# Supplemental Figure Legends

**Figure S1. Height prediction performance for non-linear GP algorithms during hyperparameter grid search. (A)** Average (line) and standard deviation (shadow) of mean squared error (MSE) over hyperparameter space for SVR based models predicting height as the penalty (*C*) (X-axis) change. SVRrbf and SVRpoly results are shown using gamma=1x10-5 and 1x10-4, respectively. **(B)** Distribution of the MSE across hyperparameter space for RF (left) and GTB (right) as the maximum features available to each tree (Max Features; X-axis) and maximum tree depth (color) change. GB results are shown using a learning rate = 0.01. **(C)** Average MSE across hyperparameter space for ANN models with different network architectures (X-axis), degrees of regularization using dropout (D.o.) or L2 regularization (L2), using either the Rectified Linear Unit (ReLU; left) or Sigmoid (right) activation function. **(D)** Distribution of the variance in MSE across the hyperparameter space for predicting height in each species using each GP algorithm. Black bar represents the median variance across the species for each GP algorithm.

**Figure S2. Comparison of feature selection algorithms and change in performance variation after feature selection. (A)** Average number of overlapping markers in the top markers (p) selected by three different feature selection algorithms for predicting height in maize across ten replicates for p=10 ~ 8,000. **(B)** Change in model performance (*r*) using five GP algorithms at predicting height in maize as the number of input markers (p) selected by three different feature selection algorithms increases. Dashed line: the mean *r* for each GP algorithm when all maize markers were used. Colored lines: mean *r* of models using features selection subsets using algorithms colored as in **(A)**. Colored areas: standard deviation around the mean. **(C)** Distribution and median of the standard deviation of model performance (*r*) across replicates for all feature selection subsets (p=10 ~ 8,000) combined across all species for each GP algorithm **(D)** Distribution and median of the standard deviation of model performance across replicates for all feature selection subsets (p=10 ~ 8,000) by species for each GP algorithm.

**Figure S3. Hyperparameter random search results from predicting height in spruce. (A)** Overview of the architecture and parameters used to train the CNN models. The parameters listed below for each layer (black) were either pre-set (value shown in purple) or the value for that parameter was selected using RandomSearchCV (values tested shown in red). **(B)** Average mean squared error (MSE) across the hyperparameter space for predicting height in rice (replicate #1). RMSprop: Root Mean Square propagation.

**Figure S4. Number of wins between each pair of GP algorithm**. Percent of replicates where one GP algorithm (y-axis) outperformed another GP algorithm (x-axis) for predicting each species/trait combination.

**Figure S5. Similarity between traits and datasets in model performance.** Hierarchical clustering of trait:species combinations based on mean predictive performance across all algorithms included in the benchmark. HT: height. DT: developmental timing. YLD: yield, GM: grain moisture. DBH: diameter at breast height. DE: wood density. ST: standability.