RESEARCH

Open Access

Combined machine learning models, docking analysis, ADMET studies and molecular dynamics simulations for the design of novel FAK inhibitors against glioblastoma

Yihuan Zhao[1,2,3*], Xiaoyu He[1,2,3] and Qianwen Wan[1,2,3]

Abstract

Gliomas, particularly glioblastoma (GBM), are highly aggressive brain tumors with poor prognosis and high recurrence rates. This underscores the urgent need for novel therapeutic approaches. One promising target is Focal adhesion kinase (FAK), a key regulator of tumor progression currently in clinical trials for glioma treatment. Drug development, however, is both challenging and costly, necessitating efficient strategies. Computer-Aided Drug Design (CADD), especially when combined with machine learning (ML), streamlines the processes of virtual screening and optimization, significantly enhancing the efficiency and accuracy of drug discovery. Our study integrates ML, docking analysis, ADMET (absorption, distribution, metabolism, elimination, and toxicity) studies to identify novel FAK inhibitors specific to GBM. Predictive models showed strong performance, with an R^2 of 0.892, MAE of 0.331, and RMSE of 0.467 using protein-level IC_{50} data in combined CDK, CDK extended fingerprints, and substructure fingerprint counts derived from 1280 FAK inhibitors. Another model, based on IC_{50} data from 2608 compounds tested on U87-MG cells, achieved an R^2 of 0.789, MAE of 0.395, and RMSE of 0.536. Using these models, we efficiently identified 275 potentially active compounds out of 5107 candidates. Subsequent ADMET analysis narrowed this down to 16 potential FAK inhibitors that meet the established drug-likeness criteria. Moreover, molecular dynamics (MD) simulations validated the stable binding interactions between the selected compounds and the FAK protein. This study highlights the effectiveness of combining ML, docking analysis, and ADMET studies to rapidly identify potential FAK inhibitors from large databases, providing valuable insights for the systematic design of FAK inhibitors.

Keywords FAK inhibitors, Machine learning, Molecular docking, ADMET

*Correspondence:
Yihuan Zhao
2225694159@qq.com
[1] Key Laboratory of Basic Pharmacology of Guizhou Province, School of Pharmacy, Zunyi Medical University, Zunyi 563006, China
[2] Key Laboratory of Basic Pharmacology of Ministry of Education and Joint International Research Laboratory of Ethnomedicine of Ministry of Education, Zunyi Medical University, Zunyi 563006, China
[3] The Key Laboratory of Clinical Pharmacy of Zunyi City, Zunyi Medical University, Zunyi 563006, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Gliomas constitute over 30% of primary brain tumors, with more than half classified as the highly aggressive glioblastoma (GBM) [1]. GBM tumor cells exhibit low differentiation, high invasiveness into normal brain tissue, and strong resistance to conventional treatments, leading to a high recurrence rate and formidable therapeutic challenges [2]. Complete surgical resection of infiltrated tumor cells is challenging, while the remaining malignant cells exhibit potent anti-apoptotic capabilities, fueling tumor relapse [3]. The resistance of GBM to conventional treatments is attributed to its internal subpopulations of stem cells and highly mutated genome, complicating treatment strategies [4]. Despite significant advancements in medical technology, including maximal surgical resection, radiotherapy, and temozolomide chemotherapy, the prognosis for GBM patients has seen only marginal improvement over the past few decades. The median survival time remains disappointingly low, typically not exceeding 15 months [5]. Hence, there is an urgent need to explore novel therapeutic avenues. With an enhanced understanding of GBM biology, innovative targeted therapies, particularly focusing on integrin downstream signaling effectors like Focal adhesion kinase (FAK), have emerged as a key research focus [6]. Early clinical trials have demonstrated promising outcomes, instilling renewed optimism for precise GBM treatment [7, 8].

Focal adhesion kinase (FAK), belonging to the protein tyrosine kinase group, is essential in controlling cellular reactions to a range of signals, including integrins, cytokines, chemokines, and growth factors, thus exerting a notable impact on tumor advancement and spread [9, 10]. Studies have linked the FAK signaling pathway to crucial physiological and pathological processes like epithelial-mesenchymal transition, angiogenesis, cell migration, and invasion, thus impacting the aggressiveness of cancers, including gliomas [11]. Consequently, targeting FAK activity has emerged as an innovative strategy in cancer therapy. Ongoing research underscores FAK as a critical target for the treatment of brain gliomas, with numerous small molecule inhibitors currently undergoing preclinical and clinical evaluations [12, 13]. Therefore, the development of FAK inhibitors is of paramount importance.

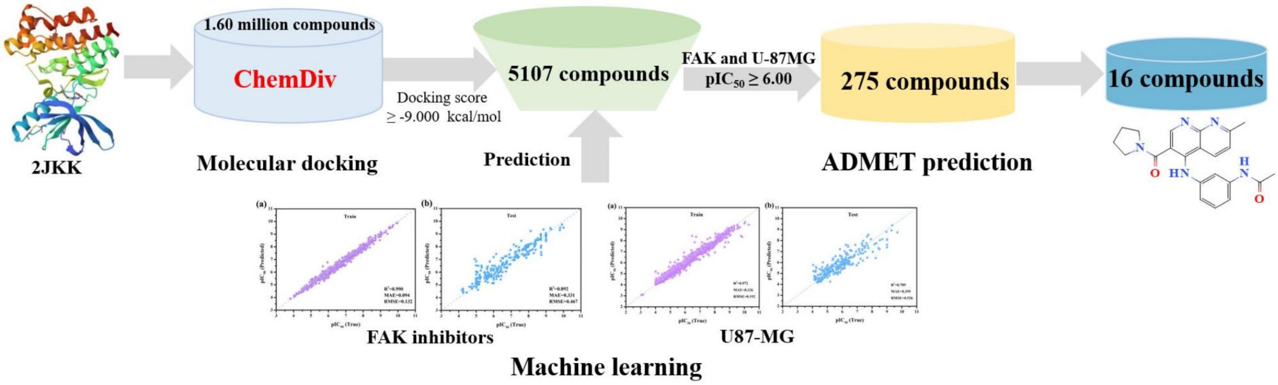
The process of developing novel drugs is not only expensive but also time-consuming. To reduce costs and enhance research efficiency, Computer-Aided Drug Design (CADD) is assuming an increasingly significant role in drug discovery [14]. Unlike traditional trial-and-error methods in the laboratory, CADD utilizes computer simulation techniques, enabling researchers to efficiently screen and optimize drugs in a virtual environment, significantly reducing unnecessary experimental steps and saving research costs [15]. In this regard, the construction of machine learning (ML) models is widely applied in CADD and has now permeated various stages of drug development [16–18]. Currently, several studies have applied machine learning to predict the activity of FAK inhibitors [19–22]. For example, Wang et al. utilized 22 FAK inhibitors to develop two types of satisfactory 3D-QSAR models [21]: the comparative molecular field analysis (CoMFA) model (R^2_{cv}=0.528, R^2_{pred}=0.7557) and the comparative molecular similarity indices analysis (CoMSIA) model (R^2_{cv}=0.757, R^2_{pred}=0.8362), for predicting the inhibitory activities of novel inhibitors. However, the current QSAR models used to predict FAK inhibition activity have relatively small construction datasets, typically comprising only tens to hundreds of compounds, which somewhat limits the widespread applicability of the models. Therefore, it is particularly important and urgent to construct a predictive model based on a relatively large dataset (with more than 1000 data points) to enable rapid screening of FAK inhibitors for glioblastoma.

In this study, we propose a comprehensive research framework that integrates machine learning models, docking analysis, ADMET studies, and molecular dynamics simulations to facilitate the design of novel FAK inhibitors against glioblastoma (Scheme 1). This integrated approach aims to accelerate the identification and development of potential FAK inhibitors, providing a valuable strategy for advancing GBM therapeutics.

Materials and methods

Development of the database

The modeling dataset, comprising molecular structures and their corresponding inhibitory activity against FAK (expressed as half-maximal inhibitory concentration IC_{50}), was retrieved from the CHEMBL database (CHEMBL2695, https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL2695/) on January 24, 2024 [23], initially consisting of 4730 entries. Specific IC_{50} values were used for compounds enabling the calculation of pIC_{50}. The study employed the base-10 logarithm of IC_{50} (represented as -logIC_{50}, denoted as pIC_{50}) as the dependent variable, rather than IC_{50}. In cases where the same compound displayed varying IC_{50} values within a narrow range (10 μm), the average was calculated as the final IC_{50} value, resulting in IC_{50} values for 1280 compounds. The distribution of pIC_{50} for the 1280 FAK inhibitors is depicted in Fig. 1a, with values ranging from 4.00 to 10.00, predominantly falling between 5.00 and 9.50. The examination of plain ring and Murcko scaffolds for the 1280 inhibitors was performed using DataWarrior software [24], which resulted in identifying 121 unique plain rings and 449 unique Murcko scaffolds. The configurations of the top 28 plain rings are presented in Fig. S1.



Scheme 1 Flowchart of combined machine learning models, docking analysis, and ADMET studies for the design of novel FAK inhibitors against human malignant glioblastoma

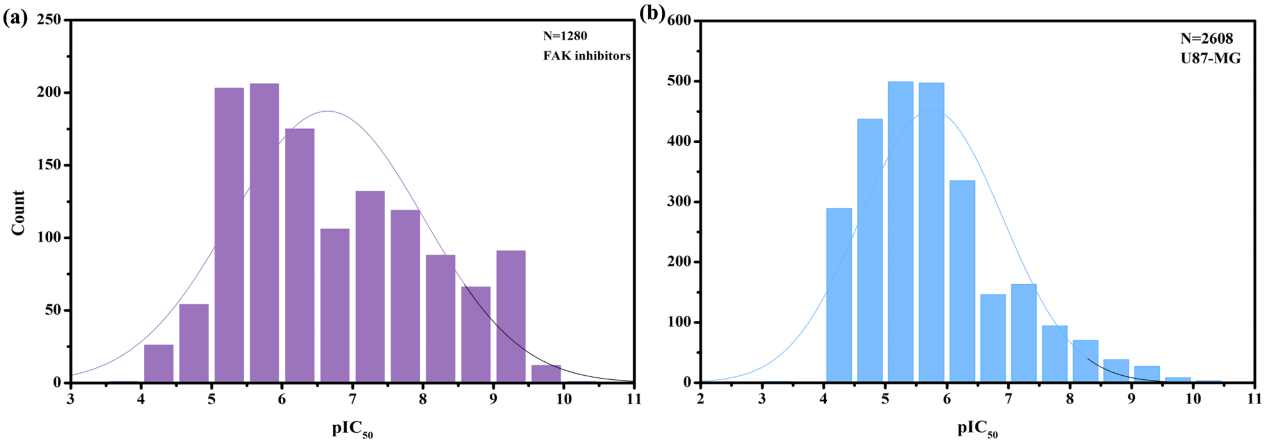


Fig. 1 The pIC_{50} distribution histogram of (**a**) 1280 FAK inhibitors (**b**) 2608 compounds against U87-MG

Notably, 35 of these plain rings were found to occur more than 10 times, with additional information available in Supporting Information Table S1. Supporting Information Table S2 offers detailed insights into the top 50 Murcko scaffolds, highlighting the structural diversity within the FAK inhibitor dataset.

The dataset of compounds targeting U-87 MG glioblastomas was also obtained from the CHEMBL database (CHEMBL3307575, https://www.ebi.ac.uk/chembl/cell_line_report_card/CHEMBL3307575/), initially comprising 11,597 entries. Specific IC_{50} values were utilized for compounds to enable the calculation of pIC_{50}. We utilized data for compounds with specific IC_{50} values (allowing the calculation of pIC_{50}). In cases where the same compound exhibited minor variations in IC_{50} values (within a range of 10 μm), the average was computed as the final IC_{50} value. Data exhibiting differences greater than 10 μM were excluded, resulting in IC_{50} values for 2608 compounds. The pIC_{50} distribution histogram for these 2608 compounds against U87-MG glioblastomas is depicted in Fig. 1b, demonstrating a predominant pIC_{50} range of 4.00-7.50. As in the previous analysis, we utilized DataWarrior software to thoroughly investigate the ring systems and Murcko backbones of 2608 compounds targeting U-87 MG glioblastomas. This analysis identified 347 distinct common rings and 2678 unique Murcko scaffolds. Fig. S2 illustrates the structures of the top 28 most frequently occurring common rings, with 84 of them appearing more than 10 times, as detailed in Supporting Information Table S3. Furthermore, Supporting Information Table S4 provides a comprehensive list of the top 50 Murcko scaffolds, showcasing not only their prevalence in the compounds but also emphasizing the structural diversity within the U-87MG dataset.

Molecular descriptions calculation and machine learning models building

Molecular fingerprints were utilized as inputs for the models to transform molecules into numerical vectors through fingerprints computation, enabling their integration into machine learning algorithms. These fingerprints encapsulate molecular structures in binary or continuous values, reflecting the diversity and similarity across molecules [25]. In this study, we successfully converted molecules into numerical vectors of fixed lengths through the application of molecular descriptors. This method of conversion ensures that key chemical properties of the molecules are retained while also enabling more efficient and standardized data processing and analysis in later stages. PaDEL software was employed in this research to compute molecular fingerprints [26]. Four categories of fingerprints were computed: CDK fingerprints, CDK extended fingerprints, substructure fingerprints, and substructure fingerprint counts. Additional details regarding these molecular descriptors can be found in Table S5. The regression models leveraged LightGBM, Random Forest, SVR, KNN, PLS, LASSO, and XGBoost algorithms. The LightGBM algorithm was prioritized in this study due to its advantages as an ensemble learning algorithm over conventional methods [27]. Training and testing with established datasets linking molecular descriptors to pIC_{50} values enabled the development of pIC_{50} machine learning models. The datasets were split into training and independent test sets, maintaining an 80:20 ratio. Ten-fold Cross-validation was implemented during model training to mitigate the impact of random data partitioning, while optimization techniques such as hyperparameter tuning were applied to enhance model performance. Grid search methodology was utilized to determine optimal parameters for the models. Tables S6 and S7 present the optimal parameters for the algorithms employed in this study. In this research, multiple predictive models were constructed utilizing the Scikit-learn within the Python 3.7 environment. To comprehensively evaluate the predictive performance of these regression models, we employed various metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2 score, abbreviated as R^2). The corresponding formula is shown below:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(y^i_{pred} - y^i_{exp}\right)^2}{\sum_{i=1}^{m}\left(y^i_{pred} - \bar{y}\right)^2} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y^i_{pre} - y^i_{exp}\right| \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y^i_{pre} - y^i_{exp}\right)^2} \quad (3)$$

In this equation, y_{exp} represents the experimental values, while y_{pred} corresponds to the predicted values. $\bar{y}$ refers to the average of the predicted values, and m indicates the sample size. All associated source code and datasets are publicly available at the following link: https://github.com/Yihuan-Zhao93/FAKML.

Molecular docking studies

The crystal structure of the FAK protein required for this study (PDB ID: 2JKK) was sourced from the internationally recognized Protein Data Bank (PDB, available at http://www.rcsb.org/pdb) [28]. To ensure accurate molecular docking, the original crystal structure of the FAK protein retrieved from the PDB was subjected to a series of preprocessing steps. Prior to docking, all crystallized water molecules and the bound ligand from the original PDB file were eliminated, and partial charges were assigned to the atoms. The ligand was subsequently converted into a three-dimensional structure and saved in pdbqt format using Autodock software [29]. Both the protein and ligand were subsequently prepared for molecular docking using AutoDock Vina. The compounds were docked into the active site of the FAK protein, defined by a grid box with dimensions of 22.5×22.5×22.5 ($Å^3$), centered at the binding location of the co-crystallized ligand. The binding energy was the key measure of compound activity, where lower values signified stronger ligand-receptor interactions. To gain a more detailed understanding of the interaction mechanisms between the ligand and FAK protein, comprehensive analysis was performed using the Discovery Studio Visualizer software [30].

ADME prediction

One of the leading reasons for failures in drug discovery is linked to poor pharmacokinetics and toxicity issues. As a result, it is crucial to evaluate the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of compounds at an early stage of drug development. In this research, we utilized the ADMETlab 2.0 [31] to predict the ADMET characteristics of FAK inhibitors identified via our predictive models. By inputting the SMILES strings of the compounds, the platform automatically generated ADMET-related predictions.

MD simulation

In this study, the Gromacs 2022 software platform [32] was used for exhaustive molecular dynamics (MD) simulations. First, based on the docking results, the proteins were separated from the small molecule ligands. Force field parameters were generated for the small-molecule ligands using the antechamber tool in the AmberTools suite, and these parameters were converted to a Gromacs software-compatible format using the ACPYPE tool. During simulations, small molecule ligands were described using the Generalized Amber Force Field (GAFF), while proteins were modeled combining the AMBER14SB force field with the TIP3P water model.

Next, the processed protein and ligand files were merged to construct the complete simulation system. The simulations were performed at constant temperature and pressure, and periodic boundary conditions were applied to ensure the stability of the system. During the simulation, the LINCS algorithm was used to constrain the hydrogen bonding and the time step was set to 2 femtoseconds. Electrostatic interactions were calculated by the Particle Mesh Ewald (PME) method with the cutoff distance set to 1.2 nm, while non-bonding interactions were calculated using a cutoff distance of 10 angstroms and the interaction list was updated every 10 steps. The system temperature was regulated at 298 K using a V-rescale thermostat, while pressure was maintained at 1 bar by a Berendsen barostat. Prior to the production molecular dynamics (MD) simulation, the system was equilibrated for 100 picoseconds under the NVT and NPT ensembles. A production MD simulation was then conducted for 100 nanoseconds, with system snapshots taken every 10 picoseconds. Finally, the simulation trajectories were analyzed using VMD and PyMOL software.

Table 1 Performance of the ML models on the ten-fold cross-validation of test set using LightGBM method

Fingerprint	Input	Test set		
	Number	R^2	MAE	RMSE
CDK	1024	0.876 ± 0.006	0.356 ± 0.010	0.500 ± 0.012
CDK extended	1024	0.881 ± 0.005	0.353 ± 0.006	0.489 ± 0.011
Substructure	307	0.785 ± 0.006	0.492 ± 0.007	0.658 ± 0.009
Substructure count	307	0.816 ± 0.006	0.452 ± 0.007	0.609 ± 0.010
CDK + CDK extended	2048	0.887 ± 0.007	0.339 ± 0.010	0.477 ± 0.014
Substructure + Substructure count	614	0.816 ± 0.006	0.452 ± 0.007	0.610 ± 0.010
CDK + Substructure count	1331	0.881 ± 0.005	0.349 ± 0.006	0.490 ± 0.010
CDK extended + Substructure count	1331	0.883 ± 0.004	0.351 ± 0.006	0.487 ± 0.008
CDK + Substructure + Substructure count	1638	0.881 ± 0.005	0.349 ± 0.006	0.490 ± 0.010
CDK extended + Substructure + Substructure count	1638	0.883 ± 0.004	0.351 ± 0.006	0.487 ± 0.008
CDK + CDK extended + Substructure	2355	0.885 ± 0.006	0.344 ± 0.009	0.482 ± 0.012
CDK + CDK extended + Substructure count	2355	0.888 ± 0.005	0.342 ± 0.006	0.476 ± 0.010
CDK + CDK extended + Substructure + Substructure count	2662	0.888 ± 0.005	0.342 ± 0.006	0.476 ± 0.010

Results and discussion

Development of ML models for FAK inhibitors

In this research, the representation of FAK inhibitors within the dataset encompassed CDK, CDK extended, substructure fingerprint, substructure fingerprint counts and their respective counts. CDK fingerprints are composed of one-dimensional arrays, 1024 bits long, organized based on the existence of distinct structural elements. Extended CDK fingerprints represent an advancement over the standard CDK iteration, incorporating supplementary ring features. Similarly, the substructure fingerprints and counts were created as 307-bit binary strings, where each bit denoted the presence of SMARTS patterns for functional group classification by Christian Laggner. Initially, an analysis was conducted to evaluate the impact of various model inputs on predictive performance to determine the optimal model input. Table 1 summarizes the performance of the test set using LightGBM-based machine learning models, presenting the averaged R^2, MAE, and RMSE values from ten-fold cross-validation. The outcomes indicated that using a single molecular fingerprint resulted in lower predictive performance compared to combined molecular fingerprints, emphasizing the complementarity of combined fingerprints in enhancing the representation of relevant molecular structural information. The model achieved the highest predictive accuracy when integrating combined CDK, CDK extended, and substructure fingerprint counts. Although combining all molecular fingerprints could yield similar results, the computational efficiency of utilizing these three fingerprints guided the selection for subsequent model development. Table S8 presents a thorough summary of the prediction outcomes for both the training and test sets, conducted without the use of cross-validation. The findings demonstrate that the model exhibits superior performance when utilizing a combination of CDK, extended CDK, and substructure fingerprint counts as input features. Fig. 2 further illustrates the predictive performance of the optimal model on both the training and independent test sets. As shown, the training set achieves an R^2 value as high as 0.990, with a mean absolute error (MAE) of 0.094 and a root mean square error (RMSE) of 0.132. On the independent test set, the model still attains an R^2 value of 0.892, while the MAE and RMSE are 0.331 and 0.467, respectively. These findings highlight the accurate predictive capability of LightGBM-based molecular fingerprint models in determining the activity (pIC_{50}) of FAK inhibitors.

In order to assist researchers in designing FAK inhibitors, we conducted an analysis of feature importance. Fig. 3a depicts the top ten significant features identified

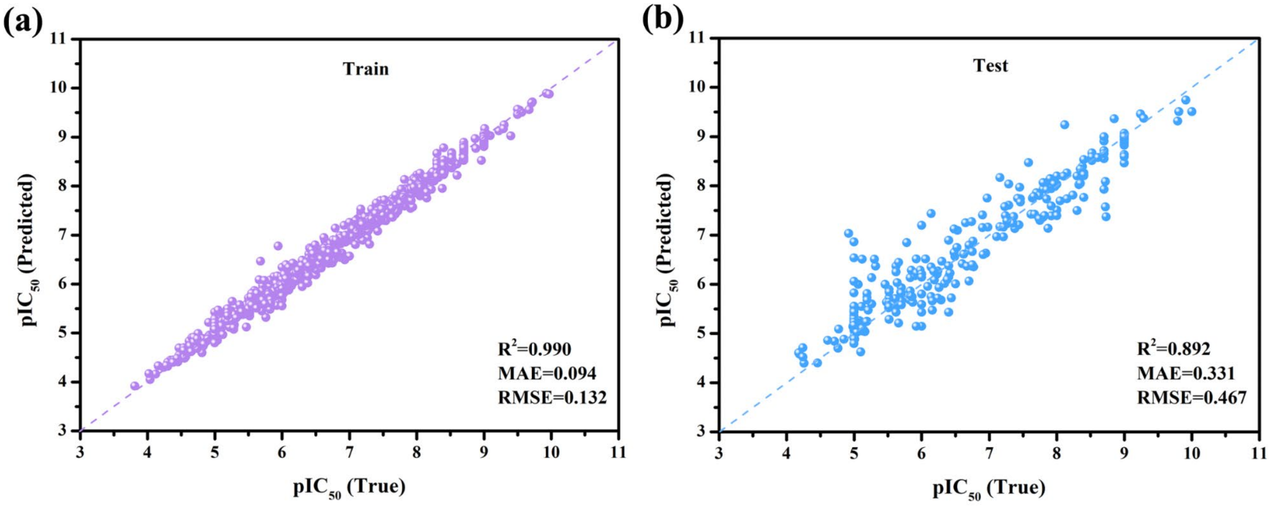


Fig. 2 The true pIC$_{50}$ value and predicted value obtained from LightGBM model of (**a**) Training set and (**b**) Independent test set

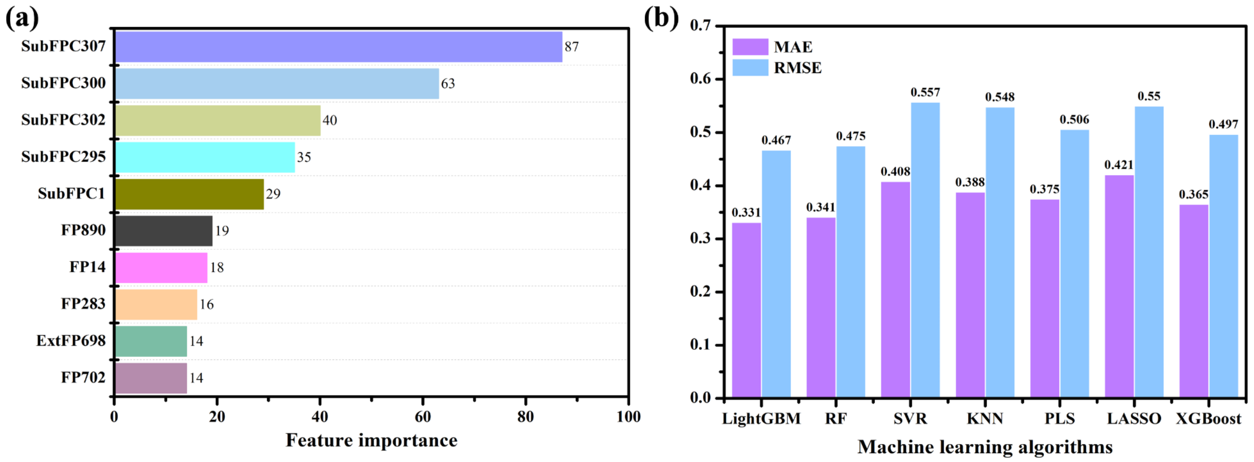


Fig. 3 (**a**) the feature importance plot and (**b**) the predictive performance of different algorithms using the optimal model parameters

through the LightGBM algorithm, with SubFPC307 emerging as the most critical. SubFPC307 indicates the presence of a specified chiral center, highlighting a specific atom in the molecule exhibiting chirality. Subsequently, SubFPC300 (1,3-Tautomerizable) relates to the potential for 1,3-tautomerization, SubFPC302 (Rotatable bond) signifies a bond capable of rotation, SubFPC295 (CONS bond) represents a conserved bond type, and SubFPC1 (Primary carbon) denotes a primary carbon atom. Moreover, the fingerprints FP890, FP14, FP283, ExtFP698, and FP702 correspond to specific structural elements or features within the molecule, providing valuable insights into their chemical properties and functionalities. Moreover, Fig. 3b compares the predictive performance of various methods on the independent test set when using optimal model parameters (see Table S6 for details). It is noteworthy that the LightGBM algorithm produces the lowest MAE and RMSE values, highlighting its superior predictive accuracy. Therefore, in the subsequent model construction, we will continue to choose this algorithm as our first option for model development.

Development of ML models for U-87 MG glioblastomas

Firstly, we selected the optimal combination of inputs for the model. Table 2 illustrates the evaluation results of the machine learning models using cross-validation on the test set with the LightGBM approach. The results demonstrate that the highest model performance is attained when CDK, CDK extended fingerprints, and substructure fingerprint counts are employed as input variables. Simultaneously, Table S9 details the prediction results for the training and test sets without cross-validation, highlighting the maximum accuracy achieved using a combination of CDK, CDK extended fingerprints, and substructure fingerprint counts. This finding aligns with

Table 2 Performance of the ML models on the ten-fold cross-validation of test set using LightGBM method

Fingerprint	Input	Test set		
	Number	R^2	MAE	RMSE
CDK	1024	0.760 ± 0.012	0.417 ± 0.009	0.572 ± 0.014
CDK extended	1024	0.756 ± 0.015	0.421 ± 0.006	0.577 ± 0.018
Substructure	307	0.650 ± 0.009	0.491 ± 0.006	0.691 ± 0.009
Substructure count	307	0.709 ± 0.013	0.462 ± 0.012	0.630 ± 0.014
CDK + CDK extended	2048	0.765 ± 0.010	0.413 ± 0.011	0.565 ± 0.012
Substructure + Substructure count	614	0.710 ± 0.013	0.461 ± 0.012	0.629 ± 0.014
CDK + Substructure count	1331	0.769 ± 0.010	0.398 ± 0.009	0.562 ± 0.012
CDK extended + Substructure count	1331	0.770 ± 0.010	0.412 ± 0.009	0.560 ± 0.012
CDK + Substructure + Substructure count	1638	0.769 ± 0.010	0.398 ± 0.009	0.562 ± 0.012
CDK extended + Substructure + Substructure count	1638	0.770 ± 0.010	0.412 ± 0.009	0.560 ± 0.012
CDK + CDK extended + Substructure	2355	0.765 ± 0.010	0.416 ± 0.010	0.566 ± 0.011
CDK + CDK extended + Substructure count	2355	0.782 ± 0.010	0.398 ± 0.006	0.545 ± 0.013
CDK + CDK extended + Substructure + Substructure count	2662	0.782 ± 0.010	0.398 ± 0.006	0.545 ± 0.013

the FAK inhibitor pIC_{50} activity model, underscoring the enhanced model performance with the incorporation of complementary molecular structural information. Fig. 4 depicts the predictive accuracy of the optimal model using the LightGBM algorithm. The training set achieved an R^2 value of 0.972, an MAE value of 0.126, and an RMSE value of 0.192, while the testing set showed an R^2 value of 0.789, an MAE value of 0.395, and an RMSE value of 0.536. These results indicate that molecular fingerprint models based on LightGBM are proficient in predicting the activity (pIC_{50}) of compounds against U-87 MG glioblastomas.

To assist researchers in designing compounds for combatting U-87 MG glioblastomas, we conducted an analysis of feature importance. Fig. 5a illustrates the top ten significant features using the LightGBM algorithm. SubFPC307 is ranked as the most crucial, representing a "Chiral center specified", indicating a specific atom within the molecule exhibiting chirality. Subsequently, SubFPC300 (1,3-Tautomerizable) relates to the potential for 1,3-tautomerization between two isomeric forms, SubFPC295 (CONS bond) signifies a conserved bond type, SubFPC302 (Rotatable bond) indicates a bond capable of rotation, SubFPC1 (Primary carbon) denotes a primary carbon atom, SubFPC18 (Alkylarylether) refers to a functional group combining alkyl and aryl moieties, SubFPC2 (Secondary carbon) signifies a secondary carbon atom, SubFPC274 (Aromatic) represents an aromatic system, while SubFPC2745 (Heterocyclic) and SubFPC2745 (Hetero N nonbasic) indicate the presence of heterocyclic structures containing non-basic nitrogen atoms. In Fig. 5b, a comparison of predictive performance among various algorithms under optimal model parameters on the independent test set (refer to Table S7) is presented. It is evident from the figure that employing the LightGBM algorithm minimizes MAE and RMSE values, indicating superior predictive performance. These results collectively highlight the efficiency of the LightGBM algorithm

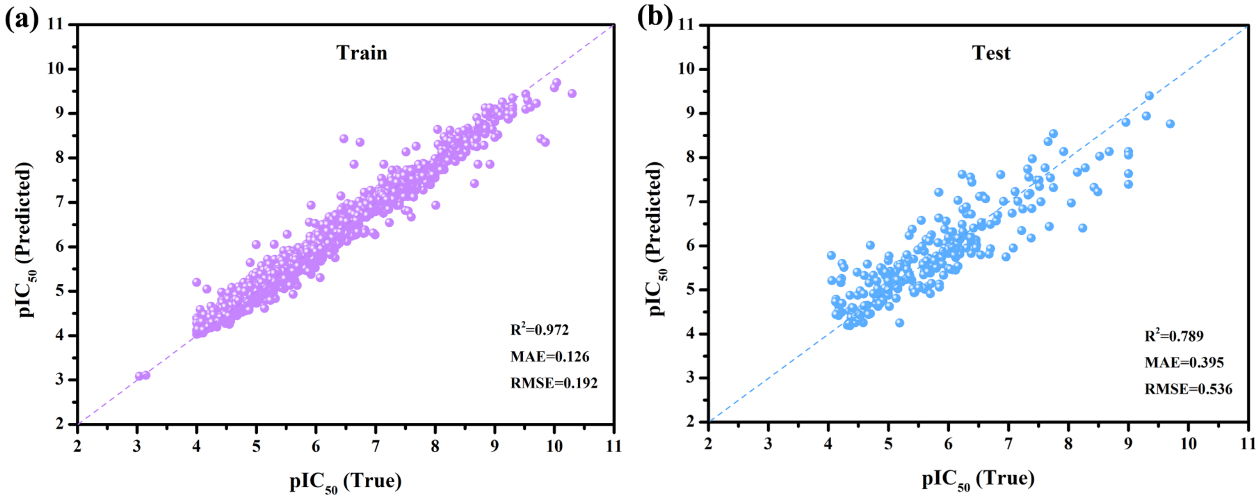


Fig. 4 The true pIC_{50} value and predicted value obtained from LightGBM model of (**a**) Train set and (**b**) Independent test set

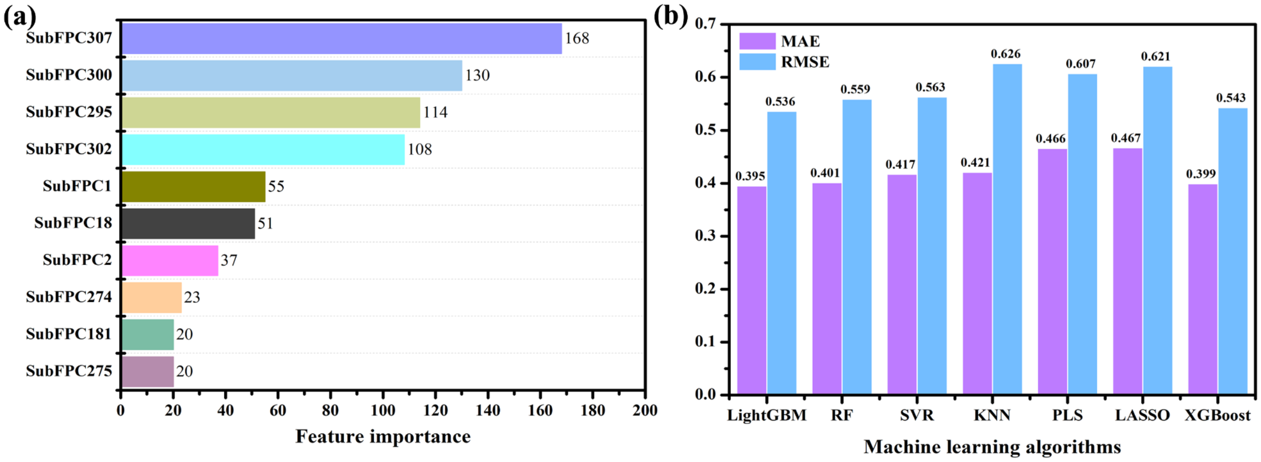


Fig. 5 (**a**) the feature importance plot (**b**) the predictive performance of different algorithms under the optimal model parameters

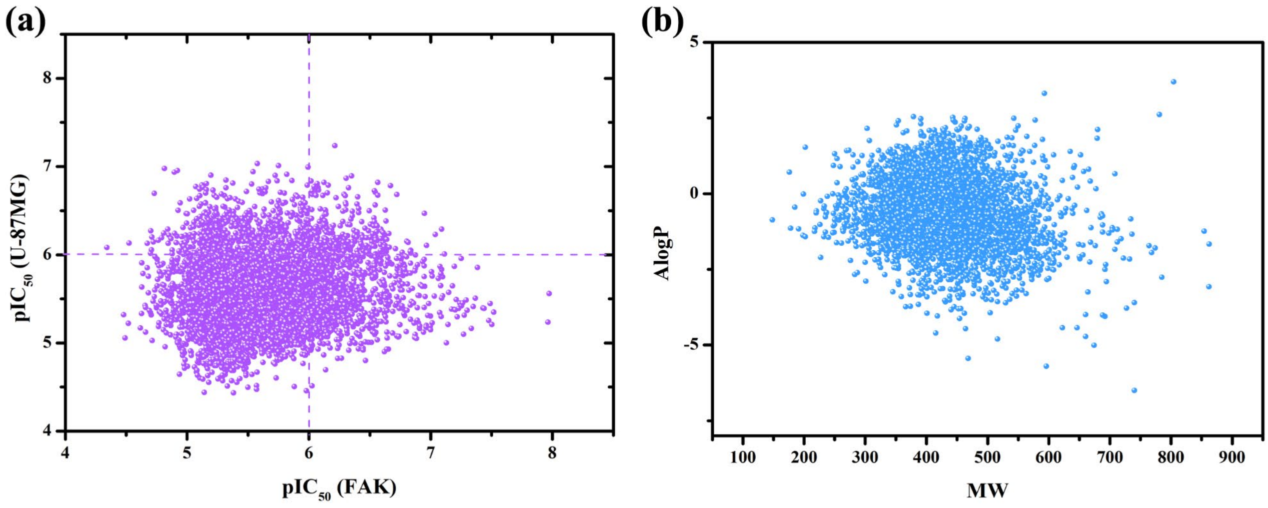


Fig. 6 (**a**) Prediction of pIC_{50} for 5107 compounds (**b**) The chemical space distribution of 5107 molecules

in accurately constructing models to predict the IC_{50} activity of organic small compounds.

Molecular docking analysis

Prior to the molecular docking procedure, the reliability of the methodology was evaluated through an initial validation phase. The crystallographically resolved ligand, TAE-226, was extracted and repositioned into the active pocket of the FAK protein. As illustrated in Supplementary Fig. S3, the superimposition of the cocrystallized and redocked structures yielded a Root Mean Square Deviation (RMSD) of 0.581 Å, indicative of precise pose recapitulation. Further docking analysis unveiled a docking scores of -9.613 kcal/mol between TAE-226 and FAK, corroborating their robust interaction. Subsequently, the ChemDiv compound library, comprising an extensive collection of 1.6 million compounds, was subjected to docking simulations within the identical binding interface utilizing the predefined algorithmic parameters. In line with reference [22], compounds exhibiting docking score values equal to or below −9.000 kcal/mol were selected for further investigation. Following the application of a docking score cut-off of -9.000 kcal/mol, 5107 compounds were selected as potential FAK inhibitors for further analysis. Fig. 6a displays the pIC_{50} values of the 5107 compounds predicted by the models, revealing a subset meeting the pIC_{50} criteria for the FAK protein and exceeding 6.00 at the cellular level for U-87MG. Among the screened compounds, 275 exhibited pIC_{50} values surpassing 6.00 for both the FAK protein and U-87MG cells, constituting 5.38% of the total compounds. The subsequent analysis will focus on predicting the ADMET properties of these 275 compounds. The DataWarrior software was employed to analyze the plain ring and Murcko scaffolds of the 5107 compounds, leading to the identification of 391 plain rings and 2277 Murcko scaffolds. Additionally, the molecular weight (MW) and AlogP values for all 5,107 compounds were computed and then represented in a two-dimensional plot to facilitate the analysis of the chemical space. Fig. 6b visually represents the chemical space distribution among the 5107 compounds, highlighting a diverse array of compounds. This broad chemical diversity enhances the potential for discovering FAK molecules with distinct pharmacological effects, presenting promising avenues for FAK inhibitor development.

ADME prediction analysis

Pharmacokinetic profiling and safety evaluation are crucial components of the drug discovery process. To thoroughly assess these essential factors, we utilized the advanced ADMETlab 2.0 platform, which provides precise predictions of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties for target compounds. Building on this, we applied four well-established drug-like evaluation criteria derived from extensive experience in drug development: Lipinski's rule of five, the Pfizer rule, the GSK rule, and the Golden Triangle. These benchmarks enabled a comprehensive evaluation of the potential efficacy and drug-likeness of our newly-designed FAK inhibitors. Out of 275 candidates, 16 compounds (Fig. 7) were identified as fully meeting all four drug-likeness criteria. Comprehensive ADMET data for these 16 compounds is presented in Table S10. With the exception of compounds 4, 6, and 14, all other compounds demonstrate excellent human intestinal absorption (HIA). Among compounds 1, 2, 4, 6, 14, and 15, which exhibit Caco-2 cell permeability exceeding −5.15, the remaining compounds showcase optimal permeability in the Caco-2 model, and all compounds display favorable permeability in the MDCK cell model. Less than half of the compounds possess appropriate plasma protein binding (PPB) values (≤90%); however, the volume of distribution (VD) values for all compounds fall within acceptable ranges. In addition to PPB and VD, evaluating blood-brain barrier (BBB) penetration is critical for understanding compound distribution. Among the compounds, only 1, 4, 5, 7, and 11 demonstrate ideal BBB penetration. CYP2D6 and CYP3A4, integral members of the cytochrome P450 (CYP) enzyme family, are involved in metabolizing over 70% of approved drugs. In this context, the FAK candidates are likely substrates of CYP2D6 or CYP3A4 enzymes, except for compounds 2, 4, 5, 12, and 16, indicating favorable metabolic characteristics. Furthermore, all candidate compounds exhibit

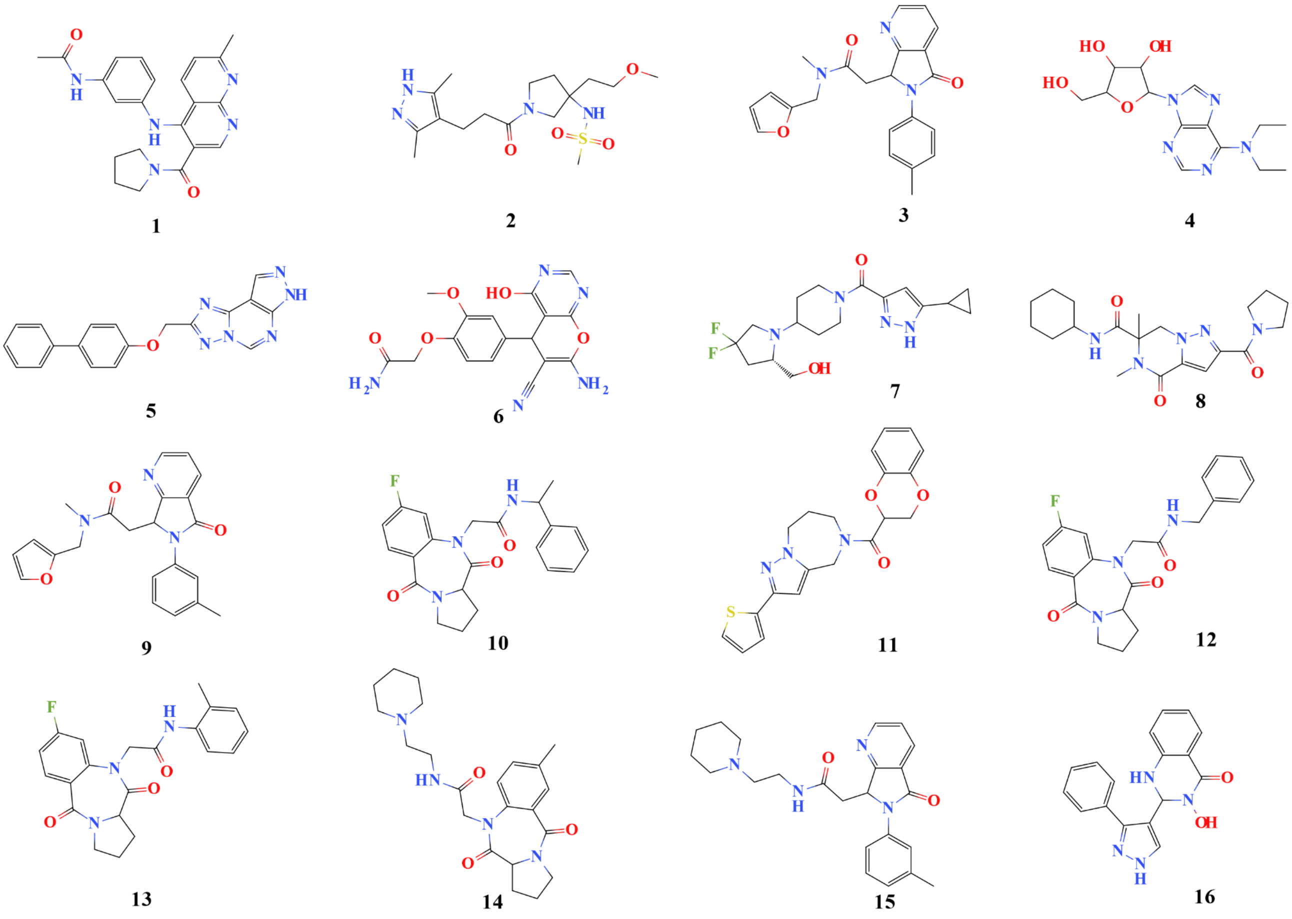


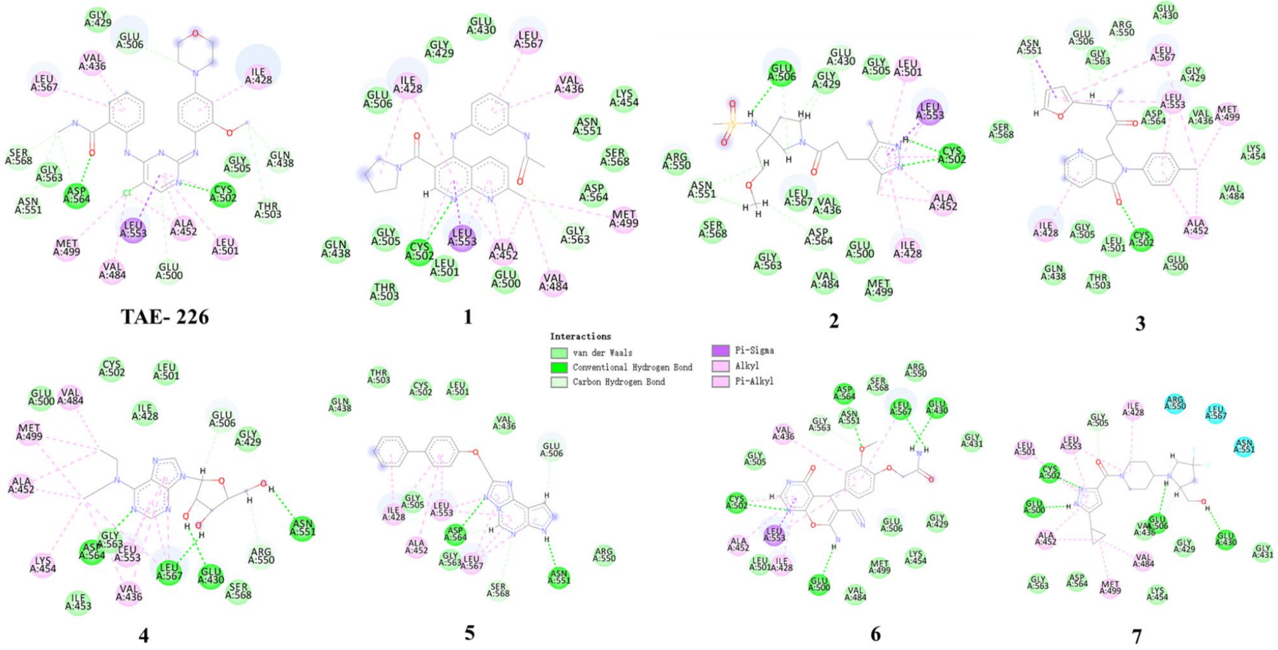
Fig. 7 The chemical structural of 16 potential new FAK inhibitors

appropriate half-lives ($T_{1/2}$), with the majority showing no toxicity in the Ames test, although compound 15 may possess mutagenic potential.

The binding patterns of the top seven potential compounds are depicted in Fig. 8. The results demonstrate that TAE-226, used as a positive control, forms strong interactions with FAK through hydrogen bonds involving the Cys502 and Asp564 residues. According to our statistical analysis (Table S11), most of the selected compounds bind to FAK by forming hydrogen bonds with Cys502 or Asp564, with some also interacting with Leu567, Glu430, Glu500, and Glu506 residues. Additionally, other types of interactions, such as Van der Waals forces, π-alkyl, and π-σ interactions, were observed between TAE-226 and FAK. Specifically, TAE-226 forms π-σ interactions with the Leu553 residue, and statistical data (Table S11) suggest that other compounds predominantly interact with Leu553 through π-σ interactions, highlighting its significant role. Several residues, including Leu501, Leu553, Leu567, Ala452, Met499, and Ile428, contribute to π-alkyl interactions. The docking analysis identifies critical residues linked to FAK's biological effects, namely Cys502, Asp564, Leu553, Leu501, Leu567, Ile428, Ala452, and Met499. Furthermore, Table S12 compares the chemical similarity between 16 FAK hits and their closest ChEMBL FAK ligands based on ECFP4 fingerprints, revealing low structural similarity between these compounds and known FAK inhibitors, underscoring the novelty of the screened compounds.

MD simulation analysis

To further explore the binding stability between specific small molecules and the FAK protein, we representatively selected the complexes of compounds 1 and 2 with the FAK protein and conducted molecular dynamics (MD) simulation experiments. Root mean square deviation (RMSD), a crucial measure for assessing the overall deviation of atomic positions from a reference structure over time, was employed as a primary indicator of system stability [33]. As depicted in Fig. 9 (a), the RMSD values for both the complex and the protein remained consistent throughout the simulation, signifying that the compound 1-FAK complex reached a stable state. Similarly, Fig. 9 (b) shows stable RMSD values for the complex and protein, indicating that compound 7 also achieved stability when bound to FAK. To further analyze the system, energy components such as solvation energy, RMSD, buried solvent-accessible surface area, and interaction energies were taken into account, and the steady-state trajectories were selected for Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) calculations. The resulting energy contributions to binding energy are summarized in Table 3. Upon examination of the data, it becomes evident that the van der Waals energy (ΔE_{vdw}) plays a dominant role in binding for both compounds 1 and 2 with FAK, surpassing electrostatic interaction energy (ΔE_{ele}), while hydrophobic and electrostatic forces provide additional contributions. The calculated binding free energy (ΔE_{MMPBSA}) for compound 1 was -77.909 ± 0.603 kJ/mol, indicating a strong binding affinity with FAK, whereas for compound 2, the value was



TAE- 226

1

2

3

4

5

6

7

Fig. 8 The binding mode of TAE-226 and seven potential FAK inhibitors with FAK

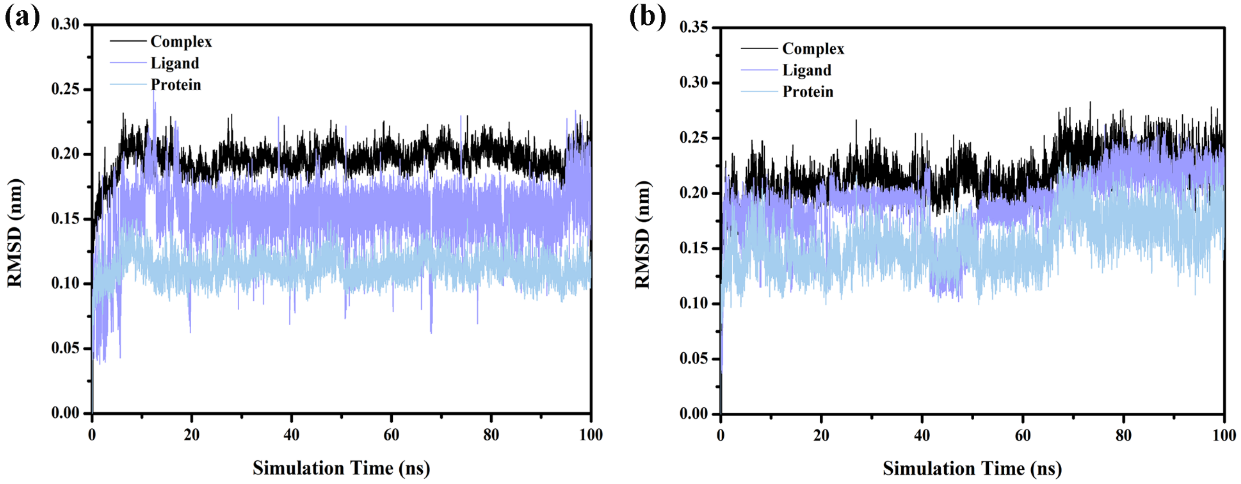


Fig. 9 RMSD values obtained from the MD simulation system for the interactions between FAK and (**a**) compound 1, and (**b**) compound 2

Table 3 The calculated binding free energy (kJ/mol) for the compound 1-FAK and compound 2-FAK complexes

Terms	compound 1-FAK	compound 7-FAK
ΔE_{vdw} (van der Waals energy)	−118.413 ± 1.109	−189.313 ± 1.377
ΔE_{ele} (electrostatic energy)	−14.905 ± 0.767	−27.539 ± 0.619
ΔE_{pol} (polar solvation free energy)	71.707 ± 1.349	123.149 ± 1.212
ΔE_{nonpol} (non-polar solvation free energy)	−16.298 ± 0.126	−23.069 ± 0.177
$\Delta E_{MMPBSA}{}^{*}$	−77.909 ± 0.603	−116.772 ± 0.447
-TΔS	19.589 ± 2.642	20.366 ± 5.081
$\Delta G_{bind}{}^{*}$ (calculated Gibbs free energy)	−58.321 ± 2.852	−96.406 ± 4.933

* $\Delta E_{MMPBSA} = \Delta E_{vdw} + \Delta E_{ele} + \Delta E_{pol} + \Delta E_{nonpol}$ $\Delta G_{bind} = \Delta E_{MMPBSA}$ -TΔS

−116.772±0.447 kJ/mol, similarly reflecting high binding affinity. The above results indicate that the selected compounds can stably bind to the FAK protein.

Conclusion

In this study, we employed a comprehensive strategy that integrates machine learning, docking analysis, and ADMET predictions to expedite the identification of FAK inhibitors specifically tailored for human malignant glioblastoma. Our predictive models exhibited high accuracy, with an R^2 of 0.892, an MAE of 0.331, and an RMSE of 0.467 for protein-level FAK inhibitors, and an R^2 of 0.789, an MAE of 0.395, and an RMSE of 0.536 for U87-MG cell-based compounds. Using these models, we efficiently identified 275 potentially active compounds out of 5107 candidates based on docking scores lower than −9.000 kcal/mol. Following ADMET assessments, 16 compounds stood out as potential FAK inhibitors. Furthermore, molecular dynamics simulations verified the stability of the binding interactions between the selected compounds and the FAK protein. This study demonstrates the utility of combining computational methods for the efficient identification of FAK inhibitors from large chemical libraries. Nevertheless, further experimental validation through in vitro and in vivo studies is required to support clinical application.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13065-024-01316-x.

Supplementary Material 1

Acknowledgements

This project is supported by Science and Technology Plan Project of Guizhou (No. Qian Science Platform Talent [2020]-041), Science and Technology Plan Project of Guizhou (No. Qian Science Platform Talent [2021]1350-017).

Author contributions

Yihuan Zhao designed the research. Yihuan Zhao performed the research and wrote the manuscript. Xiaoyu He and Qianwen Wan contributed to data analysis. Yihuan Zhao revised the manuscript. All authors reviewed the manuscript.

Funding

No. Qian Science Platform Talent [2020]-041, No. Qian Science Platform Talent [2021]1350-017.

Data availability

All source codes and dataset can be accessed at https://github.com/Yihuan-Zhao93/FAKML.

Declarations

Ethical approval consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors. We have not performed work on human/animal, therefore, there is no need of informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 May 2024 / Accepted: 3 October 2024
Published online: 18 October 2024

References

1. Clarke J, Butowski N. S.J.A.o.n. Chang, recent advances in therapy for glioblastoma, 67 (2010) 279–83.
2. Grobben B, De Deyn P. H.J.C. Slegers, t. research, rat C6 glioma as experimental model system for the study of glioblastoma growth and invasion, 310 (2002) 257–70.
3. Furnari FB, Fenton T, Bachoo RM, Mukasa A, Stommel JM, Stegh A, Hahn WC, Ligon KL, Louis DN. C.J.G. Brennan, development, malignant astrocytic glioma: genetics, biology, and paths to treatment, 21 (2007) 2683–710.
4. Garnier D, Renoult O, Alves-Guerra M-C, Paris F. C.J.F.i.o. Pecqueur, Glioblastoma stem-like cells, metabolic strategy to kill a challenging target, 9 (2019) 118.
5. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C. U.J.N.E.j.o.m. Bogdahn, Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma, 352 (2005) 987–96.
6. Liu T-J, LaFortune T, Honda T, Ohmori O, Hatakeyama S, Meyer T, Jackson D, de Groot J. W.A.J.M.c.t. Yung, Inhibition of both focal adhesion kinase and insulin-like growth factor-I receptor kinase suppresses glioma proliferation in vitro and in vivo, 6 (2007) 1357–67.
7. Shi Q, Hjelmeland AB, Keir ST, Song L, Wickman S, Jackson D, Ohmori O, Bigner DD, Friedman HS, J.N., Rich. A novel low-molecular weight inhibitor of focal adhesion kinase, TAE226, inhibits glioma growth, 46 (2007) 488–96.
8. Yang M, Li Y, Chilukuri K, Brady OA, Boulos MI, Kappes JC. D.S.J.J.o.n.-o. Galileo, L1 stimulation of human glioma cell motility correlates with FAK activation, 105 (2011) 27–44.
9. M.D.J.J.o.c.s. Schaller, Cellular functions of FAK kinases: insight into molecular mechanisms and novel functions, 123 (2010) 1007–13.
10. Serrels A, Lund T, Serrels B, Byron A, McPherson RC, von Kriegsheim A, Gomez-Cuadrado L, Canel M, Muir M, Ring JEJC. Nuclear FAK controls chemokine transcription, Tregs, and evasion of anti-tumor immunity, 163 (2015) 160–73.
11. Cornillon J, Campos L. D.J.M.S.m.s. Guyotat, Focal adhesion kinase (FAK), une protéine aux fonctions multiples, 19 (2003) 743–752.
12. Shanthi E, Krishna MH, Arunesh GM, Venkateswara Reddy K, Sooriya Kumar J. V.N.J.E.o.o.t.p. Viswanadhan, focal adhesion kinase inhibitors in the treatment of metastatic cancer: a patent review, 24 (2014) 1077–100.
13. Lv P-C, Jiang A-Q, Zhang W-M, H.-, Zhu. FAK Inhibitors Cancer Patent Rev. 2018;28:139–45.
14. Macalino SJY, Gosu V, Hong S. S.J.A.o.p.r. Choi, Role of computer-aided drug design in modern drug discovery, 38 (2015) 1686–701.
15. Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG. Structure-based virtual screening: from classical to Artificial Intelligence. Front Chem, 8 (2020).
16. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. 2021;16:949–59.
17. Cui WQ, Aouidate A, Wang SG, Yu QLY, Li YH, Yuan SG. Discovering Anti-Cancer Drugs < i > via Computational Methods, Front Pharmacol, 11 (2020).
18. Baig MH, Ahmad K, Roy S, Ashraf JM, Adil M, Siddiqui MH, Khan S, Kamal MA, Provazník I, Choi I. Computer aided Drug Design: Success and limitations. Curr Pharm Design. 2016;22:572–81.
19. Sun C-c, Feng L-j, Sun X-h, Yu R-l, Chu Y-y. -m. Kang, Study on the interactions of pyrimidine derivatives with FAK by 3D-QSAR, molecular docking and molecular dynamics simulation. New J Chem. 2020;44:19499–507.
20. Shirvani P, Fassihi A. Silico design of novel FAK inhibitors using integrated molecular docking, 3D-QSAR and molecular dynamics simulation studies. J Biomol Struct Dynamics. 2022;40:5965–82.
21. Wang F, Yang W, Li R, Sui Z, Cheng G, Zhou B. Molecular description of pyrimidine-based inhibitors with activity against FAK combining 3D-QSAR analysis, molecular docking and molecular dynamics. Arab J Chem. 2021;14:103144.
22. Tang L, Wu Z, Zhang Q, Hu Q, Dang X, Cui F, Tang L, Xiao T. A sequential light-harvesting system with thermosensitive colorimetric emission in both aqueous solution and hydrogel. Chem Commun. 2024;60:4719–22.
23. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52:D1180–92.
24. Sander T, Freyss J, Von Korff M, Rufener C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model. 2015;55:460–73.
25. Yuan Y, Zheng F, Zhan C-G. Improved prediction of blood–brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints. AAPS J. 2018;20:1–10.
26. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32:1466–74.
27. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–54.
28. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 2021;49:D437–51.
29. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.
30. Studio D. Discovery studio, Accelrys [2.1], 420 (2008).
31. Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, Yin M, Zeng X, Wu C, Lu A. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49:W5–14.
32. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem. 2005;26:1701–18.
33. Shi R, Liu Y, Ma Y, Li J, Zhang W, Jiang Z, Hou J. Insight into binding behavior, structure, and foam properties of α-lactalbumin/glycyrrhizic acid complex in an acidic environment. Food Hydrocolloids. 2022;125:107411.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.