# RESPONDER STRATIFIER ON URATE PREDICTION USING MULTI-OMICS DATA

A PREPRINT

**Zhi Ye**
University of Copenhagen
Copenhagen, Denmark
zhi@food.ku.dk

**Rasmus Riemer Jakobsen**
University of Copenhagen
Copenhagen, Denmark
rasmus@food.ku.dk

**Rasmus Riemer Jakobsen**
University of Copenhagen
Copenhagen, Denmark
rasmus@food.ku.dk

## 1 Introduction

This report details the pipeline and results for the responder stratifier using deep phenotyped data from the Human Phenotype Project (HPP) [1]. Our initiative aimed at predicting urate level in blood test using microbiome data, target SNPs, and daily diet information with other confounding variables (blood test results, BMI, waist, sex, and age). We analyzed beta-diversity in gut microbiome data with principal coordinates analysis (PCoA) based on Bray–Curtis distance [2], encoded diet information proportions with different categories, and calculated the genetic risk score (GRS) based on target SNPs. Then, machine learning models (RandomForest and XGBoost) were employed on the combined features, and we evaluated the model performance using 5-fold cross-validation. Finally, we ranked features based on their importance in our models.

## 2 Data

### 2.1 Metagenomics

The HPP had finished collecting stool samples, metagenomics sequencing, quality control, and taxonomic classification. Samples were processed to obtain MetaPhlAn 4 relative abundances, which were then separated by taxonomic level. There are 2870 relative abundances in total.

### 2.2 Blood test

Blood tests were performed on subjects during their regular medical care at their HMO (Health Maintenance Organisation).

Because the blood test had been measured multiple times per [subject, research stage], we included only the urate level closest to the gut microbiome sampling date in our analysis. Additionally, we filter the subjects who have a long gap (more than 365 days) between the blood test and gut microbiome sampling. The estimated glomerular filtration rate (eGFR) is a key indicator of kidney function and was calculated using the 2021 CKD-EPI equation [3]:

$$\text{eGFR} = 142 \times \min\left(\frac{S_{cr}}{K}, 1\right)^{\alpha} \times \max\left(\frac{S_{cr}}{K}, 1\right)^{-1.200} \times 0.9938^{\text{Age}} \times 1.012^{[\text{if female}]} \tag{1}$$

Where:

- eGFR (estimated glomerular filtration rate) = mL/min/1.73 m²
- $S_{cr}$ (serum creatinine) = mg/dL
- $K$ = 0.7 (females) or 0.9 (males)
- $\alpha$ = -0.241 (females) or -0.302 (males)

- Age is in years
- The factor 1.012 is applied only for females

### 2.3 Diets

Diet logging using a smartphone app involves collecting data on food and drink consumption. These data include information such as types of food, serving sizes, nutritional information, and the times of consumption. The dietary information was collected over 14 days. We only keep subjects with at least one week of diet information, and we removed subjects with fewer than three meals per day.

The diet information contains predefined food categories, and the corresponding gram weights were summed on a per-calendar-day basis for every participant. For each day d and food category k, we computed a within-day proportion by:

$$p_{k,d} = \frac{\text{weight}_{k,d}}{\sum_j \text{weight}_{j,d}} \tag{2}$$

Then we averaged these proportions over all logging days D available for the participant:

$$\bar{p}_k = \frac{1}{|D|} \sum_{d \in D} p_{k,d} \tag{3}$$

### 2.4 Genomics

Genomics features were computed based on six candidate single-nucleotide polymorphisms (SNPs) (*rs734553*, *rs2231142*, *rs1183201*, *rs3741414*, *rs505802*, and *rs478607*). Each variant was encoded per individual as the number of risk alleles carried. The risk score per SNP was then derived by summing these counts (0, 1, 2).

### 2.5 Data Modalities

| Modality | Amount of [Subject, Research Stage] |
|---|---|
| Urate | 4348 |
| Urate + Gut Microbiome (<365 days) | 2220 |
| Urate + Gut Microbiome (<365 days) + Diet (>=7 days) | 1868 |
| Urate + Gut Microbiome (<365 days) + Diet (>=7 days) + Genetics | 1684 |

Table 1: Data Modalities and Sample Sizes

## 3 Results

We checked the sparsity of the microbiome data by plotting the zero-prevalence of the species-level abundances data.

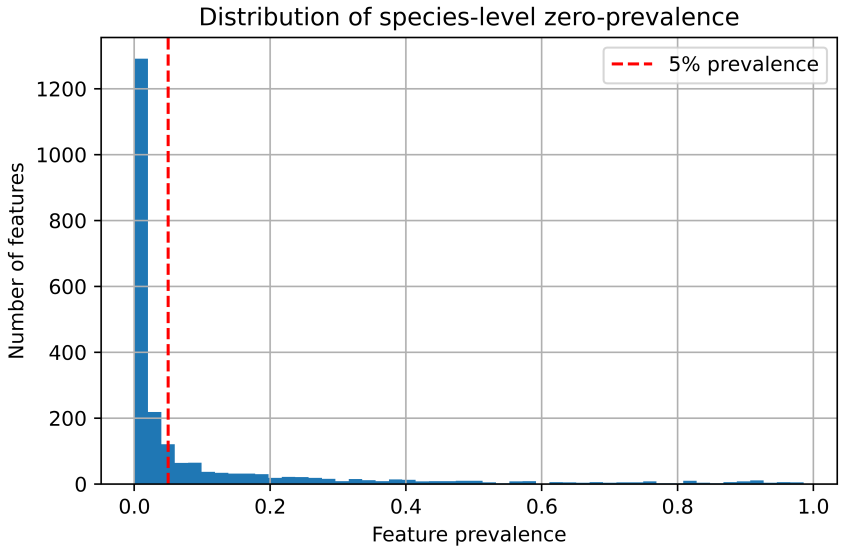



Figure 1: Sparsity of Species-Level Abundances Data

We aggregate the data to the genus level due to the high sparsity, and plot the distribution again. The number of features decreased from 2277 to 862.

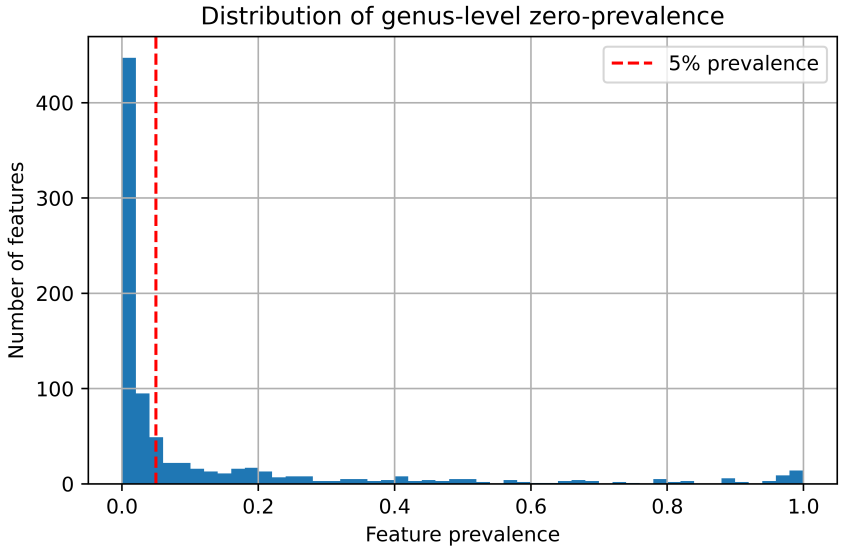



Figure 2: Sparsity of Genus-Level Abundances Data

Then we filter the rare genus by dropping the genus that is represented in less than 5% of the samples. The number of features decreased to 291.

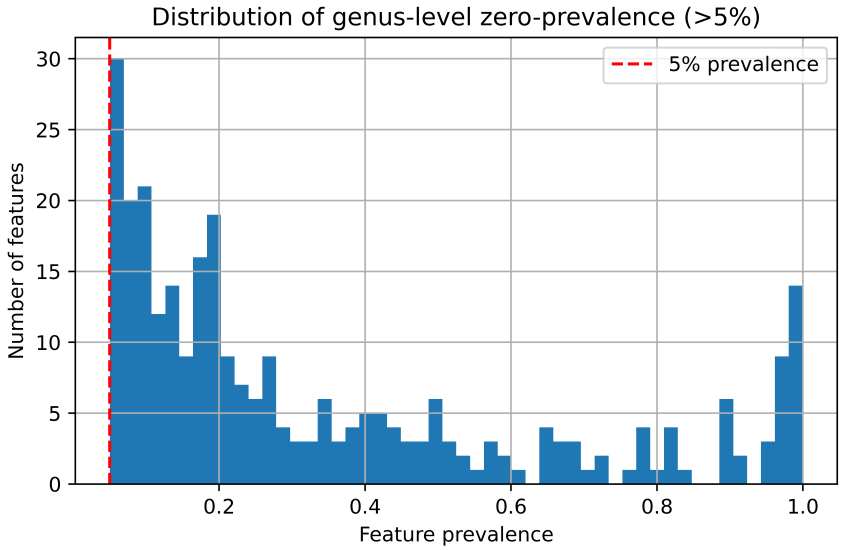



Figure 3: Sparsity of Filtered Genus-Level Abundances Data

## 4 PCoA

After that, we applied the Hellinger transformation, computed the Bray-Curtis distance matrix, and performed PCoA using the Python package scikit-bio.

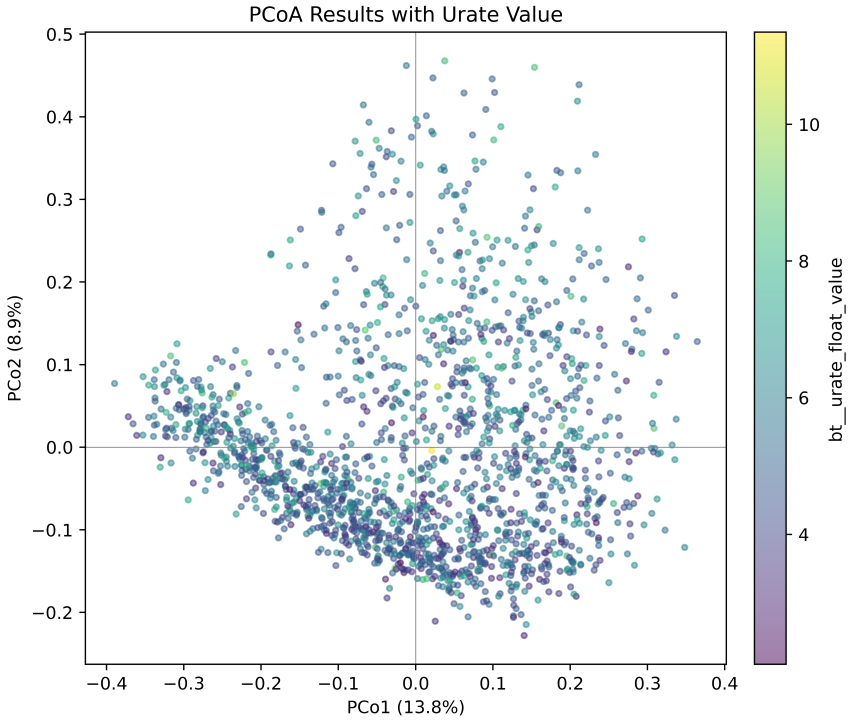



Figure 4: PCoA results labeled by Urate value

The subjects were labeled as “high”, “normal”, and “low” according to their urate level. We used a common standard that “normal” is 2.5–7.0 milligrams per deciliter (mg/dL) in males and 1.5–6.0 mg/dL in females.

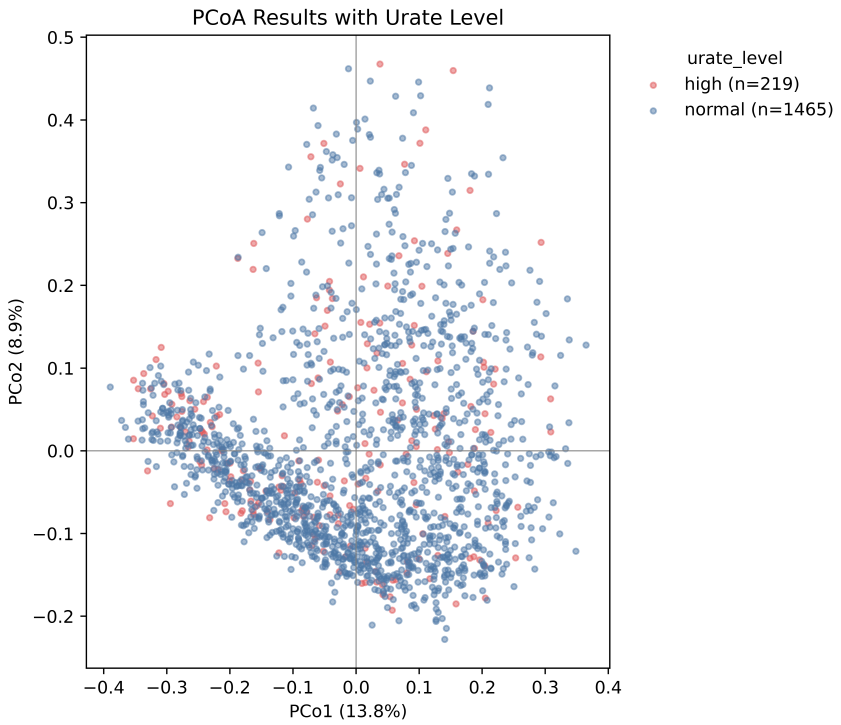



Figure 5: PCoA results labeled by eGFR value

We also plot the PCoA results with the eGFR value.

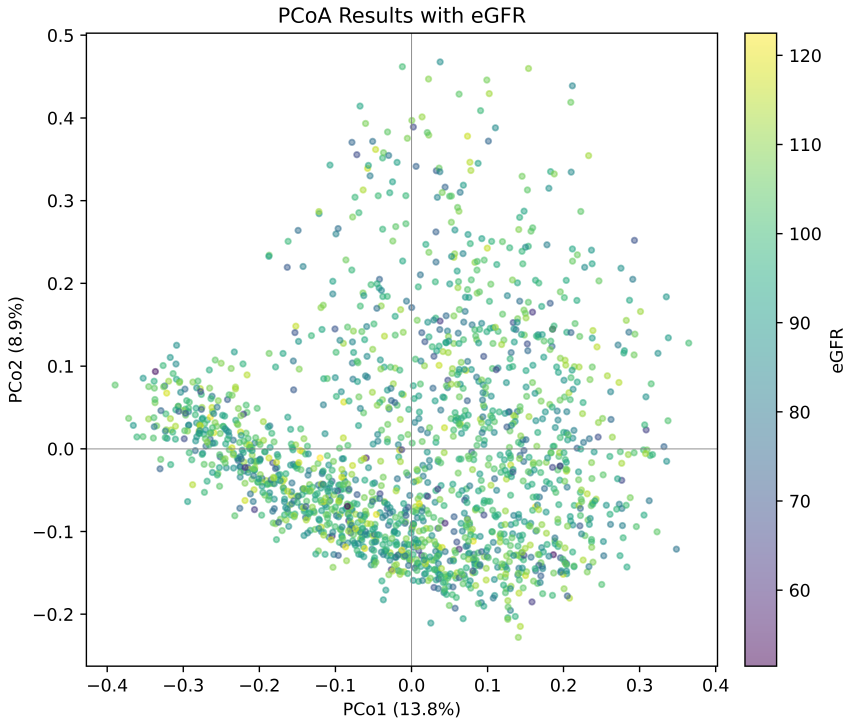



Figure 6: PCoA results labeled by eGFR value

We used the PCo1 to PCo28 as our microbiome features because they explained more than 0.5 variance.

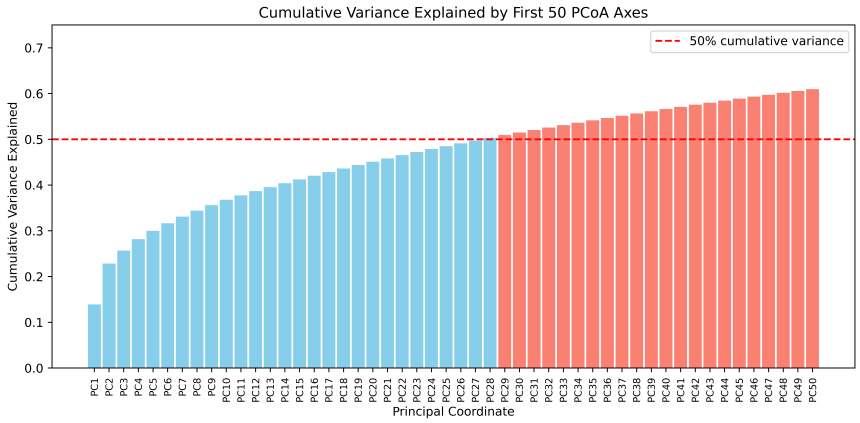



Figure 7: Variance Explained by Each PCo

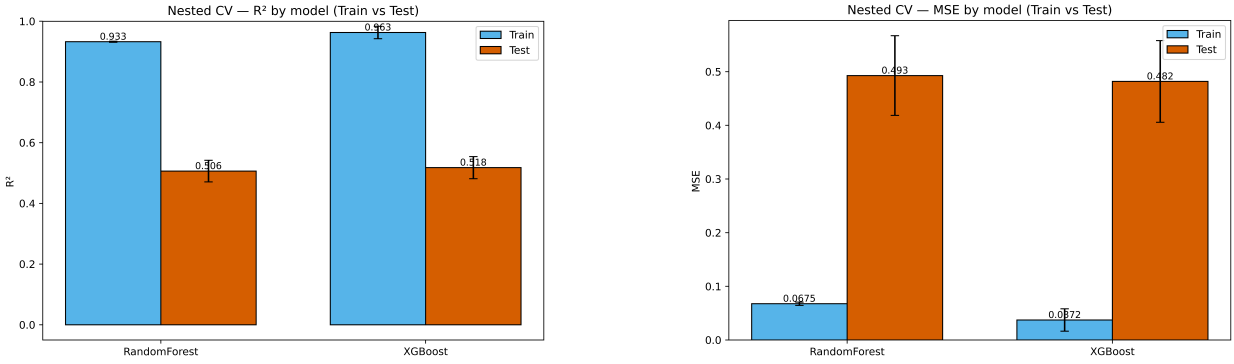
## 5 Regression

The PCoA results were combined with selected blood test biomarkers (Creatinine, Albumin, Alanine Aminotransferase, Total Bilirubin, Glucose, LDL Cholesterol), BMI, waist circumference, diet information, genetics, as well as confounding variables, sex, and age, and used as independent variables in our regression model to predict the urate value from blood test.

$$\text{Urate} \sim \text{Age} + \text{Blood\_biomarkers} + \text{eGFR} + \text{BMI} + \text{Waist} + \text{PCo\_i\_micro} + \text{Diet} + \text{SNPs} \tag{4}$$

After matching with all data modalities, the sample size decreased to 1683, with 47.0% females and 52.9% males.

We used the RandomForest regressor and XGBoost regressor as our prediction model. Both of them are tree-based ensemble methods. The models were evaluated via a 5-fold nested cross-validation. The performances of our models are moderate, achieving an R-squared score of around 0.5.

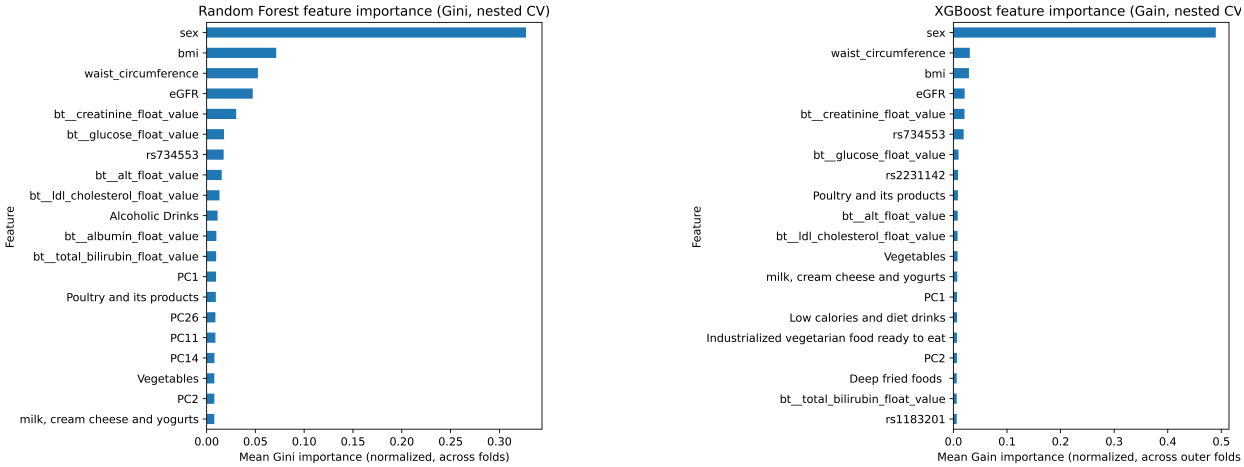



(a) Random Forest R-squared Scores

(b) XGBoost R-squared Scores

Figure 8: Model Performance Comparison

Then we get the feature importances across the CV. Sex, BMI, Creatinine, waist, and Alanine Aminotransferase are the top 5 essential features for both Random Forest and XGBoost models.

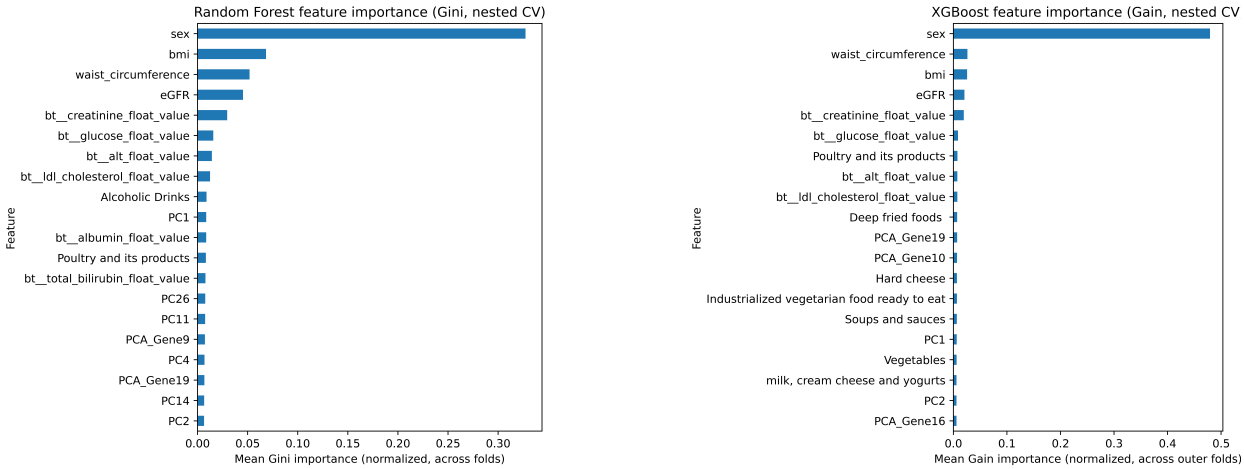



(a) Random Forest Feature Importances

(b) XGBoost Feature Importances

Figure 9: Feature Importances from Models

We re-run the regression model with the PCA results of all genes.




(a) Random Forest Feature Importances

(b) XGBoost Feature Importances

Figure 10: Feature Importances from Models

## 6 Discussion

Our models explained around 50% of the variance of the urate level in the blood test. Except for the well-known factors such as sex, BMI, waist circumference, and blood test biomarkers, which are highly associated with urate level, we also found that some microbiome PCoA features and diet features are essential in our models. This suggests that conducting interventions on the gut microbiome and daily diet patterns can help modify urate levels, thereby preventing the occurrence of related diseases.

## 7 Code Availability

Our code can be found on GitHub [1]. The PCoA analysis was performed using the Python package scikit-bio [4], and the regression models were built using scikit-learn [5] and XGBoost [6]. The pipeline was implemented in Python 3.

[1] https://github.com/zhiye9/Responder_Stratifier

## References

[1] Lee Reicher, Smadar Shilo, Anastasia Godneva, Guy Lutsker, Liron Zahavi, Saar Shoer, David Krongauz, Michal Rein, Sarah Kohn, Tomer Segev, et al. Deep phenotyping of health–disease continuum in the human phenotype project. *Nature Medicine*, pages 1–13, 2025.

[2] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[3] Andrew S Levey and Lesley A Stevens. Estimating gfr using the ckd epidemiology collaboration (ckd-epi) creatinine equation: more accurate gfr estimates, lower ckd prevalence estimates, and better risk predictions. *American Journal of Kidney Diseases*, 55(4):622–627, 2010.

[4] Jai Ram Rideout, Evan Bolyen, Daniel McDonald, Yoshiki Vázquez Baeza, Jorge Cañardo Alastuey, Anders Pitman, Jamie Morton, Qiyun Zhu, Jose Navas, Kestrel Gorlick, Matt Aton, Justine Debelius, Zech Xu, llcooljohn, Joshua Shorenstein, Laurent Luce, Will Van Treuren, John Chase, charudatta navare, Antonio Gonzalez, Colin J. Brislawn, Weronika Patena, Karen Schwarzberg, teravest, Igor Sfiligoi, Jens Reeder, Greg Caporaso, shiffer1, nbresnick, and Chris Tapo. scikit-bio/scikit-bio: scikit-bio 0.7.0, July 2025.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.