

# 1 Benchmarks

## 1.1 GUE

Table 1: The results on the GUE datasets.

Model	Promoter Detection			Core Promoter Detection		
	all	notata	tata	all	notata	tata
DNABERT (3-mer)	90.44	93.61	69.83	<b>70.92</b>	69.82	<u>78.15</u>
DNABERT (4-mer)	89.54	92.65	66.78	69.00	70.04	74.25
DNABERT (5-mer)	90.16	92.45	69.51	69.48	69.81	76.79
DNABERT (6-mer)	90.48	93.05	61.56	68.90	<u>70.47</u>	76.06
NT-500M-human	87.71	90.75	78.07	63.45	64.82	71.34
NT-500M-1000g	89.76	91.75	78.23	66.70	67.17	73.52
NT-2500M-1000g	90.95	93.07	75.80	67.39	67.46	69.66
NT-2500M-multi	<u>91.01</u>	94.00	<b>79.43</b>	<u>70.33</u>	<b>71.58</b>	72.97
DNABERT-2	86.77	<u>94.27</u>	71.59	69.37	68.04	74.17
DNABERT-2 ■	88.31	<b>94.34</b>	68.79	67.50	69.53	76.18
Grover	86.42	92.30	59.77	63.58	66.75	60.57
Enformer	85.68	92.92	69.63	60.94	66.46	46.21
SPACE	<b>91.90</b>	94.23	<u>79.13</u>	68.18	68.04	<b>79.23</b>

Model	Transcription Factor Prediction (Human)					Splice
	0	1	2	3	4	Splice
DNABERT(3-mer)	67.95	70.90	60.51	53.03	69.76	84.14
DNABERT(4-mer)	67.90	73.05	59.52	50.37	71.23	84.05
DNABERT(5-mer)	66.97	69.98	59.03	52.95	69.26	84.02
DNABERT(6-mer)	66.84	70.14	61.03	51.89	70.97	84.07
NT-500M-human	61.59	66.75	53.58	42.95	60.81	79.71
NT-500M-1000g	63.64	70.17	52.73	45.24	62.82	80.97
NT-2500M-1000g	66.31	68.30	58.70	49.08	67.59	85.78
NT-2500M-multi	66.64	70.28	58.72	51.65	69.34	<b>89.35</b>
DNABERT-2	<b>71.99</b>	<u>76.06</u>	66.52	58.54	77.43	84.99
DNABERT-2 ■	69.12	71.87	62.96	55.35	74.94	85.93
Grover	65.76	67.9	61.62	48.26	74.68	84.35
Enformer	<u>69.42</u>	72.76	<b>77.88</b>	<b>66.41</b>	<u>81.89</u>	81.55
SPACE	69.02	<b>76.49</b>	<u>76.45</u>	<u>66.08</u>	<b>82.91</b>	<u>87.48</u>

## 1.2 Nucleotide Transformer Downstream Tasks Revised

Table 2: Complete Benchmark Results of Nucleotide Transformer Downstream Tasks

Model	Chromatin profiles					
	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
BPNet (original)	0.473 $\pm$ 0.009	0.296 $\pm$ 0.046	0.543 $\pm$ 0.009	0.548 $\pm$ 0.009	0.436 $\pm$ 0.008	0.427 $\pm$ 0.036
BPNet (large)	0.487 $\pm$ 0.014	0.214 $\pm$ 0.037	0.551 $\pm$ 0.009	0.570 $\pm$ 0.009	0.459 $\pm$ 0.012	0.427 $\pm$ 0.025
DNABERT-2	0.490 $\pm$ 0.013	0.491 $\pm$ 0.010	0.599 $\pm$ 0.010	<b>0.637 <math>\pm</math> 0.007</b>	0.490 $\pm$ 0.008	0.558 $\pm$ 0.013
HyenaDNA-1KB	0.455 $\pm$ 0.015	0.423 $\pm$ 0.017	0.541 $\pm$ 0.018	0.543 $\pm$ 0.010	0.430 $\pm$ 0.014	0.521 $\pm$ 0.024
HyenaDNA-32KB	0.467 $\pm$ 0.012	0.421 $\pm$ 0.010	0.550 $\pm$ 0.009	0.553 $\pm$ 0.011	0.423 $\pm$ 0.016	0.515 $\pm$ 0.018
NT-HumanRef (500M)	0.465 $\pm$ 0.011	0.457 $\pm$ 0.010	0.589 $\pm$ 0.009	0.594 $\pm$ 0.004	0.468 $\pm$ 0.007	0.527 $\pm$ 0.011
NT-1000G (500M)	0.464 $\pm$ 0.012	0.458 $\pm$ 0.012	0.591 $\pm$ 0.007	0.581 $\pm$ 0.009	0.466 $\pm$ 0.006	0.528 $\pm$ 0.011
NT-1000G (2.5B)	0.478 $\pm$ 0.012	0.486 $\pm$ 0.023	<b>0.603 <math>\pm</math> 0.009</b>	0.632 $\pm$ 0.008	0.491 $\pm$ 0.015	0.569 $\pm$ 0.014
NT-Multispecies (2.5B)	0.503 $\pm$ 0.010	0.481 $\pm$ 0.020	0.593 $\pm$ 0.016	0.635 $\pm$ 0.016	0.481 $\pm$ 0.012	0.552 $\pm$ 0.022
Grover	0.513 $\pm$ 0.00004	0.500 $\pm$ 0.001	0.591 $\pm$ 0.001	0.596 $\pm$ 0.004	0.475 $\pm$ 0.011	0.572 $\pm$ 0.010
Enformer	0.522 $\pm$ 0.019	0.520 $\pm$ 0.015	0.552 $\pm$ 0.007	0.567 $\pm$ 0.017	0.504 $\pm$ 0.021	0.626 $\pm$ 0.015
SPACE	<b>0.548 <math>\pm</math> 0.005</b>	<b>0.547 <math>\pm</math> 0.007</b>	0.586 $\pm$ 0.010	0.602 $\pm$ 0.005	<b>0.543 <math>\pm</math> 0.009</b>	<b>0.640 <math>\pm</math> 0.007</b>

  

Model	Chromatin profiles				Regulatory elements	
	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Enhancers	Enhancers(types)
BPNet (original)	0.445 $\pm$ 0.047	0.336 $\pm$ 0.034	0.298 $\pm$ 0.030	0.531 $\pm$ 0.025	0.488 $\pm$ 0.009	0.449 $\pm$ 0.006
BPNet (large)	0.445 $\pm$ 0.049	0.298 $\pm$ 0.033	0.234 $\pm$ 0.037	0.525 $\pm$ 0.038	0.492 $\pm$ 0.008	0.454 $\pm$ 0.008
DNABERT-2	0.646 $\pm$ 0.008	0.564 $\pm$ 0.013	0.443 $\pm$ 0.025	0.655 $\pm$ 0.011	0.517 $\pm$ 0.011	0.476 $\pm$ 0.009
HyenaDNA-1KB	0.596 $\pm$ 0.015	0.484 $\pm$ 0.022	0.375 $\pm$ 0.026	0.580 $\pm$ 0.009	0.475 $\pm$ 0.006	0.441 $\pm$ 0.010
HyenaDNA-32KB	0.603 $\pm$ 0.020	0.487 $\pm$ 0.025	0.419 $\pm$ 0.030	0.590 $\pm$ 0.007	0.476 $\pm$ 0.021	0.445 $\pm$ 0.009
NT-HumanRef (500M)	0.622 $\pm$ 0.013	0.524 $\pm$ 0.013	0.433 $\pm$ 0.009	0.634 $\pm$ 0.013	0.515 $\pm$ 0.019	0.477 $\pm$ 0.014
NT-1000G (500M)	0.609 $\pm$ 0.011	0.515 $\pm$ 0.018	0.415 $\pm$ 0.019	0.634 $\pm$ 0.010	0.505 $\pm$ 0.009	0.459 $\pm$ 0.011
NT-1000G (2.5B)	0.615 $\pm$ 0.017	0.529 $\pm$ 0.012	0.483 $\pm$ 0.013	<b>0.659 <math>\pm</math> 0.008</b>	0.504 $\pm$ 0.009	0.469 $\pm$ 0.005
NT-Multispecies (2.5B)	0.618 $\pm$ 0.015	0.527 $\pm$ 0.017	0.447 $\pm$ 0.018	0.650 $\pm$ 0.014	0.527 $\pm$ 0.012	0.484 $\pm$ 0.012
Grover	0.621 $\pm$ 0.002	0.520 $\pm$ 0.023	0.421 $\pm$ 0.018	0.630 $\pm$ 0.007	0.526 $\pm$ 0.016	0.474 $\pm$ 0.003
Enformer	0.635 $\pm$ 0.019	0.593 $\pm$ 0.020	0.453 $\pm$ 0.016	0.606 $\pm$ 0.016	0.614 $\pm$ 0.010	0.573 $\pm$ 0.013
SPACE	<b>0.661 <math>\pm</math> 0.025</b>	<b>0.635 <math>\pm</math> 0.016</b>	<b>0.490 <math>\pm</math> 0.011</b>	0.650 $\pm$ 0.011	<b>0.631 <math>\pm</math> 0.007</b>	<b>0.583 <math>\pm</math> 0.008</b>

  

Model	Regulatory elements			Splicing		
	All	NoTATA	TATA	Donors	Acceptors	All
BPNet (original)	0.696 $\pm$ 0.026	0.717 $\pm$ 0.023	0.848 $\pm$ 0.042	0.859 $\pm$ 0.038	0.793 $\pm$ 0.072	0.920 $\pm$ 0.014
BPNet (large)	0.672 $\pm$ 0.023	0.672 $\pm$ 0.043	0.826 $\pm$ 0.017	0.925 $\pm$ 0.031	0.865 $\pm$ 0.026	0.930 $\pm$ 0.021
DNABERT-2	0.754 $\pm$ 0.009	0.769 $\pm$ 0.009	0.784 $\pm$ 0.036	0.837 $\pm$ 0.006	0.855 $\pm$ 0.005	0.861 $\pm$ 0.004
HyenaDNA-1KB	0.693 $\pm$ 0.016	0.723 $\pm$ 0.013	0.648 $\pm$ 0.044	0.815 $\pm$ 0.049	0.854 $\pm$ 0.053	0.943 $\pm$ 0.024
HyenaDNA-32KB	0.698 $\pm$ 0.011	0.729 $\pm$ 0.009	0.666 $\pm$ 0.041	0.808 $\pm$ 0.009	0.907 $\pm$ 0.018	0.915 $\pm$ 0.047
NT-HumanRef (500M)	0.734 $\pm$ 0.013	0.738 $\pm$ 0.008	0.831 $\pm$ 0.022	0.941 $\pm$ 0.004	0.939 $\pm$ 0.003	0.952 $\pm$ 0.003
NT-1000G (500M)	0.727 $\pm$ 0.004	0.743 $\pm$ 0.012	0.855 $\pm$ 0.041	0.933 $\pm$ 0.007	0.939 $\pm$ 0.004	0.952 $\pm$ 0.004
NT-1000G (2.5B)	0.708 $\pm$ 0.008	0.758 $\pm$ 0.007	0.802 $\pm$ 0.030	0.952 $\pm$ 0.004	0.956 $\pm$ 0.004	0.963 $\pm$ 0.001
NT-Multispecies (2.5B)	0.761 $\pm$ 0.009	0.773 $\pm$ 0.010	<b>0.944 <math>\pm</math> 0.016</b>	<b>0.958 <math>\pm</math> 0.003</b>	<b>0.964 <math>\pm</math> 0.003</b>	<b>0.970 <math>\pm</math> 0.002</b>
Grover	0.738 $\pm$ 0.012	0.754 $\pm$ 0.015	0.845 $\pm$ 0.007	0.785 $\pm$ 0.056	0.739 $\pm$ 0.002	0.784 $\pm$ 0.004
Enformer	0.745 $\pm$ 0.012	0.763 $\pm$ 0.012	0.793 $\pm$ 0.026	0.749 $\pm$ 0.007	0.739 $\pm$ 0.011	0.780 $\pm$ 0.007
SPACE	<b>0.764 <math>\pm</math> 0.012</b>	<b>0.776 <math>\pm</math> 0.011</b>	0.838 $\pm$ 0.028	0.942 $\pm$ 0.006	0.902 $\pm$ 0.004	0.906 $\pm$ 0.003

### 1.3 Genomic Benchmarks

Table 3: The results on the Genomic Benchmarks datasets, which is from [https://github.com/ML-Bioinfo-CEITEC/genomic\\_benchmarks](https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks).

Model	Mouse	Demo		drosophila	
	Enhancers	Coding VS. Intergenic	Human VS. Worm	Enhancers	
CNN	$0.715 \pm 0.087$	$0.892 \pm 0.008$	$0.942 \pm 0.002$	0.586	
HyenaDNA	$0.780 \pm 0.025$	$0.904 \pm 0.005$	$0.964 \pm 0.002$	—	
Mamba	$0.743 \pm 0.054$	$0.904 \pm 0.004$	$0.967 \pm 0.002$	—	
Caduceus-PH	$0.754 \pm 0.074$	$0.915 \pm 0.003$	<b><math>0.973 \pm 0.001</math></b>	—	
Caduceus-PS	$0.793 \pm 0.058$	$0.910 \pm 0.003$	$0.968 \pm 0.002$	—	
Enformer	$0.835 \pm 0.012$	$0.913 \pm 0.001$	$0.958 \pm 0.001$	$0.613 \pm 0.005$	
SPACE	<b><math>0.905 \pm 0.010</math></b>	<b><math>0.922 \pm 0.001</math></b>	$0.967 \pm 0.004$	<b><math>0.721 \pm 0.016</math></b>	

  

Model	Human				
	Enhancers Cohn	Enhancer Ensembl	Regulatory	OCR Ensembl	Nontata Promoters
CNN	$0.702 \pm 0.021$	$0.744 \pm 0.122$	$0.872 \pm 0.005$	$0.698 \pm 0.013$	$0.861 \pm 0.009$
HyenaDNA	$0.729 \pm 0.014$	$0.849 \pm 0.006$	$0.869 \pm 0.012$	$0.783 \pm 0.007$	$0.944 \pm 0.002$
Mamba	$0.732 \pm 0.029$	$0.862 \pm 0.008$	$0.814 \pm 0.211$	$0.815 \pm 0.002$	$0.933 \pm 0.007$
Caduceus-PH	$0.747 \pm 0.004$	$0.893 \pm 0.008$	$0.872 \pm 0.011$	$0.828 \pm 0.006$	<b><math>0.946 \pm 0.007</math></b>
Caduceus-PS	$0.745 \pm 0.007$	$0.900 \pm 0.006$	$0.873 \pm 0.007$	$0.818 \pm 0.006$	$0.945 \pm 0.010$
Enformer	$0.723 \pm 0.001$	$0.844 \pm 0.001$	$0.903 \pm 0.001$	<b><math>0.876 \pm 0.001</math></b>	$0.878 \pm 0.002$
SPACE	<b><math>0.769 \pm 0.006</math></b>	<b><math>0.919 \pm 0.014</math></b>	<b><math>0.944 \pm 0.002</math></b>	$0.854 \pm 0.001$	$0.940 \pm 0.002$

## 1.4 BEND

Table 4: Performance Comparison of Genomic Prediction Methods

Method	Genomic Tasks				
	Chromatin accessibility	Histone modification	CpG Methylation	Variant effects (expression)	Variant effects (disease)
<b>Expert method</b>	0.85	0.74	0.93	0.70	0.56
	BASSET	BASSET	BASSET	DEEPSEA	DEEPSEA
<b>Fully supervised</b>					
ResNet	–	–	–	–	–
CNN	0.75	0.76	0.84	–	–
<b>Pre-trained</b>					
ResNet-LM	0.82	0.77	0.87	0.55	0.55
AWD-LSTM	0.69	0.74	0.81	0.53	0.45
NT-H	0.74	0.76	0.88	0.55	0.48
NT-MS	0.79	0.78	<b>0.92</b>	0.54	0.77
NT-1000G	0.77	0.77	0.89	0.45	0.49
NT-V2	0.80	0.76	0.91	0.48	0.48
DNABERT	0.85	0.79	0.91	<b>0.60</b>	0.56
DNABERT-2	0.81	0.78	0.90	0.49	0.51
GENA-LM BERT	0.76	0.78	0.91	0.49	0.55
GENA-LM BigBird	0.82	0.78	0.91	0.49	0.52
HyenaDNA large	0.84	0.76	0.91	0.51	0.45
HyenaDNA tiny	0.78	0.76	0.86	0.47	0.44
GROVER	0.82	0.77	0.89	0.56	0.51
GPN-MSA	–	–	–	–	0.97
SPACE	<b>0.89</b>	<b>0.81</b>	<b>0.92</b>	0.51	0.49

## 2 Ablation

### 2.1 Ablation on NT

Table 5: Ablation on NT

**SPACE w/o decoder:** Removal of the SPACE decoder module

**SPACE w/o decoder w/ MLP:** Replacement of the decoder with a parameter-matched MLP layer

**SPACE w/o encoder:** Substitution of the MoE encoder with a standard FFN layer (retaining species embedding)

**SPACE w/o encoder and species embedding:** Further removal of species embedding from the encoder-ablated variant

Model	Chromatin profiles					
	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
SPACE - decoder	0.535	0.514	0.567	0.593	0.520	0.604
SPACE - decoder + MLP	0.551	0.528	0.577	0.580	0.534	0.637
SPACE - encoder	0.540	0.524	0.569	0.579	0.506	0.625
SPACE - encoder - species emb	0.551	0.518	0.566	0.585	0.519	0.622
SPACE	0.556	0.529	0.579	0.593	0.516	0.612

  

Model	Chromatin profiles				Regulatory elements	
	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Enhancers	Enhancers(types)
SPACE - decoder	0.661	0.601	0.452	0.627	0.598	0.563
SPACE - decoder + MLP	0.668	0.589	0.451	0.636	0.601	0.558
SPACE - encoder	0.627	0.585	0.461	0.637	0.612	0.564
SPACE - encoder - species emb	0.654	0.588	0.454	0.635	0.596	0.563
SPACE	0.637	0.582	0.457	0.644	0.607	0.564

  

Model	Regulatory elements				Splicing	
	All	NoTATA	TATA	Acceptors	All	Donors
SPACE - decoder	0.752	0.773	0.841	0.873	0.884	0.936
SPACE - decoder + MLP	0.743	0.750	0.808	0.883	0.886	0.937
SPACE - encoder	0.738	0.769	0.828	0.864	0.869	0.933
SPACE - encoder - species emb	0.739	0.767	0.828	0.869	0.876	0.942
SPACE	0.763	0.776	0.802	0.898	0.884	0.941

## 2.2 Ablation on GUE’s virus and yeast tasks

Table 6: Comparison Results on the GUE Benchmark

Model	Epigenetic Marks Prediction				
	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2
SPACE - dec +MLP	75.59	45.17	48.21	39.70	34.81
SPACE - enc	76.16	48.78	49.14	37.57	34.08
SPACE	76.40	50.76	49.18	41.30	32.83

  

Model	Epigenetic Marks Prediction					Virus
	H3K4me3	H3K79me3	H3K9ac	H4	H4ac	Covid
SPACE - dec + MLP	34.26	58.94	56.36	78.81	43.49	67.83
SPACE - enc	36.84	63.44	56.63	77.17	50.78	68.46
SPACE	37.74	61.10	57.06	79.33	51.05	68.89