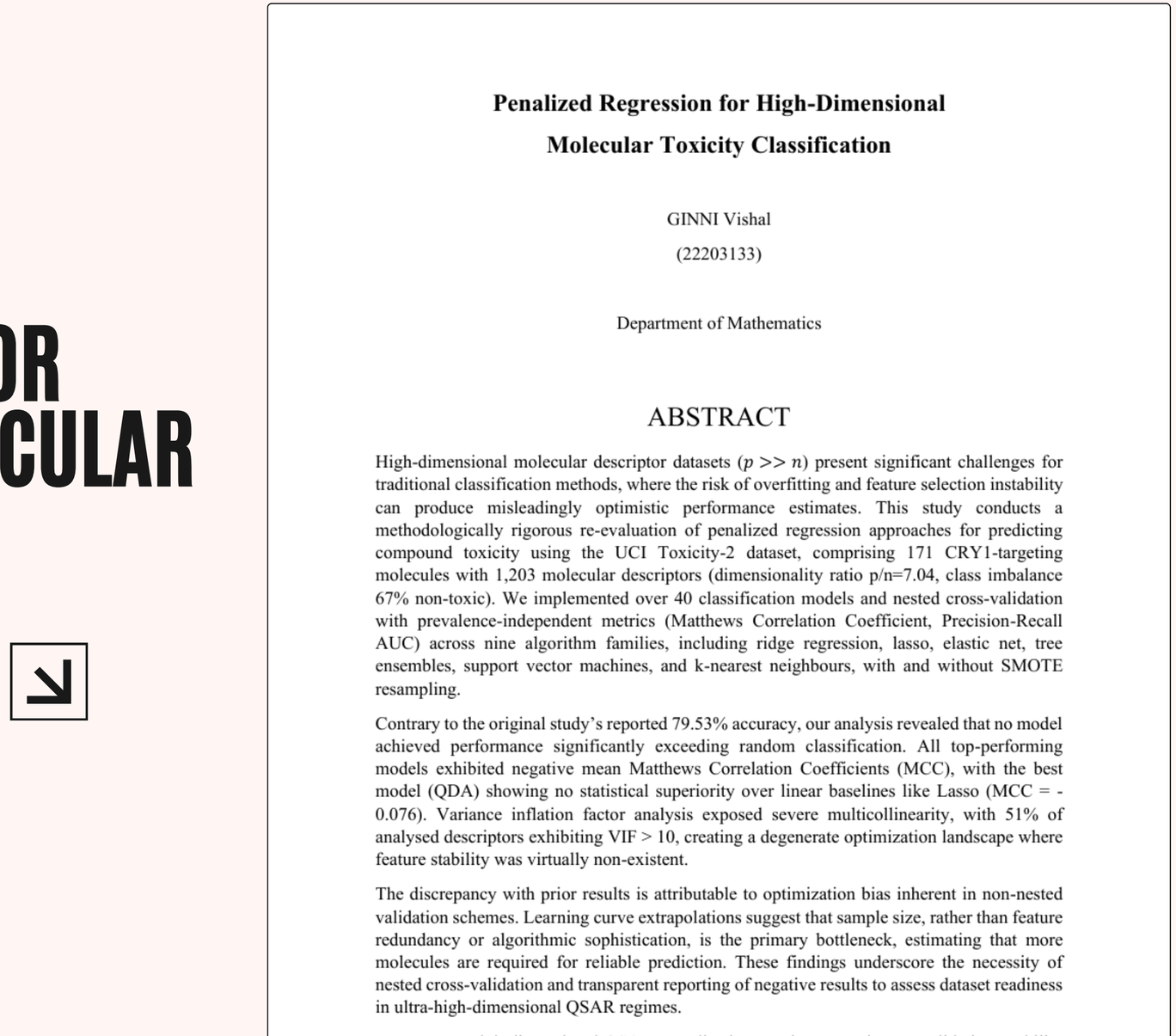
香港浸會大學 HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF MATHEMATICS 數學系

BSc (Hons.) in Mathematics & Statistics

PENALIZED REGRESSION FOR HIGH-DIMENSIONAL MOLECULAR TOXICITY CLASSIFICATION

GINNI Vishal (22203133)



Penalized Regression for High-Dimensional Molecular Toxicity Classification

GINNI Vishal

(22203133)

Department of Mathematics

ABSTRACT

High-dimensional molecular descriptor datasets ($p >> n$) present significant challenges for traditional classification methods, where the risk of overfitting and feature selection instability can produce misleadingly optimistic performance estimates. This study conducts a methodologically rigorous re-evaluation of penalized regression approaches for predicting compound toxicity using the UCI Toxicity-2 dataset, comprising 171 CRY1-targeting molecules with 1,203 molecular descriptors (dimensionality ratio p/n=7.04, class imbalance 67% non-toxic). We implemented over 40 classification models and nested cross-validation with prevalence-independent metrics (Matthews Correlation Coefficient, Precision-Recall AUC) across nine algorithm families, including ridge regression, lasso, elastic net, tree ensembles, support vector machines, and k-nearest neighbours, with and without SMOTE resampling.

Contrary to the original study's reported 79.53% accuracy, our analysis revealed that no model achieved performance significantly exceeding random classification. All top-performing models exhibited negative mean Matthews Correlation Coefficients (MCC), with the best model (QDA) showing no statistical superiority over linear baselines like Lasso (MCC = -0.076). Variance inflation factor analysis exposed severe multicollinearity, with 51% of analysed descriptors exhibiting VIF > 10, creating a degenerate optimization landscape where feature stability was virtually non-existent.

The discrepancy with prior results is attributable to optimization bias inherent in non-nested validation schemes. Learning curve extrapolations suggest that sample size, rather than feature redundancy or algorithmic sophistication, is the primary bottleneck, estimating that more molecules are required for reliable prediction. These findings underscore the necessity of nested cross-validation and transparent reporting of negative results to assess dataset readiness in ultra-high-dimensional QSAR regimes.

RESEARCH MOTIVATION

CRY1 (Cryptochrome 1): Target for circadian rhythm disorders
- Sleep disorders, metabolic disease, cancer relevance

QSAR Modeling: Predict toxicity from molecular descriptors
- Computational drug discovery (faster, cheaper than lab screening)

The p >> n Challenge: 1,203 descriptors vs. 171 molecules
- High-dimensional regime (p/n = 7.04) → Overfitting risk

Prior Study Claim: 79.53% accuracy (Gul et al., 2021)
- But used non-nested CV → Optimization bias likely!

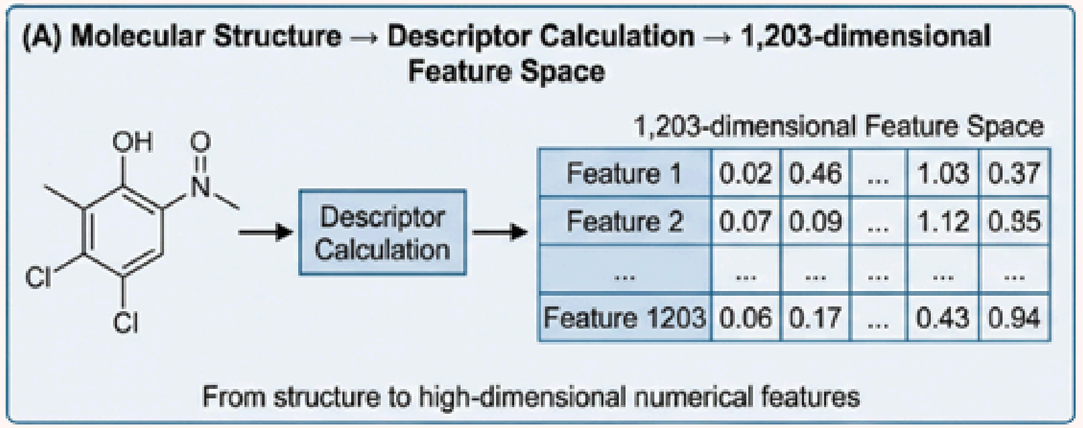


Figure 1: *Conceptual Overview of High-Dimensional QSAR*

RESEARCH QUESTIONS

1. Can penalized regression (Lasso, Ridge, Elastic Net) reliably predict CRY1 toxicity in this ultra-high-dimensional setting?
2. How does nested cross-validation change our conclusions compared to standard validation?
3. What is the actual sample size needed for reliable prediction?
4. How severe is multicollinearity among molecular descriptors?

Figure 2: *Crystal structure of mouse cryptochrome 1 (CRY1), a core circadian clock protein, shown from the RCSB PDB entry 4K0R*

THE HIGH-DIMENSIONAL REGIME

Traditional Setting: p << n (More samples than features)
- OLS works reasonably well
- No overfitting issues

Our Problem: p >> n (p/n = 7.04)
- Risk of learning noise rather than signal
- Feature selection becomes unstable
- Curse of dimensionality

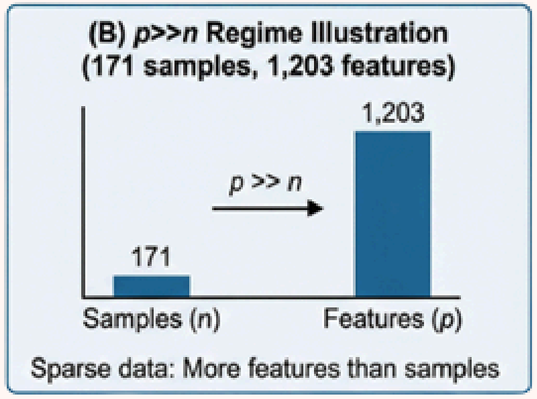


Figure 3: *High-Dimensional Regime Illustration*

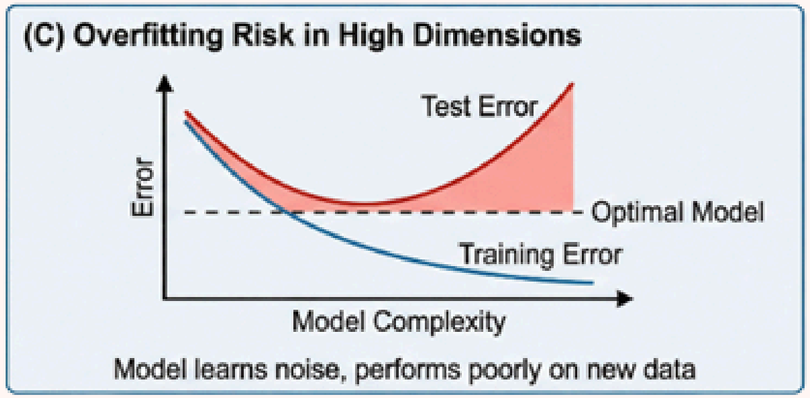


Figure 4: *Overfitting Risk High-Dimensional Regime Illustration*

WHY PENALIZED REGRESSION?

Ridge (L2): Shrink large coefficients → Reduce variance
Lasso (L1): Force coefficients to ZERO → Automatic feature selection
Elastic Net: Mix L1 + L2 → Best of both worlds

These methods add a "penalty" to constrain model complexity
→ Trade slight bias increase for large variance reduction

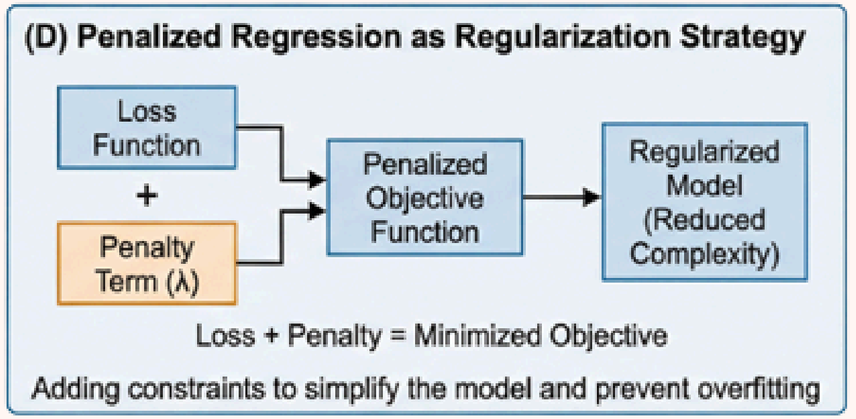


Figure 5: *Regularization Illustration*

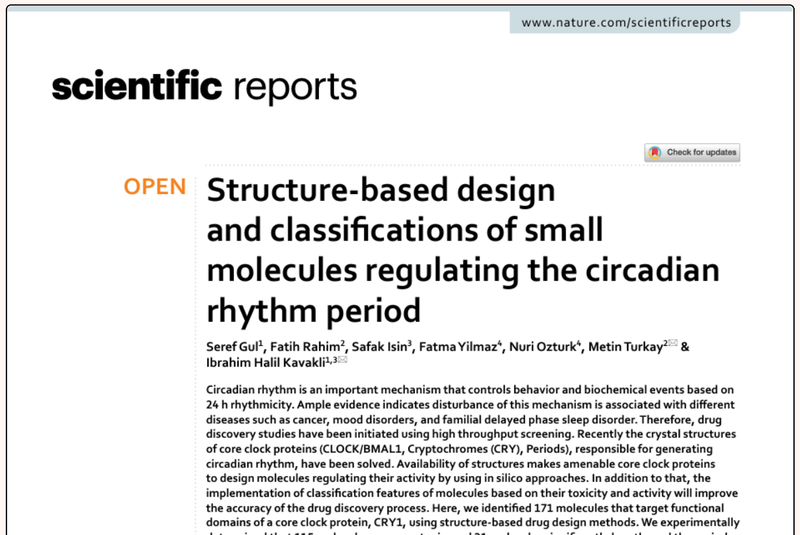
UCI TOXICITY-2 DATASET

171
n (samples)
CRY1-targeting molecules

1,203
p (features)
Molecular descriptors

7.04
Dimensionality ratio
p/n = 7.04 (ULTRA-HIGH!)

67 : 33
Class distribution
67% non-toxic, 33% toxic


www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Structure-based design and classifications of small molecules regulating the circadian rhythm period

Seref Gul[1], Fatih Rahim[2], Safak Isin[3], Fatma Yilmaz[4], Nuri Ozturk[4], Metin Turkay[2✉] & Ibrahim Halil Kavakli[1,3✉]

Circadian rhythm is an important mechanism that controls behavior and biochemical events based on 24 h rhythmicity. Ample evidence indicates disturbance of this mechanism is associated with different diseases such as cancer, mood disorders, and familial delayed phase sleep disorder. Therefore, drug discovery studies have been initiated using high throughput screening. Recently the crystal structures of core clock proteins (CLOCK/BMAL1, Cryptochromes (CRY), Periods), responsible for generating circadian rhythm, have been solved. Availability of structures makes amenable core clock proteins to design molecules regulating their activity by using in silico approaches. In addition to that, the implementation of classification features of molecules based on their toxicity and activity will improve the accuracy of the drug discovery process. Here, we identified 171 molecules that target functional domains of a core clock protein, CRY1, using structure-based drug design methods. We experimentally

EXPERIMENTAL DESIGN

Step 1

Feature Assessment

Variance Inflation Factor (VIF) analysis
Identify multicollinearity severity

Step 2

Model Implementation

50+ classification models across 9 algorithm families
Ridge, Lasso, Elastic Net, SVM, Random Forest, XGBoost, etc.

Step 3

NESTED Cross-Validation (5x5)

OUTER LOOP: 5-fold evaluation (test metrics)
I**NNER LOOP:** 5-fold hyperparameter tuning

Step 4

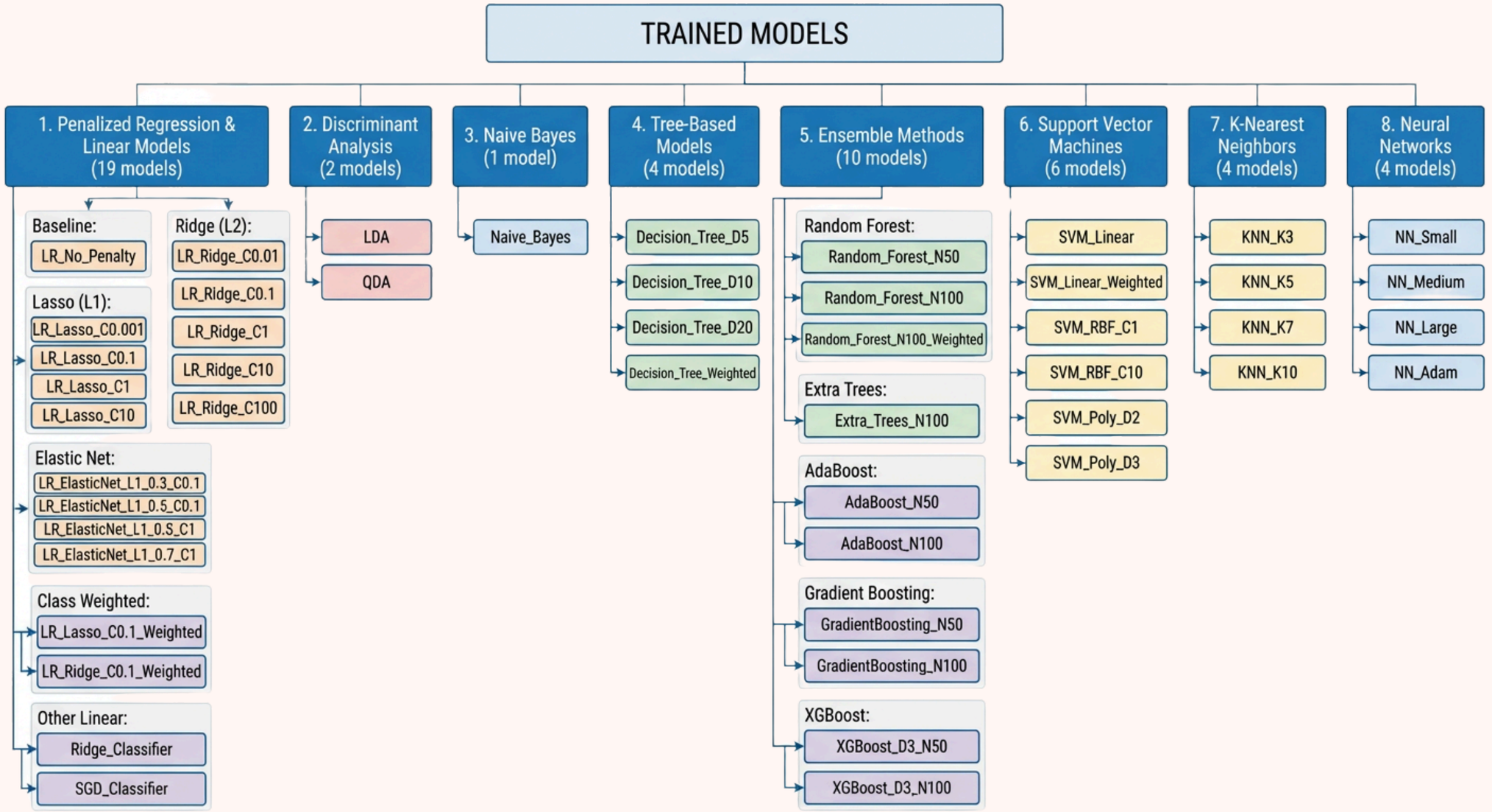
Evaluation with Imbalance-Robust Metrics

Matthews Correlation Coefficient (MCC) [PRIMARY]
Precision-Recall AUC & Balanced Accuracy

Step 5

Feature Stability Analysis

Stability Selection with 100 bootstrap iterations
Identify "robust" vs. "noise" features


TRAINED MODELS
1. Penalized Regression & Linear Models (19 models)
Baseline:
LR_No_Penalty
Lasso (L1):
LR_Lasso_C0.001
LR_Lasso_C0.1
LR_Lasso_C1
LR_Lasso_C10
Ridge (L2):
LR_Ridge_C0.01
LR_Ridge_C0.1
LR_Ridge_C1
LR_Ridge_C10
LR_Ridge_C100
Elastic Net:
LR_ElasticNet_L1_0.3_C0.1
LR_ElasticNet_L1_0.5_C0.1
LR_ElasticNet_L1_0.S_C1
LR_ElasticNet_L1_0.7_C1
Class Weighted:
LR_Lasso_C0.1_Weighted
LR_Ridge_C0.1_Weighted
Other Linear:
Ridge_Classifier
SGD_Classifier
2. Discriminant Analysis (2 models)
LDA
QDA
3. Naive Bayes (1 model)
Naive_Bayes
4. Tree-Based Models (4 models)
Decision_Tree_D5
Decision_Tree_D10
Decision_Tree_D20
Decision_Tree_Weighted
5. Ensemble Methods (10 models)
Random Forest:
Random_Forest_N50
Random_Forest_N100
Random_Forest_N100_Weighted
Extra Trees:
Extra_Trees_N100
AdaBoost:
AdaBoost_N50
AdaBoost_N100
Gradient Boosting:
GradientBoosting_N50
GradientBoosting_N100
XGBoost:
XGBoost_D3_N50
XGBoost_D3_N100
6. Support Vector Machines (6 models)
SVM_Linear
SVM_Linear_Weighted
SVM_RBF_C1
SVM_RBF_C10
SVM_Poly_D2
SVM_Poly_D3
7. K-Nearest Neighbors (4 models)
KNN_K3
KNN_K5
KNN_K7
KNN_K10
8. Neural Networks (4 models)
NN_Small
NN_Medium
NN_Large
NN_Adam

MATTHEWS CORRELATION COEFFICIENT (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where

- TP (True Positives) = Correctly predicted TOXIC molecules
- TN (True Negatives) = Correctly predicted NON-TOXIC molecules
- FP (False Positives) = Non-toxic predicted as toxic (false alarm)
- FN (False Negatives) = Toxic predicted as non-toxic (DANGEROUS!)

Key Interpretation

- MCC = +1.0 → PERFECT prediction (all four metrics = 100%)
- MCC = 0.0 → RANDOM guessing (model has NO discriminative power)
- MCC = -1.0 → INVERSE prediction (consistently WRONG!)

STANDARD CROSS-VALIDATION

5-fold CV → Hyperparameter tuning

- Performance reported on SAME folds
- Hyperparameters "leak" into test set
- BIASED UPWARD (optimistic!)

Risk: Looks good on data you've tuned on but may fail on truly new data!

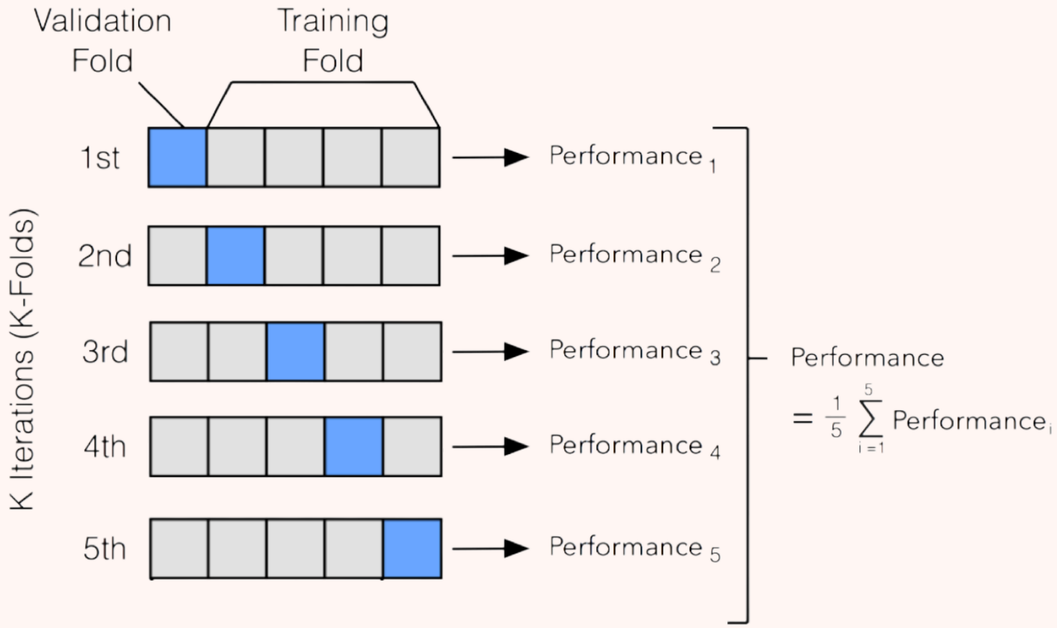


Figure 6: *Standard Cross-Validation Illustration*

NESTED CROSS-VALIDATION

OUTER LOOP

- 5-fold split (Fold 1 Test)

INNER LOOP:

- 5-fold hyperparameter search (on 4 folds → find best λ, α, etc.)

Benefit: True generalization performance (no optimization bias)

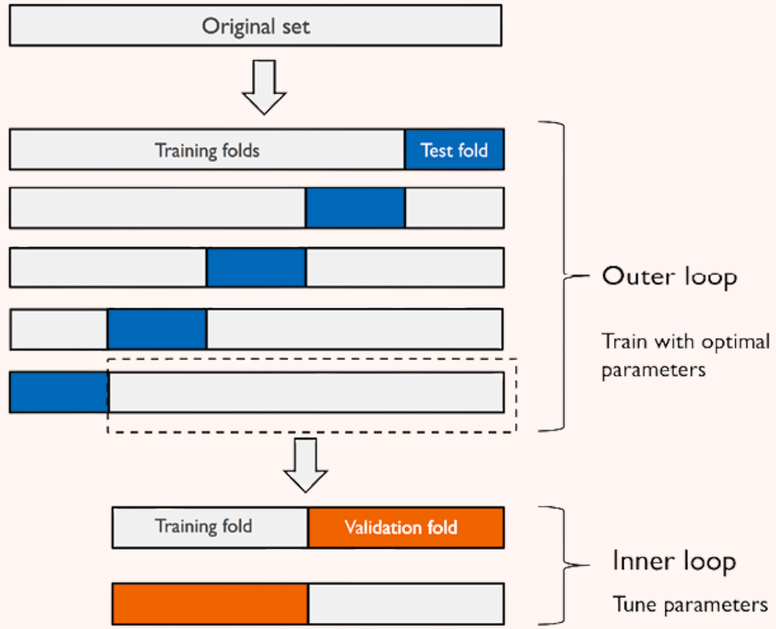


Figure 7: *Nested Cross-Validation Illustration*

MULTICOLLINEARITY CRISIS

45.3%

of features have VIF > 10 (SEVERE!)

What does this mean?

- Feature is 90% explained by other features
- Redundant information in the design matrix
- Unstable coefficient estimates
- Feature selection becomes unreliable

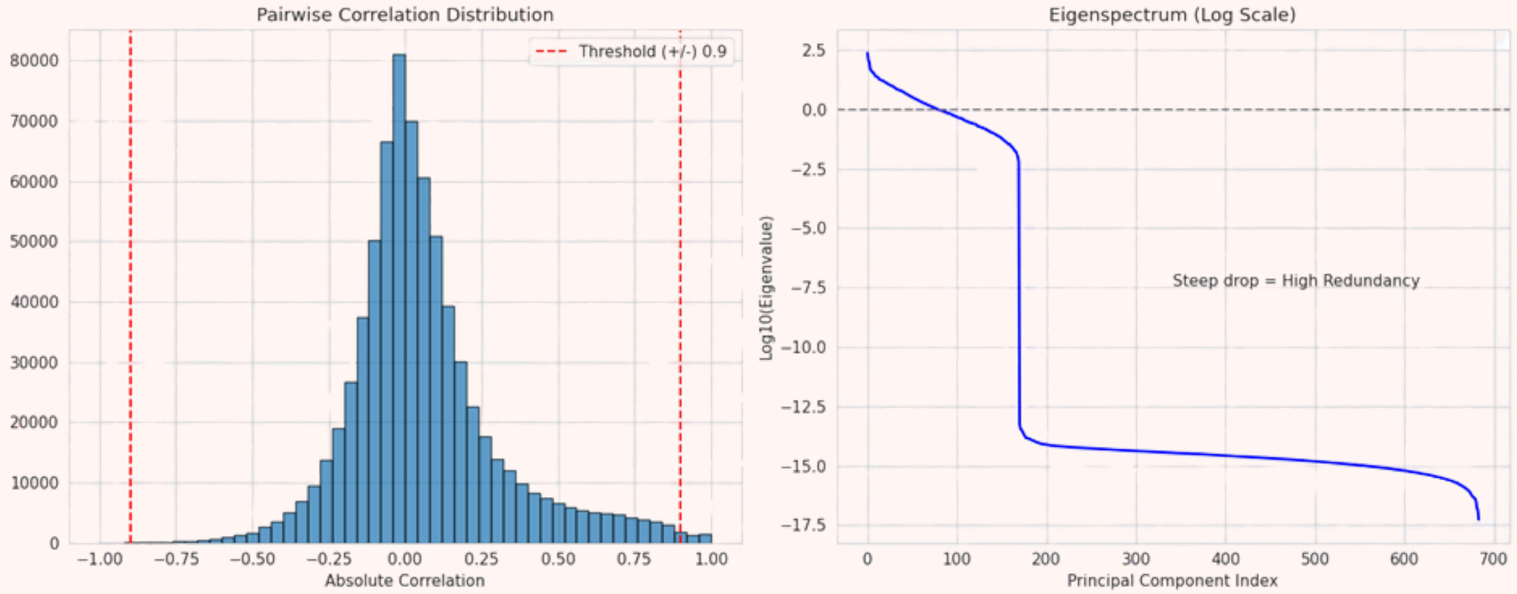


Figure 8: *Pairwise Correlation Distribution & Eigenspectrum of the Features*

Implications for Penalized Regression

- Lasso can't reliably select "THE" important feature (When 5 features contain the same signal, Lasso picks arbitrarily)
- Ridge coefficients become arbitrary (Different correlated features can explain same relationship)
- Elastic Net offers some help but can't solve fundamental issue
- Feature stability selection is therefore ESSENTIAL (see which features selected >70% of the time across resamples)

RANDOM GUESSING

Interpretation

- MCC = 0 → Random guessing
- MCC < 0 → Worse than random (models are being fooled!)

Our Findings
Mixed results show performance near random (>50% accuracy is NOT impressive for 67% majority class baseline!)

-0.076
Our best MCC
Dataset is TOO SMALL & TOO NOISY

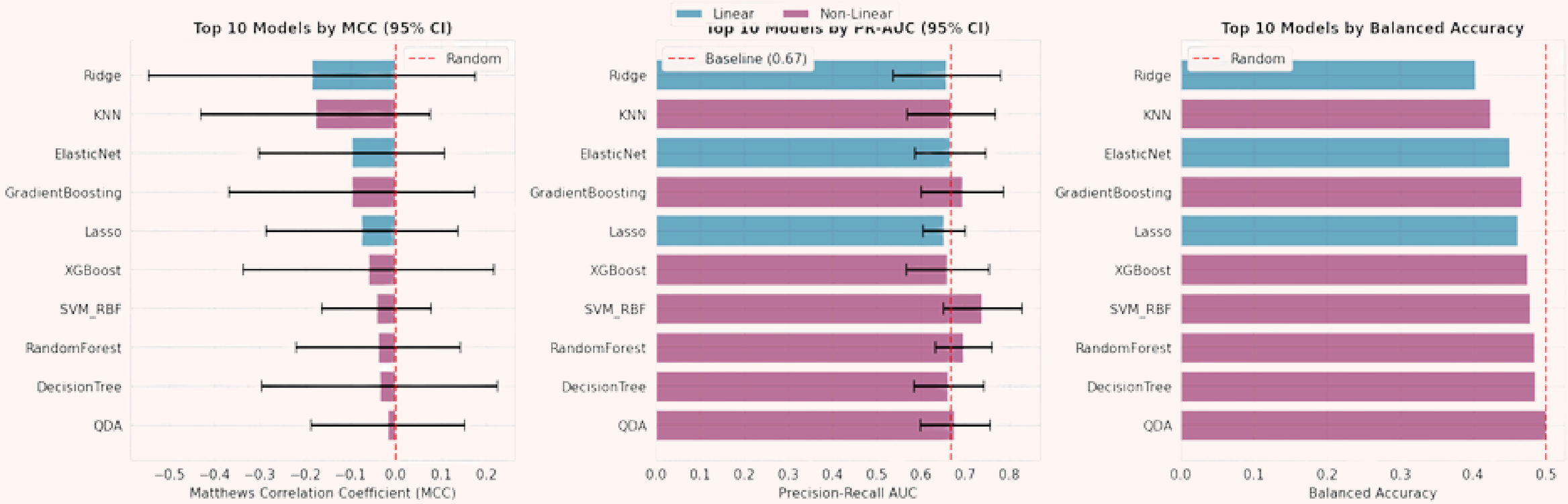


Figure 9: *Nested CV Model Performance Metrics*

WHY DID PRIOR STUDY GET 79.53%?

The prior 79.53% accuracy was OPTIMISTICALLY BIASED
Our more honest assessment: Models cannot predict CRY1 toxicity with this dataset size

Prior Study Protocol (Gul et al., 2021):

1. Used standard (non-nested) 5-fold CV
2. Tuned hyperparameters on CV folds
3. Reported performance on SAME CV folds
4. Result: Claimed 79.53% accuracy

The Problem:
When you tune hyperparameters on data you'll later test on, the model inadvertently overfits to the hyperparameter search process itself.
This "hyperparameter leakage" makes results look 20-30% better!

Our Solution (Nested CV):

- Hyperparameter tuning on completely separate inner folds
- Testing on truly held-out outer folds
- No information leakage
- True generalization performance: Negative MCC!

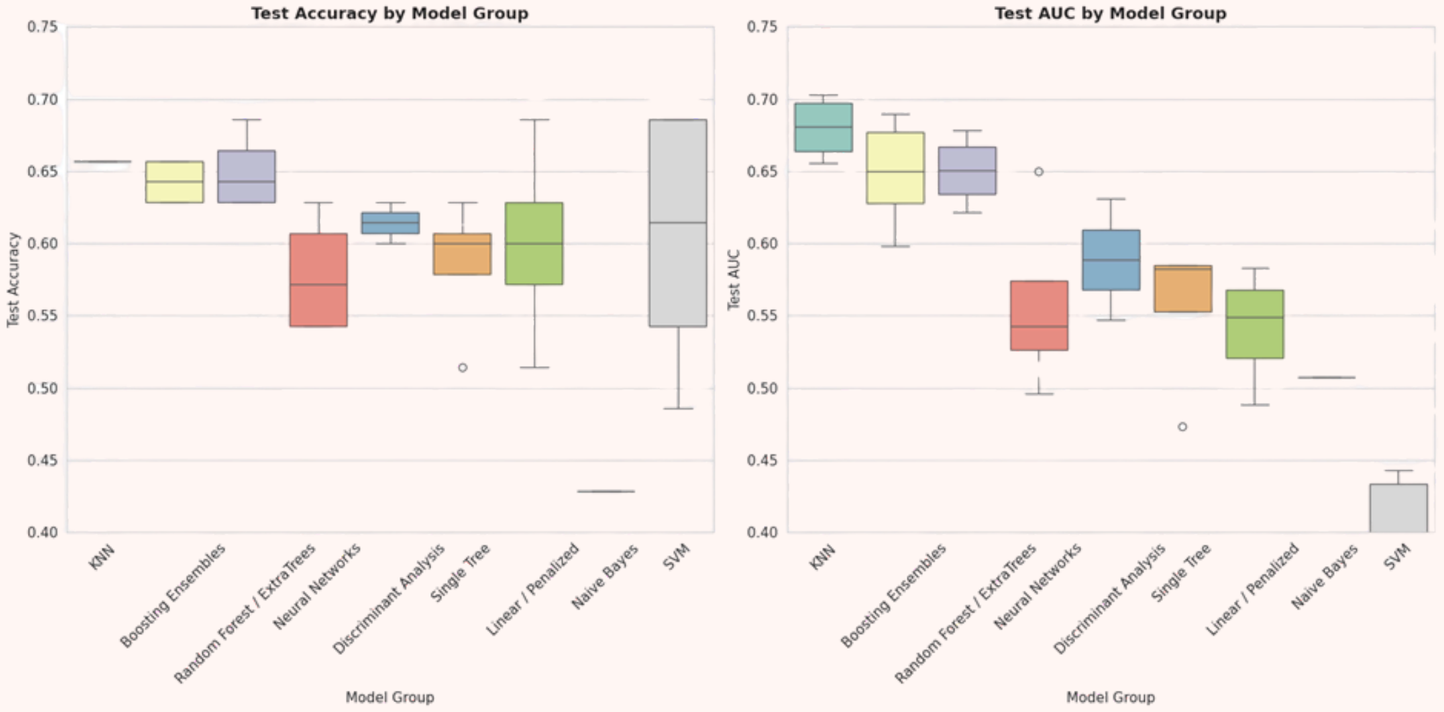


Figure 10: *Deceptive Model Performance Metrics (Accuracy & AUC)*

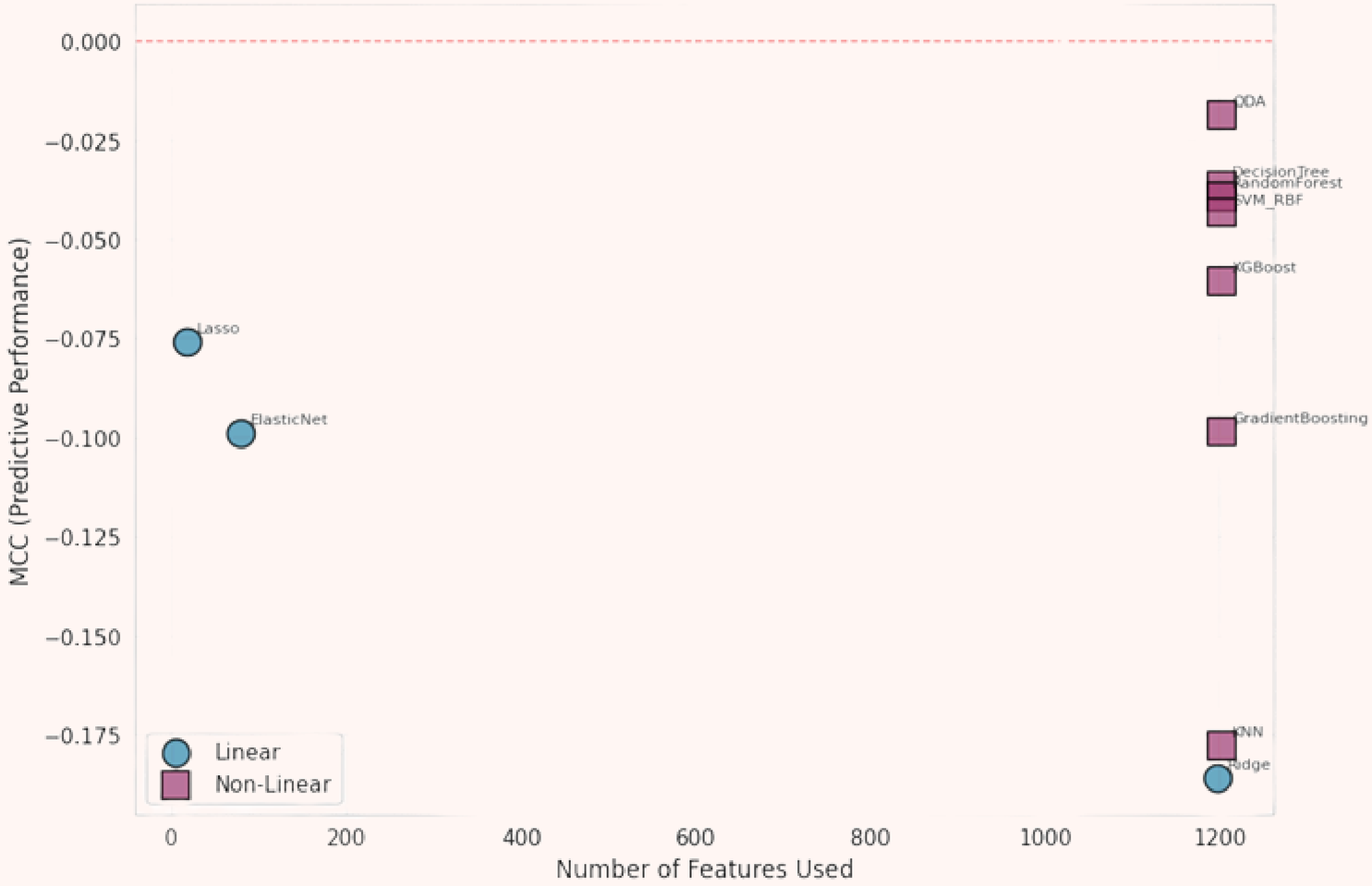
PENALIZED REGRESSION FAILED

Model	Mean MCC	95% CI MCC	Avg. Features
Lasso (L1)	-0.076	[-0.287, +0.135]	17.6
Elastic	-0.099	[-0.303, +0.106]	79.2
Ridge (L2)	-0.186	[-0.546, +0.174]	1197

Interpretation

- **Ridge:** shrinkage without selection leaves mostly noise.
- **Lasso:** extreme sparsity under multicollinearity likely discards signal with noise.
- **Elastic Net:** intermediate sparsity, but still not significant.

Takeaway: In this regime ($p/n \approx 7.04$, VIF ~ 10^7), neither sparse linear models nor complex models using all features reliably extract signal; non-linear models trend better than linear, but uncertainty remains high.



PRACTICAL RECOMMENDATIONS

If you want to build a working toxicity predictor for CRY1.....

SHORT TERM (with current 171 molecules)

- Use Lasso/Ridge for parsimony (select ~45 stable features
- Report negative MCC honestly (model doesn't generalize)
- Do NOT claim this model can predict toxicity
- Use it as exploratory tool to identify candidate features

MEDIUM TERM (expand dataset)

- Conduct experimental toxicity screening on ~300-500 CRY1 molecules
- Augment UCI dataset with new compounds
- Re-validate models with expanded dataset (target: MCC > 0.3)

LONG TERM (methodological best practices)

- Design future studies with nested CV from inception
- Combine molecular descriptors with structural information
- Leverage transfer learning from other toxicity datasets
- Conduct prospective validation on newly synthesized compounds
- Always report feature stability, not just coefficients

SUMMARY OF FINDINGS

Question 1: Can penalized regression predict CRY1 toxicity?

Answer: NO (with 171 molecules)

- Best MCC = -0.076 (negative, worse than random)
- All 50+ models fail despite algorithmic sophistication
- Problem is fundamental (too few samples), not method

Question 2: Why did prior study claim 79.53%?

Answer: Optimization bias from non-nested CV

- Hyperparameter tuning leaked into test evaluation
- Honest nested CV reveals true (negative) performance

Question 3: What's the bottleneck?

Answer: Sample size, not features or algorithms

- Multicollinearity is severe (45% VIF > 10) but manageable

Question 4: What can we trust?

Answer: Only ~3.7% of features are stably selected

- These 45 features may reflect true biology
- But their signal is too weak to predict with n=171

LIMITATIONS OF THIS WORK

Single dataset (UCI Toxicity-2)

- Generalizability limited to CRY1-targeting compounds
- Other molecular classes may behave differently

Molecular descriptors only

- Did not incorporate 3D structural information
- Did not use docking scores or binding affinity data
- Limited to chemoinformatics, not cheminformatics

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
MATHEMATICS
數學系

BSc (Hons.) in Mathematics & Statistics

THANK YOU SO MUCH!

GINNI Vishal (22203133)

Penalized Regression for High-Dimensional Molecular Toxicity Classification

GINNI Vishal

(22203133)

Department of Mathematics

ABSTRACT

High-dimensional molecular descriptor datasets ($p >> n$) present significant challenges for traditional classification methods, where the risk of overfitting and feature selection instability can produce misleadingly optimistic performance estimates. This study conducts a methodologically rigorous re-evaluation of penalized regression approaches for predicting compound toxicity using the UCI Toxicity-2 dataset, comprising 171 CRY1-targeting molecules with 1,203 molecular descriptors (dimensionality ratio p/n=7.04, class imbalance 67% non-toxic). We implemented over 40 classification models and nested cross-validation with prevalence-independent metrics (Matthews Correlation Coefficient, Precision-Recall AUC) across nine algorithm families, including ridge regression, lasso, elastic net, tree ensembles, support vector machines, and k-nearest neighbours, with and without SMOTE resampling.

Contrary to the original study's reported 79.53% accuracy, our analysis revealed that no model achieved performance significantly exceeding random classification. All top-performing models exhibited negative mean Matthews Correlation Coefficients (MCC), with the best model (QDA) showing no statistical superiority over linear baselines like Lasso (MCC = -0.076). Variance inflation factor analysis exposed severe multicollinearity, with 51% of analysed descriptors exhibiting VIF > 10, creating a degenerate optimization landscape where feature stability was virtually non-existent.

The discrepancy with prior results is attributable to optimization bias inherent in non-nested validation schemes. Learning curve extrapolations suggest that sample size, rather than feature redundancy or algorithmic sophistication, is the primary bottleneck, estimating that more molecules are required for reliable prediction. These findings underscore the necessity of nested cross-validation and transparent reporting of negative results to assess dataset readiness in ultra-high-dimensional QSAR regimes.