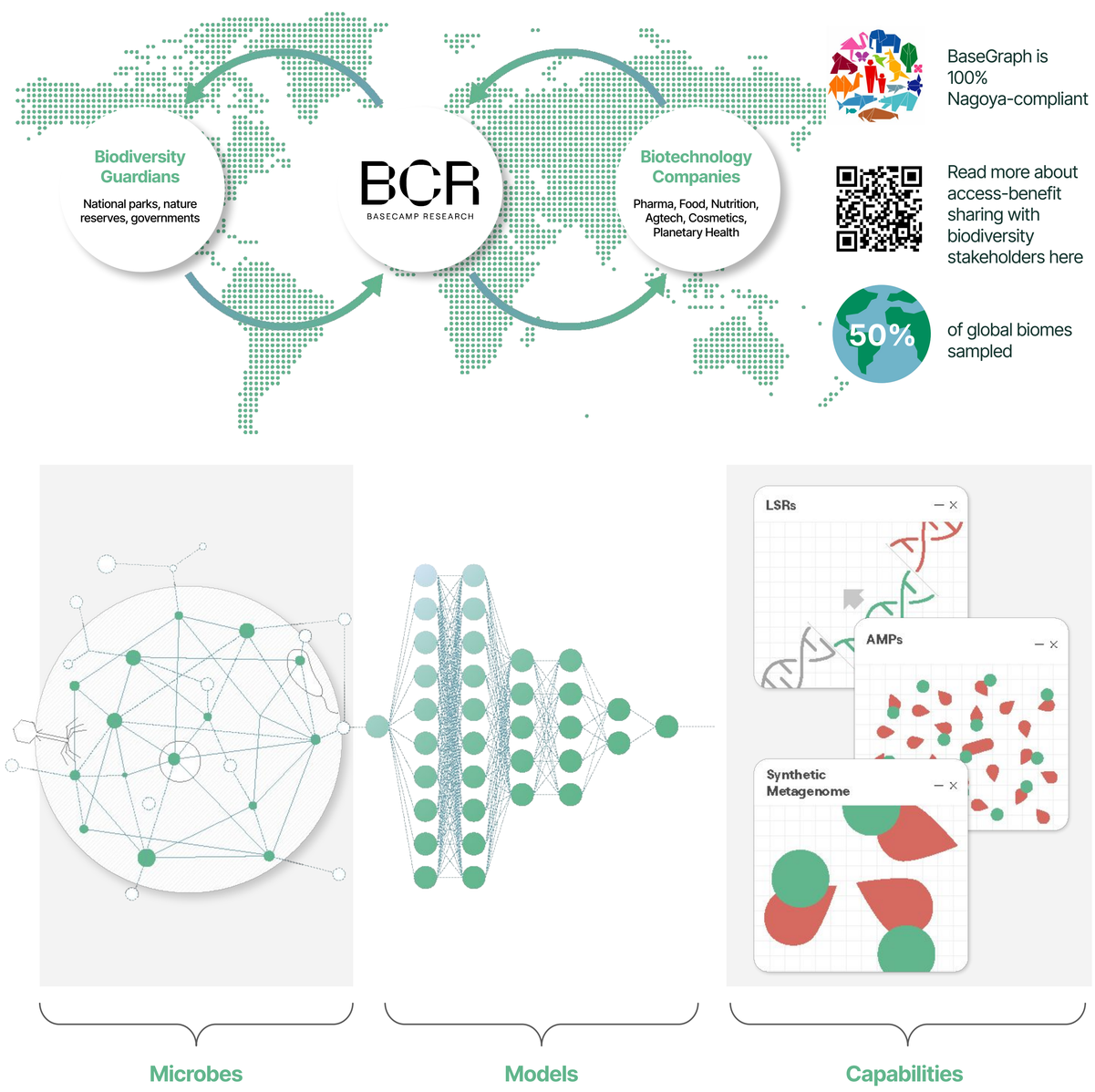
Scaling Laws and Architectural Frontiers in Metagenomic Foundation Models

BASECAMP RESEARCH

Geraldene Munsamy
geraldene@basecamp-research.com

Foundation models for genomics have the potential to revolutionize therapeutic design, yet the optimal architectural choices for modeling the vast and diverse distribution of metagenomic data remain under-explored. In this work, we present the machine learning methodology behind EDEN, a family of metagenomic foundation models scaled up to 28 billion parameters and trained on 9.7 trillion nucleotide tokens. We provide a systematic empirical study of architectural trade-offs between autoregressive Transformers (Llama-style), State-Space Models (Mamba), and Long-convolutional architectures (Hyena) for nucleotide-level modeling

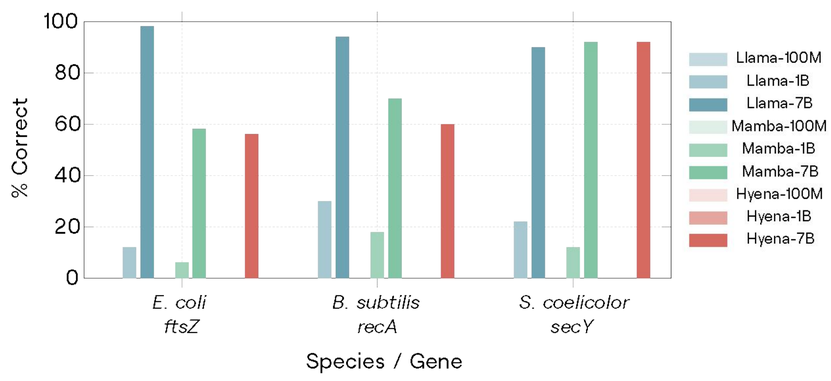
Modeling the Complexity of Biology



How Do We Scale Biology?

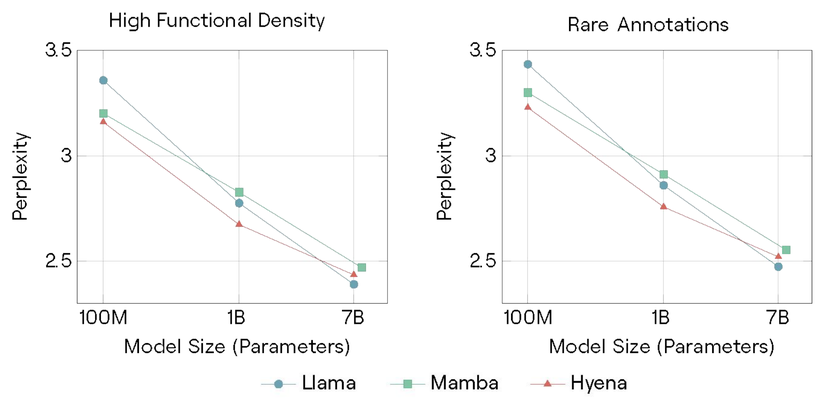
Llama (Transformer)	**Mamba** (SSM Hybrid)	**Hyena** (Convolution Hybrid)
RMSNorm (pre)	Mamba2 SSM	Short Hyena (S)
GQA + RoPE	MLP (SwiGLU)	Medium Hyena (D)
Residual add	Mamba2 SSM	Long Hyena (H)
RMSNorm (pre)	MLP (SwiGLU)	Attention (*)
SwiGLU FFN	Mamba2 SSM	Short Hyena (S)
Residual add	Attention (GQA)	Medium Hyena (D)
GQA + RoPE	MLP (SwiGLU)	Long Hyena (H)
SwiGLU FFN	Mamba2 SSM	Attention (*)

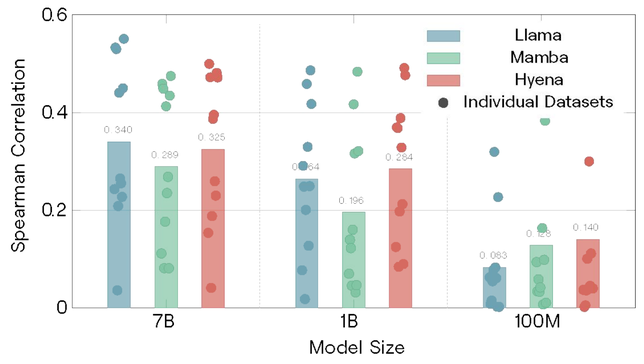
Gene Autocompletion and Fitness Prediction



Across all organisms and genes, smaller models showed limited generative capability often failing to recover the correct domain organization. Whereas 7B-parameter models substantially improved domain architecture recovery. Llama based models produced the most reliable structures with higher pLDDT values and closer agreement with expected folds.

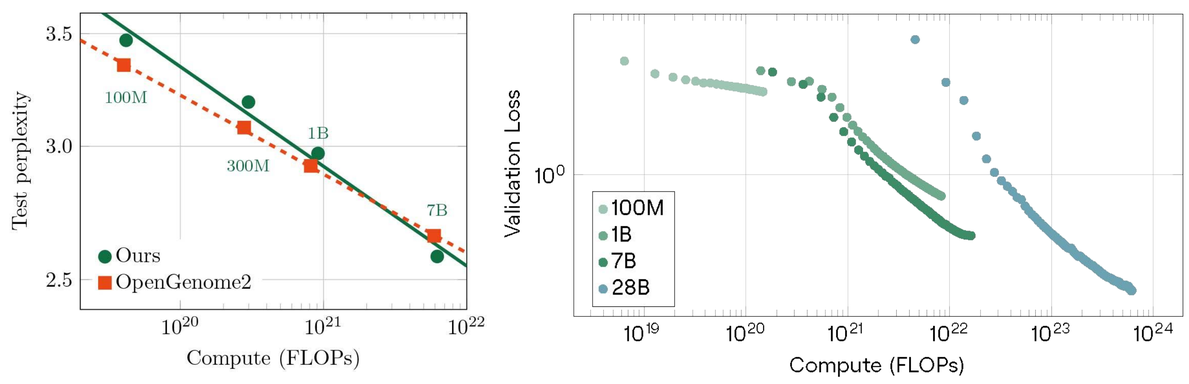
To better assess the semantic capacity of our models we evaluate how each architecture copes with increasing functional density using KEGG annotations as a ground truth for biological signal.
The Llama-based models exhibit superior scaling dynamics compared to their linear-time counterparts. While all models improve with size, Llama benefits most significantly from increased capacity, ultimately achieving best-in-class performance.





We evaluate the ability of different architectures zero-shot fitness prediction on a prokaryotic subset of RNAGym which contains deep mutational scans covering millions of variants. As illustrated, Spearman correlation with measured fitness scales with perplexity, with Llama excelling at the 7B scale.

Scaling laws and Data Quality



To isolate the impact of data quality on model performance, we conducted a scaling analysis, benchmarking our BaseData corpus against the OpenGenome2 baseline. We observe that the scaling exponent is indeed data-dependent. Models trained on our in-house dataset follow a steeper power-law trajectory ($\alpha \approx 0.058$) compared to those trained on OpenGenome2 ($\alpha \approx 0.047$). This difference indicates that our dataset possesses a higher density of learnable signal, yielding superior performance gains for every unit of added model capacity.

To the Context Length and Beyond