animals

MDPI

Article

PorcineAI-Enhancer: Prediction of Pig Enhancer Sequences Using Convolutional Neural Networks

Ji Wang [1], Han Zhang [1], Nanzhu Chen [2], Tong Zeng [1], Xiaohua Ai [1] and Keliang Wu [1,*]

[1] College of Animal Science and Technology, China Agricultural University, Beijing 100193, China; ji.wang@cau.edu.cn (J.W.); hanzhang@cau.edu.cn (H.Z.); s20223040677@cau.edu.cn (T.Z.); sy20203040845@cau.edu.cn (X.A.)
[2] Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China; nzc1417@163.com
* Correspondence: liangkwu@cau.edu.cn

Simple Summary: This study develops a deep learning framework called PorcineAI-enhancer to predict enhancer sequences in pigs. Enhancers play a key role in regulating gene expression. However, identifying enhancers experimentally remains challenging. This study constructs a reliable pig enhancer dataset by integrating multiple data sources. The PorcineAI-enhancer model employs convolutional neural networks to extract features from DNA sequences and classify them into enhancers or non-enhancers. Evaluation on an independent test set shows the model has excellent performance. It also demonstrates strong predictive capability on tissue-specific enhancers from human and pig. This tool facilitates research on gene regulation mechanisms in pigs. It provides valuable resources to understand complex traits related to agriculture and biomedicine.

Abstract: Understanding the mechanisms of gene expression regulation is crucial in animal breeding. Cis-regulatory DNA sequences, such as enhancers, play a key role in regulating gene expression. Identifying enhancers is challenging, despite the use of experimental techniques and computational methods. Enhancer prediction in the pig genome is particularly significant due to the costliness of high-throughput experimental techniques. The study constructed a high-quality database of pig enhancers by integrating information from multiple sources. A deep learning prediction framework called PorcineAI-enhancer was developed for the prediction of pig enhancers. This framework employs convolutional neural networks for feature extraction and classification. PorcineAI-enhancer showed excellent performance in predicting pig enhancers, validated on an independent test dataset. The model demonstrated reliable prediction capability for unknown enhancer sequences and performed remarkably well on tissue-specific enhancer sequences.The study developed a deep learning prediction framework, PorcineAI-enhancer, for predicting pig enhancers. The model demonstrated significant predictive performance and potential for tissue-specific enhancers. This research provides valuable resources for future studies on gene expression regulation in pigs.

Keywords: enhancer; convolutional neural networks; sequence classification

Citation: Wang, J.; Zhang, H.; Chen, N.; Zeng, T.; Ai, X.; Wu, K. PorcineAI-Enhancer: Prediction of Pig Enhancer Sequences Using Convolutional Neural Networks. *Animals* **2023**, *13*, 2935. https://doi.org/10.3390/ani13182935

Academic Editors: Jie Yang and Xueyan Zhao

Received: 20 July 2023
Revised: 21 August 2023
Accepted: 5 September 2023
Published: 15 September 2023

1. Introduction

Understanding how complex gene expression patterns are regulated is a fundamental question in biology. At the core of this question is the genome-wide identification and characterization of cis-regulatory sequences that influence the expression of protein-coding genes and long non-coding RNA genes [1]. Cis-regulatory DNA sequences play a crucial role in the regulation of gene expression. These sequences, which can be located far away from gene promoters, have been shown to have significant effects on gene expression, sometimes resulting in up to a 100-fold increase in expression [2]. Enhancers, silencers, insulators, and tethering elements are examples of cis-regulatory sequences [3]. Among

them, enhancers and their associated transcription factor proteins are particularly important in the regulation of gene expression [4]. Enhancers are genomic regions that function as major regulatory elements controlling gene expression. They play a key role in cell-type-specific gene expression programs by forming physical interactions, often over long distances, with the promoters of their target genes [5]. Multiple enhancers, located tens or hundreds of thousands of nucleotides away from their target genes, loop to their respective gene promoters and work in coordination to regulate the expression of their common target gene [5].

Enhancers can synergistically interact with functional elements such as promoters and silencers, greatly influencing the spatial and temporal expression and transcription frequency of their target genes [6]. Enhancers possess three main characteristics: firstly, enhancers exhibit specificity, meaning that each enhancer only functions in a limited number of cell types or tissues [7]. This uniqueness limits the impact of mutations on their function. Secondly, enhancers are bidirectional, meaning that enhancers can be located upstream or downstream of the promoters they activate [8]. Additionally, enhancers exhibit the feature of long-range action, where the distance between an enhancer and its activated promoter can vary [9].

These characteristics of enhancers bring about many challenges in their identification. Although high-throughput experimental techniques have achieved significant success in enhancer identification [10–12], the enormous number of experimental conditions resulting from the diverse activity of enhancers in different cell tissues presents a challenge [13–15]. It is not practical to experimentally verify the existence of all enhancers in thousands of tissues or cells, which are the two drawbacks of this method: time-consuming and expensive [16].

To overcome the limitations of high-throughput experimental techniques, computational approaches have emerged, including those based on genome comparison [17] and machine learning [18,19]. Enhancers can exist in any region of the genome, making it difficult to find a linear pattern for enhancer identification through genome comparison methods [20].

Deep learning, a hot topic in the field of machine learning, possesses powerful learning capabilities that outperform various algorithms [21–23]. As a result, it has been widely applied in cutting-edge disciplines such as computer vision [24–26] and speech recognition [27–29]. In the field of gene sequences, deep learning has been proven to be a highly effective method for enhancer prediction [16,30–35]. Unlike traditional machine learning methods, deep learning constructs multi-layer neural networks to learn feature representations, enabling the automatic learning of higher-level abstract features [36] and better handling of nonlinear data [37]. Therefore, deep learning methods can leverage more information for enhancer prediction while reducing reliance on feature engineering. In recent years, deep learning methods have achieved significant success in predicting human enhancers [38–40] and have gradually become one of the methods for enhancer prediction.

In animal husbandry, to maintain food and agricultural production while minimizing negative environmental impacts, understanding the molecular mechanisms underlying economically important complex traits in farm animals is crucial for achieving biology-driven breeding biotechnologies. The domestic pig (Sus scrofa), a cornerstone of economic food security and international trade for many countries, holds significant positions in both animal husbandry and the biomedical field. Identifying enhancer regions in pigs bears great importance for advancing livestock farming and biomedical research. Enhancers play a pivotal role in gene regulation, and linking them with genes, SNPs, SVs, or other regions of interest holds the promise of providing valuable insights into the regulation of complex production traits and adaptive characteristics. It's worth emphasizing that pigs are not only of interest due to their production traits but also due to their physiological similarities with humans. They are widely used as large animal models for preclinical research [41,42] and as xenotransplantation donors [43,44]. These features make pigs vital resources in the field of research, rendering a deeper understanding of the regulatory mechanisms of the pig genome immensely valuable for driving scientific research and applications.

Some studies [45–48] have utilized two enhancer-associated histone modifications, H3K27ac and H3K4me1 [7,49–51]. However, these studies come with high implementation costs, are limited by tissue/cell types and genetic backgrounds, and still require substantial effort and funding to determine whether these regions indeed possess actual enhancer functionality, as demonstrated through regulatory assays in transgenic mice [52].

While there have been cost-saving efforts in using deep learning to predict enhancer sequences in humans and mice [53–55], the progress in predicting enhancer sequences in livestock remains relatively limited due to the lack of a publicly accessible and reliable enhancer database for livestock species. As a result, deep learning models have not been widely applied to predict enhancer sequences in livestock, particularly in poultry and other livestock species.

In this study, we took the first step in addressing this issue by utilizing a publicly available pig enhancer database to construct a trustworthy dataset of enhancer and non-enhancer sequences. Subsequently, we employed this dataset to train a deep learning framework, named PorcineAI-enhancer, for enhancer prediction in pig genomic sequences. The main idea behind this model is to combine one-hot encoding and k-mer encoding to represent sequence data and then use CNN to extract features and perform classification, thus determining whether a sequence belongs to an enhancer region. Experimental results on an independent test dataset demonstrate the excellent performance of this method.

We have made our PorcineAI-enhancer code and data freely available on GitHub repository: https://github.com/castwj/PorcineAI-enhancer (accessed on 10 August 2023), facilitating accessibility and encouraging further research and collaboration in this field.

2. Materials and Methods

2.1. Data Preparation

Although several scientists have conducted ChIP-seq experiments to explore enhancer elements in pigs [47,56–58], most of these experiments lack collaborative support from other high-throughput data and functional validation in the laboratory. This phenomenon has resulted in a limited number of reliable enhancers in pigs, with varying quality.

To construct a deep learning model, there is an urgent need to create a high-quality and highly reliable enhancer database for pigs. In this study, we collected relevant enhancer sequence information from three different sources. Firstly, MacPhillamy et al. [59] utilized transfer learning methods and high-quality enhancer data from VISTA [52] and publicly available human and mouse ChIP-seq data to study enhancer functionality in three non-model mammalian species (cattle [45,60–62], pigs [45,47], and dogs [62]). By combining this data with species-specific ChIP-seq data, they obtained a high-confidence enhancer list. Secondly, the Functional Annotation of Animal Genomes (FAANG) project [63] is an international collaborative initiative aimed at systematically annotating animal genomes. The project employs a variety of high-throughput techniques and bioinformatics methods to comprehensively annotate the genomes of various animal species, revealing their functions and regulatory mechanisms. Recently, Pan et al. [64] integrated 223 epigenomic and transcriptomic datasets to create a comprehensive catalog of regulatory elements in pigs (Sus scrofa). We extracted enhancer information from this catalog for all tissues and merged it. Additionally, the EnhancerAtlas 2.0 database [65] is a multi-species public database that contains 13,494,603 enhancers from 16,055 datasets.

The aforementioned three datasets collectively constitute the enhancer data sources used in this study. It's worth noting that the EnhancerAtlas 2.0 database [65] uses the Sscrofa10.2 reference genome, while the other two datasets use Sscrofa11.1 [66]. Since the conversion of BED files between different reference genomes can potentially lead to the loss of some sequences, we opted to use pig iPSC and heart enhancer information from the EnhancerAtlas 2.0 database [65] as our test data. This choice aims to assess the PorcineAI-enhancer model's ability to recognize tissue-specific enhancers. Additionally, the EnhancerAtlas provides human iPSC and heart enhancer information (using hg19 as the reference genome). Considering the relatively close genetic relationship between humans

and pigs, we can compare cross-species tissue-specific enhancer prediction capabilities to explore the model's reliability.

2.2. High Confidence Sequence Acquisition

To construct an effective and robust model, we followed strict criteria to establish the dataset. To obtain high-quality and reliable enhancer sequences, we combined enhancer sequences obtained from MacPhillamy et al. [59] and Pan et al. [64]. We processed the BED files from these two datasets to obtain the overlapping fragments, which served as the initial enhancer sequences. We retained only sequences with a length more than 200 bp. Then, based on the length requirements of model, we divided the sequences into fixed-length (200 bp) fragments. Sequences shorter than 200 bp were discarded. We used Bedtools [67] to obtain the sequences, resulting in a total of 7633 enhancer sequences.

To provide an intuitive representation of the enhancer dataset used by the model within the context of the original datasets, we employed an enhancer source Venn diagram, Figure 1. This diagram effectively illustrates the overlapping and non-overlapping portions of enhancers from different sources.

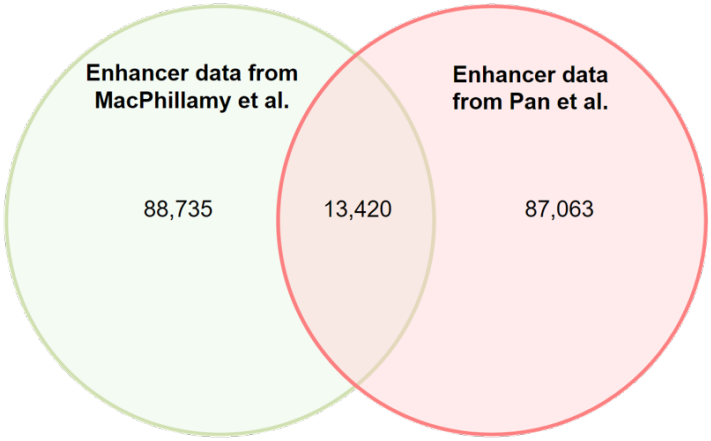

Figure 1. Enhancer Source Venn Diagram. Each circle representing a specific source, the overlapping regions indicate the common enhancers shared between the sources, while the non-overlapping regions represent the unique enhancers specific to each source. MacPhillamy et al. [59] and Pan et al. [64].

Regarding the non-enhancer sequences, previous studies often randomly extracted genomic fragments as negative samples [30], ensuring that their length distribution and quantity were the same as the enhancer sequences. However, considering the potential misclassification of some tissue-specific enhancers as non-enhancer sequences due to experimental design limitations, to ensure the reliability and practicality of the non-enhancer sequences, we initially utilized the gene annotation file of Sus11.1 [66] to extract fundamental genomic information. This encompassed gene annotations (such as protein-coding genes and long non-coding RNAs) and promoter regions (defined as 2 kb regions centered around the transcription start site of protein-coding genes). These sequences were intentionally selected as they represent regions that are unlikely to be enhancers.

We then combined these sequences with the enhancer sequence regions covered by the databases employed in our study. Using Bedtools [67], we filtered out the remaining genomic regions. Lastly, we randomly selected 7633 segments from these filtered regions to constitute the non-enhancer dataset used for training.

To reduce sequence similarity, we employed the Cd-hit [68–70] tool to remove redundant sequences with a similarity exceeding 80%. Finally, we used the resulting non-redundant sequences as the samples for the reference dataset.

2.3. Sequence Coding Method

In many deep learning algorithms used for processing biological sequences, natural language processing techniques are commonly employed to extract features from raw DNA sequences [71–73]. In our CNN model, we utilized a method called One-hot Encoding and k-mer descriptors to encode each input sequence. Each enhancer sequence in this study consists of four bases, adenine (A), guanine (G), cytosine (C), and thymine (T), with a length of 200 bp.

In the One-hot Encoding of genetic sequences, we represent each base as a one-hot vector of length four, where only one element is 1, and the rest are 0. For example, adenine (A) is represented as [1, 0, 0, 0], cytosine (C) is represented as [0, 1, 0, 0], and so on. Thus, the genetic sequence can be represented as a concatenation of a series of one-hot vectors, where each one-hot vector represents a base.

K-mer encoding is a method for converting protein or DNA sequences into vector representations. It treats every consecutive k characters (or letters) in the sequence as a unit and represents each unit as a numeric vector. When the step size is 1, a DNA sequence of length l can be divided into $(l - k + 1)$ k-mers. For example, when k = 2, the sequence 'ACGTCGACG' will be divided into seven 2-mers: "AC", "CG", "GT", "TC", "GA", "AC", "CG". This representation makes the sequence easier to compute and understand. We treat the entire DNA sequence as a sentence and the k-mer fragments as words. These vectors can be used for various bioinformatics tasks such as classification, clustering, sequence alignment, and pattern recognition.

One drawback of k-mer encoding is that it may lose some contextual information of the sequence since it divides the sequence into independent k-mer units. Additionally, k-mer encoding can be influenced by the sequence length and the chosen value of k, requiring optimization based on specific circumstances. To mitigate the impact of k-mer encoding on the results, in this study, we set the values of k to 1, 2, and 3 respectively, to strike a better balance between contextual information and computational efficiency.

To combine the One-hot and k-mer representations and form the inputs to our model, we concatenated them together, resulting in a comprehensive feature vector that captures both the nucleotide composition and sequential patterns present in the DNA sequence. This hybrid approach enables us to leverage the fine-grained information captured by the one-hot encoding and the higher-order patterns captured by the k-mer encoding simultaneously.

Specifically, the shape of the concatenated feature vector is $(4 + 1 + 2 + 1) \times$ SAMPLE LENGTH, where the first four rows correspond to the one-hot encoding of the sequence, the fifth row corresponds to the 1-mer features, the sixth and seventh rows correspond to the 2-mer features considering both left and right directions, and the last row corresponds to the 3-mer features. This comprehensive feature representation effectively captures the individual nucleotide composition and higher-level sequence patterns present in the enhancer sequence. SAMPLE LENGTH represents the chosen length of the sequence.

Finally, we convert this concatenated feature vector into a PyTorch tensor and use it as input, along with the corresponding labels, to the neural network model. This enables the model to learn from the combined information of one-hot and k-mer representations and make accurate predictions.

2.4. Sequence Analysis

Sequence analysis is a computational approach used to analyze biological sequences, such as DNA, RNA, and protein sequences. It helps researchers understand the patterns and structures of biological sequences and enables analysis and comparison of these sequences to reveal information about their functions, structures, and evolution.

SeqLogo [74] is a commonly used sequence analysis tool for visualizing conservation and variation information in DNA, RNA, or protein sequences. SeqLogo graphs typically represent the information entropy of each base or amino acid at each position in the sequence using the height of the corresponding letter. Higher information entropy indicates less conservation at that position.

In a SeqLogo graph, the height at each position reflects the distribution of different bases or amino acids at that position. If a specific base or amino acid is highly prevalent (high frequency) at a particular position, the height at that position will be higher, indicating higher conservation. Conversely, if there are multiple different bases or amino acids at a certain position, the height will be lower, indicating higher variation.

By visualizing conservation and variation through SeqLogo graphs, researchers can quickly gain insights into the conservation and variation within a sequence, aiding in the analysis and interpretation of its function and structure. SeqLogo graphs are commonly used to identify conserved motifs, functional sites, and important sequence features.

2.5. CNN Model Architecture

The PorcineAI-enhancer model we propose is a convolutional neural network designed for identifying pig genomic enhancer and non-enhancer sequences. The model consists of two convolutional blocks, each comprising three convolutional layers followed by a batch normalization layer, with a max-pooling layer after each convolutional block. The first convolutional block has 32 output channels for its convolutional layers, with a kernel size of 4×4 and a padding of 1. The second convolutional block is similar to the first one, but with an increased output channel size of 64.

After the convolutional blocks, the model flattens the output and processes it through a fully connected layer with a ReLU activation function and a size of 256. Finally, the output passes through a sigmoid activation function and a linear layer to generate scalar output.

The model utilizes binary cross-entropy loss as the training criterion. The forward method of the model takes an input tensor and outputs a tensor of the same size, representing the predicted output for each input sample. The model is trained through backpropagation and optimization algorithms to minimize the loss function and improve the accuracy of predictions. Table 1 below shows the variations in model parameters across the layers of the CNN model.

Table 1. Variations in CNN Model Parameters Across Layers.

Layer (Type)	Output Shape	Param
Conv1d-1	[−1, 32, 200]	800
BatchNorm1d-2	[−1, 32, 200]	64
Conv1d-3	[−1, 32, 200]	3104
BatchNorm1d-4	[−1, 32, 200]	64
Conv1d-5	[−1, 32, 200]	3104
BatchNorm1d-6	[−1, 32, 200]	64
MaxPool1d-7	[−1, 32, 50]	0
Conv1d-8	[−1, 64, 50]	6208
BatchNorm1d-9	[−1, 64, 50]	128
Conv1d-10	[−1, 64, 50]	12,352
BatchNorm1d-11	[−1, 64, 50]	128
Conv1d-12	[−1, 64, 50]	12,352
BatchNorm1d-13	[−1, 64, 50]	128
MaxPool1d-14	[−1, 64, 12]	0
Linear-15	[−1, 256]	196,864
Linear-16	[−1, 1]	257

These parameter variations provide insights into the number of parameters in the model and the shape changes between layers. This aids in understanding the complexity of the model and the distribution of parameters, as well as the changes that may occur during the training and optimization processes.

2.6. K-Fold Cross-Validation

K-fold cross-validation is a commonly used model evaluation method [75]. It involves splitting the dataset into k non-overlapping subsets or folds, and then iteratively using

each fold as the validation set and the remaining k-1 folds as the training set to train the model. The evaluation results from each iteration are then aggregated to obtain the average performance of the model.

K-fold cross-validation can effectively reduce overfitting by utilizing more data for model training and providing a comprehensive evaluation of the model's performance. Additionally, it helps in selecting the best model hyperparameters, such as regularization parameters and learning rates, to improve the model's generalization ability.

In the training of the PorcineAI-enhancer model, we first randomly divide the training set into five folds or partitions using stratified sampling, as illustrated in Figure 2. Each fold is used as the validation set in turn, while the remaining four folds are used as the training set for training the CNN model. Then, the five trained CNN models are combined to form an ensemble model. Next, the ensemble model is used to test the samples in an independent test set. This entire process, including data partitioning, model training, and model testing, is repeated five times to observe the variation in model performance across the five experiments.

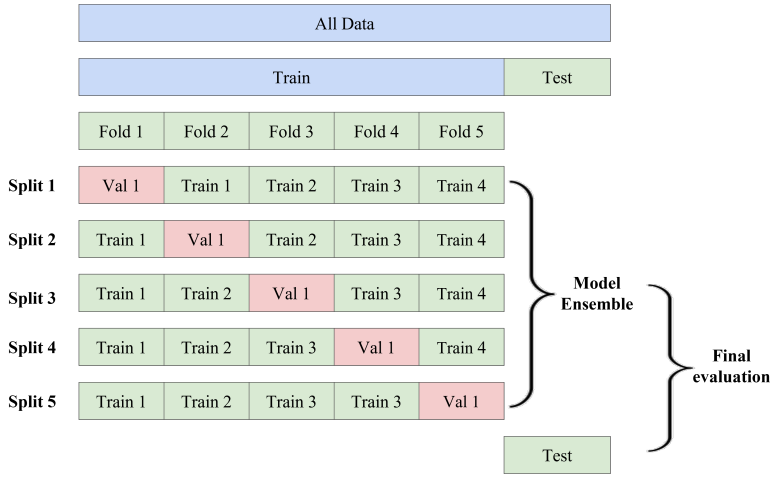

Figure 2. Training and Validation Process for PorcineAI-Enhancer Model using Stratified Sampling and Ensemble Learning. This figure illustrates the training and validation process for the PorcineAI-Enhancer model. The training set is randomly divided into five folds or partitions using stratified sampling, allowing for a balanced representation of the data in each fold. Each fold is then used as the validation set in turn, while the remaining four folds are used as the training set for training the Convolutional Neural Network (CNN) model. The five trained CNN models are combined to form an ensemble model, which is used to test the samples in an independent test set. This entire process, including data partitioning, model training, and model testing, is repeated five times to observe the variation in model performance across the five experiments. The use of stratified sampling and ensemble learning helps to improve the accuracy and robustness of the PorcineAI-Enhancer model.

By employing k-fold cross-validation, we can comprehensively evaluate the performance of the PorcineAI-enhancer model and observe how it performs with different combinations of training and validation sets. This approach helps obtain more reliable performance evaluation results and provides guidance for further improvements to the model.

3. Results

3.1. Sequence Analysis

In the SeqLogo plot, the vertical axis can be scaled using frequency or bits. When the frequency is used as the vertical axis, the SeqLogo plot displays the frequency of occurrence for each type of base or amino acid at each position. The higher the frequency, the taller the letter, indicating a more conserved base or amino acid at that position. Conversely, the lower the frequency, the lower the letter, indicating a more variable base or amino acid at that position. When bits are used as the vertical axis, the SeqLogo plot represents the

information entropy of bases or amino acids at each position. The higher the information entropy, the taller the letter, indicating a less conserved base or amino acid at that position.

Our results are presented in Figure 3. When the frequency is used as the vertical axis, the distribution of enhancer sequences and non-enhancer sequences is nearly identical. However, when bits are used as the vertical axis, they exhibit noticeable differences in their distribution.

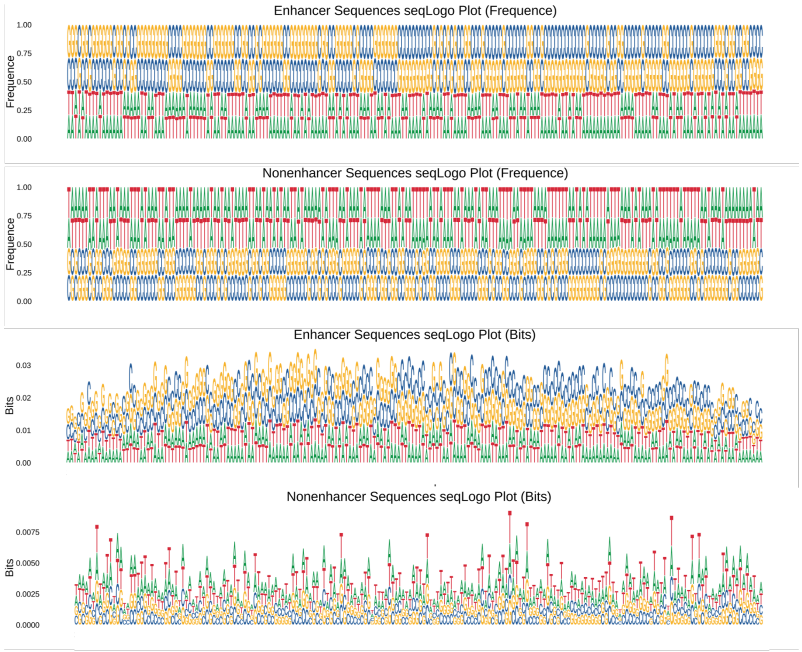

Figure 3. Differences in Information Entropy of Enhancer and Non-Enhancer Sequences Revealed by SeqLogo Analysis. In this figure, we show the results of SeqLogo analysis, which is a graphical representation of the conservation and variation of nucleotide or amino acid sequences. The vertical axis of the SeqLogo plot can be scaled using frequency or bits. Our analysis reveals that enhancer sequences and non-enhancer sequences exhibit significant differences in their information entropy when bits are used as the vertical axis. This indicates that enhancer sequences and non-enhancer sequences possess distinct characteristics in terms of sequence conservation and variation, which may be associated with their different roles in gene expression regulation. These findings provide further insights into the functional differences between enhancer and non-enhancer sequences and may have implications for understanding the mechanisms of gene expression regulation.

These findings suggest that there are significant differences in the information entropy between enhancer and non-enhancer sequences. This indicates that enhancer sequences and non-enhancer sequences possess distinct characteristics in terms of sequence conservation and variation. These characteristics may be associated with their different roles in gene expression regulation.

3.2. PorcineAI-Enhancer Model Training

We conducted model training for the PorcineAI-enhancer model. As depicted in the Figure 4, it provides a more intuitive overview of the training process for the PorcineAI-enhancer model. As illustrated in Figure 1, Model 1 refers to the training configuration where data from Fold 2–5 is employed as the training dataset, and Fold 1 serves as the validation dataset. The model is built using the parameters that exhibit the best performance on the validation set. Similarly, Model 2–5 follow this pattern, each involving a specific fold for validation while the remaining folds are utilized for training. A total of 50 epochs, where each epoch represents a complete iteration through the dataset, were carried out for training. Throughout the training process, a learning rate of 1×10^{-5} was utilized—an essential hyperparameter controlling the step size for model parameter updates. To optimize the

model's training, the Adam optimizer was chosen. Adam is a commonly used adaptive learning rate optimization algorithm that dynamically adjusts the learning rate based on estimates of the first and second moments of the gradients.

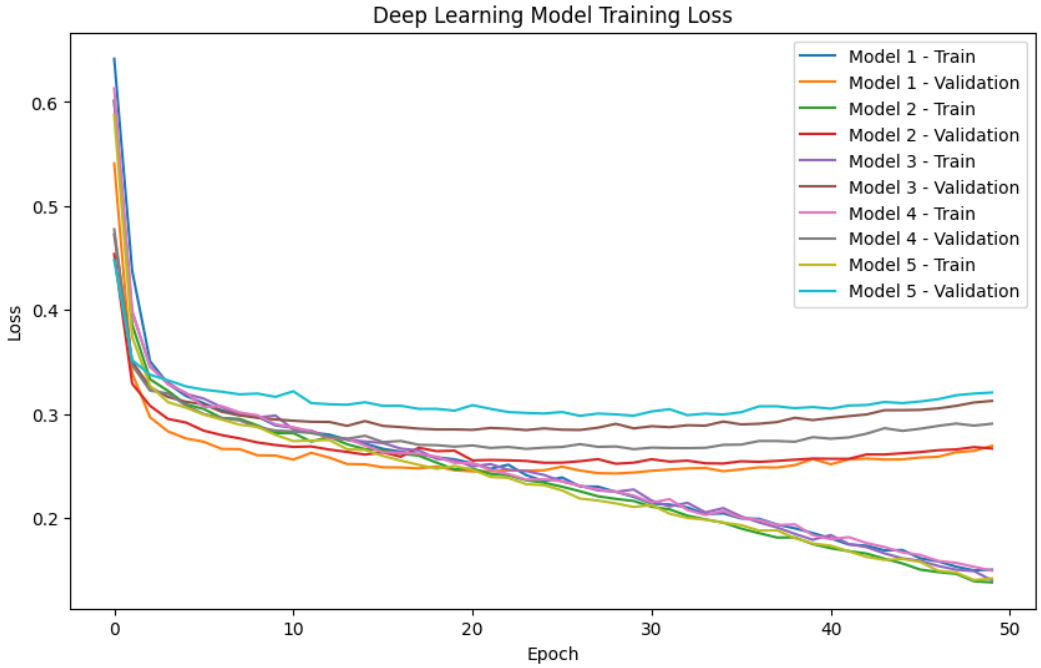

Figure 4. PorcineAI-enhancer model training loss curves. The horizontal axis represents the number of training epochs, and the vertical axis represents the model's loss value. The loss value is a metric that measures the difference between the model's predictions and the actual labels. Our goal is to minimize the loss value through training. We observe two different loss curves. One is the loss curve on the training set, which indicates the model's fit to the training data. The other curve is the loss curve on the validation set, which represents the model's performance on unseen data. We use the validation set to evaluate the model's generalization ability in real-world scenarios. Typically, we select the epoch corresponding to the minimum validation set loss as the optimal model's parameters.

From the figure, it is observable that after around 20 epochs, the model's performance on the validation set had already reached its peak. This indicates that while the model might potentially achieve better scores on the training set, further training is detrimental to performance improvement. This phenomenon suggests the occurrence of overfitting, where the model overly adapts to the training data and subsequently performs poorly on new data. To counteract overfitting, ensuring the model's generalization ability, we opted to utilize the parameters from the epoch at which each model performed best on the validation set as the parameters for the Ensemble model. This approach enables us to attain better predictive performance on previously unseen data, enhancing model stability and reliability.

3.3. Performance of the PorcineAI-Enhancer Model

Through 5-fold cross-validation on the training set, we obtained 5 validated CNN models. These models were then evaluated on independent test sets, and the evaluation parameters are presented in Table 2.

From Table 2, it can be observed that the accuracy of the models ranges from 0.905 to 0.911, with a very small standard deviation, indicating their ability to accurately classify samples. As for the AUC metric, all values exceed 0.939, with the highest AUC value being 0.946, demonstrating the models' high capability in discriminating between positive and negative samples. The higher AUC values suggest effective classification of positive and negative samples and demonstrate strong predictive performance.

Table 2. Performance Evaluation of CNN Models for Enhancer Prediction.

Model	Accuracy Score	AUC Score	Sensitivity	Specificity
Model 1 (Parts 2, 3, 4, 5 : Part 1)	0.909626719	0.939438503	0.963326785	0.855926654
Model 2 (Parts 1, 3, 4, 5 : Part 2)	0.910936477	0.944208139	0.974459725	0.847413229
Model 3 (Parts 1, 2, 4, 5 : Part 3)	0.910609037	0.94386183	0.965291421	0.855926654
Model 4 (Parts 1, 2, 3, 5 : Part 4)	0.910936477	0.940875633	0.964636542	0.857236411
Model 5 (Parts 1, 2, 3, 4 : Part 5)	0.904715128	0.94601431	0.948264571	0.861165684
Ensemble Model	0.916502947	0.948383796	0.974459725	0.858546169

By referring to Figure 5, we can observe that the evaluation metrics of the five models exhibit consistent distribution, with specificity being the lowest. Considering that our acquisition of non-enhancer sequences did not undergo experimental verification but rather aimed to remove known functional sequences, the lower specificity may be attributed to the presence of false negatives in the non-enhancer sequences. Nevertheless, the evaluation parameters of all models indicate that each model possesses sufficient capability to predict whether a sequence is an enhancer, underscoring the reliability of our construction of the original training data.

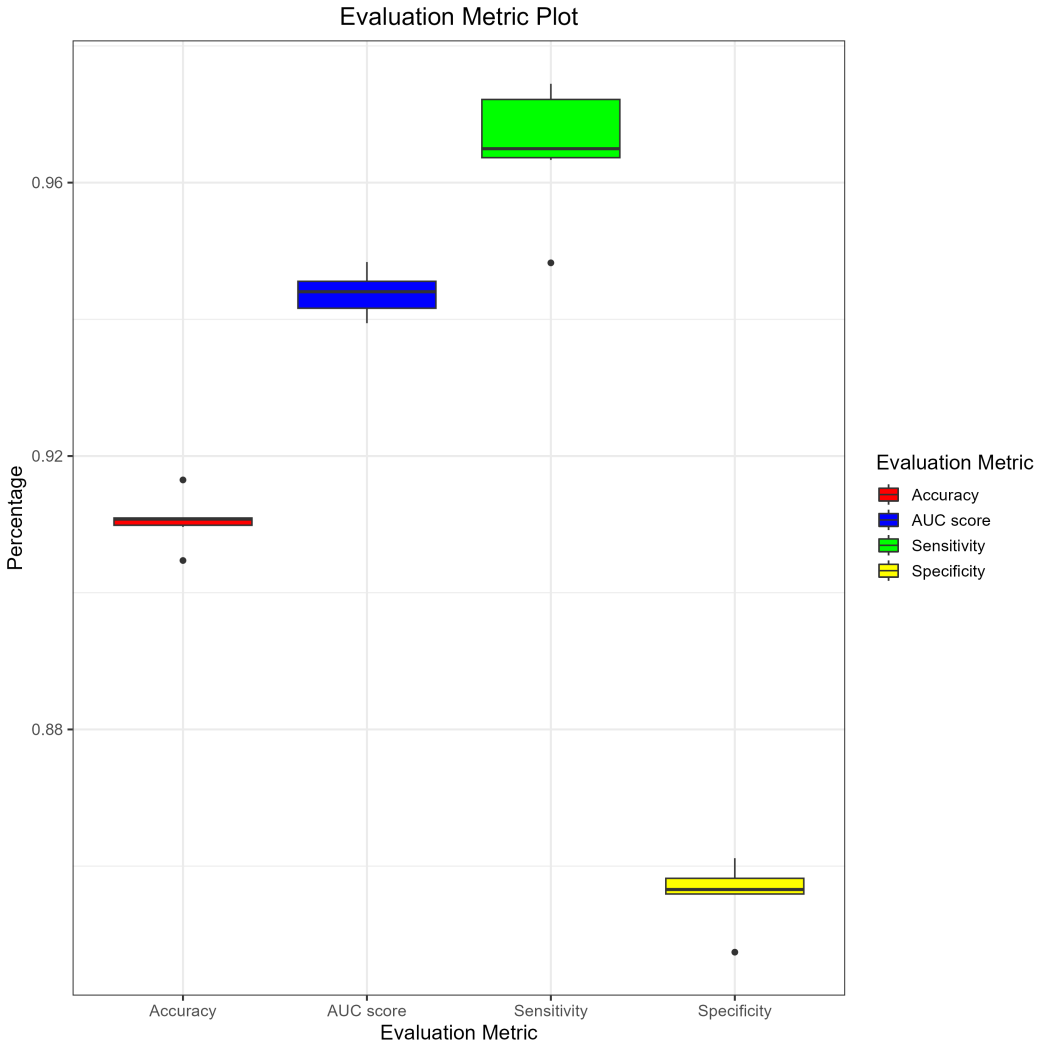

Figure 5. Robust Performance of Deep Learning Models in Predicting Enhancer Sequences. We present the evaluation metrics of five deep learning models in predicting enhancer sequences. The models demonstrate high accuracy and AUC values, indicating their capability in discriminating

between positive and negative samples. The evaluation metrics exhibit consistent distribution, with specificity being the lowest, which may be attributed to the presence of false negatives in the non-enhancer sequences. Nevertheless, all models possess sufficient capability to predict whether a sequence is an enhancer, demonstrating the reliability of our construction of the original training data. These findings support the effectiveness and feasibility of the proposed method and highlight the robustness of the features and patterns learned by the deep learning models during the training process. The robust performance of the models suggests their potential applications in predicting enhancer sequences and advancing our understanding of gene expression regulation.

These excellent evaluation metrics further substantiate the effectiveness and feasibility of the proposed method. The models achieve satisfactory results across multiple indicators, highlighting the robustness of the features and patterns learned by the deep learning models during the training process.

3.4. Comparison with Ensemble Model

Given the excellent sequence prediction capabilities exhibited by each individual model, but with some variations, we decided to further improve the predictive performance by using model ensembles. The advantage of ensemble models lies in their ability to leverage the strengths of multiple models, resulting in higher accuracy and stronger discrimination. By combining the effects of the ensemble models, we can obtain more reliable and stable prediction results.

Therefore, we constructed an ensemble model using the predictions from each individual model, and its model evaluation parameters are presented in Table 2. It is evident that the ensemble model outperforms the individual models in terms of accuracy score and AUC metrics. The ensemble model also demonstrates advantages in terms of sensitivity and specificity. The sensitivity of the ensemble model is the same as the best individual model, both achieving a value of 0.9745. This indicates that the model is highly sensitive in detecting true positive samples and avoids misclassifying them as negative samples. This is crucial in many real-world scenarios where the focus is on true positive cases. From the perspective of these evaluation metrics, the ensemble model exhibits significant advantages over the individual models, providing more reliable and accurate predictive performance.

The Figure 6 below presents the AUC curves plotted for each model on the test set. The AUC curve is a common tool for evaluating the performance of classification models. It illustrates the relationship between the true positive rate (Sensitivity) and the false positive rate (1-Specificity) at various thresholds. A value closer to 1 indicates superior performance of the model in classification tasks. Upon examining this graph, it is evident that the AUC curve of the Ensemble model slightly surpasses those of the other models. This indicates that the Ensemble model maintains a better balance between the true positive rate and the false positive rate at various thresholds.

These results further validate the effectiveness and feasibility of our proposed method. The ensemble model achieves satisfactory results across multiple metrics, showcasing the robustness of the features and patterns learned by the deep learning model during the training process. This also supports our research hypothesis and provides strong evidence for a deeper understanding of gene expression regulation.

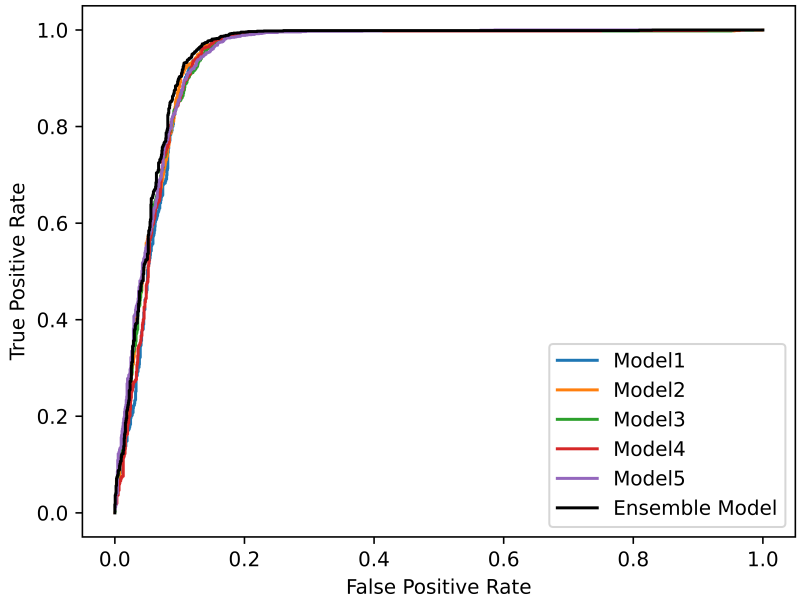

Figure 6. AUC Curves of Different Models. AUC score (Model 1 = 0.939438503, Model 2 = 0.944208139, Model 3 = 0.94386183, Model 3 = 0.940875633, Model 3 = 0.94601431, Ensemble Model = 0.948383796). A higher AUC score signifies that the model performs better across the entire range of decision thresholds, demonstrating its strong discriminative capability and overall effectiveness in distinguishing between positive and negative samples.

3.5. Comparison with Existing State-of-the-Art Methods

While the current state of enhancer prediction in pigs lacks documented research, a wealth of related studies exists in the human domain. To underscore the advancements of the PorcineAI-enhancer model, we chose to compare it with other renowned models in the realm of human enhancer prediction. Evaluation of the model's performance was conducted by analyzing several key metrics, as depicted in the Table 3 below.

Table 3. Result of comparison with existing state-of-the-art methods.

Method	ACC	AUC	SN	SP	Source
iEnhancer-2L	0.730	0.806	0.710	0.750	Liu et al., 2016 [30]
EnhancerPred	0.740	0.801	0.735	0.745	Jia and He, 2016 [76]
iEnhancer-EL	0.748	0.817	0.710	0.785	Liu et al., 2018 [77]
iEnhancer-EBLSTM	0.772	0.835	0.755	0.795	Niu et al., 2021 [78]
PorcineAI-enhancer	0.652	0.811	0.335	0.969	This study
PorcineAI-enhancer (human enhancer data)	0.769	0.832	0.785	0.752	This study

As anticipated, the model without fine-tuning exhibited noticeably lower accuracy in predicting enhancers compared to all models based on human enhancer data. While the accuracy in predicting non-enhancers was relatively higher, this is primarily due to the PorcineAI-enhancer model's inclination to label all human sequences as non-enhancers, thereby inflating its performance in non-enhancer prediction.

Upon training the model with human enhancer data, significant improvements across all metrics were observed for the PorcineAI-enhancer model, particularly in terms of AUC and MCC. Furthermore, when compared to the iEnhancer-EBLSTM method from 2021, our study's model slightly outperformed in terms of SN and MCC, but exhibited slightly lower performance in ACC, AUC, and SP. The comparative outcomes highlight that the PorcineAI-enhancer model excels over previous methods in many aspects. This superiority can be attributed to the inherent strengths of deep learning models, which are capable of more accurately capturing features and achieving higher efficiency in learning processes. The resultant model has more precise parameters, thereby achieving superior

performance outcomes. This further substantiates the effectiveness and potential of our proposed approach in enhancer prediction.

3.6. Model Performance on Tissue-Specific Enhancers

The generalization performance of deep learning models holds immense importance in practical applications. Due to the strong nonlinear fitting capability of deep learning models, issues such as training overfitting [79] and insufficient training data [80] can lead to a situation where the model performs well on the training data but poorly on new data. Overfitting causes the model to only capture the features of the training dataset without abstracting more general features, resulting in a loss of predictive ability in real-world scenarios [81]. Consequently, selecting an appropriate training set and effective model training are critical issues in the field of deep learning.

In this study, we meticulously selected enhancer sequences spanning different species and tissues as our training set to assess the model's generalization ability. These tissue-specific enhancer datasets originated from the EnhancerAtlas 2.0 database [65], encompassing enhancer sequences from human and pig iPSCs as well as heart cells. We employed the trained ensemble model to predict these sequences and compiled the prediction results in Table 4.

Table 4. Performance of the Ensemble Model on Tissue-Specific Enhancer Datasets.

Tissue	Pig	Human
Heart	0.8240	0.7031
iPSC	0.2606	0.3146

Analyzing the data in Table 4, it's evident that the model maintains remarkable generalization performance in heart tissue, achieving prediction accuracies of 0.8240 and 0.7031 for pig and human heart enhancer sequences, respectively. These outcomes explicitly demonstrate the model's generalization capacity, performing well even on test data, indicating that the model isn't merely overfitting to the training data. These results not only enhance our confidence in applying the model to a broader range of pig cell tissues but also establish a solid foundation for cross-species enhancer prediction.

However, it's noteworthy that the accuracy of the model in predicting pig and human iPSC enhancer sequences is relatively lower, at 0.2606 and 0.3146, respectively. Importantly, iPSCs are undifferentiated cell types, and their gene regulatory mechanisms might differ significantly from those of mature tissue cells. Therefore, the model's suboptimal performance on such cells doesn't necessarily reflect a weak generalization capability of the model. This further underscores the limitations of the model's applicability and provides insights for future improvements.

4. Discussion

We proposed the PorcineAI-enhancer framework, which leverages deep learning techniques and addresses the challenge of limited high-quality datasets, providing a valuable tool for predicting enhancers in pigs. The development of this framework, along with the construction of a high-quality enhancer database specifically tailored for pigs, represents a significant contribution to the field of enhancer prediction. However, the framework still has some limitations and potential areas for improvement.

Through performance evaluation on an independent test dataset, the PorcineAI-enhancer framework demonstrates excellent performance in enhancer prediction, showcasing its potential in predicting pig enhancers. These findings align with previous studies in human enhancer research [82–85], which indicate the effectiveness of deep learning in predicting enhancers across various species, including humans. Therefore, similar to model organisms such as humans, applying deep learning approaches to identify gene regulatory elements in livestock genomes could become a new paradigm in livestock breeding [86,87].

However, the dataset used in this study was constructed by integrating enhancer sequence information from various sources [59,64,65], including transfer learning methods, publicly available ChIP-seq data, and comprehensive regulatory catalogs. While we made efforts to ensure the reliability and quality of the data, potential biases and inconsistencies may still exist, particularly in the case of non-enhancer sequences where the random sampling approach we employed may have limitations. Hence, future research should consider incorporating more high-quality and well-validated enhancer datasets and non-enhancer datasets to further enhance the accuracy and generalizability of the framework.

Furthermore, we must acknowledge that although the PorcineAI-enhancer framework performs well in practical applications, this study still has its limitations. Firstly, due to the diversity of pig breeds [56], tissue specificity [57,88], and developmental stages in reality, further validation and verification of the enhancer prediction capability need to be conducted under controlled conditions to ensure the reliability of the predictions. This represents the next step for model improvement, namely fine-tuning by incorporating Chip-seq-detected enhancer sequences from different breeds, tissues, and cells, expanding its applicability to more refined application scenarios. By utilizing enhancer sequence data from different breeds, we can better understand the conservation and diversity of enhancer sequences across different species.

Additionally, our deep learning model has been widely used and performed well in previous studies. However, other deep learning model frameworks, such as attention mechanisms [89], can be employed to capture longer, more complex, and higher-level sequence features. From this perspective, further improving the PorcineAI-enhancer framework to enhance its performance represents a developmental direction for increasing the predictive capabilities of the model.

5. Conclusions

In conclusion, this study presents the development and evaluation of the PorcineAI-enhancer framework, a deep learning-based approach for enhancer prediction in pigs. The framework demonstrates excellent performance in identifying enhancer sequences and addresses the lack of high-quality datasets specific to pigs. The findings highlight the potential of deep learning techniques in enhancer prediction and contribute to the growing body of evidence supporting their effectiveness across species. The framework provides a valuable tool for researchers studying pig gene regulation and expression patterns, facilitating advancements in understanding the molecular mechanisms underlying pig traits and diseases. Despite the limitations and the need for further validation and improvement, the PorcineAI-enhancer framework represents a significant advancement in the field and sets the stage for future studies aiming to unravel the regulatory landscape of pigs and other species.

Author Contributions: Writing—original draft preparation, J.W.; Supervision, H.Z.; Data curation, N.C.; Writing—review & editing, T.Z. and X.A.; Funding acquisition and Project administration, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (award 2022ZD0115704), National Key Research and Development Program of China—A Research on the formation of important traits and environmental adaptability (award 2021YFF1000603).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and data can be freely accessed from the GitHub repository: https://github.com/castwj/PorcineAI-enhancer (accessed on 10 August 2023).

Acknowledgments: We gratefully acknowledge the support of the Natural Science Foundation of China and the National Engineering Laboratory for Animal Breeding. Their funding and resources have been instrumental in the completion of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schmitz, R.J.; Grotewold, E.; Stam, M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell* **2021**, *34*, 718–741. [CrossRef]
2. Beagan, J.A.; Pastuzyn, E.D.; Fernandez, L.R.; Guo, M.H.; Feng, K.; Titus, K.R.; Chandrashekar, H.; Shepherd, J.D.; Phillips-Cremins, J.E. Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression. *Nat. Neurosci.* **2020**, *23*, 707–717. [CrossRef] [PubMed]
3. Verheul, T.C.J.; van Hijfte, L.; Perenthaler, E.; Barakat, T.S. The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1. *Front. Cell Dev. Biol.* **2020**, *8*, 592164. [CrossRef] [PubMed]
4. Spitz, F.; Furlong, E.E.M. Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.* **2012**, *13*, 613–626. [CrossRef]
5. Schoenfelder, S.; Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **2019**, *20*, 437–455. [CrossRef]
6. Higgs, D.R. Enhancer–promoter interactions and transcription. *Nat. Genet.* **2020**, *52*, 470–471. [CrossRef] [PubMed]
7. Heintzman, N.D.; Ren, B. Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.* **2009**, *19*, 541–549. [CrossRef]
8. Bulger, M.; Groudine, M. Enhancers: The abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **2010**, *339*, 250–257. [CrossRef] [PubMed]
9. Visel, A.; Rubin, E.M.; Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **2009**, *461*, 199–205. [CrossRef]
10. Visel, A.; Blow, M.J.; Li, Z.; Zhang, T.; Akiyama, J.A.; Holt, A.; Plajzer-Frick, I.; Shoukry, M.; Wright, C.; Chen, F.; et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **2009**, *457*, 854–858. [CrossRef] [PubMed]
11. May, D.; Blow, M.J.; Kaplan, T.; McCulley, D.J.; Jensen, B.C.; Akiyama, J.A.; Holt, A.; Plajzer-Frick, I.; Shoukry, M.; Wright, C.; et al. Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **2012**, *44*, 89–93. [CrossRef] [PubMed]
12. Wang, F.; Zhang, Y.; Wu, F.; Gui, Y.; Chen, X.; Wang, Y.; Wang, X.; Gui, Y.; Li, Q. Functional assessment of heart-specific enhancers by integrating ChIP-seq data. *Pediatr. Res.* **2022**, *92*, 1332–1340. [CrossRef] [PubMed]
13. Xiong, L.; Kang, R.; Ding, R.; Kang, W.; Zhang, Y.; Liu, W.; Huang, Q.; Meng, J.; Guo, Z. Genome-wide Identification and Characterization of Enhancers Across 10 Human Tissues. *Int. J. Biol. Sci.* **2018**, *14*, 1321–1332. [CrossRef]
14. Droog, M.; Nevedomskaya, E.; Dackus, G.M.; Fles, R.; Kim, Y.; Hollema, H.; Mourits, M.J.; Nederlof, P.M.; van Boven, H.H.; Linn, S.C.; et al. Estrogen receptor α wields treatment-specific enhancers between morphologically similar endometrial tumors. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E1316–E1325. [CrossRef]
15. Farley, E.K.; Olson, K.M.; Zhang, W.; Rokhsar, D.S.; Levine, M.S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 6508–6513. [CrossRef]
16. Liao, M.; Zhao, J.P.; Tian, J.; Zheng, C.H. iEnhancer-DCLA: Using the original sequence to identify enhancers and their strength based on a deep learning framework. *BMC Bioinf.* **2022**, *23*, 480. [CrossRef] [PubMed]
17. Visel, A.; Bristow, J.; Pennacchio, L.A. Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.* **2007**, *18*, 140–152. [CrossRef] [PubMed]
18. Erwin, G.D.; Oksenberg, N.; Truty, R.M.; Kostka, D.; Murphy, K.K.; Ahituv, N.; Pollard, K.S.; Capra, J.A. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput. Biol.* **2014**, *10*, e1003677. [CrossRef]
19. Rajagopal, N.; Xie, W.; Li, Y.; Wagner, U.; Wang, W.; Stamatoyannopoulos, J.; Ernst, J.; Kellis, M.; Ren, B. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Comput. Biol.* **2013**, *9*, e1002968. [CrossRef]
20. Bissonnette, P. Extraction and Identification of Frequent Sequential Patterns in Transcription Factor Binding Site Organization of Enhancers. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 2014.
21. Robey, A.; Hassani, H.; Pappas, G.J. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv* **2020**, arXiv:2005.10247.
22. Huang, Z.; Johnson, T.S.; Han, Z.; Helm, B.; Cao, S.; Zhang, C.; Salama, P.; Rizkalla, M.; Yu, C.Y.; Cheng, J.; et al. Deep learning-based cancer survival prognosis from RNA-seq data: Approaches and evaluations. *BMC Med. Genom.* **2020**, *13*, 41. [CrossRef] [PubMed]
23. Sahoo, A.K.; Pradhan, C.; Das, H. Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In *Nature Inspired Computing for Data Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 201–212. [CrossRef]
24. Liu, D. Connecting Low-Level Image Processing and High-Level Vision via Deep Learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-ECAI-18), Stockholm, Sweden, 13–19 July 2018. [CrossRef]
25. Andrew, W.; Greatwood, C.; Burghardt, T. Visual localisation and individual identification of Holstein Friesian Cattle via deep learning. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2850–2859. [CrossRef]
26. Luongo, F.; Hakim, R.; Nguyen, J.H.; Anandkumar, A.; Hung, A.J. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery* **2021**, *169*, 1240–1244. [CrossRef]
27. Song, Z. English speech recognition based on deep learning with multiple features. *Computing* **2019**, *102*, 663–682. [CrossRef]

28. Trong, T.N.; Hautamäki, V.; Lee, K.A. Deep Language: A comprehensive deep learning approach to end-to-end language recognition. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2016), Bilbao, Spain, 21–24 June 2016. [CrossRef]
29. Cai, J.; Liu, Y. Research on English pronunciation training based on intelligent speech recognition. *Int. J. Speech Technol.* **2018**, *21*, 633–640. [CrossRef]
30. Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.C. iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudok-tuple nucleotide composition. *Bioinformatics* **2016**, *32*, 362–369. [CrossRef]
31. Cai, L.; Ren, X.; Fu, X.; Peng, L.; Zeng, X. iEnhancer-XG: Interpretable sequence-based enhancers and their strength predictor. *Bioinformatics* **2021**, *37*, 1060–1067. [CrossRef]
32. Khanal, J.; Tayara, H.; Chong, K.T. Identifying Enhancers and Their Strength by the Integration of Word Embedding and Convolution Neural Network. *IEEE Access* **2020**, *8*, 58369–58376. [CrossRef]
33. Nguyen, Q.H.; Nguyen-Vo, T.H.; Le, N.Q.K.; Do, T.T.; Rahardja, S.; Nguyen, B.P. iEnhancer-ECNN: Identifying enhancers and their strength using ensembles of convolutional neural networks. *BMC Genom.* **2019**, *20*, 951. [CrossRef] [PubMed]
34. Kim, S.G.; Harwani, M.; Grama, A.; Chaterji, S. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. *Sci. Rep.* **2016**, *6*, 38433. [CrossRef]
35. Kamran, H.; Tahir, M.; Tayara, H.; Chong, K.T. iEnhancer-Deep: A Computational Predictor for Enhancer Sites and Their Strength Using Deep Learning. *Appl. Sci.* **2022**, *12*, 2120. [CrossRef]
36. Gao, Z.; Li, Y.; Wan, S. Exploring Deep Learning for View-Based 3D Model Retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–21. [CrossRef]
37. Zhang, H.; Liu, H.; Song, R.; Sun, F. Nonlinear dictionary learning based deep neural networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July July 2016; pp. 3771–3776. [CrossRef]
38. Geng, Q.; Yang, R.; Zhang, L. A deep learning framework for enhancer prediction using word embedding and sequence generation. *Biophys. Chem.* **2022**, *286*, 106822. [CrossRef]
39. Niu, X.; Yang, K.; Zhang, G.; Yang, Z.; Hu, X. A Pretraining-Retraining Strategy of Deep Learning Improves Cell-Specific Enhancer Predictions. *Front. Genet.* **2020**, *10*, 1305. [CrossRef] [PubMed]
40. Min, X.; Chen, N.; Chen, T.; Jiang, R. DeepEnhancer: Predicting enhancers by convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 637–644. [CrossRef]
41. Rubin, C.J.; Megens, H.J.; Barrio, Á.M.; Maqbool, K.; Sayyab, S.; Schwochow, D.; Wang, C.; Carlborg, Ö.; Jern, P.; Jørgensen, C.B.; et al. Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 19529–19536. [CrossRef] [PubMed]
42. Yan, S.; Tu, Z.; Liu, Z.; Fan, N.; Yang, H.; Yang, S.; Yang, W.; Zhao, Y.; Ouyang, Z.; Lai, C.; et al. A Huntingtin Knockin Pig Model Recapitulates Features of Selective Neurodegeneration in Huntington's Disease. *Cell* **2018**, *173*, 989–1002.e13. [CrossRef]
43. Längin, M.; Mayr, T.; Reichart, B.; Michel, S.; Buchholz, S.; Guethoff, S.; Dashkevich, A.; Baehr, A.; Egerer, S.; Bauer, A.; et al. Consistent success in life-supporting porcine cardiac xenotransplantation. *Nature* **2018**, *564*, 430–433. [CrossRef]
44. Ekser, B.; Li, P.; Cooper, D.K.C. Xenotransplantation: Past, present, and future. *Curr. Opin. Organ Tran.* **2017**, *22*, 513–521. [CrossRef]
45. Kern, C.; Wang, Y.; Xu, X.; Pan, Z.; Halstead, M.; Chanthavixay, G.; Saelao, P.; Waters, S.; Xiang, R.; Chamberlain, A.; et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat. Commun.* **2021**, *12*, 1821. [CrossRef]
46. Zhou, Z.; Jiang, T.; Zhu, Y.; Ling, Z.; Yang, B.; Huang, L. A comparative investigation on H3K27ac enhancer activities in the brain and liver tissues between wild boars and domesticated pigs. *Evol. Appl.* **2022**, *15*, 1281–1290. [CrossRef]
47. Zhao, Y.; Hou, Y.; Xu, Y.; Luan, Y.; Zhou, H.; Qi, X.; Hu, M.; Wang, D.; Wang, Z.; Fu, Y.; et al. A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. *Nat. Commun.* **2021**, *12*, 2217. [CrossRef]
48. Pan, Z.; Yao, Y.; Yin, H.; Cai, Z.; Wang, Y.; Bai, L.; Kern, C.; Halstead, M.; Chanthavixay, G.; Trakooljul, N.; et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat. Commun.* **2021**, *12*, 5848. [CrossRef]
49. Heintzman, N.D.; Stuart, R.K.; Hon, G.; Fu, Y.; Ching, C.W.; Hawkins, R.D.; Barrera, L.O.; Van Calcar, S.; Qu, C.; Ching, K.A.; et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **2007**, *39*, 311–318. [CrossRef]
50. Creyghton, M.P.; Cheng, A.W.; Welstead, G.G.; Kooistra, T.; Carey, B.W.; Steine, E.J.; Hanna, J.; Lodato, M.A.; Frampton, G.M.; Sharp, P.A.; et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 21931–21936. [CrossRef]
51. Rada-Iglesias, A.; Bajpai, R.; Swigut, T.; Brugmann, S.A.; Flynn, R.A.; Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **2010**, *470*, 279–283. [CrossRef] [PubMed]
52. Visel, A.; Minovitsky, S.; Dubchak, I.; Pennacchio, L.A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **2007**, *35*, D88–D92. [CrossRef] [PubMed]

53. Oubounyt, M.; Louadi, Z.; Tayara, H.; Chong, K.T. DeePromoter: Robust Promoter Predictor Using Deep Learning. *Front. Genet.* **2019**, *10*, 286. [CrossRef]
54. Sethi, A.; Gu, M.; Gumusgoz, E.; Chan, L.; Yan, K.K.; Rozowsky, J.; Barozzi, I.; Afzal, V.; Akiyama, J.A.; Plajzer-Frick, I.; et al. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat. Methods* **2020**, *17*, 807–814. [CrossRef]
55. Yang, B.; Liu, F.; Ren, C.; Ouyang, Z.; Xie, Z.; Bo, X.; Shu, W. BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* **2017**, *33*, 1930–1936. [CrossRef]
56. Zhou, Z.; Zhu, Y.; Zhang, Z.; Jiang, T.; Ling, Z.; Yang, B.; Li, W. Comparative Analysis of Promoters and Enhancers in the Pituitary Glands of the Bama Xiang and Large White Pigs. *Front. Genet.* **2021**, *12*, 697994. [CrossRef] [PubMed]
57. Peng, Y.; Kang, H.; Luo, J.; Zhang, Y. A Comparative Analysis of Super-Enhancers and Broad H3K4me3 Domains in Pig, Human, and Mouse Tissues. *Front. Genet.* **2021**, *12*, 701049. [CrossRef]
58. Deng, D.; Tan, X.; Han, K.; Ren, R.; Cao, J.; Yu, M. Transcriptomic and ChIP-seq Integrative Analysis Reveals Important Roles of Epigenetically Regulated lncRNAs in Placental Development in Meishan Pigs. *Genes* **2020**, *11*, 397. [CrossRef]
59. MacPhillamy, C.; Alinejad-Rokny, H.; Pitchford, W.S.; Low, W.Y. Cross-species enhancer prediction using machine learning. *Genomics* **2022**, *114*, 110454. [CrossRef] [PubMed]
60. Prowse-Wilkins, C.P.; Wang, J.; Xiang, R.; Garner, J.B.; Goddard, M.E.; Chamberlain, A.J. Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. *Front. Genet.* **2021**, *12*, 664379. [CrossRef] [PubMed]
61. Fang, L.; Liu, S.; Liu, M.; Kang, X.; Lin, S.; Li, B.; Connor, E.E.; Baldwin, R.L.; Tenesa, A.; Ma, L.; et al. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* **2019**, *17*, 68. [CrossRef] [PubMed]
62. Villar, D.; Berthelot, C.; Aldridge, S.; Rayner, T.; Lukk, M.; Pignatelli, M.; Park, T.; Deaville, R.; Erichsen, J.; Jasinska, A.; et al. Enhancer Evolution across 20 Mammalian Species. *Cell* **2015**, *160*, 554–566. [CrossRef]
63. Andersson, L.; Archibald, A.L.; Bottema, C.D.; Brauning, R.; Burgess, S.C.; Burt, D.W.; Casas, E.; Cheng, H.H.; Clarke, L.; Couldrey, C.; et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **2015**, *16*, 57. [CrossRef]
64. Zhou, H.; Pan, Z.; Yao, Y.; Ying, H.; Cai, Z.; Wang, Y.; Bai, L.; Kern, C.; Halstead, M.; Chanthavixay, K.; et al. Pig genome functional annotation enhances biological interpretations of complex traits and comparative epigenomics. *Nat. Commun.* **2021**. [CrossRef]
65. Gao, T.; Qian, J. EnhancerAtlas 2.0: An updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **2019**, *48*, D58. [CrossRef]
66. Warr, A.; Affara, N.; Aken, B.; Beiki, H.; Bickhart, D.M.; Billis, K.; Chow, W.; Eory, L.; Finlayson, H.A.; Flicek, P.; et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience* **2019**, *9*, giaa051. [CrossRef]
67. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
68. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
69. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef] [PubMed]
70. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
71. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef]
72. Hamid, M.N.; Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* **2019**, *35*, 2009–2016. [CrossRef]
73. Zou, Q.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* **2019**, *25*, 205–218. [CrossRef] [PubMed]
74. Bembom, O. *seqLogo: An R Package for Plotting DNA Sequence Logos*; R Package: Vienna, Austria, 2007.
75. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2009**, *21*, 137–146. [CrossRef]
76. Jia, C.; He, W. EnhancerPred: A predictor for discovering enhancers based on the combination and selection of multiple features. *Sci. Rep.* **2016**, *6*, 38741. [CrossRef]
77. Liu, B.; Li, K.; Huang, D.; Chou, K. iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* **2018**, *34*, 3835–3842. [CrossRef]
78. Niu, K.; Luo, X.; Zhang, S.; Teng, Z.; Zhang, T.; Zhao, Y. iEnhancer-EBLSTM: Identifying Enhancers and Strengths by Ensembles of Bidirectional Long Short-Term Memory. *Front. Genet.* **2021**, *12*, 665498. [CrossRef]
79. Hasan, M.K.; Alam, M.A.; Dahal, L.; Roy, S.; Wahid, S.R.; Elahi, M.T.E.; Martí, R.; Khanal, B. Challenges of deep learning methods for COVID-19 detection using public datasets. *Informat. Med. Unlocked* **2022**, *30*, 100945. [CrossRef] [PubMed]
80. Crowther, P.S.; Cox, R.J. Accuracy of neural network classifiers as a property of the size of the data set. In Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Bournemouth, UK, 9–11 October 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1143–1149. [CrossRef]

81. Fang, J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.* **2019**, *21*, 1285–1292. [CrossRef]
82. Bu, H.; Gan, Y.; Wang, Y.; Zhou, S.; Guan, J. A new method for enhancer prediction based on deep belief network. *BMC Bioinf.* **2017**, *18*, 99–105. [CrossRef] [PubMed]
83. Nguyen, N.G.; Phan, D.; Lumbanraja, F.R.; Faisal, M.R.; Abapihi, B.; Purnama, B.; Delimayanti, M.K.; Mahmudah, K.R.; Kubo, M.; Satou, K. Applying Deep Learning Models to Mouse Behavior Recognition. *J. Biomed. Sci. Eng.* **2019**, *12*, 183–196. [CrossRef]
84. Kalinin, A.A.; Higgins, G.A.; Reamaroon, N.; Soroushmehr, S.; Allyn-Feuer, A.; Dinov, I.D.; Najarian, K.; Athey, B.D. Deep learning in pharmacogenomics: From gene regulation to patient stratification. *Pharmacogenomics* **2018**, *19*, 629–650. [CrossRef] [PubMed]
85. Taskiran, I.I.; Spanier, K.I.; Christiaens, V.; Mauduit, D.; Aerts, S. Cell type directed design of synthetic enhancers. *bioRxiv* **2022**. [CrossRef]
86. Sandhu, K.S.; Patil, S.S.; Pumphrey, M.O.; Carter, A.H. Multi-Trait Machine and Deep Learning Models for Genomic Selection using Spectral Information in a Wheat Breeding Program. *bioRxiv* **2021**. [CrossRef]
87. Telenti, A.; Lippert, C.; Chang, P.C.; DePristo, M. Deep learning of genomic variation and regulatory network data. *Hum. Mol. Genet.* **2018**, *27*, R63–R71. [CrossRef]
88. Wu, Y.; Zhang, Y.; Liu, H.; Gao, Y.; Liu, Y.; Chen, L.; Liu, L.; Irwin, D.M.; Hou, C.; Zhou, Z.A. Genome-wide identification of functional enhancers and their potential roles in pig breeding. *J. Anim. Sci. Biotechnol.* **2022**, *13*, 75. [CrossRef]
89. Zeng, R.; Liao, M. Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications. *Front. Bioeng. Biotechnol.* **2020**, *8*, 274. [CrossRef]