

Predicting microbial community structure and temporal dynamics by using graph neural network models

Corresponding Author: Professor Per H Nielsen

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Reviewers' Comments:

Remarks to the Author: In their manuscript "Predicting microbial community structure and temporal dynamics by using graph neural network models", Andersen et al use a graph neural network model to predict the future abundance dynamics of the most abundant bacteria in the microbial communities of activated sludge. With the ongoing growth in microbial sequencing data, the next frontier is using these data to build models to predict future dynamics and trends of microbial communities to better manage huge microbial resources or control microbial process. From the goal perspective, this study is a valuable contribution. In addition, this research provides the longest continuous monitoring data of activated sludge systems to date, which provides a rich database resource for the study of temporal dynamics, microbial interactions, and community assembly. Nonetheless, I believe that the current version of this manuscript provides a rough description of the results and needs more refinement and revision. Please find below some comments and questions that should be addressed.

1. The most important highlight of this paper is its predictive model "Graph neural network model". However, the description of this model in the manuscript is very general and simple. What is the framework or workflow of this model? What are the inputs and outputs as applied to this study? What are the requirements for time series data? For example, whether an equal time interval is required, whether there is a requirement for minimum sampling number, etc. Please spend at least 1 section in the Results section describing these details.
2. The authors have emphasized that they built the predictive model for each ASVs separately. In this case, I think R21:1 is an effective parameter to evaluate the prediction accuracy (Averill, C., et al. 2021, Nat Ecol Evol; X. Liu., et al. 2023, Microbiome). Please show the R21:1 results for each ASVs in your paper. Authors mentioned in their paper that they also used mean absolute error (MAE) and mean squared error (MSE) to evaluate the model performance, but there are no related results both in the main manuscript and supplementary.
3. This paper predicted the temporal dynamics of 4,800 individuals in 24 data sets in total. More specifically, the Mariagerfjord dataset is used as an example to describe the prediction situation, which is acceptable. But, since the authors also admitted that process-critical bacteria in WWTP was very important, I think there should be a group summary of the prediction of process-critical bacteria in all data sets, whether in the main text or in the supplementary. The result description given in this manuscript is too rough.
4. Only predicting the dynamics of abundant taxa may not be enough. It has been previously reported that rare taxa can also have an important impact on the structure and function of microbial communities (Lynch M D J, 2015, Nature Reviews Microbiology; Banerjee S, 2018, Nature Reviews Microbiology). What do the authors think about this?
5. The authors selected 1000 reads as the threshold for filtering samples, which is too low for 16S rRNA amplicon sequencing. Please give a reasonable reason or re-analyze with deeper sequencing depth.
6. How did you get graph network interaction strengths in this paper? Is there any difference between it and microbial interactions in traditional co-occurrence networks? If they are different, can the graph network interaction possibly replace the latter as a new method to explore microbial interactions?

7. Most of the absolute values of the graph network interaction strengths illustrated in 'Figure S4' are very small. What's the meaning of it? Is there a relationship between interaction strength values and prediction accuracy of associated ASVs?

8. If the sample size of the test dataset is larger than the sample number that the trained model can predict at one time, then do you use the actual value or the predicted value of the previous sample to predict the subsequent samples?

9. Please polish all figures in this manuscript.

(Remarks on code availability)

The code is available.

Reviewer #2

(Remarks to the Author)

In the manuscript entitled "Predicting microbial community structure and temporal dynamics by using graph neural network models", the authors leveraged an existing graph neural network model to predict the abundance dynamics of the majority of bacteria in the microbial communities of activated sludge in wastewater treatment plants (WWTPs). They trained and tested models on time series datasets from 24 different Danish full-scale WWTPs sampled 2-5 times a month over a period of 3-8 years. They were able to predict the dynamics 10 time points into the future (2-4 months) with high accuracy for most of the abundant bacteria, especially process-critical bacteria, and some even 20 time points ahead (up to 8 months). The result sounds very promising and will certainly interest some researchers in the field of microbial ecology (especially those working on microbiome data of WWTPs). However, we have serious concerns about the novelty, rigor, generalizability, and interpretability of the presented method.

1. Novelty

The presented prediction or forecasting method heavily depends on the published deep-learning model: MTSF-DG (multiple time series forecasting with dynamic graph modeling), which was cited as Ref.[24] in this manuscript. The authors did spend efforts to pre-process the microbiome data of WWTPs so that it could be fed into the MTSF-DG framework. But this significantly diminishes the novelty of this methodology work.

2. Rigor

2.1. The authors mentioned that "both stochastic factors (such as immigration) and deterministic factors (such as nutrients and predation) have a significant influence on the community structure [8–11], however, the relative contributions can vary." For temporal microbial dynamics in WWTPs, we believe that both stochastic and deterministic factors shape the community structure. But in this work, the authors didn't study the relative contributions of stochastic and deterministic factors at all. This raises a serious concern about the forecastability of the temporal microbial dynamics in WWTPs. We suggest that the authors should seriously quantify the forecastability of the temporal microbial dynamics in WWTPs before they apply any forecasting method. For example, they can compute the approximate entropy (ApEn) to quantify the forecastability of their time-series data [1].

2.2. In this work, key environmental factors (e.g., temperature, nutrient concentrations, interventions, etc) were completely ignored. The authors mentioned that including them "may improve the prediction accuracy even further, but will also complicate the approach even more, let alone require more data to be gathered consistently and reliably." This is not a satisfactory explanation. Incorporating environmental factors is essential to enhancing the model's robustness and reliability. For example, Larsen et al (2012) developed a model for predicting microbial community structure as a function of environmental parameters and microbial interactions [2]. Liu et al (2023), used environmental parameters to predict microbial community composition in AS [3].

2.3. The authors mentioned that "the exact dates or duration between sampling points were not used for the model training as any potential extrapolation would not be reliable, so they have just been sorted chronologically." Does this simply imply that the authors treated those unevenly spaced time series as evenly spaced? This seems to be problematic.

2.4. The authors mentioned that "To optimize the prediction accuracy of the model, four different pre-clustering methods were tested individually and a graph neural network model was designed (described in detail later) and then trained on each cluster". It is hard to believe that one has to perform pre-clustering and design/train the GNN for each cluster separately to achieve better prediction accuracy. Species from different clusters (regardless of how you pre-cluster them) might strongly interact with each other. Shouldn't we consider all the species in a single GNN to achieve better prediction accuracy?

2.5. The authors mentioned that "the median Bray Curtis value across all of the test datasets was 0.147 indicating a good overall prediction accuracy." Why does the median Bray Curtis value 0.147 indicate a good accuracy? With respect to what? How will traditional forecasting methods (e.g., autoregressive integrated moving average, or ARIMA) perform?

2.6. Is the temporal variations of microbial compositions within the same WWTP bigger or smaller than 0.147? Is it significant? What is the Bray-Curtis dissimilarity between true samples from the same WWTP but collected at different time points? This skepticism is further supported by Figure 4, where the distances between predicted and true samples appear comparable to those observed between samples collected at different times. It is imperative to conduct a thorough assessment of the dissimilarity among true samples over time and across different WWTPs to substantiate the judgment of the 0.147 value as indicative of prediction accuracy.

3. Generalizability

The authors trained and tested their model on time series datasets from 24 different WWTPs separately and independently of each other. Moreover, they mentioned that “During the design and implementation of the neural network model there were also many decisions made that were ultimately governed by trial and error until no more significant improvements in prediction performance were obtained across the datasets overall.” This raises a serious concern on the generalizability of their model. We don’t think the model is easily applicable to other microbial time-series datasets. There are too many trial-and-error and parameter tuning.

4. Interpretability

4.1. The authors clearly demonstrated that the predictions for some ASVs were more accurate than other ASVs. The authors also mentioned that “it was impossible to predict every single species in the community precisely with the approach and data used in this study.” But why?

4.2. The authors mentioned that “it is clear that clustering by biological function generally resulted in a lower prediction accuracy compared to the other clustering methods, except for the Ejby Mølle and Hirtshals datasets.” Why does the clustering by biological function generally result in a lower prediction accuracy?

4.3. The authors mentioned that “the interaction strengths in the graph neural network could also be interpreted as hypothetical biological interactions between the individual ASVs in the group (see Figure S4).” This is really far-fetched. The GNN is undirected or bidirectional. The interactions between ASVs do not have to be bidirectional.

4.4 As depicted in Figure 6, surprisingly, the model performance on the time series of ASV2 for 20 predicted samples into the future is even better than 15 or 10 predicted samples into the future. This is quite counter-intuitive and deserves an explanation.

References:

- [1] Pincus, S. M.; Gladstone, I. M.; Ehrenkranz, R. A. (1991). "A regularity statistic for medical data analysis". *Journal of Clinical Monitoring and Computing*. 7 (4): 335–345. doi:10.1007/BF01619355.
- [2] Larsen, P., Field, D. & Gilbert, J. Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* 9, 621–625 (2012). <https://doi.org/10.1038/nmeth.1975>
- [3] Liu, X., Nie, Y. & Wu, XL. Predicting microbial community compositions in wastewater treatment plants using artificial neural networks. *Microbiome* 11, 93 (2023). <https://doi.org/10.1186/s40168-023-01519-9>

(Remarks on code availability)

Yes. But we cannot install their code. There are more than 180 dependency packages, and most of them have incompatible issues with each other.

Also, there is an inconsistency regarding the use of sampling time as input versus merely arranging the sample points chronologically. Specifically, on page 15, lines 469-470, it is noted that samples were sorted chronologically without considering the actual dates or the duration between sampling points. However, in the GitHub documentation, one of the required inputs is metadata that must include sampling dates in the year-month-day format.

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports as part of the Nature Communications initiative to facilitate training in peer review and appropriate recognition for co-reviewers.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

The authors propose using a graph neural network (a machine-learning model) to forecast species abundance in biological wastewater treatment plants (WWTPs). I find the main merit of the paper in the dataset the authors collected for 24 full-scale WWTPs in Denmark. As the authors emphasize, this dataset is exceptional in length and coverage. On the other hand, I have several reservations about the author’s prediction approach and validation of graph neural networks. My main

reservation is that the authors ignore that the composition of the inflow to the WWTPs changes with time. For example, this might render predictions impossible if the inflow exhibits drastic changes. I comment on this below.

Main comments.

1. The ecology WWTPs depends on internal factors like the population dynamics and interactions between species (which, with some reserve, may be captured by relative abundance time-series). However, its ecology also depends on external factors like the composition of the inflow to the WWTPs. In particular, because this inflow carries both organic and inorganic materials, it can drastically change community composition. Somewhat unexpectedly, the authors perform predictions ignoring all external factors. From this perspective, some claims made by the authors, like "we are able to predict the dynamics 10 time points in the future (2-4 months) with high accuracy," seem not well-posed because the prediction accuracy depends on keeping all those external factors constant. Note these external factors must remain constant during the prediction and across the data used to train the algorithm. This is a critical limitation that must be studied and addressed in detail.

2. I have several comments regarding Figure 1:

2a. How the authors calculated a single Bray-Curtis dissimilarity needs to be clarified, considering multiple species (ASVs) and time points exist. Is this the Bray-Curtis dissimilarity across species and time?

2b. In the same line as above, it is difficult to get insights from Figure 1 because of the difficulty of interpreting the derived Bray-Curtis dissimilarity. For example, across datasets, the authors report a Bray-Curtis dissimilarity of about 0.15. What does this mean concretely for the prediction accuracy? What about the significant variations represented by error bars and outliers, which in some cases are above 0.7 (e.g., Odense NE and Ribe)? Do they correspond to some particular ASVs or time instants? It would be much more informative to present results for best, typical, and worst predictions for ASVs, one for each dataset.

3. An essential point that needs to be added to the study is how long the training time-series must be to provide accurate predictions. In particular, it would be handy to have a plot showing test prediction error as a function of the length of train time-series for each WWTP.

4. Regarding Figure 2, it would be much more informative if it presented ASVs in the top, middle, and bottom quantiles of prediction error.

5. Graph neural networks are not constructed by default to handle compositional data (i.e., relative abundances). How do the authors handle this problem to ensure that predictions remain compositional?

6. In Figure 1 and Figure 5, the authors emphasize differences in prediction when using different clustering methods. However, the boxplots in both figures are very similar across methods. Are differences statistically significant? Also, to improve the presentation, I would recommend considering only one clustering method to present the results in the main text, leaving the other methods as a supplementary discussion.

Minor comments

7. In Figure 1. It is worth noting that in the caption of Figure 1 the authors write "Bray Curtis distance". However, mathematically, the Bray-Curtis dissimilarity is not a distance because it does not satisfy the triangle inequality.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

1. I don't think the description of the model is clear enough. In fact, Figure 1 merely introduces the dataset division and provides a basic overview of the model. This is a very general overview of deep learning, applicable to almost all machine learning models. However, the training and prediction based on ASV clusters, which is the highlight of this study, is not reflected in Figure 1.

2. Although your model is trained on each cluster of 5 ASVs, you will eventually obtain prediction values for each ASV, as you mentioned in Line 129 and Figure 4. The R21:1 is suitable for evaluating the prediction accuracy of each ASV, so I believe it is applicable to your study.

3. The manuscript gives so many summary results, including some overview and boxplot graphs. Please share your detailed data and result tables as supplementary so that others can reproduce your results.

4. The manuscript should also be checked for accuracy of the language, as some parts read too colloquial.

(Remarks on code availability)

The code provides a README file with enough instructions for installing and running the application. It's clear enough. But I have no time to reproduce their results.

Reviewer #2

(Remarks to the Author)

1. The authors admitted that this study "presents a concept, or workflow, and not a universal model which can be applied to all sorts of microbial times series". This significantly diminish the impact and importance of this work. Moreover, they also claimed that their workflow, "mc-prediction", suitable for any longitudinal community dataset, not limited to the WWTP ecosystem. Is this a self-contradiction?

2. They claimed that this is the first application of predicting future community dynamics. This is certainly not true. Various deep learning models (especially those based on recurrent neural networks) have already been applied to predicting microbiome dynamics. Here are just two recent references: (1) Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics, Mayank Baranwal et al., eLife 2022; (2) Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions, Jaron C. Thompson et al., PLoS Computational Biology 2023. To our surprise, none of those efforts was mentioned in this paper. Why?

3. It is good to see Figure 1 and the added details describing the data. However, the model details should also be included in Figure 1 or in a separate figure to address questions including but not limited to the number of layers in the design and the parameters for each layer.

4. We still have reservations about the "good" performance of the model. When the Bray-Curtis dissimilarity 0.147 is arbitrarily regarded as "good or very good", in Figures 3 and 4, it is preferable to estimate numerical parameters to express the model's performance. Comparisons with previous studies, other models, and traditional co-occurrence models would also be supportive. In its current version, the evidence does not seem sufficient to justify statements in the manuscript claiming it is "good or very good."

5. We believe the authors misunderstood the question regarding the time interval. The question is not about whether the time interval should be the same or not, but rather whether time stamps should be used instead of just sequential datasets. A test on this should be added. What are the time units (10 days later? 10 weeks later? ...) for the predicted outcomes for a model like Mariagerfjord, using temporal sequential datasets ignores the time interval between points.

6. They claim their workflow, "mc-prediction", suitable for any longitudinal community dataset, not limited to the WWTP ecosystem. We don't see any clear evidence that it works to any other temporal data of microbiome. The authors should just demonstrate it, if they claim the general applicability of their workflow.

In sum, the quality of the presented work does not meet the high standards of Nature Communications. A heavily revised version of this manuscript should be submitted to a more specialized journal on WWTP.

(Remarks on code availability)

We still failed to install the tool successfully. We got the following errors (Linux system: Python 3.10.14; CentOS Linux 7):
Pip subprocess error:

ERROR: Ignored the following versions that require a different python version: 1.11.0 Requires-Python <3.13,>=3.9; 1.11.0rc1 Requires-Python <3.13,>=3.9; 1.11.0rc2 Requires-Python <3.13,>=3.9; 1.11.1 Requires-Python <3.13,>=3.9; 1.11.2 Requires-Python <3.13,>=3.9; 1.11.3 Requires-Python <3.13,>=3.9; 1.11.4 Requires-Python >=3.9; 1.12.0 Requires-Python >=3.9; 1.12.0rc1 Requires-Python >=3.9; 1.12.0rc2 Requires-Python >=3.9; 1.13.0 Requires-Python >=3.9; 1.13.0rc1 Requires-Python >=3.9; 1.13.1 Requires-Python >=3.9; 1.14.0 Requires-Python >=3.10; 1.14.0rc1 Requires-Python >=3.10; 1.14.0rc2 Requires-Python >=3.10; 1.2.0 Requires-Python >=3.9; 1.2.1 Requires-Python >=3.9; 1.2.1rc1 Requires-Python >=3.9; 1.25.0 Requires-Python >=3.9; 1.25.0rc1 Requires-Python >=3.9; 1.25.1 Requires-Python >=3.9; 1.25.2 Requires-Python >=3.9; 1.26.0 Requires-Python <3.13,>=3.9; 1.26.0b1 Requires-Python <3.13,>=3.9; 1.26.0rc1 Requires-Python <3.13,>=3.9; 1.26.1 Requires-Python <3.13,>=3.9; 1.26.2 Requires-Python >=3.9; 1.26.3 Requires-Python >=3.9; 1.26.4 Requires-Python >=3.9; 1.4.0 Requires-Python >=3.9; 1.4.0rc1 Requires-Python >=3.9; 1.4.1.post1 Requires-Python >=3.9; 1.4.2 Requires-Python >=3.9; 1.5.0 Requires-Python >=3.9; 1.5.0rc1 Requires-Python >=3.9; 1.5.1 Requires-Python >=3.9; 2.0.0 Requires-Python >=3.9; 2.0.0b1 Requires-Python >=3.9; 2.0.0rc1 Requires-Python >=3.9; 2.0.0rc2 Requires-Python >=3.9; 2.0.1 Requires-Python >=3.9; 2.1.0 Requires-Python >=3.9; 2.1.0rc0 Requires-Python >=3.9; 2.1.1 Requires-Python >=3.9; 2.1.2 Requires-Python >=3.9; 2.1.3 Requires-Python >=3.9; 2.1.4 Requires-Python >=3.9; 2.14.0 Requires-Python >=3.9; 2.14.0rc0 Requires-Python >=3.9; 2.14.1 Requires-Python >=3.9; 2.15.0 Requires-Python >=3.9; 2.15.1 Requires-Python >=3.9; 2.15.2 Requires-Python >=3.9; 2.16.0 Requires-Python >=3.9; 2.16.1 Requires-Python >=3.9; 2.16.2 Requires-Python >=3.9; 2.17.0 Requires-Python >=3.9; 2.2.0 Requires-Python >=3.9; 2.2.0rc0 Requires-Python >=3.9; 2.2.1 Requires-Python >=3.9; 2.2.2 Requires-Python >=3.9; 3.0.0 Requires-Python >=3.9; 3.0.1 Requires-Python >=3.9; 3.0.2 Requires-Python >=3.9; 3.0.3 Requires-Python >=3.9; 3.0.4 Requires-Python >=3.9; 3.0.5 Requires-Python >=3.9; 3.1.0 Requires-Python >=3.9; 3.1.1 Requires-Python >=3.9; 3.2.0 Requires-Python >=3.9; 3.2.1 Requires-Python >=3.9; 3.3.0

Requires-Python >=3.9; 3.3.1 Requires-Python >=3.9; 3.3.2 Requires-Python >=3.9; 3.3.3 Requires-Python >=3.9; 3.4.0 Requires-Python >=3.9; 3.4.1 Requires-Python >=3.9; 3.8.0 Requires-Python >=3.9; 3.8.0rc1 Requires-Python >=3.9; 3.8.1 Requires-Python >=3.9; 3.8.2 Requires-Python >=3.9; 3.8.3 Requires-Python >=3.9; 3.8.4 Requires-Python >=3.9; 3.9.0 Requires-Python >=3.9; 3.9.0rc2 Requires-Python >=3.9; 3.9.1 Requires-Python >=3.9
ERROR: Could not find a version that satisfies the requirement tensorrt-libs==8.6.1 (from versions: 9.0.0.post11.dev1, 9.0.0.post12.dev1, 9.0.1.post11.dev4, 9.0.1.post12.dev4, 9.1.0.post11.dev4, 9.1.0.post12.dev4, 9.2.0.post11.dev5, 9.2.0.post12.dev5, 9.3.0.post11.dev1, 9.3.0.post12.dev1)
ERROR: No matching distribution found for tensorrt-libs==8.6.1
failed
CondaEnvException: Pip failed

Reviewer #3

(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 #4

(Remarks to the Author)

I have read the response letter and revised manuscript of the authors with great interest. Unfortunately, in general, I find the authors have not satisfactorily addressed the comments I raised in the previous version of the manuscript.

More specifically,

1. My main concern is still about the prediction performance depending on how much external factors change, like the inevitable variations in the composition of the inflow to the WWTPs. I could not find any real efforts from the Authors to address this critical limitation. The Response Letter simply states that they "decisively taken the choice not to include them" and that "our results show that it is not necessary to include anything and everything...because we were still able to predict the community dynamics accurately". In general, this claim cannot be supported. A sufficiently large change in the inflow composition will drastically change the environmental conditions, steering the composition and function of the microbial community between different attractors. Thus, the prediction performance should be low if the neural network is trained and validated using data for one attractor and tested to predict the community response in a different attractor. The authors somehow ignored this possibility. For example, Figure 3 of the main text shows that the neural network can accurately predict one specific training-validation-test set, with data ordered in time: the training data comes first, then the validation data, and finally, the test data. There are at least two possible explanations for this result. One is that across all time considered in this dataset (2014-2023), the changes in the inflow composition were small enough to allow the community to remain in the same attractor. A second possibility is that the temporal order of the training-validation-test is essential. That is, if the test data comes first in time and the training data lasts in time, the prediction accuracy decreases. Resolving this issue is essential because users may not apply the Author's method to this specific dataset. Users would need to know under which conditions the method can produce accurate predictions (e.g., how much environmental changes degrade the prediction accuracy). In other words, this is an issue of characterizing the conditions under which the method provides accurate predictions.

2. My second concern is how long the training time series must be to provide accurate predictions (Point 3 of my previous comments). In the Response letter, the authors claim this result appears in Figure 2 of the main text, where they compare the prediction accuracy for different WWTP datasets, each with a different number of samples. This claim is partially true, as the obtained prediction accuracy depends not only on the number of training samples but also on other factors like the specific ecological dynamics of the community and environmental changes. In this sense, this issue was not addressed. I expect an analysis showing how the prediction accuracy changes with the number of samples used for training for each WWTP dataset.

Overall, I feel rather bad for not having a more favorable comment about the Author's response and revised text. I still believe the methods have the potential and data have the potential to make a significant impact in the field, but a more careful validation is still required.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed all my previous concerns and comments satisfactorily. However, I believe that the formatting of the main figures still does not fully meet the standards required for publication in Nature journals. For instance, some axis labels are unclear. I recommend revising the figures in accordance with the journal's formatting guidelines.

(Remarks on code availability)

The code is publicly available, and the installation process is clearly documented and reproducible.

Reviewer #2

(Remarks to the Author)

We have reviewed the revised manuscript "Predicting microbial community structure and temporal dynamics by using graph neural network models". We do see substantial improvement after two rounds of revisions, which should definitely be commended. Most of our previous comments have been addressed. However, two questions remain to be addressed before the manuscript can be accepted.

1: This research is highly meaningful and shows promise for predicting filamentous bacteria outbreaks in advance, thereby allowing for timely intervention and maintain a system stability. However, a critical question arises: since interventions themselves can alter the system, can we still regard the system as the same after such interventions? In other words, is it valid to use pre-intervention data for future model training, or should we treat the post-intervention system as a fundamentally different WWTP, meaning the historical data is useless anymore? Clarifying this is essential for evaluating the practical utility and value of the proposed tool.

2: It remains unclear why, as depicted in Figure 6, the model's performance on ASV2 time series appears better when predicting 20 steps into the future compared to 10 or 15 steps. The authors explain in the rebuttal letter that this is due to inherent randomness, but the explanation is unconvincing. How many repeated runs were conducted for each prediction scenario? Additionally, could the authors provide a more robust justification or analysis to explain this counterintuitive outcome?

(Remarks on code availability)

Reviewer #3

(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)

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed our remaining comments.

(Remarks on code availability)

Reviewer #3

(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)

No further comments.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Reviewers' Comments:

Remarks to the Author: In their manuscript "Predicting microbial community structure and temporal dynamics by using graph neural network models", Andersen et al use a graph neural network model to predict the future abundance dynamics of the most abundant bacteria in the microbial communities of activated sludge. With the ongoing growth in microbial sequencing data, the next frontier is using these data to build models to predict future dynamics and trends of microbial communities to better manage huge microbial resources or control microbial process. From the goal perspective, this study is a valuable contribution. In addition, this research provides the longest continuous monitoring data of activated sludge systems to date, which provides a rich database resource for the study of temporal dynamics, microbial interactions, and community assembly. Nonetheless, I believe that the current version of this manuscript provides a rough description of the results and needs more refinement and revision. Please find below some comments and questions that should be addressed.

We are happy that the reviewer overall likes the study and the goal. We are also happy for several good and constructive comments.

1. The most important highlight of this paper is its predictive model "Graph neural network model". However, the description of this model in the manuscript is very general and simple. What is the framework or workflow of this model? What are the inputs and outputs as applied to this study? What are the requirements for time series data? For example, whether an equal time interval is required, whether there is a requirement for minimum sampling number, etc. Please spend at least 1 section in the Results section describing these details.

We have added more details about the graph models in the Results section, including info about inputs and outputs. We have also included an overview figure (Figure 1) to better explain the approach. For time series, an equal time interval is required for optimal prediction performance. Unfortunately, this is rarely possible in "real life". We have collaborated closely with numerous WWTP plant operators across the country for almost a decade to gather samples, and there are a multitude of challenges that make it very difficult to obtain samples with a consistent sampling interval. However, although not perfect, our time intervals are relatively stable, mostly once a week, and the model works well using this data. We have as requested added a section about the model details (line 135-145 and 167-170).

2. The authors have emphasized that they built the predictive model for each ASVs separately. In this case, I think $R^2_{1:1}$ is an effective parameter to evaluate the prediction accuracy (Averill, C., et al. 2021, Nat Ecol Evol; X. Liu., et al. 2023, Microbiome). Please show the $R^2_{1:1}$ results for each ASVs in your paper. Authors mentioned in their paper that they also used mean absolute error (MAE) and mean squared error (MSE) to

evaluate the model performance, but there are no related results both in the main manuscript and supplementary.

The model is actually trained on groups or clusters of 5 ASVs so we are sorry if this is not clear. We have made sure this is more clear now (line 134 and line 550-560). After training we show results from individual ASVs from whichever group they were assigned by the different clustering algorithms. Hence we need to use a loss function that takes into account other ASVs like the BC dissimilarity does and not single ASVs pairwise between real- and test dataset abundance values. Thus, we don't believe the R21:1 measure is suitable. Instead we have provided a supplementary Figure S5 that also shows the errors on the test datasets using the MAE and MSE metrics.

3. This paper predicted the temporal dynamics of 4,800 individuals in 24 datasets in total. More specifically, the Mariagerfjord dataset is used as an example to describe the prediction situation, which is acceptable. But, since the authors also admitted that process-critical bacteria in WWTP was very important, I think there should be a group summary of the prediction of process-critical bacteria in all datasets, whether in the main text or in the supplementary. The result description given in this manuscript is too rough. We agree that it is important to describe the prediction accuracies better for the process-critical bacteria. We have therefore calculated the prediction accuracies for each ASV in the functional groups for each dataset and shown them in a boxplot in the supplementary (Figure S6). We have also added a paragraph in the results section describing the essential results (line 243-255).

4. Only predicting the dynamics of abundant taxa may not be enough. It has been previously reported that rare taxa can also have an important impact on the structure and function of microbial communities (Lynch M D J, 2015, Nature Reviews Microbiology; Banerjee S, 2018, Nature Reviews Microbiology). What do the authors think about this?

This is a good point and we have now addressed this better in the paper. We have by purpose selected the abundant ASV as they are most important for the processes in the plants. The top 200 ASV constitute approx. 60 percent of the total reads, thus making up a significant fraction of the reads, roughly corresponding to biomass. We agree that the conditionally rare or abundant taxa (CRAT) can have a large impact on community structure and certain processes, which is also mentioned in the discussion. But the goal of our study is to demonstrate that prediction is possible using key and abundant bacteria as examples. Hence we also mention in the discussion that if certain additional taxa are also important to include, for example if suddenly detected in high abundance in the influent wastewater, one must filter differently than just filtering to the top 200 ASVs like we did. Furthermore, across our datasets, we have calculated the median abundance of the least abundant ASVs of the top 200 across all samples in each dataset to be between 0.03 and 0.09% relative abundance, which is, arguably, overall a low abundance. Our dataset is extremely large and consists of 76,555 unique ASVs. Training that many models is not feasible within a reasonable time frame, even with more compute resources available.

5. The authors selected 1000 reads as the threshold for filtering samples, which is too low for 16S rRNA amplicon sequencing. Please give a reasonable reason or re-analyze with deeper sequencing depth.

For this study it is not needed to capture the full diversity of every single sample as we focus on the most abundant ASVs. The threshold is a bare minimum. That said, the median number of reads for all samples is 40,519 and 75% of the samples have 25,000 or more total reads. We believe that this is more than enough. We have increased the threshold to 5000 and saw only a handful more samples filtered. The results were unaffected by the additional filtering. In the ms we have described the change in threshold (line 515).

6. How did you get graph network interaction strengths in this paper? Is there any difference between it and microbial interactions in traditional co-occurrence networks? If they are different, can the graph network interaction possibly replace the latter as a new method to explore microbial interactions?

This is a good point. We obtained the interaction strengths by using the graph method from the paper referred to in the manuscript. But it's not to be directly compared with co-occurrence. As the reviewer points out, the graph network is potentially superior to traditional co-occurrence networks, since it doesn't entirely compare abundances and whether they correlate positively or negatively over time. If the values of one ASV help to predict the future values of another ASV more accurately, the graph neural network will learn a stronger interaction strength for the two (or more) ASVs during model optimization. The graph network interaction strengths may complement or replace the latter as a new method to explore microbial interactions, which can be confirmed experimentally. We have not investigated this in more detail, but it would be of interest in a follow-up study. Furthermore, in response to another similar review comment 1.1 above, we have added a section in the results which describes the model design in brief (Line 135-145).

7. Most of the absolute values of the graph network interaction strengths illustrated in 'Figure S4' are very small. What's the meaning of it? Is there a relationship between interaction strength values and prediction accuracy of associated ASVs?

There is a "self-loop" for each ASV in the graph network which denotes how the previous values of one ASV will influence its future values most. So, most of the absolute values of the graph network interaction strengths between two different ASVs are small, but some can be significant. The interaction strength influences the prediction accuracy of associated ASVs. If the previous values of one ASV help to predict the future values of another ASV more accurately, the graph neural network will learn a stronger interaction strength for the two ASVs during model optimization. We have chosen not to go into this discussion in the main paper.

8. If the sample size of the test dataset is larger than the sample number that the trained model can predict at one time, then do you use the actual value or the predicted value of the previous sample to predict the subsequent samples?

We are not sure we understand the question correctly. But the training, validation, and testing sequences are always performed on moving windows of 10 real samples, not previously predicted samples, and then the model predicts a varying number of samples (10 unless otherwise noted) ahead from there. We have described the model design better and provided an illustration of how the 3-way split works (Figure 1 and line 135-145).

9. Please polish all figures in this manuscript.

We have tried to polish all figures, so they appear more concise and clear, also with a different colorblind-friendly color scheme that is consistent across all figures, e.g., train+validate+test datasets have the same colors etc. Sometimes lines and points seemed to be overlapping randomly as well, which we have addressed. We hope this is satisfactory.

Reviewer #1 (Remarks on code availability):

The code is available.

Reviewer #2 (Remarks to the Author):

In the manuscript entitled “Predicting microbial community structure and temporal dynamics by using graph neural network models”, the authors leveraged an existing graph neural network model to predict the abundance dynamics of the majority of bacteria in the microbial communities of activated sludge in wastewater treatment plants (WWTPs). They trained and tested models on time series datasets from 24 different Danish full-scale WWTPs sampled 2-5 times a month over a period of 3-8 years. They were able to predict the dynamics 10 time points into the future (2-4 months) with high accuracy for most of the abundant bacteria, especially process-critical bacteria, and some even 20 time points ahead (up to 8 months). The result sounds very promising and will certainly interest some researchers in the field of microbial ecology (especially those working on microbiome data of WWTPs). However, we have serious concerns about the novelty, rigor, generalizability, and interpretability of the presented method.

Our manuscript presents a concept, or workflow, and not a universal model which can be applied to all sorts of microbial times series. We are sorry this was not very clearly described. We want to describe and demonstrate a well-tested method (on 24 completely different datasets!) which anyone can apply on their own community time series datasets to train neural network models with which to predict the future abundances of the individual species present in the samples. The models are not pre-trained to fit other datasets, but rather trained specifically for the datasets themselves to

enable prediction of the same dataset. We think it is a significant strength of our study, demonstrating that continuous sampling and DNA sequencing can be performed, and the near future can be predicted from there on, however, only for that particular system on which it is trained. Please see our comments below.

1. Novelty

The presented prediction or forecasting method heavily depends on the published deep-learning model: MTSF-DG (multiple time series forecasting with dynamic graph modeling), which was cited as Ref.[24] in this manuscript. The authors did spend efforts to pre-process the microbiome data of WWTPs so that it could be fed into the MTSF-DG framework. But this significantly diminishes the novelty of this methodology work.

The novelty of this study is not the model development and design itself, though it has novel elements, but rather the combination of two completely different fields of science and the subsequent application of an advanced neural network model design on a very large microbial time series dataset. This has, for the first time, allowed accurate prediction of the future community dynamics at single ASV level with high accuracy. Previously, as mentioned in the introduction, other prediction methods have only focused on predicting a single time point snapshot of the future community, and not the dynamics over multiple time points, which is more difficult and also more important ecologically.

2. Rigor

2.1. The authors mentioned that "both stochastic factors (such as immigration) and deterministic factors (such as nutrients and predation) have a significant influence on the community structure [8–11], however, the relative contributions can vary." For temporal microbial dynamics in WWTPs, we believe that both stochastic and deterministic factors shape the community structure. But in this work, the authors didn't study the relative contributions of stochastic and deterministic factors at all. This raises a serious concern about the forecastability of the temporal microbial dynamics in WWTPs. We suggest that the authors should seriously quantify the forecastability of the temporal microbial dynamics in WWTPs before they apply any forecasting method. For example, they can compute the approximate entropy (ApEn) to quantify the forecastability of their time-series data [1].

We want to look into the community structure without any prior assumptions (see comments below, to 2.2). That also means that trying to investigate the relative contributions of stochastic and deterministic factors is not relevant in this study. Furthermore, our datasets show little statistical regularity and that is exactly the challenge that we tried to overcome in this study!

2.2. In this work, key environmental factors (e.g., temperature, nutrient concentrations, interventions, etc) were completely ignored. The authors mentioned that including them "may improve the prediction accuracy even further, but will also complicate the approach even more, let alone require more data to be gathered consistently and reliably." This is not a satisfactory explanation. Incorporating environmental factors is essential to

enhancing the model's robustness and reliability. For example, Larsen et al (2012) developed a model for predicting microbial community structure as a function of environmental parameters and microbial interactions [2]. Liu et al (2023), used environmental parameters to predict microbial community composition in AS [3].

The reviewer raises an important point. Nearly all models of communities in AS include some (usually few) environmental parameters. It is, however, nearly impossible to get times series of reliable data of these parameters at full scale WWTPs, and literally impossible to get more detailed data, e.g. on wastewater composition beyond overall measures for C, N and P, which are likely important for the competition between specific taxa. And what about all unknown parameters? Instead, here we present a model that is completely independent of all environmental parameters (whether stochastic or deterministic) and it works very well! The model only takes into account the "patterns" between ASVs that are captured during the training period from that particular WWTP. There seems to be so much information hidden in these patterns so they can be used to predict so well. This way of thinking is new, and we think it is the most important outcome of this study. When that is said, it may still be possible to improve the model and prediction by including some of these environmental parameters. See also our comment to reviewer 4. All this is now better described in the paper.

2.3. The authors mentioned that "the exact dates or duration between sampling points were not used for the model training as any potential extrapolation would not be reliable, so they have just been sorted chronologically." Does this simply imply that the authors treated those unevenly spaced time series as evenly spaced? This seems to be problematic.

The samples are sorted chronologically and then used sequentially. Thus, the exact duration between the sampling points are not important as long as they are relatively stable. This is described in M&M (line 524). This is very important because years of experience have taught us that it is very difficult to collect more than 5000 environmental samples at WWTPs and ensure a consistent duration between samples. And, as demonstrated by the resulting overall prediction accuracies, it is not necessary.

2.4. The authors mentioned that "To optimize the prediction accuracy of the model, four different pre-clustering methods were tested individually and a graph neural network model was designed (described in detail later) and then trained on each cluster". It is hard to believe that one has to perform pre-clustering and design/train the GNN for each cluster separately to achieve better prediction accuracy. Species from different clusters (regardless of how you pre-cluster them) might strongly interact with each other. Shouldn't we consider all the species in a single GNN to achieve better prediction accuracy?

As for the reviewer, this was also our initial thought. We have tried to consider all the species in a single GNN, but it worked worse than pre-clustering. The reason may be that if we consider all the species in a single GNN, it will result in a big neural network and there will be too many parameters to be trained. We need more data to train a big neural network to work well. This is mentioned in the discussion (line 407-10).

2.5. The authors mentioned that “the median Bray Curtis value across all of the test datasets was 0.147 indicating a good overall prediction accuracy.” Why does the median Bray Curtis value 0.147 indicate a good accuracy? With respect to what? How will traditional forecasting methods (e.g., autoregressive integrated moving average, or ARIMA) perform?

The Bray-Curtis measure is a very common dissimilarity measure used within ecology to measure how dissimilar two multivariate samples are (<https://doi.org/10.1016/j.ecolmodel.2022.109880>). It's a range between 0-1, where 0 means completely identical. Here, we simply use it to express the difference between the true and predicted samples in the test data set for each cluster, and the models are then trained to obtain a Bray-Curtis value as low as possible. ARIMA is a different thing, it's another way to predict, not a way to measure accuracy. We do understand the concern that the reviewer expresses here that it can be a somewhat arbitrary decision exactly what BC value corresponds to a “good” accuracy. However, it's somewhat insufficient to use a single value to represent the prediction accuracy for the whole concept. The reader must be presented with different perspectives on the prediction accuracy; hence we present multiple figures to illustrate different aspects of the performance. It is a challenge to illustrate concisely how good the overall prediction accuracy is in a single figure or value, so we have rephrased in the manuscript (line 198-201).

2.6. Is the temporal variations of microbial compositions within the same WWTP bigger or smaller than 0.147? Is it significant? What is the Bray-Curtis dissimilarity between true samples from the same WWTP but collected at different time points? This skepticism is further supported by Figure 4 (now Figure 5), where the distances between predicted and true samples appear comparable to those observed between samples collected at different times. It is imperative to conduct a thorough assessment of the dissimilarity among true samples over time and across different WWTPs to substantiate the judgment of the 0.147 value as indicative of prediction accuracy.

Similarly to the previous comment, we agree that it may seem like an arbitrary decision that a BC value of 0.147 will correspond to a good prediction accuracy. As mentioned in the previous comment, we have rephrased this part in the manuscript. Regarding Figure 5, there is a small level of uncertainty between consecutive samples, however, during just a few consecutive (=chronological) samples it's clear that the model captures the overall trend of the entire community. Keeping in mind that each predicted point is 10 points into the future, we find this impressive.

3. Generalizability

The authors trained and tested their model on time series datasets from 24 different WWTPs separately and independently of each other. Moreover, they mentioned that “During the design and implementation of the neural network model there were also many decisions made that were ultimately governed by trial and error until no more

significant improvements in prediction performance were obtained across the datasets overall.” This raises a serious concern on the generalizability of their model. We don’t think the model is easily applicable to other microbial time-series datasets. There are too many trial-and-error and parameter tuning.

The reviewer is correct. A new model should be made on each WWTP separately and trained on data from that particular plant. We show in this study that this is indeed possible, and with very good results. We do, however, not think it is a big issue to optimize the model for each plant by trial and error and that is just how machine learning model tuning is done. Furthermore, we have the concept freely available on github and it can easily be adopted by anyone working with times series in WWTP or other similar systems. The intention was not to make a general model (as we don't think it is possible), only a concept and pipeline. This is described in the introduction and in the discussion (line 104-110 and 371).

4. Interpretability

4.1. The authors clearly demonstrated that the predictions for some ASVs were more accurate than other ASVs. The authors also mentioned that “ it was impossible to predict every single species in the community precisely with the approach and data used in this study.” But why?

This is a good question. We don't know, but we simply observe that some ASVs are not very well predicted. However, for the same ASV, it varies from plant to plant (dataset to dataset), so there may be irregularities. This is now better described in the paper (line 415).

4.2. The authors mentioned that “it is clear that clustering by biological function generally resulted in a lower prediction accuracy compared to the other clustering methods, except for the Ejby Mølle and Hirtshals datasets.” Why does the clustering by biological function generally result in a lower prediction accuracy?

We were not very precise in our wording. We should not have written “it is clear...”. We should have described the observation and then used the following explanation: Same functional groups do have the same overall function, e.g. nitrification or denitrification, but the different species within each group do, however, often have different ecophysiological traits. Thus, the factors controlling their presence may be very different. A good example is the PAOs, the polyphosphate accumulating organisms. Some species consume acetate, others carbohydrates, some can denitrify, others not (Singleton, et al., 2022, ISME J. doi 10.1038/s41396-022-01212-z. and Petriglieri, P., et al., . ISME J. doi.org/10.1038/s41396-021-01029-2.). From our results we can see that these groups may not interact in a way that is captured by the model and thus not superior to other groupings. This is now made more clear in the paper (line 405).

4.3. The authors mentioned that “the interaction strengths in the graph neural network could also be interpreted as hypothetical biological interactions between the individual

ASVs in the group (see Figure S4).” This is really far-fetched. The GNN is undirected or bidirectional. The interactions between ASVs do not have to be bidirectional.

We used the undirected GNN and found it helps the prediction, which only learns bidirectional interaction strengths between ASV for now. However, the undirected GNN in our code can be replaced by a directed GNN, to learn directional interactions between ASVs, and we can provide a version of codes with a directed GNN for further research if needed.

4.4 As depicted in Figure 6 (now Figure 7), surprisingly, the model performance on the time series of ASV2 for 20 predicted samples into the future is even better than 15 or 10 predicted samples into the future. This is quite counter-intuitive and deserves an explanation.

We agree that this is counter-intuitive, and we discussed this more in the paper (line 413). We believe the differences in the accuracy referred to here are marginal. These minor differences are mainly due to the inherent randomness and heuristics of neural networks. So, the same input data under the same conditions and settings will not yield completely identical results every time. ASV2 is one of the most abundant and active ASVs across all datasets and has been very easy to predict in most of the 24 datasets. In this particular example, slight differences are expected when the prediction accuracy is so similar between the 5 examples. Lastly, Figure 6 shows a clear trend between prediction window length and accuracy, however there are overlaps for some groups.

References:

[1] Pincus, S. M.; Gladstone, I. M.; Ehrenkranz, R. A. (1991). "A regularity statistic for medical data analysis". *Journal of Clinical Monitoring and Computing*. 7 (4): 335–345. doi:10.1007/BF01619355.

[2] Larsen, P., Field, D. & Gilbert, J. Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* 9, 621–625 (2012).
<https://doi.org/10.1038/nmeth.1975>

[3] Liu, X., Nie, Y. & Wu, XL. Predicting microbial community compositions in wastewater treatment plants using artificial neural networks. *Microbiome* 11, 93 (2023).
<https://doi.org/10.1186/s40168-023-01519-9>

Reviewer #2 (Remarks on code availability):

Yes. But we cannot install their code. There are more than 180 dependency packages, and most of them have incompatible issues with each other.

We have updated a few tools in the environment.yml file provided in the GitHub repository, which is used to create the conda environment for the project. We have tested the creation of the environment and ensured that it works correctly on a few platforms and versions, including Ubuntu 20, 22, and 24 as well as Windows 11. We have also

updated and tested the Docker container in the same manner and updated the readme accordingly.

Also, there is an inconsistency regarding the use of sampling time as input versus merely arranging the sample points chronologically. Specifically, on page 15, lines 469-470, it is noted that samples were sorted chronologically without considering the actual dates or the duration between sampling points. However, in the GitHub documentation, one of the required inputs is metadata that must include sampling dates in the year-month-day format.

The date information is not used at all during training, validation, and testing of the actual models. That is performed purely sequentially. It's only used for giving the predicted values an estimated time stamp depending on the sampling intervals in the real data, and thus only used when outputting the predicted data and for plotting for the manuscript. This is now better explained in the paper (line 524).

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports as part of the Nature Communications initiative to facilitate training in peer review and appropriate recognition for co-reviewers.

That is great!

Reviewer #4 (Remarks to the Author):

The authors propose using a graph neural network (a machine-learning model) to forecast species abundance in biological wastewater treatment plants (WWTPs). I find the main merit of the paper in the dataset the authors collected for 24 full-scale WWTPs in Denmark. As the authors emphasize, this dataset is exceptional in length and coverage. On the other hand, I have several reservations about the author's prediction approach and validation of graph neural networks. My main reservation is that the authors ignore that the composition of the inflow to the WWTPs changes with time. For example, this might render predictions impossible if the inflow exhibits drastic changes. I comment on this below.

Main comments.

1. The ecology WWTPs depends on internal factors like the population dynamics and interactions between species (which, with some reserve, may be captured by relative abundance time-series). However, its ecology also depends on external factors like the composition of the inflow to the WWTPs. In particular, because this inflow carries both organic and inorganic materials, it can drastically change community composition.

Somewhat unexpectedly, the authors perform predictions ignoring all external factors. From this perspective, some claims made by the authors, like "we are able to predict the dynamics 10 time points in the future (2-4 months) with high accuracy," seem not well-posed because the prediction accuracy depends on keeping all those external factors constant. Note these external factors must remain constant during the prediction and across the data used to train the algorithm. This is a critical limitation that must be studied and addressed in detail.

This is an important comment and please also see our comment to reviewer 2 (2.2). Our concept and the resulting prediction accuracy do not depend on keeping any external factors constant. We have decisively taken the choice not to include any of them, so this is one of the novel and unique things about the concept. It's simply not feasible to measure all possible factors, of which there are potentially hundreds, and if we choose to include them in the model, we would essentially only train the models to predict them instead, and not the microbial community dynamics. Secondly, the microbial species are to be considered response variables because their dynamics are directly dependent on the environment including the activity of other species also including those coming with the influent wastewater. As some of our datasets span more than a decade, drastic changes in the influent wastewater have happened many times. Our results show that it is not necessary to include anything and everything, because we were still able to predict the community dynamics accurately without including any other data than the community abundance data alone. This is what makes this concept novel, and we are fully aware that this stands in direct contrast with the traditional ecological concepts and methods, where as many environmental factors as possible are included. But for prediction, we have just shown that it is not at all necessary. This is now better described in the paper.

2. I have several comments regarding Figure 1 (now Figure 2):

2a. How the authors calculated a single Bray-Curtis dissimilarity needs to be clarified, considering multiple species (ASVs) and time points exist. Is this the Bray-Curtis dissimilarity across species and time?

The error metrics are calculated per cluster/group (5 ASVs) per time point on the test dataset. This has now been clarified in the Methods section.

2b. In the same line as above, it is difficult to get insights from Figure 1 (now Figure 2) because of the difficulty of interpreting the derived Bray-Curtis dissimilarity. For example, across datasets, the authors report a Bray-Curtis dissimilarity of about 0.15. What does this mean concretely for the prediction accuracy? What about the significant variations represented by error bars and outliers, which in some cases are above 0.7 (e.g., Odense NE and Ribe)? Do they correspond to some particular ASVs or time instants? It would be much more informative to present results for best, typical, and worst predictions for ASVs, one for each dataset.

To evaluate the model performance, we used three different loss functions to calculate the error for each 5 ASV cluster per time point between the true and predicted values. We selected the Bray-Curtis dissimilarity metric for its ecological relevance, as well as the mean absolute error (MAE) and mean squared error (MSE). Figure 2 in the main manuscript shows the model performance using Bray-Curtis's dissimilarity, while the full set of results can be found in Supplementary Figure S5.

The values obtained from these loss functions are empirical parameters without direct physical meaning. We observed that the median accuracy across datasets was 0.147 with a low spread (chosen Graph model), indicating good to very good predictions (e.g., Figures 3 and 4). Currently, we cannot determine if this threshold is generalizable due to the scarcity of multiyear, frequently sampled datasets. We have made this clear in the manuscript (lines 197)

Since the loss function is calculated per cluster of 5 ASVs, identifying specific outliers is challenging. Therefore, we evaluated and presented the overall prediction accuracy metrics. For further analysis, we select ASVs of interest and visualize the real versus predicted data in a time-series plot or heatmap (Figures 3 and 4). Please, note that the acceptable level of error depends on the study's aims. For our ASVs, the median mean absolute percentage error (MAPE) between real and predicted values on the test datasets ranged from 14.2% to 17.7%. For example, if the relative abundance in a real sample was 10.0% and the MAPE value was 15.0%, the predicted relative abundance would be between 8.5% and 11.5%. This information is also clarified in the manuscript (lines 250).

3. An essential point that needs to be added to the study is how long the training time-series must be to provide accurate predictions. In particular, it would be handy to have a plot showing test prediction error as a function of the length of train time-series for each WWTP.

We agree that this is essential information. However, we believe that we have already shown this by indicating the number of samples in each dataset (Figure 2 and Figure S5) and discussed it, so it is possible for the reader to evaluate the performance on the 24 unique and different datasets.

4. Regarding Figure 2 (now Figure 3), it would be much more informative if it presented ASVs in the top, middle, and bottom quantiles of prediction error.

The purpose of the figure is to show examples of the prediction accuracy of some of the most important bacteria in wastewater treatment, so we have not selected ASVs based on prediction error or abundance, but rather whether they are known to contribute to important biological processes.

5. Graph neural networks are not constructed by default to handle compositional data (i.e., relative abundances). How do the authors handle this problem to ensure that predictions remain compositional?

Our graph neural network implementation has fully connected neural network layers to be able to transfer all features and predict compositional data. So any characteristics of the input data will also be present in the output/predicted data, including compositionality. Furthermore, as mentioned in the Methods section, we also tested various ways of transforming and preprocessing the data before settling on division by the mean.

6. In Figure 1 (now Figure 2) and Figure 5 (now Figure 6), the authors emphasize differences in prediction when using different clustering methods. However, the boxplots in both figures are very similar across methods. Are differences statistically significant? Also, to improve the presentation, I would recommend considering only one clustering method to present the results in the main text, leaving the other methods as a supplementary discussion.

We agree that there is not always a big difference between the different clustering methods. However, we chose to still show results from all clustering methods in Figure 2 and afterwards only from graph clustering, mainly because we wanted to be able to investigate and discuss whether the prediction accuracy could be improved by grouping together species carrying out the same known biological function.

Minor comments

7. In Figure 1 (now Figure 2). It is worth noting that in the caption of Figure 1 (now Figure 2) the authors write "Bray Curtis distance". However, mathematically, the Bray-Curtis dissimilarity is not a distance because it does not satisfy the triangle inequality.

Thank you, yes, we do use BC dissimilarity, not BC distance, so we have corrected the figure caption.

** See Nature Portfolio's author and referees' website at www.nature.com/authors for information about policies, services and author benefits.

This email has been sent through the Springer Nature Tracking System NY-610A-NPG&MTS

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

1. I don't think the description of the model is clear enough. In fact, Figure 1 merely introduces the dataset division and provides a basic overview of the model. This is a very general overview of deep learning, applicable to almost all machine learning models. However, the training and prediction based on ASV clusters, which is the highlight of this study, is not reflected in Figure 1.

Answer: We agree that the figure was not very informative. We have now included a much more detailed figure (Figure 1) and legend to describe the approach and the model.

Manuscript change/addition:

- Figure 1 and figure caption, page 4, lines 155-165

2. Although your model is trained on each cluster of 5 ASVs, you will eventually obtain prediction values for each ASV, as you mentioned in Line 129 and Figure 4. The $R^2_{1:1}$ is suitable for evaluating the prediction accuracy of each ASV, so I believe it is applicable to your study.

Answer: Thank you for the advice. We have evaluated the prediction accuracy by the $R^2_{1:1}$ approach for each ASV and have now included an extra margin plot with real versus predicted values and corresponding regression lines and R^2 values to the time series example plots shown in Figure 3 for 3 ASVs, also in Figure 7 and Figure 8 (new figure testing a human gut timeseries) and supplementary Figures S10 and S11 (new figures). We agree that it is an intuitive way to show the prediction accuracy compared to real values. We would have liked to include the $R^2_{1:1}$ values in the boxplot in Figure 2, but it shows the accuracy per cluster/group of ASVs calculated during model training, and not individual ASVs, so they are not compatible with each other.

Manuscript change/addition:

- Figure 3 and figure caption, page 8, lines 285-295
- Figure 7 and figure caption, page 12, lines 380-385
- Figure 8 and figure caption, page 14, lines 420-435
- Supplementary Figures S10 and S11
- Manuscript text addition, page 7, lines 275-280

3. The manuscript gives so many summary results, including some overview and boxplot graphs. Please share your detailed data and result tables as supplementary so that others can reproduce your results.

Answer: We agree that there are many summary data in addition to some more detailed results. It has been a challenge throughout the study to present concise results for every important

aspect and show that our approach works on not just a single time series from one WWTP, but that it works for all the bacteria in 24 WWTPs, as well as the 3 other datasets we now have included from other studies and environments. For that reason, it is important to make all data available for the reader, but there is just too much data to put them into a supplementary document. From the first version of the manuscript, we have made everything available in the github code repository. We think this is the best way to provide the data and code to the reader for completely reproducibility and transparency. The github code repository already includes both the input data as well as results from all 24 WWTPs (for 200 ASVs in each plant) and code to produce all figures and statistics shown throughout the manuscript. The larger model parameters obtained through the model training for all the datasets can be downloaded separately from figshare, also as described in the readme under "article analysis".

Manuscript change/addition:

Page 20, lines 675-680

4. The manuscript should also be checked for accuracy of the language, as some parts read too colloquial.

Thank you for the advice. We have done so throughout all the manuscript.

Reviewer #1 (Remarks on code availability):

The code provides a README file with enough instructions for installing and running the application. It's clear enough. But I have no time to reproduce their results.

Reviewer #2 (Remarks to the Author):

1. The authors admitted that this study "presents a concept, or workflow, and not a universal model which can be applied to all sorts of microbial times series". This significantly diminish the impact and importance of this work. Moreover, they also claimed that their workflow, "mc-prediction", suitable for any longitudinal community dataset, not limited to the WWTP ecosystem. Is this a self-contradiction?

Answer: We did not express ourselves clearly enough, sorry about that. By concept, or workflow, we mean that we can provide any microbial time series into the pipeline and then make the predictions for that particular ecosystem (e.g., one WWTP), but not universal model for an entire ecosystem type (e.g. interchangeable for any WWTP). What we cannot do is to use a training and model and apply it to a different ecosystem (here another WWTP) for prediction. It is not possible as many (or all) bacterial species (ASVs) will be different. If the same ASVs are present in

a different site, there may be other types of dependencies between the ASVs. Thus, it is not possible to make one universal model that can be used for predictions in all 24 WWTPs – they must be trained separately on time series from each WWTP because complex communities are unique with their own dynamics depending on a multitude of other factors than the community alone.

We have now made this clearer in the manuscript. We also demonstrate how it performs on three other datasets from other studies of other environments. We have tested it on another WWTP data from Hong Kong, China and a dataset from the human gut, where our workflow works well, comparable accuracy to our own 24 WWTP datasets. Lastly, a fungal time series from a coastal marine environment, where the workflow unfortunately performed poorly, but the causes are discussed, and we list key requirements to the input data in order to obtain accurate predictions.

Manuscript change/addition:

- Page 1, line 40
- Page 3, lines 110 – 120
- Paged 13-14, lines 390 – 435
- Figure 8, page 14, lines 425-435
- Supplementary Figures S10 and S11

2. They claimed that this is the first application of predicting future community dynamics. This is certainly not true. Various deep learning models (especially those based on recurrent neural networks) have already been applied to predicting microbiome dynamics. Here are just two recent references: (1) Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics, Mayank Baranwal et al., eLife 2022; (2) Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions, Jaron C. Thompson et al., PLoS Computational Biology 2023. To our surprise, none of those efforts was mentioned in this paper. Why?

Answer: We appreciate the opportunity to clarify why papers such as Baranwal et al. (2022) and Thompson et al. (2023) did not initially appear in our citations. These studies did also employ neural networks to predict microbial systems, but there are important fundamental differences in experimental context and the objectives of the predictions.

Baranwal et al. (2022) focused on predicting microbial abundances and metabolite production during incubations of synthetic communities. These incubations are closed systems where microbial abundances change dynamically until substrate depletion and/or steady growth rates are achieved. Similarly, Thompson et al. (2023) used the data from Baranwal et al., and from

Clark et al. (2021), to develop their neural network model. Both studies applied methods that focus on predicting transient short-term dynamics.

In contrast, our work addresses long-term microbial dynamics from full-scale treatment plants, which differ conceptually and operationally from incubation experiments. Full-scale treatment plants are open, continuous systems that involve the continuous addition of nutrients and microbial populations, creating dynamic conditions that are influenced by multiple, usually unknown factors. Therefore, our objective is to predict the species dynamics over weeks (or months), solely relying on past microbial relative abundances from the same site. We hope this brief explanation is sufficient to show that our goals are inherently distinct from the transient dynamics explored in the cited studies.

While we recognize the efforts of the research community, please note that there is a growing literature on the use of neural networks for time-series predictions. Our intent in this publication is not to provide an exhaustive review of all methodologies across all experimental contexts. Instead, we focused on a novel application. To the best of our knowledge, this is the first application of neural networks to predict long-term microbial dynamics over extended timescales in real-world, full-scale continuous systems.

However, as recommended, we have now included these references in the introduction and also made clear that there are fundamental differences to our study.

Manuscript change/addition:

- Page 8, line 84

3. It is good to see Figure 1 and the added details describing the data. However, the model details should also be included in Figure 1 or in a separate figure to address questions including but not limited to the number of layers in the design and the parameters for each layer. See answer above to reviewer 1. We have now included a much more comprehensive figure.

Manuscript change/addition:

- Figure 1 and figure caption, page 4, lines 155-165

4. We still have reservations about the "good" performance of the model. When the Bray-Curtis dissimilarity 0.147 is arbitrarily regarded as "good or very good", in Figures 3 and 4, it is preferable to estimate numerical parameters to express the model's performance. Comparisons with previous studies, other models, and traditional co-occurrence models would also be supportive. In its current version, the evidence does not seem sufficient to justify statements in the manuscript claiming it is "good or very good."

Answer: We agree that it may seem like an arbitrary decision to conclude about the prediction accuracy, but we believe that we have to show at least a numerical measure to express the overall prediction accuracy. We have presented the results in several different ways to let the reader decide how accurate it really is to not rely only in Bray-Curtis dissimilarity. As recommended, we now also show another type of accuracy measure for the individual ASV timeseries, which is simply the $R^2_{1,1}$ value of a linear regression with residuals calculated relative to the 1.0-slope line. This clearly indicates the numerical difference between the observed/historical values and the predicted values in a simple 2-axis plot next to every timeseries plot.

The lack of other literature predicting studies on comparable microbial systems limits our ability to judge whether a dissimilarity threshold can be proposed to define the prediction accuracy. Moreover, what is a “good” level of accuracy is ecosystem-specific. For example, in activated sludge, it is normal to see 5-10% variations in the experimental data which do not cause any difference from an operational perspective. In that case, a prediction might be judged “good” while for a different application a more restrictive threshold can be needed.

Manuscript change/addition:

- Figure 3 and figure caption, page 8, lines 285-295
- Figure 7 and figure caption, page 12, lines 380-385
- Figure 8 and figure caption, page 14, lines 420-435
- Supplementary Figures S10 and S11
- Text addition, page 7, lines 275-280
- Text addition, page 16, lines 501 – 513

5. We believe the authors misunderstood the question regarding the time interval. The question is not about whether the time interval should be the same or not, but rather whether time stamps should be used instead of just sequential datasets. A test on this should be added. What are the time units (10 days later? 10 weeks later? ...) for the predicted outcomes for a model like Mariagerfjord, using temporal sequential datasets ignores the time interval between points.

Answer: Thank you for the clarification. We understand that the concern arises from the use of unevenly sampled datasets, where the data is ordered sequentially. Following your comment, we have tested and added an option to include the time interval between samples as an additional variable to use for model training.

We have also tested this on all 24 datasets and saw almost no differences, only a slightly decreased accuracy, see Figure S7. The median Bray Curtis values across all datasets was 0.151 compared to previously 0.147. As no practical improvement was observed, we have decided to retain the model in its original form. However, we have updated the manuscript to indicate that,

for this dataset, using timestamps or time-intervals provides similar results, but users have now the option to test and choose what might work best for their specific dataset.

We would also like to reinforce that while the sampling frequency is irregular due to the constraints and challenges of sampling environmental systems, however it was kept within a narrow variation. Additionally, based on knowledge from continuous systems, particularly wastewater treatment plants, microbial communities in such open, mixed continuous systems exhibit some inertia. The biomass retention time is typically 20-30 days, and changes within the microbial community take some time to be measurable. Therefore, variations occurring within a narrow window generally do not lead to major differences, and the potential information loss from using relative timestamps is justified by the ability to use a simpler model.

Regarding the interpretation of the outputs as you point out, the model's predictions are based on relative timestamps. However, the average sampling frequency for each site is known. For each individual site, the average sampling frequency ranges from 7 to 14 days. If we take the Mariagerfjord dataset as an example, samples were typically collected every 7 days. If the model can predict well 10 sequential relative timestamps, this would correspond to predicting outcomes for around 70 days in the future. While we acknowledge that this requires an additional step to interpret the results in terms of absolute time, we believe that using relative timestamps in combination with the average sampling frequency is a reasonable solution, given all the challenges discussed above.

Workflow change:

- Addition of time interval as optional feature in the mc-workflow

Manuscript change/addition:

- Pages 10 – 11, lines 350 – 355
- Page 16 – 17, lines 532 – 538
- Supplementary Figure S7

6. They claim their workflow, “mc-prediction”, suitable for any longitudinal community dataset, not limited to the WWTP ecosystem. We don't see any clear evidence that it works to any other temporal data of microbiome. The authors should just demonstrate it, if they claim the general applicability of their workflow.

Answer: Thank you for the suggestion. This is also questioned by reviewer 2, q1, see our response there to avoid duplication. We have now demonstrated how the workflow performs on 3 other datasets from other environments, the human gut, a coastal marine fungal ecosystem (although with some limitations), and a different wastewater treatment plant from Hong Kong.

Manuscript change/addition:

- Paged 13-14, lines 390 – 435
- Figure 8, supplementary figures S10 and S11
- Page 15, lines 455 – 461
- Page 17, lines 553 – 565

In sum, the quality of the presented work does not meet the high standards of Nature Communications. A heavily revised version of this manuscript should be submitted to a more specialized journal on WWTP.

Answer: We hope the present revised version will satisfy the reviewers. We think we have addressed all concerns, and we are grateful for the helpful and constructive advice.

Reviewer #2 (Remarks on code availability):

We still failed to install the tool successfully. We got the following errors (Linux system: Python 3.10.14; CentOS Linux 7):

Pip subprocess error:

ERROR: Ignored the following versions that require a different python version: 1.11.0 Requires-Python <3.13,>=3.9; 1.11.0rc1 Requires-Python <3.13,>=3.9; 1.11.0rc2 Requires-Python <3.13,>=3.9; 1.11.1 Requires-Python <3.13,>=3.9; 1.11.2 Requires-Python <3.13,>=3.9; 1.11.3 Requires-Python <3.13,>=3.9; 1.11.4 Requires-Python >=3.9; 1.12.0 Requires-Python >=3.9; 1.12.0rc1 Requires-Python >=3.9; 1.12.0rc2 Requires-Python >=3.9; 1.13.0 Requires-Python >=3.9; 1.13.0rc1 Requires-Python >=3.9; 1.13.1 Requires-Python >=3.9; 1.14.0 Requires-Python >=3.10; 1.14.0rc1 Requires-Python >=3.10; 1.14.0rc2 Requires-Python >=3.10; 1.2.0 Requires-Python >=3.9; 1.2.1 Requires-Python >=3.9; 1.2.1rc1 Requires-Python >=3.9; 1.25.0 Requires-Python >=3.9; 1.25.0rc1 Requires-Python >=3.9; 1.25.1 Requires-Python >=3.9; 1.25.2 Requires-Python >=3.9; 1.26.0 Requires-Python <3.13,>=3.9; 1.26.0b1 Requires-Python <3.13,>=3.9; 1.26.0rc1 Requires-Python <3.13,>=3.9; 1.26.1 Requires-Python <3.13,>=3.9; 1.26.2 Requires-Python >=3.9; 1.26.3 Requires-Python >=3.9; 1.26.4 Requires-Python >=3.9; 1.4.0 Requires-Python >=3.9; 1.4.0rc1 Requires-Python >=3.9; 1.4.1.post1 Requires-Python >=3.9; 1.4.2 Requires-Python >=3.9; 1.5.0 Requires-Python >=3.9; 1.5.0rc1 Requires-Python >=3.9; 1.5.1 Requires-Python >=3.9; 2.0.0 Requires-Python >=3.9; 2.0.0b1 Requires-Python >=3.9; 2.0.0rc1 Requires-Python >=3.9; 2.0.0rc2 Requires-Python >=3.9; 2.0.1 Requires-Python >=3.9; 2.1.0 Requires-Python >=3.9; 2.1.0rc0 Requires-Python >=3.9; 2.1.1 Requires-Python >=3.9; 2.1.2 Requires-Python >=3.9; 2.1.3 Requires-Python >=3.9; 2.1.4 Requires-Python >=3.9; 2.14.0 Requires-Python >=3.9; 2.14.0rc0 Requires-Python >=3.9; 2.14.1 Requires-Python >=3.9; 2.15.0 Requires-Python >=3.9; 2.15.1 Requires-Python >=3.9; 2.15.2 Requires-Python >=3.9; 2.16.0 Requires-Python >=3.9; 2.16.1 Requires-Python >=3.9; 2.16.2 Requires-Python >=3.9; 2.17.0 Requires-Python >=3.9; 2.2.0 Requires-Python >=3.9; 2.2.0rc0 Requires-Python >=3.9; 2.2.1

Requires-Python >=3.9; 2.2.2 Requires-Python >=3.9; 3.0.0 Requires-Python >=3.9; 3.0.1 Requires-Python >=3.9; 3.0.2 Requires-Python >=3.9; 3.0.3 Requires-Python >=3.9; 3.0.4 Requires-Python >=3.9; 3.0.5 Requires-Python >=3.9; 3.1.0 Requires-Python >=3.9; 3.1.1 Requires-Python >=3.9; 3.2.0 Requires-Python >=3.9; 3.2.1 Requires-Python >=3.9; 3.3.0 Requires-Python >=3.9; 3.3.1 Requires-Python >=3.9; 3.3.2 Requires-Python >=3.9; 3.3.3 Requires-Python >=3.9; 3.4.0 Requires-Python >=3.9; 3.4.1 Requires-Python >=3.9; 3.8.0 Requires-Python >=3.9; 3.8.0rc1 Requires-Python >=3.9; 3.8.1 Requires-Python >=3.9; 3.8.2 Requires-Python >=3.9; 3.8.3 Requires-Python >=3.9; 3.8.4 Requires-Python >=3.9; 3.9.0 Requires-Python >=3.9; 3.9.0rc2 Requires-Python >=3.9; 3.9.1 Requires-Python >=3.9
ERROR: Could not find a version that satisfies the requirement tensorrt-libs==8.6.1 (from versions: 9.0.0.post11.dev1, 9.0.0.post12.dev1, 9.0.1.post11.dev4, 9.0.1.post12.dev4, 9.1.0.post11.dev4, 9.1.0.post12.dev4, 9.2.0.post11.dev5, 9.2.0.post12.dev5, 9.3.0.post11.dev1, 9.3.0.post12.dev1)
ERROR: No matching distribution found for tensorrt-libs==8.6.1
failed
CondaEnvException: Pip failed

Answer: We are sorry that you had troubles here. We indicated in the readme file that you need to add an extra pipy index url because some libraries are proprietary nvidia libraries with a special EULA. If you run "export PIP_EXTRA_INDEX_URL='https://pypi.nvidia.com'" before building the conda environment, it should work. Otherwise, we suggest that you try using the pre-built Docker container instead, which is also what we use for producing data and results used throughout the manuscript to ensure reproducibility.

Reviewer #3 (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.

Reviewer #4 (Remarks to the Author):

I have read the response letter and revised manuscript of the authors with great interest.

Unfortunately, in general, I find the authors have not satisfactorily addressed the comments I raised in the previous version of the manuscript.

Answer: We are sorry to hear that. Based on the new advices we have got from you and the other reviewers, we have addressed all comments again, and we hope, in a satisfactory way. We agree with you that the methods and data have the potential to make a significant impact in the field.

More specifically,

1. My main concern is still about the prediction performance depending on how much external factors change, like the inevitable variations in the composition of the inflow to the WWTPs. I could not find any real efforts from the Authors to address this critical limitation. The Response Letter simply states that they "decisively taken the choice not to include them" and that "our results show that it is not necessary to include anything and everything...because we were still able to predict the community dynamics accurately". In general, this claim cannot be supported. A sufficiently large change in the inflow composition will drastically change the environmental conditions, steering the composition and function of the microbial community between different attractors. Thus, the prediction performance should be low if the neural network is trained and validated using data for one attractor and tested to predict the community response in a different attractor. The authors somehow ignored this possibility. For example, Figure 3 of the main text shows that the neural network can accurately predict one specific training-validation-test set, with data ordered in time: the training data comes first, then the validation data, and finally, the test data. There are at least two possible explanations for this result. One is that across all time considered in this dataset (2014-2023), the changes in the inflow composition were small enough to allow the community to remain in the same attractor. A second possibility is that the temporal order of the training-validation-test is essential. That is, if the test data comes first in time and the training data lasts in time, the prediction accuracy decreases. Resolving this issue is essential because users may not apply the Author's method to this specific dataset. Users would need to know under which conditions the method can produce accurate predictions (e.g., how much environmental changes degrade the prediction accuracy). In other words, this is an issue of characterizing the conditions under which the method provides accurate predictions.

Answer: The goal of our model is to provide accurate predictions for species' temporal dynamics using time-series data from full-scale/real-world systems. As you point out, and we fully agree, the chronological order of the data is a crucial factor, as the model learns the temporal patterns and microbial interactions from the time-series training set, using the logic that past events can be used to predict future events. While the idea of shuffling the data to do a test-first train-last

can reveal some aspects of temporal dependency, we believe that it is well established that models based on time-series data exploit the temporal relationships to provide accurate predictions. A deep learning model is always trained using the historical data and then inferred on the future data. In the testing, it is always tested on the most recent data.

In terms of the influence of external factors, we would like to use this opportunity to expand our explanation. As you are aware, our dataset includes 24 individual WWTP over a very long timespan (up to 8 years). During that period, the WWTPs have experienced what could be defined as “standard variations”. These include variations in the influent flow and concentrations, seasonal temperature fluctuations, rain events, occasional or regular industrial discharges (depending on the site), occasional addition of chemicals to aid denitrification, phosphate removal or to control foaming, or occasional equipment malfunctions. Nevertheless, the WWTPs have consistently achieved their main goal of producing an effluent that complies with the legislation, and it is in their best interest to avoid operating under extreme changing conditions that might degrade their performance.

Based on operational data available for two WWTP over 2-3 consecutive years (it can be found in Peces et al., 2022, <https://doi.org/10.1038/s43705-022-00098-4> supplementary material), the median variation for the measured parameters is about 98 - 127% relative to their average, with individual parameters ranging from 5 to 670% variation depending on the WWTP and parameter. While such variations might seem large enough to significantly steer microbial composition, particularly from controlled experiments where individual factors may show strong impacts, it is not that clear full-scale systems. In full-scale systems multiple variables can interact simultaneously, with synergies and antagonisms complicating the dynamics. Besides, some variables are reasonably measured, but not all of them, which could lead to a biased view on the effects of external factors. Regardless, microbial communities and their interactions seem to inherently respond and reflect these varying external factors. Therefore, our goal was to build a model that given sufficient chronologically ordered microbial data to the training-set (the model structure is generalizable, but the training is site-specific to account for the local dynamics), the model can learn these hidden patterns and provide accurate predictions without the addition of other external factors. We were also skeptical at the beginning, but that is what our results show.

If a drastic change in environmental conditions occurs, especially if absent in the site-specific training set, it should be expected that the model to perform poorly. Our model is envisioned, designed and trained to predict microbial dynamics under “standard variations,” reflecting the typical operational conditions of a WWTP. Let’s say a WWTP receives a large unexpected, unregulated discharge that disrupts the composition and dynamics of the microbial community, none of our datasets experienced such disruptive events and that is impossible to get info from full-scale plants as such event are rare and the type can be many and so-far unknown

“disasters”. This would be a different research topic in machine learning and is beyond the scope of our paper.

However, to demonstrate the applicability range of our workflow, based only on historical microbial abundance, we have found and tested three independent long-term, high-frequency longitudinal microbial datasets (described in the manuscript with new results added in the main body and supplementary information). From these additional datasets, we observed that for sparsely assembled communities, such as those in fungal time-series from marine coastal environments (Christmas et al., 2023 - <https://doi.org/10.1098/rspb.2022.2129>) where certain species exhibit transient blooms, the model struggles and its accuracy is notably bad. Based on these new findings, we have revised the manuscript to carefully specify the data requirements that are best for our model. We think this way the users can make informed decisions whether they want to try our model in their specific context.

Manuscript change/addition:

- Page 3, lines 97 - 104
- Paged 13-14, lines 390 – 435
- Figure 8, supplementary figures S10 and S11
- Page 15, lines 455 – 470
- Page 17, lines 553 – 565

2. My second concern is how long the training time series must be to provide accurate predictions (Point 3 of my previous comments). In the Response letter, the authors claim this result appears in Figure 2 of the main text, where they compare the prediction accuracy for different WWTP datasets, each with a different number of samples. This claim is partially true, as the obtained prediction accuracy depends not only on the number of training samples but also on other factors like the specific ecological dynamics of the community and environmental changes. In this sense, this issue was not addressed. I expect an analysis showing how the prediction accuracy changes with the number of samples used for training for each WWTP dataset.

Answer: This question is important and we agree that there may be some variation between each WWTP as there may be latent variables in the different WWTP which impact other variables. We do, however, not think that differences in factors such as the specific ecological dynamics of the community and environmental changes are that important when the goal is prediction of community dynamics – these will, overall be within the “standard” variations WWTPs are exposed to and thus that we can reasonably well extrapolate from an analysis of 1-2 plants. Secondly, a unique feature of our approach is that the neural network learns from any

pattern present in the data, and that includes any underlying and hidden variable, of which there can be thousands, many of which are practically impossible to measure. That is why we focus only on the response variables (here the relative abundances of every species), and instead only learn patterns from those, because they are ultimately affected by other variables which we cannot measure.

To test how much the number of samples impact prediction accuracy, we have now done a supplementary investigation, where we took our longest dataset (Aalborg W with 344 samples) and separated it into 3 additional smaller subsets, where we took each 2nd, 3rd, and 4th sample and ran the workflow on each. We observed that the prediction accuracy increased with more samples in the training dataset. We also tried searching the literature for studies with datasets with more samples than that to supplement our results but couldn't find any. We have now described these results in the manuscript and added a supplementary Figure S5.

Manuscript change/addition:

- Supplementary Figure S5
- Page 5, lines 205 - 209

Overall, I feel rather bad for not having a more favorable comment about the Author's response and revised text. I still believe the methods have the potential and data have the potential to make a significant impact in the field, but a more careful validation is still required.

Answer: We hope this revised version has handled all the issues raised. We do also think that our new approach has the potential to make a significant impact in the field.

Reviewer #1 (Remarks to the Author):

The authors have addressed all my previous concerns and comments satisfactorily. However, I believe that the formatting of the main figures still does not fully meet the standards required for publication in Nature journals. For instance, some axis labels are unclear. I recommend revising the figures in accordance with the journal's formatting guidelines.

Thank you for the observation. We have checked and updated all figures where needed and ensured that any text and axis + legend titles are readable with a font size that is proportional to other content within each figure. Note that the images used in the manuscript are not high-resolution images because of the word file format. We have uploaded high resolution (high DPI) files separately.

Reviewer #1 (Remarks on code availability):

The code is publicly available, and the installation process is clearly documented and reproducible.

Reviewer #2 (Remarks to the Author):

We have reviewed the revised manuscript “Predicting microbial community structure and temporal dynamics by using graph neural network models”. We do see substantial improvement after two rounds of revisions, which should definitely be commended. Most of our previous comments have been addressed. However, two questions remain to be addressed before the manuscript can be accepted.

1: This research is highly meaningful and shows promise for predicting filamentous bacteria outbreaks in advance, thereby allowing for timely intervention and maintain a system stability. However, a critical question arises: since interventions themselves can alter the system, can we still regard the system as the same after such interventions? In other words, is it valid to use pre-intervention data for future model training, or should we treat the post-intervention system as a fundamentally different WWTP, meaning the historical data is useless anymore? Clarifying this is essential for evaluating the practical utility and value of the proposed tool.

Answer: We are very happy that the reviewer, overall, is satisfied with our revisions. We agree that the remaining issue deserves more attention, so we have included a clarification in the main manuscript (Lines 552 – 555) and added a supplementary

figure (Figure S12) to show that the model retains its prediction accuracy in the event of a significant intervention.

The issue: Can we still regard the system as the same after such interventions?

Yes, we can. Unless the invention is extremely drastic, so the community completely changes overnight (which would never happen as the plant performance would deteriorate), lots of “normal” variations in conditions are captured by the model during the timeseries (> 100 time points). A realistic intervention would merely be a change in plant operation, and would never result in specific species being eliminated, but change their relative abundance. In WWTPs, the microbial community composition is highly robust, while only the relative abundances change over time as reported in Nierychlo et al. (<https://doi.org/10.1016/j.watres.2020.115955>), and Peces et al. (<https://doi.org/10.1038/s43705-022-00098-4>)

In one of the WWTPs, they have in fact had a major operational change, and we have looked into this dataset more closely. In late 2020 the Viby WWTP implemented the S::Select® technology to improve the biomass settling, along with the installation of hydrocyclones to separate the lighter and heavier biomass fractions. This can be deemed a strong intervention. However, we could not detect a loss in prediction performance, even though some species showed changing abundances after this plant design change, see new supplementary Figure S12. This has now been described in lines 552 – 555.

To further support this, we can also refer to our previously published study by Wagner et al. (<https://doi.org/10.1016/j.watres.2022.118340>), conducted at Damhusåen WWTP. Damhusåen WWTP operates four parallel, independent treatment lines. Line B served as a control (no interventions), while the others underwent various interventions for control of filamentous bacteria: addition of flocculants (Line A), ozone treatment (Line D), and hydrocyclone installation (Line C). The main conclusion was that seasonal dynamics were the dominant driver of microbial community composition. As explained in lines 208 – 210, and Figure 2 – main manuscript – the prediction accuracy for the four lines was very similar, advocating that typical strong operational interventions do not invalidate pre-intervention data. We have not described this further in the main text as we think the above example is sufficient.

2: It remains unclear why, as depicted in Figure 6, the model’s performance on ASV2 time series appears better when predicting 20 steps into the future compared to 10 or 15 steps. The authors explain in the rebuttal letter that this is due to inherent randomness, but the explanation is unconvincing. How many repeated runs were conducted for each prediction scenario? Additionally, could the authors provide a more

robust justification or analysis to explain this counterintuitive outcome?

Alternative answer:

Answer: Overall, the model behaves as expected, with longer prediction windows resulting in lower accuracy (see Fig. 6, main manuscript). As the reviewer points out, there was a case in Figure 7, where 20 steps into the future appeared to perform better than 15 steps. We explained this with inherent randomness. We have now re-run everything again to ensure that this wasn't an error, however, it turns out that it in fact was an error. Thus, we have updated Figure 7 with the new data that now shows the expected pattern correctly and we have not done further in the main text. We want to thank the reviewer many times for observing this!

Reviewer #3 (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.