Deep Learning accelerated correlation of genotypic and phenotypic data

M.Tech Thesis

by

Aayush Dhanesh Agrawal

DEPARTMENT OF COMPUTER SCIENCE

AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

INDORE

June 2025

Deep Learning accelerated correlation of genotypic and phenotypic data

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

in partial fulfillment of the requirements for

the award of the degree

of

Master of Technology

By

Aayush Agrawal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

June 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Deep Learning accelerated correlation of genotypic and phenotypic data** in the partial fulfillment of the requirements for the award of the degree of **Master of Technology** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the period from July 2023 to May 2025 under the supervision of Prof. Kapil Ahuja, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

20/May/2025
Signature of the Student with Date

(Aayush Dhanesh Agrawal)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

30/06/2025
Signature of Thesis Supervisor with Date

(Prof. Kapil Ahuja)

Aayush Dhanesh Agrawal has successfully given his M.Tech. Oral Examination held on **30/April/2025**.

30/06/2025
Signature(s) of Supervisor(s) of M.Tech. thesis

Date:

Subhra Mazumdar
Signature of Chairman, PG Oral Board

Date: 23.05.2025

Signature of HoD

Date: 23 May 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has supported me throughout this journey, making it a truly enriching and memorable experience.

First and foremost, I am deeply thankful to my supervisor, **Prof. Kapil Ahuja**, for his invaluable guidance, continuous encouragement, and unwavering support throughout the course of this work. His insightful suggestions and constructive feedback have been instrumental in shaping the direction and depth of my research.

I am also grateful to **Mr. Kuldeep Pathak**, PhD Scholar at IIT Indore, for their technical guidance and mentorship. Their constant support has helped me grow both technically and personally, and I remain truly appreciative of their contributions.

My heartfelt thanks to **Dr. Ranveer Singh**, Head of the Department of Computer Science and Engineering, for his kind support and encouragement during the course of my study.

I also extend my sincere appreciation to **Prof. Suhas S. Joshi**, Director, Indian Institute of Technology Indore, for providing a research-conducive environment and the opportunity to pursue my work at such a prestigious institute.

Lastly, I am forever indebted to my parents for their unconditional love, encouragement, and emotional support, which have been the backbone of my journey.

I am thankful to everyone who has directly or indirectly contributed to this work and stood by me throughout.

Aayush Agrawal

ABSTRACT

The prediction of phenotypic values based on genetic data is referred to as genomic prediction (GP). Genome-wide association studies (GWAS), on the other hand, look for correlations between genotypic markers (single nucleotide polymorphisms, SNPs) and phenotypic traits like grain yield and plant height in order to discover the key SNPs responsible for those traits. This study aims to address the distinct challenges of both GP and SNP identification. The rrBLUP and BLINK models are widely used for GP and GWAS, respectively. However, rrBLUP can only model simple linear relationships between genotype and phenotype, and BLINK often results in false positives when identifying SNPs. To address these challenges, we use machine learning approaches capable of capturing complicated, non-linear patterns, hence improving genomic prediction performance and SNP identification.

In this study, we evaluate popular ML model support vector regression (SVR) and its variants as well as the transformer-based GPformer, for their ability to improve predictive performance. Motivated by the difficulty of identifying significant SNPs in high dimensionalty low sample size SNP data, we initially create a hybrid model that combines the regression power of SVR with the feature interaction strength of self attention. Building on this breakthrough, we then reimagine the SNP sequence as a two dimensional, image like representation, a strategy that reveals spatial patterns in genomic variation by taming the curse of dimensionality and enabling potent image-based learning models.

Finally, our proposed model, ResGene18, builds on the ResNet18 architecture which is one of the most popular convolutional neural networks for image based tasks. ResGene18 is evaluated on two soybean datasets, exclusive ICAR and publicly available USDA. It consistently outperforms traditional statistical methods. On ICAR, it delivers a 51% improvement over rrBLUP, while on USDA it achieves an average gain of around 1%. Furthermore our model uncovers more significant SNPs for each trait across both datasets, identifying 57% more markers in ICAR and 34% more in USDA compared to the BLINK model.

By combining a deep learning backbone with a novel genomic to image data transformation, ResGene18 effectively addresses the dual challenges of genomic prediction and SNP identification. It not only improves phenotypic prediction performance but also uncovers meaningful genetic markers with higher precision, demonstrating its potential as a powerful tool for advancing genomic research in crops like soybean.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Introduction

Genotypic data, characterized by deoxyribonucleic acid (DNA) sequence variations such as single nucleotide polymorphisms (SNPs), provides a molecular level view of genetic diversity. These sequences are composed of four nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C), which serve as the fundamental building blocks of genetic information. Phenotypic data, on the other hand, captures measurable traits such as plant height, grain yield. The task of accurately predicting phenotypic traits from underlying genetic data has long been a central challenge in plant genomics. This process, commonly known as genomic prediction (GP), involves developing statistical or machine learning models to estimate trait values such as plant height or grain yield based solely on an individual's genotype, typically encoded as high-dimensional single nucleotide polymorphism (SNP) data. In parallel, genome-wide association studies (GWAS) have been widely used to uncover associations between specific genetic markers and observable traits, enabling researchers to identify the most influential SNPs driving phenotypic variation. While GP aims to maximize predictive accuracy, GWAS focuses on biological interpretability by SNP Identification. Traditional methods such as rrBLUP and BLINK have been widely adopted for GP and GWAS, respectively. rrBLUP, a ridge regression-based method, has demonstrated efficiency and ro-

bustness in handling large-scale genotype data; however, it is limited by its linear modeling assumption, making it less effective when the genotype–phenotype relationship is non-linear or influenced by complex interactions. On the other hand, BLINK, an iterative fixed-effect model tailored for GWAS, has improved statistical power and speed over earlier GWAS algorithms, yet it remains prone to producing false positives,

In this study, we have systematically evaluated the performance of popular machine learning models, including Support Vector Regression (SVR) and its variants, as well as the recently proposed transformer-based model *GPformer*, in the context of genomic prediction. These models have been assessed for their ability to improve predictive accuracy when dealing with high-dimensional SNP data. Motivated by the persistent challenge of identifying significant SNPs from such high-dimensional, low-sample datasets, we have initially proposed a hybrid framework that integrates the robust regression capabilities of SVR with the feature extraction power of a self-attention mechanism. This hybrid model has demonstrated improved performance by effectively capturing complex, non-linear interactions among SNPs. Building upon this breakthrough, we have further introduced a novel transformation strategy that reshapes the one-dimensional SNP sequence into a two-dimensional, image-like representation. Specifically, the SNP features have been reorganized into a 3D tensor where each channel corresponds to one of the 20 soybean chromosomes. This biologically meaningful restructuring has enabled the use of image based learning techniques, such as ResNet inspired convolutional architectures, to uncover spatial dependencies in the genomic data—thereby mitigating the curse of dimensionality and enhancing trait prediction performance.

Our proposed ResGene18 has been evaluated on two soybean datasets, first on the exclusive ICAR and then on the publicly available USDA dataset. It has consistently outperformed traditional statistical models delivering a striking 51% boost in prediction accuracy on ICAR and a more modest, yet meaningful, 1% gain on USDA compared to rrBLUP. However when it comes to uncovering trait-associated SNPs, ResGene18 has identified 57% more significant markers in the ICAR and 34% more in USDA than the BLINK model. This

highlights its powerful combination of prediction strength and discovery capability.

By integrating a deep learning backbone with a novel genomic-to-image transformation, **ResGene18** has successfully tackled the twin challenges of genomic prediction and SNP identification. It has enhanced phenotypic prediction accuracy and identified key genetic markers with improved precision, highlighting its potential as a valuable tool for advancing genomic research in crops such as soybean.

Chapter 2

Literature Survey

In this section, we divide our discussion into two key areas. The first focuses on traditional statistical techniques that have been widely employed to explore and quantify the correlation between genotypic variations and phenotypic traits. These techniques aim to uncover significant genetic markers associated with observable traits. The second area delves into the growing body of research on machine learning approaches developed for genomic prediction and the identification of informative single nucleotide polymorphisms (SNPs). These data-driven methods are designed to enhance predictive accuracy and detect meaningful patterns within high-dimensional genomic datasets.

2.1 Previous Studies Based on Statistical Methods

Kaler et al. [2020] have employed previously reported datasets comprising 346 soybean accessions with 31,260 SNPs and 279 maize accessions with 48,833 SNPs to evaluate association mapping models. They have compared eight statistical methods, spanning single-locus to multilocus approaches, across traits with varying heritability. The FarmCPU model has emerged as the most effective, accurately identifying SNPs near known genomic regions while minimizing both false positives and false negatives. In simulated datasets, FarmCPU has detected QTLs closer to the true number compared to other models. Unlike MLM based

methods that have proven overly conservative, FarmCPU paired with less stringent multiple testing corrections has produced more balanced and biologically relevant results. This work has highlighted FarmCPU's robustness across species with contrasting LD decay patterns, reinforcing its utility for genomic studies.

Wang et al. [2021] have conducted a comprehensive GWAS using a dataset of 259 re-sequenced rice accessions, generating 1,371.65 Gb of raw sequencing data and identifying 2.8 million SNPs. They have analyzed 13 agronomic traits including grain size, plant height, panicle length, and heading date by leveraging phenotypic data and BLUP values collected over two years. Principal component analysis (PCA) has been performed using GCTA to account for population structure, and breeding values have been estimated using the lme4 package in R. Their GWAS has identified 816 significant SNPs, with candidate genes located within 200 kb of these loci based on linkage disequilibrium patterns. Further haplotype analysis has helped refine the identification of genomic regions associated with key traits. This work has provided valuable candidate regions and SNPs for future gene validation and marker-assisted selection, offering a rich genomic resource for breeding high yielding rice cultivars.

Yoosefzadeh-Najafabadi et al. [2023] have carried out a comparative GWAS study using 227 soybean genotypes, selected after excluding 23 accessions due to high missing data. From an initial pool of 40,712 SNPs, they have retained 17,958 high-quality markers, which were mapped across all 20 soybean chromosomes. Their study has evaluated seed quality traits—namely protein content, oil percentage, and 100 seed weight using both the conventional FarmCPU method and a machine learning-based support vector regression (SVR) approach. Significant negative correlation between protein and oil contents has been observed, with respective heritability values of 0.69 and 0.67. SVR mediated GWAS has identified 13 SNPs linked to seed oil content (on chromosomes 3, 12, 13, 14, 15, and 16), while FarmCPU has identified 12 SNPs (on chromosomes 7, 8, 13, 15, and 19). The results have demonstrated that SVR is more effective in capturing trait relevant QTLs, particularly

by considering potential interactions often missed in traditional models. This study has emphasized the promising role of machine learning in advancing the accuracy of GWAS for marker-assisted breeding in soybean.

Paper	**Dataset Used**	**Methods Used**	**Conclusion**
Kaler et al. [2020]	346 soybean accessions (31,260 SNPs), 279 maize accessions (48,833 SNPs)	8 statistical models including MLM, FarmCPU, GLM, SUPER	FarmCPU outperformed other models by minimizing false positives/negatives and consistently identifying SNPs near known genes.
Wang et al. [2021]	259 rice accessions, 2.8 million SNPs	GWAS, BLUP, PCA (GCTA), haplotype analysis	Identified 816 SNPs significantly associated with 13 agronomic traits; provided valuable regions and candidate genes for marker-assisted selection.
Yoosefzadeh-Najafabadi et al. [2023]	227 soybean genotypes, 17,958 SNPs (filtered from 40,712)	GWAS using FarmCPU and SVR	SVR mediated GWAS has identified more relevant QTLs than FarmCPU, highlighting ML's potential in enhancing GWAS precision for seed traits.

Table 2.1: Summary of GWAS and Genomic Prediction Studies

2.2 Previous Studies Based on Machine Learning

Ma et al. [2018] have developed DeepGS, a deep convolutional neural network for predicting phenotypes from genotypes in genomic selection. They have used a dataset of 2,000 Iranian bread wheat accessions genotyped with 33,709 DArT markers. DeepGS has incorporated convolution, sampling, and dropout techniques to handle high-dimensional data and capture complex genotype-phenotype relationships. Compared to RR BLUP, DeepGS has achieved better prediction accuracy and shown complementary strengths. An ensemble of DeepGS and RR-BLUP has further improved selection performance, demonstrating that deep learning can enhance genomic selection outcomes.

Liu et al. [2019] have proposed a deep learning framework using convolutional neural networks (CNNs) to predict quantitative traits from SNP data in soybean. They have employed a dataset from the SoyNAM project, which includes over 5,000 recombinant inbred lines and 4,236 high-quality SNPs. By treating missing SNP values as a separate genotype class, their model has bypassed the need for imputation. The deep learning approach has achieved higher prediction accuracy than traditional statistical methods. Importantly, they have used saliency maps for feature selection, successfully identifying the most relevant SNPs and SNP combinations associated with the traits. This framework has demonstrated strong potential for both genomic prediction and interpretable marker discovery.

Grinberg et al. [2020] have explored the effectiveness of machine learning methods in genomic prediction using datasets from yeast, wheat, and rice. The yeast dataset has included 1,008 haploid strains with 11,623 Boolean markers, while the wheat dataset has involved 254 breeding lines genotyped with 33,516 SNPs. They have compared common machine learning models such as elastic net, lasso, ridge regression, random forest, GBM, and SVM with traditional statistical genetics approaches like genomic BLUP and a two step linear regression method. Their results have shown that machine learning models, particularly GBM and lasso for yeast, and SVM and BLUP for wheat and rice, have generally outperformed

classical methods. Random forest has emerged as the most robust model under conditions of noise and missing data, while BLUP has performed well in datasets with strong population structure. This study has highlighted both the promise and the complexity of applying machine learning to phenotype prediction tasks in genomics.

Wang et al. [2023] have introduced DNNGP, a deep neural network-based method for genomic prediction that integrates multi-omics data to improve trait prediction in plants. They have evaluated DNNGP using four datasets, including wheat2000 and tomato332, and compared its performance against five established models: GBLUP, LightGBM, SVR, DeepGS, and DLGWAS. Unlike traditional linear models, DNNGP has leveraged a hierarchical deep learning architecture with batch normalization and early stopping to dynamically learn complex genotype–phenotype relationships. The model has demonstrated superior accuracy, especially on large-scale datasets, while maintaining competitive performance even with smaller datasets. DNNGP has also outperformed DeepGS in computational efficiency, running up to 10 times faster, and has offered flexible hyperparameter tuning on local systems. These results have shown DNNGP to be a robust and scalable method for genomic prediction, particularly suited for modern breeding platforms involving high-dimensional omics data.

Wu et al. [2024] have introduced a novel Transformer-based genomic prediction model, *GPformer*, designed to capture long-range dependencies across SNPs for improved phenotype prediction. They have evaluated GPformer on diverse crop datasets, including soybean999 (7,883 SNPs), Maize282 (282 inbred lines with 3,093 SNPs), and Rice469 (469 indica rice accessions with 5,291 SNPs). Unlike traditional models, GPformer has leveraged an autocorrelation attention mechanism to extract relevant genomic signals regardless of physical SNP distance. As a key innovation, they have also developed a knowledge guided module (KGM) that integrates GWAS derived information into the model as prior knowledge. This combination GPformer + KGM has consistently outperformed mainstream methods such as RR-BLUP, SVR, LightGBM, and DNNGP across multiple evaluation metrics, including

MAE, PCC, and a novel Consistent Index (CI). The study has shown that GPformer is not only highly accurate but also robust to hyperparameter variations, making it well suited for practical breeding applications.

Wang et al. [2025] have proposed a novel genomic prediction framework called *WheatGP*, which integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to improve phenotype prediction in wheat. They have evaluated WheatGP using two datasets: Wheat599, consisting of 599 CIMMYT varieties genotyped with 1,279 markers, and Wheat2000, which includes 2,000 Iranian bread wheat varieties with 33,709 markers. The CNN component has captured short-range genomic dependencies, while the LSTM module has modeled long-range interactions between gene loci. Compared to rrBLUP, XGBoost, SVR, and DNNGP, WheatGP has demonstrated superior prediction accuracy, achieving up to 0.73 for wheat yield and between 0.62 and 0.78 for other agronomic traits. Additionally, they have employed Shapley Additive explanations (SHAP) to interpret the model by identifying the most influential genomic features. These results have shown that WheatGP is both accurate and interpretable, making it a robust and efficient tool for genomic selection and wheat breeding.

Paper	Dataset Used	Methods Used	Conclusion
Ma et al. [2018]	2000 Iranian wheat accessions, 33,709 DArT markers	DeepGS (CNN based), RR-BLUP	DeepGS has achieved better prediction than RR-BLUP; ensemble of both has further enhanced performance.
Liu et al. [2019]	SoyNAM: 5000+ RILs, 4236 SNPs	CNN with saliency maps, no imputation	DL model has outperformed statistical methods; saliency maps have identified key SNPs for trait prediction.
Grinberg et al. [2020]	Yeast (1008 strains, 11,623 markers), Wheat (254 lines, 33,516 SNPs)	ML models: GBM, RF, SVM, EN, Lasso, Ridge; vs. BLUP, 2-step LR	ML methods (esp. GBM, lasso, SVM) have outperformed classical methods; RF most robust to noise; BLUP good with population structure.
Wang et al. [2023]	Wheat2000, Tomato332 + 2 more datasets	DNNGP (deep neural net) vs. GBLUP, LightGBM, SVR, DeepGS, DLGWAS	DNNGP has shown superior accuracy and speed; scalable and efficient for high-dimensional multi-omics prediction.
Wu et al. [2024]	Soybean999 (7883 SNPs), Maize282 (3093 SNPs), Rice469 (5291 SNPs)	GPformer (Transformer + KGM), vs. RR-BLUP, SVR, LightGBM, DNNGP	GPformer + KGM has outperformed all baselines in MAE, PCC, CI; robust across datasets and phenotypes.
Wang et al. [2025]	Wheat599 (1279 SNPs), Wheat2000 (33,709 SNPs)	WheatGP (CNN + LSTM) + SHAP, vs. RR-BLUP, SVR, DNNGP, XGBoost	WheatGP has achieved high accuracy (up to 0.73); SHAP used for model interpretability; suitable for breeding pipelines.

Table 2.2: Summary of Machine Learning Based Genomic Prediction Studies

Chapter 3

Methods & Experiments

This section provides a detailed account of the experiments, including the methodologies employed, the datasets used, and the model architectures tested

3.1 Dataset

The following section has discussed two datasets: the first, referred to as the *ICAR dataset*, while the second is the publicly available *USDA dataset*, which is described in the subsequent subsection.

3.1.1 Dataset specification

The dataset used in this study has been exclusively obtained from ICAR-IISR[1], and contains both genotypic and phenotypic data for 269 soybean varieties cultivated across central India. Phenotyping has been conducted during the summer season (mid-June to mid-October) over three consecutive years, from 2019 to 2021. The genotypic data includes 66,589 Single Nucleotide Polymorphisms (SNPs), which serve as genetic markers to analyze variation among the soybean varieties. As shown in Table 3.1, columns represent SNPs distributed across all 20 chromosomes of the soybean genome, while rows correspond to the

[1]ICAR-Indian Institute of Soybean Research, Indore, M.P., India

individual varieties used for analysis.

Varieties	S1_60978	S1_62104	$\cdots$	S20_50278263	S20_50287457
MP_7_GW-1	A	K	$\cdots$	T	W
MP_4_SOY-523	A	K	$\cdots$	C	W
MP_4_SOY-520	A	K	$\cdots$	C	T
MP_7_GW-6	R	K	$\cdots$	T	T
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
MP_3_SOY-403	R	G	$\cdots$	T	W

Table 3.1: SNP (Genotype) Matrix

The phenotypic dataset have four important phenotype traits such as Plant Height (PH), Number of Nodes (NN), Grain Yield (GY), and Canopy Temperature (CT). The short summary of the phenotypic dataset is given in the Table 3.2.

Varieties	PH	NN	GY	CT
MP_7_GW-1	48.83	6.86	5.45425	27.37
MP_4_SOY-523	54.04	10.65	1.45425	28.83
MP_4_SOY-520	54.24	9.45	2.07425	29.00
MP_7_GW-6	54.64	6.45	2.07425	27.99
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
MP_3_SOY-403	42.93	3.425	1.11125	30.76

Table 3.2: Phenotypic Trait Values for Soybean Varieties

Furthermore, we have used another publicly available soybean dataset to evaluate our model's performance on. This dataset comprises $20,087$ varieties from the USDA Soybean Germplasm Collection, genotyped with the SoySNP50K, which we have downloaded from SoyBase (`https://www.soybase.org/dlpages/#snp50k`; accessed April 13, 2025). From these data, we have extracted Height, Oil, Protein and Yield measurements for the $1,170$ varities reported by Hill et al. [2008]. After filtering to retain only common varities, we

proceed to evaluate both existing and our proposed model on this dataset as well. In following sections, we refer to the ICAR-IISR dataset simply as the ICAR dataset, and the USDA Soybean dataset as the USDA dataset.

3.1.2 Dataset preprocessing

Before training the models, we preprocess the genotypic dataset by encoding the SNP values into a numerical format suitable for machine learning algorithms. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the original genotypic data matrix, where n represents the number of soybean varieties (i.e., 269), and d denotes the number of single nucleotide polymorphisms (SNPs) Markers or features (i.e., $66,589$). Correspondingly, let $\mathbf{y} \in \mathbb{R}^{n \times 1}$ represent the phenotypic vector associated with each soybean phenotypic trait. However, each phenotypic trait corresponds to a distinct $\mathbf{y}$, while maintaining the same dimensionality.

Following the encoding scheme described in Lipka et al. [2012], homozygous alleles[2] A and T are denoted by 0, while C and G are denoted by 2 to reflect genetic similarity, heterozygous alleles[3] (R, Y, S, W, K, and M) are denoted by 1 to represent their mixed genetic nature. The character 'N', which indicates a missing value in the SNP data, is denoted by -1. This transformation yields the encoded genotypic matrix $\tilde{\boldsymbol{X}}$, which is more computationally tractable and suitable for subsequent analyses.

[2]Homozygous alleles refer to identical versions of a gene inherited from both parent, located at the same locus on homologous chromosomes.

[3]Heterozygous alleles refer to different versions of a gene inherited from each parent at the same locus, resulting in genetic variation at that position.

Varieties	S1_60978	S1_62104	$\cdots$	S20_50278263	S20_50287457
MP_7_GW-1	0	1	$\cdots$	0	1
MP_4_SOY-523	0	1	$\cdots$	2	1
MP_4_SOY-520	0	1	$\cdots$	2	0
MP_7_GW-6	1	1	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
MP_3_SOY-403	1	2	$\cdots$	0	1

Table 3.3: Encoded SNP Matrix

3.2 State of the art Statistical model with application to Genomic Prediction and GWAS

Statistical methods have long been at the core of genomic prediction, and numerous studies have employed and compared them across diverse datasets and traits. These methods have provided a strong foundation for evaluating genetic potential, especially in early-stage breeding. The following section highlights one of the most widely used and effective statistical approaches for genomic prediction.

3.2.1 The ridge regression best linear unbiased prediction (rrBLUP)

Before the widespread adoption of machine learning techniques in genomic studies, statistical models such as the Mixed Linear Model (MLM) and its compressed variant (CMLM) have been commonly used for Genome-Wide Association Studies (GWAS) Zhang et al. [2010]. Among these, rrBLUP method has gained prominence as a standard approach for both genomic prediction and GWAS. Its strength lies in effectively handling high-dimensional SNP data while accounting for population structure. By combining ridge regression with the BLUP framework, rrBLUP is able to estimate marker effects efficiently within a mixed

model setup Endelman [2011].

In our study, we have applied rrBLUP to both the ICAR and USDA soybean datasets using the `rrBLUP` package in R. We have followed a 10-fold cross-validation strategy to assess the model's predictive performance consistently across the datasets. This has allowed us to benchmark rrBLUP against advanced machine learning models under a unified evaluation framework.

3.2.2 Bayesian information and Linkage disequilibrium(BLINK)

Over time, Genome-Wide Association Studies (GWAS) have advanced through a series of methodological improvements. The Mixed Linear Model (MLM) was initially favored for its ability to control false positives, followed by the Compressed MLM (CMLM) for better scalability. Later, FarmCPU improved statistical power by separating fixed and random effects. The most recent and widely adopted method is BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway) Huang et al. [2019], which enhances speed and accuracy by using Bayesian Information Criterion (BIC) and removing the assumption of normally distributed markers.

In our study, we have used the BLINK method to identify trait-associated SNPs, where lower p-values correspond to higher statistical significance. As shown in Section 4.4, BLINK has effectively identified several SNPs associated with our traits of interest. We have implemented BLINK using the GAPIT library, an R package specifically designed for genome-wide association analysis, on both the ICAR and USDA soybean datasets. BLINK has offered superior power and computational efficiency by leveraging linkage disequilibrium and Bayesian Information Criterion (BIC) for model selection Huang et al. [2019]. This has allowed us to detect true associations more accurately, especially in the presence of complex population structures. As a result, BLINK has served as a reliable and state-of-the-art approach in our GWAS analysis.

3.3 Basic Machine Learning Models with direct application to Genomic Prediction

3.3.1 Support Vector Regression (SVR)

We have used Support Vector Regression (SVR) as a baseline model for genomic prediction, owing to its proven ability to handle high-dimensional SNP datasets with strong generalization capabilities. Prior studies have demonstrated that SVR performs competitively in genomic contexts by effectively managing the curse of dimensionality and providing robust predictive accuracy Basak et al. [2007].

SVR works by mapping input features into a high-dimensional space using kernel functions and learns a linear function within an ϵ-insensitive margin, allowing the model to tolerate small deviations while maintaining predictive sharpness. We have employed a linear kernel. The SVR model has been trained using 10-fold cross-validation, ensuring that each sample is tested once and trained nine times. Additionally, we have implemented an 90-10 data split, where 90% of the data is used for training and the remaining 10% is reserved for evaluating the model's generalization performance.

3.3.2 Variant SVR

To further enhance prediction performance and address the challenges of high-dimensional SNP data, we have implemented a customized variant of SVR. This variant incorporates L2 regularization with an ϵ-insensitive loss function, which has made the model more robust to noise and well-suited for datasets with limited samples but large numbers of features. Inspired by recent studies that introduced semismooth optimization techniques in SVR frameworks Yin and Li [2019], we have adopted a convex optimization approach to solve this formulation.

The variant has been implemented using `CVXPY`, a Python-based modeling tool that

enables efficient specification of objective functions and constraints. This framework has allowed us to define the optimization problem clearly and solve it with high precision. We have applied 10-fold cross-validation for model reliability, and like the baseline SVR, the dataset has been divided using an 90-10 train-test split. The model has demonstrated improved control over overfitting while still capturing meaningful genotype–phenotype relationships, particularly in the presence of noisy and sparse signal patterns across SNPs.

3.3.3 GPFormer Model

In addition to SVR-based approaches, we have employed the GPFormer model, a Transformer-based deep learning architecture specifically designed for genomic prediction tasks Lu et al. [2025]. This model has outperformed traditional machine learning methods across multiple trait predictions in prior research. GPFormer adopts an encoder–decoder architecture and integrates critical deep learning components such as multi-head self-attention, residual connections, positional encoding, batch normalization, and dropout.

These design elements have enabled GPFormer to effectively capture long-range dependencies among SNPs, which is a limitation in many conventional models. The self-attention mechanism has allowed the model to focus on relevant SNP regions without relying on feature selection heuristics. For training, we have followed the same 10-fold cross-validation procedure and retained an 90-10 split for evaluation. The final phenotype prediction has been obtained from the output of a fully connected layer applied to the final hidden state of the encoder. This design has led to high model stability, reduced overfitting, and consistent performance across different trait prediction scenarios.

3.4 Advance ML models with novel application to Genomic Prediction and GWAS

3.4.1 Attention with Support Vector Regression

We have faced a major challenge due to the high dimensionality of SNP data, which makes the direct application of self-attention computationally expensive. Traditional self-attention mechanisms scale quadratically with input size, which is impractical for large genomic datasets.

To address this, we have designed a lightweight self-attention mechanism where the entire SNP sequence of a sample is treated as a single token. This approach reduces the computational complexity and allows self-attention to operate in linear time, generating an self-attention score matrix across SNPs. The resulting matrix has been used as input to a Support Vector Regression (SVR) model.

This hybrid model has shown slightly better prediction performance than standalone SVR and has identified a greater number of informative SNPs, indicating that the self-attention mechanism has effectively captured relevant patterns in the data.

3.4.2 Attention with Variant of Support Vector Regression

To further explore the effectiveness of self-attention, we have also applied the same self-attention mechanism in combination with a variant of SVR. This experiment was intended to evaluate whether the benefits of self-attention persist across different SVR formulations. The performance of this model has been found to be comparable to the original Attention with SVR model, confirming the generalizability of the self-attention based feature extraction approach.

3.4.3 ResGene18

High dimensionality of SNP data has remained a persistent challenge in genomic prediction tasks. To address this, we have looked into existing literature and found that Muneeb et al. [2022] have transformed SNP data into a 2D matrix format to apply convolutional neural networks. However, their dataset involved low-dimensional SNPs and was used for classification, making the approach unsuitable for our high-dimensional regression-based problem.

Motivated by this, we have proposed a novel transformation technique tailored for high-dimensional SNP data. Specifically, we have reshaped the SNP input into a 3D format, We have represented the SNP data using 20 channels, where each channel corresponds to one of the 20 soybean chromosomes, as illustrated in Figure 3.1. This chromosome-wise decomposition has not only preserved biological relevance but has also enabled the use of deep learning techniques for effective feature extraction. Additionally, this multi-channel strategy aligns with the concept of input channels in image-based models, allowing us to break the high-dimensional SNP data into manageable chunks. Our decision to segment the data in this way has been motivated by the biological distribution of SNPs across 20 chromosomes, thereby introducing a meaningful bias that supports both interpretability and model efficiency. To the best of our knowledge, we are the first to apply such a genomic-to-image transformation for both genomic prediction and GWAS tasks. Owing to the unique nature of this transformation and its integration with a lightweight deep learning backbone, we have named our architecture **ResGene18**, where the '18' corresponds to the 18 layers of the ResNet18 model chosen for being the simplest and least complex variant in the ResNet family, yet sufficiently powerful for our application. We have trained the ResNet18 model using 10-fold cross-validation to ensure robust and unbiased evaluation across different subsets of the data. For optimization, we have used Stochastic Gradient Descent (SGD), which is known to perform well on image-based datasets by enabling smoother convergence and

better generalization. This combination of architecture and training strategy has allowed the model to learn meaningful spatial dependencies from the SNP image representations.

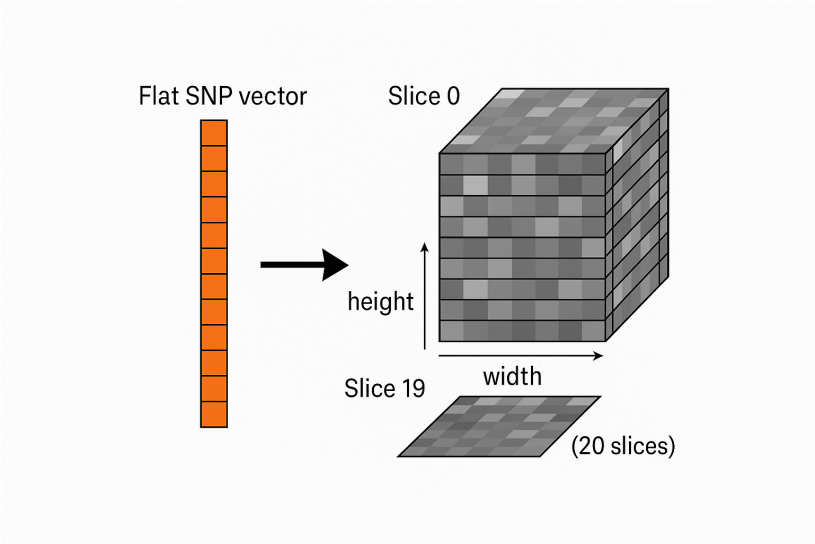

Figure 3.1: Transformation of sequential SNP data into 2D multi-channel image format.

As a result, our proposed model, ResGene18, has achieved the highest Pearson's correlation coefficient (PCC) among all models, including traditional statistical models, baseline machine learning methods, and our other proposed hybrids as demonstrated in sections 4.2. Additionally, ResGene18 has identified a greater number of informative SNPs compared to all other models, highlighting its effectiveness in both Genomic prediction and SNP Identification as demonstrated in section 4.3 and 4.4

Chapter 4

Results

4.1 Experimental Setup

For our machine learning experiments, we have utilized the Kaggle platform, running Python version 3.11.11. The ResNet18 architecture has been implemented using the PyTorch library to process SNP image representations for genomic prediction. On the other hand, statistical analyses including rrBLUP and the BLINK model have been carried out using the `rrBLUP` and `GAPIT` packages within RStudio. The computational environment has included an Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100-PCIe-16GB GPU, providing sufficient processing power for deep learning and statistical modeling tasks.

4.2 Evaluation Metric

For evaluating our model, we have used the following well-known metrics, which are described below.

Pearson Correlation Coefficient (PCC): This metric evaluates the linear relationship between the predicted and actual values. It indicates how closely the two sets of values align on a straight line. A PCC value close to 1 suggests a strong positive correlation, meaning the predictions closely follow the actual trend. A value near 0 implies no linear

relationship, while a negative value indicates an inverse correlation. This metric is especially useful in assessing how well a model captures the direction and strength of the relationship between input features and the target trait Cohen et al. [2009].

The mathematical formulation of PCC metrics As shown in Equation (4.1), where y_i indicates the actual values for the i^{th} sample, y_i' represents the predicted values, and $\bar{y}$ is the mean of the actual values. Similarly, $\bar{y}'$ denotes the mean of the predicted values (i.e., $\bar{y}' = \frac{1}{n}\sum_{i=1}^{n} y_i'$).

$$\text{PCC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i' - \bar{y}')}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i' - \bar{y}')^2}} \tag{4.1}$$

In soybean Genome-Wide Association Studies (GWAS), PCC (Pearson Correlation Coefficient) is especially helpful. GWAS seeks to identify genetic variant/trait associations. Using PCC in this type of analysis permits researchers to evaluate the linear correlation between predicted and actual trait values based on genetic markers. A higher PCC value indicates a stronger linear relationship between specific genetic variants and the target trait, thereby aiding in identification of significant genetic components influencing soybean traits. Additionally, genomic prediction studies frequently adopt PCC as a primary evaluation metric due to its interpretability and effectiveness in assessing model performance Sehrawat et al. [2023].

4.3 Model Evaluation

4.3.1 Model Evaluation: ICAR Dataset

The ICAR dataset has comprised approximately 66,000 SNP markers, 269 soybean accessions, and four phenotypic traits. A comparative analysis of all the methods is presented in the Table below.

Trait	**Statistical**	**Existing ML Models**			**Our Proposed ML Models**			**% Gain**
	rrBLUP	SVR	Variant SVR	GPFormer	ATT+SVR	ATT+Variant SVR	ReGene18	
PH	0.331	0.421	0.4105	0.4642	0.436	0.394	**0.485**	46.53
NN	0.3422	0.376	0.3237	0.3892	0.3778	0.241	**0.4094**	19.64
GY	0.3208	0.201	0.174	0.2845	0.224	0.211	**0.3221**	0.41
CT	0.079	0.097	0.0997	0.1044	0.072	0.012	**0.1878**	137.72
Average % Gain between rrBLUP and ResGene18								**51.07**

Table 4.1: Comparison of PCC values across different models and traits on ICAR dataset. The % Gain column represents the relative improvement of ResGene18 over rrBLUP.

As demonstrated in Table 4.1, machine learning-based approaches have outperformed the statistical method rrBLUP. Notably, among all the methods tested, our proposed model ResGene18 has achieved the highest accuracy across all traits, outperforming statistical techniques and other machine learning models, thereby highlighting its strong potential for genomic prediction tasks.

4.3.2 Model Evaluation: USDA Dataset

Similar algorithms have also been evaluated on the USDA soybean dataset, which includes 1,173 accessions, approximately 42,000 SNP markers, and four phenotypic traits: plant height, oil content, protein content, and grain yield. On this dataset, the traditional

statistical method rrBLUP has outperformed all machine learning models, including SVR, its variant, GPFormer, and our proposed models Attention+SVR and Attention+Variant SVR. However, among all approaches, only our final proposed model ResGene18 has consistently delivered the highest prediction accuracy, achieving the best Pearson correlation coefficients (PCC) across all four traits. This demonstrates the robustness and superior performance of ResGene18 for genomic prediction, even on large and complex datasets, as evidenced in the results shown in Table 4.2

Trait	**Statistical**	**Existing ML Models**			**Our Proposed ML Models**			**% Gain**
	rrBLUP	SVR	Variant SVR	GPFormer	ATT+SVR	ATT+Variant SVR	ResGene18	
HEIGHT	0.7341	0.687	0.7242	0.6485	0.703	0.704	**0.7392**	0.69
OIL	0.7098	0.652	0.6006	0.6585	0.657	0.606	**0.7112**	0.20
PROTEIN	0.6662	0.611	0.6587	0.6453	0.636	0.662	**0.6786**	1.86
YIELD	0.7553	0.736	0.7378	0.7076	0.737	0.726	**0.7604**	0.68
Average % Gain between rrBLUP and ResGene18								**0.86**

Table 4.2: Comparison of PCC values across different models and traits on USDA dataset. The % Gain column represents the relative improvement of ResGene18 over rrBLUP.

4.4 SNP Identification

Single Nucleotide Polymorphism (SNP) detection plays a crucial role in Genome-Wide Association Studies (GWAS), as it helps uncover genetic variations linked to complex traits and diseases Fang et al. [2017]. Identifying significant SNPs allows researchers to pinpoint genomic regions influencing traits of interest, thereby supporting marker-assisted selection (MAS) in breeding programs and contributing to advancements in personalized medicine. To further demonstrate the effectiveness of our model, we compared the SNPs it identified with those reported as ground truth on the SoyBase database Grant et al. [2010]. Specifically, we assessed the top 20 SNPs identified by our model ResGene18 against the top 20 discovered by BLINK, a leading statistical method widely recognized for its GWAS performance Huang et al. [2019]. For a fair comparison, we considered only those SNPs where the genomic position difference between the identified SNP and the SoyBase ground truth was less than 3 million base pairs.

4.4.1 SNP Identification on ICAR Dataset

Trait	BLINK	ResGene18	% Difference
PH	9	11	22.22
NN	4	6	50.00
GY	6	11	83.33
CT	4	7	75.00
Average % Difference			**57.14**

Table 4.3: Comparison of number of SNPs identified by BLINK and ResGene18 on the ICAR dataset.

As demonstrated in Table 4.3 our proposed model ResGene18 and the statistical model BLINK is employed to identify significant SNPs associated with four traits in the ICAR dataset. ResGene18 has consistently identified a higher number of SNPs. On average, ResGene18 has achieved a **57.14%** improvement in SNP identification compared to BLINK.

Chr	Identified Position	Identified Position by BLINK	Difference
S18	8,658,515	9,263,941	605,426
S18	5,908,800	5,271,342	637,458
S18	5,908,800	5,244,535	664,265
S3	42,527,196	41,769,136	758,060
S3	42,527,196	41,769,127	758,069
S3	42,527,196	41,699,688	827,508
S5	4,300,531	2,899,164	1,401,367
S18	50,198,744	51,620,945	1,422,201
S13	13,259,705	15,903,319	2,643,614
		Total SNPs identified by BLINK	**9**
Chr	**Identified Position**	**Identified Position by ResGene18 (bp)**	**Difference**
S18	4,702,682	4,522,238	180,444
S13	34,916,010	34,676,862	239,148
S20	46,223,443	45,694,464	528,979
S18	53,267,229	52,532,933	734,296
S20	46,223,443	45,100,008	1,123,435
S15	25,719,454	24,648,725	1,070,729
S5	4,300,531	2,778,975	1,521,556
S18	50,198,744	52,532,914	2,334,170
S15	26,551,533	29,011,688	2,460,155
S12	36,038,398	39,034,317	2,995,919
S12	36,038,398	39,034,638	2,996,240
		Total SNPs identified by ResGene18	**11**

Table 4.4: SNPs identified for the trait **Plant Height** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chr	Ground Position	Identified Position - BLINK	Difference
S16	37,079,553	37,059,512	20,041
S16	37,079,553	36,999,780	79,773
S19	43,990,450	43,165,754	824,696
S11	4,980,454	3,837,243	1,143,211
		Total SNPs identified by BLINK	**4**
Chr	**Ground Position**	**Identified Position - ResGene18 (bp)**	**Difference**
S19	43,990,450	43,454,424	536,026
S11	4,980,454	6,060,734	1,080,280
S19	43,990,450	42,881,920	1,108,530
S13	32,115,483	33,952,562	1,837,079
S13	32,115,483	34,654,698	2,539,215
S18	55,620,032	52,932,922	2,687,110
S18	55,808,363	52,832,933	2,975,430
		Total SNPs identified by ResGene18	**7**

Table 4.5: SNPs identified for the trait **Number of Nodes** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chr	Ground Position	Identified Position - BLINK	Difference
S7	6,396,861	6,152,396	244,465
S16	33,786,653	32,831,101	955,552
S16	33,786,653	32,792,281	994,372
S6	47,750,740	49,061,490	1,310,750
S6	47,750,740	49,076,176	1,325,436
S13	37,033,440	38,828,253	1,794,813
		Total SNPs identified by BLINK	**6**
Chr	**Ground Position**	**Identified Position - ResGene18**	**Difference**
S10	176,869	193,685	16,816
S19	43,022,543	42,863,274	159,269
S10	1,020,657	1,111,345	90,688
S7	1,300,300	1,697,640	397,340
S7	1,300,300	1,697,673	397,373
S19	43,022,543	43,847,204	824,661
S10	1,020,657	184,336	836,321
S18	5,153,977	3,883,552	1,270,425
S11	7,897,730	8,913,591	1,015,861
S12	36,324,462	38,308,336	1,983,874
S12	36,324,462	39,250,317	2,925,855
		Total SNPs identified by ResGene18	**11**

Table 4.6: SNPs identified for the trait **Grain Yield** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chr	Ground Position	Identified Position - BLINK (bp)	Difference
S11	7,251,966	6,988,207	263,759
S11	7,251,966	8,185,858	933,892
S11	7,251,966	6,240,051	1,011,915
S11	7,251,966	6,189,928	1,062,038
		Total SNPs identified by BLINK	**4**
Chr	**Ground Position**	**Identified Position - ResGene18**	**Difference**
S13	34,845,629	34,676,862	168,767
S13	34,845,629	34,654,698	190,931
S17	7,536,244	6,603,695	932,549
S6	12,426,395	10,774,194	1,652,201
S17	7,536,244	5,665,674	1,870,570
S6	12,426,395	9,819,431	2,606,964
S6	12,426,395	9,804,189	2,622,206
		Total SNPs identified by ResGene18	**7**

Table 4.7: SNPs identified for the trait **Canopy Temprature** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Remarkably, our model identified a higher number of SNPs overlapping with SoyBase-verified loci compared to the widely used statistical method BLINK, highlighting its superior ability for SNP identification in genomic studies. Across all four traits in the ICAR dataset — Plant Height (PH), Number of Nodes (NN), Grain Yield (GY), and Canopy Temperature (CT) — our model consistently outperformed BLINK, identifying 2, 2, 5, and 3 additional SNPs respectively. This demonstrates the model's robustness and effectiveness in capturing meaningful genetic signals.

4.4.2 SNP Identification on USDA Dataset

Trait	BLINK	ResGene18	% Difference
HEIGHT	10	13	30.00
OIL	13	14	7.69
PROTEIN	8	11	37.50
YIELD	7	11	57.14
Average % Difference			**33.58**

Table 4.8: Comparison of number of SNPs identified by BLINK and ResGene18 on the USDA dataset. % Difference is calculated between ResGene18 and BLINK.

In the case of the USDA dataset, the comparison between the baseline statistical approach BLINK and our proposed deep learning model ResGene18 has revealed that ResGene18 has successfully detected a greater number of SNPs. This improvement is reflected in an average percentage increase of **33.58%**.

Chromosome	Identified Position	Identified Position - BLINK	Difference
S07	6,598,470	6,875,359	276,889
S08	42,402,751	42,090,911	311,840
S10	46,373,097	46,729,408	356,311
S10	46,373,097	46,730,455	357,358
S09	46,006,470	45,242,813	763,657
S04	5,241,170	4,239,539	1,001,631
S10	7,750,656	6,196,008	1,554,648
S03	35,565,679	33,636,101	1,929,578
S07	15,350,606	14,106,459	1,244,147
S09	46,006,470	48,543,094	2,536,624
		Total SNPs identified by BLINK	**10**
Chromosome	**Identified Position**	**Identified Position - ResGene18**	**Difference**
S7	4,535,039	4,623,018	87,979
S18	5,908,800	5,735,470	173,330
S2	41,290,024	41,544,629	254,605
S18	50,325,432	50,590,190	264,758
S1	55,166,202	54,750,659	415,543
S3	43,679,222	43,025,036	654,186
S2	41,290,024	40,730,110	559,914
S18	50,325,432	49,301,920	1,023,512
S20	40,857,494	42,125,878	1,268,384
S9	46,006,470	47,670,973	1,664,503
S18	50,325,432	48,132,642	2,192,790
S5	4,300,531	2,238,811	2,061,720
S8	43,005,079	45,517,454	2,512,375
		Total SNPs identified by ResGene18	**13**

Table 4.9: SNPs identified for the trait **Height** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chromosome	Identified Position	Identified Position - BLINK	Difference
S5	41,780,982	41,855,235	74,253
S9	43,972,043	43,303,153	668,890
S6	49,834,614	48,850,059	984,555
S5	2,661,929	3,755,641	1,093,712
S4	7,855,555	9,014,045	1,158,490
S7	9,339,240	8,021,155	1,318,085
S3	37,272,628	35,837,462	1,435,166
S03	43,679,222	45,151,144	1,471,922
S2	46,806,494	45,316,774	1,489,720
S9	10,384,779	11,970,660	1,585,881
S5	2,661,929	4,904,466	2,242,537
S11	22,911,236	25,114,851	2,203,615
S8	8,281,543	10,700,780	2,419,237
S4	7,855,555	9,104,210	1,248,655
		Total SNPs identified by BLINK	**14**
Chromosome	**Identified Position**	**Identified Position - ResGene18**	**Difference**
S1	54,711,960	54,698,255	13,705
S7	4,986,841	4,883,153	103,688
S10	1,020,657	1,432,529	411,872
S5	2,661,929	2,207,089	454,840
S7	4,986,841	5,457,246	470,405
S15	3,828,587	3,319,227	509,360
S3	43,679,222	44,343,929	664,707
S6	6,291,064	3,313,294	2,977,770
S2	42,237,432	40,804,255	1,433,177
S7	6,623,116	5,456,442	1,166,674
S2	42,237,432	40,730,110	1,507,322
S17	6,418,985	4,955,428	1,463,557
S20	40,857,494	42,763,762	1,906,268
S17	6,418,985	3,924,278	2,494,707
		Total SNPs identified by ResGene18	**14**

Table 4.10: SNPs identified for the trait **Oil** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chromosome	Identified Position	Identified Position - BLINK	Difference
S7	37,126,884	36,936,795	190,089
S9	2,238,015	2,669,838	431,823
S6	5,666,361	6,214,894	548,533
S3	37,272,628	37,895,336	622,708
S9	46,053,138	45,099,143	953,995
S7	37,126,884	36,237,935	888,949
S3	37,272,628	35,507,990	1,764,638
S4	7,855,555	10,193,474	2,337,919
		Total SNPs identified by BLINK	**8**
Chromosome	**Identified Position**	**Identified Position - ResGene18**	**Difference**
S13	18,150,116	18,211,337	61,221
S7	6,623,116	6,198,644	424,472
S15	4,045,527	4,475,844	430,317
S15	4,045,527	4,481,538	436,011
S2	39,587,099	39,936,971	349,872
S8	46,415,576	45,529,194	886,382
S15	4,045,527	3,319,227	726,300
S18	50,815,522	49,747,507	1,068,015
S20	40,857,494	42,125,878	1,268,384
S19	40,540,660	38,209,047	2,331,613
S20	40,857,494	37,971,425	2,886,069
		Total SNPs identified by ResGene18	**11**

Table 4.11: SNPs identified for the trait **Protein** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Chromosome	Identified Position	Identified Position - BLINK	Difference
S7	42,849,437	42,750,874	98,563
S7	4,535,039	4,429,773	105,266
S9	46,006,470	45,885,099	121,371
S7	42,849,437	43,190,165	340,728
S7	12,644,089	13,176,087	531,998
S9	46,006,470	44,884,482	1,121,988
S9	37,962,563	39,358,464	1,395,901
		Total SNPs identified by BLINK	**7**
Chromosome	**Identified Position**	**Identified Position - ResGene18**	**Difference**
S7	4,535,039	4,883,153	348,114
S20	42,395,006	43,193,377	798,371
S18	5,908,800	6,790,349	881,549
S20	42,395,006	43,293,034	898,028
S18	5,908,800	7,278,846	1,370,046
S18	51,521,459	49,747,507	1,773,952
S18	51,521,459	49,301,920	2,219,539
S18	51,521,459	49,292,394	2,229,065
S18	51,521,459	49,289,323	2,232,136
S15	48,666,451	51,634,506	2,968,055
S15	48,666,451	51,647,162	2,980,711
		Total SNPs identified by ResGene18	**11**

Table 4.12: SNPs identified for the trait **Yield** using **BLINK** and **ResGene18**, along with corresponding ground truth positions and positional differences.

Consistent with the ICAR dataset, our model also demonstrated superior SNP detection on the USDA dataset. Specifically, it identified a greater number of SNPs than the BLINK statistical model across all four traits *Height* (3 more SNPs), *Oil* (1), *Protein* (3), and *Yield* (4). This further highlights the robustness and effectiveness of our approach in uncovering biologically meaningful genetic associations.

4.5 Potential SNP positions with specific phenotypic traits

Building upon the above results, we propose a set of novel, potentially significant SNPs in the Table 4.13 and 4.14 that demonstrate strong associations with the four phenotypic traits for each dataset. Given their predictive importance and alignment with established genomic knowledge, we suggest that these candidate SNPs may contribute meaningfully to trait variation and could serve as valuable additions to the existing catalog of trait-associated markers. Their inclusion could enhance the comprehensiveness of current genomic resources and support future efforts in marker-assisted selection and functional genomics in soybean.

Plant Height		Number of Nodes		Grain Yield		Canopy Temperature	
Chr	**Location**	**Chr**	**Location**	**Chr**	**Location**	**Chr**	**Location**
S4	412,338	S11	8,917,811	S1	53,392,386	S14	35,715,633
S8	8,363,522	S8	50,391,345	S13	33,952,562	S2	53,110,518
S20	45,098,114	S1	54,115,102	S3	48,690,949	S18	52,532,914
S10	1,111,345	S16	25,843,791	S8	8,362,116	S9	310,406
S19	43,847,204	S2	52,465,611	S1	54,115,102	S14	31,623,739

Table 4.13: Potential SNPs associated with phenotypic traits identified using our ResGene18 model on the ICAR dataset.

Height		Oil		Protein		Yield	
Chr	**Location**	**Chr**	**Location**	**Chr**	**Location**	**Chr**	**Location**
S14	1,732,902	S12	3,230,818	S9	50,008,825	S12	3,230,818
S13	19,083,710	S18	50,610,764	S18	6,830,722	S13	18,520,168
S8	6,196,069	S14	1,416,586	S9	50,007,447	S1	53,041,644
S13	18,520,168	S8	44,096,773	S5	1,358,556	S10	51,426,325
S13	18,211,337	S18	6,790,349	S12	2,464,843	S1	53,067,596

Table 4.14: Potential SNPs associated with phenotypic traits identified using our ResGene18 model on the USDA dataset.

Chapter 5

Conclusion

In this study, we have introduced **ResGene18**, a novel deep learning model that integrates a genomic-to-image transformation with the ResNet18 architecture. By leveraging this biologically inspired representation and a lightweight convolutional network, ResGene18 has effectively addressed the dual challenges of genomic prediction and SNP identification.

The success of ResGene18 stems from a series of progressive innovations. Starting with support vector regression (SVR) and its variants, we have evaluated their predictive capacity alongside the transformer-based GPFormer. Recognizing the limitations of existing methods in handling high-dimensional, low-sample SNP data, we developed a hybrid model that combines the regression strength of SVR with the attention mechanism's ability to model complex feature interactions. After that, we have reimagined SNP data as two-dimensional, chromosome-wise image-like structure, a novel approach that has facilitated spatial pattern recognition and mitigated the curse of dimensionality.

To the best of our knowledge, we are the first to apply such a transformation for both genomic prediction and GWAS tasks. In ResGene18 architecture, the "18" signifies the layers chosen for their simplicity and efficiency, making the model well-suited for high dimensional genomic datasets.

Through extensive evaluation on two soybean datasets namely the ICAR and USDA, our model has consistently outperformed traditional statistical approaches such as rrBLUP and BLINK. It has demonstrated an average performance improvement of approximately 51% on the first dataset and around 1% on the second dataset. Additionally, in terms of SNP discovery, ResGene18 has identified 57% more significant markers on the ICAR dataset and 34% more on the USDA dataset, further highlighting its effectiveness for trait-marker identification.

Looking ahead, the ResGene18 framework holds promise for broader applications. In future work, this architecture can be validated on multiple publicly available genomic datasets and extended to a wider range of crops. Such evaluations will help further establish the generalizability and robustness of our approach in diverse genomic prediction and GWAS scenarios.

Bibliography

Debasish Basak, Srimanta Pal, Dipak Chandra Patranabis, et al. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224, 2007.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

Jeffrey B Endelman. Ridge regression and other kernels for genomic selection with r package rrblup. *The plant genome*, 4(3), 2011.

Chao Fang, Yanming Ma, Shiwen Wu, Zhi Liu, Zheng Wang, Rui Yang, Guanghui Hu, Zhengkui Zhou, Hong Yu, Min Zhang, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, 18:1–14, 2017.

David Grant, Rex T Nelson, Steven B Cannon, and Randy C Shoemaker. Soybase, the usda-ars soybean genetics and genomics database. *Nucleic acids research*, 38(suppl_1): D843–D846, 2010.

Nastasiya F Grinberg, Oghenejokpeme I Orhobor, and Ross D King. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*, 109(2):251–277, 2020.

J. J. Hill, E. K. Peregrine, G. L. Sprau, C. R. Cremeens, R. L. Nelson, J. H. Orf, and D. A.

Thomas. Evaluation of the usda soybean germplasm collection: Maturity groups 000-iv (pi 578371 - pi 612761). Technical Bulletin 1919, U.S. Department of Agriculture, 2008.

Meng Huang, Xiaolei Liu, Yao Zhou, Ryan M Summers, and Zhiwu Zhang. Blink: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience*, 8(2):giy154, 2019.

Avjinder S Kaler, Jason D Gillman, Timothy Beissinger, and Larry C Purcell. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Frontiers in plant science*, 10:1794, 2020.

Alexander E Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J Bradbury, Michael A Gore, Edward S Buckler, and Zhiwu Zhang. Gapit: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399, 2012.

Yang Liu, Duolin Wang, Fei He, Juexin Wang, Trupti Joshi, and Dong Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in genetics*, 10:1091, 2019.

Xiaolian Lu, Changhua Liu, and Jing Wang. 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)*, volume 13635, pages 227–233. SPIE, 2025.

Wenlong Ma, Zhixu Qiu, Jie Song, Jiajia Li, Qian Cheng, Jingjing Zhai, and Chuang Ma. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248:1307–1318, 2018.

Muhammad Muneeb, Samuel F Feng, and Andreas Henschel. Can we convert genotype sequences into images for cases/controls classification? *Frontiers in Bioinformatics*, 2: 914435, 2022.

Shivani Sehrawat, Keyhan Najafian, and Lingling Jin. Predicting phenotypes from novel genomic markers using deep learning. *Bioinformatics Advances*, 3(1):vbad028, 2023.

Aijun Wang, Yuqi Jiang, Xinyue Shu, Zhongping Zha, Desuo Yin, Yao Liu, Danhua Zhang, Deze Xu, Chengzhi Jiao, Xiaomei Jia, et al. Genome-wide association study-based identification genes influencing agronomic traits in rice (oryza sativa l.). *Genomics*, 113(3): 1396–1406, 2021.

Chunying Wang, Di Zhang, Yuexin Ma, Yonghao Zhao, Ping Liu, and Xiang Li. Wheatgp, a genomic prediction method based on cnn and lstm. *Briefings in Bioinformatics*, 26(2): bbaf191, 2025.

Kelin Wang, Muhammad Ali Abid, Awais Rasheed, Jose Crossa, Sarah Hearne, and Huihui Li. Dnngp, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, 16(1):279–293, 2023.

Cuiling Wu, Yiyi Zhang, Zhiwen Ying, Ling Li, Jun Wang, Hui Yu, Mengchen Zhang, Xianzhong Feng, Xinghua Wei, and Xiaogang Xu. A transformer-based genomic prediction method fused with knowledge-guided module. *Briefings in Bioinformatics*, 25(1):bbad438, 2024.

Juan Yin and Qingna Li. A semismooth newton method for support vector classification and regression. *Computational Optimization and Applications*, 73(2):477–508, 2019.

Mohsen Yoosefzadeh-Najafabadi, Sepideh Torabi, Dan Tulpan, Istvan Rajcan, and Milad Eskandari. Application of svr-mediated gwas for identification of durable genetic regions associated with soybean seed quality traits. *Plants*, 12(14):2659, 2023.

Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed

linear model approach adapted for genome-wide association studies. *Nature genetics*, 42 (4):355–360, 2010.