# Peer Review Information

**Journal:** Nature Methods
**Manuscript Title:** Fast and flexible analysis of linked microbiome data with mako
**Corresponding author name(s):** Karoline Faust

## Reviewer Comments & Decisions:

**Decision Letter, initial version:**

| | |
|---|---|
| Date: | 8th Apr 21 21:54:30 |
| From: | Lin.tang@nature.com |
| To: | karoline.faust@kuleuven.be |
| Subject: | Decision on Nature Methods submission NMETH-BC45289 |
| Message: | 8th Apr 2021 |

Dear Professor Faust,

Your Brief Communication entitled "Fast and flexible analysis of linked microbiome data with mako" has now been seen by 3 reviewers, whose comments are attached. While they find your work of potential interest, they have raised serious concerns which in our view are sufficiently important that they preclude publication of the work in Nature Methods, at least in its present form.

As you will see, the reviewers raise concerns about target users, use cases and many technical points. Should further work allow you to fully address these criticisms, including facilitating wide use by a large audience and adding stronger use cases, we would be willing to look at a revised manuscript (unless, of course, something similar has by then been accepted at Nature Methods or appeared elsewhere). This includes submission or publication of a portion of this work somewhere else. We hope you understand that until we have read the revised paper in its entirety we cannot promise that it will be sent back for peer-review.

If you are interested in revising this manuscript for submission to Nature Methods in the future, please contact me to discuss your appeal before making any revisions. Otherwise, we hope that you find the reviewers’ comments helpful when preparing your paper for submission elsewhere.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In this manuscript, Röttjers and Faust described a software mako for fast and flexible analysis of linked microbiome data. The manuscript is well-written and easy-to-follow. My main comments are listed below:

1) Motivation
Regarding the gap mako aims to fill, the authors stated that "a static network database risk becoming outdated .... ". This reviewer does not see how mako can address this, as even with the help of mako, one has to re-run the analysis periodically to keep the database updated.

2) Technology innovation
One of the key contributions of mako seem to stem from using Neo4J and Cypher. However, this technology has been around for over 10 years. All the described procedures are somehow "standard" in order to use it. This is certainly not straightforward, and requires some tweaking, but it is quite achievable giving the documentations and examples for Neo4J. Authors should be aware of the cost comes with the choice of this platform. This reviewer agrees with the capacity of handling the big network query (Neo4J is mainly designed for this). Long-term maintenance and updating, customization of the Neo4J-based system could be a big issue. On the other hand, SQLite + R + igraph are so popular and flexible for network manipulation and analysis, with many solutions for large-scale analysis.

3) Content
The inclusion of the precomputed networks from 60 datasets is very well appreciated. However, it seems the main credit should go to FlashWeave which allows such large-scale computing. Importing or exporting the data to/from Neo4J are relatively straightforward

4) Target audience
It is not clear who are the target users. From developers' perspective, mako will help a lot to accelerate the initial setup. However, they will soon need to learn Cypher in order to customize the queries. For general researchers (i.e. no coding experience), it is too complicated.

Overall, this reviewer feels it is insufficient as a new software for this Journal. A resource (i.e. an online platform maintained and updated by the authors' group) with strong use cases could be well appreciated by the community

Reviewer #2:
Remarks to the Author:
The authors present mako, a tool that supports network queries on very large microbiome data sets. This seems novel, accessible (e.g., a CLI, GUI, and API are available), well-tested, and well-documented. Overall I'm excited about this software, but I ran into some usability issues that prevented me from testing it.

Notes on software:

The software looks to be extensively unit-tested. That's excellent! The manual is also very nice.

I ran into issues installing mako on macOS following the instructions in the README. This seems to be related to installation of the biom-format dependency:

ERROR: Could not build wheels for biom-format which use PEP 517 and cannot be installed directly

This problem might be averted if the authors distribute using conda/bioconda, and that might faciliate adoption by users who are already comfortable with conda installations.

I was able to get the Neo4j docker container running, and was able to connect to the server. I ran into some issues working through the tutorial which I think is a result of the formatting of the manual. On page 10 of the manual (section 2.5), where the first command is entered, it wasn't clear how to enter the command. I first tried entering this line-by-line (i.e., enter the text associated with a single bullet, then hit Enter) and got an error. It worked when I joined the four lines into one command separated by spaces, which surprised me - I thought I'd have to separate them with semi-colons. The issue though is that this was confusing to follow, and will be an issue for new-comers. I recommend providing more detail on how this first command should be entered. For example, formatting as a code-block somehow could be very helpful for this.

In the next bit of the tutorial, the formatting of the commands again posed an issue. For example, 'Acorn bar- nacles' was pasted with the hyphen in it. I wasn't able to get through the rest of this demo as a result of subsequent issues with copy-paste.

I tried to move on to the next section, but wasn't able to continue because I didn't have mako installed locally due to the issue I ran into above.

I recommend working on these documentation issues, and then doing some basic user-testing. For example, identify a colleague who can test this out (or solicit someone to serve as a test user from the Internet) and see what they struggle with from the manual.

Notes on manuscript:

It would be helpful to discuss FlashWeave as this approach to developing the association networks impacts everything that follows. How are multiple comparisons handled with FlashWeave? Is compositional data handled appropriately?

Can association network data be loaded into mako using text files, or only pre-loaded in a network database? Loading from text files seems like it will be much easier for users who are not experienced with network databases.

I recommend that in addition to BIOM tables being acceptable as input, QIIME 2 feature table are allowed as well. This would be very straight-forward as QIIME 2 feature tables can be viewed as biom files with QIIME 2 (there are other options for accessing this data as well if this isn't possible). This would allow the large QIIME 2 user community to immediately start using mako with very little additional development effort on the mako end.

Line 58: It's not clear exactly how OWL is being used here. I think an example would help to illustrate this.

Line 73: 'act as "AND-gates" in the control of gene expression' I don't know the function of "AND-gates" will be generally accessible to the Nature Methods audience - I think a more generally accessible description would make sense here.

The focus of this work is on microbiome networks. Could mako be applied to multi-omics data, such as microbes and metabolites from the same samples? If so, I think this should be highlighted as this is increasingly an area of interest. If not, this is just a suggestion for future work (not a prerequiste for publication).

Line 158: It seems that "aquatic microbiome" would be a subset of "earth microbiome".

Line 179: Please clarify "all other abundances were binned".

Reviewer #3:
Remarks to the Author:
PAPER REVIEW:

Mako is a nice tool for users to load microbiome data into a neo4j database, and then perform network-based queries or analysis of the data. I think mako will be a very interesting tool for communities to explore Qiita data sets. However, I would argue that unless the users expect to run exactly the same queries and analysis as provided, certain programming knowledge is still required. For example, "run_motifs.py" codes propionate associations. If a user wishes to check a different function, he/she will need to know how to modify the code.

The authors curated 60 BIOM files as their test example. It would be interesting to see what the limitation of this system is. How many BIOM files can be reasonable used in an analysis without exceeding the system limitation or incurring a very long wait?

I would suggest that the 60-network database example in Section 2.1 to be presented as a way to exploring the data instead of a scientific investigation. The reason is that I think an example of 60 data sets (only a tiny subset from all the Qiita datasets) is too small to present a general picture. It pretty much depends on the data sets you have selected. For example, I randomly selected a few files from the provided 60 data sets, and my Animal graph was actually very similar to the Plant graph! Moreover, many of your Plant examples are about roots and rhizosphere, while many of the Non-saline examples are about soil. This makes me wonder whether that contributed to the similarity between your Plant and Non-saline graphs.

Also, I have difficulty viewing Figure 1(c) and Figure S1. I wonder whether the quality of these two figures can be improved.

CODE REVIEW using Code Ocean:

I did have fun trying different examples using Code Ocean.

Re. using own data: The authors explained in the Methods section how to prepare biom and network influence files. Since I didn’t have my own data, I simply took one of the example data sets, and made some modifications to the data. The tool worked fine with my modified data set.

By the way, there's a minor bug: When a Cypher query did not return any results, the code gave a Python TypeError (see below).

```
==========
Data Sets: 1721, 1792, 2104, 10097, 11947
Cypher query (one of the example queries from README.md): MATCH p=(:Order {name:
'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name: 'o__Clostridiales'}) RETURN p--(:Taxon)--
(:Edge) RETURN p

MATCH p=(:Order {name: 'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name:
'o__Clostridiales'}) RETURN p--(:Taxon)--(:Edge) RETURN p
Traceback (most recent call last):
File "../code/run_query.py", line 117, in <module>
main(sys.argv)
File "../code/run_query.py", line 46, in main
query_counts = process_query(query_results)
File "../code/run_query.py", line 83, in process_query
for value in results[0]['p']:
TypeError: 'NoneType' object is not subscriptable
```

Although we cannot offer to publish your paper in Nature Methods, the work may be appropriate for another journal in the Nature Research portfolio. If you wish to explore suitable journals and transfer your manuscript to a journal of your choice, please use our manuscript transfer portal. If you transfer to Nature-branded journals or to the Communications journals, you will not have to re-supply manuscript metadata and files. This link can only be used once and remains active until used.

All Nature Research journals are editorially independent, and the decision to consider your manuscript will be taken by their own editorial staff. For more information, please see our manuscript transfer FAQ page. Note that any decision to opt in to In Review at the original journal is not sent to the receiving journal on transfer. You can opt in to *In Review* at receiving journals that support this service by choosing to modify your manuscript on transfer. In Review is available for primary research manuscript types only.

## Author Rebuttal to Initial comments

# Rebuttal plan

## Responses to the reviewers

**Reviewer #1:**

**Remarks to the Author:**

**In this manuscript, Röttjers and Faust described a software mako for fast and flexible analysis of linked microbiome data. The manuscript is well-written and easy-to-follow. My main comments are listed below:**

**1) Motivation**

**Regarding the gap mako aims to fill, the authors stated that "a static network database risk becoming outdated .... ". This reviewer does not see how mako can address this, as even with the help of mako, one has to re-run the analysis periodically to keep the database updated.**

We agree with the reviewer that mako does not remove the need to re-run analyses. However, the flexibility of setting up the database means that this can be done quickly and locally, depending on the research question at hand. We will rephrase this in the manuscript.

**2) Technology innovation**

**One of the key contributions of mako seem to stem from using Neo4J and Cypher. However, this technology has been around for over 10 years. All the described procedures are somehow "standard" in order to use it. This is certainly not straightforward, and requires some tweaking, but it is quite achievable giving the documentations and examples for Neo4J. Authors should be aware of the cost comes with the choice of this platform. This reviewer agrees with the capacity of handling the big network query (Neo4J is mainly designed for this). Long-term maintenance and updating, customization of the Neo4J-based system could be a big issue. On the other hand, SQLite + R + igraph are so popular and flexible for network manipulation and analysis, with many solutions for large-scale analysis.**

While Neo4J and Cypher have been around for a long time, microbiome research almost never uses them, which implies the existence of technological hurdles. Mako's key contribution is the removal of these hurdles through i) a new OWL-based data schema for microbiome data and ii) optimized functions for rapid import and export of large biological data sets. In this way, mako enables researchers to populate and interact with a Neo4j database in a matter of hours rather than weeks and also to re-use standard, documented Cypher queries. We will improve the description of mako's key contribution in the manuscript.

In addition, the reviewer raises the point that SQLite and other relational databases are also suitable for large-scale data analysis. However, typical network queries are much harder to implement in SQL than in Cypher. We will use a PostgreSQL alternative to mako to illustrate this point with examples.

**3) Content**

**The inclusion of the precomputed networks from 60 datasets is very well appreciated. However, it seems the main credit should go to FlashWeave which allows such large-scale computing. Importing or exporting the data to/from Neo4J are relatively straightforward**

We respectfully disagree with the reviewer that importing and exporting data to/from Neo4j is straightforward. Without an accessible method to import and export data, researchers would store data according to different data schemas, which means that Cypher queries would not be interchangeable between them. By developing an accessible API to port several standard formats, we make it possible for researchers to share their Cypher queries.

Concerning our network collection, it could not have been made without weeks of checking EMPO terms and pre-processing the data. Beyond our own work of curation and network construction, credit should go to QIITA rather than to FlashWeave. There are other fast network construction tools available (SPIEC-EASI, fastLSA, bnlearn), but we gladly admit that our collection depends on QIITA's unique capabilities as a microbiome database and will point this out more clearly in the manuscript.

**4) Target audience**

**It is not clear who are the target users. From developers' perspective, mako will help a lot to accelerate the initial setup. However, they will soon need to learn Cypher in order to customize the queries. For general researchers (i.e. no coding experience), it is too complicated.**

We will better clarify that mako is intended for bioinformaticians and biologists with rudimentary computational skills. We will also take several steps to make mako more user-friendly, including:

- Developing a mako web page with an improved manual and tutorial
- Providing a collection of videos illustrating example queries and a collection of use cases (2 already in the manuscript and 3 new ones)
- Supporting conda installation for mako

**Overall, this reviewer feels it is insufficient as a new software for this Journal. A resource (i.e. an online platform maintained and updated by the authors' group) with strong use cases could be well appreciated by the community**

Mako is designed to easily and quickly build a local database from the user's microbiome and network data. One reason for favouring a local solution over a shared resource is the rapid development of network inference tools. There are now dozens of different tools, each of which with a set of configurations, and it is unlikely that a standard will emerge in the next years. In addition, microbiome data nowadays often combine taxa and functions derived from sequencing with other omics data such as metabolomics, not to mention sample metadata. In our opinion, an online platform will never be flexible enough to store all microbiome data of interest and build networks with the user's preferred tool(s) and tool settings, and so a local solution is preferable. We will better clarify these important points in the manuscript.

To make it easier for novices to query our custom network data, we will expand the Code Ocean representation of mako, so example and test queries can be run without the need to install mako.

Concerning strong use cases, we plan to add the following three:

- Exploration of the HMA-LMA (high and low microbial abundance) dichotomy in sponge microbial networks
- Enumeration of paths between nodes of interest in a curated gut microorganism-metabolite network to see for instance whether gut bacteria detected in a faecal sample can convert starch to butyrate
- Identification of genera with the largest number of associations to metabolites known to be affected by IBD using metabolite-microorganism networks constructed from a recent meta-omics IBD study

**Reviewer #2:**

**Remarks to the Author:**

**The authors present mako, a tool that supports network queries on very large microbiome data sets. This seems novel, accessible (e.g., a CLI, GUI, and API are available), well-tested, and well-documented. Overall I'm excited about this software, but I ran into some usability issues that prevented me from testing it.**
We are glad that the reviewer classifies this tool as novel, accessible and well-documented and we apologize for the remaining bugs. We will intensify beta testing. In addition, we will develop a website to present the manual in a more attractive manner and to better explain how to use Neo4j and mako.

**Notes on software:**

**The software looks to be extensively unit-tested. That's excellent! The manual is also very nice.**
**I ran into issues installing mako on macOS following the instructions in the README. This seems to be related to installation of the biom-format dependency:**

**ERROR: Could not build wheels for biom-format which use PEP 517 and cannot be installed directly**
**This problem might be averted if the authors distribute using conda/bioconda, and that might faciliate adoption by users who are already comfortable with conda installations.**
We thank the reviewer for their helpful comment. We will write conda recipes and will enable the distribution of mako via conda/bioconda.

**I was able to get the Neo4j docker container running, and was able to connect to the server. I ran into some issues working through the tutorial which I think is a result of the formatting of the manual. On page 10 of the manual (section 2.5), where the first command is entered, it wasn't clear how to enter the command. I first tried entering this line-by-line (i.e., enter the text associated with a single bullet, then hit Enter) and got an error. It worked when I joined the four lines into one command separated by spaces, which surprised me - I thought I'd have to separate them with semi-colons. The issue though is that this was confusing to follow, and will be an issue for new-comers. I recommend providing more detail on how this first command should be entered. For example, formatting as a code-block somehow could be very helpful for this.**
**In the next bit of the tutorial, the formatting of the commands again posed an issue. For example, 'Acorn bar- nacles' was pasted with the hyphen in it. I wasn't able to get through the rest of this demo as a result of subsequent issues with copy-paste.**
We will port the tutorial from a PDF format to a website format that supports code blocks.

**I tried to move on to the next section, but wasn't able to continue because I didn't have mako installed locally due to the issue I ran into above.**
We will add conda support to avoid this problem.

**I recommend working on these documentation issues, and then doing some basic user-testing. For example, identify a colleague who can test this out (or solicit someone to serve as a test user from the Internet) and see what they struggle with from the manual.**
We apologize for our insufficient beta testing and will intensify our testing efforts.

**Notes on manuscript:**

**It would be helpful to discuss FlashWeave as this approach to developing the association networks impacts everything that follows. How are multiple comparisons handled with FlashWeave? Is compositional data handled appropriately?**
FlashWeave uses the centred log-ratio transformation to handle compositional data and corrects for multiple testing. We will briefly mention this in the manuscript.

**Can association network data be loaded into mako using text files, or only pre-loaded in a network database? Loading from text files seems like it will be much easier for users who are not experienced with network databases.**
Yes, mako supports the use of edge lists, which are text files. We will include a section on the website that addresses all supported file formats, including example files.

**I recommend that in addition to BIOM tables being acceptable as input, QIIME 2 feature table are allowed as well. This would be very straight-forward as QIIME 2 feature tables can be viewed as biom files with QIIME 2 (there are other options for accessing this data as well if this isn't possible). This would allow the large QIIME 2 user community to immediately start using mako with very little**

**additional development effort on the mako end.**
We will add support for Qiime 2 feature tables (both count and taxonomy files).

**Line 58: It's not clear exactly how OWL is being used here. I think an example would help to illustrate this.**
We will add a more detailed figure on the website that explains the data schema and how it is used by mako to confirm database integrity. In a nutshell, mako takes the relationships defined in the OWL file and queries the database to find relationships that do not connect to nodes described in that file. If this is the case, mako raises an error that the relationship is used incorrectly. When we replace Figure S1 with an online version, we will include a detailed explanation of the OWL file and its use in mako.

**Line 73: 'act as "AND-gates" in the control of gene expression' I don't know the function of "AND-gates" will be generally accessible to the Nature Methods audience - I think a more generally accessible description would make sense here.**
We altered the text as follows:
"Network motifs may indicate the presence of specific dynamic behaviours. For instance, in microbial communities, a negative circuit implementing a rock-paper-scissors game can promote diversity (e.g., Kerr et al. 2002)."

**The focus of this work is on microbiome networks. Could mako be applied to multi-omics data, such as microbes and metabolites from the same samples? If so, I think this should be highlighted as this is increasingly an area of interest. If not, this is just a suggestion for future work (not a prerequiste for publication).**
Currently, the database has been designed to support a flexible node format – Property – which is able to represent any sort of sample or taxon metadata, including metabolites. This could accommodate natural language processing algorithms, meaning that it becomes possible to learn relationships between metabolites and microbes. We will amend the manuscript to highlight this as a future area of interest. In addition, we will add a use case involving a metabolite-microorganism network to demonstrate this capability.

**Line 158: It seems that "aquatic microbiome" would be a subset of "earth microbiome".**
We will replace "earth microbiome" with "soil microbiome, including terrestrial plants". We chose to combine these since many plant studies include soil samples.

**Line 179: Please clarify "all other abundances were binned".**
Taxa that fell below the prevalence threshold were merged into a single synthetic taxon so that the total sum of abundances was not disturbed. We will amend this in the manuscript.

**Reviewer #3:**

**Remarks to the Author:**

**PAPER REVIEW:**

**Mako is a nice tool for users to load microbiome data into a neo4j database, and then perform network-based queries or analysis of the data. I think mako will be a very interesting tool for communities to explore Qiita data sets. However, I would argue that unless the users expect to run exactly the same queries and analysis as provided, certain programming knowledge is still required. For example, "run_motifs.py" codes propionate associations. If a user wishes to check a different function, he/she will need to know how to modify the code.**

We thank the reviewer for the kind comments. We agree that users need to learn how to modify Cypher queries to run their own analysis. To make this easy, we will expand the tutorial and include videos to demonstrate how to use Cypher queries in the Neo4j Browser. We will also expand the manual with an introduction on how to access the mako API to run custom queries.

**The authors curated 60 BIOM files as their test example. It would be interesting to see what the limitation of this system is. How many BIOM files can be reasonable used in an analysis without exceeding the system limitation or incurring a very long wait?**

This is an interesting question, and we will investigate the relationship between BIOM number and runtime. For graph databases, the relationship between size and time is more complicated than for relational databases, because the graph traversals necessary to carry out the queries become larger. Consequently, the main limitation is not necessarily the size of the database, but the profile of the query plan in relation to the size of the database. Since mako implements several checks that support meta-analyses, the load of those “checking” queries is what becomes prohibitive rather than the size of the database itself.

**I would suggest that the 60-network database example in Section 2.1 to be presented as a way to exploring the data instead of a scientific investigation. The reason is that I think an example of 60 data sets (only a tiny subset from all the Qiita datasets) is too small to present a general picture. It pretty much depends on the data sets you have selected. For example, I randomly selected a few files from the provided 60 data sets, and my Animal graph was actually very similar to the Plant graph! Moreover, many of your Plant examples are about roots and rhizosphere, while many of the Non-saline examples are about soil. This makes me wonder whether that contributed to the similarity between your Plant and Non-saline graphs.**

We agree with the reviewer that we only use a small subset of the QIITA datasets and that the EMPO classification we relied on is not a perfect description of the studies and could lead to misleading conclusions. However, some of the patterns we found are in agreement with results from other studies. In particular, motif density is related to edge density in general, which has been found to differ across biomes (Faust *et al.* 2015; Ma *et al.* 2020). We will discuss these limitations and the agreement with previous findings in the manuscript.

**Also, I have difficulty viewing Figure 1(c) and Figure S1. I wonder whether the quality of these two figures can be improved.**

We will simplify Figure 1c and improve its resolution in the manuscript. In addition, we will remove Figure S1 from the manuscript and represent it on the planned mako website instead, using a graph visualization method that better accommodates the size of Figure S1.

**CODE REVIEW using Code Ocean:**

**I did have fun trying different examples using Code Ocean.**

**Re. using own data: The authors explained in the Methods section how to prepare biom and network influence files. Since I didn't have my own data, I simply took one of the example data sets, and made some modifications to the data. The tool worked fine with my modified data set.**

**By the way, there's a minor bug: When a Cypher query did not return any results, the code gave a Python TypeError (see below).**

**==========**

**Data Sets: 1721, 1792, 2104, 10097, 11947**

**Cypher query (one of the example queries from README.md): MATCH p=(:Order {name: 'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name: 'o__Clostridiales'}) RETURN p--(:Taxon)--(:Edge) RETURN p**

```
MATCH p=(:Order {name: 'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name:
'o__Clostridiales'}) RETURN p--(:Taxon)--(:Edge) RETURN p
Traceback (most recent call last):
File "../code/run_query.py", line 117, in <module>
main(sys.argv)
File "../code/run_query.py", line 46, in main
query_counts = process_query(query_results)
File "../code/run_query.py", line 83, in process_query
for value in results[0]['p']:
TypeError: 'NoneType' object is not subscriptable
```

This error is due to the two return statements in the query. We will amend the Code Ocean capsule to report a more informative error message stating that this type of error is caused by empty query results that can stem from malformed queries. When using mako with the Neo4J Browser, errors in queries are highlighted directly.

## References

Kerr B, Riley MA, Feldman M & Bohannan BJM. "Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors" *Nature* 2002;**418**:171-174.

Faust K, Lima-Mendez G, Lerat J-S *et al.* Cross-biome comparison of microbial association networks.

*Front Microbiol* 2015;**6**.

Ma B, Wang Y, Ye S *et al.* Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* 2020;**8**:1–12.

## Revision Plan

We plan to carry out the following revisions to the manuscript, the presented software and the documentation:

- Implement Qiime 2 file support
- Create conda distributions
- Create a documentation website containing five use cases
    - Use cases 1 and 2 (motif discovery and propionate production) are already discussed in the main
    - Use case 3 will explore differences between microbial networks for high- and low microbial abundance sponges
    - Use case 4 will showcase path finding in a previously published curated gut microorganism-metabolite network intersected with microorganisms from a faecal sample
    - Use case 5 will identify genera associated to metabolites that significantly differ in IBD and healthy samples using data from a recent meta-omics IBD study
- Include documentation on using the API to write a Python script with custom queries
- Include documentation on how the OWL file is used to check database structure
- Include additional example Cypher queries in the documentation
- Enhance tutorial with video and graph visualizations
- Expand Code Ocean capsule to better support custom queries
- Simplify complex figures or move to the website in a more appropriate format

We will also carry out the following additional analyses:

- Compare SQL queries to Cypher queries
- Plot runtime against number of BIOM files

### Decision Letter, first revision:

| | |
|---|---|
| Date: | 3rd Jun 21 05:08:50 |
| From: | Lin.tang@nature.com |
| To: | karoline.faust@kuleuven.be |
| Subject: | Decision on appeal of Nature Methods submission NMETH-BC45289A-Z |
| Message: | 3rd Jun 2021 |

Dear Professor Faust,

Thank you for your letter asking us to reconsider our decision on your Brief Communication, "Fast and flexible analysis of linked microbiome data with mako". After careful consideration we have decided that we are willing to consider an extensively revised version of your manuscript that fully addressed reviewers' concerns..

When revising your paper:

* include a point-by-point response to our referees and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

***[REDACTED]***

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within eight weeks. We are very aware of the difficulties caused by the COVID-19 pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please submit reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see:
http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID
Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit http://www.springernature.com/orcid.

We look forward to hearing from you soon.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

**Author Rebuttal, first revision:**

## Response to referees

**Reviewer #1:**

**Remarks to the Author:**

**In this manuscript, Röttjers and Faust described a software mako for fast and flexible analysis of linked microbiome data. The manuscript is well-written and easy-to-follow.**

We thank the reviewer for the helpful and kind comments and have made significant changes to the software's documentation to increase the value of our contribution, as detailed below.

**My main comments are listed below:**

**1) Motivation**

**Regarding the gap mako aims to fill, the authors stated that "a static network database risk becoming outdated .... ". This reviewer does not see how mako can address this, as even with the help of mako, one has to re-run the analysis periodically to keep the database updated.**

We agree with the reviewer that mako does not remove the need to re-run analyses. However, the flexibility of setting up the database means that this can be done quickly and locally. We do not think that a static database, even if updated frequently, would be suitable for these analyses since association inference generates different networks depending on the experimental design, the tool used to construct the network and its specific setting.

We have rephrased this as follows (line 19-22):
"However, a static network database does not accommodate the flexibility required for microbial associations, since inferred associations change depending on the experimental design used to generate the abundance data and the parameters of the tools used to infer the networks."

**2) Technology innovation**

**One of the key contributions of mako seem to stem from using Neo4J and Cypher. However, this technology has been around for over 10 years. All the described procedures are somehow "standard" in order to use it. This is certainly not straightforward, and requires some tweaking, but it is quite achievable giving the documentations and examples for Neo4J. Authors should be aware of the cost comes with the choice of this platform. This reviewer agrees with the capacity of handling the big network query (Neo4J is mainly designed for this). Long-term maintenance and updating, customization of the Neo4J-based system could be a big issue. On the other hand, SQLite + R + igraph are so popular and flexible for network manipulation and analysis, with many solutions for large-scale analysis.**

While Neo4J and Cypher have been around for a long time, microbiome research almost never uses them, which implies the existence of technological hurdles. Mako's key contribution is the removal of these hurdles through i) a new OWL-based data schema for microbiome data and ii) optimized functions for rapid import and export of large biological data sets. In this way, mako enables researchers to populate and interact with a Neo4j database in a matter of hours rather than weeks and also to re-use standard, documented Cypher queries.

We have improved the description of mako's key contribution in the manuscript (line 40-41):
"The mako software package fills this gap by providing tools to rapidly start working with Neo4j and Cypher."

In addition, the reviewer raises the point that SQLite and other relational databases are also suitable for large-scale data analysis. However, typical network queries are much harder to implement in SQL than in Cypher. We have used a PostgreSQL alternative to mako to illustrate this point with examples.

The two queries below find patterns of edges that have the same taxonomic labels,
i.e. g__Eschericha—g__Roseburia:

```
SELECT string_agg(source::varchar, ','), string_agg(target::varchar, ','),
p.Genus as source, q.Genus as target, sign(WEIGHT)
FROM network as e
JOIN taxonomy as p ON e.source = p.taxon
JOIN taxonomy as q on e.target = q.taxon
WHERE p.Genus IS NOT NULL AND q.Genus IS NOT NULL
GROUP BY p.Genus, q.Genus, SIGN(e.weight) HAVING COUNT(*) > 1 LIMIT 1;
```

The above SQL query joins the network table together with the taxonomy table twice, so that both source and target nodes have their taxonomy added. It then groups by the Genus column of both nodes and returns one result that has a count larger than 1.

```
MATCH p=(a:Edge)--()--(x:Genus)--()--(b:Edge)--()--(y:Genus)--()--(a:Edge)
WHERE sign(a.weight)=sign(b.weight)
RETURN p LIMIT 1
```

The above Cypher query returns one pattern containing two edges which are connected by the same genus labels.

The two queries below find edges that occur only in one network.

```
SELECT string_agg(networkID::varchar, ',') AS networks,
source, target, SIGN(weight), COUNT(*) FROM edges
GROUP BY source, target, SIGN(weight) HAVING COUNT(*) = 1
```

The above SQL query groups associations by source and target and counts how many edges exist; only ones that exist once are returned.

```
MATCH (n:Edge)-->(x:Network)
WITH n MATCH (n)-[r]->(y:Network)
WITH n, count(r) as num
WHERE num=1 RETURN n
```

The above Cypher query matches edges and connects these to other networks; only edges are returned that are connected to a single network.

The two queries below find 3-node cliques.

```
SELECT * FROM edges AS a
JOIN edges AS b ON b.source = a.target
JOIN edges AS c ON c.source = b.target
WHERE c.target = a.source;
```

The above SQL query finds three edges that start and end at the same node. However, these table joins are costly and therefore limited to small motifs only. Hence, this query cannot easily be extended to include taxonomic information, which could require joining each edge table to a taxonomy table first.

```
MATCH p=(a:Taxon)--(x:Edge)--(:Taxon)--(y:Edge)--(:Taxon)--(z:Edge)--(a)
WHERE x.weight < 0 AND y.weight < 0 AND z.weight < 0 RETURN p
```

The above Cypher query finds three edges that start and end at the same taxon node. Adding taxonomic information can be done by using a WHERE clause and specifying the taxonomy there and does not lead to significant computational hurdles.

In conclusion, we agree with the reviewer that it is possible to do with SQLite what can be done with Neo4j; however, we believe Cypher is especially valuable when it comes to writing more complex queries that would otherwise require multiple JOIN and GROUP BY statements in SQL.

**3) Content**

**The inclusion of the precomputed networks from 60 datasets is very well appreciated. However, it seems the main credit should go to FlashWeave which allows such large-scale computing. Importing or exporting the data to/from Neo4J are relatively straightforward**

We respectfully disagree with the reviewer that importing and exporting data to/from Neo4j is straightforward. Without an accessible method to import and export data, researchers would store data according to different data schemas, which means that Cypher queries would not be interchangeable between them. By developing an accessible CLI and API to port several standard formats, we make it possible for researchers to share their Cypher queries.

Concerning our network collection, it could not have been made without weeks of checking EMPO terms and pre-processing the data. Beyond our own work of curation and network construction, credit should go to QIITA rather than to FlashWeave. There are other fast network construction tools available (SPIEC-

EASI, fastLSA, bnlearn), but we gladly admit that our collection depends on QIITA's unique capabilities as a microbiome database.

We have therefore rephrased this as follows (line 35-38):
"Additionally, mako includes a curated database derived from 60 separate data sets downloaded from Qiita, a platform for hosting microbial studies that facilitates meta-analyses of this scale."

To illustrate that network inference can be carried out with another approach than FlashWeave's, we have included a case study with microbe-metabolite links that are inferred with Spearman correlation: https://ramellose.github.io/mako_docs/examples/ibd/intro/. Here, we use Neo4j to query microbial families with the most associations to metabolites of interest, a feat that is not possible with FlashWeave or with Qiita alone.

**4) Target audience**

**It is not clear who are the target users. From developers' perspective, mako will help a lot to accelerate the initial setup. However, they will soon need to learn Cypher in order to customize the queries. For general researchers (i.e. no coding experience), it is too complicated.**

To suggest that users need some expertise to work with Neo4j, we have rephrased the introduction as follows (line 33-35):
"This software package includes a range of methods to interact with Neo4j databases and the pattern-based query language Cypher, requiring only rudimentary computational skills."

Additionally, we have taken several steps to make mako more user-friendly, including:

- Development of a mako web page with an improved manual and tutorial (https://ramellose.github.io/mako_docs/)
- Provided a quickstart page composed of videos (https://ramellose.github.io/mako_docs/quickstart/guide/)
- Included a collection of use cases, 2 already in the manuscript, 3 new ones (https://ramellose.github.io/mako_docs/examples/intro/)
- Supported conda installation for mako (https://ramellose.github.io/mako_docs/manual/introduction/conda/)

**Overall, this reviewer feels it is insufficient as a new software for this Journal. A resource (i.e. an online platform maintained and updated by the authors' group) with strong use cases could be well appreciated by the community**

Mako is designed to easily and quickly build a local database from the user's microbiome and network data. One reason for favouring a local solution over a shared resource is the rapid development of network inference tools. There are now dozens of different tools, each of which with a set of configurations, and it is unlikely that a standard will emerge in the next years. In addition, microbiome data nowadays often combine taxa and functions derived from sequencing with other omics data such as metabolomics, not to mention sample metadata. In our opinion, an online platform will never be flexible enough to store all microbiome data of interest and build networks with the user's preferred tool(s) and tool settings, and so a local solution is preferable.

Secondly, association networks are not static and can change depending on the included samples and experimental designs, as is shown by the inclusion of metadata variables in the FlashWeave publication (Tackmann, Rodrigues and von Mering 2019). Researchers may need to re-run network inference depending on the process of interest. For instance, consider a soil data set spanning several topsoil types. In the original analysis, a single network is constructed for this data set, but if researchers subsequently wish to compare networks specific to topsoil types, the construction needs to be repeated on sample subsets.

We have briefly summarized these issues in the manuscript as follows (line 20-24):

"However, a static network database does not accommodate the flexibility required for microbial associations, since inferred associations change depending on the experimental design used to generate the abundance data and the parameters of the tools used to infer the networks. Additionally, microbial network inference is a rapidly developing field and no standard protocols exist for the inference of these networks."

To make it easier for novices to query our custom network data, we have expanded the Code Ocean capsule. The capsule now reports Cypher errors and returns results as a json file that can be investigated in more detail. Consequently, example queries can be run from the capsule without the need to install mako.

Concerning strong use cases, we have added the following three:

- Exploration of the HMA-LMA (high and low microbial abundance) dichotomy in sponge microbial networks (https://ramellose.github.io/mako_docs/examples/sponges/intro/)
- Enumeration of paths between nodes of interest in a curated gut microorganism-metabolite network to see for instance whether gut bacteria detected in a faecal sample can convert starch to butyrate (https://ramellose.github.io/mako_docs/examples/metalit/intro/)
- Identification of genera with the largest number of associations to metabolites known to be affected by IBD using metabolite-microorganism networks constructed from a recent meta-omics IBD study (https://ramellose.github.io/mako_docs/examples/ibd/intro/)

**Reviewer #2:**

**Remarks to the Author:**

**The authors present mako, a tool that supports network queries on very large microbiome data sets. This seems novel, accessible (e.g., a CLI, GUI, and API are available), well-tested, and well-**

**documented. Overall I'm excited about this software, but I ran into some usability issues that prevented me from testing it.**

We are glad that the reviewer classifies this tool as novel, accessible and well-documented and we apologize for the remaining bugs. We have intensified beta testing to resolve this, so that the tool and accompanying documentation has now been tested by five individuals. In addition, we have developed a website to present the manual in a more attractive manner and to better explain how to use Neo4j and mako: https://ramellose.github.io/mako_docs/

**Notes on software:**

**The software looks to be extensively unit-tested. That's excellent! The manual is also very nice.**

**I ran into issues installing mako on macOS following the instructions in the README. This seems to be related to installation of the biom-format dependency:**

**ERROR: Could not build wheels for biom-format which use PEP 517 and cannot be installed directly**

**This problem might be averted if the authors distribute using conda/bioconda, and that might faciliate adoption by users who are already comfortable with conda installations.**

We thank the reviewer for the helpful comment. We have written conda recipes so mako can be distributed via conda: https://ramellose.github.io/mako_docs/manual/introduction/conda/

**I was able to get the Neo4j docker container running, and was able to connect to the server. I ran into some issues working through the tutorial which I think is a result of the formatting of the manual. On page 10 of the manual (section 2.5), where the first command is entered, it wasn't clear how to enter the command. I first tried entering this line-by-line (i.e., enter the text associated with a single bullet, then hit Enter) and got an error. It worked when I joined the four lines into one command separated by spaces, which surprised me - I thought I'd have to separate them with semi-colons. The issue though is that this was confusing to follow, and will be an issue for new-comers. I recommend providing more detail on how this first command should be entered. For example, formatting as a code-block somehow could be very helpful for this.**

**In the next bit of the tutorial, the formatting of the commands again posed an issue. For example, 'Acorn bar- nacles' was pasted with the hyphen in it. I wasn't able to get through the rest of this demo as a result of subsequent issues with copy-paste.**

We have ported the tutorial from a PDF format to a website format that supports code blocks, see for example https://ramellose.github.io/mako_docs/demo/pisaster/create/.

**I tried to move on to the next section, but wasn't able to continue because I didn't have mako installed locally due to the issue I ran into above.**

As suggested above, we have added conda support to avoid this problem.

**I recommend working on these documentation issues, and then doing some basic user-testing. For example, identify a colleague who can test this out (or solicit someone to serve as a test user from the**

**Internet) and see what they struggle with from the manual.**

We apologize for our insufficient beta testing and have intensified our testing efforts by including a total of five beta testers covering MacOS Big Sur, Ubuntu 18.04, Ubuntu 21.04 and Windows 10.

**Notes on manuscript:**

**It would be helpful to discuss FlashWeave as this approach to developing the association networks impacts everything that follows. How are multiple comparisons handled with FlashWeave? Is compositional data handled appropriately?**

FlashWeave uses the centred log-ratio transformation to handle compositional data and corrects for multiple testing.

We have clarified this in the Methods section of the manuscript (line 191-193):

"Since Flashweave incorporates a clr transformation, it handles compositional data appropriately without additional pre-processing steps."

**Can association network data be loaded into mako using text files, or only pre-loaded in a network database? Loading from text files seems like it will be much easier for users who are not experienced with network databases.**

Yes, mako supports the use of edge lists, which are text files. We have included additional documentation describing the CLI: https://ramellose.github.io/mako_docs/manual/cli/io/.

Additionally, we have included a case study that includes text files (https://ramellose.github.io/mako_docs/examples/ibd/setup/) and one that includes XML files (https://ramellose.github.io/mako_docs/examples/metalit/setup/).

**I recommend that in addition to BIOM tables being acceptable as input, QIIME 2 feature table are allowed as well. This would be very straight-forward as QIIME 2 feature tables can be viewed as biom files with QIIME 2 (there are other options for accessing this data as well if this isn't possible). This would allow the large QIIME 2 user community to immediately start using mako with very little additional development effort on the mako end.**

We have added support for Qiime 2 feature tables (both count and taxonomy files) in one of the modules (https://ramellose.github.io/mako_docs/manual/cli/neo4biom/).

**Line 58: It's not clear exactly how OWL is being used here. I think an example would help to illustrate this.**

We have added a more detailed section on the website that explains the data schema and how it is used by mako to confirm database integrity (https://ramellose.github.io/mako_docs/cypher/schema/intro/).

Additionally, we included a dynamic figure that replaces Figure S1 and includes the NCIt definitions (https://ramellose.github.io/mako_docs/cypher/schema/overview/). In a nutshell, mako takes the relationships defined in the OWL file and queries the database to find relationships that do not connect to nodes described in that file. If this is the case, mako raises an error that the relationship is used incorrectly.

**Line 73: 'act as "AND-gates" in the control of gene expression' I don't know the function of "AND-gates" will be generally accessible to the Nature Methods audience - I think a more generally accessible description would make sense here.**

We adjusted the text as follows (line 77-80):
“Network motifs may indicate the presence of specific dynamic behaviours. For instance, in microbial communities, a negative circuit implementing a rock-paper-scissors game can promote diversity (Kerr et al. 2002).”

**The focus of this work is on microbiome networks. Could mako be applied to multi-omics data, such as microbes and metabolites from the same samples? If so, I think this should be highlighted as this is increasingly an area of interest. If not, this is just a suggestion for future work (not a prerequiste for publication).**

Currently, the database has been designed to support a flexible node format that can represent any sort of sample or taxon metadata, including metabolites.

We have amended the manuscript to highlight this future area of interest (line 27-31):

“To simplify this task, we present mako (microbial associations catalog), a software package for the rapid and simple construction and use of network databases from microbiome data, including diverse metadata such as literature-curated metabolic relationships.”

Additionally, we have added two use cases that describe relationships between microbes and metabolites:

https://ramellose.github.io/mako_docs/examples/ibd/intro/
https://ramellose.github.io/mako_docs/examples/metalit/intro/

Both use cases demonstrate that mako is already able to work with diverse data, including multi-omics data.

**Line 158: It seems that "aquatic microbiome" would be a subset of "earth microbiome".**

We have replaced “earth microbiome” with “soil microbiome, including terrestrial plants” (line 164). We chose to combine these since many plant studies include soil samples.

**Line 179: Please clarify "all other abundances were binned".**

We have rephrased this to (line 186-188): “Taxa that fell below the prevalence threshold were merged into a single synthetic taxon so that the total sum of abundances was not disturbed.”

**Reviewer #3:**

**Remarks to the Author:**

**PAPER REVIEW:**

**Mako is a nice tool for users to load microbiome data into a neo4j database, and then perform network-based queries or analysis of the data. I think mako will be a very interesting tool for communities to explore Qiita data sets. However, I would argue that unless the users expect to run exactly the same queries and analysis as provided, certain programming knowledge is still required. For example, "run_motifs.py" codes propionate associations. If a user wishes to check a different function, he/she will need to know how to modify the code.**
We thank the reviewer for the kind comments. We agree that users need to learn how to modify Cypher queries to run their own analysis.

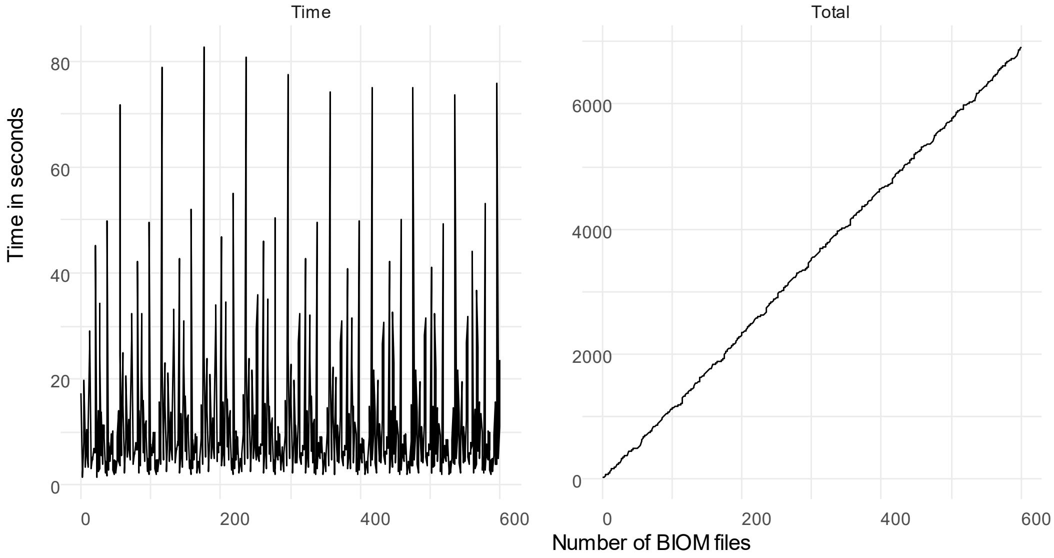
To make this easy, we have expanded the tutorial and included videos to demonstrate how to use Cypher queries in the Neo4j Browser: https://ramellose.github.io/mako_docs/quickstart/guide/

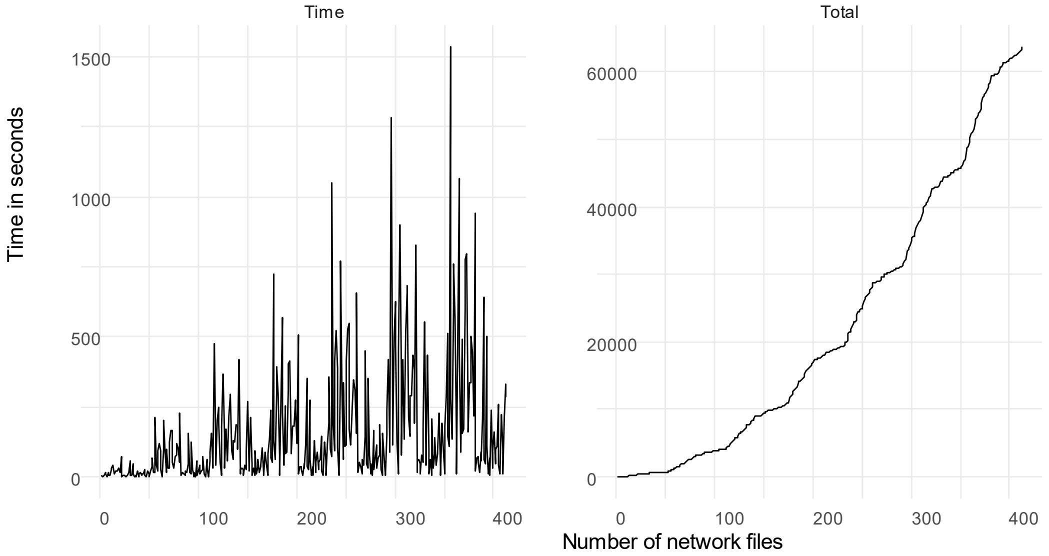
We have also expanded the manual with an introduction on how to access the mako API to run custom queries: https://ramellose.github.io/mako_docs/demo/query/intro/

**The authors curated 60 BIOM files as their test example. It would be interesting to see what the limitation of this system is. How many BIOM files can be reasonable used in an analysis without exceeding the system limitation or incurring a very long wait?**

This is an interesting question, and we have investigated the relationship between BIOM number and runtime. For graph databases, the relationship between size and time is more complicated than for relational databases, because the graph traversals necessary to carry out the queries become larger. Consequently, the main limitation is not necessarily the size of the database, but the profile of the query plan in relation to the size of the database. Since mako implements several checks that support meta-analyses, the load of those "checking" queries is what becomes prohibitive rather than the size of the database itself.

We copied our 60 files and renamed them to have unique taxon identifiers and file names. As we hypothesized, the relationship between time and file depends on the type of query that is required to write the file to the database. The network query plan increases in size as the database grows larger, but the BIOM query plan does not. As a result, the BIOM files display a linear relationship between the number of files and the total time consumed, but the network files do not.





**I would suggest that the 60-network database example in Section 2.1 to be presented as a way to exploring the data instead of a scientific investigation. The reason is that I think an example of 60 data sets (only a tiny subset from all the Qiita datasets) is too small to present a general picture. It pretty much depends on the data sets you have selected. For example, I randomly selected a few files from the provided 60 data sets, and my Animal graph was actually very similar to the Plant graph! Moreover, many of your Plant examples are about roots and rhizosphere, while many of the Non-saline examples are about soil. This makes me wonder whether that contributed to the similarity between your Plant and Non-saline graphs.**

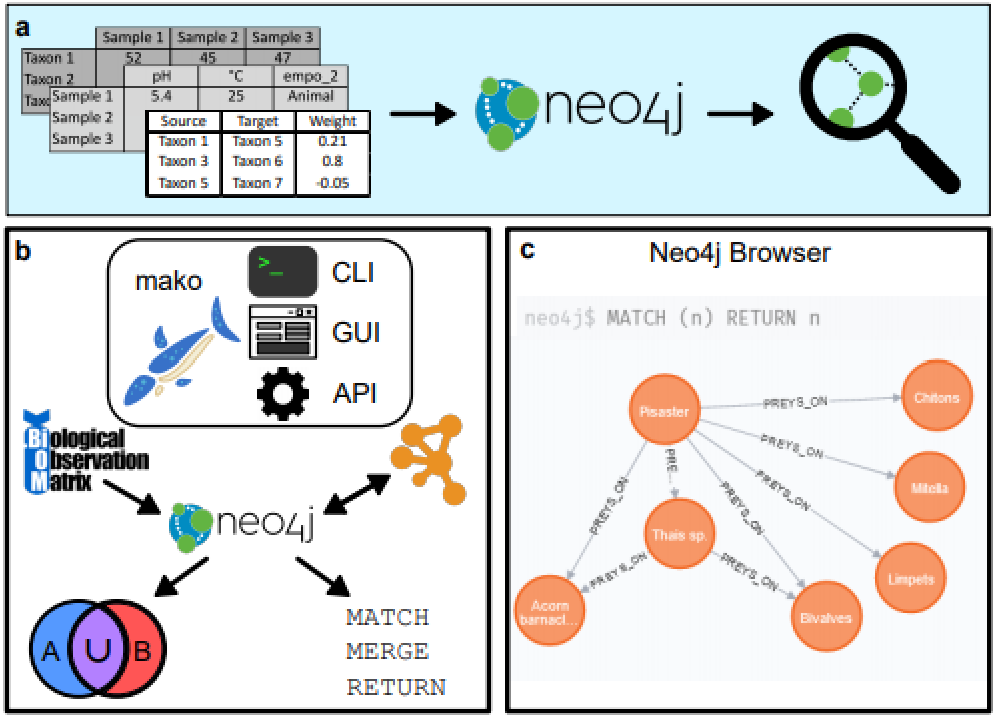
We agree with the reviewer that we only use a small subset of the QIITA datasets and that the EMPO classification we relied on is not a perfect description of the studies and could lead to misleading conclusions. However, some of the patterns we found are in agreement with results from other studies. In particular, motif density is related to edge density in general, which has been found to differ across biomes (Faust *et al.* 2015; Ma *et al.* 2020).

We have discussed these limitations and the agreement with previous findings in the manuscript (line 87-97):

“We used our 60-network database to explore certain combinations of weights in maximally-connected subgraphs across four EMPO (Earth Microbiome Project Ontology) terms (Thompson *et al.* 2017). Our results suggest that there are distinct differences between the four EMPO terms. Firstly, there are more associations attributed to non-saline and animal networks, as could be expected given that 33 and 15 of our 60 data sets had the Animal or Non-saline EMPO terms respectively. Secondly, animal-derived networks appear to have higher motif density. These results are supported by previous studies that found differences in edge density across biomes (Faust *et al.* 2015; Ma *et al.* 2020). However, our approach did not allow us to fully explore differences between plant and non-saline microbiomes, since both studies may contain soil samples."

**Also, I have difficulty viewing Figure 1(c) and Figure S1. I wonder whether the quality of these two figures can be improved.**

We have simplified Figure 1c and improved its resolution in the manuscript:



```
neo4j$ MATCH (n) RETURN n
```

Figure 1: Mako features. **a** The mako software supports uploading of tables and other file formats (in particular networks) to a Neo4j database, which can then be used to carry out meta-analyses. **b** The software includes a command line interface (CLI), graphical user interface (GUI) and application programming interface (API) which can all be used to run the mako functionality. For example, mako can port from BIOM files to a Neo4j database, export to Cytoscape, carry out set operations and several other query-based tasks. **c** Screenshot of the Neo4j browser, which can be used to run queries and access the database. This screenshot displays a part Paine's food web including the keystone species *Pisaster ochraceus*.

In addition, we have removed Figure S1 from the manuscript and represented it on the mako website instead, using a graph visualization method that better accommodates the size of Figure S1: https://ramellose.github.io/mako_docs/cypher/schema/overview/

**CODE REVIEW using Code Ocean:**

**I did have fun trying different examples using Code Ocean.**

**Re. using own data: The authors explained in the Methods section how to prepare biom and network influence files. Since I didn't have my own data, I simply took one of the example data sets, and made some modifications to the data. The tool worked fine with my modified data set.**

**By the way, there's a minor bug: When a Cypher query did not return any results, the code gave a Python TypeError (see below).**

**==========**

**Data Sets: 1721, 1792, 2104, 10097, 11947**

**Cypher query (one of the example queries from README.md): MATCH p=(:Order {name: 'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name: 'o__Clostridiales'}) RETURN p--(:Taxon)--(:Edge) RETURN p**

```
MATCH p=(:Order {name: 'o__Bacillales'})--(:Taxon)--(b:Edge)--(:Taxon)--(:Order {name:
'o__Clostridiales'}) RETURN p--(:Taxon)--(:Edge) RETURN p
Traceback (most recent call last):
File "../code/run_query.py", line 117, in <module>
main(sys.argv)
File "../code/run_query.py", line 46, in main
query_counts = process_query(query_results)
File "../code/run_query.py", line 83, in process_query
for value in results[0]['p']:
TypeError: 'NoneType' object is not subscriptable
```

This error is due to the two return statements in the query. We have amended the Code Ocean capsule to report a more informative error message stating that this type of error is caused by empty query results that can stem from malformed queries. When using mako with the Neo4J Browser, errors in queries are highlighted directly. For the Code Ocean capsule however, we now refer to the mako.log file for a more detailed overview of errors.

**References**

Kerr B, Riley MA, Feldman M & Bohannan BJM. "Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors" *Nature* 2002;**418**:171-174.

Faust K, Lima-Mendez G, Lerat J-S *et al.* Cross-biome comparison of microbial association networks. *Front Microbiol* 2015;**6**.

Ma B, Wang Y, Ye S *et al.* Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* 2020;**8**:1–12.

Tackmann J, Rodrigues JFM, von Mering C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst* 2019;**9**:286–96.

Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**.

**Revision Summary**

We have carried out the following revisions to the manuscript, the presented software and the documentation:

- Implement Qiime 2 file support (https://ramellose.github.io/mako_docs/manual/cli/neo4biom/)
- Created conda distributions (https://ramellose.github.io/mako_docs/manual/introduction/conda/)
- Create a documentation website containing five case studies (two from the main and three additional ones)
    - Motif case study (https://ramellose.github.io/mako_docs/examples/motifs/intro/)
    - Propionate case study (https://ramellose.github.io/mako_docs/examples/propionate/intro/)
    - HMA-LMA in sponge microbial networks (https://ramellose.github.io/mako_docs/examples/sponges/intro/)
    - Enumeration of paths between nodes of interest in a curated gut microorganism-metabolite network to see for instance whether gut bacteria detected in a faecal sample can convert starch to butyrate (https://www.nature.com/articles/ncomms15393) (https://ramellose.github.io/mako_docs/examples/metalit/intro/)
    - Identification of genera with the largest number of associations to metabolites known to be affected by IBD using metabolite-microorganism networks constructed from a recent meta-omics IBD study (https://pubmed.ncbi.nlm.nih.gov/30531976/) (https://ramellose.github.io/mako_docs/examples/ibd/intro/)
- Include documentation on using the API to write a Python script with custom queries (https://ramellose.github.io/mako_docs/demo/query/intro/)
- Include documentation on how the OWL file is used to check database structure (https://ramellose.github.io/mako_docs/cypher/schema/intro/)
- Include additional example Cypher queries in the documentation (https://ramellose.github.io/mako_docs/cypher/introduction/intro/)
- Expand Code Ocean capsule to better support custom queries (https://codeocean.com/capsule/0482418/tree)
- Replace/complement complex figures and enhance tutorial with video and graph visualizations (Quickstart: https://ramellose.github.io/mako_docs/quickstart/guide/, Figure S1 alternative: https://ramellose.github.io/mako_docs/cypher/schema/overview/)

We have also carried out the following additional analyses:

- Compared SQL queries to Cypher queries
- Plotted runtime against number of BIOM files

## Decision Letter, second revision:

Date: 22nd Jul 21 22:42:56
From: Lin.tang@nature.com
To: karoline.faust@kuleuven.be
CC: methods@us.nature.com;ziqian.li@nature.com
Subject: Decision on Nature Methods submission NMETH-BC45289B
Message: 22nd Jul 2021

Dear Professor Faust,

Your Brief Communication, "Fast and flexible analysis of linked microbiome data with mako", has now been seen again by 3 reviewers. As you will see from their comments below, although the reviewers find your paper has been improved, they still raised a number of important concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to address these concerns. Among other revisions, we think it would be necessary to fully address Reviewer 2's concerns on the documentation/software.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

***[REDACTED]***

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within 6 weeks. We are very aware of the difficulties caused by the COVID-19 pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please submit reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
Please include a “Data availability” subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: “The data that

support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID
Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on

'Modify my Springer Nature account'. For more information please visit please visit http://www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Lin

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
General comments:
I am happy to review this revised version of mako. The description, documentation and case studies have been improved a lot with regarding to the motivation, main contributions, target audience and how to use the tool. The demo also works out of the box. This reviewer appreciate the efforts made by the authors.

Overall, mako has addressed a difficult technical challenge in a way that can produce meaningful knowledge. The inclusion of a conceptual model written in OWL that can be used to check the integrity of the Neo4j database is a non-obvious contribution, but a good start pointing for the community

Specific comments:
Although I agree with the general comparisons (SQL vs Neo4J), it is important to point out the serious performance issues associated with adding nodes to a neo4j graph. For instance, there is no details of why the code starts a database rather than contributes to a database, assuming the users will use the tool which calls the command line constructor for a neo4j database: not the query method to do so. In

our experience, at half million relationships, the cost of doing so already several times slower than creating the database from scratch.

Reviewer #2:
Remarks to the Author:

I appreciate the effort that the authors put into building their documentation and improving the install process. I ran into several issues trying to follow the documentation. I am a software engineer, so I was able to debug some of these, but most of your users are not going to be software engineers. I think these issues in the documentation are going to be a major barrier to adoption of this software. I recommend sitting down with a biologist (who is not a bioinformatician!) and watching them try to follow the instructions to start using this software, without you offering any input on how to use it.

The conda install instructions aren't working (or the conda environment is incomplete). I created the conda environment according to the instructions, but got a "command not found" error when I called "mako -h". To address this I cloned the repository and pip installed mako from the cloned repository.

The quickstart should also include the instructions for setting up neo4j. Otherwise the user has to go back to installing when they are trying to run their analysis.

On trying to connect to the database, I received the following error: "ServiceUnavailable: WebSocket connection failure. Due to security constraints in your web browser, the reason for the failure is not available to this Neo4j Driver" This required me to change the port referenced in the documentation.

On trying to connect to the database I had to parse the username and password from the command.

Make the commands easier to copy/paste (the user shouldn't have to edit them). For example:
- "mako io -fp local_filepath -cf -net demo_1.graphml demo_2.graphml" results in an error because "local_filepath" doesn't exist. In the text just before this, you told the user to navigate to the directory containing the downloaded files, so you know what directory the user is in - you don't need to make them change the command (just have them specify '.' or '$PWD').
- "The Docker command below is given as a multi-line command for clarity" - Backslashes could be appended to the end of each line so the multi-line command could be copy/pasted.

The documentation should help users interpret the results. It shows how to generate some network diagrams, but I don't know how to interpret what they mean.

In the "Calculate metadata correlations" you note that "that this is not a robust way to estimate correlations between taxa and metadata" but it's what you're showing the user how to do. This seems dangerous - why are you illustrating how to do something they shouldn't be doing?

In the "Setting up the database" section, no instruction is given on how to access the biom files or how to generate the network_filepaths. I'm therefore stuck again trying to work through this documentation, and I can't give a complete review of the software.

Reviewer #3:
Remarks to the Author:
Mako is a nice tool for users to load microbiome data into a neo4j database, and then perform network-based queries or analysis of the data. I think mako will be a very interesting tool for communities to explore Qiita data sets. Since the data is loaded into a neo4j database, it opens the possibility for those users who are familiar with neo4j to expand the database schema and load additional data to broaden their research.

The new user guide and examples are a big improvement, though users still need to have programming knowledge and skills to use the tool. Therefore, the target audience for the tool will not be general biology researchers, but bioinformatics experts with programming training (though I have noticed that many next generation biologists are tech-savvy). A possible scenario is to have bioinformatics experts using mako to develop a set of applications for regular biology researchers to use.

The points this reviewer raised previously have all been properly addressed. This reviewer does not have other concerns except for maybe the complexity of mako.

There are a couple of typos in the manuscript:

Page 6, line 105 ".." ==> "."
Page 10, line 191 "Flashweave" ==> "FlashWeave"

## Author Rebuttal, second revision:

Reviewer #1:

Remarks to the Author:

General comments:

I am happy to review this revised version of mako. The description, documentation and case studies have been improved a lot with regarding to the motivation, main contributions, target audience and how to use the tool. The demo also works out of the box. This reviewer appreciate the efforts made by the authors.

We thank the reviewer for the kind remarks.

Overall, mako has addressed a difficult technical challenge in a way that can produce meaningful knowledge. The inclusion of a conceptual model written in OWL that can be used to check the integrity of the Neo4j database is a non-obvious contribution, but a good start pointing for the community

Specific comments:

Although I agree with the general comparisons (SQL vs Neo4J), it is important to point out the serious performance issues associated with adding nodes to a neo4j graph. For instance, there is no details of why the code starts a database rather than contributes to a database, assuming the users will use the tool which calls the command line constructor for a neo4j database: not the query method to do so. In our experience, at half million relationships, the cost of doing so already several times slower than creating the database from scratch.

We agree with the reviewer that starting up the database is a significant computational effort. However, the only place where the code always starts a database is in the Code Ocean capsule. This is not a limitation of the mako toolbox, but rather a fundamental issue with compute capsules not being persistent; for each run, the database has to be recreated. This is not the case when using mako locally.

We are not fully certain to what other parts of the code the reviewer could refer. There is a command line constructor in mako that sets up a database but using it is optional and the documentation refers to other methods for setting up databases (namely, via Docker or via running Neo4j Community Server).

Creating a database from scratch can be done rapidly by exporting a database to a CSV file and then restoring the database from this CSV file. However, this requires the database to have been populated according to the appropriate data schema already, since this method of constructing databases does not check whether the database conforms to the data schema.

To clarify the issue of the database being restarted, we now emphasize this at the top of the Code Ocean page:
"Each time the capsule is run, the Neo4j database is set up from scratch. To save time (and Code Ocean compute time) and reuse the same database for multiple queries, please install Neo4j locally or through Docker: https://ramellose.github.io/mako_docs/quickstart/guide/."

Reviewer #2:

Remarks to the Author:

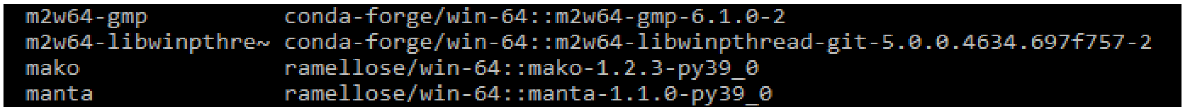
I appreciate the effort that the authors put into building their documentation and improving the install process. I ran into several issues trying to follow the documentation. I am a software engineer, so I was able to debug some of these, but most of your users are not going to be software engineers. I think these issues in the documentation are going to be a major barrier to adoption of this software. I recommend sitting down with a biologist (who is not a bioinformatician!) and watching them try to follow the instructions to start using this software, without you offering any input on how to use it.

We appreciate the suggestion from the reviewer and intend to continuously improve the documentation as required. We previously had five beta testers with two having a very limited computational background. To improve the documentation, we doubled the number of beta testers; the recruited testers had no bioinformatics background and limited computational skills. From our interactions with the new beta testers, we found that some instructions were easy to overlook so we added these in duplicate where appropriate. Our new testers ran into issues when they forgot to download and unzip files, could not find the connection information for the database or did not navigate to the correct folder. Consequently, we updated our quickstart instructions accordingly (https://ramellose.github.io/mako_docs/quickstart/guide/) and added a FAQ (https://ramellose.github.io/mako_docs/faq/). In addition to describing frequently experienced issues, the FAQ directs users to the Github issues pages.

The conda install instructions aren't working (or the conda environment is incomplete). I created the conda environment according to the instructions, but got a "command not found" error when I called "mako -h". To address this I cloned the repository and pip installed mako from the cloned repository.

We apologize for the issues the reviewer experienced during installation and thank the reviewer for going through the effort of installing from the cloned repository. We checked our macOS, Windows and Ubuntu environments and the conda environment works there. However, we were able to reproduce the “command not found” error when conda incorrectly prioritized channels, leading the installation command to install a different environment where the “mako –h” command does not work. Without additional information, we can unfortunately not be confident that this is indeed the error the reviewer ran into. However, we have added the screenshot and debugging text below to the website (https://ramellose.github.io/mako_docs/installation/instructions/conda/). This screenshot can be used to diagnose whether the installation command is installing mako from the correct source.

“If you are unable to run the mako –h command, please check your installation logs to see whether mako was downloaded from the correct source. You may need to configure your conda channel order in case the source is not similar to the source shown below. First try running conda config –add channels ramellose again: if this does not work, please see the conda webpage (https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-channels.html) for additional instructions.”



```
 m2w64-gmp           conda-forge/win-64::m2w64-gmp-6.1.0-2
 m2w64-libwinpthre~ conda-forge/win-64::m2w64-libwinpthread-git-5.0.0.4634.697f757-2
 mako               ramellose/win-64::mako-1.2.3-py39_0
 manta              ramellose/win-64::manta-1.1.0-py39_0
```

The quickstart should also include the instructions for setting up neo4j. Otherwise the user has to go back to installing when they are trying to run their analysis.

We have copied the Docker instructions from the Neo4j section to the quickstart page (https://ramellose.github.io/mako_docs/quickstart/guide/).

On trying to connect to the database, I received the following error: "ServiceUnavailable: WebSocket connection failure. Due to security constraints in your web browser, the reason for the failure is not available to this Neo4j Driver" This required me to change the port referenced in the documentation.

We apologize for the inconvenience. For the quickstart, we assume users follow the instructions for setting up Docker (https://ramellose.github.io/mako_docs/neo4j/docker/docker/). We have now clarified this in the quickstart section. Installing Docker changes the default ports so that the Docker configuration does not interfere with a local installation using the same ports. The Browser instructions detail these alternative ports: https://ramellose.github.io/mako_docs/neo4j/browser/browser/. We have added a reference to this page in the demo section.

On trying to connect to the database I had to parse the username and password from the command.

We apologize for the lack of clarity. We are not certain where the reviewer ran into this issue; we included Docker parameters wherever possible. For example, the Pisaster demo explains where to find the username and password (https://ramellose.github.io/mako_docs/demo/pisaster/connect/), while the Browser documentation provides an overview of alternative configurations (https://ramellose.github.io/mako_docs/neo4j/browser/browser/). Similarly, the Sponge case study, all other case studies and the demos includes these parameters in the commands already (https://ramellose.github.io/mako_docs/examples/sponges/setup/, https://ramellose.github.io/mako_docs/demo/vignette/biom/). We have now explicitly included these parameters on the quickstart page as well (https://ramellose.github.io/mako_docs/quickstart/guide/).

Make the commands easier to copy/paste (the user shouldn't have to edit them). For example:

- "mako io -fp local_filepath -cf -net demo_1.graphml demo_2.graphml" results in an error because "local_filepath" doesn't exist. In the text just before this, you told the user to navigate to the directory containing the downloaded files, so you know what directory the user is in - you don't need to make them change the command (just have them specify '.' or '$PWD').

We have replaced the "local_filepath" parameter with "." throughout the documentation, for example at https://ramellose.github.io/mako_docs/demo/vignette/biom/.

- "The Docker command below is given as a multi-line command for clarity" - Backslashes could be appended to the end of each line so the multi-line command could be copy/pasted.

We thank the reviewer for the helpful suggestion. We have adapted the white space of the commands and added the backslashes, for example at https://ramellose.github.io/mako_docs/neo4j/docker/docker/.

The documentation should help users interpret the results. It shows how to generate some network diagrams, but I don't know how to interpret what they mean.

The interpretation of the results will depend on the data set and the query. For instance, the result of a query in a network with known metabolite-bacteria links has to be interpreted differently from the result of a query in an association network where links are inferred. For the study cases, we already discuss the interpretation of the results. For example, we use Cypher queries to find taxonomic groups linked to sphingolipids, a metabolite found to be involved in IBD (https://ramellose.github.io/mako_docs/examples/ibd/data/). However, since the demo uses a highly simplified synthetic data set, any biological interpretation is limited. We have added a brief paragraph that discusses this (https://ramellose.github.io/mako_docs/demo/vignette/networks/):

“In this case, the shown edges are part of the imported synthetic networks, the demo_1.graphml and the demo_2.graphml files. The interpretation of these edges therefore depends on the type of data that the grapmhl files were based on. If these were association networks, the orange Edge nodes would represent statistically significant links between different OTUs (and therefore do not represent observations of interactions). However, if they were metabolic interaction networks, each Edge node could represent an exchange of metabolites.”

We also briefly discuss the interpretation of the metadata correlations (https://ramellose.github.io/mako_docs/demo/vignette/metadata/):

“The query ‘MATCH p=(n:Taxon)--(:Property) RETURN p LIMIT 50’ returns 50 patterns consisting of Taxon nodes connected to Property nodes. In this case, the only possible connections matching this pattern were previously generated by the hypergeometric test, as the name of the relationships indicates (Figure 4). Here, the query result therefore shows that OTU_1 was the only taxon linked to a certain pH and body site, as it was connected to those nodes via a hypergeometric test.”

In addition, we now better explain the logic behind the Cypher queries in the demo (for example here: https://ramellose.github.io/mako_docs/demo/vignette/biom/).

In the "Calculate metadata correlations" you note that "that this is not a robust way to estimate correlations between taxa and metadata" but it's what you're showing the user how to do. This seems dangerous - why are you illustrating how to do something they shouldn't be doing?

We apologize for the lack of clarity of this statement. Spearman correlations and hypergeometric tests are often suitable ways to estimate associations, but the implementations in mako are unable to explicitly handle experimental structure. For example, it is not possible to calculate correlations only for a subset of samples (i.e. leave out blanks or calculate correlations per group). We should have stated that the mako toolbox cannot assess whether these statistical approaches are suitable for the user's purpose. Therefore, we have amended the documentation to clarify that this is for exploratory purposes only.

We have rephrased this section of the documentation as follows:
"Note that this functionality is meant for exploratory analysis only and does not include multiple-testing corrections. Depending on the type of metadata available to you and the structure of your experimental design, more appropriate statistical methods may be available."

Although this functionality may not be optimal, it is mainly intended to be a showcase of how relatively simple statistical operations such as Spearman correlations can be carried out using the mako toolbox. The modular design makes it relatively straightforward to adopt other types of analyses. We therefore still prefer to present this to users to demonstrate that these sorts of operations are possible.

In the "Setting up the database" section, no instruction is given on how to access the biom files or how to generate the network_filepaths. I'm therefore stuck again trying to work through this documentation, and I can't give a complete review of the software.

We apologize for the lack of clarity of the documentation. We originally intended for users to download these files from the Code Ocean capsule as stated in the introduction of this case study. However, since this was not clear enough, we have adapted the case study to provide separate download links (https://ramellose.github.io/mako_docs/demo/query/intro/).

Reviewer #3:

Remarks to the Author:

Mako is a nice tool for users to load microbiome data into a neo4j database, and then perform network-based queries or analysis of the data. I think mako will be a very interesting tool for communities to explore Qiita data sets. Since the data is loaded into a neo4j database, it opens the possibility for those users who are familiar with neo4j to expand the database schema and load additional data to broaden their research.

We thank the reviewer for the kind comments and hope the mako tool will indeed help users explore their data in this way.

The new user guide and examples are a big improvement, though users still need to have programming knowledge and skills to use the tool. Therefore, the target audience for the tool will not be general biology researchers, but bioinformatics experts with programming training (though I have noticed that many next generation biologists are tech-savvy). A possible scenario is to have bioinformatics experts using mako to develop a set of applications for regular biology researchers to use.

The reviewer's comments indeed describe one intended scenario for mako; the API was specifically developed with advanced users in mind, so they can easily adapt the toolbox for their own needs. We hope that the toolbox will therefore promote adoption of graph database technologies in microbiology.

The points this reviewer raised previously have all been properly addressed. This reviewer does not have other concerns except for maybe the complexity of mako.

There are a couple of typos in the manuscript:

Page 6, line 105 ".." ==> "."

Page 10, line 191 "Flashweave" ==> "FlashWeave"

We thank the reviewer for noticing these issues and have fixed them in the manuscript.

**Decision Letter, third revision:**

| | |
|---|---|
| Date: | 23rd Sep 21 22:08:24 |
| From: | Lin.tang@nature.com |
| To: | karoline.faust@kuleuven.be |
| CC: | methods@us.nature.com;ziqian.li@nature.com |
| Subject: | AIP Decision on Manuscript NMETH-BC45289C |
| Message: | Our ref: NMETH-BC45289C |

22nd Sep 2021

Dear Dr. Faust,

Thank you for submitting your revised manuscript "Fast and flexible analysis of linked microbiome data with mako" (NMETH-BC45289C). It has now been seen by the Reviewer 2 and their comments are below. The reviewer finds that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referee's final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Methods Please do not hesitate to contact me if you have any questions.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #2 (Remarks to the Author):

The authors have addressed the issues that I ran into in my previous tests, and I have now been able to follow the install and usage instructions. I have only a few minor comments at this time which should be considered optional to address before publication (I do think addressing these will help with adoption of the software).

Minor points:

The sponge_biomfiles.zip and sponge_networks.zip inflate to the current diretory when running (for example) "unzip sponge_networks.zip". This results in an error when trying to run the mako commands to write data to the neo4j database because the referenced directory sponge_networks doesn't exist. This was easy to work-around but will confuse a novice user.

The query in the quickstart guide contains line breaks and therefore results in an error when it's copy/pasted into neo4j browser. This will also confuse a novice user.

I still found interpretation of results to be lacking a bit in the tutorial. I think providing more detail on this for users will help with getting users to adopt the system.

## Author Rebuttal, fourth revision:

**Reviewer #2 (Remarks to the Author):**

**The authors have addressed the issues that I ran into in my previous tests, and I have now been able to follow the install and usage instructions. I have only a few minor comments at this time which should be considered optional to address before publication (I do think addressing these will help with adoption of the software).**
We thank the reviewer again for their helpful comments and have addressed the remaining comments.

**Minor points:**

**The sponge_biomfiles.zip and sponge_networks.zip inflate to the current diretory when running (for example) "unzip sponge_networks.zip". This results in an error when trying to run the mako commands to write data to the neo4j database because the referenced directory sponge_networks doesn't exist. This was easy to work-around but will confuse a novice user.**
We added an additional instruction – "Unzip these files into two folders, sponge_biomfiles and sponge_networks respectively.
In your command-line interface, navigate to the location that contains both folders."

**The query in the quickstart guide contains line breaks and therefore results in an error when it's**

**copy/pasted into neo4j browser. This will also confuse a novice user.**
Horizontal scroll boxes have been added for all code sections with line breaks in the tutorial (see https://ramellose.github.io/mako_docs/quickstart/guide/).

**I still found interpretation of results to be lacking a bit in the tutorial. I think providing more detail on this for users will help with getting users to adopt the system.**
The following explanation has been added to the tutorial (https://ramellose.github.io/mako_docs/quickstart/guide/):
"You should now have a running Neo4j database that contains a number of networks derived from different taxonomic orders of sponges. This can be used to find associations linked to specific bacterial orders, like in the query above. This query finds all associations between taxa assigned to the orders Nitrospirales and Cenarchaeales, across all networks in the database.

It is also possible to ask questions such as 'Which associations are found in at least three sponge orders?' or 'Are there associations that form a motif which are recovered in multiple networks?'. Additionally, you can add extra data to the database, such as sponge LMA or HMA status (low or high microbial abundance) and use this to query associations linked to those statuses. The benefit of using a Neo4j database is that the queries look relatively similar to schematic overviews you might make of the information you are interested in. This makes writing those queries a lot more intuitive."

## Final Decision Letter:

Date: 1st Nov 21 00:12:08
From: Lin.tang@nature.com
To: karoline.faust@kuleuven.be
CC: ziqian.li@nature.com,methods@us.nature.com
BCC: rjsproduction@springernature.com,communities@nature.com
Subject: Decision on Nature Methods submission NMETH-BC45289D
Message:

1st Nov 2021

Dear Professor Faust,

I am pleased to inform you that your Brief Communication, "Fast and flexible analysis of linked microbiome data with mako", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our January print issue, and will be published online prior to that. The received and accepted dates will be 19th Feb 2021 and 1st Nov 2021. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-BC45289D and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

Please note that *Nature Methods* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

**Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates.** For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF. As soon as your article is published, you will receive an automated email with your shareable link.

Please note that you and your coauthors may order reprints and single copies of the issue containing your article through Springer Nature Limited's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points. Thank you very much again for publishing your paper in our journal!

Best regards,

Lin

Lin Tang, PhD
Senior Editor
Nature Methods