Big Data and Information Analytics

BDIA, 10: 177–188.
DOI: 10.3934/bdia.2026009
Received: 12 July 2026
Revised: 29 July 2026
Accepted: 31 July 2026
Published: 06 August 2026

https://www.aimspress.com/journal/BDIA

Research article

A multi-dimensional feature selection and stacking ensemble framework for QSAR prediction

Chengkai Tu, Weifeng Jiang* and Ruoyi Liu

College of Science, China Jiliang University, Hangzhou 310018, China

*** Correspondence:** Email: casujiang89@gmail.com.

Abstract: Quantitative structure-activity relationship (QSAR) modeling suffers from high-dimensional feature redundancy and individual algorithm inductive bias. To address these bottlenecks, this paper proposed a two-stage multidimensional integrated feature selection and heterogeneous stacking ensemble framework (MIFS-Stacking). Evaluated on 1974 estrogen receptor alpha (ERα) inhibitor samples from DrugBank, the MIFS module first compressed the 729-dimensional descriptor space to 20 core features using heuristic filtering and a multi-algorithm weighted ensemble scoring scheme across five heterogeneous evaluators. Next, a Level 1 primary layer combining five mathematically diverse base learners (ElasticNet, KNN, SVR, random forest, and XGBoost) captured complex nonlinear patterns, while a Level 2 L2-regularized RidgeCV meta-learner synthesized predictions via out-of-fold (OOF) cross-validation to suppress secondary overfitting. Global hyperparameter optimization was efficiently conducted using the Optuna Bayesian framework. Empirical results demonstrated that MIFS-Stacking achieved superior generalization performance on the test set (R^2 = 0.7375, RMSE = 0.7339, MAE = 0.5344), significantly outperforming all standalone baselines. The proposed MIFS-Stacking paradigm offered a robust, scalable, and highly accurate computational strategy for high-dimensional chemoinformatics regression tasks.

Keywords: quantitative structure-activity relationship; multi-dimensional feature selection; heterogeneous stacking ensemble; Bayesian optimization

1. Introduction

Quantitative structure-activity relationship (QSAR) modeling serves as a fundamental

computational methodology in chemoinformatics and computational pharmacology, aiming to establish quantitative mapping relationships between the microscopic molecular descriptors of compounds and their macroscopic biological activities [1,2]. In recent years, driven by the rapid development of high-throughput screening (HTS) technologies and data-intensive science, the dimensionality of molecular databases has experienced explosive growth [3]. Physical, chemical, topological, and quantum chemical descriptors extracted from a single compound frequently reach hundreds or even thousands of dimensions [3]. While such high-dimensional chemoinformatics data offers rich microscopic structural insights for molecular representation, it also introduces severe methodological challenges to traditional data mining and statistical modeling approaches [4].

First, high-dimensional molecular descriptor sets generally suffer from the "curse of dimensionality", high data sparsity, and severe multicollinearity [5]. Directly utilizing all features for modeling not only easily triggers model overfitting, but also significantly increases computational complexity and impairs the extrapolation and generalization ability of the predictive model [6]. Traditional feature selection techniques, such as linear correlation filtering or single regularization methods (e.g., LASSO), are typically built upon the simplistic assumptions of a single algorithm [7]. However, the mapping between molecular descriptors and biological activity exhibits both linear and complex nonlinear characteristics [8]. A single feature evaluation algorithm is prone to specific inductive bias, making it difficult to comprehensively capture key pharmacophore factors in high-dimensional space, and may even mistakenly eliminate core features possessing nonlinear synergistic effects [9].

Second, regarding the choice of prediction algorithms, individual machine learning models exhibit inherent limitations when tackling complex QSAR regression tasks [10]. For instance, linear models (such as elastic net) struggle to capture high-order nonlinear interactions [11]; kernel methods (such as support vector regression, SVR) are highly sensitive to hyperparameters and face scalability bottlenecks under large sample sizes [12]; and tree-based ensemble models (such as random forest and XGBoost), despite their strong nonlinear fitting capabilities, are prone to high predictive variance or local overfitting in sparse feature spaces [13,14]. To overcome the inductive bias of individual models, stacking ensemble architectures combine multiple heterogeneous base learners and perform secondary induction at the meta-learner level, which has been proven effective in reducing predictive variance and extending fitting boundaries [15,16]. Nevertheless, when applied to high-dimensional QSAR predictions, most existing stacking architectures face secondary overfitting caused by high correlations among base learners' predictions, as well as convergence bottlenecks due to vast hyperparameter search spaces [17].

To address these methodological bottlenecks, this paper utilizes the bioactivity (pIC_{50}) prediction of estrogen receptor alpha (ERα) inhibitors—a core target in anti-breast cancer therapy—as an application vehicle, and proposes a two-stage multidimensional integrated feature selection and heterogeneous stacking ensemble framework (MIFS-stacking). The primary methodological contributions of this study are as follows:

- Two-stage multidimensional integrated feature selection (MIFS) mechanism: overcoming the evaluation bias of single algorithms, a two-stage feature selection architecture combining "heuristic filtering" and "multi-algorithm heterogeneous weighted evaluation" is designed. By integrating five heterogeneous criteria—Spearman rank correlation, random forest and gradient boosting decision tree importance scores, ElasticNet sparsity, and recursive feature elimination (RFE)—a high signal-to-noise ratio compression of the high-dimensional feature space is achieved.
- Deep stacking integration of heterogeneous base learners and regularized meta-learner: in

Level 1, five mathematically complementary base learners (RF, XGBoost, SVR, KNN, ElasticNet) are integrated in parallel to cover linear, topological-geometric, and nonlinear feature spaces; in Level 2, RidgeCV with L2 regularization is introduced as the meta-learner, coupled with an out-of-fold (OOF) cross-validation mechanism to effectively suppress secondary feature overfitting and data leakage.

2. Theoretical principles and proposed framework

To overcome the inherent limitations of high-dimensional sparsity and the inductive bias of individual algorithms in QSAR modeling, this study proposes a comprehensive computational framework termed MIFS-Stacking. The architecture strictly separates feature selection from predictive modeling to ensure high extrapolation capability and robustness.

2.1. Overall architecture of the MIFS-stacking framework

The proposed framework fundamentally consists of two sequential phases. The first phase acts as the input layer, constructing a two-stage MIFS model to extract the most predictive core feature subset from the high-dimensional molecular descriptor space. The second phase operates as the core predictive layer, employing a stacking architecture that integrates five mathematically complementary base learners and a regularized meta-learner, optimized globally via the Optuna Bayesian framework.

2.2. Two-stage MIFS

The objective of the MIFS module is to extract a feature subset possessing the strongest representational capacity for the target biological activity *y* from the original high-dimensional feature space $X \in \mathbb{R}^{n\times p}$ (where $p = 729$).

2.2.1. Stage I: heuristic noise reduction

The initial stage employs heuristic filtering to eliminate redundant and low-information features. First, a variance filter is applied. For each feature X_j, its sample variance is defined as:

$$Var(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \tag{1}$$

A threshold $\tau = 0.05$ is established to retain features satisfying $Var(X_j) > \tau$. Subsequently, the Spearman rank correlation coefficient is utilized to detect and eliminate severe multicollinearity among features:

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

where d_i represents the difference in ranks between two variables. If the correlation between two features satisfies $|\rho_s| > 0.9$, the feature demonstrating a lower correlation with the target variable y is discarded.

2.2.2. Stage II: multidimensional ensemble evaluation

To eliminate the sensory bias of individual algorithms, a multi-algorithm weighted ensemble scoring model is constructed. The evaluation algorithm set is defined as $\mathcal{M} = \{m_1, m_2, \ldots, m_5\}$, comprising Spearman correlation, RF, gradient boosting decision tree (GBDT), ElasticNet, and RFE. For the k-th algorithm, its raw score for feature X_j is denoted as S_{kj}, calculated based on the following heterogeneous logics:

- Spearman rank correlation (m_1): captures the nonlinear monotonic relationship between the feature rank $R(X_j)$ and the bioactivity rank $R(y)$,where S_{1j} represents the absolute value of the Spearman rank correlation coefficient between the j-th descriptor and the target property, ensuring that only the magnitude of the correlation is considered.

$$S_{1j} = |\rho_s| = \left|1 - \frac{6\sum(R(x_{ij}) - R(y_i))^2}{n(n^2-1)}\right| \tag{3}$$

- RF mean decrease impurity (m_2): evaluates the sum of Gini impurity decrease when feature X_j is selected as the splitting node s across all decision trees T [9]:

$$S_{2j} = \frac{1}{N_{tree}} \sum_{T \in RF} \sum_{\substack{s \in T \\ v(s)=j}} \Delta Gini(s, X_j) \tag{4}$$

- GBDT global loss reduction (m_3): measures the cumulative contribution of feature X_j to the reduction of the global loss function (e.g., mean squared error (MSE)) across sequential regression trees:

$$S_{3j} = \sum_{t=1}^{T} \sum_{\substack{s \in tree_t \\ v(s)=j}} Gain(s) \tag{5}$$

- ElasticNet sparsification (m_4): imposes dual L1 (Lasso) and L2 (Ridge) penalties. The feature score is determined by the absolute value of its regularized regression coefficient:

$$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 + \alpha\lambda \|\beta\|_1 + \frac{\alpha(1-\lambda)}{2} \|\beta\|_2^2 \right) \tag{6}$$

$$S_{4j} = |\beta_j| \tag{7}$$

- Recursive feature elimination (m_5): a wrapper-based strategy that iteratively eliminates the least important features. Given a total feature count P, the score is inversely related to its elimination rank ($Rank_j$):

$$S_{5j} = P - Rank_j + 1 \tag{8}$$

To eliminate dimensional discrepancies across heterogeneous algorithms, a min-max normalization is applied:

$$Norm(S_{kj}) = \frac{S_{kj} - min(S_k)}{max(S_k) - min(S_k)} \tag{9}$$

Ultimately, the comprehensive score for each feature, $Score_j$, is synthesized using an equal-weight ensemble strategy to form the core feature vector $\mathbf{x}^*$:

$$Score_j = \sum_{k=1}^{5} w_k \cdot Norm(S_{kj}), \sum w_k = 1 \quad (10)$$

2.3. Heterogeneous stacking ensemble modeling

Single-model approaches often suffer from inherent inductive bias, especially when modeling complex drug structure-activity relationships. Therefore, a stacking heterogeneous ensemble architecture is implemented to maximize predictive accuracy.

2.3.1. Level 1: heterogeneous base learners

The primary layer integrates five mathematically diverse models $\mathscr{H} = \{h_1, h_2, \ldots, h_5\}$, ensuring comprehensive mapping from linear to complex topological spaces:

- ElasticNet: provides a smoothed linear baseline through dual regularization:

$$\hat{y}_{ENet} = \sum X_j \beta_j + b \quad (11)$$

- KNN: a nonparametric spatial proximity approach, calculating the mean bioactivity of the k nearest spatial neighbors $N_k(x)$ [18]:

$$\hat{y}_{KNN} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \quad (12)$$

- SVR: utilizes the $\epsilon -$insensitive loss function and kernel trick $\kappa(x_i, x_j)$ to find the optimal hyperplane in a high-dimensional space [12]:

$$\hat{y}_{SVR} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \kappa(x_i, x) + b \quad (13)$$

- RF: a bagging-based parallel ensemble that mitigates local noise by averaging the outputs of T independent regression trees $h_t(x)$ [9]:

$$\hat{y}_{RF} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (14)$$

- XGBoost: an extreme gradient boosting framework that approximates complex indicators by additively superimposing weak learners $f_t(x)$ while employing a second-order Taylor expansion:

$$\hat{y}_{XGB} = \sum_{t=1}^{T} f_t(x) \quad (15)$$

To strictly prevent data leakage, a 5-fold OOF cross-validation strategy is adopted to generate the meta-features. The training set is split into $K = 5$ subsets. Each base model h_m is trained on $K - 1$ subsets and predicts on the hold-out validation set to construct the meta-feature column z_m.

2.3.2. Level 2: regularized meta-learner

The predictions from the primary layer form a highly correlated secondary meta-feature matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_5]$. To suppress the multicollinearity among base learners' outputs, RidgeCV, equipped with an L2 regularization penalty, is selected as the meta-learner f_{meta}:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - f_{meta}(z_i; \beta))^2 + \lambda \|\beta\|_2^2 \quad (16)$$

Equation (16) presents the objective function of ridge regression, where λ controls the penalty imposed on the coefficient vector β.

The final synthesized prediction output of the Stacking architecture is formulated as:

$$\hat{y} = f_{meta}(h_1(x), \ldots, h_5(x)) \quad (17)$$

2.4. Optuna Bayesian hyperparameter optimization

To prevent arbitrary manual tuning, the Optuna framework, founded on Bayesian optimization theory, is integrated. By treating the cross-validated error of the stacking model as the objective function, Optuna dynamically constructs a surrogate probability model of the hyperparameter space. This enables the algorithm to efficiently balance exploration and exploitation, dynamically navigating the hyperparameter search space to locate the global optimum under regularization constraints.

3. Empirical evaluation and data analytics

3.1. Dataset overview, descriptive analytics, and preprocessing

The empirical analysis is conducted using compound data retrieved from the DrugBank database, targeting the compounds, characterized across 729 molecular descriptors, a continuous bioactivity target variable (pIC_{50}), and five ADMET pharmacokinetic properties. The structural composition of the dataset is detailed in Table 1.

Table 1. Dataset dimension, nature, and research application.

Data dimension	Variable nature	Research application
1974 × 729	Continuous independent variables	Provides 1D, 2D, and 3D structural features representing microscopic physicochemical properties.
1974 × 1	Continuous dependent variable	Contains bioactivity indicator pIC_{50} ($-\log_{10} IC_{50}$), serving as the target variable for quantitative regression.
1974 × 5	Discrete dependent variables	Encompasses absorption (Caco-2), distribution (CYP3A4), metabolism (hERG), excretion (HOB), and toxicity (MN).

To eliminate dimensional disparities and ensure numerical stability during optimization, data cleaning confirmed zero missing values across all 729 descriptors. Subsequently, min-max normalization was applied to scale feature descriptors and target values into the [0,1] range:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (18)$$

Descriptive statistical analysis was performed on the primary target variable, pIC_{50}. As shown in Table 2, the mean value of pIC_{50} is 6.586, which is virtually identical to its median (6.581), indicating

a clear central tendency. The variable exhibits a standard deviation of 1.423, ranging from a minimum of 2.456 to a maximum of 10.337.

Table 2. Descriptive statistics of biological activity.

Statistical metric	Numerical value	Statistical metric	Numerical value
Sample size	1974	Median	6.581
Mean	6.586	Maximum	10.337
Standard deviation	1.423	Skewness	0.183
Minimum	2.456	Kurtosis	–0.785

As shown in Table 2,the skewness of 0.183 (slight right-skew) and kurtosis of –0.785 confirm that pIC_{50} follows a continuous, quasi-normal distribution suitable for regression modeling. Regarding feature properties, the 729 molecular descriptors consist of 370 discrete/integer variables (50.8%) and 359 continuous/float variables (49.2%).

3.2. Two-stage integrated feature selection (MIFS) results

The high-dimensional feature space was systematically compressed through the two-stage MIFS framework:

- Stage I (heuristic noise reduction): variance filtering ($\tau = 0.05$) removed low-variance descriptors, reducing the dimensionality from 729 to 360. Spearman collinearity filtering ($|\rho_s| > 0.9$) further pruned redundant features, scaling the space down to 150 descriptors.
- Stage II (multidimensional ensemble scoring): the 150 descriptors were evaluated across five heterogeneous models. As shown in Table 3, the weighted integration scoring method has selected the top 20 core molecular descriptors.

To confirm that the MIFS module eliminated multicollinearity, a Spearman correlation matrix was constructed for the Top 20 descriptors. As shown in the Figure 1, inter-feature correlation coefficients are predominantly low, exhibiting clear block independence. This confirms that the selected 20 core descriptors provide high predictive information density with minimal redundancy.

Table 3. Top 20 core molecular descriptors selected by MIFS.

Core descriptor	Final score	Core descriptor	Final score
MDEC-23	0.8024	maxsOH	0.386
LipoaffinityIndex	0.6178	minssO	0.3609
nHBAcc	0.5716	AMR	0.3464
C1SP2	0.4801	minHBint10	0.3322
maxHsOH	0.4789	MDEC-22	0.3271
BCUTp-1h	0.4664	minHBint5	0.3251
CrippenLogP	0.4293	MDEC-33	0.3232
C3SP2	0.4261	SwHBa	0.3176
MLFER_A	0.4235	XLogP	0.3174
minsOH	0.397	nsssCH	0.3068

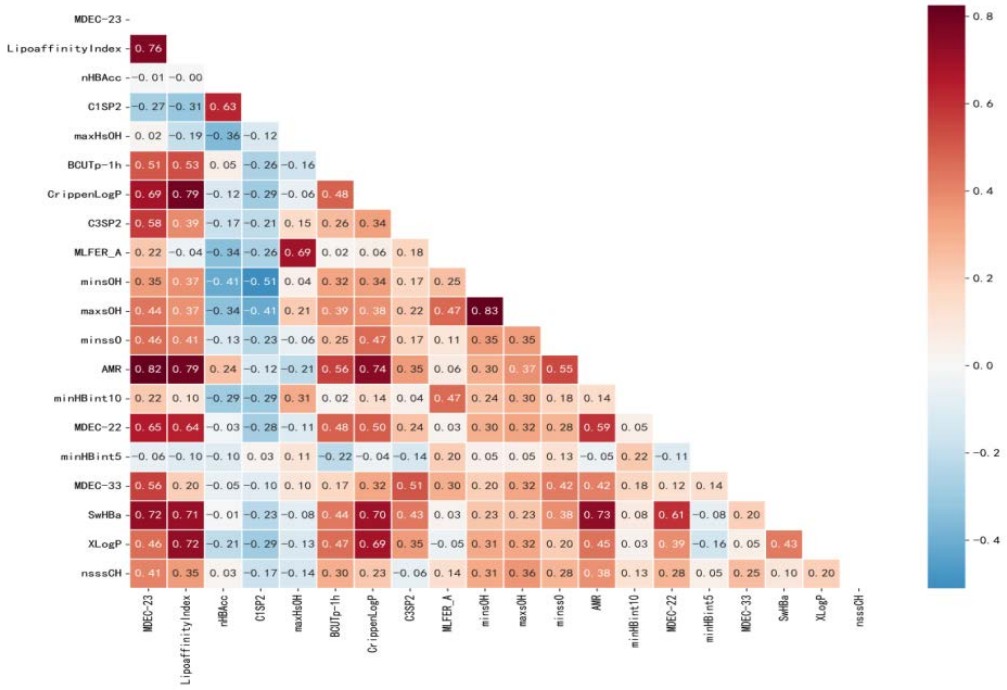

Figure 1. Spearman correlation matrix Heatmap.

3.3. Optuna hyperparameter optimization analytics

Hyperparameter tuning was conducted using the Optuna framework across 30 trial iterations. Figure 2 shows that the cross-validation R^2 score increased rapidly within the first 15 iterations and converged at iteration 25, reaching a peak R^2 of 0.7487. The final parameter settings are shown in Table 4.

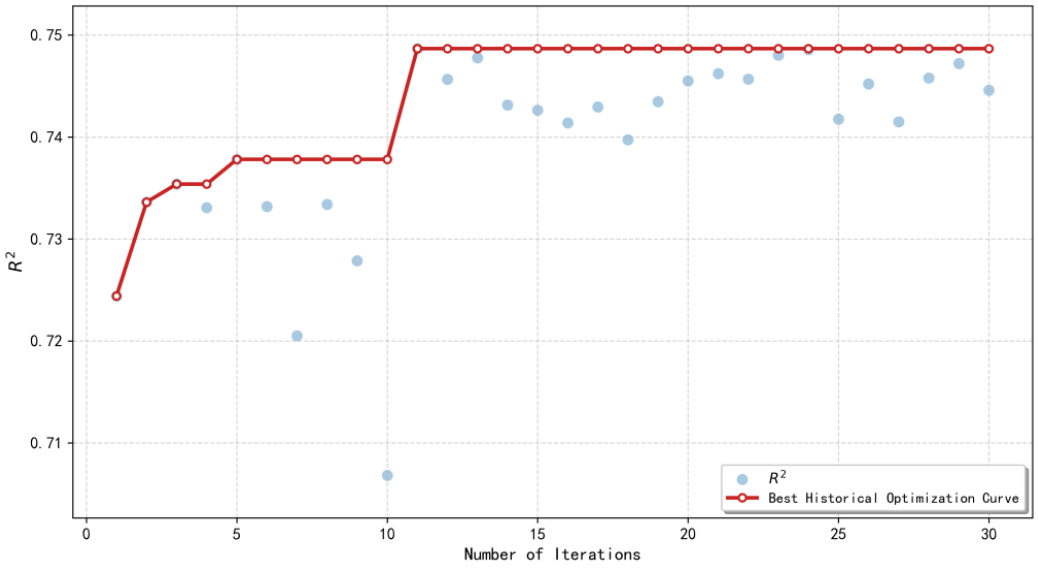

Figure 2. Hyperparameter convergence curve.

Table 4. Optimal hyperparameter configurations.

Learner	Parameter	Value
Random Forest	n_estimators	100
	max_depth	3
	min_samples_leaf	7
XGBoost	learning_rate	0.0856
	n_estimators	150
	max_depth	6
	subsample	0.6002
	colsample_bytree	0.6903
	reg_lambda	8.1853
SVR	C	0.6041
	gamma	0.0022
KNN	n_neighbors	11
	weights	uniform
ElasticNet	alpha	0.0087
	l1_ratio	0.868

3.4. Predictive performance comparison and error analytics

The predictive accuracy was evaluated using four standard statistical metrics: coefficient of determination (R^2), MSE, root mean squared error (RMSE), and mean absolute error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{19}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{21}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{22}$$

The dataset was split into training and testing sets in a 70:30 ratio. Comparative results between the proposed stacking framework and five standalone base models are summarized in Table 5.

As demonstrated in Table 5, the proposed MIFS-Stacking framework consistently outperforms all single base learners across both training and testing datasets. Specifically, on the test set, MIFS-Stacking achieves the highest coefficient of determination (R^2= 0.7375) and the lowest error metrics, with RMSE, MAE, and MSE reaching 0.7339, 0.5344, and 0.5386, respectively.

Overall superiority over baseline models: while linear and kernel-based models like ElasticNet (R^2 = 0.5091) and SVR (R^2 = 0.5043) struggle to adequately capture nonlinear relationships in the descriptor space, tree-based models show better adaptability. Although XGBoost exhibits strong performance on the training set (R^2 = 0.9512), it experiences a performance drop on the test set (R^2 = 0.7269, RMSE = 0.7486), indicating a mild tendency toward overfitting. In contrast, MIFS-stacking balances high fitting capacity (R^2 = 0.9323 on train) with superior generalization performance on unseen data.

Table 5. Predictive performance comparison across models.

Model	Dataset	R^2	RMSE	MAE	MSE
Elastic Net	Train	0.5727	0.9268	0.7354	0.859
	Test	0.5091	1.0037	0.79	1.0074
KNN	Train	0.7433	0.7183	0.5266	0.516
	Test	0.6692	0.8239	0.5994	0.6789
SVR	Train	0.5702	0.9296	0.7332	0.8641
	Test	0.5043	1.0086	0.7875	1.0173
Random Forest	Train	0.6245	0.8688	0.6818	0.7548
	Test	0.5463	0.9649	0.7416	0.9311
XGBoost	Train	0.9512	0.3133	0.2219	0.0982
	Test	0.7269	0.7486	0.5409	0.5604
MIFS-Stacking	Train	0.9323	0.3690	0.2690	0.1362
	Test	0.7375	0.7339	0.5344	0.5386

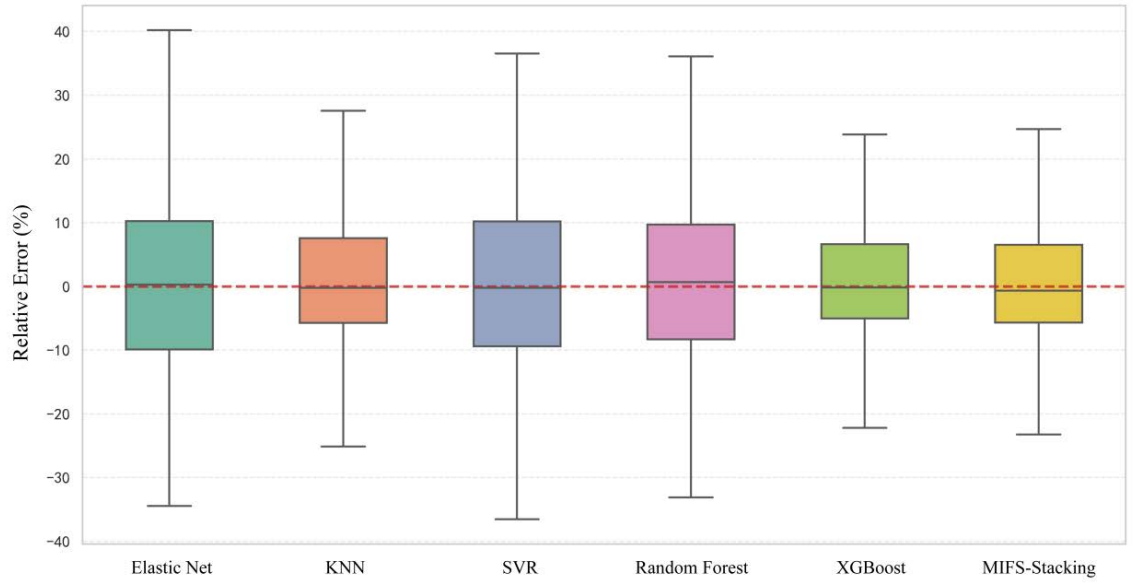

Figure 3. Scatter plot matrix and relative error boxplots.

Error distribution and variance suppression: the relative error boxplots in Figure 3 further substantiate the stability and robustness of the MIFS-stacking model. The median relative error for MIFS-Stacking is closely aligned with the 0% reference baseline, featuring a significantly narrow interquartile range (IQR). Compared to traditional models such as ElasticNet and SVR—which display broad error dispersion and substantial upper/lower extreme whiskers—MIFS-stacking effectively compresses prediction variance and minimizes severe estimation outliers.

In conclusion, the empirical evidence from Table 5 and Figure 3 confirms that the second-level meta-learner in the MIFS-Stacking framework effectively coordinates the complementary strengths of heterogeneous base learners, thereby mitigating individual bias and delivering highly precise and stable bioactivity predictions.

4. Discussion and conclusions

In this study, we proposed the MIFS-Stacking framework to effectively address the challenges of high-dimensional feature redundancy and individual model bias in QSAR modeling [5,6]. By coupling a two-stage feature selection architecture with a regularized heterogeneous ensemble model, our approach achieves superior prediction accuracy, lower prediction variance, and enhanced extrapolation capacity on high-dimensional chemoinformatics data [3,15].

The empirical results validate the methodological efficacy of the proposed framework. The two-stage MIFS module successfully compressed the molecular descriptor space from 729 to 20 core variables [1]. By combining heuristic filtering with a weighted ensemble scoring mechanism across five heterogeneous algorithms, MIFS eliminated single-algorithm evaluation bias and removed severe multicollinearity while preserving critical representational information [1,7,9]. Building upon these core descriptors, the Level 1 primary layer integrated five diverse base learners (RF, XGBoost, SVR, KNN, and ElasticNet) covering linear, topological, and kernelized spatial domains [1,11,12,14]. Coupled with an L2-regularized RidgeCV meta-learner and OOF cross-validation in Level 2, the stacking model effectively suppressed secondary overfitting and captured complex nonlinear relationships [11,15]. Evaluated on 1974 compound samples, MIFS-stacking achieved the highest test set R^2 (0.7375) and lowest RMSE (0.7339), significantly outperforming all standalone baselines [1,16]. Additionally, the integration of Optuna Bayesian optimization enabled rapid parameter convergence within 25 iterations, ensuring an automated and reproducible tuning process.

In summary, the MIFS-stacking framework provides a robust, scalable, and highly accurate computational strategy for high-dimensional QSAR regression. The proposed workflow is inherently generalizable and can be readily extended to broader virtual screening pipelines and ADMET property estimations in computational pharmacology. Future research will explore adapting this framework to multi-target joint predictions and incorporating deep geometric graph learning representations.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by College Students' Innovative Entrepreneurial Training Plan Program under Grant No. 202510356033.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Hansch C, Fujita T, (1964) p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 86: 1616–1626. https://doi.org/10.1021/ja01062a035
2. Franke R, Gruska A, Devillers J, Chessel D, Dunn III WJ, Wold S, et al. (1995) Multivariate data analysis of chemical and biological data. In: *Chemometric Methods in Molecular Design*, H. van de Waterbeemd (Ed.). https://doi.org/10.1002/9783527615452.ch4

3. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al., (2006) DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34: D668–D672. https://doi.org/10.1093/nar/gkj067
4. Lin X, Li X, Lin X, (2020) A review on applications of computational methods in drug screening and design. *Molecules* 25: 1375. https://doi.org/10.3390/molecules25061375
5. Goh GB, Hodas NO, Vishnu A, (2017) Deep learning for computational chemistry. *J Comput Chem* 38: 1291–1307. https://doi.org/10.1002/jcc.24764
6. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 9: 203–214. https://doi.org/10.1038/nrd3078
7. Zou H, Hastie T, (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
8. Guyon I, Weston J, Barnhill S, Vapnik V, (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46: 389–422. https://doi.org/10.1023/A:1012487302797
9. Breiman L, (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324
10. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP, (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 43: 1947–1958. https://doi.org/10.1021/ci034160g
11. Wang JD, Xu YS, Peng F, Xiao BS, (2020) Infrared cooperative localization optimization algorithm based on ridge regression. *J Beijing Univ Aeronaut Astronaut* 46: 563–570. https://doi.org/10.13700/j.bh.1001-5965.2019.0238
12. Smola AJ, Schölkopf B, (2004) A tutorial on support vector regression. *Stat Comput* 14: 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
13. Chen T, Guestrin C, (2016) XGBoost: A scalable tree boosting system, In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
14. Zhou JY, He PF, Qiu RF, Chen G, Wu WG, (2021) Research on intrusion detection using a hybrid approach combining random forests and gradient boosting trees. *J Software* 32: 3254–3265. http://dx.doi.org/10.13328/j.cnki.jos.006062
15. Wolpert DH, (1992) Stacked generalization. *Neural Networks* 5: 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
16. He B, Gan JY, (2024) Prediction model of anti-breast cancer drug activity based on machine learning. *Comput Sci Appl* 14: 298–307. https://doi.org/10.12677/CSA.2024.142030
17. Xu N, La L, (2019) Prediction of new retail coupon usage behavior based on XGBoost. *J Southwest China Norm Univ Nat Sci Ed* 44: 101–105. https://doi.org/10.13718/j.cnki.xsxb.2019.03.017
18. Wu XY, Wang SH, Zhang YD, (2017) A review of the theory and applications of the k-nearest neighbors algorithm. *Comput Eng Appl* 53: 1–7. https://doi.org/10.3778/j.issn.1002-8331.1707-0202

AIMS Press