Briefings in Bioinformatics, 2026, 27, bbag319
https://doi.org/10.1093/bib/bbag319
Published: 19 June 2026
Problem Solving Protocol

OXFORD

Porcine MutBERT: a family of lightweight genomic foundation models for functional element prediction in pigs

Weicai Long[1,‡], Rong Zhou[2,‡], Wenkang Wei[3], Xiaoai Zhang[3], Shanshan Wu[4], Kui Li[4], Yanlin Zhang[1,*], Zishuai Wang[3,4,*]

[1]Data Science and Analytics Thrust, The Hong Kong University of Science and Technology (Guangzhou), No. 1 Du Xue Road, Nansha District, Guangzhou 511453, China
[2]The State Key Laboratory of Animal Biotech Breeding, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, No. 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China
[3]Agro-biological Gene Research Center of Guangdong Academy of Agricultural Sciences, State Key Laboratory of Swine and Poultry Breeding Industry, No. 29 Jinying Road, Tianhe District, Guangzhou 510640, China
[4]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, No. 1 Funong Road, Dapeng District, Shenzhen 518000, China

*Corresponding authors: E-mail: yanlinzhang@hkust-gz.edu.cn; E-mail: zishuwang2-c@my.cityu.edu.hk.
‡Weicai Long and Rong Zhou contributed equally to this work.

Abstract

The pig (*Sus scrofa*) is both an economically important livestock species and a valuable biomedical model . Its genome bears regulatory features shaped by domestication and selection that are often poorly captured by genomic language models (gLMs) trained on human or model organism data. To address these challenges, we developed Porcine MutBERT, a suite of lightweight gLMs with 86 million parameters that employs a probabilistic masking strategy targeting evolutionarily informative single-nucleotide polymorphisms. This design captures population-specific variation while reducing computational cost. We further propose PorcineBench, a benchmark that evaluates gLM performance across porcine functional genomics tasks, including chromatin accessibility (ATAC-seq), CTCF binding, and histone modifications (H3K27ac, H3K4me1, and H3K27me3). Results show that Porcine MutBERT family achieves highly competitive performance on PorcineBench relative to substantially larger models, while providing an explicitly porcine-adapted alternative for downstream functional genomics in pigs. These findings underscore the advantages of species-adapted, efficient architectures in agricultural genomics and demonstrate that compact gLMs can expand accessibility and impact in resource-constrained settings. The code and data are available at https://github.com/ai4nucleome/pigmutbert.

Keywords Sus scrofa, genomic language model, lightweight architecture, functional genomics, porcine benchmark

Introduction

The pig (*Sus scrofa*) is a crucial livestock species [1] with significant agricultural and biomedical relevance [2]. The pig's genome has undergone complex selective pressures during domestication, resulting in a genomic architecture that is distinct from that of humans and other model organisms [3]. Therefore, understanding its genome is vital to advancing agricultural productivity, disease control, and biomedical research, particularly in areas such as organ transplantation. Traditional experimental approaches are resource-intensive, limited by sample size, and often constrained in coverage, motivating the development of computational methods that can efficiently infer functional genomic features.

Genomic language models (gLMs) offer a promising alternative by efficiently predicting gene expression and other phenotypic features. Through self-supervised learning on large genomic datasets, gLMs can capture rich sequence representations that enable accurate predictions, thus significantly reducing both the time and cost associated with experimental approaches [4, 5]. Although several prominent gLMs, such as DNABERT [6], DNABERT-2 [7], HyenaDNA [8], DNAGPT [9], GENA-LM [10], EVO [11], and Nucleotide Transformer (NT) [12] have achieved strong results across functional genomic tasks [13, 14], they are trained primarily on data from humans or models and often contain billions of parameters. Consequently, their applicability to non-model species such as pigs is limited, and their high computational demands hinder practical deployment in agricultural genomics.

To address these challenges, we developed Porcine MutBERT, a suite of lightweight gLMs specifically designed for pig genomes. All the Porcine MutBERT models contain 86 million parameters, making it computationally efficient and deployable on standard hardware. Unlike traditional random masking strategies, Porcine MutBERT-Var

Received: November 4, 2025. **Revised:** April 24, 2026. **Accepted:** May 18, 2026

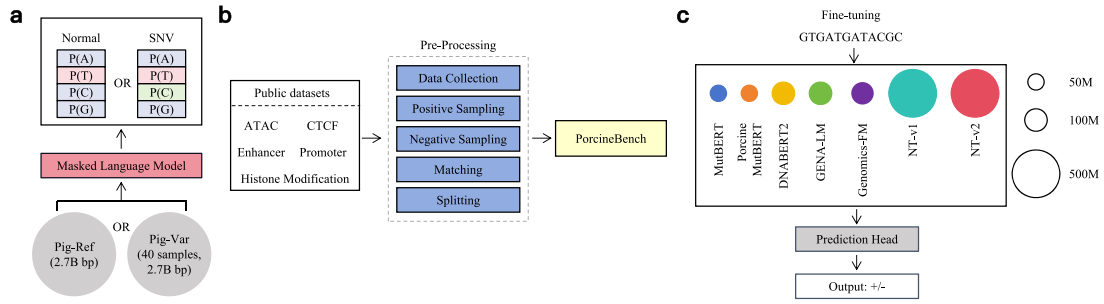

Figure 1 Overview of Porcine MutBERT. (a) Porcine MutBERT architecture. The input data consist of two sources. Pig-Ref is derived from the Sscrofa11.1 reference assembly, while Pig-Var includes sequences from 40 pig samples. Both datasets contain ∼2.7 billion base pairs. We trained the model using a masked language modeling (MLM) approach, where prediction is conditioned on the provided labels, and the pretraining objective is to minimize the cross-entropy loss over the masked positions. In the variant representation, the nucleotide entries with non-zero probabilities correspond to the reference and alternative alleles, respectively. (b) The process of constructing PorcineBench. The details of five steps are in section 3.6. (c) Fine-tuning strategy. We used a supervised approach on PorcineBench.

employs a probabilistic, mutation-informed masking strategy [15] that prioritizes evolutionarily informative single-nucleotide polymorphisms derived from population genomics data. This design enables the model to capture regulatory dependencies shaped by selection pressures, enhancing its ability to decode functional non-coding elements and predict genomic activity.

Furthermore, we established PorcineBench, a comprehensive benchmarking suite covering key functional genomics tasks, including chromatin accessibility (ATAC-seq), CTCF binding, and histone modifications (H3K27ac, H3K4me1, and H3K27me3). Comparative results show that Porcine MutBERT family achieves competitive performance against larger human-trained models and recently developed cross-species gLMs. These results establish that a compact, evolution-informed architecture can serve as a powerful tool for functional genomic prediction in pigs, advancing the genomic foundation for precision breeding, and agricultural biotechnology.

Materials and methods

MutBERT architecture

As shown in Fig. 1a, we used MutBERT [15] to train a masked language model based on Pig data. MutBERT is based on the Transformer architecture [16], specifically designed as an encoder-only model, and incorporates several key innovations such as RoPE [17] and flash attention [18] to effectively capture genomic sequence features. The architecture is composed of 86 million trainable parameters and designed to process genomic sequences represented probabilistically, as opposed to traditional one-hot encoding. Unlike models that use overlapping k-mers, MutBERT tokenizes DNA sequences using single nucleotide. The input sequences are tokenized into individual nucleotides (A, T, C, G), with additional special tokens for handling ambiguous nucleotides (N) and other sequence-specific information. The model processes input sequences consisting of 512 tokens, including two special tokens ([CLS], [SEP]) for classification tasks. Each token is mapped to a 768-D embedding space via an embedding layer. These embeddings are then augmented with rotary positional encodings to capture the relative positions of nucleotides within the input sequence, which is crucial for learning long-range dependencies in genomic sequences. In summary, the architecture of MutBERT is tailored for genomic sequences, leveraging Transformer-based attention mechanisms to learn contextual representations from DNA sequences. The model's capacity to process flexible length contexts and capture variation across the genome makes it well suited for tasks in functional genomics and population-based studies.

Table 1 Pretraining dataset. The four rows represent the splitting dataset, the number of base pairs, the number of SNVs, and the range of probabilities in the datasets, respectively.

	Pig-Ref		Pig-Var	
Split	**Train (M)**	**Test (M)**	**Train (M)**	**Test (M)**
Total bp	2349	56	2349	56
No. SNVs	0	0	42	1.3
Prob. of SNV	0 or 1	0 or 1	[0.0125, 0.9875]	[0.0125, 0.9875]

Models and pretraining dataset

For the pretraining of our porcine genomic models, we use two primary datasets: Pig-Ref and Pig-Var. The datasets used for training are summarized in Table 1. The overall preprocessing steps followed a similar approach to the one used in MutBERT, where we handled the genomic sequences by focusing on continuous stretches of DNA with a minimum sequence length of 510 base pairs, ensuring high-quality inputs for our models. Unless otherwise specified, Porcine MutBERT refers to the Porcine MutBERT family, which includes two models: Porcine MutBERT-Ref and Porcine MutBERT-Var.

Porcine MutBERT-Ref

The reference genome for pigs was obtained from the NCBI RefSeq database [19]. This dataset includes 20 chromosomes in total, with sequences from chromosomes 1–18, X, and Y. We used chromosome 18 for the validation set and the remaining chromosomes (1–17, X, and Y) for training. The total base pairs (bp) for these 20 chromosomes are estimated to be ∼2.7 billion bp, with the specifics of each chromosome's length derived from the RefSeq data. These sequences were processed by removing regions containing ambiguous bases (N) and assembling continuous sequences of at least 510 bp. Adjacent sequences without N bases were concatenated using a special separator token [SEP], resulting in a sequence with length of L, where L is the length of the sequence after padding. This sequence was then converted into a one-hot encoding matrix, with a shape of $9 \times L$, where 9 is the vocabulary size, including four canonical nucleotides

A, T, C, G, and five special tokens [PAD], [CLS], [SEP], [UNK], and [MASK].

Porcine MutBERT-Var

For the mutation dataset, we selected 40 different pig samples, including both wild and domesticated ones [20]. Raw sequencing data from these samples were aligned to the reference genome, and variant call format (VCF) files were generated. These VCF files contain the information necessary to derive mutation probabilities at each genomic locus. The mutation probabilities were integrated into the same $9 \times L$ matrix format as the reference sequences, but for each base, the matrix was modified according to the probability of each possible nucleotide (A, T, C, G) at that locus, as indicated in the VCF data. This allows us to capture the population-specific and breed-specific variations in the porcine genome. Detailed metadata for all selected samples are provided in Supplementary Table S1.

Pretraining strategy

We followed self-supervised recipe of MutBERT and pretrained with an MLM objective. The masking policy is identical to MutBERT: for each input, we first prioritize loci that harbor variants in the mutation-aware inputs when constructing the mask set, and only then backfill with non-variant positions as needed to meet the target mask budget. Optimizer uses AdamW [21] with $\beta_1 = 0.9, \beta_2 = 0.99$ and weight decay equals to 0.01. Pretraining was carried out with a sequence length of 512 tokens and an effective batch size of one million tokens for 200 000 update steps, resulting in the model training on a total of 200 billion tokens. We adopted a custom learning rate (lr) schedule with 10 000 linear warmup steps from an initial lr of 1e-8 to a peak of 2e-4, followed by cosine decay back to 1e-8 over the remaining steps. To construct inputs, we used random sampling over the prepared $9 \times L$ matrices. At each step, we sampled a start index uniformly at random and extract a 510-token contiguous window. We then prepended [CLS] and appended [SEP] to form a 512-token sequence for the model. Because the corpus was built by concatenating non-N contigs with a [SEP] delimiter, any randomly sampled 510-token window that contains an internal [SEP] (i.e. would span discontiguous genomic fragments) was discarded and resampled to ensure all training sequences were truly contiguous in the underlying genome.

Alternative models

We benchmarked Porcine MutBERT against five widely used genomic foundation model families, as shown in Table 2. We note architecture traits, parameter scale, and whether porcine sequences were included during pretraining.

DNABERT-2

DNABERT-2 replaces k-mer tokenization with Byte Pair Encoding (BPE) [22] and removes the hard context limit by adopting ALiBi [23] positional biases. Additionally, it also integrates flash attention for efficiency. The official paper reports evaluation on inputs up to 10k bp on GUE/GUE+ tasks and emphasizes efficiency at 117 M parameters for a common checkpoint. Pretraining was performed on a multi-species corpus (135 species) without *S. scrofa*.

Nucleotide Transformer (NT-v1)

NT-v1 introduced large-scale, multi-species datasets spanning 850 genomes for pretraining with 6-mer tokenization and models ranging from 500 million to 2.5 billion parameters. It can process DNA sequences up to 1000 tokens (6 kb). The released NT-v1 multi-species checkpoints did not include porcine-specific pretraining.

Nucleotide Transformer (NT-v2)

NT-v2 focuses on parameter-efficient variants from 50 million to 500 million parameters and updates the stack with RoPE positional embeddings and swiGLU activations [24]. The 500 M multi-species model was pretrained on 850 genomes, using 900 billion tokens in total. The model can support input length up to 2048 tokens (12 kb). As with the first generation, the public NT-v2 multi-species checkpoints had no porcine-specific data for pretraining.

MutBERT family

The original MutBERT series (MutBERT, MutBERT-Human-Ref, and MutBERT-Multi) used an 86 M encoder-only architecture with character-level tokens and a probabilistic genome input. MutBERT and MutBERT-Human-Ref were pretrained exclusively on the 1000 Genomes Project [25] and the human reference genome, while MutBERT-Multi was pretrained on multi-species alignments anchored to human [26]. There was no porcine pretraining in these human-centric checkpoints.

GENA-LM

It is a transformer-based foundational DNA language model series designed for long-sequence genomics and cross-species generalization. GENA-LM family comprises numerous variants and was pretrained on a corpus exceeding one trillion bp derived from human and multispecies genomes utilizing BPE tokenization. This model integrates the architectures of BERT [27] and BigBird [28], incorporating sparse attention mechanisms, and is further augmented with a recurrent memory mechanism to facilitate the processing of contexts extending to 36 000 bp. For benchmarking purposes, the gena-lm-bert-base-t2t-multi model was selected, as its released weights include sequences from porcine genomes during the pretraining phase.

Genomics-FM

Genomics-FM is a genomics foundation model trained on large-scale multi-species genomic data, including *S. scrofa*. We included it as an additional baseline because it is one of the few published models with porcine sequences in pretraining. For fair comparison, we evaluated Genomics-FM under the same downstream fine-tuning protocol as the other models on PorcineBench.

Fine-tuning strategy

We adopted a unified protocol across all 21 downstream tasks and vary only the adaptation regime by model scale. We performed full-parameter fine tuning for all models. The fine-tuning process is shown in Fig. 1c. Optimization uses AdamW, and the learning rate increases linearly from 0 to 3e-5 over the first 50 steps and then decreases linearly to 0 for the remainder of training. Fine-tuning runs for three epochs with a mini-batch size of 32. All other hyperparameters, schedules, and data splits across different models are the same. As the MutBERT family was pretrained with a 512-token context, tasks that require longer context, specifically the histone mark prediction task, used a 1024-bp sequence during fine tuning and evaluation. We realize this extension through RoPE-based length extrapolation.

Table 2 The statistics and performance of each model. The four columns represent the number of model parameters, species in pretraining, the number of being top-1 among all the models, and the average Matthews correlation coefficient (MCC) scores on the 21 datasets of the PorcineBench, respectively.

Model	Params.	Species	No. Top-1	Ave. scores
DNABERT-2	117 M	135 Multi species	1	50.94
GenomicsFM	117 M	1160 Multi species (with *S. scrofa*)	1	50.60
GENA-LM	110 M	312 Multi species (with *S. scrofa*)	0	44.85
MutBERT-Human-Ref	86 M	*Homo sapiens*	0	48.61
MutBERT	86 M	*H. sapiens*	0	48.76
MutBERT-Multi	86 M	100 Multi species	0	48.75
NTv1-500 M-1000G	480 M	*H. sapiens*	0	41.20
NTv1-500 M-Human-Ref	480 M	*H. sapiens*	0	39.43
NTv2-50 M-Multi	56 M	850 Multi species	0	45.58
NTv2-100 M-Multi	98 M	850 Multi species	0	48.19
NTv2-250 M-Multi	235 M	850 Multi species	0	50.46
NTv2-500 M-Multi	498 M	850 Multi species	7	**52.01**
Porcine MutBERT-Var	86 M	*S. scrofa*	4	51.56
Porcine MutBERT-Ref	86 M	*S. scrofa*	**8**	51.49

Bold values indicate the highest values, and underlined values indicate the second-highest values.

Porcine benchmark

In this section, we detail each benchmark task and its corresponding dataset, with key statistics provided in Table 3. The whole pipeline can be found in Fig. 1b. The processed BED files containing peak regions of ATAC-seq, CTCF binding sites, enhancer and promoter predictions, as well as data for three histone marks (H3K27ac, H3K27me3, and H3K4me1) were obtained from a publicly available dataset [29]. The enhancer and promoter benchmarks were constructed directly from released pig chromatin-state annotation BED files, and the annotations were produced with a 15-state model. Following the source annotation [29], we mapped E1 (TssA) to promoter and E6 (EnhA) to enhancer. All regulatory assays were ingested from compressed BED files aligned to the S. scrofa11.1 assembly. GSM accessions in file names were preserved to maintain sample-level provenance. We used Adipose, Cerebellum, Lung for CTCF, and Cortex, Muscle, and Liver for the remaining tasks to ensure adequate samples. Across tasks, processing follows a shared blueprint with task-specific window sizes: (i) positive sampling. We centered each positive interval on its midpoint and expanded symmetrically to the target length. After that, we required the expanded window to be completely contained within the original interval. We discarded any window whose extracted sequence contains N. In addition, we computed the counts of G and C (GC content) and recorded the chromosome for subsequent matching. (ii) Negative sampling. We drew 10-folds as many non-overlapping, fixed-length negatives with bedtools [30] constrained to the same genome build. Similar to methods of constructing positive samples, we extracted sequences for all negatives and applied the same N-free and GC computations. (iii) Matching. We established one-to-one positive–negative pairs by prioritizing same chromosome, identical GC matches, then relaxing GC to ± 1 and ± 2, and finally allowing cross-chromosome matches with GC priorities $(0, \pm 1)$. (iv) Splitting. In the last step, we split pairs into train, validation and test sets in an 8:1:1 ratio, then flattened and shuffled so that each split is perfectly class balanced.

Table 3 Overview of PorcineBench.

Task	Tissue	Length	Train	Val	Test
ATAC	Cortex	500	50 000	6250	6250
	Muscle	500	50 000	6250	6250
	Liver	500	50 000	6250	6250
CTCF	Adipose	500	13 994	1750	1750
	Cerebellum	500	18 970	2372	2372
	Lung	500	28 120	3514	3516
Enhancer	Cortex	400	50 000	6250	6250
	Muscle	400	46 804	5850	5852
	Liver	400	28 208	3526	3628
Promoter	Cortex	400	11 542	1442	1444
	Muscle	400	15 008	1876	1878
	Liver	400	10 932	1366	1368
H3K27ac	Cortex	1000	28 728	3592	3592
	Muscle	1000	26 304	3288	3288
	Liver	1000	15 376	1922	1922
H3K27me3	Cortex	1000	50 000	6250	6250
	Muscle	1000	50 000	6250	6250
	Liver	1000	25 704	3214	3214
H3K4me1	Cortex	1000	29 294	3662	3662
	Muscle	1000	8498	1062	1064
	Liver	1000	42 808	5352	5352

Chromatin profiles prediction

The focus on open chromatin aligns with the observation that adaptive regions in pigs are strongly enriched in ATAC Islands and enhancer states. Positive samples were drawn from BED files corresponding to open-chromatin peaks. For each peak, we took the midpoint and expanded $-250/+249$ to 500 bp, enforcing full containment within the peak interval. GC content and chromosome were recorded per retained window. Negatives were generated at 10 times with fixed length 500 bp and matched one-to-one as described. For each tissue, final subsets were produced by an 8:1:1 pair-wise split and class balancing, including 50 000, 6250, 6250 samples in the training, validation, and testing sets, respectively.

CTCF binding site prediction

CTCF binding sites were compiled from GSM-labeled BEDs on S. scrofa11.1, providing peak coordinates for insulator-like elements. We centered each peak and expanded to 500 bp. As a result, we achieved a dataset with 17 494, 23 714, and 35 150 samples for the tissue of

Adipose, Cerebellum, and Lung, respectively. The datasets was then partitioned with the same ratio as ATAC at the pair level and flattened to balanced splits.

Enhancer annotation

The emphasis on enhancer prediction is motivated by the porcine evidence that candidate adaptive signals are particularly concentrated in enhancer and open-chromatin contexts. Positive samples of Enhancer aggregated across enhancer subclasses were curated from the porcine multi-tissue chromatin state BEDs on S. scrofa11.1. We constructed 400 bp windows by midpoint expansion −200/+199 bp within each enhancer interval and applied the N-free filter. Negative samples of length 400 bp were matched using the GC content and chromosome. After an 8:1:1 pair-wise split, class-balanced splits were obtained by flattening and shuffling. Finally, we achieved a dataset with 62 500, 58 506, and 35 362 instances for the tissue of Cortex, Muscle, and Liver, respectively.

Promoter prediction

The promoter state definitions and tissue modules follow the porcine annotation schema used throughout this work. Promoter-proximal positives were sourced from promoter states. Windows of 400 bp were derived by midpoint expansion −200/+199 bp while requiring full containment and N-free status. We achieved a dataset with 14 428, 18 762, and 13 666 samples for three tissues, respectively. A ratio of 8:1:1 was applied to split the dataset, resulting in training, validation, and test set for fine-tuning.

Histone modification prediction

For each mark, ChIP-seq peak BEDs on S. scrofa11.1 were used to define positives per tissue. We centered each peak and expanded −500/+499 bp to a window of 1000 bp, enforcing complete containment. Negatives samples were also matched by the GC content and chromosome policy. After an 8:1:1 pair-wise split, we produced balanced splits by flattening and shuffling, resulting in 88 012 samples for H3K27ac, 157 132 samples for H3K27me3 and 100 754 samples for H3K4me1.

Hardware

Model pretraining was carried out using NVIDIA L20 accelerators, specifically a machine containing four devices. Since the model was able to fit on a single L20 device, pretraining was carried out in a data parallel fashion where model parameters were replicated and the effective batch sharded across the four devices with subsequent gradient accumulation. We trained for a total of ∼7 days.

Results

Pretraining convergence

To assess whether the porcine models were sufficiently trained, we monitored the training and evaluation losses throughout the full pretraining run. As shown in Fig. 2, both Porcine MutBERT-Ref and Porcine MutBERT-Var exhibit rapid loss reduction in the early stage and gradually approach a stable plateau toward the end of training, indicating clear convergence after 200k update steps.

After pretraining, Porcine MutBERT-Ref reached a final training loss of 0.935 and an evaluation loss of 0.985, whereas Porcine MutBERT-Var achieved lower final losses of 0.893 and 0.931, respectively. These

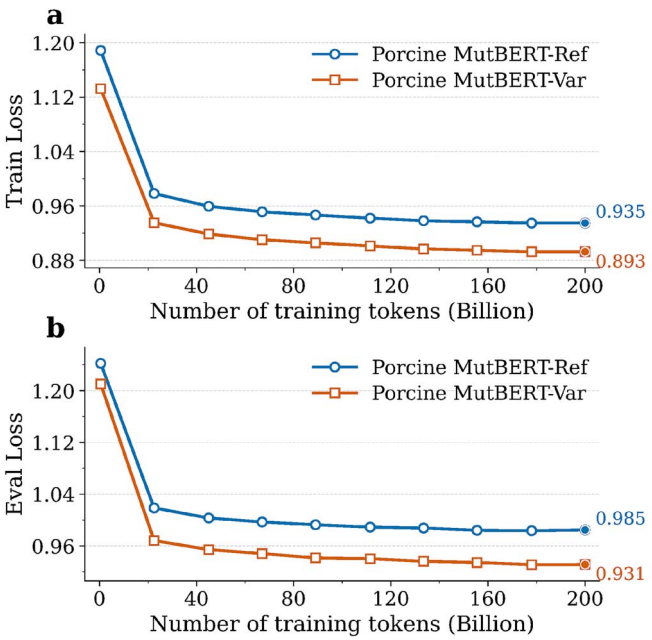

Figure 2 Pretraining loss convergence curves of Porcine MutBERT. (a) Training loss of Porcine MutBERT-Ref and Porcine MutBERT-Var during pretraining. (b) Evaluation loss of Porcine MutBERT-Ref and Porcine MutBERT-Var during pretraining. The x-axis shows the number of training tokens in billions.

results suggest that both models were sufficiently trained, and that the variant-aware model achieved better language modeling performance than the reference-only model.

Overall performance

As shown in Tables 4 and 5, on PorcineBench, the macro-averaged MCC [31] leaderboard places NTv2-500M-Multi first (52.01), with Porcine MutBERT-Var (51.56), and Porcine MutBERT-Ref (51.49) close behind. Additional robustness analyses are provided in Supplementary Tables S3–S8, including chromosome-held-out evaluation, five-fold fine-tuning results, and paired statistical tests. These supplementary analyses show that the main conclusions are broadly stable beyond the default random split. Table 2 summarizes the top-1 counts across tasks. Porcine MutBERT-Ref achieves the most wins, leading on 8 of 21 tasks, followed by NTv2-500M-Multi with 7 of 21. Porcine MutBERT-Var ranks next with 4 of 21, while DNABERT2 and Genomics-FM each lead on 1 of 21 tasks. In all, Porcine MutBERT family achieves top results in 12 of 21 tasks. These results indicate that a compact, species-adapted encoder with 86 M parameters attains competitive performance while covering a broad swath of tasks, thereby achieving a favorable accuracy-scale balance when compared with larger generalist baselines and pig-relevant multi-species baselines. The small gap to NTv2-500M-Multi suggests residual benefits from model size, but the proximity of porcine models highlights their efficiency, despite a six times parameter difference. A detailed runtime and resource comparison with NTv2-500M-Multi is provided in Supplementary Table S2.

Dissecting the seven task families clarifies what the models learn about *S. scrofa* regulation and why winners shift. For ATAC-seq, the models pretrained on porcine datasets capture these signals similarly, with Porcine MutBERT-Ref sometimes edging due to deterministic

Table 4 The first part of the results: performance of seven models on PorcineBench using MCC metric. The bold value indicates the highest scores across all the models. Avg. MutBERT denotes the average score across MutBERT, MutBERT-Human-Ref, and MutBERT-Multi.

Task	DNABERT2	GENA-LM	MutBERT	MutBERT-Human-Ref	MutBERT-Multi	NTv1-500M-1000G	NTv1-500M-Human-Ref	Avg. MutBERT
ATAC_Cortex	43.18	41.04	42.87	43.64	42.22	34.50	33.69	42.91
ATAC_Liver	39.70	39.41	39.60	39.14	39.49	31.52	31.59	39.41
ATAC_Muscle	**39.81**	33.92	36.87	37.29	37.20	28.40	27.27	37.12
Avg. ATAC	40.90	38.12	39.78	40.02	39.64	31.47	30.85	39.81
CTCF_Adipose	80.47	52.59	76.36	70.43	78.98	64.98	49.35	75.26
CTCF_Cerebellum	74.75	56.72	66.53	64.21	69.28	60.31	46.87	66.67
CTCF_Lung	33.27	26.33	28.39	29.87	29.98	25.04	20.56	29.41
Avg. CTCF	62.83	45.21	57.09	54.84	59.41	50.11	38.93	57.11
Enhancer_Cortex	51.89	48.34	50.16	49.67	50.57	41.68	41.60	50.13
Enhancer_Liver	48.06	42.57	45.29	45.34	46.43	38.61	39.86	45.69
Enhancer_Muscle	38.99	34.43	36.01	36.90	36.84	27.72	25.80	36.58
Avg. Enhancer	46.31	41.78	43.82	43.97	44.61	36.00	35.75	44.13
Promoter_Cortex	75.50	73.87	75.65	76.60	76.23	68.84	69.95	76.16
Promoter_Liver	70.70	67.57	69.41	68.65	68.65	61.49	60.13	68.90
Promoter_Muscle	69.25	64.67	67.38	67.85	67.91	59.86	63.33	67.71
Avg. Promoter	71.82	68.70	70.81	71.03	70.93	63.40	64.47	70.93
H3K27ac_Cortex	61.48	59.02	61.24	60.95	59.36	55.32	53.75	60.52
H3K27ac_Liver	52.45	47.41	53.81	52.94	49.68	46.45	43.61	52.14
H3K27ac_Muscle	54.82	52.25	54.09	53.28	51.40	44.61	42.83	52.92
Avg. H3K27ac	56.25	52.89	56.38	55.72	53.48	48.79	46.73	55.19
H3K27me3_Cortex	36.48	31.70	34.19	34.20	31.76	28.16	27.86	33.38
H3K27me3_Liver	27.90	25.33	27.03	26.91	25.05	19.20	20.85	26.33
H3K27me3_Muscle	34.40	31.27	32.01	31.42	32.58	26.57	25.93	32.00
Avg. H3K27me3	32.93	29.43	31.08	30.84	29.80	24.64	24.88	30.57
H3K4me1_Cortex	52.88	46.24	48.90	51.43	49.49	39.11	38.88	49.94
H3K4me1_Liver	39.00	33.11	38.64	36.48	38.73	29.75	28.76	37.95
H3K4me1_Muscle	44.80	34.03	39.49	43.54	41.84	32.99	35.55	41.62
Avg. H3K4me1	45.56	37.79	42.34	43.82	43.35	33.95	34.40	43.17
Average	50.94	44.85	48.76	48.61	48.75	41.20	39.43	48.70

targets. In the meantime, NTv2-500M-Multi can overtake them when extended context captures distal cooperative patterns beyond the 512-token pretraining length. For CTCF, the samples are fewer than that in ATAC task. For instance, the training set of Adipose comprised only 13 994 samples in total, corresponding to 6997 positive and 6997 negative samples. Despite this limited data availability, the strongest model varies across tissues, but Porcine MutBERT-Var still remains competitive, especially on Lung. In contrast, Porcine MutBERT-Ref exhibited considerable variability and its overall performance was markedly lower than that of Porcine MutBERT-Var. Furthermore, although the sample size was also relatively small, all models achieved comparatively high scores on the promoter prediction task, suggesting that this task is intrinsically less challenging. In the enhancer prediction task, models pretrained on a single species exhibited similarly comparable performance. Specifically, the differences within the MutBERT family as well as among Porcine MutBERT variants were relatively small. Within this context, Porcine MutBERT-Var frequently ranked among the strongest models, particularly when compared against the first-generation NT, which was exclusively trained on human data. For the histone modifications prediction task, which requires sequence inputs of 1000 bp, the MutBERT family models had to rely on RoPE-based length extrapolation to accommodate such extended sequences. Despite this challenge, Porcine MutBERT-Ref consistently achieved the best scores, further emphasizing that even the strongest genomic language models remain difficult to directly extend to species for which they were not explicitly pretrained.

Across all families, Porcine MutBERT-Var delivers aggregate performance within 0.45 points of NTv2-500M-Multi while decisively outperforming GENA-LM, DNABERT2, and Genomics-FM on the overall average score. These results collectively highlight a critical limitation of current genomic language models: even modern gLMs exhibit substantial challenges when directly transferred to species outside their pretraining domain. This observation underscores the importance of developing species-specific pretraining strategies for achieving robust cross-species generalization.

Comparison with models trained on pigs

In this section, we compared Porcine MutBERT family against GENA-LM and Genomics-FM. Under the unified protocol, both Porcine MutBERT-Var and Porcine MutBERT-Ref exceeded GENA-LM and Genomics-FM on the overall macro score, as shown in Table 2. As illustrated in Tables 4 and 5, the results on PorcineBench further emphasize the advantages of our model architecture and pretraining paradigm. On this benchmark, Porcine MutBERT family achieved the best scores in 18 of 21 sub-tasks among these pig-relevant models. Notably, among the 21 tasks in total, GENA-LM recorded the lowest score on 20 of 21 sub-tasks, suggesting that even when trained with data from the target species, existing genomics language models may fail to capture sufficiently robust species-specific regulatory knowledge, thereby limiting their applicability in downstream biological scenarios. Within the Porcine MutBERT family, Porcine MutBERT-Ref and Porcine MutBERT-Var exhibited a complementary relationship: Porcine MutBERT-Ref consistently obtained the best scores on most histone modification and promoter tasks, whereas Porcine MutBERT-Var performed better on several CTCF and enhancer

Table 5 The second part of the results: Performance of seven Models on PorcineBench Using MCC Metric. The bold values indicate the highest scores across all the models. Avg. Poricne MutBERT denotes the average score across Porcine MutBERT-Var and Porcine MutBERT-Ref.

Task	NTv2-50M-Multi	NTv2-100M-Multi	NTv2-250M-Multi	NTv2-500M-Multi	Genomics-FM	Porcine MutBERT-Var	Porcine MutBERT-Ref	Avg. Porcine MutBERT
ATAC_Cortex	41.11	42.12	42.98	**44.23**	42.64	43.68	43.93	43.81
ATAC_Liver	35.62	37.79	39.72	41.18	40.87	41.57	**42.08**	41.83
ATAC_Muscle	33.44	34.62	37.92	38.48	39.51	38.77	38.52	38.65
Avg. ATAC	36.72	38.18	40.21	41.30	41.01	41.34	41.51	41.43
CTCF_Adipose	53.66	74.75	79.33	83.68	**83.77**	81.96	77.92	79.94
CTCF_Cerebellum	66.97	70.55	72.44	**76.92**	76.39	73.15	67.06	70.11
CTCF_Lung	27.89	34.70	34.08	35.32	34.15	**35.33**	30.34	32.84
Avg. CTCF	49.51	60.00	61.95	65.31	64.77	63.48	58.44	60.96
Enhancer_Cortex	47.38	49.07	50.74	52.46	52.01	**53.11**	52.23	52.67
Enhancer_Liver	43.58	44.71	49.52	**50.78**	50.26	50.58	46.86	48.72
Enhancer_Muscle	32.88	34.09	36.76	**40.22**	25.96	36.02	38.84	37.43
Avg. Enhancer	41.28	42.62	45.67	47.82	42.74	46.57	45.98	46.27
Promoter_Cortex	71.19	74.40	75.97	75.81	75.90	**77.76**	77.54	77.65
Promoter_Liver	66.09	66.37	69.32	71.17	69.06	71.94	**72.11**	72.03
Promoter_Muscle	64.89	64.80	67.76	67.97	69.85	69.47	**71.40**	70.44
Avg. Promoter	67.39	68.52	71.02	71.65	71.60	73.06	73.68	73.37
H3K27ac_Cortex	57.88	59.08	61.35	61.42	61.53	**63.22**	62.67	62.95
H3K27ac_Liver	51.46	51.85	51.82	52.64	53.13	54.63	**56.30**	55.47
H3K27ac_Muscle	51.95	51.42	53.58	**55.67**	53.94	53.99	54.23	54.11
Avg. H3K27ac	53.76	54.12	55.58	56.58	56.20	57.28	57.73	57.51
H3K27me3_Cortex	33.21	32.95	35.38	**37.07**	35.72	34.40	36.49	35.45
H3K27me3_Liver	26.00	26.61	28.57	28.40	24.08	27.42	**29.13**	28.28
H3K27me3_Muscle	30.44	31.36	34.96	**37.49**	35.37	36.28	35.97	36.13
Avg. H3K27me3	29.88	30.31	32.97	34.32	31.72	32.70	33.86	33.28
H3K4me1_Cortex	47.79	51.63	51.75	53.28	53.47	54.30	**55.22**	54.76
H3K4me1_Liver	34.75	37.97	39.13	41.31	40.73	41.04	**42.65**	41.85
H3K4me1_Muscle	38.99	41.23	46.52	46.66	44.32	44.19	**49.73**	46.96
Avg. H3K4me1	40.51	43.61	45.80	47.08	46.17	46.51	49.20	47.86
Average	45.58	48.19	50.46	**52.01**	50.60	51.56	51.49	51.52

tasks. This observation highlights that solving the full spectrum of domain-specific challenges may ultimately require leveraging multiple task-specialized models rather than relying on a single general-purpose architecture. Tables 4 and 5 also summarize the macro-averaged results across tissues for each task. Among all other tasks, the Porcine MutBERT family consistently ranked among the top-performing models. Of particular note, on the promoter annotation task, which is intrinsically less challenging, Porcine MutBERT-Ref still surpassed GENA-LM by approximately five points and outperformed Genomics-FM across all three promoter tissues, demonstrating a clear, and significant performance advantage.

Cross-species transfer under matched architecture

Tables 4 and 5 compare the performance of the Porcine MutBERT family and the original MutBERT family, which share the same model architecture but differ in the species data used during pretraining. As summarized in Tables 4 and 5, the highest average score achieved by the original MutBERT family was 48.76 (MutBERT), whereas the Porcine MutBERT family reached a higher score of 51.56 (Porcine MutBERT-Var). Both top-performing models rely on probabilistic genome representations. The Porcine MutBERT family covered all 21 datasets, with Porcine MutBERT-Ref and Porcine MutBERT-Var consistently occupying the top two positions across all tasks. These findings provide compelling evidence that modern genomic language models cannot be directly extended across species to achieve competitive performance. The largest discrepancy appeared in H3K4me1, where the performance gap reached 5.38 points. This divergence can be attributed to two factors. First, the Human-centric MutBERT family, being trained mainly on human data, failed to acquire transferable cross-species knowledge. Second, the H3K4me1 task contained relatively few samples in Muscle tissue, which increased the reliance on few-shot robustness. Consequently, the lack of both cross-species adaptability and few-shot capability compounded the performance difference.

Embedding visualization

To gain insight into what Porcine MutBERT learns across layers, we visualized Porcine MutBERT-Var embeddings of held-out promoter and enhancer sequences using t-SNE (Fig. 3). In the early layer, the two sequence types remain substantially mixed. As depth increases, their representations become progressively more structured, and by the final layer promoter and enhancer sequences are much more clearly separated. This qualitative pattern suggests that Porcine MutBERT-Var gradually transforms raw porcine sequence inputs into higher-level representations that capture regulatory distinctions relevant to promoter and enhancer elements. As a complementary analysis, the *in silico* mutation results are provided in Supplementary Figs S1 and S2.

Discussion

In this work, we presented two species-adapted porcine models: Porcine MutBERT-Var and Porcine MutBERT-Ref and a multi-omics benchmark (PorcineBench). PorcineBench contains 7 task families across 6 tissues in total, resulting in 21 datasets. On

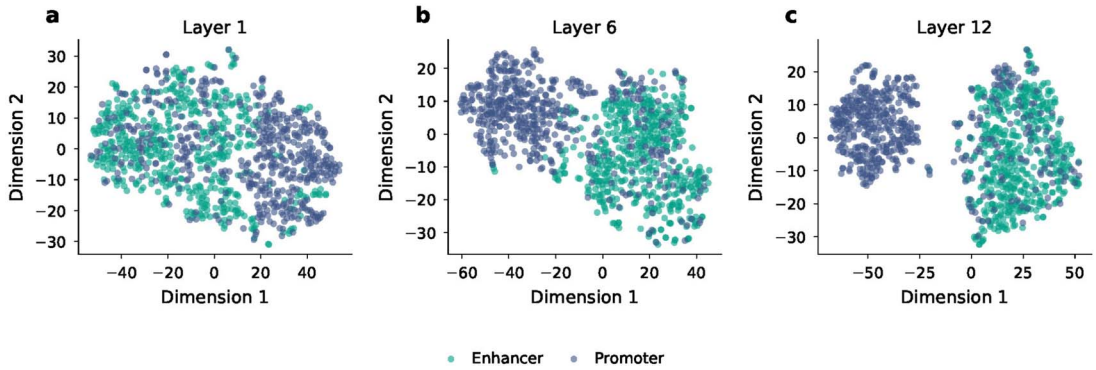

Figure 3 Layer-wise t-SNE visualization of Porcine MutBERT-Var embeddings for held-out promoter and enhancer sequences at three representative layers: (a) Layer 1, (b) Layer 6, and (c) Layer 12.

PorcineBench, Porcine MutBERT family secures 12/21 top-1 results and remains competitive with substantially larger generalist models. By contrast, existing gLMs exhibited limited performance, with even state-of-the-art models excelling only on a small subset of tasks.

PorcineBench provides a systematic and reliable framework for evaluating future pig-specific models. The benchmark covers seven tasks and six tissues, and incorporates strict data filtering and balancing strategies, including removal of ambiguous sequences containing N, as well as samples matching by chromosome and GC content. Finally, we developed a five-step pipeline that transforms raw porcine regulatory element data into fair and trustworthy downstream datasets. Leveraging the MutBERT architecture with porcine-specific pretraining, the Porcine MutBERT family delivered consistently strong results. In PorcineBench, Porcine MutBERT-Var achieved an average MCC of 51.56, while Porcine MutBERT-Ref reached 51.49. Remarkably, despite being only one-sixth the size of the largest 500 M-parameter model, Porcine MutBERT narrowed the performance gap to within 0.45 points, striking an optimal balance between accuracy and model scale. Although NTv2-500M-Multi remains a strong generalist baseline, Porcine MutBERT offers a more favorable practical tradeoff for porcine applications by requiring smaller checkpoint storage and less GPU memory while maintaining broadly comparable speed. These findings not only emphasize the effectiveness of the MutBERT architecture but also highlight the superiority of species-specific pretraining. Compared with models exclusively pretrained on single species, Porcine MutBERT consistently achieved top scores across all tasks.

Conclusion

Our findings highlight a persistent challenge in genomic foundation modeling: cross-species generalization remains difficult. The pig genome differs from human and other commonly used model organisms in repetitive content, structural variation, and breed-specific regulatory landscapes shaped by domestication and selection. These biological differences, in conjunction with model inductive biases (e.g. tokenization granularity, positional encoding, and context length), create a fundamental mismatch between human or multi-species pretraining and porcine downstream distributions. This mismatch explains why large-scale human or multi-species models fail to reliably transfer to pigs (*S. scrofa*) under standardized evaluation.

Key Points

- We developed Porcine MutBERT, a suite of lightweight genomic language models with 86 million parameters that employs a probabilistic masking strategy targeting evolutionarily information.
- The proposed PorcineBench is a comprehensive evaluation suite, spanning seven task families across six tissues, with balanced positive–negative sampling and GC content matching.
- Results show that Porcine MutBERT models achieve highly competitive performance on PorcineBench.

Author contributions

W.L., Y.Z., and Z.W. conceived the study. W.L., R.Z., W.W., S.W., and K.L. performed analysis. Z.W. supervised the project. W.L. and Z.W. wrote the article. All authors read and approved the final article.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

The authors declare no competing interests.

Funding

This work is funded by Biological Breeding-National Science and Technology Major Project (2023ZD04076), Guangdong Provincial Project (No. 2024QN11N085), the National Natural Science Foundation of China (32472855, 32441082), Fund of State Key Laboratory of Swine and Poultry Breeding Industry (No. GDNKY-ZQQZ-K6, GDNKY-ZQQZ-K19), Provincial Rural Revitalization Strategy Special Fund Project for Seed Industry Revival Action (No. 2024-XPY-00-015) and Basic Research Center for Livestock and Poultry Sciences (CAAS-BRC-LP-2025-01).

Data availability

The datasets and codebase used in this work can be accessed at https://github.com/ai4nucleome/pigmutbert.

References

1. Balogh P. Global and national economic importance of pig meat production. *Acta Agrar Debrecen* 2017;**73**:13–20.
2. Lunney JK, Van Goor A, Walker KE *et al.* Importance of the pig as a human biomedical model. *Sci Transl Med* 2021;**13**:eabd5758.
3. Groenen MAM, Archibald AL, Uenishi H *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 2012;**491**:393–8.
4. Benegas G, Ye C, Albors C *et al.* Genomic language models: opportunities and challenges. *Trends Genet* 2025;**41**:286–302.
5. Consens ME, Dufault C, Wainberg M *et al.* Transformers and genome language models. *Nat Mach Intell* 2025;**7**:346–62.
6. Ji Y, Zhou Z, Liu H *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20.
7. Zhou Z, Ji Y, Li W *et al.* DNABERT-2: efficient foundation model and benchmark for multi-species genomes. In: *The Twelfth International Conference on Learning Representations*, Vienna, Austria: OpenReview.net. 2024.
8. Nguyen E, Poli M, Faizi M *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv Neural Inf Proces Syst* 2023;**36**:43177–43201.
9. Zhang D, Zhang W, He B *et al.* DNAGPT: a generalized pre-trained tool for multiple DNA sequence analysis tasks. *bioRxiv* 2023;**2023**.07.11.548628.
10. Fishman V, Kuratov Y, Shmelev A *et al.* GENA-LM: a family of open-source foundational DNA language models for long sequences. *Nucleic Acids Res***53**:gkae1310.
11. Nguyen E, Poli M, Durrant MG *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* 2024;**386**:eado9336.
12. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;**22**:287–97.
13. Wang K, Zeng X, Zhou J *et al.* BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning. *Brief Bioinform* 2024;**25**:bbae195.
14. Di Pierro M, Cheng RR, Aiden EL *et al.* De novo prediction of human chromosome structures: epigenetic marking patterns encode genome architecture. *Proc Natl Acad Sci USA* 2017;**114**:12126–12131.
15. Long W, Houcheng S, Xiong J *et al.* MutBERT: probabilistic genome representation improves genomics foundation models. *Bioinformatics***41**:i294–303.
16. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: Guyon I, VonLuxburg U, Bengio S *et al.* (eds.), *Advances in Neural Information Processing Systems*, Vol.**30**. Red Hook, NY: Curran Associates, Inc., 2017.
17. Jianlin S, Murtadha Ahmed YL, Pan S *et al.* RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* 2024;**568**:127063.
18. Dao T, Dan F, Ermon S *et al.* FlashAttention: fast and memory-efficient exact attention with IO-awareness. *Adv Neural Inf Proces Syst* 2022;**35**:16344–16359.
19. Goldfarb T, Kodali VK, Pujar S *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2025;**53**:D243–57.
20. Wang Z, Li Z, Huang T *et al.* Genomic insights into the demographic history and local adaptation of wild boars across eurasia. *Cell. Genomics* 2025;**5**:100954.
21. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Rush A, Levine S, Livescu K, Mohamed S (eds.), *International Conference on Learning Representations*, New Orleans, LA, USA: OpenReview.net, 2019.
22. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*. Erk K, Smith NA (eds.), pp. 1715–25. Berlin, Germany: Association for Computational Linguistics, 2016.
23. Press O, Smith N, Lewis M. Train short, test long: attention with linear biases enables input length extrapolation. In: Liu Y, Finn C, Choi Y, Deisenroth M (eds.), *International Conference on Learning Representations*, OpenReview.net. 2022.
24. Chowdhery A, Narang S, Devlin J *et al.* PaLM: scaling language modeling with pathways. *J Mach Learn Res* 2023;**24**:1–113.
25. Byrska-Bishop M, Evani US, Zhao X *et al.* High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* 2022;**185**:3426–40.
26. Benegas G, Albors C, Aw AJ *et al.* A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 2025;**43**:1960–65.
27. Devlin J, Chang M-W, Lee K *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds.), *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. 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics, 2019.
28. Zaheer M, Guruganesh G, Dubey KA *et al.* Big bird: Transformers for longer sequences. In: Larochelle H, Ranzato M, Hadsell R *et al.* (eds.), *Advances in Neural Information Processing Systems*, Vol.**33**, pp. 17283–97. Red Hook, NY: Curran Associates, Inc., 2020.
29. Pan Z, Yao Y, Yin H *et al.* Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun* 2021;**12**:5848.
30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics***26**:841–2.
31. Jurman G, Riccadonna S, Furlanello C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLoS ONE* 2012;**7**:e41882.

Briefings in Bioinformatics, 2026, **27**, bbag319
https://doi.org/10.1093/bib/bbag319
Problem Solving Protocol