B.Tech Project Report

Determination of SNP Effect-size of Soybean NAM Population Using Deep Learning

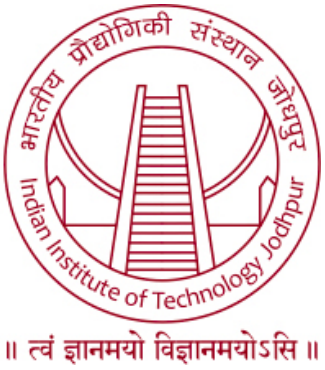
Submitted by

Archit Dwivedi (B18BB005)

in partial fulfilment of the requirements of obtaining the degree of
Bachelors of Technology
in
Biotechnology

Under the guidance of

Dr Sushmita Paul



Department of Bioscience and Bioengineering
Indian Institute of Technology
Jodhpur, India

"The best solution is not where everyone is looking."

1. Introduction

To reveal the genetic risk factors involved in the manifestation of a common trait (or disease) is the primal goal of any genetic study. The knowledge of the prevalent genetic factors associated with the trait under consideration could lead to deeper insights into its pathomechanism and susceptibility. Subsequently, this knowledge could drive the effective drug discovery and early prognosis of the disease. In agronomics applications, in particular, GWAS is increasingly popular in determining the genetic markers associated with important traits such as yield, oil content, protein content, etc. [1]. It has proved to be a valuable tool for sustainable food production. There are various technologies curated in the literature that facilitate genetic risk factor discovery[2] but GWAS (Genome-Wide Association Study) stands out among all of them. GWAS, through its ability to unveil the prevailing genotype-phenotype association in individuals, has revolutionized the way complex-trait genetics studies are done[3]. During the statistical test of association (referred as the '*conventional methods of association'* hereon), GWAS uses the genome-wide SNP (Single Nucleotide Polymorphism) profiles of a huge population of individuals to statistically determine the effect size of each SNP on the trait. The effect size of an SNP facilitates the understanding of the extent to which that SNP might be influencing a trait. The conventional methods of GWAS most often utilise a type of statistical modelling technique called GLM (Generalised Linear Model)[4] to estimate the relationship between the trait (dependent variable) and the genotype (predictor variable). Consider m instances of a continuous trait $\boldsymbol{Y}=(y_1, y_2, \ldots, y_m)^T$, a genotype matrix $\boldsymbol{G}=(g_1, g_2, \ldots, g_m)^T$ and a vector of covariates $\boldsymbol{X}=(x_1, x_2, \ldots, x_p)^T$ as an identifier for properties such as ancestry indicator or the Principal Components[5]. The multivariate GLM is formulated as the following equation.

$$Y = \beta_0 + \beta_X X + G\beta_1 + \epsilon$$

Here $\beta_0 \in \mathbb{R}^m$ represents the intercept vector, $\beta_x \in \mathbb{R}^{m \times p}$ represents the matrix of regression parameter for p covariates, $\beta_1 \in \mathbb{R}^m$ represents regression parameter of m instances of the genotype of a certain SNP. $\epsilon \in \mathbb{R}^m$ here represents random error and is assumed to be contained within a normal distribution $\epsilon \sim N(0, \Sigma)$. After establishing the relationship of association between the SNP genotype and phenotype, the effect size of that SNP can be calculated through the wald test. We subject the relationship under the following hypothesis test-

$$H_0 : \beta_1 = 0$$

$$H_1 : H_0 \text{ is not true}$$

Under the wald test, the hypothesised value of the parameter (β_1) is compared against its estimated value. If **Z**=(1_n, **X**, **G**), then the estimated value of the parameter through maximum likelihood estimation would be given by

$$\boldsymbol{\beta}_1 = \mathbf{Y}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$$

Hence, the value of the Wald statistic is given by

$$\mathbf{W} = \frac{\boldsymbol{\beta}_1^2}{var(\boldsymbol{\beta}_1)}$$

The magnitude of the Wald statistic is proportional to the genetic effect size of the corresponding SNP. Therefore, in this study, $\mathbf{W}$ has been taken as the direct measure of genetic effect size.

This study is performed on the experimental soybean nested association mapping dataset curated under the SoyNAM project[6][7]. In the aforementioned study, more than 5000 RIL (Recombinant Inbred Lines) have been genotyped over 4236 SNPs across all its 20 chromosomes. Acquisition of the dataset was duly followed by a quality control procedure. This includes removal of missing values, removal of duplicates, removal of variants with 100% LD (linkage disequilibrium) and removal of variants below the set MAF (Minimum Allele Frequency) threshold (taken as 5%). The following table shows the summary of the processed data.

Phenotype	**Environment**	**Number of Samples**	**Heritability**
Protein	Illinois (2012)	5128	0.545
Oil	Illinois (2012)	5128	0.617
Moisture	Illinois (2012)	5128	0.582
Height	Illinois (2013)	5138	0.667

2. Hypothesis

Most of the genome-wide association studies are done using statistical methods where it relies on the tendency of a genetic marker to inherit along with the causal gene through LD. So far, statistically driven GWA studies have been shown effective in decoding rare traits and diseases. In rare phenotypes, the usual scenario is that the SNP association to the phenotype follows a linear model. However, there are several issues in GWAS for complex traits that undermine the robustness of the methodology to a great extent[8]. The statistical GWAS consider each SNP independently for establishing the genotype-phenotype association. Whereas, complex traits

are governed by high-order genetic interactions. Therefore, the statistical methods of GWAS are fundamentally unequipped for the association studies of complex traits. There are many methods enlisted in the literature capable of performing statistical multi-factor association tests on complex traits. However, in most polygenic traits the unknown number of interacting SNPs lays a challenge for them. With the advent of computational superior algorithms and hardware, it is today possible to carry of studies, such as complex-trait GWAS, where a large number of parameters are involved. Moreover, statistical GWAS is based on several assumptions about the distribution of genotypes and their contribution to the phenotype. However, assuming such a thing is not always practical on the ground. In this study, a deep learning network has been devised to mitigate the need for prior assumptions about the data and still fit over it to produce reliable SNP effect sizes.

---

3. Proposed Method

The raw genotype dataset contains genotypes encoded as (*AA, AB, BB*) which is computationally uninterpretable. Therefore, in the first step, the homozygous reference allele genotype was designated as 0, the heterozygous alternate allele genotype was designated as 1, and the homozygous alternate allele genotype was designated as 2. Thereafter, the genotypes were re-encoded in the on-hot format. A deep neural network has been devised to learn the phenotype from the associated genotype profiles of the sample. The implemented network consists of a residual block[8], creating a skip connection from the input node to further into the network, as shown in Appendix D. The residual block is followed by dense layers. The following table curates some of the prominent features of the network.

```
HOM_REF    = [100]
HET        = [010]
HOM_ALT    = [001]
```

Parameters	**Types/Values**
Activation function	Conv: Linear | FC: Inverse Square Root Unit
Dropout Rate	0.75
Optimization	Adam
Learning Rate	0.001
Loss Function	Mean Squared Error
Early stop	MAE with patience 5
Regularisation	L2 regularisation
Epochs	1000
Batch Size	250

After the model training, the SNP effect size (referred to as *SNP Propensity*) is determined as described in [9].

---

4. Results

Manhattan plots of SNP effect size by wald test and SNP propensity derived from the model have been plotted for each trait (see Appendix B). Also, the learning curve of the model for each of the traits was plotted (see Appendix C).

Following is a brief contrast between the significant SNPs obtained from both methods (Based on the results from the Soybase repository[10]). Note that the verdict about the association with the phenotype as given below is based on the available QTL literature. They might still be involved in pathways leading to that trait.

Height

~ Based on SNP Propensity: [***Gm04_18306789**, Gm12_2659260, Gm12_2894203 and Gm13_23782754*] All of them are considered an influential genetic marker in determining the height of the plant [11].

~ Based on the statistical method of GWAS: [***Gm04_18306789**, Gm07_40760824* and *Gm10_50834461*] *Gm04_18306789, Gm07_40760824* are found to be directly influencing the height [11].

~ Note that both methods have found *Gm04_18306789* as significant.

Moisture

~ Based on SNP Propensity: [***Gm2_48371970***, *Gm17_6781998* and *Gm18_8451185*]. Only *Gm2_48371970* [12] was found to have a direct influence on the trait.

~ Based on the statistical method of GWAS: [***Gm2_48371970**, Gm14_8175721* and *Gm19_42137460*] *Gm2_48371970* [12], *Gm14_8175721* [13] and *Gm19_42137460* [14] are found to be associated with the trait.

~ Note that both the methods identified *Gm2_48371970* as significant.

Oil

~ Based on SNP Propensity: [***Gm04_8184443**, Gm15_48737423* and *Gm16_756426*] *Gm15_48737423* [15] and *Gm16_756426* [16] are found to be associated with the trait.

~ Based on the statistical method of GWAS: [***Gm04_8184443**, Gm02_7439248* and *Gm18_1685024*] Only *Gm18_1685024* [17] was found to be associated with the trait

Protein

~ Based on SNP Propensity: [*Gm02_5299205,* ***Gm07_7832406*** and *Gm20_29976653*] Only *Gm20_29976653* [18] was found to be associated with the seed protein content.

~ Based on the statistical method of GWAS: [***Gm07_7832406**, Gm18_2102506* and *Gm18_1685024*] None of the SNPs was found to be associated with the trait.

Note: The verdict about the association of the genetic marker with the phenotype as given above is solely based on the available literature[10]. They might still be involved in pathways leading to that trait.

---

6. References

See Appendix A.

Appendix A

[1] Yuki Nakano, Yuriko Kobayashi, Genome-wide Association Studies of Agronomic Traits Consisting of Field- and Molecular-based Phenotypes, Reviews in Agricultural Science, 2020, Volume 8, Pages 28-45, Released March 27, 2020, Online ISSN 2187-090X, https://doi.org/10.7831/ras.8.0_28

[2] Kumar S, Yadav N, Pandey S, Thelma BK. Advances in the discovery of genetic risk factors for complex forms of neurodegenerative disorders: contemporary approaches, success, challenges and prospects. J Genet. 2018 Jul;97(3):625-648. PMID: 30027900.

[3] Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22. doi:10.1016/j.ajhg.2017.06.005

[4] Chu, Benjamin & Keys, Kevin & Sinsheimer, Janet & Lange, Kenneth. (2019). Multivariate GWAS: Generalized Linear Models, Prior Weights, and Double Sparsity. 10.1101/697755.

[5] Zhao, Huaqing et al. “A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS).” Statistical applications in genetics and molecular biology vol. 17,6 /j/sagmb.2018.17.issue-6/sagmb-2017-0054/sagmb-2017-0054.xml. 4 Dec. 2018, doi:10.1515/sagmb-2017-0054

[6] Xavier, A., Beavis, W. D., Specht, J. E., Diers, B., Muir, W. M., and Rainey, K. M. (2015). SoyNAM: Soybean nested association mapping dataset

[7] Song, Q., Yan, L., Quigley, C., Jordan, B. D., Fickus, E., Schroeder, S., et al. (2017). Genetic characterization of the soybean nested association mapping population. Plant Genome. 10 (2). doi: 10.3835/plantgenome2016.10.0109

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[9] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR, abs/1312.6034.

[10] Soybase Repository, https://www.soybase.org/

[11] Sungwoo Lee, T. H. Jun, Andrew P. Michel, M. A. Rouf Mian, SNP markers linked to QTL conditioning plant height, lodging, and maturity in soybean, Euphytica, 10.1007/s10681-014-1252-8, 203, 3, (521-532), (2014).

[12] Cornelious, B., Chen, P., Chen, Y. et al. Identification of QTLs Underlying Water-Logging Tolerance in Soybean. Mol Breeding 16, 103–112 (2005). https://doi.org/10.1007/s11032-005-5911-2

[13] Mian, M., Ashley, D., Boerma, H. An additional QTL for water use efficiency in soybean. Crop Sci. 1998, 38(2):390-393

[14] Specht, J.E., Chase, K., Macrander, M., Graef, G.L., Chung, J., Markwell, J.P., Germann, M., Orf, J.H., Lark, K.G. Soybean Response to Water:A QTL Analysis of Drought Tolerance. Crop Sci. 2001, 41(2):493-509

[15] Lee, S.H., Bailey, M.A., Mian, M.A.R., Carter, T.E. Jr., Shipe, E.R., Ashley, D.A., Parrott, W.A., Hussey, R.S., Boerma, H.R. RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. Appl. Genet. 1996, 93(5-6):649-657

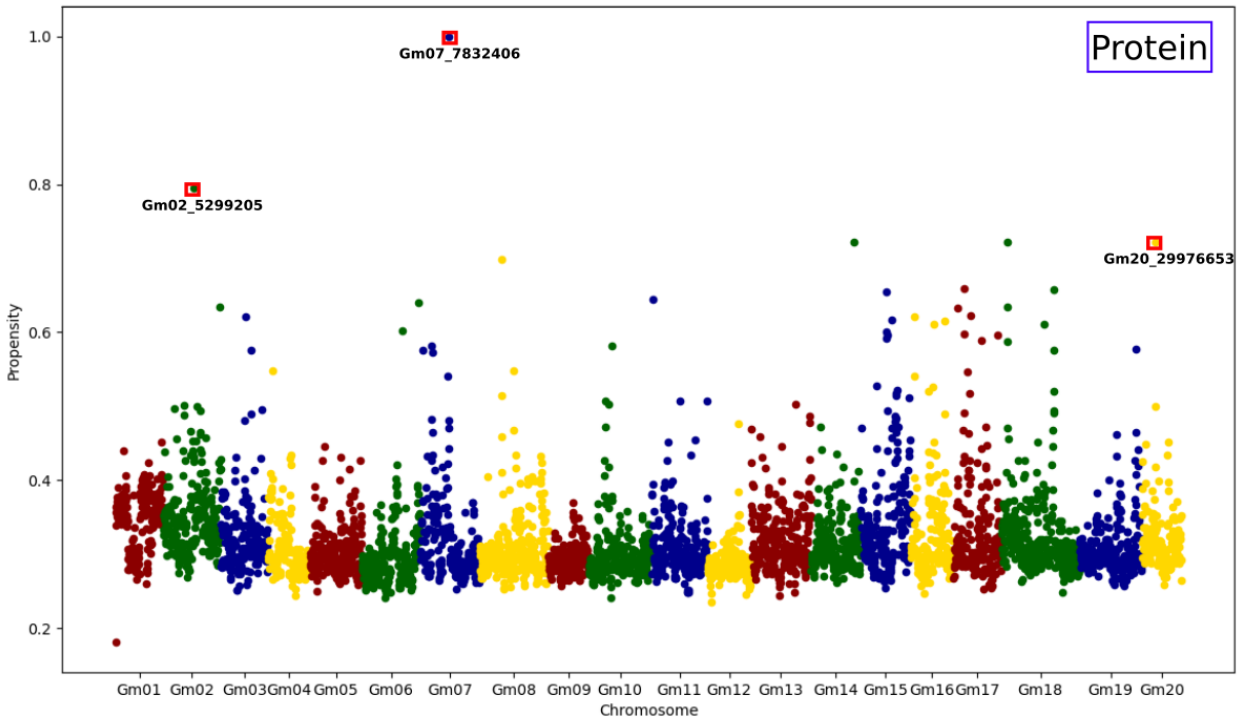
[16] Cardinal AJ, Whetten R, Wang S, Auclair J, Hyten D, Cregan P, Bachlava E, Gillman J, Ramirez M, Dewey R, Upchurch G, Miranda L, Burton JW. Mapping the low palmitate fap1 mutation and validation of its effects in soybean oil and agronomic traits in three soybean populations. Theor Appl Genet. 2014 Jan;127(1):97-111. doi: 10.1007/s00122-013-2204-8. Epub 2013 Oct 17. PMID: 24132738.

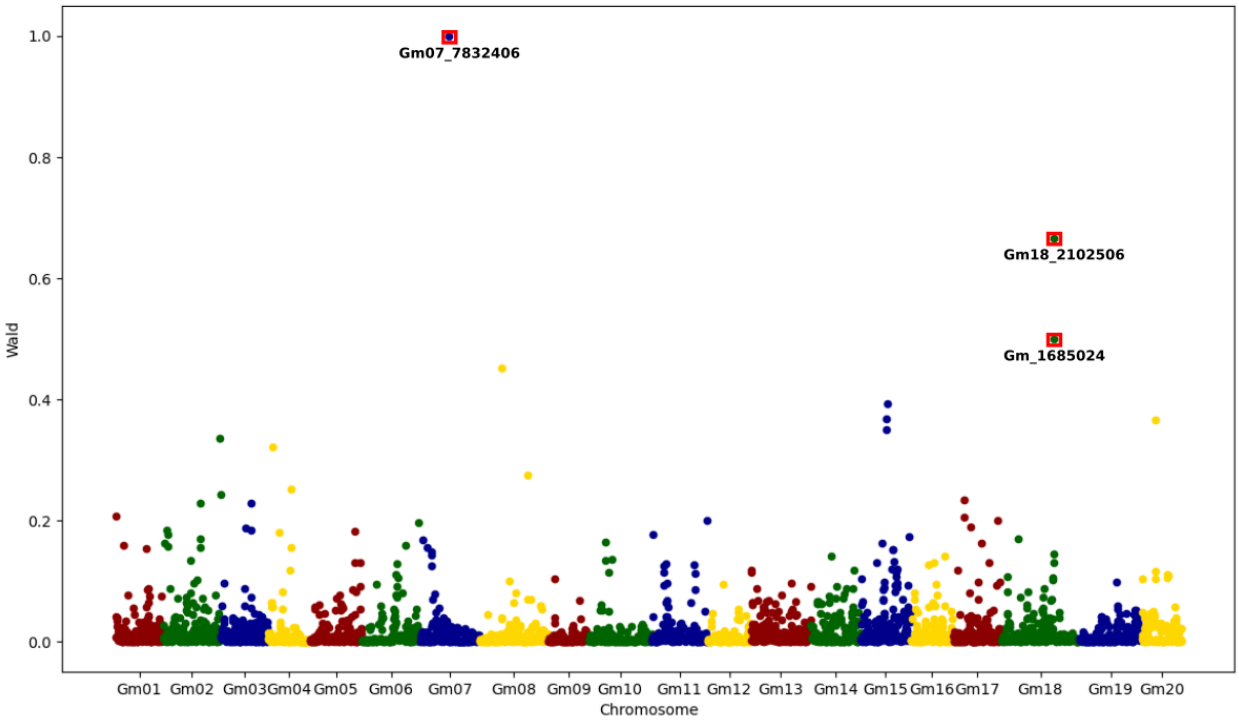
[17] E. C. Brummer, A. D. Nickell, J. R. Wilcox, R. C. Shoemaker, Mapping the Fan Locus Controlling Linolenic Acid Content in Soybean Oil, Journal of Heredity, Volume 86, Issue 3, May 1995, Pages 245–247, https://doi.org/10.1093/oxfordjournals.jhered.a111572

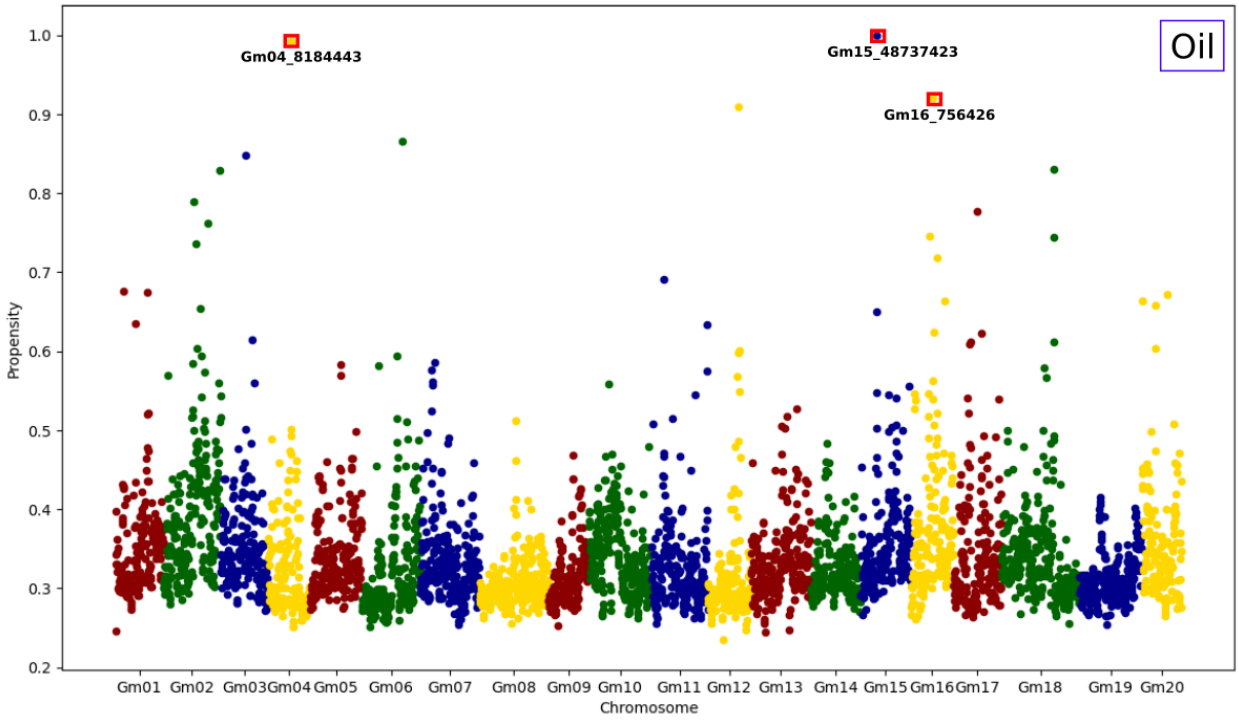
[18] Bolon, YT., Joseph, B., Cannon, S.B. et al. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10, 41 (2010). https://doi.org/10.1186/1471-2229-10-41

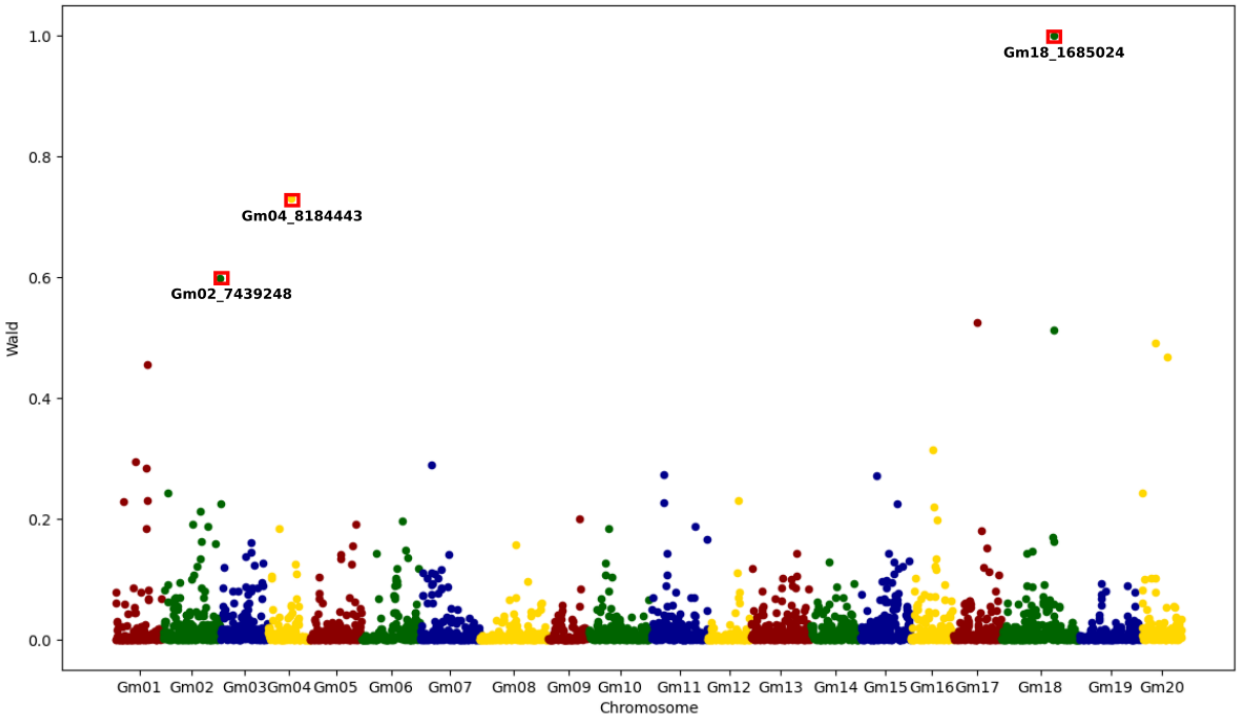
Appendix B

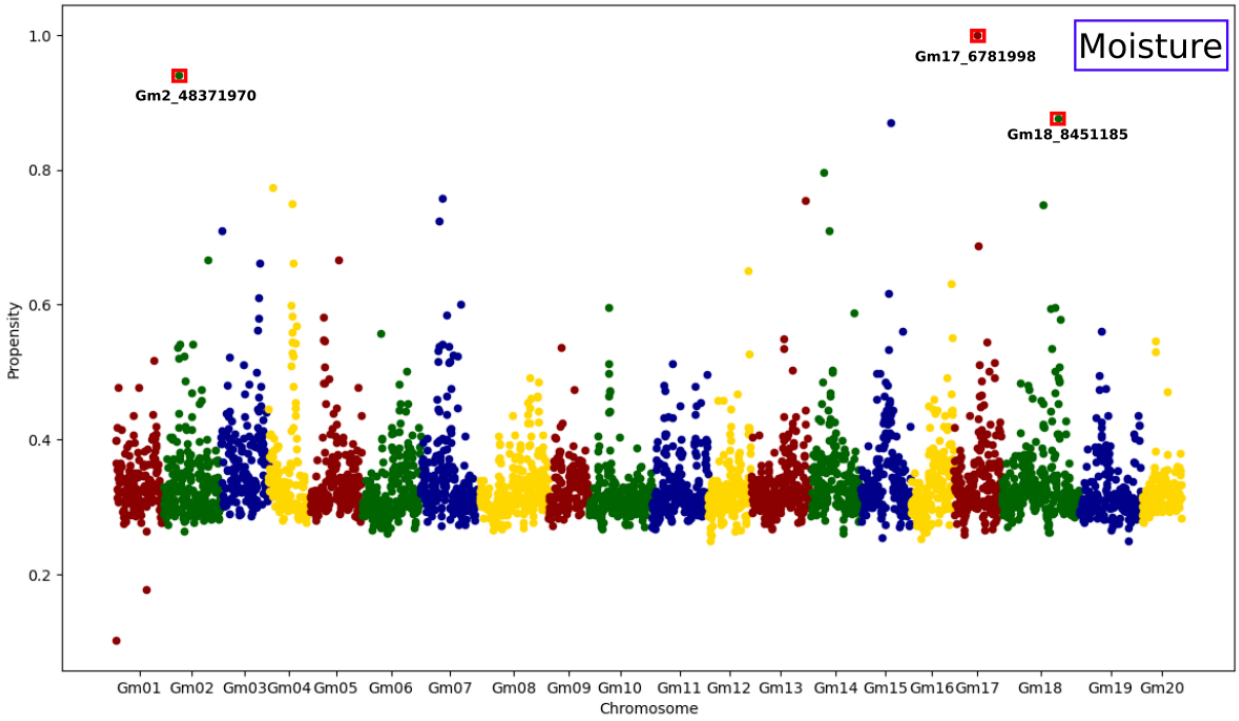
Manhattan plots of SNP effect size by wald test, and SNP propensity derived from the model.

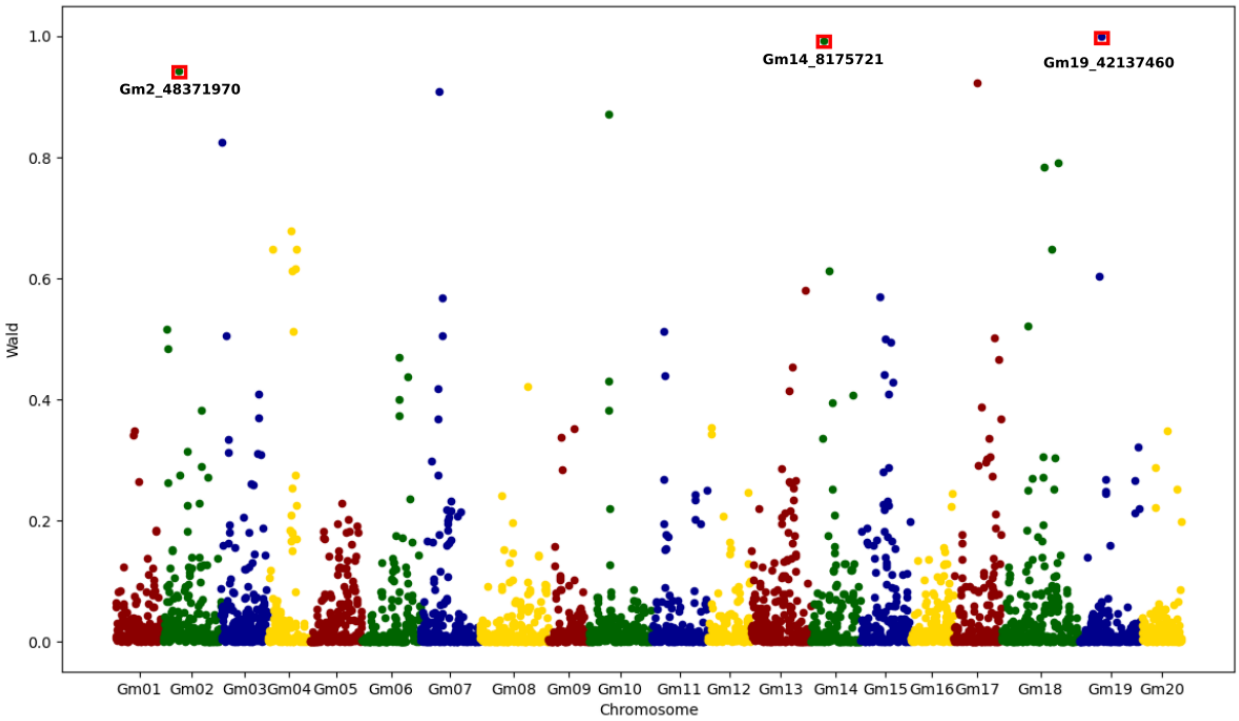


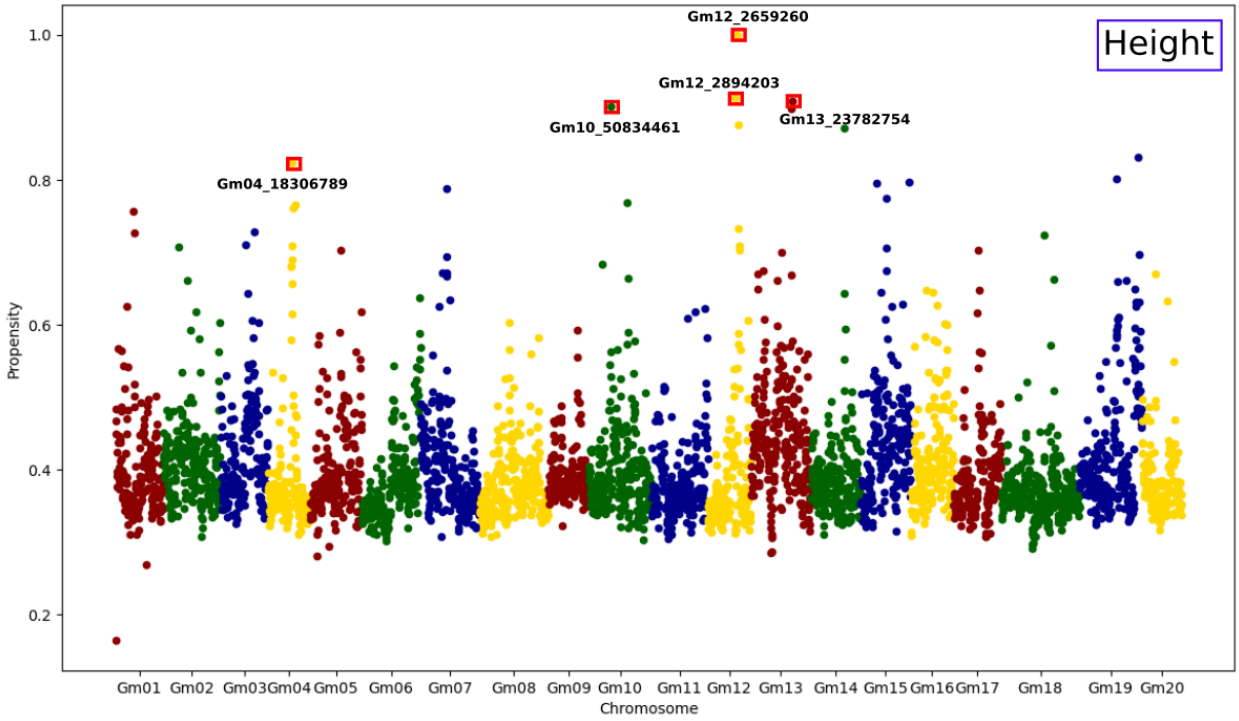


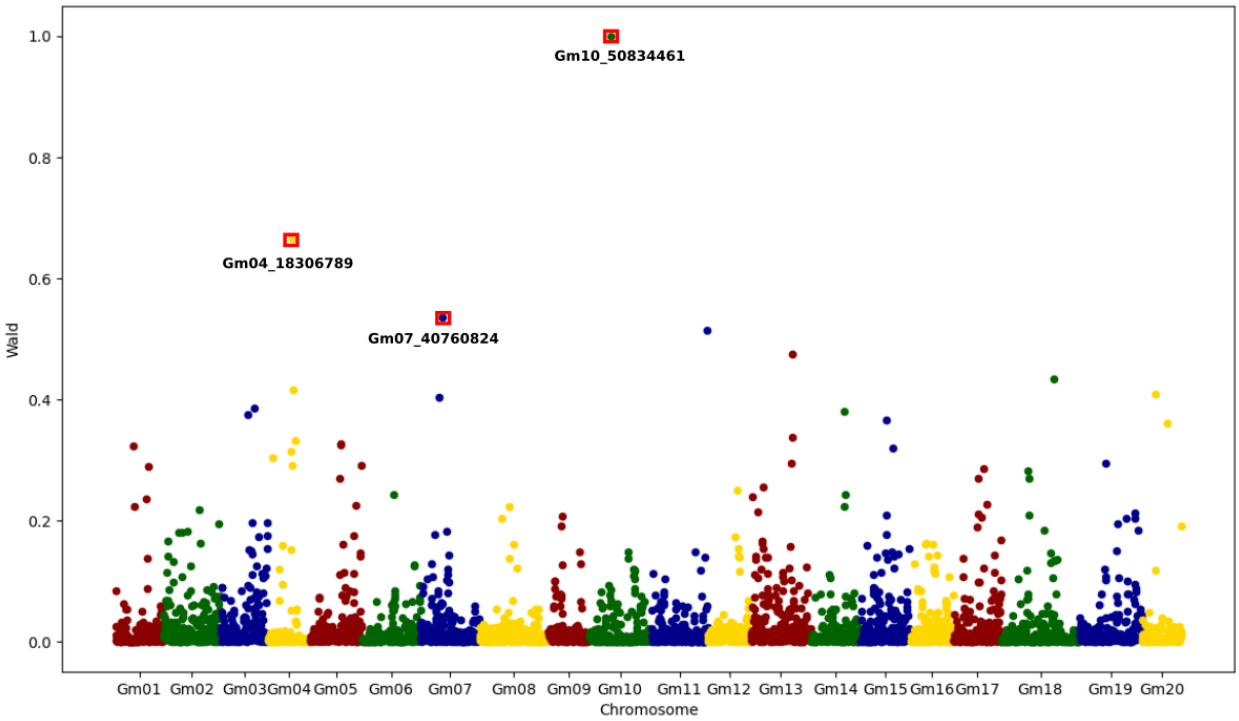

Oil
Gm04_8184443
Gm15_48737423
Gm16_756426
Propensity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome


Gm18_1685024
Gm04_8184443
Gm02_7439248
Wald
1.0
0.8
0.6
0.4
0.2
0.0
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome


Moisture
Gm2_48371970
Gm17_6781998
Gm18_8451185
Propensity
1.0
0.8
0.6
0.4
0.2
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome

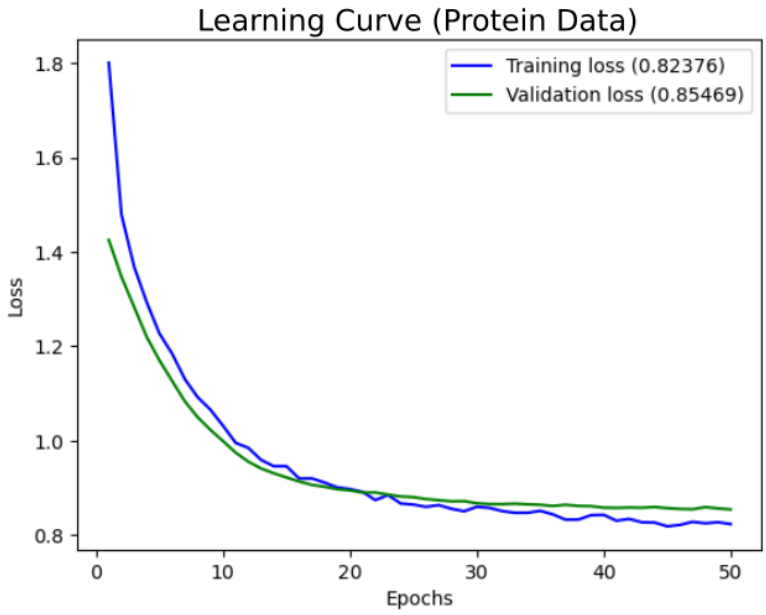

Gm2_48371970
Gm14_8175721
Gm19_42137460
Wald
1.0
0.8
0.6
0.4
0.2
0.0
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome

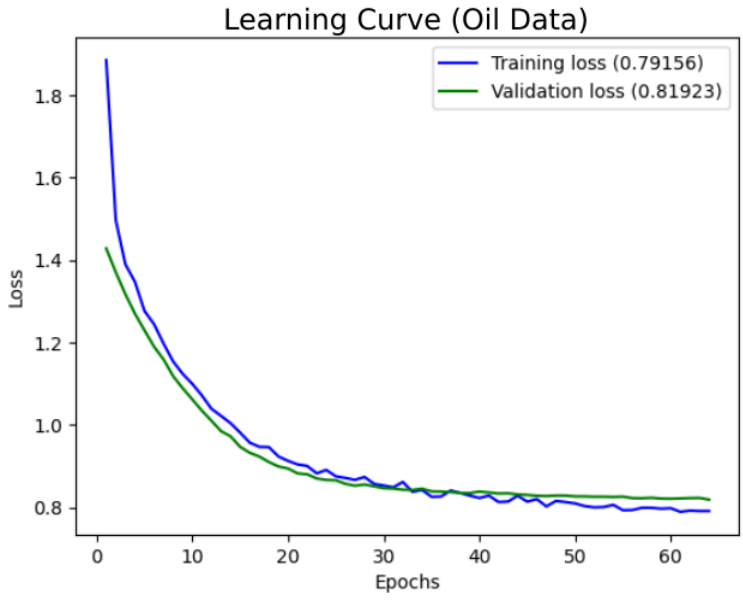

Height
Gm12_2659260
Gm12_2894203
Gm13_23782754
Gm10_50834461
Gm04_18306789
Propensity
1.0
0.8
0.6
0.4
0.2
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome

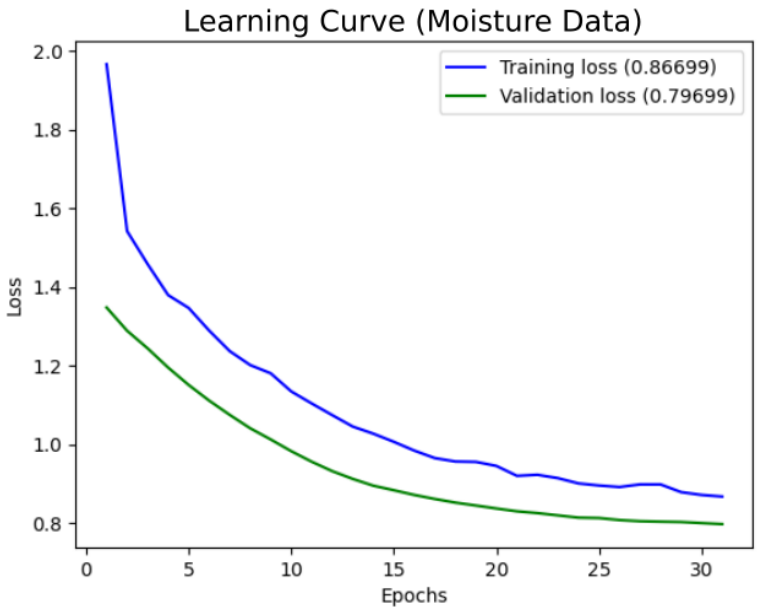

Gm10_50834461
Gm04_18306789
Gm07_40760824
Wald
1.0
0.8
0.6
0.4
0.2
0.0
Gm01 Gm02 Gm03 Gm04 Gm05 Gm06 Gm07 Gm08 Gm09 Gm10 Gm11 Gm12 Gm13 Gm14 Gm15 Gm16 Gm17 Gm18 Gm19 Gm20
Chromosome

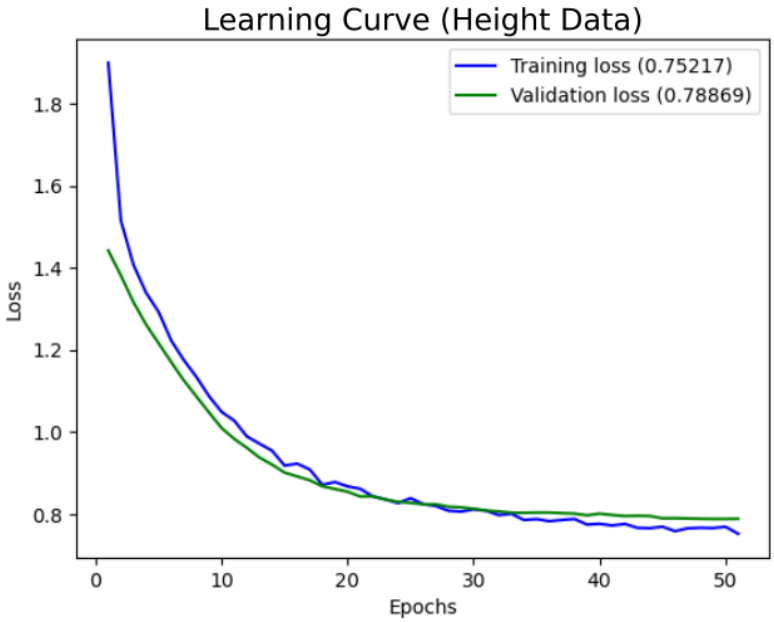
Appendix C

Learning curves for each trait.



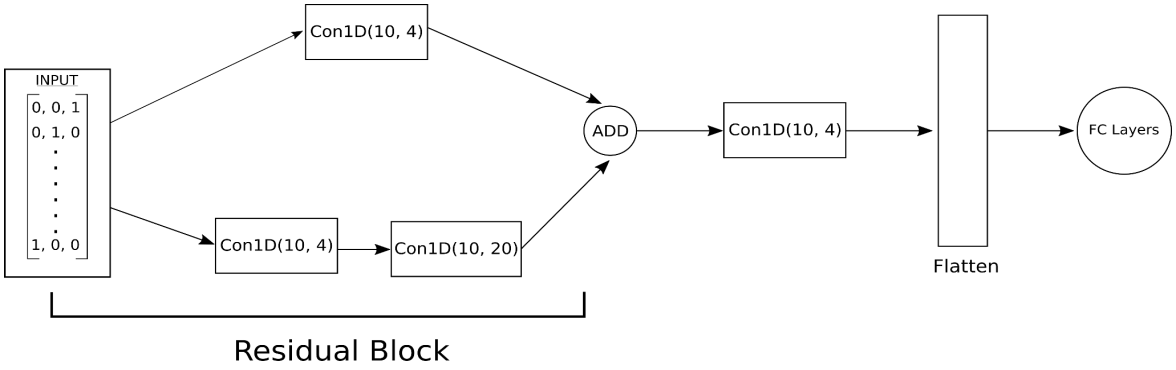



Learning Curve (Moisture Data)
Training loss (0.86699)
Validation loss (0.79699)
Loss
Epochs
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
5
10
15
20
25
30


Learning Curve (Height Data)
Training loss (0.75217)
Validation loss (0.78869)
Loss
Epochs
1.8
1.6
1.4
1.2
1.0
0.8
0
10
20
30
40
50

Appendix D

Proposed network:



Hardware used for training:

Model trained on *Tesla V100-PCIE-32GB GPU*

```
Last login: Tue Nov 30 17:23:23 2021 from gateway
(base) ┌─[xf@master] - [~] - [493]
└─[$] nvidia-smi
Wed Dec  1 10:25:49 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.57.02    Driver Version: 470.57.02    CUDA Version: 11.4     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-PCIE...  Off  | 00000000:01:00.0 Off |                    0 |
| N/A   43C    P0    36W / 250W |      0MiB / 32510MiB |      5%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
```

Model Summary:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #     Connected to
=================================================================
input_3 (InputLayer)         [(None, 4236, 4)]         0
_________________________________________________________________
conv1d_8 (Conv1D)            (None, 4236, 10)          170         input_3[0][0]
_________________________________________________________________
conv1d_9 (Conv1D)            (None, 4236, 10)          2010        conv1d_8[0][0]
_________________________________________________________________
conv1d_10 (Conv1D)           (None, 4236, 10)          170         input_3[0][0]
_________________________________________________________________
dropout_6 (Dropout)          (None, 4236, 10)          0           conv1d_9[0][0]
_________________________________________________________________
add_2 (Add)                  (None, 4236, 10)          0           conv1d_10[0][0]
                                                                   dropout_6[0][0]
_________________________________________________________________
conv1d_11 (Conv1D)           (None, 4236, 10)          410         add_2[0][0]
_________________________________________________________________
dropout_7 (Dropout)          (None, 4236, 10)          0           conv1d_11[0][0]
_________________________________________________________________
flatten_2 (Flatten)          (None, 42360)             0           dropout_7[0][0]
_________________________________________________________________
dropout_8 (Dropout)          (None, 42360)             0           flatten_2[0][0]
_________________________________________________________________
out (Dense)                  (None, 1)                 42361       dropout_8[0][0]
=================================================================
Total params: 45,121
Trainable params: 45,121
Non-trainable params: 0
```