# Research Proposal

## Feature Engineering for Enhanced Genomic Prediction.

Patrick Gibbs 1083438
Word Count: 3298

## Summary

Genomic prediction (GP) is the use of genome-wide data to predict complex traits. GP has wide use in agricultural breeding programs and is of growing interest to predict disease disposition in humans. The primary limitation facing GP is that for any real-world dataset the number of markers, usually single nucleotide polymorphisms (SNPs), is far greater than the number of sequenced organisms. This high dimensionality means model parameters cannot be estimated precisely, limiting predictive power. Furthermore, GP is only tractable for populations with vast sequence information such as model organisms and agricultural crops. This proposal aims to address these key limitations of GP: high dimensionality and tractability limited to large datasets. First, a novel dimensionality reduction technique for increased GP performance coined *kinship partitioning* is proposed. Kinship partitioning will be benchmarked for prediction of human height and *Arabidopsis thaliana (A. thaliana)* flowering time using popular GP models. Second, it is proposed that kinship partitioning will increase the transferability of GP models between populations. To test this hypothesis I will use kinship partitioning to develop GP models within *A. thaliana* and human populations, then test transferability of these models to novel populations.

## Background

Genomic prediction is the use of dense genome-wide markers to predict traits, and strikes at the fundamental goal of genetics – linking genotype to phenotype. As it uses dense genome-wide markers, GP is particularly suited to predict complex traits which are controlled by numerous small effect elements. Therefore, genomic prediction is widely used to estimate breeding values in agriculture, allowing farmers to propagate crops/livestock with the highest putative yield (Heffner *et al*., 2009; Lorenz *et al*., 2011; Desta & Oritz, 2014). There is additionally growing research using genomic prediction for predisposition to complex pathologies such as heart disease (Abraham *et al.,* 2016; Lello *et al*., 2019). The number of markers (typically SNPs) required to map effects throughout the genome is large, meaning GP has high feature dimensionality. Therefore, the major challenge facing genomic prediction is estimating the effects of every marker (*p*) given relatively few phenotyped individuals (*n*). In current datasets, the number of markers is typically much greater than the number of sequenced individuals, making it challenging to fit predictive models. Moreover, many markers throughout the genome have negligible effects on the phenotype, meaning datasets for genomic prediction are noisy.

**Current methods for genomic prediction**

Given the vast number of SNPs available from large-scale sequencing experiments, GP is canonically approached through penalised regression (Meuwissen *et al.,* 2001). Penalised regression is favoured as it effectively shrinks the weight of uninformative features, giving the best performance in p >> n scenarios. Such linear statistical models assume each genetic element acts additively on the phenotype. This relaxation allows for minimal parameterization requiring few training examples to fit, however it prevents integration of complex genetic interactions such as dominance and epistasis. Recent studies have also explored more sophisticated machine learning (ML) models which can parameterize interaction between features, particularly artificial neural networks (ANN) and ensemble learning algorithms. Such ML methods have seen enormous success in other fields such as speech recognition (Hinton *et al.,* 2012), computer vision (Ciregan *et al.,* 2012), and putative protein folding (Jumper *et al.,* 2021). Given that molecular studies have highlighted interactive genetic architectures underpinning traits, models which have the capacity to parameterise this interaction should dominate GP performance.

However, it has been reported that ANNs perform consistently worse than penalised regression for GP problems (Abdollahi-Arpanahi *et al.,* 2020; Azodi *et al*., 2019; Bellot *et al., 2018*). The disparity in performance is attributed to the fact that ANNs have far more parameters than linear models. Therefore, ANNs require either additional training data or reduced feature dimensionality to effectively fit each of these parameters. Moreover, ANNs are sensitive to noise – signal from uninformative markers is

detrimental. Ensemble learning algorithms used in prior studies such as *Gradient Boosting* or *Random Forest* have shown promising performance comparable to linear statistical models (Abdollahi-Arpanahi *et al.,* 2020; Azodi *et al*., 2019). However, similar to ANN, ensemble algorithms are also understood to suffer from overfitting to uninformative features (Bishop & Nasrabadi, 2006). It is therefore widely suggested that GP can only be progressed through feature engineering for dimensionality reduction (Ramstein *et al.,* 2019); more informative features will improve performance, especially as model complexity increases.

**Current feature engineering strategies**

Two means of GP dimensionality reduction currently exist: feature selection (filtering out uninformative SNPs), or feature engineering (combining groups of SNPs into new features). A widely applied filtering approach is to incorporate annotation data where researchers have manually labelled the function of each genetic element. This method allows the dataset to be filtered for only functionally relevant elements, removing synonymous SNPs and intergenic regions, and has seen performance benefits in GP (Gao *et al.,* 2017; Teng *et al.,* 2022; Do *et al*., 2015). A popular feature engineering strategy is to utilise kinship (overall similarity between genotypes) between organisms as a predictor – individuals sharing genetic features also typically share similar phenotypes (Durel *et al*., 1998; Hayes & Goddard 2008). The predictive capacity of kinship for the phenotype of organisms is well established and has long been utilised in mixed-linear models (Henderson, 1963; Henderson, 1976; Henderson, & Quaas, 1976). However, using overall kinship between organisms as a predictive feature loses all positional information about the genetic variation between them. Recent GP studies also use principal component decomposition to combine co-linear features. Here only the top principle components which describe a majority of SNP variance are used for predictions.

**Current lack of model transferability**

Given the enormous amount of genomic data required to fit penalised regression models, genomic prediction is only accessible on large datasets. Canonically, models are usually fit to individual populations and lack transferability due to unseen allele effects when predicting in novel populations (Guo *et al.*, 2014; De Roos *et al., 2009*). Therefore, genomic prediction remains a tool exclusive to highly resourced projects, such as industrial agriculture. The development of transferable models between populations, or even closely related species would democratise the applicability of genomic prediction.

## Aims

In this project I will develop a novel feature engineering approach, coined *kinship partitioning,* in order to reduce feature dimensionality and increase model transferability between populations. In order to contextualise the specific aims of this project, a high-level description of this method is presented here. In this approach the genome will be partitioned into segments and features will be constructed as the kinship between individuals for each segment across the genome. For $k$ partitions and $n$ individuals, kinship partitioning will generate $n \cdot k$ features, allowing dimensionality reduction if $k < \frac{p}{n}$. This strategy loses some specific positional information about SNPs in return for shrinkage of the feature set. Features constructed through kinship partitioning are unlikely to be independent due to linkage. Therefore, correlated features can be combined by applying principal component decomposition yielding further dimensionality reduction consistent with prior approaches (Azodi *et al.,* 2020). Additionally, annotation data will be used to remove putatively inconsequential variation such as silent mutations. Here partitions will be done for functionally labelled regions of the genome such as genes, meaning features have a discrete functional basis. Utilising this novel approach, this project aims to answer two main questions:

1. **Is kinship inferred from subsets of the genome a more informative feature than individual SNPs in terms of GP performance?**
2. **Can genomic prediction models be optimised utilising kinship to make accurate predictions across populations?**

These specific aims will be investigated using human height and arabidopsis flowering time. These traits along with their pertaining genomic sequence data will be sourced from the UK Biobank and 1001 Genomes database respectively (Bycroft *et al*., 2018 ; Seren *et al.,* 2016). Both traits are ideal for GP being polygenic with a variety of effect sizes (Wood *et al.,* 2014) and have been the subject of prior GP studies (Bellot *et al.,* 2018*;* Lello *et al* 2018). I choose to use a human trait as the UK Biobank is the largest public database of sequenced organisms, giving statistical power to this study. *A. Thaliana* data from 1001 Genomes has the advantage of individuals being grown under controlled laboratory conditions implying high heritability of traits. Moreover, this *A. thaliana* dataset contains genetically distinct populations which have been labelled in prior literature (Seren *et al.,* 2016), making it ideal to investigate aim 2.

### Aim 1: Kinship partitioning vs full SNP data

The dimensionality reduction facilitated by kinship partitioning is expected to allow GP models to be fit more precisely increasing performance, particularly for complex machine learning models. Therefore, the performance of GP utilising kinship partitioning will be tested against the current practice of using individual markers as features for the prediction of human height and *A. thaliana* flowering time. This comparison will be made using the following popular GP models: *linear fixed effect model, gradient boosting,* and *dense neural networks*. Linear effect models represent the canonical non-interactive means of GP. Dense neural networks and gradient boosting can parameterize interactive complexity and have been highlighted in the literature as promising avenues for improved genomic prediction (Abdollahi-Arpanahi *et al.,* 2020; Azodi *et al.,* 2019; Bellot *et al.,* 2018). I will additionally implement the most prevalent GP model in prior literature, *genomic best linear unbiased predictor* (GBLUP). As GBLUP typically uses overall kinship as a random effect, it is not amenable to implement kinship partitioning (Habier *et al., 2007*). Therefore, GBLUP will be used as an additional performance baseline to evaluate kinship partitioning.

If kinship partitioning improves performance above both null cases, models taking raw SNP features and GBLUP, then the feature engineering strategy is validated. If kinship partitioning realises substantial benefit in the ML models it additionally indicates that given informative features, the investigated phenotypes are best understood through an interactive genetic model.

### Aim 2: GP model transferability

As kinship partitioning captures a relaxed representation of genomic variation, I propose that it will allow the construction of more transferable models between populations compared to using individual SNPs. Kinship partitioning will not be confounded by SNPs not present in the training set. Use of individual SNPs may also develop models focused on specific linkage disequilibrium rather than functional underpinning, making them highly specific to the training population. However, when annotation data is employed, kinship partitioning features represent functional units which I predict will generate more general models.

A common ML technique to deal with limited training data is to first develop a model on a related task where training data is abundant, then adapt this model to the desired task with the limited training data (Devlin *et al.,* 2019). Analogously, I will test whether GP models underpinned by kinship partitioning can

be retrained on novel populations. Specifically, I will investigate this by clustering sub populations within the human and *A. thaliana* datasets. From these clusters, pairs of sub datasets will then be generated, one pertaining to an individual population (denoted the *novel population)*, the second consisting of all other populations (denoted the *larger dataset*). I will train each model specified in aim 1 on the larger dataset, generating a *prior model* which can be used to predict the phenotypes of the *novel population*. I will test if updating the prior model weights with a subset of the *novel population* data increases performance when predicting the rest of the phenotypes in the *novel population*. These cross population predictions will be repeated using differing amounts of data from the novel population to adapt the *prior model.* This will allow me to quantify model performance on novel populations in terms of the amount of the data used to update the model.

I will test if this method provides better results than training on an individual population from scratch. I predict that prior learning will enhance the performance of GP when it is applied to a smaller dataset. If this is the case it will demonstrate that kinship partitioning provides increased transferability of GP models, and allow GP to be expanded into smaller datasets in future studies.

# Methods

## Datasets

Human height data of 488,377 individuals each with 805,426 SNPs will be sourced from the UK Biobank (Bycroft *et al* 2018). Consistent with prior GP studies on this dataset, SNPs with minor allele frequencies of less than 0.1% will be filtered out, as well as poor quality genomes with missing call rates >10%, resulting in 645,589 SNPs and 488,371 individuals (Lello *et al., 2018)*. *Arabidopsis thaliana* flowering time data will be sourced from the Arapheno database (Seren *et al., 2016)*, and the pertaining sequence data of 1135 individuals with up to 10,707,430 SNPs from the 1001 genome project (Alonso-Blanco *et al.,* 2016). A pipeline will be developed to integrate genome annotation files into the sequence data, allowing functional categorisation of each SNP in the dataset. This will allow the generation of two datasets: genome wide SNPs and SNPs only pertaining to genes. Consistent with prior studies, partitions on genes will include 5kb upstream/downstream regions to account for regulatory elements (Gao *et al*., 2017; Tang *et al*., 2022).

## Feature engineering

**Kinship Partitioning**

Kinship partitioning will be conducted on chosen k partitions across the genome. For each individual, the feature set is constructed as the kinship to every other individual for each partition, resulting in $n \cdot k$ features. The following partitioning strategies will be tested: equal sized partitions with arbitrary boundaries, and partitions based on genes. For equal sized partitions, multiple values of *k* will be tested. These values will span from k = 1 where a partition makes up the full genome, to the highest computationally viable value of *k,* noting that the size of the feature set grows to $n^2$ as $k \rightarrow n$.

**Principal Component Decomposition**

Using both kinship partitioning and individual SNPs, principal component decomposition will be applied for further dimensionality reduction. In line with prior literature (Azodi *et al.,* 2020), the optimal number of principal components will be estimated iteratively, whereby the number of principal components utilised for prediction will be increased until the model accuracy plateaus.

## Proposed Models

**GBLUP baseline**

GBLUP (Habier *et al., 2007;* VanRaden, 2008) is one of the most widely utilised genomic prediction algorithms. Predictions (**y**) are made based on the following mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

where **y** is a vector of the predicted phenotypes, **X** is a design matrix, and β is a vector of fixed effects. Commonly $\mathbf{X}\boldsymbol{\beta}$ is set to the population mean $\boldsymbol{\mu}$ (Habier *et al., 2007;* VanRaden, 2008). **Z** is a design matrix, which in my implementation will specify kinship. $\boldsymbol{u}$ is a vector of random effects where $\boldsymbol{u} \sim N(O, \mathbf{G})$, where *G* is the genomic covariance matrix. $\boldsymbol{\epsilon}$ pertains to residual effects, $\boldsymbol{\epsilon} \sim N(O, \boldsymbol{I}\sigma^2_R)$ where $\sigma^2_R$ is the residual variance.

**Linear Fixed Effect Model**

In the linear fixed effect phenotypes (***y***) will be predicted with the following equation:

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

In my implementation, **X** is a design matrix specifying the principal components of either genomic markers or partitioned kinship for each individual. $\boldsymbol{\beta}$ is a vector of random effects and will be estimated using GLM elastic net penalised regression (Friedman *et al.,* 2010). $\boldsymbol{\epsilon}$ pertains to residual effects $\boldsymbol{\epsilon} \sim N(O, \boldsymbol{I}\sigma^2_R)$.

**Gradient Boosting**

Gradient boosting is an ensemble method which combines the results of many simple models known as *weak learners* to generate prediction, and is an established means of GP (González-Recio *et al., 2013)*. *Weak learners* are implemented sequentially utilising the whole dataset, where each weak learner is fitted to the residuals of its predecessor. Rather than updating weights continually, the model is fitted by adding weak learners until the residuals across the whole dataset are minimised. Specifically XGBoost (Friedman, 2001), a popular gradient boosting algorithm which utilises shallow decision trees as weak learners, will be utilised given high performance in previous GP studies (Abdollahi-Arpanahi *et al* 2020). XGBoost requires optimization of the number of decision trees (*weak learners),* the depth of each tree, and the learning rate which scales the amount that each decision tree reduces residuals.

**Dense Artificial Neural Networks**

Dense Neural Networks consist of a collection of perceptrons, defined as a unit which takes a number of weighted inputs and computes one output based on a predefined *activation function.* Perceptrons are organised into layers: an input layer, a number of hidden layers and an output layer. In the input layer each perceptron takes one feature value as input and each subsequent hidden layer takes the output of every prior layer. Finally, the output layer consists of one perceptron which maps all the outputs of the preceding layer to the prediction provided by the model. Weights which scale the input/output of each perceptron are fitted through *backpropagation* and *gradient descent* (Rumelhart & Hilton, 1986) which iteratively cycles through the training dataset adjusting weights throughout the network to minimise a loss function. The number of hidden layers will be optimised; the activation will be optimised among:

- The identity function: $f(x) = x$
- the logistic sigmoid function: $f(x) = \frac{1}{1 + exp(-x)}$
- the hyperbolic tan function: $f(x) = tanh(x)$
- the rectified linear unit function: $f(x) = max(0, x)$

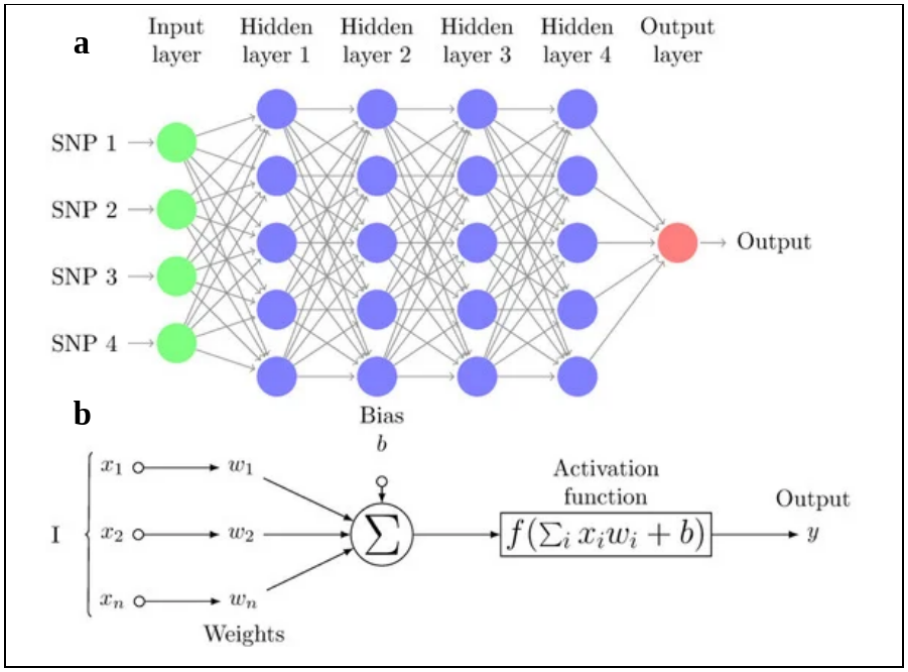


**Figure 1: Diagram of *Dense Neural Network*.** (a) shows *Dense Neural Network Structure* constructed from connected perceptrons (represented as circles). (b) Diagram of an individual perceptron which takes the output of each perceptron in the prior layer scaled by weights (model parameters). The perceptron then computes an output based on an activation function, this output is transmitted to all perceptrons in the following layer.
*Note.* Imaged sourced from http://1001genomes.github.io/admixture-map/ based on the publication Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.

M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491.

**Model implementation**

GBLUP will be implemented using the rrBLUP package in R (Endelman, 2011) All other models will be implemented with python3 using the package sklearn (Pedregosa *et al., 2011)*. Specifically, the *ElasticNet* module will be used for the fixed effect model*, GradientBoostingRegressor* for gradient boosting, and *MLPRegressor*for Dense Neural Networks.

## Building Transferable Models

To measure transferability of GP models across populations, populations within *A. thaliana* and Human datasets first need to be identified. For *A. thaliana,* I will leverage prior studies which have already identified populations in the 1001 genomes dataset (Figure 2) (Alonso-Blanco *et al.,* 2016). Prior studies have not published amenable population clustering for the UK Biobank data. Thus, I will cluster the human data manually using the Bayesian model based algorithm *fastStructure (*Pritchard *et al., 2000;* Seren *et al.,* 2014). The number of clusters (*k*) to decompose the dataset into will be chosen utilising the rate of change in average the log probability of the computed clusters between successive k values as demonstrated by Evanno and colleagues (2005).

Transferability will be measured as model performance in cross-validation across clusters, with some refitting to novel populations. I will testTransferability for kinship partitioning and using individual SNPs as predictors. In the case of kinship partitioning, additional training instances will be generated as the kinship to individuals in the original training set for each partition. In the cases of individual SNPs, additional instances will be constructed only from SNPs present in the prior model.

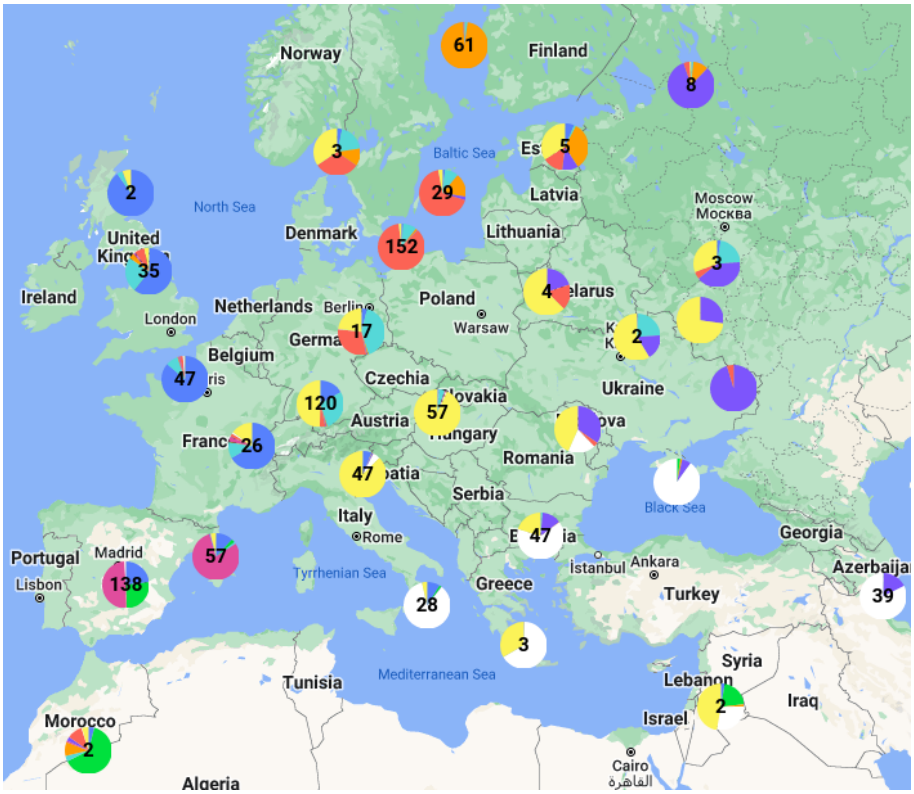


**Figure 2: Admixture map of *Arabidopsis thaliana* individuals in the 1001 Genomes database.** Each pie chart displayed represents the number of individuals pertaining to a geographical region. Colours in each pie chart represent the population each individual belongs to after genomic clustering was conducted.
*Note.* Imaged sourced from http://1001genomes.github.io/admixture-map/ based on the publication Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491.

## Benchmarking Process

The data set will be broken into two parts: a development part for optimising model meta parameters and a test/train part for validation. The development set will be used to deduce the number of partitions for kinship partitioning, the number principle components used for model prediction as well as all metaparameters for each GP model. Parameters may be chosen simultaneously utilising the *Gibbs* sampling algorithm (Rouchka, 1997).

Once ideal metaparameters are identified, models will be independently trained and tested utilising K-Fold cross validation on the remaining data for the highest computationally viable value of K (Figure 3). Performance of a model on a dataset is then measured as the correlation between the expected and

observed phenotypes using the Pearson correlation coefficient (*r)*, and the root mean squared error (RMSE) defined below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}, \quad r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Where $x_i$ is each prediction, $y_i$ is each observed value and *N* is the total predictions. $r \in [-1, 1]$ quantifies the degree of correlation between predicted and observed values where $r = 0$ implies no correlation and $r = \pm 1$ is a perfect positive/negative correlation respectively. *RMSE* measures standardised error between expected and observed. Thus strong models maximise *r* and minimise *RMSE*.

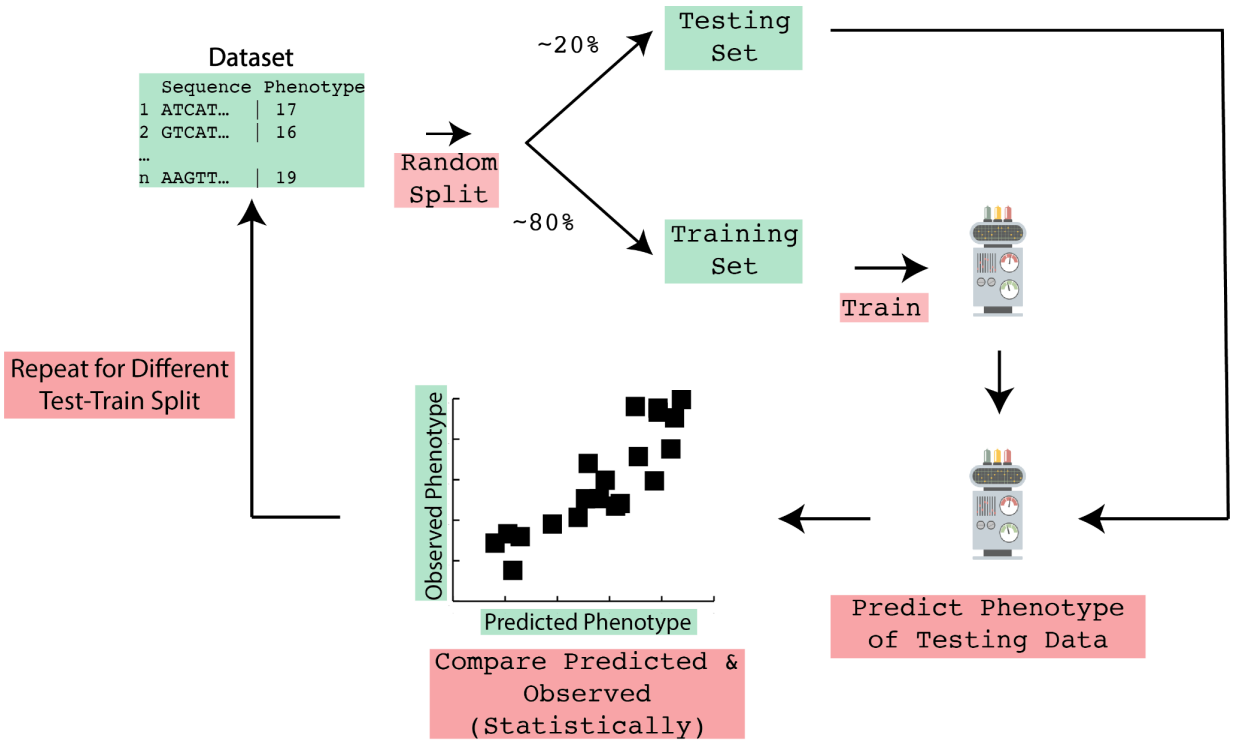


**Figure 3: Cross validation process that will be used to test GP models and feature engineering techniques.** Dataset split into a testing group and a training group. The training group is used to develop a GP model, then predict the phenotypes of the testing group. The correlation between the testing group's predicted and observed phenotypes is measured. This process is repeated multiple test train splits of the data.

# Timeline

| Task | feature engineering techniques | | Semester 2 2022 | | | | Summer 2022-2023 | | Semester 1 2023 | | | | Winter 2023 | | Semester 2 2023 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | First Half | Secound Half | Q1 | Q2 | Q3 | Q4 | First Half | Secound Half | Q1 | Q2 | Q3 | Q4 | First Half | Secound Half | Q1 | Q2 | Q3 | Q4 |
| Parse data from online databases | | | | | | | | | | | | | | | | | | |
| Develope Model implementation | | | | | | | | | | | | | | | | | | |
| Develop testing inferstructure | | | | | | | | | | | | | | | | | | |
| Implementation of feature engineering | | | | | | | | | | | | | | | | | | |
| Cluster Populations | | | | | | | | | | | | | | | | | | |
| Test performance for feature engineering techniques | | | | | | | | | | | | | | | | | | |
| Test performance for across populations | | | | | | | | | | | | | | | | | | |
| Finalise testing and benching marking models | | | | | | | | | | | | | | | | | | |
| Thesis Write up | | | | | | | | | | | | | | | | | | |
| Prepare Presentation | | | | | | | | | | | | | | | | | | |
| Thesis Finalisation | | | | | | | | | | | | | | | | | | |

Link to chart: https://docs.google.com/spreadsheets/d/1LVTMI5hW3BlCFpgvIG-W56VZ1y6sVtrjX9mf3UeFQ2k/edit?usp=sharing

## References

Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(1), 1-15.

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., De Livera, A. M., Yetukuri, L., ... & Inouye, M. (2016). Genomic prediction of coronary heart disease. *European heart journal*, *37*(43), 3267-3278.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491.

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., & Shiu, S.-H. (2019). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3: Genes|Genomes|Genetics*, *9*(11), 3691–3702.

Azodi, C. B., Pardo, J., VanBuren, R., de Los Campos, G., & Shiu, S. H. (2020). Transcriptome-based prediction of complex traits in maize. *The plant cell*, *32*(1), 139-151.

Bellot, P., de Los Campos, G., & Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits?. *Genetics*, *210*(3), 809-819.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203-209.
Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in statistics-theory and methods*, *3*(1), 1-27.

Ciregan, D., Meier, U., & Schmidhuber, J. (2012, June). Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, 3642-3649.
Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in plant science*, *19*(9), 592-601.

De Roos, A. P. W., Hayes, B. J., & Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics*, *183*(4), 1545-1553.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American*

*Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Do, D. N., Janss, L. L., Jensen, J., & Kadarmideen, H. N. (2015). SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. *Journal of animal science*, *93*(5), 2056-2063.

Durel, C. E., Laurens, F., Fouillet, A., & Lespinasse, Y. (1998). Utilization of pedigree information to estimate genetic parameters from large unbalanced data sets in apple. *Theoretical and applied genetics*, *96*(8), 1077-1085.

Elliot L. Heffner, E. L. Heffner, Mark E. Sorrells, M. E. Sorrells, & Jean-Luc Jannink, J. Jannink. (2009). Genomic Selection for Crop Improvement. *Crop science, 49*, 1-12.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The plant genome*, *4*(3), 250-255.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, *14*(8), 2611-2620.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1-22.

Gao, N., Martini, J. W., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., & Li, J. (2017). Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics*, *207*(2), 489-501.

González-Recio, O., Jiménez-Montero, J. A., & Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *Journal of dairy science*, *96*(1), 614-624.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., ... & Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and applied genetics*, *127*(3), 749-762.

Habier, D., Fernando, R. L., & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389-2397.

Hayes, B. J., & Goddard, M. E. (2008). Prediction of breeding values using marker-derived relationship matrices. *Journal of animal science*, *86*(9), 2089-2092.

Henderson, C.R. (1963) Selection Index and Expected Genetic Advance. *Statistical genetics and plant breeding*, National Academy of Sciences, No. 982, National Research Council Publication, Washington DC, 141-163.

Henderson, C. R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics*, *32*(1), 69–83. https://doi.org/10.2307/2529339

Henderson, C. R., & Quaas, R. L. (1976). Multiple trait evaluation using relatives' records. *Journal of animal science*, *43*(6), 1188-1197.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE signal processing magazine*, *29*(6), 82-97.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de Los Campos, G., & Hsu, S. D. (2018). Accurate genomic prediction of human height. *Genetics*, *210*(2), 477-497.

Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C., & Hsu, S. D. (2019). Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific reports*, *9*(1), 1-16.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., ... & Jannink, J. L. (2011). Genomic selection in plant breeding: knowledge and prospects. *Advances in agronomy*, 110, 77-123.

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825-2830.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Ramstein, G. P., Jensen, S. E., & Buckler, E. S. (2019). Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theoretical and applied genetics*, *132*(3), 559–567.

Rouchka, E. C. (1997). A brief overview of gibbs sampling. *Bioinformatics Technical Report Series, No. TR-ULBL-2008-02, University of Louisville*, *9*-15.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., & Korte, A. (2016). AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic Acids Research*, 45(1), 1054–1059.

Teng, J., Ye, S., Gao, N., Chen, Z., Diao, S., Li, X., Yuan, X., Zhang, H., Li, J., Zhang, X., & Zhang, Z. (2022). Incorporating genomic annotation into single-step genomic prediction with imputed whole-genome sequence data. *Journal of integrative agriculture*, 21(4), 1126–1136.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, *91*(11), 4414-4423.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... & Lichtner, P. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, *46*(11), 1173-1186.