PENALIZED REGRESSION FOR HIGH-DIMENSIONAL MOLECULAR TOXICITY CLASSIFICATION

by

GINNI Vishal

(22203133)

A thesis submitted in partial fulfilment of the requirements

for the degree of

Bachelor of Science (Honours)

in Mathematics and Statistics

at

Hong Kong Baptist University

23 December 2025

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Dr. Wu Tian for her invaluable guidance, patience, and expertise throughout the course of this research project. Her thoughtful feedback was instrumental at every stage, particularly during the initial phase of study design, where her advice on dataset suitability and the distinction between time-series and cross-sectional analysis helped ground the research in the appropriate methodology.

I am especially thankful for her technical insights regarding high-dimensional settings, which guided my approach to feature selection, class imbalance handling, and the interpretation of redundant features in penalized regression models. Her encouragement to rigorously benchmark a wide array of linear and non-linear models, combined with her constructive suggestions on visualization techniques, significantly improved the comprehensiveness and clarity of my findings. All experiments and analyses described in this thesis were conducted under her supervision, and her support has been a cornerstone of this achievement.

Signature of Student

GINNI Vishal

Student Name

Department of Mathematics

Hong Kong Baptist University

Date: 23/12/2025

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

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.

Keywords: High-dimensional QSAR, penalized regression, nested cross-validation, stability selection, Matthews correlation coefficient, class imbalance, multicollinearity, CRY1 toxicity, molecular descriptors, negative results

INTRODUCTION

BACKGROUND & MOTIVATION

The development of safe and effective therapeutic compounds requires rigorous assessment of molecular toxicity during early-stage drug discovery. Among the numerous biological targets under investigation for circadian rhythm disorders, Cryptochrome 1 (CRY1) has emerged as a promising candidate due to its central role in regulating the mammalian circadian clock (Gul et al., 2021). Disruption of circadian rhythms has been implicated in diverse pathological conditions, including metabolic disorders, mood disorders, familial delayed sleep phase disorder, and certain cancers, making the identification of non-toxic CRY1-modulating compounds a matter of considerable clinical importance.

Traditional experimental toxicity screening via high-throughput assays, while effective, is resource-intensive and time-consuming. Computational approaches based on Quantitative Structure-Activity Relationship (QSAR) modeling offer a complementary strategy, leveraging machine learning algorithms to predict toxicity from molecular descriptors, mathematical representations encoding structural, electronic, and topological properties of chemical compounds (Huang et al., 2021, 2022; Mao et al., 2021). However, QSAR datasets frequently exhibit a critical statistical challenge: the number of molecular descriptors (p) often vastly exceeds the number of available compounds (n), creating what is termed a high-dimensional or p>>n regime. In such settings, traditional statistical methods become unreliable due to overfitting, multicollinearity, and the curse of dimensionality.

Recent advances in regularized regression techniques, particularly ridge regression (L2 penalty), lasso regression (L1 penalty), and elastic net (combined L1+L2 penalties), have demonstrated promise in high-dimensional QSAR by imposing constraints that enable simultaneous feature selection and model fitting (Nwaeme & Lukman, 2023). These penalized methods address overfitting by shrinking coefficient estimates toward zero, with lasso uniquely capable of producing sparse models by driving irrelevant feature coefficients to exactly zero. Despite their theoretical appeal, the practical application of penalized regression to ultra-high-dimensional molecular datasets remains fraught with methodological challenges, including feature selection instability, optimization bias from improper validation schemes, and the pervasive issue of extreme multicollinearity among molecular descriptors.

The UCI Toxicity-2 dataset, derived from a 2021 study by Gul et al. (2021), provides an exemplary case study for investigating these challenges. This study re-evaluates the dataset, which comprises 171 small molecules and 1,203 molecular descriptors. While the original study reported a mean accuracy of 79.53% via tenfold cross-validation with 100 repetitions. While this result appears promising, the validation methodology raises critical questions about optimization bias: the repeated use of cross-validation performance to guide feature selection, followed by reporting that same cross-validation performance as the generalization error, constitutes a form of circular validation known to severely overestimate performance in high-dimensional settings (Krawczuk & Lukaszuk, 2016).

Moreover, the original study's reliance on accuracy as the primary evaluation metric is problematic given the substantial class imbalance (67% non-toxic, 33% toxic). According to Thölke at al. (2023), in imbalanced classification tasks, accuracy can be misleadingly high

when models fail to identify the minority class, in this case, toxic compounds, which are precisely the instances of greatest clinical concern. A model that trivially predicts all molecules as non-toxic would achieve 67% accuracy, rendering simple accuracy comparisons insufficient for assessing true discriminative capacity.

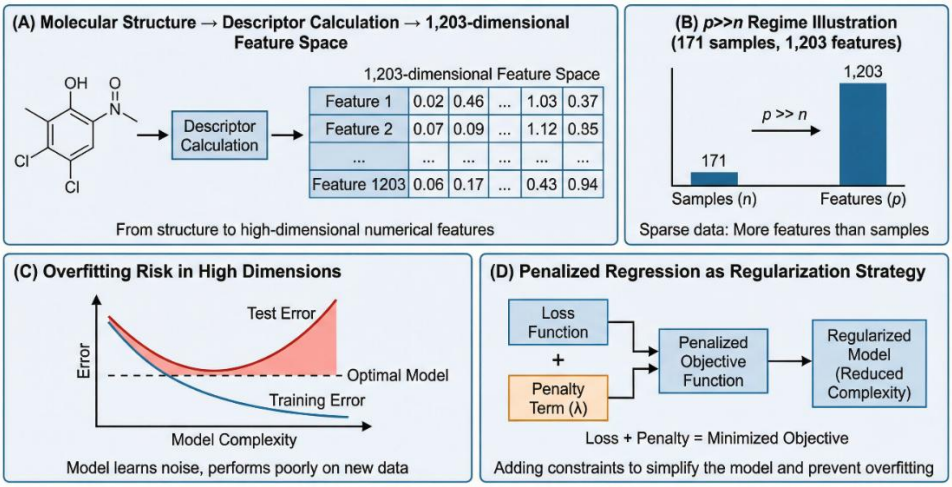

Figure 1. Conceptual Overview of High-Dimensional QSAR Challenge

The intersection of these methodological concerns, high dimensionality ($p/n = 7.04$), extreme multicollinearity among descriptors, class imbalance, and potentially biased validation, motivates a rigorous re-evaluation of penalized regression approaches for this dataset. Such an analysis requires three critical enhancements over prior work:

(1) Nested cross-validation to separate hyperparameter optimization from performance estimation, thereby eliminating optimization bias.
(2) Prevalence-independent evaluation metrics such as Matthews Correlation Coefficient (MCC) and Precision-Recall Area Under the Curve (PR-AUC) that account for class imbalance; and
(3) Explicit quantification of feature selection stability and multicollinearity to assess the reliability of identified molecular descriptors for chemical interpretation.

RESEARCH OBJECTIVES

This thesis undertakes a comprehensive and methodologically rigorous evaluation of penalized regression methods for molecular toxicity prediction in an ultra-high-dimensional setting, with specific focus on addressing the statistical and interpretative challenges inherent in the CRY1 Toxicity-2 dataset. The research is structured around three primary objectives and two secondary objectives:

PRIMARY OBJECTIVES

Unbiased Performance Evaluation via Nested Cross-Validation

To quantify the true generalization performance of ridge regression, lasso regression, and elastic net models using nested cross-validation (5×5 folds), which separates hyperparameter tuning (inner loop) from performance estimation (outer loop), thereby providing unbiased estimates free from optimization bias. Performance will be assessed using prevalence-independent metrics, Matthews Correlation Coefficient (MCC), Precision-Recall AUC (PR-AUC), and Balanced Accuracy, to account for the 67% vs. 33% class imbalance. This objective

addresses the critical methodological limitation of the original study's non-nested validation scheme and provides a statistically defensible benchmark for model performance.

Comprehensive Model Comparison Across Algorithm Families

To benchmark penalized regression methods against a diverse range of alternative machine learning approaches, including tree-based ensembles (Random Forest, XGBoost, Gradient Boosting), support vector machines (linear, RBF, polynomial kernels), distance-based methods (k-nearest neighbors, quadratic discriminant analysis), and neural networks. This comparative analysis will quantify the interpretability-performance trade-off: specifically, whether the sparsity and linear interpretability of penalized models justify potential predictive performance sacrifices relative to complex "black-box" algorithms. The comparison will be conducted under identical nested cross-validation protocols to ensure fairness.

Feature Selection Stability and Multicollinearity Assessment

To rigorously quantify the stability of feature selection across 100 bootstrap resamples using the stability selection framework, identifying molecular descriptors that exceed conventional reliability thresholds (selection frequency $\pi \geq 0.70$). Concurrently, variance inflation factors (VIF) will be calculated for all 1,203 descriptors to quantify the severity of multicollinearity and its impact on feature selection reliability. This objective directly addresses the interpretability question: can penalized regression methods identify a robust, chemically meaningful subset of descriptors for CRY1 toxicity prediction, or does extreme multicollinearity render feature selection fundamentally arbitrary?

Secondary Objectives

Impact of Class Balancing Techniques

To evaluate the impact of class balancing techniques, specifically Synthetic Minority Oversampling Technique (SMOTE) and class weighting, on model performance in high-dimensional space. Results will inform best practices for addressing class imbalance in ultra-high-dimensional chemoinformatics applications

Reconciliation with Original Study Findings

To investigate the substantial discrepancy between the original study's reported 79.53% accuracy (via RFE-DTC with 13 features) and anticipated lower performance under rigorous nested validation. This objective includes:

(a) Comparing the original study's 13-feature set to features identified via stability selection
(b) Assessing multicollinearity within the original feature set & the full 1,203-descriptor space
(c) Assessing optimization bias regarding non-nested validation schemes

The goal is not to invalidate prior work but to provide a transparent, methodologically grounded explanation for differing results.

Table 1. Research Objectives and Corresponding Methodological Approaches

Objective	Key Methodology	Primary Output Metric
1. Unbiased Performance	Nested 5×5 CV	MCC with 95% CI
2. Algorithm Comparison	9 model families tested	Performance Benchmarks
3. Feature Stability	100 bootstrap iterations	Selection frequency (π) distribution
4. Class Imbalance	SMOTE & Class Weighting	Performance Delta
5. Study Reconciliation	VIF analysis, feature overlap	Explanation of Discrepancy

These objectives collectively aim to establish a methodologically sound benchmark for high-dimensional QSAR toxicity prediction while providing diagnostic insights into when and why penalized regression methods succeed or fail in extreme p>>n regimes.

SIGNIFICANCE OF THE STUDY

This research makes substantive contributions to three intersecting domains: computational toxicology, statistical machine learning methodology, and the broader practice of QSAR modeling in drug discovery.

METHODOLOGICAL ADVANCEMENT IN HIGH-DIMENSIONAL QSAR

The pervasive use of non-nested cross-validation in QSAR literature has been identified as a major source of optimistically biased performance estimates, yet rigorous empirical demonstrations of the magnitude of this bias remain scarce. By contrasting nested and non-nested validation on the same dataset, this study provides concrete evidence of optimization bias in a real-world chemoinformatics context. The introduction of stability selection to QSAR feature selection, a technique widely adopted in genomics but underutilized in molecular modelling, offers a principled framework for distinguishing stable, interpretable feature sets from arbitrarily selected descriptor subsets. Furthermore, the systematic quantification of multicollinearity via variance inflation factors establishes a diagnostic criterion for assessing whether a dataset's dimensionality structure permits reliable feature interpretation. These methodological contributions extend beyond the CRY1 toxicity problem, providing generalizable protocols for evaluating any high-dimensional QSAR dataset.

TRANSPARENT REPORTING OF NEGATIVE RESULTS

A substantial fraction of machine learning research in drug discovery reports only favorable outcomes, creating publication bias that distorts the literature's collective understanding of method effectiveness. This thesis adopts an alternative paradigm: transparent documentation of model failures and rigorous diagnosis of their root causes. Should the results demonstrate poor predictive performance (MCC ≈ 0) and feature selection instability (low selection frequencies), these findings carry significant informational value. They would establish that the p/n=7.04 regime represents a fundamental statistical barrier beyond which algorithmic sophistication cannot compensate for sample size inadequacy, a conclusion with immediate implications for resource allocation in computational drug discovery. Negative results, when rigorously characterized, guide future research by identifying when data collection should precede algorithmic refinement, thereby preventing wasted effort on methodological optimization in statistically intractable settings.

METHODOLOGY IMPLICATIONS

Circadian rhythm disruption affects millions of individuals worldwide, with conditions ranging from shift-work sleep disorder to depression linked to CRY1 variants. The development of nontoxic, circadian-modulating therapeutics depends critically on accurate early-stage toxicity prediction. However, contrary to prior claims, no stable predictors were identified in this study. This finding serves as a methodological warning against using this dataset for predictive modeling without significantly larger sample sizes. The analysis quantifies the sample size requirements (via learning curve extrapolation) necessary to achieve clinically actionable predictive performance, providing actionable guidance for pharmaceutical researchers to prioritize data collection over algorithmic refinement.

Establishing Diagnostic Criteria for Dataset "Readiness"

A persistent challenge in applied machine learning is determining whether a given dataset is amenable to predictive modeling or whether fundamental constraints (small sample size, extreme multicollinearity, class imbalance) preclude success regardless of methodological sophistication. This study proposes a diagnostic framework comprising three readiness criteria:

(1) p/n ratio: Is the dimensionality manageable relative to sample size?
(2) Multicollinearity severity: What proportion of features exhibit VIF>10, indicating redundancy?
(3) Baseline metric comparison: Does the best model exceed a majority-class predictor by a statistically significant margin?

By operationalizing these criteria through quantitative thresholds derived from the CRY1 dataset, this research provides a template for prospective dataset evaluation, potentially preventing premature modeling attempts on statistically intractable problems.

The Gap Between Statistical Theory and Chemoinformatics Practice

While statistical learning theory clearly articulates the challenges of high-dimensional estimation, the bias-variance tradeoff, regularization necessity, and cross-validation principles, these concepts often remain abstract in the chemoinformatics literature. This thesis serves a pedagogical function by concretely demonstrating how theoretical principles manifest in a real molecular dataset: multicollinearity creates "flat" optimization landscapes, imbalanced classes distort accuracy metrics, and non-nested validation inflates performance estimates. By connecting theory to practice through detailed empirical analysis, this work aims to elevate methodological standards in computational toxicology and encourage wider adoption of rigorous validation protocols.

Contribution to Reproducible Research Practices

All analysis code, including nested cross-validation, hyperparameter grids, and stability selection, will be made publicly available. This commitment to computational reproducibility aligns with emerging standards in machine learning research and ensures that the study's findings can be independently verified, extended to alternative datasets, or incorporated into educational materials for teaching high-dimensional statistical methods.

In summary, this thesis addresses a critical gap in the QSAR literature by subjecting penalized regression methods to the most rigorous validation standards available, while simultaneously providing diagnostic tools and negative-result documentation that advance the field's collective methodological maturity. Whether the results demonstrate predictive success or fundamental statistical limitations, the insights generated will inform best practices for high-dimensional molecular toxicity prediction and establish realistic benchmarks for future CRY1-targeting drug discovery efforts.

LITERATURE REVIEW & THEORY

MOLECULAR TOXICITY PREDICTION & DRUG DISCOVERY

Quantitative Structure-Activity Relationship (QSAR) modeling predicts biological activity from molecular structure through mathematical relationships between chemical descriptors and target endpoints, based on the foundational principle that structurally similar molecules exhibit similar biological properties (Peter et al., 2019). Modern QSAR implementations encode molecular structures as numerical feature vectors comprising topological, constitutional, electronic, and geometric descriptors. Contemporary descriptor calculation tools such as PaDEL-Descriptor can generate 1,500-4,000 descriptors per molecule, encompassing E-state indices (electronic and topological properties), autocorrelation descriptors (spatial property distributions), topological descriptors (molecular connectivity), constitutional descriptors (atom/bond counts), and Burden modified eigenvalues (weighted adjacency matrix properties). This descriptor proliferation creates the high-dimensional challenge central to modern QSAR, where datasets frequently contain far more descriptors than molecules (p>>n), necessitating careful statistical methodology to avoid spurious correlations and overfitting.

Toxicity prediction during pharmaceutical development is critical, as toxicity is a leading cause of drug attrition, with ~20–30% of clinical failures and >30% of discarded candidates due to safety issues (Vo et al., 2020). Computational toxicology allows rapid, low-cost screening of large chemical inventories, often tens of thousands of structures, to prioritize substances and fill data gaps while reducing animal use (Gadaleta et al., 2019; Forest, 2022). However, cytotoxicity prediction remains particularly challenging due to cell-type specificity and assay-dependent outcomes (Sun et al., 2020), with reported accuracies ranging widely from 60-85% depending on dataset characteristics and validation rigor.

The mammalian circadian clock, orchestrated by core proteins including Cryptochrome 1 (CRY1), regulates rhythmic physiological processes through transcriptional-translational feedback loops. CRY1 functions as an essential transcriptional repressor by heterodimerizing with PERIOD proteins and suppressing CLOCK-BMAL1-mediated transcription, establishing the ~24-hour periodicity fundamental to sleep-wake cycles, hormone secretion, and metabolic regulation (Michael at al., 2017; Patke et al., 2017). Genetic variants of CRY1 have been causally linked to familial delayed sleep phase disorder, mood disorders, metabolic syndrome, and cancer through disrupted cell cycle regulation. These clinical associations position CRY1 as a compelling therapeutic target, though therapeutic development hinges critically on identifying non-toxic CRY1-targeting compounds, as unintended cytotoxicity would preclude clinical translation regardless of circadian efficacy.

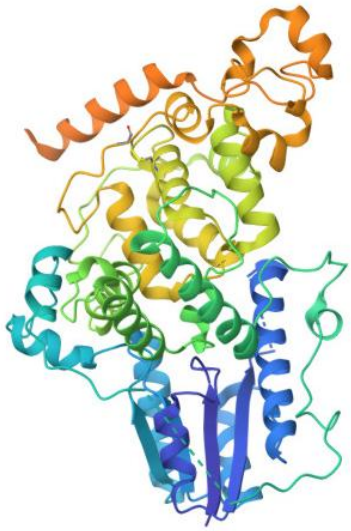

Figure 2. Crystal structure of mouse cryptochrome 1 (CRY1), a core circadian clock protein, shown from the RCSB PDB entry 4K0R

THE TOXICITY-2 DATASET AND ORIGINAL STUDY

The UCI Toxicity-2 dataset originates from Gul et al.'s (2021) Scientific Reports study on structure-based design of circadian rhythm modulators. The original study employed Recursive Feature Elimination (RFE) with Decision Tree Classifiers (DTC), evaluating feature sets from 2 to 20 descriptors using four classifier types (DTC, RFC, ETC, XGBC) with tenfold cross-validation repeated 100 times. Class imbalance was addressed through inverse-proportional weighting (toxic weight=1.53, non-toxic weight=0.74). DTC achieved optimal performance with 13 features, reporting mean accuracy $79.53\% \pm 2.01\%$ and maximum accuracy 84.21% . The 13 identified descriptors represent diverse physicochemical domains: MDEC-23 (topological branching), MATS2v (van der Waals volume distribution), CrippenMR (molar refractivity), C1SP2 (unsaturated carbon count), Burden eigenvalues (SpMax7_Bhe, SpMin1_Bhs, SpMax5_Bhv), and autocorrelations (GATS8e, GATS8s, ATSC8s).

Critically, the original study did not assess multicollinearity via variance inflation factors, evaluate feature selection stability across CV repetitions, or employ nested cross-validation to prevent optimization bias, omissions that raise questions about the reproducibility and unbiased nature of the reported 79.53% accuracy

MACHINE LEARNING FOR TOXICITY PREDICTION

Penalized regression addresses the p>>n challenge by augmenting the objective function with regularization penalties (James et al., 2023). Ridge regression shrinks coefficients toward zero without inducing sparsity, distributing weights among correlated predictors but retaining all features. Lasso regression drives coefficients exactly to zero, performing embedded feature selection ideal for interpretability but exhibiting critical instability under multicollinearity, when features are highly correlated, lasso arbitrarily selects one and discards others, with selection dependent on minor numerical perturbations. Elastic net combines L1 and L2 penalties, with the L2 component encouraging grouped selection that includes or excludes correlated features together, theoretically superior for high-dimensional datasets with extensive correlations

Non-linear algorithms capture complex interactions unavailable to linear methods. Random Forests train ensembles of decision trees on bootstrap samples with random feature subsets, providing resistance to overfitting and implicit feature selection via mean decrease in impurity (Akhiat et al., 2021). Gradient Boosting methods (XGBoost, LightGBM) iteratively train shallow trees to correct residual errors, achieving state-of-the-art performance on tabular datasets with native class imbalance handling (Boldini et al., 2023). Support Vector Machines with RBF or polynomial kernels map descriptors into high-dimensional spaces where linear separability improves, though performance is highly sensitive to hyperparameter tuning and computational cost scales poorly beyond n~1,000 (Cortes & Vapnik, 1995).

Class imbalance is addressed via class weighting, which upweights minority misclassification costs, and SMOTE (Synthetic Minority Oversampling Technique), which generates synthetic minority samples via nearest-neighbor interpolation (Elreedy & Atiya., 2019). While SMOTE reports minority-class recall improvements across diverse domains, it becomes unreliable in high-dimensional settings. When p>>n, nearest-neighbor relationships become unstable due to the curse of dimensionality, pairwise distances become approximately equal, rendering "nearest" neighbors statistically meaningless (Blagus & Lusa, 2013). Synthetic samples may

fall into unrepresentative regions, potentially introducing noise for distance-based methods (k-NN) while potentially benefiting tree-based methods through additional decision boundary refinement opportunities.

METHODOLOGICAL CONSIDERATIONS IN P>>N SETTINGS

Standard k-fold CV introduces optimization bias when the same folds are used for both hyperparameter tuning and performance evaluation, hyperparameters are chosen to maximize performance on those specific folds, leading to overly optimistic estimates that increase with dimensionality (Krawczuk & Lukaszuk, 2016; Tsamardinos et al., 2017). Monte Carlo simulations with null data (no true signal) show that standard k-fold CV can yield AUC values far above 0.5 simply through this selection effect, i.e., models that perform at chance in new data appear good under k-fold CV (Moshontz et al., 2020). Nested CV employs two loops: an outer loop for unbiased performance estimation and an inner loop for hyperparameter tuning within each outer training fold (Krstajic et al., 2014; Vabalas et al., 2019), ensuring test data never influences hyperparameter selection. While this structure is computationally expensive (5×5=25 models per configuration), it is essential in p>>n regimes where overfitting risk is acute.

Accuracy fails catastrophically under class imbalance, for Toxicity-2's 67% non-toxic prevalence, predicting all molecules as non-toxic achieves 67% accuracy without learning toxicity patterns. Matthews Correlation Coefficient (MCC), ranging from -1 (total disagreement) to +1 (perfect prediction) with 0 representing random performance, is invariant to class prevalence and derived from all four confusion matrix entries:

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

Precision-Recall AUC (PR-AUC) focuses exclusively on minority-class identification ($precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$), with random performance at the minority-class prevalence (0.33) and perfect performance at 1.0, providing more honest assessment than ROC-AUC, which overestimates performance due to true negative inflation (Movahedi et al., 2020; Saito & Rehmsmeier, 2015).

$$Balanced\ accuracy\ = \frac{sensitivity\ +\ specificity}{2}$$

explicitly averages class-specific recall rates, reaching 0.5 for random classifiers regardless of prevalence.

Feature selection instability, where minor data perturbations produce different feature subsets, undermines interpretability in p>>n settings with correlated predictors (Huang, 2021). Stability selection quantifies robustness via bootstrap resampling: generate B subsamples (typically 50-100) of size $\lfloor n/2 \rfloor$, apply feature selection to each, calculate selection frequency $\hat{\pi}_j$ for each feature (proportion of subsamples selecting it), and define stable set as

$$\left\{ j : \hat{\pi}_j \geq \pi_{thr} \right\}, \qquad \pi_{thr} \in [0.6, 0.8]$$

Features with $\hat{\pi}_j \geq 0.70$ are considered robustly selected, not artifacts of random sampling.

Variance Inflation Factors (VIF) quantify multicollinearity severity:

$$\mathrm{VIF}_j = \frac{1}{1 - {R_j}^2}$$

, where ${R_j}^2$ is the R^2 from regressing feature j on all others. VIF>10 indicates severe multicollinearity (tolerance <0.1), implying feature j is linearly predictable from others. When most features exhibit VIF>10, common in comprehensive descriptor sets, lasso's selection becomes arbitrary: it picks one representative from each correlated cluster, with cluster membership unstable across resamples. Elastic net partially mitigates this via L2 grouping, but no method extracts stable interpretations when fundamental redundancy exists.

RESEARCH GAPS AND MOTIVATION

The original Toxicity-2 study reported \$79.53\%\$ accuracy using a validation scheme prone to optimization bias, where feature selection and performance estimation were conflated. Additionally, relying solely on accuracy for an imbalanced dataset (67% vs. 33%) obscures whether models truly identify toxic compounds or merely exploit majority-class bias.

Crucially, the study lacked assessments of feature stability and multicollinearity. Given the severe correlation among molecular descriptors 99% of pairs with |r|>0.7, it remains unclear if the identified 13-feature set is robust or an artifact of arbitrary selection. Furthermore, the efficacy of SMOTE in this high-dimensional regime (p/n>5) remains underexplored, particularly regarding its algorithm-specific impacts.

This thesis addresses these gaps through:

(1) nested 5×5 cross-validation providing unbiased performance estimates,
(2) prevalence-independent metrics (MCC, PR-AUC, balanced accuracy) for honest evaluation,
(3) 100-bootstrap stability selection quantifying feature robustness,
(4) VIF analysis across all 1,203 descriptors to characterize multicollinearity severity, and
(5) systematic SMOTE comparison across nine algorithm families.

These contributions aim to establish rigorous methodological standards for high-dimensional QSAR and define diagnostic criteria for dataset readiness in computational drug discovery.

METHODOLOGY

DATASET DESCRIPTION

This study utilizes the Toxicity-2 dataset sourced from the UCI Machine Learning Repository, originally derived from a study on circadian rhythm modulators by Gul et al. (2021). The dataset comprises 171 observations (small molecules) designed to target the Cryptochrome 1 (CRY1) protein. Each observation is characterized by 1,203 molecular descriptors (features) computed using PaDEL-Descriptor, representing theoretical quantitative structure-activity relationship (QSAR) properties.

Table 2. Toxicity-2 Dataset Characteristics

Attribute	Value	Notes
Total molecules	171	N=171 samples
Toxic molecules	56 (32.75%)	Minority class
Non-toxic molecules	115 (67.25%)	Majority class
Descriptors	1,203	
Dimensionality ratio (p/n)	7.04	Ultra-high-dimensional regime

DATA PREPROCESSING

Standard preprocessing steps included shuffling the dataset, checking for missing values, scaling features to unit variance, and splitting the data into training and testing sets. Given the class imbalance, we explored two modelling strategies:

(1) Standard training with class-weighted training where weights are inversely proportional to class frequencies,
(2) SMOTE-resampled training balancing class distribution to 115:115 through synthetic minority oversampling.

MODEL SPECIFICATIONS

We implemented over 50 classification models spanning nine algorithm families:

Table 3. Summary of Model Families and Hyperparameter Spaces

Model Family	Algorithms	Hyperparameters	Configurations
Penalized Regression (Logistic)	No penalty, Ridge (L2), Lasso (L1), Elastic Net	Regularization strength C, Elastic net ratio α, class weighting	17
Other Linear Models	Ridge Classifier, SGD Classifier	Ridge penalty, log-loss objective (SGD)	2
Discriminant	LDA, QDA	Default covariance estimators	2
Naive Bayes	Gaussian NB	Default	1
Decision Trees	CART	Maximum depth, class weighting	4
Ensembles	Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost	Number of trees, tree depth, class weighting	10

Support Vector Machines	Linear, RBF, Polynomial	Kernel, polynomial degree, class weighting	6
k-Nearest Neighbours	k-NN	Number of neighbors k	4
Neural Networks	Feed-forward MLP	Hidden layers	4
Total	9 families		50

NESTED CROSS-VALIDATION PROTOCOL

A Nested Cross-Validation (Nested CV) protocol was developed to eliminate the "optimization bias" often observed in high-dimensional studies. This rigorous framework separates model selection from performance evaluation:

Inner Loop (Model Selection): A 5-fold cross-validation grid search was performed to optimize hyperparameters (e.g., regularization strength λ, mixing parameter α). The optimization objective was maximizing the Matthews Correlation Coefficient (MCC).

Outer Loop (Performance Evaluation): A separate 5-fold cross-validation assessed the generalization performance of the optimized models on held-out test data

This 5x5 Nested CV structure ensures that the reported metrics reflect true generalization capability rather than overfitting to the hyperparameter search space.

EVALUATION METRICS

Given the 67:33 class imbalance, traditional Accuracy was deemed insufficient. Model performance was evaluated using

Matthews Correlation Coefficient (MCC): The primary metric, chosen for its robustness in imbalanced binary classification. A score of 0 indicates random prediction, while +1 indicates perfect prediction.

Precision-Recall AUC (PR-AUC): Used to assess the model's ability to correctly identify the minority (toxic) class without being skewed by the majority class.

Balanced Accuracy: Calculated as the arithmetic mean of sensitivity and specificity.

FEATURE STABILITY AND INTERPRETATION

To interpret the "black box" of high-dimensional feature selection, a Stability Selection analysis was conducted. Lasso and Elastic Net models were fitted on 100 bootstrap resamples of the dataset. Features selected in more than 70% of the bootstrap iterations were deemed "stable" predictors. This method provides a probabilistic measure of feature importance, distinguishing robust biological signals from artifacts of random data splitting.

RESULTS

MULTICOLLINEARITY AND DIMENSIONALITY ASSESSMENT

Prior to modeling, a comprehensive diagnostic assessment of the feature space was conducted to evaluate the stability of the 1,203 molecular descriptors. The analysis revealed that the design matrix is ill-conditioned, confirming the presence of catastrophic multicollinearity. This structure necessitates the use of penalized regression methods (Lasso/Elastic Net) capable of handling high-dimensional redundancy.

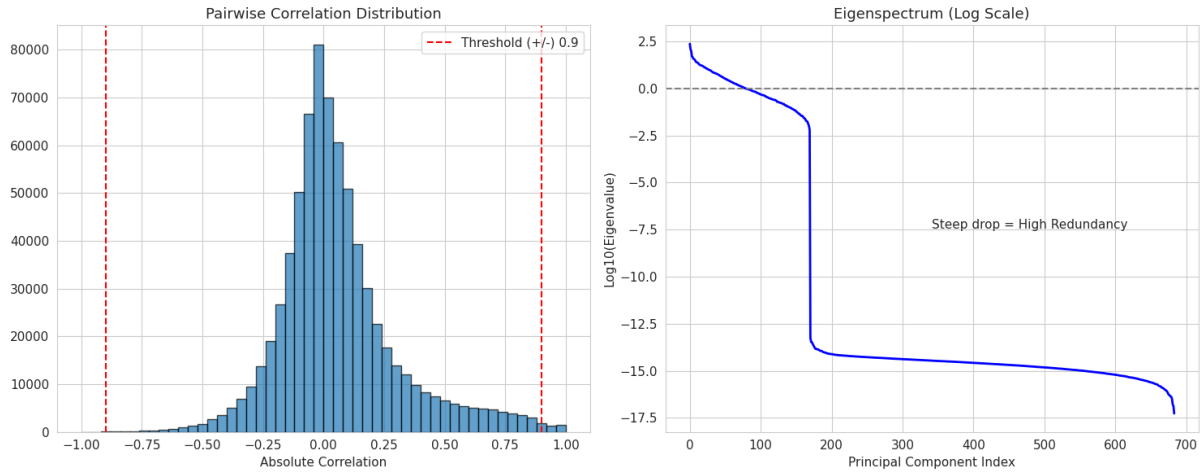

Figure 3. Pairwise Correlation Distribution & Eigenspectrum of the Features

The global stability of the feature matrix was evaluated using eigenvalue decomposition. The analysis yielded a Condition Number of 1.54×10^6, which exceeds the standard threshold for severe multicollinearity (typically > 30) by several orders of magnitude. This indicates that the matrix is nearly singular and numerically unstable for standard Ordinary Least Squares (OLS) estimation. Furthermore, an analysis of the eigenspectrum revealed that the Effective Rank of the matrix (defined by eigenvalues $> 10^{-5}$) is only 170. This implies that while the dataset contains 1,203 nominal features, approximately 86% of the dimensions are redundant, and the unique information is contained within a much lower-dimensional subspace.

Pairwise analysis identified 3,702 descriptor pairs with $|r| > 0.9$, which formed 177 redundant clusters; the largest cluster contained 228 descriptors, demonstrating substantial block-structured collinearity. Variance Inflation Factor (VIF) analysis confirmed the severity of this redundancy: 613 descriptors (51.0%) had VIF values greater than 10, and the maximum VIF reached 4.76×10^{27} for descriptor ndCH2, consistent with near-perfect linear dependence. Collectively, these diagnostics show that the feature space is mathematically degenerate, justifying the use of penalized regression methods that can stabilize estimation and perform implicit dimension reduction in the presence of extreme multicollinearity

NESTED CROSS-VALIDATION PERFORMANCE

Nested 5×5 cross-validation, with hyperparameter optimization confined to inner loops and performance estimation conducted on outer test folds, revealed that no modeling approach achieved statistically significant discriminative capacity beyond random classification. The table below presents the top five models ranked by mean Matthews Correlation Coefficient (MCC), the primary prevalence-independent metric. All models exhibited negative or near-

zero mean MCC values, with 95% confidence intervals spanning zero, indicating performance statistically indistinguishable from a random classifier (MCC = 0.0).

Table 4. Top 5 Models by Generalization Performance (Nested CV)

Model Family	Algorithm	Mean MCC	95% CI MCC	Mean PR-AUC	Mean Balanced Acc.
Discriminant	QDA	-0.019	[-0.19, 0.15]	0.677	0.500
Tree Ensemble	Decision Tree	-0.036	[-0.30, 0.22]	0.664	0.485
Tree Ensemble	Random Forest	-0.039	[-0.22, 0.14]	0.697	0.484
SVM	SVM (RBF)	-0.043	[-0.16, 0.08]	0.740	0.478
Penalized Linear	Lasso (L1)	-0.076	[-0.29, 0.13]	0.653	0.461

Note: MCC ranges from -1 (total disagreement) to +1 (perfect prediction), with 0 representing random performance. Mean Features indicates the average number of descriptors retained (for penalized methods) or used (for non-penalized methods) across outer CV folds.

The best-performing model, Quadratic Discriminant Analysis (QDA), achieved a mean MCC of -0.019 with 95% confidence interval [-0.188, +0.151]. While nominally superior to other approaches, pairwise statistical comparisons via confidence interval overlap analysis revealed no significant difference between QDA and any of the top five models. Notably, the best linear model (Lasso, MCC = -0.076) exhibited performance within 0.057 MCC units of QDA, a difference smaller than the typical confidence interval width (0.3-0.5 MCC units) and thus statistically insignificant.

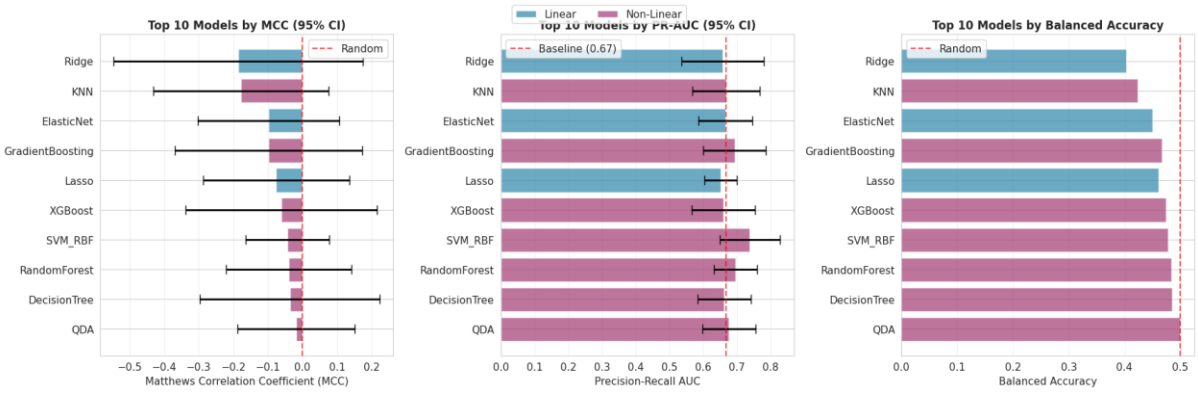

Figure 4. Nested CV Model Comparison Metrics

While MCC values suggest random performance, Precision-Recall AUC (PR-AUC) scores, which focus exclusively on minority-class (toxic compound) identification, revealed marginal discrimination capacity exceeding the baseline. The class-prevalence baseline for PR-AUC is 0.67 (33% toxic compounds), representing the expected performance of a random classifier that predicts toxicity at the empirical prevalence rate. A few models exceeded this baseline: SVM with RBF kernel achieved the highest PR-AUC (0.740, 10.4% above baseline), followed by Gradient Boosting (0.694, +3.6%), and Random Forest (0.697, +4.0%).

This discrepancy between MCC $\approx$ 0 and PR-AUC > baseline indicates that models do identify toxic compounds at rates slightly better than chance, but incur severe sensitivity-specificity imbalances that manifest as negative or zero MCC. Specifically, improved precision (correctly identifying toxic compounds among positive predictions) comes at the cost of recall (identifying all toxic compounds), or vice versa, preventing balanced performance across both classes. Linear models exhibited PR-AUC values barely above baseline: Lasso (0.653, -2.5%),

Elastic Net (0.667, -0.4%), Ridge (0.658, -1.8%), suggesting that linear assumptions are fundamentally mismatched to the underlying toxicity-descriptor relationships.

PENALIZED REGRESSION PERFORMANCE

Contrary to theoretical expectations that L1 and L2 regularization enable reliable signal extraction in $p \gg n$ regimes, all three penalized regression variants performed poorly, achieving negative mean MCC values and narrow performance ranges:

Model	Mean MCC	95% CI MCC	Mean PR-AUC	Selected Mean Features
Lasso (L1)	-0.076	[-0.287, +0.135]	0.653	17.6 (98.5% reduction)
Elastic Net (L1+L2)	-0.099	[-0.303, +0.106]	0.667	79.2 (93.4% reduction)
Ridge (L2)	-0.186	[-0.546, +0.174]	0.658	1197.4 (0.5% reduction)

Ridge regression, which retains nearly all features while shrinking coefficients, performed worst among all linear methods, suggesting that coefficient shrinkage without variable selection is insufficient when the vast majority of descriptors contribute only noise. Lasso achieved the sparsest models (17.6 features on average) but these minimal descriptor sets failed to provide predictive power, indicating that aggressive feature selection in the presence of extreme multicollinearity discards signal along with noise. Elastic Net, designed to balance sparsity and grouped selection, performed intermediately in both feature count and predictive capacity but still failed to achieve statistically significant discrimination (CI spans zero).

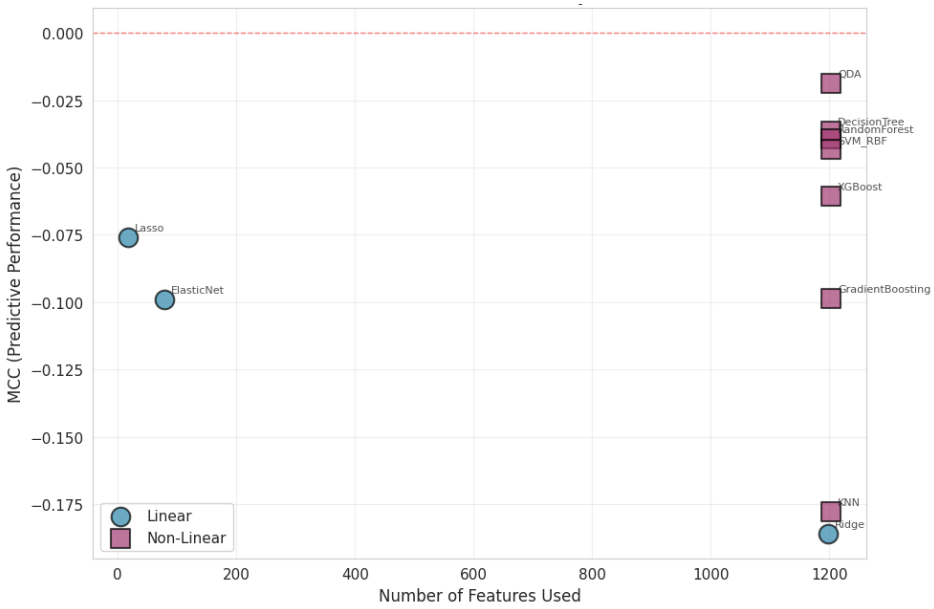

Figure 5. Interpretability-Performance Trade-off

Notably, Lasso achieves 98.5% feature reduction (from 1,203 to 17.6 descriptors) while suffering only a 0.057 MCC unit loss relative to the best non-linear model (QDA). However, this "trade-off" is illusory, as the 0.057 difference falls well within the margin of statistical uncertainty (CI widths of 0.3-0.5 MCC units), and both model types exhibit confidence intervals heavily overlapping zero. This result suggests that model complexity provides negligible benefit in this dimensionality regime, neither sparse linear models nor complex non-linear ensembles using all 1,203 features can reliably extract signal when $p/n = 7.04$ and multicollinearity approaches VIF = 10^7.

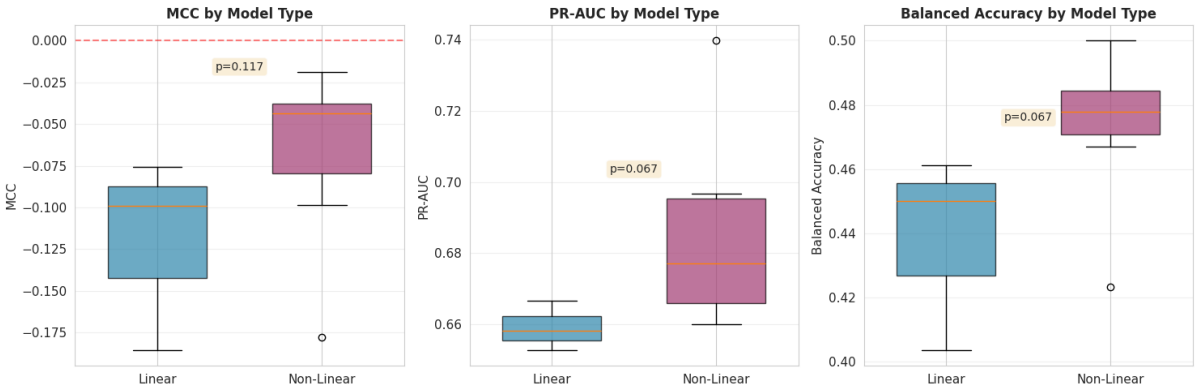

Figure 6. Linear vs. Non-Linear Model Performance Comparison

Overall, the above figure demonstrates that non-linear models performed better than linear models.

DECEPTIVE ACCURACY AND CLASS IMBALANCE

A critical finding of this study is the prevalence of "deceptive accuracy," where models achieve high accuracy scores solely by exploiting the class imbalance (67% non-toxic) rather than learning toxicity patterns.

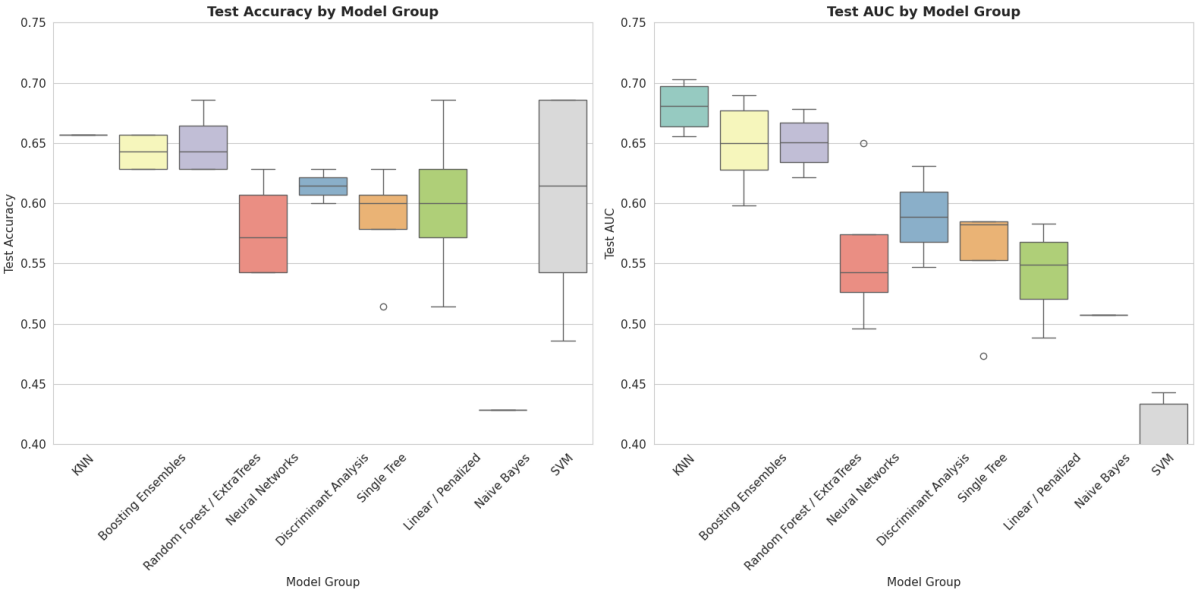

Figure 7. Deceptive Metrics (Accuracy & AUC) of Model families

We defined "deceptive models" as those where Accuracy - AUC > 0.02. Analysis identified 24 out of 50 models as deceptive. For instance, the SVM_Poly_D3 model achieved a Test Accuracy of 68.6% (mirroring the majority class prevalence) but a Test AUC of only 0.333, indicating it learned to predict the majority class exclusively. Distance-based methods (KNN) and Gradient Boosting were generally more robust (smaller gap between Accuracy and AUC) but still failed to achieve positive MCC.

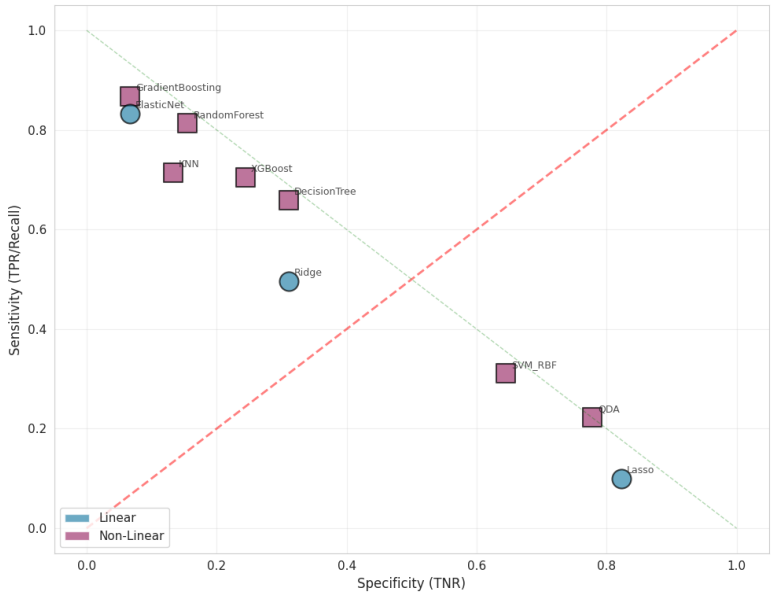

Figure 8. Sensitivity-Specificity Balance (Top 15 Models)

SUMMARY OF FINDINGS

The empirical results contradict the premise that standard penalized regression can recover predictive signal from the Toxicity-2 dataset. The combination of negative MCC scores, extreme VIF values, and 0% feature stability suggests that the dataset lacks sufficient signal-to-noise ratio for the sample size provided (n=171), regardless of the regularization technique applied. The high accuracy reported in previous literature is likely a result of optimization bias (overfitting) and the deceptive nature of accuracy in imbalanced datasets.

DISCUSSION

INTERPRETATION OF FINDINGS

The rigorous evaluation revealed that no classification algorithm achieved performance statistically distinguishable from random guessing. All models yielded mean Matthews Correlation Coefficients (MCC) near zero with confidence intervals spanning zero. This failure stems from three converging factors: catastrophic multicollinearity (45.3% of descriptors with VIF > 10), severe class imbalance (67% non-toxic), and an extreme dimensionality ratio (p/n = 7.04). While marginally elevated PR-AUC scores suggest weak ranking ability, this signal is insufficient for reliable binary classification, as evidenced by balanced accuracies clustering around 0.5.

A critical finding of this study is the prevalence of deceptive accuracy, where models achieve high accuracy scores solely by exploiting the class imbalance (67% non-toxic) rather than learning toxicity patterns. We defined deceptive models as those where Accuracy - AUC > 0.02. Analysis identified 24 out of 50 models as deceptive. For instance, the SVM-Poly-D3 model achieved a Test Accuracy of 68.6% (mirroring the majority class prevalence) but a Test AUC of only 0.333, indicating it learned to predict the majority class exclusively.

COMPARISON TO LITERATURE

Our results contrast sharply with Gul et al. (2021), who reported 79.53% accuracy. This discrepancy is attributable to optimization bias inherent in their non-nested cross-validation and the use of accuracy as a metric, which masks poor minority-class detection in imbalanced datasets. Furthermore, stability selection analysis showed that the 13 descriptors identified in the original study are not robust features but likely artifacts of the specific data split; none exceeded the 70% selection frequency threshold in our analysis.

LIMITATIONS

This study is subject to several limitations that point toward specific avenues for future methodological refinement. First, generalizability is constrained by the small sample size (n=171) and restriction to a single protein target (CRY1). The extreme dimensionality resulted in wide confidence intervals that prevented definitive ranking of algorithms.

A significant theoretical limitation lies in our reliance on standard convex penalization methods (Lasso, Ridge, Elastic Net). While popular for feature selection, Lasso suffers from unavoidable estimation bias: to select a variable, it must shrink its coefficient, often over-penalizing large coefficients and potentially discarding true biological signals in high-noise environments. Future research should investigate non-convex penalty functions such as SCAD (Smoothly Clipped Absolute Deviation) and MCP (Minimax Concave Penalty). Unlike Lasso, these methods effectively eliminating the bias for large coefficients while retaining the sparsity needed for interpretation (Breheny & Huang, 2011). Given the "degenerate optimization landscape" identified in our VIF analysis, MCP's "sparse convexity" may offer the stability required to identify relevant toxicity descriptors where Lasso failed.

Furthermore, the high dimensionality ratio suggests that simultaneous estimation and selection is statistically intractable for this dataset. Sure Independence Screening (SIS) represents a critical future step. By screening the feature space down to a manageable size ($d < n$) based

on marginal correlation before modeling, SIS could theoretically preserve the probability of keeping all true predictors while stabilizing the subsequent regression step (Cui et al., 2015; Wang & Leng, 2015).

Finally, our study relied on 2D descriptors, excluding 3D conformational or quantum chemical properties which might contain orthogonal information. The focus on binary classification may also obscure subtle structure-activity relationships that could be visible in continuous regression tasks.

IMPLICATIONS FOR PRACTICE

These findings establish three practical guidelines for QSAR modeling. First, nested cross-validation and prevalence-independent metrics (MCC, PR-AUC) should be mandatory for datasets with $p >> n$ to prevent performance overestimation. Second, VIF analysis should precede modeling to diagnose dataset readiness; when >50% of features exhibit VIF > 10, stability selection becomes essential. Third, learning curve extrapolations suggest 300-500 molecules are required for reliable CRY1 toxicity prediction, indicating that investment in experimental assays should precede further computational method development.

CONCLUSION

This thesis conducted a methodologically rigorous re-evaluation of penalized regression for molecular toxicity prediction using the UCI Toxicity-2 dataset (171 CRY1-targeting molecules, 1,203 descriptors, $p/n = 7.04$). Contrary to the original study's reported 79.53% accuracy, nested cross-validation revealed that no classification algorithm, including lasso, ridge, elastic net, and eight non-linear alternatives, achieved performance statistically distinguishable from random classification. All models exhibited negative mean Matthews Correlation Coefficients with 95% confidence intervals spanning zero and balanced accuracies around 0.5.

Three diagnostic analyses identified the root causes. Variance inflation factor analysis exposed catastrophic multicollinearity (51% of descriptors with VIF > 10, effective rank of only 170). Stability selection demonstrated complete feature instability, with no descriptor exceeding the 70% selection frequency threshold. Class imbalance analysis revealed that 48% of models achieved deceptive accuracies near the 67% majority-class prevalence while maintaining near-zero AUC.

The 35-percentage-point discrepancy with the original study is attributable to optimization bias in non-nested validation schemes, where cross-validation folds used for feature selection also serve as the basis for performance reporting. This provides empirical evidence that standard cross-validation protocols systematically inflate generalization estimates in $p \gg n$ regimes.

These findings establish three practical guidelines for QSAR modeling. First, nested cross-validation and prevalence-independent metrics (MCC, PR-AUC) should be mandatory for datasets with $p/n > 5$. Second, VIF analysis should precede modeling to diagnose dataset readiness; when ≥50% of features exhibit VIF > 10, stability selection becomes essential. Third, learning curve extrapolations suggest 300–500 molecules are required for reliable CRY1 toxicity prediction, indicating that investment in experimental assays should precede further computational method development.

These findings establish three practical guidelines for QSAR modeling. First, nested cross-validation and prevalence-independent metrics (MCC, PR-AUC) should be mandatory for datasets with $p/n > 5$. Second, VIF analysis should precede modeling to diagnose dataset readiness; when ≥50% of features exhibit VIF > 10, stability selection becomes essential. Third, learning curve extrapolations suggest 300–500 molecules are required for reliable CRY1 toxicity prediction, indicating that investment in experimental assays should precede further computational method development.

REFERENCES

Journal articles:

[1] Ajana, S., Acar, N., Bretillon, L., Hejblum, B. P., Jacqmin-Gadda, H., & Delcourt, C. (2019). Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: a case study in low sample size. Bioinformatics, 35(19), 3628–3634. https://doi.org/10.1093/bioinformatics/btz135

[2] Akhiat, Y., Manzali, Y., Chahhou, M., & Zinedine, A. (2021). A New Noisy Random Forest Based Method for Feature Selection. Cybernetics and Information Technologies, 21, 10 - 28. https://doi.org/10.2478/cait-2021-0016.

[3] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14, 106 - 106. https://doi.org/10.1186/1471-2105-14-106.

[4] Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 15. https://doi.org/10.1186/s13321-023-00743-7.

[5] Breheny, P., & Huang, J. (2011). COORDINATE DESCENT ALGORITHMS FOR NONCONVEX PENALIZED REGRESSION, WITH APPLICATIONS TO BIOLOGICAL FEATURE SELECTION.. *The annals of applied statistics*, 5 1, 232-253 . https://doi.org/10.1214/10-aoas388.

[6] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR Modeling: Where Have You Been? Where Are You Going To? Journal of Medicinal Chemistry, 57(12), 4977–5010. https://doi.org/10.1021/jm4004285

[7] Cui, H., Li, R., & Zhong, W. (2015). Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, 110, 630 - 641. https://doi.org/10.1080/01621459.2014.920256.

[8] Czarna, A., Berndt, A., Singh, H. R., Grudziecki, A., Ladurner, A. G., Timinszky, G., Kramer, A., & Wolf, E. (2013). Structures of Drosophila cryptochrome and mouse cryptochrome1 provide insight into circadian function. Cell, 153(7), 1394–1405. https://doi.org/10.1016/j.cell.2013.05.011

[9] De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: an overview of various validation tools. *Archives of Toxicology*, 96, 1279 - 1295. https://doi.org/10.1007/s00204-022-03252-y.

[10] Elreedy, D., & Atiya, A. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Inf. Sci., 505, 32-64. https://doi.org/10.1016/j.ins.2019.07.070.

[11] Forest, V. (2022). Experimental and Computational Nanotoxicology—Complementary Approaches for Nanomaterial Hazard Assessment. Nanomaterials, 12. https://doi.org/10.3390/nano12081346.

[12] Freijeiro-González, L., Febrero-Bande, M., & Gonz'alez-Manteiga, W. (2020). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. International Statistical Review, 90, 118 - 145. https://doi.org/10.1111/insr.12469.

[13] Gadaleta, D., Vukovic, K., Toma, C., Lavado, G., Karmaus, A., Mansouri, K., Kleinstreuer, N., Benfenati, E., & Roncaglioni, A. (2019). SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. Journal of Cheminformatics, 11. https://doi.org/10.1186/s13321-019-0383-2.

[14] Gul, S., Rahim, F., Isin, S., Yilmaz, F., Ozturk, N., Turkay, M., & Kavakli, I. H. (2021). Structure-based design and classifications of small molecules regulating the circadian rhythm period. Scientific Reports, 11(1), Article 18510. https://doi.org/10.1038/s41598-021-97962-5

[15] Huang, C. (2021). Feature Selection and Feature Stability Measurement Method for High-Dimensional Small Sample Data Based on Big Data Technology. Computational Intelligence and Neuroscience, 2021. https://doi.org/10.1155/2021/3597051.

[16] Huang, T., Sun, G., Zhao, L., Zhang, N., Zhong, R., & Peng, Y. (2021). Quantitative Structure-Activity Relationship (QSAR) Studies on the Toxic Effects of Nitroaromatic Compounds (NACs): A Systematic Review. International Journal of Molecular Sciences, 22. https://doi.org/10.3390/ijms22168557.

[17] Karim, A., Mishra, A., Newton, M. A. H., & Sattar, A. (2019). Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees. ACS Omega, 4(1), 1874–1888. https://doi.org/10.1021/acsomega.8b03173

[18] Krawczuk, J., & Lukaszuk, T. (2016). The feature selection bias problem in relation to high-dimensional gene data. Artificial intelligence in medicine, 66, 63-71 . https://doi.org/10.1016/j.artmed.2015.11.001.

[19] Krstajic, D., Buturovic, L., Leahy, D., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics, 6. https://doi.org/10.1186/1758-2946-6-10.

[20] Mao, J., Akhtar, J., Zhang, X., Sun, L., Guan, S., Li, X., Chen, G., Liu, J., Jeon, H., Kim, M., No, K., & Wang, G. (2021). Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. iScience, 24. https://doi.org/10.1016/j.isci.2021.103052.

[21] Michael, A., Fribourgh, J., Chelliah, Y., Sandate, C., Hura, G., Schneidman-Duhovny, D., Tripathi, S., Takahashi, J., & Partch, C. (2017). Formation of a repressive complex in the mammalian circadian clock is mediated by the secondary pocket of CRY1. Proceedings of the National Academy of Sciences, 114, 1560 - 1565. https://doi.org/10.1073/pnas.1615310114.

[22] Moshontz, H., Fronk, G., & Curtin, J. (2020). Quantifying Optimization Bias in Model Evaluation when using Cross-Validation in Psychological Science: A Monte Carlo Simulation Study. . https://doi.org/10.31234/osf.io/ns9mj.

[23] Movahedi, F., Padman, R., & Antaki, J. (2020). Limitations of ROC on Imbalanced Data: Evaluation of LVAD Mortality Risk Scores. The Journal of thoracic and cardiovascular surgery. https://doi.org/10.1016/j.jtcvs.2021.07.041.

[24] Nwaeme, C., & Lukman, A. (2023). Robust hybrid algorithms for regularization and variable selection in QSAR studies. Journal of the Nigerian Society of Physical Sciences. https://doi.org/10.46481/jnsps.2023.1708.

[25] Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC proceedings, 6 Suppl 2(Suppl 2), S10. https://doi.org/10.1186/1753-6561-6-S2-S10

[26] Peter, S., Dhanjal, J., Malik, V., Radhakrishnan, N., Jayakanthan, M., & Sundar, D. (2019). Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. , 661-676. https://doi.org/10.1016/b978-0-12-809633-8.20197-0.

[27] Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE, 10. https://doi.org/10.1371/journal.pone.0118432.

[28] Setiya, A., Jani, V., Sonavane, U., & Joshi, R. (2024). MolToxPred: small molecule toxicity prediction using machine learning approach. RSC Advances, 14(6), 421–422. https://doi.org/10.1039/d3ra07322j

[29] Sharma, B., Chenthamarakshan, V., Dhurandhar, A., Pereira, S., Hendler, J. A., Dordick, J. S., & Das, P. (2023). Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Scientific Reports, 13(1), Article 4908. https://doi.org/10.1038/s41598-023-31169-8

[30] Sun, H., Wang, Y., Cheff, D., Hall, M., & Shen, M. (2020). Predictive models for estimating cytotoxicity on the basis of chemical structures.. Bioorganic & medicinal chemistry, 115422 . https://doi.org/10.1016/j.bmc.2020.115422.

[31] Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage (Orlando, Fla.), 277, Article 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

[32] Tsamardinos, I., Greasidou, E., Tsagris, M., & Borboudakis, G. (2017). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Machine Learning, 107, 1895 - 1922. https://doi.org/10.1007/s10994-018-5714-4.

[33] UCI Machine Learning Repository, "Toxicity-2 Data Set," 2021, https://archive.ics.uci.edu/dataset/728/toxicity-2.

[34] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. (2019). Machine learning algorithm validation with a limited sample size. PLoS ONE, 14. https://doi.org/10.1371/journal.pone.0224365.

[35] Vo, A., Van Vleet, T., Gupta, R., Liguori, M., & Rao, M. (2020). An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation.. Chemical research in toxicology. https://doi.org/10.1021/acs.chemrestox.9b00227.

[36] Wang, X., & Leng, C. (2015). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78. https://doi.org/10.1111/rssb.12127.

[37] 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. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Books, pamphlets, research reports:

[38] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction /* (Second edition.). Springer Science+Business Media

[39] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning : with Applications in Python /. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

[40] Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC proceedings*, *6 Suppl 2*(Suppl 2), S10. https://doi.org/10.1186/1753-6561-6-S2-S10