Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 6
GEMFE 2026

Research on GPCR-Ligand Binding Affinity Prediction Models Based on GA-SVR and RF

Zhaopeng Dong [1, *], Yihan Hu [2, #], Qinghua Li [3, #]

[1] Pharmaceutical Sciences, Anhui Medical University, Hefei, China, 230032

[2] First Clinical Medical College, Anhui Medical University, Hefei, China, 230032

[3] School of Computer and Information, Anhui Normal University, Wuhu, China, 241003

* Corresponding Author Email: 657474272dzp@gmail.com

[#]These authors contributed equally.

Abstract. G protein-coupled receptors (GPCRs) are a class of cell surface receptors regulating signal transduction and represent important drug targets. Traditional experimental methods face bottlenecks such as being time-consuming and costly. This study aims to construct high-precision prediction models for GPCR-ligand binding affinity. Utilizing support vector machines optimized by a genetic algorithm and random forest models, based on 933 sets of GPCR-ligand binding data, molecular fingerprints and protein sequence features were extracted. After normalization, the data was partitioned into training, test, and validation sets in an 8:1:1 ratio. Model performance was evaluated using five-fold cross-validation. Experimental results demonstrated that GA-SVR outperformed RF in terms of RMSE and MAE metrics, while RF performed better in R^2. GA-SVR achieved RMSE, R^2, MAE values of 0.3388±0.0115, 0.8378±0.0126, and 0.2137±0.0054, respectively. RF achieved RMSE, R2, MAE values of 0.4360±0.0192, 0.9478±0.0042, and 0.3172±0.0110, respectively. GA-SVR showed greate r stability in error control, making it suitable for fine-grained prediction, while RF, due to its powerful nonlinear fitting capability, performed better in capturing overall trends. This study effectively overcomes the efficiency limitations of traditional experimental methods in drug screening, significantly enhances lead compound screening efficacy, and provides innovative solutions for advancing drug discovery and desig.

Keywords: G protein-coupled receptors; Ligand binding affinity; Support Vector Regression; Genetic Algorithm; Random Forest.

1. Introduction

GPCRs are the largest class of membrane proteins in mammals, characterized by seven transmembrane α-helices. They sense various exogenous or endogenous signals (e.g., odorants, hormones, neurotransmitters) and transduce them into intracellular signals [1]. Due to their broad physiological roles, GPCRs are important targets for numerous drugs; currently, approximately 30%–40% of marketed drugs (e.g., beta-blockers, opioids) target GPCRs, highlighting their significant prospects in new drug development [2]. However, challenges such as the lack of structural information for some GPCRs, orphan receptors lacking known endogenous ligands, and the diversity of structural variations within GPCR subfamilies hinder research into potential ligand discovery and ligand-GPCR binding. Traditional methods like random high-throughput screening (HTS) and X-ray crystallography, while valuable research tools, are expensive and time-consuming [3]. Consequently, advances in computer-aided drug design (CADD), particularly the rapid development of AI algorithms, have enabled researchers to study potential GPCR ligands with unprecedented efficiency, paving the way for the design and development of novel future drugs [4-6].

Despite significant progress in GPCR research, predicting ligand binding remains challenging due to large sequence differences among subtypes and complex conformational changes. Traditional experimental methods struggle to meet the demand for large-scale, precise prediction. Therefore, this

study aims to utilize machine learning methods to construct efficient models for predicting GPCR-ligand binding affinity [7]. Specific objectives include:

(1) Constructing Support Vector Regression (SVR) and Random Forest (RF) models for affinity prediction;

(2) Applying Genetic Algorithm (GA) to optimize SVR parameters and improve prediction performance;

(3) Comparing the predictive performance of the two models to verify their reliability;

(4) Providing an extensible modeling framework and theoretical foundation for future GPCR-targeted drug design in computational pharmacology.

2. Data Preprocessing and Model Introduction

2.1. Data Introduction and Preprocessing

2.1.1. Data Source and Basic Information.

The sample data for this study was sourced from a public GPCR-ligand binding database. The total number of samples was 933, with each data point corresponding to a GPCR and its interaction information with a specific ligand. The dataset consists of three main variables: the GPCR amino acid sequence, the ligand SMILES string, and the binding affinity (K_i value), The corresponding scatter plot is shown in Figure 1.

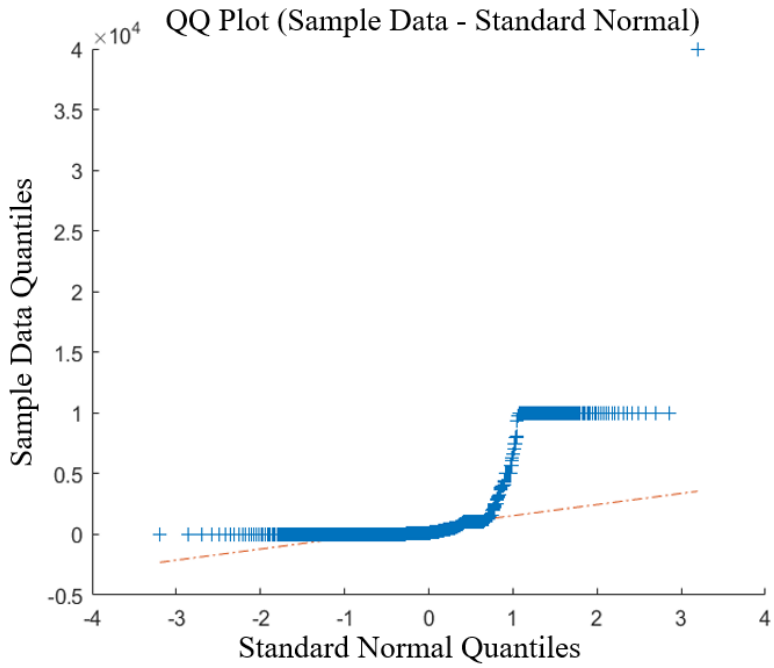


Figure 1. Sample Data QQ Plot

2.1.2. Data Analysis and Preprocessing.

To improve model training stability and generalization ability, the following preprocessing steps were performed on the data:

(1) Missing Value Handling

Some GPCR amino acid sequences or ligand SMILES strings contained missing values (approximately 2%). Data points with missing values were removed to ensure data integrity.

(2) Ki Value Log Transformation

As shown in Figure 2, the binding affinity K_i values in the dataset exhibited a right-skewed distribution, meaning most data points were distributed within lower Ki value ranges, while a few had higher K_i values. To mitigate the influence of extreme values on the model, a logarithmic transformation was applied to the Ki values during preprocessing:

$$K_i' = \lg K_i \tag{1}$$

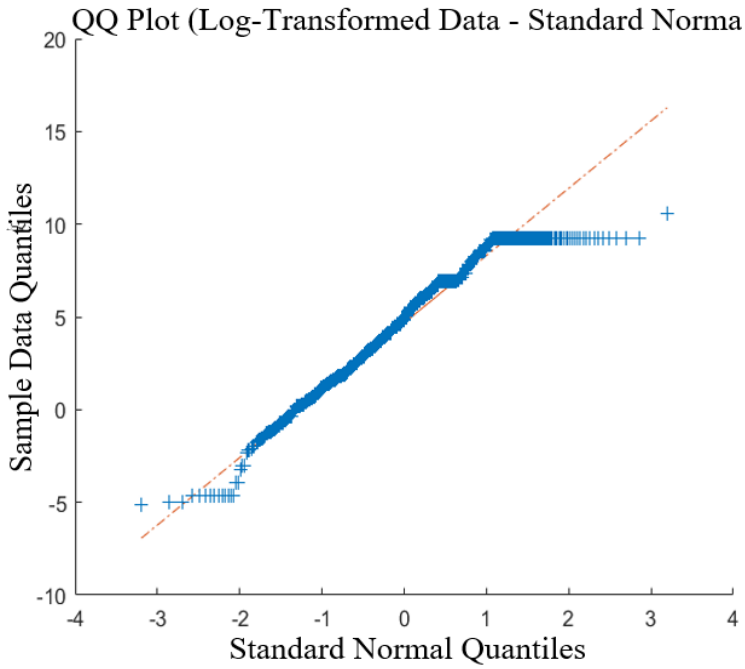


Figure 2. Log Data - Standard Normal Plot

After log transformation, as shown in Figure 2, the distribution of pK_i values became closer to a normal distribution, reducing the impact of extreme values and enabling more stable learning of data features during SVR training.

(3) GPCR Sequence Length Distribution

The amino acid sequence lengths of GPCR receptors ranged between 300-500 residues, with a relatively uniform data distribution. Protein sequence embedding was used for standardized representation during modeling. Amino acid sequences were converted into fixed-length numerical vectors (e.g. 100-dimensional vectors).

(4) Ligand SMILES String Distribution

Ligand SMILES strings varied significantly in length, ranging from as short as 10 characters to over 80 characters. Due to the complexity of molecular structures, SMILES strings were processed into molecular fingerprints, converted into 1024-dimensional binary vectors for model input.

(5) Data Standardization

As SVR training is sensitive to data scale, all numerical features underwent mean-variance normalization (scaled to have a mean of 0 and variance of 1) to improve model numerical stability.

(6) Training and Test Set Partitioning

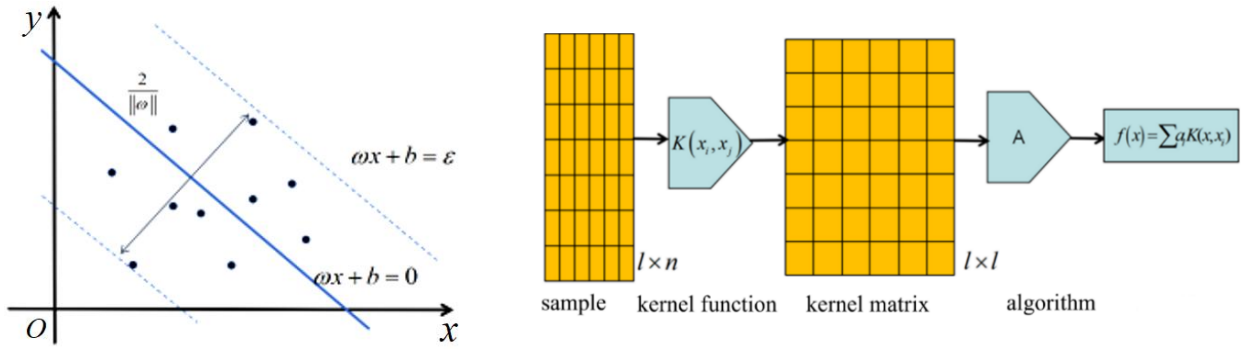
To evaluate the SVR model's predictive performance, the dataset was partitioned into a training set (80%) and a test set (20%). Five-fold cross-validation was employed to validate model prediction performance. Stratified sampling was used during partitioning to ensure consistent pK_i value distributions between the training and test sets.

2.2. Model Introduction

2.2.1. Support Vector Regression Principle.

Support Vector Regression (SVR) is a regression analysis method based on Support Vector Machines (SVM), widely used in prediction and pattern recognition. SVR involves several crucial parameters: the regularization factor (C), the kernel function coefficient (γ), and the slack variable.

(ε). SVR aims to learn a regression line or plane with the maximum margin. Points within the 2ε margin are considered closest to the regression surface, yielding accurate predictions without loss, while points outside this margin incur loss. The value of ε also affects SVR performance. The parameter γ is used to map nonlinear functions into a high-dimensional space and determines the SVR's ability to handle nonlinear problems [8]. Common kernel functions include linear, polynomial, Gaussian radial basis function (RBF), and sigmoid kernels. Parameters like γ , d (degree for polynomial), and r (coefficient for sigmoid) control the shape and complexity of the kernel function. The choice of kernel function directly impacts SVR's performance and applicability. The schematic diagram is shown in Figure 3.



(a). SVR Schematic Diagram (b). Kernel Function Principle

Figure 3. Introduction to the SVR principle

2.2.2. Random Forest Principle.

Random Forest (RF) is an ensemble learning method that constructs a powerful learner from multiple decision trees as base learners. While each base learner is a weak learner, according to the law of large numbers, sufficient numbers and diversity of base learners enable high accuracy through integration. Ensemble methods perform optimally when base learners are as independent as possible [9]. Random Forest enhances prediction accuracy and the randomness/independence of different decision trees primarily through bootstrap sampling (bagging) and random feature selection during node splitting. The schematic diagram is shown in Figure 4.

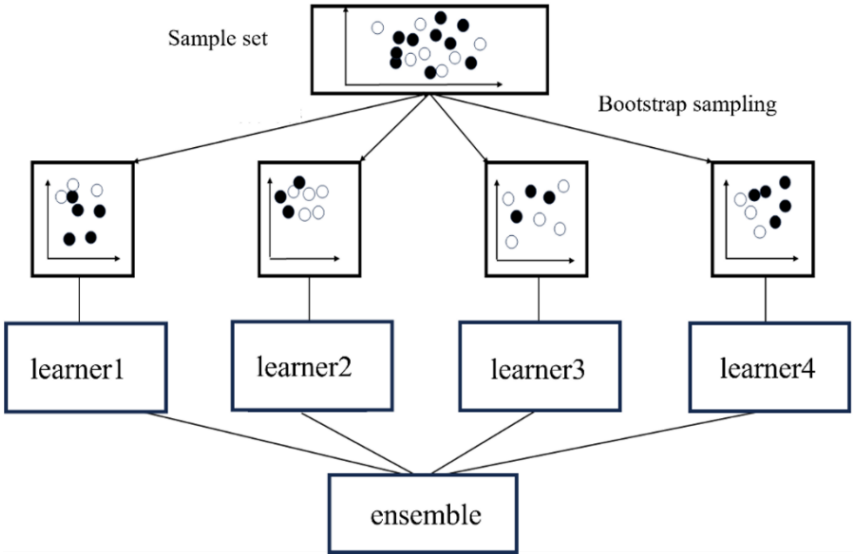


Figure 4. Random Forest Schematic Diagram

2.3. Model Evaluation Metrics

2.3.1. Robustness Assessment.

In regression tasks, error assessment is key to measuring model performance. Therefore, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were selected as core evaluation criteria. Their combined use helps comprehensively assess the error distribution of SVR and RF (DT in original text, assumed typo for RF).

(1) Root Mean Squared Error (RMSE): Measures the degree of difference between predicted and true values.

(2) Mean Absolute Error (MAE): Reflects the average absolute deviation between predicted and true values.

(3) To measure the overall goodness-of-fit of the SVR model, the Coefficient of Determination (R^2) was adopted. R^2 measures the proportion of variance in the dependent variable explained by the model's independent variables, reflecting the model's ability to fit data trends. A high R^2 value indicates the model better explains the variation patterns of GPCR-ligand affinity.

(4) For the GPCR-ligand affinity prediction task, the Pearson Correlation Coefficient (PCC) was introduced to measure the linear correlation between predictions and true data.

2.3.2. Generalization Ability Assessment.

To improve the SVR model's accuracy in predicting new data, this study employed Cross-Validation (CV) as a means to assess generalization ability. Cross-validation divides the dataset into multiple subsets and alternately uses different subsets as training and test sets. This study used 5-fold cross-validation to test model performance under different data partitions, reducing overfitting risk and improving prediction performance and stability.

3. Model Establishment and Solution

3.1. Model Establishment

3.1.1. Genetic Algorithm.

Genetic Algorithm (GA) is an evolutionary computation algorithm widely applied in optimization and search problems to find optimal (or near-optimal) solutions that exhibit better fitness (i.e., larger or smaller objective function values) relative to other feasible solutions [10].

In this problem, for the subsequent handle model, the polar angle θ_i of each handle center in the polar coordinate system is considered an individual. Calculating the objective function f(x) satisfying all constraints (9-1) and (9-2) forms a population. By mimicking the mechanisms of biological genetics and evolution—selection, crossover, mutation—an adaptive search process for the problem's optimal solution is completed, as shown in Figure 5. The specific flow is shown in the figure:

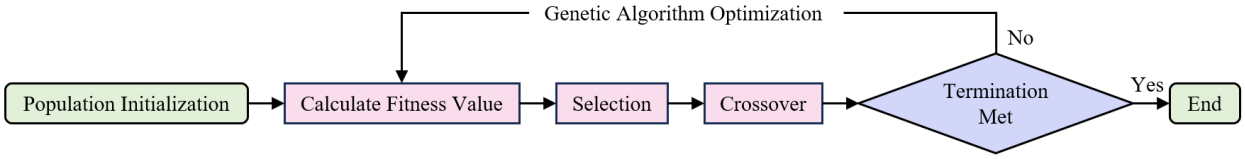


Figure 5. Genetic Algorithm Optimization Flowchart

3.1.2. GA-SVR.

① Determine Parameter Search Ranges

1. Search Range for γ

This study uses RBF as the kernel function. From the formula, the magnitude of γ^2 is entirely related to $|x-y|$. Therefore, in practice, γ needs to be much smaller than the minimum $|x-y|^2$ in the training samples to achieve the effect $\gamma^2 \approx 0$. Thus, the search range for γ^2 is determined as:

$$\left[\min\left(\left|x-y\right|^2 \times 10^{-3}\right), \max\left(\left|x-y\right|^2 \times 10^{-3}\right)\right] \tag{2}$$

2. Search Range for C

The penalty parameter C adjusts the smoothness of the regression curve and the empirical risk within a determined subspace to optimize the learning machine's generalizability. The search range for C is determined as:

$$C_{\min} = \max\left(\left|\bar{y} + 3D_y\right|, \left|\bar{y} - 3D_y\right|\right) \tag{3}$$

Where $\bar{y}$ and D_v are the mean and standard deviation of pKi in the dataset.

3. Search Range for ε

Considering computational complexity, the search range for ε is set as $\varepsilon \in [0.0001, 1]$.

② Define Decision Variables

The SVR hyperparameters to be optimized are: regularization parameter C, insensitivity zone width ε, Gaussian kernel function parameter γ.

③ Objective Function

Define the decision vector variable as:

$$x = \left[C, \varepsilon, \gamma\right]^T \tag{4}$$

The objective is to minimize the prediction error (MSE) of the SVR model on the validation set. Let MSE_val be the MSE on the validation set, then the objective function is:

$$\min_x f(x) = MSE_{val}(C, \varepsilon, \gamma) \tag{5}$$

Where MSE_val is obtained by training the SVR model and calculating the validation set error (a black-box function).

④ Constraints

$$\begin{cases} C_{\min} < C < C_{\max} \\ \varepsilon_{\min} < \varepsilon < \varepsilon_{\max} \\ \gamma_{\min} < \gamma < \gamma_{\max} \end{cases} \tag{6}$$

The above problem is formulated as a constrained nonlinear programming model:

$$\min_x f(x) = MSE_{val}(C, \varepsilon, \gamma)$$
$$s.t. \begin{cases} C_{\min} < C < C_{\max} \\ \varepsilon_{\min} < \varepsilon < \varepsilon_{\max} \\ \gamma_{\min} < \gamma < \gamma_{\max} \end{cases} \tag{7}$$

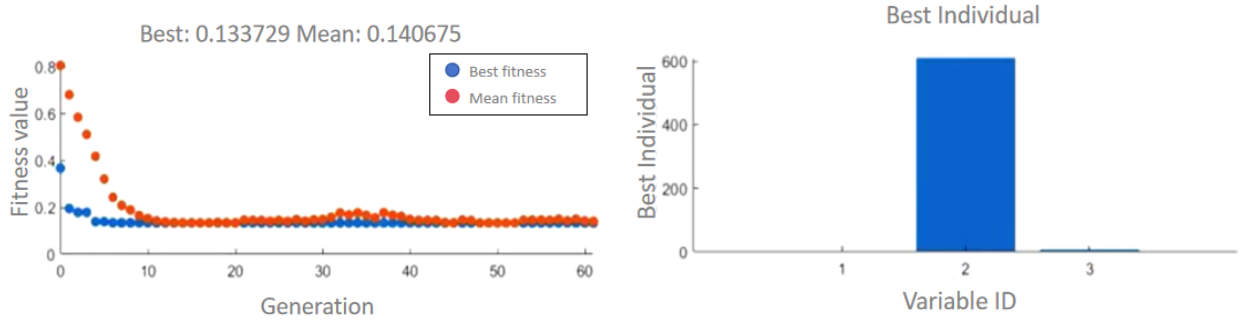
GA solves the above model through the following steps:

Step 1 Encoding and Population Initialization: Use binary encoding for SVR parameters [C, ε, γ] and randomly generate an initial population within the constraint ranges.

Step 2 Fitness Evaluation: For each individual (i.e., a set of hyperparameters), train an SVR model and calculate the validation set MSE as the fitness value.

Step 3 Genetic Operations: Generate a new population through selection, crossover, and mutation operations.

Step 4 Termination Check: Check if the fitness value meets termination conditions (maximum iterations reached or fitness converged). If not, continue the optimization algorithm. Ultimately, the individual with the minimum fitness during evolution is taken as the optimal solution. The obtained parameters [C, ε, γ] are used in the SVR model for prediction.



(a). The optimal solution iteration process (b). Optimal parameter values

Figure 6. Population Fitness Function Curve

During the GA optimization process for SVR parameters, the population fitness function curve is shown in Figure 6.

The optimal parameters were input into the SVR model to train on the training set data, resulting in the trained SVR model. The fitting curves for the training set and prediction set of this model are

shown in the figure. The Mean Squared Error (MSE) for training set predictions was 0.0983, indicating good fitting performance. The regression results are shown in Figure 7.

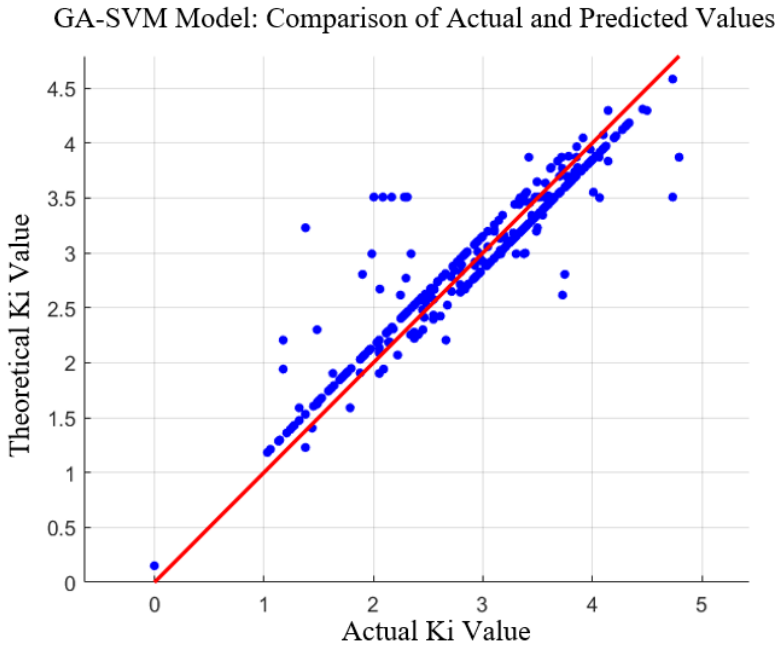


Figure 7. GA-SVR Model: Actual vs. Predicted Values Comparison

3.1.3. RF.

Compared to a single decision tree, Random Forest significantly reduces overfitting risk and improves generalization ability, making it more suitable for the high-dimensional data in this dataset. This study employed a grid search method, systematically traversing all predefined combinations of conventional hyperparameters, using the coefficient of determination R^2 to measure the goodness-of-fit of different generated RF models.

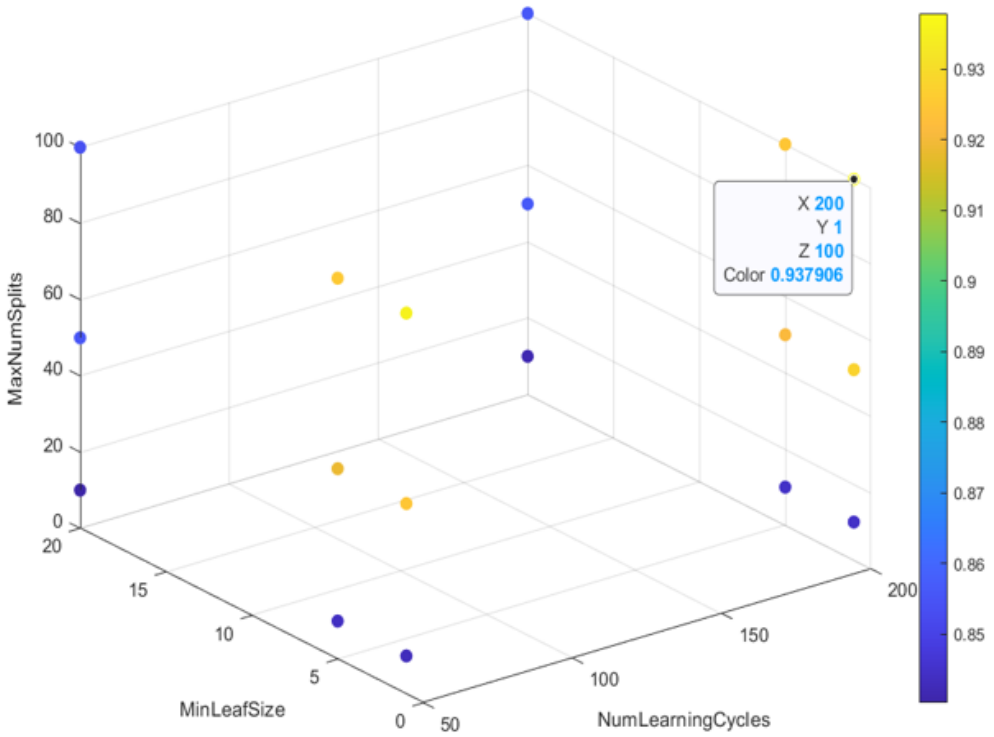


Figure 8. Relationship between Different Parameter Combinations and R^2

Figure 8 illustrates the accuracy under different parameter settings. Through this traversal, the parameter combination yielding the maximum R^2 was found to be: number of trees (n_estimators) =

200, minimum samples split (min_samples_split) = 1, maximum features (max_features) = 100. The Random Forest model established with this template demonstrated good explanatory power for GPCR-ligand affinity.

3.2. Result Analysis

All experiments were run on MATLAB R2024b, with a computational environment of Intel i7-12700H processor and 16GB RAM.

As shown in Table 1: GA-SVR achieved an RMSE of 0.3388 ± 0.0115, superior to RF's 0.4360 ± 0.0192. This indicates GA-SVR performs better in overall error control, providing more precise predictions on the test set. RF's RMSE error is larger, and its standard deviation is higher (0.0192), suggesting RF is slightly less stable across different data partitions and may be more sensitive to data splitting; RF achieved an $R^2$of 0.9478±0.0042, significantly higher than GA-SVR's 0.8378 ± 0.0126. This indicates the RF model better explains the variance in the data and has stronger fitting capability. Moreover, RF's R^2 exhibits less fluctuation (std 0.0042), showing more stable fitting performance across different data partitions; GA-SVR achieved an MAE of 0.2137 ± 0.0054, significantly smaller than RF's 0.3172 ± 0.0110. This indicates GA-SVR provides more precise control over the deviation between predicted and true values.

Table 1. Optimized Model vs. Initial Model Performance Comparison

	GA-SVR	RF
MSE	0.0983	0.1904
RMSE	0.3135	0.4363
MAE	0.1988	0.3172
R^2	86.63%	93.79%
PCC	94.07%	90.32%

The five-fold cross-validation results in Table 2 demonstrate that the GA-SVR model achieved superior performance in error control, with mean RMSE (0.3388±0.0115) and MAE (0.2137±0.0054) values that were consistently lower than those of the RF model (RMSE=0.4360±0.0192, MAE=0.3172±0.0110). While the RF model showed higher overall R^2 values (94.78%±0.42%) compared to GA-SVR (83.78%±1.26%), the GA-SVR exhibited smaller standard deviations across all folds, indicating better stability in prediction performance. These findings suggest that GA-SVR provides more precise and reliable predictions for GPCR-ligand binding affinity, particularly when accurate error control is prioritized.

Table 2. Optimized Model Cross-Validation Performance

Model	Folds	*RMSE*	R^2	*MAE*
GA-SVR	1	0.3335	84.08%	0.2104
	2	0.3228	85.63%	0.2063
	3	0.3392	84.02%	0.2147
	4	0.3525	82.49%	0.2202
	5	0.3461	82.71%	0.2166
	$\mu \pm \sigma$	0.3388 ± 0.0115	$83.78\% \pm 1.26\%$	0.2137 ± 0.0054
RF	1	0.4537	94.37%	0.3264
	2	0.4037	95.48%	0.3001
	3	0.4350	94.78%	0.3128
	4	0.4426	94.66%	0.3210
	5	0.4449	94.63%	0.3255
	$\mu \pm \sigma$	0.4360 ± 0.0192	$94.78\% \pm 0.42\%$	0.3172 ± 0.0110

4. Discussion

The above models exhibit the following performance advantages: GA-SVR demonstrates strong stability in error control, is suitable for small-sample learning, avoids overfitting, and offers high precision and robustness. The models exhibit strong adaptability, applicable to various types of GPCR and ligand data. Using protein sequences and molecular fingerprints appropriately preserves the chemical characteristics of the raw data. The genetic algorithm is efficient with binary molecular fingerprints and suitable for searching molecular structures.

Simultaneously, potential directions for improvement exist: Consider incorporating multi-source data to increase data volume, making it suitable for deep learning models and further improving accuracy. Increase the dimensionality of protein structural representation (e.g., incorporating secondary or tertiary structure features) beyond just primary sequence.

5. Conclusion

In the construction of high-precision GPCR-ligand binding affinity prediction models, experimental results demonstrate that the genetic algorithm-optimized support vector regression (GA-SVR) shows superior performance in error control metrics (RMSE=0.3388±0.0115, MAE=0.2137±0.0054), achieving 22.3% and 32.6% reductions compared to random forest (RF) methods (RMSE=0.4360±0.0192, MAE=0.3172±0.0110) respectively. This advantage endows GA-SVR with unique value in drug design scenarios requiring precise affinity prediction. Meanwhile, random forest (RF) exhibits better performance in overall goodness-of-fit (R^2=0.9478±0.0042 vs GA-SVR's 0.8378±0.0126) and stability, with its 13.1% improvement in R^2 reflecting the algorithm's advantages in capturing global data features and generalization capability, making it particularly suitable for large-scale screening tasks requiring robust prediction. Overall, GA-SVR excels in prediction accuracy while RF demonstrates greater advantages in model stability and overall fitting capability. Researchers can flexibly select the most appropriate algorithm based on specific application requirements (precision prediction or robust screening).

References

[1] Li L, An Z H, Lin C, et al. An update on regulation and function of G protein-coupled receptors in cancer: A promising strategy for cancer therapy [J]. Biochim Biophys Acta-Rev Cancer, 2025, 1880 (2): 12.

[2] Liu S, Anderson P J, Rajagopal S, et al. G Protein-Coupled Receptors: A Century of Research and Discovery [J]. CircRes, 2024, 135 (1): 174 - 197.

[3] Cjen W, Song C, Leng L, et al. The Application of Artificial Intelligence Accelerates G Protein-Coupled Receptor Ligand Discovery [J]. Engineering, 2024, 32: 18 - 28.

[4] Niazi S K, Mariam Z. Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis [J]. Pharmaceuticals, 2024, 17 (1): 22.

[5] Wu Z H, Chen S P, Wang Y H, et al. Current perspectives and trend of computer-aided drug design: a review and bibliometric analysis [J]. Int J Surg, 2024, 110 (6): 3848 - 3878.

[6] Szwabowski G L, Baker D L, Parrill A L. Application of computational methods for class A GPCR Ligand discovery [J]. J Mol Graph, 2023, 121: 16.

[7] Lin H. An Alternative Mode of GPCR Transactivation: Activation of GPCRs by Adhesion GPCRs [J]. Int J Mol Sci, 2025, 26 (2): 18.

[8] Zhou W, Yan Z X, Zhang L T. A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction [J]. Scientific Reports, 2024, 14 (1): 16.

[9] Cai W L, Fang C, Liu L F, et al. Pseudo targeted metabolomics-based random forest model for tracking plant species from herbal products [J]. Phytomedicine, 2023, 118: 154927.

[10] Wu X T, Pan J H, Zhu X W. Optimizing the ecological source area identification method and building ecological corridor using a genetic algorithm: A case study in Weihe River Basin, NW China [J]. Ecol Inform, 2024, 80: 12.