BMC Chemistry

RESEARCH

Open Access

Discovery of novel NLRP3 inhibitors based on machine learning and physical methods

Tao Jiang[1†], Shijing Qian[2†], Jinhong Xu[1], Shuihong Yu[3], Yang Lu[1], Linsheng Xu[1*] and Xiaosi Yang[3*]

Abstract

The NLRP3 inflammasome plays a crucial role in inflammatory responses, particularly in alcohol-related liver disease (ALD). Given that NLRP3 has emerged as a potential therapeutic target for ALD, the development of effective inhibitors is of great importance. In this study, we trained 11 regression models, and the results showed that LightGBM, Random Forest, and XGBoost performed the best, achieving R^2 values of 0.774, 0.755, and 0.719, respectively. Using machine learning models and physical methods, we screened more than 11.5 million compounds from Asinex, Princeton, UkrOrgSynthesis, Chemdiv, Chembridge, Alinda, Enamine, and Lifechemicals, which led to the identification of 26 potential NLRP3 inhibitors. Furthermore, molecular dynamics simulations and MMGBSA binding energy calculations confirmed the stability of the interactions between NLRP3 and three key molecules: 19,655,631 (source Chembridge), 38,214,692 (source Chembridge), and Z1180203703 (source Enamine). Additionally, ADMET analysis revealed their favorable pharmacokinetic properties. This study provides insights and candidate molecules for discovering NLRP3 inhibitors, potentially applicable in treating related diseases.

Keywords NLRP3 inflammasome, NLRP3 inhibitors, Machine learning, Molecular docking, Molecular dynamics simulation, Drug discovery

Introduction

Inflammatory responses are common physiological and pathological processes that occurs in response to external invaders. Inflammasomes play a crucial role in this process, with NLRP3 (NOD-like receptor family pyrin domain-containing protein 3) being one of the key proteins involved. NLRP3 is a critical member of the NOD-like receptor (NLR) family, which includes NLRP2, NLRP4, NLRP6, NLRP7, and others [1]. Among these, NLRP3 is the most prominent member of the NLR family. It consists of a pyrin domain (PYD), a NACHT domain required for ATPase activity, and a leucine-rich repeat (LRR) motif [2–4]. Upon detection of activation signals, NLRP3 transitions from an inactive homotypic oligomer to an active oligomeric inflammasome, promoting the assembly of the adaptor molecule ASC, activating caspase-1, and inducing the proteolytic cleavage and activation of pro-inflammatory cytokines from the IL-1 family and gasdermin D [5]. Aberrant activation of NLRP3 can trigger the onset of inflammatory diseases. Growing evidence indicates that NLRP3 significantly contributes to alcohol-related liver disease (ALD), making it a potential therapeutic target. ALD encompasses a range of alcohol-induced liver disorders, including alcoholic steatohepatitis

†Tao Jiang and Shijing Qian contributed to the work equally and should be regarded as co-first authors.

*Correspondence:
Linsheng Xu
987854635@qq.com
Xiaosi Yang
yxsi@aqmc.edu.cn
[1]Anqing 116 Hospital, No.150 Shuangjing Street, Yingjiang District, Anqing City, Anhui Province, China
[2]Tongji Hospital of Tongji University, No. 389 Xincun Road, Putuo District, Shanghai City, China
[3]School of Basic Medical Sciences, Anqing Medical College, No.1588, Jixian North Road,Yixiu District, Anqing City, Anhui Province, China

(ASH), cirrhosis, and hepatocellular carcinoma (HCC) [6]. The liver regulates a wide range of critical physiological processes and plays an essential role in activating the innate immune system, which initiates inflammatory events. Chronic ethanol exposure disrupts hepatic inflammatory mechanisms and leads to the release of pro-inflammatory mediators, such as chemokines, cytokines, and the activation of inflammasomes. The mechanisms underlying liver fibrosis/cirrhosis involve the activation of the NLRP3 inflammasome [7]. Studies have shown that prolonged alcohol exposure activates CYP2E1 in hepatocytes, leading to antioxidant system dysfunction and excessive production of reactive oxygen species (ROS) and inducible nitric oxide synthase (iNOS) [8, 9]. This results in endoplasmic reticulum stress and activates the inflammatory response through the TLR4/MyD88/NF-κB signaling axis, significantly promoting NLRP3 inflammasome activation [10, 11]. Once activated, the NLRP3 inflammasome exacerbates the inflammatory signaling cascade through its sustained release of pro-inflammatory mediators. An in vivo study demonstrated that ethanol-fed mice exhibited significantly higher expression of inflammasome components, including NLRP3, ASC, and caspase-1, compared to control mice [12]. In contrast, the absence of NLRP3 inflammasome components reversed the increase in pro-inflammatory cytokine-mediated steatosis and hepatic injury in alcohol-exposed mice [13]. These studies underscore the critical role of NLRP3 in the pathogenesis of ALD. While current research has advanced our understanding of the mechanisms underlying ALD, effective pharmacological treatments remain lacking. Over the past decade, significant progress has been made in elucidating the role of NLRP3 inflammasome formation and activation in liver injury and the specific contributions of upstream and downstream signaling pathways involved. As such, the development of NLRP3-targeted drugs holds great potential for the treatment of ALD.

In recent years, the development of machine learning (ML) techniques has garnered significant attention from drug developers due to their efficiency and cost-effectiveness. ML plays a crucial role in drug discovery and development [14, 15]. Recently, Maryam Zulfat and others constructed a classification model to discover small molecule inhibitors targeting NLRP3 for the treatment of Alzheimer's disease, achieving an accuracy of up to 94% for the best model. Additionally, Cheng Shi and colleagues successfully screened the molecule CSC-6 through the development of a machine learning classification model and structure-based drug discovery methods, with an IL-1β inhibition effect of 2.3±0.38 μM in PMA-THP-1 cells. These reports indicate the feasibility of machine learning and structure-based drug discovery methods. However, there have been no reports on constructing machine learning regression models for the discovery of small molecule inhibitors targeting NLRP3. In this study, we describe an approach that integrates machine learning regression models with molecular docking, molecular dynamics (MD) simulations, ADMET predictions, and MM-PBSA calculations to predict novel and potential inhibitors targeting NLRP3. This approach aims to provide insights for drug development in the clinical treatment of ALD.

Materials and methods

Data collection

The dataset of NLRP3 molecules along with their activity data for machine learning modeling was sourced from the CHEMBL database [16] and patents. The data collection process from the CHEMBL database was as follows: we searched for the keyword "NLRP3" and selected the human NLRP3 target. On the target page, we downloaded the IC50 distribution data from the Activity Charts. Based on the downloaded CSV file, rows where the "Standard Relation" was not "=" were removed. We excluded non-IL-1β activity types based on the descriptions in the "Assay Description" column, and converted "Standard Units" to "nM." Finally, we retained the columns for SMILES and Standard Value. A total of 398 molecular structures and their NLRP3 activity data were collected.

For molecular data from patents, we retrieved data from nine patents containing NLRP3 molecules with IL-1β inhibition activity: WO2021214284A1, WO2018015445, WO20181674681, WO2020234715A1,WO2021150574A1,WO2021209539A1,WO2021209552A1, WO2021214284A1, and WO2020021447A1. Molecular structures from the patents were sketched using ChemDraw 20.0, and IC50 values were extracted. This provided 825 additional data points. In total, we obtained 1,223 data points from both sources. IC50 values were then converted to pIC50.

Descriptor calculation and data splitting

In this study, molecules were represented using Morgan fingerprints (with a radius of 3 and a length of 2048 bits), calculated using the RDKit 2022.09.5 package. These fingerprints encode molecular structures by traversing each atom and its bonding relationships, with hashing operations ensuring uniform vector lengths. As a result, molecular fingerprints are typically high-dimensional, sparse binary vectors (0/1), which are well-suited for machine learning models such as support vector machines and fully connected neural networks that handle high-dimensional sparse vectors effectively.

In addition, since molecular properties such as molecular weight, solubility, and surface area are often related to activity and drug-likeness. The following descriptors were also computed using the RDKit 2022.09.5 package: mol_weight, log_p, num_h_donors, num_h_acceptors, tpsa, num_rota1_bonds, num_aromatic_rings,

num_aliphatic_rings, num_saturated_rings, num_heteroatoms, num_valence_electrons, num_radical_electrons, and qed. These additional descriptors supplemented the information provided by molecular fingerprints, addressing gaps in domain knowledge in pharmaceutical science. Both were combined to form the feature vectors.

We randomly split the dataset of 1,223 molecules into an 80:20 ratio using scikit-learn's 'train_test_split' function, with 80% for training and 20% for testing.

Regression model training and testing

To explore the performance of various regression models in predicting pIC50 values, we compared a range of classical regression algorithms, including linear regression, ridge regression, Lasso regression, ElasticNet regression, support vector regression (SVR), K-nearest neighbor (KNN) regression, decision tree regression, random forest regression, gradient boosting regression, XGBoost regression, LightGBM regression, and multilayer perceptron (MLP) neural network regression.

The selected regression models span linear models, regularized models, nonlinear models, and ensemble learning models. Linear regression and its regularized versions (Ridge, Lasso, ElasticNet) served as baseline models to assess their performance on simple linear relationships. Nonlinear models such as SVR and KNN were employed to capture complex nonlinear relationships between molecular structures and pIC50 values. Ensemble models, including random forest, gradient boosting, XGBoost, and LightGBM, were chosen for their ability to handle large, complex datasets and exploit high-dimensional features effectively. Additionally, the MLP model, as a neural network, learned complex nonlinear mappings through multiple hidden layers. XGBoost regression was implemented using the "xgboost 2.1.0" package, LightGBM regression with the "lightgbm 4.4.0" package, and the remaining algorithms were implemented using the scikit-learn 1.0.2 package.

Evaluation metrics

For model evaluation, the study employed Mean Squared Error (MSE) and the coefficient of determination (R^2) as primary performance metrics. These metrics reflect the model's prediction error and goodness-of-fit, providing an intuitive understanding of model performance in predicting pIC50 values.

Mean Squared Error (MSE) is a metric used to measure the difference between predicted and true values, calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

where n is the number of samples, y_i is the predicted value for the i-th sample, and $\widehat{y_i}$ is the true value. A lower MSE indicates smaller prediction error and higher accuracy. In this study, MSE was used to assess the error in predicting pIC50 values, with lower MSE values indicating that the model better captured the relationship between molecular descriptors and pIC50.

Coefficient of Determination (R^2) measures the goodness-of-fit of a model, ranging from 0 to 1. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the true values. Higher R^2 values, closer to 1, indicate stronger explanatory power of the model. An R^2 of 1 indicates perfect fit, while an R^2 of 0 means the model has no explanatory power. In this study, R^2 was used to evaluate how well the model fit the pIC50 data. Higher R^2 values suggest the model captured the variance in the target variable effectively. By comprehensively analyzing both MSE and R^2, we could evaluate the predictive performance of various regression models and select the optimal model for pIC50 prediction.

Database preparation

In this study, a combined dataset of 11,526,814 molecules was used, sourced from various commercial compound databases, including Asinex, Princeton, UkrOrgSynthesis, Chemdiv, Chembridge, Alinda, Enamine, and Lifechemicals. This extensive compound collection greatly increases the possibility of discovering novel NLRP3 inhibitors.

Virtual screening

First, the LightGBM, Random Forest, and XGBoost models were used to predict the pIC50 values for the database compounds. Compounds predicted to have a pIC50 greater than 6 by all three models were considered potential candidates. These selected molecules were then subjected to subsequent molecular docking and binding affinity prediction analyses.

Before docking, the 3D structures of the molecules were prepared using the LigPrep module in the Schrödinger software suite. This process involved energy minimization of the molecules under the OPLS4 force field, protonation state prediction using the Epik method [17], and removal of any salts or ions from the molecular library. The cleaned molecules were then used for structure-based virtual screening. The crystal structure of the NLRP3 protein (PDB ID: 7ALV [18]) retrieved from the PDB database (https://www.rcsb.org/) was used as the basis for virtual screening. The Protein Preparation Wizard module in Maestro 13.0 was employed to prepare the protein, including adjustments to bond orders, charge assignments,

removal of water molecules, addition and optimization of hydrogen atoms, protonation state prediction of amino acids at pH 7.4, and energy minimization under the OPLS4 force field. The receptor grid files for the docking site were generated using the Receptor Grid Generation module in Schrödinger 2022-3, with the co-crystal ligand from the prepared structure selected as the center of the active site.

Docking and affinity calculations were performed using the Virtual Screening Workflow (VSW) module in Schrödinger 2022-3, employing HTVS, SP, and XP algorithms in sequence to screen the prepared molecules. Finally, binding energies were further refined using the prime MMGBSA method to select the most promising compounds.

Molecular dynamics simulation (MD)

Molecular dynamics (MD) simulations were performed using the AMBER 22 software suite [19, 20]. Before the simulation, BCC charges for the small molecules were calculated using the antechamber module [21]. Small molecules and proteins were described using the GAFF2 [22] and ff14SB force fields [23], respectively. Hydrogen atoms were added using the LEaP module, and the system was solvated in a TIP3P octahedral water box [24], with periodic boundaries set to 10 Å. Na+/Cl- ions were added to neutralize the system, and the topology and parameter files were generated for simulation.

Energy minimization was performed using 2,500 steps of steepest descent and 2,500 steps of conjugate gradient minimization. The system was then heated from 0 K to 298.15 K over 200 ps under constant volume. This was followed by 500 ps of NVT ensemble simulation at 298.15 K to allow for uniform solvent distribution. Lastly, 500 ps of NPT ensemble equilibration was performed, followed by a 100 ns NPT production run. Non-bonded interactions were truncated at 10 Å, and long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method [25]. SHAKE was used to constrain hydrogen bond lengths [26], while temperature control was maintained using the Langevin thermostat [27], with a collision frequency of 2 ps^{-1}. The pressure was maintained at 1 atm, and the integration time step was set to 2 fs. Trajectories were saved every 10 ps for subsequent analysis.

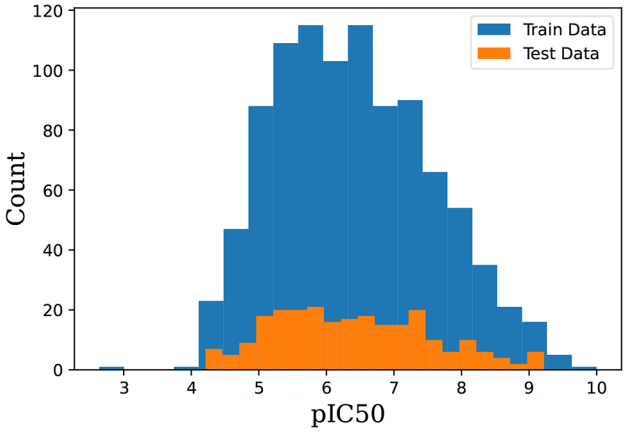

Fig. 1 Distribution of activity data in the training and test sets

MM/GBSA binding free energy calculation

Binding free energies between proteins and ligands were calculated using the MM/GBSA method [28, 29]. The MD trajectories from 90 to 100 ns were used for the calculation, and the binding free energy (ΔG_{bind}) was determined using the following equation:

$$\Delta G_{bind} = \Delta E_{internal} + \Delta E_{VDW} + \Delta E_{elec} + \Delta G_{GB} + \Delta G_{SA}$$

In this equation, $\Delta E_{internal}$ represents internal energy, ΔE_{VDW} denotes van der Waals interactions, and ΔE_{elec} refers to electrostatic interactions. ΔG_{GB} and ΔG_{GA} represent the solvation free energy, where $\Delta G_{GB}$is the polar solvation energy and ΔG_{GA} is the non-polar solvation energy. ΔG_{GB} was calculated using the GB model developed by Nguyen et al. [30], and ΔG_{GA} was calculated as the product of surface tension (γ) and solvent-accessible surface area (SASA) [31], with $\Delta G_{SA} = 0.0072 \times SASA$. Entropic contributions were omitted due to their high computational cost and low precision [28].

ADMET prediction

ADMET properties were predicted using the QikProp module in Schrödinger 2022-3. The QikProp module was used to predict properties such as QPlogPo/w, QPlogS, QPlogHERG, QPPCaco, QPlogBB, and human oral absorption.

Results and discussion

Data distribution and model training

A total of 1,223 NLRP3 inhibitor molecules, each with IC50 data for IL-β inhibition, were collected from the CHEMBL database and nine patents. Before modeling, IC50 values were converted to pIC50 and randomly split into a training set and test set in a 4:1 ratio. The data distribution after splitting is shown in Fig. 1. The distribution of the test and training sets follows an approximately normal distribution, and both sets display similar patterns. This indicates that our random split is uniform, providing a solid foundation for accurate modeling.

Molecular representation refers to numerical depictions of molecular properties, such as molecular descriptors, fingerprints, SMILES strings, and potential energy functions [32]. However, when predicting molecular properties for biochemical mechanisms that remain unclear, it can be challenging for scientists to design effective molecular descriptors, leading to failures in constructing QSAR models. Since molecular properties are largely determined by molecular structure, including

functional groups, researchers aim to incorporate molecular bonding relationships into QSAR modeling. In this study, we used Morgan fingerprints (with a radius of 3 and a length of 2048 bits) as one of the molecular representation methods. Additionally, given that we are studying drug molecules, we included drug-likeness-related molecular descriptors, such as molecular weight, solubility, and surface area, to complement the molecular representation. Based on the molecular representation and pIC50 data, we followed the machine learning workflow shown in Fig. 2 to train models using linear regression, ridge regression, Lasso regression, ElasticNet regression, support vector regression, K-nearest neighbor regression, decision tree regression, random forest regression, gradient boosting regression, XGBoost regression, LightGBM regression, and multilayer perceptron (MLP) neural network regression.

Model testing and selection of the best model

We calculated the residuals between predicted and experimental data for each model on the test set. The results are shown in Fig. 3, where a smaller residual (Abs

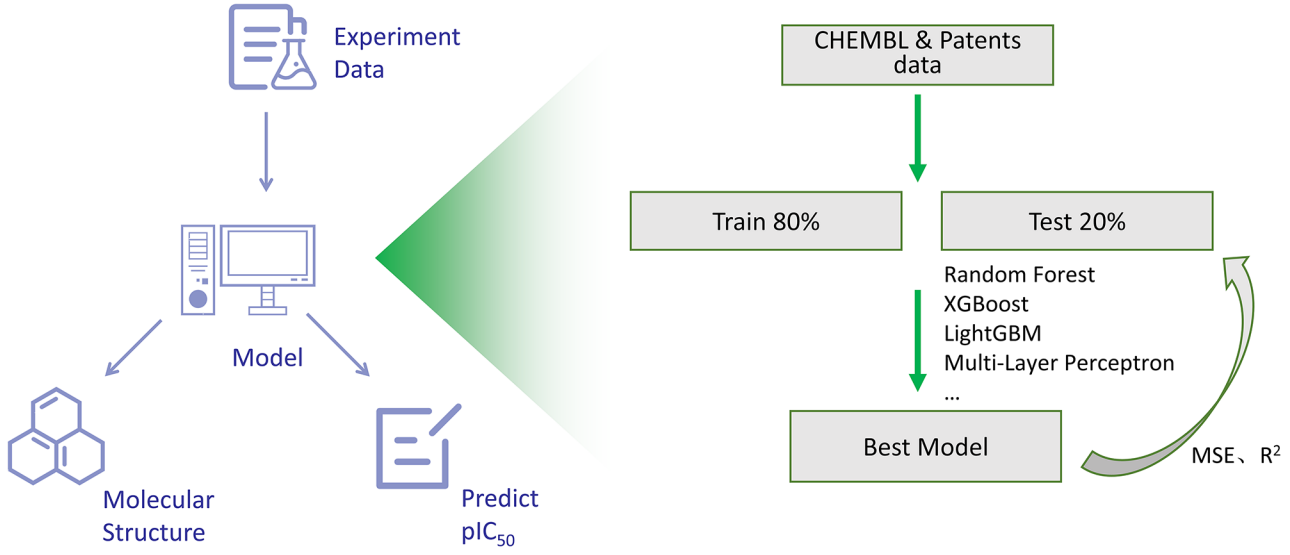

Fig. 2 Machine learning workflow

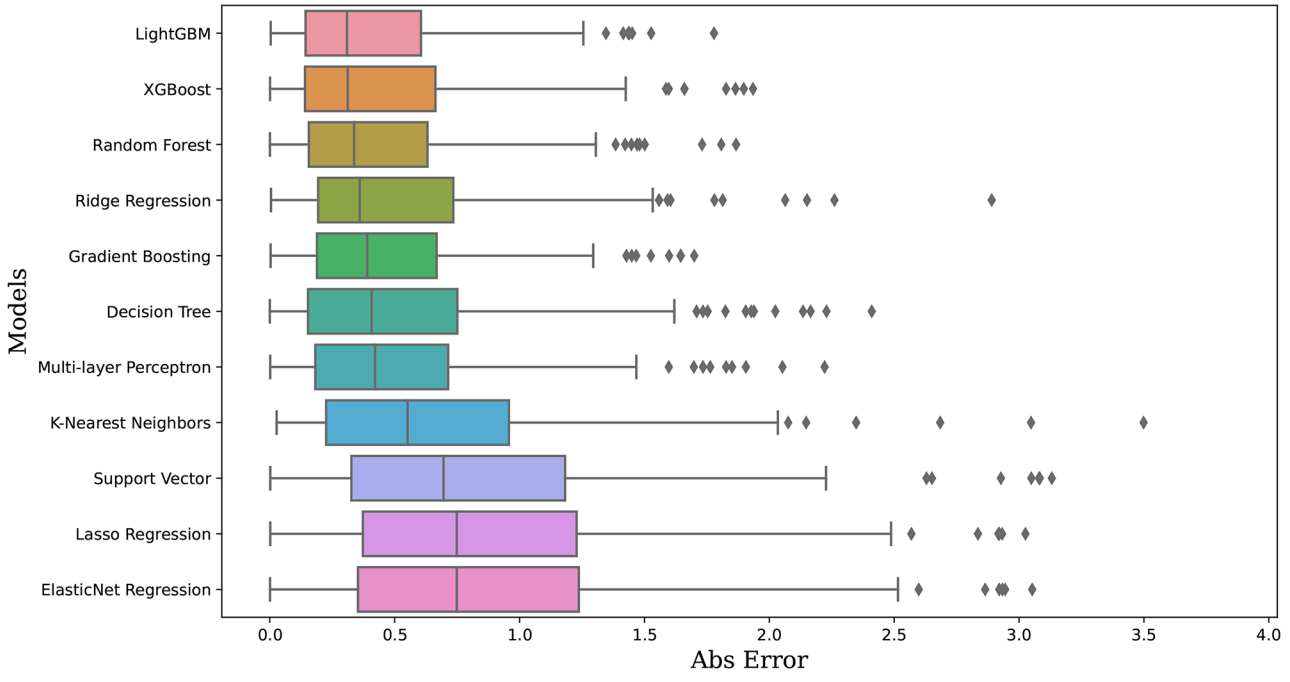

Fig. 3 Residual plot based on the test set

Table 1 MSE and R^2 calculated for each model on the test set

Model name	MSE	R^2
LightGBM	0.298	0.774
Random Forest	0.323	0.755
XGBoost	0.370	0.719
Gradient Boosting	0.372	0.716
Multi-layer Perceptron	0.436	0.669
Ridge Regression	0.492	0.627
Decision Tree	0.572	0.566
K-Nearest Neighbors	0.786	0.403
Support Vector	1.150	0.128
ElasticNet Regression	1.196	0.093
Lasso Regression	1.197	0.092
Linear Regression	1.84E+17	-1.40E+17

Error) indicates a closer match between the predicted and experimental results. As shown in Table 1, the results show that LightGBM, Random Forest, and XGBoost are the top three models, with MSE values of 0.298, 0.323, and 0.370, and R^2 values of 0.774, 0.755, and 0.719, respectively. Additionally, Gradient Boosting, Multi-layer Perceptron, Ridge Regression, and Decision Tree models also performed well, with MSE values below 0.5 and R^2 values above 0.5. In contrast, the K-Nearest Neighbors, Support Vector, ElasticNet Regression, Lasso Regression, and Linear Regression models showed poor performance in both MSE and R^2. Notably, the Linear Regression model exhibited particularly high error values, indicating that it is unsuitable for modeling NLRP3 inhibitors. Consequently, further analysis will focus on the performance of the LightGBM, Random Forest, and XGBoost models.

Figure 4 shows scatter plots of the predicted versus experimental results for the LightGBM, Random Forest, and XGBoost models on the training and test sets. For all three models, the training set data fits the experimental results well, with points distributed symmetrically along the central diagonal line, indicating no signs of overfitting and confirming that our models are robust and reliable.

In the test set, the predicted results of LightGBM, Random Forest, and XGBoost also show a high degree of correlation with the experimental results, with data points clustered near the central diagonal line. This suggests that the models perform well on unseen molecules. In summary, the LightGBM, Random Forest, and XGBoost models we developed can be effectively used for screening NLRP3 inhibitors.

Machine learning-based screening of a large-scale compound library

Machine learning models offer significant advantages in drug discovery, not only due to their accuracy but also

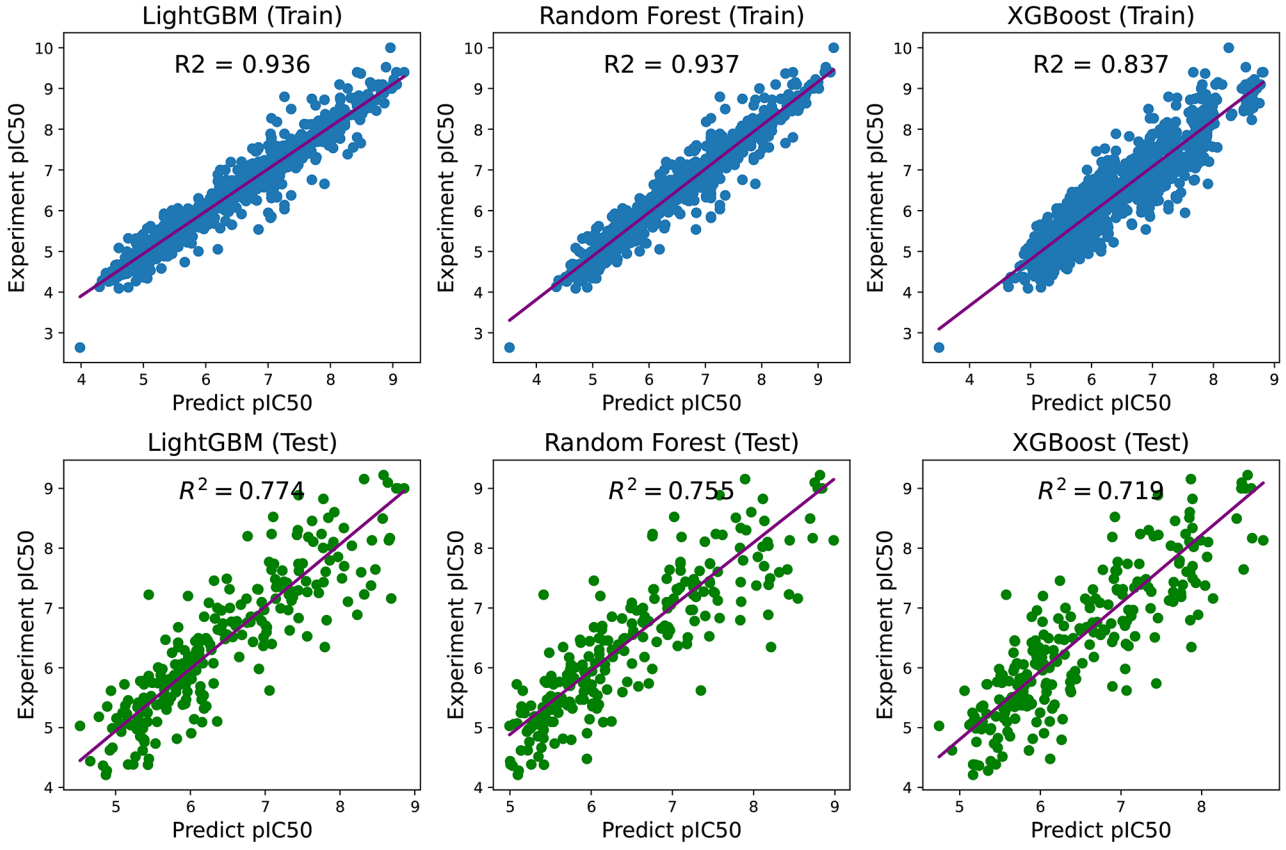

Fig. 4 Scatter plots of the predicted versus experimental results for LightGBM, Random Forest, and XGBoost on the training and test sets

their speed in screening large compound libraries. In this study, we collected a total of 11,526,814 compounds from eight commercially available molecular databases. As shown in Fig. 5, we used LightGBM, Random Forest, and XGBoost models in parallel to screen the dataset, selecting molecules predicted by each model to have a pIC50 greater than 6 (indicating potential IL-1β inhibition with activity stronger than 500 nM). This initial screening yielded 43,276 compounds.

Next, we applied the HTVS, SP, and XP modes of the Glide software to further filter these molecules, resulting in 346 compounds. We then employed the Prime MMGBSA method to calculate the binding free energy of these compounds with NLRP3, selecting those with binding energies better than −50 kcal/mol, narrowing the list to 41 compounds (see Table S1). Finally, we performed ADMET predictions and identified 26 compounds with favorable ADMET properties. The predicted pIC50 values, docking scores, and Prime MMGBSA results for these 26 hits are summarized in Table 2, with the ADMET prediction results available in the Table S2.

As shown in the table above, the predicted activity values and binding energy data for the 26 hits identified through the screening process are presented. Additionally, we included the reference compound MCC950 for comparison. Our calculations show that MCC950's predicted pIC50 values in the LightGBM, XGBoost, and Random Forest models were 8.042, 8.034, and 8.029, respectively. Previous reports indicate that MCC950's actual activity is 28 nM [33] and 8.5 nM [34], corresponding to pIC50 values of 7.55 and 8.070, respectively. Clearly, our predictions closely match the experimental results, further confirming the reliability of the models we developed.

For the 26 selected compounds, the predicted pIC50 values across the three models ranged between 6 and 8.6, suggesting that these molecules exhibit activity superior to 500 nM. Additionally, MCC950's docking score and MMGBSA binding energy were −6.935 kcal/mol and −39.78 kcal/mol, respectively. Negative binding energy values indicate binding potential, with smaller values reflecting stronger binding affinity. In comparison, most

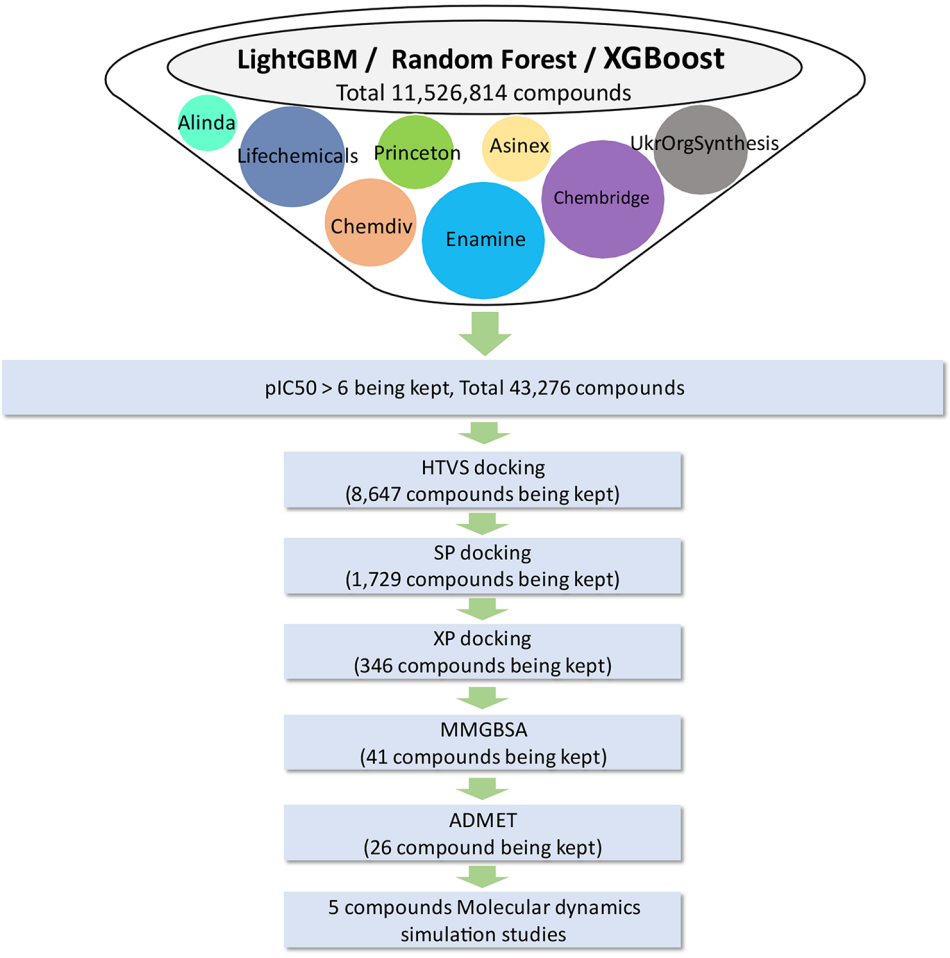

Fig. 5 Virtual screening workflow

Table 2 Hits identified through virtual screening

Name	LightGBM (pIC50)	XGBoost (pIC50)	Random_Forest (pIC50)	docking score (kcal/mol)	MMGBSA dG Bind (kcal/mol)
MCC950[1]	8.042	8.034	8.029	-6.935	-39.78
Chembridge:19,655,631	8.251	8.366	8.232	-8.373	-60.79
Enamine: Z1180203703	8.304	8.515	7.684	-7.505	-57.99
Chembridge:38,214,692	6.296	6.344	7.22	-7.102	-56.82
Enamine: Z4263586645	7.436	7.215	7.802	-7.391	-56.21
Enamine: Z192478440	7.352	6.836	7.647	-7.458	-56.2
Enamine: Z32463764	6.998	6.332	6.282	-6.727	-55.94
UkrOrgSynthesis: PB70887122	7.857	7.722	8.108	-7.306	-55.44
Enamine: Z92441463	7.66	7.532	7.186	-6.806	-55.27
Chemdiv: C594-0115	7.463	7.618	7.802	-7.694	-54.95
Enamine: Z2371440972	7.319	7.531	7.892	-7.469	-54.81
Lifechemicals: F6200-4395	6.432	6.233	6.969	-7.223	-54.65
Chembridge:45,249,460	6.653	6.74	7.567	-7.241	-54.44
IBS: STOCK6S-90,691	6.264	6.004	6.468	-8.521	-53.31
Chembridge:64,720,146	6.272	6.681	7.486	-7.145	-52.21
Enamine: Z3289682378	6.447	6.062	7.197	-7.036	-52.15
Enamine: Z17930893	6.293	6.137	6.045	-6.945	-51.93
Enamine: Z1213669791	6.642	7.449	7.443	-7.077	-51.68
Alinda: IBS-L0209972	6.761	6.654	6.754	-7.953	-51.52
Enamine: Z1942531317	7.487	8.366	7.886	-7.259	-51.52
Chembridge:48,372,456	8.096	7.053	8.135	-7.113	-51.47
Enamine: Z2335263608	6.334	6.257	6.002	-7.483	-51.35
Chembridge:59,927,739	7.52	7.701	6.994	-6.975	-51.2
Chembridge:61,131,069	6.176	6.608	7.125	-8.612	-51.13
Chembridge:56,193,240	6.377	6.26	6.986	-7.825	-50.66
Chembridge:43,388,692	6.392	6.191	7.701	-7.612	-50.52
IBS: STOCK6S-92,717	6.177	6.343	6.594	-7.515	-50.47

of the 26 selected compounds exhibited better scores and binding energies than MCC950, suggesting that they also have an advantage in binding affinity, which is fundamental for exerting biological activity. Notably, compounds 19,655,631 and 38,214,692 from the Chembridge database, and compounds Z1180203703, Z4263586645, and Z192478440 from the Enamine database demonstrated the best activity. As shown in Fig. 6, these compounds feature diverse molecular scaffolds, indicating that they are promising novel NLRP3 inhibitors. We will further investigate their binding modes with NLRP3 and explore their dynamic binding properties.

ADMET prediction

[1] Predicted octanol/water partition co-efficient log p (acceptable range: -2.0 to 6.5).

[2] Predicted aqueous solubility; S in mol/L (acceptable range: -6.5 to 0.5).

[3] Predicted IC50 value for blockage of HERG K+ channels (concern below −7).

[4] Predicted Caco-2 cell permeability in nm/s (acceptable range: <25 is poor and >500 is great).

[5] Predicted brain/blood partition coefficient (acceptable range: -3.0 to 1.2).

[6] Predicted Human Oral Absorption: (acceptable range: >75 is great).

[7] Calculated value of RuleOfFive: (acceptable range: ≤1 is great).

In addition to possessing binding affinity for the target, the favorable ADMET properties of small chemical molecules contribute to their pharmacological efficacy in vivo. We conducted ADMET calculations, and the results are shown in Table 3. The QPlogPo/w values indicate that our five hit molecules exhibit good oil-water distribution characteristics and solubility, while QPlogHERG suggests that these molecules do not pose cardiotoxicity risks. The QPPCaco values demonstrate the good permeability of these molecules across cell membranes. The QPlogBB values range between −3.0 and 1.2, suggesting favorable distribution properties. Furthermore, the prediction results of Human Oral Absorption indicate that these molecules possess good absorption profiles in vivo. None of the molecules violate the "Rule of Five," suggesting good drug-likeness.

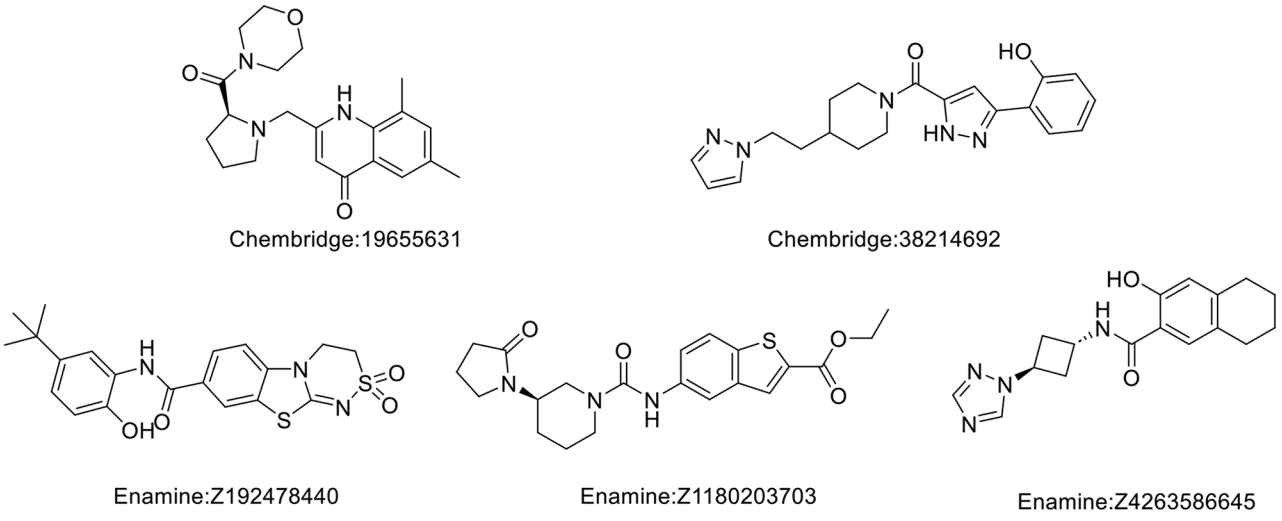

Fig. 6 Structures of compounds 19,655,631, 38,214,692, Z1180203703, Z4263586645, Z192478440

Table 3 ADMET prediction for hits

Compound ID	QPlogPo/w[1]	QPlogS[2]	QPlogHERG[3]	QPPCaco[4]	QPlogBB[5]	HumanOralAbsorption[6]	RuleOfFive
mcc950	3.161	-5.603	-3.743	218.84	-1.398	87.339	0
Chembridge-19,655,631	1.2	-2.116	-3.887	262.487	-0.002	77.27	0
Enamine-Z1180203703	2.663	-4.788	-2.957	196.638	-0.941	83.589	0
Chembridge-38,214,692	3.364	-5.164	-6.095	579.141	-1.045	96.091	0
Enamine-Z4263586645	2.897	-4.894	-5.207	557.578	-0.917	93.061	0
Enamine-Z192478440	2.611	-5.445	-5.617	135.354	-1.602	80.381	0

Binding mode analysis

In this study, we docked these five most promising molecules to NLRP3, with their binding modes shown in Fig. 7A. It can be observed that all these molecules bind consistently between the NBD, HD1, WHD, and HD2 domains. This binding site is consistent with the one revealed by Dekker et al. [18]. using cryo-electron microscopy to elucidate the interaction between the small molecule MCC950 and the NLRP3 protein. This suggests that the selected molecules exert their inhibitory effect on NLRP3 by stabilizing the inactive conformation of the NBD, HD1, WHD, and HD2 domains.

As shown in Fig. 7B, Chembridge:19,655,631 forms hydrogen bonds with A228 and cation-π interactions with R578. Additionally, the molecule's dimethyl group forms hydrophobic interactions with V414, I411, M408, and M661 within the protein. The binding mode of Chembridge:38,214,692 with the protein is depicted in Fig. 7C, showing hydrogen bonding with A228, R578, and S626, along with cation-π interactions with R578. Its benzene ring also forms hydrophobic contacts with V414 and I411. Figure 7D illustrates the binding of Enamine: Z192478440 with hydrogen bonds formed with A228 and S626, and π-π interactions with R578. The tert-butyl group of the molecule fits into a hydrophobic pocket formed by M408, V414, and I411. As shown in Fig. 7E, Enamine: Z1180203703 forms hydrogen bonds with R578, Q624, and S626 in the NLRP3 protein. Similarly, in Fig. 7F, Enamine: Z4263586645 binds through hydrogen bonds with A228, Q624, and S626, while its benzyl group interacts hydrophobically with M408, V414, and I411. These interactions form the basis for the stable binding between the small molecules and the protein, providing valuable insights for subsequent structure-based molecular modifications.

Molecular dynamics simulation analysis

The RMSD in molecular dynamics simulations reflects the mobility of the ligand. Greater RMSD values and fluctuations indicate higher mobility, while smaller and stable RMSD values reflect stable motion. As shown in Fig. 8A, the RMSD of the small molecules reaches stability in the early stages of the simulation, fluctuating within a narrow range of 0–2 Å. This indicates that the small molecules bind tightly to the protein, maintaining high binding stability. Additionally, we analyzed the changes in RoG for the six complexes during the dynamic simulation, as shown in Fig. 8B. RoG represents the compactness of the system, reflecting the tightness of the NLRP3 protein. The calculation results show that the RoG of the six complexes fluctuates stably between 23.5 and 24.25 Å throughout the simulation, suggesting that MCC950,

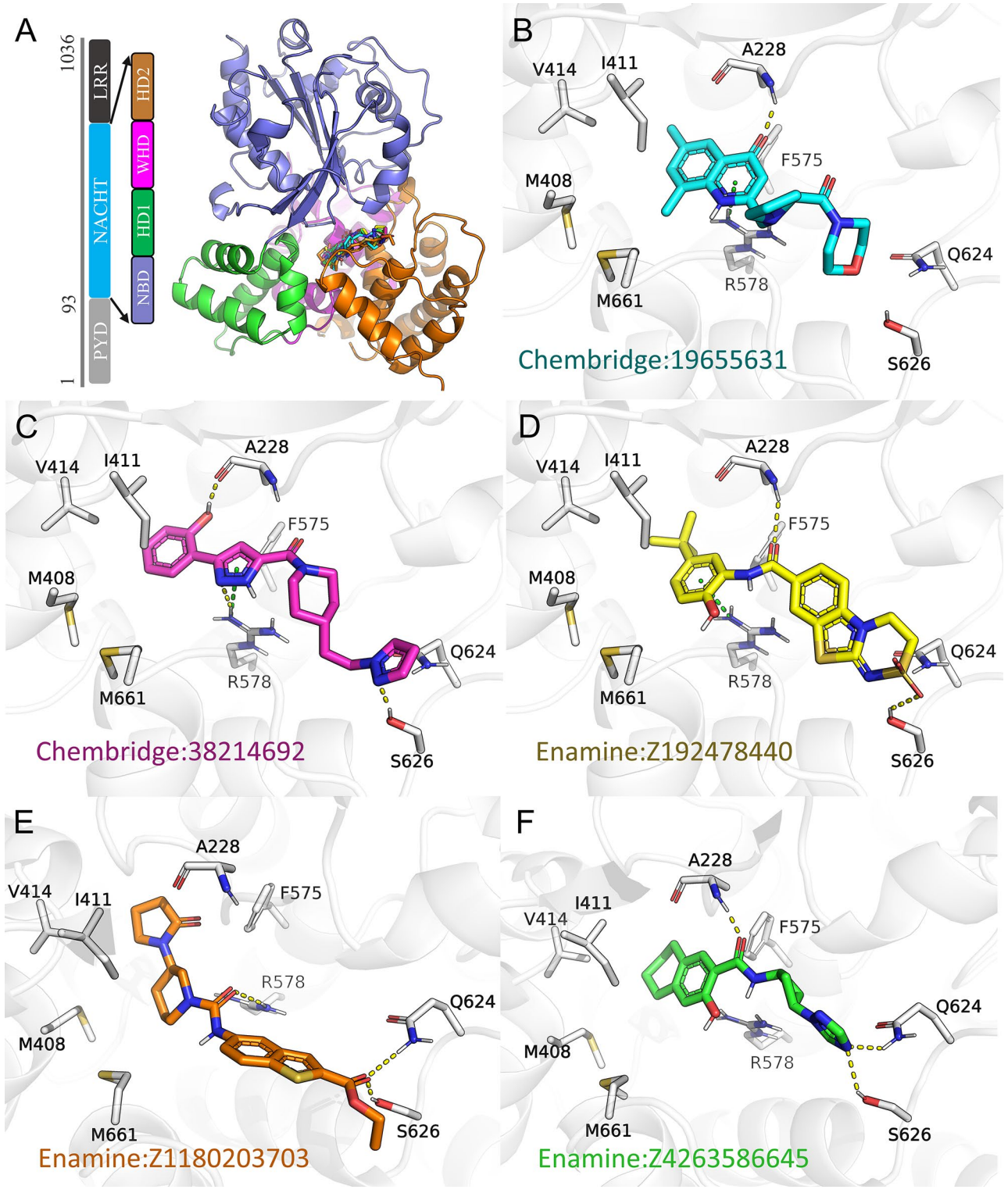

Fig. 7 The binding site information (**A**) and the binding modes of Chembridge:19,655,631 (**B**), Chembridge:38,214,692 (**C**), Enamine: Z192478440 (**D**), Enamine: Z1180203703 (**E**), and Enamine: Z4263586645 (**F**) with the NLRP3 protein. In the figure, yellow dashed lines represent hydrogen bonds, while green dashed lines denote cation-π or π-π interactions, plot by PyMOL 2.5.1 [35]

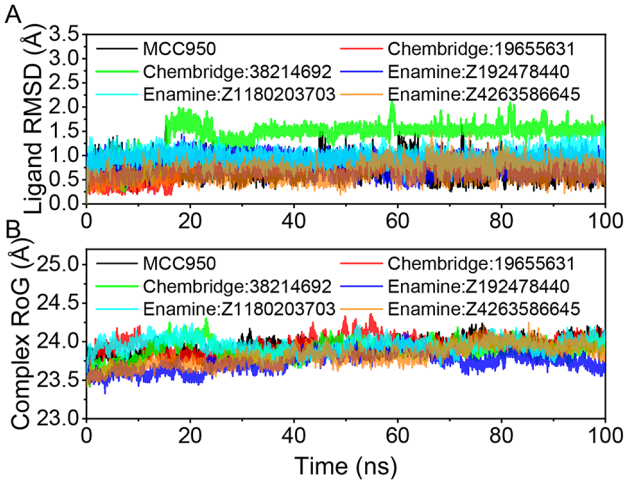

Fig. 8 (**A**) The changes in the root-mean-square deviation (RMSD) of ligands over the simulation time and (**B**) the changes in the radius of gyration (RoG) of the complexes over time

along with 19,655,631, 38,214,692, Z1180203703, Z4263586645, and Z192478440, maintains the compactness of the NACHT domain, thereby stabilizing the inactive state of NLRP3 and exerting its inhibitory effect.

Based on the molecular dynamic simulation trajectories, we calculated the binding energies using the MM-GBSA method, which provides a more accurate assessment of the binding effects between small molecules and the target protein. As shown in the Table 4, the binding energies of MCC950, 19,655,631, 38,214,692, Z192478440, Z1180203703, and Z4263586645 with the protein are −27.93±4.30, -28.88±1.53, -28.15±1.63, -26.55±3.76, -37.91±1.79, and −25.42±1.00 kcal/mol, respectively. The negative values indicate that these molecules exhibit binding affinity to the target protein, with lower values representing stronger binding. Our calculations clearly show that 19,655,631, 38,214,692, and Z1180203703 exhibit better binding effects with NLRP3 than MCC950, underscoring their superior binding performance. Additionally, energy decomposition reveals that van der Waals forces and electrostatic interactions are the primary contributors to the binding, while non-polar solvation energies make weaker contributions.

The above energy calculations reveal that 19,655,631, 38,214,692, and Z1180203703 exhibit optimal dynamic binding effects. Using MM-GBSA energy decomposition, we identified the top 10 key residues contributing to the binding. Moreover, we sampled the binding conformations of the molecules and proteins at 0 ns, 30 ns, 60 ns, and 100 ns from the simulation trajectories to visually observe the conformational stability of the small molecule-protein complexes. As shown in Fig. 9, for 19,655,631 and 38,214,692, the key amino acid is TYR-631, while for Z1180203703, the key amino acids are ARG-577, TYR-631, and LEU-627, all with binding energy contributions less than −2 kcal/mol. No significant conformational changes were observed for these three molecules over the simulation time, indicating that they all stably bind to the active site.

Conclusion

In this study, we collected structural and activity data of NLRP3 inhibitors from literature and patents and employed several machine learning methods, including LightGBM, Random Forest, and XGBoost, to build quantitative structure-activity relationship (QSAR) models. These models showed strong performance in predicting NLRP3 activity, with R^2 values of 0.774, 0.755, and 0.719, respectively.

To fully leverage these models, we applied them to predict NLRP3 activity for a large dataset of 11,526,814 compounds from commercially available databases. For compounds with predicted pIC50 values greater than 6, we further performed molecular docking and MMGBSA calculations, identifying 41 compounds with high binding affinity to NLRP3. However, in addition to binding affinity, compounds must exhibit favorable ADMET properties to be considered potential drugs. Therefore, we conducted ADMET calculations and screened out 26 high-potential NLRP3 inhibitors.

Subsequently, we employed molecular dynamics simulations to explore the binding stability and interaction

Table 4 Binding free energies and energy components predicted by MM/GBSA (kcal/mol)

Ligand name	MCC950	19,655,631	38,214,692	Z192478440	Z1180203703	Z4263586645
ΔE_{vdw}	-45.63±3.05	-50.14±1.32	-43.45±2.77	-50.53±2.60	-49.48±1.99	-45.49±1.14
ΔE_{elec}	-110.37±5.59	-37.02±1.61	-44.44±4.07	-34.86±4.04	-31.50±3.55	-25.78±3.26
ΔG_{GB}	134.46±4.35	64.79±2.52	65.56±4.09	65.12±4.23	49.11±3.38	51.77±2.31
ΔG_{SA}	-6.41±0.17	-6.50±0.09	-5.82±0.12	-6.27±0.12	-6.04±0.17	-5.91±0.00
ΔG_{bind}	-27.93±4.30	-28.88±1.53	-28.15±1.63	-26.55±3.76	-37.91±1.79	-25.42±1.00

ΔE_{vdW}: van der Waals energy

ΔE_{elec}: electrostatic energy

ΔG_{GB}: electrostatic contribution to solvation

ΔG_{SA}: non-polar contribution to solvation

ΔG_{bind}: binding free energy

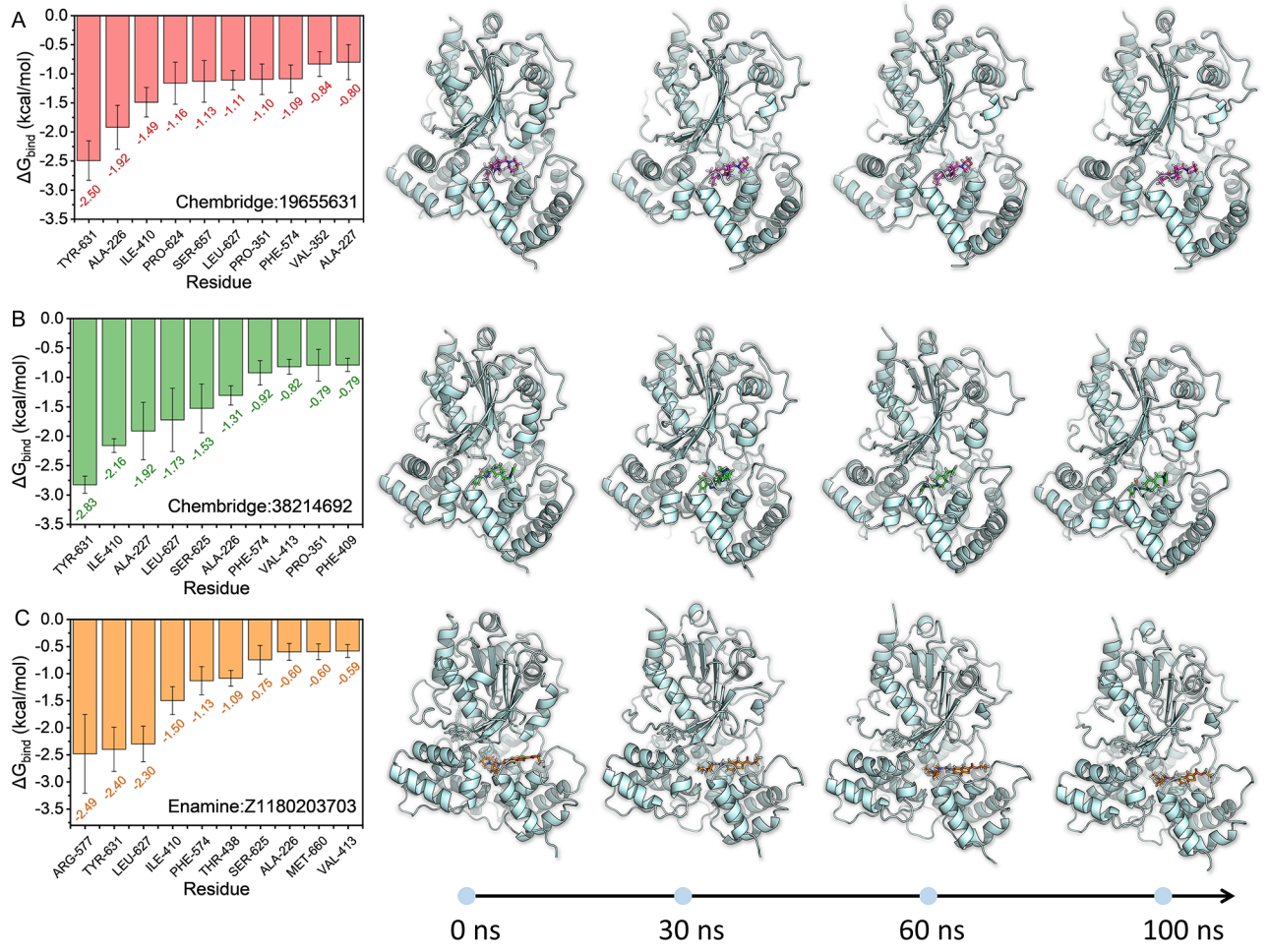

Fig. 9 illustrates the top 10 residues contributing to small molecule-protein binding, as well as the changes in drug binding conformations over simulation time

details of these potential NLRP3 inhibitors with the protein. The results revealed that compounds 19,655,631, 38,214,692, and Z1180203703 exhibited comparable binding stability to MCC950 and demonstrated stronger binding energies. Notably, the key amino acids in the protein binding pocket—TYR-631, ARG-577, and LEU-627—played a pivotal role in inhibitor binding, offering valuable insights for further structure-based development of NLRP3 inhibitors.

Our study demonstrates that machine learning-based QSAR models are highly efficient for screening compound libraries of millions of molecules, underscoring the value of building accurate models. Constructing effective QSAR models requires large, high-quality datasets of biochemical data. However, in the past, we only obtained a small amount of data sets from literature for modeling. In this research, in addition to sourcing data from literature (such as ChemBL), we also acquired data from patents, resulting in a model with superior performance.

Looking ahead, we aim to further expand our dataset by collaborating with commercial data sources, which will enhance the generalizability and predictive accuracy of our models. Additionally, improving the interpretability of machine learning models will be crucial for future drug design efforts, as it will aid in understanding the relationship between compound structure and activity and guide the design of new molecules.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13065-024-01323-y.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
TJ and SJQ contributed to Conceptualization, methodology, investigation, visualization and writing. JHX contributed to supervision, and Writing – review & editing. SHY and YL contributed to resources. XSY and LSX contributed to

project, resources, and supervision, Funding acquisition.All authors reviewed the manuscript and agreed to publish it.

Funding
This work was supported by the Academic sponsorship project for top talents in university disciplines (majors), (Grant No. gxbjZD2022105).

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

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Received: 11 September 2024 / Accepted: 16 October 2024
Published online: 28 October 2024

References

1. Gupta L, Ahmed S, Singh B, Prakash S, Phadke S, Aggarwal A. Novel NLRP12 variant presenting with familial cold autoimmunity syndrome phenotype. Ann Rheum Dis. 2021;80:e117. https://doi.org/10.1136/annrheumdis-2019-216158.
2. Riaz M, Rehman AU, Shah SA, Rafiq H, Lu S, Qiu Y, Wadood A. Predicting Multi-interfacial binding mechanisms of NLRP3 and ASC Pyrin Domains in Inflammasome activation. ACS Chem Neurosci. 2021;12:603–12. https://doi.org/10.1021/acschemneuro.0c00519.
3. Cassel SL, Sutterwala FS. Sterile inflammatory responses mediated by the NLRP3 inflammasome. Eur J Immunol. 2010;40:607–11. https://doi.org/10.1002/eji.200940207.
4. Palumbo L, Carinci M, Guarino A, Asth L, Zucchini S, Missiroli S, Rimessi A, Pinton P, Giorgi C. The NLRP3 inflammasome in neurodegenerative disorders: insights from epileptic models. Biomedicines. 2023;11. https://doi.org/10.3390/biomedicines11102825.
5. Akbal A, Dernst A, Lovotti M, Mangan MSJ, McManus RM, Latz E. How location and cellular signaling combine to activate the NLRP3 inflammasome. Cell Mol Immunol. 2022;19:1201–14. https://doi.org/10.1038/s41423-022-00922-w.
6. Mackowiak B, Fu Y, Maccioni L, Gao B. Alcohol-associated liver disease. J Clin Invest. 2024;134. https://doi.org/10.1172/jci176345.
7. Brahadeeswaran S, Dasgupta T, Manickam V, Saraswathi V, Tamizhselvi R. NLRP3: a new therapeutic target in alcoholic liver disease. Front Immunol. 2023;14:1215333. https://doi.org/10.3389/fimmu.2023.1215333.
8. Harjumäki R, Pridgeon CS, Ingelman-Sundberg M. CYP2E1 in alcoholic and non-alcoholic Liver Injury. Roles of ROS, reactive intermediates and lipid overload. Int J Mol Sci. 2021;22. https://doi.org/10.3390/ijms22158221.
9. McVicker BL, Tuma PL, Kharbanda KK, Lee SM, Tuma DJ. Relationship between oxidative stress and hepatic glutathione levels in ethanol-mediated apoptosis of polarized hepatic cells. World J Gastroenterol. 2009;15:2609–16. https://doi.org/10.3748/wjg.15.2609.
10. Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on toll-like receptors. Nat Immunol. 2010;11:373–84. https://doi.org/10.1038/ni.1863.
11. Liu J, Ren F, Cheng Q, Bai L, Shen X, Gao F, Busuttil RW, Kupiec-Weglinski JW, Zhai Y. Endoplasmic reticulum stress modulates liver inflammatory immune response in the pathogenesis of liver ischemia and reperfusion injury. Transplantation. 2012;94:211–7. https://doi.org/10.1097/TP.0b013e318259d38e.
12. Valles SL, Blanco AM, Azorin I, Guasch R, Pascual M, Gomez-Lechon MJ, Renau-Piqueras J, Guerri C. Chronic ethanol consumption enhances interleukin-1-mediated signal transduction in rat liver and in cultured hepatocytes. Alcohol Clin Exp Res. 2003;27:1979–86. https://doi.org/10.1097/01.Alc.0000099261.87880.21.
13. Petrasek J, Bala S, Csak T, Lippai D, Kodys K, Menashy V, Barrieau M, Min SY, Kurt-Jones EA, Szabo G. IL-1 receptor antagonist ameliorates inflammasome-dependent alcoholic steatohepatitis in mice. J Clin Invest. 2012;122:3476–89. https://doi.org/10.1172/jci60777.
14. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021;25:1315–60. https://doi.org/10.1007/s11030-021-10217-3.
15. Priya S, Tripathi G, Singh DB, Jain P, Kumar A. Machine learning approaches and their applications in drug discovery and design. Chem Biol Drug Des. 2022;100:136–53.
16. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40:D1100–7. https://doi.org/10.1093/nar/gkr777.
17. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des. 2007;21:681–91.
18. Dekker C, Mattes H, Wright M, Boettcher A, Hinniger A, Hughes N, Kapps-Fouthier S, Eder J, Erbel P, Stiefl N, Mackay A, Farady CJ. Crystal structure of NLRP3 NACHT Domain with an inhibitor defines mechanism of Inflammasome Inhibition. J Mol Biol. 2021;433:167309. https://doi.org/10.1016/j.jmb.2021.167309.
19. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. Wiley Interdisciplinary Reviews: Comput Mol Sci. 2013;3:198–210. https://doi.org/10.1002/wcms.1121.
20. Case DA, Aktulga HM, Belfon K, Cerutti DS, Cisneros GA, Cruzeiro VWD, Forouzesh N, Giese TJ, Götz AW, Gohlke H, Izadi S, Kasavajhala K, Kaymak MC, King E, Kurtzman T, Lee TS, Li P, Liu J, Luchko T, Luo R, Manathunga M, Machado MR, Nguyen HM, O'Hearn KA, Onufriev AV, Pan F, Pantano S, Qi R, Rahnamoun A, Risheh A, Schott-Verdugo S, Shajan A, Swails J, Wang J, Wei H, Wu X, Wu Y, Zhang S, Zhao S, Zhu Q, Cheatham TE 3rd, Roe DR, Roitberg A, Simmerling C, York DM, Nagan MC, Merz KM Jr. AmberTools. J Chem Inf Model. 2023;63:6183–91. https://doi.org/10.1021/acs.jcim.3c01153.
21. Wang J, Wang W, Kollman PA, Case DA. Antechamber: an accessory software package for molecular mechanical calculations. J Am Chem Soc. 2001;222:U403.
22. He X, Man VH, Yang W, Lee T-S, Wang J. (2020) A fast and high-quality charge model for the next generation general AMBER force field. J Chem Phys 153.
23. Tian C, Kasavajhala K, Belfon KA, Raguette L, Huang H, Migues AN, Bickel J, Wang Y, Pincay J, Wu Q. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. J Chem Theory Comput. 2019;16:528–52.
24. Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A. 2001;105:9954–60. https://doi.org/10.1021/jp003020w.
25. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995;103:8577–93.
26. Kräutler V, Van Gunsteren WF, Hünenberger PH. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. J Comput Chem. 2001;22:501–8. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.
27. Larini L, Mannella R, Leporini D. Langevin stabilization of molecular-dynamics simulations of polymers by means of quasisymplectic algorithms. J Chem Phys. 2007;126:104101. https://doi.org/10.1063/1.2464095.
28. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model. 2010;51:69–82.
29. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10:449–61.
30. Nguyen H, Roe DR, Simmerling C. Improved generalized born Solvent Model parameters for protein simulations. J Chem Theory Comput. 2013;9:2020–34. https://doi.org/10.1021/ct3010485.
31. Weiser J, Shenkin PS, Still WC. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Comput Chem. 1999;20:217–30.
32. Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, Wei Z, Lei M. Machine learning in materials science. InfoMat. 2019;1:338–58.
33. Harrison D, Boutard N, Brzozka K, Bugaj M, Chmielewski S, Cierpich A, Doedens JR, Fabritius CRY, Gabel CA, Galezowski M, Kowalczyk P, Levenets O, Mroczkowska M, Palica K, Porter RA, Schultz D, Sowinska M, Topolnicki G, Urbanski P, Woyciechowski J, Watt AP. Discovery of a series of

ester-substituted NLRP3 inflammasome inhibitors. Bioorg Med Chem Lett. 2020;30:127560. https://doi.org/10.1016/j.bmcl.2020.127560.

34. Coll RC, Robertson AA, Chae JJ, Higgins SC, Muñoz-Planillo R, Inserra MC, Vetter I, Dungan LS, Monks BG, Stutz A, Croker DE, Butler MS, Haneklaus M, Sutton CE, Núñez G, Latz E, Kastner DL, Mills KH, Masters SL, Schroder K, Cooper MA, O'Neill LA. A small-molecule inhibitor of the NLRP3 inflammasome for the treatment of inflammatory diseases. Nat Med. 2015;21:248–55. https://doi.org/10.1038/nm.3806.

35. DeLano WL. Pymol: an open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr. 2002;40:82–92.

Publisher's note