frontiers | Frontiers in Pharmacology

TYPE Methods
PUBLISHED 15 November 2023
DOI 10.3389/fphar.2023.1260349

OPEN ACCESS

EDITED BY
Wenying He,
Hainan Normal University, China

REVIEWED BY
Ailin Zhao,
Sichuan University, China
Amit Kumar Halder,
University of Porto, Portugal

*CORRESPONDENCE
Peijian Zhang,
peijianzh@126.com

RECEIVED 17 July 2023
ACCEPTED 30 October 2023
PUBLISHED 15 November 2023

CITATION
Wang Y and Zhang P (2023), Prediction of histone deacetylase inhibition by triazole compounds based on artificial intelligence.
Front. Pharmacol. 14:1260349.
doi: 10.3389/fphar.2023.1260349

Prediction of histone deacetylase inhibition by triazole compounds based on artificial intelligence

Yiran Wang and Peijian Zhang*

College of Computer Science and Technology, Qingdao University, Qingdao, Shandong Province, China

A quantitative structure-activity relationship (QSAR) study was conducted to predict the anti-colon cancer and HDAC inhibition of triazole-containing compounds. Four descriptors were selected from 579 descriptors which have the most obvious effect on the inhibition of histone deacetylase (HDAC). Four QSAR models were constructed using heuristic algorithm (HM), random forest (RF), radial basis kernel function support vector machine (RBF-SVM) and support vector machine optimized by particle swarm optimization (PSO-SVM). Furthermore, the robustness of four QSAR models were verified by K-fold cross-validation method, which was described by Q^2. In addition, the R^2 of the four models are greater than 0.8, which indicates that the four descriptors selected are reasonable. Among the four models, model based on PSO-SVM method has the best prediction ability and robustness with R^2 of 0.954, root mean squared error (RMSE) of 0.019 and Q^2 of 0.916 for the training set and R^2 of 0.965, RMSE of 0.017 and Q^2 of 0.907 for the test set. In this study, four key descriptors were discovered, which will help to screen effective new anti-colon cancer drugs in the future.

KEYWORDS

cancer, HDAC inhibition, quantitative structure-activity relationship, support vector machine, particle swarm optimization

1 Introduction

Colon and rectal cancers are the most common gastrointestinal tumors (Chen et al., 2019; Pan et al., 2023). Currently, colon cancer is one of the most common solid malignancies, which is the third leading cause of cancer-related new cases and deaths worldwide (Chen et al., 2019; Shi et al., 2022a; Shi et al., 2022b). Most patients with colon cancer present with advanced disease, whose survival rate is very low. More than 95% of the patients with colorectal cancer in the diagnosis of aged 50 or older, and the overall survival rate for advanced, metastatic, and recurrent colon cancer is less than 50%. (Wang et al., 2015; Shi et al., 2022a). Due to the numerous factors in the development and progression of colon cancer, its pathogenesis is still unclear. Current treatment options for colon cancer include surgery, chemotherapy, and molecularly targeted therapy. Treatments for colon cancer are limited, and a large percentage of patients develop resistance to current treatments (Schmoll and Stein, 2014). Therefore, the study of new therapeutic strategies and the development of new drugs are the key points in the research of colon cancer (Xiong et al., 2023).

At present, an emerging therapeutic approach for colon cancer is the use of corresponding histone deacetylase (HDAC) inhibitors (Tavares et al., 2017). HDAC is a kind of epigenetic antitumor drug targets (Choi et al., 2023). Because of the important role of HDAC in various biological processes such as cell proliferation, metastasis and apoptosis,

HDAC inhibitors has been widely studied as a novel anticancer drug target (Gillette, 2021; Roy et al., 2023). Though HDAC inhibitors have not been approved by FDA to treat colon cancer, some preclinical studies have discovered its efficacy to treat colon cancer *in vitro* and *in vivo* (Kang et al., 2009; Asklund et al., 2012; Yao et al., 2014). However, there are some limitations of current approved HDAC inhibitors, such as pan-inhibition, etc. Thus, it is in urgent need to develop novel HDAC inhibitors to improve colon cancer treatment (Place et al., 2005; Sang et al., 2023) Nan Sun et al. designed and synthesized a series of triazole-containing compounds as novel HDAC inhibitors, which have significant anti-proliferation effect on murine and human colon cancer cell lines MC38 and HCT116 (Sun et al., 2023). In the studies of discovering new HDAC inhibitors, the measurement of HDAC inhibition IC_{50} values of compounds has great influence for the design of new effective anti-colon cancer drugs. Since numerous chemical experiments are costly and time-consuming, a new and effective method for predicting the IC_{50} of untested compounds should be found (Zhao et al., 2020).

In 1964, the concept of the quantitative structure-activity relationship (QSAR) was first proposed by Free et al. and then was widely used (Free and Wilson, 1964; Hansch and Steward, 1964; Myint and Xie, 2010). QSAR is based on the general principle of medicinal chemistry that the biological activities of a ligand or compound is related to its molecular structure or properties, and molecules with similar structures may have similar biological activities (Myint and Xie, 2010). Model established based on QSAR can reveal quantitative structure-activity relationship between biological activities and part of descriptors of set of known compounds with similar structure (Huang et al., 2021). Then QSAR model can be used to predict the activities of unknown compounds that have similar structure with the previous known compounds, which has widely used in the process of screening out efficient and novel drugs (Sun et al., 2023).

Therefore, four QSAR models were established in this study to predict the HDAC inhibition IC_{50} of 60 selected compounds based on descriptors selected by the heuristic method (HM). The methods establishing models in study are HM, random forest (RF), support vector machine with radial basis kernel function (RBF-SVM) and RBF kernel function support vector machine with particle swarm optimization (PSO-SVM). In addition, the HM method was also used to select descriptors. Among the four models, the model constructed by PSO-SVM has the best performance and strongest robustness. In addition, the R^2 of the established four models meet the requirement of predicting IC_{50} of compounds, which indicates the descriptors used in the models were enough and good for drug design. Overall, this study will provide efficient guidance and help for the screening of a new type of colon cancer drug.

2 Material and methods

2.1 Data set

The 60 compounds containing triazole studied in this paper are all from the same literature, which eliminates unforeseen problems due to data from different sources[17]. The IC_{50} value of HDAC inhibition was determined by Sun et al. after exposing the compound to the same experimental environment for 72 h to ensure the accuracy of the experiment[17]. The compounds used in this study and their IC_{50} values are shown in Supplementary Table S1. All compounds were randomly divided into the training set and test set in the ratio of 4:1, of which 48 compounds in training set were used to construct models and the remaining 12 compounds in test set were used to evaluate the performance of the models (Fan et al., 2018).

2.2 Calculating descriptors

The calculation of molecular descriptors and the selection of appropriate molecular descriptors are the key prerequisites of establishing QSAR models, which directly affects the performance of QSAR models, such as accuracy of prediction. Comprehensive descriptors for structural and statistical analysis (CODESSA) is currently the more common package to calculate molecular descriptors and perform statistical analysis (Katritzky et al., 2005; Zhang et al., 2013). The steps for calculating and selecting molecular descriptors are as follows. Firstly, ChemDraw was used to draw the structure of the compounds to get mol file and skc file (Evans, 2014). Secondly, the compounds in mol format were optimized using HyperChem software under the guidance of the theory that the lower the energy of a molecule is, the more stable its structure is. In the optimization process, MM+ molecular mechanical force field was used to preliminarily optimize the compound, and then semi-empirical AM1 method was used to further optimize the compound to obtain the most stable structure, which is beneficial to improve the calculation accuracy of molecular descriptors (Lima et al.). The optimized structures by HyperChem were stored in hin and zmt formats as the input of MOPAC. After that, MOPAC was used to generate mno files as the input files of CODESSA to calculate descriptors. Five categories of molecular descriptors were obtained using CODESSA, which are structural, topological, geometric, electrostatic, and quantum chemical (Coi et al., 2006; Madugula and Yarasi, 2017).

2.3 Linear model by HM

HM is a common and effective method for selecting descriptors which was widely used at present (Wang et al., 2020; Gao et al., 2022). This method is not limited by the size of data set, and is highly efficient (Yang et al., 2023; Li et al., 2023). Therefore, HM was used to select the appropriate descriptor from the specific descriptors computed by CODESSA. This method can select the descriptors responsible for activity from the descriptor set by building a multiple linear regression models. The following descriptors should be excluded before building linear regression models by HM. 1) Special descriptors that not all compounds possess. 2) Descriptors with correlation coefficients greater than 0.8, namely collinear descriptors.

In this study, square of correlation coefficient (R^2) and root mean square error (RMSE) were used to evaluate and analyze the performance of models established by HM, RF, SVM and PSO-SVM. In addition, the robustness of the models were verified by K-fold cross-validation.

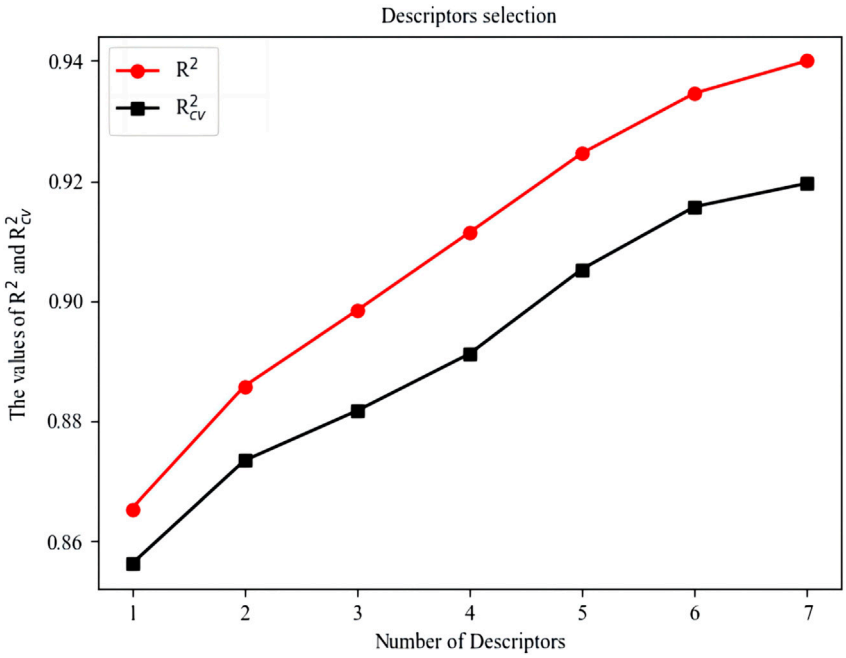

FIGURE 1
Influence of the Number of descriptors on R^2 and R_{cv}^2.

2.4 Nonlinear model by RF

The IC_{50} value of HDAC inhibition is influenced by many factors, so the linear model cannot accurately predict the IC_{50} of HDAC. Therefore, three kinds of nonlinear models were established by RF, SVM and PSO-SVM. Random forest is a supervised machine learning algorithm based on ensemble learning (Feng et al., 2019). It can effectively reduce the risk of overfitting and is more conducive to obtain a robust model. Therefore, it is a novel and efficient method to establish QSAR nonlinear models (Zhang et al., 2015a; Fang et al., 2022.).

The steps to build a RF regression model are: 1) First, build a dataset containing the descriptor values and -lg (IC_{50}) values for training set and test set. 2) Build RF model. Set the number of decision trees to control the RF's behavior (Li et al., 2012; Song, 2015; Ao et al., 2019). 3) Train the RF using the training set. RF method constructed multiple decision trees based on descriptor data and IC_{50} values in training set, and performs feature selection and partitioning in each tree. 4) Use the trained RF model to predict the samples in the test set. In this model, the prediction results of each decision tree were weighted to obtain the predicted IC_{50} value. 5) Use a variety of performance indicators to evaluate prediction accuracy and generalization ability of RF model.

2.5 Nonlinear model by RBF-SVM

Support vector machine (SVM) proposed by Vapnik and colleagues in 1964 is a generalized linear classifier that classifies data through supervised learning (Girosi, 1998). SVM can also be used for regression, which is called support vector regression (SVR) (Santamaria-Bonfil et al., 2016). The main idea of SVR can be summarized as transforming the linearly inseparable samples of the low-dimensional input space into the high-dimensional feature space by using the nonlinear mapping algorithm, so that the linear regression can be performed in the high-dimensional feature space (Collobert and Bengio, 2001; Khemchandani and Jayadeva, 2009; Huang and Zhao, 2018).

By introducing ε-insensitive loss function, regularization constant C, relaxation variable ξ_i, $\hat{\xi}_i$, normal vector w, displacement b to be determined, the SVR is formalized as follows in Eq. 1, and the constraint as follows in Eq. 2:

$$f(x) = \min_{w,b,\xi_i,\hat{\xi}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \hat{\xi}_i\right) \tag{1}$$

$$s.t.\begin{Bmatrix} w^T \cdot x_i + b - y_i \le \varepsilon + \xi_i, \\ y_i - w^T \cdot x_i - b \le \varepsilon + \hat{\xi} \\ \xi_i \ge 0, \hat{\xi} \ge 0, i = 1, 2...m \end{Bmatrix} \tag{2}$$

Therefore, the final linear regression function of SVR is obtained as shown in Eq. 3:

$$f(x) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)\kappa\left(x_i^T x\right) + b \tag{3}$$

In Eq. 3, $\kappa(x_i^T x)$ is the kernel function, α_i and $\hat{\alpha}_i$ are Lagrange multipliers. Several kernel functions commonly used in support vector machines include linear kernel function, polynomial kernel function, radial basis kernel function and so on. The core idea of RBF is to map each sample point to an infinite dimensional feature space, so as to make linearly indivisible data linearly separable. It is the most commonly used kernel function and shown in Eq. 4:

$$\kappa(x_i, x) = \exp\left(-\gamma\|x_i - x\|^2\right) \tag{4}$$

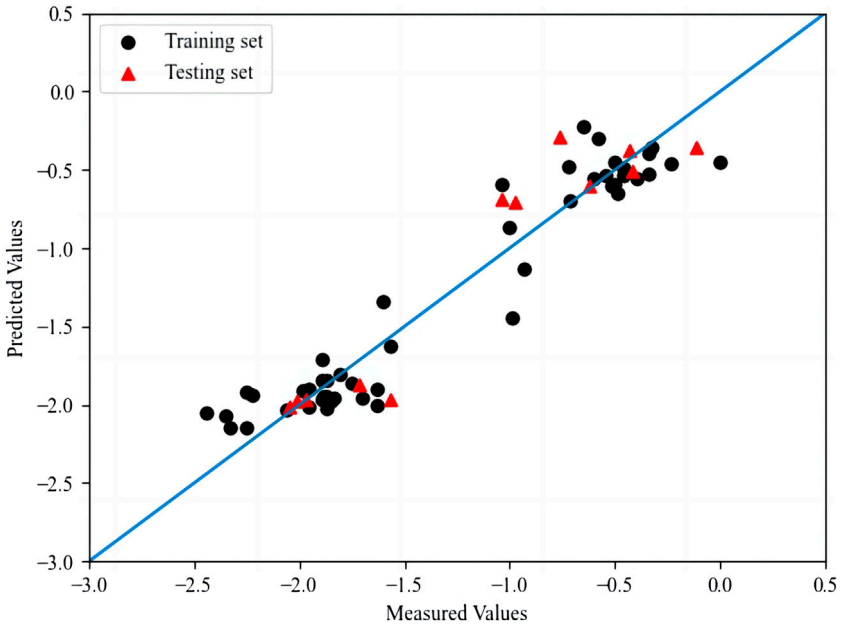

FIGURE 2
The plot of measured and predicted -lg (IC_{50}) by HM.

where γ is a hyperparameter of a radial basis kernel function.

2.6 Nonlinear model by PSO-SVM

Because complexity of optimizing many parameters in RBF-SVM is high, particle swarm optimization (PSO) algorithm was introduced in the parameter optimization process, which can converge to the global optimal solution with high efficiency. PSO was proposed by Kennedy and Eberhart in 1995, which is one of the most widely used optimization algorithm (Eberhart and Shi, 2004; Zhang et al., 2015b).

The basic concept of PSO is derived from the study of the foraging behavior of birds (Gong et al., 2016; Bonyadi and Michalewicz, 2017; Pervaiz et al., 2021). PSO can be expressed as: each particle can be regarded as a search individual in the N-dimensional search space, which iterates continuously, updates the speed and position, and finally obtains the optimal solution satisfying the termination condition (Li et al., 2002; Lin et al., 2008; Subasi, 2013). The formula for PSO update speed and location is shown below:

$$V_{id} = \omega V_{id} + C_1 random(0,1)(P_{id} - X_{id}) + C_2 random(0,1)\left(P_{gd} - X_{id}\right) \quad (5)$$

$$X_{id} = X_{id} + V_{id} \quad (6)$$

where ω is the inertial factor that specifies search step size. C_1 and C_2 are acceleration constants, X_{id} represents the d-dimensional position of each particle i, and P_{gd} represents the D-dimension of the global optimal solution.

3 Result

3.1 Results of HM

579 descriptors were calculated by CODESSA. To obtain several descriptors most relevant to the HDAC inhibition, the number of molecular descriptors in linear models were increased from 1 to 7, and the corresponding R^2 and R_{cv}^2 were recorded. Considering that excessive selection of descriptors is not conducive to drug screening and design, the number of descriptors were determined to be 4 of which R^2 and R_{cv}^2 both reached about 0.9. The influence of the number of descriptors on R^2 and R_{cv}^2 is shown in Figure 1.

The four selected molecular descriptors and their physicochemical meanings are shown in Supplementary Table S2. Their correlation coefficients are shown in Supplementary Table S3, all of which are less than 0.8.

The multiple regression linear model established by HM, is shown in Eq. 7.

$$-\lg(IC_{50}) = -0.351132 - 2.165127*d_1 + 0.940460*d_2 - 0.669793*d_3 + 0.800716*d_4 \quad (7)$$

where d_1, d_2, d_3 and d_4 represented MERHN, MREHN, MNRIN and MVO, respectively.

The R^2 and RMSE of the training set and the test set in the model are 0.917, 0.832 and 0.044, 0.056 respectively. The plot for HM model is shown in Figure 2. In addition, the Q^2 of the training set and the test set are 0.832 and 0.804 respectively through K-fold cross-validation, where the K value is 5.

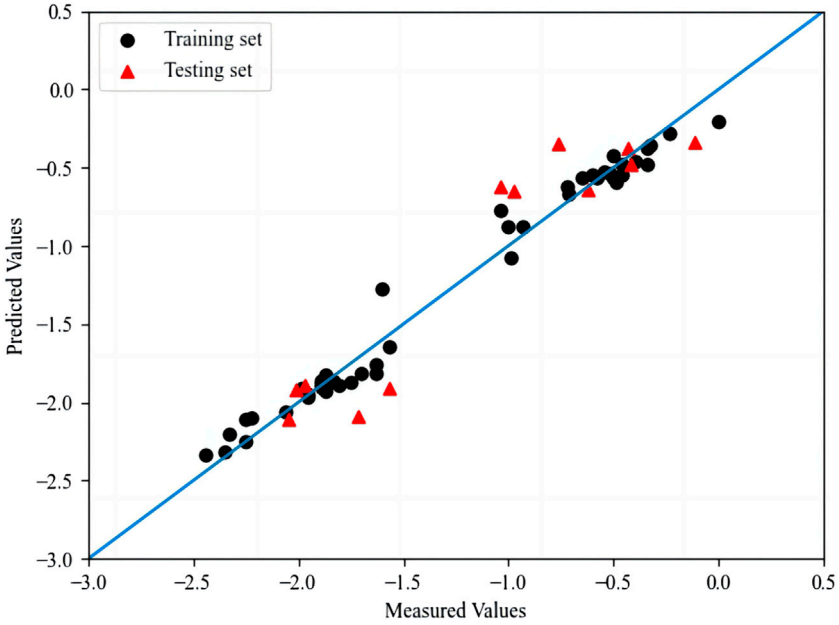

FIGURE 3
The plot of measured and predicted -lg (IC_{50}) by RF.

3.2 Results of RF

In order to ensure that the results of HM, RF, RBF-SVM and PSO-SVM can be compared with each other, the 4 descriptors selected by HM were used to build other three models by RF, RBF-SVM and PSO-SVM.

When using the RF method, it is necessary to determine the values of some parameters in RF, such as the number of decision trees *n*, the maximum depth of the tree *d*, the number of samples contained in each internal node s_1, and the number of samples contained in each leaf node s_2. Among them, the larger *n* is, the better the effect of the model tends to be. However, when *n* is larger to a certain extent, the decision boundary is reached, and the accuracy of the random forest usually stops rising or starts to fluctuate.

Set *n* to 100, d to the default value of the python library, s_1 to 2, and s_2 to 1. The R^2 and RMSE for the training set are 0.982 and 0.009, and the R^2 and RMSE for the test set are 0.841 and 0.063. The optimal prediction results of RF are shown in Figure 3. The results Q^2 of K-fold cross-validation for training set and test set are 0.886 and 0.823 respectively, where the *K* value is 5.

3.3 Results of RBF-SVM

The RBF-SVM method needs to determine the values of some parameters, such as the penalty coefficient *C*, ε-insensitive loss function, and the parameter *γ* of the kernel function. As the penalty coefficient used to control the loss function, if *C* is too large, the penalty for false regression predictions is too large, which is easy to lead to overfitting. If *C* is too small, the penalty for false regression predictions is too small, which easily leads to underfitting. As a parameter of the radial basis kernel function, the larger the *γ* is, the easier it is to overfit, and the smaller the *γ* is, the easier it is to underfit. The greater the ε is, the less the support vector is, and the greater the support vector is.

Grid search is the simplest and most widely used hyperparameter search algorithm, which determines the optimal value by looking for all the points within the search range. The optimal values of C and γ were determined by using grid search. The optimal values of *C*, γ and ε are 9.11,10.24 and 0.1, respectively. The optimal prediction results of RBF-SVM are shown in Figure 4. The R^2 and RMSE of the training set and prediction set using RBF-SVM are 0.957,0.022 and 0.944,0.025, respectively. When K-fold ($K = 5$) cross-validation was executed, Q^2 for the training set and the test set are 0.897 and 0.871, respectively.

3.4 Results of PSO-SVM

Many parameters need to be determined during the modeling process using RBF-SVM, and PSO algorithm with character of easy implement, high precision, fast convergence can quickly find the optimal solution. Therefore, PSO algorithm was used instead of grid search to find the optimal parameters. The particle swarm size and number of iterations were set to 600 and 1,500, respectively.

The optimal values of *C*, *γ* and ε are 9.13, 1.82 and 0.0, respectively. The R^2 and RMSE of the training and prediction sets using RBF-SVM were 0.966,0.018 and 0.975,0.012, and Q^2 were 0.903 and 0.896, respectively. The optimal prediction results of PSO-SVM are shown in Figure 5.

4 Comparison of different results

The prediction results of the four models are shown in Supplementary Table S4, and the K-fold cross-validation results

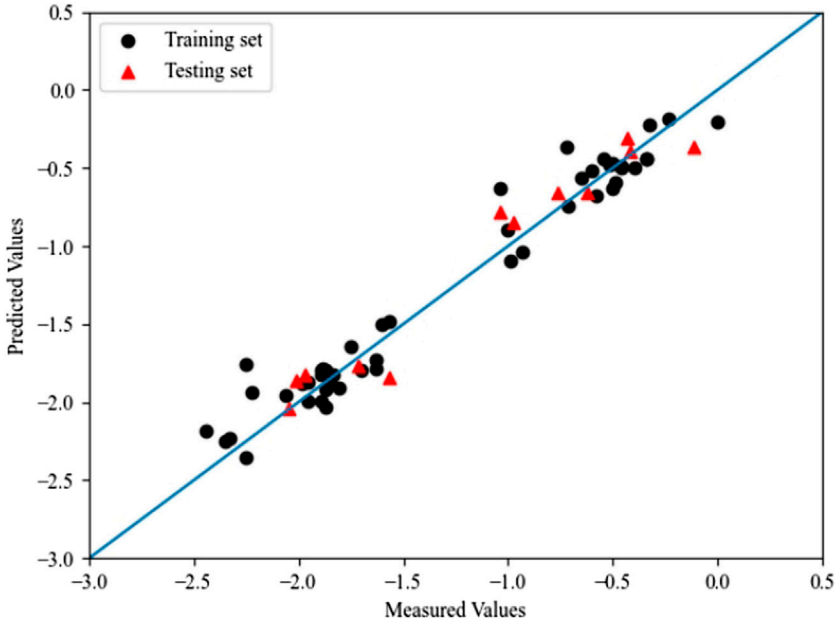

FIGURE 4
The plot of measured and predicted -lg (IC_{50}) by RBF-SVM.

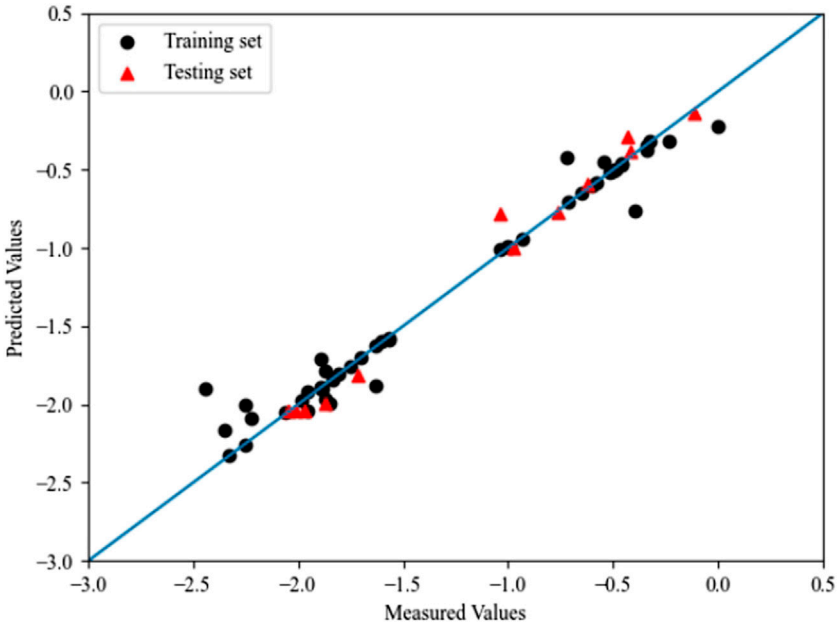

FIGURE 5
The plot of measured and predicted -lg (IC_{50}) by PSO-SVM.

are shown in Supplementary Table S5, where the value of K was set to 5. Moreover, it can be seen from Supplementary Table S4 that the RMSE of PSO-SVM is the smallest, indicating the best degree of fitting.

It can be seen from Supplementary Table S4 that the prediction accuracy and robustness of the nonlinear model established by RF, RBF-SVM and PSO-SVM methods are stronger than that of the linear model HM. However, the prediction accuracy of the model built by RF method on the training set was very high, but the prediction accuracy of the test set was relatively ordinary, which indicates that the model built by RF method is overfitting. The prediction accuracy of the models built by RBF-SVM and PSO-SVM

are very high. Among the four models, the model established by PSO-SVM has the strongest prediction ability and stability, for the PSO algorithm can efficiently find more optimized parameters.

5 Discussion

In this study, 4 descriptors were selected from 579 descriptors of 60 compounds containing triazole, and 4 QSAR models were established using HM, RF, SVM and PSO-SVM methods to predict the HDAC inhibition. Among the four models, the prediction ability and stability of PSO-SVM are the best, indicating that the model established by PSO-SVM method has a broad application prospect in searching for compounds with significant anti-colon cancer effect, and can be used as an effective method to assist drug design. In addition, this study also revealed four descriptors with significant inhibitory effects on HDAC: Max e-e repulsion for a H-N bond, Max resonance energy for a H-N bond, Max nucleoph react index for a N atom and Min valency of a O atom.

Among the four descriptors, Max e-e repulsion for a h-n bond is the one that has the most significant inhibiting effect on HDAC. Because it is always selected as the top node and has the highest Gini coefficient in the model built by RF. It reflects the repulsion force between electrons and plays a key role in forming the momentum distribution of the final correlated double electrons. The second descriptor is Max resonance energy for an H-N bond. The bond resonance energy represents the contribution of a given bond in a molecule to the topological resonance energy. If a molecule has one or more bonds with a large negative bond resonance energy, the molecule is very chemically reactive. The third descriptor Max nucleoph react index for a N atom is a quantum chemical descriptor, which indicates the strength of covalent bond in the molecule, represents the maximum nuclear reaction index of N atoms. Min valency of a O atom is a quantum chemical descriptor whose scope goes beyond the strength of intramolecular adhesion and accounts for the stability of the molecule and its conformational flexibility. Attempts to increase the valence of the O atom in the substituent may help to reduce the IC50 value of HDAC inhibition.

The compounds numbered 20 and 27 in Supplementary Table S1 have lower IC_{50} values, so other similar compounds with the above descriptors may be novel anti-colon cancer inhibitors that could be designed as potential drugs. Overall, this study revealed four descriptors with significant HDAC inhibition, which could help in the design of novel anti-colon cancer drugs in the future.

6 Conclusion

PyCharm Community Edition 2022.2.1 was used for experiments, and the scikit learn library was used to build machine learning models. The models established by RBF-SVM and PSO-SVM have good prediction performance and strong robustness, indicating that the models constructed by RBF-SVM and PSO-SVM have a broad application prospect in the study of the inhibitory effect of triazole-containing compounds on colon cancer. In addition, this study revealed 4 key descriptors that influence the inhibition of HDAC: Max e–e repulsion for a H–N bond, Max resonance energy for a H-N bond, Max nucleoph react index for a N atom and Min valency of a O atom.

In addition to some traditional and common QSAR models, this study also used particle swarm optimization to optimize the SVM model, which greatly improved the prediction accuracy, making the accuracy of the test set increased to 0.975, which will provide guidance and help for the future research on anti-colon cancer drugs.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

YW: Data curation, Methodology, Validation, Writing–original draft.
PZ: Formal Analysis, Supervision, Validation, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1260349/full#supplementary-material

References

Ao, Y. L., Li, H. Q., Zhu, L. P., Ali, S., and Yang, Z. G. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J. PETROLEUM Sci. Eng.* 174, 776–789. doi:10.1016/j.petrol.2018.11.067

Asklund, T., Kvarnbrink, S., Holmlund, C., Wibom, C., Bergenheim, T., Henriksson, R., et al. (2012). Synergistic killing of glioblastoma stem-like cells by bortezomib and HDAC inhibitors. *ANTICANCER Res.* 32 (7), 2407–2413. doi:10.1093/annonc/mds166

Bonyadi, M. R., and Michalewicz, Z. (2017). Particle swarm optimization for single objective continuous space problems: a review. *Evol. Comput.* 25 (1), 1–54. doi:10.1162/EVCO_r_00180

Chen, N., Chen, J., Yao, B., and Li, Z. G. (2018). QSAR study on antioxidant tripeptides and the antioxidant activity of the designed tripeptides in free radical systems. *MOLECULES* 23 (6), 1407. doi:10.3390/molecules23061407

Chen, Z., Li, K., Yin, X., Li, H., Li, Y., Zhang, Q., et al. (2019). Lower expression of gelsolin in colon cancer and its diagnostic value in colon cancer patients. *J. Cancer* 10 (5), 1288–1296. doi:10.7150/jca.28529

Choi, J., Hwang, J., Ramalingam, M., Jeong, H. S., and Jang, S. (2023). Effects of HDAC inhibitors on neuroblastoma SH-SY5Y cell differentiation into mature neurons via the Wnt signaling pathway. *BMC Neurosci.* 24 (1), 28. doi:10.1186/s12868-023-00798-0

Coi, A., Massarelli, I., Murgia, L., Saraceno, M., Calderone, V., and Bianucci, A. M. (2006). Prediction of hERG potassium channel affinity by the CODESSA approach. *Bioorg. Med. Chem.* 14 (9), 3153–3159. doi:10.1016/j.bmc.2005.12.030

Collobert, R., and Bengio, S. (2001). SVMTorch: support vector machines for large-scale regression problems. *J. Mach. Learn. Res.* 1 (2), 143–160. doi:10.1162/15324430152733142

Eberhart, R. C., and Shi, Y. H. (2004). Guest editorial special issue on particle swarm optimization. IEEE Trans. Evol. Comput. *JUN* 8 (3), 201–203. doi:10.1109/tevc.2004.830335

Evans, D. A. (2014). History of the harvard ChemDraw project. *Angew. CHEMIE-INTERNATIONAL Ed.* 53 (42), 11140–11145. doi:10.1002/anie.201405820

Fan, T. J., Sun, G. H., Zhao, L. J., Cui, X., and Zhong, R. G. (2018). QSAR and classification study on prediction of acute oral toxicity of N-nitroso compounds. *Int. J. Mol. Sci.* 19 (10), 3015. doi:10.3390/ijms19103015

Fang, Z. J., Yu, X. L., and Zeng, Q. (2022). Random forest algorithm-based accurate prediction of chemical toxicity to Tetrahymena pyriformis. *TOXICOLOGY*, 480. doi:10.1016/j.tox.2022.153325

Feng, W., Boukir, S., and Huang, W. (2019). "Ieee. MARGIN-BASED random forest for imbalanced land cover classification," in *Ieee international geoscience and remote sensing symposium* (China: IGARSS), 3085–3088.

Free, S. M., Jr., and Wilson, J. W. (1964). A MATHEMATICAL CONTRIBUTION TO STRUCTURE-ACTIVITY STUDIES. *J. Med. Chem.* 7, 395–399. doi:10.1021/jm00334a001

Gao, Z., Xia, R. Z., and Zhang, P. J. (2022). Prediction of anti-proliferation effect of [1,2,3]Triazolo[4,5-d]pyrimidine derivatives by random forest and mix-kernel function SVM with PSO. *Chem. Pharm. Bull.* 70 (10), 684–693. doi:10.1248/cpb.c22-00376

Gillette, T. G. (2021). HDAC inhibition in the heart: erasing hidden fibrosis. *Circulation* 143 (19), 1891–1893. doi:10.1161/CIRCULATIONAHA.121.054262

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *NEURAL Comput.* 10 (6), 1455–1480. doi:10.1162/089976698300017269

Gong, Y. J., Li, J. J., Zhou, Y. C., Chung, H. S. H., Shi, Y. H., et al. (2016). Genetic learning particle swarm optimization. *IEEE Trans. Cybern.* 46 (10), 2277–2290. doi:10.1109/TCYB.2015.2475174

Hansch, C., and Steward, A. R. (1964). THE USE OF SUBSTITUENT CONSTANTS IN THE ANALYSIS OF THE STRUCTURE-ACTIVITY RELATIONSHIP IN PENICILLIN DERIVATIVES. *J. Med. Chem.* 7, 691–694. doi:10.1021/jm00336a001

Huang, T., Sun, G., Zhao, L., Zhang, N., Zhong, R., and Peng, Y. (2021). Quantitative structure-activity relationship (QSAR) studies on the toxic effects of nitroaromatic compounds (NACs): a systematic review. Int. J. Mol. Sci. *Aug 9* (16), 22. doi:10.3390/ijms22168557

Huang, Y., and Zhao, L. (2018). Review on landslide susceptibility mapping using support vector machines. *CATENA* 165, 520–529. doi:10.1016/j.catena.2018.03.003

Kang, M. R., Kang, J. S., Han, S. B., Kim, J. H., Kim, D. M., Lee, K., et al. (2009). A novel delta-lactam-based histone deacetylase inhibitor, KBH-A42, induces cell cycle arrest and apoptosis in colon cancer cells. *Biochem. Pharmacol.* 78 (5), 486–494. doi:10.1016/j.bcp.2009.05.010

Katritzky, A. R., Kuanar, M., Fara, D. C., Karelson, M., Acree, W. E., Solov'ev, V. P., et al. (2005). QSAR modeling of blood: air and tissue: air partition coefficients using theoretical descriptors. *Bioorg. Med. Chem.* 13 (23), 6450–6463. doi:10.1016/j.bmc.2005.06.066

Khemchandani, R., and Jayadeva, C. S. (2009). Regularized least squares fuzzy support vector regression for financial time series forecasting. *EXPERT Syst. Appl.* 36 (1), 132–138. doi:10.1016/j.eswa.2007.09.035

Li, A., Qin, Z., Bao, F., and He, S. (2002). Particle swarm optimization algorithms. *Comput. Eng. Appl.* 38 (21), 1–3.

Li, G. L., Wang, X. Q., Li, A. Q., and Zhang, P. J. (2023). QSAR study on the IC_{50} of thiosemicarbazone derivatives as PC-3 inhibitors based on mixed kernel function support vector machine. *Lat. Am. J. Pharm.* 42 (3), 543–553.

Li, Z., Zhang, T., and Wu, X. (2012). Methodology of regression by random forest and its application on metabolomics. *Chin. J. Health Statistics* 29 (2), 158–160.

Lima, N., Rocha, G., Freire, R., and Simas, A. (2018). RM1 semiempirical model: chemistry, pharmaceutical research, molecular biology and materials science. *J. Braz. Chem. Soc.* doi:10.21577/0103-5053.20180239

Lin, S.-W., Ying, K.-C., Chen, S.-C., and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Syst. Appl.* 35 (4), 1817–1824. doi:10.1016/j.eswa.2007.08.088

Madugula, S. S., and Yarasi, S. (2017). Molecular design of porphyrin dyes for dye sensitized solar cells: a quantitative structure property relationship study. *Int. J. QUANTUM Chem.* 117 (14), e25385. doi:10.1002/qua.25385

Myint, K. Z., and Xie, X. Q. (2010). Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *Int. J. Mol. Sci.* 11 (10), 3846–3866. doi:10.3390/ijms11103846

Pan, G., Zhang, P., Chen, A., Deng, Y., Zhang, Z., Lu, H., et al. (2023). Aerobic glycolysis in colon cancer is repressed by naringin via the *HIF1A* pathway. J. Zhejiang Univ. Sci. B. *Mar.* 24 (3), 221–231. doi:10.1631/jzus.B2200221

Pervaiz, S., Ul-Qayyum, Z., Bangyal, W. H., Gao, L., and Ahmad, J. (2021). A systematic literature review on particle swarm optimization techniques for medical diseases detection. *Comput. Math. METHODS Med.* 2021, 2021–2110. doi:10.1155/2021/5990999

Place, R. F., Noonan, E. J., and Giardina, C. (2005). HDAC inhibition prevents NF-kappa B activation by suppressing proteasome activity: down-regulation of proteasome subunit expression stabilizes I kappa B alpha. *Biochem. Pharmacol.* 70 (3), 394–406. doi:10.1016/j.bcp.2005.04.030

Roy, R., Ria, T., RoyMahaPatra, D., and Uh, Sk (2023). Single inhibitors versus dual inhibitors: role of HDAC in cancer. *ACS Omega* 8 (19), 16532–16544. doi:10.1021/acsomega.3c00222

Sang, D. M., Na, I. H., Anh, D. T., Thi Mai Dung, D., Thi Thu Hang, N., Phuong-Anh, N. T., et al. (2023). Novel (E)-3-(3-Oxo-4-substituted-3,4-dihydro-2H-benzo [b] [1,4]oxazin-6-yl)-N-hydroxypropenamides as histone deacetylase inhibitors: design, synthesis and bioevaluation. *Chem. Biodivers.* 20. doi:10.1002/cbdv.202201030

Santamaria-Bonfil, G., Reyes-Ballesteros, A., and Gershenson, C. (2016). Wind speed forecasting for wind farms: a method based on support vector regression. *Renew. ENERGY* 85, 790–809. doi:10.1016/j.renene.2015.07.004

Schmoll, H. J., and Stein, A. (2014). COLORECTAL CANCER IN 2013 towards improved drugs, combinations and patient selection. *Nat. Rev. Clin. Oncol.* 11 (2), 79–80. doi:10.1038/nrclinonc.2013.254

Shi, Y., Li, J., Tang, M., Liu, J., Zhong, Y., and Huang, W. (2022a). CircHADHA-augmented autophagy suppresses tumor growth of colon cancer by regulating autophagy-related gene via miR-361. *Front. Oncol.* 12, 937209. doi:10.3389/fonc.2022.937209

Shi, Y., Li, J. Y., Tang, M., Liu, J. W., Zhong, Y. L., and Huang, W. (2022b). CircHADHA-augmented autophagy suppresses tumor growth of colon cancer by regulating autophagy-related gene via miR-361. *Front. Oncol.* 12, 12. doi:10.3389/fonc.2022.937209

Song, J. (2015). Bias corrections for Random Forest in regression using residual rotation. *J. KOREAN Stat. Soc.* 44 (2), 321–326. doi:10.1016/j.jkss.2015.01.003

Subasi, A. (2013). Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. Comput. Biol. Med. *Jun* 43 (5), 576–586. doi:10.1016/j.compbiomed.2013.01.020

Sun, N., Yang, K., Yan, W., Yao, M., Yu, C., Duan, W., et al. (2023). Design and synthesis of triazole-containing HDAC inhibitors that induce antitumor effects and immune response. *J. Med. Chem.* 66 (7), 4802–4826. doi:10.1021/acs.jmedchem.2c01985

Tavares, M. T., Shen, S., Knox, T., Hadley, M., Kutil, Z., Bařinka, C., et al. (2017). Synthesis and pharmacological evaluation of selective histone deacetylase 6 inhibitors in melanoma models. *ACS Med. Chem. Lett.* 8 (10), 1031–1036. doi:10.1021/acsmedchemlett.7b00223

Wang, J., Du, Y., Liu, X. M., Cho, W. C., and Yang, Y. X. (2015). MicroRNAs as regulator of signaling networks in metastatic colon cancer. *BIOMED Res. Int.* 2015, 823620. doi:10.1155/2015/823620

Wang, Y., Liu, Z., Qu, A. L., Zhang, P. J., Si, H. Z., and Zhai, H. L. (2020). Study of tacrine derivatives for acetylcholinesterase inhibitors based on artificial intelligence. *Lat. Am. J. Pharm.* 39 (6), 1159–1170.

Xiong, X., Wang, S., Gao, Z., and Ye, Y. (2023). C6orf15 acts as a potential novel marker of adverse pathological features and prognosis for colon cancer. *Pathol. Res. Pract.* 245, 154426. doi:10.1016/j.prp.2023.154426

Yao, Y. W., Liao, C. H., Li, Z., Wang, Z., Sun, Q., Liu, C., et al. (2014). Design, synthesis, and biological evaluation of 1, 3-disubstituted-pyrazole derivatives as new class I and IIb histone deacetylase inhibitors. *Eur. J. Med. Chem.* 86, 639–652. doi:10.1016/j.ejmech.2014.09.024

Yang, X., Qiu, H., Zhang, Y., and Zhang, P. (2023). Quantitative structure–activity relationship study of amide derivatives as xanthine oxidase inhibitors using machine learning. *Front. pharmacol.*, 14.

Zhang, H., Li, J., and Kim, C. K. (2013). Quantitative structure-properties relationship studies on physicochemical properties of organic molecules using CODESSA. *ASIAN J. Chem.* 25 (10), 5670–5672. doi:10.14233/ajchem.2013.oh58

Zhang, L., Wang, P. K., Jiang, T. Y., Fan, G. H., and Dan, C. H. (2015a). Ieee. Prediction of Torpedo initial velocity based on random forests regression. *Int. Conf. INTELLIGENT HUMAN-MACHINE Syst. Cybern. IHMSC* I, 337–339. doi:10.1109/IHMSC.2015.17

Zhang, Y. D., Wang, S. H., and Ji, G. L. (2015b). A comprehensive survey on particle swarm optimization algorithm and its applications. *Math. PROBLEMS Eng.* 2015, 1–38. doi:10.1155/2015/931256

Zhao, Y. T., Liu, X. Q., Ouyang, J., Wang, Y., Xu, S., Tian, D., et al. (2020). Studies on the IC50 of metabolically stable 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas as human CCR5 receptor antagonists based on QSAR. *Lett. DRUG Des. Discov.* 17 (8), 1036–1046. doi:10.2174/1570180817666200320105725