SPACE: Your Genomic Profile Predictor is a Powerful DNA Foundation Model

Zhao Yang [* 1 2 3]   **Jiwei Zhu** [* 1 2 3]   **Bing Su** [1 2 3]

Abstract

Inspired by the success of unsupervised pre-training paradigms, researchers have applied these approaches to DNA pre-training. However, we argue that these approaches alone yield suboptimal results because pure DNA sequences lack sufficient information, since their functions are regulated by genomic profiles like chromatin accessibility. Here, we demonstrate that supervised training for genomic profile prediction serves as a more effective alternative to pure sequence pre-training. Furthermore, considering the multi-species and multi-profile nature of genomic profile prediction, we introduce our **S**pecies-**P**rofile **A**daptive **C**ollaborative **E**xperts (SPACE) that leverages Mixture of Experts (MoE) to better capture the relationships between DNA sequences across different species and genomic profiles, thereby learning more effective DNA representations. Through extensive experiments across various tasks, our model achieves state-of-the-art performance, establishing that DNA models trained with supervised genomic profiles serve as powerful DNA representation learners. The code is available at https://github.com/ZhuJiwei111/SPACE.

1. Introduction

DNA sequences, encoded by four nucleotide bases (`A`, `C`, `G`, and `T`), serve as biology's fundamental language that carries genetic instructions. Understanding the syntax and regulatory grammar of this molecular language is promising for diverse applications, including disease diagnosis (Kernohan & Boycott, 2024; Sermon et al., 2004), drug discovery (Pe-

terson & Liu, 2023; Trajanoska et al., 2023), and protein engineering (Gosai et al., 2024; Yang et al., 2025).

Due to the complexity of DNA sequences, gaining a clear understanding of DNA is not easy. Inspired by the success of unsupervised pre-training paradigms in NLP, such as masked language modeling (Devlin et al., 2019) (MLM) and next-token prediction (Brown et al., 2020) (NTP), several DNA foundation models (DFMs) have recently emerged following similar pre-training approaches. DFMs (Ji et al., 2021; Zhou et al., 2024; Dalla-Torre et al., 2024; Nguyen et al., 2024a;b) aim to learn transferable representations by leveraging large-scale pre-training on massive DNA sequences, which can be adapted for exploring the functions and mechanisms of DNA. Through fine-tuning, these models have demonstrated promising performance in tasks such as regulatory element identification, splice site recognition, and epigenetic modification prediction.

Although DFMs have made some progress, we argue that solely applying unsupervised pre-training techniques to DNA sequences alone cannot learn high-level semantic representations with strong generalization capabilities (Tang et al., 2023; Tang & Koo, 2024). Natural language sequences themselves can express their meaning relatively completely. In contrast, the functional roles of DNA sequences in organisms are far more complex and are regulated by numerous genomic profiles in a cell-type-specific manner (Fu et al., 2025), including various epigenetic modifications (Portela & Esteller, 2010), chromatin accessibility (Tan et al., 2023), and transcription factor binding (Peterson & Liu, 2023). Therefore, without incorporating these additional biological contexts, DFMs pre-trained only on raw DNA sequences may struggle to generalize to diverse cell environments or provide meaningful insights into the underlying biology.

Given that DNA's functional roles are regulated by various biological factors beyond sequence alone, we revisit supervised genomic profile prediction models (GPPMs) as an alternative to unsupervised DFMs for learning DNA sequence representations. These models (Zhou & Troyanskaya, 2015; Kelley et al., 2018; Zhou et al., 2018; Chen et al., 2022; Avsec et al., 2021) are trained to predict experimentally measurable genomic profiles that directly encode regulatory and functional information in a cell-type-specific manner. By learning to map sequences to these biologically meaning-

---

[*]Equal contribution [1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Research on Large Models and Intelligent Governance [3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE. Correspondence to: Bing Su <bingsu@ruc.edu.cn>.

ful profiles that reflect the complex regulatory mechanisms described above, supervised models may capture more functionally relevant representations compared to unsupervised DFMs that rely solely on sequence patterns. Such biological context-aware representations could be better aligned with downstream genomic applications, such as identifying regulatory elements or understanding gene expression patterns. While Dalla-Torre et al. (2024) has demonstrated through preliminary experiments that Enformer (Avsec et al., 2021) can learn effective DNA representations, there has been no systematic study on improving the representation learning capabilities of GPPMs.

However, current GPPMs employ oversimplified architectures, using a shared encoder for DNA sequences from different species and independent prediction heads for different genomic profiles. This design has two major limitations. First, the species-shared encoder fails to capture species-specific characteristics, as regulatory mechanisms and their influences often vary across species (Karollus et al., 2024). These distinct features are crucial for understanding subtle genomic variations and context-dependent expression patterns. Second, genomic profile prediction inherently involves multiple interrelated tasks (Fu et al., 2025), as different profiles influence each other and are often regulated by common mechanisms. The independent prediction heads, however, prevent the model from capturing these cross-profile dependencies and their variations across species.

To effectively model both cross-species and cross-profile relationships, we introduce our **S**pecies-**P**rofile **A**daptive **C**ollaborative **E**xperts (SPACE), which consists of two key components: (1) a species-aware encoder module and (2) a profile-grouped enhancement decoder module, both built upon Mixture of Experts (MoE). The species-aware encoder employs sparse routing to dynamically balance species-specific and shared biological features, while the profile-grouped enhancement decoder uses dual-gated expert weighted aggregation to capture the intricate dependencies between different genomic profiles. This design enables our model to effectively learn both species-specific patterns and shared regulatory mechanisms across profiles.

The major contributions of this paper include:

- We revisit the supervised pre-training paradigm for DNA sequence foundation models through genomic profile prediction as the pre-training objective, demonstrating how function-related biological contextual information can be effectively encoded into the learned representations.

- We propose SPACE, a novel architecture that leverages MoE to better capture the relationships between DNA sequences across different species and genomic

profiles, thereby learning more effective DNA representations compared to current GPPMs, which simply employ a shared encoder and independent prediction heads.

- Through extensive experiments across a wide range of tasks, our SPACE achieves state-of-the-art (SOTA) performance, demonstrating that supervised pre-training for genomic profile prediction serves as a more effective and powerful alternative to pure sequence pre-training.

2. Related Work

Supervised genomic profile models are trained to predict functional genomic profiles from DNA sequences (Kathail et al., 2024). DeepSEA (Zhou & Troyanskaya, 2015) pioneered this paradigm by leveraging convolutional neural networks (CNNs) to extract DNA sequence features for multitask prediction. Subsequent works (Kelley et al., 2018; Zhou et al., 2018; Chen et al., 2022) have continued to advance this direction through either more advanced architectures or larger-scale training data. Enformer (Avsec et al., 2021), widely recognized as the SOTA method, achieved superior prediction performance through a hybrid Transformer-CNN architecture. While these methods primarily focus on *ab initio* prediction of genomic profiles from DNA sequences and directly utilize these profiles for downstream tasks such as variant effect prediction, few studies (Dalla-Torre et al., 2024) have explored whether their intermediate representations capture meaningful biological patterns. Moreover, these models, which typically adopt a shared encoder coupled with independent profile prediction heads, have not thoroughly explored more effective architectural designs that could potentially enhance both prediction performance and representation learning.

Unsupervised DNA foundation models draw from the success of unsupervised pre-training in NLP. DNABERT (Ji et al., 2021) pioneered this approach, maintaining nearly identical training methods to BERT (Devlin et al., 2019) while adapting the tokenization scheme to 6-mers (Celikkanat et al., 2024) for DNA sequences. Subsequent works have continued along this direction, employing either MLM (Zhou et al., 2024; Dalla-Torre et al., 2024; Sanabria et al., 2024) or NTP (Nguyen et al., 2024a;b) as unsupervised training objectives. Although these methods have made effective optimizations in terms of training data, model architectures, and tokenization strategies, they still adhere to the assumption that unsupervised pre-training on pure DNA sequences alone is sufficient for learning effective representations. Moreover, there has been little systematic comparison between these models and genomic profile prediction models in terms of their representation learning capabilities.
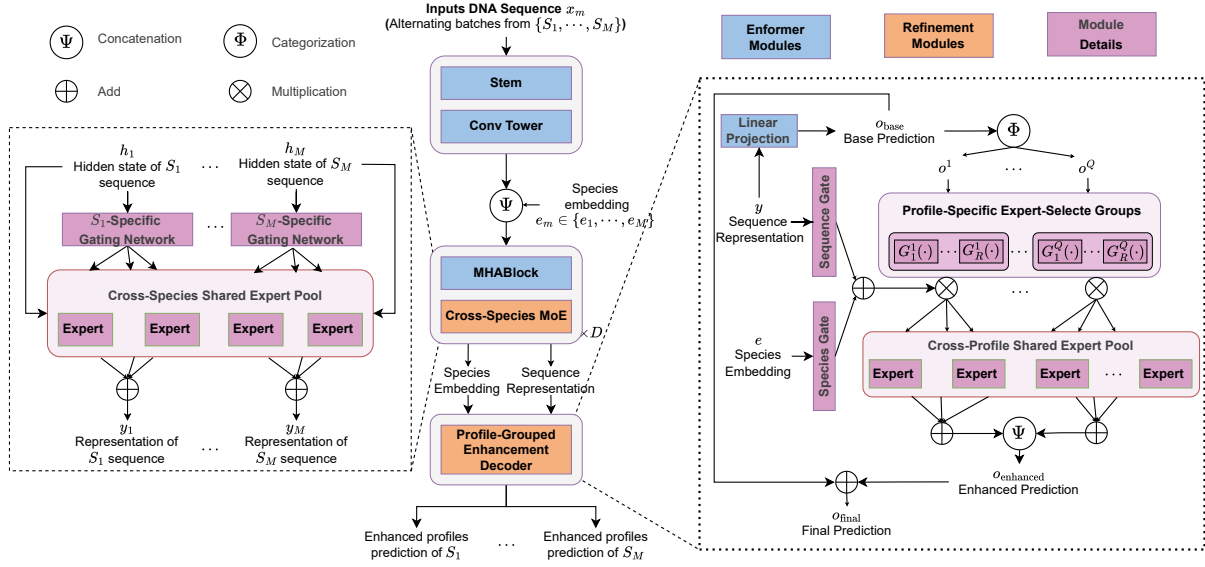
Figure 1. **Overview of our SPACE architecture.** It processes the input DNA sequence with three stages: (1) spatial compression and local context aggregation via a CNN-based aggregation module; (2) latent representation learning via a species-aware sparse MoE-based encoding module; (3) multi-profile prediction decoder via the dual-gated expert weighted prediction enhancement module. The detailed structures of the encoding module and the dual-layer gated prediction enhancement module are shown in the left and right, respectively.

The MoE framework is a conditional computation technique that selectively activates different expert networks for different inputs through sparse routing (Jacobs et al., 1991; Shazeer et al., 2017). In Transformer-based large language models (LLMs), MoE is typically applied to feedforward networks (FFNs) to achieve better parameter efficiency while maintaining model capacity (Fedus et al., 2022; Jiang et al., 2023; Liu et al., 2024). This adaptive routing mechanism is particularly well-suited for our genomic modeling task, as it enables the model to dynamically balance between learning species-specific patterns and shared biological features, while also capturing the complex dependencies between different genomic profiles. Following common practice in Transformer architectures, we also implement MoE by replacing the FFNs in our model.

3. Method

3.1. Problem Formulation

Consider DNA sequences from M species $\{S_1, \ldots, S_M\}$. For each sequence x_m from species S_m, we predict C_m genomic profile values. We train with interleaved batches across all M species to facilitate cross-species knowledge transfer (Kelley, 2020; Avsec et al., 2021). Through this supervised pre-training, the learned representations are expected to capture rich biological and regulatory information.

3.2. Overview

To better capture cross-species and cross-profile representations, we present SPACE. As illustrated in Figure 1, our architecture consists of three key stages: (1) CNN-based Local Context Aggregation following Enformer (Avsec et al., 2021); (2) Species-aware Transformer Encoder and (3) Profile-Grouped Enhancement Decoder for genomic profile prediction.

3.3. Local Context Aggregation

Given an input DNA sequence x_m, we first follow Enformer (Avsec et al., 2021) to compress and aggregate the raw nucleotides through 1D-CNNs, generating hidden states $h_m \in \mathbb{R}^{L \times d_h}$ at 128bp resolution, where L denotes the compressed sequence length and d_h is the hidden dimension.

3.4. Species-aware Encoder

Previous approaches to cross-species modeling (Kelley, 2020; Avsec et al., 2021) typically employ a shared encoder for all species, lacking fine-grained modeling of species relationships. To address this limitation, we propose a novel cross-species modeling framework consisting of Species-specific Embedding and Cross-species MoE layers.

Species-specific embedding. We augment the aggregated

3

hidden states h_m with a trainable species-specific embedding $e_m \in \mathbb{R}^{1 \times d_h}$ by concatenation. The combined representation then passes through D transformer layers with our Sparse Cross-species MoE for further transformation. This design is analogous to the source tokens used in recent language models (Jiang et al., 2023), where document-level embeddings are prepended to provide explicit context about the content source. In our case, the species-specific embedding serves as an explicit signal to guide the model in distinguishing and handling species-specific characteristics. Among DNA models, Karollus et al. (2024) employed a similar approach.

Cross-species MoE. Furthermore, we introduce a sparse MoE encoding module that enables adaptive species-aware representation learning through dynamic parameter routing. For the M species $\{S_1, ..., S_M\}$, each MoE layer consists of two core components: (1) a set of N shared expert networks $\{E_1, ..., E_N\}$, and (2) M species-specific gating networks $\{G_1, ..., G_M\}$, where each G_m is associated with species S_m to dynamically weight expert contributions based on species-specific patterns.

For an aggregated hidden state h_m from species S_m, the output representation y_m is computed as:

$$\hat{h}_m = \text{MHAttention}([e_m, h_m]),$$

$$y_m = \sum_{k=1}^{N} \underbrace{G_m(\hat{h}_m)_k}_{\text{the k-th value of } G_m(\hat{h}_m)} \cdot E_k\left(\hat{h}_m\right), \quad (1)$$

where $e_m \in \mathbb{R}^{d_h}$ denotes the species embedding vector, and $[\cdot]$ represents concatenation, $\hat{h}_m$ is the hidden state after attention. The gating weights are computed through:

$$G_m(\hat{h}_m) = \text{Softmax}\left(\text{TopK}\left(g(\hat{h}_m) + \mathcal{R}_{\text{noise}}\right)\right), \quad (2)$$

where $g(\cdot)$ is the gating function and $\mathcal{R}_{\text{noise}}$ is a noise injection term for training stability enhancement (Fedus et al., 2022). Moreover, to explicitly guide expert networks in learning both conserved and species-specific patterns, we introduce an expert-species mutual information loss inspired by Mod-Squad (Chen et al., 2023). We maximize the mutual information between species identity S and expert selection E to encourage species-specific expert specialization:

$$\mathcal{L}_{\text{MI}} = -MI(S; E) = -H(S) - H(E) + H(S, E)$$
$$= \sum_{i=m}^{M} P(S_m) \log P(S_m) + \sum_{n=1}^{N} P(E_n) \log P(E_n)$$
$$- \sum_{m=1}^{M} \sum_{n=1}^{N} P(S_m, E_n) \log P(S_m, E_n), \quad (3)$$

where S_m denotes the species probability and E_n represents the selection weight of each expert. The detailed derivations are provided in Appendix A.1.

After the encoding stage, we obtain the sequence representation $y \in \mathbb{R}^{L \times d_h}$ that captures both species-specific and shared biological features. Note that the sequence length remains L as we exclude the species embedding token from the final representation, retaining only the original sequence positions for downstream genomic profile prediction.

3.5. Profile-grouped Enhancement Decoder

Current GPPMs treat profile prediction as independent multitasks, overlooking inherent relationships between genomic profiles. This limitation motivates our decoder design based on the following biological principles:

Biological principles for profile relationships. Genomic profile prediction should account for two key biological relationships: **(P1)** *Evolutionary conservation*: Shared regulatory mechanisms exist across homologous profiles in different species (Schmidt et al., 2010). **(P2)** *Functional interdependencies*: Different genomic profiles often share regulatory mechanisms and exhibit mutual influences, e.g., regions with high gene expression typically exhibit increased chromatin accessibility (Fu et al., 2025).

To leverage these biological insights, we propose a dual-layer gated prediction enhancement module that enables systematic knowledge sharing across profiles. For clarity, we present the formulation for a single species S_m and omit the subscript m in subsequent notation.

Genomic profiles can be categorized based on their experimental assays: for instance, DNase and ATAC-seq measure chromatin accessibility, while CAGE quantifies gene expression levels. Profiles from the same experimental type typically share similar functional mechanisms, enabling knowledge transfer within each category. Given Q distinct profile types $\{T_1, ..., T_Q\}$ with specific biological interpretations, for the DNA sequence representation $y \in \mathbb{R}^{L \times d_h}$ and the species embedding $e \in \mathbb{R}^{d_h}$, the enhancement module operates through the following sequential steps.

Profile categorization for initial predictions. We first perform a linear projection on y to obtain the initial base prediction o_{base}, which represents the final profile predictions from previous GPPMs (Kelley, 2020; Avsec et al., 2021) that do not incorporate biological insights. Based on biological priors, o_{base} is categorized into Q independent parts $\{o^1, ..., o^Q\}$, as follows.

$$o_{\text{base}} = (\text{Linear}(y))^T \in \mathbb{R}^{d_{\text{out}} \times L},$$
$$\{o^1, ..., o^Q\} = \Phi(o_{\text{base}}), \quad (4)$$

where d_{out} denotes the dimension specifying the total number of genomic profiles (i.e., d_{out} equals C_m for species S_m). The category operator $\Phi(\cdot)$ is constructed based on domain-specific biological knowledge, which decomposes the base prediction into Q profile types $\{o^q\}_{q=1}^{Q}$ where $o^q \in \mathbb{R}^{d_q \times L}$

corresponds to biological profile type T_q, with d_q indicating the number of profiles categorized to T_q.

Dual-gated expert weighted aggregation. Each dimension of o^q represents the base predicted sequence for a specific profile track. To capture shared regulatory patterns, we employ K cross-profile-type shared experts $\{E_k\}_{k=1}^K$, where each expert $E_k : \mathbb{R}^{d_q \times L} \to \mathbb{R}^{d_q \times L}$ enhances the categorized base prediction o^q.

For adaptive expert selection, we introduce a dual-gated mechanism with two complementary components. **The first layer (group-level gating)** addresses principle **(P1)** by dynamically assigning weights to R expert groups based on species and sequence context, capturing evolutionary conservation patterns. The group weights $\hat{G}^q$ are computed as:

$$\hat{G}^q = \text{Softmax}\left(G^q_{\text{species}}(e) + G^q_{\text{sequence}}(\text{Pool}(y)) \right), \quad (5)$$

where $G^q_{\text{species}}(\cdot)$ and $G^q_{\text{sequence}}(\cdot)$ map $\mathbb{R}^{d_h} \to \mathbb{R}^R$, weighting the R expert-selected groups from species and sequence perspectives, respectively. Here, $\text{Pool}(\cdot)$ denotes global average pooling applied along the sequence length L and q indicates the profile type. **The second layer (expert-level gating)** addresses principle **(P2)** by selecting specific experts based on profile prediction patterns, capturing functional interdependencies between profiles. For each group r and profile type q, we define the expert selection function $G^q_r(o^q) \in \mathbb{R}^K$, where $G^q_r(o^q)_k$ represents the weight of the k-th expert within the r-th group, determined by the base prediction patterns. The computation of G^q_r follows a similar gating mechanism as in Equation 2.

Combining both gating layers, the enhanced prediction for profile type T_q is formulated as:

$$o^q_{\text{enhanced}} = \sum_{r=1}^R \underbrace{\hat{G}^q_r}_{\text{Group weight}} \cdot \left(\sum_{k=1}^K \underbrace{G^q_r(o^q)_k}_{\text{Expert weight}} \cdot E_k(o^q) \right), \quad (6)$$

where the dual-gated mechanism first weights relevant expert groups through $\hat{G}^q_r$ (based on species and sequence context), then within each selected group, chooses appropriate experts through $G^q_r(o^q)_k$ (based on prediction patterns).

The final predictions are computed through connections between enhanced and base predictions:

$$o_{\text{enhanced}} = \Psi\left(\{o^1_{\text{enhanced}}, ..., o^Q_{\text{enhanced}}\} \right), \quad (7)$$
$$o_{\text{final}} = o_{\text{base}} + o^T_{\text{enhanced}},$$

where $\Psi(\cdot)$ is the inverse operator of $\Phi(\cdot)$, concatenating the enhanced predictions across different profile types, and the residual connection ensures stable training.

3.6. Training Objective

Following Enformer (Avsec et al., 2021), we adopt the Poisson negative log-likelihood as the primary loss function. To further refine species-aware expert selection in Section 3.4 by maximizing mutual information between species proportion and expert activations, we introduce an auxiliary mutual information loss. The composite loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Poisson}} - \alpha \sum_{d=1}^D MI(S; E_d), \quad (8)$$

where $\alpha = 0.01$ controls the mutual information regularization strength, D denotes the number of transformer layers, S represents the species identifier, and E_d indicates the shared expert pool at layer d, the Poisson loss $\mathcal{L}_{\text{Poisson}}$ is mathematically formulated in Appendix A.2.

4. Experiments

4.1. Experiment Setup

Pre-training dataset. The pre-training dataset aligns with that used in Enformer (Kelley, 2020; Avsec et al., 2021), containing DNA sequences and corresponding genomic profiles for human and mouse genomes. Both species shared four conserved profile types: chromatin accessibility (DNase/ATAC-seq), transcription factor binding (TF ChIP-seq), histone modifications (Histone ChIP-seq), and transcriptional activity (CAGE). The number of profiles varies among different profile types in different species, with detailed dataset specifications provided in Appendix B.

Implementation details. Our model was pre-trained using supervised genomic profile prediction, maintaining the same prediction targets and genomic intervals as implemented in Enformer (Avsec et al., 2021). For cross-species joint modeling, we implemented an alternating training strategy using eight NVIDIA A40 GPUs. Training proceeded for 50,000 steps (approximately 8 days) with a global batch size of 64, achieved through 8 gradient accumulation steps (1 sample per GPU). Optimization employed AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 5×10^{-4}, linearly ramped from 0 during the first 5,000 steps followed by cosine decay. Gradient norms were clipped at 0.2 to maintain stability.

4.2. Nucleotide Transformer Downstream Tasks

We conducted rigorous benchmarking against the suite of 18 genomic datasets established in NT (Dalla-Torre et al., 2024), encompassing three fundamental task categories: (1) histone modification marker prediction, (2) cis-regulatory element annotation, and (3) splice site recognition. Following the evaluation protocol from NT, we employed the Matthews Correlation Coefficient (MCC)

Table 1. **MCC performance of Nucleotide Transformer downstream tasks.** This benchmark includes three categories of downstream tasks, comprising a total of 18 datasets derived from human samples. The term 'NT downstream tasks' will be used to refer to these tasks.

Model	Chromatin profiles					
	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
DNABERT-2	0.490 ± 0.013	0.491 ± 0.010	0.599 ± 0.010	**0.637 ± 0.007**	0.490 ± 0.008	0.558 ± 0.013
HyenaDNA-32KB	0.467 ± 0.012	0.421 ± 0.010	0.550 ± 0.009	0.553 ± 0.011	0.423 ± 0.016	0.515 ± 0.018
NT-HumanRef (500M)	0.465 ± 0.011	0.457 ± 0.010	0.589 ± 0.009	0.594 ± 0.004	0.468 ± 0.007	0.527 ± 0.011
NT-1000G (500M)	0.464 ± 0.012	0.458 ± 0.012	0.591 ± 0.007	0.581 ± 0.009	0.466 ± 0.006	0.528 ± 0.011
NT-1000G (2.5B)	0.478 ± 0.012	0.486 ± 0.023	**0.603 ± 0.009**	0.632 ± 0.008	0.491 ± 0.015	0.569 ± 0.014
NT-Multispecies (2.5B)	0.503 ± 0.010	0.481 ± 0.020	0.593 ± 0.016	0.635 ± 0.016	0.481 ± 0.012	0.552 ± 0.022
Enformer	0.522 ± 0.019	0.520 ± 0.015	0.552 ± 0.007	0.567 ± 0.017	0.504 ± 0.021	0.626 ± 0.015
SPACE	**0.548 ± 0.005**	**0.547 ± 0.007**	0.586 ± 0.010	0.602 ± 0.005	**0.543 ± 0.009**	**0.640 ± 0.007**

Model	Chromatin profiles				Regulatory elements	
	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Enhancers	Enhancers(types)
DNABERT-2	0.646 ± 0.008	0.564 ± 0.013	0.443 ± 0.025	0.655 ± 0.011	0.517 ± 0.011	0.476 ± 0.009
HyenaDNA-32KB	0.603 ± 0.020	0.487 ± 0.025	0.419 ± 0.030	0.590 ± 0.007	0.476 ± 0.021	0.445 ± 0.009
NT-HumanRef (500M)	0.622 ± 0.013	0.524 ± 0.013	0.433 ± 0.009	0.634 ± 0.013	0.515 ± 0.019	0.477 ± 0.014
NT-1000G (500M)	0.609 ± 0.011	0.515 ± 0.018	0.415 ± 0.019	0.634 ± 0.010	0.505 ± 0.009	0.459 ± 0.011
NT-1000G (2.5B)	0.615 ± 0.017	0.529 ± 0.012	0.483 ± 0.013	**0.659 ± 0.008**	0.504 ± 0.009	0.469 ± 0.005
NT-Multispecies (2.5B)	0.618 ± 0.015	0.527 ± 0.017	0.447 ± 0.018	0.650 ± 0.014	0.527 ± 0.012	0.484 ± 0.012
Enformer	0.635 ± 0.019	0.593 ± 0.020	0.453 ± 0.016	0.606 ± 0.016	0.614 ± 0.010	0.573 ± 0.013
SPACE	**0.661 ± 0.025**	**0.635 ± 0.016**	**0.490 ± 0.011**	0.650 ± 0.011	**0.631 ± 0.007**	**0.583 ± 0.008**

Model	Regulatory elements			Splicing		
	All	NoTATA	TATA	Donors	Acceptors	All
DNABERT-2	0.754 ± 0.009	0.769 ± 0.009	0.784 ± 0.036	0.837 ± 0.006	0.855 ± 0.005	0.861 ± 0.004
HyenaDNA-32KB	0.698 ± 0.011	0.729 ± 0.009	0.666 ± 0.041	0.808 ± 0.009	0.907 ± 0.018	0.915 ± 0.047
NT-HumanRef (500M)	0.734 ± 0.013	0.738 ± 0.008	0.831 ± 0.022	0.941 ± 0.004	0.939 ± 0.003	0.952 ± 0.003
NT-1000G (500M)	0.727 ± 0.004	0.743 ± 0.012	0.855 ± 0.041	0.933 ± 0.007	0.939 ± 0.004	0.952 ± 0.004
NT-1000G (2.5B)	0.708 ± 0.008	0.758 ± 0.007	0.802 ± 0.030	0.952 ± 0.004	0.956 ± 0.004	0.963 ± 0.001
NT-Multispecies (2.5B)	0.761 ± 0.009	0.773 ± 0.010	**0.944 ± 0.016**	**0.958 ± 0.003**	**0.964 ± 0.003**	**0.970 ± 0.002**
Enformer	0.745 ± 0.012	0.763 ± 0.012	0.793 ± 0.026	0.749 ± 0.007	0.739 ± 0.011	0.780 ± 0.007
SPACE	**0.764 ± 0.012**	**0.776 ± 0.011**	0.838 ± 0.028	0.942 ± 0.006	0.902 ± 0.004	0.906 ± 0.003

as the primary performance metric across all tasks to ensure methodological consistency. The formal definition of MCC, along with its theoretical properties, is comprehensively detailed in Appendix A.3. Our comparative analysis includes both unsupervised pre-training approaches (DNABERT (Ji et al., 2021), DNABERT2 (Zhou et al., 2024), and NT (Dalla-Torre et al., 2024)) and supervised baseline (Enformer (Avsec et al., 2021)). In alignment with NT's methodology, we implemented 10-fold cross-validation with fixed random seeds (0-9) and early stopping based on validation performance. All benchmark performance metrics for the compared models in downstream tasks are directly sourced from the original experimental results reported in NT, ensuring consistent evaluation protocols and dataset configurations. As detailed in Table 1, our model achieves SOTA performance on 11 out of 18 predic-

tion tasks. Notably, this superior performance persists even when compared to the parameter-intensive NT-Multispecies variant (2.5B parameters), demonstrating that our supervised pre-training paradigm enables the acquisition of more robust DNA sequence representations. Moreover, our architectural improvements consistently outperform Enformer's original implementation across all tasks, empirically confirming the effectiveness of our modules. The specific details and complete results of the tasks are presented in Appendix C.

4.3. Cross-species Validation on GUE Benchmark

To rigorously evaluate the cross-species generalization capacity of our architectural refinements to Enformer, we employed the GUE benchmark (Zhou et al., 2024). While the benchmark encompasses 7 prediction tasks across 4 taxonomic groups, our experimental design strategically pri-

Table 2. **Comparison Results with Enformer on the GUE Benchmark.** MCC was used in EMP tasks, while F1-score was employed in the CVC task. The upward arrow (↑) denotes SPACE's relative improvement over Enformer in novel species.

Model	Epigenetic Marks Prediction				
	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2
Enformer	70.65	37.87	42.41	34.00	29.65
SPACE	79.53 (↑ 8.88)	54.12 (↑ 16.25)	54.82 (↑ 12.41)	50.92(↑ 16.92)	43.80 (↑ 14.15)

Model	Epigenetic Marks Prediction					Virus
	H3K4me3	H3K79me3	H3K9ac	H4	H4ac	Covid
Enformer	22.19	55.69	49.35	76.32	32.90	61.33
SPACE	49.47 (↑ 27.28)	66.93 (↑ 11.24)	59.29 (↑ 9.94)	81.25 (↑ 4.93)	53.09 (↑ 20.19)	70.26 (↑ 8.93)

oritizes yeast and viral genomic contexts — evolutionarily distant lineages characterized by marked nucleotide-level divergence from the mammalian species used during model training. These phylogenetically distinct evaluations include Epigenetic Mark Prediction (EMP) across 10 yeast datasets and COVID Variants Classification (CVC) in viral genomes. Following the evaluation protocol established in DNABERT2 (Zhou et al., 2024), we adopted the MCC for EMP and the F1-score for CVC.

As quantified in Table 2, our refined architecture demonstrates significant improvements over the original Enformer in these tasks. This systematic evaluation provides empirical evidence that our architectural modifications enhance cross-species generalization capabilities, particularly in identifying evolutionarily conserved regulatory features compared to Enformer's baseline implementation. Cross-architecture benchmarking against DNABERT2 and other established baselines (Appendix D) confirms the universality of these improvements, with non-Enformer baseline results rigorously reproduced from DNABERT2's original experimental protocol to ensure methodological consistency. All evaluations strictly adhered to benchmark specifications, including standardized train-test splits and hyperparameter configurations, to maintain reproducibility and fairness.

4.4. Genomic Benchmarks

To further validate the capabilities of our model, we performed extended benchmarking using the Genomic Benchmarks (Grešová et al., 2023) dataset, which represents the only mainstream benchmark encompassing species beyond those investigated in our previous experiments, including Human-or-worm classification and Drosophila enhancer classification. Following a methodology similar to Caduceus (Schiff et al., 2024), we evaluated Enformer and SPACE, adopting the baseline model results reported in that paper. It is worth noting that Caduceus did not measure the enhancer prediction task for Drosophila melanogaster, so

we referenced the CNN results from Genomic Benchmarks. The results are presented in the Table 3.

4.5. Analysis of the MoE architecture

Analysis of the species-aware encoder module. As shown in Figure 2a, this part reveals the implicit species-characteristic learning mechanism through visual analysis of expert selection frequencies in the final Transformer layer from the Enformer test dataset. In our experiments, we adopt a 4-expert architecture with a top-3 selection mechanism ($k = 3$). Quantitative analysis demonstrates significant functional differentiation in the expert system: Expert 1 and Expert 3 exhibit species-specific learning capabilities for single species (human and mouse), while Expert 0 and Expert 2 primarily participate in cross-species conserved feature extraction. This hierarchical division successfully achieves spatial decoupling of species-specific representations and evolutionarily conserved features, providing an interpretable solution for multi-species joint modeling.

Analysis of the profile-grouped enhancement decoder module. In our experiments, we employed 8 cross-profile-type shared experts ($K = 8$) with 2 expert-selected groups per profile type ($R = 2$), where each group dynamically integrated the top 3 most contributory experts through a dual-gated expert weighted aggregation. Through hierarchical weighting, we derived final expert selection probabilities. Normalized expert selection frequencies across Enformer test datasets were systematically quantified and visualized in Figure 2b, revealing distinct biological patterns.

Notably, TF binding (TF ChIP-seq) and histone modifications (Histone ChIP-seq) profiles exhibited high expert specialization, reflecting their biologically inherent complexity: TF binding involves combinatorial interactions among diverse transcription factor families, while histone modifications require interpretation of multilayered epigenetic codes governed by cooperative post-translational modifications. Conversely, Chromatin accessibility (DNase/ATAC-

Table 3. **The results on the Genomic Benchmarks datasets.** The results for Drosophila enhancers were obtained from Genomic Benchmarks (Grešová et al., 2023), while all other baseline results were sourced from the Caduceus model (Schiff et al., 2024).

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

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

seq signals) and transcription initiation (CAGE signals) profiles showed substantial expert overlap with differential weighting, mirroring their mechanistic and positional interdependence. Chromatin accessibility creates permissive 3D topological environments essential for transcription initiation, with transcription start site (TSS)-associated promoter and enhancer regions exhibiting substantial spatial overlap with chromatin accessibility domains. This "functional dependency-spatial coupling" drives the prediction enhancement decoder module to develop coordinated feature extraction strategies. These findings collectively demonstrate that divergent genomic profiles converge on shared regulatory architectures and engage in reciprocal regulatory interactions – a phenomenon that fundamentally informed our development of the profile-grouped enhancement decoder to systematically leverage such interconnectivity.

4.6. Comparative Analysis with Enformer in Gene Expression Prediction

We conducted a comparative analysis based on the core task of the baseline Enformer model, which aims to predict human and mouse genomic profiles at 128-bp resolution from input DNA sequences. We computed the average Pearson correlation coefficients across all positions for genomic profiles in the test set and performed stratified visualization by species and profile types, as illustrated in Figure 2c. The results demonstrate that our approach significantly enhances the prediction accuracy for mouse genomic profiles while maintaining the prediction performance for human genomic profiles. It should be noted that our training steps were only

1/3 of those used by Enformer, yet we were still able to achieve good results on the training objective tasks.

4.7. Ablation Study

We conducted ablation experiments with a half-scale model (hidden_dim=768) on five configurations: (1) baseline without the prediction-enhanced decoder, (2) decoder replacement with a parameter-matched MLP, (3) substitution of MoE layers with standard FFNs in the encoder, (4) additional removal of species embeddings from configuration (3), and (5) our complete dual-module architecture. The results are presented in the Table 4. SPACE demonstrates superior performance across most tasks, with the notable exception of the TATA box dataset (see Table 14) – due to its exclusive focus on simple sequence motifs rather than complex regulatory mechanisms. This indicates that while our decoder doesn't directly boost chromatin profile prediction accuracy, the MoE architecture implicitly models cross-profile regulatory dependencies, offering significant advantages for tasks requiring integrated profile understanding. Cross-species evaluation on the GUE benchmark (yeast and virus tasks, detailed in Table 15) further demonstrates that the MLP-based decoder variant exhibits substantially weaker generalization to new species compared to SPACE's enhancement decoder architecture.

5. Limitations

This work has limitations in data coverage and model scale. First, SPACE has only been trained on two species. While
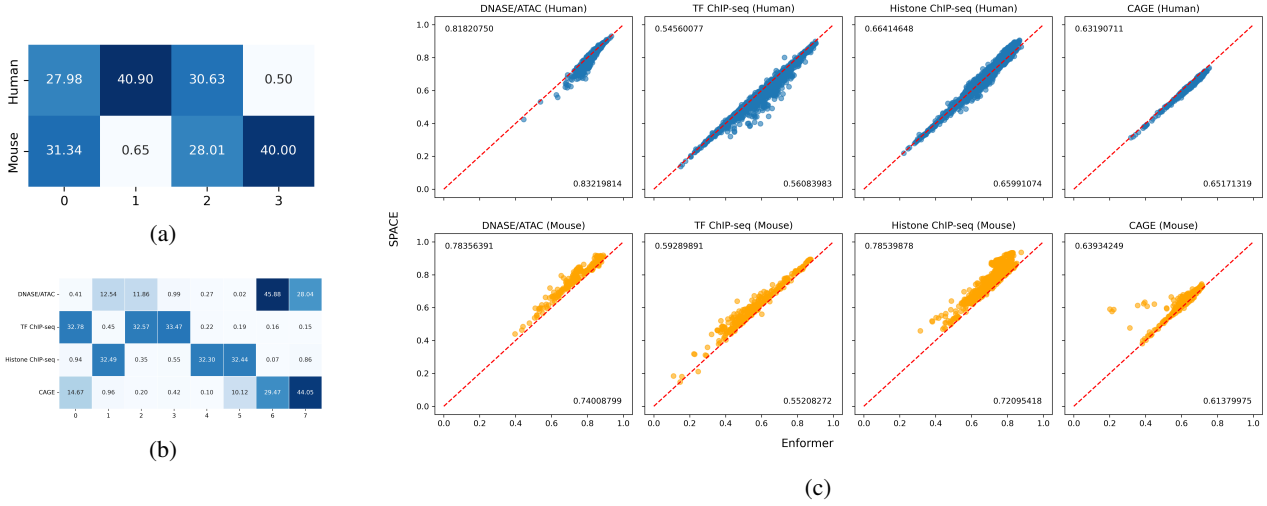
Figure 2. **Expert selection visualizations and prediction results.** (a) Visualization of expert selection in the final cross-species MoE. (b) Expert selection in the profile-grouped enhancement decoder module. (c) Pearson correlation coefficients across all positions per profile on the test set. Each point represents the average correlation of predicted genomic profiles across all 128-bp binned genomic positions.

Table 4. **Ablation Studies on NT downstream tasks and GUE benchmarks.** The results include the average outcomes of the three major categories of downstream tasks in NT and the average results of the EMP tasks and the CVC task in the GUE benchmark experiments.

Model	NT			GUE	
	Chromatin	Regulatory	Splicing	EMP	CVC
SPACE w/o decoder	0.5674	0.7054	0.8977	0.5339	0.6866
SPACE w/o decoder w/ MLP	0.5651	0.6920	0.9020	0.5153	0.6783
SPACE w/o encoder	0.5653	0.7022	0.8887	0.5346	0.6846
SPACE w/o encoder w/o species emb	0.5692	0.6986	0.8957	0.5322	0.6856
SPACE	0.5705	0.7024	0.9077	0.5368	0.6889

this study demonstrates the advantages of our design, we believe extending to more species would yield benefits as additional data becomes available (Vandereyken et al., 2023). Second, constrained by computational resources, our model (588M, sparse-activated) is significantly smaller than the largest variant of NT (2.5B, dense) (the detail is in Appendix H. Given the observed scaling laws in DFMs (Dalla-Torre et al., 2024; Nguyen et al., 2024a), we believe increasing our model scale would lead to improvements.

6. Conclusion

In this work, we demonstrate that supervised pre-training via genomic profile prediction serves as a moreeffective and targeted alternative to pure sequence pre-training for DNA foundation models. To distinguish unique characteristics of different species and profiles while fully utilizing the transferrable knowledge among species and profiles, we introduce SPACE that provides biological insights into the model architecture. Through extensive evaluations, we show that our supervised pre-training with the proposed SPACE sets a new standard for DNA representation learning, paving the way for future developments in DFMs.

Impact Statement

This paper presents work whose goal is to advance the field at the intersection of Machine Learning and Genomics. There are many potential societal consequences of our work in this interdisciplinary area, none of which we feel must be specifically highlighted here.

Acknowledgements

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Benegas, G., Albors, C., Aw, A. J., Ye, C., and Song, Y. S. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, pp. 1–6, 2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Celikkanat, A., Masegosa, A. R., and Nielsen, T. D. Revisiting k-mer profile for effective and scalable genome representation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Chen, K. M., Wong, A. K., Troyanskaya, O. G., and Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, 54(7): 940–949, 2022.

Chen, Z., Shen, Y., Ding, M., Chen, Z., Zhao, H., Learned-Miller, E. G., and Gan, C. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11828–11837, 2023.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Fu, X., Mo, S., Buendia, A., Laurent, A. P., Shao, A., Alvarez-Torres, M. d. M., Yu, T., Tan, J., Su, J., Sagatelian, R., et al. A foundation model of transcription across human cell types. *Nature*, pp. 1–9, 2025.

Gosai, S. J., Castro, R. I., Fuentes, N., Butts, J. C., Mouri, K., Alasoadura, M., Kales, S., Nguyen, T. T. L., Noche, R. R., Rao, A. S., et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, pp. 1–10, 2024.

Grešová, K., Martinek, V., Čechák, D., Šimeček, P., and Alexiou, P. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Karollus, A., Hingerl, J., Gankin, D., Grosshauser, M., Klemon, K., and Gagneur, J. Species-aware dna language models capture regulatory elements and their evolution. *Genome Biology*, 25(1):83, 2024.

Kathail, P., Bajwa, A., and Ioannidis, N. M. Leveraging genomic deep learning models for non-coding variant effect prediction. *arXiv preprint arXiv:2411.11158*, 2024.

Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS computational biology*, 16(7):e1008050, 2020.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Kernohan, K. D. and Boycott, K. M. The expanding diagnostic toolbox for rare genetic diseases. *Nature Reviews Genetics*, 25(6):401–415, 2024.

Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Marin, F. I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., and Boomsma, W. Bend: Benchmarking dna language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024a.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024b.

Peterson, A. A. and Liu, D. R. Small-molecule discovery through dna-encoded libraries. *Nature Reviews Drug Discovery*, 22(9):699–722, 2023.

Portela, A. and Esteller, M. Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057–1068, 2010.

Sanabria, M., Hirsch, J., Joubert, P. M., and Poetsch, A. R. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8): 911–923, 2024.

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, 2010.

Sermon, K., Van Steirteghem, A., and Liebaers, I. Preimplantation genetic diagnosis. *The Lancet*, 363(9421): 1633–1641, 2004.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., Thandapani, P., Skok, J., Aifantis, I., Fenyö, D., et al. Cell-type-specific prediction of 3d chromatin organization enables high-throughput in silico genetic screening. *Nature biotechnology*, 41(8):1140–1150, 2023.

Tang, Z. and Koo, P. K. Evaluating the representational power of pre-trained dna language models for regulatory genomics, march 2024, 2024.

Tang, Z., Toneyan, S., and Koo, P. K. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nature Genetics*, 55(12):2021–2022, 2023.

Trajanoska, K., Bhérer, C., Taliun, D., Zhou, S., Richards, J. B., and Mooser, V. From target discovery to clinical drug development with human genetics. *Nature*, 620 (7975):737–745, 2023.

Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8):494–515, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yang, Z., Su, B., Cao, C., and Wen, J.-R. Regulatory DNA sequence design with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024.

A. Derivation of Mathematical Formulations for Key Functions

A.1. Mutual Information Analysis

The Mutual Information defined in Equation (3) is:

$$
\begin{aligned}
\mathcal{L}_{\text{MI}} = -MI(S; E) &= -H(S) - H(E) + H(S, E) \\
&= \sum_{i=m}^{M} P(S_m) \log P(S_m) + \sum_{n=1}^{N} P(E_n) \log P(E_n) \\
&\quad - \sum_{m=1}^{M} \sum_{n=1}^{N} P(S_m, E_n) \log P(S_m, E_n),
\end{aligned}
$$

where S_m denotes the species probability and E_n represents the selection weight of each expert.

We split the formulae to analyse them separately. The mutual information decomposition exhibits three fundamental components:

Species Entropy:

$$
-\sum_{i=1}^{M} P(S_i) \log P(S_i) = H(S).
$$

This term represents the inherent diversity of species distribution in training data. As $P(S_i)$ constitutes a fixed prior, $H(S)$ remains constant during optimization.

Expert Diversity Regularization:

$$
-\sum_{j=1}^{N} P(E_j) \log P(E_j) = H(E).
$$

Maximizing this entropy term encourages balanced utilization of experts, preventing expert collapse where few experts dominate computations. Formally, this ensures:

$$
\lim_{H(E) \to \log N} P(E_j) = \frac{1}{N} \quad \forall j.
$$

Conditional Specialization Objective:

$$
\sum_{i=1}^{M} \sum_{j=1}^{N} P(S_i, E_j) \log P(S_i, E_j) = -H(S, E).
$$

Minimizing this joint entropy (equivalent to maximizing $-H(S, E)$) sharpens the conditional distribution $P(E_j|S_i)$, thereby promoting:

$$
\lim_{H(S,E) \to 0} P(E_j|S_i) = \begin{cases} 1 & \text{if } j = \arg\max_k G_k^{S_i}(x) \\ 0 & \text{otherwise} \end{cases}.
$$

This objective ensures that, for a given species, the model preferentially activates a fixed subset of k experts.

In this way, the sparse MoE-based encoding module encourages different expert combinations to handle different species, while some shared experts in the pool can capture common knowledge across species.

A.2. Poisson negative log-likelihood

The Poisson negative log-likelihood function is defined as

$$
\mathcal{L}_{\text{Poisson}} = \frac{1}{N} \sum_{i=1}^{N} (p_i - t_i \ln p_i),
$$

whree p denotes the prediction vector and t represents the target vector.

A.3. Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a statistically rigorous metric for evaluating classification models. Its definition and generalization to multi-class problems are formally outlined below.

Binary Classification Case For binary classification, let TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. The MCC is defined as:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Here, TP, TN, FP, and FN correspond to entries in the confusion matrix for two classes.

Multi-class Classification Case

For K-class classification ($K \geq 2$), let C be the $K \times K$ confusion matrix, where C_{ij} represents the number of samples from class i predicted as class j. The MCC generalizes to:

$$\text{MCC} = \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{m=1}^{K} C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\left(\sum_{k=1}^{K} \sum_{l=1}^{K} C_{kl} \sum_{\substack{m=1 \\ m \neq k}}^{K} C_{mk}\right)\left(\sum_{k=1}^{K} \sum_{l=1}^{K} C_{lk} \sum_{\substack{m=1 \\ m \neq k}}^{K} C_{km}\right)}}.$$

This formulation quantifies the covariance between all class pairs, ensuring robustness to imbalanced data distributions.

The MCC ranges in $[-1, 1]$, where 1, 0, and -1 correspond to perfect prediction, random guessing, and total disagreement, respectively.

B. Pre-training Dataset

Table 5. Genomic Dataset Statistics

Species	Train	Val	Test	Sequence Length
Human	34,021	2,213	1,937	131,072 bp
Mouse	29,295	2,209	2,017	131,072 bp

Our model was pretrained on the same dataset as Enformer (Avsec et al., 2021), with detailed composition statistics provided in Table 5. To address the pronounced species imbalance between human and mouse genomic data, we implemented balanced batch sampling through randomized minority-class augmentation, ensuring equal representation of both species in every batch. This strategy mitigates species bias while preserving sequence diversity through stochastic resampling.

The dataset comprises DNA sequences paired with genomic profiles as prediction targets. These genomic profiles are categorized into four functional classes: chromatin accessibility (DNase/ATAC-seq), transcription factor binding (TF ChIP-seq), histone modifications (Histone ChIP-seq), and transcriptional activity (CAGE). The species-specific distribution of profile types is quantified in Table 6, which details the number of available tracks per category for each organism.

Table 6. Distribution of Genomics profiles

species	DNase/ATA	TF ChIP	Histone ChIP	CAGE	Total
Human	684	2131	1860	638	5313
Mouse	228	308	750	357	1643

C. Nucleotide Transformer Downstream Tasks Revised

C.1. Datasets

The benchmark dataset comprises 18 downstream tasks originally proposed in Nucleotide Transformer (NT) (Dalla-Torre et al., 2024), accessible via `https://huggingface.co/datasets/InstaDeepAI/nucleotide_transformer_downstream_tasks_revised`. These tasks establish a unified genomics benchmarking framework encompassing both binary and multi-class classification challenges. All data is exclusively derived from human samples, organized into three biologically meaningful categories: Chromatin Profiles, Regulatory Elements and Splicing. The complete dataset composition, including sequence numbers, class distributions and sequence length statistics, is detailed in Table 7.

Table 7. Details of the NT downstream tasks

Task	Number of train sequences	Number of test sequences	Number of labels	Sequence length
promoter_all	30,000	1,584	2	300
promoter_tata	5,062	212	2	300
promoter_no_tata	30,000	1,372	2	300
enhancers	30,000	3,000	2	400
enhancers_types	30,000	3,000	3	400
splice_sites_all	30,000	3,000	3	600
splice_sites_acceptor	30,000	3,000	2	600
splice_sites_donor	30,000	3,000	2	600
H2AFZ	30,000	3,000	2	1,000
H3K27ac	30,000	1,616	2	1,000
H3K27me3	30,000	3,000	2	1,000
H3K36me3	30,000	3,000	2	1,000
H3K4me1	30,000	3,000	2	1,000
H3K4me2	30,000	2,138	2	1,000
H3K4me3	30,000	776	2	1,000
H3K9ac	23,274	1,004	2	1,000
H3K9me3	27,438	850	2	1,000
H4K20me1	30,000	2,270	2	1,000

C.2. Implementation

We maintained identical hyperparameter configurations across all tasks. Our systematic hyperparameter search included learning rates of 5×10^{-5}, 3×10^{-5}, and 5×10^{-4}, combined with batch sizes of 8, 16, and 32. Through empirical validation, we identified the optimal configuration employing a learning rate of 5×10^{-5} with batch size 8. The training protocol utilized the AdamW optimizer (Loshchilov & Hutter, 2019) over 3 epochs, while retaining default parameter settings from the HuggingFace Transformer Trainer implementation (Wolf et al., 2020).

C.3. Results

The complete benchmark results of the downstream tasks for NT are presented in Table 8. All baseline results are sourced from NT (Dalla-Torre et al., 2024). Performance per task was calculated as the median of the 10 cross-validation folds ($\pm$ standard deviation). The best results for each task are highlighted in **bold**.

Table 8. Complete Benchmark Results of Nucleotide Transformer Downstream Tasks

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

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

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

D. GUE

D.1. Dataset

GUE is a comprehensive benchmark for genome understanding consising of 28 distinct datasets across 7 tasks and 4 species, downloaded from https://github.com/MAGICS-LAB/DNABERT_2. The complete dataset composition, including sequence numbers, class distributions and sequence length statistics, is detailed in Table 9

Table 9. The Composition of GUE Datasets

Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

D.2. Implementation

Building upon DNABERT2's downstream task hyperparameter framework, we systematically evaluated learning rates from $5 \times 10^{-6}, 5 \times 10^{-5}, 6 \times 10^{-5}, 7 \times 10^{-5}, 8 \times 10^{-5}, 3 \times 10^{-4}$ while maintaining a consistent batch size of 32 across all tasks. Task-specific learning rates were empirically determined through validation set performance. The optimization process employed the AdamW algorithm (Loshchilov & Hutter, 2019) with 10,000 training steps, while retaining default parameter configurations from the HuggingFace Transformer Trainer implementation (Wolf et al., 2020).

Table 10. The results on the GUE datasets

Model	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	78.77	56.20	**61.99**	**55.30**	36.49	40.34
DNABERT-2	78.27	52.57	56.88	50.52	31.13	36.27
DNABERT-2 ■	**80.17**	**57.42**	61.90	53.00	39.89	41.20
Enformer	70.65	37.87	42.41	34.00	29.65	22.19
SPACE	79.53	54.12	54.82	50.92	**43.80**	**49.47**

D.3. Results

The results on the GUE datasets are presented in Table 10 and Table 11. In accordance with the implementation protocol of DNABERT2 (Zhou et al., 2024), all benchmark tasks utilized the Matthews Correlation Coefficient (MCC) for performance evaluation, with the singular exception of viral sequence analysis where F1-score metrics were employed. The notation DNABERT2 ■ specifically denotes the model variant that underwent additional masked language modeling (MLM) pre-training on the training sets of the Genomic Understanding and Evaluation (GUE) benchmark, as detailed in the DNABERT2 methodology.

Table 11. The results on the GUE datasets.

Model	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	78.23
NT-2500M-1000g	61.20	52.36	79.76	41.46	90.95	93.07	75.80
NT-2500M-multi	64.70	56.01	<u>81.67</u>	49.13	<u>91.01</u>	94.00	**79.43**
DNABERT-2	**67.39**	55.63	80.71	<u>50.43</u>	86.77	<u>94.27</u>	71.59
DNABERT-2 ■	65.46	<u>57.07</u>	**81.86**	50.35	88.31	**94.34**	68.79
grover					86.42	92.3	59.77
Enformer	55.69	49.35	76.32	32.90	85.68	92.92	69.63
SPACE	<u>66.93</u>	**59.29**	81.25	**53.09**	**91.90**	94.23	<u>79.13</u>

Model	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	<u>78.15</u>
DNABERT(4-mer)	67.90	73.05	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	76.79
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**	<u>76.06</u>	66.52	58.54	77.43	69.37	68.04	74.17
DNABERT-2 ■	69.12	71.87	62.96	55.35	74.94	67.50	69.53	76.18
grover	65.76	67.9	61.62	48.26	74.68	63.58	66.75	60.57
Enformer	<u>69.42</u>	72.76	**77.88**	**66.41**	<u>81.89</u>	60.94	66.46	46.21
SPACE	69.02	**76.49**	<u>76.45</u>	<u>66.08</u>	**82.91**	68.18	68.04	**79.23**

Model	Transcription Factor Prediction (Mouse)					Virus	Splice
	0	1	2	3	4	Covid	Splice
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	63.31	83.76	71.52	69.44	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	<u>73.49</u>	50.80	68.49	85.93
grover							84.35
Enformer	**67.15**	81.56	<u>85.99</u>	67.88	44.03	61.33	81.55
SPACE	<u>65.94</u>	84.91	**90.30**	**86.72**	<u>50.66</u>	70.26	<u>87.48</u>

E. Results on BEND Benchmark

Table 12. Results on all tasks of BEND

Method	Genomic Tasks				
	Chromatin accessibility	Histone modification	CpG Methylation	Variant effects (expression)	Variant effects (disease)
Expert method	0.85 BASSET	0.74 BASSET	0.93 BASSET	0.70 DEEPSEA	0.56 DEEPSEA
Fully supervised					
ResNet	–	–	–	–	–
CNN	0.75	0.76	0.84	–	–
Pre-trained					
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	**0.92**	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	**0.60**	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	**0.89**	**0.81**	**0.92**	0.51	0.49

Experiments in the main paper involve complete fine-tuning of our Encoder parameters. In this section, we validate the effectiveness of SPACE's embeddings on the Bend Benchmark (Marin et al., 2024) (i.e., directly using frozen SPACE embeddings for downstream tasks). All our experimental settings strictly follow the official configurations of Bend (Marin et al., 2024). The results are shown in Figure 12. We observe that SPACE achieves SOTA performance on chromatin accessibility, histone modification, and CpG methylation tasks, with chromatin accessibility surpassing the second-best method by 0.04. It is worth noting that although our supervised pre-training tasks also include chromatin accessibility and histone modification representations, which may provide potential advantages, their data processing approaches are not entirely identical. However, on the two variant effects tasks, SPACE shows limited effectiveness, similar to most DFMs. We hypothesize that masked language modeling tasks may be necessary to achieve good performance on variant effect prediction (Benegas et al., 2025).

F. Genomic Benchmarks

F.1. Dataset

Genomic Benchmarks currently comprises nine datasets focusing on regulatory elements (promoters, enhancers, and open chromatin regions) from three model organisms: Homo sapiens (human), Mus musculus (mouse), and Caenorhabditis elegans (nematode). All data were downloaded from https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks. The detailed composition of these datasets is presented in Table 13.

Table 13. Composition of Genomic Benchmarks.

Name	sequences	classes	Class ratio
dummy_mouse_enhancers_ensembl	1210	2	1.0
demo_coding_vs_intergenomic_seqs	100000	2	1.0
demo_human_or_worm	100000	2	1.0
drosophila_enhancers_stark	6914	2	1.0
human_enhancers_cohn	27791	2	1.0
human_enhancers_ensembl	154842	2	1.0
human_ensembl_regulatory	289061	3	1.2
human_nontata_promoters	36131	2	1.2
human_ocr_ensembl	174756	2	1.0

F.2. Implementation

We systematically evaluated learning rates $5 \times 10^{-6}, 5 \times 10^{-5}, 6 \times 10^{-5}, 7 \times 10^{-5}, 8 \times 10^{-5}, 3 \times 10^{-4}$ and batch sizes 8, 16, 32, 64. The optimal learning rate and batch size for each task were determined through validation set performance experiments. The optimization process employed the AdamW algorithm (Loshchilov & Hutter, 2019) with 3 training epochs, while maintaining the default parameter configuration from the HuggingFace Transformer Trainer implementation (Wolf et al., 2020).

G. Ablation Study

SPACE demonstrates comparable or superior performance to the decoder-removed variant in 14/18 tasks, with 11/18 tasks still outperforming even when replaced by a parameter-matched MLP. Notably, for regulatory element classification tasks, SPACE achieves better results in 4/5 datasets, with the only exception being the TATA box dataset—which primarily examines sequence motifs of TATA boxes and does not require complex regulatory mechanism understanding. This suggests that while our decoder does not explicitly improve direct chromatin profile prediction accuracy, the MoE architecture implicitly captures cross-profile regulatory interactions by modeling their dependencies. This capability provides critical advantages for tasks requiring integrated understanding of multiple profiles, such as regulatory element prediction.

H. Model Parameter Counts

We present the parameter counts of SPACE and its ablation variants in Table 16. The SPACE (large) configuration represents our primary model with complete architectural components for comparative analysis, while the other variants correspond to reduced-scale models specifically designed for ablation studies. These smaller models employ 131 KB input sequences with a compressed hidden dimension of 768 and operate under a batch size of 32.

Table 16. Model Parameter Counts of SPACE and its ablation variants

	SPACE (large)	SPACE w/o enhancement	SPACE w/o species MoE	SPACE (small)
param counts	588.75M	150.96M	105.19M	183.19M
hidden dim	1536	768	768	768

It should be particularly noted that, based on the sparse architecture design of the MoE, our model activates only a partial subset of parameters during a single forward computation. This selective parameter activation mechanism makes the number of effective parameters actually involved in the computation significantly lower than the total number of parameters in the model, thus significantly reducing the computational resource consumption while maintaining the model capacity.

Table 14. Ablation study on NT downstream tasks.

Model	Chromatin profiles					
	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
SPACE w/o decoder	0.535	0.514	0.567	0.593	0.520	0.604
SPACE w/o decoder w/ MLP	0.551	0.528	0.577	0.580	0.534	0.637
SPACE w/o encoder	0.540	0.524	0.569	0.579	0.506	0.625
SPACE w/o encoder and 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 w/o decoder	0.661	0.601	0.452	0.627	0.598	0.563
SPACE w/o decoder w/ MLP	0.668	0.589	0.451	0.636	0.601	0.558
SPACE w/o encoder	0.627	0.585	0.461	0.637	0.612	0.564
SPACE w/o encoder and 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 w/o decoder	0.752	0.773	0.841	0.873	0.884	0.936
SPACE w/o decoder w/ MLP	0.743	0.750	0.808	0.883	0.886	0.937
SPACE w/o encoder	0.738	0.769	0.828	0.864	0.869	0.933
SPACE w/o encoder and 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

Table 15. Ablation study on GUE benchmarks.

Model	Epigenetic Marks Prediction				
	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2
SPACE w/o dec	76.76	46.75	50.09	39.56	34.80
SPACE w/o dec w/ MLP	75.59	45.17	48.21	39.70	34.81
SPACE w/o enc	76.16	48.78	49.14	37.57	34.08
SPACE w/o enc and species emb	76.94	48.77	42.46	43.01	34.33
SPACE	76.40	50.76	49.18	41.30	32.83

Model	Epigenetic Marks Prediction					Virus
	H3K4me3	H3K79me3	H3K9ac	H4	H4ac	Covid
SPACE w/o dec	34.85	57.85	55.38	79.78	49.05	68.66
SPACE w/o dec w/ MLP	34.26	58.94	56.36	78.81	43.49	67.83
SPACE w/o enc	36.84	63.44	56.63	77.17	50.78	68.46
SPACE w/o enc and species emb	37.13	63.84	56.27	78.29	51.14	68.56
SPACE	37.74	61.10	57.06	79.33	51.05	68.89