

A EXPERIMENTS

A.1 ALL EXPERIMENT RESULTS

	Epigenetic Marks Prediction					
	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3
DNABERT (3-mer)	74.15	42.07	48.49	42.95	31.34	28.92
DNABERT (4-mer)	73.03	41.88	48.03	41.06	30.66	25.31
DNABERT (5-mer)	73.40	40.68	48.29	40.65	30.67	27.10
DNABERT (6-mer)	73.10	40.06	47.25	41.44	32.27	27.81
NT-500M-human	69.67	33.55	44.14	37.15	30.87	24.06
NT-500M-1000g	72.52	39.37	45.58	40.45	31.05	26.16
NT-2500M-1000g	74.61	44.08	50.86	43.10	30.28	30.87
NT-2500M-multi	<u>78.77</u>	<u>56.20</u>	61.99	55.30	<u>36.49</u>	<u>40.34</u>
DNABERT-2	78.27	52.57	56.88	50.52	31.13	36.27
DNABERT-2♦	80.17	57.42	<u>61.90</u>	<u>53.00</u>	39.89	41.20

	Epigenetic Marks Prediction				Promoter Detection		
	H3K79me3	H3K9ac	H4	H4ac	all	notata	tata
DNABERT (3-mer)	60.12	50.48	78.27	38.60	90.44	93.61	69.83
DNABERT (4-mer)	59.77	51.44	78.28	36.40	89.54	92.65	66.78
DNABERT (5-mer)	59.61	51.11	77.27	37.48	90.16	92.45	69.51
DNABERT (6-mer)	61.17	51.22	79.26	37.43	90.48	93.05	61.56
NT-500M-human	58.35	45.81	76.17	33.74	87.71	90.75	78.07
NT-500M-1000g	59.33	49.29	76.29	36.79	89.76	91.75	<u>78.23</u>
NT-2500M-1000g	61.20	52.36	79.76	41.46	<u>90.95</u>	93.07	75.80
NT-2500M-multi	64.70	<u>56.01</u>	<u>81.67</u>	49.13	91.01	94.00	79.43
DNABERT-2	67.39	55.63	80.71	50.43	86.77	<u>94.27</u>	71.59
DNABERT-2♦	<u>65.46</u>	57.07	81.86	<u>50.35</u>	88.31	94.34	68.79

	Transcription Factor Prediction (Human)					Core Promoter Detection		
	0	1	2	3	4	all	notata	tata
DNABERT (3-mer)	67.95	70.90	60.51	53.03	69.76	70.92	69.82	78.15
DNABERT (4-mer)	67.90	<u>73.05</u>	59.52	50.37	71.23	69.00	70.04	74.25
DNABERT (5-mer)	66.97	69.98	59.03	52.95	69.26	69.48	69.81	<u>76.79</u>
DNABERT (6-mer)	66.84	70.14	61.03	51.89	70.97	68.90	<u>70.47</u>	76.06
NT-500M-human	61.59	66.75	53.58	42.95	60.81	63.45	64.82	71.34
NT-500M-1000g	63.64	70.17	52.73	45.24	62.82	66.70	67.17	73.52
NT-2500M-1000g	66.31	68.30	58.70	49.08	67.59	67.39	67.46	69.66
NT-2500M-multi	66.64	70.28	58.72	51.65	69.34	<u>70.33</u>	71.58	72.97
DNABERT-2	71.99	76.06	66.52	58.54	77.43	69.37	68.04	74.17
DNABERT-2♦	<u>69.12</u>	71.87	<u>62.96</u>	<u>55.35</u>	<u>74.94</u>	67.50	69.53	76.18

	Transcription Factor Prediction (Mouse)					Virus	Splice
	0	1	2	3	4	Covid	Reconstruct
DNABERT (3-mer)	42.31	79.10	69.90	55.40	41.97	62.23	84.14
DNABERT (4-mer)	49.42	79.95	72.62	51.79	44.13	59.87	84.05
DNABERT (5-mer)	42.45	79.32	62.22	49.92	40.34	50.46	84.02
DNABERT (6-mer)	44.42	78.94	71.44	44.89	42.48	55.50	84.07
NT-500M-human	31.04	75.04	61.67	29.17	29.27	50.82	79.71
NT-500M-1000g	39.26	75.49	64.70	33.07	34.01	52.06	80.97
NT-2500M-1000g	48.31	80.02	70.14	42.25	43.40	66.73	85.78
NT-2500M-multi	<u>63.31</u>	83.76	71.52	<u>69.44</u>	47.07	73.04	89.35
DNABERT-2	56.76	84.77	79.32	66.47	52.66	<u>71.02</u>	84.99
DNABERT-2♦	64.23	86.28	81.28	73.49	<u>50.80</u>	68.49	<u>85.93</u>

Table 4: This table presents the performance of all the models on the GUE benchmark. ♦: perform further pre-training on the training sets of the GUE benchmark.

	EMP	TF-M	CVC	TF-H	PD-tata	PD-o	CPD-tata	CPD-o	SSP
Epochs	3	1k	8	3	10	4	10	4	5

Table 5: The number of training steps we used for the following tasks: Epigenetic Marks Prediction (EMP), Transcription Factor Prediction on the Human genome and the Mouse genome (TF-H and TF-M), Covid Variants Classification (CVC), *tata* dataset of Promoter Detection (PD-tata), *notata* and *all* datasets of Promoter Detection (PD-o), *tata* dataset of Core Promoter Detection (CPD-tata), *notata* and *all* datasets of Core Promoter Detection (CPD-o), and Splice Site Prediction (SSP). In the task of Transcription Factor Prediction on the Mouse genome, we train the model for 1000 steps on each dataset.

A.2 IMPLEMENTATION

We pre-train DNABERT-2 with the Masked Language Modeling (MLM) loss with a mask ratio of 15%. Notably, we independently mask every token instead of masking spans of continuous tokens like Ji et al. (2021). We use a batch size of 4096 and a max sequence length of 128. We train the model for 500000 steps using the AdamW (Loshchilov & Hutter, 2019) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-6$ and weight decay of $1e-5$. The learning rate linearly increases from 0 to $5e-4$ during the first 30000 steps while linearly decreasing to 0 in the last 470000 steps.

A.3 HYPERPARAMETERS

This section presents the hyperparameters we used in the fine-tuning stage on each model. Table 5 shows the number of training steps we used for each task. We use AdamW (Loshchilov & Hutter, 2019) as optimizer. We keep most of the other hyperparameters the same for all the models across all the datasets, including a batch size of 32, a warmup step of 50, and a weight decay of 0.01. For DNABERT and DNABERT-2, we perform standard fine-tuning with a learning rate of $3e-5$, while for the Nucleotide Transformers, we perform parameter efficient fine-tuning (PEFT) using Low-Rank Adaptation (LoRA) with a learning rate of $1e-4$, a LoRA alpha of 16, a LoRA dropout of 0.05, and a LoRA r of 8. The hyperparameters are selected based on grid searches over commonly used ones in preliminary experiments. The pre-training stage takes approximately 14 days using eight Nvidia RTX 2080Ti GPUs. To train the model, we used the Transformers library (Wolf et al., 2020) and the Composer library (Team, 2021).

A.4 PRELIMINARY EXPERIMENTS ON NUCLEOTIDE TRANSFORMER

Since there is no official fine-tuning code of Nucleotide Transformer (Dalla-Torre et al., 2023), we use its open-sourced checkpoints in Huggingface Modelhub² and train it with our code base using LoRA. For a fair comparison with this model, in this section, we present preliminary experiments that compare the results reported in their paper with the performance of this model under our implementation. We select the epigenetic marks prediction task for benchmarking since it is the only shared task among Dalla-Torre et al. (2023) and GUE. The task contains 10 datasets. For each dataset, we randomly split it into training and test sets with a ratio of 9:1. As shown in Table 6, our LoRA implementation leads to slightly better results than the results reported in the original paper, making our comparison to the model fair and convincing despite the fact that we do not have access to its official fine-tuning implementation.

²<https://huggingface.co/InstaDeepAI>

	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3
500M-human*	72.00	37.00	45.00	36.00	27.00	24.00
500M-human	69.67	33.55	44.14	37.15	30.87	24.06
500M-1000g*	74.00	38.00	47.00	38.00	26.00	24.00
500M-1000g	72.52	39.37	45.58	40.45	31.05	26.16
2500M-1000g*	75.00	45.00	53.00	42.00	28.00	31.00
2500M-1000g	74.61	44.08	50.86	43.10	30.28	30.87
2500M-multi*	79.00	54.00	62.00	54.00	32.00	41.00
2500M-multi	78.77	56.20	61.99	55.30	36.49	40.34

	H3K79me3	H3K9ac	H4	H4ac	Average
500M-human*	57.00	45.00	75.00	33.00	45.10
500M-human	58.35	45.81	76.17	33.74	45.35
500M-1000g*	56.00	48.00	76.00	34.00	46.10
500M-1000g	59.33	49.29	76.29	36.79	47.68
2500M-1000g*	57.00	49.00	79.00	41.00	50.00
2500M-1000g	61.20	52.36	79.76	41.46	50.86
2500M-multi*	62.00	55.00	81.00	49.00	56.90
2500M-multi	64.70	56.01	81.67	49.13	58.06

Table 6: This table presents the performance of the Nucleotide Transformer on ten datasets of epigenetic marks prediction on the Yeast genome. As shown in the table, our implementation achieves better performance than the results reported in the paper, indicating the fairness of comparison in our experiments. *: Results taken from [Dalla-Torre et al. \(2023\)](#).

B DATA

B.1 MULTI-SPECIES GENOME FOR PRE-TRAINING

Table 7 lists the 135 species in 7 categories that we randomly selected for genome foundation model pre-training and presents the number of nucleotides we achieved from each species.

Category	Species	Num. of Nucleotides (M)
Fungi	Ceratobasidium	655.37
	Claviceps Maximensis	329.79
	Fusarium Annulatum	449.98
	Melampsora	699.52
	Metschnikowia	109.36
	Mucor Saturninus	391.17
	Penicillium Chermesinum	275.81
	Saccharomyces Cerevisiae	121.54
	Sporopachydermia Quercuum	155.71
	Tranzscheliella Williamsii	184.77
Xylariales	399.96	
Protozoa	Phytophthora Sojae	792.65
	Pythium Apiculatum	450.99
Mammalian	Bubalus Bubalis	28768.00
	Camelus Dromedarius	19757.02
	Human	31372.10
	Macaca Assamensis	27593.76
	Macaca Nigra	28217.13
Mus Musculus	26545.98	

(Continued on next page)

(Continued from previous page)

Category	Species	Nucleotides (M)
	Peromyscus Californicus	24677.56
Invertebrate	Brachionus Rubens	1327.37
	Ceroptres Masudai	12.95
	Cotesia Typhae	1866.62
	Croniades Pieria	3889.85
	Drosophila Athabasca	1221.16
	Emesis Russula	4848.08
	Hydra Oligactis	12597.75
	Meganola Albula	3604.25
	Oscheius	383.21
Rutpela Maculata	20213.33	
Other Vertebrate	Anas Zonorhyncha	11697.08
	Coregonus Clupeaformis	26824.02
	Gnathonemus Longibarbis	7314.74
	Myxocyprinus Asiaticus	23407.19
	Rhipidura Dahli	10112.96
Bacteria	Aeromonas	47.33
	Agrobacterium	97.22
	Alcaligenaceae Bacterium	20.88
	Aliivibrio	46.48
	Alphaproteobacteria Bacterium	14.22
	Amycolatopsis Antarctica	63.43
	Anaerostipes Faecis	32.00
	Arthrobacter	36.27
	Atopobium	28.63
	Bacillus Bc15	57.34
	Bacillus Bs3 2021	43.51
	Bacterium	7.54
	Bacteroidetes Bacterium Qs	8.99
	Breoghaniania Corrubedonensis	53.32
	Caldicoprobacter Oshimai	27.25
	Candidatus Cryptobacteroides Excrementipullorum	27.63
	Candidatus Dadabacteria Bacterium Rbg Combo	11.49
	Candidatus Dwaynia Gallinarum	16.82
	Candidatus Falkowbacteria Bacterium	13.88
	Candidatus Geothermincola Secundus	24.76
	Candidatus Gottesmanbacteria Bacterium	11.08
	Candidatus Nomurabacteria Bacterium Full	6.29
	Candidatus Portnoybacteria Bacterium Big Fil Rev	8.17
	Candidatus Regiella Insecticola	20.62
	Candidatus Roizmanbacteria Bacterium Combo All	11.13
	Candidatus Rokubacteria Bacterium	22.06
	Candidatus Saccharibacteria Bacterium	6.55
	Candidatus Staskawiczbacteria Bacterium Full	6.79
	Christensenella	18.75
	Clostridiaceae Bacterium	29.62
	Clostridiales Bacterium	16.59
	Clostridium Cag 505	21.26
	Clostridium Mcc328	36.43
	Clostridium Nexile	38.43
	Clostridium Uba3521	25.99
	Collinsella Urealyticum	19.45
Coprobacillus Cateniformis	38.38	

(Continued on next page)

(Continued from previous page)

Category	Species	Nucleotides (M)
	Cyanobium	40.33
	Dehalococcoidia Bacterium	17.59
	Enterobacteriaceae Bacterium	41.46
	Evtapia Gabavorous	24.94
	Firmicutes Bacterium	36.66
	Fulvirirga	65.24
	Jeongeupia Chitinilytica	39.11
	Legionella Endosymbiont Of Polyplax Serrata	5.30
	Listeria Ilorinensis	30.31
	Maribacter Cobaltidurans	46.40
	Marinomonas	37.73
	Mesorhizobium	65.15
	Methyloceanibacter Caenitepidi	34.25
	Microvirga	68.63
	Mycolicibacter Engbaekii	45.21
	Novosphingobium	46.18
	Omnitrophica Wor Bacterium Rbg	12.52
	Pantoea	43.14
	Paraburkholderia Edwinii	82.99
	Parerythrobacter Lutipelagi	30.98
	Paulownia Witches Phytoplasma	8.92
	Polaromonas Eurypsychrophila	41.61
	Prevotella Ag 487 50 53	29.63
Bacteria	Prevotella Uba3619	31.72
	Prevotella Uba634	18.51
	Prochlorococcus Ag-321-I09	3.29
	Prochlorococcus Ag-363-B18	15.54
	Prochlorococcus Ag-402-L19	11.17
	Prochlorococcus Scb243 498N4	14.12
	Providencia	41.89
	Pseudomonas 35 E 8	63.56
	Pseudomonas Bigb0408	59.52
	Pseudomonas P867	62.01
	Pseudomonas Promysalinigenes	50.47
	Roseobacter	44.14
	Salinicola Peritrichatus	46.19
	Salmonella S096 02912	48.09
	Salmonella Zj-F75	47.87
	Sinorhizobium	65.53
	Sodalis Ligni	63.85
	Sphaerochaeta	28.61
	Sphingobacterium	36.55
	Sphingomonas Carotinifaciens	37.53
	Sphingomonas Mesophila	22.91
	Sporosarcina Jiandibaonis	36.30
	Sporosarcina Ureilytica	34.37
	Staphylococcus Gdq20D1P	28.50
	Staphylococcus M0911	24.38
	Streptococcus	22.18
	Streptomyces 8401	88.39
	Streptomyces Di166	88.71
	Streptomyces Durbertensis	59.24
	Streptomyces Neau-Yj-81	118.84
	Streptomyces Rk74B	87.36
	Thermopetrobacter	26.06

(Continued on next page)

(Continued from previous page)

Category	Species	Nucleotides (M)
Bacteria	Uncultured Kushneria	35.31
	Uncultured Phascolarctobacterium	17.95
	Uncultured Proteus	35.66
	Verrucomicrobiales Bacterium	3.15
	Vibrio	41.47
	Victivallis Lenta	55.45
	Virgibacillus Salexigens	44.18
Xanthomonadales Bacterium	37.47	

Table 7: Details statistics of the multi-species genome dataset for pre-training.

B.2 GENOME UNDERSTANDING EVALUATION (GUE)

Task	Metric	Datasets	Train / Dev / Test
Core Promoter Detection	mcc	tata	4904 / 613 / 613
		notata	42452 / 5307 / 5307
		all	47356 / 5920 / 5920
Promoter Detection	mcc	tata	4904 / 613 / 613
		notata	42452 / 5307 / 5307
		all	47356 / 5920 / 5920
Transcription Factor Prediction (Human)	mcc	wgEncodeEH000552	32378 / 1000 / 1000
		wgEncodeEH000606	30672 / 1000 / 1000
		wgEncodeEH001546	19000 / 1000 / 1000
		wgEncodeEH001776	27294 / 1000 / 1000
		wgEncodeEH002829	19000 / 1000 / 1000
Splice Site Prediction	mcc	reconstructed	36496 / 4562 / 4562
Transcription Factor prediction (Mouse)	mcc	Ch12Nrf2Iggrab	6478 / 810 / 810
		Ch12Znf384hpa004051Iggrab	53952 / 6745 / 6745
		MelJundIggrab	2620 / 328 / 328
		MelMafkDm2p5dStd	1904 / 239 / 239
		MelNelfeIggrab	15064 / 1883 / 1883
Epigenetic Marks Prediction	mcc	H3	11971 / 1497 / 1497
		H3K14ac	26438 / 3305 / 3305
		H3K36me3	27904 / 3488 / 3488
		H3K4me1	25341 / 3168 / 3168
		H3K4me2	24545 / 3069 / 3069
		H3K4me3	29439 / 3680 / 3680
		H3K79me3	23069 / 2884 / 2884
		H3K9ac	22224 / 2779 / 2779
		H4	11679 / 1461 / 1461
H4ac	27275 / 3410 / 3410		
Virus	f1	Covid variant classification	77669 / 7000 / 7000

Table 8: Statistics of tasks in the GUE benchmark, including the name and the number of training, validation, and test samples in each dataset.

The proposed benchmark Genome Understanding Evaluation (GUE) contains 28 datasets of 7 biological important genome analysis tasks for 4 different species. To comprehensively evaluate the genome foundation models in modeling variable-length sequences, we select tasks with input lengths ranging from 70 to 1000. Table 8 presents the details statistics of each evaluation dataset. The following tasks are included in the GUE benchmark.

Promoter detection (Human) focuses on identifying (proximal) promoter regions, crucial sequences in the human genome responsible for instigating transcription. As many primary regulatory elements are located in this region, accurately detecting these sites is instrumental in advancing

our grasp of gene regulation mechanisms and pinpointing the genomic underpinnings of numerous diseases. The dataset is divided twofold, TATA and non-TATA, based on whether a TATA box motif is present in the sequence. We extract -249 +50 bp around the transcription start site (TSS) from TATA and non-TATA promoters downloaded from Eukaryotic Promoter Database (EPDnew) (Dreos et al., 2013) and use it as our promoter class. Meanwhile, we construct the non-promoter class with equal-sized randomly selected sequences outside of promoter regions but with TATA motif (TATA non-promoters) or randomly substituted sequences (non-TATA, non-promoters). We also combine the TATA and non-TATA datasets to obtain a combined dataset named *all*.

Core promoter detection (Human) is similar to proximal promoter detection with a focus on predicting the core promoter region only, the central region closest to the TSS and start codon. A much shorter context window (center -34 +35 bp around TSS) is provided, making this a more challenging task than proximal promoter prediction.

Transcription factor binding site prediction (Human) predicts binding sites of transcription factors (TF), the key proteins that regulate gene expression in the human genome. Their accurate prediction is key to deciphering complex genetic interactions and identifying potential targets for gene therapies. We accessed the legacy 690 ENCODE ChIP-seq experiments (Consortium et al., 2012) via the UCSC genome browser, which encompasses 161 TF binding profiles in 91 human cell lines. We extracted a 101-bp region around the center of each peak as TFBS class and nonoverlapping sequences with the same length and GC content as non-TFBS class. Finally, we randomly select 5 datasets out of a subset of 690 that we curated by heuristically filtering out tasks that are either too trivial (e.g., over 0.95 F1) or too challenging (e.g., less than 0.50 F1) for existing language models.

Splice site prediction (Human) predicts splice donor and acceptor sites, which are the exact locations in the human genome where alternative splicing occurs. This prediction is crucial to understanding protein diversity and the implications of aberrant splicing in genetic disorders. The dataset (Wang et al., 2019) consists of 400-bp-long sequences extracted from Ensembl GRCh38 human reference genome. As suggested by Ji et al. (2021), existing models can achieve almost perfect performance on the original dataset, containing 10,000 splice donors, acceptors, and non-splice site sequences, which is overly optimistic on detecting non-canonical sites in reality. As such, we reconstruct the dataset by iteratively adding adversarial examples (unseen false positive predictions in hold-out set) in order to make this task more challenging.

Transcription factor binding site prediction (Mouse) predicts the binding site of transcription factors on mouse genomes. Similar to human binding site data, we obtain mouse ENCODE ChIP-seq data (Stamatoyannopoulos et al., 2012), which is the largest available collection on the UCSC genome browser (n=78). This time, the negative examples are created using dinucleotide shuffling while preserving relative frequencies, while all other settings stay the same as the human TFBS prediction dataset. We also randomly select 5 datasets out of the 78 datasets using the same process described above.

Epigenetic marks prediction (Yeast) predicts epigenetic marks in yeast, modifications on the genetic material that influence gene expression without altering the DNA sequence. Precise prediction of these marks aids in elucidating the role of epigenetics in yeast. We download the 10 datasets from <http://www.jaist.ac.jp/~tran/nucleosome/members.htm> and randomly split each dataset into training, validation, and test sets with a ratio of 8:1:1.

Covid variant prediction (Virus) aims to predict the variant type of the SARS_CoV_2 virus based on 1000-length genome sequences. We download the genomes from the EpiCoV database (Khare et al., 2021) of the Global Initiative on Sharing Avian Influenza Data (GISAID). We consider 9 types of SARS_CoV_2 variants, including *Alpha*, *Beta*, *Delta*, *Eta*, *Gamma*, *Iota*, *Kappa*, *Lambda* and *Zeta*.