# G2S: a new deep learning tool for predicting stool microbiome structure from oral microbiome data

**Simone Rampelli** ( simone.rampelli@unibo.it )
Universita di Bologna https://orcid.org/0000-0002-5655-6695
**Marco Candela**
University of Bologna
**Elena Biagi**
University of Bologna
**Patrizia Brigidi**
University of Bologna
**Silvia Turroni**
University of Bologna

**Software article**

**Keywords:** gut microbiome, oral microbiome, deep learning

**Posted Date:** March 19th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-18048/v1



**Version of Record:** A version of this preprint was published at Frontiers in Genetics on April 9th, 2021. See the published version at https://doi.org/10.3389/fgene.2021.644516.

# Abstract

Background Deep learning methodologies have revolutionized prediction in many fields and show the potential to do the same in microbial metagenomics. However, deep learning is still unexplored in the field of microbiology, with only a few software designed to work with microbiome data. In the frame of meta-community theory, we foresee new perspectives for the development and application of deep learning algorithms in microbiology, with a great potential in the field of human microbiome.

Results G2S is a bioinformatic tool for the taxonomic prediction of the human stool microbiome directly from oral microbiome data of the same individual. The tool uses a deep convolutional neural network trained on data of the Human Microbiome Project, allowing to infer the stool microbiome at the family level more accurately than other approaches. G2S was validated on already characterized oral and fecal sample pairs, and then applied to ancient microbiome data from dental calculi, to derive putative intestinal components in medieval subjects.

Conclusions G2S infers the family-level taxonomic configuration of the stool microbiome mirroring the real composition with exceptional performance. G2S can be used with modern samples, allowing to predict the eubiotic/dysbiotic state of the gut microbiome when fecal sampling is missing, and especially with ancient samples, as a unique opportunity in the field of paleomicrobiology to recover data related to ancient gut microbiome configurations.


# Background

Deep learning is increasingly being used for predicting features and solving complex problems. Unlike traditional algorithms, in which the expertise and rules are already coded, the deep learning algorithms are built to automatically detect patterns in the data [1, 2], by also embedding the computation of variables into the models themselves to yield end-to-end models [3]. In particular, the construction and training of deep learning algorithms have been enabled by the increasingly availability of big data and by the rapid growth in the number and size of public available databases. So far, deep neural networks have been instrumental in advances in modern artificial intelligence, with applications as facial recognition, speech recognition, and self-driving vehicles. More recently, new applications in the field of molecular biology and metagenomics has been pioneered. Indeed, the same deep learning approaches are beginning to be applied to genetics, agriculture, and medicine [4–10]. However, deep learning is still unexplored in the field of microbial metagenomics, with only a few approaches suitable for microbiome data [11–13], and a huge untapped potential still unexplored.

The human microbiome, i.e. the sum of the different microbial ecosystems that colonize the niches of the human body, plays an important role for human physiology and its dysbiotic variations can impact our health [14]. Shifts in the composition of the microbial communities inhabiting the oral cavity and the gastrointestinal tract have been associated with the onset and/or progression of several conditions, such as periodontitis [15], and a series of modern chronic disorders, including inflammatory bowel disease [16],

obesity [17], cardiovascular disease [18] and some forms of cancer [19–21]. The importance of the human microbiome in health and disease makes it imperative to understand the drivers of its variation. In this context, a new frontier is represented by the meta-community theory, according to this the symbiont human microbial ecosystems are in intimate connection, showing reciprocal influences and exchanges [22, 23]. Supporting a meta-community vision of the human microbial ecology, a close link between oral and intestinal microbiome has recently been hypothesized, with the former reflecting changes in the latter, in both healthy and diseased individuals [24–27].

Another scale of human microbiome variation is represented by its change across evolutionary timeline. Particularly, a large body of literature indicates that the current human gut microbiome has evolved towards at least two different configurations, rural and urban, both associated with the corresponding subsistence strategy. Respect to the first, generally considered as the pristine human gut microbiome, the urban configuration is characterized by an overall compression of microbial biodiversity, a wholescale loss of commensal microbial groups, an increased presence of genes related to antibiotic resistance and xenobiotics metabolism [28–33]. However, principally because of paucity ancient stool samples, the ancestral human gut microbiome is still unknown and the evolutionary trajectories and the drivers leading to its contemporary configurations are still to be described, living important gaps in the knowledge on the gut microbiome human host co-evolutionary trajectories. Contrary to ancient fecal samples, dental ones are more common and well preserved, allowing the extraction of the ancient oral microbiome from the ancient DNA conserved in dental tartar. Coherently with the meta-community vision, the ancient oral microbiome configuration can somehow mirror structural features of the gut one because of inherent connections between the two ecosystems. In this scenario, here we developed a new deep learning-based tool, G2S, which infers the gut microbiome configuration from oral microbiome data of a given individual. G2S is based on a model trained and tested on 171 paired samples of gingival and stool microbiome retrieved from the Human Microbiome Project (HMP) [34]. Our approach can be relevant to predict the eubiotic/dysbiotic state of the gut microbiome when fecal data are not available, and particularly suitable for human archaeological records, where coprolites and fecal sediments are really uncommon compared to dental calculi and other human remains. G2S is available on the website https://github.com/simonerampelli/g2s.

## Implementation

# Implementation of the G2S software

G2S adapted a deep convolutional neural network (ConvNet) to generate prediction of gut microbiome configurations from oral microbiome data. ConvNet was structured with two hidden layers, each with 64 units, and a final linear layer with 7 units and no activation function. We selected the mean square error as the loss function, and the mean absolute error as the metric to evaluate the differences between the predictions and the targets during training. In order to minimize the overfitting problems due to the small number of samples within the dataset, we also included a weight regularization step, by adding to the

loss function a cost associated with having high weights. The cost was proportional to the square of the weight coefficient value (L2 regularization or weight decay).

For training and testing of ConvNet, we downloaded all the paired samples available (i.e. gingival and stool samples from the same individual) from the HMP project [34] (see also Additional File 1). Specifically, 342 samples were analyzed using the QIIME pipeline [35] and the Greengenes database [36] in order to obtain the microbiome classification at different taxonomic levels. The genus-level abundance table of 171 gingival microbiome samples was normalized feature-wise prior to its usage for deep learning. In particular, the data were centered on the mean of each specific genus and scaled based on their standard deviation. For the analysis, only 39 genera present in more than 4 samples with relative abundance greater than 0.1% were retained. The 6 bacterial families of the stool microbiome dataset with subject prevalence of at least 80%, including Bacteroidaceae, Porphyromonadaceae, Lachnospiraceae, Ruminococcaceae, Veillonellaceae and Erysipelotrichaceae, were selected as features to be predicted by the ConvNet analysis. An additional variable, called “Other” (i.e. the percentage remaining to reach 100%), was also considered as a feature to be inferred. Training and test datasets were separated to contain 80% and 20% of all profiles, i.e. 137 and 34 paired samples, respectively.

In order to better evaluate the model, we used a k-fold cross-validation approach with 4 partitions and 500 epochs. We got the best performance after the 5th epoch, with a mean absolute error of 7.3%. To increase the predictive performance of ConvNet, the results were then transformed as follows: (i) negative predictions were set to 0, and (ii) the sum of the value for each sample was rescaled to 100%. Finally, based on the result on the training dataset, we also built a confusion matrix to adjust the predictions of those families with recurring over- or underestimation.

G2S includes all of these steps in a single R script, and requires only a relative abundance table of the oral microbiome (between 0 and 1) at the genus level with samples in the columns and the full taxonomy following the Greengens_05_2013 style in the rows as input file. For each sample analyzed, the predicted microbiome is summarized in a table as the relative abundance of the most abundant bacterial families. Additionally, histograms of the same families are provided, using the “graphics” and “base” R packages. The schematic overview of the G2S framework is provided in Fig. 1.

# Ascertaining the performance of G2S on the test dataset

As reported above, 34 paired microbiome samples were included in the test dataset and analyzed to evaluate the G2S performance in predicting the stool microbiome configuration from the oral microbiome sample of the same individual. The performance of G2S was compared with that of other available approaches, including Random Forest [37] and a stochastic algorithm, i.e. a customized method that generates mock profiles of the stool microbiome by randomly imputing the abundances of bacterial families in the range of the training dataset (see Additional File 2 for the script source).

# Case study: using G2S in paleomicrobiology to predict the stool microbiome profile from ancient dental calculi

Microbiome data from dental calculi of 4 adult human skeletons (G12, B17, B61 and B78), characterized by sequencing of the V5 and V6 regions of the 16S rRNA gene (8 samples in total) [38], were used to illustrate the potential and results of G2S. No ethics committee approval was required to perform the analysis included in this study.

## Results

We first applied G2S to the test dataset to evaluate its cross-validated predictions. In particular, mean absolute errors for each family scaled to one standard deviation of real data (maes) < 1 were considered as reference parameters for a good quality of the prediction. As expected, G2S predicts relative abundances with an average maes of 0.78, ranging from the best score for Veillonellaceae (maes = 0.3) to the worst case for Erysipelotrichaceae (maes = 1).

To gain insights into the predictive performance of G2S, we globally compared, sample by sample, the inferred microbiome configurations with the real data by means of bar plots (Fig. 2). Pearson correlations between the predicted microbiome and real data were used to evaluate predictions for each subject. In particular, we considered excellent those predictions with $r > 0.9$ (53% of predictions), good those with r between 0.71 and 0.9 (23% of predictions), discrete with r between 0.41 and 0.7 (12% of predictions), and incorrect with $r \leq 0.4$ (12% of predictions). When we analyzed the cases in which G2S inferred an incorrect prediction, we found that the stool microbiome configurations were very peculiar, with the relative abundances of at least three out of seven bacterial families at the extreme limits of the range of the training dataset (i.e. below or above the 10th or 90th percentile, respectively). In addition, these samples were dominated by other families, usually included in the “Other” category, such as Rikenellaceae and Alcaligenaceae, and showed a large distance overall from the median configuration of the training dataset (maes > 1). Therefore, G2S worked properly when the stool microbiome configurations to be predicted were closer to the median configuration of the training dataset (maes < 1); conversely, it did not work as expected in samples whose microbiome structure markedly deviated from the training dataset (maes > 1) (Fig. 3).

G2S showed a better mimicry of the abundance of microbiomes in the test dataset than other methods, including Random Forest and a stochastic method developed specifically for this comparison, which generates mock profiles of the stool microbiome in the range of the training dataset (Fig. 4). Random Forest under- or overestimated bacterial families with a global maes of 1.19, ranging from 0.99 for Lachnospiraceae to 1.89 for Veillonellaceae. The performance of our custom predictor was even more inaccurate, with a total of one hundred permutational predictions showing maes between 1.18 and 1.4 (mean = 1.29). The best performance of G2S in predicting the stool microbiome structure is probably due to the predicting power of deep learning that automatically detects patterns in the data, by also embedding the computation of variables into the models themselves to yield end-to-end models.

In the second part of our analysis, we used G2S to infer the stool microbiome from the oral microbiome data of four adult human skeletons with evidence of mild to severe periodontal disease, of the medieval monastic site of Dalheim, Germany (c. 950–1200 CE)[38]. G2S inferred the stool microbiome structure at the family level, estimating the abundance of the 7 features, i.e. the 6 bacterial families and the category “Other” including all the other families (Fig. 5A). Interestingly, Bacteroidaceae was the predicted dominant component in the feces of the four subjects, using both the V5 and V6 regions as targets of the 16 rRNA gene (relative abundance > 40%). On the other hand, the family Erysipelotrichaceae showed the lowest relative abundance (< 1%) in all eight samples. Significant differences in taxon relative abundance were found with respect to the stool microbiome of modern healthy subjects from the HMP cohort, including a higher relative abundance of Porphyromonadaceae, Lachnospiraceae and Veillonellaceae in the predicted ancient microbiome configurations (p value < 0.05, Wilcoxon test) (Fig. 5B). However, this is not unexpected given the profoundly different lifestyles of ancient individuals of the Middle Ages and modern people, in terms of diet, contact with the environment and sanitization practices [34, 38]. Future studies in larger worldwide cohorts, including paired samples of oral and intestinal microbiome, are needed to refine the accuracy of the G2S software and predict a higher number of bacterial families as well as taxa at different phylogenetic levels, such as genera and species.

# Discussion

G2S is specifically designed to predict the structure of the human stool microbiome from oral microbiome data. In particular, it uses relative abundance tables of the oral microbiome generated by next-generation sequencing, and a deep learning approach that allows high-speed prediction of the stool microbiome without any downstream process. It could be used with both modern and ancient samples, providing a good prediction of the fecal microbiome with a net saving of time and costs. This is particularly relevant in the context of paleomicrobiology, where human coprolites and fecal sediments are very rare compared to dental calculi. On the other hand, it should be noted that G2S is blind to bacterial families that are not included in the algorithm (i.e. those with a lower relative abundance compared to the HMP cohort); therefore it is not suitable for stool microbiome configurations that strongly deviate from those of the training dataset. This stresses the importance of collecting paired samples (i.e. oral and fecal) in future studies, possibly from cohorts from different geographic locations, in order to extend the range of the training dataset and thus the applicability of G2S. Finally, other future implementations could include predictions at different taxonomic levels, as well as functional predictions thanks to the recent expansion of shotgun metagenomics.

# Conclusion

G2S opens up new possibilities in bioinformatics approaches related to metagenomics, extending in silico procedures to predict the human stool microbiome from oral microbiome data. Starting from modern and ancient oral microbiome samples, the tool infers the stool microbiome with a family resolution. Its main field of application is probably paleomicrobiology, as a tool that can help understand

how the gut microbiome of the past was structured, and its implications for human evolution. An update of the G2S tool will be periodically performed to incorporate newly released microbiome studies.

# Abbreviations

HMP
Human microbiome project

# Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability and requirements

Project name: G2S

Project home page: https://github.com/simonerampelli/g2s

Operating systems: command line on Linux or OS X or Windows

Programming language: R

Other requirements: R package: keras including tensorflow

Licence: FreBSD

Any restriction to use by non-academics: No

### Availability of data and materials

The dataset used for setting up G2S is available at the Human Microbiome Project website https://www.hmpdacc.org/HMQCP/. Microbiome data from the ancient samples were taken from the study conduct by Warinner and collegues [38].

### Competing interests

The authors declare that they have no competing interests.

### Abbreviations

HMP: Human microbiome project

### Authors' contributions

SR built, released, maintained the G2S software and ran the benchmarks. SR, MC and ST wrote the manuscript. EB and PB revised and edited the text. All authors read and approved the final manuscript.

## References


1. Murphy KP. Machine learning: a probabilistic perspective. MIT Press; 2012.
2. Bishop CM. Pattern recognition and machine learning. Springer, New York; 2016.
3. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016
4. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831-838.
5. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE. 2016;104:176-197.
6. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15.
7. Demirci S, Peters SA, de Ridder D, van Dijk ADJ. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. Plant J. 2018;13979.
8. Webb S. Deep learning for biology. Nature. 2018;554:555-557.
9. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat Biotechnol. 2018;36:829-838.
10. Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40-47.
11. Geman O, Chiuchisan I, Covasa M, Doloc C, Milici MR, Milici LD. Deep learning tools for human microbiome big data. In: Balas V, Jain L, Balas M editors. Soft Computing Applications. SOFA 2016. Advances in Intelligent Systems and Computing, vol 633. Springer, Cham; 2016. p. 265-275.

12. Reiman D, Metwally A, Yang Dai. Using convolutional neural networks to explore the microbiome. Conf Proc IEEE Eng Med Biol Soc. 2017;4269-4272.
13. Galkin F, Aliper A, Putin E, Kuznetsov I, Gladyshev VN, Zhavoronkov A. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. bioRxiv. 2018;507780.
14. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. Nature. 2011;474:327-36.
15. Griffen AL, Beall CJ, Campbell JH, Firestone ND, Kumar PS, Yang ZK, et al. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. ISME J. 2012;6:1176-85.
16. Glassner KL, Abraham BP, Quigley EMM. The microbiome and inflammatory bowel disease. J Allergy Clin Immunol. 2020;145:16-27.
17. Rampelli S, Guenther K, Turroni S, Wolters M, Veidebaum T, Kourides Y, et al. Pre-obese children's dysbiotic gut microbiome and unhealthy diets may predict the development of obesity. Commun Biol. 2018;1:222.
18. Pietiäinen M, Liljestrand JM, Kopra E, Pussinen PJ. Mediators between oral dysbiosis and cardiovascular diseases. Eur J Oral Sci. 2018;126 Suppl 1:26-36.
19. Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nat Rev Gastroenterol Hepatol. 2019;16:690-704.
20. Karpiński PM. Role of oral microbiota in cancer development. Microorganisms. 2019;7:20.
21. Helmink BA, Khan MAW, Hermann A, Gopalakrishnan V, Wargo JA. The microbiome, cancer, and cancer therapy. Nature Medicine. 2019;25:377-388.
22. Koskella B, Hall LJ, Metcalf CJE. The microbiome beyond the horizon of ecological and evolutionary theory. Nat Ecol Evol. 2017;1:1606-1615.
23. Miller ET, Svanbäck R, Bohannan BJM. Microbiomes as metacommunities: understanding host-associated microbes through metacommunity ecology. Trends Ecol Evol. 2018;33:926-935.
24. Prodan A, Levin E, Nieuwdorp M. Does disease start in the mouth, the gut or both? 2019;e45931.
25. Iwauchi M, Horigome A, Ishikawa K, Mikuni A, Nakano M, Xiao JZ, et al. Relationship between oral and gut microbiota in elderly people. Immun Inflamm Dis. 2019;7:229-236.
26. Bajaj JS, Betrapally NS, Hylemon PB, Heuman DM, Daita K, White MB, et al. Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. Hepatology. 2015;62:1260-71.
27. Schmidt TSB, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, et al. Extensive transmission of microbes along the gastrointestinal tract. eLife. 2019;8:e42693.
28. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486:222-227.

29. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun. 2014;5:3654.
30. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun. 2015;6:6505.
31. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, et al. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr Biol. 2015;25:1682-1693.
32. Jha AR, Davenport ER, Gautam Y, Bhandari D, Tandukar S, Ng KM, et al. Gut microbiome transition across a lifestyle gradient in Himalaya. PLoS Biol. 2018;16:e2005396.
33. Ayeni FA, Biagi E, Rampelli S, Fiori J, Soverini M, Audu HJ, et al. Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. Cell Rep. 2018;23:3056-3067.
34. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207-214.
35. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335-336.
36. The greengenes database. 2013. https://greengenes.secondgenome.com. Accessed 21 Sept 2018.
37. Breiman L. Random Forests. Mach Learn. 2001;45:5-32.
38. Warinner C, Rodrigues JF, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336-344.

# Figures

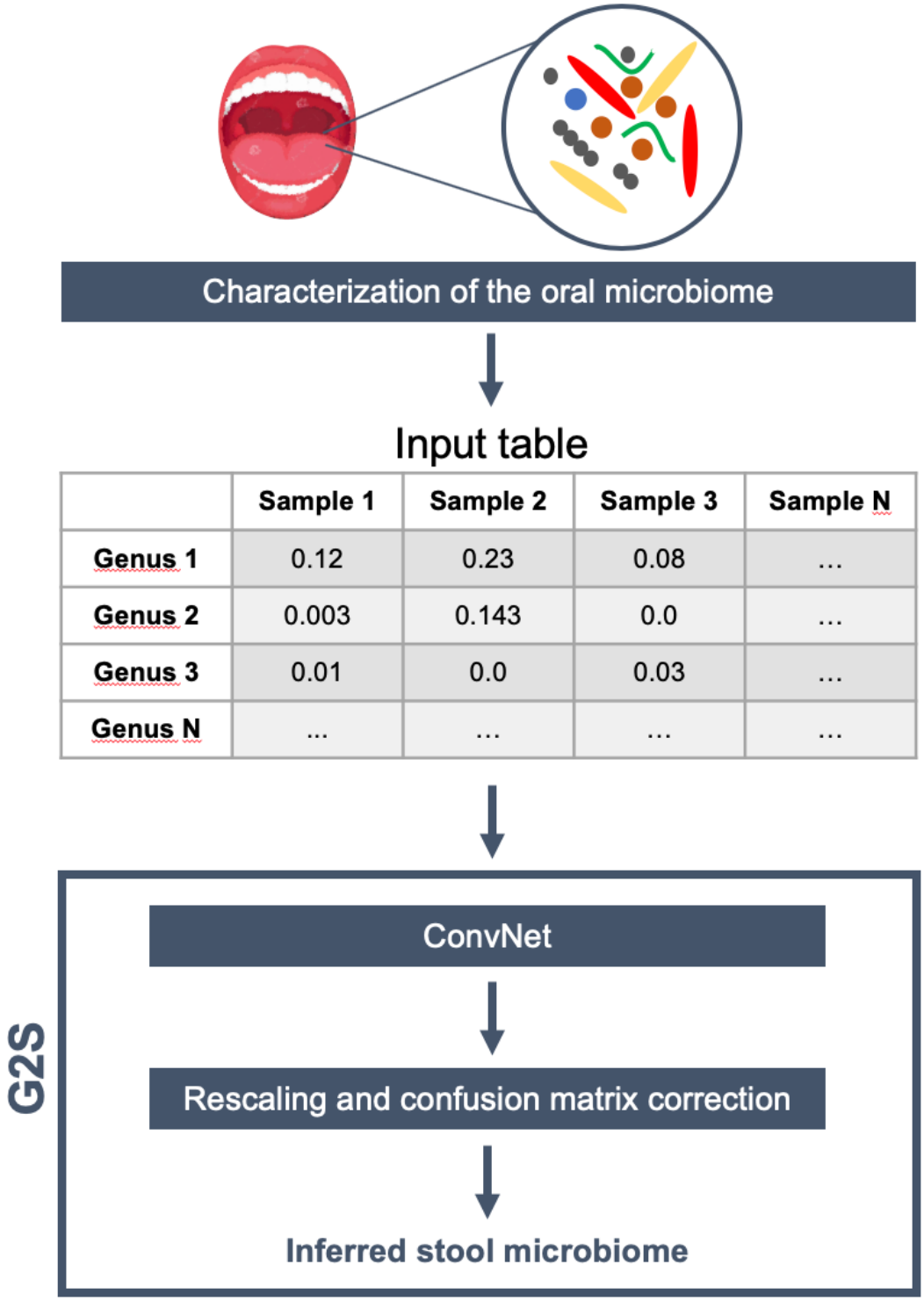
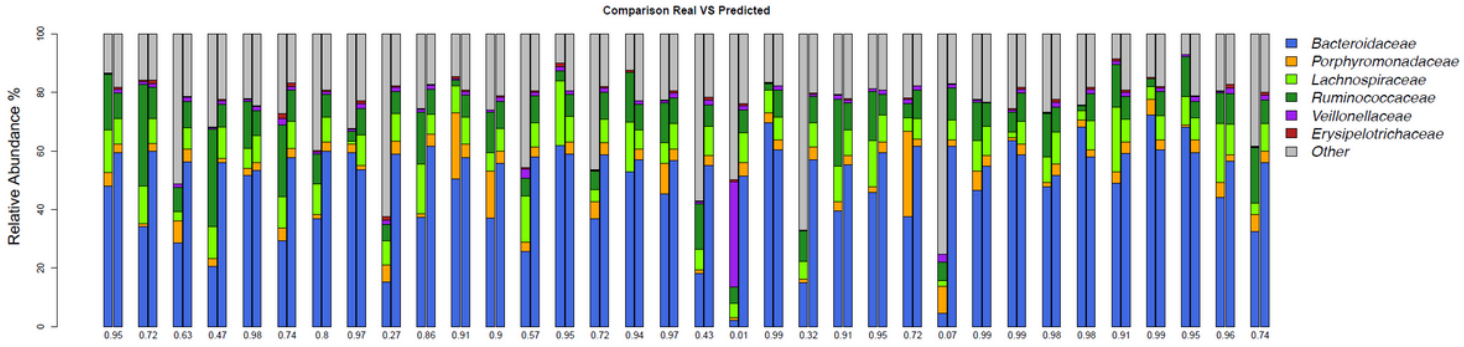


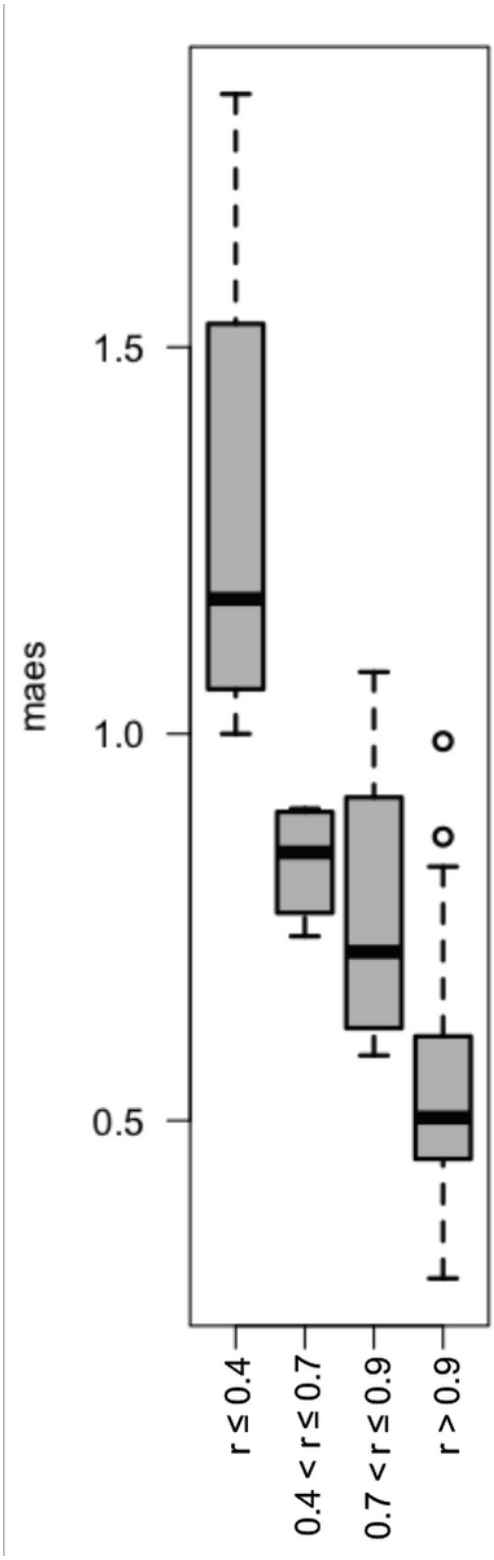

**Figure 1**

G2S workflow. The input file is a genus-level relative abundance table (.tsv format), obtained from the characterization of human oral microbiome samples. The stool microbiome is inferred using a deep convolutional neural network (ConvNet) adjusted by a confusion matrix and rescaled to 100%. The results are tabulated as relative abundance.




## Figure 2

Comparison between G2S predictions and real data from the test dataset. The family-level bar plots of the 34 stool samples of the test dataset from the HMP project are visualized next to their inferred configurations obtained by G2S. Pearson correlation coefficients (r) are provided below each pair of bar plots.




## Figure 3

G2S predictions are more accurate when the configurations to be inferred fall within the range of the training dataset. Box plots of the mean absolute error scaled to one standard deviation (maes) between the real stool microbiome configuration of the samples in the test dataset and the median configuration of the training dataset. Samples were divided into four groups based on the quality of the G2S

predictions (i.e. the Pearson correlation coefficients between the real values and the inferred configurations).

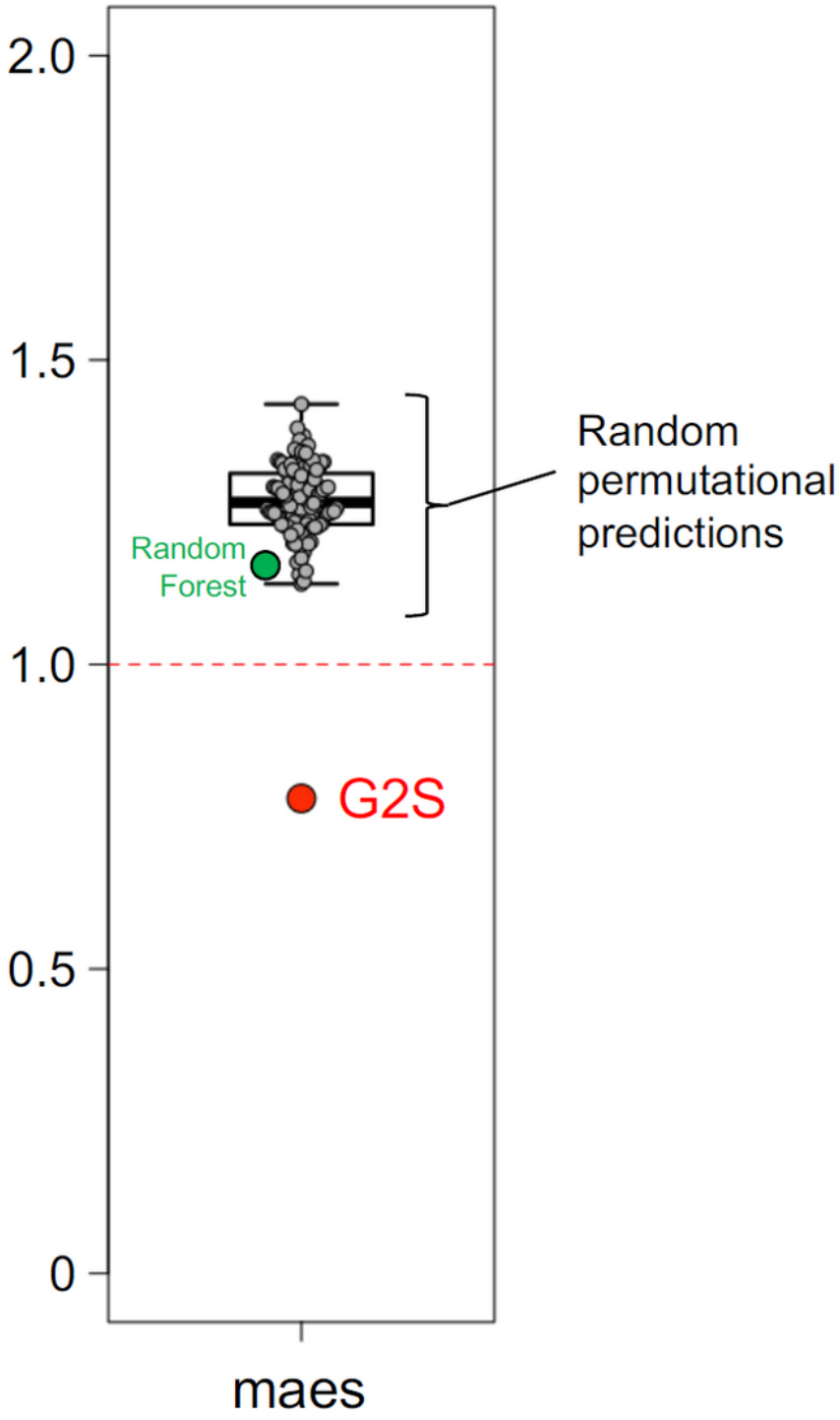



**Figure 4**

G2S predicts the stool microbiome configuration with better performance than other methods. The mean absolute errors scaled to one standard deviation (maes) between the real data of the samples from the

test dataset and the configurations inferred by G2S, Random Forest and a stochastic permutational method (100 predictions), are reported in the dot plot.

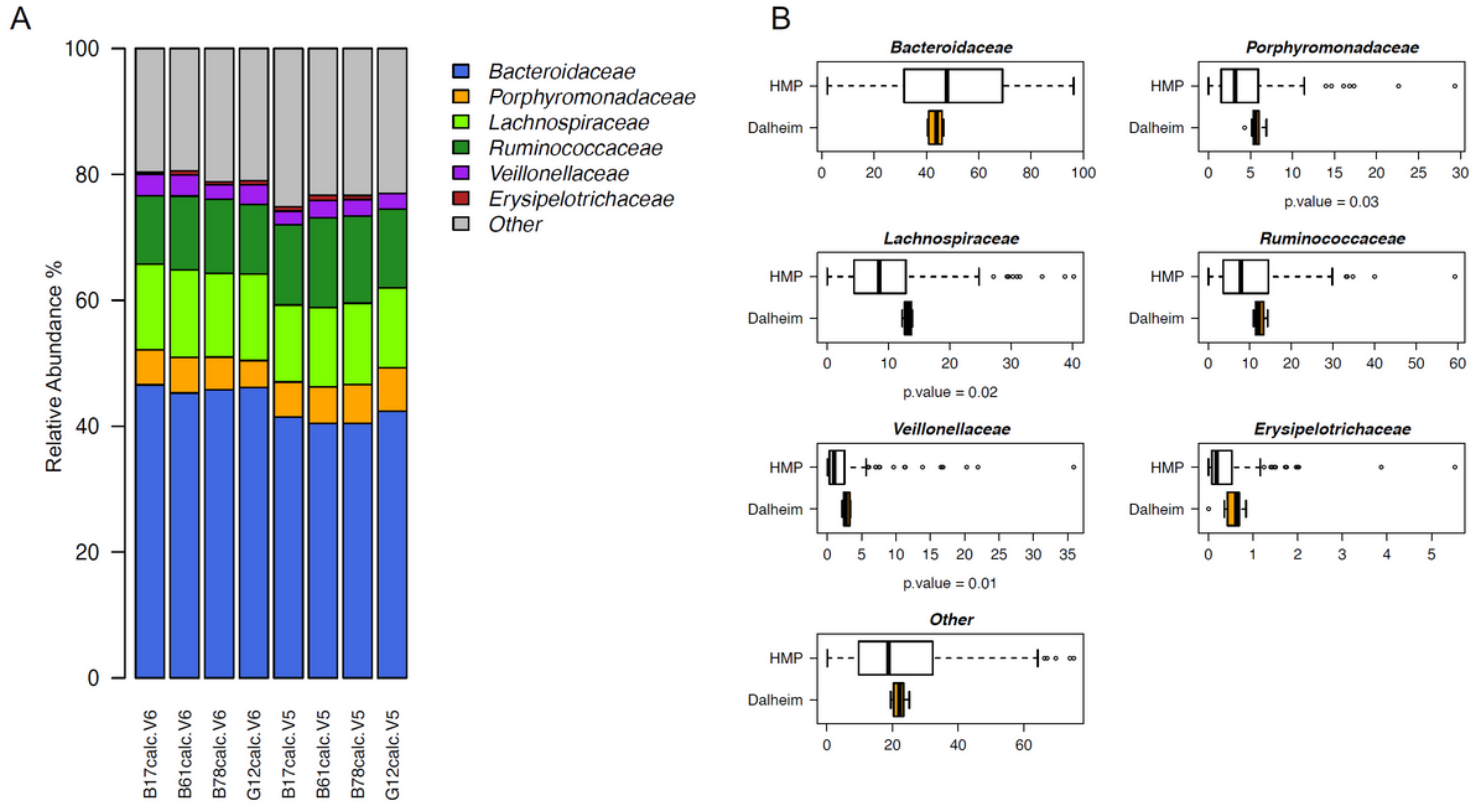



**Figure 5**

Reconstructing the ancient stool microbiome of adult medieval individuals. A, bar plots of stool microbiome configurations inferred from 16S rRNA gene (V5 and V6 regions) sequencing data of ancient microbiomes (i.e. dental calculi from the medieval monastic site of Dalheim, Germany, c. 950–1200 CE) [38]. B, comparison between the predicted ancient microbiome configurations and the modern stool microbiome of subjects from the HMP cohort (p values were determined by Wilcoxon test).

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile2.R
- AdditionalFile1.xlsx