nature portfolio

Peer Review File

Knowledge and Data-driven Two-layer Networking for Accurate Metabolite Annotation in Untargeted Metabolomics

Corresponding Author: Professor Zheng-Jiang Zhu

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Zhang et al., presented the work entitled Knowledge and Data-driven Two-layer Networking for Accurate 2 Metabolite Annotation in Untargeted Metabolomics where they describe the construction and validation of a network-based approach to perform metabolite annotations. The work is well written, and the authors provide great detail in each validation step, providing data, code and multiple supplementary materials. However, the work needs clarifications, especially in the benchmarking and validation sections, to enable the users to better understand the efficiency of the method and how to better employ it.

The three main issues I believe should be addressed are:
1) Benchmarking: from what I understood, the authors used a “concordance” of the tools to annotate 4302 features observed in multiple datasets. As all tools provide in silico predictions, one can not rule out that all predictions are wrong. The ideal ground truth for benchmarking would be to compare the accuracy of all tools for standard annotation. This has to be done carefully, taking into account random seed metabolites for MetDNA3 and comparing the remaining metabolites accordingly for each tool. From the methods section, I could not understand what structure databases were used to search with the in silico tools, this should be clarified and standardized for the benchmark.
2) I could not understand the rationale behind the validation of the metabolites γ-Glu-Thr-Gly and N-glycolyltaurine out of thousands of predictions. That is to say, how are the users going to be able to avoid false positives? Would be much believable if the authors explained a filtering strategy (maybe I missed it in my reading) or tried to validate a large number of predictions, thus estimating the false positive rates.

This is also a general concern with the author’s use of the words ‘confidece’ and ‘significance’, maybe this wording should be used with more caution, as the ‘Knowledge layer’ was expanded with predictions not validated experimentally, and the matching provided with mass spectrometry data is inherently subjected to errors. The author should consider in future versions having more appropriate estimates of significance, in the lines of
Scheubert, K., Hufsky, F., Petras, D. et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8, 1494 (2017). https://doi.org/10.1038/s41467-017-01318-5
Palmer, A., Phapale, P., Chernyavsky, I. et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods 14, 57–60 (2017). https://doi.org/10.1038/nmeth.4072

3) Is MetDNA3 designed for animal cells/tissues only? Sorry if that was implicit. If yes, change the title accordingly. I believe the need of clarification stems from the fact that transformations provided by BioTransformer have what the authors called ‘Human Super Transformer’ component, and if the transformations mapped are indeed biased for animals or humans, users could increase false positives trying to use it for other organisms.

Another issue is that although the code, data, and web interface are available for review purposes, it would be important to

provide a worked example of data pre-processing and metabolite annotation to facilitate the review process. I tried to install the package on standard Google Colab (https://colab.research.google.com/notebook#create=true&language=r) following the instructions of the github page and was unable to install it. The option to subscribe to the web tool and try it could breach the anonymity of the review.

“installation of package ‘/tmp/RtmpW0b815/file1075f20d1e4/MrnAnnoAlgo3_3.1.1.tar.gz’ had non-zero exit status”

>session_info()
$platform
$version
'R version 4.4.3 (2025-02-28)'
$os
'Ubuntu 22.04.4 LTS'
$system
'x86_64, linux-gnu'
$ui
'X11'
$language
'en_US'
$collate
'en_US.UTF-8'
$ctype
'en_US.UTF-8'
$tz
'Etc/UTC'
$date
'2025-04-15'
$pandoc
'NA'
$quarto

Are the code and model for the graph neural network available?

Minor issues

a) What was the chemical taxonomy criterion used to classify the compounds? Was it that of ClassyFire (Djoumbou Feunang et al., 2016)?

b) From the following statement, between lines 120-121: “Additionally, evaluation of the topological properties of the curated MRN revealed a higher global clustering coefficient and an improved degree distribution relative to knowledge databases”. What is the global clustering coefficient based on and how was it obtained, as well as the degree distribution relative to knowledge databases? i) is the number of nodes that have a given degree? Legend or text could be more descriptive.

c) As shown between the lines 140-141: “MS2 similarity between features was calculated and applied as a filtering constraint to eliminate unwanted nodes, further refining the network structure.” What metric was used to determine the similarity between the spectra in MS2? Cosine score, modified cosine score, spectral entropy?

d) As shown between the lines 206-208: “Furthermore, network-based annotation propagation expanded the annotated metabolome to 12,508 metabolites (level 3 annotations), including 9,410 known and 3,098 unknown metabolites (Figure 3c), highlighting the substantial increase in metabolite coverage.” What is the annotation rate in terms of the total number of features captured in pre-processing for both Orbitrap and Astral data?

e) Do the authors intend to submit the fragmentation spectra of the new compounds characterized in this work to the databases used by the metabolomics community, such as MoNA?

f) Do the authors consider submitting the raw data do metabolights, metabolomics workbench or gnps?

(Remarks on code availability)
Code, data, and web interface are available for review purposes, it would be important to provide a worked example of data pre-processing and metabolite annotation to facilitate the review process. I tried to install the package on standard Google Colab (https://colab.research.google.com/notebook#create=true&language=r) following the instructions of the github page and was unable to install it. The option to subscribe to the web tool and try it could breach the anonymity of the review.

“installation of package ‘/tmp/RtmpW0b815/file1075f20d1e4/MrnAnnoAlgo3_3.1.1.tar.gz’ had non-zero exit status”

>session_info()
$platform

$version
'R version 4.4.3 (2025-02-28)'
$os
'Ubuntu 22.04.4 LTS'
$system
'x86_64, linux-gnu'
$ui
'X11'
$language
'en_US'
$collate
'en_US.UTF-8'
$ctype
'en_US.UTF-8'
$tz
'Etc/UTC'
$date
'2025-04-15'
$pandoc
'NA'
$quarto

Are the code and model for the graph neural network available?

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
In this manuscript, Zhang et al report a new computational algorithm, MetDNA3, for metabolite identification in LC-MS mass spectrometry metabolomics. The work is a further development of the authors' previous systems, MetDNA and KGMN / MetDNA2, and the recursive algorithm used to annotate features appear to be similar. A novel feature in this version is a deep learning method for extending known metabolic networks by additional predicted reactions. The methodology is interesting, and I appreciated the experimental validation of several predicted metabolites suggesting that MetDNA3 does enable real discoveries. This new version also seems to improve substantially upon MetDNA2, with increased coverage and accuracy. However, the systematic evaluation of the method is not easy to understand, and some of the accuracy metrics used might exaggerate the performance of the system. The presentation and some of the method details needs clarification, as detailed below.

** Major points **

1. Regarding accuracy of the metabolic reaction network (MRN; Fig 1c), the reported AUC of 0.9947 might give readers the impression that the MRN is almost free of errors, but I am not convinced that AUC is the relevant measure here. In their cross-validation scheme, the authors choose ~15k true reaction pairs and ~15k non-pairs (random pairs) among 11,532 metabolites. From Fig 1c it looks like the MRN achieves 100% TP rate at about 1% FP rate (false positives / negatives), but the classification problem is extremely unbalanced. There are 66M pairs total among the 11,532 metabolites, most of which are likely to be negatives, so with 1% FP rate one would expect a total of ~660k predicted pairs in the cross-validated MRN (I did not find the actual number), of which 15k were true positives, which means the the false discovery rate (false positives / predicted positives) could be as high as 97%.

Although FDR estimated this way is conservative since the known network is incomplete, it would seem that most of the pairs in the large MRN are likely to be false positives. In this context, citing "improvements of 66.4-fold and 162.8-fold" seems misleading. Certainly the MRN has higher coverage, but likely at the cost of high FDR. I think the authors should provide the FDR as well, so that the uninitiated reader can understand the amount of false positives in the MRN. I would appreciate a discussion of these points.

2. In light of the above, I also wonder how to interpret the observation that when reconciling the MRN with MS2-based molecular networks (the "two-layer topology", Fig 2), most of the MRN is discarded: only 2.3% of reactions are kept, which is close to my rough FDR estimate of 97% above. Is it the case that most of the novel, GNN-predicted part of the MRN was discarded? If so, I wonder to what extent the MRN contributes to the final results. It would be useful to see a direct comparison of the method performance (as in Fig 5d) against (i) using only the knowledge databases instead of the full

MRN, and (ii) using only the MS2-based network (molecular networking), to better understand the effect of the "knowledge layer".

3. In the evaluation of metabolite annotation accuracy (which is the key performance metric) in Fig 3, the authors report "coverage" which I assume is TP/P (I think "recall" is the standard term), and accuracy defined as 1 - total error rate for "Top 3". The results seem impressive, but again it is not clear how to interpret the numbers. Specifically:

a) How were the Top N annotations selected? I assume the authors counted the N highest-scoring putative metabolite annotations for each peak as predicted positives, but it is not clear how predicted annotations were scored -- I didn't see this described in methods?

b) Was coverage also computed using the same Top 3 method as accuracy, so that the numbers are comparable?

c) Given that the prediction problem is unbalanced, the authors should show the FP and FN rates separately in Fig 5 rather than just accuracy, so that it is clear where the improvement is coming from. Also, as discussed above, I would like to see the false discovery rate as well (with the understanding that is a conservative estimate).

4. MetDNA3 seems to rely heavily on MS2 similarity (molecular networking) between molecules, since new connections will only be made for peaks X and Y that have similar MS2 spectra, and reaction pairs are also constrained to agree with MS2 similarity (Fig 2a). With this in mind, I'm wondering if evaluating correctness of the method against MS2-based compound identification (from standards) is a circular or biased to some extent?

5. A key piece of evidence for accuracy of the MetDNA3 method is the independent validation of predictions by MS2. I appreciated this effort, and it does indicate that the method is capable of identifying both "known unknowns" and truly novel compounds. However, it would be most useful to discuss how these candidates were selected, and whether certain compound classes (such as peptides) are more likely to be discovered by the algorithm.

6. Finally, MS2-based matching (against standards) is throughout assumed to be the gold standard for correct metabolite identification. While MS2 can certainly be strong evidence when used correctly, MS2 spectra often cannot distinguish between closely related compounds, for example isomers such as glucose, fructose, and galactose; also, MS2 is underpowered for small molecules that don't fragment well. This aspect should also be discussed.

** Minor points **

7. I find that the main results are not clearly stated in the abstract:
a) Does "10-fold improved accuracy" refers to computational efficiency (reduced computation time) ? please clarify.
b) The statement "It annotated over 1,600 metabolites" refers to MS2 matching against the authors' standard library, which is unrelated to the algorithm.
c) "12,000 new metabolites" - for clarity, these should be called putative annotations.

8. In Fig 1c, the Tanimoto coefficient is substantially higher in MRN and especially in the biotransformer network, suggesting that the predicted network focuses on structurally similar molecules. This is understandable given the methodology, but nevertheless seems like a weakness, since reactions that yield products that are structurally different from the substrates would be biased against. Could be discussed.

9. It would be interesting to see how accuracy (Fig 3) depends on the number of recursive iterations taken by the algorithm. Intuitively, one might expect that after several steps of inference, each of which has some error probability, the annotation error rate will shoot up.

10. Methods page 26, RT match tolerance was specified as "30%"; does this mean 30% relative error, e.g. 1 min +/- 0.3 min? If so this seems quite wide and would not give much information for matching?

11. For MS2 matching, the methods state a dot product cutoff of 0.8, but it is not clear how many MS2 fragments were required? Low-quality spectra with few fragments can easily yield false matches even at high dot product values.

12. For MS2 spectral similarity between features, the authors state a dot product cutoff of 0.5, but don't explain how mass differences were accounted for. Directly comparing MS2 spectra between molecules typically requires adjusting for the precursor mass differences. Please clarify.

13. The term "reaction pair" was a bit confusing to me, as it seems to suggest a pair of two reactions rather than a pair of metabolites; would suggest "metabolite pair" or "reaction".

Roland Nilsson
Karolinska Institutet

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Reviews follow attached to preserve formatting.

(Remarks on code availability)
The code can be used to perform some of the tasks available through the web tool, not all.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have made several improvements to the manuscript, but a couple of issues still remain that the authors may wish to address.

1. I still have doubts about the error rate estimates for the MRN. If I understand correctly, the FPR estimate of 2.1% measures the ability of the GRN model to distinguish a true reaction pair from a randomly chosen pair (Fig 1b). From their rebuttal, it seems like the authors agree that the corresponding FDR on the full set of pairs would be very high, as the number of negative pairs vastly outnumber the positive ones. This is because the reaction pair prediction problem is highly unbalanced, which makes it inherently difficult.

To get around this problem, the authors filter the reaction pairs by mass difference and by molecular structure overlap. These are stringent filters, reducing the number of candidate pairs from 1.44 billion to only 1.66 million (almost 1000x). Then, the authors directly apply the above 2.1% FPR estimate to the filtered 1.66M pairs to come up with an estimate of 2.4% FDR. I don't think this reasoning is correct, because the 2.1% FPR was estimated from an entirely different test set: the GNN is a molecular structure classifier, and FPR as low as 2.1% in the test scheme of Fig 1b is not surprising because the randomly chosen negative pairs consist of completely unrelated structures, so the classification problem here is easy. In contrast, when applying the GNN to the 1.66M pairs that have already been stringently filtered for structural similarity, the classification problem is much harder (since all pairs are now structurally similar) and consequently I would expect the actual FPR on this set to be much higher. So I think this calculation mixes apples and oranges. Moreover, the 2.4% FPR describes only the GNN step, but the interesting error rate is of course that of the MRN as a whole.

Instead, I would suggest that the authors estimate the FDR in the MRN by cross-validating their entire procedure: simply apply the full computational pipeline shown in Suppl Fig 1i to the test set from Fig 1b, count the number of resulting pairs among the positive (RP) and negative (non-RP) classes, calculate the per-pair FPR and then correct for unbalancing to get the resulting FDR for reaction pairs in the MRN.

I realize that the MRN is not the final output of the MetDNA3 algorithm, but as the MRN is a valuable resource in itself I think an accurate error rate estimate is important. As it stands, the manuscript claims to expand existing biochemistry knowledgebases more than "162.8-fold" with an FDR of 2.4%, which sounds rather unlikely.

3. The newly added FDR estimation using decoy spectra is interesting, but I don't think it qualifies as a method validation. The decoy-based method is an FDR *estimation* technique, used to assess the rate of false discovering in the absence of a test set; it is only an approximate method, and as Scheubert et al 2017 clearly show, the decoy method can be optimistic, particularly for small q-values. However, in this study the authors actually have a test set, so I think the true FDR should be directly computed from the test set rather than estimated, as done in Figure 3.

2. Regarding comment #4 in my previous review, I understand that the metric used to evaluate MS2 matches differs between the MRN algorithm and the validation, as the former compares molecules of different mass, while the latter performs spectral matching. Still, it seems that there could be some circularity / bias here, since both predict metabolite identity based on MS2 similarity. For example, say we have two structurally related metabolites A, B and two LCMS features X, Y with similar MS2

spectra where X is known to be metabolite A while Y is unknown; and say MetDNA3 makes the connection X-Y and predicts Y to be metabolite B. Even if this prediction is wrong, it is likely that we can successfully "validate" it by matching the spectra Y to a reference spectrum of compound B, because Y was selected by similarly to X=A, and B is structurally similar to A. The underlying problem is that MS2 matching is in itself not infallible. The authors treat MS2 matching as ground truth for validation, but if we consider that MS2 matching also has an error rate, and that MetDNA3 and MS2 matching will tend to make the same errors since MetDNA3 relies heavily on MS2 information, then it seems that validating MetDNA3 by MS2 matching will overestimate accuracy. I don't know of a way to address this problem, but I think it is worth discussing this caveat in the text. Note that Scheubert et al 2017 also pointed out this issue and observed that it leads to optimistic FDR estimates.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to congratulate the authors on the great work and thank them for their politeness and effort to answer all queries. I think the manuscript is ready for publication. I would only advise the authors to be careful with their FDR approach; it still seems very weak. As I don't have much time to look at the simulations with the attention they need, I will trust the author's work.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
With this latest revision, I feel the authors have addressed all my comments. The revised manuscript text is now more balanced and appropriately points out the potential pitfalls of this complex software system, in particular the difficulty of estimating error rates in metabolite identification. I recommend the manuscript for publication.

(Remarks on code availability)

Response to the reviewers:

We sincerely thank the reviewers for their valuable comments, which significantly strengthened the manuscript. Detailed revisions were given as follows.

Reviewer #1:

General Comment: *"Zhang et al., presented the work entitled Knowledge and Data-driven Two-layer Networking for Accurate Metabolite Annotation in Untargeted Metabolomics where they describe the construction and validation of a network-based approach to perform metabolite annotations. The work is well written, and the authors provide great detail in each validation step, providing data, code and multiple supplementary materials. However, the work needs clarifications, especially in the benchmarking and validation sections, to enable the users to better understand the efficiency of the method and how to better employ it."*

Response: We sincerely appreciate the reviewer's positive feedback. In response to the reviewer's concern about the clarity of the benchmarking and validation sections, we have revised these parts accordingly. We believe these changes have substantially improved both the manuscript and the MetDNA3 tool. Please refer to the following responses and revisions for details.

The three main issues I believe should be addressed are:

Comment #1: *"1) Benchmarking: from what I understood, the authors used a "concordance" of the tools to annotate 4302 features observed in multiple datasets. As all tools provide in silico predictions, one can not rule out that all predictions are wrong. The ideal ground truth for benchmarking would be to compare the accuracy of all tools for standard annotation. This has to be done carefully, taking into account random seed metabolites for MetDNA3 and comparing the remaining metabolites accordingly for each tool. From the methods section, I could not understand what structure databases were used to search with the in silico tools, this should be clarified and standardized for the benchmark."*

Response: Thank you for the reviewer's thoughtful comments on our benchmarking strategy. In our study, we evaluated the accuracy of MetDNA3 using two complementary approaches: (1) for annotations with chemical standards, we withheld 70% for annotation and validation with MetDNA3 (**Figures 3d-g**); (2) for annotations without chemical standards, we applied a cross-tool concordance strategy (**Figures 4a-c**). Features annotated consistently by MetDNA3 and multiple in-silico tools were considered reliable. Your interpretation of "concordance" is correct. As all tools provide only in-silico predictions in the absence of standards, it remains possible that all predictions are incorrect. We fully agree that such cross-validation may overestimate performance.

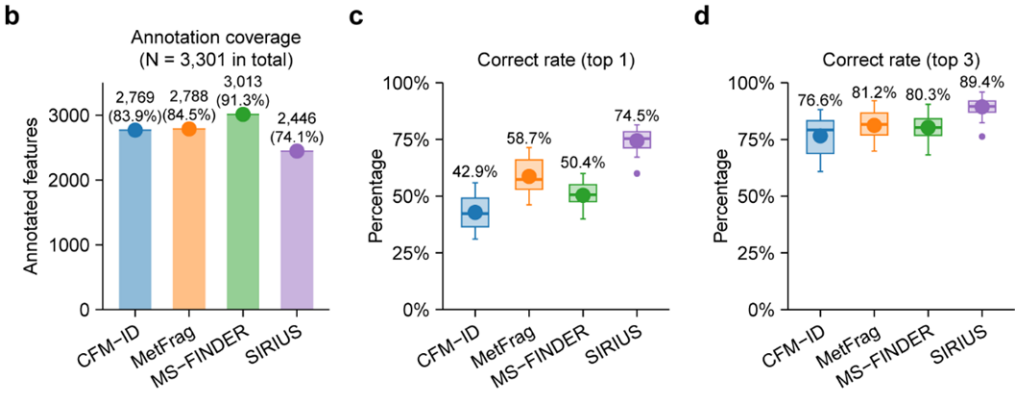
To address this concern, we followed your suggestion and evaluated the accuracy of multiple in-silico tools using features annotated by chemical standards (data from Figure 3). Specifically, we used the validation metabolites (N = 3,301) from Figure 3 as a benchmark set to assess the performance of CFM-ID, MetFrag, MS-FINDER, and SIRIUS (**Supplementary Figure 12**). Most tools showed high

correct rates, with Top 3 annotation rates ranging from 76.6% to 89.4% (**Supplementary Figure 12d**). We note that these tools may have included part of validation metabolites in their training, potentially contributing to the high accuracy. **Together, while cross-tool concordance may overestimate performance due to the lack of chemical standards, it remains a reasonable and practical approach for assessing annotation reliability.**

We have added this point into the discussion of the revised manuscript as follows:

*"To evaluate the accuracy of these methods, we used the validation metabolites (N = 3,301) from Figure 3 as a benchmark set to assess the performance of CFM-ID, MetFrag, MS-FINDER, and SIRIUS (**Supplementary Figure 12**). Most tools showed high correction rates, with Top 3 annotation rates ranging from 76.6% to 89.4%. We note that these tools may have included part of validation metabolites in their training, potentially contributing to the high accuracy. Together, while cross-tool concordance may overestimate performance due to the lack of chemical standards, it remains a reasonable and practical approach for assessing annotation reliability."*

[editorial note: panel redacted]



Supplementary Figure 12. Evaluation of multiple in-silico tools used for cross-tool concordance. (**a**) Workflow for performance validation. (**b**) Annotation coverage. (**c**) Correct rate for Top 1 annotations. (**d**) Correct rate for Top 3 annotations.

Regarding the structure databases used in benchmarking, we clarify that all in-silico tools, including MetDNA3, were configured to search against **a standardized compound database of 53,583 unique metabolites**, integrated from KEGG, MetaCyc, and HMDB. This ensures consistency and fairness across all tools. We have updated the Methods section to reflect this clarification:

*"**Corroborations of annotated metabolites using other bioinformatic tools**. Four bioinformatic tools were employed to validate the annotated metabolites by MetDNA3, including CFM-ID (version 4.0), MetFrag (version 2.4.6-CL), MS-FINDER (version 3.61), and SIRIUS (version 6.0.6). A standardized compound database of 53,583 unique metabolites, integrated from KEGG, MetaCyc, and HMDB, used for annotation is the same as MetDNA3. Data formatting and parameter settings were adjusted according to the specifications of each tool. Detailed parameter configurations are provided in the **Supplementary Table 4**."*

Comment #2: *"2) I could not understand the rationale behind the validation of the metabolites γ-Glu-Thr-Gly and N-glycolyltaurine out of thousands of predictions. That is to say, how are the users going to be able to avoid false positives? Would be much believable if the authors explained a filtering strategy (maybe I missed it in my reading) or tried to validate a large number of predictions, thus estimating the false positive rates.*

This is also a general concern with the author's use of the words 'confidece' and 'significance', maybe this wording should be used with more caution, as the 'Knowledge layer' was expanded with predictions not validated experimentally, and the matching provided with mass spectrometry data is inherently subjected to errors. The author should consider in future versions having more appropriate estimates of significance, in the lines of

Scheubert, K., Hufsky, F., Petras, D. et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8, 1494 (2017). https://doi.org/10.1038/s41467-017-01318-5

Palmer, A., Phapale, P., Chernyavsky, I. et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods 14, 57–60 (2017). https://doi.org/10.1038/nmeth.4072"

Response: Thank you for the insightful comment. In our study, we employed a stepwise filtering strategy to prioritize unknown candidates for validation. First, we applied cross-validation using multiple in-silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, we prioritized candidates recurrent across multiple biological datasets, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, we manually inspected the MS2 spectra to select reliable candidates. Using this strategy, we successfully validated two unknown metabolites not found in current human metabolome databases. While not exhaustive, this demonstrates the potential of MetDNA3 for novel metabolite discovery.

We have added the following description to the section '**Corroborations of annotated metabolites using other bioinformatic tools**' in the Methods of the revised manuscript:

"A stepwise filtering strategy was applied to prioritize unknown candidates for validation. First, cross-validation was performed using multiple in silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, candidates recurrently detected across multiple biological datasets were then prioritized, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, MS2 spectra were manually inspected to select reliable candidates."

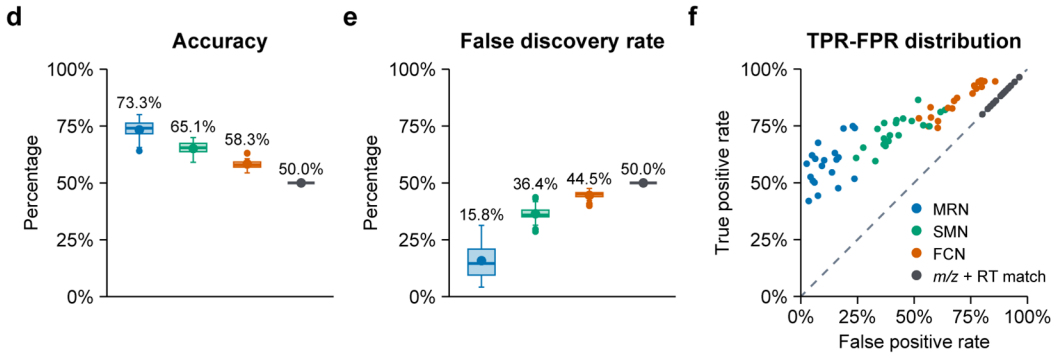
Second, we sincerely thank the reviewer for the suggestion to estimate the false discovery rate (FDR) for network-based annotation in MetDNA3. Following your recommendation, we adopted the spectrum-based method proposed by Scheubert et al. (**Nat. Commun., 2017, 8:1494**). Specifically, we generated decoy MS2 spectra by randomizing the experimental MS2 spectra (**Supplementary Figure 11**), as described in the original study, and incorporated them into the MetDNA3 data layer for annotation

propagation. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, we required that the dot product similarity between the decoy MS2 and the original MS2 be less than 0.5. The resulting annotations from decoy spectra served as negative controls for evaluating false positives. Based on generated decoy MS2 spectra, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR of 15.8% (see **revised Figure 5c-e**). This accuracy is comparable to the validation using chemical standards shown in Figure 3, which reported 84.4% accuracy for Top 3 annotations. To the best of our knowledge, no existing method has employed decoy MS2 spectra to evaluate false positives in network-based annotation propagation. We hope this approach provides valuable insights for improving future network-based annotation strategies.

[editorial note: figure redacted]

Supplementary Figure 11. Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation. The workflow of experimental and decoy MS2 spectra-based annotation propagation using MetDNA3. Decoy MS2 spectra were generated by spectrum-based method proposed by Scheubert et al. Details were described in the Methods section.

[editorial note: panel redacted]



Revised Figure 5c-f. Benchmarking different knowledge-driven network topologies. (c) Schematic workflow for evaluating annotation performances using experimental and decoy MS2 spectra. **(d-f)** Comparative evaluations of accuracy (d), false discovery rate (FDR) (e), and true positive rate (TPR)-false positive rate (FPR) distribution (f) for different networks in network-based annotation propagation. MRN, metabolic reaction network; SMN, structure-guided molecular network; FCN, fully connected network.

Third, to demonstrate the superiority of our curated MRN for metabolite annotation (requested by **Rev#3**), we used the same decoy MS2 spectra to evaluate the accuracy and FDR of two additional knowledge-driven networks with distinct topologies: a structure-guided molecular network (SMN) and a fully connected network (FCN). The results showed that as network connectivity increased (from MRN to SMN to FCN), annotation accuracy decreased from 73.3% to 65.1% and then to 58.3% (**revised Figure 5d**). Correspondingly, the FDR increased from 15.8% to 36.4% and 44.5% (**revised Figure 5e**). For comparison, annotation based solely on *m/z* + RT matching with decoy MS2 spectra yielded 50% accuracy and 50% FDR. The performance of the FCN closely resembled that of *m/z* + RT matching, suggesting that FCN-based annotation relied mainly on *m/z* + RT matches rather than network-based propagation. From the distribution of TPR and FPR, it is evident that the MRN exhibits both lower TPR and FPR compared to other methods (**revised Figure 5f**). These results highlight that indiscriminately increasing network connectivity does not improve annotation quality. Instead, carefully curating knowledge networks with high specificity is essential for accurate network-based metabolite annotation and propagation.

We have updated the manuscript to include the new data and methodological details as outlined below.
1) Revision in the Methods: *"**Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation.** Decoy MS2 spectra were generated from experimental MS2 spectra by spectrum-based method proposed by Scheubert et al.[39], and incorporated them into the MetDNA3 data layer for annotation propagation. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, the dot product similarity between the decoy MS2 and the original MS2 was required less than 0.5. For both experimental and decoy MS2 spectra, 30% of them were used as seed metabolites to initiate propagation, and the remaining 70% were used to validate the performance of network-based annotation propagation. The annotation results were classified as true positives (TP) if annotated within rank ≤ 3, and false negatives (FN) otherwise. For decoy MS2 spectra, the same propagation process was applied, where the results are considered false positives (FP) if annotated within rank ≤ 3, and true negatives (TN) otherwise."*

2) Revision in main text related to new Figure 5: *"Next, we adopted the spectrum-based method proposed by Scheubert et al.[39] to generate decoy MS2 spectra by randomizing the experimental MS2 spectra (**Supplementary Figure 11**). The resulting annotations from decoy spectra served as negative controls for evaluating annotation accuracy and false discovery rate (FDR) of network-based annotation propagation (**Figure 5c and Supplementary Data 4**). Based on generated decoy MS2 spectra, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR of 15.8% (**Figure 5d and 5e)**. In addition, we used the same decoy MS2 spectra to evaluate the performances of SMN and FCN. The results showed that as network connectivity increased (from MRN to SMN to FCN), annotation accuracy decreased from 73.3% to 65.1% and then to 58.3% (**Figure 5d**). Correspondingly, the FDR increased from 15.8% to 36.4% and 44.5% (**Figure 5e**). The performance of FCN closely resembled that of m/z+RT matching, suggesting that FCN-based annotation was primarily driven by m/z+RT matches rather than network-based propagation. From the distribution of TPR and FPR, it is evident that our curated MRN exhibits both lower TPR and FPR compared to other methods*

*(**Figure 5f**). These results highlight that indiscriminately increasing network connectivity does not improve annotation quality. Instead, carefully curating knowledge networks with high specificity is essential for accurate network-based metabolite annotation and propagation."*

3) Regarding the use of the terms "confidence" and "significance," we acknowledge the reviewer's point and have revised the manuscript to minimize their usage. Additionally, we have clarified that the knowledge layer is based on predictions and has not yet been experimentally validated in the main text as follows:

"Notably, the curated MRN was constructed based on predictive models and has not yet been experimentally validated. Its primary purpose is to enhance network-based annotation propagation rather than to directly discover novel reactions."

Comment #3: *"3) Is MetDNA3 designed for animal cells/tissues only? Sorry if that was implicit. If yes, change the title accordingly. I believe the need of clarification stems from the fact that transformations provided by BioTransformer have what the authors called 'Human Super Transformer' component, and if the transformations mapped are indeed biased for animals or humans, users could increase false positives trying to use it for other organisms."*

Response: Thank you for the insightful comment. In MetDNA3, the known metabolite database integrates KEGG, MetaCyc, and HMDB, covering a broad range of common metabolites across species. However, the generation of unknown metabolites relies on BioTransformer's default pipeline, which includes the "Human Super Transformer" module. Consequently, the predicted metabolites may indeed be biased toward human or mammalian metabolism. While the methodological framework of MetDNA3 is generally applicable to other organisms such as plants, bacteria, and fungi, users are encouraged to customize the unknown metabolite database and metabolic reaction network according to their specific needs.

We have added this point in the discussion and recommend that users interpret the results with consideration of the biological context of their target organism. Detailed revisions in the revised manuscript were given as follows:

"Meanwhile, BioTransformer-based generation of unknown metabolites may be limited by its predefined reaction types, potentially introducing species-specific biases. For example, in MetDNA, unknown metabolite generation relied heavily on mammalian-derived reactions, which may bias results toward human or mammalian metabolism. We recommend users interpret annotation results with caution and consider constructing customized unknown metabolite databases and metabolic reaction networks tailored to their specific biological context."

Comment #4: *"Another issue is that although the code, data, and web interface are available for review purposes, it would be important to provide a worked example of data pre-processing and metabolite annotation to facilitate the review process. I tried to install the package on standard Google Colab*

(https://colab.research.google.com/notebook#create=true&language=r) following the instructions of the github page and was unable to install it. The option to subscribe to the web tool and try it could breach the anonymity of the review.

“installation of package ‘/tmp/RtmpW0b815/file1075f20d1e4/MrnAnnoAlgo3_3.1.1.tar.gz’ had non-zero exit status”

>session_info()

$platform

$version

'R version 4.4.3 (2025-02-28)'

$os

'Ubuntu 22.04.4 LTS'

$system

'x86_64, linux-gnu'

$ui

'X11'

$language

'en_US'

$collate

'en_US.UTF-8'

$ctype

'en_US.UTF-8'

$tz

'Etc/UTC'

$date

'2025-04-15'

$pandoc

'NA'

$quarto

Are the code and model for the graph neural network available?”

Response: We sincerely thank the reviewers for their valuable suggestions and for taking the time to test our software and code. MetDNA3 includes not only the core algorithm but also supporting modules

such as the metabolite database and spectrum matching, some of which cannot be made publicly available due to intellectual property constraints. To support transparency, we have uploaded the core algorithm code, specifically the two-layer interactive network topology for recursive metabolite annotation, to GitHub (https://github.com/ZhuMetLab/MrnAnnoAlgo3). The full functionality of MetDNA3 remains accessible via our website (http://metdna.zhulab.cn/). For anonymous access during the review process, we have created a dedicated test account:

Username: metdna

Password: metdna@123

This account allows reviewers to explore the platform and validate key functionalities without registration. Demo datasets are available for download from Zenodo repository (https://doi.org/10.5281/zenodo.15589872). Detailed instructions are also provided alongside the demo datasets.

Additionally, for the graph neural network module, we have made the model and full source code publicly available on GitHub (https://github.com/ZhuMetLab/ReactionPredictor), along with detailed instructions. The related information has been added into the "Code Availability" section of the revised manuscript:

"The algorithm of two-layer networking topology was mainly developed using R and is executed in MetDNA3. The source code was provided in GitHub [https://github.com/ZhuMetLab/MrnAnnoAlgo3]. The completed functions are provided in the MetDNA3 webserver [http://metdna.zhulab.cn/] via a free registration. The GNN-based model for predicting of reaction relationships was developed in Python. The source code was provided in GitHub [https://github.com/ZhuMetLab/ReactionPredictor]."

Minor issues

Comment #5: *"a) What was the chemical taxonomy criterion used to classify the compounds? Was it that of ClassyFire (Djoumbou Feunang et al., 2016)?"*

Response: We thank the reviewer for this comment. Yes, we used ClassyFire to perform chemical taxonomy classification of the compounds, and we have added this information and the citation to the captions of Figure 5 and Supplementary Figure 8 for clarity:

*"**Revised Figure 5.** (b) Proportions of metabolites propagated within the same metabolite superclass for different network-based metabolite annotation. Superclass taxonomy was calculated using ClassyFire."*

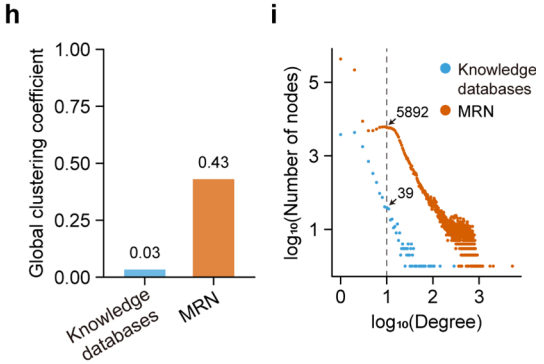
*"**Revised Supplementary Figure 8.** (d, i) Superclass distribution of annotated metabolites in HILIC–MS(+) (d) and HILIC–MS(-) (i). Superclass taxonomy was calculated using ClassyFire."*

Comment #6: *"b) From the following statement, between lines 120-121: "Additionally, evaluation of the topological properties of the curated MRN revealed a higher global clustering coefficient and an improved*

degree distribution relative to knowledge databases". What is the global clustering coefficient based on and how was it obtained, as well as the degree distribution relative to knowledge databases? i) is the number of nodes that have a given degree? Legend or text could be more descriptive."

Response: We thank the reviewer for this insightful question. To clarify, the global clustering coefficient measures the overall interconnectedness of a network. It is calculated as the ratio of the number of closed triplets to the total number of triplets in the network. A triplet refers to a group of three nodes connected by at least two edges (Wasserman and Faust, Social Network Analysis, Cambridge University Press, 1994; https://doi.org/10.1017/CBO9780511815478). This metric was computed using the transitivity() function from the igraph R package, under the "global" setting.

The degree of a node refers to the number of directly connected neighbor metabolites. Your understanding of degree distribution is correct; it represents the number of nodes in the network corresponding to a given degree value. For example, in our curated MRN, 5,892 metabolites (nodes) have a degree of 10, whereas the knowledge databases-based network has only 39 such nodes. As shown in **revised Figure 1i**, the degree distribution clearly illustrates that our curated MRN has significantly higher connectivity than the network constructed from knowledge databases.



Revised Figure 1. (h-i) Global clustering coefficient of network (h), and degree distribution of nodes (i) in knowledge databases and curated MRN. The global clustering coefficient is calculated as the ratio of the number of closed triplets to the total number of connected triplets within the network. The degree of a node refers to the number of directly connected neighbor metabolites.

In the revised manuscript, we have updated the **Figure 1i** and corresponding text to provide clearer explanations. Additionally, we have clarified the definitions of the global clustering coefficient and degree distribution in the "**Calculation of network topological properties**" section of the Methods. The detailed revisions are summarized as follows:

1) Revisions in the main text: *"Additionally, evaluation of the topological properties of the curated MRN revealed a higher global clustering coefficient and an improved degree distribution relative to knowledge databases (**Figures 1h and 1i, and Supplementary Table 1**). Degree distribution describes the number of nodes in the network corresponding to a given degree value. For example, in MRN, 5,892 metabolites (nodes) have a degree of 10, whereas the knowledge databases-based network has only 39 such nodes (**Figures 1i**)."*

2) Revisions in the Methods: *"**Calculation of network topological properties.** To characterize the structural and functional features of the networks, a series of topological properties were computed using the R package igraph (version 2.0.3) and Python package networkit (version 11.1). These properties were categorized into three groups: Information, Connectivity, and Community. Information-level metrics include the number of nodes and edges, the number of connected components, and the number of nodes in the largest connected component. Network components were obtained using the components() function in R. Connectivity-level metrics include average degree, network density, global clustering coefficient, average shortest path length, and network diameter. Node degree was computed using the degree() function, and the average degree was calculated as the mean degree across all nodes. Network density was computed using the edge_density() function in R. The global clustering coefficient was calculated as the ratio of the number of closed triplets to the total number of triplets in the network, where a triplet is defined as three nodes connected by at least two edges. It was calculated using the transitivity() function with the setting type="global" in R. The average shortest path length was computed using the mean_distance() function, and the network diameter was obtained via the diameter() function in R. Community-level metrics were derived using the Louvain method, a modularity-based community detection algorithm, implemented via networkit.community.detectCommunities() function in Python."*

Comment #7: *"c) As shown between the lines 140-141: "MS2 similarity between features was calculated and applied as a filtering constraint to eliminate unwanted nodes, further refining the network structure." What metric was used to determine the similarity between the spectra in MS2? Cosine score, modified cosine score, spectral entropy?"*

Response: We thank the reviewer for this valuable question. A modified dot-product function was used to calculate MS2 spectral similarity between the seed feature and tis neighbor feature. Specifically, when the precursor *m/z* of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with *m/z* values exceeding that of the neighbor precursor are excluded. Conversely, when the neighbor has a higher precursor *m/z*, fragment ions in the MS2 spectrum with *m/z* values greater than that of the seed are removed. This spectral similarity calculation followed the same method as described in our previous MetDNA publication (Shen et al., *Nat. Commun.*, 2019, https://doi.org/10.1038/s41467-019-09550-x).

We have clarified this point in "**Two-layer interactive networking topology strategy implemented in MetDNA3**" section of Methods the revised manuscript as follows:

"Then, the MS2 spectral similarity between linked features in the feature network was calculated and applied as a constraint (dot product score ≥ 0.5), resulting in a knowledge-constrained feature network. A modified dot-product function was used to calculate MS2 spectral similarity and followed the same method as described in our previous MetDNA publication[19]. Specifically, when the precursor m/z of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with m/z values exceeding that of the neighbor are excluded. Conversely, when the neighbor has a higher precursor m/z, fragment ions in the MS2 spectrum with m/z values greater than that of the seed are

removed. The resulting MS2 spectral similarity-based edge relationships were then mapped back to the knowledge layer, generating a data-constrained MRN."

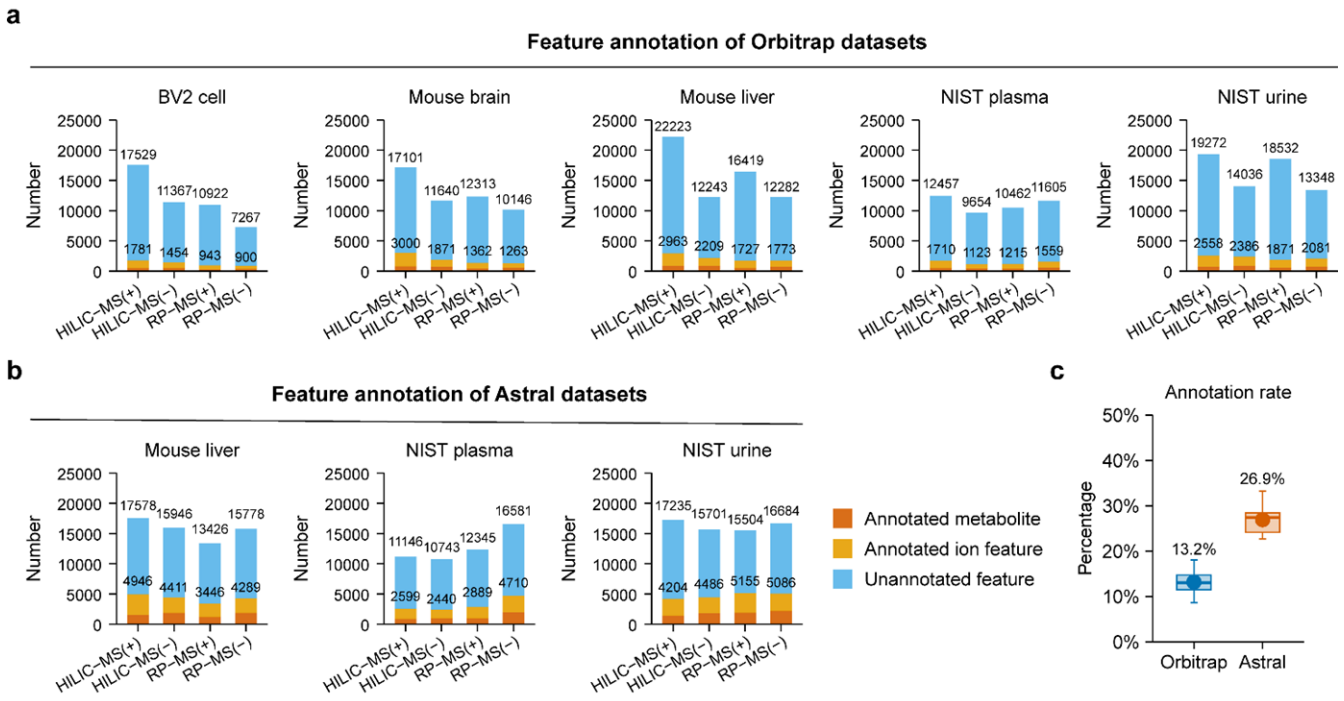
Comment #8: *"d) As shown between the lines 206-208: "Furthermore, network-based annotation propagation expanded the annotated metabolome to 12,508 metabolites (level 3 annotations), including 9,410 known and 3,098 unknown metabolites (Figure 3c), highlighting the substantial increase in metabolite coverage." What is the annotation rate in terms of the total number of features captured in pre-processing for both Orbitrap and Astral data?"*

Response: We appreciate the reviewer's thoughtful comment. We have calculated the annotation rate, defined as the proportion of annotated features relative to the total number of detected features obtained from data pre-processing. These results are now presented in **Supplementary Figure 6**. Specifically, in the Orbitrap datasets, we detected between 7,267 and 22,223 features (median = 12,298) in different biological samples. Among these, 900 to 3,000 features (median = 1,750) were annotated using MetDNA3 (**Supplementary Figure 6a**), resulting in an average annotation rate of 13.2% (**Supplementary Figure 6c**). In the Astral datasets, we detected between 10,743 and 17,578 features (median = 15,740) in different biological samples, among which 2,440 to 5,155 features (median = 4,350) were annotated using MetDNA3 (**Supplementary Figure 6b**), yielding an average annotation rate of 26.9% (**Supplementary Figure 6c**). Notably, the total number of features detected in the Orbitrap and Astral datasets was comparable; however, the Astral instrument's higher MS2 acquisition coverage led to substantially improved annotation coverage.

The feature annotations in MetDNA3 include both the identified metabolites and their corresponding ion features, such as isotopes, neutral losses, adducts, and in-source fragments. In general, metabolite features were first annotated through network-based propagation. Subsequently, the global peak correlation network, originally developed in MetDNA2 (Zhou et al., ***Nat. Commun.***, 2022, https://doi.org/10.1038/s41467-022-34537-6), was employed to further annotate the various ion features associated with each metabolite. This approach uses chromatographic coelution correlations to recognize related ion features, including adducts, isotopes, neutral losses, and in-source fragments.

In the revised manuscript, we have updated the **Supplementary Figure 6** and corresponding text to describe annotation rate as follows:

*"Detailed metabolite annotations for all samples were provided in **Supplementary Figures 5 and 6**, and **Supplementary Data 1**. Notably, data acquired using the Astral instrument demonstrated superior annotation coverage compared to the Orbitrap (**Figures 3b and 3c**). Additionally, the Astral data exhibited a significantly higher annotation rate (26.9%) compared to the Orbitrap (13.2%) (**Supplementary Figure 6**)."*



Supplementary Figure 6. Feature annotation in Orbitrap and Astral datasets using MetDNA3. (**a**) Feature annotations for Orbitrap datasets. (**b**) Feature annotations for Astral datasets. (**c**) Feature annotation rate, defined as the percentage of annotated features (including annotated metabolites and ion features) relative to the total number of detected features. HILIC–MS(+): positive MS mode with HILIC separation; HILIC–MS(-): negative MS mode with HILIC separation; RP–MS(+): positive MS mode with RP separation; RP–MS(-): negative MS mode with RP separation.

Comment #9: *"e) Do the authors intend to submit the fragmentation spectra of the new compounds characterized in this work to the databases used by the metabolomics community, such as MoNA?"*

Response: We thank the reviewer for the suggestion. We have submitted the MS2 spectra of the newly characterized compounds to the MoNA database (https://mona.fiehnlab.ucdavis.edu/). The deposited spectra can be found under the IDs from MoNA_0003569 to MoNA_0003577.

We have also included this information in the "Data Availability" section of the revised manuscript as follows:

"The MS2 spectra of the newly characterized compounds can be accessed at MoNA database [https://mona.fiehnlab.ucdavis.edu/] via IDs from MoNA_0003569 to MoNA_0003577."

Comment #10: *"f) Do the authors consider submitting the raw data do metabolights, metabolomics workbench or gnps?"*

Response: We thank the reviewer for the helpful suggestion. We have deposited the raw data in the MassIVE repository (https://massive.ucsd.edu/). The datasets are available under the accession IDs MSV000097913 and MSV000097914.

We have also included this information in the "Data Availability" section of the revised manuscript as follows:

"The raw data files of BV2 cells, mouse brain tissue, mouse liver tissue, NIST human plasma, and NIST human urine from the Orbitrap instrument can be accessed at the MassIVE repository [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000097913] or the National Omics Data Encyclopedia under Accession Code OEP00006095 [https://www.biosino.org/node/project/detail/OEP00006095]. The raw data files of mouse liver tissue, NIST human plasma, and NIST human urine from the Astral instrument can be accessed at the MassIVE repository [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000097914] or National Omics Data Encyclopedia under Accession Code OEP00006083 [https://www.biosino.org/node/project/detail/OEP00006083]."

Reviewer #2:

"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

Response: We appreciate Nature Communications' initiative to involve and recognize Early Career Researchers in peer review. We sincerely thank the reviewer for the valuable comments, which have greatly strengthened the manuscript.

Reviewer #3:

General Comment: *"In this manuscript, Zhang et al report a new computational algorithm, MetDNA3, for metabolite identification in LC-MS mass spectrometry metabolomics. The work is a further development of the authors' previous systems, MetDNA and KGMN / MetDNA2, and the recursive algorithm used to annotate features appear to be similar. A novel feature in this version is a deep learning method for extending known metabolic networks by additional predicted reactions. The methodology is interesting, and I appreciated the experimental validation of several predicted metabolites suggesting that MetDNA3 does enable real discoveries. This new version also seems to improve substantially upon MetDNA2, with increased coverage and accuracy. However, the systematic evaluation of the method is not easy to understand, and some of the accuracy metrics used might exaggerate the performance of the system. The presentation and some of the method details needs clarification, as detailed below."*

Response: We sincerely thank the reviewer for the constructive evaluation of our work. We are pleased that the methodological improvements and experimental validations were found valuable. In response to the reviewer's concerns about clarity, we have made substantial revisions to improve the presentation and provide clearer explanations. Please refer to the following responses and revisions for details.

*** Major points ***

Comment #1: *"1. Regarding accuracy of the metabolic reaction network (MRN; Fig 1c), the reported AUC of 0.9947 might give readers the impression that the MRN is almost free of errors, but I am not convinced that AUC is the relevant measure here. In their cross-validation scheme, the authors choose ~15k true reaction pairs and ~15k non-pairs (random pairs) among 11,532 metabolites. From Fig 1c it looks like the MRN achieves 100% TP rate at about 1% FP rate (false positives / negatives), but the classification problem is extremely unbalanced. There are 66M pairs total among the 11,532 metabolites, most of which are likely to be negatives, so with 1% FP rate one would expect a total of ~660k predicted pairs in the cross-validated MRN (I did not find the actual number), of which 15k were true positives, which means the the false discovery rate (false positives / predicted positives) could be as high as 97%.*

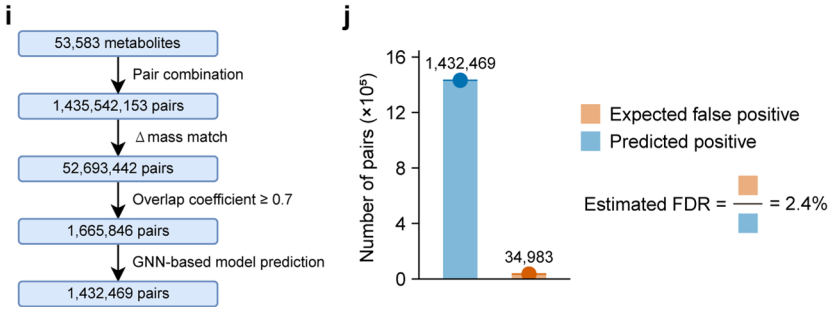
Although FDR estimated this way is conservative since the known network is incomplete, it would seem that most of the pairs in the large MRN are likely to be false positives. In this context, citing "improvements of 66.4-fold and 162.8-fold" seems misleading. Certainly the MRN has higher coverage, but likely at the cost of high FDR. I think the authors should provide the FDR as well, so that the uninitiated reader can understand the amount of false positives in the MRN. I would appreciate a discussion of these points."

Response: We thank the reviewer for the insightful comment and apologize for the insufficient details on metabolic reaction network (MRN) construction in the original manuscript, which may have led to misunderstanding. Here, we provide a more detailed description of our workflow for constructing the MRN. The metabolite knowledge databases contain a total of 53,583 metabolites, and pairing every two metabolites generates 1,435,542,153 (1,435 million) possible metabolite pairs. As the reviewer pointed

out, the vast majority of these pairs are likely non-reaction pairs (non-RPs), and using the full set for prediction would indeed result in a very high number of false positives.

To address this issue, in our work, we implemented a two-step pre-screening strategy to substantially reduce the number of metabolite pairs input into the prediction model. First, we matched the mass difference (Δmass) between two metabolites to the frequently observed mass differences derived from 14,974 known reactions recorded in the metabolite knowledge databases (KEGG and MetCyc), encompassing 2,045 unique Δmass values. Only metabolite pairs with matched Δmass values were retained. This filtering step reduced the number of candidate pairs to 52,963,442 (52.9 million). Next, we further filtered these pairs by the structural overlap coefficient for each pair (calculated by RDKit). Pairs with a coefficient ≥ 0.7 were retained, resulting in a final set of 1,665,846 (1.66 million) pairs. These pairs were then used as input for the graph neural network (GNN)-based model to predict reaction relationships. Among them, 1,432,469 metabolite pairs were predicted as possible reaction pairs and used to curate the MRN. The detailed workflow is summarized in **new Supplementary Figure 1i**, and corresponding descriptions have been incorporated into the Methods section of the revised manuscript.

According to the results shown in **Supplementary Figure 1h**, our GNN-based prediction model exhibited a false positive rate (FPR) of 2.1% in the testing set. Based on this value, the expected number of false positives among the 1,665,846 candidate metabolite pairs is 34,983 (1,665,846 × 2.1% = 34,983). As indicated above, in the final MRN, we retained 1,432,469 predicted reaction pairs. Accordingly, the estimated false discovery rate (FDR) is 2.4% (34,983 / 1,432,469 = 2.4%; **Supplementary Figure 1j)**. Taken altogether, our two-step pre-screening strategy significantly reduced the number of metabolite pairs input into the GNN-based prediction, resulting in **an estimated FDR of only 2.4%** for the predicted reaction pairs.



Revised Supplementary Figure 1i and 1j. (**i**) Two-step pre-screening workflow to reduce metabolite pairs input into the GNN-based prediction. (**j**) Estimated false discovery rate (FDR) of the predicted reaction pairs in the curated MRN.

The detailed revisions are summarized as follows:

1) Revisions in the main text: *"To control potential false positives, a two-step pre-screening strategy was applied prior to the prediction, resulting in an estimated false discovery rate (FDR) of 2.4% (**Supplementary Figures 1i and 1j**)."*

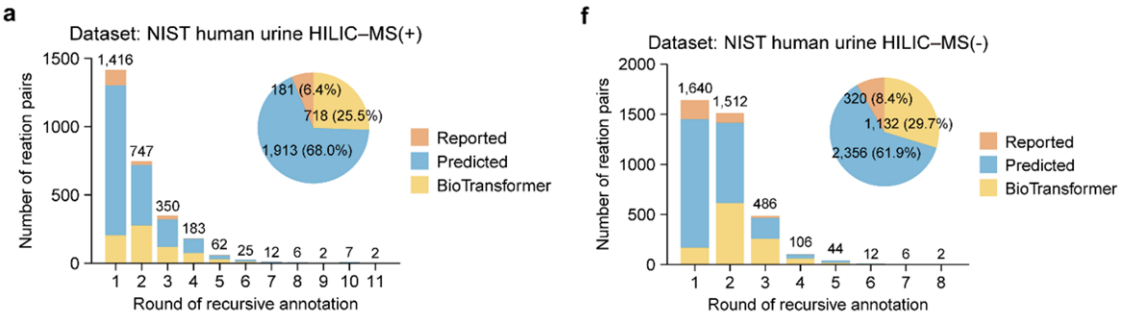
2) Revisions in the Methods of "Curation of knowledge-based metabolic reaction network": *"Specifically, we applied this model to predict potential reaction pairs within a metabolite database*

*containing 53,583 metabolites. Exhaustively pairing all metabolite combinations would yield 1,435,542,153 (1.44 billion) possible pairs. However, most of these are likely non-reaction pairs (non-RPs), and direct prediction on this full set would result in a high number of false positives. To address this, we implemented a two-step pre-screening strategy to significantly reduce the number of candidate metabolite pairs (**Supplementary Figure 1i**). First, we matched the mass difference (Δmass) between two metabolites to the frequently observed mass differences derived from 14,974 reported reaction pairs in the metabolite knowledge databases (KEGG and MetCyc), encompassing 2,045 unique Δmass values. Only metabolite pairs with matched Δmass values were retained. This filtering step reduced the number of candidate pairs to 52,963,442 (52.9 million). Second, we further filtered these pairs by the structural overlap coefficient for each pair (calculated by RDKit). Pairs with a coefficient ≥ 0.7 were retained, resulting in a final set of 1,665,846 (1.66 million) pairs. These pairs were then used as input for the graph neural network (GNN)-based model to predict reaction relationships. Among them, 1,432,469 metabolite pairs were predicted as possible reaction pairs. According to model evaluation on the testing set (**Supplementary Figure 1h**), the GNN exhibited a false positive rate (FPR) of 2.1%. Therefore, among the 1,665,846 candidate pairs, the expected number of false positives is estimated to be 34,983 (1,665,846 × 2.1%). In the final MRN, we retained 1,432,469 predicted reaction pairs. Accordingly, the estimated false discovery rate (FDR) is 2.4% (34,983 / 1,432,469 = 2.4%; **Supplementary Figure 1j**). Finally, we combined the 1,432,469 GNN-predicted reaction pairs with 14,974 reported reaction pairs from databases, resulting in a curated MRN comprised a total of 47,748 metabolites and 1,447,443 reaction pairs, including both those from knowledge databases and those predicted by the GNN model (**Supplementary Table 1**)."*

Comment #2: *"2. In light of the above, I also wonder how to interpret the observation that when reconciling the MRN with MS2-based molecular networks (the "two-layer topology", Fig 2), most of the MRN is discarded: only 2.3% of reactions are kept, which is close to my rough FDR estimate of 97% above. Is it the case that most of the novel, GNN-predicted part of the MRN was discarded? If so, I wonder to what extent the MRN contributes to the final results. It would be useful to see a direct comparison of the method performance (as in Fig 5d) against (i) using only the knowledge databases instead of the full MRN, and (ii) using only the MS2-based network (molecular networking), to better understand the effect of the "knowledge layer"."*

Response: We thank the reviewer for the thoughtful insights. As demonstrated in our response to Comment #1, the estimated FDR of our MRN is approximately 2.4%. Indeed, a large proportion of reaction pairs in the MRN were filtered out during integration with the two-layer network topology. This outcome is primarily due to two reasons: (1) many metabolite nodes in the MRN were not detected in the experimental dataset; and (2) for detected metabolites, the corresponding features often did not meet the MS2 similarity threshold required for network construction. Therefore, the low retention rate of 2.3% does not imply a high FDR or poor quality of the predicted MRN. Rather, it reflects the selective activation of MRN components by the experimental data under stringent matching criteria.

Despite the small fraction of predicted reactions retained, they play a substantial role in annotation propagation. As shown in our original manuscript (**Supplementary Figures 6a and 6f**), 68.0% and 61.9% of propagated annotations were supported by GNN-predicted reaction pairs, underscoring their significant contribution to the overall annotation performance. For your reference, we have included the two figures again below.



Copy of Supplementary Figure 6a and 6f. Characterization of recursive-based metabolite annotation and propagation. (**a, f**) Distributions and proportions of reported, predicted, and BioTransformer-generated reaction pairs during annotation propagation in HILIC–MS(+) (a) and HILIC–MS(-) (f). Human urine samples were tested (n=6 technical replicates).

We sincerely appreciate the reviewer's suggestion for a comparison of method performances among our curated MRN, knowledge databases-based network, and the MS2 similarity-based network. However, in the original Figure 5c–d, the validation dataset was highly imbalanced, containing 3,275 true positive cases and 28,508 false cases. Under such conditions, a naïve method predicting all structures as negatives would yield an accuracy of 89.7%, which is clearly misleading.

To address this, we revised the manuscript by following Reviewer #1's suggestion and constructed decoy MS2 spectra using the spectrum-based method reported by Scheubert et al. (***Nat. Commun.***, 2017, https://www.nature.com/articles/s41467-017-01318-5) to more appropriately evaluate the FDR values of each network method.

Specifically, we generated decoy MS2 spectra by randomizing the experimental MS2 spectra, as described in the original study, and incorporated them into the MetDNA3 data layer for annotation propagation. The resulting annotations from decoy spectra served as negative controls for evaluating false positives. The complete workflow is shown in **Supplementary Figure 11** and detailed in the revised Methods section. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, we required that the dot product similarity between the decoy MS2 and the original MS2 be less than 0.5. Based on generated decoy MS2 spectra and 20 biological datasets acquired on orbitrap instruments, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR of 15.8% (see **revised Figure 5d**). This accuracy is comparable to the validation using chemical standards shown in Figure 3, which reported 84.4% accuracy for Top 3 annotations. To the best of our knowledge, no existing method has employed decoy MS2 spectra to

evaluate false positives in network-based annotation propagation. We hope this approach provides valuable insights for improving future network-based annotation strategies.

[editorial note: figure redacted]

Supplementary Figure 11. Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation. The workflow of experimental and decoy MS2 spectra-based annotation propagation using MetDNA3. Decoy MS2 spectra were generated by spectrum-based method proposed by Scheubert et al. (Nat. Commun., 2017, 8:1494). Details were described in the Methods section.

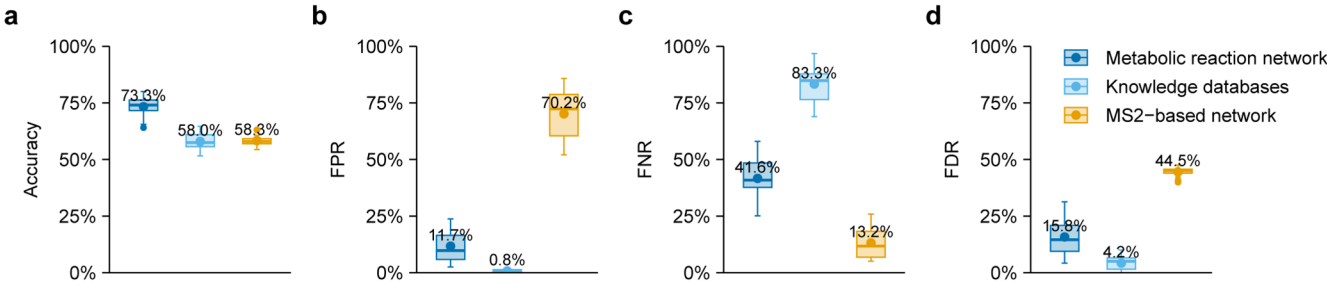


Figure R1. Decoy MS2 spectra-based performance benchmark using metabolic reaction network, knowledge databases-based network, and MS2 similarity-based network. (a-d) Comparative evaluations of accuracy (a), false positive rate (FPR) (b), false negative rate (FNR) and false discovery rate (FDR) for different networks.

Then, we used the same decoy MS2 spectra to evaluate the method performances of two additional networks: the knowledge databases-based network and the MS2 similarity-based network. As shown in **Figure R1a**, our MRN achieved higher accuracy (73.3%) compared to the knowledge databases-based network (58.8%) and the MS2 similarity-based network (58.3%). Due to limited coverage, the knowledge databases-based network maintained a very low false positive rate (FPR; 0.8%) but exhibited a high false negative rate of 83.3% (FNR; **Figure R1b and R1c**), indicating that annotation propagation was heavily constrained. In contrast, the MS2 similarity-based network showed a high FPR of 70.2% and a FNR of 13.2%, suggesting that annotation propagation without knowledge-based constraints leads to a large number of false positives (**Figure R1b and R1c**). In terms of FDR, the knowledge databases-based network achieved a very low value (4.3%), while the MS2 similarity-based network showed a high FDR of 44.5%. By comparison, our MRN reported an FDR of 15.8% (**Figure R1d**). Taken together, considering accuracy, FPR, FNR, and FDR, MRN demonstrated a more balanced performance across both positive and negative annotations.

The details of Decoy MS2 spectra-based assessment are added into the revised manuscript as follows:

*"**Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation.** Decoy MS2 spectra were generated from experimental MS2 spectra by spectrum-based method proposed by Scheubert et al.[39], and incorporated them into the MetDNA3 data layer for annotation propagation. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, the dot product similarity between the decoy MS2 and the original MS2 was required less than 0.5. For both experimental and decoy MS2 spectra, 30% of them were used as seed metabolites to initiate propagation, and the remaining 70% were used to validate the performance of network-based annotation propagation. The annotation results were classified as true positives (TP) if annotated within rank ≤ 3, and false negatives (FN) otherwise. For decoy MS2 spectra, the same propagation process was applied, where the results are considered false positives (FP) if annotated within rank ≤ 3, and true negatives (TN) otherwise."*

Comment #3: *"3. In the evaluation of metabolite annotation accuracy (which is the key performance metric) in Fig 3, the authors report "coverage" which I assume is TP/P (I think "recall" is the standard term), and accuracy defined as 1 - total error rate for "Top 3". The results seem impressive, but again it is not clear how to interpret the numbers. Specifically:*

a) How were the Top N annotations selected? I assume the authors counted the N highest-scoring putative metabolite annotations for each peak as predicted positives, but it is not clear how predicted annotations were scored -- I didn't see this described in methods?

b) Was coverage also computed using the same Top 3 method as accuracy, so that the numbers are comparable?

c) Given that the prediction problem is unbalanced, the authors should show the FP and FN rates separately in Fig 5 rather than just accuracy, so that it is clear where the improvement is coming from. Also, as discussed above, I would like to see the false discovery rate as well (with the understanding that is a conservative estimate)."

Response: We thank the reviewer for the detailed comments regarding the evaluation metrics. Your interpretation of 'coverage' in Figure 3 is not entirely accurate. In our manuscript, "coverage" refers to whether a feature receives an annotation, regardless of the correctness of that annotation. The calculation is as shown in Eq. 2:

$$Coverage = \frac{\#\ Annotated\ features}{\#\ Validated\ features} \quad (2)$$

Your understanding of "accuracy" in Figure 3 is correct. It refers to the proportion of annotated features for which the correct structure appears within the Top 3 ranked annotations. However, we noticed that the term "accuracy" was defined differently in Figure 3 and the new Figure 5 (based on decoy MS2 spectra), which may cause confusion to readers. **To avoid this confusion, we have renamed "accuracy" in Figure 3 as "correct rate"** and have clearly updated its definition in the Methods section.

Notably, the term of "correct rate" were also commonly used in MetDNA and MetDNA2 with the same definition. The calculation is as shown in Eq. 3:

$$Correct\ rate = \frac{\#\ Features\ with\ a\ correct\ structure\ in\ Top\ N\ annotations}{\#\ Annotated\ features} \quad (3)$$

In response to **Comment #3a**, we think the reviewer is correct in their understanding. The "Top N" annotations were selected based on the rank of annotation scores. The top N ranked structures for each feature were retained as Top N annotations. The score was calculated by a combination of *m/z* match score, RT match score and MS2 dot product similarity score using Eq 4:

$$Score\ of\ metabolite\ annotation = Score_{m/z}w_{m/z} + Score_{RT}w_{RT} + Score_{spec}w_{spec} \quad (4)$$

$$Score_{m/z} = 1 - \frac{|m/z\ error|}{m/z\ tolerance} \quad (5)$$

$$Score_{RT} = 1 - \frac{|Relative\ RT\ error|}{RT\ tolerance} \quad (6)$$

$$Score_{spec} = \frac{\sum W_A W_B}{\sqrt{\sum W_A^2 W_B^2}} \quad (7)$$

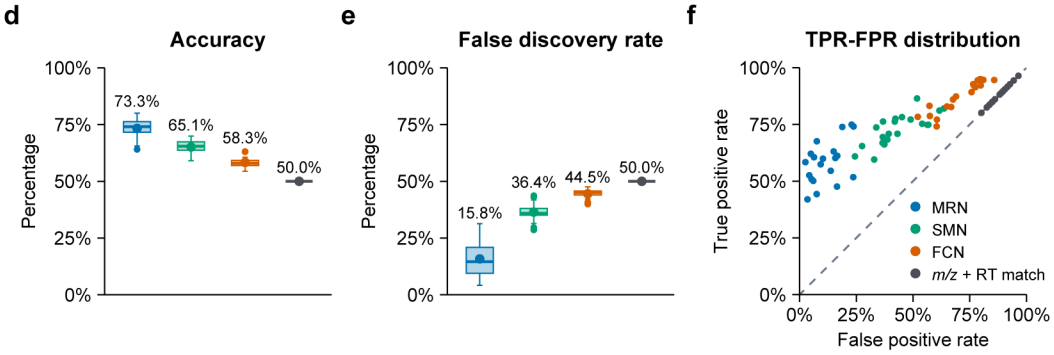
where $Score_{m/z}$ represents the *m/z* match score and is calculated as indicated in Eq. 5. $Score_{RT}$ represents the RT match score and is calculated using Eq. 6. $Score_{spec}$ represents the MS2 spectral similarity score and is calculated as indicated in Eq. 7. A modified dot-product function was used to calculate MS2 spectral similarity and followed the same method as described in our previous MetDNA publication (Shen et al., **Nat. Commun.**, 2019, https://doi.org/10.1038/s41467-019-09550-x). *w* represents weight, the default values of $w_{m/z}$, w_{RT}, and w_{spec} are 0.25, 0.25, and 0.5, respectively. The scoring function follows the previously published MetDNA publication[19].

In response to **Comment #3b**, we clarify that "coverage" refers to whether a feature receives an annotation, regardless of the correctness of that annotation. As defined in Equation 1, this metric is independent of annotation ranking and is therefore directly comparable across different methods.

In response to **Comment #3c**, we sincerely thank the reviewer for the suggestion to estimate the false discovery rate (FDR) for network-based annotation in MetDNA3. Accordingly, we revised the manuscript by following Reviewer #1's suggestion and constructed decoy MS2 spectra using the spectrum-based method reported by Scheubert et al. (***Nat. Commun.***, 2017, https://www.nature.com/articles/s41467-017-01318-5) to more appropriately evaluate the FDR values of each network method. Specifically, we generated decoy MS2 spectra by randomizing the experimental MS2 spectra (**Supplementary Figure 11**), and incorporated them into the MetDNA3 data layer for annotation propagation. The resulting annotations from decoy spectra served as negative controls for evaluating annotation accuracy and false discovery rate (FDR) of network-based annotation propagation (**revised Figure 5c and Supplementary Data 4**). Based on generated decoy MS2 spectra, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR

of 15.8% (**revised Figure 5d and 5e**). In addition, we used the same decoy MS2 spectra to evaluate the performances of SMN and FCN. The results showed that as network connectivity increased (from MRN to SMN to FCN), annotation accuracy decreased from 73.3% to 65.1% and then to 58.3% (**revised Figure 5d**). Correspondingly, the FDR increased from 15.8% to 36.4% and 44.5% (**revised Figure 5e**). The performance of FCN closely resembled that of *m/z*+RT matching, suggesting that FCN-based annotation was primarily driven by *m/z*+RT matches rather than network-based propagation. From the distribution of TPR and FPR, it is evident that our curated MRN exhibits both lower TPR and FPR compared to other methods (**revised Figure 5f**). These results highlight that indiscriminately increasing network connectivity does not improve annotation quality. Instead, carefully curating knowledge networks with high specificity is essential for accurate network-based metabolite annotation and propagation.

[editorial note: panel redacted]



Revised Figure 5c-f. Benchmarking different knowledge-driven network topologies. (c) Schematic workflow for evaluating annotation performances using experimental and decoy MS2 spectra. **(d-f)** Comparative evaluations of accuracy (d), false discovery rate (FDR) (e), and true positive rate (TPR)-false positive rate (FPR) distribution (f) for different networks in network-based annotation propagation. MRN, metabolic reaction network; SMN, structure-guided molecular network; FCN, fully connected network.

Comment #4: *"4. MetDNA3 seems to rely heavily on MS2 similarity (molecular networking) between molecules, since new connections will only be made for peaks X and Y that have similar MS2 spectra, and reaction pairs are also constrained to agree with MS2 similarity (Fig 2a). With this in mind, I'm wondering if evaluating correctness of the method against MS2-based compound identification (from standards) is a circular or biased to some extent?"*

Response: We thank the reviewer for this thoughtful comment. We did not fully understand your comment and apologize if we have misunderstood any part of your comment. As you correctly noted, in our annotation propagation, a new connection between two features is established only when both MS2

spectral similarity and reaction pair (RP) constraints are satisfied. To clarify the potential concern regarding circular reasoning: MS2 spectrum-based metabolite identification generally refers to direct MS2 spectral matching between experimental MS2 spectra and reference MS2 spectra obtained from chemical standards. In contrast, the MS2 similarity employed in our annotation propagation refers to the spectral similarity between two experimental features (mostly with different precursor *m/z* values). Although both approaches involve MS2 spectra, their purposes and implementations are fundamentally different. Therefore, evaluating our method's performance using standard-based MS2 spectral match serves as an independent external benchmark, rather than constituting circular reasoning.

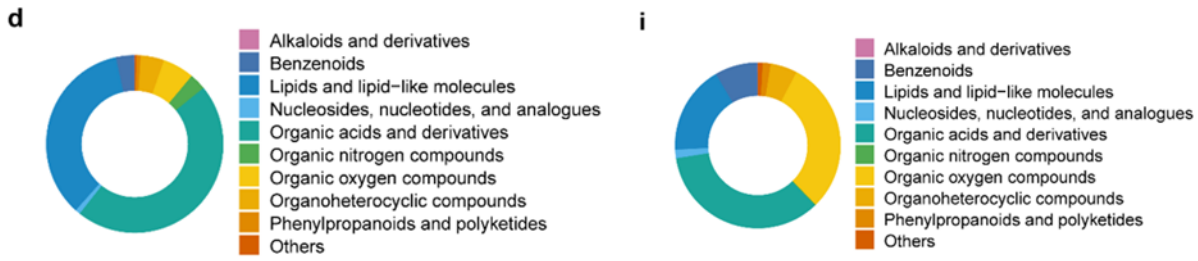
Comment #5: *"5. A key piece of evidence for accuracy of the MetDNA3 method is the independent validation of predictions by MS2. I appreciated this effort, and it does indicate that the method is capable of identifying both "known unknowns" and truly novel compounds. However, it would be most useful to discuss how these candidates were selected, and whether certain compound classes (such as peptides) are more likely to be discovered by the algorithm."*

Response: We thank the reviewer for the insightful comment regarding the validation of newly annotated metabolites, which also aligns with the Comment#2 from Reviewer 1. Specifically, in our study, we employed a stepwise filtering strategy to prioritize unknown candidates for validation. First, we applied cross-validation using multiple in-silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, we prioritized candidates recurrent across multiple biological datasets, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, we manually inspected the MS2 spectra to select reliable candidates. Using this strategy, we successfully validated two unknown metabolites not found in current human metabolome databases. While not exhaustive, this demonstrates the potential of MetDNA3 for novel metabolite discovery.

We have added the following description to the section '**Corroborations of annotated metabolites using other bioinformatic tools**' in the Methods of the revised manuscript:

"A stepwise filtering strategy was applied to prioritize unknown candidates for validation. First, cross-validation was performed using multiple in silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, candidates recurrently detected across multiple biological datasets were then prioritized, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, MS2 spectra were manually inspected to select reliable candidates."

Regarding the potential bias of the algorithm toward specific compound classes, we conducted a preliminary analysis of the chemical class distribution of the annotation results (**Supplementary Figures 6d and 6i)** and did not observe any clear bias. However, we acknowledge that other factors—such as sample type, acquisition conditions, and instrument platform—may also influence annotation outcomes. Therefore, we recommend that users interpret the results with appropriate caution and critical consideration.



Copy of Supplementary Figure 6d and 6i. Characterization of recursive-based metabolite annotation and propagation. (d, i) ClassyFire superclass distribution of annotated metabolites in HILIC–MS(+) (d) and HILIC–MS(-) (i). Human urine samples were tested (n=6 technical replicates).

Comment #6: *"6. Finally, MS2-based matching (against standards) is throughout assumed to be the gold standard for correct metabolite identification. While MS2 can certainly be strong evidence when used correctly, MS2 spectra often cannot distinguish between closely related compounds, for example isomers such as glucose, fructose, and galactose; also, MS2 is underpowered for small molecules that don't fragment well. This aspect should also be discussed."*

Response: We thank the reviewer for this important observation. We fully agree that while MS2-based spectral matching is widely regarded as a gold standard in metabolite identification. To address this, we have added a discussion of these limitations in the revised manuscript:

"However, it should be noted that while MS2 spectrum-based match is widely considered the gold standard for metabolite identification, it has inherent limitations. For small molecules, MS2 spectra often lack sufficient fragment information, and isomeric metabolites may be indistinguishable based on MS2 spectra alone. Therefore, incorporating orthogonal approaches such as retention time (RT) and collision cross section (CCS) is increasingly important for improving small molecule discrimination. Integrating these complementary data into network-based annotation propagation can also help control false positives, ultimately enhancing the accuracy of metabolite annotation in untargeted metabolomics."

*** Minor points ***

Comment #7: *"7. I find that the main results are not clearly stated in the abstract:*

a) Does "10-fold improved accuracy" refers to computational efficiency (reduced computation time) ? please clarify.

b) The statement "It annotated over 1,600 metabolites" refers to MS2 matching against the authors' standard library, which is unrelated to the algorithm.

c) "12,000 new metabolites" - for clarity, these should be called putative annotations."

Response: We appreciate the reviewer's helpful suggestions regarding the abstract. Here are our responses:

7a) Yes, the “10-fold improved efficiency” refers to computational performance, specifically a reduction in runtime.

7b) The statement “over 1,600 metabolites” has been clarified to indicate that these were seed metabolites annotated based on chemical standards.

7c) The term “12,000 new metabolites” has been revised to “12,000 putatively annotated metabolites” to avoid overstatement and better reflect the nature of level 3 annotations.

Together, we have updated abstract in the revised manuscript: *“The generated networking topology enabled interactive annotation propagation with over 10-fold improved computational efficiency. In common biological samples, it annotated over 1,600 seed metabolites with chemical standards and >12,000 putatively annotated metabolites through network-based propagation.”*

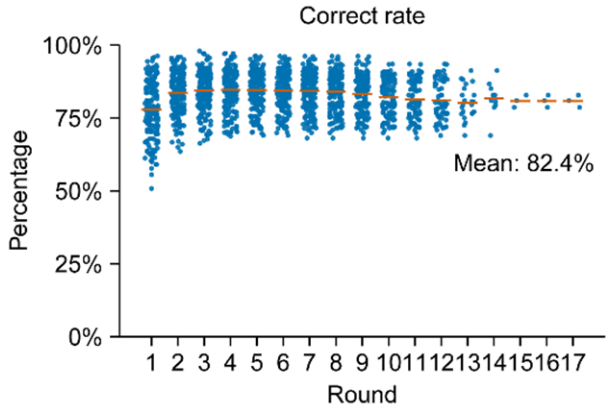
Comment #8: *“8. In Fig 1c, the Tanimoto coefficient is substantially higher in MRN and especially in the biotransformer network, suggesting that the predicted network focuses on structurally similar molecules. This is understandable given the methodology, but nevertheless seems like a weakness, since reactions that yield products that are structurally different from the substrates would be biased against. Could be discussed.”*

Response: We thank the reviewer for this insightful comment. We agree that certain biochemical reactions may yield products that are structurally dissimilar from their substrates, which may not be well captured by our current network expansion approach. We have added a discussion point in the manuscript:

“In addition, many existing knowledge-driven networks tend to retain reaction or metabolite pairs with high structural similarity as the basis for network construction. This approach may overlook metabolite pairs that are structurally dissimilar but exhibit high MS2 spectral similarity. In the future, it will be important to explore more specific and tailored methods for identifying reaction or metabolite pairs that are better suited for network-based annotation propagation.”

Comment #9: *“9. It would be interesting to see how accuracy (Fig 3) depends on the number of recursive iterations taken by the algorithm. Intuitively, one might expect that after several steps of inference, each of which has some error probability, the annotation error rate will shoot up.”*

Response: We appreciate the reviewer’s insightful suggestion. To address this point, we conducted a detailed analysis of annotation correct rates across different rounds of recursive propagation, as shown in **Supplementary Figure 7**. The results indicate that the correct rates for Top 3 annotations remain relatively stable across different rounds, with an average of 82.4%. This suggests that error accumulation is well controlled in our network-based annotation propagation.



Supplementary Figure 7. Correct rates for Top 3 annotations across different recursive propagation rounds.

We have added this new figure and the corresponding description in the main text as follows: *"Across different recursive propagation rounds, the annotation correction rates remained stable, with a mean value of 82.4% for Top 3 annotations (**Supplementary Figure 7**)."*

Comment #10: *"10. Methods page 26, RT match tolerance was specified as "30%"; does this mean 30% relative error, e.g. 1 min +/- 0.3 min? If so this seems quite wide and would not give much information for matching?"*

Response: Thank you for this valuable comment. Yes, the "30%" refers to the relative error of retention time match. This setting followed our previous MetDNA publication (Shen et al., **Nat. Commun., 2019, 10:1516**, https://doi.org/10.1038/s41467-019-09550-x). Given that RT prediction accuracy is still limited, we adopted a relatively wide tolerance to account for variability across platforms and conditions. This parameter is user-adjustable and can be refined based on specific experimental setups or user preferences. We have updated "**Two-layer interactive networking topology strategy implemented in MetDNA3**" section of Methods to clarify:

"The curated MRN was pre-mapped to the experimental data (metabolic features) through MS1 m/z matching and predicted RT matching with tolerances set at 15 ppm for MS1 m/z matching and 30% for RT matching. Given that RT prediction accuracy is still limited, we adopted a relatively wide tolerance to account for variability across platforms and conditions."

Comment #11: *"11. For MS2 matching, the methods state a dot product cutoff of 0.8, but it is not clear how many MS2 fragments were required? Low-quality spectra with few fragments can easily yield false matches even at high dot product values."*

Response: Thank you for the valuable comment. In the current implementation, we did not set a minimum number of MS2 fragments required for seed metabolite matching. As noted, some metabolites inherently produce fewer fragments, which is a known limitation in fragmentation of small molecules. To

ensure annotation accuracy, seed metabolite annotation in MetDNA3 is based on the combined matching of *m/z*, RT, and MS2 spectrum against the standard library. We have updated "**Two-layer interactive networking topology strategy implemented in MetDNA3**" section of Methods to clarify:

*"**(1) Annotation of seed metabolites.** Seed metabolites were first annotated through matching their experimental values such as MS1, retention time and MS2 spectra to the standard libraries. These seed metabolites were classified as level 1 annotation in MetDNA3. The match tolerances were set as MS1 match, 15 ppm; RT match, 25 s; MS2 spectral match, 0.8 (dot product score). MS2 spectral matching has no restriction on the number of matched fragment ions."*

Comment #12: *"12. For MS2 spectral similarity between features, the authors state a dot product cutoff of 0.5, but don't explain how mass differences were accounted for. Directly comparing MS2 spectra between molecules typically requires adjusting for the precursor mass differences. Please clarify."*

Response: We thank the reviewer for this valuable question. A modified dot-product function was used to calculate MS2 spectral similarity between the seed feature and tis neighbor feature. Specifically, when the precursor *m/z* of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with *m/z* values exceeding that of the neighbor precursor are excluded. Conversely, when the neighbor has a higher precursor *m/z*, fragment ions in the MS2 spectrum with *m/z* values greater than that of the seed are removed. This spectral similarity calculation followed the same method as described in our previous MetDNA publication (Shen et al., *Nat. Commun.*, 2019, https://doi.org/10.1038/s41467-019-09550-x).

We have clarified this point in "**Two-layer interactive networking topology strategy implemented in MetDNA3.**" section of Methods the revised manuscript as follows:

"Then, the MS2 spectral similarity between linked features in the feature network was calculated and applied as a constraint (dot product score ≥ 0.5), resulting in a knowledge-constrained feature network. A modified dot-product function was used to calculate MS2 spectral similarity and followed the same method as described in our previous MetDNA publication[19]. Specifically, when the precursor m/z of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with m/z values exceeding that of the neighbor are excluded. Conversely, when the neighbor has a higher precursor m/z, fragment ions in the MS2 spectrum with m/z values greater than that of the seed are removed. The resulting MS2 spectral similarity-based edge relationships were then mapped back to the knowledge layer, generating a data-constrained MRN."

Comment #13: *"13. The term "reaction pair" was a bit confusing to me, as it seems to suggest a pair of two reactions rather than a pair of metabolites; would suggest "metabolite pair" or "reaction"."*

Response: We thank the reviewer for this valuable question. We define a reaction pair (or reactant pair, RP) by linking a substrate metabolite with its product metabolite displaying similar chemical structures in a specific metabolic reaction. The concept originates from the definition of "reaction class" in KEGG

(https://www.genome.jp/kegg/reaction/), and was first introduced in the original MetDNA publication. To avoid potential confusion, the term “reactant pair” may be preferable and can be used interchangeably in this context. We have added the “reactant pair” in the main text where the “reaction pair” is first introduced.

In addition, we have also clarified this terminology in the discussion of the revised manuscript as follows:

“It should be noted that, in a broader conceptual sense, a reaction pair (or reactant pair) can also be interpreted as a metabolite pair, which represents the relationship between two metabolites—particularly in the context of knowledge networks or metabolite annotation methods that rely on relationships.”

Response to the reviewers:

We sincerely thank the reviewers for their valuable comments, which significantly strengthened the manuscript. Detailed revisions were given as follows.

Reviewer #1:

General Comment: *"Dear Editor and Authors. I acknowledge all the work the authors have done to improve the manuscript. I believe that modifications promoted significant improvement, but, maybe due to a miscommunication issue, I still think there are a few points that may have an important impact on the message of the manuscript as well as on the correct use of the tool by less experienced users. I recommend further review to close this gap."*

Response: We sincerely thank the reviewer for the thoughtful reassessment and kind acknowledgment of the improvements made. We also appreciate the reviewer's continued concern regarding potential miscommunication and user guidance. In response, we have carefully addressed the remaining points and revised the manuscript accordingly to further clarify key messages and enhance usability for less experienced users. We hope these reversions effectively close the gap.

Comment #1: *Do you have standards for those 3,301? If yes state is in the text instead of writing 'validation from fig 3'. Where the number 3,301 comes from? Reading the section 'Improved coverage and correct rate of metabolite annotation.' that discusses the fig 3, I could not figure it out. On the text it seams that you have level 1 annotation for 1652:*

"Additionally, two MS instrument types, Orbitrap and Astral, were tested. Across all datasets, MetDNA3 successfully annotated 1,652 unique level 1 metabolites by matching MS1, RT, and MS2 spectra to the chemical standards, with approximately 600–1,000 level 1 metabolites identified per biological sample on average (Figure 3b)."

Again, for me this comparison is very misleading, on the discussion of figure 4 the authors state

"Integrating results from all tools yielded an overall annotation consistency of 90.1% across 4,302 metabolite features in all 20 datasets." Which for what I understand is the union of the annotations for all tools, what should be highlighted in the figure was the

"Notably, 1,680 features (39.1%) were consistently supported by all four tools", that 39% is a much more believable number.

Honestly, I suggest this analysis be removed to supplementary and make suppositions only about the standards. These results are too inflated, the authors can not state such things: "validated by independent tools" this is not validation.

"These findings indicate that majority of network-based metabolite annotations in MetDNA3 could be validated by independent tools. While discrepancies among different annotation tools are inevitable, employing a multi-tool validation strategy enhances confidence in metabolite annotations."

Response: We sincerely thank the reviewer for this insightful comment, which helped us clarify the presentation and interpretation of our validation analysis.

Regarding the number "3,301", to clarify, we evaluated five types of biological samples, and each sample was acquired under four methods with orbitrap (HILIC–MS(+), HILIC–MS(−), RP–MS(+), RP–MS(−); **Figure 3a**). There were 20 datasets in total. Across these datasets, we selected 4,705 level 1 metabolite annotations with adduct forms of $[M+H]^+$ or $[M-H]^-$ for validation experiment in **Figure 3d**. We have added the new data in the **revised Supplementary Table 2**. All of these annotations based on matches to the standard library (matching MS1, RT, and MS2 spectra). When performing the validation experiment in **Figure 3d**, we divided all 4,705 level 1 metabolite annotations by 30% of each dataset for seed metabolites, and the remaining 70% of each dataset for validation. **Therefore, a total of 3,301 level 1 metabolite annotations were used for validation.** It is important to note that a single metabolite may appear in multiple datasets and be identified multiple times. The numbers of unique metabolites were also provided in the **revised Supplementary Table 2**.

Regarding the number "1,652", as shown in **Figure 3b**, a total of 6,763 level 1 metabolite annotations were obtained from eight biological samples acquired using Orbitrap and Astral instruments (resulting in 32 datasets, as each sample was analyzed using four methods). However, one metabolite may appear in multiple samples and be identified multiple times. **When these 6,763 annotations are unified to count unique metabolites, the total number is 1,652.** Therefore, we made the related statement in the manuscript.

Supplementary Table 2. Number of metabolites used for validation in Figure 3d-g from Orbitrap datasets.

Sample	Data acquisition	# annotated metabolites[a]	# Seed metabolites (30%)	# Validation metabolites (70%)
BV2 cell	HILIC–MS(+)	205	61	144
BV2 cell	HILIC–MS(-)	229	68	161
BV2 cell	RP–MS(+)	124	37	87
BV2 cell	RP–MS(-)	201	60	141
Mouse brain	HILIC–MS(+)	287	86	201
Mouse brain	HILIC–MS(-)	273	81	192
Mouse brain	RP–MS(+)	171	51	120
Mouse brain	RP–MS(-)	247	74	173
Mouse liver	HILIC–MS(+)	304	91	213

Mouse liver	HILIC–MS(-)	351	105	246
Mouse liver	RP–MS(+)	150	45	105
Mouse liver	RP–MS(-)	304	91	213
NIST plasma	HILIC–MS(+)	201	60	141
NIST plasma	HILIC–MS(-)	236	70	166
NIST plasma	RP–MS(+)	153	45	108
NIST plasma	RP–MS(-)	249	74	175
NIST urine	HILIC–MS(+)	238	71	167
NIST urine	HILIC–MS(-)	281	84	197
NIST urine	RP–MS(+)	234	70	164
NIST urine	RP–MS(-)	267	80	187
Sum		4705	1404	3301
Unique[b]		1019	671	931

Note: [a] We selected metabolite annotations corresponding to the $[M+H]^+$ or $[M-H]^-$ adduct forms. [b] Duplicated metabolite structures and stereoisomers were removed when counting unique metabolites.

We have revised the main text to explicitly state these numbers for transparency. The revisions are highlighted in red:

*"To evaluate the coverage and correct rate of MetDNA3 in network-based annotation propagation, we analyzed 20 datasets acquired using the Orbitrap instrument (**Figures 3a-c**) and followed the validation framework outlined in **Figure 3d.** Specifically, 30% of the level 1 metabolite annotations were designated as seed metabolites for network-based annotation propagation, while the remaining 70% were reserved for validation. The detailed metabolite numbers are summarized in **Supplementary Table 2**. This process was repeated 10 times for each dataset using a 10-fold cross-validation approach (**Figure 3d**). Detailed results are provided in the **Supplementary Data 2**."*

*"These tools have been increasingly adopted to enhance the structural elucidation of metabolites. To evaluate the accuracy of these methods, we also used the validation metabolites (N=3,301, from 20 Orbitrap datasets, **Supplementary Table 2**) as a benchmark set to assess the performance of CFM-ID, MetFrag, MS-FINDER, and SIRIUS (**Supplementary Figure 13**). Most tools showed high correct rates, with Top 3 annotation rates ranging from 76.6% to 89.4%."*

Regarding the use of the term "validation" in Figure 4, we fully understand the reviewer's concern that true validation ideally involves comparisons against chemical standards. However, most level 3 annotations generated by MetDNA3 either lack available chemical standards or represent truly unknown metabolites, making such validation infeasible. Instead, in Figure 4, we employed four widely used metabolite annotation tools—CFM-ID, MetFrag, MS-FINDER, and SIRIUS—to corroborate the consistency of MetDNA3's level 3 annotations (N=4,302 features with MS2 spectra). We agree that this does not constitute true validation; therefore, in **Figure 4**, we have revised the term "validation" to "corroboration" to more accurately reflect the nature of the analysis. In addition, we have also followed the suggestion for the reviewer, and revised the related statement in the manuscript as follows (revisions are highlighted in red):

1) Legend of Figure 4: *"**Figure 4. Corroboration of network-based metabolite annotations.** (a) Schematic illustration of corroboration between MetDNA3 and other bioinformatic tools for metabolite annotations. (b) Numbers of annotated metabolites corroborated between different bioinformatic tools and MetDNA3. (c) Statistics on the number of bioinformatic tools that generate consistent annotations with MetDNA3."*

2) Main text: *"Multiple bioinformatic strategies have been developed for metabolite annotation in untargeted metabolomics, and integrating multiple approaches provides deeper insights. In particular, most level 3 annotations generated by MetDNA3 either lack available chemical standards, making validation using chemical standards infeasible. To this end, we employed four widely used metabolite annotation tools—CFM-ID[35], MetFrag[36], MS-FINDER[37], and SIRIUS[14]—to corroborate the consistency of network-based metabolite annotations in MetDNA3 (**Figure 4a**). None of these tools rely on network-based annotation strategies. In the 20 Orbitrap datasets, a total of 4,302 features with level 3 annotations were further analyzed using four independent bioinformatic tools for structural annotation and compared against MetDNA3 results (**Figure 4b and Supplementary Data 3**). Integrating results from all tools showed that the consistency of individual tools with MetDNA3 ranged from 60% to 80% across the 4,302 features (**Figure 4b**). In total, 3,894 features (90.1%) had at least one consistent annotation between MetDNA3 and one of the four tools. Notably, 1,680 features (39.1%) were supported by all four tools, while only 428 features (9.9%) lacked corroboration from any tool (**Figure 4c**). These findings indicate that majority of network-based metabolite annotations in MetDNA3 could be corroborated by independent tools. While discrepancies among different annotation tools are inevitable, employing a multi-tool corroboration strategy enhances confidence in metabolite annotations."*

3) Revisions in the Methods of "Corroborations of annotated metabolites using other bioinformatic tools.": *"Four bioinformatic tools were employed to corroborate the annotated metabolites by MetDNA3, including CFM-ID (version 4.0), MetFrag (version 2.4.6-CL), MS-FINDER (version 3.61), and SIRIUS (version 6.0.6).*

For known metabolites (level 3.1; structures derived from KEGG, MetaCyc, and HMDB), a standardized compound database of 53,583 unique metabolites, integrated from KEGG, MetaCyc, and HMDB, used for annotation is the same as MetDNA3. Data formatting and parameter settings were adjusted according to the specifications of each tool. Detailed parameter configurations are provided in

*the **Supplementary Table 5**. A total of 4,302 features with level 3.1 annotations were collected from 20 Orbitrap datasets and subjected to corroboration using the above tools (**Supplementary Data 3**).*

Comment #2: *"Again, that does not explain the choice for the metabolites metabolites γ-Glu-Thr-Gly and N-glycolyltaurine. Where is the list of metabolites filtered by this procedure? How many of this list could be confirmed? Why did you chose γ-Glu-Thr-Gly and N-glycolyltaurine? If you told me, I filtered 10% of my top rank annotated metabolites and was able to validate 2 out of 10, this would be much more believable for me."*

Response: We apologize for not clearly explaining the detail information behind the selection of γ-Glu-Thr-Gly and N-glycolyltaurine. To clarify, we first collected all annotated unknown metabolites (level 3.2 with putative structures generated from BioTransformer) from the 20 Orbitrap datasets, resulting in a total of 1,456 annotation candidates. Then, we employed a two-step filtering strategy to prioritize unknown candidates for validation. First, we applied a scoring-based filtering strategy using multiple in silico tools (MetFrag, MS-FINDER, and SIRIUS). Second, we prioritized candidates that were recurrently detected across multiple biological datasets. The two-step filtering narrowed the list from 1,456 to 5 candidate metabolites (**Supplementary Table 6**). We checked their MS2 spectra, and synthesized two standards for validation, including γ-Glu-Thr-Gly and N-glycolyltaurine. The complete list of 5 candidate metabolites and corresponding information is provided in new **Supplementary Table 6**.

Supplementary Table 6. Five putative unknown metabolites filtered for validation.

Sample and dataset	Feature	Putative annotation	MetFrag score	MS-FINDER score	SIRIUS score
BV2 cell HILIC–MS(+)	M168T345	4-(2-aminoallyl)-1H-imidazole-1-carboxylic acid	1	6.9866	-144.362
Mouse liver HILIC–MS(+)	M168T351		1	7.0026	-161.606
Mouse brain HILIC–MS(-)	M182T222		1	7.4138	-95.755
Mouse liver HILIC–MS(-)	M182T223	N-glycolyltaurine	1	6.1993	-84.234
NIST urine HILIC–MS(-)	M182T223		0.9975	6.2455	-84.334
Mouse brain HILIC–MS(+)	M306T438	γ-Glu-Thr-Gly	1	5.8417	-61.298

Mouse liver HILIC–MS(+)	M306T438		1	6.8446	-46.619
BV2 cell HILIC–MS(+)	M313T406	N-acetyl-Asp-His	1	5.2773	-119.134
Mouse brain HILIC–MS(+)	M313T407		1	6.868	-69.778
BV2 cell HILIC–MS(-)	M610T486	N5-(3-((2-(4-carboxy-4-oxobutanamido)-3-((carboxymethyl)amino)-3-oxopropyl)disulfaneyl)-1-((carboxymethyl)amino)-1-oxopropan-2-yl)glutamine	1	5.8591	-80.571
Mouse liver HILIC–MS(-)	M610T487_1		1	5.9215	-81.546

We have revised the manuscript to clarify the validation processing (revisions are highlighted in red):

1) Revisions in the main text: *"We further validated some metabolite annotations in MetDNA3 using chemical standards (**Figure 4d-g and Supplementary Figure 10**). To prioritize candidates for validation, we applied a stepwise filtering strategy that integrated corroborated annotation from multiple in silico tools and checked their recurrences across multiple biological datasets (**see Methods**). The most notable examples are the discovery of two novel metabolites that are absent from human metabolome databases."*

2) Revisions in the Methods of "Corroborations of annotated metabolites using other bioinformatic tools": *For unknown metabolites annotated by MetDNA3 (level 3.2; structures generated from BioTransformer), we corroborated these annotations using multiple in silico tools. A stepwise filtering strategy was applied to prioritize unknown candidates for validation. First, cross-validation was performed using multiple in silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Second, candidates recurrently detected across multiple biological datasets were then prioritized, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. We collected all annotated unknown metabolites from the 20 Orbitrap datasets, resulting in a total of 1,456 annotation candidates. The filtering process narrowed the list from 1,456 to 5 candidate metabolites (**Supplementary Table 6**). We checked their MS2 spectra, and synthesized two standards for validation, including γ-Glu-Thr-Gly and N-glycolyltaurine.*

Comment #3: *"It is really great that the authors implemented a MS2 decoy strategy. However, I'm not sure simple fixing a ranking threshold would be a good strategy to find false positives. In Scheubert et*

al., as far as I understand, the authors used the standard creation of a null distribution with the score distribution of decoy spectra matching and:

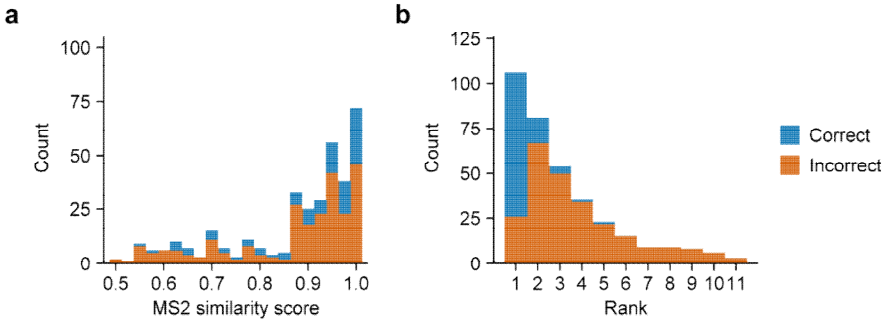
"The p-value of a spectrum match is the probability to randomly draw a result of this or better quality, under the null hypothesis for which a spectrum has been randomly generated."

This would be the preferred strategy."

Response: We sincerely appreciate the reviewer's insightful comment. We fully acknowledge the decoy-based null distribution strategy proposed by Scheubert et al. However, in the context of network-based annotation propagation, MS2 similarity is calculated between feature pairs, rather than between a query spectrum and a fixed spectral library, as in conventional annotation workflows. This fundamental difference makes it challenging to construct a meaningful null distribution for MS2 similarity scores based on randomly generated spectra. To illustrate this limitation, we used the NIST urine HILIC–MS(+) Orbitrap dataset employed in **Figure 3d** to examine **MS2 similarity score distributions** between correct annotations and incorrect candidates. As shown in new **Supplementary Figure 14a**, the distributions of MS2 similarity scores substantially overlap between true and false annotations, limiting their discriminative power and rendering them unsuitable for p-value estimation under a decoy-based null hypothesis.

Instead, ranking-based evaluation has been widely adopted in recent studies on metabolite annotation (e.g., Stravs et al., **Nat. Methods, 2022, 19:865-870**, https://doi.org/10.1038/s41592-022-01486-3; Hoffmann et al., **Nat. Biotechnol., 2022, 40:411-421**, https://doi.org/10.1038/s41587-021-01045-9), where the focus is on whether the correct structure appears within the top N candidates. In our approach, candidate annotations are ranked based on a composite score (*m/z* error, RT error, and MS2 similarity). Again, using the same NIST urine dataset, we observed that **rank distributions** offer better separation between correct and incorrect annotations compared to MS2 similarity scores alone (new **Supplementary Figure 14b**).

In our decoy-based evaluation (**Supplementary Figure 12**), we applied the same propagation process using decoy MS2 spectra and used the top 3 ranked annotations as the default threshold. This allows us to assess the likelihood of erroneous annotations emerging from decoy MS2 spectra. While this approach does not provide a formal p-value, it supports estimation of the false discovery rate (FDR) under a consistent ranking framework.



Supplementary Figure 14. Comparison of distributions between correct annotations and incorrect candidates in the NIST urine HILIC–MS(+) Orbitrap dataset. (**a**) Distribution of MS2 similarity scores.

(**b**) Distribution of candidate ranks. A total of 100 correct annotations and 249 incorrect candidates were included in the comparison.

We have updated the Methods section accordingly under "**Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation.**": *The annotation results were classified as true positives (TP) if annotated within rank ≤ 3, and false negatives (FN) otherwise. For decoy MS2 spectra, the same propagation process was applied, where the results are considered false positives (FP) if annotated within rank ≤ 3, and true negatives (TN) otherwise (**Supplementary Figure 12**). Notably, since MS2 similarity is computed between pairwise features in the network, a traditional null distribution of scores cannot be directly constructed. Therefore, a ranking-based approach was adopted to find false positive and estimate FDR (**Supplementary Figure 14**).*

Comment #4: *"Thank for acknowledged the problems. I think that is still very dangerous, allowing users to make mistakes very easily. I would recommend have a different network for non-animal and also explicitly warning users on the tool's web page, at least."*

Response: Thank you for the suggestion. We have added **a clear warning message both on the MetDNA3 homepage and on the parameter setting page**. These messages explicitly remind users to exercise caution when applying MetDNA3 to non-animal datasets. An example of the warning message is shown in **Figures R1 and R2**.

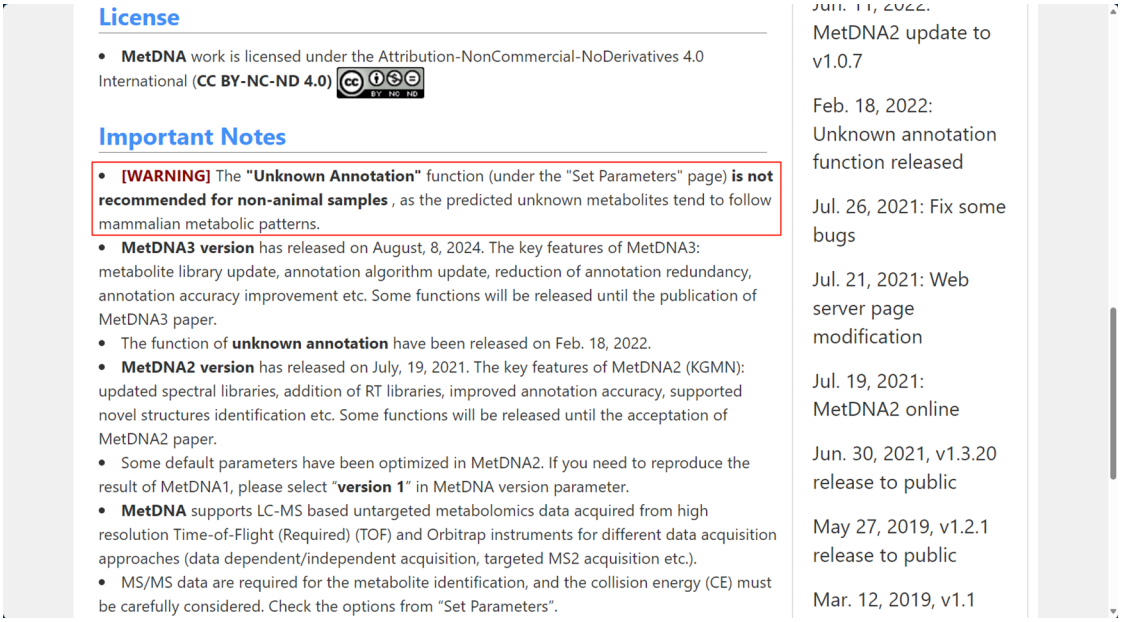


Figure R1. Screenshot showing the updated warning message on homepage of MetDNA.

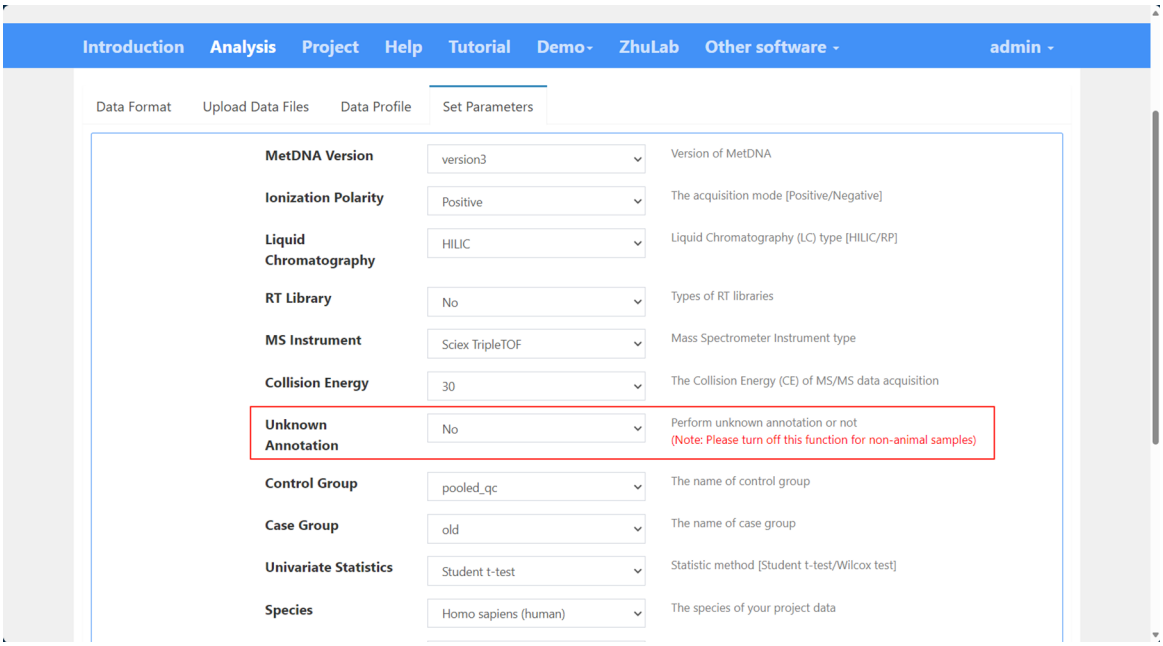


Figure R2. Screenshot showing the updated warning message of "Unknown Annotation" function (under the "Set Parameters" page).

Comment #5: *"The authors have to declare a conflict of interest and state clearly what the property constraints are. It is really a shame the R package is not fully functional."*

Response: Thank you for the suggestion. All functionalities of MetDNA3 are fully accessible through the web platform (http://metdna.zhulab.cn/), which is freely available to all users. Hosting MetDNA3 as a webserver facilitates easier maintenance and updates on our end. Over the past six years, the MetDNA webserver has attracted more than 3,000 registered users worldwide. In addition, the source code for the core algorithms described in this study is fully available on GitHub (https://github.com/ZhuMetLab/MrnAnnoAlgo3). We believe these resources provide sufficient transparency and accessibility to support the publication of our manuscript.

Finally, MetDNA3 has been filed for a patent application (CN202510593995.5). We have disclosed the conflict of interest in the manuscript as follows: *"Z.-J.Z. and H.Z. are inventors on a patent application (CN202510593995.5) that covers part of the technology described in this study."*

Reviewer #2:

"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

Response: We appreciate Nature Communications' initiative to involve and recognize Early Career Researchers in the peer-review process. We sincerely thank the reviewer and the co-reviewer for their valuable comments and constructive feedback, which have significantly improved the quality of our manuscript.

Reviewer #3:

General Comment: *"The authors have made several improvements to the manuscript, but a couple of issues still remain that the authors may wish to address."*

Response: We sincerely thank the reviewer for the continued feedback and thoughtful suggestions. We appreciate the recognition of the improvements made and have carefully addressed the remaining concerns in this revised version of the manuscript. Point-by-point responses and corresponding revisions are detailed below.

Comment #1: *"I still have doubts about the error rate estimates for the MRN. If I understand correctly, the FPR estimate of 2.1% measures the ability of the GRN model to distinguish a true reaction pair from a randomly chosen pair (Fig 1b). From their rebuttal, it seems like the authors agree that the corresponding FDR on the full set of pairs would be very high, as the number of negative pairs vastly outnumber the positive ones. This is because the reaction pair prediction problem is highly unbalanced, which makes it inherently difficult.*

To get around this problem, the authors filter the reaction pairs by mass difference and by molecular structure overlap. These are stringent filters, reducing the number of candidate pairs from 1.44 billion to only 1.66 million (almost 1000x). Then, the authors directly apply the above 2.1% FPR estimate to the filtered 1.66M pairs to come up with an estimate of 2.4% FDR. I don't think this reasoning is correct, because the 2.1% FPR was estimated from an entirely different test set: the GNN is a molecular structure classifier, and FPR as low as 2.1% in the test scheme of Fig 1b is not surprising because the randomly chosen negative pairs consist of completely unrelated structures, so the classification problem here is easy. In contrast, when applying the GNN to the 1.66M pairs that have already been stringently filtered for structural similarity, the classification problem is much harder (since all pairs are now structurally similar) and consequently I would expect the actual FPR on this set to be much higher. So I think this calculation mixes apples and oranges. Moreover, the 2.4% FPR describes only the GNN step, but the interesting error rate is of course that of the MRN as a whole.

Instead, I would suggest that the authors estimate the FDR in the MRN by cross-validating their entire procedure: simply apply the full computational pipeline shown in Suppl Fig 1i to the test set from Fig 1b, count the number of resulting pairs among the positive (RP) and negative (non-RP) classes, calculate the per-pair FPR and then correct for unbalancing to get the resulting FDR for reaction pairs in the MRN.

I realize that the MRN is not the final output of the MetDNA3 algorithm, but as the MRN is a valuable resource in itself I think an accurate error rate estimate is important. As it stands, the manuscript claims to expand existing biochemistry knowledgebases more than "162.8-fold" with an FDR of 2.4%, which sounds rather unlikely."

Response: We sincerely thank the reviewer for the insightful and constructive feedback. We fully acknowledge the concerns raised regarding our estimation of the false discovery rate (FDR) for the predicted metabolic reaction network (MRN), and agree that our initial approach lacked sufficient rigor.

Following your suggestion, we attempted to apply the full MRN construction pipeline to the GNN test set (N=2,995) to evaluate the FDR in an integrated manner. However, after applying the same pre-screening steps in **Supplementary Figure 1i**, only 2 non-reaction pairs (non-RPs) were retained. This makes it infeasible to derive a reliable FDR estimate. Rather than focusing solely on the limitations of FDR estimation, we believe it is equally important to highlight the value of MRN construction in supporting network-based metabolite annotation propagation. Indeed, we acknowledge that our initial MRN prediction strategy lacked sufficient consideration, and that estimating FDR for such a large-scale and complex network is significantly more challenging than we initially anticipated.

In light of these limitations, **we have removed the 2.4% FDR estimate from the manuscript to avoid overstatement.** We have revised the Discussion section to clearly acknowledge the difficulty of accurately estimating the FDR in this context and to better explain our rationale for the two-step pre-screening strategy. We also explicitly emphasize the need for caution in interpreting MRN predictions and recognize that further improvements—such as refining the composition of non-RPs or enhancing the GNN model—are needed to increase prediction robustness. Finally, we believe that despite the lack of accurate FDR estimation, the MRN still represents a valuable resource for uncovering potential biochemical relationships. We are grateful to the reviewer for prompting us to adopt a more critical and transparent evaluation of this resource.

We have revised the related statement in the manuscript as follows (revisions are highlighted in red):

1) Main text: *"To control potential false positives, a two-step pre-screening strategy was applied prior to the prediction, ~~resulting in an estimated false discovery rate (FDR) of 2.4%~~ (**Supplementary Figure 1i ~~and 1j~~**)."*

*"The curated metabolic reaction network substantially enhanced both the coverage and topological connectivity of the knowledge databases. In comparison to knowledge databases, the curated MRN comprised a total of 765,755 metabolites and 2,437,884 potential reaction pairs~~—representing substantial improvements of 66.4-fold and 162.8-fold, respectively~~ (**Figures 1f and 1g**)."*

2) Discussion section: *"The proper curation of a knowledge network is crucial for supporting network-based metabolite annotation. We curated a comprehensive metabolic reaction network using graph neural network-based predictions of reaction relationships, which enhanced both network coverage, connectivity, and specificity. Our results demonstrated that the predicted reaction pairs in MRN significantly improved annotation propagation, exhibiting performance comparable to reported reaction pairs (**Supplementary Figure 9**). However, it is foreseeable that the direct prediction of reaction pairs may include a considerable number of false positives. To mitigate this, we employed a two-step pre-screening strategy aimed at reducing potential false positives of MRN (**Supplementary Figure 1i**). Nevertheless, due to the lack of clearly defined non-RPs under these screening conditions, and the incompatibility of applying GNN-based performance metrics to the filtered candidate set, we refrain from*

reporting an FDR for the MRN. This limitation underscores the need for cautious interpretation of the predicted reaction pairs and highlights several directions for future methodological improvement—including refining the GNN model, expanding the training set, and improving the selection and definition of negative examples. While we recognize these challenges, we consider the MRN a valuable exploratory resource for identifying potential biochemical relationships, and emphasize the need for ongoing efforts to rigorously assess and enhance its predictive accuracy. It should be noted that, in a broader conceptual sense, a reaction pair (or reactant pair) can also be interpreted as a metabolite pair, which represents the relationship between two metabolites—particularly in the context of knowledge networks or metabolite annotation methods that rely on relationships."

Comment #2: *"Regarding comment #4 in my previous review, I understand that the metric used to evaluate MS2 matches differs between the MRN algorithm and the validation, as the former compares molecules of different mass, while the latter performs spectral matching. Still, it seems that there could be some circularity / bias here, since both predict metabolite identity based on MS2 similarity. For example, say we have two structurally related metabolites A, B and two LCMS features X, Y with similar MS2 spectra where X is known to be metabolite A while Y is unknown; and say MetDNA3 makes the connection X-Y and predicts Y to be metabolite B. Even if this prediction is wrong, it is likely that we can successfully "validate" it by matching the spectra Y to a reference spectrum of compound B, because Y was selected by similarly to X=A, and B is structurally similar to A. The underlying problem is that MS2 matching is in itself not infallible. The authors treat MS2 matching as ground truth for validation, but if we consider that MS2 matching also has an error rate, and that MetDNA3 and MS2 matching will tend to make the same errors since MetDNA3 relies heavily on MS2 information, then it seems that validating MetDNA3 by MS2 matching will overestimate accuracy. I don't know of a way to address this problem, but I think it is worth discussing this caveat in the text. Note that Scheubert et al 2017 also pointed out this issue and observed that it leads to optimistic FDR estimates."*

Response: We thank the reviewer for raising this important point. While the principles behind network-based annotation propagation and validation through reference MS2 matching differ, we acknowledge that both approaches rely on MS2 spectral similarity. As such, there is a risk of bias, particularly in cases where structurally similar metabolites produce highly similar MS2 spectra. We agree that MS2 matching is not infallible and may lead to optimistic FDR estimates. To reflect this limitation, we have added the following discussion to the revised manuscript (revisions are highlighted in red):

"However, it should be noted that while MS2 spectrum-based match is widely considered the gold standard for metabolite identification, it has inherent limitations. For small molecules, MS2 spectra often lack sufficient fragment information, and isomeric metabolites may be indistinguishable based on MS2 spectra alone. Consequently, both network-based annotation propagation and validation steps that rely on MS2 spectral matching may lead to optimistic FDR estimates. This limitation warrants careful consideration, and strategies to address it should be explored in future work. Therefore, incorporating orthogonal approaches such as retention time (RT) and collision cross section (CCS) is increasingly important for improving small molecule discrimination. Integrating these complementary data into

network-based annotation propagation can also help control false positives, ultimately enhancing the accuracy of metabolite annotation in untargeted metabolomics."

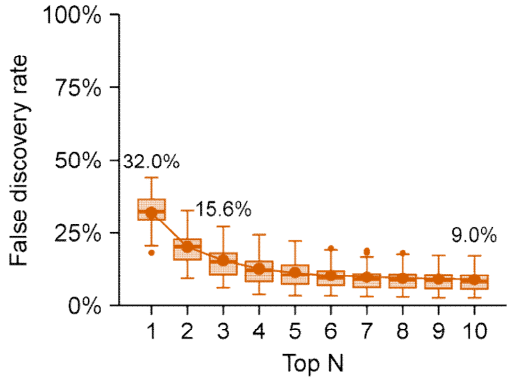
Comment #3: *"The newly added FDR estimation using decoy spectra is interesting, but I don't think it qualifies as a method validation. The decoy-based method is an FDR *estimation* technique, used to assess the rate of false discovering in the absence of a test set; it is only an approximate method, and as Scheubert et al 2017 clearly show, the decoy method can be optimistic, particularly for small q-values. However, in this study the authors actually have a test set, so I think the true FDR should be directly computed from the test set rather than estimated, as done in Figure 3."*

Response: We thank the reviewer for the constructive comment. Following the suggestion from Reviewer #1, we implemented a decoy MS2-based strategy to estimate the false discovery rate (FDR) and evaluate potential false positives in the network-based annotation propagation process.

Here, we also followed your suggestion and use the validation set in **Figure 3d** to calculate the FDR directly. The detailed information of the validation set is provided in the **revised Supplementary Table 2**. Specifically, we define a true positive (TP) as a correct annotation ranked within the top N candidates, and a false positive (FP) as an annotation that does not match the ground truth within the top N. The FDR is then calculated using Eq. 8:

$$False\ discovery\ rate = \frac{FP}{TP+FP} = \frac{\#\ Feature\ without\ a\ correct\ structure\ in\ Top\ N\ annotations}{\#\ Annotated\ features} \quad (8)$$

The results were shown in **revised Supplementary Figure 8**. When *N* = 1, 3, and 10, the corresponding FDRs are 32.0%, 15.6% and 9.0%. These values naturally depend on the choice of N—a smaller N applies a stricter criterion, leading to a higher FDR, while a larger N is more lenient and results in a lower FDR. As a balanced reference point, we recommend reporting the FDR at N = 3 (15.6%) as representative.



Supplementary Figure 8. False discovery rate of Top N annotations in network-based annotation achieved by MetDNA3.

To avoid confusion, we now refer to the FDR calculated from the decoy MS2 spectra as the **estimated FDR**, and have revised the manuscript accordingly. Notably, the estimated FDR calculated from decoy propagation was 15.8% (**Figure 5e**), which is highly consistent with the direct FDR obtained from the validation set in **revised Supplementary Figure 8** (15.6%). This close agreement supports the reliability and practical value of the decoy-based estimation approach in this context.

We have updated the Methods section and added a new figure, along with the corresponding description in the main text as follows: *"In terms of Top N correct rate, MetDNA3 achieved annotation correct rate of 68.0%, 84.4%, and 91.0% for N = 1, 3, and 10, respectively (**Figure 3g**). Across different recursive propagation rounds, the annotation correction rates remained stable, with a mean value of 82.4% for Top 3 annotations (**Supplementary Figure 7**). FDR was also evaluated, yielding 15.6% for Top 3 annotations (**Supplementary Figure 8**)."*

Response to the reviewers:

We appreciate the reviewers' positive comments toward publication.

Reviewer #1:

"I would like to congratulate the authors on the great work and thank them for their politeness and effort to answer all queries. I think the manuscript is ready for publication. I would only advise the authors to be careful with their FDR approach; it still seems very weak. As I don't have much time to look at the simulations with the attention they need, I will trust the author's work."

Response: We sincerely thank the reviewer for the encouraging and supportive comments, and for the recommendation for publication.

Reviewer #3:

"With this latest revision, I feel the authors have addressed all my comments. The revised manuscript text is now more balanced and appropriately points out the potential pitfalls of this complex software system, in particular the difficulty of estimating error rates in metabolite identification. I recommend the manuscript for publication."

Response: We are grateful for the reviewer's positive assessment and recommendation for publication.

Dear Editor and Authors. I acknowledge all the work the authors have done to improve the manuscript. I believe that modifications promoted significant improvement, but, maybe due to a miscommunication issue, I still think there are a few points that may have an important impact on the message of the manuscript as well as on the correct use of the tool by less experienced users. I recommend further review to close this gap.

Response to the reviewers:

We sincerely thank the reviewers for their valuable comments, which significantly strengthened the manuscript. Detailed revisions were given as follows.

Reviewer #1:

General Comment: *"Zhang et al., presented the work entitled Knowledge and Data-driven Two-layer Networking for Accurate Metabolite Annotation in Untargeted Metabolomics where they describe the construction and validation of a network-based approach to perform metabolite annotations. The work is well written, and the authors provide great detail in each validation step, providing data, code and multiple supplementary materials. However, the work needs clarifications, especially in the benchmarking and validation sections, to enable the users to better understand the efficiency of the method and how to better employ it."*

Response: We sincerely appreciate the reviewer's positive feedback. In response to the reviewer's concern about the clarity of the benchmarking and validation sections, we have revised these parts accordingly. We believe these changes have substantially improved both the manuscript and the MetDNA3 tool. Please refer to the following responses and revisions for details.

The three main issues I believe should be addressed are:

Comment #1: *"1) Benchmarking: from what I understood, the authors used a "concordance" of the tools to annotate 4302 features observed in multiple datasets. As all tools provide in silico predictions, one can not rule out that all predictions are wrong. The ideal ground truth for benchmarking would be to compare the accuracy of all tools for standard annotation. This has to be done carefully, taking into account random seed metabolites for MetDNA3 and comparing the remaining metabolites accordingly for each tool. From the methods section, I could not understand what structure databases were used to search with the in silico tools, this should be clarified and standardized for the benchmark."*

Response: Thank you for the reviewer's thoughtful comments on our benchmarking strategy. In our study, we evaluated the accuracy of MetDNA3 using two complementary approaches: (1) for annotations with chemical standards, we withheld 70% for annotation and validation with MetDNA3 (**Figures 3d-g**); (2) for annotations without chemical standards, we applied a cross-tool concordance strategy (**Figures 4a-c**). Features annotated consistently by MetDNA3 and multiple in-silico tools were

considered reliable. Your interpretation of "concordance" is correct. As all tools provide only in-silico predictions in the absence of standards, it remains possible that all predictions are incorrect. We fully agree that such cross-validation may overestimate performance.

To address this concern, we followed your suggestion and evaluated the accuracy of multiple in-silico tools using features annotated by chemical standards (data from Figure 3). Specifically, we used the validation metabolites (N = 3,301) from Figure 3 as a benchmark set to assess the performance of CFM-ID, MetFrag, MS-FINDER, and SIRIUS (**Supplementary Figure 12**). Most tools showed high correct rates, with Top 3 annotation rates ranging from 76.6% to 89.4% (**Supplementary Figure 12d**). We note that these tools may have included part of validation metabolites in their training, potentially contributing to the high accuracy. **Together, while cross-tool concordance may overestimate performance due to the lack of chemical standards, it remains a reasonable and practical approach for assessing annotation reliability.**

We have added this point into the discussion of the revised manuscript as follows:

*"To evaluate the accuracy of these methods, we used the validation metabolites (N = 3,301) from Figure 3 as a benchmark set to assess the performance of CFM-ID, MetFrag, MS-FINDER, and SIRIUS (**Supplementary Figure 12**). Most tools showed high correction rates, with Top 3 annotation rates ranging from 76.6% to 89.4%. We note that these tools may have included part of validation metabolites in their training, potentially contributing to the high accuracy. Together, while cross-tool concordance may overestimate performance due to the lack of chemical standards, it remains a reasonable and practical approach for assessing annotation reliability."*

Do you have standards for those 3,301? If yes state is in the text instead of writing 'validation from fig 3'. Where the number 3,301 comes from? Reading the section 'Improved coverage and correct rate of metabolite annotation.' that discusses the fig 3, I could not figure it out. On the text it seams that you have level 1 annotation for 1652:

"Additionally, two MS instrument types, Orbitrap and Astral, were tested. Across all datasets, MetDNA3 successfully annotated 1,652 unique level 1 metabolites by matching MS1, RT, and MS2 spectra to the chemical standards, with approximately 600–1,000 level 1 metabolites identified per biological sample on average (Figure 3b)."

Again, for me this comparison is very misleading, on the discussion of figure 4 the authors state

"Integrating results from all tools yielded an overall annotation consistency of 90.1% across 4,302 metabolite features in all 20 datasets." Which for what I understand is the union of the annotations for all tools, what should be highlighted in the figure was the

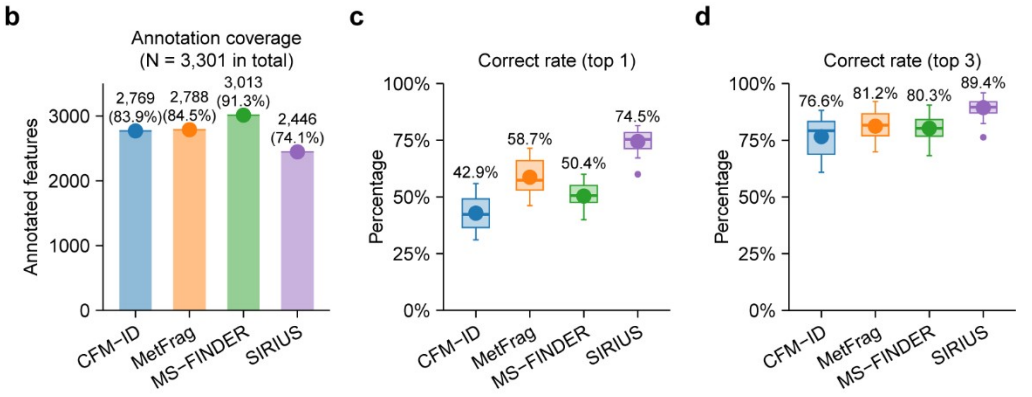
"Notably, 1,680 features (39.1%) were consistently supported by all four tools", that 39% is a much more believable number.

Honestly, I suggest this analysis be removed to supplementary and make suppositions only about the standards. These results are too inflated, the authors can not state such things: "validated by independent tools" this is not validation.

"These findings indicate that majority of network-based metabolite annotations in MetDNA3 could be validated by independent tools. While discrepancies among different annotation tools are inevitable, employing a multi-tool validation strategy enhances confidence in metabolite annotations."

a

[editorial note: panel redacted}



Supplementary Figure 12. Evaluation of multiple in-silico tools used for cross-tool concordance. (**a**) Workflow for performance validation. (**b**) Annotation coverage. (**c**) Correct rate for Top 1 annotations. (**d**) Correct rate for Top 3 annotations.

Regarding the structure databases used in benchmarking, we clarify that all in-silico tools, including MetDNA3, were configured to search against **a standardized compound database of 53,583 unique metabolites**, integrated from KEGG, MetaCyc, and HMDB. This ensures consistency and fairness across all tools. We have updated the Methods section to reflect this clarification:

*"**Corroborations of annotated metabolites using other bioinformatic tools**. Four bioinformatic tools were employed to validate the annotated metabolites by MetDNA3, including CFM-ID (version 4.0), MetFrag (version 2.4.6-CL), MS-FINDER (version 3.61), and SIRIUS (version 6.0.6). A standardized compound database of 53,583 unique metabolites, integrated from KEGG, MetaCyc, and HMDB, used for annotation is the same as MetDNA3. Data formatting and parameter settings were adjusted according to the specifications of each tool. Detailed parameter configurations are provided in the **Supplementary Table 4**."*

Comment #2: *"2) I could not understand the rationale behind the validation of the metabolites γ-Glu-Thr-Gly and N-glycolyltaurine out of thousands of predictions. That is to say, how are the users going to be able to avoid false positives? Would be much believable if the authors explained a filtering strategy (maybe I missed it in my reading) or tried to validate a large number of predictions, thus estimating the false positive rates.*

This is also a general concern with the author's use of the words 'confidece' and 'significance', maybe this wording should be used with more caution, as the 'Knowledge layer' was expanded with predictions not validated experimentally, and the matching provided with mass spectrometry data is

inherently subjected to errors. The author should consider in future versions having more appropriate estimates of significance, in the lines of

Scheubert, K., Hufsky, F., Petras, D. et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8, 1494 (2017). https://doi.org/10.1038/s41467-017-01318-5

Palmer, A., Phapale, P., Chernyavsky, I. et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods 14, 57–60 (2017). https://doi.org/10.1038/nmeth.4072"

Response: Thank you for the insightful comment. In our study, we employed a stepwise filtering strategy to prioritize unknown candidates for validation. First, we applied cross-validation using multiple in-silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, we prioritized candidates recurrent across multiple biological datasets, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, we manually inspected the MS2 spectra to select reliable candidates. Using this strategy, we successfully validated two unknown metabolites not found in current human metabolome databases. While not exhaustive, this demonstrates the potential of MetDNA3 for novel metabolite discovery.

We have added the following description to the section '**Corroborations of annotated metabolites using other bioinformatic tools**' in the Methods of the revised manuscript:

"A stepwise filtering strategy was applied to prioritize unknown candidates for validation. First, cross-validation was performed using multiple in silico tools, retaining candidates with MetFrag scores ≥ 0.5, MS-FINDER scores ≥ 5, and SIRIUS scores ≥ -200. Next, candidates recurrently detected across multiple biological datasets were then prioritized, as such recurrence suggests higher biological relevance and reduces the likelihood of false positives. Finally, MS2 spectra were manually inspected to select reliable candidates."

Again, that does not explain the choice for the metabolites metabolites γ-Glu-Thr-Gly and N-glycolyltaurine. Where is the list of metabolites filtered by this procedure? How many of this list could be confirmed? Why did you chose γ-Glu-Thr-Gly and N-glycolyltaurine? If you told me, I filtered 10% of my top rank annotated metabolites and was able to validate 2 out of 10, this would be much more believable for me.

Second, we sincerely thank the reviewer for the suggestion to estimate the false discovery rate (FDR) for network-based annotation in MetDNA3. Following your recommendation, we adopted the spectrum-based method proposed by Scheubert et al. (**Nat. Commun., 2017, 8:1494**). Specifically, we generated decoy MS2 spectra by randomizing the experimental MS2 spectra (**Supplementary Figure 11**), as described in the original study, and incorporated them into the MetDNA3 data layer for annotation propagation. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, we required that the dot product similarity between the decoy MS2 and the original MS2 be less than 0.5. The resulting annotations from decoy spectra served as negative controls for evaluating false positives. Based on generated decoy MS2 spectra, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR of 15.8% (see **revised**

Figure 5c-e). This accuracy is comparable to the validation using chemical standards shown in Figure 3, which reported 84.4% accuracy for Top 3 annotations. To the best of our knowledge, no existing method has employed decoy MS2 spectra to evaluate false positives in network-based annotation propagation. We hope this approach provides valuable insights for improving future network-based annotation strategies.

[editorial note: figure redacted}

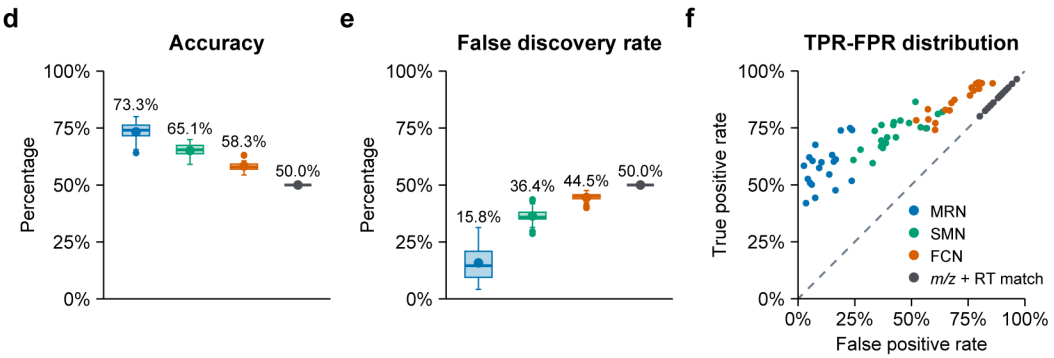
Supplementary Figure 11. Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation. The workflow of experimental and decoy MS2 spectra-based annotation propagation using MetDNA3. Decoy MS2 spectra were generated by spectrum-based method proposed by Scheubert et al. Details were described in the Methods section.

It is really great that the authors implemented a MS2 decoy strategy. However, I'm not sure simple fixing a ranking threshold would be a good strategy to find false positives. In Scheubert et al., as far as I understand, the authors used the standard creation of a null distribution with the score distribution of decoy spectra matching and:

"The p-value of a spectrum match is the probability to randomly draw a result of this or better quality, under the null hypothesis for which a spectrum has been randomly generated."

This would be the preferred strategy.

[editorial note: panel redacted}



Revised Figure 5c-f. Benchmarking different knowledge-driven network topologies. (c) Schematic workflow for evaluating annotation performances using experimental and decoy MS2 spectra. **(d-f)** Comparative evaluations of accuracy (d), false discovery rate (FDR) (e), and true positive rate (TPR)-false positive rate (FPR) distribution (f) for different networks in network-based annotation propagation. MRN, metabolic reaction network; SMN, structure-guided molecular network; FCN, fully connected network.

Third, to demonstrate the superiority of our curated MRN for metabolite annotation (requested by **Rev#3**), we used the same decoy MS2 spectra to evaluate the accuracy and FDR of two additional knowledge-driven networks with distinct topologies: a structure-guided molecular network (SMN) and a fully connected network (FCN). The results showed that as network connectivity increased (from MRN to SMN to FCN), annotation accuracy decreased from 73.3% to 65.1% and then to 58.3% (**revised Figure 5d**). Correspondingly, the FDR increased from 15.8% to 36.4% and 44.5% (**revised Figure 5e**). For comparison, annotation based solely on *m/z* + RT matching with decoy MS2 spectra yielded 50% accuracy and 50% FDR. The performance of the FCN closely resembled that of *m/z* + RT matching, suggesting that FCN-based annotation relied mainly on *m/z* + RT matches rather than network-based propagation. From the distribution of TPR and FPR, it is evident that the MRN exhibits both lower TPR and FPR compared to other methods (**revised Figure 5f**). These results highlight that indiscriminately increasing network connectivity does not improve annotation quality. Instead, carefully curating knowledge networks with high specificity is essential for accurate network-based metabolite annotation and propagation.

We have updated the manuscript to include the new data and methodological details as outlined below.
1) Revision in the Methods: *"**Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation.** Decoy MS2 spectra were generated from experimental MS2 spectra by spectrum-based method proposed by Scheubert et al.[39], and incorporated them into the MetDNA3 data layer for annotation propagation. To ensure that the generated decoy MS2 spectra sufficiently simulate incorrect spectrum, the dot product similarity between the decoy MS2 and the original MS2 was required less than 0.5. For both experimental and decoy MS2 spectra, 30% of them were used as seed metabolites to initiate propagation, and the remaining 70% were used to validate the performance of network-based annotation propagation. The annotation results were classified as true positives (TP) if annotated within rank ≤ 3, and false negatives (FN) otherwise. For decoy MS2 spectra, the same propagation process was applied, where the results are considered false positives (FP) if annotated within rank ≤ 3, and true negatives (TN) otherwise."*

2) Revision in main text related to new Figure 5: *"Next, we adopted the spectrum-based method proposed by Scheubert et al.[39] to generate decoy MS2 spectra by randomizing the experimental MS2 spectra (**Supplementary Figure 11**). The resulting annotations from decoy spectra served as negative controls for evaluating annotation accuracy and false discovery rate (FDR) of network-based annotation propagation (**Figure 5c and Supplementary Data 4**). Based on generated decoy MS2 spectra, network-based annotation in MetDNA3 achieved 73.3% accuracy for Top 3 annotations, corresponding to an FDR of 15.8% (**Figure 5d and 5e)**. In addition, we used the same decoy MS2 spectra to evaluate the performances of SMN and FCN. The results showed that as network connectivity increased (from MRN to SMN to FCN), annotation accuracy decreased from 73.3% to 65.1% and then to 58.3% (**Figure 5d**). Correspondingly, the FDR increased from 15.8% to 36.4% and 44.5% (**Figure 5e**). The performance of FCN closely resembled that of m/z+RT matching, suggesting that FCN-based annotation was primarily driven by m/z+RT matches rather than network-based propagation. From the distribution of TPR and FPR, it is evident that our curated MRN exhibits both*

*lower TPR and FPR compared to other methods (**Figure 5f**). These results highlight that indiscriminately increasing network connectivity does not improve annotation quality. Instead, carefully curating knowledge networks with high specificity is essential for accurate network-based metabolite annotation and propagation."*

3) Regarding the use of the terms "confidence" and "significance," we acknowledge the reviewer's point and have revised the manuscript to minimize their usage. Additionally, we have clarified that the knowledge layer is based on predictions and has not yet been experimentally validated in the main text as follows:

"Notably, the curated MRN was constructed based on predictive models and has not yet been experimentally validated. Its primary purpose is to enhance network-based annotation propagation rather than to directly discover novel reactions."

That was a good improvement.

Comment #3: *"3) Is MetDNA3 designed for animal cells/tissues only? Sorry if that was implicit. If yes, change the title accordingly. I believe the need of clarification stems from the fact that transformations provided by BioTransformer have what the authors called 'Human Super Transformer' component, and if the transformations mapped are indeed biased for animals or humans, users could increase false positives trying to use it for other organisms."*

Response: Thank you for the insightful comment. In MetDNA3, the known metabolite database integrates KEGG, MetaCyc, and HMDB, covering a broad range of common metabolites across species. However, the generation of unknown metabolites relies on BioTransformer's default pipeline, which includes the "Human Super Transformer" module. Consequently, the predicted metabolites may indeed be biased toward human or mammalian metabolism. While the methodological framework of MetDNA3 is generally applicable to other organisms such as plants, bacteria, and fungi, users are encouraged to customize the unknown metabolite database and metabolic reaction network according to their specific needs.

We have added this point in the discussion and recommend that users interpret the results with consideration of the biological context of their target organism. Detailed revisions in the revised manuscript were given as follows:

"Meanwhile, BioTransformer-based generation of unknown metabolites may be limited by its predefined reaction types, potentially introducing species-specific biases. For example, in MetDNA, unknown metabolite generation relied heavily on mammalian-derived reactions, which may bias results toward human or mammalian metabolism. We recommend users interpret annotation results with caution and consider constructing customized unknown metabolite databases and metabolic reaction networks tailored to their specific biological context."

Thank for acknowledged the problems. I think that is still very dangerous, allowing users to make mistakes very easily. I would recommend have a different network for non-animal and also explicitly warning users on the tool's web page, at least.

Comment #4: *"Another issue is that although the code, data, and web interface are available for review purposes, it would be important to provide a worked example of data pre-processing and metabolite annotation to facilitate the review process. I tried to install the package on standard Google Colab (https://colab.research.google.com/notebook#create=true&language=r) following the instructions of the github page and was unable to install it. The option to subscribe to the web tool and try it could breach the anonymity of the review.*

"installation of package '/tmp/RtmpW0b815/file1075f20d1e4/MrnAnnoAlgo3_3.1.1.tar.gz' had non-zero exit status"

>session_info()

$platform

$version

'R version 4.4.3 (2025-02-28)'

$os

'Ubuntu 22.04.4 LTS'

$system

'x86_64, linux-gnu'

$ui

'X11'

$language

'en_US'

$collate

'en_US.UTF-8'

$ctype

'en_US.UTF-8'

$tz

'Etc/UTC'

$date

'2025-04-15'

$pandoc

'NA'

$quarto

Are the code and model for the graph neural network available?"

Response: We sincerely thank the reviewers for their valuable suggestions and for taking the time to test our software and code. MetDNA3 includes not only the core algorithm but also supporting modules such as the metabolite database and spectrum matching, some of which cannot be made publicly available due to intellectual property constraints. To support transparency, we have uploaded the core algorithm code, specifically the two-layer interactive network topology for recursive metabolite annotation, to GitHub (https://github.com/ZhuMetLab/MrnAnnoAlgo3). The full functionality of MetDNA3 remains accessible via our website (http://metdna.zhulab.cn/). For anonymous access during the review process, we have created a dedicated test account:

Username: metdna
Password: metdna@123

This account allows reviewers to explore the platform and validate key functionalities without registration. Demo datasets are available for download from Zenodo repository (https://doi.org/10.5281/zenodo.15589872). Detailed instructions are also provided alongside the demo datasets.

Additionally, for the graph neural network module, we have made the model and full source code publicly available on GitHub (https://github.com/ZhuMetLab/ReactionPredictor), along with detailed instructions. The related information has been added into the "Code Availability" section of the revised manuscript:

"The algorithm of two-layer networking topology was mainly developed using R and is executed in MetDNA3. The source code was provided in GitHub [https://github.com/ZhuMetLab/MrnAnnoAlgo3]. The completed functions are provided in the MetDNA3 webserver [http://metdna.zhulab.cn/] via a free registration. The GNN-based model for predicting of reaction relationships was developed in Python. The source code was provided in GitHub [https://github.com/ZhuMetLab/ReactionPredictor]."

The authors have to declare a conflict of interest and state clearly what the property constraints are. It is really a shame the R package is not fully functional.

Minor issues

Great work on all minor issues.

Comment #5: *"a) What was the chemical taxonomy criterion used to classify the compounds? Was it that of ClassyFire (Djoumbou Feunang et al., 2016)?"*

Response: We thank the reviewer for this comment. Yes, we used ClassyFire to perform chemical taxonomy classification of the compounds, and we have added this information and the citation to the captions of Figure 5 and Supplementary Figure 8 for clarity:

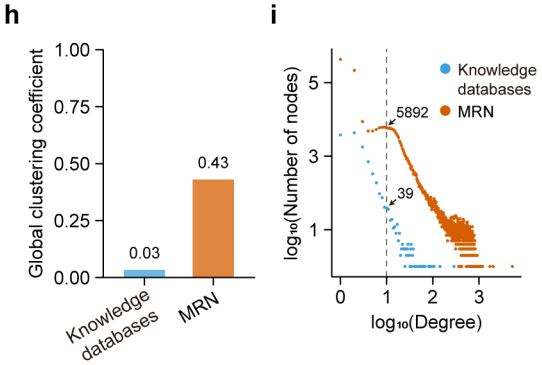
*"**Revised Figure 5.** (b) Proportions of metabolites propagated within the same metabolite superclass for different network-based metabolite annotation. Superclass taxonomy was calculated using ClassyFire."*

*"**Revised Supplementary Figure 8.** (d, i) Superclass distribution of annotated metabolites in HILIC–MS(+) (d) and HILIC–MS(-) (i). Superclass taxonomy was calculated using ClassyFire."*

Comment #6: *"b) From the following statement, between lines 120-121: "Additionally, evaluation of the topological properties of the curated MRN revealed a higher global clustering coefficient and an improved degree distribution relative to knowledge databases". What is the global clustering coefficient based on and how was it obtained, as well as the degree distribution relative to knowledge databases? i) is the number of nodes that have a given degree? Legend or text could be more descriptive."*

Response: We thank the reviewer for this insightful question. To clarify, the global clustering coefficient measures the overall interconnectedness of a network. It is calculated as the ratio of the number of closed triplets to the total number of triplets in the network. A triplet refers to a group of three nodes connected by at least two edges (Wasserman and Faust, Social Network Analysis, Cambridge University Press, 1994; https://doi.org/10.1017/CBO9780511815478). This metric was computed using the transitivity() function from the igraph R package, under the "global" setting.

The degree of a node refers to the number of directly connected neighbor metabolites. Your understanding of degree distribution is correct; it represents the number of nodes in the network corresponding to a given degree value. For example, in our curated MRN, 5,892 metabolites (nodes) have a degree of 10, whereas the knowledge databases-based network has only 39 such nodes. As shown in **revised Figure 1i**, the degree distribution clearly illustrates that our curated MRN has significantly higher connectivity than the network constructed from knowledge databases.



Revised Figure 1. (h-i) Global clustering coefficient of network (h), and degree distribution of nodes (i) in knowledge databases and curated MRN. The global clustering coefficient is calculated as the ratio of

the number of closed triplets to the total number of connected triplets within the network. The degree of a node refers to the number of directly connected neighbor metabolites.

In the revised manuscript, we have updated the **Figure 1i** and corresponding text to provide clearer explanations. Additionally, we have clarified the definitions of the global clustering coefficient and degree distribution in the "**Calculation of network topological properties**" section of the Methods. The detailed revisions are summarized as follows:

1) Revisions in the main text: *"Additionally, evaluation of the topological properties of the curated MRN revealed a higher global clustering coefficient and an improved degree distribution relative to knowledge databases (**Figures 1h and 1i, and Supplementary Table 1**). Degree distribution describes the number of nodes in the network corresponding to a given degree value. For example, in MRN, 5,892 metabolites (nodes) have a degree of 10, whereas the knowledge databases-based network has only 39 such nodes (**Figures 1i**)."*

2) Revisions in the Methods: *"**Calculation of network topological properties.** To characterize the structural and functional features of the networks, a series of topological properties were computed using the R package igraph (version 2.0.3) and Python package networkit (version 11.1). These properties were categorized into three groups: Information, Connectivity, and Community. Information-level metrics include the number of nodes and edges, the number of connected components, and the number of nodes in the largest connected component. Network components were obtained using the components() function in R. Connectivity-level metrics include average degree, network density, global clustering coefficient, average shortest path length, and network diameter. Node degree was computed using the degree() function, and the average degree was calculated as the mean degree across all nodes. Network density was computed using the edge_density() function in R. The global clustering coefficient was calculated as the ratio of the number of closed triplets to the total number of triplets in the network, where a triplet is defined as three nodes connected by at least two edges. It was calculated using the transitivity() function with the setting type="global" in R. The average shortest path length was computed using the mean_distance() function, and the network diameter was obtained via the diameter() function in R. Community-level metrics were derived using the Louvain method, a modularity-based community detection algorithm, implemented via networkit.community.detectCommunities() function in Python."*

Comment #7: *"c) As shown between the lines 140-141: "MS2 similarity between features was calculated and applied as a filtering constraint to eliminate unwanted nodes, further refining the network structure." What metric was used to determine the similarity between the spectra in MS2? Cosine score, modified cosine score, spectral entropy?"*

Response: We thank the reviewer for this valuable question. A modified dot-product function was used to calculate MS2 spectral similarity between the seed feature and tis neighbor feature. Specifically, when the precursor *m/z* of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with *m/z* values exceeding that of the neighbor precursor are excluded.

Conversely, when the neighbor has a higher precursor *m/z*, fragment ions in the MS2 spectrum with *m/z* values greater than that of the seed are removed. This spectral similarity calculation followed the same method as described in our previous MetDNA publication (Shen et al., *Nat. Commun.*, 2019, https://doi.org/10.1038/s41467-019-09550-x).

We have clarified this point in "**Two-layer interactive networking topology strategy implemented in MetDNA3**" section of Methods the revised manuscript as follows:

"Then, the MS2 spectral similarity between linked features in the feature network was calculated and applied as a constraint (dot product score ≥ 0.5), resulting in a knowledge-constrained feature network. A modified dot-product function was used to calculate MS2 spectral similarity and followed the same method as described in our previous MetDNA publication[19]. Specifically, when the precursor m/z of the seed feature is greater than that of the neighboring feature, fragment ions in the seed MS2 spectrum with m/z values exceeding that of the neighbor are excluded. Conversely, when the neighbor has a higher precursor m/z, fragment ions in the MS2 spectrum with m/z values greater than that of the seed are removed. The resulting MS2 spectral similarity-based edge relationships were then mapped back to the knowledge layer, generating a data-constrained MRN."

Comment #8: *"d) As shown between the lines 206-208: "Furthermore, network-based annotation propagation expanded the annotated metabolome to 12,508 metabolites (level 3 annotations), including 9,410 known and 3,098 unknown metabolites (Figure 3c), highlighting the substantial increase in metabolite coverage." What is the annotation rate in terms of the total number of features captured in pre-processing for both Orbitrap and Astral data?"*

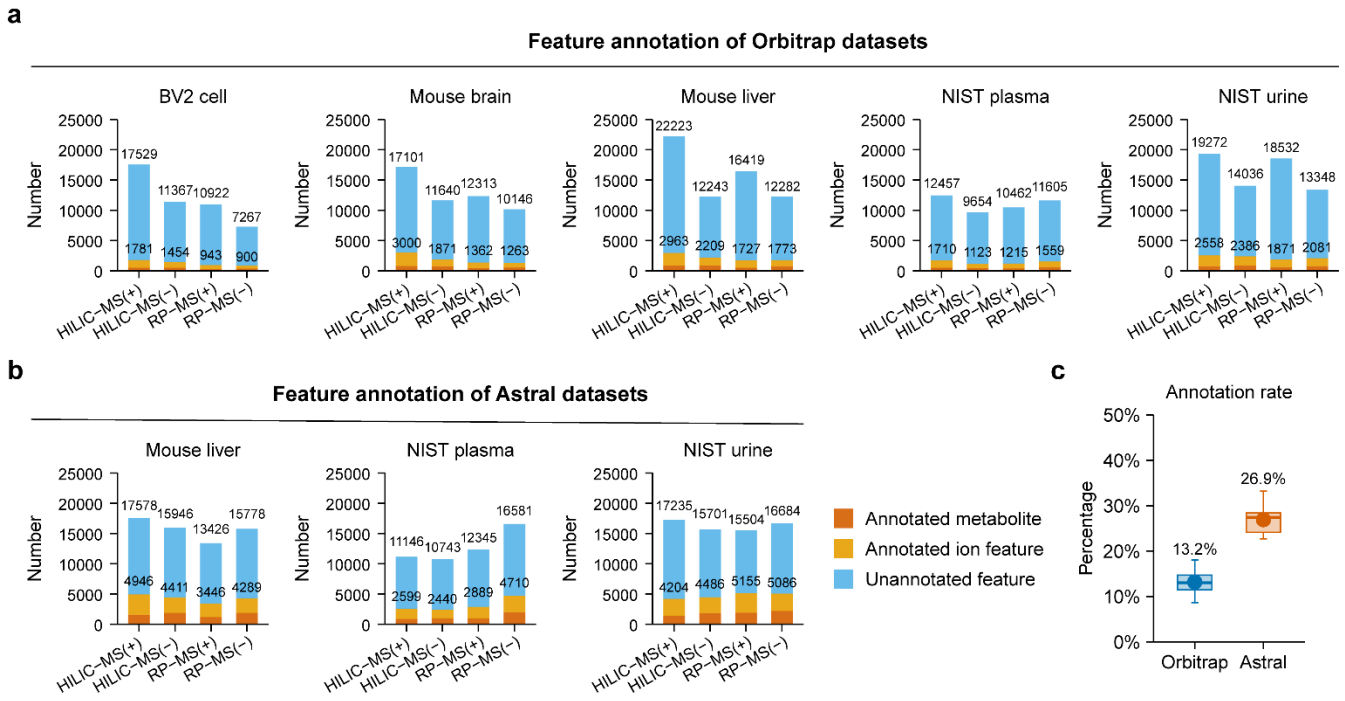
Response: We appreciate the reviewer's thoughtful comment. We have calculated the annotation rate, defined as the proportion of annotated features relative to the total number of detected features obtained from data pre-processing. These results are now presented in **Supplementary Figure 6**. Specifically, in the Orbitrap datasets, we detected between 7,267 and 22,223 features (median = 12,298) in different biological samples. Among these, 900 to 3,000 features (median = 1,750) were annotated using MetDNA3 (**Supplementary Figure 6a**), resulting in an average annotation rate of 13.2% (**Supplementary Figure 6c**). In the Astral datasets, we detected between 10,743 and 17,578 features (median = 15,740) in different biological samples, among which 2,440 to 5,155 features (median = 4,350) were annotated using MetDNA3 (**Supplementary Figure 6b**), yielding an average annotation rate of 26.9% (**Supplementary Figure 6c**). Notably, the total number of features detected in the Orbitrap and Astral datasets was comparable; however, the Astral instrument's higher MS2 acquisition coverage led to substantially improved annotation coverage.

The feature annotations in MetDNA3 include both the identified metabolites and their corresponding ion features, such as isotopes, neutral losses, adducts, and in-source fragments. In general, metabolite features were first annotated through network-based propagation. Subsequently, the global peak correlation network, originally developed in MetDNA2 (Zhou et al., ***Nat. Commun.***, 2022, https://doi.org/10.1038/s41467-022-34537-6), was employed to further annotate the various ion

features associated with each metabolite. This approach uses chromatographic coelution correlations to recognize related ion features, including adducts, isotopes, neutral losses, and in-source fragments.

In the revised manuscript, we have updated the **Supplementary Figure 6** and corresponding text to describe annotation rate as follows:

*"Detailed metabolite annotations for all samples were provided in **Supplementary Figures 5 and 6**, and **Supplementary Data 1**. Notably, data acquired using the Astral instrument demonstrated superior annotation coverage compared to the Orbitrap (**Figures 3b and 3c**). Additionally, the Astral data exhibited a significantly higher annotation rate (26.9%) compared to the Orbitrap (13.2%) (**Supplementary Figure 6**)."*



Supplementary Figure 6. Feature annotation in Orbitrap and Astral datasets using MetDNA3. (**a**) Feature annotations for Orbitrap datasets. (**b**) Feature annotations for Astral datasets. (**c**) Feature annotation rate, defined as the percentage of annotated features (including annotated metabolites and ion features) relative to the total number of detected features. HILIC–MS(+): positive MS mode with HILIC separation; HILIC–MS(-): negative MS mode with HILIC separation; RP–MS(+): positive MS mode with RP separation; RP–MS(-): negative MS mode with RP separation.

Comment #9: *"e) Do the authors intend to submit the fragmentation spectra of the new compounds characterized in this work to the databases used by the metabolomics community, such as MoNA?"*

Response: We thank the reviewer for the suggestion. We have submitted the MS2 spectra of the newly characterized compounds to the MoNA database (https://mona.fiehnlab.ucdavis.edu/). The deposited spectra can be found under the IDs from MoNA_0003569 to MoNA_0003577.

We have also included this information in the “Data Availability” section of the revised manuscript as follows:

“The MS2 spectra of the newly characterized compounds can be accessed at MoNA database [https://mona.fiehnlab.ucdavis.edu/] via IDs from MoNA_0003569 to MoNA_0003577.”

Comment #10: *“f) Do the authors consider submitting the raw data do metabolights, metabolomics workbench or gnps?”*

Response: We thank the reviewer for the helpful suggestion. We have deposited the raw data in the MassIVE repository (https://massive.ucsd.edu/). The datasets are available under the accession IDs MSV000097913 and MSV000097914.

We have also included this information in the “Data Availability” section of the revised manuscript as follows:

“The raw data files of BV2 cells, mouse brain tissue, mouse liver tissue, NIST human plasma, and NIST human urine from the Orbitrap instrument can be accessed at the MassIVE repository [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000097913] or the National Omics Data Encyclopedia under Accession Code OEP00006095 [https://www.biosino.org/node/project/detail/OEP00006095]. The raw data files of mouse liver tissue, NIST human plasma, and NIST human urine from the Astral instrument can be accessed at the MassIVE repository [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000097914] or National Omics Data Encyclopedia under Accession Code OEP00006083 [https://www.biosino.org/node/project/detail/OEP00006083].”

Reviewer #2:

"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

Response: We appreciate Nature Communications' initiative to involve and recognize Early Career Researchers in peer review. We sincerely thank the reviewer for the valuable comments, which have greatly strengthened the manuscript.