RESEARCH ARTICLE

Development of a novel mTOR inhibitor prediction model and prediction platform based on machine learning

Chuanxin Yu, Zhao Zhang, Qingqing Wang, Jianyu Xiao*

School of Computer Science and Technology, Engineering Research Center of Cognitive Behavioral Intelligent Computing and Application, Huaibei Normal University, Huaibei, Anhui, China.

Received: March 17, 2025; **accepted:** July 6, 2025.

Mammalian Target of Rapamycin (mTOR), a key component of the PI3K/AKT/mTOR signaling pathway, plays a central role in tumor development. Although mTOR inhibitors such as rapamycin have been used clinically, limitations remain in terms of efficacy and side effects. This study integrated five machine learning algorithms including random forest (RF), support vector machine (SVM), extreme gradient boosting (XGBoost), k-nearest neighbor (KNN), naive Bayes (NB) and five deep learning algorithms including deep neural networks (DNN), graph convolutional networks (GCN), graph attention networks (GAT), message passing neural networks (MPNN), Attentive Fingerprint (Attentive FP) and combined molecular descriptors, molecular fingerprints, and molecular images to construct 28 prediction models for exploring the interactions and mechanisms between biomolecules. The models were comprehensively evaluated using seven evaluation indicators including F1-score, accuracy (ACC), balanced accuracy (BA), sensitivity (SE), specificity (SP), Matthew's correlation coefficient (MCC), area under the ROC curve (AUC) to select the optimal model. In-depth analysis of the model using the SHapley additive exPlanations (SHAP) method revealed the key structural fragments affecting mTOR inhibition. The results showed that the random forest algorithm combined with the Morgan molecular fingerprint had high accuracy in predicting mTOR inhibitor activity and could accurately identify relevant key structural fragments. In addition, this study presented a user-friendly prediction platform based on the random forest algorithm, providing researchers with a convenient tool for quickly evaluating the mTOR inhibitory activity of compounds. These results provided reliable references for the discovery of new mTOR inhibitors and could facilitate new breakthroughs in the field of cancer treatment.

Keywords: mTOR inhibitors; machine learning; deep learning; cancer treatment; molecular fingerprint.

***Corresponding author:** Jianyu Xiao, School of Computer Science and Technology, Engineering Research Center of Cognitive Behavioral Intelligent Computing and Application, Huaibei Normal University, Huaibei, Anhui, China. Phone: +86 18856193108. Email: xjy@chnu.edu.cn.

Introduction

Mammalian Target of Rapamycin (mTOR) inhibitors have become the focus of research in the field of tumor therapy because of their critical roles in regulating cell growth, proliferation, and metabolism. mTOR, a serine/threonine kinase, is a core component of the phosphoinositide 3-kinase (PI3K)/Protein Kinase B (AKT)/mTOR signaling pathway and plays a crucial role in the occurrence and development of various tumors [1]. Although mTOR inhibitors such as rapamycin and its derivatives have been used in clinical practice, certain limitations remain in terms of their efficacy and side effects [2]. Long-term rapamycin use may lead to immune suppression

and the development of tumor resistance [3]. Further, owing to the complexity of the mTOR signaling pathway, traditional drug design methods often struggle to accurately predict the efficacy and side effects of drugs [4]. Therefore, the development of novel mTOR inhibitors with improved efficacy and fewer side effects is important for research and clinical applications.

In recent years, machine learning technology has shown significant potential in drug design and discovery. By analyzing a large amount of compound data and biological activity information in detail, machine learning models can accurately identify potential drug candidates and predict their biological activities against specific targets [5]. This method significantly accelerates the drug discovery process and improves the accuracy and efficiency of drug design [6]. Stokes *et al*. successfully identified a lead compound, halicin, with broad antibacterial activity using a neural network model by analyzing phenotypic screening data [7]. Further, a series of machine learning-based models have been developed to identify novel antibacterial compounds that target multiple pathogens including those developed by Wang *et al*. for methicillin-resistant *Staphylococcus aureus* [8], Fields *et al*. for *Pseudomonas aeruginosa* [9], Ashdod *et al*. for *Plasmodium falciparum* [10], and Zheng *et al*. for *Schistosomiasis* [11]. In addition, Kumari *et al*. combined random forest and an autoencoder to identify 20 key molecular descriptors for predicting mTOR inhibitors, which was the first time that traditional machine learning combined with deep learning feature extraction methods [12]. Wang *et al*. used support vector machine (SVM) models to screen potential mTOR inhibitors and provide new candidate drugs for cancer treatment [13]. Zhang *et al*. used machine learning models to analyze gastric cancer samples and found that Acute Myeloid Leukemia (AML) cells with Tumor Protein p53 (TP53) mutations were insensitive to rapamycin, whereas Mouse Double Minute 2 homolog (MDM2)-dependent cells were sensitive [14]. DS-5272 is a small-molecule MDM2 inhibitor that blocks the interaction between MDM2 and p53, restores the tumor inhibitory function of p53, enhances the efficacy of anticancer drugs such as rapamycin, and is primarily used in tumor treatment research. The combined treatment effect of rapamycin and DS-5272 was studied and further evaluated. The Joint Non-negative Matrix Factorization (JNMF) algorithm was used to identify AML subtypes sensitive to mTOR inhibitors for predicting the response of patients with gastric cancer to the treatment. These studies demonstrated the diverse applications of machine learning in drug discovery and highlighted its important role in improving the efficiency and accuracy of drug development.

This study collected data from 2,074 mTOR inhibitors, of which 1,962 compounds were active, while 112 compounds were inactive. By comprehensively applying various machine learning algorithms and deep learning techniques combined with three molecular representation methods including molecular descriptors, molecular fingerprints, and molecular images, 28 prediction models were constructed to explore the interactions and mechanisms between biomolecules in depth. The SHAP method was used to gain a deeper understanding of the underlying mechanisms of the constructed model and explore important structural fragments in the model that had guiding significance for optimizing lead compounds and designing new reagents. The best model was obtained based on a comprehensive comparison of multiple evaluation indicators. To further promote the practical application of the model, a user-friendly prediction platform was developed to conveniently predict the mTOR inhibitory activities of the compounds. This study provided a reliable reference for discovering new mTOR inhibitors and brought new breakthroughs in cancer treatment.

Materials and methods

Data collection and cleaning

The raw dataset was obtained from the CHEMBL 24 (https://www.ebi.ac.uk/chembl/) database downloaded in April 2024 [15]. Only data relevant to human mTOR inhibitor experiments were selected. The raw data were processed by removing nulls and metal ions, standardizing each compound in the dataset to a generic representation using the Python Standardizer package (https://github.com/flatkinson/standardiser) with default parameters including removal of counter ions, solvent components, and salts, addition of hydrogen atoms, and neutralization of charge by addition or subtraction of atoms, removing duplicate values, retaining compounds with a clear bioassay value (assay type B) such as half maximal inhibitory concentration (IC_{50}), half maximal effective concentration (EC_{50}), dissociation constant (Kd), and inhibition constant (Ki) while removing compounds lacking any bioassay data, removing duplicates and molecules with molecular weights greater than 1,000 Da, and taking the average of these reported bioassay values as the final value if a molecule had multiple bioactivity data points. The compounds with criteria values less than or equal to 10,000 were set to 1 and labeled as active substances, whereas the others were set to 0 and labeled as inactive substances. These steps were designed to clean and process the data for obtaining reliable and analyzable compound data, which formed the basis for further research and analysis. A schematic diagram of research on the prediction of mTOR inhibitors was shown in Figure 1.

Molecular representation calculations

Molecular representation plays a crucial role in developing reliable and robust quantitative structure–activity relationship (QSAR) models, and its quality greatly affects the accuracy of the predictions made by these models. To fully explore the chemical information of molecules, three different types of features were used to describe them including molecular descriptors, molecular fingerprints, and molecular graphs. Molecular fingerprints included the Morgan fingerprint that was similar to extended-connectivity fingerprints (ECFP) (1,024 bits) [16], the Molecular ACCess System (MACCS) key (166 bits) [17], the Atomparis fingerprint (2,048 bits) [18], and the RDKIT descriptor [19]. For a given molecule, the molecule graph comprised two matrices including an N × N adjacency matrix A, which was used to represent the graphical structure of the molecule, and an N × F node feature matrix X, where N was the number of nodes and F was the number of node features. The node feature matrix contained atomic features such as atom type, formal charge, hybridization, number of bound hydrogens, aromaticity, degree, hydrogen number, chirality, and partial charge. Edge features included information such as bond type, whether the paired atoms were in the same ring, whether the bonds were conjugated, and the stereo configuration of the bonds [20]. These features were typically represented as unique heat codes to represent the information in the molecular graph for further analysis and prediction in the QSAR model. The combined use of different types of molecular features could improve the model performance and accuracy, leading to a better understanding and prediction of the biological activities and properties of compounds. In this study, DeepChem (https://deepchem.io/) (version 2.5.0) was used to generate molecular graph-based representations.

Machine learning algorithms and model construction

Five classical machine learning algorithms including RF [21], SVM [22], XGBoost, KNN [23], NB [24] and five deep learning algorithms including DNN [25], GCN [26], GAT [27], MPNN [28], Attentive FP [29] were used to construct a predictive model for distinguishing between activity and inactivity regarding mTOR. The classical machine learning models, RF, SVM, KNN, and NB were constructed using the Scikit-Learn framework (https://scikit-learn.org/) [30], while the XGBoost model was developed using the XGBoost Python package [31]. All models based on molecular descriptors and fingerprints as well as deep learning models for graph neural network (GNN) were trained on a 11th Gen Intel

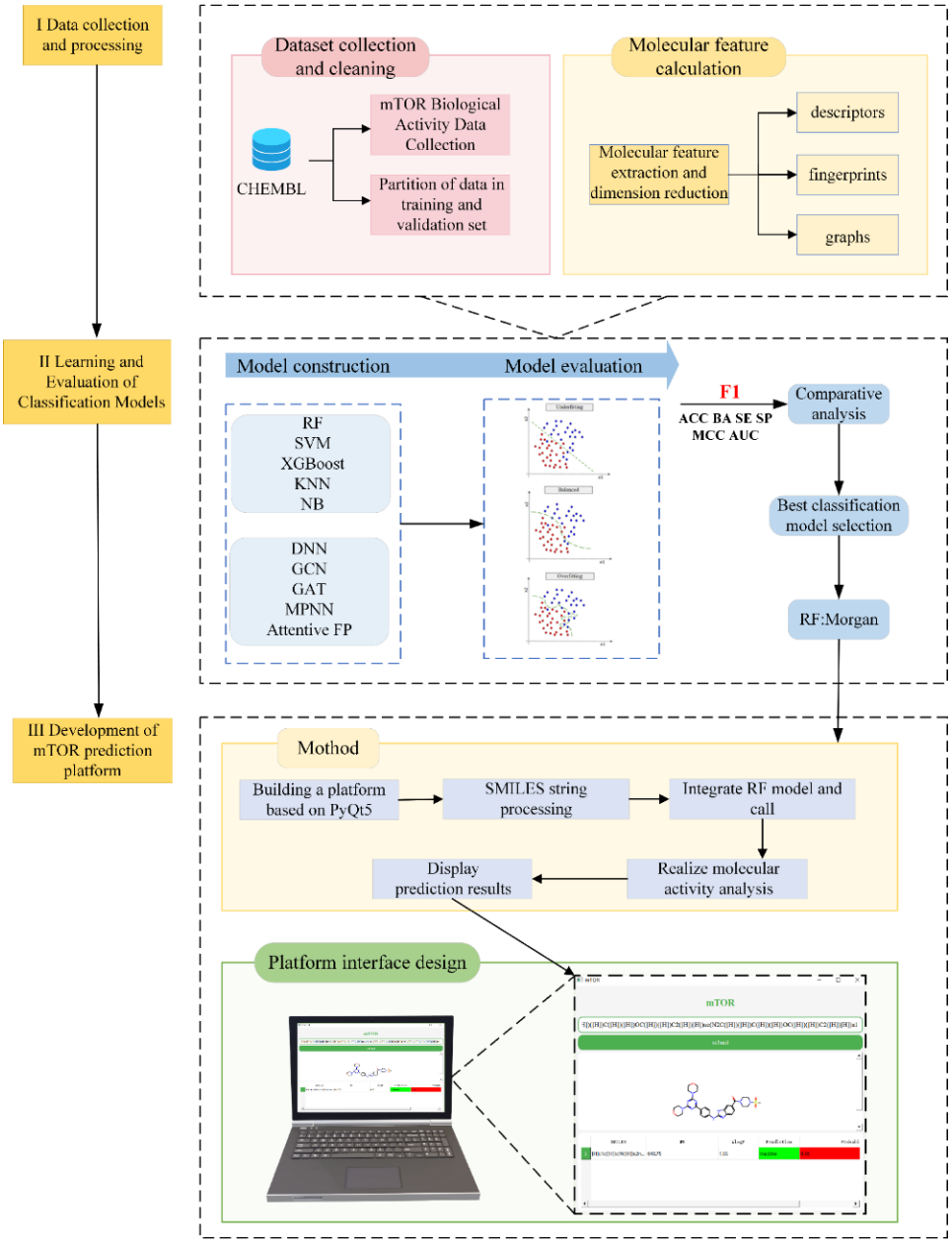

Figure 1. Schematic diagram of research on predicting novel mTOR inhibitors.

Core™ i5 computer with 16 GB RAM and 512 GB SSD hard drive under WIN10 operating system and integrated graphics. 28 predictive models were developed based on three different types of molecular features including molecular descriptors, molecular fingerprints, molecular graphs, five selected deep learning algorithms, and five machine learning algorithms (Figure 2). All models were selected based on F1 scores [32].

Model performance evaluation

The performance of the classification models was evaluated using the evaluation metrics of specificity (SP), sensitivity (SE), accuracy (ACC), F1-score, Matthew's correlation coefficient (MCC), balanced accuracy (BA), area under the ROC curve (AUC) as follows.

$$SP = \frac{TN}{TN+FP} \quad (1)$$

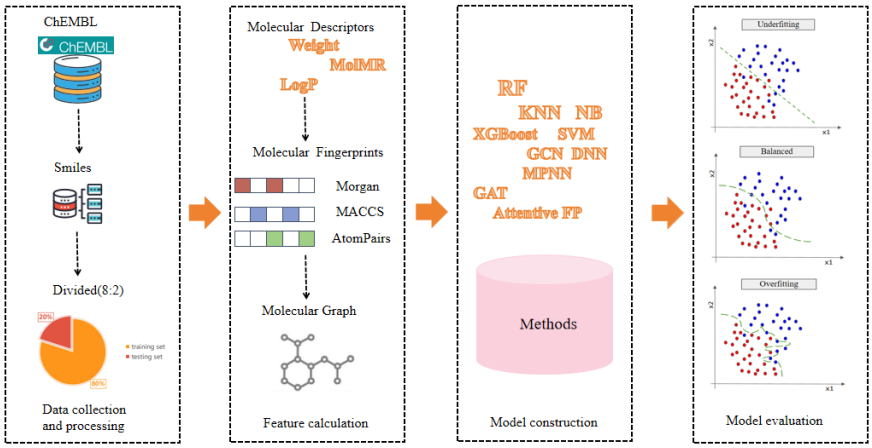

Figure 2. Model construction process.

$$SE = \frac{TN}{TP+FN} \quad (2)$$

$$ACC = \frac{TP+TN}{TP+TN+FT+FN} \quad (3)$$

$$F1 = \frac{Precision \times Sensitivity}{Precision+Sensitivity} = \frac{2 \times TP}{2 \times TP+FP+FN} \quad (4)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \quad (5)$$

$$BA = \frac{TPR+TNR}{2} = \frac{SE+SP}{2} \quad (6)$$

where the numbers of true positives, true negatives, false positives, and false negatives were represented as TP, TN, FP, and FN, respectively.

Prediction platform design

The mTOR inhibition activity prediction platform developed in this study was based on the PyQt5 (https://www.riverbankcomputing.com/software/pyqt/intro) framework and implemented a user-friendly and easy-to-use interface. This platform allowed users to input simplified molecular input linear expression (SMILES) strings, which were converted into molecular objects using the MolFromSmiles (smi) function in the RDKit library (https://www.rdkit.org/), laying the foundation for subsequent processing. Furthermore, by using RDKit's AllChem, the GetMorgan FingerprintAsBitVect function, the molecular objects were converted into molecular fingerprints, and the key chemical features of molecules were extracted to provide rich structural information for machine learning models. These molecular fingerprints were input as feature vectors into a pretrained random forest model to predict the mTOR inhibitory activity. This model was optimized to effectively identify the key structural fragments related to mTOR inhibition and predict the probability of molecular activity. To further enhance the user experience and assist in molecular activity analysis, the platform integrated the function of generating molecular structure images, allowing users to intuitively understand molecular features, thereby enhancing the user experience. Through model calculations, the system output predicted activity probabilities, providing users with a comprehensive mTOR inhibition activity prediction tool. These designs improved the accuracy of predictions as well as enhanced the user's analytical experience through intuitive molecular structure images.

Results and discussion

Dataset analysis

In the processed dataset, 1,962 compounds were identified as active (inhibitors), while 112 compounds were identified as inactive (non-inhibitors). The proportion of active compounds in the total sample was as high as 94.6%,

reflecting the significant advantage of the active compounds in the sample. Considering the natural distribution, the dataset was not artificially balanced by adding theoretical pseudomolecules. The compounds in the dataset exhibited a wide range of molecular weights (MW) from 144.176 to 966.226, and AlogP values were from -2.870 to 8.263. LogP is the logarithm of the partition coefficient of a compound between octanol and water, which is an important indicator of molecular lipophilicity (hydrophobicity). AlogP is a calculation method to predict logP using the atom additive method. It estimates the logP of the entire molecule by summing up the contribution values of each atom and combining them with the empirical parameters. The larger the logP (or Alogp), the more lipophilic and hydrophobic is the molecule. The smaller the logP, the more hydrophilic is the molecule. In drug development, logP is usually used to evaluate the absorption, permeability, and drug formation of molecules. The obtained molecular weights and AlogP values indicated that the compounds in the modeling dataset covered a broad chemical space. This chemical diversity was crucial for constructing accurate predictive models. Each dataset was randomly divided into two subsets including the training set (80%) and testing set (20%). The molecular weight and AlogP values, which together defined the dimensions of the chemical space, were calculated using the RDKit software. After cleaning, the dataset for subsequent model construction was chosen.

Performance of molecular descriptor-based predictive models

Six predictive models were constructed based on RDKIT descriptors. In this study, five machine learning algorithms and one deep learning algorithm (RF, SVM, XGBoost, KNN, NB, and DNN) were used. For classical machine learning algorithms, the SelectPercentile module in the Scikit-Learn package with the percentage set to 30% was used to select the optimized RDKIT descriptors as input features for the model. Each model comprised a combination of a specific molecular representation and a machine learning algorithm (e.g., RF: Rdkit). The results showed that RF and SVM demonstrated good performance on multiple evaluation metrics. In particular, the random forest algorithm achieved an ACC of 0.962, an F1 score of 0.980, and an AUC as high as 0.997 (Table 1). Considering various indicators, the random forest algorithm was the best performing choice.

Performance of molecular fingerprint-based predictive models

Based on three molecular fingerprints (Morgan, MACCS, and AtomPairs), five machine learning algorithms and one deep learning algorithm (RF, SVM, XGBoost, KNN, NB, and DNN) were used to construct 18 predictive models. The RF algorithm exhibited excellent performance for the three fingerprint-based models when applied to the mTOR dataset. Compared with the other five machine learning algorithms, the RF algorithm achieved an average F1 score of 0.979, an average BA of 0.607, and an average AUC of 0.892, yielding excellent results for all metrics. The results showed that RF outperformed the other models in classification on this dataset. Notably, the RF algorithm performed particularly well on the Morgan fingerprint-based model, achieving the highest ACC of 0.959 and a very high F1 score of 0.979 on the modeled dataset, which indicated that the model possessed a high rate of prediction accuracy and exhibited outstanding performance in terms of BA and SE. Further, the performance of the XGBoost algorithm achieved an average F1 score of 0.978, which was lower only than that of the RF algorithm, while its average AUC was as high as 0.881 (Table 1).

Performance of molecular graph-based predictive models

Recently, GNN and their variants have been widely used in tasks related to drug discovery [33]. A significant advantage of GNN over traditional predefined molecular descriptors or fingerprints is the ability to automatically learn task-specific molecular representations through the graph-convolution mechanism [34]. Various GNN models and their variants including GNN,

Table 1. Evaluation of the test dataset.

Molecular Features	Algorithms	ACC	F1 scores	BA	SE	SP	MCC	AUC
Morgan	RF	0.959	0.979	0.622	0.995	0.250	0.407	0.882
	SVM	0.954	0.977	0.525	1.000	0.050	0.218	0.840
	XGBoost	0.957	0.978	0.574	0.997	0.150	0.323	0.881
	KNN	0.947	0.972	0.664	0.977	0.350	0.364	0.843
	Nb	0.911	0.953	0.597	0.944	0.250	0.169	0.597
	DNN	0.952	0.975	0.975	0.990	0.961	0.296	0.906
	Mean	0.947	0.972	0.660	0.984	0.335	0.296	0.825
MACCS	RF	0.959	0.979	0.599	0.997	0.200	0.388	0.892
	SVM	0.954	0.977	0.525	1.000	0.050	0.218	0.840
	XGBoost	0.957	0.978	0.574	0.997	0.150	0.323	0.881
	KNN	0.957	0.978	0.550	1.000	0.100	0.309	0.847
	Nb	0.200	0.275	0.580	0.159	1.000	0.095	0.799
	DNN	0.952	0.975	0.976	1.000	0.952	0.000	0.987
	Mean	0.830	0.860	0.634	0.859	0.409	0.222	0.874
AtomPairs	RF	0.959	0.979	0.599	0.997	0.200	0.388	0.902
	SVM	0.954	0.977	0.525	1.000	0.050	0.218	0.840
	XGBoost	0.957	0.978	0.574	0.997	0.150	0.323	0.881
	KNN	0.959	0.979	0.694	0.987	0.400	0.476	0.874
	Nb	0.706	0.821	0.703	0.706	0.700	0.188	0.703
	DNN	0.952	0.975	0.975	0.990	0.961	0.296	0.904
	Mean	0.914	0.951	0.678	0.946	0.410	0.315	0.851
Molecular Graph	AttentiveFP	0.952	0.975	0.976	1.000	0.952	0.000	0.745
	GAT	0.957	0.978	0.978	0.997	0.959	0.323	0.782
	GCN	0.940	0.969	0.969	0.977	0.960	0.218	0.798
	MPNN	0.952	0.975	0.976	1.000	0.952	0.000	0.499
	Mean	0.950	0.974	0.975	0.994	0.956	0.135	0.706
Rdkit	RF	0.962	0.980	0.624	0.997	0.250	0.444	0.881
	SVM	0.954	0.977	0.525	1.000	0.050	0.218	0.840
	XGBoost	0.957	0.978	0.574	0.997	0.150	0.323	0.881
	KNN	0.957	0.978	0.621	0.992	0.250	0.378	0.654
	Nb	0.954	0.977	0.525	1.000	0.050	0.218	0.780
	DNN	0.952	0.975	0.975	1.000	0.952	0.000	0.767
	Mean	0.956	0.978	0.641	0.998	0.284	0.264	0.801

GAT, MPNN, Attentive FP) have demonstrated excellent state-of-the-art accuracy in various molecular property prediction tasks and are often used to compare the performance of different algorithms or models on the same dataset or task. In the tasks of drug discovery and molecular property prediction [35]. In this research, four deep learning algorithms of GCN, GAT, MPNN, and Attentive FP were employed to construct four molecular graph-based models (Figure 3). Overall, the GAT model exhibited excellent or near-optimal performance for most evaluation metrics. Specifically, this model achieved a high mean F1 score of 0.978 and BA of 0.978, demonstrating its power in classification tasks. In contrast, although the Attentive FP achieved a perfect 1.000 on the SE metric and the GCN was similar or slightly superior to the GAT in terms of SP as 0.960 and AUC as 0.798, their overall performance remained slightly inferior. The performance of MPNN was relatively weak for all metrics, especially when dealing with more complex tasks, in which the performance gap was more pronounced.

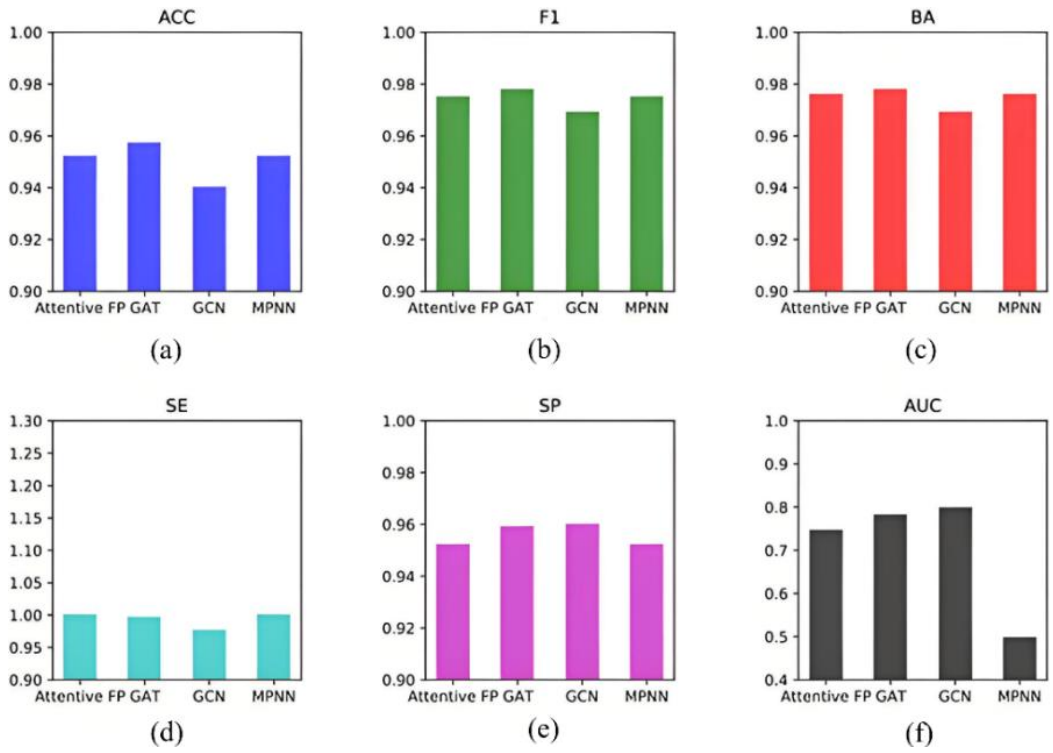

Figure 3. Metrics of deep learning models.

Optimal models

The comparison of the constructed molecular descriptor-, molecular fingerprint-, and molecular graph-based predictive models showed that the RF algorithm outperformed the other five machine learning methods in the descriptor- and molecular fingerprint-based models with significantly higher mean metrics for the F1 score, BA, and AUC. Further, the extreme gradient boosting algorithm achieved excellent results for this dataset (Table 1). Among the constructed graph-based models, GAT outperformed three other deep learning methods of GCN, MPNN, and Attentive FP. In this dataset, descriptor- and fingerprint-based models typically exhibited a superior performance compared to that of graph-based models, which suggested that deep learning methods failed to outperform classical machine learning methods in this scenario, particularly XGBoost and RF, which were the two most effective algorithms. This result was consistent with the results of a recent systematic comparative study [36].

Interpretations of the optimal model

The SHAP method was used to explore important structural fragments in the model to gain a deeper understanding of the intrinsic mechanisms of the constructed model. Given that the combination of RF and the Morgan fingerprint model exhibited excellent predictive performance, SHAP's TreeExplainer method was used to identify the key local structures in the model. By calculating the SHAP values, the top 20 structural fragments with significant favorable or unfavorable effects on mTOR inhibition were successfully screened and visualized. The results showed that some of the fingerprints were mainly concentrated in the positive region, whereas the others were mainly distributed in the negative region. Specifically, fingerprints of Morgan_926, Morgan_896, Morgan_831, Morgan_361, and Morgan_935 were in the positive region, suggesting that molecules containing these structural fragments had a higher probability of exhibiting anti-tumor activity. In contrast, fingerprints of Morgan_64, Morgan_548, Morgan_435, Morgan_175, and

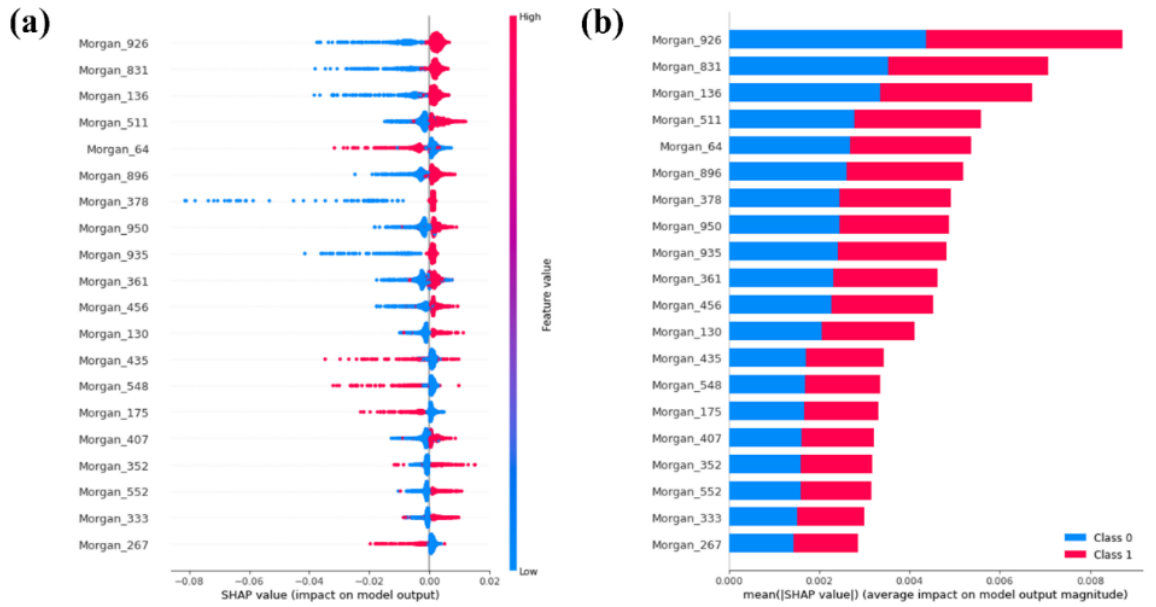

Figure 4. Random forest: Morgan model of the top 20 most important features. **(a)** SHAP values for each molecular substructure. Red and blue indicated the magnitude of the eigenvalue. The darker the red color is, the larger the eigenvalue is. **(b)** the average of the absolute SHAP values for each molecular substructure. Red indicated the presence of the structural fragment with the feature value of 1. Blue indicated the absence of the structural fragment with the feature value of 0.

Morgan_267 were in the negative region, implying a lower probability that the structural fragments predicted by the model had an inhibitory effect on tumor cells (Figure 4). These results provided important insights into the relationship between molecular structure and anti-tumor activity. By analyzing the chemical structures of two common mTOR inhibitors, WYE-687 and KU-0063784, the results found that both contained the Morgan_935, Morgan_361, and Morgan_896 fingerprints, which were favorable for anti-tumor activity, as well as the Morgan_926, Morgan_831, and Morgan _361 fingerprints. Moreover, they did not contain the Morgan_64, Morgan_548, Morgan_435, Morgan_175, and Morgan_267 fingerprints, which were unfavorable for anti-tumor activity (Figure 5). These results suggested that both inhibitors could effectively inhibit mTOR. These results were highly consistent with the predictions of proposed model and experimental data of this study, further validating the model accuracy. These findings provided important guidance for the selection and optimization of mTOR inhibitors.

Conclusion

This study combined five machine learning algorithms and five deep learning algorithms to successfully construct 28 prediction models based on molecular descriptors, molecular fingerprints, and molecular graphs. By analyzing the data of 2,074 mTOR inhibitors and comparing the models using seven evaluation metrics, the results showed that the random forest algorithm combined with the Morgan molecular fingerprint model outperformed other models based on molecular descriptors and molecular graphs in terms of predictive performance, demonstrating its potential for identifying the key structural fragments related to mTOR inhibition. For a deeper understanding of the mechanisms underlying the constructed model, the SHAP method was used to explore important structural fragments in the model that had guiding significance for optimizing lead compounds and designing new reagents. A novel mTOR inhibitor prediction platform was developed using the random forest algorithm, providing researchers with a convenient tool for rapidly evaluating the

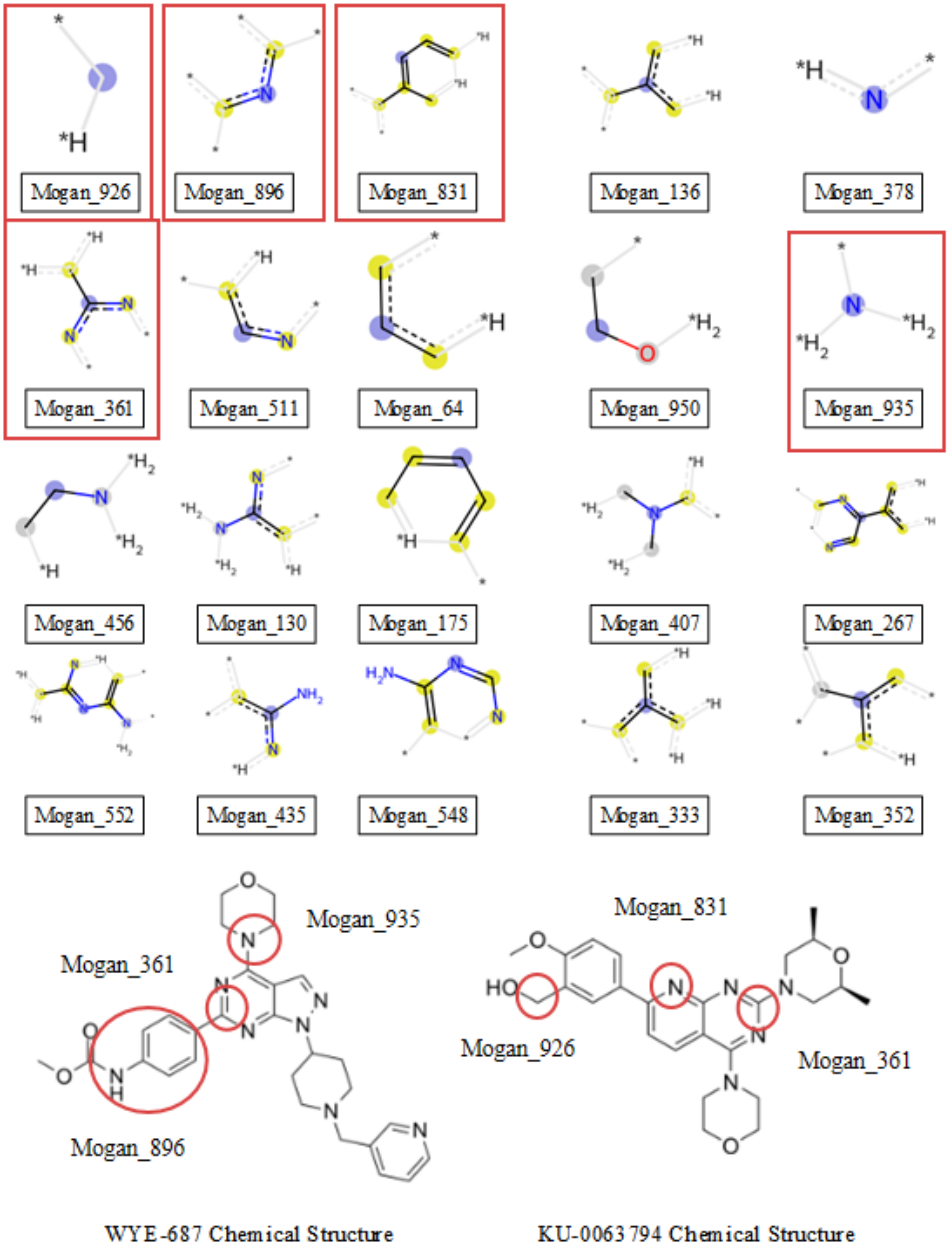

Figure 5. Random forest: Important molecular substructures of Morgan's model and chemical structures of common mTOR inhibitors.

mTOR inhibitory activities of the compounds. This research not only laid the foundation for the discovery of new mTOR inhibitors but also provided new ideas for personalized and precise cancer treatment. Future research is needed to further optimize the model and expand the dataset for improved accuracy and reliability of predictions, thereby promoting the translation of mTOR inhibitors to clinical applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62006092), the Natural Science Research Project of the Anhui Educational Committee (Grant No. 2023AH030081), the 2023 New Era Education Provincial Quality Engineering Project (Graduate Education) (Grant No. 2023cxcysj103), and the

2024 New Era Education Provincial Quality Engineering Project (Graduate Education).

Data availability statement

The data supporting the findings of this study are openly available at the following URL/DOI: https://data.mendeley.com/drafts/84jbmvzck5.

References

1. Laplante M, Sabatini DM. 2012. mTOR signaling in growth control and disease. Cell. 149(2):274-293.
2. Hidalgo M, Buckner JC, Erlichman C, Pollack MS, Boni JP, Dukart G, *et al*. 2006. A phase I and pharmacokinetic study of temsirolimus (CCI-779) administered intravenously daily for 5 days every 2 weeks to patients with advanced cancer. Clin Cancer Res. 12(19):5755-5763.
3. Vicary GW, Roman J. 2016. Targeting the mammalian target of rapamycin in lung cancer. Am J Med Sci. 352(5):507-516.
4. Knight ZA, Lin H, Shokat KM. 2010. Targeting the cancer kinome through polypharmacology. Nat Rev Cancer. 10(2):130-137.
5. Tropsha A. 2010. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 29(6-7):476-488.
6. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, *et al*. 2019. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 18(6):463-477.
7. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, *et al*. 2020. A deep learning approach to antibiotic discovery. Cell. 180(4):688-702.
8. Wang L, Li YC, Xu MY, Pang XQ, Liu ZH, Tan W, *et al*. 2016. Chemical fragment-based CDK4/6 inhibitors prediction and web server. RSC Adv. 6(21):16972-16981.
9. Fields FR, Freed SD, Carothers KE, Hamid MN, Hammers DE, Ross JN, *et al*. 2020. Novel antimicrobial peptide discovery using machine learning and biophysical selection of minimal bacteriocin domains. Drug Develop Res. 81(1):43-51.
10. Ashdown GW, Dimon M, Fan M, Sánchez-Román Terán F, Witmer K, Gaboriau DC, *et al*. 2020. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. Sci Adv. 6(39):eaba9338.
11. Zheng JX, Xia S, Lv S, Zhang Y, Bergquist R, Zhou XN, *et al*. 2021. Infestation risk of the intermediate snail host of *Schistosoma japonicum* in the Yangtze River Basin: Improved results by spatial reassessment and a random forest approach. Infect Dis Poverty. 10(03):34-46.
12. Kumari C, Abulaish M, Subbarao N. 2020. Exploring molecular descriptors and fingerprints to predict mTOR kinase inhibitors using machine learning techniques. IEEE ACM T Comput Bi. 18(5):1902-1913.
13. Wang PP, Xu XB, Li YH, Li B, Pei QL, Yu P, *et al*. 2021. Discovery of novel mammalian target of rapamycin (mTOR) inhibitors by support vector machine. IOP Conf Ser: Earth Environ Sci. 692(3):032028.
14. Zhang H, Zhuo HQ, Hou JJ, Cai JC. 2023. Machine learning models predict the mTOR signal pathway-related signature in the gastric cancer involving 2063 samples of 7 centers. Aging (Albany NY). 15(13):6152-6162.
15. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, *et al*. 2019. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 47(D1):D930-D940.
16. Zhong SF, Guan XH. 2023. Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties. Environ Sci Technol. 57(46):18193-18202.
17. Xia ZH, Yan AX. 2017. Computational models for the classification of mPGES-1 inhibitors with fingerprint descriptors. Mol Divers. 21:661-675.
18. Chen YT, Yu XX, Li WH, Tang Y, Liu GX. 2023. *In silico* prediction of hERG blockers using machine learning and deep learning approaches. J Appl Toxicol. 43(10):1462-1475.
19. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, *et al*. 2020. An open source chemical structure curation pipeline using RDKit. J Cheminformatics. 12:1-16.
20. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. 2016. Molecular graph convolutions: moving beyond fingerprints. J Comput Aid Mol Des. 30:595-608.
21. 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(6):1947-1958.
22. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. 2003. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. J Chem Inf Comput Sci. 43(6):2048-2056.
23. Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE T Inform Theory. 13(1):21-27.
24. Duda RO, Hart PE: Pattern classification and scene analysis. Volume 3. 1st edition. New York: Wiley. 1973:731-739.
25. McCulloch WS, Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 5:115-133.
26. Kipf TN, Welling M. 2016. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907.
27. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. 2018. Graph attention networks. arXiv:1710.10903.
28. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. 2017. Neural message passing for quantum chemistry. ICML. 70:1263-1272.
29. Xiong ZP, Wang DY, Liu XH, Zhong FS, Wan XZ, Li XT, *et al*. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 63(16):8749-8760.
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. 2011. Scikit-learn: Machine Learning in Python. arXiv:1201.0490.
31. Chen T, Guestrin C. 2016. Xgboost: A scalable tree boosting system. In Proc 2016 SIGKDD:785-794.

32. Kc GB, Bocci G, Verma S, Hassan MM, Holmes J, Yang JJ, *et al*. 2021. A machine learning platform to estimate anti-SARS-CoV-2 activities. Nat Mach Intell. 3(6):527-535.
33. Wu ZQ, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, *et al*. 2018. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 9(2):513-530.
34. Yang K, Swanson K, Jin WD, Coley C, Eiden P, Gao H, *et al*. 2019. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 59(8):3370-3388.
35. Wu ZH, Pan SR, Chen FW, Long GD, Zhang CQ, Yu PS. 2020. A comprehensive survey on graph neural networks. IEEE T Neur Net Lear. 32(1):4-24.
36. Jiang DJ, Wu ZX, Hsieh CY, Chen GY, Liao B, Wang Z, *et al*. 2021. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminformatics. 13:1-23.