Whole Genome Transformer for Gene Interaction Effects in Microbiome Habitat Specificity

Zhufeng Li[1,2,3], **Sandeep S Cranganore**[4,5], **Nicholas Youngblut**[6], **Niki Kilbertus**[1,2,3]
[1]Technical University of Munich
[2]Helmholtz Munich
[3]Munich Center for Machine Learning (MCML)
[4]Forschungszentrum Jülich
[5]Technical University of Vienna
[6]Max Planck Institute for Biology, Tübingen
{zhufeng.li,niki.kilbertus}@tum.de

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.

Code — https://github.com/zhufengli/prokaformer
Extended version — https://arxiv.org/abs/2405.05998

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 and Moore 2011; Wan et al. 2010). Modern high-throughput sequencing techniques such as massively parallel methods (Ronaghi et al. 1996; Nyren, Pettersson, and Uhlen 1993; Nayfach et al. 2021) or single cell RNA sequencing (Hwang, Lee, and Bang 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) demonstrated 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 and Lee (2023) provided 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 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 works like ESM (Lin et al. 2023) and LookingGlass (Hoarfrost et al. 2022) included a large degree of known microbial diversity, these 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 much more genotypic complexity than models trained on single genes or 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 and 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, the cellular functioning of most microbial genes and non-coding elements is unknown, which has led 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, Kirk, and Ley 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.

Existing 'genotype to phenotype' methods. A number of approaches have been 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 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 and Didelot 2018; Lees et al. 2020; Yang and Jiang 2023).

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 and Salzberg 2014; Youngblut et al. 2020; Wood, Lu, and Langmead 2019). For example, Traitar (Weimann et al. 2016a) uses support vector machines with a sparsity penalty to predict phenotypes based on Pfam annotations (Mistry et al. 2021). Those features can be aggregated over large collections of genes to use as input for machine learning methods (Weimann et al. 2016b; Barash et al. 2018; Wheeler, Gardner, and Barquist 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 and Gallins 2019; Huang et al. 2023). Djemiel et al. (2022) provides 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 and Rashid 2022; Deschênes et al. 2023; Hammack and 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. On a methodological level, operating on (collections of) entire genomes at the sequence level remains difficult (Alharbi and Rashid 2022). Even recent approaches to scale transformers to longer sequences via linear attention models (Dai et al. 2019; Sukhbaatar et al. 2019; Rae et al. 2019; Child et al. 2019; Beltagy, Peters, and Cohan 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.

Contributions. In this work, we focus on habitat specificity, i.e., predicting and understanding a complex phenotype—where a microbiome sample was collected—directly from the collection of genomes of the organisms in the sample. First, we phrase this as a classification task via the following steps: (a) comprehensive gene identification from a sample of genomic sequences, (b) compute fixed-size vector representations for each gene using the multi-billion parameter ESM-2 model (Lin et al. 2023), (c) train our encoder-only transformer on sequences of gene embeddings with a classification head for the habitat.

Next, we develop attribution techniques to extract highly habitat-predictive pairs of genes by: (a) retrieving pairs of genes with high attention scores, (b) clustering pairs by similarity, (c) looking up the genes from pairs within a cluster in existing databases (Cantalapiedra et al. 2021b), and (d) constructing gene interaction networks for individual genomes. We train our model on a large subset of ProGenomes v3, a dataset of almost 1 million high-quality prokaryotic genomes (Fullam et al. 2023).

Our empirical evaluations provide multiple insights. (a) Given the complexity of the phenotype, we obtain strong classification performance. (b) Our attribution is among the first to assess the importance of gene co-occurrence across entire genomes for phenotype prediction. It recovers some known interactions, and we hypothesize that it proposes good candidates for experimental follow up. (c) Our findings indicate that exploiting sequence level information is beneficial compared to functional or taxonomic annotations when predicting phenotype from genotype—in line with recently stated conjectures (Deschênes et al. 2023; Hammack and Blaby-Haas 2023). In summary, studying how interactions among large collections of genes/proteins relate to complex phenotypes (such as habitat) directly from sequence level

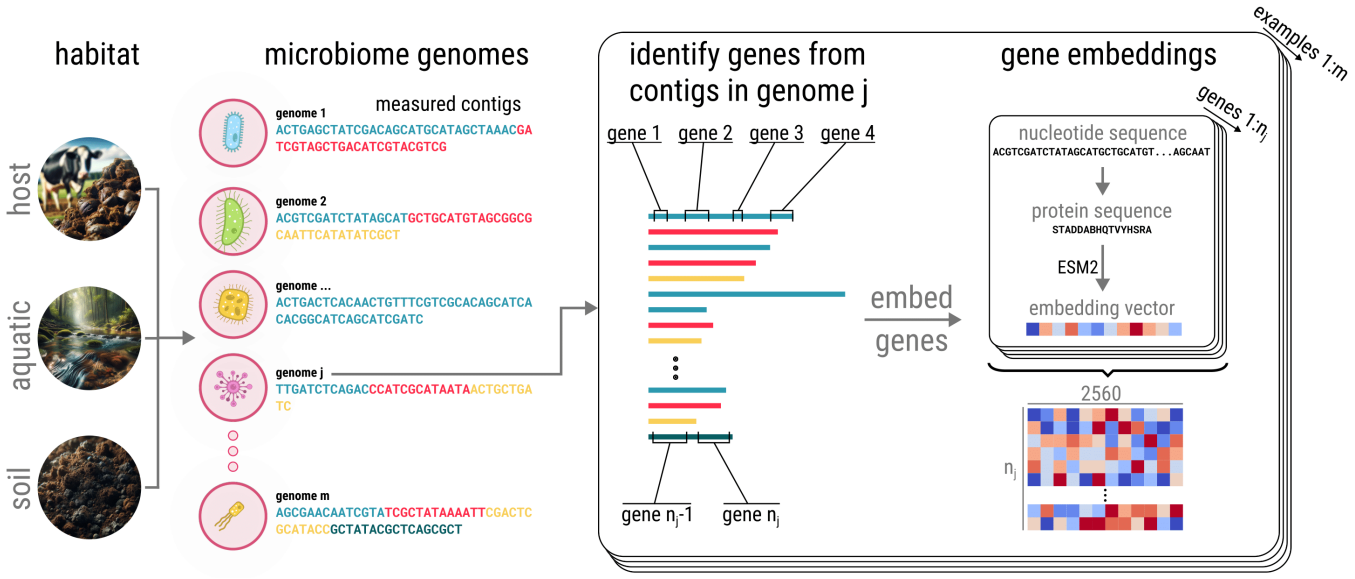

Figure 1: A conceptual overview of our data preprocessing pipeline. Each sample stands for an entire genome, reconstructed from shotgun sequencing in terms of contiguous consensus regions (contigs). We identify all genes within each contig (using Prodigal) and embed the corresponding protein sequences using an existing protein large language model (ESM-2) into a d_{emb}-dimensional vector space. A single 'input example', corresponding to an entire genome, is ultimately represented by a $(n_j \times d_{\text{emb}})$-dimensional tensor.

data holds great promise to advance our understanding of how the microbiome interacts with hosts and environments alike. While we focus on habitat specificity, we highlight that our methodology is not limited to such broad classification tasks from microbial data, but extends to other tasks and domains.

2 Methodology

Microbiome data. Various peculiarities arise from the prevailing sequencing technology (Ghurye, Cepeda-Espinoza, and Pop 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. 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 to account 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 the 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. The distribution of the number of contigs per sample is highly skewed, with most samples containing relatively few contigs. Similarly, the number of genes per contig shows a pronounced skew towards small contigs, where the majority contain only a few genes. In contrast, the total number of genes per sample demonstrates a wider range, with many samples containing a substantial number of genes, reflecting the overall variability across samples. The common peak at around 4,000 genes per sample 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 a 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 (cf., attribution paragraph in Section 2). Hence, we propose an encoder-only BERT-like architecture (Devlin et al. 2019) for classification (using the standard cross-entropy loss) with 15 layers, a single attention head, 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, we truncate them. 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 (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 and Hutter 2019) with linear learning rate decay for 16 epochs on 4 NVIDIA A100 (40GB) GPUs until convergence of the out-of-sample 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) that was classified correctly with 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 to 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
$$\begin{aligned}\text{dist}(p_i, p_j) = \min\{&2 - S_c(x_i^1, x_j^1) - S_c(x_i^2, x_j^2),\\ &2 - S_c(x_i^1, x_j^2) - S_c(x_i^2, x_j^1)\}\,,\end{aligned}$$
where S_c is the cosine similarity and we are agnostic about the order of the genes within 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. 2021a), 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.

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

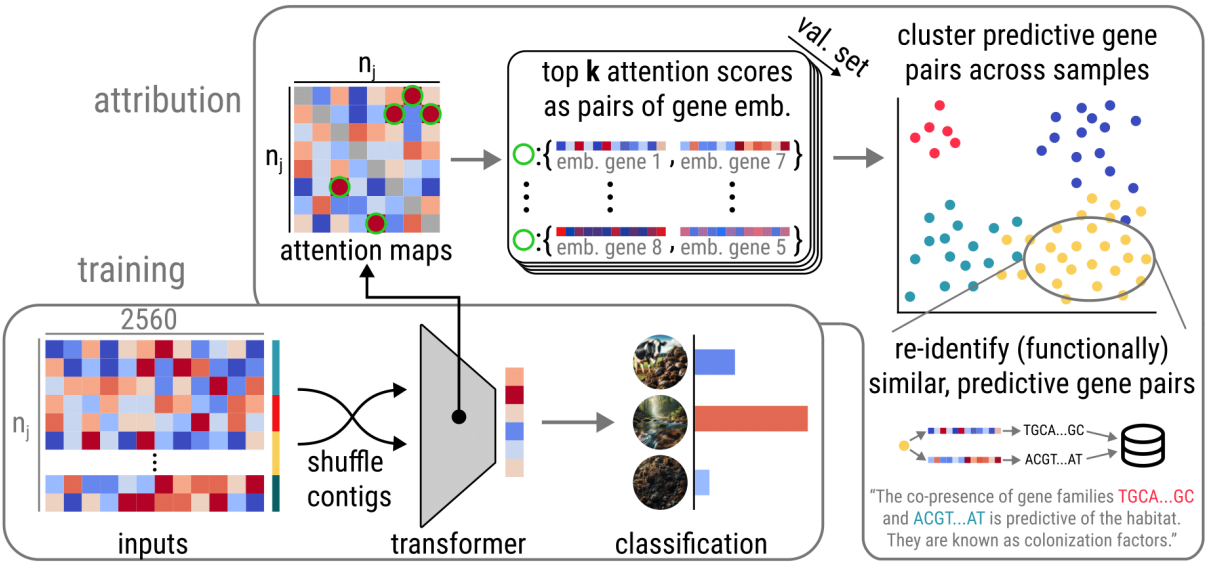

Figure 2: A conceptual overview of our training and attribution pipelines. **Training:** We feed the $(n_j \times d_{\text{emb}})$-dimensional inputs to our transformer, interpreted as a sequence of n_j 'tokens', each already represented by a fixed d_{emb}-dimensional embedding. We randomly shuffle the contigs within each sample, since the 'correct' order is unknown. Our model is then trained with the cross-entropy loss for classification. **Attribution:** After training, we extract the last-layer attention maps for all validation samples. We find the indices of the top-k attention scores in each map, i.e., which gene embedding attends strongly to which other gene embedding. We cluster these pairs and visualize the clustering via non-linear dimensionality reduction. Within each cluster, we then re-identify the nucleotide sequences of all genes within all pairs and match them against gene annotation databases.

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 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 on gene interaction networks that may be involved in colonization (Stephens et al. 2015; Powell et al. 2016a; 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 with 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. 2016b; Ellison et al. 2011).

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 1). Given the complexity of the task (see Section 1), this is a strong performance for our 3-way classification

class	samples	precision	recall	F1
test set				
host	488	0.84	0.80	0.82
soil	412	0.63	0.43	0.51
aquatic	553	0.66	0.84	0.74
pseudo-samples				
host	488	0.58	0.82	0.68
soil	412	0.58	0.16	0.24
aquatic	553	0.58	0.69	0.63

Table 1: One-vs-rest performance of our model.

task. For more detailed comparisons with baselines and ablations of our model we refer to the extended version of this manuscript. Table 1(top) shows how performance varies across habitats: while host samples are identified well, samples from the 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 and 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 and Huang 2022; Moreno-Gámez 2022). There

is likely also a 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.

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 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. Given those limitations, we would expect classification performance to drop to essentially random guessing unless the genes contained in the 'pseudo-examples' are highly predictive for the habitat. Our model still achieves an overall accuracy of 58%, substantially better than random guessing. Table 1(bottom) shows that the model can still extract useful information from host and aquatic 'pseudo-examples'. This provides strong evidence that gene pairs identified by our attribution contain a significant number (and important combinations) of habitat-specific genes.

Clustering. The purpose of running gene pair clustering is two-fold: it serves as additional validation that allows us to judge the consistency of gene pair prediction across all genomes from the same environment. At the same time, it can provide us with new perspectives on understanding the function of genes and the relationship between genotype and phenotype. We expect gene pairs within the same cluster to have similar functions, and that a cluster reflects common gene families shared by microbes from a given habitat.

In Figure 3 we illustrate the gene pair clusters using UMAP (McInnes, Healy, and Melville 2018).[1] The pairs of genes cluster well, indicating that gene pairs within a cluster are functionally similar as measured by the distance of their ESM-2 embeddings. Further, different clusters are well-separated, indicating that we have identified different 'hubs' of gene interactions that are individually predictive of the habitat.

We further verify that within most found clusters, gene families are quite uniform. From the extracted functional and taxonomic annotations, we find that the clusters 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. All found orthologs in all of the clusters for the three classes are shown in the extended version. Great care must be taken when associating biochemical functions of single-gene coded proteins with complex phenotypes. However, we believe that surfacing interpretable pointers toward potentially relevant interactions from full genome data is a promising tool to guide hypothesis formation for experimental colonization studies.

Gene interaction networks. We present an example of one of the gene interaction networks constructed by our attribution technique in Figure 4. The genome from which this network is 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 and 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).

These examples of gene annotations demonstrate our model's capability to predict not only habitat-specific genes but also how and with which other genes interact to become highly predictive of the habitat. We hope to demonstrate with this example how our framework could be used by biologists to investigate concrete scientific questions around the relevance of gene interactions in complex phenotypes. Further, we highlight that besides confirmatory evidence, our model can also be used to extract highly connected hubs across a large number of samples that are not found in existing databases, i.e., that are part of the "functional microbial dark matter". Such genes may be particularly well suited for experimental study in the quest of uncovering "microbial dark matter". Examples of gene interaction networks for the other two habitats can be found in the extended version.

[1] We omit 'outliers', i.e., points that do not belong 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.

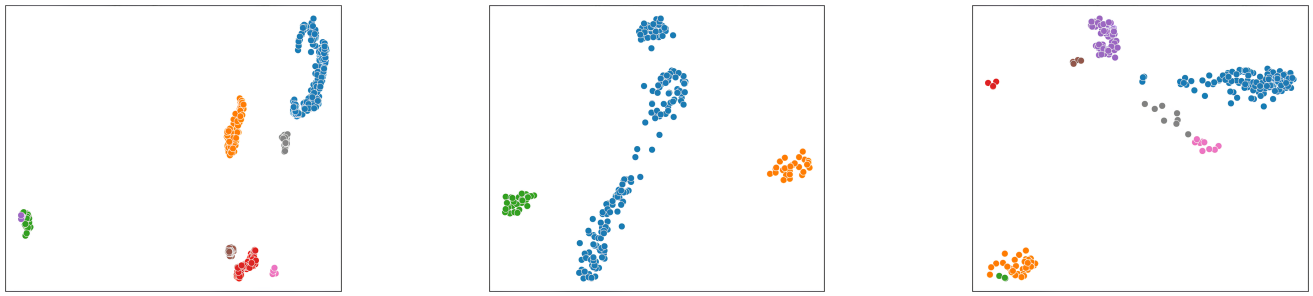

Figure 3: Two-dimensional visualization of the clusters for host (left), soil (middle), and aquatic (right) samples via UMAP (McInnes, Healy, and Melville 2018), omitting points not belonging to any cluster.

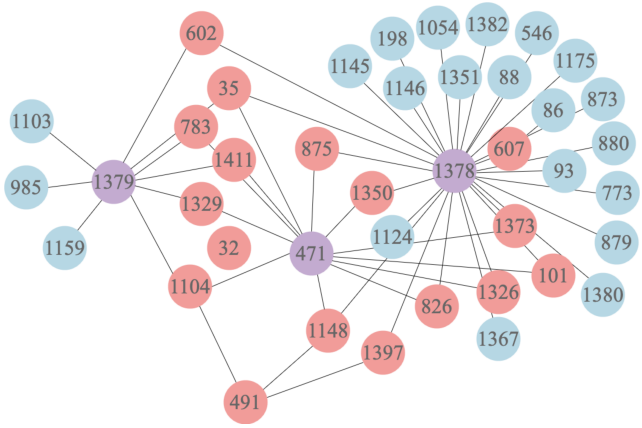

Figure 4: Gene interaction network for the sample 1311.SAMN14644158; coral/violet: genes with more than one neighbor (hubs); blue: genes with one neighbor (peripheral). Violet hubs are described in the text. Numbers are in order of appearance on the genome.

4 Discussion and Outlook

Summary. We introduced a model predicting complex phenotypes, such as habitat, from entire genomes on the sequence level of microbial sequencing data. Our attribution technique extracts pairs (and collections) of genes whose co-presence is highly predictive of the phenotype. We trained our model on high-quality prokaryotic genomes from ProGenomes v3 and demonstrated state of the art classification performance. Internal and external validations evidence the usefulness of our method in uncovering habitat-specific gene pairs and generating interpretable gene interaction networks that can serve as powerful hypothesis generators.

Limitations and future work. Our method handles large effective context lengths while preserving genes as meaningful input units. However, the context length is ultimately still memory limited due to full attention computations. Moreover, in line with our attribution goals, our analysis is limited to the coding regions of genomes. Assessing how non-coding regions affect classification and attribution is an interesting direction for future work. ESM-2 has primarily been trained on eukaryote proteins. While our results indicate that ESM-2 still provides informative embeddings for prokaryotes, further analysis is required to assess whether existing models are "general purpose" enough to capture microbial diversity. We believe that training similar foundation models specifically for microbiome research is a worthwhile endeavor. Finally, we only used habitat as a phenotype.

Other directions for future work include applying our general framework to more fine-grained classification tasks such as predicting host range (Ji et al. 2023), geographic distributions, virulence, or industrially important metabolic products. For example, when predicting antimicrobial resistance, our attribution may uncover gene networks involved in developing certain types of antimicrobial resistance. Ultimately, experimental follow ups are required to confirm the potential impact of our hypothesis generator on biological practice. Finally, while not necessarily novel, the broader framework of representing variable length collections of variable length sequences by replacing inner sequences via fixed-size embeddings from large sequence models holds great promise for future multi-omics data analysis.

Acknowledgements

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre.

References

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

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. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, 39(1): 105–114.

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

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. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10): 1196–1203.

Barash, E.; Sal-Man, N.; Sabato, S.; and Ziv-Ukelson, M. 2018. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics*, 35(12): 2001–2008.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Besrour-Aouam, N.; Mohedano, M. L.; Fhoula, I.; Zarour, K.; Najjari, A.; Aznar, R.; Prieto, A.; Ouzari, H.-I.; and López, P. 2019. Different Modes of Regulation of the Expression of Dextransucrase in Leuconostoc lactis AV1n and Lactobacillus sakei MN1. *Frontiers in Microbiology*, 10.

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

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

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

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

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

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

Cantalapiedra, C. P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; and Huerta-Cepas, J. 2021b. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12): 5825–5829.

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

Cheifet, B. 2019. Where is genomics going next? *Genome Biology*, 20(1): 17.

Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509.

Chklovski, A.; Parks, D. H.; Woodcroft, B. J.; and Tyson, G. W. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8): 1203–1212.

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

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

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

Consens, M. E.; Dufault, C.; Wainberg, M.; Forster, D.; Karimzadeh, M.; Goodarzi, H.; Theis, F. J.; Moses, A.; and Wang, B. 2023. To Transformers and Beyond: Large Language Models for the Genome. *arXiv preprint arXiv:2311.07621*.

Cui, H.; Zhou, G.; Ruan, H.; Zhao, J.; Hasi, A.; and Zong, N. 2023. Genome-Wide Identification and Analysis of the Maize Serine Peptidase S8 Family Genes in Response to Drought at Seedling Stage. *Plants*, 12(2).

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv:1901.02860.

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. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023–01.

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

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

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

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

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

Dörr, T.; Lam, H.; Alvarez, L.; Cava, F.; Davis, B. M.; and Waldor, M. K. 2014. A Novel Peptidoglycan Binding Protein Crucial for PBP1A-Mediated Cell Wall Biogenesis in Vibrio cholerae. *PLOS Genetics*, 10(6): 1–14.

D'Elia, D.; Truu, J.; Lahti, L.; Berland, M.; Papoutsoglou, G.; Ceci, M.; Zomer, A.; Lopes, M. B.; Ibrahimi, E.; Gruca, A.; et al. 2023. Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Frontiers in Microbiology*, 14.

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

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

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. 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, 226–231. AAAI Press.

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. 2023. proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic acids research*, 51(D1): D760–D766.

Ghurye, J. S.; Cepeda-Espinoza, V.; and Pop, M. 2016. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J Biol Med*, 89(3): 353–362.

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

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

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

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

Hsu, C.; Verkuil, R.; Liu, J.; Lin, Z.; Hie, B.; Sercu, T.; Lerer, A.; and Rives, A. 2022. Learning inverse folding from millions of predicted structures. *ICML*.

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

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

Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; and Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1): 119.

Ji, Y.; Shang, J.; Tang, X.; and Sun, Y. 2023. HOTSPOT: hierarchical host prediction for assembled plasmid contigs with transformer. *Bioinformatics*, 39(5): btad283.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2020. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *bioRxiv*.

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

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. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; and Tanabe, M. 2015. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1): D457–D462.

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. 2019. Genetic determinants of gut microbiota composition and bile acid profiles in mice. *PLoS Genetics*, 15(8): e1008073.

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

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

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

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

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

Li, Z.; Das, A.; Beardall, W. A. V.; Zhao, Y.; and Stan, G.-B. 2023. Genomic Interpreter: A Hierarchical Genomic Deep Neural Network with 1D Shifted Window Transformer. arXiv:2306.05143.

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. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030.

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. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, 550(7674): 61–66.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Marsh, J. W.; Kirk, C.; and Ley, R. E. 2023. Toward Microbiome Engineering: Expanding the Repertoire of Genetically Tractable Members of the Human Gut Microbiome. *Annual Review of Microbiology*, 77.

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

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

Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; and Rives, A. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*.

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

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. 2019. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, 48(D1): D621–D625.

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. 2021. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1): D412–D419.

Moreno-Gámez, S. 2022. How bacteria navigate varying environments. *Science*, 378(6622): 845–845.

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

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

Nyren, P.; Pettersson, B.; and Uhlen, M. 1993. Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry*, 208(1): 171–175.

Parks, D. H.; Chuvochina, M.; Rinke, C.; Mussig, A. J.; Chaumeil, P.-A.; and Hugenholtz, P. 2022. 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.

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. 2023. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983): 594–602.

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

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

Rae, J. W.; Potapenko, A.; Jayakumar, S. M.; and Lillicrap, T. P. 2019. Compressive Transformers for Long-Range Sequence Modelling. arXiv:1911.05507.

Rao, R.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J. F.; Abbeel, P.; Sercu, T.; and Rives, A. 2021. MSA Transformer. *bioRxiv*.

Rao, R. M.; Meier, J.; Sercu, T.; Ovchinnikov, S.; and Rives, A. 2020. Transformer protein language models are unsupervised structure learners. *bioRxiv*.

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

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; and Fergus, R. 2019. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *PNAS*.

Rojas-Carulla, M.; Tolstikhin, I.; Luque, G.; Youngblut, N.; Ley, R.; and Schölkopf, B. 2019. GeNet: Deep Representations for Metagenomics. arXiv:1901.11015.

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

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

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

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

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

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

Sukhbaatar, S.; Grave, E.; Bojanowski, P.; and Joulin, A. 2019. Adaptive Attention Span in Transformers. arXiv:1905.07799.

Tay, Y.; Dehghani, M.; Abnar, S.; Shen, Y.; Bahri, D.; Pham, P.; Rao, J.; Yang, L.; Ruder, S.; and Metzler, D. 2021. Long Range Arena : A Benchmark for Efficient Transformers. In *International Conference on Learning Representations*.

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

Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N. L.; and Yu, W. 2010. 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.

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

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

Weimann, A.; Mooren, K.; Frank, J.; Pope, P. B.; Bremges, A.; and McHardy, A. C. 2016b. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. *mSystems*, 1(6): 10.1128/msystems.00101–16.

Wheeler, N. E.; Gardner, P. P.; and Barquist, L. 2018. Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica. *PLOS Genetics*, 14(5): 1–20.

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

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

Xu, H.; Liu, J.-J.; Liu, Z.; Li, Y.; Jin, Y.-S.; and Zhang, J. 2019. Synchronization of stochastic expressions drives the clustering of functionally related genes. *Science Advances*, 5(10): eaax6525.

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

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. 2020. 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.

Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; and Ahmed, A. 2021. Big Bird: Transformers for Longer Sequences. arXiv:2007.14062.

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

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