

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☒ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☒ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒ | ☐ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

For bioinformatic processing of the raw DNA sequencing data we have used usearch11, which is closed source (though will soon be open source according to the website <https://drive5.com/usearch/>)

Data analysis

There are many open source tools and packages used in this project. Please see the GitHub repository, where exact versions of each and every dependency are listed. They are listed in the `renv.lock` and `environment.yml` files.

GitHub: <https://github.com/kasperskytte/mc-prediction>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw DNA sequencing data has been submitted to the Sequence Read Archive under the bioproject ID: PRJNA1021975. All downstream datasets and results can

be produced from this data using the code provided in the GitHub repository (https://github.com/kasperskytte/asv_pipeline).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender Not applicable

Reporting on race, ethnicity, or other socially relevant groupings Not applicable

Population characteristics Not applicable

Recruitment Not applicable

Ethics oversight Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study develops and applies a prediction model to predict microbial species abundance in wastewater treatment plants from multiyear longitudinal sampling. The study evaluates 24 wastewater treatment plants sampled every 1-2 weeks over 3-8 years (total of 5307 independent samples). We also study a published dataset from another WWTP, a marine site and a human gut.
Research sample	The samples were collected from the aerated tank in activated sludge full-scale wastewater treatment plants in Denmark
Sampling strategy	Samples were collected every 1-2 weeks from the aerated tank in activated sludge full-scale wastewater treatment, for a period of at least 3 consecutive years. This sampling frequency is based on previous findings that capture microbial dynamics and recurrent temporal patterns.
Data collection	Labelled tubes and metadata sheets were provided to the wastewater treatment plant operators. The plant operators fill the basic metadata in provided sample sheets (corresponding SampleDate to each SampleID) and send it back to the Center of Microbial Communities together with the samples for sequencing and processing.
Timing and spatial scale	For some of the wastewater treatment plants sampling started in 2015 and completed in 2022, with samples collected every 1-2 weeks. Other wastewater treatment plants joined the project mostly during 2018 with the same sampling frequency.
Data exclusions	none of the data has been excluded
Reproducibility	Each prediction model is applied independently to each wastewater treatment plant dataset, ensuring that the results are significant. Additionally, all code and data is made public on GitHub at https://github.com/kasperskytte/mc-prediction licensed under the GNU General Public License v3.0. The raw DNA sequencing data from the 5307 environmental samples from activated sludge used for this study has been published to the NCBI SRA under accession ID PRJNA1021975 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1021975/).
Randomization	Not necessary for this prediction model. The dataset was partitioned in 3 groups, the training set, validation set and prediction set based on consecutive samples
Blinding	Not necessary for this prediction model, it is blind by nature.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Aerated tank in activated sludge full-scale wastewater treatment plants in Denmark in multi-year longitudinal study. Wastewater treatment plants operated without major disturbances with wastewater temperatures oscillating between 10-18°C during the year.
------------------	--

Location	Samples were collected from 24 municipal WWTPs in Denmark. Specific metadata and geolocation of each site is found in https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1021975
Access & import/export	Not necessary for samples within national territory. All samples are from Danish wastewater treatment plants used for microbial identification
Disturbance	The wastewater treatment plant operators collect samples for biochemical analyses with the same (or more) frequency. Or study did not cause any disturbance to the ecosystem or the working schedules of plant personnel

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	not applicable
Novel plant genotypes	not applicable
Authentication	not applicable