GenomeOcean: An Efficient Genome Foundation Model Trained on Large-Scale Metagenomic Assemblies

Zhihan Zhou[1], Robert Riley[2], Satria Kautsar[2], Weimin Wu[1], Rob Egan[2], Steven Hofmeyr[2], Shira Goldhaber-Gordon[3], Mutian Yu[1], Harrison Ho[2,4], Fengchen Liu[2,5], Feng Chen[7], Rachael Morgan-Kiss[6], Lizhen Shi[1], Han Liu[1], and Zhong Wang[†2,4]

[1]Northwestern University, Evanston, IL, USA, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [3]John Hopkins University, Baltimore, MD, USA, [4]University of California at Merced, Merced, CA, USA, [5]University of California at Berkeley, Berkeley, CA, USA, [6]Miami University, Oxford, Ohio, USA, [7]Illumina Inc., Foster City, CA, USA

ABSTRACT

Genome foundation models hold transformative potential for precision medicine, drug discovery, and understanding complex biological systems. However, existing models are often inefficient, constrained by suboptimal tokenization and architectural design, and biased toward reference genomes, limiting their representation of low-abundance, uncultured microbes in the rare biosphere. To address these challenges, we developed **GenomeOcean**, a 4-billion-parameter generative genome foundation model trained on over 600 Gbp of high-quality contigs derived from 220 TB of metagenomic datasets collected from diverse habitats across Earth's ecosystems. A key innovation of GenomeOcean is training directly on large-scale co-assemblies of metagenomic samples, enabling enhanced representation of rare microbial species and improving generalizability beyond genome-centric approaches. We implemented a byte-pair encoding (BPE) tokenization strategy for genome sequence generation, alongside architectural optimizations, achieving up to 150× faster sequence generation while maintaining high biological fidelity. GenomeOcean excels in representing microbial species and generating protein-coding genes constrained by evolutionary principles. Additionally, its fine-tuned model demonstrates the ability to discover novel biosynthetic gene clusters (BGCs) in natural genomes and perform zero-shot synthesis of biochemically plausible, complete BGCs. GenomeOcean sets a new benchmark for metagenomic research, natural product discovery, and synthetic biology, offering a robust foundation for advancing these fields.

Keywords: Genome Foundation Model, Metagenomics

1 Introduction

The success of language models (Brown et al., 2020; Dubey et al., 2024; Kenton and Toutanova, 2019), where representations learned from massive text corpora have enabled breakthroughs in tasks ranging from translation (Sutskever, 2014) to conversation (Ouyang et al., 2022), has inspired a recent surge in the development of foundation models for biology. These biological foundation models have demonstrated significant potential in capturing relevant patterns across diverse modalities: they can predict genomic elements (Dalla-Torre et al., 2023; Ji et al., 2021; Liu et al., 2025; Nguyen et al., 2023, 2024; Sanabria et al., 2024; Zhou et al., 2023; Zvyagin et al., 2022), infer RNA structures, and optimize codons (Chen et al., 2022; Li et al., 2023), identify cell type-specific epigenetic signals (Gao et al., 2023), annotate single-cell transcriptomic profiles (Cui et al., 2024a,b), and even model or design novel protein structures (Hayes et al., 2024; Jumper et al., 2021; Lin et al., 2023a; Nijkamp et al., 2023; Watson et al., 2023). Given the inherent interconnectedness of these modalities, efforts are underway to develop foundation models that integrate multiple data types to enhance their capabilities. Examples include aligning protein sequences, 3D structures, and literature text to reduce false positives (Zhang et al., 2024), and leveraging transfer learning across DNA, RNA, and protein to improve predictions of transcript isoform expression (Garau-Luis et al., 2024).

Since DNA is the central repository of genetic information in most living organisms, and it encodes the instructions to produce RNA and proteins, foundation models trained with only DNA sequences have been shown to be capable of modeling other modalities. Embedding from genome foundation models (gFMs) can discriminate various types of RNA genes they encode (Dalla-Torre et al., 2023; Ji et al., 2021; Nguyen et al., 2023), and perform protein-level tasks such as predicting essential genes in bacteria (Nguyen et al., 2024). Compared with protein language models, gFMs capture a wider range of biological functions: not only protein coding regions including the "dark proteome" or non-canonical open reading frames (Deutsch et al., 2024) but also noncoding regulatory elements, repeats, mobile

[†]**Corresponding to** Zhong Wang (zhongwang@lbl.gov)

elements, and other genomic "dark matter" that influences gene expression and phenotypes (Avsec et al., 2021; Linder et al., 2023). In addition, most protein foundation models are limited to one protein at a time, and have difficulty modeling higher-order functions from coordinated gene networks or pathways. Moreover, standard protein models implicitly assume canonical codon usage and ignore widespread outliers such as codon expansion (Shandell et al., 2021), condensation (Bloomfield, 1996) , or stop-codon suppression (Ivanova et al., 2014).

Unlike large natural language models (Brown et al., 2020; Dubey et al., 2024; Kenton and Toutanova, 2019) that use subwords as tokens, the optimal tokenization strategy for gFMs remains an open question. GenSLMs employ genetic codons as tokens (Zvyagin et al., 2022), a strategy that may be particularly effective for understanding virus and prokaryotic genomes, given that the majority of microbial genomes consist of protein-coding regions. Alternatively, fixed-size k-mers have been used successfully to represent genomic sequences (Dalla-Torre et al., 2023; Ji et al., 2021), despite their lack of direct biological meaning, as they can effectively represent noncoding DNA elements. Using single nucleotides as tokens provides the most granular representation of DNA sequences (Nguyen et al., 2023, 2024; Schiff et al., 2024), offering maximum flexibility in sequence modeling. However, using single nucleotide tokens leads to increased computational costs, making the models less efficient and unsuitable for large inference projects. Another frequently used tokenization is Byte-Pair Encoding (BPE), which leverages high-frequency DNA segments found in genomes to reduce the number of tokens required to efficiently represent a DNA sequence (Sennrich et al., 2016). Despite its efficiency, BPE has primarily been used in BERT-style architectures for sequence classification tasks (Sanabria et al., 2024; Zhou et al., 2023, 2024). It is unknown whether it is suitable for GPT-style generation, as it may break biologically meaningful units (e.g., codons or regulatory motifs) into suboptimal fragments, potentially disrupting the model's ability to learn or generate functionally coherent sequences.

Another open problem is the selection of training sets for gFMs, a decision that significantly impacts model performance and generalizability. Predicted genes focus on coding regions, offering insights into functional aspects, but they may introduce prediction errors and neglect non-coding elements. Reference genomes provide high-quality, curated data, yet they are often biased toward well-studied species, limiting their ability to generalize across diverse organisms, especially uncultured, low-abundance organisms known as the rare biosphere (Pascoal et al., 2021; Sogin et al., 2006). Evo leveraged the Genome Taxonomy Database (GTDB) datasets, a taxonomically diverse and manually curated collection of genomes (Parks et al., 2022), to enhance the diversity of its training data (Nguyen et al., 2024). However, this approach may still overlook functional diversity and other novel genomic features, especially those can only effectively reached by metagenome analysis (Pavlopoulos et al., 2023). Environmental metagenome sequencing, on the other hand, provide a comprehensive and unbiased representation of genomic diversity of habitats, encompassing both coding and non-coding regions from a wide range of species, making them a promising candidate for training gFMs. However, half of the phylogenetic diversity is estimated to be missed from the metagenome-assembled genomes (MAGs), especially those species from environmental samples (Wu et al., 2025). In addition, to assemble large metagenome datasets face many challenges such as assembly quality, potential noise, and the substantial computational resources required to produce them, limiting their adoption as training sets in gFMs.

Here, we introduce GenomeOcean, a BPE-based generative genome foundation model (gFM) trained directly on metagenome assemblies. Together with several architectural and computational optimizations (Ainslie et al., 2023; Dao, 2023; Kwon et al., 2023) on the Transformer Decoder (Vaswani et al., 2017), GenomeOcean generates sequences two orders of magnitude faster than existing gFMs (Nguyen et al., 2024; Zvyagin et al., 2022) of comparable size. To construct a high-quality pre-training dataset that represents extensive microbial diversity, we compiled a set of published large-scale metagenomics assemblies (Har, 2018; Oliver et al., 2024; Riley et al., 2023), and conducted three more terabase-scale coassemblies (Peterson et al., 2009; Sunagawa et al., 2020; Wang et al., 2019), yielding a training set integrating data from multiple environmental metagenome studies spanning oceans, lakes, mammalian gut, forests, and soils. This effort produced approximately 645.4 Gbp of high-quality contigs, derived from 219 TB of raw environmental data, ensuring a broad and representative training set for GenomeOcean.

Surprisingly, we found that BPE-based, decoder-only models are capable of representing biological functions, as GenomeOcean excels at both DNA- and protein-level tasks. It generates DNA embeddings that effectively cluster and segregate different microbial species, achieving performance comparable to gold-standard tetra-nucleotide frequency (TNF)-based approaches. GenomeOcean also demonstrates a profound understanding of protein functions, responding appropriately to synonymous and non-synonymous mutations in prompts while generating sequences that encode full-length proteins aligned with known protein structures, despite the prevalence of fragmented genes in the training data. Furthermore, when fine-tuned on a set of multi-genic biosynthetic gene cluster (BGC) sequences, the model can not only discover novel BGCs in natural genomes, but also generate the complete set of genes required for the synthesis of specific natural products, showcasing its ability to comprehend higher-order genomic functional modules.

2 Results

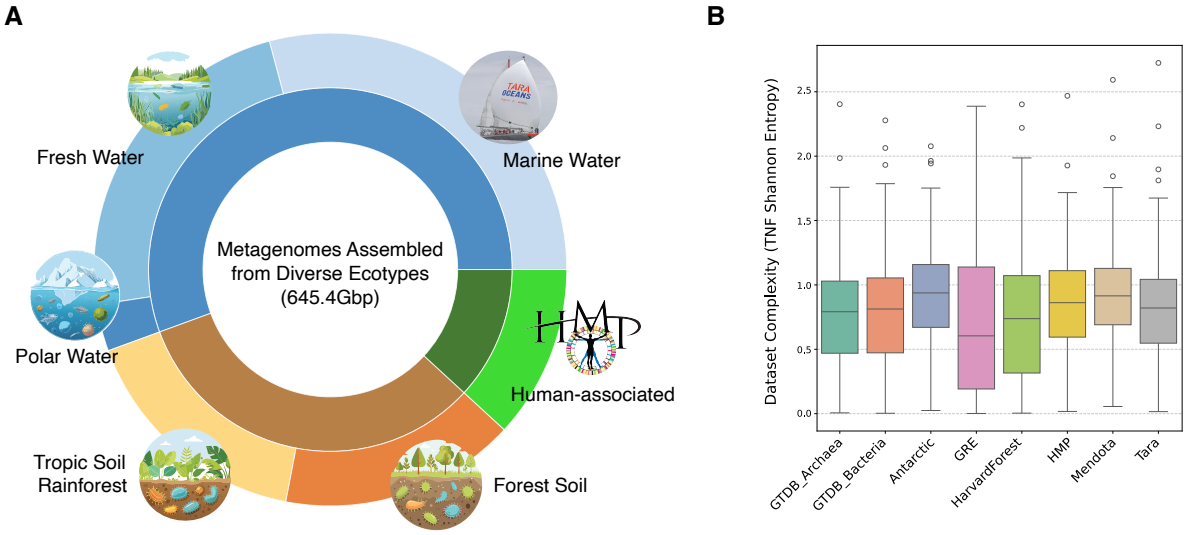

Figure 1: Compiling a diverse metagenomics dataset for training a microbial genome foundation model. (**A**) Overview of the environmental sources used for metagenome assemblies. These include time-series freshwater samples spanning decades, spatial-series marine samples across global oceans, thousands of human-associated microbiome samples, tropical and forest soils, and Antarctic subglacial lake communities. (**B**) Shannon entropy derived from tetra-nucleotide frequencies as a proxy for community diversity. The diversity for assembled environmental datasets are compared against the GTDB bacterial and archaeal genome collections. Except for the tropical soil sample (GRE), environmental datasets generally show similar or higher diversity than GTDB.

2.1 Assembly of a Diverse Earth Microbiome Representation

To construct a training dataset that represents a broad diversity of microbial life, we used the assemblies of six large, environmental metagenome datasets spanning aquatic, terrestrial, and host-associated communities (**Figure** 1**A**, **Table** 1). As demonstrated previously, combining a large number of samples from a single metagenome project into one coassembly increases the representation of relatively rare microbial species while achieving lower error rates compared to single-sample assemblies (Hofmeyr et al., 2020). Our selection included freshwater microbial communities from Lake Mendota (Oliver et al., 2024), marine environments collected globally by the Tara Oceans expedition (Bork et al., 2015; Sunagawa et al., 2020), human-associated microbial communities from the Human Microbiome Project (HMP) (Peterson et al., 2009), a tropical soil enrichment from the Great Redox Experiment (GRE) (Riley et al., 2023), a forest soil community from Harvard Forest (Alteio et al., 2020), and a polar microbial community from Antarctic subglacial lakes (Wang et al., 2019). In total, these six assemblies comprise approximately 645 Gbp of high-quality contigs derived from over 219 TB of raw metagenomic data.

The assembly statistics in Table 1 reveal the large scale of these datasets with input FASTQ sizes ranging from 3.4 TB to 98.0 TB, and resulting assembly sizes from 14.7 Gbp to 323.1 Gbp. Our training set totals 645.4 Gbp. L50 values ranging from 1.76 million to 104.1 million, and N50 values from 0.71 Kbp to 1.65 Kbp. Both metrics indicate that a large number of assembled contigs are smaller than an average microbial gene (∼3kb), reflecting the fragmented nature of metagenome assembly of large, complex communities (Meyer et al., 2022) . Even thought it is theoretically possible, training a model to learn full-length gene information from non-overlapping gene fragments has not been previously demonstrated.

Within each assembly experiment, 87%–95% of reads could be mapped back to their respective assemblies, consistent with the observation that these large metagenome assemblies captured the majority of the underlying community diversity in each ecosystem (**Table** 1).

Quantifying species-level diversity in such complex and largely uncharacterized communities is inherently challenging. To address this, we used Shannon entropy of tetra-nucleotide frequency (TNF) distributions as a proxy for alpha diversity (**Methods**). When compared to GTDB, a reference compendium of over half a million phylogenetically diverse bacterial and archaeal genomes, our environmental assemblies exhibited similar or greater TNF-based diversity. The sole exception was the tropical soil data (GRE), which showed lower diversity (**Figure** 1B), likely reflecting the source communities enriched by stable isotope probing (SIP) experiments (Riley et al., 2023).

Table 1: Metagenome datasets for GenomeOcean training. L50 is defined as count of smallest number of contigs whose length sum makes up half of the total assembly size. N50 is a median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value.

Dataset*	# raw reads (billions)	Input size (TB)	Assembly L50 (millions)	Assembly N50 Kbp	% reads mapped	# Contigs >500bp (millions)	Assembly size (Gbp)
Lake Mendota	72.1	24.3	20.87	1.17	94	96	105.5
Tara Oceans	352.1	71.6	104.1	0.85	93	363	323.1
HMP	453.1	98.0	22.42	0.71	87	69	54.0
GRE	22.7	8.0	9.1	1.65	88	56	75.1
Harvard Forest	8.9	3.4	16.8	1.05	95	71	73.0
Antarctica	29.2	9.7	1.76	1.59	94	11	14.7

*Tara Oceans, HMP, and Antarctica datasets were assembled in this study. The other assemblies were previously published.

2.2 Pre-training

We combined the aforementioned metagenome assemblies for pre-training. Contigs shorter than 1024 tokens (approximately 5 kb) were concatenated with a delimiter special token before being fed into the model. We employed a Transformer Decoder architecture and a Byte-Pair Encoding (BPE) tokenizer to efficiently represent DNA sequences (**Figure** 2**A**). Most existing gFMs show promising results with only a few hundred millions parameters, but the datasets they were trained on have limited diversity (Dalla-Torre et al., 2023; Zhou et al., 2023). To understand the effects of scaling in the generative genome foundation model on our dataset, we experimented with three model sizes of increasing parameter counts (100M, 500M, and 4B) on the same dataset. All models were initially trained with a maximum sequence length of 1024 tokens, a global batch size of 4×10^6 tokens, and a peak learning rate of 4×10^{-4}. The learning rate was linearly increased from 0 to 4×10^{-4} over the first 2,000 steps and then gradually decayed to 4×10^{-5} by the end of training. The models were first trained for $50,000$ steps. We observe model performance improved consistently with increasing parameter counts (**Figure** 2**B**). The 4B model achieved significantly lower training losses compared to the 100M and 500M variants, reflecting its superior capacity to capture genomic complexity. Due to budget constraints we did not experiment with larger models in this study. Consequently, we selected the 4B model for further analysis.

We continually trained the 4B model with the same settings for $5,000$ steps to further reduce the training loss. To extend the model's capacity for long-context modeling essential for capturing high-order genomic architectures such as operons and biosynthetic gene clusters, we used the above dataset and continued training the 4B model with a maximum sequence length of $10,240$ tokens (equivalent to ∼51 kb) for an additional 1,600 steps. This training phase maintained a total batch size of 4×10^6 tokens and used a learning rate of 1×10^{-4}.

2.3 Inference Efficiency Benchmarking

We implemented several optimization techniques to enhance the computational efficiency of GenomeOcean. Byte-Pair Encoding (BPE) tokenization converts DNA sequences into variable-length tokens, reducing sequence length by an average factor of five and significantly improving computational efficiency without compromising model performance. Architecturally, GenomeOcean integrates FlashAttention-2 to minimize memory and computation bottlenecks, Group-Query Attention to alleviate memory bandwidth constraints, and vLLM for dynamic batching and efficient memory management (**Methods**). To evaluate its performance, we benchmarked GenomeOcean against Evo-7B (Nguyen et al., 2024) and GenSLMs-2.5B (Zvyagin et al., 2022), two leading generative gFMs of similar sizes, focusing on two common tasks: sequence embedding and sequence generation.

For sequence embedding, we assessed both memory usage and speed across a range of input lengths. We compare the peak GPU memory required by each model to encode a single sequence (**Figure** 3**A**) and embedding throughput by measuring the time needed to process 128 sequences of variable length (**Figure** 3**B**). To ensure fair comparisons, we dynamically adjusted the batch size for each model with different sequence lengths to fully utilize GPU memory without exceeding capacity. For sequence generation, we evaluated throughput as the number of base pairs produced per second (**Figure** 3**C**). Each model generated 1 kb sequences from 1 kb prompts, and we recorded the sustained generation speed. All the benchmarking was done on a single NVIDIA A100 80GB GPU.

Across all benchmarks, GenomeOcean consistently achieved higher efficiency than models with similar or even smaller sizes. For embedding, GenomeOcean required less than half the memory Evo used to embed a 0.5 kb sequence while processing 32 kb sequences. Evo encountered out-of-memory errors at 4 kb, whereas GenomeOcean handled 32 kb sequences within just one-sixth of the same GPU's memory. Likewise, GenomeOcean embedded a 32 kb sequence in less time than Evo took to process a 2 kb sequence or GenSLMs needed for 4 kb. For sequence generation, GenomeOcean

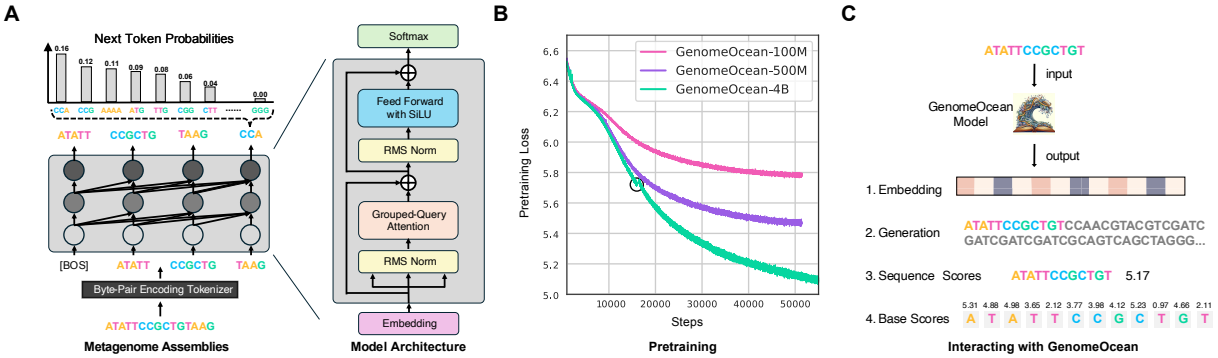

Figure 2: Overview of GenomeOcean pre-training. (**A**) GenomeOcean uses a Transformer-based decoder architecture and a BPE tokenizer to efficiently represent DNA sequences, converting them into variable-length token units. This approach balances representational flexibility with computational efficiency. The model is trained to predict the next token given a sequence context. (**B**) Training loss curves for three GenomeOcean variants (100M, 500M, and 4B parameters). As model size increases, the training loss decreases, demonstrating improved capacity to capture genomic complexity. The transient jump in the 4B model's loss is due to reinitializing the optimizer when adjusting the training configuration. (**C**) GenomeOcean provides multiple capabilities given an input DNA sequence: it can generate new sequences, produce embeddings, and score individual bases. These base functionalities can be used to support a wide range of downstream genomic analyses.

was 87-150× faster, producing over 12 kb per second. These efficiency gains allow GenomeOcean to scale to longer sequences and higher throughput, making it more accessible for large-scale genomic analysis projects while lowering computational costs.

2.4 GenomeOcean Embeddings Represent Microbial Species Sequence Composition

Microbial taxonomy classification is still an outstanding challenge, with various approaches, from 16S rRNA gene sequencing to whole-genome strategies, having been proposed for classifying different species (Achtman and Wagner, 2008). In metagenomics, tetra-nucleotide frequencies (TNF) are frequently used as coarse-grained genomic signatures to differentiate species within complex communities (Pride et al., 2003). Together with species abundance variation, TNF has been effectively used in metagenome binning (Kang et al., 2015, 2019; Meyer et al., 2022; Nissen et al., 2021), a process to form metagenome assembled genomes (MAGs). In low-sample datasets where abundance information is less useful, binning accuracy deteriorates. We hypothesize that embedding sequences into a high-dimensional feature space using a foundation model like GenomeOcean may capture subtle compositional signals, offering a higher accuracy on species discrimination.

We evaluated GenomeOcean's ability to differentiate species on a synthetic metagenomic dataset designed for species classification where 10 known species were strategically selected (ZymoBIOMICS Microbial Community Standards, **Methods**). Figure 4 compares two-dimensional UMAP projections of sequence embeddings derived from GenomeOcean, GenSLM, Evo, and TNF-based vectors. Clusters formed by GenomeOcean embeddings are more faithful to underlying species identities, matching the well-established TNF approach. Species that are closely related, such as *E. coli* and *Salmonella*, cluster together, and other related taxa (e.g., *Enterococcus faecalis*, *Listeria monocytogenes*, and *Staphylococcus aureus*) form nearby but distinct groupings. To quantify this observation, we applied HDBSCAN clustering to the UMAP embeddings and computed the Adjusted Rand Index (ARI). GenomeOcean achieved an ARI of 0.92, slightly outperforming TNF (0.81) and substantially surpassing GenSLM (0.064) and Evo (0.521).

Notably, Evo and GenSLMs, despite being trained on large and diverse sets of microbial sequences, did not match the species-level resolution obtained by GenomeOcean or even TNF in this dataset. The reasons for this difference are not entirely clear and may reflect distinct model architectures, tokenization schemes, or likely, the greater diversity of the training datasets.

2.5 GenomeOcean Learns Protein-coding Principles from DNA Alone

Unlike large eukaryote genomes, the majority of a microbe's genome sequence encodes functional proteins. Training a model solely on nucleotide sequences does not explicitly teach it about protein structure or function. Compared

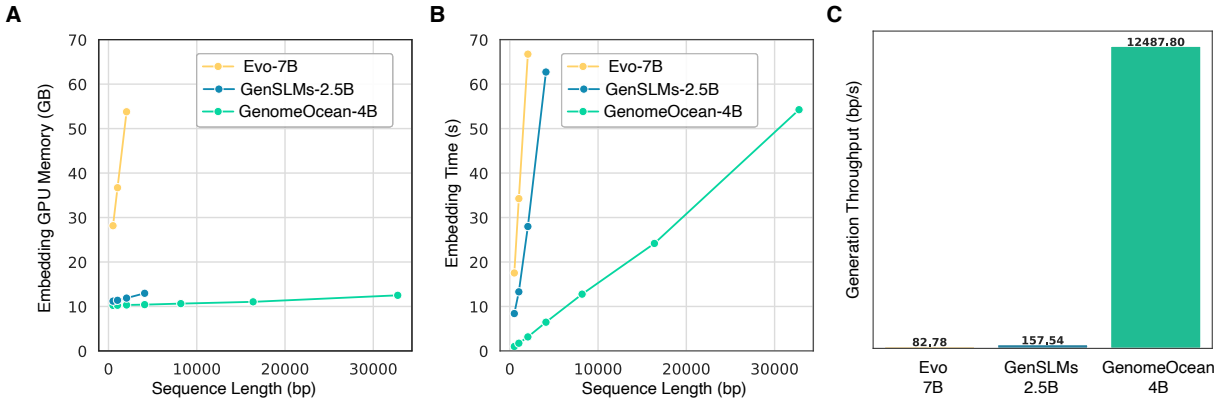

Figure 3: GenomeOcean delivers efficient inference performance. (**A**) Memory usage for embedding. Evo runs out of memory on an 80GB GPU at 4 kb input lengths, while GenomeOcean requires less than 13 GB for 32 kb inputs. (**B**) Embedding speed. GenomeOcean processes 32 kb sequences faster than Evo and GenSLMs can handle much shorter sequences, illustrating its high embedding throughput. (**C**) Sequence generation speed. GenomeOcean generates new DNA at a rate 150× faster than Evo and 87× faster than GenSLMs, producing over 12 kb per second.

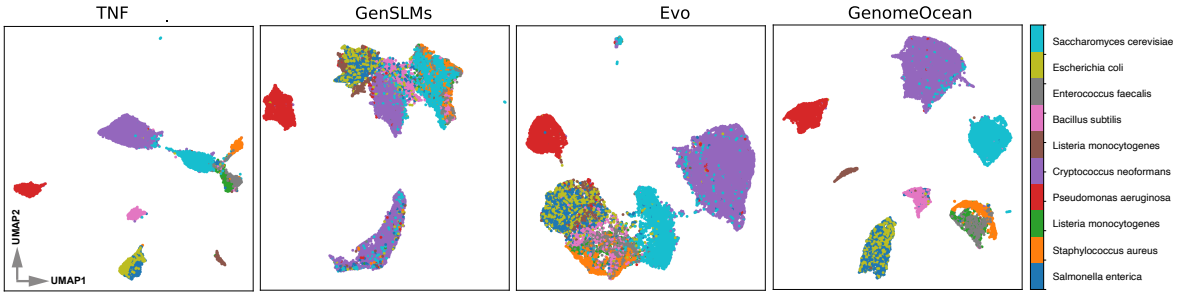

Figure 4: GenomeOcean embeddings delineate microbial species. Left to right: UMAP projections of embeddings from TNF, GenSLMs, Evo, and GenomeOcea on the Zymo Mock synthetic metagenome dataset. Data points represent DNA fragments, colored by their source organism. GenomeOcean embeddings faithfully reflect the phylogenetic relationship of known species: they place closely-related species together (E.coli and Salmonella), while separating others. Related species are also close to each other (*Enterococcus faecalis*, *Listeria monocytogenes*, and *Staphylococcus aureus*), and mixed ups only happen between the two Eukaryotic species (*Saccharomyces cerevisiae* to *Cryptococcus neoformans*).

with using codons or single nucleotides that are "natural", using BPE tokens may pose an additional challenge for the model when generating genes that code functional proteins.

To assess GenomeOcean's capability to generate genomic sequences that code for diverse and biologically meaningful proteins, we first generated a synthetic dataset using GenomeOcean and functionally annotated the predicted proteins using standard metagenome annotation pipelines (**Methods**). We found that these synthetic sequences encode proteins spanning major functional categories (**Figure** 5**A**). These functional predictions reflect a broad repertoire, ranging from metabolic enzymes to regulatory proteins, suggesting that the model learns generalizable rules for translating nucleotide patterns into protein-coding genes.

GenomeOcean can also "autocomplete" full-length proteins when given partial prompts. For example, starting with the first 600 bp of the *Staphylococcus aureus* GMP synthetase (*guaA*) gene, GenomeOcean extended the sequence to produce a full-length coding sequence (*guaA-gen*). We predicted the three-dimensional structures of both the original and generated proteins using Chai-1 (Discovery et al., 2024), a state-of-the-art structure prediction method. The resulting structures share strikingly similar folds, indicating that GenomeOcean-generated sequences preserve critical residues that dictate protein folding (**Figure** 5**B**).

To further demonstrate the biological plausibility of the above generated *guaA* variants, we compared multiple generations from the same prompt to a diverse set of natural homologs belonging to the PRK00074 protein family. The sequence logos of the generated set align closely with those from natural homologs, reproducing many conserved

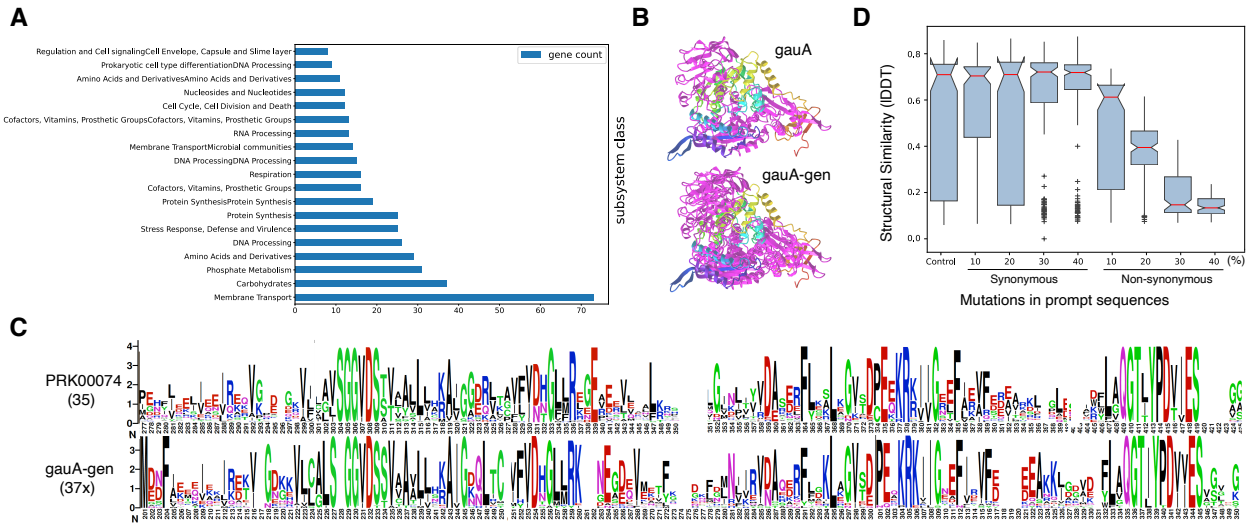

Figure 5: GenomeOcean recapitulates protein functionality from DNA. (**A**) Functional distribution of predicted proteins from a synthetic metagenome generated by GenomeOcean, illustrating diverse protein categories identified. (**B**) Predicted 3D structures of a *Staphylococcus aureus guaA* GMP synthetase and a GenomeOcean-generated variant (*guaA-gen*) based on a 600 bp *guaA* prompt. Both were modeled with Chai-1, showing similar overall folds. The prompt was included for modeling. (**C**) Sequence logos comparing 35 most diverse natural members of the PRK00074 protein family (*PRK00074 (35)*) and 37 GenomeOcean-generated variants (*gauA-gen (37x)*) prompted from the same first 600 bp of *guaA*. The height of each residue represents its conservation. The prompts was excluded as they are identical. (**D**) Prompt mutation analysis using a sponge-derived TRAP protein. Synonymous and non-synonymous codon changes were introduced into a 500 bp prompt, and resulting structures from generated sequences were evaluated using lDDT (local Distance Difference Test, a method for evaluating the similarity of protein structure models compared to a reference structure) scores.

residues and patterns (**Figure** 5**C**). This alignment suggests that GenomeOcean captures the fine-grained residue-level constraints imposed by evolution or structure.

It is possible that GenomeOcean simply emits tokens at a frequency mimicking those the training data, without any true "understanding" of the coding potential of the genomic sequence. If this is true, we would expect GenomeOcean to be sensitive to prompt perturbations. We chose a different protein, a sponge-derived TRAP protein as a reference to test this hypothesis. We created mutated prompts (first 500 bp of the ORF) by introducing synonymous (codon changes that do not alter the amino acid sequence) or nonsynonymous (codon changes that alter amino acids) substitutions, and measured the generated proteins' structural similarity to the wild-type reference using lDDT scores (**Figure** 5**D**, see **Methods** for details). Even with 40% codons randomized synonymously in the prompts, GenomeOcean still can produce coherent protein variants. In contrast, with only 10% nonsynonymous mutations the structural similarity rapidly decreased. These results further support GenomeOcean's understanding of coding constraints.

Together, these findings show that GenomeOcean, trained exclusively on raw DNA with an unnatural vocabulary, internalizes an implicit grammar of protein coding. It captures fundamental relationships between nucleotide sequences and protein structure/function, enabling zero-shot generation of diverse, biologically plausible protein-coding genes.

2.6 Model Safety

With the capability of generating realistic sequences, it is essential to ensure artificial sequences do not contaminate public databases containing natural sequences. Therefore, we evaluated whether existing genome foundation models could distinguish artificial sequences generated by GenomeOcean from natural sequences in public databases. We formulated this as a binary classification problem, where models predict whether an input sequence is natural or artificial. We construct a balanced dataset with 18,000/2,000/20,000 2kbp sequences for training/validation/test (**Methods**). We evaluated three genome foundation models on this task: two discriminative models, DNABERT-2 (Zhou et al., 2023) and Nucleotide Transformer V2 (Dalla-Torre et al., 2023), and GenomeOcean itself. We performed standard fine-tuning for DNABERT-2 and Nucleotide Transformer while adapting Low-Rank Adaptation (LoRA) for

efficient fine-tuning on GenomeOcean.

Table 2 presents the models' accuracy in distinguishing GenomeOcean-generated artificial sequences from natural ones. DNABERT-2 achieved marginally better performance than Nucleotide Transformer, though both models misclassified approximately 15% of sequences, indicating their limited ability to consistently differentiate between natural and artificial sequences. In contrast, GenomeOcean demonstrated superior discrimination capability with an F1 score exceeding 99%. This performance may be attributed to GenomeOcean's ability to recognize subtle patterns in its generated sequences that other models cannot detect. The high classification accuracy provides additional confidence in the model's deployment and public release. Together with the pre-trained checkpoint, we also release the GenomeOcean-based artificial sequence detector for safe usage of GenomeOcean.

Table 2: Distinguish Natural and Artificial Sequences.

Model	Precision	Recall	F1
DNABERT-2	85.23	85.02	85.12
Nucleotide Transformers V2	83.31	82.97	83.14
GenomeOcean	99.03	99.03	99.03

2.7 GenomeOcean's Long-context Enables Modeling Higher-order Genomic Functional Modules Such as Biosynthetic Gene Clusters

Prokaryotic genome sizes vary greatly, from ultra-small archaeal genomes of only a few tens of kilobases to enormous bacterial genomes exceeding 10 megabases. Although these genomes generally lack global organizational patterns, one prominent feature is the presence of operons and biosynthetic gene clusters (BGCs), where genes involved in a single metabolic pathway are located in close proximity (Medema et al., 2015). Genome mining of BGCs has paved the way for discovering useful bioactive molecules that work as antibiotics, anticancer drugs, and organic pesticides, among other uses (Blin et al., 2017; Dinglasan et al., 2025). Significant efforts have been made on cultivating microbes (Steele et al., 2019) for new drug discovery, and through mining the large number of publicly available genomes (Gavriilidou et al., 2022), it has been estimated that only less than 3% of BGCs can be associated with known pathways.

BGCs can span tens of kilobases and contain numerous enzymatic genes as well as regulatory elements. We tested GenomeOcean's ability to learn high-order genomic functional modules by further training it using twelve million BGC sequences from the SMC database (Udwary et al., 2024), resulting in a specialized model named bgcFM (**Methods**).

Without prompting, bgcFM can generate realistic, long BGC sequences (Supplementary Data X). For example, when generating type I polyketide synthase (T1PKS) clusters, bgcFM arranges core genes with the appropriate series of catalytic domains in a linear, modular assembly-line architecture (Figure 6**A**). This organization, essential for biosynthetic functionality, is challenging to recreate without explicitly modeling extensive contextual dependencies. None of the generated sequences have significant nucleotide matches to the sequences in the training set, suggesting that bgcFM "invented" new BGC variants—distinct from those in the training data not by memorizing the training data, but by learning the underlying principles of BGC organization and combinatorial domain arrangements.

State-of-art BGC prediction tools heavily rely on hand-crafted rules, which could be biased and incomplete Hannigan et al. (2019). bgcFM offers an opportunity to include other patterns that are associated with known rules, thereby enable the discovery of new BGCs missed by the current annotation pipelines. To test this possibility, we implemented a contrastive approach by subtracting the likelihood score of the foundation model from bgcFM, to suppress noise and irrelevant information the model learned from BGCs (Methods). By scanning a microbial genome using this approach, one can identify candidate regions containing BGCs.

As shown in Figure 6**B**, many of the candidate regions overlap with existing BGC annotations with very few false positives. Interestingly, some of the candidate regions overlap BGCs not included in the training data, as these BGCs were recently predicted by the latest version of AntiSMASH (Blin et al., 2023), supporting the generalization ability of bgcFM to unseen BGCs. Interestingly, some of the bgcFM candidate regions are not annotated at all, but bear several features of BGCs. Figure 6**B** shows one example of these. This region contains a transcription factor binding site related to antibiotic production, AfsR, followed by a short protein reminiscence of a ribosomal peptide natural product (RiPP) precursor. It also contains several putative accessory genes including a P450, an acyl-AMP ligase, and a trans-peptidase. As many BGCs, the region also comprise several transporters and regulators, including a multidrug-efflux transporter gene, three different ABC transporter genes, and a TetR family regulatory gene, which is strongly associated with small-molecule productions.

3 Discussion

This study presents GenomeOcean, a genome foundation model capable of zero-shot generation of biologically meaningful sequences at multiple scales. From producing protein-coding genes that respect evolutionary constraints to

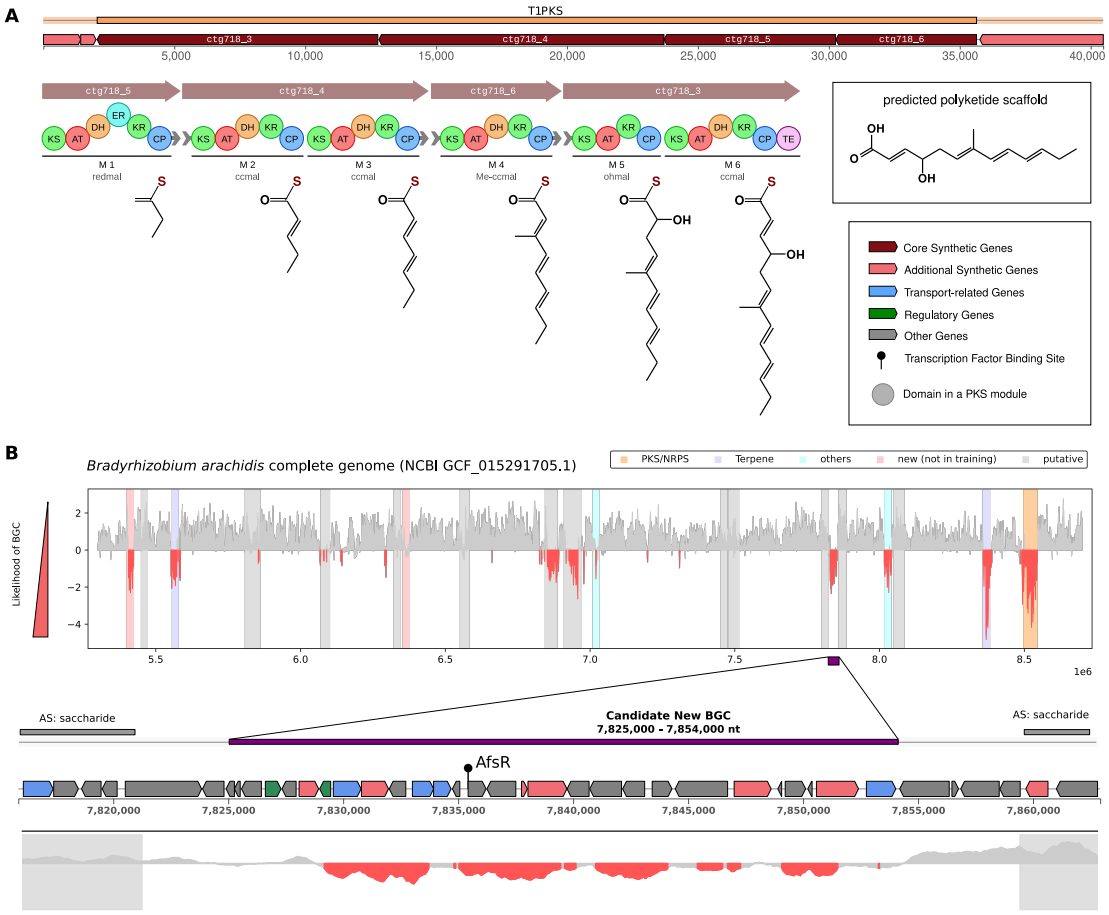

Figure 6: bgcFM captures higher-order biosynthetic gene architectures.
(A) 40-kb BGC generated *de novo* by bgcFM showcasing the correct domain arrangements and a full-length assembly-line architecture characteristic of the Type-I Polyketide Synthase pathways (KS: Ketosynthase, AT: Acetyltransferase, DH: Dehydratase, ER: Enoylreductase, KR: Ketoreductase, CP: carrier protein, TE: Thioesterase termination domain). The domain organization and functional completeness support the model's capacity to infer long-range genomic relationships. **(B)** Calculating bgcFM's loss scores over the entire genome to identify candidate BGC regions. The 3.5Mb segment shows that the dip in loss scores correlates with antiSMASH-predicted BGCs and, in some cases, also covers new classes not originally present in the training data, and even potential new BGCs. Shown in this example is a 29-kb region lies between two antiSMASH-predicted 'saccharide' clusters. Some interesting features of this region include a transcription factor binding site related to antibiotic production, AfsR, followed directly downstream by a 324-bp short protein reminiscence of a ribosomal peptide natural product (RiPP) precursor. A wide array of putative accessory genes also populate the region, including a P450, an acyl-AMP ligase, and a trans-peptidase. The region also comprise a multidrug-efflux transporter gene, three different ABC transporter genes, and a TetR family regulatory gene, which was widely associated with small-molecule productions.

synthesizing large genomic modules such as biosynthetic gene clusters, GenomeOcean demonstrates the capacity to learn and represent complex genomic elements. Compared to previous genome language models, GenomeOcean achieves superior computing efficiency both in memory usage and inference speed. The reduction in input length afforded by byte-pair encoding and the incorporation of computationally efficient architectural components enabled the model to handle input sequences of at least 50 kb. Although we have not fully explored its extrapolation capabilities, methods such as position interpolation can potentially extend the context length by an order of magnitude. With the current design, GenomeOcean already meets the requirements for a broad range of genomic analysis tasks.

By coassembling terabases of metagenome data, our metagenomics assemblies dataset demonstrates better species diversity than well-curated reference compendium such as Genome Taxonomy Database (GTDB). Learning from an un-biased and diverse dataset, GenomeOcean produces embeddings that are faithfully to the underlying species distribution while produces sequences with a diverse set of functions. The training strategy developed in this work can be further expanded to include larger datasets, for insurance, all assembled metagenomes that are publicly available. As of the time of this work, there are 25 Tb assembled metagenomes in the IMG database, and there are another 25 Tb in NCBI. Many of these datasets, however, may have poor qualities, especially those generated in the early days of metagenomics. Significant effort will be required to filter out poor data and reduce redundancy. Additionally, a substantial compute budget is needed for a training set of this size.

It is possible to greatly improve the generation quality of GenomeOcean by test time fine-tuning considering the model's efficiency. For example, one can filter model outputs by integrating known constraints to improve generation quality with "Best-of-N" sampling. Generating multiple sequences from the same prompt and selecting the best one with a proxy objective (e.g., tetra-nucleotide frequency consistency) can lead to better outputs. However, for general genomic analysis, the lack of standardized, efficient evaluation and filtering strategies complicates this fine-tuning process. Generating new variants of a known protein family, for example, involves intensive computing to predict protein structures. We only scratched the surface of the massive information contained in the GenomeOcean models. The training set does contain numerous Eukaryotic and viral species, we are hoping the research community can take advantage of the models and tools we developed to enable a deep dive in those areas.

We also demonstrated that GenomeOcean's long context capacity can be used to reconstruct and predict complex metabolic genome architectures. The fine-tuned model with biosynthetic gene clusters paves the way for more efficient discovery and engineering of natural product pathways, potentially accelerating the identification of novel bioactive compounds. It should be possible to similarly fine-tune GenomeOcean to model other high-order genome functional modules such as operons. The length of its context, however, does not support modeling microbial genomes as a whole, as most of them are in the range of megabases. A different model architecture may be needed for whole genome modeling, as the linear sequence of the genes on the genomes is often not essential.

Despite the progress demonstrated here, several limitations remain. The completeness and uniformity of the training data present challenges. For example, GenomeOcean assigns high loss to certain training sets such as Tara Oceans, despite its large representation in the data. This discrepancy suggests that not all patterns or functional elements have been equally learned. Similarly, some abundant protein families, such as those involved in carbon fixation, are not successfully autocompleted by the model. It is unclear whether this limitation is due to insufficient model capacity, training incompleteness, or inherent biological diversity. Our model, with only four billion parameters, may not be large enough to fully capture the extraordinary complexity of environmental metagenomes, raising the question of how much larger models need to be in order to achieve more comprehensive functional modeling.

Another limitation lies in the prompting strategy. Currently, prompts must be provided as DNA sequences, limiting the user's ability to guide generation toward specific biological goals. Future work will be needed to integrate additional modalities or control signals that can direct the model toward desired functions or niches.

4 Methods

4.1 Model and Training

4.1.1 Tokenization

Tokenization is the first step of modeling DNA sequences with foundation models. It transforms a DNA sequence into a series of predefined tokens (e.g., `CG` and `AATGC`), which are then converted to numerical vectors as input for the foundation model. GenomeOcean adapts a SentencePiece (Kudo and Richardson, 2018) tokenizer with Byte-Pair Encoding (BPE) to tokenize each input DNA sequence into a set of non-overlapping tokens, as proposed in Zhou et al. (2023). The vocabulary of the tokens is determined by iteratively selecting on those frequently occurring sequence k-mers (k ranges from 1 to 12 in our case) in the pre-training corpus until the desired size (in this case 4096) is reached. The same tokenizer was used for all the models in this work.

4.1.2 Architecture and Pre-training

We optimize the Transformer Decoder (Vaswani et al., 2017) architecture with an emphasis on model efficiency. We incorporate FlashAttention-2 (Dao, 2023), which refines the computation process and GPU work partitioning to accelerate attention—the core module of Transformer models. To further improve computational and memory efficiency, we adopt Group Query Attention (GQA) (Ainslie et al., 2023). Unlike standard multi-head attention, GQA partitions query heads into groups, where each group shares a single key head and value head. In addition, we use Root Mean Square (RMS) layer normalization (Zhang and Sennrich, 2019), the Sigmoid Linear Unit (SiLU) activation function (Elfwing et al., 2018) for enhanced representational capacity, and Rotary Positional Embedding (RoPE) (Su et al., 2024) for more flexible positional encoding. During inference, we integrate GenomeOcean with vLLM (Kwon et al., 2023), an efficiency-focused LLM inference library that employs optimizations such as efficient memory management and dynamic sequence batching to increase generation throughput. The model hyperparameters are presented in Table 3. GenomeOcean is pre-trained on the Perlmutter supercomputer at the National Energy Research Scientific Computing Center (NERSC)[1]. We scaled the model training on 64 NVIDIA A100 GPUs across 16 compute nodes. The first stage cost 14 days, and the second stage cost 1 day. We implemented efficient multi-node training with DeepSpeed (Rajbhandari et al., 2020).

Table 3: Hyperparameters of GenomeOcean.

Model Size	100M	500M	4B
feed forward dropout	0.1	0.1	0.1
hidden size	768	1536	3072
intermediate size	3072	6144	16384
num. attention heads	8	8	12
num. query and key heads	8	8	4
num. layers	12	14	24
rms eps	1e-5	1e-5	1e-5
rope theta	1e6	1e6	1e6
learning rate	4e-4	4e-4	4e-4

4.1.3 Fine-tune GenomeOcean as Biosynthetic Gene Clusters Foundation Model (bgcFM)

We started the fine-tuning with the pre-trained GenomeOcean-4B model. We split the fine-tuning into 2 phases with different sequence lengths. The first phase maintained a total batch size of 2×10^6 tokens and a maximum sequence length of 1024, while the second phase maintained a batch size of 1.3×10^6 tokens and a maximum sequence length of 10240. We trained the model for 16,000 steps in the first phase and 1,600 steps in the second phase.

4.2 Training/Evaluation Datasets

4.2.1 Assembled Metagenome Datasets

All of the six coassemblies used for training GenomeOcean were performed using MetaHipMer (Hofmeyr et al., 2020), a distributed metagenome assembler that scales efficiently on supercomputers, enabling it to use hundreds or thousands of compute nodes to coassemble datasets larger and more complex than is possible on shared memory computers. The assemblies were performed on three national laboratory supercomputers. Table 4 provides further details about the assemblies, including resources and time used.

Table 4: Assembly details for metagenome datasets used for GenomeOcean training.

Dataset	Date Assembled	Assembly Time (mins)	Nodes Used	System Used
Lake Mendota	8/4/2022	35	1500	Frontier
Tara Oceans	5/10/2023	95	9000	Frontier
HMP	8/23/2023	47	9000	Frontier
GRE	11/2/2020	84	512	Summit
Harvard Forest	11/18/2021	11	1600	Cori
Antarctica	6/16/2023	22	1024	Frontier

4.2.2 Evaluation of Dataset Complexity

We used a simple metric to measure the Shannon Entropy of a dataset. For each dataset, we took a random set of 10,000 contigs, and only took the first 3,000 bases of each contig to ensure their lengths are comparable. We then

[1] https://www.nersc.gov

calculated TNF vectors from these contigs to obtain a TNF matrix. Finally, we computed the Shannon's entropy of each of the 136 columns of the TNF matrix by binning the 10,000 values in a row into 100 bins. These datasets were also sampled proportionally according to their size to form 10,000 contigs to represent a sample from the pre-training dataset (`meta`). For comparison with GTDB reference genomes, all the genomes from Archaea and Bacteria were combined first, and 10,000 contigs were sampled the same way as the other datasets.

4.2.3 Biosynthetic Gene Cluster (BGC) Collection

11,063,856 predicted BGCs totaling 242,122,742,143 bases were derived from JGI's Secondary Metabolite Collection Database (Udwary et al., 2024). BBMap (Bushnell, 2022) Version 39.06 was used to cluster these sequences at a minimum sequence identity of 97%, and only the longest sequence from each cluster was retained in the training set. After deduplication, 1,716,441 BGCs (15.51%) or 43,541,503,146 bases (17.98%) of the original 11 million BGCs were used for further training of the GenomeOcean model.

4.2.4 ZymoBiomics Microbial Community Standard Datasets

The ZymoBiomics Microbial Community Standard (Zymo Research Corp., Irvine, CA, United States), referred to as Zymo dataset, contains the following 10 species: *Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella eterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, and *Cryptococcus neoformans*. The original WGS reads can be found with the NCBI Accession number: SRX15657751.

4.2.5 Generated Sequence Discrimination

We constructed the dataset from two public databases: the Genome Taxonomy Database (GTDB) (Parks et al., 2022) and sequences from the Critical Assessment of Metagenome Interpretation (CAMI2) challenge (Meyer et al., 2022). To construct the dataset, we randomly sampled 10,000 2kbp sequences from each database as prompts and generated one corresponding 2kbp sequence from each prompt. This resulted in four sets of sequences: 10,000 natural sequences from CAMI2, 10,000 artificial sequences generated from CAMI2 prompts, 10,000 natural sequences from GTDB, and 10,000 artificial sequences generated from GTDB prompts. To prevent data leakage, we used CAMI2 sequences for training and validation, while reserving GTDB sequences for testing. The CAMI2 sequences were split into training and validation sets with a ratio of 90:10, resulting in a final split of 18,000/2,000/20,000 sequences for train/validation/test sets.

4.3 Model Evaluation

4.3.1 Sequence Composition

The 10 genomes in the ZymoMock dataset were divided into 3,000 non-overlapping segments, resulting in a total of 24,499 segments. Embeddings for each segment were generated using TNF, GenomeOcean, Evo, and GenSLM, creating four distinct sets of embeddings. To evaluate the effectiveness of these embeddings in capturing sequence composition information, UMAP was applied to reduce their dimensionality to two dimensions for visualization. This was followed by HDBSCAN clustering to quantitatively analyze patterns using the Adjusted Rand Index (ARI). Both UMAP and HDBSCAN were applied with default settings.

4.3.2 Generating and Evaluating Synthetic Metagenomes

We randomly generated 3,400 sequences using the following parameters (min_length=1024, max_length=10240, temperature=1.0, top_k=-1, top_p=0.95, presence_penalty=0.5, frequency_penalty=0.5, repetition_penalty=1.0). Sequences with more than 20% simple repeats (low-complexity) were filtered out, leaving 1,414 sequences for annotation.

The sequences were then annotated with Rapid Annotations using Subsystems Technology toolkit (RASTtk) v1.073 using default parameters, and the results were analyzed at the subsystem class level (Aziz et al., 2008).

4.3.3 Generating Artificial Genomic Sequences for Structural Modeling

The GMP synthetase *Staphylococcus aureus* strain MRSN967703 *gau* gene (gauA) was selected as a reference. Its first 600bp of genomic sequence was used as the prompt to generate 100 sequences. The 3D structures of the generated sequences were predicted using ESM Atlas v2023_02 (Lin et al., 2023b), and then aligned with Foldmason (Gilchrist et al., 2024) to estimate its local distance difference test (lDDT, Mariani et al. (2013)), a similarity score to the reference. One of the generated sequences that showed the highest lDDT similarity score to the reference was selected for structural modeling using Chai-1. All of the generated sequences that encoded a long ORF, totaling 37, were used

for multiple sequence alignment and protein sequence logo generation. For comparison, 35 most diverse natural gauA proteins were obtained from NCBI [2].

TRAP is a transporter solute binding protein encoded by an uncultured marine bacterium A marine protein with known structure (https://www.rcsb.org/structure/5I5P). A blast search using the protein identified a homologous gene from a sponge metagenome (NCBIid: OY729418). This gene was used as a reference and its first 500bp was used as the prompt for prompt mutation analysis.

For each experiment, 100 sequences were generated, and those encode an long ORF that is close to the reference were selected to calculate the LDDT score to the reference. For synonymous mutations, a percentage of codons were randomly swapped without changing the protein sequence. For nonsynonymous mutations, a percentage of codons were randomly mutated and the encoded protein sequences were changed.

4.3.4 Zero-shot BGC generation experiment

A total of 258,260 sequences were generated using bgcFM using a custom script (https://github.com/jgi-genomeocean/gmeval/tree/main), using parameters repetition-penalty=[1.0-1.5] and temperature=[0.7-1.1]. Each ten sequences took an average of 100 seconds to generate on a single 4 x Nvidia A40 machine. Each sequence generation step is immediately followed by gene prediction in the same python script using pyrodigal Larralde (2022). The gene sequences were then annotated for biosynthetic functionalities using Apache Spark and the Axolotl library (https://github.com/JGI-Bioinformatics/axolotl/tree/main/axolotl), and antiSMASH version 7.0 Blin et al. (2023) was run in parallel to assign BGC membership, upon which 11,123 returned a positive hit. 1,459 PKS-type BGCs are filtered down to 1,044 T1PKS to give a closer inspection of its modules. All mentioned data is made available in the Supplementary Data.

4.3.5 Loss score BGC scanning

We wrote a custom script (https://github.com/jgi-genomeocean/gmeval/tree/main) to generate a per-token loss score for any input genome. The script works by splitting the full-length genome into overlapping fragments of 50kb, then calculate the per-token perplexity scores. To get the final loss scores for BGC scanning purposes, loss scores from the pre-trained model and the BGC fine-tuned are subtracted. We applied a 1,000-bp sliding window to reduce noise and smoothened out the loss scores within the window by taking the mean of the scores. A significant dip in loss score is defined as a score lower than the 5th percentile of the entire genome.

5 Data & Code

5.1 Public Metagenome Assemblies

Three public metagenome assemblies were downloaded from JGI's IMG/M databases:

1. Lake Mendota: Freshwater microbial communities from Lake Mendota, Crystal Bog Lake, and Trout Bog Lake in Wisconsin, United States (multi-decadal time-series). IMG Submission ID: 288555, doi:10.46936/10.25585/60001198.
2. Great Redox Experiment (GRE): Lab enrichment of tropical soil microbial communities from Luquillo Experimental Forest, Puerto Rico (95 samples). IMG Submission ID: 255440, doi: 10.46936/10.25585/60000880.
3. Harvard Forest Soil: Forest soil microbial communities from Barre Woods Harvard Forest LTER site, Petersham, Massachusetts, United States (28 samples). IMG Submission ID: 264502, doi: 10.46936/fics.proj.2016.49483/60006003.

5.2 Metagenome Raw Reads Datasets Used in Assembly

The raw read datasets are listed below:

1. HMP: A subset of Human Microbiome Project metagenomes (13,836 samples), the data was described in a previous study (Peterson et al., 2009).
2. Tara Oceans: Multi-depth, world-wide sampling of Ocean environments, constituting the largest modern-day worldwide collection of plankton sampled 'end to end' around the world (1211 samples) (Bork et al., 2015; Sunagawa et al., 2020).
3. Antarctic: Water samples and derived enriched culture from Lake Fryxell and Lake Bonny in Antarctica (21 samples), described in a previous study (Wang et al., 2019).

[2] https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=234614)

5.3 Code

GenomeOcean is open source and publicly available at https://github.com/jgi-genomeocean/genomeocean.

6 Funding Statement

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC awards BER-ERCAP0026482 and BER-ERCAP0030547.

This research used the Lawrencium computational cluster resource provided by the IT Division at the Lawrence Berkeley National Laboratory (Supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231)

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory (https://ror.org/01qz5mb56), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research was also supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of two U.S. Department of Energy organizations (Office of Science and the National Nuclear Security Administration) responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering, and early testbed platforms, in support of the nation's exascale computing imperative.

HL was supported by NIH R01LM013722. There was no additional external funding received for this study.

7 Author Contributions

ZW, HL, and ZZ conceived the research idea. ZW supervised the project. RR, SH, and RE performed metagenome assemblies. WW built the tokenizer. ZZ designed and trained the models. FL developed inference infrastructure. SK performed the BGC analysis. RMK, FC performed the Antarctic experiment and generated data. ZW, ZZ, SK, WW, SGG, MY, HH, and LS performed model evaluation. ZW, ZZ, RR, SH, RE, and SK wrote the manuscript with input from all co-authors.

8 Conflict of Interests

FC was an employee of Illumina Inc. All other authors declare no conflict of interest.

9 Acknowledgment

The authors thank Alex Copeland for his feedback throughout the duration of this project.

References

National center for biotechnology information. barre woods harvard forest lter site sequencing reads. https://www.ncbi.nlm.nih.gov/sra/, 2018. PRJNA441471, PRJNA441472, PRJNA441473, PRJNA441474, PRJNA441475, PRJNA441476, PRJNA441477, PRJNA441478, PRJNA441479, PRJNA441480, PRJNA441481, PRJNA441482, PRJNA441483, PRJNA441484, PRJNA441485, PRJNA441486, PRJNA441487, PRJNA441488, PRJNA441489, PRJNA441490, PRJNA441491, PRJNA441492, PRJNA441493, PRJNA441494, PRJNA441495, PRJNA441496, PRJNA441497, PRJNA441498.

Mark Achtman and Michael Wagner. Microbial diversity and the genetic nature of microbial species. *Nature reviews microbiology*, 6(6):431–440, 2008.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

L. V. Alteio, F. Schulz, R. Seshadri, N. Varghese, W. Rodriguez-Reillo, E. Ryan, D. Goudeau, S. A. Eichorst, R. R. Malmstrom, R. M. Bowers, L. A. Katz, J. L. Blanchard, and T. Woyke. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems*, 5(2), April 2020. doi:10.1128/msystems.00768-19. URL https://doi.org/10.1128/msystems.00768-19.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, et al. The rast server: rapid annotations using subsystems technology. *BMC genomics*, 9:1–15, 2008.

Kai Blin, Thomas Wolf, Marc G Chevrette, Xiaowen Lu, Christopher J Schwalen, Satria A Kautsar, Hernando G Suarez Duran, Emmanuel LC De Los Santos, Hyun Uk Kim, Mariana Nave, et al. antismash 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic acids research*, 45(W1):W36–W41, 2017.

Kai Blin, Simon Shaw, Hannah E Augustijn, Zachary L Reitz, Friederike Biermann, Mohammad Alanjary, Artem Fetter, Barbara R Terlouw, William W Metcalf, Eric J N Helfrich, Gilles P van Wezel, Marnix H Medema, and Tilmann Weber. antismash 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research*, 51(W1):W46–W50, 05 2023. ISSN 0305-1048. doi:10.1093/nar/gkad344. URL https://doi.org/10.1093/nar/gkad344.

Victor A Bloomfield. Dna condensation. *Current opinion in structural biology*, 6(3):334–341, 1996.

P Bork, C Bowler, Colomban de Vargas, G Gorsky, E Karsenti, and P Wincker. Tara oceans studies plankton at planetary scale, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Brian Bushnell. Bbmap, Jul 2022. URL https://sourceforge.net/projects/bbmap/.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024a.

Zhanbei Cui, Tongda Xu, Jia Wang, Yu Liao, and Yan Wang. Geneformer: Learned gene compression using transformer-based context modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8035–8039. IEEE, 2024b.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01, 2023.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Eric W Deutsch, Leron W Kok, Jonathan M Mudge, Jorge Ruiz-Orera, Ivo Fierro-Monti, Zhi Sun, Jennifer G Abelin, M Mar Alba, Julie L Aspden, Ariel A Bazzini, et al. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *BioRxiv*, 2024.

Jaime Lorenzo N Dinglasan, Hiroshi Otani, Drew T Doering, Daniel Udwary, and Nigel J Mouncey. Microbial secondary metabolites: advancements to accelerate discovery towards application. *Nature Reviews Microbiology*, pages 1–17, 2025.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, pages 2024–10, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Zijing Gao, Qiao Liu, Wanwen Zeng, Rui Jiang, and Wing Hung Wong. Epigept: a pretrained transformer model for epigenomics. *bioRxiv*, 2023.

Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Masa Roller, Bernardo P de Almeida, Lorenz Hexemer, Christopher Blum, Stefan Laurent, Jan Grzegorzewski, Maren Lang, et al. Multi-modal transfer learning between biological foundation models. *arXiv preprint arXiv:2406.14150*, 2024.

Athina Gavriilidou, Satria A Kautsar, Nestor Zaburannyi, Daniel Krug, Rolf Müller, Marnix H Medema, and Nadine Ziemert. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature microbiology*, 7(5):726–735, 2022.

Cameron LM Gilchrist, Milot Mirdita, and Martin Steinegger. Multiple protein structure alignment at scale with foldmason. *bioRxiv*, pages 2024–08, 2024.

Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, Michael Wurst, Jakub Kotowski, Dan Chang, Rurun Wang, Grazia Piizzi, Gergely Temesi, Daria J Hazuda, Christopher H Woelk, and Danny A Bitton. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18):e110–e110, 08 2019. ISSN 0305-1048. doi:10.1093/nar/gkz654. URL https://doi.org/10.1093/nar/gkz654.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model, july 2024. *URL https://www. biorxiv. org/content/10.1101/2024.07*, 1:v1, 2024.

Steven Hofmeyr, Rob Egan, Evangelos Georganas, Alex C Copeland, Robert Riley, Alicia Clum, Emiley Eloe-Fadrosh, Simon Roux, Eugene Goltsman, Aydın Buluç, et al. Terabase-scale metagenome coassembly with metahipmer. *Scientific reports*, 10(1):10689, 2020.

Natalia N Ivanova, Patrick Schwientek, H James Tripp, Christian Rinke, Amrita Pati, Marcel Huntemann, Axel Visel, Tanja Woyke, Nikos C Kyrpides, and Edward M Rubin. Stop codon reassignments in the wild. *Science*, 344(6186): 909–913, 2014.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7: e7359, 2019.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

Martin Larralde. Pyrodigal: Python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72):4296, 2022. doi:10.21105/joss.04296. URL https://doi.org/10.21105/joss.04296.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, pages 2023–09, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Biorxiv*, pages 2023–08, 2023.

Ollie Liu, Sami Jaghouar, Johannes Hagemann, Shangshang Wang, Jason Wiemels, Jeff Kaufman, and Willie Neiswanger. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene De Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, et al. Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9):625–631, 2015.

Fernando Meyer, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, Mohammed Alser, Dmitry Antipov, Francesco Beghini, et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature methods*, 19(4):429–440, 2022.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Jakob Nybo Nissen, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, Henrik Bjørn Nielsen, Thomas Nordahl Petersen, Ole Winther, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nature biotechnology*, 39(5): 555–560, 2021.

Tiffany Oliver, Neha Varghese, Simon Roux, Frederik Schulz, Marcel Huntemann, Alicia Clum, Brian Foster, Bryce Foster, Robert Riley, Kurt LaButti, Robert Egan, Patrick Hajek, Supratim Mukherjee, Galina Ovchinnikova, T. B. K. Reddy, Sara Calhoun, Richard D. Hayes, Robin R. Rohwer, Zhichao Zhou, Chris Daum, Alex Copeland, I-Min A. Chen, Natalia N. Ivanova, Nikos C. Kyrpides, Nigel J. Mouncey, Tijana Glavina del Rio, Igor V. Grigoriev, Steven Hofmeyr, Leonid Oliker, Katherine Yelick, Karthik Anantharaman, Katherine D. McMahon, Tanja Woyke, and Emiley A. Eloe-Fadrosh. Coassembly and binning of a twenty-year metagenomic time-series from lake mendota. *Scientific Data*, 11(1), September 2024. ISSN 2052-4463. doi:10.1038/s41597-024-03826-8. URL http://dx.doi.org/10.1038/s41597-024-03826-8.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 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.

Francisco Pascoal, Rodrigo Costa, and Catarina Magalhães. The microbial rare biosphere: current concepts, methods and ecological principles. *FEMS Microbiol. Ecol.*, 97(1), January 2021.

Georgios A Pavlopoulos, Fotis A Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, Simon Roux, Lee Call, Natalia N Ivanova, et al. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983):594–602, 2023.

Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The nih human microbiome project. *Genome research*, 19(12): 2317–2323, 2009.

David T Pride, Richard J Meinersmann, Trudy M Wassenaar, and Martin J Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome research*, 13(2):145–158, 2003.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

Robert Riley, Robert M. Bowers, Antonio Pedro Camargo, Ashley Campbell, Rob Egan, Emiley A. Eloe-Fadrosh, Brian Foster, Steven Hofmeyr, Marcel Huntemann, Matthew Kellom, Jeffrey A. Kimbrel, Leonid Oliker, Katherine Yelick, Jennifer Pett-Ridge, Asaf Salamov, Neha J. Varghese, and Alicia Clum. Terabase-scale coassembly of a tropical soil microbiome. *Microbiology Spectrum*, 11(4), August 2023. ISSN 2165-0497. doi:10.1128/spectrum.00200-23. URL http://dx.doi.org/10.1128/spectrum.00200-23.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi:10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Mia A Shandell, Zhongping Tan, and Virginia W Cornish. Genetic code expansion: a brief history and perspective. *Biochemistry*, 60(46):3455–3469, 2021.

Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.*, 103(32):12115–12120, August 2006.

Andrew D Steele, Christiana N Teijaro, Dong Yang, and Ben Shen. Leveraging a large microbial strain collection for natural product discovery. *Journal of Biological Chemistry*, 294(45):16567–16576, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shinichi Sunagawa, Silvia G Acinas, Peer Bork, Chris Bowler, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, et al. Tara oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, 18(8):428–445, 2020.

I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

Daniel W Udwary, Drew T Doering, Bryce Foster, Tatyana Smirnova, Satria A Kautsar, and Nigel J Mouncey. The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*, page gkae1060, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhong Wang, Harrison Ho, Rob Egan, Shijie Yao, Dongwan Kang, Jeff Froula, Volkan Sevim, Frederik Schulz, Jackie E Shay, Derek Macklin, et al. A new method for rapid genome classification, clustering, visualization, and novel taxa discovery from metagenome. *bioRxiv*, page 812917, 2019.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Dongying Wu, Rekha Seshadri, Nikos C Kyrpides, and Natalia N Ivanova. A metagenomic perspective on the microbial prokaryotic genome census. *Science Advances*, 11(3):eadq2166, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Li Zhang, Han Guo, Leah V Schaffer, Young Su Ko, Digvijay Singh, Hamid Rahmani, Danielle Grotjahn, Elizabeth Villa, Michael Gilson, Wei Wang, et al. Proteinaligner: A multi-modal pretraining framework for protein foundation models. *bioRxiv*, pages 2024–10, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv*, 2024.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *bioRxiv*, 2022.