WHOLE GENOME TRANSFORMER FOR GENE INTERACTION EFFECTS IN MICROBIOME HABITAT SPECIFICITY

Zhufeng Li
Technical University of Munich
Helmholtz Munich
Munich Center for Machine Learning (MCML)

Sandeep S Cranganore
Forschungszentrum Jülich
Technical University of Vienna

Nicholas Youngblut
Arc Institute

Niki Kilbertus
Technical University of Munich
Helmholtz Munich
Munich Center for Machine Learning (MCML)

ABSTRACT

Leveraging the vast genetic diversity within microbiomes offers unparalleled insights into complex phenotypes, yet the task of accurately predicting and understanding such traits from genomic data remains challenging. We propose a framework taking advantage of existing large models for gene vectorization to predict habitat specificity from entire microbial genome sequences. Based on our model, we develop attribution techniques to elucidate gene interaction effects that drive microbial adaptation to diverse environments. We train and validate our approach on a large dataset of high quality microbiome genomes from different habitats. We not only demonstrate solid predictive performance, but also how sequence-level information of entire genomes allows us to identify gene associations underlying complex phenotypes. Our attribution recovers known important interaction networks and proposes new candidates for experimental follow up.

1 INTRODUCTION AND RELATED WORK

Machine learning (ML) on genetic data. Determining how gene-gene interactions influence certain traits, health, and disease has been a longstanding challenge for biologists and medical researchers (Gilbert-Diamond & Moore, 2011; Wan et al., 2010). Modern high-throughput sequencing techniques such as massive parallel methods (Ronaghi et al., 1996; Nyren et al., 1993; Nayfach et al., 2021) or single cell RNA sequencing (Hwang et al., 2018; Jovic et al., 2022) together with recent developments in transformer-based models (Vaswani et al., 2017), which nowadays operate on sequences lengths up to 100,000 (Avsec et al., 2021) or even 1 million (Nguyen et al., 2023) base pairs, allow for modeling highly complex sequence diversity spanning large sections of the genome.

Within this paradigm, Jumper et al. (2021) achieved state of the art in protein folding predictions, Avsec et al. (2021) identified enhancer-promoter interactions with unprecedented accuracy, and Li et al. (2023); Avsec et al. (2021) demonstrate promising results on gene regulatory network inference. The potential impact on human health has also inspired large-scale concerted industry efforts into building large transformer models that can perform multiple relevant tasks at once. For instance, in a sequence of papers (Rives et al., 2019; Rao et al., 2020; 2021; Meier et al., 2021; Hsu et al., 2022; Lin et al., 2022; 2023), a collection of models was released – dubbed Evolutionary Scale Modeling (ESM) – that perform tasks from protein design (beyond natural proteins) and (inverse) protein folding to variant-, function-, and property-prediction. Consens et al. (2023); Choi & Lee (2023) provide detailed overviews of recent deep-learning (in particular transformer) based models for the genome and what they are capable of.

Importance of the microbiome. Bacteria and archaea are often heavily underrepresented in deep learning models trained on genetic data (Zhou et al., 2023; Dalla-Torre et al., 2023). While modeling human genetic diversity has many direct implications for human health (Sapoval et al., 2022; Clapp et al., 2017), developing models that incorporate the vast genetic diversity across the microbial tree of

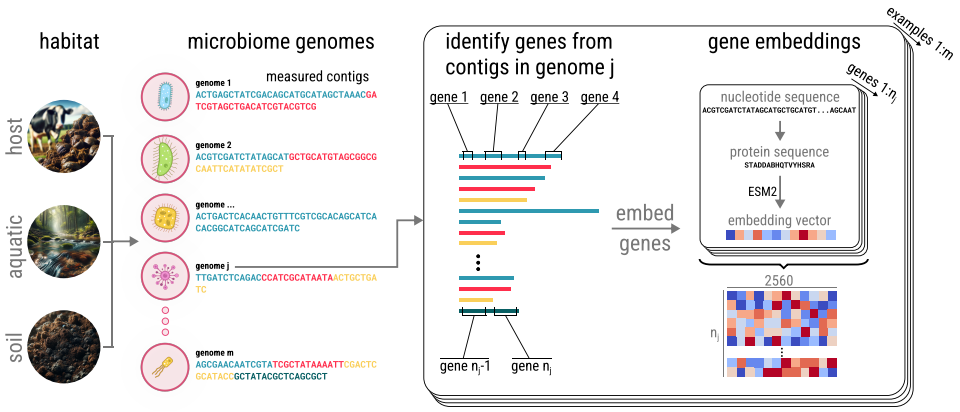

Figure 1: A conceptual overview of our data preprocessing pipeline. Each sample stands for an entire genome covered by individual contigs. We identify all genes within each contig (using Prodigal) and embed the corresponding protein sequences with a protein large language model (ESM-2) into a d_{emb}-dimensional vector space. A single 'input example' is then represented by a $(n_j \times d_{\text{emb}})$-dimensional tensor.

life may lead to similar benefits, such as the development of novel microbiome therapeutics, inferring the health benefits of microbe-produced metabolites, and predicting the evolution of antibiotic resistance (Hernández Medina et al., 2022). Unlike the relatively static nature of the human genome, the microbiome is highly dynamic, adapting to environmental changes and interactions with its host or environment (Lloyd-Price et al., 2017; Ducarmon et al., 2023). The plasticity of the microbiome could be harnessed to treat disease more easily via microbiome interventions versus gene- or immuno-therapy (Schupack et al., 2022; Ratiner et al., 2023).

While some work, such as ESM (Lin et al., 2023) and LookingGlass (Hoarfrost et al., 2022), have included a large degree of known microbial diversity, such models are limited to single genes or short DNA sequences of 100 to 200 base pairs (Hoarfrost et al., 2022). Moreover, microbial genes are often arranged in operons that are co-regulated and often form protein complexes (Cao et al., 2019). Modeling large segments of the genome can thus incorporate more genotypic complexity than models trained on short DNA sequences (Wei et al., 2024; Nguyen et al., 2023; Cheifet, 2019).

Predicting phenotype from genotype is quite challenging in the context of the microbiome. First, the majority of microbial genome assemblies are not complete (Parks et al., 2022; Chklovski et al., 2023), and instead comprise 10's to 1000's of genome fragments (contigs). Even among individual genomes belonging to the same species, genomes can differ substantially in genomic content and arrangement (Rouli et al., 2015; Lapierre & Gogarten, 2009); thus, the ordering of contigs usually cannot be inferred from closely related, completely assembled genomes. Second, microbial genome databases under-represent microbial diversity, especially microbes that are rare in well-studied environments or microbes only found in understudied environments (Brewster et al., 2019; Pavlopoulos et al., 2023). Third, cellular functioning of most microbial genes and non-coding elements is unknown, which has lead to initiatives to uncover this "microbial dark matter" (Hoarfrost et al., 2022; Pavlopoulos et al., 2023); however, much work is still needed. This work is especially challenging, given that many microbes cannot be cultivated (Almeida et al., 2021), and genetic tools only exist for a small subset of cultivatable microbes (Marsh et al., 2023). Fourth, microbial phenotypes are often difficult to measure, given the challenge to isolate and measure the traits of individual strains. Complex phenotypes, such as microbial habitat may involve a number of factors, including many cellular processes produced by a multitude of genes and regulatory elements.

We provide an in-depth overview of existing 'genotype to phenotype' methods with a comparison of the different characteristics of existing models in Appendix D.

2 Methodology

Microbiome data. Various peculiarities arise from the prevailing sequencing technology Ghurye et al. (2016) used for large scale microbial DNA sequencing screens as collected by ProGenomes (Mende et al., 2016; 2019; Fullam et al., 2023). For example, instead of obtaining entire genomes, one typically only reconstructs so-called 'contigs', i.e., contiguous consensus regions of DNA that have been recovered from the short sequenced snippets. While different chromosomes are expected to produce different contigs, even circular, single-chromosome genomes may lead to multiple contigs.

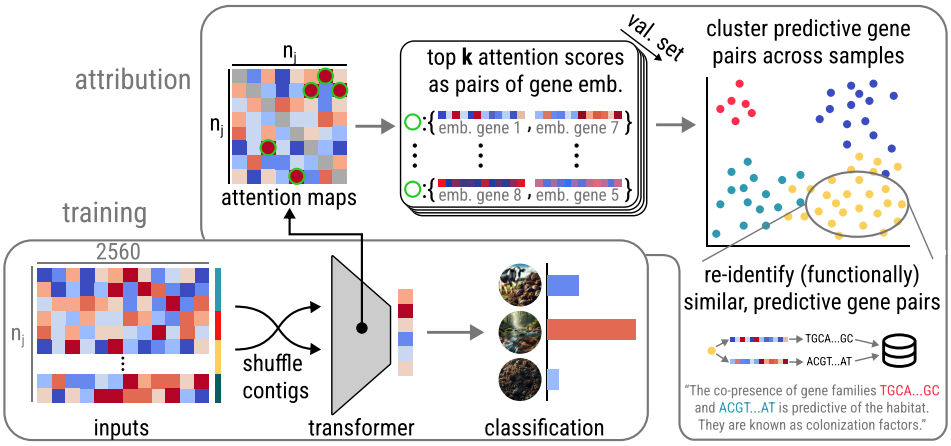

Figure 2: A conceptual overview of our training and attribution pipelines. **Training:** We feed the $(n_j \times d_{\text{emb}})$-dimensional inputs to our transformer, interpreted as a sequence of embeddings. We randomly shuffle contigs within samples, since the 'correct' order is unknown. The model is trained with the cross-entropy loss. **Attribution:** After training, we run extract the last-layer attention maps for all validation samples and find the indices of the top-k attention scores. The corresponding pairs of embeddings are clustered and visualized via non-linear dimensionality reduction. Within each cluster, we re-identify the all gene sequences and match them against databases for annotations.

While genes appear in the right order within a contig, we typically cannot determine the order in which contigs appear within the full genome. We limit our attention to coding genes, requiring us to identify individual genes from within each contig. Our tailored data-preprocessing aims at accounting for these task-specific aspects. Figure 1 provides an overview of the first stage of our framework.

Dataset. We obtain all genomic data from ProGenomes v3, an open-source database comprising over 900,000 consistently annotated bacterial and archaeal genomes from over 40,000 species. Collectively, the genomes contain 4 billion genes; for reference, the human genome contains about 20,000 coding genes. Consistent phenotypic data across all genomes in the database is limited, so we focus on habitat classification in order to comprehensively utilize the available genomic data and assess prediction performance for a complex phenotype. We select the three habitats with the most associated genomic data: *host* (symbiotic or parasitic microbiome, which relies on a host organism, typically collected from animal feces), *soil* (generally free-living microbiome collected from the soil), and *aquatic* (free-living microbiome collected from natural water bodies). In total, our genome dataset comprised $m = 29,089$ genomes (soil: 8,248; host: 9,770; aquatic: 11,070) and 3,056,557 contigs with a mean length of 3445 ± 1632 genes.

Gene embeddings. The high variability of contig lengths in our dataset challenges direct application of existing deep learning approaches. We therefore deploy a multi-gene approach that leverages an existing protein large language model to produce fixed-sized embeddings as input to our model. Our workflow consists of identifying coding genes in each contig with Prodigal (Hyatt et al., 2010), which results in 33 ± 179 genes per contig. Figure 3(left) shows the distributions of how many contigs are contained in a sample with a clear skew towards few contigs per sample (note the logarithmic y-axis), and Figure 3(right) shows the distribution of the overall number of genes extracted per sample. For completeness, we show the distribution of the number of genes per contig in Figure 4 in Appendix A. The common peak at around 4,000 genes aligns well with expectations of average gene counts in bacteria and archaea. We then use ESM-2 (3B) (Lin et al., 2023) to embed each amino acid sequence identified by Prodigal into a fixed-dimensional ($d_{\text{emb}} = 2560$) vector space. Ultimately, for each sample j (i.e., each genome) we stack all n_j gene embeddings belonging to that sample into a $(n_j \times d_{\text{emb}})$-dimensional tensor, where n_j still varies across samples and which comprises one 'input example' for our model. For the roughly 8k, 10k, and 11k samples from soil, host, and aquatic habitats (a total of $m = 29,089$ training examples), respectively this yields a total of almost 1TB of pre-computed ESM-2 gene embeddings as the final dataset for our transformer model. Figure 1 provides an conceptual overview of our data preparation process.

Model architecture and training. Since individual genes are typically shared by many organisms within and across habitats, we hypothesize that habitat specificity heavily depends on the co-presence and interaction effects of multiple genes. For these interactions, the local context is relevant because functionally related genes tend to be clustered in local neighborhoods on the genome (Xu et al., 2019). The attention mechanism in transformer architectures (Vaswani et al., 2017) is not only well

suited to capture such associations in making predictions but also allows for attribution techniques to extract relevant pair-wise interaction effects. Hence, we propose an encoder-only BERT-like architecture (Devlin et al., 2019) for classification (using the standard cross-entropy loss) with 15 layers, 1 attention heads, and a hidden dimension of 640. To reduce the memory footprint during training, we feed the original embeddings of dimension $d_{\text{emb}} = 2560$ obtained from ESM-2 into a single linear layer to obtain a reduced hidden dimension of 640.

We set the maximum input sequence length to 4096, reaching beyond the average number of genes within a genome. Because some samples in our dataset contain more genes than that (c.f., Figure 3, we truncate them.[1] Here, we make use of the fact that the order of genes is preserved within contigs, but not across contigs. Specifically, in each epoch we randomly permute the contigs within every input example before potentially truncating (c.f., Figure 2). Over multiple epochs, this procedure allows the model to learn dependencies between all possible pairs of genes even for the longest examples despite the limited maximum sequence length. Moreover, the permutation may encode our prior knowledge that there is no intrinsic (known) order among the contigs within an example as an invariance in the model. While various techniques for sparse and/or linear attention (Tay et al., 2021) may allow us to extend the maximum input sequence, it would impede attention-based attribution, as we would not obtain comparable attention scores for all pairs of genes. Similarly, recent techniques scaling transformers to millions of base pairs such as Hyena (Nguyen et al., 2023) rely on dilated convolutions on the input sequence, rendering attribution to interactions difficult. Therefore, we opted for full attention using FlashAttention (Dao et al., 2022) during training, which still allows us to extract complete attention scores during attribution/validation.

Overall, our model consists of over 68 million trainable parameters. We used AdamW (Loshchilov & Hutter, 2019) with linear learning rate decay and trained for 16 epochs on 4 NVIDIA A100 GPUs until convergence of the out-of-sample classification performance on the validation set.

Attribution techniques. During training, we hold out $n_{\text{val}} = 1453$ samples for validation and our attribution analysis. The goal of our attribution technique is to extract gene-pairs or even larger collections of genes whose co-presence in a given sample is predictive of the habitat. While genes within a pair need not necessarily physically interact as in protein complexes, we posit that they 'interact' in being jointly specific to the habitat. We propose the following procedure for attribution, which we depict in Figure 2.

1. For each sample in the validation set (each consisting of a collection of fixed-size gene embeddings grouped into contigs; c.f., Figure 2) that was classified correctly with a certain confidence (top softmax value above 0.85), compute all last-layer attention maps and extract the positions (indices) of the top-k scores for a fixed $k \in \mathbb{N}$. Following common practice in the literature (starting with Vaswani et al. (2017)), we interpret high attention scores as relevant for the prediction task. Each of the extracted $n_{\text{val}} \cdot k$ indices corresponds to a pair of input gene embeddings $\{p_i := (x_i^1, x_i^2)\}_{i=1}^{n_{\text{val}} \cdot k}$ for $x_i^j \in \mathbb{R}^{d_{\text{emb}}}$.
2. In this step we use DBSCAN (Ester et al., 1996) as a clustering algorithm, which has the advantage of inferring the number of clusters by itself, and cluster via the following custom distance function

$$\text{dist}(p_i, p_j) = \min\{2 - S_c(x_i^1, x_j^1) - S_c(x_i^2, x_j^2), 2 - S_c(x_i^1, x_j^2) - S_c(x_i^2, x_j^1)\} ,$$

   where S_c is the cosine similarity and we are agnostic about the order of the genes with in the pair.
3. For each point p_i in each cluster, we recover the two gene sequences that produced the gene embeddings x_i^1, x_i^2. We then perform sequence similarity search on all these genes in the databases EggNOG (Cantalapiedra et al., 2021), KEGG orthologs (Kanehisa et al., 2015), and NCBI Blast (Altschul et al., 1990; Boratyn et al., 2019; Camacho et al., 2023) to extract functional and taxonomic annotations.
4. We propose gene interaction networks loosely inspired by gene pathways. If a certain gene appears in more than one of the pairs *within a sample*, we use these overlaps in the extracted k pairs of genes to construct a gene network. Genes that are hubs in these networks have many highly predictive interactions with other genes and may thus be of particular functional importance.

[1] Note that each of these 4096 'tokens' represents an entire gene, each of which can consist of thousands of base-pairs. Therefore, the 'effective' context window approaches 10^7 base-pairs.

3 RESULTS

Why habitat classification? The reason we focus on the seemingly 'simple' three-way classification of habitats (host, soil, and aquatic) is three-fold. First, habitat is a broad and highly complex phenotype, which is difficult to predict directly from genotype. Hence, strong performance on this task indicates that our general framework may apply equally to other phenotypes. Second, it is straightforward to compare feature attributions among all three classes in order to help validate our approach. Third, habitat annotations are typically reliable and widely available for microbiome samples. In the remainder of this section, we particularly focus on extensive internal and external validation results demonstrating that our modeling approach indeed manages to pick up on the importance of the co-presence of genes.

We conjecture that gene pairs (or collections/networks) found by our attribution technique are of biological interest in various ways. For example, when predicting host-related habitats, such gene clusters may shed light not only on specific genes, but also gene interaction networks that may be involved in colonization (Stephens et al., 2015; Powell et al., 2016b; Kemis et al., 2019). When the identified gene pairs are found in gene annotation databases and have known functional annotations, we can directly point to interactions of functional aspects associated to the predicted phenotype and potential colonization properties. On the contrary, when the found genes are part of the "microbial functional dark matter", we hypothesize they are good candidates to follow up on experimentally. For example, one could knock out the predicted genes and measure the abundance of the mutant versus wild type in a model habitat (Powell et al., 2016a; Ellison et al., 2011; Brouwer et al., 2020).

Classification performance. We evaluate our model on $n_{\text{val}} = 1453$ held out samples from the ProGenomes v3 dataset. It achieves an overall accuracy of 71% (Table 2). Given the complexity of the task (see Section 1), this is a strong performance for our 3-way classification task. Table 1 shows how performance varies across habitats: while host samples are identified well, samples from soil are often misclassified as aquatic. Biologically, host microbiomes are mostly symbiotic or parasitic, where they tend to lose unneeded portions of their genome due to deletional bias in bacterial genomes (McCutcheon & Moran, 2012; Boscaro et al., 2017). This arguably leads to substantial genomic differences from free-living microbiomes in soil or aquatic environments, which conversely can have strong adaptability due to their versatile metabolic pathway and, therefore, can survive in a variety of environments (Shu & Huang, 2022; Moreno-Gámez, 2022). There is likely also more direct mixing of microbiomes inhabiting soil and aquatic environments, rendering distinguishing soil from aquatic examples incredibly difficult. Finally, the sample imbalance in our training set is slightly skewed towards aquatic examples. In Appendix C we perform an ablation study of how the number of layers, size of feedforward layers, and embedding dimension affect our model performance.

Table 1: One-vs-rest classification performance of our method on the test set.

class	samples	precision	recall	F1
host	488	0.84	0.80	0.82
soil	412	0.63	0.43	0.51
aquatic	553	0.66	0.84	0.74

Baselines. Since our modeling approach is the only of its type, there exist large gaps in terms of data types, capability, and interoperability between existing works and ours. Table 5 in Appendix D provides a detailed comparison to existing methods, highlighting in which way most of them fall short in our problem setting. To compare raw predictive performance, we put together a strong baseline using k-mer counts as features with traditional machine learning classifiers. This is a widely popular and typically highly effective approach to supervised ML on sequence data (Dubinkina et al., 2016; Benoit et al., 2016; Wood & Salzberg, 2014). It avoids the necessity of annotations (highly incomplete for prokaryotes) and scales well to entire genomes. The best performing traditional ML models in the literature on k-mer counts and in bioinformatics more broadly are often random forests (Bi et al., 2023; Wheeler et al., 2018) and SVMs (Weimann et al., 2016a; Bi et al., 2023; Barash et al., 2018). Table 2 shows that our method achieves higher accuracy than these algorithms trained on k-mer counts for different typical values of k. Finally, we highlight that by design (using k-mer

Table 2: Accuracy for random forests and SVMs using linear and RBF kernels.

method	k-mer	acc
random forest	3	57
	5	58
	8	59
SVM linear	3	57
	5	62
	8	56
SVM rbf	3	63
	5	67
	8	68
ours	–	**71**

counts as features), these methods cannot be interpreted in terms of single gene or gene interaction importance. In Appendix B we provide an internal validation of the effectiveness of our attribution techniques using 'pseudo-samples' providing strong evidence that our method indeed identifies habitat-specific gene pairs.

Clustering. In Figure 5 we illustrate the gene pair clusters using UMAP (McInnes et al., 2018).[2] The pairs of genes indeed cluster well, indicating that gene pairs within a cluster are indeed functionally similar as measured by distance of their ESM-2 embeddings. Further, different clusters are well-separated, indicating that we have indeed identified different 'hubs' of gene interactions that are individually predictive of the habitat. For completeness, we provide similar plots using t-SNE (van der Maaten & Hinton, 2008) instead of UMAP in Figure 6 in Appendix A, showing that the clear separation of clusters is not specific to the choice of dimensionality reduction technique.

We further verified that within most found clusters, gene families are quite uniform. From the extracted functional and taxonomic annotations, we found that the clusters indeed recover biologically plausible 'functional factors'. For example, in the largest (blue) cluster from host samples, most of the pairs share the KEGG orthologs (Kanehisa et al., 2015) K01992 and K11051. The latter is known as multidrug/hemolysin transport system permease, a protein that plays an important role in bacterial infection of animal hosts. In the largest (blue) cluster from aquatic samples, most gene pairs share the K08226 functional ortholog. Genes from this ortholog code chlorophyll transporter. This matches our knowledge that most photosynthetic bacteria, such as Cyanobacteria and Chlorobi, live in water. In the largest (blue) cluster from soil samples, we found the following frequent orthologs: K01535, K01531, K17686, K01533, and K17686. These gene families are all involved in ion transport. For completeness, we provide all found orthologs in all of the clusters for the three classes in Appendix E.

Gene interaction networks. We present an example of one of the gene interaction networks constructed by our attribution technique in Figure 7. The genome from which this network was constructed belongs to Streptococcus agalactiae, a commensal bacterium. Although it colonizes the gastrointestinal and genitourinary tract of up to 30% of healthy human adults, it is still poorly understood. We could only find functional annotations for 14 of the 41 genes in the network. The rest of the genes have no annotation via our methodology. In particular, the gene with the most connections, gene 1378, is identified as a peptidoglycan bound protein that can have various functions, including roles in cell wall synthesis, cell division, and interaction with the environment. In the context of bacterial colonization, peptidoglycan-bound proteins can contribute to the adherence of bacteria to host tissues, evasion of the host immune response, and establishment of infection (Dörr et al., 2014). Further, gene 1379, another highly connected hub in our network, is involved in dextransucrase activity. Dextransucrase is an enzyme that catalyzes the formation of dextran, which can contribute to the formation of biofilms, which are communities of bacteria that adhere to surfaces. Biofilms play a crucial role in bacterial colonization, as they can protect bacteria from environmental stresses and enhance their survival and growth (Besrour-Aouam et al., 2019; Lee & Park, 2015). Finally, gene 471, yet another highly connected hub, belongs to peptidase S8 family 5, also known as subtilases. This enzyme plays important roles in colonization, including the degradation of host tissues and evasion of the host immune system (Cui et al., 2023).

Finally, we provide further validation of our attribution technique using the STRING database (Szklarczyk et al., 2023), which was specifically set up to systematically collect and integrate protein-protein interactions that contain both physical and functional associations in Appendix G.

4 Conclusion

We introduced a model that predicts complex phenotypes, such as habitat, from entire genomes on the sequence level of microbial sequencing data. Our attribution techniques allow us to extract pairs (and collections) of genes whose interaction or co-presence is highly predictive of the chosen phenotype. We train our model on a large subset of the ProGenomes v3 dataset containing high-quality prokaryotic genomes and demonstrate state of the art classification performance. Our internal and external validation evidence the usefulness of our attribution techniques in uncovering habitat-specific gene pairs and generating interpretable gene interaction networks that can serve as powerful hypotheses generators for the underlying mechanisms of complex biological processes.

[2]We omit 'outliers', i.e., points that did not belonging to any cluster after DBSCAN finished for a clearer illustration. These outliers are bound to exist due to the breadth of habitat as a phenotype.

IMPACT STATEMENT

This work presents methodological advances in the use of machine learning for predicting complex phenotypes from microbial genomic data, with potentially far-reaching implications for both the field of computational biology and society at large. By enabling more accurate predictions of habitat specificity from the genetic makeup of microbiomes and especially understanding the underlying drivers in terms of gene interactions, our research may aid innovative applications in environmental conservation, sustainable agriculture, and personalized medicine. The ability to understand and predict the interactions between microbial genes and their environments could lead to breakthroughs in the development of new biomarkers for health conditions, the creation of targeted microbiome therapies, and the enhancement of biodiversity conservation strategies.

Ethically, while the potential for positive impact is vast, we recognize the importance of considering potential downsides, especially in applications related to human and planetary health. The same understanding that may be leveraged for improved treatments, may also be used to discover or engineer particularly resistant pathological organisms. Generally, the adoption of advanced machine learning techniques in genomics must be accompanied by efforts to prevent misuse and ensure equitable access to the benefits they bring. We advocate for a continued democratic dialogue to address these challenges and ensure that the advancements in computational biology contribute positively to society and the environment.

REFERENCES

Alam, I., Kamau, A. A., Ngugi, D. K., Gojobori, T., Duarte, C. M., and Bajic, V. B. Kaust metagenomic analysis platform (kmap), enabling access to massive analytics of re-annotated metagenomic data. *Scientific reports*, 11(1):11511, 2021.

Alharbi, W. S. and Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. *Human Genomics*, 16(1):1–20, 2022.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, 39(1):105–114, 2021.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021.

Baltoumas, F. A., Karatzas, E., Liu, S., Ovchinnikov, S., Sofianatos, Y., Chen, I.-M., Kyrpides, N. C., and Pavlopoulos, G. A. Nmpfamsdb: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Research*, 52(D1):D502–D512, 2024.

Barash, E., Sal-Man, N., Sabato, S., and Ziv-Ukelson, M. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics*, 35(12):2001–2008, 11 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty928. URL https://doi.org/10.1093/bioinformatics/bty928.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020.

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023.

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. Multiple comparative metagenomics using multiset k-mer counting, 2016.

Besrour-Aouam, N., Mohedano, M. L., Fhoula, I., Zarour, K., Najjari, A., Aznar, R., Prieto, A., Ouzari, H.-I., and López, P. Different modes of regulation of the expression of dextransucrase in leuconostoc lactis av1n and lactobacillus sakei mn1. *Frontiers in Microbiology*, 10, 2019. ISSN

1664-302X. doi: 10.3389/fmicb.2019.00959. URL https://www.frontiersin.org/articles/10.3389/fmicb.2019.00959.

Bi, X., Liang, W., Zhao, Q., and Wang, J. SSLpheno: a self-supervised learning approach for gene–phenotype association prediction using protein–protein interactions and gene ontology data. *Bioinformatics*, 39(11):btad662, 11 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad662. URL https://doi.org/10.1093/bioinformatics/btad662.

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T. L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, 20(1):405, July 2019.

Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D. H., and Keeling, P. J. Parallel genome reduction in symbionts descended from closely related free-living bacteria. *Nature Ecology & Evolution*, 1(8):1160–1167, August 2017.

Brewster, R., Tamburini, F. B., Asiimwe, E., Oduaran, O., Hazelhurst, S., and Bhatt, A. S. Surveying gut microbiome research in africans: toward improved diversity and representation. *Trends in microbiology*, 27(10):824–835, 2019.

Brouwer, S., Barnett, T. C., Ly, D., Kasper, K. J., De Oliveira, D. M. P., Rivera-Hernandez, T., Cork, A. J., McIntyre, L., Jespersen, M. G., Richter, J., Schulz, B. L., Dougan, G., Nizet, V., Yuen, K.-Y., You, Y., McCormick, J. K., Sanderson-Smith, M. L., Davies, M. R., and Walker, M. J. Prophage exotoxins enhance colonization fitness in epidemic scarlet fever-causing streptococcus pyogenes. *Nature Communications*, 11(1):5018, October 2020.

Calle, M. L. Statistical analysis of metagenomics data. *Genomics & informatics*, 17(1), 2019.

Camacho, C., Boratyn, G. M., Joukov, V., Vera Alvarez, R., and Madden, T. L. ElasticBLAST: accelerating sequence search via cloud computing. *BMC Bioinformatics*, 24(1):117, March 2023.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. eggnog-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, 38(12):5825–5829, 2021.

Cao, H., Ma, Q., Chen, X., and Xu, Y. Door: a prokaryotic operon database for genome analyses and functional inference. *Briefings in bioinformatics*, 20(4):1568–1577, 2019.

Cheifet, B. Where is genomics going next?, 2019.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers, 2019.

Chklovski, A., Parks, D. H., Woodcroft, B. J., and Tyson, G. W. Checkm2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20 (8):1203–1212, 2023.

Choi, S. R. and Lee, M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology*, 12(7):1033, 2023.

Clapp, M., Aurora, N., Herrera, L., Bhatia, M., Wilen, E., and Wakefield, S. Gut microbiota's effect on mental health: The gut-brain axis. *Clinics and practice*, 7(4):987, 2017.

Collins, C. and Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS computational biology*, 14(2):e1005958, 2018.

Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F. J., Moses, A., and Wang, B. To transformers and beyond: Large language models for the genome. *arXiv preprint arXiv:2311.07621*, 2023.

Cui, H., Zhou, G., Ruan, H., Zhao, J., Hasi, A., and Zong, N. Genome-wide identification and analysis of the maize serine peptidase s8 family genes in response to drought at seedling stage. *Plants*, 12(2), 2023. ISSN 2223-7747. doi: 10.3390/plants12020369. URL https://www.mdpi.com/2223-7747/12/2/369.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

de Los Campos, G., Vazquez, A. I., Hsu, S., and Lello, L. Complex-trait prediction in the era of big data. *Trends in Genetics*, 34(10):746–754, 2018.

Deschênes, T., Tohoundjona, F. W. E., Plante, P.-L., Di Marzo, V., and Raymond, F. Gene-based microbiome representation enhances host phenotype classification. *Msystems*, 8(4):e00531–23, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Djemiel, C., Maron, P.-A., Terrat, S., Dequiedt, S., Cottin, A., and Ranjard, L. Inferring microbiota functions from taxonomic genes: a review. *Gigascience*, 11:giab090, 2022.

Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V., and Alexeev, D. G. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17(1): 38, January 2016.

Ducarmon, Q. R., Grundler, F., Le Maho, Y., de Toledo, F. W., Zeller, G., Habold, C., and Mesnage, R. Remodelling of the intestinal ecosystem during caloric restriction and fasting. *Trends in Microbiology*, 2023.

Dörr, T., Lam, H., Alvarez, L., Cava, F., Davis, B. M., and Waldor, M. K. A novel peptidoglycan binding protein crucial for pbp1a-mediated cell wall biogenesis in vibrio cholerae. *PLOS Genetics*, 10(6):1–14, 06 2014. doi: 10.1371/journal.pgen.1004433. URL https://doi.org/10.1371/journal.pgen.1004433.

D'Elia, D., Truu, J., Lahti, L., Berland, M., Papoutsoglou, G., Ceci, M., Zomer, A., Lopes, M. B., Ibrahimi, E., Gruca, A., et al. Advancing microbiome research with machine learning: key findings from the ml4microbiome cost action. *Frontiers in Microbiology*, 14, 2023.

Ellison, C. E., Hall, C., Kowbel, D., Welch, J., Brem, R. B., Glass, N. L., and Taylor, J. W. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences*, 108(7):2831–2836, 2011. doi: 10.1073/pnas.1014971108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1014971108.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

Fullam, A., Letunic, I., Schmidt, T. S., Ducarmon, Q. R., Karcher, N., Khedkar, S., Kuhn, M., Larralde, M., Maistrenko, O. M., Malfertheiner, L., et al. progenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic acids research*, 51(D1):D760–D766, 2023.

Ghurye, J. S., Cepeda-Espinoza, V., and Pop, M. Metagenomic assembly: Overview, challenges and applications. *Yale J Biol Med*, 89(3):353–362, September 2016.

Gilbert-Diamond, D. and Moore, J. H. Analysis of gene-gene interactions. *Curr Protoc Hum Genet*, Chapter 1:Unit1.14, July 2011.

Hammack, A. T. and Blaby-Haas, C. E. Machine learning sheds light on microbial dark proteins. *Nature Reviews Microbiology*, pp. 1–1, 2023.

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., and Rasmussen, S. Machine learning and deep learning applications in microbiome research. *ISME Communications*, 2(1):98, 2022.

Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature communications*, 13(1):2606, 2022.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

Huang, S., Ailer, E., Kilbertus, N., and Pfister, N. Supervised learning and model analysis with compositional data. *PLOS Computational Biology*, 19(6):e1011240, 2023.

Hwang, B., Lee, J. H., and Bang, D. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11 (1):119, Mar 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186/1471-2105-11-119.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *bioRxiv*, 2020. doi: 10.1101/2020.09.17.301879. URL https://www.biorxiv.org/content/early/2020/09/19/2020.09.17.301879.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., and Luo, Y. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, 2022.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 10 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1070. URL https://doi.org/10.1093/nar/gkv1070.

Kemis, J. H., Linke, V., Barrett, K. L., Boehm, F. J., Traeger, L. L., Keller, M. P., Rabaglia, M. E., Schueler, K. L., Stapleton, D. S., Gatti, D. M., et al. Genetic determinants of gut microbiota composition and bile acid profiles in mice. *PLoS Genetics*, 15(8):e1008073, 2019.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

Lapierre, P. and Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends in genetics*, 25(3):107–110, 2009.

Lee, C. G. and Park, J. K. Comparison of inhibitory activity of bioactive molecules on the dextransucrase from streptococcus mutans. *Applied Microbiology and Biotechnology*, 99(18):7495–7503, September 2015.

Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., and Corander, J. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, 11(4):10–1128, 2020.

Li, H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

Li, Z., Das, A., Beardall, W. A. V., Zhao, Y., and Stan, G.-B. Genomic interpreter: A hierarchical genomic deep neural network with 1d shifted window transformer, 2023.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *bioRxiv*, 2023. doi: 10.1101/2023.08.30.555582. URL https://www.biorxiv.org/content/early/2023/09/01/2023.08.30.555582.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., et al. Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550(7674):61–66, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Marsh, J. W., Kirk, C., and Ley, R. E. Toward microbiome engineering: Expanding the repertoire of genetically tractable members of the human gut microbiome. *Annual Review of Microbiology*, 77, 2023.

McCutcheon, J. P. and Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26, January 2012.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://www.biorxiv.org/content/10.1101/2021.07.09.450648v1.

Mende, D. R., Letunic, I., Huerta-Cepas, J., Li, S. S., Forslund, K., Sunagawa, S., and Bork, P. progenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res*, 45(D1):D529–D534, October 2016.

Mende, D. R., Letunic, I., Maistrenko, O. M., Schmidt, T. S. B., Milanese, A., Paoli, L., Hernández-Plaza, A., Orakov, A. N., Forslund, S. K., Sunagawa, S., Zeller, G., Huerta-Cepas, J., Coelho, L. P., and Bork, P. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, 48(D1):D621–D625, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz1002. URL https://doi.org/10.1093/nar/gkz1002.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C., Paladin, L., Raj, S., Richardson, L. J., et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

Moreno-Gámez, S. How bacteria navigate varying environments. *Science*, 378(6622):845–845, 2022. doi: 10.1126/science.adf4444. URL https://www.science.org/doi/abs/10.1126/science.adf4444.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. A genomic catalog of earth's microbiomes. *Nature biotechnology*, 39(4):499–509, 2021.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*, 2023.

Nyren, P., Pettersson, B., and Uhlen, M. Solid phase dna minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*, 208(1):171–175, 1993. ISSN 0003-2697. doi: https://doi.org/10.1006/abio.1993.1024. URL https://www.sciencedirect.com/science/article/pii/S0003269783710249.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, 50(D1):D785–D794, 2022.

Pavlopoulos, G. A., Baltoumas, F. A., Liu, S., Selvitopi, O., Camargo, A. P., Nayfach, S., Azad, A., Roux, S., Call, L., Ivanova, N. N., et al. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983):594–602, 2023.

Powell, J. E., Leonard, S. P., Kwong, W. K., Engel, P., and Moran, N. A. Genome-wide screen identifies host colonization determinants in a bacterial gut symbiont. *Proceedings of the National Academy of Sciences*, 113(48):13887–13892, 2016a. doi: 10.1073/pnas.1610856113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1610856113.

Powell, J. E., Leonard, S. P., Kwong, W. K., Engel, P., and Moran, N. A. Genome-wide screen identifies host colonization determinants in a bacterial gut symbiont. *Proceedings of the National Academy of Sciences*, 113(48):13887–13892, 2016b.

Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling, 2019.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1.

Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.

Ratiner, K., Ciocan, D., Abdeen, S. K., and Elinav, E. Utilization of the microbiome in personalized medicine. *Nature Reviews Microbiology*, pp. 1–18, 2023.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L. J., et al. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759, 2023.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Rojas-Carulla, M., Tolstikhin, I., Luque, G., Youngblut, N., Ley, R., and Schölkopf, B. Genet: Deep representations for metagenomics, 2019.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1):84–89, November 1996.

Rouli, L., Merhej, V., Fournier, P.-E., and Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes and new infections*, 7:72–85, 2015.

Salaverria, I., Philipp, C., Oschlies, I., Kohler, C. W., Kreuz, M., Szczepanowski, M., Burkhardt, B., Trautmann, H., Gesk, S., Andrusiewicz, M., Berger, H., Fey, M., Harder, L., Hasenclever, D., Hummel, M., Loeffler, M., Mahn, F., Martin-Guerrero, I., Pellissery, S., Pott, C., Pfreundschuh, M., Reiter, A., Richter, J., Rosolowski, M., Schwaenen, C., Stein, H., Trümper, L., Wessendorf, S., Spang, R., Küppers, R., Klapper, W., Siebert, R., Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe, German High-Grade Lymphoma Study Group, and Berlin-Frankfurt-Münster-NHL trial group. Translocations activating IRF4 identify a subtype of germinal center-derived b-cell lymphoma affecting predominantly children and young adults. *Blood*, 118(1):139–147, April 2011.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., et al. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, 2022.

Schmidt, T. S., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O. M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., Sunagawa, S., et al. Spire: a searchable, planetary-scale microbiome resource. *Nucleic Acids Research*, 52(D1):D777–D783, 2024.

Schupack, D. A., Mars, R. A., Voelker, D. H., Abeykoon, J. P., and Kashyap, P. C. The promise of the gut microbiome as part of individualized treatment strategies. *Nature Reviews Gastroenterology & Hepatology*, 19(1):7–25, 2022.

Shu, W.-S. and Huang, L.-N. Microbial diversity in extreme environments. *Nature Reviews Microbiology*, 20(4):219–235, April 2022.

Stephens, W. Z., Wiles, T. J., Martinez, E. S., Jemielita, M., Burns, A. R., Parthasarathy, R., Bohannan, B. J., and Guillemin, K. Identification of population bottlenecks and colonization factors during assembly of bacterial communities within the zebrafish intestine. *MBio*, 6(6):10–1128, 2015.

Sukhbaatar, S., Grave, E., Bojanowski, P., and Joulin, A. Adaptive attention span in transformers, 2019.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., and von Mering, C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, 51(D1):D638–D646, January 2023.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., and Yu, W. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2010.07.021. URL https://www.sciencedirect.com/science/article/pii/S0002929710003782.

Wei, X., Tan, H., Lobb, B., Zhen, W., Wu, Z., Parks, D. H., Neufeld, J. D., Moreno-Hagelsieb, G., and Doxey, A. C. Annoview enables large-scale analysis, comparison, and visualization of microbial gene neighborhoods. *bioRxiv*, pp. 2024–01, 2024.

Weimann, A., Mooren, K., Frank, J., Pope, P. B., Bremges, A., and McHardy, A. C. From genomes to phenotypes: Traitar, the microbial trait analyzer. *mSystems*, 1(6):10.1128/msystems.00101–16, 2016a. doi: 10.1128/msystems.00101-16. URL https://journals.asm.org/doi/abs/10.1128/msystems.00101-16.

Weimann, A., Mooren, K., Frank, J., Pope, P. B., Bremges, A., and McHardy, A. C. From genomes to phenotypes: Traitar, the microbial trait analyzer. *MSystems*, 1(6):e00101–16, 2016b.

Wheeler, N. E., Gardner, P. P., and Barquist, L. Machine learning identifies signatures of host adaptation in the bacterial pathogen salmonella enterica. *PLOS Genetics*, 14(5):1–20, 05 2018. doi: 10.1371/journal.pgen.1007333. URL https://doi.org/10.1371/journal.pgen.1007333.

Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, March 2014.

Wood, D. E., Lu, J., and Langmead, B. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13, 2019.

Xu, H., Liu, J.-J., Liu, Z., Li, Y., Jin, Y.-S., and Zhang, J. Synchronization of stochastic expressions drives the clustering of functionally related genes. *Science Advances*, 5(10):eaax6525, 2019. doi: 10.1126/sciadv.aax6525. URL https://www.science.org/doi/abs/10.1126/sciadv.aax6525.

Yang, Y. and Jiang, X. Evolink: a phylogenetic approach for rapid identification of genotype–phenotype associations in large-scale microbial multispecies data. *Bioinformatics*, 39(5):btad215, 2023.

Youngblut, N. D., de la Cuesta-Zuluaga, J., Reischer, G. H., Dauser, S., Schuster, N., Walzer, C., Stalder, G., Farnleitner, A. H., and Ley, R. E. Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. *mSystems*, 5(6):10.1128/msystems.01045–20, 2020. doi: 10.1128/msystems.01045-20. URL https://journals.asm.org/doi/abs/10.1128/msystems.01045-20.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big bird: Transformers for longer sequences, 2021.

Zhou, Y.-H. and Gallins, P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in genetics*, 10:579, 2019.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A Additional visualizations

For completeness, besides the distributions of contigs per sample and genes per sample in Figure 3, we also present the distribution of genes per contig across the three classes in Figure 4. Moreover, Figure 6 provides a visualization akin to the one in Figure 5, using t-SNE (van der Maaten & Hinton, 2008) for non-linear dimensionality reduction instead of UMAP (McInnes et al., 2018). In both visualizations, the same clusters are clearly visible and separated, indicating robustness of the found clusters to the specific dimensionality reduction technique.

B Internal validation

To provide some internal validation of the effectiveness of our attribution technique, we construct 'pseudo-examples', inputs to our model that consist only of genes that were identified by the attribution to be part of highly-predictive pairs for $k = 100$. We randomly concatenate

Table 3: Internal validation on 'pseudo-samples'.

class	samples	precision	recall	F1
host	488	0.58	0.82	0.68
soil	412	0.58	0.16	0.24
aquatic	553	0.58	0.69	0.63

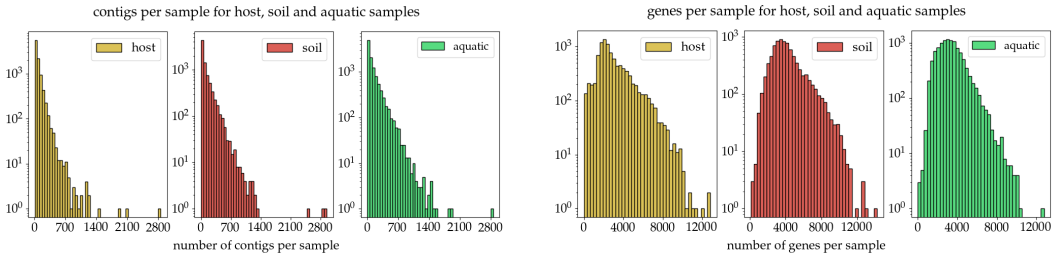

Figure 3: **Left:** Histogram of the number of contigs per sample (genome). **Right:** Histogram of the number of genes per sample (genome).

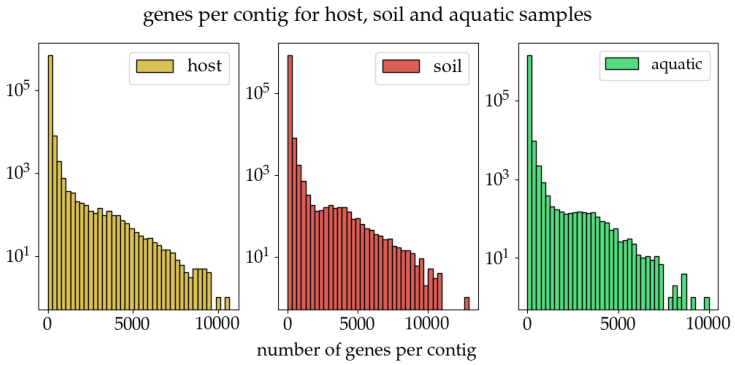

Figure 4: Histogram of the number of genes per contig.

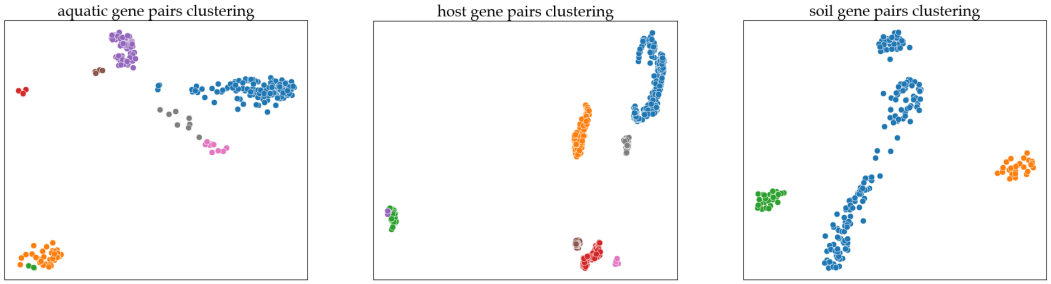

Figure 5: Two-dimensional visualization of the clusters for aquatic (left), host (middle), and soil (right) samples via UMAP (McInnes et al., 2018), omitting points not belonging to any cluster.

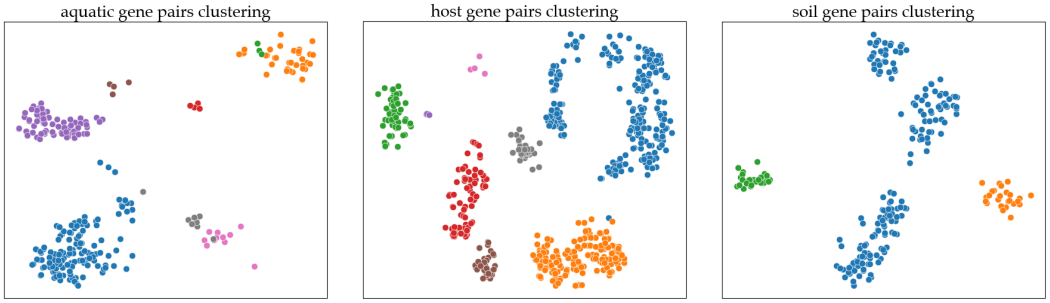

Figure 6: Two-dimensional visualization of the clusters for each of the three habitats aquatic (left), host (middle), and soil (right) separately via t-SNE (van der Maaten & Hinton, 2008), where do not show points not belonging to any cluster.

Table 4: Comparison of model configurations with different layers, layer sizes, and embedding dimension (d_{emb}). The first model configuration has been used in this paper.

layers	layer size	d_{emb}	acc (%)	class	precision	recall	F1
15	2048	2560	71.2	host	0.84	0.80	0.82
				soil	0.63	0.43	0.51
				aquatic	0.66	0.84	0.74
15	1024	2560	66.2	host	0.77	0.80	0.78
				soil	0.49	0.45	0.47
				aquatic	0.65	0.67	0.66
10	2048	2560	65.5	host	0.77	0.81	0.79
				soil	0.51	0.54	0.52
				aquatic	0.67	0.61	0.64
15	2048	1280	64.8	host	0.79	0.78	0.78
				soil	0.50	0.53	0.51
				aquatic	0.64	0.62	0.63

the respective gene embeddings from each validation example (without repetitions) to form 'pseudo-examples' which consist on average of only about 100 genes. These pseudo-examples (a) present only about 3% of the original genomes, and (b) only serve as a bag of genes in that the true order of genes on the genome (or within contigs) is lost—typically crucial information (Salaverria et al., 2011). Given those limitations, we would expect classification performance to drop to essentially random guessing unless the genes contained in the 'pseudo-examples' are indeed highly predictive for the habitat. Our model still achieves an overall accuracy of 58%, substantially better than random guessing. Table 3 shows that again, the model can apparently still extract useful information from host and aquatic 'pseudo-examples'. This provides strong evidence that gene pairs identified by our attribution, indeed contain a significant number (and important combinations) of habitat-specific genes.

C Ablation Study

We now show ablations of varying the number of attention layers, feedforward layer size, and embedding dimension separately to understand their individual contribution to our model's performance. Table 4 compares different models in terms of overall accuracy as well as one-vs-rest classification metrics for each class. These results highlight that downsizing the model in any way (10 instead of 15 layers, halving the feedforward layer size, or halving the embedding dimension using ESM2-650M instead of ESM-3B) drastically reduces performance roughly to the level of the baselines shown in Table 2. We highlight that the third and most impactful ablation of reducing the embedding dimension implies using a weaker embedding model, i.e., one may still achieve better performance by optimally compressing ESM-3B embeddings. Overall, these results may indicate that our model is not yet "larger than necessary", i.e., one may still obtain moderate performance improvements by using an even larger model.

D 'Genotype to Phenotype' Overview

Existing 'genotype to phenotype' methods. A number of approaches have be used to determine microbial phenotypes from genomic data. The most prominent are homology-based methods in which the function of a gene (or other genetic element) is inferred by a sequence similarity search to references with characterized functions. This approach is challenged by a number of factors such as a lack of characterized references and the often incorrect assumption that sequence similarity predicts functional similarity. A similar approach is genome-wide association (GWAS) of nucleotide-level variations among very closely related organisms to infer phenotype based on how genetic variation correlates to characterized phenotypic variation (de Los Campos et al., 2018; Collins & Didelot, 2018; Lees et al., 2020; Yang & Jiang, 2023). Such methods often require many closely related individuals

(e.g., intra-species) with matched high quality genome assemblies and characterized phenotypes. Another approach is the use of phylogenies to infer phenotypes of characterized sections of the evolutionary tree based on relatedness to characterized representatives. This approach is challenged by the difficulties of inferring accurate phylogenies, obtaining adequate numbers of phenotypically characterized representatives, and assuming that evolutionary relatedness correlates strongly with phenotypic similarity.

Given the often complex associations between genotype and phenotype, recent work has often leveraged machine learning to produce intricate models trained on empirical data. Traditionally, the focus has been on feature-based approaches, using genetic annotations from which phenotypes are inferred (Wood & Salzberg, 2014; Youngblut et al., 2020; Wood et al., 2019). For example, Traitar (Weimann et al., 2016b) uses support vector machines with a sparsity penalty to predict phenotypes based on Pfam annotations (Mistry et al., 2021).[3] Those features can be aggregated over large collections of genes to use as input for machine learning methods (Weimann et al., 2016a; Barash et al., 2018; Wheeler et al., 2018; Hernández Medina et al., 2022; D'Elia et al., 2023). A different approach is to ignore gene-level information and directly work on taxonomic compositional count data (Li, 2015; Calle, 2019; Knight et al., 2018; Zhou & Gallins, 2019; Huang et al., 2023). Djemiel et al. (2022) provide a high-level overview of existing work on functional inference from microbiota.

Despite the impressive progress achieved by these efforts, recent advances suggest that incorporating long stretches of genome sequences can enhance our understanding of genotype-phenotype relationships (Eraslan et al., 2019; Alharbi & Rashid, 2022; Deschênes et al., 2023; Hammack & Blaby-Haas, 2023). Deep learning applied to raw DNA data, such as CNNs for taxonomy prediction (Rojas-Carulla et al., 2019) or unsupervised training of transformers on k-mers as tokens (Ji et al., 2020), has indeed shown promise in this regard, offering a more nuanced view of the genetic underpinnings of complex phenotypes. Recently, several ML-based methods have also offered to prioritize non-coding variants; still, the recognition of disease-associated variants in complex traits, such as cancers, is challenging (Alharbi & Rashid, 2022). On a methodological level, operating on (collections of) entire genomes at the sequence level remains difficult (Alharbi & Rashid, 2022). Even recent approaches to scale transformer models to longer sequences via linear attention models (Dai et al., 2019; Sukhbaatar et al., 2019; Rae et al., 2019; Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2021) or reducing sequence lengths up front by stacked shifted window transformers (Liu et al., 2021) cannot directly be scaled to entire (collections of) genomes.

In summary, studying how interactions among large collections of genes/proteins relate to complex phenotypes (such as habitat) directly from sequence level data holds great promise to advance our understanding of how the microbiome interacts with hosts and environments alike.

Overview of existing methods. We compiled Table 5 summarizing the capabilities of most of the other potentially competing existing methods that we described in the related work. We assess them along the key requirements we set for our method, namely a) whether functional or taxonomic annotations/matches in existing databases are needed (often scarce for microbial life), b) whether they make use of full (coding) sequence information and scale to the full (coding) genome, and c) whether they allow for gene (interaction) attribution (requiring some sort of assessment of the influence or importance of all possible gene pairs on the prediction).

In essence, existing approaches primarily fall short in at least one of the following two ways: a) They do not take into account the full sequence information, but only highly abstracted annotations of individual genes. In this work, we consider all coding regions of a genome to qualify as "full sequence" as well. These approaches typically do not have enough information for strong prediction performance, especially for prokaryotes where much less is known about a much larger fraction of organisms, c.f. "microbial dark matter". b) They do not allow for pair-wise attribution, either because they have an excessively fine-grained granularity (Nguyen et al., 2023; Rojas-Carulla et al., 2019), which makes gene-level identification intractable, or they accommodate large sequences via "incomplete" attention computations (Zaheer et al., 2021; Beltagy et al., 2020). Most approaches to increase the maximally allowed input sequence length of transformers is by reducing the attention

[3]There is a wide array of resources and platforms for computational microbiome research, such as the MGnify platform for microbiome sequence data analysis (Richardson et al., 2023), SPIRE for searchable database integrating diverse information derived from metagenomes including many modalities (Schmidt et al., 2024), online analysis platforms (Alam et al., 2021), and more traditional protein family databases/mappers like NMPFamsDB (Baltoumas et al., 2024) or eggNOG (Cantalapiedra et al., 2021).

Table 5: Comparison of existing models in the literature with respect to the relevant aspects in our problem setting. The "partial" symbol (✓) refers to the following. *using full sequence*: HyenaDNA up to 1m bps; Borzoi up to 524k bps. *attribution*: only for fragments of genes that cannot be pre-selected; *classification*: adaptations to the original model are required; *Using full sequence* means working with sequence level information directly and includes both the full genome as well as all coding regions.

model	annotations required	using full sequence	gene attribution	reference
ours	✗	✓	✓	–
baselines	✗	✓	✗	described in Section 3
HyenaDNA	✗	(✓)	(✓)	(Nguyen et al., 2023)
Enformer	✗	✗	(✓)	(Avsec et al., 2021)
Genomic Interpreter	✗	✗	(✓)	(Li et al., 2023)
Kraken 2	✓	✓	✗	(Wood et al., 2019)
Traitar	✓	✗	✓	(Weimann et al., 2016b)
BacPaCS	✓	✗	✓	(Barash et al., 2018)
Genet	✗	✗	(✓)	(Rojas-Carulla et al., 2019)
DNABERT	✗	✗	✗	(Ji et al., 2020)
Geneformer	✗	✗	✗	(Theodoris et al., 2023)
Borzoi	✗	(✓)	(✓)	(Linder et al., 2023)

computation such that the quadratic cost is reduced (typically) to scale roughly linearly in the sequence length. This inevitably means that we do not get all pairwise attention scores within any given sample, which is what our attribution method is built on. If we relied on one of the linear attention methods, we would have to work with approximate/incomplete attributions as well—leading to missing relevant interactions only present in a subset of examples. Another way of reducing the computational cost is by compressing the input sequences in the first place, e.g., via strided convolutions or other techniques to compress sequences (Avsec et al., 2021; Benegas et al., 2023; Linder et al., 2023). Instead of concatenating all gene sequences and compressing them jointly (thereby typically losing information about gene boundaries), our approach leverages existing large protein models to preserve genes as individual entities (but in a fixed-size vector representation instead of the base pair sequence).

E Cluster ortholog annotations

In Tables 6 to 8 we list all found orthologs from all the clusters in the three different habitats shown in Figure 5. We provide this list as it demonstrates how our method can produce compact results that can be used by domain experts to inform their experiments and provide hypotheses for relevant interactions. For concrete instances, one can swiftly look up these orthologs in databases (with usable online tools available) to get an idea of which genes have been clustered and which are important hubs within our gene interaction networks.

F Gene interaction networks

We provide two additional examples of gene interaction networks from the aquatic and soil habitats. The network in Figure 8 is from an aquatic genome sample of Prochlorococcus marinus. The network in Figure 9 is from a soil genome sample of class Acidimicrobiia (unknown species). Comparably,

Table 6: Host cluster gene orthologs.

cluster	KEGG orthologs
blue	K01992, K11051, K01095, K02950, K02887, K03628, K02992, K02952, K03438, K02986, K02874, K02358, K03686, K06168, K02913, K02988, K06217, K04077, K01338, K03544
orange	K01537, K03043, K01624, K02945, K03553, K00611, K03496, K00088, K02976, K14623, K07254, K00549, K18929, K03621
green	K02913, K02945, K03665, K06958
red	K02935, K00817, K02950
purple	K06898, K01937, K01677, K04565, K03628, K01939, K00052 ,K07246, K00097, K22024, K06334, K00937, K02996, K03816, K00533, K03070, K02431, K18843, K01571, K09124, K01892, K00335, K03658, K03086, K00773, K00640
brown	K04751, K04752, K03628, K00573
pink	K02886, K07448, K03106
grey	K06996, K03856

Table 7: Aquatic cluster gene orthologs.

cluster	KEGG orthologs
blue	K08226, K02200, K07712, K01578, K00937, K10716, K06916, K17226, K00567, K00873, K07304, K07313, K01947, K03525, K02045, K11712, K10943, K01104, K02844, K14335, K03628, K06929, K03684, K00570, K03753, K04096, K01430, K01939, K08483, K09984, K01259, K03825, K07068, K10912, K08311, K03806, K08929, K05982, K18092, K17227, K01772, K00077, K02498, K00052, K02313, K08963, K07636, K00097, K22024, K00147, K01972, K07667, K09888, K01525
orange	K01537, K03644, K03665, K00937, K01533K17686, K06916, K12297, K01626, K03695, K02017, K10763, K01578, K00254, K00931, K01534, K00574, K00773, K00325, K06921, K06954, K03723, K03466, K03786, K03655, K03656, K03694, K03555, K02669, K14682, K13821
green	K00548,K01533, K17686, K01534, K01649
red	no annotation
purple	K03321, K14518, K02711, K00627, K00645, K01572, K02160, K09966
brown	K01535, K01537, K01537
pink	K15012, K02112, K01561, K01626, K06861, K02010, K02017, K00554, K03644, K06217, K00937, K01996
grey	no annotation

Table 8: Soil cluster gene orthologs.

cluster	KEGG orthologs
blue	K01535, K01531, K17686, K01533, K17686, K00937, K03644, K00567, K22319, K00873, K16329, K07568, K14415, K03657, K03750, K07219, K13599, K07146, K02428, K03495, K12132, K11212, K00574, K08256, K00226, K00254, K01006, K01921, K01588, K15371, K06442, K00641, K07020, K14414, K03183, K01939, K07646, K01812, K01835, K01840, K07566, K14652, K00260, K00261, K01972, K00471, K00955, K05838, K06949, K00794, K14941, K01903, K03526, K07738, K00548, K01338
orange	K21020, K01768, K07712, K07713, K07588, K02584, K06714, K07659, K05962
green	K04750, K01246

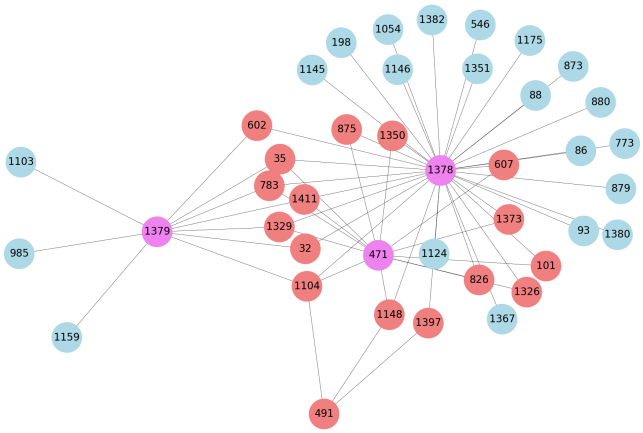

Figure 7: Gene interaction network constructed for the sample 1311.SAMN14644158. Coral color indicates genes with more than one neighbor (hub), while blue indicates genes with only one connection (peripheral). Genes are numbered by the order of their appearance on the genome. Purple hubs are described in the text.

little can be said about the precise meaning and function of the key hubs in these networks, which highlights the fact that much less is known about free-living bacteria compared to the ones living in a host as they are comparably more relevant for human health and disease. We thus leave these as two examples of relatively understudied and potentially interesting hypotheses to be followed up on experimentally.

G VALIDATION ON THE STRING DATABASE.

In an attempt to further validate the biological relevance of our attribution technique, we turn to the recently released STRING database (Szklarczyk et al., 2023). This database was specifically set up to systematically collect and integrate protein-protein interactions that contain both physical and functional associations. Unfortunately, a majority of prokaryotic genes in our dataset are not found in the STRING database. However, we could still identify some genes in our validation set with matches in the STRING database. We now manually compare the results of our attribution technique with entries in the STRING database for genes related to the survival of prokaryotes, such as DNA replication. In sample 91844.SAMEA2820670, which is identified as Candidatus Portiera, our model identifies gene 249 as a hub. This gene is annotated as DNA polymerase III beta subunit (dnaN), which is correctly found to interact with genes annotated as DNA polymerase III delta' subunit (gene 36, holB), DNA polymerase III epsilon subunit (gene 54, dnaQ) and type IIA topoisomerase (DNA

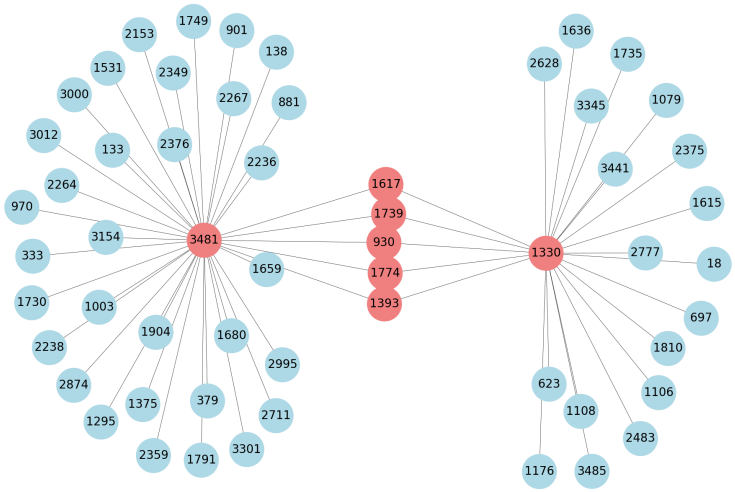

Figure 8: Gene interaction network constructed for the sample 159733.SAMEA6070310. Coral color indicates genes with more than one neighbor (hub), while blue indicates genes with only one connection (peripheral). Genes are numbered by the order of their appearance on the genome.

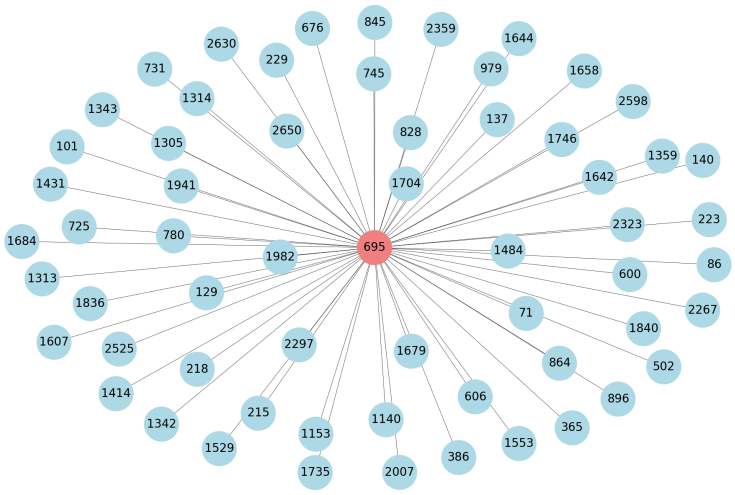

Figure 9: Gene interaction network constructed for the sample 2024894.SAMN08179843. Coral color indicates genes with more than one neighbor (hub), while blue indicates genes with only one connection (peripheral). Genes are numbered by the order of their appearance on the genome.

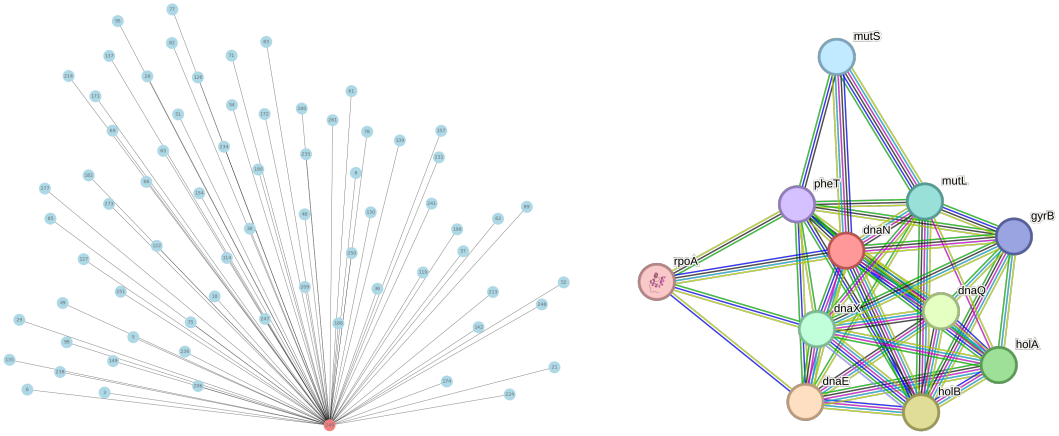

Figure 10: **Left:** Gene interaction network constructed for the sample 91844.SAMEA2820670. Coral color indicates genes with more than one neighbor (hub), while blue indicates genes with only one connection (peripheral). Genes are numbered by the order of their appearance on the genome. **Right:** Protein-protein interactions extracted from the STRING database (Szklarczyk et al., 2023) around dnaN. Only edges in magenta color are experimentally verified. Edges in other colors are predictions from the database.

gyrase/topo II, topoisomerase IV) B subunit (gene 250, gyrB), see Figure 10(left). The final DNA polymerase III is a result of pairwise interactions of the subunits. In comparison, the similar (albeit more difficult to interpret) complex shown in Figure 10(right) is obtained from the STRING database.