Reply to reviewers

Reviewer(s)’ Comments to Author:

**Reviewer: 1**

Recommendation: Publish either as is or subject to minor revisions as indicated.

Comments:

In this study, a framework called MCnebula was established to facilitate mass spectrometry data analysis process by focusing on critical chemical classes and visualization. MCnebula can be applied to explore classification and structural characteristic of unknown compounds that beyond the limit of spectral library, which was very interesting. The concept of MCnebula was similar with molecular network in some degree since its function of visualization. The research design was appropriated, and the methods were adequately described. In my view, I recommend publication in Analytical Chemistry, but there are two main concerns regarding the manuscript.

1. The first one is related to the dependence on the MCnebula tool. I think that this complex workflow will reduce the number of potential users. It’s better to simplify the workflow and include the analysis platform in R package. The workflow of MCnebula only included ‘Feature detection’, but the R package didn’t include it. There are many good processing tools in R such as XCMS.

**Response：**The R package of XCMS is commonly used for pre-processing LC-MS data. According to Reviewers’ suggestions, we have encapsulated the related functions and methods of XCMS into the R package of MCnebula2, which enables fast ‘Feature detection’ in R with preset steps and custom parameters. We introduced its functions and usage of this module on our website: <https://mcnebula.org/docs/prologue/quick_start/#option-2-with-xcms> In addition, we introduce this module in the methods section of the article.

1. The second main concern is related to the usage guidelines. Although reference documentation for the R package was available, there was lack of more detailed examples and instructions on how to use the workflow as a whole, including installation for MAC and Windows, specific methods for bridging the various modules of the workflow etc.

**Response：**We built a website with instructions on MCnebula workflow in detail. The website not only provides the documentations of installation for Linux, Windows and MacOS but also contains the demo analysis with codes (both metabolic and chemical workflow). Furthermore, the module of quick start is beginner-friendly and user without bioinformatics background can get start with MCnebula quickly. See the website below with more details: <https://mcnebula.org/>.

Other comments :

1. I agree with the authors that in recent years there have been huge developments in the field of *in silico* fragmentation. However, these *in silico* predictions have been seriously flawed in the past, for example, in the lipids field. Please, discuss.

**Response:** As we discussed in the section of introduction, the spectral library matching is still the main method for LC-MS/MS data analysis and compounds annotations with fragmentation spectra due to its high accuracy. But automated interpretation of tandem mass spectrometry (MS/MS) data is often limited to searching in spectral libraries, so that the user can identify only compounds for which a reference sample has been measured and its spectrum stored. Recently, *in silico* tools were developing quickly since its unparalleled superiority of searching a broader Molecular Structure Database, which greatly expanded the application of mass spectrometry in biological researches1–7. Among these *in silico* tools, SIRIUS 4 was a rapid and robust tool. It can identify the molecular formula of a query compound with very high accuracy (more than 70%) and no spectral databases are required for this step of the analysis8. Unfortunately, the application of SIRIUS to lipidomics has not been as successful as metabolomics, which may be related to the structure of lipid compounds, such as the difficulty to generate discriminative fragmentations (as lipids are based on fundamental building blocks, their fragmentation leads to reoccurring and fairly predictive patterns9). We covered this shortcoming in discussion section of article.

1. In the section of introduction, some misuse of tense should be corrected. For example, in the line 34, ‘…several strategies have been developed…’

**Response：**We revised the sentence in line 34 in the manuscript according to the suggestion. In addition, other misuses of tense were corrected in the revised manuscript.

1. Some sentences could be more precise and concise. For example, in the Line 48 Page 2, suggestion: ’ Currently, the cutting-edge technology, called SIRIUS 4, integrates many advanced artificial intelligence algorithms and has achieved an accuracy rate of 70% when retrieving structures from libraries.’

**Response：**Thank you for your suggestion. In the Line 48 Page 2, we revised the sentence according to your suggestion.

1. In Page 13, Line 25, the stability comparison between MCnebula and GNPS were 10.5%, 19.8% and 41.7%, 50.1%, respectively (Fig. 2a). Please confirm if the figure citation here is right.

**Response：**We did make a mistake in the sentence of Page 13, Line 25. ‘The stability comparison between MCnebula and GNPS were 10.5%, 19.8% and 41.7%, 50.1%, respectively (Fig. 2a).’ In this sentence, the ‘Fig. 2a’ was modified to ‘Fig. 3a’ instead.

1. In the section of discussion (Page 23 Line15), please explain in detail what could the Compound 1642 turn into? What kind of factor could be responsible for the chemical transformation?

**Response：**We uploaded the wrong version of the table (Tab. S4), resulting in the disappearing of Compound 1642 in the table. Now we uploaded the latest version of the table instead of the old one. For the transformation of the compound, we speculated that hydrolysis, oxidation, and differential isomerization may occur after salt or heat treatment during the processing.

**Reviewer: 2**

Recommendation: May eventually be publishable, but requires major revisions as indicated.

Comments:

In this manuscript, the authors presented a computer tool (also called MCnebula) for analyzing and visualizing untargeted LC-MS data. This tool allowed the characterization of sample unknown compounds by using their fragmentation properties to determine their classes, structures, and substructures. The performance of the MCnebula tool was demonstrated using two examples: a human serum metabolomics analysis and a herbal analysis. In my opinion, the MCnebula tool presented in this manuscript could be useful, particularly for metabolic pathway analysis by visualization. The manuscript described its operation and performance appropriately. The authors integrated a new abundance-based classes (ABC) selection algorithm with the MCnebula tool for chemical class selection, which was available for selecting representative classes to achieve the subsequent analysis. The manuscript contains many good ideas and has some novel developments in algorithms. Therefore, I recommend its publication in Analytical Chemistry.

Here are a few general comments regarding the manuscript:

1. For this work to be published, I think the tutorial associated with MCnebula has to work properly. I was able to download and install MCnebula, along with the associated SIRIUS 4 dependency. Unfortunately, the tutorial didn’t work for me. I tried to follow the instructions provided at <https://github.com/Cao-lab-zcmu/MCnebula2>.

**Response：**Sorry for the bad experience when you use the tutorial of MCnebula. Indeed, our introduction on the Github project homepage (<https://github.com/Cao-lab-zcmu/MCnebula2>) is brief so it may lead user being confused at some point. Thus, we built a website with instructions on MCnebula workflow in detail (<https://mcnebula.org/>). The tutorials were updated and provided a comprehensive instruction of MCnebula, which was an easy-to-use workflow.

1. Fig.3 showed the comparison between the two methods but I am not sure if ‘classified numbers’ could represent the accuracy of chemical classes. Do higher values mean higher accuracy? I recommended general terms like precision and recall instead of classified numbers.

**Response：** Thanks for the suggestions. ‘Classified numbers’ in Fig 3 was represent the number of all ‘features’ that are predicted to be in a chemical class, including true positive (TP) and false positive (FP) prediction. In cases where the intended meaning isn’t clear, we added a note of ‘TP + FP’ in bracket after the ‘Classified numbers’ in the caption of Fig.3. According to the suggestion of reviewers, we added two common terms——precision and recall in our evaluation (see Fig. 2a).

1. Please explain the function of Parent-nebula since I didn’t get it from the manuscript.

**Response：** We mentioned Child-Nebulae and Parent-Nebula in the manuscript, but we focused on the use of Child-Nebulae since the Parent-Nebula only showed the overall distribution map of all the compounds in the dataset. We added the functional explanation of Parent-Nebula in our revised manuscript (In subsection of ‘overview’ of section ‘results’). In fact, the generation and visualization of Child-Nebulae is the characteristic of the MCnebula. The application of Child-Nebulae for LC-MS/MS data analysis highlighted the advantages of the ABC selection algorithm for fast filtering and selection of chemical classes. For Parent-Nebula, it is a graphical visualization of a mixture of all the features, which contains too much information for the user to extract useful information intuitively. In some degree, Parent-Nebula can be regarded as a network graph obtained by the traditional Molecular Networking method, while Child-Nebula is an improvement of Parent-Nebula. Child-Nebula provides visualizing multiple views split features from Parent-Nebula in a chemical taxonomy perspective.

1. In visualization, many tools or websites (GNPS and MetaboAnalyst) contain the function of network and clustering analysis for the LC-MS dataset. The characteristics and advantages of MCnebula should be added in the manuscript compared with other tools.

**Response：**In supplementary Table 1, we compared MCnebula workflow with some popular public tools (SIRIUS, GNPS, MZmine, XCMS, MetaboAnalyst, MS-DIAL) to indicate the characteristics and advantages of MCnebula. Considering the visualization of clustering for LC-MS datasets, we showed three aspects of indicators (Spectral based; Classes based; In depth annotation) to compare the differences with these visualizations methods such as MCnebula and GNPS. Furthermore, visualization function of MCnebula incorporated statistical analysis, which could achieve the screening of chemical classes or similar molecular structure. For example, MCnebula contained a module called ‘Tracing top features’，which could focus on the interesting compounds according to the dataset. Compared to other tools, the extensive functions and characteristics of MCnebula were showed in the supplementary Table 1.

1. In fig.6, how to get heatmaps from the diagram of child-nebula? I didn’t get it well from the section of Method.

**Response：** The heat map we presented in the manuscript was related to the screening of chemical classes and ‘features’. Firstly, when we focused on one class, we were able to obtain the ‘features’ belonging to this class with Nebula-Index. Subsequently, we filtered the ‘features’ with Q-value (also called adjust P-value) and log2(FC), which were generated from the statistical analysis results of the related dataset. In order to introduce the generation of heatmaps in MCnebula from the diagram of child-nebula, we added the related contents in the section of method. Besides, the sentence “The ‘features’ are select by either in infection groups versus control groups or HM versus HS group: Q-value < 0.05, |log2(FC)| ≥ 0.3.” was added in the legend of Fig 6 (heat map).

1. In the section of method, how to draw nebula diagrams, select the child-nebula from parent-nebula and match features ID with nebula-index? These questions still do not make sense to me even after reading method. Please explain them in details.

**Response：**The questions above are related to ABC selection algorithm. We described the ABC selection algorithm in detail in the section of method. This algorithm filtered and selected a series of chemical classes, and divided into groups of ‘features’ according to their chemical classes to form Nebula-Index. Nebula-Index contained the information of chemical classes and the ID of ‘features’ belonging to these classes. When Nebula-Index merged with the annotation data (such as molecular formula, molecular structure) of all these ‘features’, the data for plotting Child-Nebulae was available. The subsequent visualization involved some programming in detail of data organization and plotting, which could be achieved by encapsulating existed R packages (e.g., ‘ggplot2’, ‘igraph’, ‘ggraph’).

1. Page 21, line 31-32, I didn’t find the features of ID 1642, 1785 and 2321 in the supplementary tables.

**Response：**We uploaded the wrong version of the table, resulting in the disappearing of Compound 1642, 1785 and 2321 in the table. Now we uploaded the latest version of the table instead of the old one. Please note that this table only showed compounds with Tanimoto Similarity > 0.5.

1. All codes used in the manuscript should be provided. Also, check the links for the MASSIVE dataset. Both links are not pointing to the correct dataset.

**Response：**In sub-section ‘Data and code availability’ of section ‘Experimental section’, we introduced all the codes for data analysis and how to obtain all the codes. Due to our mistake, the wrong MASSIVE link was given in the manuscript and it is now fixed in the link below:

(<https://massive.ucsd.edu/ProteoSAFe/QueryMSV?id=MSV000079949>).

1. Some concepts should be explained well when are mentioned in the manuscript for the first time, such as ‘feature level’, ‘chemical classified level’ and ‘Regex match’ etc.

**Response：**Thanks for kind suggestions. We checked all the concepts that were mentioned in the manuscript for the first time and the explanation of those concepts were added in the revised manuscript. **‘Feature level’** meant relative quantification of the peak area relating to some ‘features’. **‘Chemical classified level’** in the manuscript referred to the classifying of ‘features’ with the knowledge of systematic chemical classification. **‘Regex Match’**, also termed regular expression, introduced a pattern of string matches that can be used to check whether a string contained a certain substring, which was used to replace a matching substring, or to remove a substring from a string that matches a certain condition.

1. The manuscript would benefit greatly by a thorough revision specifically focused on the English language. At times, the high amount of English-associated errors makes it difficult to interpret the message of the paper.

**Response：We looked through the manuscript and revised the errors in English language.**

A few general English notes:

1. Check the repetition of punctuations and words.

Page 3, line 19, there were two ‘.’ in one sentence.

Page 6, line 30, the word ‘without’ was duplicated.

**Response:** The repetition of punctuations and words were revised (Page 3, line 19; Page 6, line 30).

2. Check both the phrase and the verb tenses.

Page 8, line 30, ‘allows the analysis process…’ should be replaced by ‘allowed the analysis process…’

Page 8, line 39, ‘In order that…’should be replaced by ’In case that…’

**Response:** The phrase and the verb tenses were revised (Page 8, line 30; Page 8, line 39).

3. Check several typos in the manuscript.

**Response:** Some typos in the manuscript were revise.

# Reference

1. Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. [Metabolite identification and molecular fingerprint prediction through machine learning](https://doi.org/10.1093/bioinformatics/bts437). *Bioinformatics (Oxford, England)* **28**, 2333–2341 (2012).

2. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. [Searching molecular structure databases with tandem mass spectra using CSI:FingerID](https://doi.org/10.1073/pnas.1509788112). *Proceedings of the National Academy of Sciences* **112**, 12580–12585 (2015).

3. Ludwig, M., Dührkop, K. & Böcker, S. [Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints](https://doi.org/10.1093/bioinformatics/bty245). *Bioinformatics (Oxford, England)* **34**, i333–i340 (2018).

4. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. [In silico fragmentation for computer assisted identification of metabolite mass spectra](https://doi.org/10.1186/1471-2105-11-148). *BMC Bioinformatics* **11**, 148 (2010).

5. Allen, F., Greiner, R. & Wishart, D. [Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification](https://doi.org/10.1007/s11306-014-0676-4). *Metabolomics* **11**, 98–110 (2015).

6. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. [MetFrag relaunched: Incorporating strategies beyond in silico fragmentation](https://doi.org/10.1186/s13321-016-0115-9). *Journal of Cheminformatics* **8**, 3 (2016).

7. Blaženović, I. *et al.* [Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: Database boosting is needed to achieve 93% accuracy](https://doi.org/10.1186/s13321-017-0219-x). *Journal of Cheminformatics* **9**, 32 (2017).

8. Dührkop, K. *et al.* [SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information](https://doi.org/10.1038/s41592-019-0344-8). *Nature Methods* **16**, 299–302 (2019).

9. Krettler, C. A. & Thallinger, G. G. [A map of mass spectrometry-based *in Silico* fragmentation prediction and compound identification in metabolomics](https://doi.org/10.1093/bib/bbab073). *Briefings in Bioinformatics* **22**, bbab073 (2021).