RESEARCH ARTICLE

# Transformer Fusing Chromosome Conformation and Genomic Information for Soybean Trait Prediction

**Ailing Chen**[1], **Qingke Zou**[1], **Xidi Yang**[2] **and Jie Zhou**[1,*]

[1] College of Mathematics, Sichuan University, Chengdu 610064, China
[2] State Key Laboratory of Swine and Poultry Breeding Industry, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China

## Abstract

**Genomic information is increasingly leveraged for the precise prediction of crop traits, with the adoption of advanced genomic prediction techniques resulting in substantial improvements in both crop yield and quality. However, traditional genomic prediction methods exhibit notable limitations in capturing long-range dependencies and fully utilizing prior information from chromosome structure. In this work, two novel Transformer models fusing chromosome conformation and genomic information have been proposed. One is the chromosomal self-attention fusion model, which captures cross-chromosomal interactions more precisely by introducing chromosomal conformation information into the self-attention mechanism of the Transformer. The other is the chromatin interaction squeeze excitation model, which extracts global features of the chromosome from all single nucleotide polymorphism sites on each chromosome. It then employs the chromatin interaction matrix to perform a weighted fusion of these global features, enabling the effective utilization of inter-chromosomal information. In addition, two novel metrics are introduced to comprehensively assess the effectiveness of the internal self-attention mechanism. They quantify the concentration of attention while measuring the alignment between the attention distribution and the chromosomal interaction priors. Experiments show that the two proposed models exhibit significant advantages in predicting soybean oil content and protein.**

**Keywords**: transformer, information fusion, chromosome interaction, genomic prediction, soybean traits.

## 1 Introduction

Soybeans, ranking among the world's three major staple crops, serve not only as a crucial food source but also as an important economic crop, with its cultivation spanning numerous countries across the globe [1]. According to data from the Food and Agriculture Organization of the United Nations, the global total soybean output reached 422 million tons in 2024, with approximately 80% of it being utilized to produce plant-based proteins and vegetable oils to meet the

**Academic Editor:**
Jian Lan

**Submitted:** 20 June 2025
**Accepted:** 27 December 2025
**Published:** 08 February 2026

**Vol.** 3, **No.** 1, 2026.
10.62762/CJIF.2025.226807

***Corresponding author:**
✉ Jie Zhou
jzhou@scu.edu.cn

**Citation**
Chen, A., Zou, Q., Yang, X., & Zhou, J. (2026). Transformer Fusing Chromosome Conformation and Genomic Information for Soybean Trait Prediction. *Chinese Journal of Information Fusion*, 3(1), 31–45.

ever-increasing food demands. Soybeans are rich in high-quality proteins and fats, and are extensively used in food processing and feed production. They are more easily digestible by animals compared to crops like corn, sorghum, and oats. Additionally, the unsaturated fatty acids found in soybeans, particularly linoleic acid which is beneficial to human health, have also garnered significant attention. However, soybean production is constrained by various environmental factors such as biotic stresses, heatwaves, droughts, and floods, which result in inconsistent product quality. Consequently, enhancing both the yield and quality of soybeans has emerged as a crucial objective in the development of modern agriculture. Currently, leveraging genetic engineering breeding techniques to enhance the oil and protein content of soybeans, thereby boosting both their yield and nutritional value, has emerged as a prominent area of focus in academic research and agricultural production [2].

Trait prediction based on genotype data stands as a pivotal approach in soybean breeding, with numerous studies already having demonstrated its immense potential [3]. As high-throughput genome sequencing technologies become more widespread, accessing genotype information for soybeans has become increasingly convenient. Consequently, the research focus in this field has shifted towards efficiently and accurately transforming these data into phenotypic predictions [4].

Traditional statistical-based genomic prediction methods can be broadly classified into two main categories: best linear unbiased prediction (BLUP) and Bayesian models. The former encompasses approaches such as genomic BLUP (GBLUP) [5] and single-step BLUP (SSBLUP) [6]. GBLUP stands as one of the most classical algorithms, which employs a mixed linear model to solve a linear system incorporating genomic information for estimating breeding values. Despite its effectiveness, the algorithm's assumption of a linear relationship between genes and traits inherently limits its capability in handling complex traits. The latter includes methods such as BayesA [7], BayesB [7], BayesC [8], BayesC$\pi$ [8], and BayesLASSO [9], among others. Although these approaches are theoretically capable of better capturing nonlinear relationships, they generally entail higher computational complexity and impose stricter data requirements.

In recent years, the rapid advancement of machine learning has facilitated the application of numerous nonlinear models in genomic prediction, yielding promising predictive outcomes. Examples of such models include support vector machine (SVM) [10], random forest (RF) [12], and gradient boosting machine (GBM) [11], etc. Recently, SVM has been effectively employed to construct predictive models for the developmental stages and yield of rice in China [13]. RF has been utilized to accurately predict the flowering time of six traits in Arabidopsis thaliana [14]. Additionally, RF has been applied to forecast the backfat thickness of a Brazilian beef cattle breed, resulting in a model that is computationally efficient during the training phase and exhibits high accuracy [15]. Furthermore, GBM has been used to make precise yield predictions for the ZhengDan958 maize hybrid variety [16].

Despite their effectiveness, current machine learning methods in genomic prediction still face major challenges related to inadequate feature design and poor generalization capabilities [17]. Many of these methods rely on manually constructed features, making it difficult to fully leverage the key information present in large-scale genotypic data. Moreover, when the number of training samples is significantly lower than that of the prediction samples, their performance improvements are not substantial [18, 19].

To address these challenges, an increasing number of studies have begun to explore the application of neural networks, such as deep neural networks (DNN) [20] and convolutional neural networks (CNN) [21], in genomic prediction. Compared with traditional machine learning models, neural networks not only possess a stronger capacity for nonlinear fitting but also can more directly reflect the intricate relationships between genotypes and phenotypes. Besides, they simplify the learning rules through backpropagation [22]. In recent studies, DNN has been utilized to construct the DNNGP framework, thereby expediting the progress in improving plant agronomic traits [23]. A DualCNN architecture, based on a two-stream convolutional neural network framework, has been designed to predict the quantitative trait effects of single nucleotide polymorphisms (SNPs) and elucidate the contributions of genotypes to phenotypic variations [24]. Additionally, a Bi-GRU model, built upon the gated recurrent unit (GRU), has been proposed to capture long-term dependencies within sequences and better depict the intrinsic spatiotemporal dynamic characteristics of the sequences [25].

Furthermore, a model integrating CNN with extended

one-hot encoding has been proposed to predict economic traits such as growth rate and lean meat percentage in pig populations, thereby enhancing breeding efficiency [26]. Meanwhile, long short-term memory network (LSTM) [27] has been employed to accurately identify DNA-protein binding sites from DNA sequences [28]. Subsequently, another classic model that utilizes the Transformer encoder, DNABERTT [32], is designed to capture the contextual information in the human genome sequence and learn the patterns and structures of genetic sequences [29]. Similarly, the Enformer model, also based on the Transformer architecture, has been applied to the study of human DNA sequences, significantly improving the accuracy of predictions derived from genomic sequences [30]. Recent research has further expanded the application of Transformer in genomic prediction: the EBMGP model is proposed with elastic net feature selection and Transformer embeddings to enhance predictive performance [35]; an improved Transformer scheme incorporating batch normalization and cosine annealing algorithms is developed specifically for soybean breeding data [36]. Additionally, a GPformer model combining genome-wide association studies with deep learning has been devised for genomic prediction [31].

Notably, multimodal learning and interpretability research are emerging as new trends in genomic prediction. First, a systematic review of multimodal deep learning integration in plant breeding is conducted [37], which lays the groundwork for the field. Building on this foundation, a framework for analyzing attention mechanisms in genomic Transformer models is proposed to provide new perspectives for biological interpretation [38]. Furthermore, a multimodal Transformer model enabling cell-type-agnostic regulatory prediction is developed [39]. These advances collectively signify the field's evolution from unimodal prediction toward multimodal interpretable analysis.

With the continuous advancement of genomic technologies, trait prediction based on genomic information has emerged as a crucial tool for precision breeding. However, current methodologies still fall short in capturing long-range dependencies and integrating prior knowledge of chromosomal structures. To address these limitations, we leverage soybean SNP data along with chromosomal interaction information for trait prediction, resulting in the design of two innovative Transformer models. The primary contributions of this work are outlined as follows:

- A chromosomal self-attention fusion model (CSAFM) has been proposed to fully leverage chromosomal interaction information. By introducing inter-chromosomal correlation weights into the self-attention mechanism, it more accurately captures interactions across different chromosomes;
- A chromosomal interaction squeeze excitation model (CISEM) has been proposed to effectively capture interactions across different chromosomes. By extracting and reweighting global features at the chromosomal level, it can adaptively capture regulatory information across chromosomes, thereby enhancing the accuracy of genomic predictions;
- Two novel evaluation metrics have been introduced to comprehensively assess the effectiveness of the self-attention mechanism within the model.

The rest of this article is organized as follows. Section 2 introduces two genomic prediction models designed with Transformer. Section 3 presents two evaluation metrics to comprehensively assess the effectiveness of the self-attention mechanism. Section 4 provides two examples to evaluate the performance of the proposed methods on real-world data. Section 5 gives the conclusion.

## 2 Methodology

In this section, we introduce two predictive models, CSAFM and CISEM, which incorporate chromosomal interaction information at distinct stages to enhance their ability to capture long-range dependencies and interactions across different chromosomes.

### 2.1 CSAFM

To fully leverage the interaction information between chromosomes, we propose the CSAFM based on the Transformer architecture, with its overall framework illustrated in Figure 1. CSAFM takes into account both the local information of chromosomes and the global interaction information across chromosomes when calculating the attention scores. Specifically, we introduce a matrix $\boldsymbol{B} \in \mathbb{R}^{n\times d_k}$ that contains chromosomal interaction information, where $n$ is the number of SNP locus and $d_k$ is the feature dimension of each SNP locus, to represent global correlations into the self-attention mechanism of the Transformer. This matrix is loaded as a trainable bias term to the key matrix $\boldsymbol{K} \in \mathbb{R}^{n\times d_k}$. Thus, the attention mechanism is

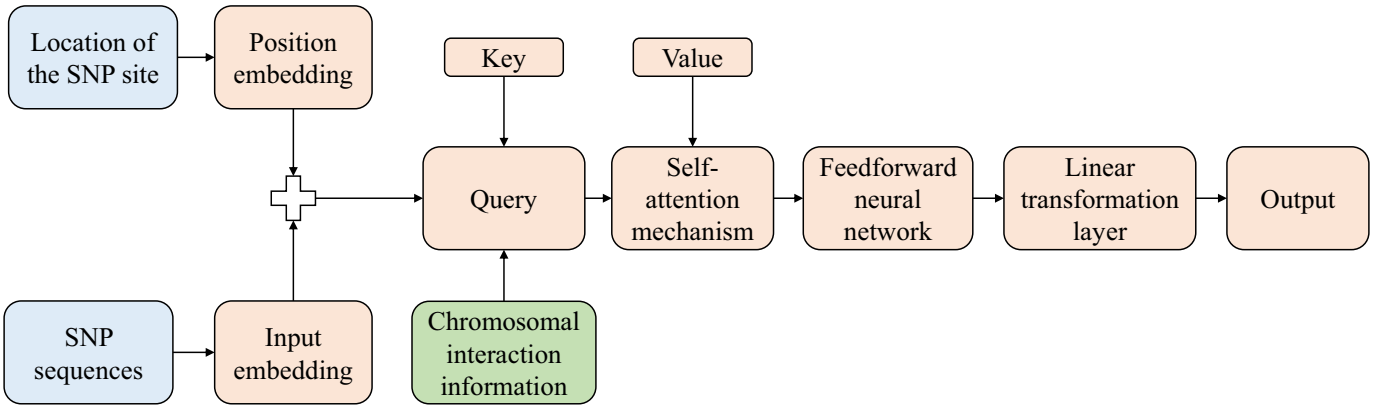



**Figure 1.** CSAFM architecture.

redesigned as follows:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V};\boldsymbol{B}) = \text{softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{K}\oplus\boldsymbol{B})^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \quad (1)$$

where $\boldsymbol{Q} = [\boldsymbol{Q}_1^{\mathrm{T}},\ldots,\boldsymbol{Q}_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n\times d_k}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times d_k}$ are the query matrix and the value matrix in the self-attention mechanism, respectively. And $\text{softmax}(\cdot)$ is an activation function that normalizes the weights. Additionally, $\oplus$ represents the operation of loading chromosomal interaction information into the corresponding SNP site keys. Specifically, for the query $\boldsymbol{Q}_i$ at the $i$-th position, we have $\boldsymbol{Q}_i(\boldsymbol{K}\oplus\boldsymbol{B})^{\mathrm{T}} = \boldsymbol{Q}_i(\boldsymbol{K}+\boldsymbol{B}_i)^{\mathrm{T}}$ with $\boldsymbol{B}_i = [\boldsymbol{b}_{i1}^{\mathrm{T}},\ldots,\boldsymbol{b}_{in}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n\times d_k}$, here, the bias term $\boldsymbol{b}_{ij}$ $(j = 1,\ldots,n)$ is derived from the chromosomal interaction matrix $\boldsymbol{H} \in \mathbb{R}^{G\times G}$, and the specific expression given by

$$\boldsymbol{b}_{ij} = \left[\boldsymbol{H}_{c(i)c(j)},\boldsymbol{H}_{c(i)c(j)},\ldots,\boldsymbol{H}_{c(i)c(j)}\right] \in \mathbb{R}^{1\times d_k}, \quad (2)$$

where $c(i)$, $c(j)$ denote the chromosome numbers of tokens $i$ and $j$, respectively.

In fact, the key $\boldsymbol{K}$ in the self-attention mechanism typically represents the static content of input tokens, whereas global prior information, such as chromosomal interactions, is more suitable for integration into the key as memory-like information. Loading this global prior onto the query $\boldsymbol{Q}$ could potentially disrupt the representation of the queries themselves, thereby degrading the matching effectiveness. By applying the bias to the key $\boldsymbol{K}$ instead, we can fully leverage the global prior while preserving the stability of the query matrix $\boldsymbol{Q}$, ultimately enabling better regulation of the attention scores.

In the task of soybean trait prediction, each token can be regarded as representing a single SNP locus. The following lemma and theorem demonstrate how CSAFM adaptively enhances the attention weights of important tokens.

**Lemma 1.** *Let $\boldsymbol{a} = [\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_n] \in \mathbb{R}^n$ be a vector, and $\boldsymbol{w}_i = \exp(\boldsymbol{a}_i)/\sum_{j=1}^n \exp(\boldsymbol{a}_j)$ $(i = 1,\ldots,n)$ be its softmax weights. Define $\boldsymbol{b} = [\boldsymbol{b}_1,\boldsymbol{b}_2,\ldots,\boldsymbol{b}_n] \in \mathbb{R}^n$ such that $\boldsymbol{b}_i \geq \boldsymbol{b}_j$ if and only if $\boldsymbol{a}_i \geq \boldsymbol{a}_j$ (i.e., $\boldsymbol{b}$ preserves the order of $\boldsymbol{a}$). Let $\tilde{\boldsymbol{a}} = \boldsymbol{a} + \boldsymbol{b}$, with softmax weights $\tilde{\boldsymbol{w}}_i = \exp(\tilde{\boldsymbol{a}}_i)/\sum_{j=1}^n \exp(\tilde{\boldsymbol{a}}_j)$. Then, the maximum component of $\tilde{\boldsymbol{w}}_i$ strictly exceeds the maximum component of $\boldsymbol{w}_i$, i.e.,*

$$\max_i \tilde{\boldsymbol{w}}_i > \max_i \boldsymbol{w}_i. \quad (3)$$

*Proof.* See Appendix A. □

**Theorem 1.** *Introducing a bias term into the keys within the Transformer architecture enables the amplification of the ratio of attention weights among distinct SNP sites, provided that the bias term fulfills specific conditions. Furthermore, this incorporation concurrently reinforces the attention weight allocated to the most pivotal tokens.*

*Proof.* See Appendix B. □

From Theorem 1, we observe that CSAFM enhances each token's latent representation by directly superimposing a bias on the key $\boldsymbol{K}$, thereby incorporating global contextual information from the chromosome interaction matrix (CIM) $\boldsymbol{H}$. Therefore, this model has the following three advantages:

- When a token's chromosome exhibits strong interactions with other chromosomes, its corresponding bias value increases, directly elevating the attention score between this token and the query $\boldsymbol{Q}$. This means the model pays more attention to the features of SNPs located on chromosomes with stronger interactions,

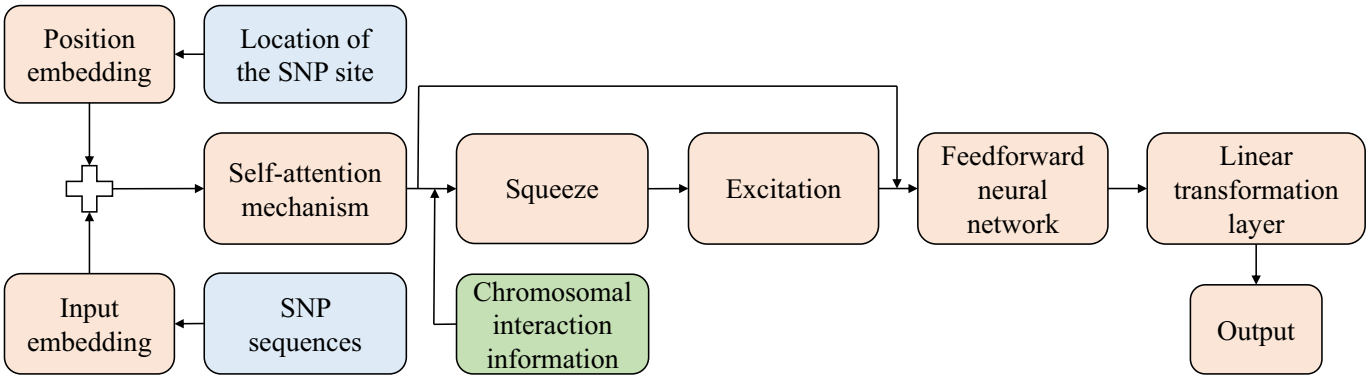



**Figure 2.** CISEM architecture.

facilitating the learning of cross-chromosomal regulatory information;

- When two tokens are similar in content but their respective chromosomes do not interact strongly, their attention scores decrease relatively. This design effectively mitigates the issue of weight misallocation that can arise from relying solely on local semantic similarity, preventing the model from overemphasizing irrelevant tokens due to superficial feature similarities. Consequently, it reduces erroneous attention assignments and enhances the robustness;
- By integrating the CIM into the key $\boldsymbol{K}$, the attention distribution after softmax normalization becomes more concentrated. Intuitively, this is akin to the model placing greater emphasis on significant inter-chromosomal interactions rather than giving equal attention to all SNPs. This approach minimizes noise interference and improves the quality of modeling gene regulatory information.

### 2.2 CISEM

In this subsection, we will introduce another approach, CISEM, whose architecture is illustrated in Figure 2. CISEM incorporates a squeeze excitation (SE) module combined with a CIM after the self-attention module in the Transformer. This approach aims to enhance the ability to adaptively capture cross-chromosome regulatory information through global feature extraction and re-weighting at the chromosome level, thereby improving prediction accuracy.

For the input SNP feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d_k}$, after passing through the multi-head self-attention module, a residual connection is applied to obtain:

$$\widetilde{\boldsymbol{X}} = \text{Self-Attention}(\boldsymbol{X}) + \boldsymbol{X}, \tag{4}$$

Then, after layer normalization, the result is obtained as follows:

$$\boldsymbol{Y} = \text{LayerNorm}(\widetilde{\boldsymbol{X}}), \tag{5}$$

where $\text{LayerNorm}(\cdot)$ refers to the layer normalization operation.

In Equation (5), $\boldsymbol{Y}$ has captured the global dependency information within the input sequence, but the prior information regarding the inter-chromosomal structure has not yet been introduced.

Assume that the soybean genome consists of $G$ pairs of chromosomes, with each pair containing $L$ SNP sites. Let $\boldsymbol{M} \in \mathbb{R}^{G \times G}$ be the trainable CIM, where its initial values are derived from $\boldsymbol{H}$. During training, the matrix $\boldsymbol{M}$ is optimized, with the constraint that each element in $\boldsymbol{M}$ cannot deviate from its initial value by more than $\gamma$, i.e., for any $p, q \in [1, G]$, we have

$$\boldsymbol{M}_{pq} = \min\left(\boldsymbol{H}_{pq} + \gamma, \max\left(\boldsymbol{H}_{pq} - \gamma, \boldsymbol{M}_{pq}\right)\right), \tag{6}$$

where $\boldsymbol{M}_{pq}$ represents the correlation between the $p$-th and $q$-th pairs of chromosomes.

By using $\boldsymbol{M}$ to fuse the global features at the soybean chromosome level, we consider the $g$-th pair of chromosomes on the chromosome:

$$\tilde{\boldsymbol{z}}_g = \left[\sum_{q=1}^{G} \boldsymbol{Y}_{(i,:)} \left(\sum_{j \in \mathcal{L}_q} \boldsymbol{Y}_{(j,:)}\right)^{\mathrm{T}} \cdot \boldsymbol{M}\,(g,q)\right]_{i \in \mathcal{L}_g}, \tag{7}$$

where $\mathcal{L}_g$ denotes the index set of SNPs on the $g$-th pair of chromosomes. $\boldsymbol{Y}_{(i,:)}$ refers to the features corresponding to the $i$-th SNP locus. Therefore, $\tilde{z}_g$ integrates information from other chromosomes. At the same time, $\boldsymbol{M}$ can continuously optimize during training, enabling it to better learn the conformational information between chromosomes.

Then, we employ an SE module with two fully connected layers (with reduction ratio $r$) to generate channel-wise recalibration vectors for each chromosome pair. The weight vector computation follows:

$$\boldsymbol{\omega}^{(g)} = \sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1\tilde{\boldsymbol{z}}_g\right)\right) \in \mathbb{R}^{1\times L}, \tag{8}$$

where, $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{L}{r}\times L}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{L\times\frac{L}{r}}$ are learnable parameters, $\delta(\cdot)$ denotes ReLU activation, and $\sigma(\cdot)$ represents sigmoid function for range normalization $[0,1]$

The chromosome-specific weights are aggregated into a diagonal calibration matrix:

$$\boldsymbol{W} = \operatorname{diag}\left(\omega_1,\omega_2,\ldots,\omega_L\right), \tag{9}$$

$$\boldsymbol{E} = \boldsymbol{W}\boldsymbol{Y}, \tag{10}$$

where $\omega_i$ corresponds to the $i$-th element of the concatenated weight vector $\left[\boldsymbol{\omega}^{(1)},\boldsymbol{\omega}^{(2)},\ldots,\boldsymbol{\omega}^{(G)}\right]$. The recalibrated matrix $\boldsymbol{E}$ serves as enhanced input for subsequent network layers, enabling chromosome-level global context modulation of SNP features.

Given the original feature covariance $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$, with the introduction of chromosomal conformation information, the covariance matrix is updated to:

$$\boldsymbol{\Sigma}_{\boldsymbol{E}} = \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} = \boldsymbol{W}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}. \tag{11}$$

Under the Frobenius norm constraint $\|\boldsymbol{W}\|_{\mathrm{F}} \le 1$, we derive:

$$\|\boldsymbol{\Sigma}_{\boldsymbol{E}}\|_{\mathrm{F}} \le \|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_{\mathrm{F}}. \tag{12}$$

This norm reduction indicates effective suppression of spurious feature correlations. The model thereby retains critical SNP interactions while mitigating:

- False positive correlations from feature redundancy;
- Overemphasis on weakly associated SNPs;
- Potential overfitting to noisy covariations.

This recalibration mechanism enhances discriminative capability by focusing on biologically meaningful SNP interactions.

## 3 Attention-based model evaluation metrics

In genetic regulation, different loci exhibit varying degrees of importance with only key regulatory factors playing dominant roles in specific contexts. To evaluate the ability to capture critical loci, we propose a normalized attention entropy (NAE) metric based on attention weights:

$$\mathrm{NAE} = -\frac{1}{n\ln n}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}_{ij}\ln\boldsymbol{\alpha}_{ij}, \tag{13}$$

where $\boldsymbol{\alpha}_{ij}$ represents the attention weights from the final attention layer.

In fact, the design of NAE is based on entropy. Specifically, the attention entropy for the $i$-th locus is given by:

$$S(i) = -\sum_{j=1}^{n}\boldsymbol{\alpha}_{ij}\ln\boldsymbol{\alpha}_{ij}, \tag{14}$$

which quantifies the concentration of attention distribution at locus $i$. The theoretical maximum entropy $\ln n$ occurs when attention is uniformly distributed. To eliminate sequence length bias, we normalize the entropy values and average across all SNP loci to obtain NAE.

A lower NAE indicates highly concentrated attention distributions, analogous to biological systems where few regulatory factors dominate expression control. This reflects the enhanced selectivity of the model in identifying critical information. Conversely, higher NAE suggests diffuse attention patterns, implying the model may fail to focus on important loci, similar to dispersed regulatory mechanisms.

For the second evaluation metric, we first construct a CIM $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{G\times G}$ to incorporate prior interaction information between SNPs:

$$\widetilde{\boldsymbol{H}}_{ij} = \boldsymbol{H}_{g(i),g(j)}, \tag{15}$$

where $g(i)$ maps SNP $i$ to its chromosome index. After row-wise normalization to form valid probability distributions, we measure the alignment between attention weights $\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_{i1},\boldsymbol{\alpha}_{i2},\ldots,\boldsymbol{\alpha}_{in}]$ and the normalized interaction vector $\widetilde{\boldsymbol{H}}_{(i,:)}$ using cosine similarity:

$$\cos(\boldsymbol{\alpha}_i,\widetilde{\boldsymbol{H}}_{(i,:)}) = \frac{\boldsymbol{\alpha}_i\cdot\widetilde{\boldsymbol{H}}_{(i,:)}}{\|\boldsymbol{\alpha}_i\|_2\|\widetilde{\boldsymbol{H}}_{(i,:)}\|_2}, \tag{16}$$

where $\widetilde{\boldsymbol{H}}_{(i,:)}$ represents the $i$-th row vector of the matrix $\widetilde{\boldsymbol{H}}$, and $\|\cdot\|_2$ is the Euclidean norm of the vector.

Averaging across all SNPs yields our second metric, the attention alignment score (AAS):

$$\mathrm{AAS} = \frac{1}{n}\sum_{i=1}^{n}\cos(\boldsymbol{\alpha}_i,\widetilde{\boldsymbol{H}}_{(i,:)}). \tag{17}$$

Higher AAS values indicate better alignment between the learned attention patterns and the chromosomal 3D structure, and suggest that the model's biological plausibility is enhanced by better adherence to chromosomal interaction patterns observed in vivo.

## 4 Experiments

In this section, we will predict soybean traits using real data to evaluate the effectiveness of the proposed models. All experiments are implemented in Python 3.8 on a computer equipped with an NVIDIA RTX 3090 GPU with 24GB of memory.

### 4.1 Experiment setup

To validate the effectiveness of the proposed method, several classic and advanced approaches were compared, including CNN [21], LSTM [27], GRU [25] and Transformer [32].

#### *4.1.1 Data preprocessing*

The experiment utilized the WM82.A1 dataset from the publicly accessible SoyBase database for soybeans. This dataset can be accessed at the website https://www.soybase.org/tools/snp50k. The SoySNP50K iSelect BeadChip was employed to genotype 20087 soybean accessions. After excluding samples with a heterozygosity rate of alleles exceeding 10%, SNP data for 19648 soybean accessions were obtained. These accessions comprise 1168 wild soybeans and 18480 cultivated soybeans.

Subsequently, we implement the following three steps for uniform quality control of the original dataset: 1) use Plink V1.9 [34] to remove samples with an allele missing rate greater than 10%; 2) filter out SNP loci with a minor allele frequency lower than 5%; 3) use Beagle [33] to infer and impute missing genotype data. After these steps, the WM82.A1 dataset retains 11779 samples, each containing 42453 SNP loci and 84906 bases.

In addition, the phenotypic trait data of soybeans are sourced from the US National Plant Germplasm System (NPGS), accessible at the website https://npgsweb.ars-grin.gov/gringlobal/search. By referencing the soybean PI (Plant Introduction) numbers, one can locate corresponding phenotypic information, such as flowering time, 100-seed weight, oil content, protein content, and more. For this experiment, oil content and protein content were selected as the target traits for prediction.

#### *4.1.2 Chromosome interaction matrix*

In this subsection, we employ the Hi-C contact matrix to compute the CIM. The Hi-C data used in the experiment are sourced from the PRJCA009364 project at the National Genomics Data Center (NGDC), accessible at https://ngdc.cncb.ac.cn/gsa/search?searchTerm=%22PRJCA009364%22. This dataset comprises a total of 27 samples, and Hi-C data for each sample are generated using the Illumina paired-end sequencing technology.

We conduct three-dimensional structural analysis of chromosomes using the HiC-Pro tool, normalize the valid interaction pairs, and thereby construct a Hi-C contact matrix. This matrix records the contact frequencies between different genomic fragments.

We then assign the genomic fragments in the Hi-C contact matrix to their respective chromosomal regions. Subsequently, within each chromosomal region and between different chromosomal regions, we calculate the sum of interaction frequencies for all fragments, thus generating the CIM. Each element in this matrix represents the overall interaction intensity between two chromosomal regions. Figure 3 displays a heatmap of the CIM for 20 pairs of chromosomes, where the shade of color reflects the interaction intensity between different regions.

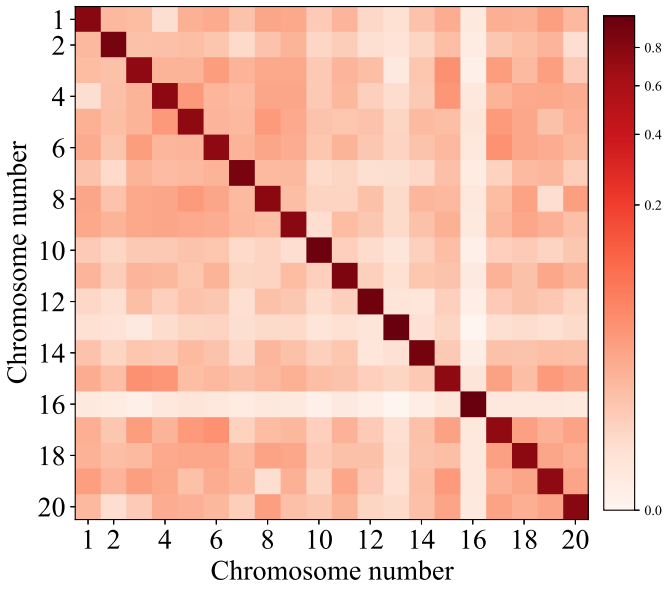



**Figure 3.** Chromosome interaction matrix.

#### *4.1.3 Evaluation metrics*

In genomic prediction tasks, selecting appropriate evaluation metrics is crucial for assessing model performance. In addition to the attention-based evaluation metric proposed in Section 3, we also

employed two commonly used evaluation metrics from regression models: mean absolute error (MAE) and Pearson correlation coefficient (PCC), along with the consistency index (CI). The definitions of MAE and PCC are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i|, \tag{18}$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\hat{\boldsymbol{y}}_i - \bar{\hat{\boldsymbol{y}}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2}\cdot\sqrt{\sum_{i=1}^{n}(\hat{\boldsymbol{y}}_i - \bar{\hat{\boldsymbol{y}}})^2}}, \tag{19}$$

where $\boldsymbol{y}_i$ represents the true value, $\hat{\boldsymbol{y}}_i$ denotes the predicted value, $n$ is the number of samples in the test set, and $\bar{\boldsymbol{y}}$ and $\bar{\hat{\boldsymbol{y}}}$ are the mean values of the true and predicted traits in the test set, respectively. A smaller MAE indicates lower prediction error and better model performance. A PCC value closer to 1 suggests a stronger positive linear correlation between the predictions and the true values.

For different models, MAE and PCC may not always show consistent performance (both being good or both being poor), making it difficult for a single metric to comprehensively reflect the performance in terms of numerical accuracy and trend prediction. To address this, the CI combines MAE and PCC, with the following expression:

$$\mathrm{CI} = \frac{\mathrm{PCC}}{\frac{\mathrm{MAE}}{\mathrm{mean}|\boldsymbol{y}|} + 1}, \tag{20}$$

where mean$|\boldsymbol{y}|$ is the mean absolute value of the true values of the sample characteristics in the test set. It is used to normalize MAE, ensuring that the CI values fall within a reasonable range. A CI closer to 1 indicates that the model not only aligns well with the true trends but also has low prediction errors.

#### *4.1.4 Parameter Settings*

In the experiments, the self-attention mechanism of the Transformer employs multi-head attention with 8 heads, and the encoder consists of 6 layers. Furthermore, we utilize Bayesian optimization to fine-tune the hyperparameters of the model, aiming to enhance predictive performance. The parameter search space for Bayesian optimization is defined as follows: the learning rate ranges from $10^{-6}$ to $10^{-4}$, the batch size varies between 16 and 64, and the dropout rate is set within the range of 0.1 to 0.3. The proposed methods, CSAFM and CISEM, adhere to the same parameter search space. Notably, the latter method incorporates an additional parameter, $\gamma = 10^{-3}$, which imposes constraints on the update magnitude of the CIM.

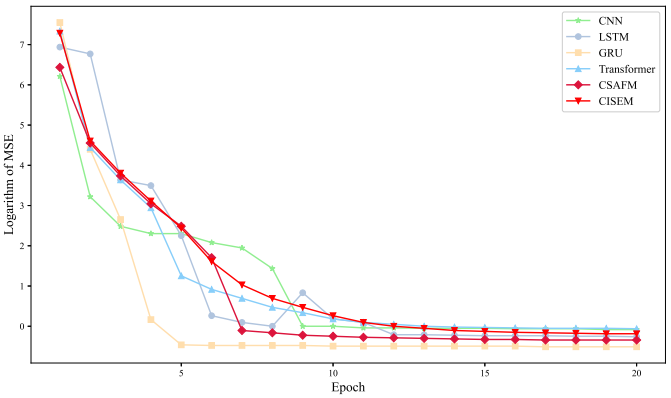



**Figure 4.** Training loss for oil content prediction.

For CNN, the input data's feature dimension is set to 1. It employs 64 convolutional kernels, each with a size of 3, for feature extraction, and the output dimension is 1. The network comprises 2 convolutional layers, with the ReLU activation function applied after each convolution operation to introduce non-linearity.

Regarding LSTM, the number of input features per time step is set to 1, and the hidden layer dimension is configured as 64, with the output dimension being 1. Dropout is applied between each layer of the LSTM, with a dropout rate of 0.2.

For GRU, the SNP data is segmented into chunks of length 1024 for training, addressing the issue of insufficient video memory caused by excessively long input sequences. The entire network architecture consists of 2 stacked GRU layers followed by a fully connected layer.

Additionally, the dataset is partitioned into training, validation, and test sets in a 7:2:1 ratio, ensuring no data overlap between these sets. In the prediction task, the number of sampling iterations is fixed at 10. The mean squared error (MSE) is selected as the loss function.

### 4.2 Experimental results and analyses

#### *4.2.1 Soybean trait prediction*

Figures 4 and 5 illustrate the loss reduction trends during the training process of different models for the tasks of predicting oil content and protein content, respectively. Observing the overall trends, it is evident that the loss for all models gradually decreases as the number of training epochs increases, and they tend to converge around the 12-th epoch. This indicates that the models have sufficiently learned the information

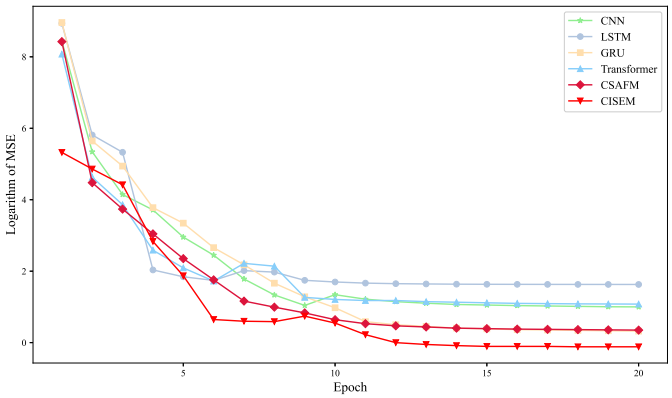



**Figure 5.** Training loss for protein content prediction.

embedded in the training data.

Tables 1 and 2 present the performance of various models in predicting oil content and protein content (hereinafter, the value closest to the ideal in each row will be highlighted in bold.), respectively. It is evident that Transformer outperforms LSTM, GRU, and CNN in terms of prediction accuracy and consistency, demonstrating that the self-attention mechanism enhances the predictive capability for soybean traits. However, Transformer does not fully leverage chromosome conformation information, resulting in limitations in modeling complex chromosomal associations.

In summary, the exceptional performance of CSAFM and CISEM models in soybean trait prediction stems from their effective integration of chromosome conformation and interaction information. This advantage can be attributed to three key factors:

1) By incorporating chromosome conformation information into the attention mechanism, both models demonstrate significant improvements in predictive performance, particularly in accurately modeling cross-chromosome SNP associations. This demonstrates that chromosome interaction priors play a critical role in complex trait prediction by effectively enhancing attention concentration on key SNP loci;
2) The attention mechanism proves effective as it allocates differentiated weights to SNP loci based on their relevance. Through chromosome conformation integration, our models can prioritize structurally proximal genomic regions even when they are distant in linear sequence. This capability enables the capture of long range dependencies that conventional methods struggle to represent, leading to improved prediction accuracy;
3) The performance difference between oil and protein content predictions results from their distinct genetic architectures. Oil content is mainly influenced by a limited set of large effect loci with strong structural interactions, making the CISEM model more suitable as it aggregates chromosome level features and employs SE based recalibration to achieve lower prediction error. Conversely, protein content shows high polygenicity and is regulated by numerous dispersed loci, where the CSAFM model demonstrates superior correlation and consistency through its fine grained key vector attention adjustment.

Based on this analysis, we propose the following practical guidance for model selection: the CISEM model is better suited for traits governed by a limited number of strong genomic signals (e.g., oil content), whereas the CSAFM model demonstrates superior performance for complex polygenic traits (e.g., protein content).

#### 4.2.2 *Comparison of computational cost*

To comprehensively evaluate model computational efficiency, we compare the parameter count, relative computational complexity measured in floating point operations (FLOPs) with the baseline Transformer as reference, and per-sample inference time under batch size setting of 1. As summarized in Table 3, which presents a comparison of these metrics across different models, both CSAFM and CISEM modules maintain competitive prediction performance while introducing only modest computational overhead, manifested in two aspects: i) In terms of parameter count and computational complexity, CSAFM and CISEM increase the parameters and FLOPs by only approximately 20%–30% compared to the baseline Transformer; ii) Regarding inference efficiency, the per-sample inference time of all models remains within the practical range of 6.0–6.5ms on modern GPU hardware.

The experimental results demonstrate that the two proposed enhancement modules achieve significant improvements in prediction performance while introducing only limited computational complexity, indicating promising application potential.

#### 4.2.3 *Effectiveness analysis of internal mechanisms*

In this subsection, we conduct ablation studies to evaluate the effectiveness of key mechanisms in the

**Table 1.** Comparison of oil content prediction performance.

| Method | CNN | GRU | LSTM | Transformer | CSAFM | CISEM |
|---|---|---|---|---|---|---|
| MAE | 4.2514 | 2.6972 | 1.8794 | 1.3525 | 0.8902 | **0.4198** |
| PCC | 0.1455 | 0.3314 | 0.3212 | 0.3239 | 0.3268 | **0.3653** |
| CI | 0.1178 | 0.2884 | 0.2910 | 0.3014 | 0.3122 | **0.3600** |
| NAE | - | - | - | 0.7457 | 0.6963 | **0.6811** |
| AAS | - | - | - | 0.1486 | **0.4593** | 0.3133 |

**Table 2.** Comparison of protein content prediction performance.

| Method | CNN | GRU | LSTM | Transformer | CSAFM | CISEM |
|---|---|---|---|---|---|---|
| MAE | 2.1391 | 4.3712 | 2.4345 | 1.9474 | 1.7078 | **1.4473** |
| PCC | 0.1801 | 0.1655 | 0.1544 | 0.1855 | **0.2704** | 0.2083 |
| CI | 0.1739 | 0.1543 | 0.1484 | 0.1797 | **0.2429** | 0.2158 |
| NAE | - | - | - | 0.8057 | **0.7818** | 0.7882 |
| AAS | - | - | - | 0.0739 | 0.2421 | **0.2870** |

**Table 3.** Comparison of computational cost.

| Method | CNN | GRU | LSTM | Transformer | CSAFM | CISEM |
|---|---|---|---|---|---|---|
| Parameters (M) | 2.0 | 5.0 | 5.0 | 15.0 | 18.0 | 19.0 |
| Relative FLOPs | 2.0 | 5.0 | 5.0 | 1.0 | 1.2 | 1.3 |
| Time (ms/sample) | 1.5 | 3.0 | 3.0 | 5.0 | 6.0 | 6.5 |

**Table 4.** Effectiveness evaluation of internal mechanisms in the CSAFM.

| Trait prediction | CIM | GAU | MAE | PCC | CI | NAE | AAS |
|---|---|---|---|---|---|---|---|
| Oil content | × | × | 1.3525 | 0.3239 | 0.3014 | 0.7457 | 0.1486 |
| | ✓ | × | 1.0501 | 0.3255 | 0.3072 | 0.7200 | 0.3100 |
| | ✓ | ✓ | **0.8902** | **0.3268** | **0.3122** | **0.6963** | **0.4593** |
| Protein content | × | × | 1.9474 | 0.1855 | 0.1797 | 0.8057 | 0.0739 |
| | ✓ | × | 1.8202 | 0.2357 | 0.2150 | 0.7953 | 0.1551 |
| | ✓ | ✓ | **1.7078** | **0.2704** | **0.2429** | **0.7818** | **0.2421** |

**Table 5.** Effectiveness evaluation of internal mechanisms in the CISEM.

| Trait prediction | CIM | SE | MAE | PCC | CI | NAE | AAS |
|---|---|---|---|---|---|---|---|
| Oil content | × | × | 1.3525 | 0.3239 | 0.3014 | 0.7457 | 0.1486 |
| | ✓ | × | 0.6511 | 0.3542 | 0.3452 | 0.7053 | 0.2657 |
| | ✓ | ✓ | **0.4198** | **0.3653** | **0.3600** | **0.6811** | **0.3133** |
| Protein content | × | × | 1.9474 | 0.1855 | 0.1797 | 0.8057 | 0.0739 |
| | ✓ | × | 1.5607 | 0.2051 | 0.2074 | 0.7923 | 0.2105 |
| | ✓ | ✓ | **1.4473** | **0.2083** | **0.2158** | **0.7882** | **0.2870** |

CSAFM and CISEM models. Specifically, for CSAFM, we investigate the contributions of both the CIM and the Gated Attention Unit (GAU). The CIM introduces a bias integration mechanism by incorporating the Hi-C interaction matrix as an additive term to the Key vectors in the self-attention module. Biologically, this reflects that SNP loci in spatial proximity within the 3D genome architecture are more likely to engage in cooperative regulation—even when linearly distant. By integrating this bias, CSAFM leverages chromosomal conformation information to guide attention toward structurally meaningful interactions.

**Table 6.** Effectiveness evaluation of the internal mechanisms in CISEM with different $\gamma$ values.

| Trait prediction | CIM | No integration ($\gamma = 0$) | Optimal ($\gamma = 0.0002$) | Overweighting ($\gamma = 0.001$) |
|---|---|---|---|---|
| Oil content | MAE | 0.7042 | **0.4198** | 0.5102 |
| | PCC | 0.3351 | **0.3653** | 0.3581 |
| Protein content | MAE | 1.7201 | **1.4473** | 1.5604 |
| | PCC | 0.1954 | **0.2083** | 0.2029 |

The GAU, on the other hand, consists of two core components: multi-head attention and a feed-forward neural network. As for CISEM, we focus on assessing the importance of both the CIM and the SE module. All ablation experiments are carried out in the context of predicting soybean oil and protein content.

As demonstrated in the evaluation results of Tables 4 and 5, each internal mechanism contributes positively to the predictive performance for soybean soybean traits. In the CSAFM model (see Table 4), the sole introduction of the CIM already leads to comprehensive improvements across all evaluation metrics for both oil and protein content prediction tasks, outperforming the baseline model that lacks any mechanisms. Furthermore, when the GAU is combined with CIM, the model achieves even better performance, indicating a synergistic enhancement effect between the two components. This trend is similarly validated in the CISEM model (see Table 5): the integration of CIM and the SE module yields the most accurate and robust prediction outcomes. These experimental results confirm that both the chromosome interaction matrix and their corresponding architecture-specific modules are essential and effective components for capturing the complex genetic determinants of soybean trait.

Moreover, in the CISEM model, the parameter $\gamma$ governs the balance between the original self-attention outputs and the chromosome interaction-based recalibration. From a biological perspective, $\gamma$ can be interpreted as a factor modulating the intensity of structural prior integration. An excessively small $\gamma$ may lead to underutilization of inter-chromosomal information, while an overly large $\gamma$ could overemphasize global characteristics at the expense of local dependencies. To address this, we conducted experiments to validate this hypothesis.

Table 6 systematically evaluates how different $\gamma$ parameter settings affect prediction performance for soybean oil and protein content in the CISEM model, comparing three representative scenarios: no chromosomal interaction integration ($\gamma = 0$), optimal integration strength ($\gamma = 0.0002$), and excessive interaction weighting ($\gamma = 0.001$). The results demonstrate that $\gamma = 0.0002$ yields the best performance for both traits, reducing MAE by 40.4% and improving PCC by 9.0% for oil content while achieving 15.9% lower MAE and 6.6% higher PCC for protein content compared to the baseline ($\gamma = 0$). Notably, when using an excessively large $\gamma$ value, prediction accuracy for both traits declines significantly, confirming our hypothesis that over-reliance on global interaction information compromises the model's ability to capture local dependencies. These findings collectively indicate that an appropriate chromosomal interaction integration strength effectively leverages structural priors while preserving local feature sensitivity, thereby substantially enhancing the accuracy and robustness of soybean trait prediction.

## 5 Conclusion

In this work, we have proposed two innovative methods for soybean genome prediction based on the Transformer framework, namely CSAFM and CISEM. These methods fuse chromosome conformation information into the model, achieving better fusion with genomic data and more accurately capturing the interactions between SNP sites, thereby enhancing the prediction performance for soybean traits. Experiments on real-world data demonstrated the clear advantages of the proposed model in capturing long-range dependencies and gene regulation. Additionally, we introduced two new metrics, NAE and AAS, to quantify the attention dispersion and its alignment with biological priors, providing intuitive and quantitative tools for model evaluation. Future work will focus on integrating multi-omics data, such as transcriptomics, epigenomics, and proteomics, into the Transformer framework to further improve genome prediction accuracy.

## Data Availability Statement

Data will be made available on request.

## Funding

This work was supported in part by Sichuan Science and Technology Program under Grant 2024NSFSC0444, and in part by the Fundamental Research Funds for the Central Universities under Grant SCU2023D008.

## Conflicts of Interest

The authors declare no conflicts of interest.

## AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

## Ethical Approval and Consent to Participate

Not applicable.

## References

[1] Cabanos, C., Matsuoka, Y., & Maruyama, N. (2021). Soybean proteins/peptides: A review on their importance, biosynthesis, vacuolar sorting, and accumulation in seeds. *Peptides, 143*, 170598. [CrossRef]

[2] Vargas-Almendra, A., Ruiz-Medrano, R., Núñez-Muñoz, L. A., Ramírez-Pool, J. A., Calderón-Pérez, B., & Xoconostle-Cázares, B. (2024). Advances in Soybean Genetic Improvement. *Plants, 13*(21), 3073. [CrossRef]

[3] Ravelombola, W., Qin, J., Shi, A., Song, Q., Yuan, J., Wang, F., … & Zhang, M. (2021). Genome-wide association study and genomic selection for yield and related traits in soybean. *PLOS ONE, 16*(8), e0255761. [CrossRef]

[4] Gao, P., Zhao, H., Luo, Z., Lin, Y., Feng, W., Li, Y., … & Wang, X. (2023). SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. *Briefings in Bioinformatics, 24*(6), bbad349. [CrossRef]

[5] VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science, 91*(11), 4414–4423. [CrossRef]

[6] Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics, 128*(6), 409–421. [CrossRef]

[7] Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics, 157*(4), 1819–1829. [CrossRef]

[8] Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2009). Genomic Selection Using Low-Density Marker Panels. *Genetics, 182*(1), 343–353. [CrossRef]

[9] Kärkkäinen, H. P., & Sillanpää, M. J. (2012). Back to Basics for Bayesian Model Building in Genomic Selection. *Genetics, 191*(3), 969–987. [CrossRef]

[10] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. [CrossRef]

[11] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. [CrossRef]

[12] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. [CrossRef]

[13] Su, Y., Xu, H., & Yan, L. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in China. *Saudi Journal of Biological Sciences, 24*(3), 537–547. [CrossRef]

[14] Wang, P., Lehti-Shiu, M. D., Lotreck, S., Segura Abá, K., Krysan, P. J., & Shiu, S. H. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications, 15*(1), 6856. [CrossRef]

[15] Mokry, F., Higa, R., de Alvarenga Mudadu, M., Oliveira de Lima, A., Meirelles, S. L., Barbosa da Silva, M. V., … & Correia de Almeida Regitano, L. (2013). Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. *BMC Genetics, 14*(1), 47. [CrossRef]

[16] Ennaji, O., Baha, S., Vergutz, L., & El Allali, A. (2024). Gradient boosting for yield prediction of elite maize hybrid ZhengDan 958. *PLOS ONE, 19*(12), e0315493. [CrossRef]

[17] Lourenço, V. M., Ogutu, J. O., Rodrigues, R. A. P., Posekany, A., & Piepho, H. (2024). Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC Genomics, 25*(1), 152. [CrossRef]

[18] Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., … & Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science, 22*(11), 961–975. [CrossRef]

[19] Tong, H., & Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology, 257*, 153354. [CrossRef]

[20] Hinton, G. E., & Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science, 313*(5786), 504–507. [CrossRef]

[21] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. [CrossRef]

[22] Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. *Briefings in Bioinformatics, 18*(5), 851–869. [CrossRef]

[23] Wang, K., Abid, M. A., Rasheed, A., Crossa, J., Hearne, S., & Li, H. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant, 16*(1), 279–293. [CrossRef]

[24] Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Frontiers in Genetics, 10*, 1091. [CrossRef]

[25] Li, W., Guo, Y., Wang, B., & Yang, B. (2023). Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction. *Pattern Recognition, 136*, 109234. [CrossRef]

[26] Wang, Z., Li, W., & Tang, Z. (2024). Enhancing the genomic prediction accuracy of swine agricultural economic traits using an expanded one-hot encoding in CNN models. *Journal of Integrative Agriculture*. [CrossRef]

[27] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780. [CrossRef]

[28] Zhang, Y., Qiao, S., Ji, S., & Li, Y. (2019). DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. *International Journal of Machine Learning and Cybernetics, 11*(4), 841–851. [CrossRef]

[29] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics, 37*(15), 2112–2120. [CrossRef]

[30] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., … & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods, 18*(10), 1196–1203. [CrossRef]

[31] Wu, C., Zhang, Y., Ying, Z., Li, L., Wang, J., Yu, H., … & Xu, X. (2023). A transformer-based genomic prediction method fused with knowledge-guided module. *Briefings in Bioinformatics, 25*(1), bbad438. [CrossRef]

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Kaiser, Ł. (2017). Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

[33] Browning, B. L., & Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics, 84*(2), 210–223. [CrossRef]

[34] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics, 81*(3), 559–575. [CrossRef]

[35] Ji, L., Hou, W., Zhou, H., Xiong, L., Liu, C., Yuan, Z., & Li, L. (2025). EBMGP: a deep learning model for genomic prediction based on Elastic Net feature selection and bidirectional encoder representations from transformer's embedding and multi-head attention pooling. *Theoretical and Applied Genetics, 138*(5), 1-15. [CrossRef]

[36] Lu, X., Liu, C., & Wang, J. (2025, May). Soybean genomic phenotype prediction method based on improving the transformer model with batch normalization and cosine annealing algorithm. In *International Conference on Artificial Intelligence and Machine Learning Research (CAIMLR 2024)* (Vol. 13635, pp. 227-233). SPIE. [CrossRef]

[37] Montesinos-López, O. A., Chavira-Flores, M., Kiasmiantini, Crespo-Herrera, L., Saint Piere, C., Li, H., … & Crossa, J. (2024). A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. *GENETICS, 228*(4), iyae161. [CrossRef]

[38] Consens, M. E., Diaz-Navarro, A., Chu, V., Stein, L., He, H. H., Moses, A., & Wang, B. (2025). Interpreting attention mechanisms in genomic transformer models: a framework for biological insights. *bioRxiv*, 2025-06. [CrossRef]

[39] Javed, N., Weingarten, T., Sehanobish, A., Roberts, A., Dubey, A., Choromanski, K., & Bernstein, B. E. (2025). A multi-modal transformer for cell type-agnostic regulatory predictions. *Cell Genomics, 5*(2), 100762. [CrossRef]

## Appendix

### A Proof of Lemma 1

*Proof.* Let $\boldsymbol{a}_{\max} = \max_i \boldsymbol{a}_i$, achieved at index $k$ ($k \in \{1, \ldots, n\}$). By construction, $\tilde{\boldsymbol{a}}_{\max} = \boldsymbol{a}_{\max} + \boldsymbol{b}_k$. For any $j \neq k$,

$$\begin{aligned} \tilde{\boldsymbol{a}}_{\max} - \tilde{\boldsymbol{a}}_j &= (\boldsymbol{a}_k + \boldsymbol{b}_k) - (\boldsymbol{a}_j + \boldsymbol{b}_j) \\ &= (\boldsymbol{a}_k - \boldsymbol{a}_j) + (\boldsymbol{b}_k - \boldsymbol{b}_j) \\ &\geq \boldsymbol{a}_k - \boldsymbol{a}_j, \end{aligned} \tag{A1}$$

since $\boldsymbol{b}_k \geq \boldsymbol{b}_j$. Thus, $\tilde{\boldsymbol{a}}_{\max} \geq \tilde{\boldsymbol{a}}_j$, with strict inequality if $\boldsymbol{a}_j < \boldsymbol{a}_{\max}$ or $\boldsymbol{b}_j < \boldsymbol{b}_k$.

The softmax function is monotonic, so $\boldsymbol{w}_k = \max_i \boldsymbol{w}_i$. To compare $\tilde{\boldsymbol{w}}_k$ and $\boldsymbol{w}_k$, observe:

$$\tilde{\boldsymbol{w}}_k = \frac{\exp(\tilde{\boldsymbol{a}}_k)}{\sum_{j=1}^{n} \exp(\tilde{\boldsymbol{a}}_j)} = \frac{\exp(\boldsymbol{a}_k + \boldsymbol{b}_k)}{\sum_{j=1}^{n} \exp(\boldsymbol{a}_j + \boldsymbol{b}_j)}, \tag{A2}$$

and

$$\boldsymbol{w}_k = \frac{\exp(\boldsymbol{a}_k)}{\sum_{j=1}^{n}\exp(\boldsymbol{a}_j)}. \tag{A3}$$

From (A2) and (A3), we have

$$\begin{aligned}\frac{\tilde{\boldsymbol{w}}_k}{\boldsymbol{w}_k} &= \frac{\exp(\boldsymbol{b}_k)\sum_{j=1}^{n}\exp(\boldsymbol{a}_j)}{\sum_{j=1}^{n}\exp(\boldsymbol{a}_j+\boldsymbol{b}_j)} \\ &= \frac{\exp(\boldsymbol{b}_k)}{\sum_{j=1}^{n}\exp(\boldsymbol{b}_j)\cdot\frac{\exp(\boldsymbol{a}_j)}{\sum_{i=1}^{n}\exp(\boldsymbol{a}_i)}}.\end{aligned} \tag{A4}$$

Let $\mathbb{E}_{\boldsymbol{w}}(\exp(\boldsymbol{b}_j)) = \sum_{j=1}^{n}\boldsymbol{w}_j\exp(\boldsymbol{b}_j)$, where $\mathbb{E}_{\boldsymbol{w}}(\cdot)$ represents the operation of taking the expectation of $\boldsymbol{w}$ with respect to it being treated as a random variable. The denominator in (A4) becomes $\sum_{j=1}^{n}\exp(\boldsymbol{b}_j)\boldsymbol{w}_j = \mathbb{E}_{\boldsymbol{w}}[\exp(\boldsymbol{b}_j)]$. Since $\boldsymbol{b}_k \geq \boldsymbol{b}_j$ for all $j$, and $\boldsymbol{b}_k > \boldsymbol{b}_j$ for $j \neq k$ (as $\boldsymbol{a}_k > \boldsymbol{a}_j$ for $j \neq k$ and $\boldsymbol{b}$ preserves order), we obtain

$$\mathbb{E}_{\boldsymbol{w}}[\exp(\boldsymbol{b}_j)] < \sum_{j=1}^{n}\boldsymbol{w}_j\exp(\boldsymbol{b}_k) = \exp(\boldsymbol{b}_k). \tag{A5}$$

Thus,

$$\frac{\tilde{\boldsymbol{w}}_k}{\boldsymbol{w}_k} = \frac{\exp(\boldsymbol{b}_k)}{\mathbb{E}_{\boldsymbol{w}}[\exp(\boldsymbol{b}_j)]} > 1 \implies \tilde{\boldsymbol{w}}_k > \boldsymbol{w}_k. \tag{A6}$$

Since $\tilde{\boldsymbol{w}}_k = \max_i \tilde{\boldsymbol{w}}_i$ and $\boldsymbol{w}_k = \max_i \boldsymbol{w}_i$, Lemma 1 holds. □

## B Proof of Theorem 1

*Proof.* For the $i$-th location query $\boldsymbol{Q}_i$, the attention scores of the $j$-th and $l$-th tokens are respectively given by

$$\boldsymbol{S}_{ij} = \frac{\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_k}},\ \boldsymbol{S}_{il} = \frac{\boldsymbol{Q}_i\boldsymbol{K}_l^{\mathrm{T}}}{\sqrt{d_k}}, \tag{A7}$$

where $\boldsymbol{K}_j$ and $\boldsymbol{K}_l$ are the key vectors at the $j$-th and $l$-th positions, respectively.

The attention weights obtained after applying the $\mathrm{softmax}(\cdot)$ for the $j$-th and $l$-th tokens are respectively

$$\boldsymbol{\alpha}_{ij} = \frac{\exp(\boldsymbol{S}_{ij})}{\sum_m \exp(\boldsymbol{S}_{im})},\ \boldsymbol{\alpha}_{il} = \frac{\exp(\boldsymbol{S}_{il})}{\sum_m \exp(\boldsymbol{S}_{im})}. \tag{A8}$$

Loading the bias terms $\boldsymbol{b}_{ij}$ and $\boldsymbol{b}_{il}$ onto the keys, we get

$$\tilde{\boldsymbol{K}}_j = \boldsymbol{K}_j + \boldsymbol{b}_{ij},\ \tilde{\boldsymbol{K}}_l = \boldsymbol{K}_l + \boldsymbol{b}_{il}, \tag{A9}$$

where

$$\boldsymbol{b}_{il} = \left[\boldsymbol{H}_{c(i)c(l)}, \boldsymbol{H}_{c(i)c(l)}, \ldots, \boldsymbol{H}_{c(i)c(l)}\right] \in \mathbb{R}^{1\times d_k}. \tag{A10}$$

From (A7), (A8) and (A9), the new attention scores are given by

$$\tilde{\boldsymbol{S}}_{ij} = \frac{\boldsymbol{Q}_i\tilde{\boldsymbol{K}}_j^{\mathrm{T}}}{\sqrt{d_k}} = \frac{\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_k}} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{ij}^{\mathrm{T}}}{\sqrt{d_k}}, \tag{A11}$$

$$\tilde{\boldsymbol{S}}_{il} = \frac{\boldsymbol{Q}_i\tilde{\boldsymbol{K}}_l^{\mathrm{T}}}{\sqrt{d_k}} = \frac{\boldsymbol{Q}_i\boldsymbol{K}_l^{\mathrm{T}}}{\sqrt{d_k}} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{il}^{\mathrm{T}}}{\sqrt{d_k}}. \tag{A12}$$

According to (A8), the corresponding new attention weights are

$$\tilde{\boldsymbol{\alpha}}_{ij} = \frac{\exp\left(\boldsymbol{S}_{ij} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{ij}^{\mathrm{T}}}{\sqrt{d_k}}\right)}{\sum_m \exp\left(\boldsymbol{S}_{im} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{im}^{\mathrm{T}}}{\sqrt{d_k}}\right)}, \tag{A13}$$

$$\tilde{\boldsymbol{\alpha}}_{il} = \frac{\exp\left(\boldsymbol{S}_{il} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{il}^{\mathrm{T}}}{\sqrt{d_k}}\right)}{\sum_m \exp\left(\boldsymbol{S}_{im} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{im}^{\mathrm{T}}}{\sqrt{d_k}}\right)}. \tag{A14}$$

Let $\Delta\boldsymbol{S}_{ij,l} = \boldsymbol{S}_{ij} - \boldsymbol{S}_{il}$ represent the difference in self-attention scores. Then, the score difference after introducing the bias can be expressed as

$$\begin{aligned}\Delta\tilde{\boldsymbol{S}}_{ij,l} &= \left(\boldsymbol{S}_{ij} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{ij}^{\mathrm{T}}}{\sqrt{d_k}}\right) - \left(\boldsymbol{S}_{il} + \frac{\boldsymbol{Q}_i\boldsymbol{b}_{il}^{\mathrm{T}}}{\sqrt{d_k}}\right) \\ &= \Delta\boldsymbol{S}_{ij,l} + \frac{\boldsymbol{Q}_i\left(\boldsymbol{b}_{ij}-\boldsymbol{b}_{il}\right)^{\mathrm{T}}}{\sqrt{d_k}}.\end{aligned} \tag{A15}$$

If for any $l \in \{1,\ldots,n\}$ and $l \neq j$,

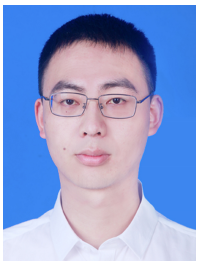
$$\frac{\boldsymbol{Q}_i\left(\boldsymbol{b}_{ij}-\boldsymbol{b}_{il}\right)^{\mathrm{T}}}{\sqrt{d_k}} > 0, \tag{A16}$$

then we have $\Delta\tilde{\boldsymbol{S}}_{ij,l} > \Delta\boldsymbol{S}_{ij,l}$.

From (A13), (A14) and (A15), it immediately follows that

$$\begin{aligned}\frac{\tilde{\boldsymbol{\alpha}}_{ij}}{\tilde{\boldsymbol{\alpha}}_{il}} &= \exp\left(\Delta\tilde{\boldsymbol{S}}_{ij,l}\right) \\ &= \frac{\boldsymbol{\alpha}_{ij}}{\boldsymbol{\alpha}_{il}}\cdot\exp\left(\frac{\boldsymbol{Q}_i\left(\boldsymbol{b}_{ij}-\boldsymbol{b}_{il}\right)^{\mathrm{T}}}{\sqrt{d_k}}\right) \\ &> \frac{\boldsymbol{\alpha}_{ij}}{\boldsymbol{\alpha}_{il}}.\end{aligned} \tag{A17}$$

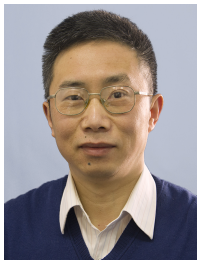
Additionally, if the token at the $j$-th position is the most important token, i.e., $\boldsymbol{\alpha}_{ij} > \boldsymbol{\alpha}_{il}$ for any $l \in \{1,\ldots,n\}$ and $l \neq j$. Then, according to Lemma 1, after adding the bias term, the corresponding attention weight will be further amplified, that is, $\tilde{\boldsymbol{\alpha}}_{ij} > \boldsymbol{\alpha}_{ij}$ holds. □



**Ailing Chen** received his B.S. and M.S. degrees in 2021 and 2025 in College of Mathematics respectively from Sichuan University, China. He is currently working in Ziyang, Sichuan, China. (Email: 13398306133@163.com)

**Qingke Zou** received the B.S. degree in information and computing science from Taiyuan University of Science and Technology, Taiyuan, China, in 2019, and the M.S. degree in mathematical statistics from Sichuan University, Chengdu, China, in 2022, where he is currently pursuing the Ph.D. degree in statistics. His research interests include information fusion, high-dimensional statistical inference, random matrices, and hyperspectral image processing. (Email: m18208258544@163.com)

**Xidi Yang** is currently working toward the Ph.D. degree in College of Animal Genetics and Breeding from Sichuan Agricultural University. He is Director of the Information Branch of the China Animal Husbandry Association and Director of the Sichuan Animal Husbandry and Veterinary Medicine Society. (Email: yangxidi@stu.sicau.edu.cn)



**Jie Zhou** received the B.S. and M.S. degrees in pure mathematics and the Ph.D. degree in probability theory and mathematical statistics from Sichuan University, Chengdu, China, in 1989, 1992, and 2003, respectively. He has been with the College of Mathematics, Sichuan University, since 1992, where he is currently a Professor. He has authored or co-authored three books and over 110 journals and conference proceedings papers. His research interests include detection and estimation, high-dimensional statistical inference, random matrices, information geometry, information fusion, statistical signal processing, nonlinear filtering, and target tracking. He serves as an Associate Editor for the Chinese Journal of Information Fusion. (Email: jzhou@scu.edu.cn)