DOI：10.3969/j.issn.2096-8248.2026.01.010
引用格式：黄池，魏荣斌，马少杰 . 整合机器学习与分子动力学模拟挖掘抗癌靶点 MELK 抑制剂［J］. 江苏海洋大学学报（自然科学版），2026，35（1）：83-94.

整合机器学习与分子动力学模拟挖掘抗癌靶点 MELK 抑制剂

黄　池[a]，魏荣斌[a, b]，马少杰[a, b]

（江苏海洋大学 a. 药学院；b. 江苏省海洋药物活性分子筛选重点实验室，江苏 连云港　222005）

摘　要：母系胚胎亮氨酸拉链激酶（maternal embryonic leucine zipper kinase，MELK）因其在多种癌症的发生发展中发挥关键作用，已成为一个备受关注的抗肿瘤药物靶点。为高效挖掘 MELK 抑制剂，结合机器学习方法筛选大规模化合物库，力求发现高活性候选分子。首先，基于已知的 MELK 抑制剂数据构建并优化机器学习模型，其中随机森林回归为最佳模型（R^2=0.800 4，RMSE=0.496 8）。随后基于该模型，从筛选出的预测得分最高的 5 个候选化合物出发，进一步开展分子对接及分子动力学模拟研究，结果显示这些化合物在活性口袋中具有稳定的构象和良好的结合自由能，显示出潜在的 MELK 抑制活性。最后利用 SwissADME 对化合物进行全面的 ADME 性质预测。该研究验证了机器学习辅助虚拟筛选方法在 MELK 抑制剂发现中的有效性，为新型 MELK 抑制剂的设计和开发提供了重要候选分子及理论基础。

关键词：MELK；机器学习；激酶抑制剂；分子对接
中图分类号：R914.2　　**文献标志码：**A　　**文章编号：**2096-8248（2026）01-0083-12

Integrating machine learning with molecular dynamics simulations to mine inhibitors of the anticancer target MELK

HUANG Chi[a], WEI Rongbin[a, b], MA Shaojie[a, b]

（a. School of Pharmacy; b. Jiangsu Key Laboratory of Marine Drug Active Molecular Screening, Jiangsu Ocean University, Lianyungang 222005, China）

Abstract: Maternal embryonic leucine zipper kinase（MELK）has become a focus of anti-tumor drugs because it plays a key role in the occurrence and development of many cancers. In order to mine MELK inhibitors efficiently, this study combined with machine learning method to screen large-scale compound libraries and tried to find high-activity candidate molecules. Firstly, the machine learning model was constructed and optimized based on the known MELK inhibitor data, in which Random Forest Regression is the best model（R^2=0.800 4, RMSE=0.496 8）. Then, based on this model, five candidate compounds were selected with the highest prediction scores, and further carried out molecular docking and molecular dynamics simulation research. The research shows that these compounds have stable conformation and good binding free energy in the active pocket, showing potential MELK inhibition activity. Finally,

收稿日期：2025-08-31；**修订日期：**2025-10-29
基金项目：江苏海洋大学江苏省海洋药物活性分子筛选重点实验室开放基金资助项目（HY202302）
作者简介：黄池，硕士研究生，研究方向为计算机辅助药物设计，（E-mail）1186988083@qq.com。
通信作者：马少杰，副教授，博士，研究方向为计算机辅助技术和人工智能的药物研发，（E-mail）mashaojie@jou.edu.cn。

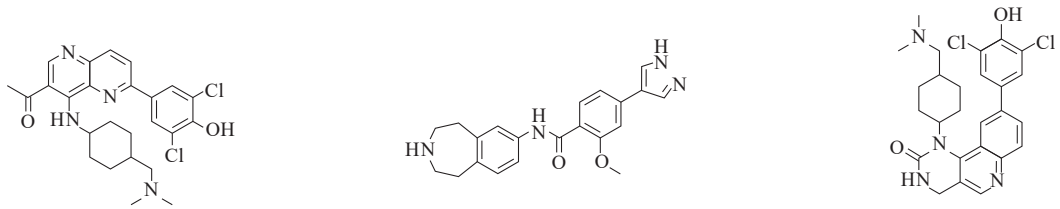
SwissADME was used to predict the ADME properties of the compounds. This study verified the effectiveness of virtual screening method assisted by machine learning in the discovery of MELK inhibitors, and provided important candidate molecules and theoretical basis for the design and development of new MELK inhibitors.

Key words: MELK; machine learning; kinase inhibitor; molecular docking

0 引言

癌症是全球范围内导致人类死亡的主要原因之一，严重威胁着人类的生命健康与生存质量。蛋白激酶作为细胞信号转导网络中的关键调控节点，其异常活化或表达与多种癌症的发生发展密切相关，已成为抗肿瘤药物研发的热点靶点。母系胚胎亮氨酸拉链激酶（maternal embryonic leucine zipper kinase，MELK）是一种细胞周期依赖性蛋白激酶，隶属于蔗糖非发酵-1/amp活化蛋白激酶（Snf1/AMPK）家族，在细胞内信号转导、细胞周期、细胞增殖、细胞凋亡、转录后修饰、胚胎发育等方面发挥着重要作用[1-2]。大量研究表明，MELK在多种恶性肿瘤中呈现异常高表达状态，如胶质母细胞瘤、乳腺癌、肺癌和卵巢癌等，并且其高表达水平往往与肿瘤的侵袭性增强、患者预后不良以及化疗抵抗相关。因此，靶向MELK被认为是开发新型抗癌疗法的一个颇具前景的策略。

目前已有研究报道了一些MELK抑制剂（例如OTSSP167，MELK-T1，HTH-01-091，见图1），但它们可能存在选择性不足、药代动力学特性不佳或易产生耐药性等问题[3-5]。此外，当前的MELK抑制剂在化学结构多样性方面仍有提升空间，急需发现具有新颖骨架和更优成药性的候选药物。



a OTSSP167　　b MELK-T1　　c HTH-01-091

图1　已报道的MELK抑制剂结构

Fig. 1　Structure of reported MELK inhibitors

传统的药物发现流程，尤其是高通量筛选（HTS），虽然是发现先导化合物的重要手段，但其成本高昂、周期漫长且筛选效率有限。近年来，以计算机辅助药物设计（computer-aided drug design，CADD）为代表的计算方法，在加速药物研发进程、降低研发成本方面展现出巨大潜力[6]。其中，机器学习（machine learning，ML）技术凭借其从大规模数据中学习复杂模式并进行准确预测的能力，已成功应用于虚拟筛选，且显著提高了苗头化合物的发现效率和命中率[7-8]。其中之一是定量构效关系（quantitative structure-activity relationships，QSAR）分子建模，这是一种用于药物发现的技术，旨在建立分子化学结构与其生物活性之间的数学关系[9-10]。QSAR使用一系列分子描述符，量化分子的化学结构，然后采用统计技术来模拟这些描述符与生物活性之间的关系，通过构建基于已知活性化合物的机器学习模型，可以对海量化合物数据库进行快速、低成本的活性预测[11-12]。本研究利用QSAR的计算逻辑，以机器学习（ML）为工具，使用数据库中的化合物进行计算模型的构建与训练，随后为了进一步精确评估潜在候选分子的结合能力和作用机制，利用分子对接（molecular docking）技术用于预测配体与靶蛋白的结合模式和亲和力，分子动力学（molecular dynamics，MD）模拟从原子层面揭示配体—蛋白复合物在生理条件下的动态行为、结合稳定性以及关键相互作用，为筛选结果提供更可靠的理论支持。如Das等[13]采用计算机对接，通过使用Glide（薛定谔公司）的高通量虚拟筛选来

鉴定针对 MELK 的小分子。Pasala 等[14]最近的一项研究利用 CADD 预测幽门螺杆菌菌株的潜在新靶点。Cobre 等[15]在 ZINC-22 database 中通过机器学习模型预测，确定了 124 种类似于马拉韦洛克的新型抗 hiv 候选药物。

鉴于此，本研究利用了一种综合策略，将基于 QSAR 的机器学习模型、分子对接和分子动力学模拟相结合，从 Enamine 化学数据库中筛选潜在的 MELK 抑制剂。本研究期望能发现具有新颖化学结构和良好成药前景的 MELK 抑制剂，为后续的实验验证和新型抗癌药物的开发提供有价值的先导化合物。

1　材料和方法

1.1　数据收集和预处理

数据集中的所有用于模型训练的活性抑制剂分子均来源于 ChEMBL 数据库（https：//www.ebi.ac.uk/chembl/），从 ChEMBL 中共检索到 959 组对 MELK 具有抑制作用的活性分子数据条目。为了进一步提高数据集的质量，对这些数据进行以下处理：① 只有来自人类的分子具有可用的半抑制浓度 IC_{50} 值，将 IC_{50} 转换为 pIC_{50}［$pIC_{50}=-\log(IC_{50})$］。② 去除重复分子以及含有空白数据的分子。③ 分子相似性阈值设为 0.9，如果两个分子的相似性得分达到或超过 0.9，则认为是近乎相同的变体或结构类似物；生物活性阈值设为 2，筛选出活性值 $pIC_{50}\geqslant 2$ 的分子，排除低活性或无活性的数据点[16]。④ 根据 Simeon 等[17]的研究进行 IC_{50} 值的分类，并将生物活性分为 3 组：活性化合物，IC_{50}<100 nmol/L；具有中间活性的化合物，IC_{50} 在 100~1 000 nmol/L 之间；非活性化合物，IC_{50}>1 000 nmol/L。经过这些处理，共剩余 717 个活性分子。从 Enamine 化学数据库（https：//enamine.net/）下载激酶数据库，该数据库专为发现新型蛋白激酶抑制剂而设计，共计 64 960 种化合物，在该研究中作为筛选 MELK 抑制剂的数据库。随后，使用指纹描述符（PubChem Fingerprint）对选定的数据库分子进行表征[18]。

1.2　机器学习

建立回归模型，旨在根据生物活性化合物的化学结构（以 PubChem 描述符的形式表示）预测生物活性（以 pIC_{50} 的形式表示）。在本分析中，响应变量是 pIC_{50}，预测变量是 PubChem 指纹描述符。

机器学习算法的选择分两个阶段进行。首先，将数据集随机分为 80% 的训练集和 20% 的测试集，使用 Python 中的 Lazy Predict 库对各种算法进行筛选，使用决定系数（R^2）和均方根误差（RMSE）作为选择标准[19]。随后，使用 Python Sci-kit-Learn 库训练和测试识别出的最佳算法，并且进行超参数网格优化，用最终训练好的机器学习模型来预测 MELK 抑制剂[20]。

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i-y_i)^2}, \quad (1)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}。 \quad (2)$$

式中：$\hat{y}_i$ 和 y_i 分别为第 i 个样本的预测值和真实值，$\bar{y}_i$ 为测试集中样本的平均值。

下载 Enamine 化学数据库中的激酶数据库，并为每个分子计算 881 个 PubChem 子结构指纹。这些指纹编码了每种化合物是否含有某种亚结构，将指纹提交给训练好的机器学习模型，获得每种输出的预测 pIC_{50}，根据预测的 pIC_{50} 评分对化合物进行排序。

1.3　分子对接

分子对接是预测配体—蛋白质结合相互作用的常用方法，在识别活性分子方面越来越重要；也是评估药物—大分子相互作用的重要方法，由于其能够提高研究效率并降低实验室成本而被广泛应用。本研究进行分子对接模拟，以确定从 QSAR 模型中获得的化合物和 MELK 晶体结构的结合位点。

根据机器学习模型预测出的 MELK 抑制剂排序，从化合物数据库（https：//pubchem.ncbi.nlm.nih.gov/）下载排序靠前的 5 个分子。从蛋白质数据库（http：//www.rcsb.org/）中检索 MELK 的晶体结构（PDB ID：4IXP）。用 ChemDraw 3D 软件将分子结构进行能量最小化，并转换成 Mol 2 格式，用 PyMOL 3.2 对关键蛋白进行脱水、氢化等预处理，用 AutoDockTools-1.5.7 进行分子对接[21]，以 x：−48.256，y：47.906，z：22.837 为中心，通过创建尺寸为 2.25 nm×2.25 nm×2.25 nm 的网格框来定义结合位点。使用 AutoDock Vina 生成 10 个对接结果，以确保结合位点的彻底探索。每个生成结果的结合亲和力由 AutoDock Vina 的评分函数进行估计。随后将生成的 pdbqt 文件导入 PyMOL 软件转换为 PDB 格式，并使用 PyMOL 3.2 可视化具有最佳结合能的蛋白质—配体复合物结构。

1.4　分子动力学模拟

用 Amber 24[22]软件对母系胚胎亮氨酸拉链激

酶（MELK）复合物进行 200 ns 分子动力学（MD）模拟。蛋白质—配体复合物的初始构型取自上一步分子对接的结果。用 Amber Tools 中的 tleap 模块制备该系统，对蛋白质使用 ff14SB 力场，对配体使用 GAFF 力场。蛋白质—配体复合物被 Cl^- 离子中和，并在截断的八面体 TIP3P 水盒中明确溶解，在溶质和盒子边缘之间保持 1.2 nm 的最小缓冲距离。然后，通过能量最小化来松弛系统的结构并保证系统的几何形状。在 300 K 和 1 bar 的恒定温度下的 NPT 系综中，对蛋白质与配体的复合物进行 200 ns 的分子动力学模拟。

1.5 结合自由能计算 MM-GBSA

为了评估筛选出小分子与 MELK 之间的结合亲和力，并阐明关键相互作用残基，提取分子动力学模拟最后 50 ns 的轨迹，使用 Amber 24 软件中的 MM-GBSA 程序计算蛋白质—配体的结合自由能[23]。计算中使用了与分子动力学模拟相同的力场（Amber ff14SB）所定义的原子半径和电荷，

$$\Delta G_{bind}=G_{complex}-(G_{protein}+G_{ligand})。\quad (3)$$

式中：ΔG_{bind} 为结合自由能的变化，$G_{complex}$ 为蛋白质—配体复合物的总自由能，$G_{protein}$ 为未结合蛋白质的自由能，G_{ligand} 为孤立配体的自由能。

为了进一步探究单个残基对结合自由能的贡献，执行每残基能量分解。该方法将总的有效结合自由能分解为每个残基的贡献，包括其分子力学相互作用能（范德华能和静电能）和溶剂化自由能的贡献。分解计算同样基于上述的 MM-GBSA 方案。详细的分解结果有助于识别对 MELK 复合物相互作用起关键作用的关键残基。

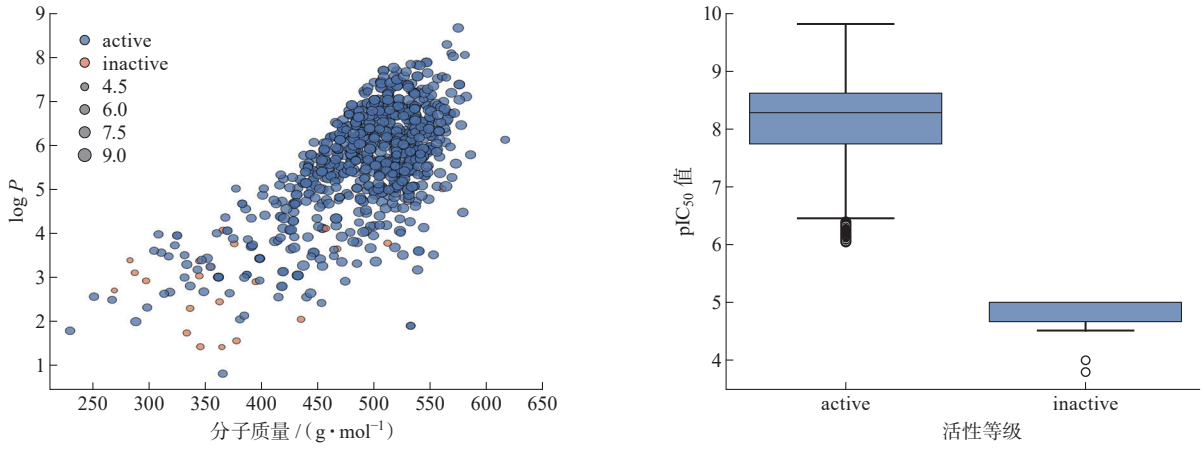
1.6 ADME 性质预测

为了评估从虚拟筛选中脱颖而出的前 5 个候选化合物的药代动力学特性和类药性，使用 SwissADME 在线服务器（http: //www.swissadme.ch/）[24]，将每个候选化合物的 SMILES 格式输入到服务器中进行计算。主要关注以下几个方面的性质：① 药物理化性质：包括分子质量（molecular weight, MW），计算得出的共识对数分配系数（consensus log *P*）、拓扑极性表面积（TPSA）、氢键供体（HBD）和氢键受体（HBA）的数量。② 药代动力学：重点评估胃肠道（GI）吸收的预测结果和血脑屏障（BBB）的渗透能力。③ 类药性：依据经典的 Lipinski 五规则（Lipinski's Rule of Five）[25]对化合物进行评估，筛选标准为违反的条目数不多于 1 条。④ 药物化学友好性：检查化合物是否含有 PAINS（pan-assay interference compounds）和 Brenk 结构警报，并排除任何含有此类警报基团的分子[26]。

2 结果

2.1 数据的预处理和描述

在数据处理阶段，首先对原始化合物数据进行筛选和整理。筛选标准包括去除缺失、重复及异常值，并计算每个化合物的分子质量（MW）、脂水分配系数（log *P*）以及半抑制浓度的对数（pIC_{50}），根据 pIC_{50} 的数值，将化合物分为 “active”（高活性）和 “inactive”（低活性）两类。为了直观展示数据的分布情况，使用散点图和箱线图，分别展示分子质量和 log *P* 与生物活性之间的关系，如图 2 所示。



a 基于分子质量和 log *P* 的化学空间分析　　b 高活性与低活性化合物的 pIC_{50} 值分布箱线图

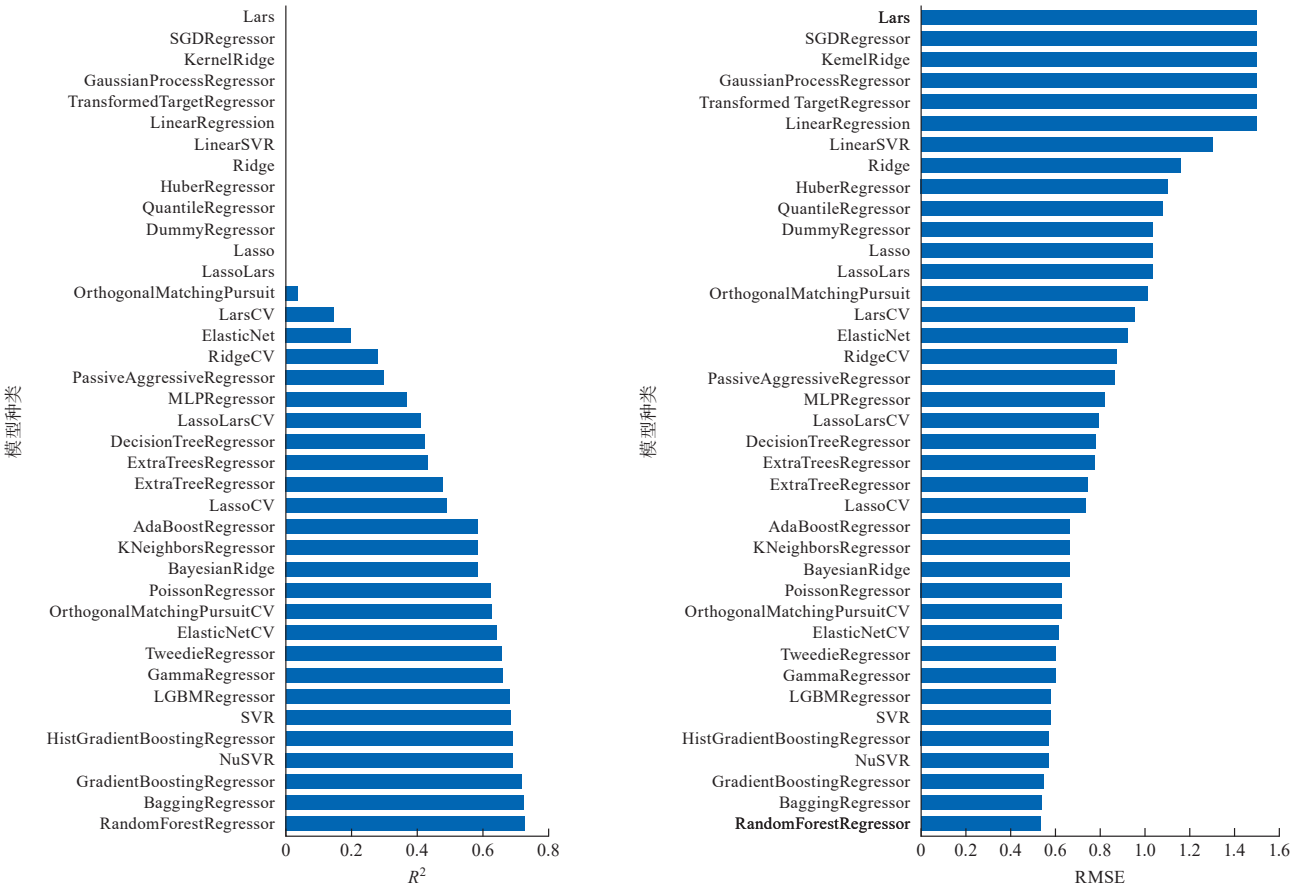
图 2 收集数据集中化合物的分析

Fig. 2 Analysis of compounds in the collection data set

从图 2a 可见，active（高活性）组与 inactive（低活性）组在分子质量及 log *P* 分布上存在一定的区分。大部分 active 化合物集中分布在较高分子量（450~600 g/mol）及较高 log *P*（5~8）区域，呈现一定的正相关趋势。同时，点的大小反映 pIC_{50} 值，显示高活性化合物主要聚集在分子质量较大和 log *P* 较高的区域。相比之下，inactive 组化合物多分布在分子质量较低及 log *P* 较低区域，且其 pIC_{50} 值普遍较小。此结果表明，亲脂性较强的分子可能具有较好的生物活性。如图 2b 所示，active 和 inactive 两组化合物的 pIC_{50} 值分布差异明显。active 组的 pIC_{50} 值整体较高，分布范围大致在 6.5 至 10 之间，中位数约为 8，且存在部分高离群点；inactive 组的 pIC_{50} 值则集中在 4 至 5.5 之间，分布较为集中且无明显高值。该结果进一步验证了基于 pIC_{50} 对化合物生物活性分级的有效性。

2.2　机器学习

本研究对 32 种主流机器学习回归模型的预测性能进行了系统对比，包括线性模型、集成学习模型及核方法等，旨在利用 PubChem 分子描述符预测生物活性（pIC_{50}）。模型评估指标为决定系数（R^2）及均方根误差（RMSE）。综合 R^2 和 RMSE 指标，本研究选取 Random Forest Regression（随机森林回归）为最佳回归模型（见图 3）。



a R^2 值　　b 均方根误差（RMSE）值

图 3　不同机器学习回归模型在测试集上的特征比较

Fig. 3　Comparison of characteristics of different machine learning regression models on test sets

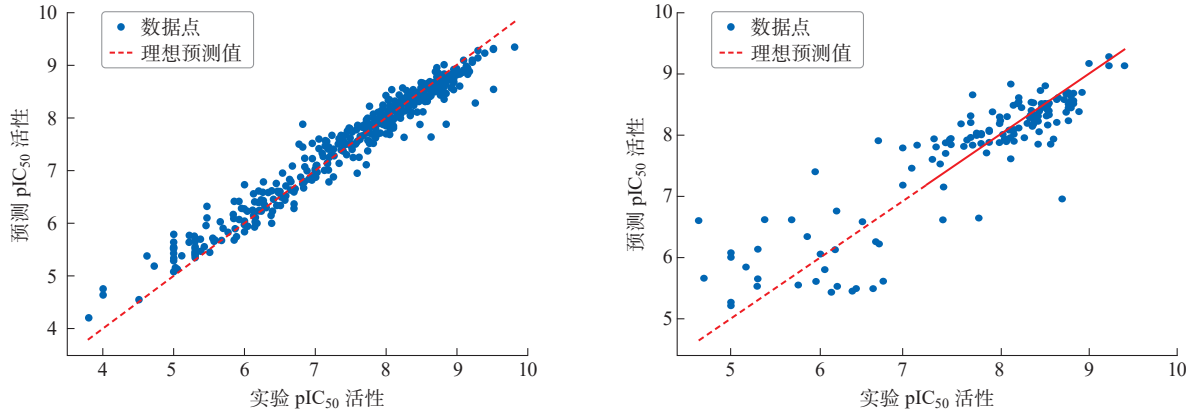
随后针对随机森林回归模型进行超参数网格优化，以进一步提升模型的预测性和泛化能力。基于 4 个选定的重要特征，使用随机参数和 5 重交叉验证来调整模型的超参数（如表 1 所示）。该过程旨在系统地探索模型的潜在配置空间，避免手动调参的低效性和主观性。

表 1　随机森林回归模型的超参数优化网格

Table 1　Super-parameter optimization grid of Random Forest Regression model

参数名称	候选值 1	候选值 2	候选值 3
n estimators	200	300	400
max_depth	None	5	10
min_samples split	2	5	10
min_samples leaf	1	2	4

优化后的随机森林回归模型的参数为：‘max_depth’=None，‘min_samples leaf’=1，‘min_samples split’=2，‘n estimators’=200。在训练集中 R^2=0.956 5，RMSE=0.234 3；在测试集中 R^2=0.800 4，RMSE=0.496 8。数据表明该模型在训练集上表现出较强的预测能力，测试集 R^2 值约为 0.80，表明该模型可以解释未见数据的 pIC_{50} 值方差的很大一部分（80%），表明具有良好的通用性[27]。为了具体展示模型预测效果，绘制了其在训练集和测试集上的预测散点图（见图 4）。



b 训练集　　a 测试集

图 4 Random Forest Regression 模型分别在训练集和测试集上的实验 pIC_{50} 值与预测 pIC_{50} 值的相关性图

Fig. 4 Correlation diagram of experimental pIC_{50} value and predicted pIC_{50} value of Random Forest Regression model on training set and test set respectively

使用从 Enamine 化学数据库（https：//enamine.net/）下载的激酶数据库进行 MELK 抑制剂的预测，结果如表 2 所示。本研究取排序前 5 个的化合物进行后续的分子对接和分子动力学模拟。

表 2 基于 64 960 个候选化合物的 MELK 抑制活性预测结果前 10 个

Table 2 Top 10 prediction results of MELK inhibitory activity based on 64 960 candidate compounds

化合物编号	脂水分配系数	分子质量（脱盐后）/（g·mol^{-1}）	旋转键数	水溶解度对数	pIC_{50} 预测值
Z30082877	2.139	374.394	4	−5.323	8.003 302
Z217682380	2.210	348.399	3	−3.015	8.001 675
Z226447674	3.203	372.421	2	−5.400	7.986 586
Z26973524	2.903	373.406	2	−4.580	7.978 261
Z184476194	5.000	494.522	7	−5.644	7.968 434
Z2143985785	4.111	354.406	4	−5.670	7.957 537
Z1702891961	4.760	514.960	7	−6.680	7.955 056
Z46636102	5.219	375.418	3	−5.840	7.950 935
Z30085094	1.610	360.367	3	−5.000	7.926 913
Z838303034	3.152	398.458	4	−5.410	7.902 861

2.3 分子对接

在机器学习模型预测结果基础上，选取 pIC_{50} 值排序前 5 的化合物（分别命名为 I~V，结构如图 5 所示），对它们与目标蛋白的结合能力进行分子对接分析，以评估各自潜在的生物活性和靶向特异性。

对接结果如图 6 所示，5 个化合物与靶点蛋白的对接结果均取最优结果。对接结果显示，这 5 个化合物都很好地占据 MELK 的结合口袋，且占据了口袋中的大部分可用空间，而不会与原子发生冲突。所有分子在活性口袋内与关键残基（如 VAL-24，LEU-138，CYX-88，ARG-39 等）形成了稳定的氢键和疏水相互作用，其中 CYX-88，GLU-86 和 ARG-39 多次与配体形成氢键，LEU-138，ILE-148 和 VAL-24 则是多次与配体形成疏水作用，且结合距离合理（0.2~0.36 nm）。表明它们确实是潜在的 MELK 抑制剂化合物。

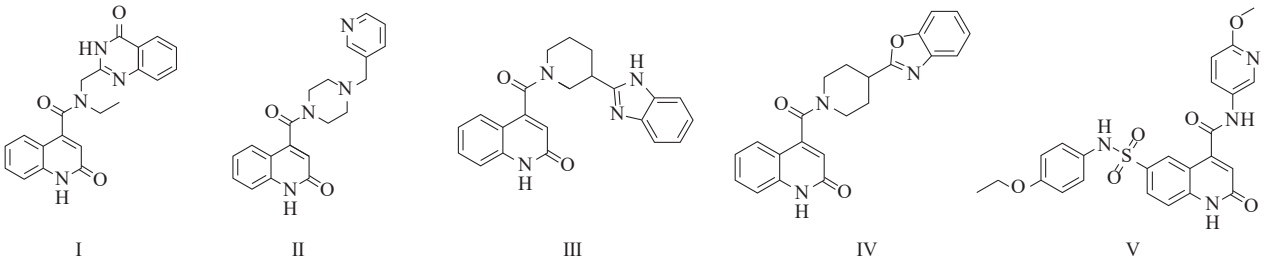
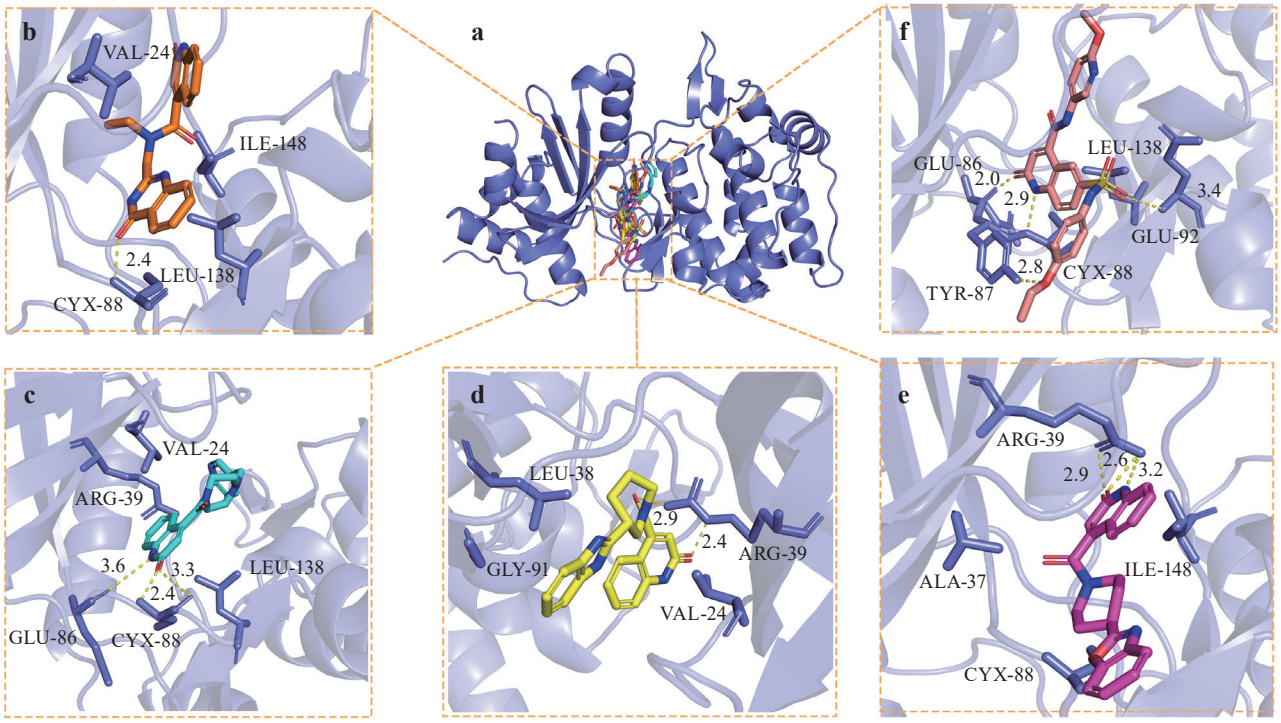


图 5　预测 pIC_{50} 值排序前 5 的化合物结构（I~V）

Fig. 5　Predict the structure of the top 5 compounds with pIC_{50} value (I—V)



注：a 为 5 个配体在结合口袋中的总览图。b~f 分别为化合物 I~V 的详细三维相互作用图。

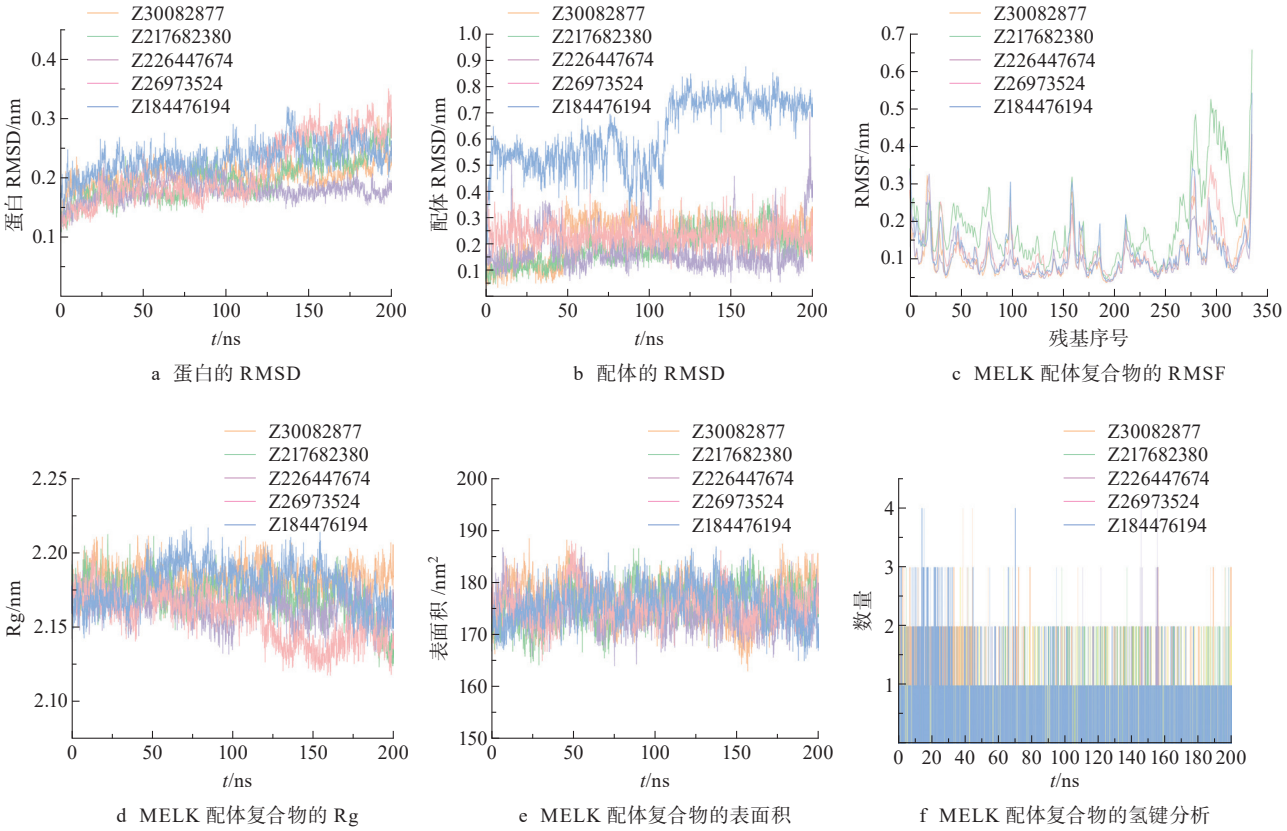
图 6　排序前 5 的预测化合物与 MELK 的分子对接结果

Fig. 6　Molecular docking results of the top 5 predicted compounds with MELK

2.4　分子动力学模拟

采用分子动力学模拟方法进一步研究排序前 5 的化合物与 MELK 活性位点的结合行为及其相互作用。在正式分析之前，采用均方根偏差（RMSD）、均方根涨落（RMSF）、回旋半径（Rg）、溶剂可及表面积（SASA）及氢键（H-bond）5 项指标，对蛋白 — 配体复合物的分子动力学模拟轨迹进行结构稳定性量化评估（见图 7）。首先，RMSD 评估模拟结果与初始结构最终总体偏差的大小，如图 7a 和 b 所示，5 个体系中蛋白的 RMSD 波动平稳，基本保持在 0.1~0.3 nm，配体的 RMSD 也波动平稳，基本保持在 0.15~0.3 nm，这表明 MELK 口袋内的 5 个配体与初始结构没有太大差异。其次，RMSF 是原子位置变化在时间上的平均值，它可以表征整个模拟过程中蛋白质氨基酸的灵活性和运动剧烈程度，对 5 种复合物体系的 RMSF 值的分析表明，5 种复合物残基的波动值非常相似（见图 7c）。蛋白质的回旋半径（Rg）是衡量蛋白质致密性的指标，如果蛋白质折叠稳定，则 Rg 将保持在一个相对稳定的值。计算 5 种复合物的回旋半径，发现 5 条曲线基本平滑，Rg 值基本收敛于 2.125~2.2 nm（见图 7d），这表明蛋白质折叠是稳定的。SASA 为溶剂可及表面积，计算结果表明，5 种化合物的可及表面积相似，均保持在 165~185 nm^2 之间（见图 7e）。最后，氢键在复合物配体与蛋白

质之间的结合中起着至关重要的作用。候选配体与 MELK 之间形成氢键的时间演化分析表明，配体在过程中始终形成 0~4 个氢键，表现出很强的亲和力，特别是排序第 1 的 Z30082877。总地来说，从图 7 可以看出，5 个系统的整体结构都是稳定的，在分子动力学模拟过程中没有发生明显的结构变化[28]。



a 蛋白的 RMSD　b 配体的 RMSD　c MELK 配体复合物的 RMSF

d MELK 配体复合物的 Rg　e MELK 配体复合物的表面积　f MELK 配体复合物的氢键分析

图 7　排序前 5 的预测化合物与 MELK 的模拟结果

Fig. 7　Top 5 predicted compounds and MELK simulation results

2.5 结合自由能计算 MM-GBSA

使用 MM-GBSA 方法计算了 5 种复合物体系中 MELK 与配体之间的结合自由能，如表 3 所示。

表 3　前 5 种化合物的 MM-GBSA 结合能计算结果

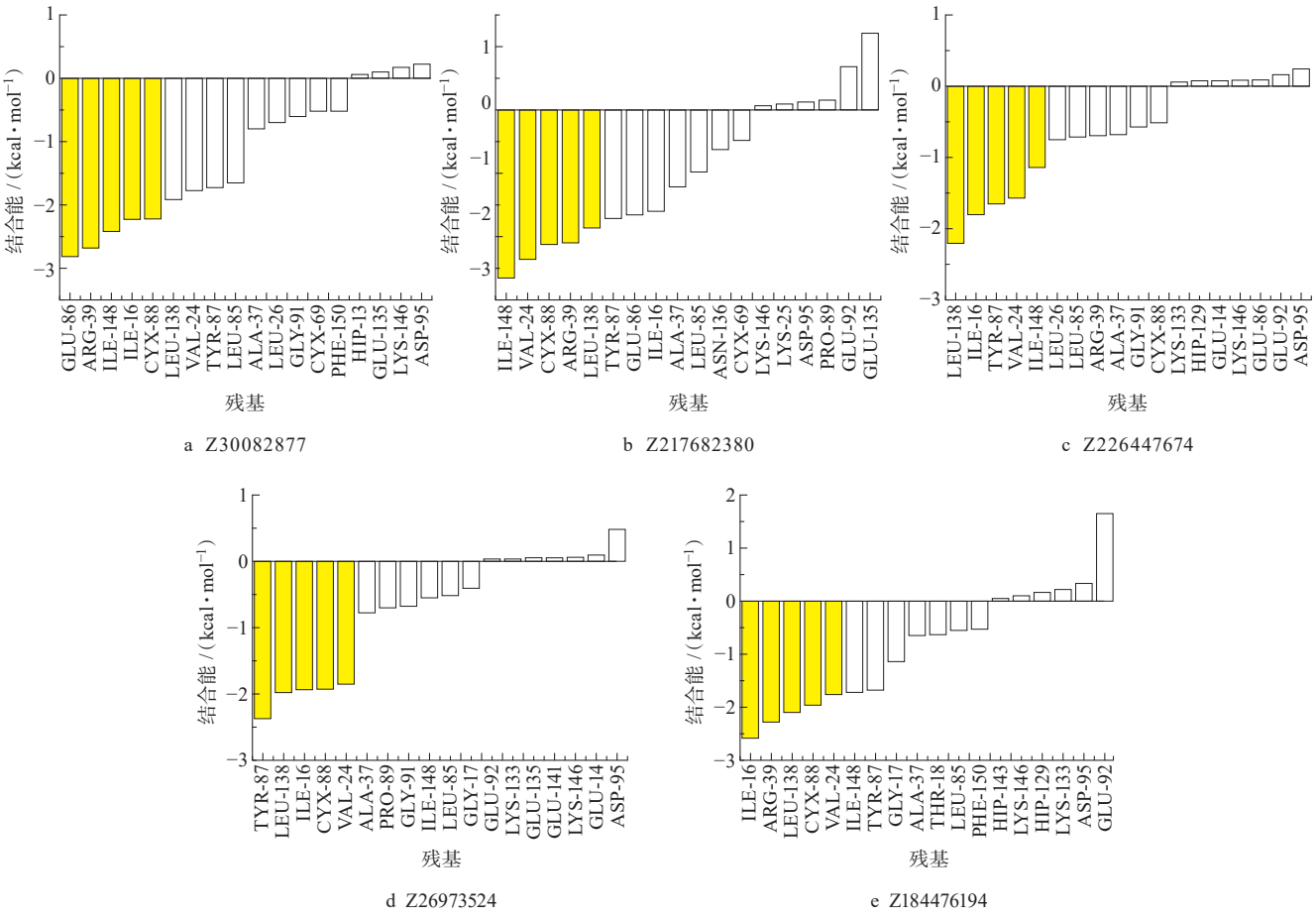
Table 3　Calculation results of MM-GBSA binding energy of the top five compounds

化合物编号	pIC_{50} 预测值	MM-GBSA 结合自由能 /（kcal·mol^{-1}）
OTSSP167	9.130 562	−35.589 6
MELK-T1	6.098 004	−20.298 9
HTH-01-091	8.361 115	−57.672 5
Z30082877	8.003 302	−39.538 5
Z217682380	8.001 675	−36.439 5
Z226447674	7.986 586	−25.831 3
Z26973524	7.978 261	−26.685 1
Z184476194	7.968 434	−31.314 5

注：1 kcal=4.186 kJ。下同。

所有复合物的总结合自由能均为负值，这 5 种化合物的结合效果虽然不及最好的已知的 MELK 抑制剂 HTH-01-091，但均超过 MELK-T1，并且 Z30082877 与 Z217682380 的结合效果均超过 OTSSP167，同时结合构建的机器学习模型对已知抑制剂的预测结果，表明这 5 种化合物存在成为 MELK 抑制剂的潜力。说明靶点蛋白与高分化合物之间能自发形成稳定的复合体。为了探索蛋白质—配体结合界面上对复合物结合有重要贡献的关键残基，对 5 个系统分别采用 MM-GBSA 对复合物体系进行单个残基自由能分解分析（见图 8）。结果表明，活性位点周围的氨基酸残基，尤其是 CYX-88 和 LEU-138，在每一个系统中都作出了很大的贡献。对于每个单独系统而言，GLU-86，ILE-148，VAL-24，TYR-87 和 ILE-16 也都在一个或多个系统中作出很大贡献，进一步稳定了配体在结合口袋中的定位。综上所述，结合自由能计算及能量分解分析进一步确认了筛选的高分化合物与 MELK 之间的强相互作用。其中，LEU-138 和 CYX-884 等关键残基及它们

之间的协同作用，是维持配体高亲和结合的主要结构基础。



a Z30082877　　b Z217682380　　c Z226447674

d Z26973524　　e Z184476194

图 8　排序前 5 的预测化合物的 MM-GBSA 残基能量分解图

Fig. 8　MM-GBSA residue energy decomposition diagram of the top 5 predicted compounds

2.6 排序前 5 化合物的 ADME 预测

为了评估这前 5 种候选化合物的成药潜力，利用 SwissADME 对其进行了全面的 ADME（吸收、分布、代谢、排泄）性质预测，预测结果汇总于表 4 中。

表 4　通过 SwissADME 计算受试化合物的物理化学性质和 ADME 性质

Table 4　Physical and chemical properties and ADME properties of the tested compounds calculated by SwissADME

性质	参数	分子名称				
		Z30082877	Z217682380	Z226447674	Z26973524	Z184476194
理化性质	分子质量 (MW, g/mol)	374.39	348.40	372.42	373.40	494.52
	Consensus log *P* *	2.61	1.89	3.13	3.27	2.78
	TPSA	99.18	69.56	82.11	79.46	148.12
类药性	Lipinski 违规数	0	0	0	0	0
	Bioavailability Score **	0.55	0.55	0.55	0.55	0.55
药代动力学	GI 吸收	High	High	High	High	Low
	BBB 渗透	No	Yes	No	No	No
	P-gp 底物	No	Yes	Yes	Yes	No
	CYP1A2 抑制剂	No	Yes	Yes	Yes	Yes
	CYP2C19 抑制剂	No	No	Yes	Yes	Yes
	CYP2C9 抑制剂	Yes	No	Yes	Yes	Yes
	CYP2D6 抑制剂	No	Yes	Yes	Yes	Yes
	CYP3A4 抑制剂	No	Yes	Yes	Yes	Yes
合成	合成可及性	2.66	2.45	3.01	2.99	3.41

注：*Consensus log *P* 是多种算法预测 log *P* 的平均值。**Bioavailability Score 为 0.55，表示化合物有良好的口服生物利用度概率。

值得注意的是，所有化合物均未违反Lipinski五规则，这预示着它们具有成为口服药物的良好潜力。此外，所有化合物均未触发PAINS和Brenk警报，表明其结构中不包含已知的滥竽充数或有毒的化学基团。

综合来看，Z30082877表现出最为均衡和理想的ADMET特性：高GI吸收、非P-gp底物、非BBB渗透、最低的CYP抑制风险以及良好的类药性。相比之下，Z217682380虽吸收良好，但BBB渗透和P-gp底物是其主要缺陷。Z226447674和Z26973524同样受到P-gp底物和广泛CYP抑制的困扰。Z184476194则因低GI吸收和广泛的CYP抑制而成为5个中前景最弱的候选者。因此，Z30082877被确定为进行后续生物活性测试和优化的首要候选化合物。

3 讨论

本研究旨在通过计算方法筛选和发现针对MELK的潜在活性化合物。首先对原始数据集进行了细致的预处理，包括数据清洗、特征计算（分子质量、log P）和基于pIC_{50}值的活性分类，构建了一个适合机器学习建模的数据集。数据集特征分析显示，高活性与低活性化合物在pIC_{50}值上存在显著差异，且活性化合物倾向于具有特定的理化性质范围。在机器学习建模阶段，系统地比较了多种回归算法的性能。结果表明，Random Forest Regression在预测化合物pIC_{50}值方面表现最优，其在独立测试集上获得了比较好的R^2（约0.80）和RMSE（约0.50）。这表明随机森林模型能够有效捕捉分子描述符与生物活性之间的复杂非线性关系，并具有良好的泛化能力。

为了进一步验证模型预测并从结构层面理解化合物的活性机制，选取了预测pIC_{50}值排序前5的化合物进行了分子对接研究。对接结果揭示了这些高预测活性化合物与MELK活性口袋的潜在结合模式。值得注意的是，所有选定化合物均能有效嵌入活性位点，并通过与关键氨基酸残基（如CYX-88，ARY-39）形成氢键和疏水作用等方式稳定结合。特别是CYS-88，它与所有5个化合物均形成了氢键，提示其在配体识别和结合中的核心作用。后续采用分子动力学模拟方法预测了排序前5的化合物与MELK活性位点的结合行为及其相互作用。评估了RMSD，RMSF，Rg，SASA和H-bond 5个参数，显示出5个系统的整体结构都是稳定的，在分子动力学模拟过程中没有发生明显的结构变化。使用MM-GBSA方法计算了5种复合物体系中MELK与配体之间的结合自由能，表明靶点蛋白与高分化合物之间能自发形成稳定的复合体，其中，LEU-138和CYX-884等关键残基及它们之间的协同作用，是维持配体高亲和结合的主要结构基础。最后为了评估前5种候选化合物的成药潜力，利用SwissADME对其进行了全面的ADME（吸收、分布、代谢、排泄）性质预测，结果表明，Z30082877表现出最为均衡和理想的ADME特性：高GI吸收、非P-gp底物、非BBB渗透、最低的CYP抑制风险以及良好的类药性，被确定为进行后续生物活性测试和优化的首要候选化合物。

尽管本研究取得了一些有意义的成果，但也存在一定的局限性。首先，QSAR模型的预测性能受限于所用描述符的种类和数据集的化学空间覆盖度。未来可以考虑引入更全面的分子描述符（如3D描述符、药效团指纹等）或整合更大、更多样化的数据集来提升模型性能。其次，分子对接本质上是一种理论计算方法，其结果依赖于打分函数的准确性和构象搜索的完备性，真实的结合模式和亲和力仍需实验验证。最后，本研究尚未对预测的化合物进行实验活性测试，这是验证模型和对接结果可靠性的最终手段。尽管存在上述局限，本研究成功整合了机器学习和分子对接方法，为MELK抑制剂的发现提供了一个有效的计算流程。所识别的关键相互作用模式和高预测活性的化合物骨架为后续的药物设计和优化提供了有价值的线索。例如，可以针对CYS-88或ARG-39设计更强的氢键供体和受体，或在疏水空腔内引入合适的疏水基团以增强结合。

4 结论

本研究成功构建并验证了一个基于随机森林算法的QSAR模型，该模型能够有效预测化合物对MELK的pIC_{50}值（测试集R^2约为0.80，RMSE约为0.50）。进一步的分子对接研究揭示了预测活性最高的化合物与MELK活性位点的潜在结合模式，强调了CYX-88和LEU-138等关键残基在配体结合中的重要作用。分子动力学模拟和结合自由能计算，表明了预测化合物在MELK结合口袋中具有很高的稳定性。最后通过ADME预测，综合评定出Z30082877为进行后续生物活性测试和优化的首要

候选化合物。这些发现为模型预测提供了结构层面的支持，并为未来针对 MELK 的新型抑制剂的合理设计和虚拟筛选提供了有价值的计算工具和结构信息。建议对预测活性最高的化合物进行后续的实验合成与生物活性评价，以进一步验证本研究的计算结果。

参考文献：

[1] SEONG H A, MANOHARAN R, HA H. Smad proteins differentially regulate obesity-induced glucose and lipid abnormalities and inflammation via class-specific control of AMPK-related kinase MPK38/MELK activity[J]. Cell Death & Disease, 2018, 9(5): 471.

[2] LIN M L, PARK J H, NISHIDATE T, et al. Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family[J]. Breast Cancer Research, 2007, 9(1): R17.

[3] ZHANG Ya, ZHOU Xiangxiang, LI Ying, et al. Inhibition of maternal embryonic leucine zipper kinase with OTSSP167 displays potent anti-leukemic effects in chronic lymphocytic leukemia[J]. Oncogene, 2018, 37(41): 5520-5533.

[4] JOHNSON C N, BERDINI V, BEKE L, et al. Fragment-based discovery of type I inhibitors of maternal embryonic leucine zipper kinase[J]. ACS Medicinal Chemistry Letters, 2015, 6(1): 25-30.

[5] HUANG H T, SEO H S, ZHANG Tinghu, et al. MELK is not necessary for the proliferation of basal-like breast cancer cells[J]. Elife, 2017, 6: e26693.

[6] YU Wenbo, MACKERELL A D JR. Computer-aided drug design methods[J]. Methods in Molecular Biology, 2017, 1520: 85-106.

[7] CHING T, HIMMELSTEIN D S, BEAULIEU-JONES B K, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of the Royal Society, Interface, 2018, 15(141): 20170387.

[8] 史大华, 韩抒彤, 张钊源, 等. 噻唑衍生物的合成、表征及其乙酰胆碱酯酶抑制活性研究[J]. 江苏海洋大学学报(自然科学版), 2025, 34(3): 66-73.

[9] LO Y C, RENSI S E, TORNG W, et al. Machine learning in chemoinformatics and drug discovery[J]. Drug Discovery Today, 2018, 23(8): 1538-1546.

[10] 孙婷, 刘洋, 魏宠芝, 等. 水中芳香化合物与臭氧反应活性的可解释性机器学习模型[J/OL]. 环境化学, 1-11[2025-07-25].https: /link.cnki.net/urlid/11.1844.X.20250725.1306.006.

[11] ZHANG Lu, TAN Jianjun, HAN Dan, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery[J]. Drug Discovery Today, 2017, 22(11): 1680-1685.

[12] 尚雅欣, 雷小洁, 方子牛, 等. 基于 GA-BP 神经网络模型的抗乳腺癌候选药物活性预测[J]. 数学理论与应用, 2024, 44(2): 103-125.

[13] DAS A, PRAJAPATI A, KARNA A, et al. Structure-based virtual screening of chemical libraries as potential MELK inhibitors and their therapeutic evaluation against breast cancer[J]. Chemico-Biological Interactions, 2023, 376: 110443.

[14] PASALA C, CHILAMAKURI C S R, KATARI S K, et al. An *in silico* study: novel targets for potential drug and vaccine design against drug resistant *H. pylori*[J]. Microbial Pathogenesis, 2018, 122: 156-161.

[15] COBRE A D F, ARA A, ALVES A C, et al. Identifying 124 new anti-HIV drug candidates in a 37 billion-compound database: an integrated approach of machine learning (QSAR), molecular docking, and molecular dynamics simulation[J]. Chemometrics and Intelligent Laboratory Systems, 2024, 250: 105145.

[16] KRAMER C, KALLIOKOSKI T, GEDECK P, et al. The experimental uncertainty of heterogeneous public K(i) data[J]. Journal of Medicinal Chemistry, 2012, 55(11): 5165-5173.

[17] SIMEON S, ANUWONGCHAROEN N, SHOOMBUATONG W, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking[J]. PeerJ, 2016, 4: e2322.

[18] SRISONGKRAM T, KHAMTANG P, WEERAPREEYAKUL N. Prediction of KRAS (G12C) inhibitors using conjoint fingerprint and machine learning-based QSAR models[J]. Journal of Molecular Graphics & Modelling, 2023, 122: 108466.

[19] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination *R*-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Computer Sci-

ence, 2021, 7: e623.

[20] HUBER N R, MISSERT A D, GONG Hao, et al. Random search as a neural network optimization strategy for convolutional-neural-network (CNN) -based noise reduction in CT[J]. Proceedings of SPIE—the International Society for Optical Engineering, 2021, 11596: 115961U.

[21] TROTT O, OLSON A J. AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading[J]. Journal of Computational Chemistry, 2010, 31(2): 455-61.

[22] CASE D A, AKTULGA H M, BELFON K, et al. The AmberTools[J]. Journal of Chemical Information and Modeling, 2023, 63(20): 6183-6191.

[23] YLILAURI M, PENTIKÄINEN O T. MMGBSA as a tool to understand the binding affinities of filamin-peptide interactions[J]. Journal of Chemical Information and Modeling, 2013, 53(10): 2626-2633.

[24] DAINA A, MICHIELIN O, ZOETE V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules[J]. Scientific Reports, 2017, 7: 42717.

[25] LIPINSKI C A, LOMBARDO F, DOMINY B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[J]. Advanced Drug Delivery Reviews, 2001, 46(1/2/3): 3-26.

[26] QIAN Jingjing, ZOU Jingpei, LIU Shanming, et al. Synthesis, characterization, crystal structure, and cholinesterase inhibitory activity of 2-phenylthiazole derivatives[J]. Journal of Molecular Structure, 2023, 1282: 135248.

[27] PIRZADA R H, YASMEEN F, HASEEB M, et al. Small molecule inhibitors of IL-1R1/IL-1beta interaction identified via transfer machine learning QSAR modelling[J]. International Journal of Biological Macromolecules, 2024, 282(Pt 5): 137295.

[28] HALIM S A, WAQAS M, ASIM A, et al. Discovering novel inhibitors of P2Y(12) receptor using structure-based virtual screening, molecular dynamics simulation and MMPBSA approaches[J]. Computers in Biology and Medicine, 2022, 147: 105743.

(责任编辑: 褚金红, 李琴)