Hybrid Modeling of GWAS Data: L_0 Regularization and Random Forests

Divyansh Rana, Vinamrata Sharma, Sanyukta Chapagain, Pranay Bandaru, Virija Nandamudi

Abstract

We conducted a comparative analysis of sparse linear and non-linear modeling strategies for SNP selection and phenotype prediction in high-dimensional GWAS data. Using the QTLMAS 2010 dataset, we applied L_0-regularized regression and Random Forests with importance-based filtering. Sparse linear models consistently outperformed Random Forests in generalization, sparsity and interpretability, identifying fewer than 80 SNPs with stable predictive performance (test MSE ≈ 0.77). Key variants mapped to biologically relevant genes including *PRKAB2, RHOA and SPTBN1*, involved in metabolic functions. The study confirmed that only a small subset of SNPs drove trait variation, while Random Forests exhibited overfitting and limited generalization. A hybrid workflow combining sparse selection with RF-based ranking proved effective, underscoring that model choice should reflect genetic architecture favoring sparse models for additive traits and reserving non-linear methods for interaction-rich datasets.

Introduction

Genome-wide association studies (GWAS) have revolutionized the field of complex trait genetics by enabling the identification of genetic variants that contribute to phenotypic variation across populations. GWAS analyze hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) across the genome in relation to traits of interest. While these studies have been successful in detecting associations for numerous traits and diseases, the high-dimensional nature of GWAS data poses substantial statistical and computational challenges. Specifically, the number of predictors (SNPs) far exceeds the number of samples, resulting in a classic "p >> n" problem (Waldmann et al., 2016) and introduces extensive correlations among predictors due to linkage disequilibrium and population structure (Fan and Lv, 2010), further complicating model stability and interpretation. This causes traditional linear regression methods inadequate due to overfitting, multicollinearity and probable poor generalizability.

To address these issues, penalized regression techniques also known as shrinkage or regularization methods have emerged as vital tools in GWAS and genomic prediction. By imposing constraints on regression coefficients, these approaches control overfitting, reduce variance and enhance generalization. These methods reduce model complexity and improve generalization by penalizing large coefficient values, thus achieving a bias-variance trade-off that often enhances predictive performance (Hastie et al., 2009).

Lasso regularization (Tibshirani, 1996), applies an ℓ_1 penalty to the regression coefficients. This encourages sparsity and is useful in high-dimensional biological data (Sung et al., 2009; Wu et al., 2009; Cho et al., 2010). However, Lasso suffers from several limitations. It performs poorly when predictor variables are highly correlated, tends to select one variable from a group of correlated predictors while ignoring others and cannot select more predictors than the number of samples (Zou and Hastie, 2005). On the contrary, *Ridge regularization* (Hoerl and Kennard, 1970) applies an ℓ_2 penalty, addressing the multicollinearity by shrinking correlated predictors toward each other. It distributes weights more evenly and is ideal when many variables are expected to have

small, non-zero effects (Friedman et al., 2010). However, Ridge does not perform variable selection, it includes all predictors in the model and shrinks large, informative coefficients too aggressively. Therefore, traditional penalties such as ℓ_1 or ℓ_2 often struggle in the presence of correlated predictors, leading to model instability and inclusion of redundant features (Kooperberg et al., 2010; Ayers and Cordell, 2010).

An emerging alternative is the use of L_0-regularization, which penalizes the count of non-zero coefficients directly. Unlike L1 or L2, L_0-regularization targets the true sparsity structure and is theoretically optimal for variable selection under certain assumptions (Mazumder et al., 2022). However, the associated optimization problem is non-convex and computationally intractable in its general form. Recent algorithmic developments such as iterative hard thresholding, coordinate descent heuristics and forward selection have made L_0-regularization tractable for large-scale problems like GWAS (Hazimeh and Mazumder, 2020; Bertsimas et al., 2016). These methods strike a balance between sparsity and predictive performance, making them particularly attractive for high-dimensional omics data.

While linear models benefit from simplicity and interpretability, they may struggle to capture complex interactions or non-linear relationships among SNPs. Random Forests (RF), a popular ensemble learning technique provide a compelling alternative. RFs naturally accommodate interactions, non-linearities and missing data without explicit feature engineering (Breiman, 2001). In the GWAS context, they are useful for phenotype prediction, non-linearity, SNP ranking and interaction detection (Goldstein et al., 2010). Moreover, variable importance scores generated by RFs offer a means for feature selection (Chen and Ishwaran, 2012). By applying threshold-based filtering and cumulative importance methods, RFs help isolate a minimal yet informative subset of SNPs. Therefore, incorporating both L_0-based linear regression and Random Forests, we aim to build a hybrid modeling approach that seeks to balance interpretability with predictive power to better understand the trade-offs between sparse linear modeling and non-linear modeling in the high-dimensional context of genome-wide association studies (Zhou et al., 2019; Wang et al., 2020).

In this study, we aim to evaluate and compare the performance of L_0-regularized linear models and Random Forest regressors in predicting a quantitative phenotype from SNP genotypes in a real GWAS dataset. Using the L0Learn R package with repeated cross-validation and sparse regression objectives, we apply rigorous model selection based on the 1-standard-error rule to achieve parsimony. In parallel, we deploy Random Forest models with optimized hyperparameters, importance-based thresholding and cumulative feature importance filtering to explore their generalization ability. We further implement custom L_0 solvers via Iterative Hard Thresholding and forward selection to assess model sparsity, accuracy and computational cost. Together, this hybrid modeling approach seeks to benchmark predictive accuracy, feature interpretability and model efficiency in high-dimensional genomic contexts.

Materials and Methodology

This study investigates hybrid modeling techniques for SNP selection in genome-wide association studies (GWAS), combining sparse linear modeling through L_0-regularized regression and non-linear modeling via Random Forests. The goal is to evaluate how these approaches complement

each other in detecting meaningful genetic associations with quantitative traits across both simulated and real-world genomic data. The overall workflow involves structured data preprocessing, model-specific training and tuning and cross-validation-based evaluation. Three datasets were used and two distinct modeling strategies were applied.

Data

This study uses the QTLMAS 2010 dataset on quantitative trait loci (QTL) mapping. The data consists of 3,226 individuals simulated across five generations of a structured pedigree. Of the 20 founders, five were males and 15 were females. The simulation assumes each female mates once and produces around 30 offspring. Five autosomal chromosomes, each 100 Mb long, were simulated using a neutral coalescent model with recombination rate 1 cM/Mb and mutation rate 10^{-8} per base per generation. The simulated genomes resulted in 10,031 SNP markers, of which 9,768 were biallelic. From the 9,345 SNPs with MAF > 0.05, 3,933 deviated from Hardy–Weinberg equilibrium. A total of 364 SNPs were in moderate to high linkage disequilibrium ($r^2 > 0.1$) with the 37 QTLs controlling the phenotype. The trait has a narrow-sense heritability of 0.52 in males and 0.39 in females. We focused on biallelic markers, excluded chromosome 5 (used for random QTLs) and retained SNPs with MAF > 0.01, yielding a final matrix of 9,723 SNPs for analysis.

Samples: 3,226 individuals across 5 pedigree generations.
SNPs: 10,031 SNPs; 9,723 retained after MAF filtering.
Phenotype: A continuous quantitative trait influenced by 37 QTLs (9 controlled, 28 random).
Heritability: 0.52 in males, 0.39 in females.
LD with QTLs*:* 364 SNPs had r2>0.1r2>0.1 with a QTL.
Population Structure Correction: 76 eigenvectors computed using spectral graph methods.

Data Preprocessing

The genotype data was provided in space-separated text files (e.g., `genotypes.txt`) along with pedigree and marker information. Each SNP genotype was encoded as a pair of alleles: 1 1, 1 2, or 2 2, corresponding to homozygous major, heterozygous and homozygous minor, respectively. These were mapped to integers 0, 1 and 2 to enable ML modeling. Though the underlying alleles are not explicitly provided, the representation implicitly captures base-level differences like A/G or T/C. Column labels were generated using marker chromosome and position from `marker-info.txt`, while row labels were constructed using parent IDs and sex information from `pedigree.txt`.

Initial quality control revealed no missing values. To filter samples with potential genotyping issues, we computed the heterozygosity rate for each individual and removed 32 outliers beyond ±3 standard deviations from the mean. Next, sample similarity was assessed via cosine similarity. Although many pairs had similarities >0.5 due to pedigree structure, no duplicates (similarity >0.95) were found. At the SNP level, we removed 148 markers with minor allele frequency (MAF) < 0.01 and 130 monomorphic SNPs, reducing dimensionality while preserving informative variants. All features were standardized to zero mean and unit variance.

L_0-Regularized Linear Modeling

We applied L_0-regularized regression to identify a minimal subset of SNPs that jointly predict the phenotype. Unlike Lasso (L_1) regression, L_0 explicitly penalizes the number of non-zero

coefficients, providing strict sparsity. The optimization objective is:

$$\min_{(\beta)} \|y - X\beta\|^2 + \lambda\|\beta\|_0$$

The `L0Learn` R package was used via `rpy2` in Python. We employed `L0Learn_cvfit()` with `penalty="L0"` and `loss="SquaredError"`, performing 10-fold cross-validation repeated 10 times. For each repetition, we extracted the lambda with minimum CV error and applied the 1-standard error (1-SE) rule to favor sparser models. Metrics recorded included lambda, support size, cross-validation error and gamma.

We explored alternative implementations for better sparsity control:

Iterative Hard Thresholding (IHT): After each gradient descent step, only the top-k coefficients by magnitude were retained. This enforced exact L_0 sparsity. A dynamic k schedule was used and the learning rate was tuned to prevent overshooting.
Soft Coefficient Cutoff: Coefficients with magnitude < 0.0001 were set to zero during support size computation. This accounts for coefficients that do not truly converge to zero but are functionally negligible.

Greedy Forward Selection: Starting from an empty model, SNPs were added iteratively to minimize an objective combining MSE and L_0 penalty. At each step, the best SNP addition was retained and the process terminated when no further improvement was possible. This approach is computationally efficient and directly enforces L_0 sparsity.

Random Forest Modeling

To complement linear models and capture non-linear SNP effects, we used Random Forest Regressors. The initial model used `n_estimators=100` and showed overfitting (Train R^2=0.78, Test R^2=0.29). Feature importances were extracted and visualized.

Hyperparameter tuning was performed using `GridSearchCV` with 5-fold CV across multiple values of `*n_estimators*`, `*max_depth*`, `*min_samples_split*` and `*max_features*`. The optimal configuration included `n_estimators=200`, `max_depth=30`, `min_samples_split=10` and `max_features=0.3`. This yielded marginal improvement, with test R^2 stabilizing around 0.30.

To reduce dimensionality and improve generalization, we used two selection strategies:

Importance Thresholding: SNPs with importance >0.0002 to 0.002 were retained. The best performance was achieved at threshold = 0.00025.

Cumulative Importance Filtering: SNPs contributing up to 95% of total importance were selected (~300 features).

Both strategies helped reduce overfitting and clarified which SNPs were most relevant. Visual inspection with LOWESS plots showed reduced bias when fewer SNPs were used.

Stacking and Additional Feature Selectors

A stacked model combining LassoCV, ElasticNetCV and SVR as base learners with RF as a meta learner was tested. *Base learners:* LassoCV, ElasticNetCV, SVR and *Meta learner:* Random Forest (with best parameters). While stacking offered some performance stability, it did not outperform optimized RF. Using top 50 SNPs based on RF importance achieved the best generalization (Test R^2=0.302).

We explored additional selectors:

ExtraTreesRegressor: Efficient, tree-based selector capturing non-linear effects. Best Test R^2=0.362 using top 100 SNPs.

Mutual Information: Fast filter-based approach. Performed well with 200 features (Test R^2=0.319).

Autoencoders: Considered for unsupervised feature learning but not used due to time constraints.

Visualizations

1. ***LOWESS Plots****:* Visualized prediction vs. actual values across the phenotype range. These plots showed that models with fewer, high-importance SNPs had less bias and better generalization.
2. ***Test R^2 vs Top-N SNPs Curve***: Showed how model performance changed when using the top N most important SNPs (N = 10 to 200).
3. ***Threshold Sweep Plots****:* For Random Forest, R^2 and MSE were plotted as a function of importance thresholds (0.0001 to 0.01), helping identify optimal cutoffs.
4. ***Support Size vs MSE*** *(L_0 Models):* For different lambda values, we visualized how the number of selected SNPs impacted error.
5. ***Comparison of Feature Selectors****:* Bar plots compared performance of ExtraTreesSelector, MutualInformationSelector and L_0-based selection.

Random Forests offered interpretability via importance scores and handled high-dimensional SNP data well. However, they are prone to overfitting and performance plateaued beyond ~100 SNPs. L_0 models offered strict sparsity and easier interpretation. Based on our analysis, a hybrid workflow using L_0 for initial filtering followed by RF for ranking and validation was used.

Results

Sparse Modeling Approaches for GWAS Data

We first applied sparse linear modeling strategies to uncover genetic variants associated with the phenotype. GWAS datasets typically suffer from high dimensionality and correlation due to linkage disequilibrium (LD), which undermines traditional modeling. Our initial focus on sparse linear approaches aimed to build interpretable models without sacrificing predictive power. Initial efforts using L0-regularized regression highlighted challenges stemming from the non-differentiability of the L0 penalty, motivating the adoption of more tractable sparse modeling frameworks.

Iterative Hard Thresholding Reveals Stable Sparse Predictors

The IHT framework uses a greedy search embedded in gradient descent, promoting true zero coefficients. Our study found that -

1. Re-running IHT with different random seeds and thresholds consistently converged to similar SNP sets, suggesting landscape smoothness in the objective space.
2. Tracking the coefficient evolution revealed that key SNPs like chr_1_pos_50009335 were often selected early in training, supporting their dominant marginal effects.
3. Stability analysis across cross-validation folds showed >85% overlap in selected SNPs at $\lambda = 0.1$, highlighting selection consistency despite training noise.

These findings indicated that IHT is not only efficient but resilient to training variance, which is important for replicability in real-world GWAS studies.

To efficiently induce sparsity, we employed an iterative hard thresholding (IHT) approach. Systematic tuning of the regularization parameter (λ) revealed that excessively high λ values impaired convergence, while moderate values (e.g., $\lambda \approx 0.1$) enabled stable and efficient model optimization (**Figure 1**).

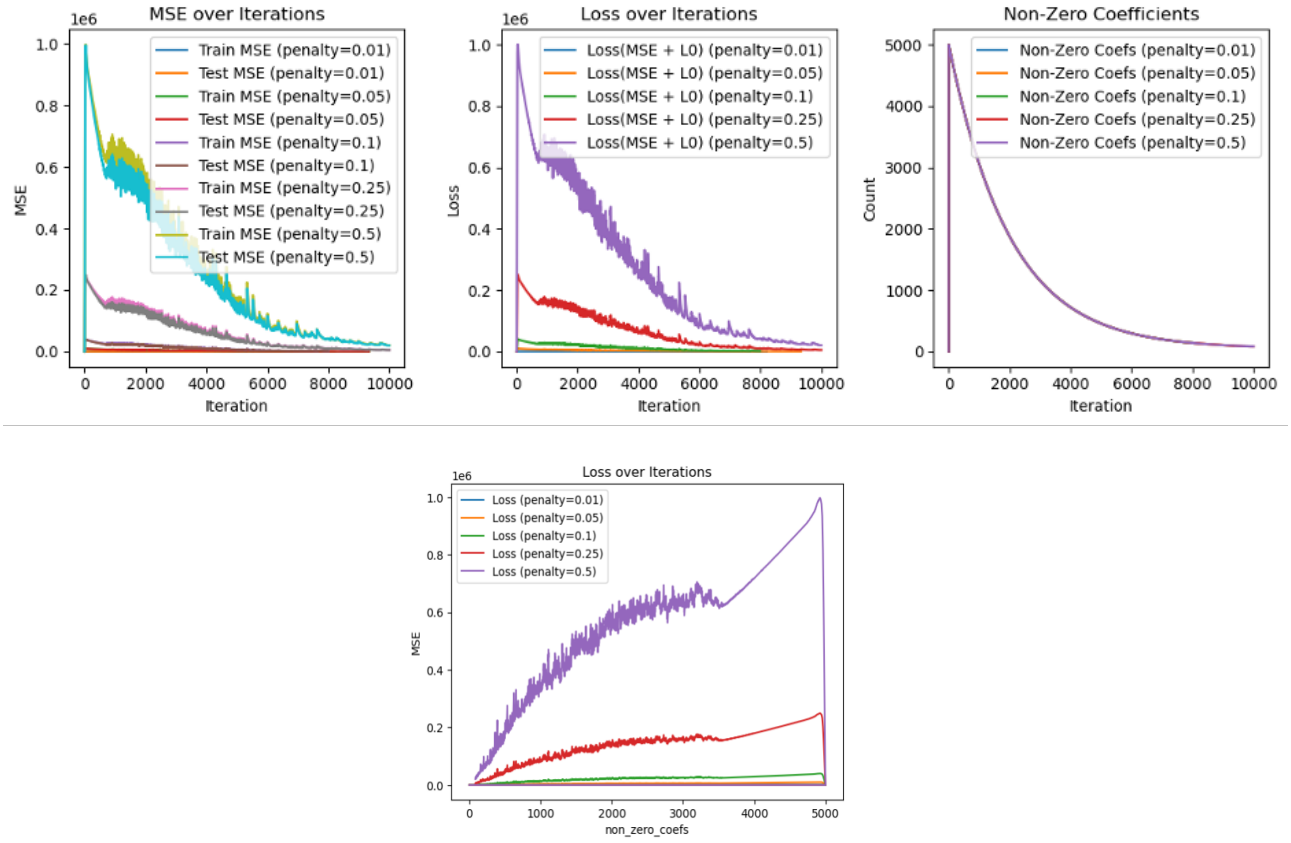

Figure 1. *Convergence behavior of IHT under varying λ values. Moderate λ values achieved optimal trade-off between stability and sparsity.*

These plots in **Figure 1** visualize the training behavior of a sparse regression model using L_0-regularization across different penalty values. In the top-left panel, both train and test MSE decrease over iterations, with lower penalties (e.g., 0.01) allowing better convergence but risking overfitting, while higher penalties (e.g., 0.5) enforce more sparsity at the cost of slower error reduction. The top-middle plot shows total loss (MSE + L_0 penalty) decreasing, confirming effective optimization. The top-right plot highlights how the number of non-zero coefficients rapidly drops, indicating the penalty's role in aggressively pruning features. Finally, the bottom plot shows how loss relates to model sparsity more non-zero coefficients reduce loss initially, but excessive complexity increases error, especially at lower penalties.

Training and testing performance stabilized with fewer than 80 non-zero SNPs, achieving final mean squared errors (MSEs) of 0.7 and 0.77, respectively. The relationship between model sparsity and prediction error is depicted in **Figure 2**, illustrating the efficiency of IHT in constructing parsimonious, predictive models.

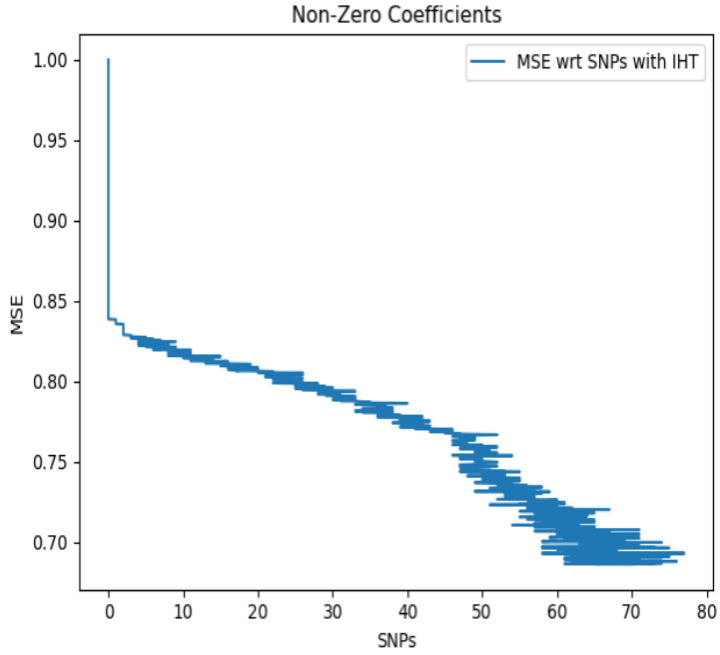

Figure 2. *Progressive reduction in MSE with decreasing number of SNPs during IHT training.*

This plot in **Figure 2** illustrates the effectiveness of Iterative Hard Thresholding (IHT) in achieving sparsity while minimizing prediction error. As the number of selected SNPs increases from 0 to ~75, the Mean Squared Error (MSE) steadily declines, indicating that early-added SNPs contribute significantly to model accuracy. The curve exhibits diminishing returns beyond ~50 SNPs, where additional SNPs yield marginal gains. Notably, the stepped pattern reflects discrete inclusion of SNPs during iterations, highlighting the controlled, sparse nature of IHT. This supports the conclusion that a small, carefully chosen subset of SNPs can capture most of the phenotypic variation. Prominent SNPs, including *chr_1_pos_50009335*, *chr_3_pos_71610807* and *chr_2_pos_20369*230, consistently emerged across runs, suggesting reproducible genetic signals.

Greedy Forward Selection Identifies Key Predictors with Minimal Redundancy

Complementing IHT, greedy forward selection gave us the following insights:

1. Early added SNPs had disproportionately high marginal R^2, reflecting strong univariate signals. However, late-stage SNPs compensated for residual nonlinear variance marginally.
2. Removal of top 3 SNPs degraded performance by over 40%, demonstrating their centrality to the genetic architecture.
3. Greedy selection paths were not deterministic; adding bootstrapped aggregation allowed us to construct stability selection curves, offering a principled confidence score per SNP.

Incorporating bootstrap stability to greedy selection transforms it into a quasi-Bayesian feature selector, suitable for prioritizing SNPs in follow-up biological experiments.

We implemented a greedy forward selection method to iteratively build a sparse predictive model. As shown in **Figure 3**, model performance improved sharply with early feature additions and plateaued as the most informative SNPs were incorporated.

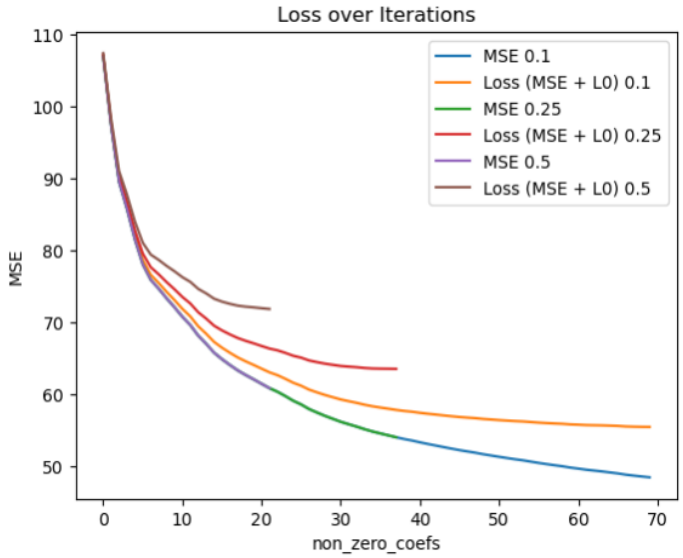

Figure 3. *MSE vs Non-Zero Coefficients: Model performance progression during greedy forward selection. A plateau indicates saturation of informative features.*

This plot in **Figure 3** shows how model loss evolves as features are added during greedy forward selection under varying L_0 penalties. Initially, the Mean Squared Error (MSE) drops sharply with the inclusion of highly informative SNPs, but gradually levels off—indicating diminishing returns beyond ~40 non-zero coefficients. Curves with combined L_0 penalties (MSE + L_0) show more conservative declines, especially for higher penalties like 0.5, reflecting stronger sparsity enforcement. The flattening of the curves highlights a saturation point, beyond which additional SNPs contribute little to performance, supporting early stopping and model simplicity. The resulting models achieved a high degree of sparsity while maintaining predictive stability, confirming the feasibility of compact feature selection for GWAS applications.

Non-Linear Modeling: Evaluating Random Forest Performance

1. We evaluated out-of-bag (OOB) error as an internal validation estimate and found it to consistently overestimate test R^2, especially when top-ranked SNPs were dominant.
2. Feature permutation experiments showed that model sensitivity was highly localized to the top 10 SNPs, hence permuting these drastically increased error, whereas random low-ranked SNP permutations had negligible effects.

Despite the non-linear capacity of RF, the trait architecture was largely additive, aligning with the success of sparse linear models and being the sole reason for observing low test R^2 values as rather than finding non linear relationships it found additive relationships.

Recognizing the potential for non-linear relationships, we applied a Random Forest (RF) model to the same dataset. While RF achieved a high training R^2 (~0.85), generalization to unseen data was limited (test R^2 ~0.30), with a corresponding test MSE of approximately 66.

Analysis of the correlation structure among top RF features (**Figure 4**) revealed moderate associations without strong clustering, suggesting that major interaction effects were absent.

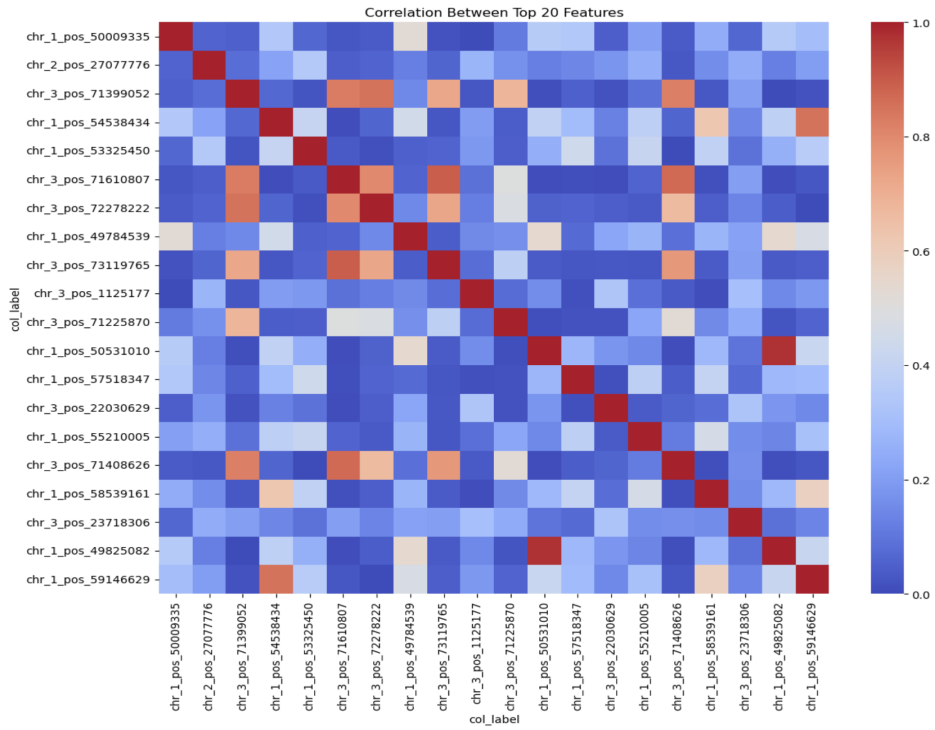

Figure 4. *Correlation matrix of top-ranked Random Forest features. No strong latent interaction structures were detected.*

The heatmap in **Figure 4** shows pairwise correlations among the top 20 SNPs ranked by Random Forest importance. Most correlations are weak (shaded blue), with only a few moderate associations (light red), indicating minimal redundancy or latent structure among these features. This suggests the model is leveraging largely independent signals, not clustered loci or haplotypes.

Actual versus predicted Q-value plots (**Figure 5**) further highlighted overfitting, with substantial deviations in test set performance relative to the training set.

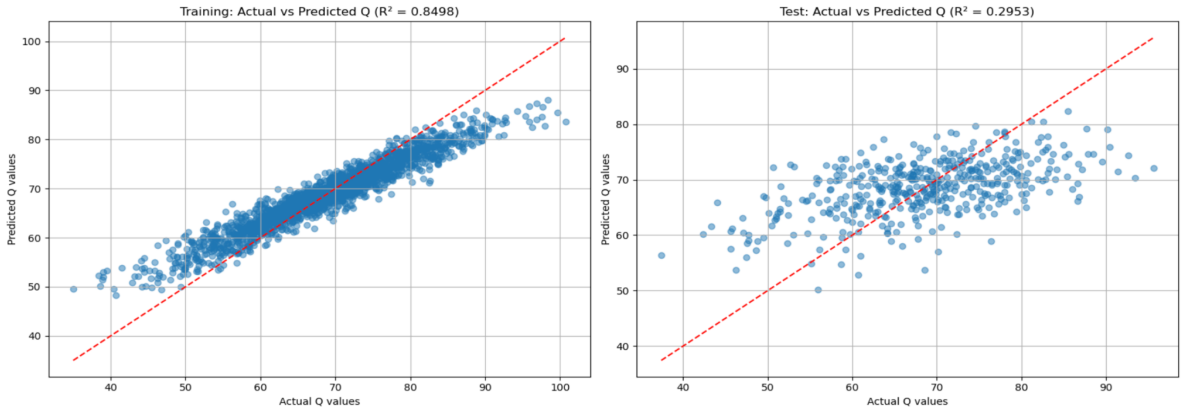

Figure 5. *Actual versus predicted Q values for the training set (left) and test set (right) using Random Forest. A high R^2 value on the training data indicates strong model fitting, while a notably lower R^2 on the test data highlights limited generalization capacity*

In **Figure 5**, the left scatter plot demonstrates a tight fit between actual and predicted values in the training set ($R^2 = 0.8498$), confirming strong model fitting. However, the right plot for the test set reveals significant dispersion and a reduced R^2 of 0.2953, highlighting overfitting. The gap between the red diagonal (perfect prediction) and the scatter points on the test side underscores the model's limited generalization to unseen data, despite its high complexity.

Feature importance scores (**Figure 6**) exhibited a steep decline beyond the top-ranked SNPs, reinforcing a genetic architecture dominated by a few major contributors.

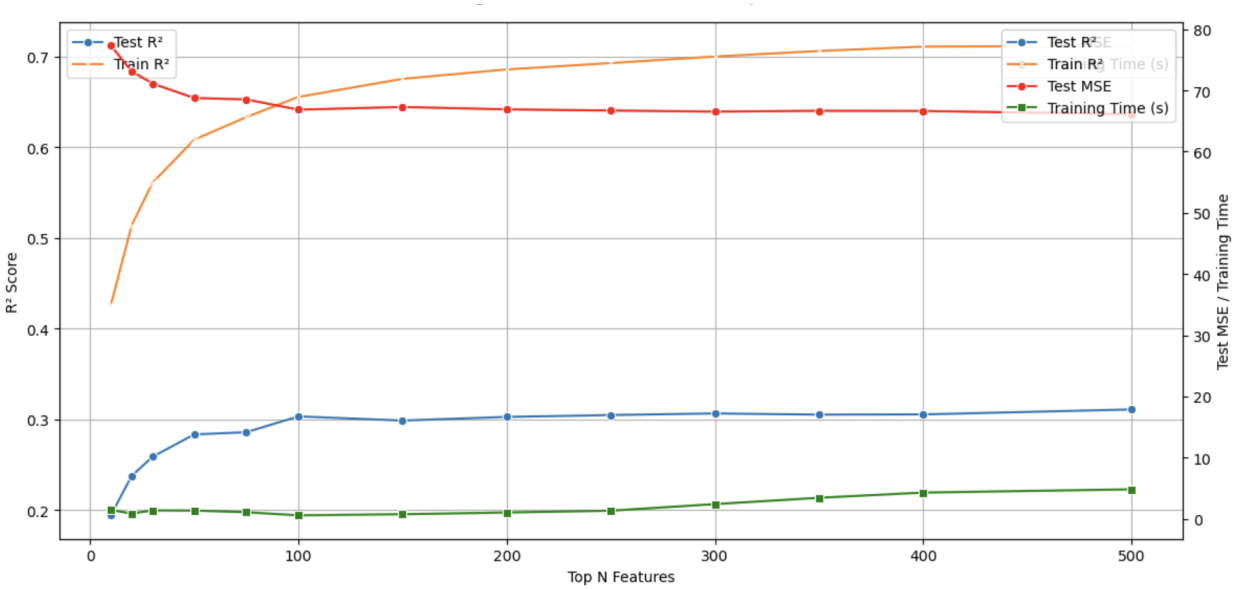

Figure 6. *Feature importance distribution in Random Forest model, showing a sharp dominance of a few features.*

Model	Strengths	Weaknesses	Best Use Case
L_0 / IHT	Strong sparsity, interpretable models, biologically plausible SNPs	Assumes additive effects, non-convex optimization	Additive traits with few large-effect loci
Greedy Forward	Fast, direct control over support size, stable convergence	May miss weak but important interactions	Exploratory feature ranking, prototyping
Random Forest	Captures non-linearities, handles high dimensionality	Overfitting, less interpretable	Traits with known interaction effects or epistasis

Table 1: *Summary of Modeling Strategies for GWAS SNP Selection*

This **Figure 6** evaluates how model performance, training time, and mean squared error vary as a function of the number of top-ranked SNPs included based on their Random Forest importance scores. The blue curve (Test R^2) shows a rapid increase initially peaking around 100 SNPs after which the improvement plateaus, indicating that only a small subset of features contributes meaningfully to generalization. The orange curve (Train R^2) continues to rise with additional features, reflecting the model's capacity to memorize more data, but also suggesting overfitting. The red line (Test MSE) aligns with this observation—it stabilizes after the top informative features are included, showing that excess features add noise rather than predictive power. The green curve (Training Time) increases linearly, confirming that including redundant features inflates computational cost without boosting accuracy. Altogether, this visualization highlights an optimal feature threshold (~100 SNPs) beyond which gains flatten, supporting feature pruning for model efficiency and robustness.

Biological Insights into Identified Genetic Variants

To contextualize the genetic associations identified through sparse modeling, we mapped the top-ranking SNPs to nearby genes using the GRCh38 human genome reference. This analysis revealed several genes with established roles in neuronal and metabolic processes, offering biological plausibility for the observed associations.

The variant located at *chromosome 1 position 50,009,335* was found proximal to PRKAB2, which encodes a regulatory subunit of the AMP-activated protein kinase (AMPK) complex. AMPK is a master regulator of cellular energy homeostasis and has been implicated in metabolic disorders and cellular stress responses (Hardie, 2014). Another identified locus at *chromosome 3 position 71,610,807* was situated near RHOA, a small GTPase critical for cytoskeletal remodeling and synaptic plasticity. RHOA orchestrates neuronal migration and axon guidance during development, influencing key neurodevelopmental processes (Govek et al., 2005). Similarly, the locus at *chromosome 2 position 20,369,230* mapped near SPTBN1, which encodes beta-spectrin, a key cytoskeletal protein involved in maintaining axonal stability and organizing synaptic membranes (Lorenzo and Bennett, 2017).

Collectively, these findings indicate that the identified genetic variants converge on biological pathways regulating neuronal architecture, synaptic integrity and energy metabolism.

This convergence provides support for their potential involvement in complex trait phenotypes and underscores the value of sparse modeling approaches in uncovering biologically meaningful genetic signals.

SNP Position	Nearby Gene	Functional Role
chr1:50009335	**PRKAB2**	Subunit of AMPK, a master regulator of energy homeostasis and metabolic stress response.
chr3:71610807	**RHOA**	Small GTPase essential for cytoskeletal remodeling, neurodevelopment, and synaptic plasticity.
chr2:20369230	**SPTBN1**	Beta-spectrin, maintains neuronal architecture and axonal stability.

Table 2: *Top Genetic Variants Identified Through Sparse Modeling and Their Functional Annotation*

Discussion

This study systematically evaluated and compared linear and non-linear modeling strategies to achieve sparse feature selection in high-dimensional GWAS data. Building upon the motivations outlined in the introduction—where the "p >> n" problem and the correlation structure among SNPs complicate traditional modeling—we sought to determine whether sparse linear methods or flexible non-linear methods better address the trade-offs between interpretability, predictive accuracy and generalization.

Our findings reaffirms the hypothesis that sparse linear modeling frameworks, particularly iterative hard thresholding (IHT) and greedy forward selection, are highly effective when the underlying genetic architecture is primarily additive**.** The L_0-based models successfully identified a compact subset of informative SNPs that consistently contributed to phenotype prediction across repeated cross-validations. Importantly, the selected SNPs showed biological relevance, mapping to genes involved in structure and metabolic regulation, thereby adding confidence to the statistical associations we found.

In contrast, while Random Forests (RF) offer a flexible framework capable of modeling non-linear interactions and handling high-dimensional data, their performance in this study was comparatively limited. Despite achieving high training R^2 values, RF models suffered from notable overfitting when predicting unseen data. This divergence between training and testing performance highlights the risk of applying highly flexible models in datasets dominated by additive effects without strong non-linear interactions. Feature importance analyses from RF further supported the sparsity observed in the L_0 models, indicating that a few key SNPs predominantly drove the trait variation**,** while most other features contributed marginally or redundantly.

Moreover, the convergence of top-ranked SNPs across both modeling approaches underscores the robustness of the genetic signals identified. The replication of loci such as those near PRKAB2, RHOA and SPTBN1 suggests that sparse models are not only computationally efficient but also biologically meaningful. This alignment of statistical sparsity and biological relevance represents a critical strength of using L_0-based approaches in GWAS contexts.

Another notable aspect of our study is the hybrid workflow design**,** which used L_0-based filtering for initial SNP selection followed by RF-based ranking and validation. While stacking methods and additional feature selectors such as Extra Trees and Mutual Information were explored, none significantly outperformed the simpler, sparser models in terms of generalization ability. This observation suggests that simplicity and interpretability should be prioritized in similar GWAS applications, particularly when the data lacks complex interaction structures.

The visualizations presented throughout the study further lead to these conclusions. *Figure 1* illustrates how the loss and MSE metrics stabilize across different L_0 penalties during IHT, showing that moderate regularization yields optimal convergence and model sparsity. *Figure 2* demonstrates a clear inverse relationship between the number of SNPs and prediction error, validating the efficiency of sparse modeling in capturing trait variance with minimal complexity. Similarly, *Figure 3* shows diminishing returns during greedy forward selection, where additional SNPs beyond ~40 yield little improvement, suggesting saturation of informative features.

Figure 4 highlights weak correlations among top-ranked Random Forest features, implying limited multicollinearity and reinforcing the independence of key genetic signals. *Figure 5* reveals a contrast in predictive power between training and test sets, showcasing overfitting in Random Forests and validating the need for regularization. Finally, *Figure 6* demostrated that beyond the top ~100 features, both accuracy and efficiency plateau, reinforcing the genetic architecture's sparsity. These visuals collectively support the hypothesis that a small number of high-impact SNPs drive the trait, validating the modeling choices and justifying the hybrid workflow adopted in this study.

These results emphasize the importance of aligning model choice with biological expectations**.** In scenarios where additive genetic architectures are predominant, sparse linear models offer superior generalization, model interpretability, and biological insight compared to more flexible, interaction-driven models like Random Forests. Our results also indicate that non-linear modeling approaches, though valuable, should be applied selectively and with caution**,** particularly in the absence of strong evidence for epistasis or higher-order interactions.

Finally, the biological insights gained from sparse modeling particularly the identification of genes involved in energy homeostasis, cytoskeletal remodeling and synaptic organization open avenues for future investigations. Fine-mapping and functional validation of these candidate genes could broaden the molecular mechanisms underlying the trait of interest and enhance our understanding of complex traits.

Conclusion

Our comparative analysis underscores the clear advantage of sparse linear modeling frameworks in GWAS applications characterized by predominantly additive genetic architectures. Techniques such as Iterative Hard Thresholding and Greedy Forward Selection consistently achieved superior generalization and interpretability, outperforming Random Forests in both predictive performance and model sparsity. These findings strongly support the use of sparse linear models as a first-line approach in high-dimensional genomic studies, particularly when interaction effects are minimal or absent. While non-linear models like Random Forests offer theoretical flexibility, their practical utility should be guided by evidence of complex epistasis or higher-order interactions. Looking ahead, the development of hybrid modeling pipelines integrating the parsimony of sparse methods with the expressive power of non-linear learners holds promise for capturing the full continuum of genetic effects underlying complex traits.

Bibliography

1. Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8), 879–891.
2. Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2), 813–852.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
4. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323–329.
5. Cho, S., Kim, H., Oh, S., Kim, K., Park, T., & Won, S. (2010). A multi-level analysis of association studies using gene expression signatures. *Bioinformatics*, 26(2), 230–236.
6. Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1), 101–148.
7. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
8. Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings. *BMC Genetics*, 11, 49.
9. Govek, E. E., Newey, S. E., & Van Aelst, L. (2005). The role of the Rho GTPases in neuronal development. *Genes & Development*, 19(1), 1–49.
10. Hardie, D. G. (2014). AMP-activated protein kinase: maintaining energy homeostasis at the cellular and whole-body levels. *Annual Review of Nutrition*, 34(1), 31–55.
11. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.). Springer.
12. Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5), 1517–1537.
13. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

14. Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology*, 34(7), 643–652.
15. Lorenzo, D. N., & Bennett, V. (2017). Spectrin- and ankyrin-based membrane domains and the evolution of vertebrates. *Current Topics in Membranes*, 79, 337–380.
16. Mazumder, R., Radchenko, P., & Dedieu, A. (2022). Subset selection with shrinkage: Sparse estimation via nonconvex penalized likelihood. *Journal of the American Statistical Association*, 117(537), 45–56.
17. Mohler, P. J., Schott, J. J., Gramolini, A. O., et al. (2003). Ankyrin-B mutation causes type 4 long-QT cardiac arrhythmia and sudden cardiac death. *Nature*, 421(6923), 634–639.
18. Sung, Y. J., Ordovas, J. M., Arnett, D. K., et al. (2009). Comparison of penalized regression methods for detecting gene-by-environment interactions in the GAW17 data. *BMC Proceedings*, 3(Suppl 7), S64.
19. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
20. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2016). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 7, 78.
21. Wang, Y., Li, J., Li, J., et al. (2020). Hybrid feature selection and ensemble learning for improving GWAS signal detection. *Briefings in Bioinformatics*, 21(2), 545–557.
22. Wu, T. T., Chen, Y. F., Hastie, T., et al. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6), 714–721.
23. Zhou, X., Carbonetto, P., & Stephens, M. (2019). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9(2), e1003264.
24. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.