the CCS information of different adduct forms with experimental CCS (if exists) and predicted CCS (Figure 2.4). The CCS reported here is the unified CCS values. We defined unified CCS as the average CCS value with definitive confidence level. The definition of confidence level can find in Table 2.2. CCS values of confidence 1, 2, 3 are all experimental values. The definition of confidence level:

Figure 2.2: Download interested compounds

Figure 2.3: Download interested compounds

- **Confidence level 1** represents the CCS value of the specie which is acquired with DTIMS and has been reported at least twice in different labs with the maximum relative error less than 1%.
- **Confidence level 2** represents the CCS value of specie which is acquired with DTIMS, TWIMS or TIMS and has been reported at least twice in different labs with the maximum relative error less than 3%.
- **Confidence level 3** represents the CCS value of specie is acquired with DTIMS, TWIMS or TIMS and only reported by one lab.
- **Confidence level 4** represents the predicted CCS value.
- **Conflict** means the CCS value of specie which is acquired with DTIMS, TWIMS or TIMS and has been reported at least twice in different labs, but the maximum relative error is more than 3%.

2.2.3 Experimental CCS records

This part records the detailed information of experimental CCS values (Figure 2.5). Compounds that have experimental CCS records contains the basic information of adduct form, m/z , experimental CCS values and charge. Besides, we also provide the information of used instrument platform and the type of ion mobility mass spectrometry. Detail information can be found in Table 2.3. The measured approach is also provided, including single-field, multiple-fields, and empirical method. Corresponding reference literature is



Figure 2.4: Unified CCS



Figure 2.5: Experimental CCS records

listed in the DOI column. If compounds don't have experimental CCS record, it would have no information in this part.

2.2.4 Predicted CCS records

This part records the detailed information of predicted CCS values (Figure 2.6). It provides the basic information of adduct forms, m/z, charge, corresponding predicted CCS values using AllCCS_V1 tool, users can reference our in pressing paper (XXXX) for detailed information. Here we define representative structure similarity (RSS) to represent the similarity between this compound and the training set.

Figure 2.6: Predicted CCS records

Figure 2.7: Database link

2.2.5 Database link

This part provides the link to databases that contain the compound (Figure 2.7).

2.3 Advanced Search

For advanced search, there are two modes for users to search the compounds in our unified CCS database, including “**single mode**” and “**batch mode**”.


Figure 2.8: Advanced search – single mode. This figure is not visible in the provided image, but its caption indicates it shows the single mode search interface.

Figure 2.8: Advanced search – single mode

2.3.1 Single mode

As named, users can search for one compounds at one time in single mode. Here, we provide several optional identifiers for users to choose (Figure 2.8), including compound's name, database ID, formula, SMILES, InChI, InChIKey.

Note:

- If you don't have confirmed identifier, keep it as null.
- If there are contradictory identifier, it will return no available data.

2.3.2 Batch mode

If there are multiple compounds, users can use batch mode to search (Figure 2.9). Currently, this function supports identifiers including "Database ID", SMILES, InChIKey. While choosing database ID as identifier, there are several choice of databases in the right panel close to identifier option.

Note:

- In searching panel, you can enter one item per line.
- It should not contain extra space.
- It supports up to 100 query items per request.

Figure 2.9: Advanced search – batch mode

Chapter 3

CCS Prediction

This part provides a machine-learning based CCS prediction function with the input of SMILES structures. The prediction error is estimated as low as ~2% (Median relative error), and users could predict CCS values for novel structures. Detail of prediction is provided in our AllCCS article [ref10](#).

3.1 Data preparation

For CCS prediction, users should provide the SMILES and the unique identifier for each SMILES. And there are two approaches for users to search in the interface (Figure [3.1](#)).

3.1.1 Direct input in the panel

Users can directly search in the panel with one entry per line containing one identifier and one SMILES. When the input is complete, click submit to get the predicted results.

Note:

- The line must be tab-separated.
- The identifiers must be unique in one submission.
- Due to the limit of computational resource, the maximum item is limited as 50 for one submission.

Figure 3.1: CCS prediction



Figure 3.2: CCS prediction results

3.1.2 Prediction with uploading CSV file

It is also available for prediction by uploading a CSV file. Users can download the CSV demo file, and the data format of the CSV file is showed in the Figure 3.1. The first column contains the identifier of each SMILES and the second column is the corresponding SMILES. The format of CSV file is same as the inputting panel. Please note that the identifiers must be unique in one file, and the maximum item is limited as **50** for one file.

3.2 Result

Results of user submissions are showed in the “My projects” panel (Figure 3.2).

With preview conditions, users can get the detailed result of inputted SMILES (Figure 3.3). Compounds entries would be sorted by different adduct information. In below texts, it contains brief information for each compound.

- **Name:** Consistent with the identifiers you input.

- **SMILES**: SMILES structures.
- **Monoisotopic mass**: Monoisotopic mass of structure
- **Adduct**: The adduct form. AllCCS provides 7 adducts forms (Positive mode: [M+H]⁺, [M+Na]⁺, [M-H₂O+H]⁺, [M+NH₄]⁺; Negative mode: [M-H]⁻, [M+Na-2H]⁻, [M+HCOO]⁻).
- **m/z**: The ratio of mass and charge
- **Predicted CCS**: CCS value for the specific structure and adduct.
- **RSS**: Representative structure similarity. See Section 2.2.4 for more information.
- **Status**:
 - **Valid**: Successful prediction
 - **Error1**: Invalid SMILES structure
 - **Error2**: The mass range is out of the limitation (AllCCS only supports small compound CCS prediction with mass between 60-1200).

Users can also click download to obtain CSV table which contains the same information as preview results (Figure 3.4).

Note:

- Due to the computation resource limitation, it allows up to 10 projects one time in “*My projects panel*”. If users want to execute more projects, please delete previous projects.

Figure 3.3: Preview results

Figure 3.4: Download results

Chapter 4

Metabolite Annotation

The part provides an easy-to-use annotation function for unknown features or compounds. It consists of two functions in this part. One is feature match, which supports searching database with measured m/z and CCS for annotation. The second function is candidate rank, which can be in conjunct with any other annotation tools, such as MetFrag, CFM-ID, MS-Finder, and SIRIUS etc. With the results from previously mentioned tools, and inputting experimental m/z and experimental CCS, users can filter/re-rank candidates for more accurate annotation.

4.1 Feature to candidates (Feature match)

Feature match function provides a simple and straightforward way to annotate features with experimental m/z and CCS in the AllCCS database. There are several annotation conditions set in this part (Figure 4.1).

- **m/z**: Experimental m/z of your interested peaks.
- **CCS**: Experimental CCS of your interested peaks.
- **m/z tolerance (\pm)**: The tolerance between experimental m/z and theoretical m/z in database. Users can choose ppm or Dalton as unit.
- **CCS tolerance (\pm)**: The tolerance between experimental CCS and theoretical CCS in database. Users can choose percentage or Å² as unit.
- **Adduct**: Match with the checked adduct forms that exist in the

Figure 4.1: Feature annotation

database.

- **Type:** Match with the checked CCS type in the database, including Experimental CCS, Predicted CCS (See Section 2.1).
- **Database:** Match with the checked database. It includes 8 compound structure sources (See Section 2.1).

The match results are as follows (Figure 4.2). The results contain brief information for each compound as section 2.1 Compound Browser mentioned.

Users can check the interested compounds in the last column. Click the download option, you could download a CSV table containing the information of you checked compounds (Figure 4.2).

Note:

- Download function supports up to 100 items for one time.

4.2 Complement to MS/MS result (Candidate rank)

The candidate rank function supports filtering/re-rank candidates with inputted candidate list. This function could be in conjunct with any other in-silico MS/MS tools like MetFrag, CFM-ID, MS-Finder, and SIRIUS etc.

Figure 4.2: Feature match results

Specifically, it has two rank types: **CCS filtering** and **CCS scoring**. CCS filtering excludes candidates which CCS errors (comparing with predicted CCS values) beyond the pre-defined cutoff value. CCS scoring generates an integrated score for each candidate with custom weights of CCS score (The detail is some with our previous publication - LipidIMMS Analyzer9) and MS/MS score. With CCS filtering or scoring, it can provide more credible results.

4.2.1 Data preparation

The MS/MS results (CSV file) generated from other tools should be modified as specific format. A demo data is showed in Figure 4.3. It should include 7 columns with specific names (“rank”, “name”, “smiles”, “inchikey”, “adduct”, “score”). The definitions of each column are given as follows:

- **rank**: The first column in the CSV table. It is the score rank from other tools (e.g. MetFrag, CFM-ID, MS-FINDER, SIRIUS etc).
- **name**: The second column in the CSV table. Compound names of candidates.
- **smiles**: The third column in the CSV table. SMILES structures of candidates.
- **inchikey**: The fourth column in the CSV table. InChIKey identifier of candidates.
- **adduct**: The fifth column in the CSV table. Adduct form of candidates. Please note that it only supports 7 common adducts (Positive mode, [M+H]⁺, [M+Na]⁺, [M+NH₄]⁺, [M-H₂O+H]⁺; Negative mode, [M-H]⁻, [M+Na-2H]⁻, [M+HCOO]⁻).
- **score**: The sixth column in the CSV table. The MS/MS match score generated in in-silico MS/MS tools.

Note:

- The inputted CSV file should have the same column name with demo data.
- The column order should be keep same with demo data.

Figure 4.3: Candidate rank function

4.2.2 Parameter setting

In the candidates rank function, users can get more reliable results by adjusting parameters according to their experiments. The candidate rank function contains parameters as follows:

- **Measured m/z**: Experimental m/z of corresponding feature.
- **Measured CCS**: Experimental CCS of corresponding feature.
- **Candidate list**: A CSV file of candidate (See Section 4.2.1).
- **Rank type**: It consists of CCS filtering and CCS scoring. When choosing CCS filtering, the next option is CCS tolerance (Figure 4.3). It will filter the results that out of the pre-defined CCS tolerance. If you choose CCS scoring, the follow options are showed as Figure Figure 4.4. It consists of Minimum tolerance, Maximum tolerance, CCS weight, and MS/MS weight.
- **CCS tolerance (\pm)**: This parameter is available in CCS filtering. The tolerance between experimental CCS and theoretical CCS in database. Users can choose percentage or Å2 as unit.
- **Minimum tolerance (%)**: This parameter is available in CCS scoring. If error is within the tolerance, CCS match score equals to 1. Range: 0-10.
- **Maximum tolerance (%)**: This parameter is available in CCS scoring. If error is larger than the tolerance, CCS match score equals to 0, and lipid candidates will be removed. Range: 0-20.
- **CCS weight**: This parameter is available in CCS scoring. The CCS match score weight to calculate integrated score.
- **MS/MS weight**: This parameter is available in CCS scoring. The MS/MS match score weight to calculate integrated score.

4.2.3 Results

Results of user submission are showed in the “My projects” panel (Figure 4.2).

4.2.3.1 CCS filtering

Users can get the detailed information of candidates by clicking the browser button (Figure 4.5). In below text, it contains brief information for each candidate.

Figure 4.4: CCS scoring

- **Rank:** The new rank after filtering with CCS.
- **Name:** Consistent with the name in candidate list (see Section 4.2.1).
- **SMILES:** Consistent with the smiles in candidate list (see Section 4.2.1).
- **InChIKey:** Consistent with the inchikey in candidate list (see Section 4.2.1).
- **Adduct:** Consistent with the adduct in candidate list (see Section 4.2.1).
- **MS/MS score:** Consistent with the score in candidate list (see Section 4.2.1).
- **MS/MS rank:** Consistent with the rank in candidate list (see Section 4.2.1).
- **Predicted CCS (\AA^2):** The predicted CCS with SMILES and InChIKey.

Users can also click download to obtain CSV table which contains the same information as preview results (Figure 4.5).

4.2.3.2 CCS scoring

The results generated with CCS scoring function are similar with results from CCS filtering (Figure 4.6). The difference is that CCS scoring has two

Figure 4.5: CCS filtering result of candidates

Figure 4.6: CCS scoring result of candidates

additional columns as CCS score and integrated score.

- **CCS score:** CCS score generated by comparing experimental CCS to predicted CCS values. CCS match is scored using a trapezoidal function. (The detail of trapezoidal function LipidIMMS Analyzer)
- **Integrated score:** The integrated score is calculated using a linear weighting function according to the user-defined weight for each match score.