Supplementary Information: Dataset naming

We have collected four genomic classification tasks including 17 datasets. We named the datasets in a manner that clearly shows their features. Specifically:

For the **4mc sites detection in multiple species** task, we named the six datasets as: A.Thaliana 4mC, C.Elegans 4mC, D.Melanogaster 4mC, E.Coli 4mC, G.Pickeringii 4mC, G.Subterraneus 4mC.

For the **DNase-I hypersensitive sites detection** task, we named the one dataset as DNase_I Hypersensitive.

For the **5mC and 6mA modifications detection** task, we named the two datasets as: 5-methylcytosin (5mC), N6-methyladenosine (6mA)

For the **Promoter identification in multiple species** task, we named the eight datasets as: Promoter GM12878, Promoter HUVEC, Promoter Hela-S3, Promoter NHEK, Promoter B_amyloliquefaciens, Promoter R_capsulatus, Promoter Arabidopsis NonTATA, Promoter Arabidopsis TATA.

Besides, we have renamed the datasets adopted from DNABERT-2, NT-v2, and HyenaDNA in our study for better overall clarity. This document provides a mapping between the original dataset names and the nomenclature in our study. It can serve as a reference for tracing back to the original dataset sources.

1. Datasets adopted from DNABERT-2

Detailed descriptions of these datasets can be found in the original DNABERT-2 publication.

Our Dataset Naming	Original Reference
Promoter NonTATA 300 bps Promoter TATA 300 bps Promoter All 300 bps	Promoter detection (Human)
Promoter NonTATA 70 bps Promoter TATA 70 bps Promoter All 70 bps	Core promoter detection (Human)
Human TFBS 1 Human TFBS 2 Human TFBS 3 Human TFBS 4 Human TFBS 5	Transcription factor binding site prediction (Human)
Mouse TFBS 1 Mouse TFBS 2 Mouse TFBS 3 Mouse TFBS 4 Mouse TFBS 5	Transcription factor binding site prediction (Mouse)
Yeast H3 Yeast H3K79me3 Yeast H3K9ac Yeast H3K14ac Yeast H3K4me3 Yeast H3K36me3 Yeast H3K4me2	Epigenetic marks prediction (Yeast)

Yeast H4 Yeast H3K4me1 Yeast H4ac	
Covid variants	Covid variant prediction (Virus)
Splice Site Type DNABERT	Splice site prediction (Human)

2. Datasets adopted from NT-v2

Detailed descriptions can be found in the corresponding sections of the original NT-v2 publication.

Our Dataset Naming	**Original Reference**
Enhancer Enhancer strength	Enhancer sequence prediction, Section A 4.3
Splice Site Type NT Donors Acceptors	Splice site prediction, Section A 4.4

3. Datasets adopted from HyenaDNA

Detailed descriptions of these datasets can be found in the original DNABERT-2 publication.

Our Dataset Naming	**Original Reference**
Enhancer ensembl	Human Enhancers Ensembl
Promoter NonTATA 251bps	Human Nontata Promoters
Human vs worm	Human vs worm
Regulatory region type	Human Regulatory
Open chromatin region	Human OCR Ensembl
Coding	Coding vs Intergenomic
Enhancer cohn	Human Enhancers Cohn

Supplementary Tables and Figures

Supplementary Table 1: Details of each model. The bolded ones are the selected configurations in this study.

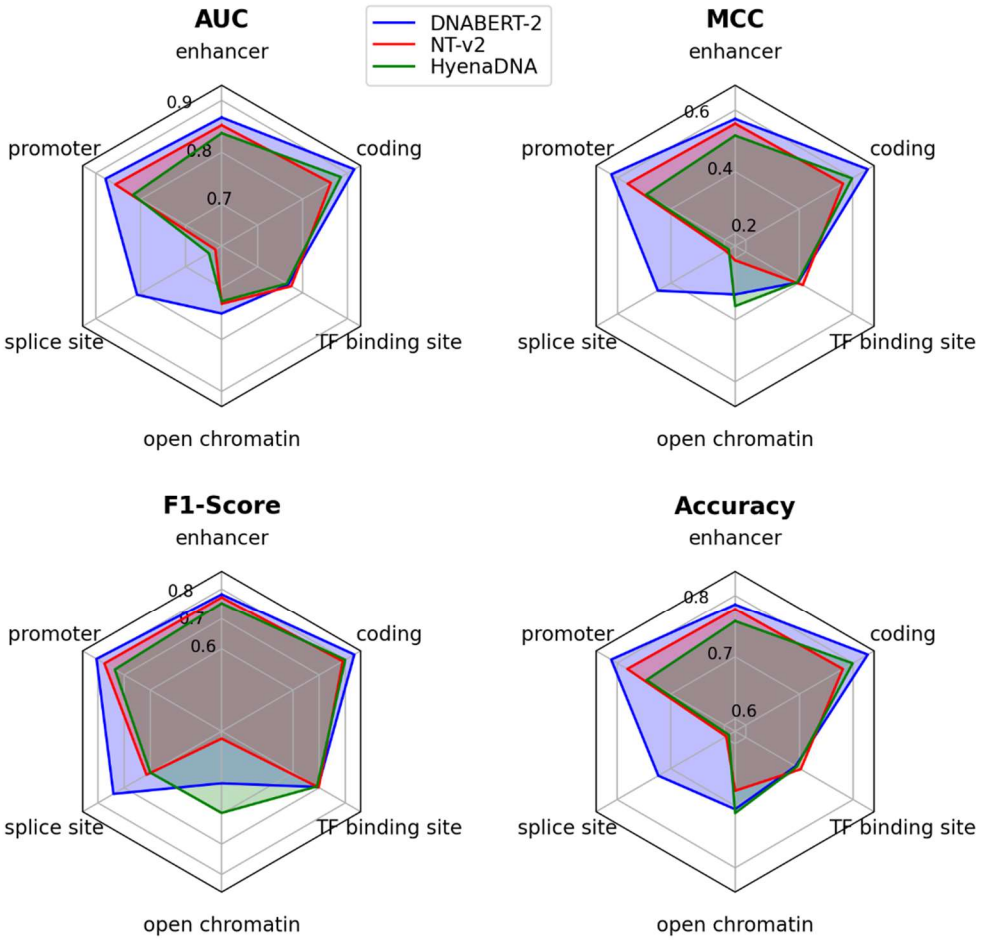
Model Configuration	Number of Parameters	Output Embedding Dimension
DNABERT-2	117M	768
NT-v2-50m	50M	512
NT-v2-100m	100M	512
NT-v2-250m	250M	768
NT-v2-500m	500M	1024
Hyena-tiny-1k	436K	128
Hyena-tiny-1k-d256	436K	256
Hyena-tiny-16k-d128	436K	128
Hyena-small-32k	3M	256
Hyena-medium-160k	6M	256
Hyena-medium-450k	6M	256
Hyena-large-1m	6M	256

Supplementary Table 2: Details of each dataset.

Data	Training Size	Testing Size	Maximum Length	Median Length
4mC_E.coli	8681	3721	41	41
4mC_C.elegans	84926	36398	41	41
4mC_G.pickeringii	24053	10309	41	41
4mC_G.subterraneus	63567	27243	41	41
4mC_D.melanogaster	126466	54200	41	41
4mC_A.thaliana	156697	67157	41	41
DNase_I	711	306	275	243
5-methylcytosin (5mC)	2344	2344	41	41
N6-methyladenosine (6mA)	18336	18334	41	41
Promoter GM12878	10992	2750	2999	1622
Promoter R_capsulatus	7406	3175	40	40
Promoter Arabidopsis_TATA	3063	1313	251	251
Promoter NHEK	8170	2044	2400	400
Promoter Arabidopsis_NonTATA	8267	3543	251	251
HUVEC	11928	2982	2997	1267
Promoter B_amyloliquefaciens	1483	636	40	40
Promoter Hela-S3	11736	2936	2999	1113
Mouse TFBS 1	6478	810	101	101

Mouse TFBS 2	53952	6745	101	101
Mouse TFBS 3	2620	328	101	101
Mouse TFBS 4	1904	239	101	101
Mouse TFBS 5	15064	1883	101	101
Promoter NonTATA 70 bps	42452	5307	300	300
Promoter TATA 70 bps	42452	5307	70	70
Promoter All 70 bps	4904	613	70	70
Promoter NonTATA 300 bps	47356	5920	300	300
Promoter TATA 300 bps	47356	5920	70	70
Promoter All 300 bps	4904	613	300	300
Human TFBS 1	32378	1000	101	101
Human TFBS 2	30672	1000	101	101
Human TFBS 3	19000	1000	101	101
Human TFBS 4	27294	1000	101	101
Human TFBS 5	19000	1000	101	101
Enhancer ensembl	123872	30970	573	269
Promoter NonTATA 251bps	27097	9034	251	251
Human vs worm	75000	25000	200	200
Regulatory region type	150000	57713	802	401
Enhancer cohn	20843	6948	500	500
Coding	75000	25000	200	200
Open chromatin region	139804	34952	593	315
Covid variants	73335	9168	999	999
Donors	19775	2198	600	600
Splice Site Type, DNABERT	36496	4562	400	400
Acceptors	19961	2218	600	600
Splice Site Type, NT	27000	3000	400	400
Enhancer	14968	400	200	199
Enhancer strength	14968	400	200	199
Yeast H3	11971	1497	500	500
Yeast H3K79me3	23069	2884	500	500
Yeast H3K9ac	22224	2779	500	500
Yeast H3K14ac	26438	3305	500	500
Yeast H3K4me3	29439	3680	500	500
Yeast H3K36me3	27904	3488	500	500
Yeast H3K4me2	24545	3069	500	500
Yeast H4	11679	1461	500	500
Yeast H3K4me1	25341	3168	500	500
Yeast H4ac	27275	3410	500	500

Supplementary Figure 1: The radar plots showing the performance of the models on human DNA sequence classification datasets, based on 4 metrics. Due to the vast number of datasets in this category, we averaged the performance scores across datasets of similar kinds of problems. There are six kinds of problems in total: identification of enhancer region, promoter region, splice site, open chromatin region, transcription factor binding site, and coding region.



Supplementary Table 3: Comparison of AUCs of DNABERT-2, NT-v2 and HyenaDNA, all experimented using mean pooling. ***DeLong Test significance < 0.01. Bolded value: DeLong Test significance < 0.05.*

Data	DNABERT-2	NT-v2	HyenaDNA
Promoter GM12878	0.985	0.982	0.976
Promoter HUVEC	**0.99****	0.987	0.982
Promoter Hela-S3	**0.989****	0.984	0.981
Promoter NHEK	**0.95****	0.933	0.927
Promoter NonTATA 251 bps	0.93	0.89	0.932
Promoter NonTATA 70 bps	0.856	**0.867****	0.852
Promoter TATA 70 bps	0.788	**0.803****	0.78
Promoter All 70 bps	0.837	**0.848****	0.837
Promoter NonTATA 300 bps	0.972	0.971	0.963
Promoter TATA 300 bps	0.736	0.753	**0.794****
Promoter All 300 bps	0.936	0.939	0.935
Coding	**0.944****	0.929	0.941
Donor	**0.899****	0.802	0.804
Acceptor	**0.91****	0.815	0.804
Enhancer	0.87	0.865	0.833
Enhancer Cohn	**0.822****	0.789	0.776
Enhancer Ensembl	0.937	**0.939****	0.936
TFBS Data 1	0.834	0.812	0.830
TFBS Data 2	0.9	0.883	0.889
TFBS Data 3	0.803	0.79	0.797
TFBS Data 4	0.724	0.701	0.731
TFBS Data 5	0.866	0.831	0.844
Open chromatin region	**0.724**	0.719	0.719
DNase_I Hypersensitive	0.864	0.853	0.832
Promoter B_amyloliquefaciens	0.851	0.826	0.862
Promoter R_capsulatus	0.687	0.676	**0.712****
Promoter Arabidopsis NonTATA	0.946	0.939	**0.955****
Promoter Arabidopsis TATA	0.951	0.951	**0.961****
Human vs worm	**0.98****	0.978	0.95
Mouse TFBS	0.801	0.766	0.726
5-methylcytosin (5mC)	0.69	**0.723****	0.68
N6-methyladenosine (6mA)	0.744	**0.765****	0.747
A.Thaliana 4mC	0.599	**0.632****	0.593
C.Elegans 4mC	0.596	**0.648****	0.598
D.Melanogaster 4mC	0.614	**0.652****	0.609
E.Coli 4mC	0.548	0.608	0.615
G.Pickeringii 4mC	0.595	0.632	0.631
G.Subterraneus 4mC	0.579	0.614	0.609
Yeast Epigenetic Marks	**0.778**	0.759	0.742

Supplementary Table 4: DNABERT-2 Mean pooling AUCs compared to CLS token pooling. *Bolded value: row maximum.*

Data	Mean Pooling	CLS Token
Promoter GM12878	**0.985**	0.964
Promoter HUVEC	**0.99**	0.974
Promoter Hela-S3	**0.989**	0.971
Promoter NHEK	**0.95**	0.912
Donors	**0.899**	0.823
Acceptors	**0.91**	0.793
Enhancer	**0.87**	0.863
Coding	**0.944**	0.915
Enhancer cohn	**0.822**	0.792
Enhancer ensembl	0.937	**0.947**
Human TFBS 1	**0.866**	0.785
Human TFBS 2	**0.724**	0.66
Human TFBS 3	**0.9**	0.834
Human TFBS 4	**0.879**	0.817
Human TFBS 5	**0.803**	0.744
Promoter NonTATA 251bps	**0.93**	0.861
Open chromatin region	**0.724**	0.685
Promoter NonTATA 70 bps	**0.856**	0.816
Promoter TATA 70 bps	0.788	**0.809**
Promoter All 70 bps	**0.837**	0.803
Promoter NonTATA 300 bps	**0.972**	0.938
Promoter TATA 300 bps	**0.736**	0.698
Promoter All 300 bps	**0.936**	0.897
Promoter B_amyloliquefaciens	0.851	0.856
Promoter R_capsulatus	**0.687**	0.661
Promoter Arabidopsis_NonTATA	**0.946**	0.891
Promoter Arabidopsis_TATA	**0.951**	0.903
Human vs worm	**0.98**	0.946
Mouse TFBS	**0.801**	0.700
5-methylcytosin (5mC)	**0.69**	0.678
N6-methyladenosine (6mA)	**0.744**	0.731
DNase_I	**0.864**	0.815
4mC_A.thaliana	**0.599**	0.59
4mC_C.elegans	**0.596**	0.587
4mC_D.melanogaster	**0.614**	0.604
4mC_E.coli	0.548	**0.567**
4mC_G.pickeringii	**0.595**	0.587
4mC_G.subterraneus	0.579	**0.588**
Yeast Epigenetic Marks	**0.778**	0.734

Supplementary Table 5: NT-v2 Mean pooling AUCs compared to CLS token pooling. *Bolded value: row maximum.*

Data	Mean Pooling	CLS Token
Promoter GM12878	**0.982**	0.878
Promoter HUVEC	**0.987**	0.912
Promoter Hela-S3	**0.984**	0.909
Promoter NHEK	**0.933**	0.855
Donors	0.802	**0.636**
Acceptors	**0.815**	0.632
Enhancer	0.865	**0.879**
Coding	**0.929**	0.863
Enhancer cohn	**0.789**	0.728
Enhancer ensembl	0.939	**0.95**
Human TFBS 1	**0.831**	0.801
Human TFBS 2	**0.701**	0.663
Human TFBS 3	**0.883**	0.836
Human TFBS 4	**0.856**	0.824
Human TFBS 5	**0.79**	0.751
Promoter NonTATA 251bps	**0.89**	0.834
Open chromatin region	**0.719**	0.657
Promoter NonTATA 70 bps	**0.867**	0.838
Promoter TATA 70 bps	0.803	**0.872**
Promoter All 70 bps	**0.848**	0.822
Promoter NonTATA 300 bps	**0.971**	0.91
Promoter TATA 300 bps	**0.753**	0.694
Promoter All 300 bps	**0.939**	0.875
Promoter B_amyloliquefaciens	**0.826**	0.797
Promoter R_capsulatus	**0.676**	0.668
Promoter Arabidopsis_NonTATA	**0.939**	0.85
Promoter Arabidopsis_TATA	**0.951**	0.855
Human vs worm	**0.978**	0.919
Mouse TFBS	**0.766**	0.722
5-methylcytosin (5mC)	**0.723**	0.713
N6-methyladenosine (6mA)	**0.765**	0.752
DNase_I	**0.853**	0.806
4mC_A.thaliana	**0.632**	0.6
4mC_C.elegans	**0.648**	0.594
4mC_D.melanogaster	**0.652**	0.611
4mC_E.coli	**0.608**	0.579
4mC_G.pickeringii	**0.632**	0.607
4mC_G.subterraneus	**0.614**	0.581
Yeast Epigenetic Marks	**0.759**	0.643

Supplementary Table 6: HyenaDNA Mean pooling AUCs compared to EOS token pooling. *Bolded value: row maximum.*

Data	Mean Pooling	EOS Token
Promoter GM12878	**0.976**	0.884
Promoter HUVEC	**0.982**	0.906
Promoter Hela-S3	**0.981**	0.9
Promoter NHEK	**0.927**	0.854
Donors	**0.804**	0.626
Acceptors	**0.804**	0.67
Enhancer	**0.833**	**0.833**
Coding	**0.941**	0.885
Enhancer cohn	**0.776**	0.733
Enhancer ensembl	0.936	**0.944**
Human TFBS 1	**0.844**	0.787
Human TFBS 2	**0.731**	0.624
Human TFBS 3	**0.889**	0.842
Human TFBS 4	**0.888**	0.83
Human TFBS 5	**0.797**	0.741
Promoter NonTATA 251bps	**0.932**	0.853
Open chromatin region	**0.719**	0.665
Promoter NonTATA 70 bps	**0.852**	0.79
Promoter TATA 70 bps	**0.78**	0.732
Promoter All 70 bps	**0.837**	0.769
Promoter NonTATA 300 bps	**0.963**	0.818
Promoter TATA 300 bps	**0.794**	0.717
Promoter All 300 bps	**0.935**	0.797
Promoter B_amyloliquefaciens	**0.862**	0.688
Promoter R_capsulatus	**0.712**	0.602
Promoter Arabidopsis_NonTATA	**0.955**	0.814
Promoter Arabidopsis_TATA	**0.961**	0.82
Human vs worm	**0.95**	0.837
Mouse TFBS	**0.624**	0.726
5-methylcytosin (5mC)	**0.68**	0.604
N6-methyladenosine (6mA)	**0.747**	0.681
DNase_I	**0.832**	0.787
4mC_A.thaliana	**0.593**	0.557
4mC_C.elegans	**0.598**	0.583
4mC_D.melanogaster	**0.609**	0.57
4mC_E.coli	**0.615**	0.579
4mC_G.pickeringii	**0.631**	0.603
4mC_G.subterraneus	**0.609**	0.577
Yeast Epigenetic Marks	**0.742**	0.665