JanusDNA: A Powerful Bi-directional Hybrid DNA Foundation Model

Qihao Duan, Bingding Huang, Zhenqiao Song, Irina Lehmann, Lei Gu, Roland Eils, Benjamin Wild

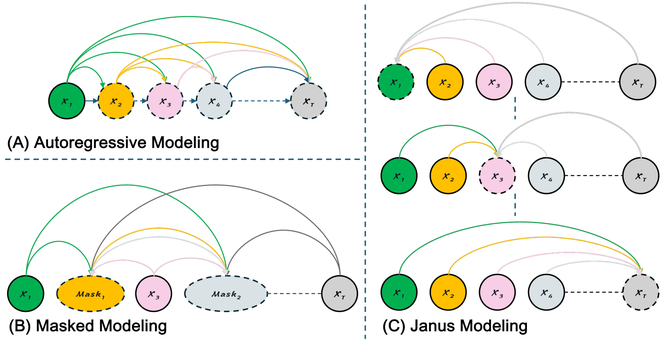
qihao.duan, benjamin.wild@bih-charite.de

BIH Berlin Institute of Health @Charité · FUDAN UNIVERSITY · MAX PLANCK INSTITUTE FOR HEART AND LUNG RESEARCH · Carnegie Mellon University · SZTU 深圳技术大学

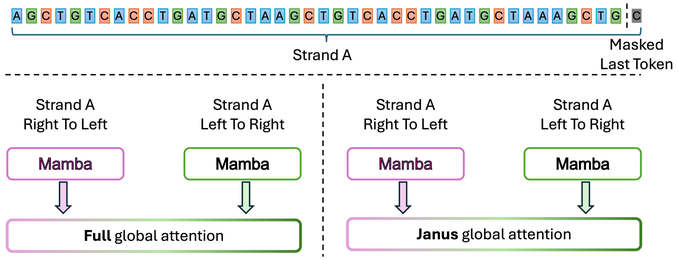
NEURAL INFORMATION PROCESSING SYSTEMS

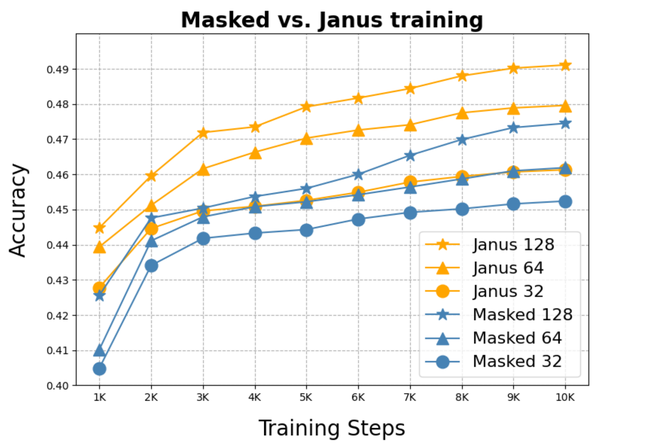
Janus Modeling

Efficiently learn from every token with bidirectional understanding.



Ablation Experiments



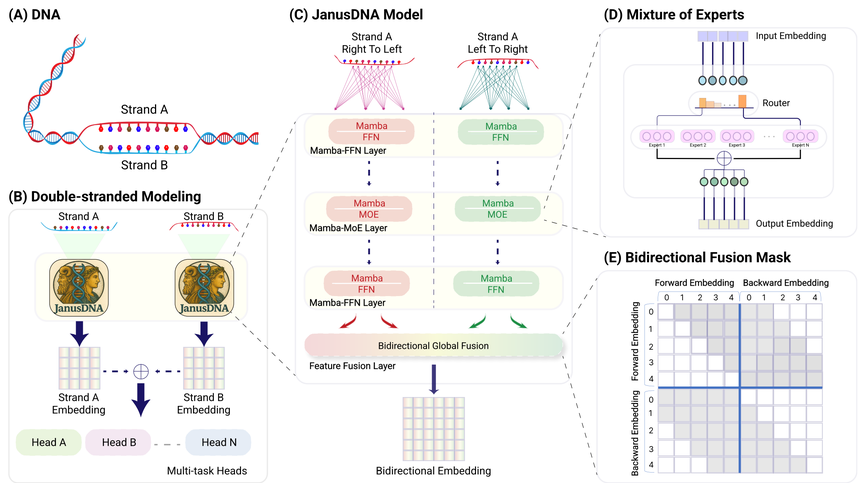


Motivation

1. *Global attention is stronger in modeling long-range dependencies in DNA, but it is memory-intensive —> leading to limited length coverage and low resolution.*
2. *DNA functions bidirectionally —> while unidirectional understanding introduces knowledge bias.*
3. *Traditional masked modeling learns from only about 15% of input tokens per training step —> resulting in low learning efficiency.*

JanusDNA Architecture and Workflow

Integrating ***bidirectional understanding into training*** while integrating ***bi-strand understanding into inference*** with ***global attention***.



Short-range Benchmarks Genomic Benchmark, Nucleotide Transformer

- Achieves performance comparable to or exceeding all previous SOTA models of similar scale.
- Surpasses models with **250×** more activated parameters on 12 out of 18 NT benchmark tasks.

Long-range Benchmarks DNALONGBENCH

- Outperforms models **30×** larger, including expert and SOTA models.

Contributions

- **JanusDNA** introduces the first bidirectional DNA foundation model, leveraging a novel **Janus Modeling** paradigm that integrates autoregressive optimization efficiency with masked modeling's bidirectional comprehension.
- Its **Hybrid Mamba–Attention–MoE** architecture enables **global long-range dependency modeling** and **single-nucleotide resolution** processing of sequences **up to 1 million base pairs on a single 80 GB GPU.**
- JanusDNA outperforms models up to **250×** larger and achieves state-of-the-art results across diverse short- and long-range genomic tasks.

Short-range Benchmarks 200 – 4,776 base pairs

Models activated param	CNN (264k)	HyenaDNA (436k)	Mamba (468k)	Caduceus Ph (470k)	Caduceus PS (470k)	ConvNova (386k)	JanusDNA mlp w/ Mid-Attn (426k)	JanusDNA mlp w/o Mid-Attn (431k)
Mouse Enhancers	0.715±0.087	0.780±0.025	0.743±0.054	0.754±0.074	**0.793**±0.058	0.784±0.009	0.770±0.048	0.769±0.029
Coding vs. Intergenomic	0.892±0.008	0.904±0.005	0.904±0.004	0.915±0.003	0.910±0.003	**0.943**±0.001	0.912±0.003	0.911±0.001
Human vs. Worm	0.942±0.002	0.964±0.002	0.967±0.002	**0.973**±0.001	0.968±0.002	0.967±0.002	0.971±0.001	0.971±0.001
Human Enhancers Cohn	0.702±0.021	0.729±0.014	0.732±0.029	**0.747**±0.004	0.745±0.007	0.743±0.005	0.741±0.005	0.742±0.006
Human Enhancer Ensembl	0.744±0.122	0.849±0.006	0.862±0.008	0.893±0.008	**0.900**±0.006	**0.900**±0.004	0.897±0.004	0.899±0.004
Human Regulatory	0.872±0.005	0.869±0.012	0.814±0.211	0.872±0.011	0.873±0.007	0.873±0.002	**0.877**±0.005	0.868±0.008
Human OCR Ensembl	0.698±0.013	0.783±0.007	0.815±0.002	**0.828**±0.006	0.818±0.006	0.793±0.004	0.822±0.003	0.824±0.001
Human NonTATA Promoters	0.861±0.009	0.944±0.002	0.933±0.007	0.946±0.007	0.945±0.010	0.951±0.003	**0.957**±0.004	0.954±0.010

Genomic Benchmark

	> 100M activated Param. Models			< 2M activated Param. Models				
	Enformer (252M)	DNABERT-2 (117M)	NT-v2 (500M)	HyenaDNA (1.6M)	Caduceus-Ph (1.9M)	ConvNova (1.7M)	JanusDNA mlp w/ midattn (2.001M)	JanusDNA mlp w/o midattn (2.009M)
Histone Markers								
H3	0.719±0.048	0.785±0.033	0.784±0.047	0.779±0.037	0.815±0.048	0.812±0.017	**0.835**±0.009	0.831±0.023
H3k14ac	0.288±0.077	0.516±0.028	0.551±0.021	0.612±0.065	0.631±0.026	0.644±0.009	**0.729**±0.022	0.718±0.026
H3k36me3	0.344±0.055	0.591±0.020	0.625±0.013	0.613±0.041	0.601±0.129	0.661±0.019	**0.702**±0.015	0.699±0.025
H3k4me1	0.291±0.061	0.511±0.028	0.550±0.021	0.512±0.024	0.523±0.039	0.554±0.023	0.615±0.035	**0.616**±0.018
H3k4me2	0.211±0.069	0.336±0.040	0.319±0.045	0.455±0.095	0.487±0.170	0.485±0.032	**0.589**±0.023	0.586±0.019
H3k4me3	0.158±0.072	0.352±0.077	0.410±0.033	0.549±0.056	0.544±0.045	0.566±0.027	**0.688**±0.026	0.675±0.014
H3k79me3	0.496±0.042	0.613±0.030	0.626±0.026	0.672±0.048	0.697±0.077	0.700±0.007	**0.747**±0.013	0.743±0.009
H3K9ac	0.420±0.063	0.542±0.029	0.562±0.040	0.581±0.061	0.622±0.030	0.658±0.011	**0.673**±0.014	0.661±0.027
H4	0.732±0.076	0.796±0.027	0.799±0.025	0.763±0.044	0.811±0.022	0.808±0.008	0.812±0.011	**0.813**±0.013
H4ac	0.273±0.063	0.463±0.041	0.495±0.032	0.564±0.038	0.621±0.054	0.636±0.011	0.698±0.013	**0.705**±0.023
Regulatory Annotation								
Enhancer	0.451±0.108	0.516±0.098	0.548±0.144	0.517±0.117	0.546±0.073	**0.586**±0.038	0.559±0.042	0.542±0.044
Enhancer types	0.309±0.134	0.423±0.051	0.424±0.132	0.386±0.185	0.439±0.054	0.500±0.018	**0.503**±0.038	0.492±0.096
Promoter: all	0.954±0.006	0.971±0.006	**0.976**±0.006	0.960±0.005	0.970±0.004	0.967±0.001	0.970±0.002	0.970±0.003
NonTATA	0.955±0.010	0.972±0.005	**0.976**±0.005	0.959±0.008	0.969±0.011	0.968±0.003	0.971±0.004	0.971±0.003
TATA	0.960±0.023	0.955±0.021	0.966±0.013	0.944±0.040	0.953±0.016	**0.969**±0.003	0.958±0.007	0.960±0.008
Splice Site Annotation								
All	0.848±0.019	0.939±0.009	**0.983**±0.008	0.956±0.011	0.940±0.027	0.965±0.004	0.967±0.005	0.943±0.020
Acceptor	0.914±0.028	0.975±0.006	**0.981**±0.011	0.958±0.010	0.937±0.033	0.971±0.003	0.957±0.012	0.961±0.009
Donor	0.906±0.027	0.963±0.006	**0.985**±0.022	0.949±0.024	0.948±0.025	0.965±0.003	0.948±0.008	0.935±0.016

Nucleotide Transformer

Long-range Benchmarks 450,000 base pairs

Models activated param	expert model (252M)	Caduceus-Ph (7.7M)	JanusDNA w/o mid-Attn (7.662M)	JanusDNA mlp w/o mid-Attn (7.745M)
Artery Tibial	0.741	0.690	0.803	**0.852**
Adipose Subcutaneous	0.736	0.759	0.741	**0.769**
Cells Cultured fibroblasts	0.639	0.690	0.771	**0.802**
Muscle Skeletal	0.621	0.789	0.803	**0.864**
Nerve Tibial	0.683	0.842	0.877	**0.914**
Skin Not Sun Exposed Suprapubic	0.710	0.812	0.875	**0.903**
Skin Sun Exposed Lower leg	0.700	0.692	0.706	**0.846**
Thyroid	0.612	0.703	0.752	**0.793**
Whole Blood	0.689	0.769	0.794	**0.821**

DNALONGBENCH

References

1. Yair Schiff, et al. Caduceus: Bi-directional equivariant long-range dna sequence modeling, ICML 2024.
2. Hugo Dalla-Torre, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(2):287–297, 2025.
3. Nguyen, E., et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. Neurips, 2024.
4. Žiga Avsec, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods, 18(10):1196–1203, 2021.
5. Gu, A. and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
6. Zhou, Z., et al. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006.
7. Cheng, Wenduo, et al. "Dnalongbench: a benchmark suite for long-range dna prediction tasks." bioRxiv (2025).

Code and Paper

WeChat