# Oceanic phnJ Gene prediction and Environmental Predictor Interpolation

This repository contains the code and data processing pipeline for:

- Interpolating global oceanographic predictors to a unified 1° × 1° resolution grid (example: Phosphate, Depth).

- Building a Random Forest model to predict the concentration of the `phnJ` gene abundance and performing model validation, partial dependence analysis, and Monte Carlo uncertainty quantification.

- Generating visualizations and figures presented in the manuscript.

- Processing raw metagenomic reads to quantify `phnJ` and `mpnS` gene abundance, including quality control, mapping, gene detection, and normalization.

## Files

- `1\_interpolation\_predictors.R`: Interpolates environmental predictors to 1° × 1° grid. Includes both CSV-based and NetCDF-based interpolation examples.

- `2\_random\_forest\_modeling.R`: Builds and validates the Random Forest model to predict `phnJ` gene abundance globally.

- `3\_Codes\_for\_figure\_generation.R`: Contains the code and software used to generate the figures presented in the manuscript.

- `4\_quantify\_phnJ\_abundance.sh`: Processes raw metagenomic reads to quantify phnJ and mpnS gene abundance, including quality control, mapping, gene detection, coverage estimation, and normalization.

- `5\_boundary\_shapefile/: Contains the **boundary shapefiles** used to constrain interpolation areas during environmental predictors' gridding.

- `6\_Raw\_data\_for\_random\_forest\_model/: Includes **gene abundance** data and associated **environmental predictors** used to train the Random Forest model. All environmental predictors used for prediction can be found in *1\_interpolation\_predictors.R* or are detailed in the main manuscript.

- `7\_sequences\_for\_phnJ\_quantification/: Contains curated **phnJ gene** sequences used for homolog search, mapping, and quantification during metagenomic analyses.

- `8\_sequences\_for\_mpnS\_quantification/: Contains curated **mpnS gene** sequences used for homolog search (e.g., using HMMER), mapping, and quantification.

## Requirements

- R version 4.0.0 or higher

- Required R packages: "sf", "sp", "raster", "gstat", "ncdf4", "maptools", "tidyverse", "tmap", "randomForest", "caret", "ggplot2", "Metrics", "minpack.lm", "reshape2"

- Required bash tools for shell script (`4\_quantify\_phnJ\_abundance.sh`): `fastqc`, `bwa`, `samtools`, `bedtools`, `metaphlan`, `kaiju`

## Additional Notes

- Be sure to adjust file paths in the scripts according to your local setup or repository structure.

- For more details on how the data are used in the model, refer to the associated article (Zhuang and Mao et al., 2025 Under review in NC).