

Schematic for two-layer cross-validation machine learning model.

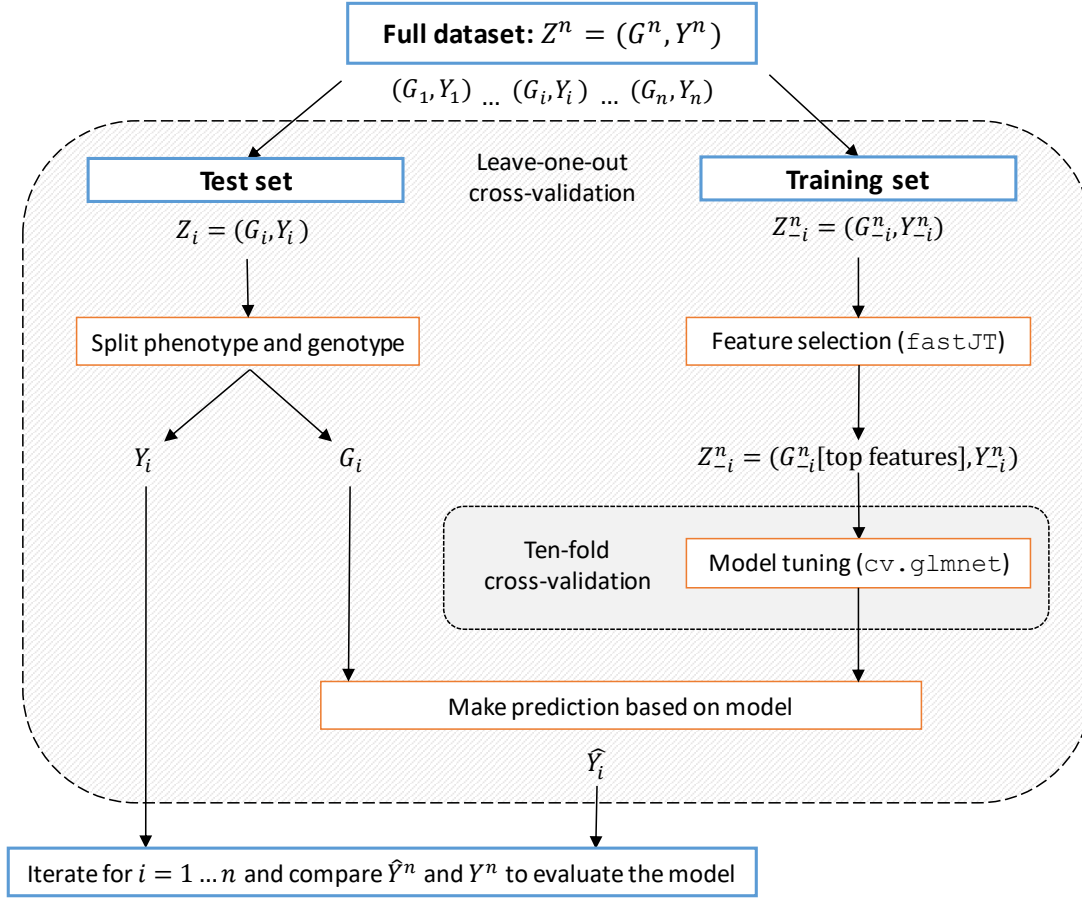


Figure S1: Here $Z^n = (G^n, Y^n)$ represents the feature (G^n) and phenotype (Y^n) data for all n samples. $Z_{-i}^n = (G_{-i}^n, Y_{-i}^n)$ represents these data for $n - 1$ samples remaining after excluding the i^{th} sample. $Z_i = (G_i, Y_i)$ represents all the feature and phenotype data for the i^{th} sample. Z_{-i}^n and Z_i constitute the training and testing data sets for the i^{th} iteration of the outer leave-one-out cross-validation layer. The top features are selected by applying `fastJT` to the training data $Z_{-i}^n = (G_{-i}^n, Y_{-i}^n)$. On the basis of these selected features and the training data, an elastic net model is tuned using ten-fold cross-validation with `cv.glmnet`. The tuned model is applied to G_i , the feature data of sample i , to obtain a prediction $\hat{Y}_i$. This process is repeated for samples $i = 1 \dots n$ to get $\hat{Y}^n = (\hat{Y}_1, \dots, \hat{Y}_n)$. The model is evaluated by comparing the n cross-validated predictions to the corresponding observed values $Y^n = (Y_1, \dots, Y_n)$.