高等学校化学学报
CHEMICAL JOURNAL OF CHINESE UNIVERSITIES
研究论文

基于机器学习与分子动力学模拟发现CDK2抑制剂

谭英佳[1]，陈　亮[2]，刘聿琳[1]，那日松[3]，赵　熹[1]
（1. 吉林大学化学学院，理论化学研究所，长春 130023；
2. 吉林省农业科学院大豆研究所，长春 130033；
3. 河南农业大学植物保护学院，郑州 450046）

摘要　通过机器学习和分子动力学模拟方法发现了细胞周期蛋白依赖性激酶2（CDK2）潜在的抑制剂. 首先，利用现有的大型活性数据库和机器学习算法，建立了针对CDK2抑制剂的分类模型. 采用圆形指纹（ECFP6）的极端梯度提升树模型（XGBoost）筛选Enamine数据库，并选出了1152个新型化合物. 通过分子对接和打分函数对这些潜在化合物在CDK2中的亲和力进行了排序，并采用指纹聚类的方法将化合物分为4类. 分别从4类中选择1种对接评分较高的化合物，然后对4种化合物进行了类药性分析和分子动力学模拟. 类药性分析结果表明，筛选出的4种潜在的CDK2抑制剂（Z1766368563，Z363564868，Z1891240670和Z2701273053）具有良好的成药性，并在分子动力学模拟结果中具有较高的结合自由能. 这4种化合物可作为CDK2的先导化合物进行后续的改造和优化.

关键词　CDK2抑制剂；机器学习；分子动力学；结合自由能
中图分类号　O641　　**文献标志码**　A　　doi：10.7503/cjcu20240442

Discovery of CDK2 Inhibitors Based on Machine Learning and Molecular Dynamics Simulations

TAN Yingjia[1], CHEN Liang[2], LIU Yulin[1], NA Risong[3], ZHAO Xi[1*]
(*1. College of Chemistry, Institute of Theoretical Chemistry, Jilin University, Changchun* 130023, *China*;
2. Soybean Research Institute, Jilin Academy of Agricultural Sciences, Changchun 130033, *China*;
3. College of Plant Protection, Henan Agricultural University, Zhengzhou 450046, *China*)

Abstract　Four potential cyclin-dependent kinase 2 (CDK2) inhibitors were discovered through machine learning and molecular dynamics simulation methods. First, a classification model for CDK2 inhibitors was established using existing large-scale activity databases and machine learning algorithms. The extreme gradient boosting (XGBoost) model with extended-connectivity fingerprints (ECFP6) was used to screen the Enamine database, identifying 1152 novel compounds. These potential compounds were then ranked based on their affinity for CDK2 using molecular docking and scoring functions. The compounds were clustered into four categories using fingerprint clustering methods, and one compound with a high docking score was selected from each category. Subsequently, the four selected compounds underwent drug-likeness analysis and molecular dynamics simulations. The four potential CDK2 inhibitors (Z1766368563, Z363564868, Z1891240670 and Z2701273053) demonstrated good drug-likeness properties and high binding free energy in molecular dynamics simulation results. The findings suggest that these four

收稿日期：2024-09-25. 网络首发日期：2024-11-18.
联系人简介：赵　熹，男，博士，副教授，主要从事生物大分子的计算及药物筛选方面的研究. E-mail：zhaoxi@jlu.edu.cn
基金项目：国家自然科学基金(批准号：32472613)和河南省杰出青年基金项目(批准号：232300421008).
Supported by the National Natural Science Foundation of China(No.32472613) and the Excellent Youth Foundation of Henan province, China (No.232300421008).

compounds can serve as lead compounds for subsequent modification and optimization as CDK2 inhibitors.

Keywords Cyclin-dependent kinase 2 (CDK2) inhibitor; Machine learning; Molecular dynamics; Binding free energy

细胞周期蛋白依赖性激酶2(Cyclin-dependent kinase 2, CDK2)[1~3]是一种在细胞周期调控中起关键作用的蛋白激酶, 其与细胞周期蛋白(Cyclins)结合后可形成活性复合物, 从而调控细胞从G1期进入S期以及S期的进程. CDK2通过调控DNA复制和细胞分裂确保细胞周期的正常进行[4,5]. 在许多癌症中, CDK2的活性异常增加, 导致细胞不受控制地增殖[6]. 抑制CDK2的活性可以阻止癌细胞的增殖, 诱导细胞凋亡. CDK2的异常活性与多种癌症的发生和发展密切相关[7~11]. 例如, 在乳腺癌、肺癌和黑色素瘤中, CDK2的过表达或活性增加常常导致细胞周期失控, 促进肿瘤的生长和扩散. 因此, CDK2被认为是一个潜在的抗癌药物靶点.

CDK2早期的抑制剂[12,13]被归类为I型抑制剂, 结合区域位于激酶的保守ATP的结合位点, 这与CDK家族相似, 意味着CDK2抑制剂很难获得特异性. 目前, 大多数已进入临床试验的抑制剂都属于I型抑制剂. II型抑制剂是优先结合无活性的CDK2构象与活化细胞周期蛋白的结合位点的具有竞争性的化合物. II型抑制剂具有开发CDK2活性抑制剂的更大潜力, 但此类抑制剂的研究较少. 同样, 靶向CDK2其它结合区域的III型变构抑制剂也处于初步的开发阶段. 研究发现, CDK2与其它CDK抑制剂联合对于一些癌症可以取得更好的疗效, 这极大地拓宽了药物开发的范围. CDK2/CDK7抑制剂可以治疗CDK2依赖性癌症和细胞周期蛋白E1扩增引起的癌症, 其首先直接抑制CDK2, 然后通过抑制CDK7来防止残留的CDK2的活化[13]. CDK2和PI3K双靶点抑制剂已被证明在结肠癌细胞系中具有较好的体外疗效[14]. CDK1/CDK2双靶点抑制剂的组合(例如, roscovitine)与PI3K相互作用, 可使人神经胶质瘤细胞死亡[15]. 除了上述作用外, 有证据表明CDK2活性也会影响细胞分化和适用性免疫表达[15], 因此, 新一代CDK2抑制剂的开发具有重要意义.

CDK2抑制剂的结合位点是三磷酸腺苷(ATP)结合位点(图1), 具有明显的疏水区域和设计CDK2抑制剂的关键位点. CDK2蛋白含有β-折叠的N末端和由螺旋组成的C末端, 呈双叶形状. ATP结合位点位于两个结构域界面处的裂痕上. 根据ATP中的化学部分, 该裂痕可分为3个区域: 其中一个疏水区域包括氨基酸ILe10, Ala31, Val64, Phe80, Glu81, Phe82, Leu83, Leu134和Ala144, 形成ATP的腺嘌呤部分的结合环境[16]; 第二个区域包含ATP的核糖部分相互作用的3个氨基酸(Val18, Asp86和Gln131), 与核糖的羟基形成氢键; Lys33, Asp145和Asn132与ATP的三磷酸链构象的极性氨基酸组成第三区.

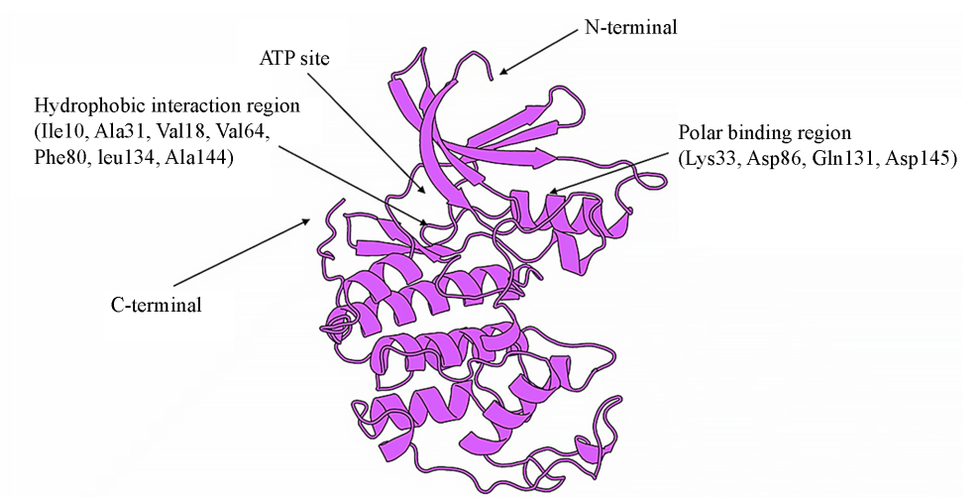

Fig. 1 Structure and binding sites of CDK2

本文利用现有的活性数据和机器学习算法, 建立了针对CDK2抑制剂的分类模型. 通过指标分析, 发现基于扩展连通性指纹[17](ECFP6)的极端梯度提升树模型(XGBoost)、随机森林(RF)模型以及

图神经网络指纹算法(Attentive FP)在CDK2数据集中表现出良好的预测性能. 最终, 选择XGBoost模型在Enamine数据库中筛选出4种潜在的CDK2抑制剂(Z1766368563, Z363564868, Z1891240670和Z2701273053), 其具有良好的成药性, 并能稳定存在于CDK2中, 可作为先导化合物进行后续的改造和优化.

1 材料与方法

1.1 数据集的收集和表示

使用的CDK2数据集来自BindingDB数据库(https: //www. bindingdb. org/). 为保证数据集的质量, 对CDK2数据集进行了如下处理: (1) 删除缺失活性的化合物; (2) 除去化合物中的盐和络合物; (3) 删除重复的分子; (4) 除去分子量低于30或高于1000的化合物. 经过以上步骤处理后, CDK2数据集中有5680个化合物(https: //github. com/yulinliu111/DATA), 活性与非活性化合物分别为3139和2541个.

为了更好地评估分类模型的性能和泛化能力, 将CDK2数据集随机分割为3个子集: 训练集(80%)、验证集(10%)和测试集(10%). 其中, 训练集用来训练二分类模型, 验证集用来评估模型和进行超参数调优, 测试集则用来最终评估模型的预测能力和泛化性.

1.2 化学空间可视化

将数据集拆分后, 对数据集中每个化合物生成了1024位ECFP6分子指纹. 并使用t-分布式随机邻域嵌入(t-SNE)[18,19]算法将指纹嵌入到二维向量中, 以便实现对化合物的可视化. 为更全面地描述化合物, 使用RDKit生成了一些简单的化学描述符, 包括分子量(MW)、拓扑极性表面积(TPSA)、氢键供体和受体数量(HBD/HBA)等. 在使用化学描述符时, 需要将描述符进行归一化处理, 可以避免过大的分子描述符对稀疏向量产生偏移, 避免影响t-SNE图的可视化结果.

1.3 模型构建与超参数调优

采用两种传统的机器学习算法[随机森林(RF)和极端梯度提升树模型(XGBoost)][20,21]和3种深度学习图算法[图卷积神经网络(GCN)、图注意力网络(GAT)和图注意力网络指纹(Attentive FP)][22,23]对CDK2分别建立分类模型, 来预测CDK2抑制剂的活性. 所有分子表示均使用Python语言环境实现. 为解决随机分割导致数据不平衡的问题, 将传统的机器学习模型使用10折交叉验证来训练数据, 图神经网络每个模型运行10次, 并将10次预测的评估指标的平均值作为最终的评价结果.

超参数在深度神经网络中是非常重要的. 为了更好地训练深度图神经网络模型, 通过训练集与验证集的评估指标, 采用了贝叶斯优化进行超参数搜索, 探索的通用图神经网络的超参数为隐藏层数(Number_layer)、学习率(Learning_rate)、全连接层丢弃率(Predictor_dropout)、权值衰减(Weight_decay)和批处理大小(Batch_size).

1.4 模型评估

使用特异性(SP)、灵敏度(SE)、马修斯相关系数(MCC)、准确度(ACC)以及工作特异性曲线(AUC)对模型进行评估. MCC取值在-1~1之间, -1代表完美负相关, 1代表完美正相关, 0代表随机分类器. 其它评估指标的取值范围都在0~1之间, AUC为1时为完美分类器, 为0. 5时为随机分类器. 评估指标计算公式如下:

$$SP = \frac{TN}{TN + FP} \tag{1}$$

$$SE = \frac{TP}{TP + FN} \tag{2}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{3}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

其中，TP，FN，TN和FP分别为真阳性、假阴性、真阴性和假阳性的数量.

1.5 分子对接药物筛选

将机器学习模型初步筛选的1152个CDK2潜在抑制剂使用AutoDock Vina[24]软件进行分子对接，中心网格通过CDK2的ATP结合位点确定，对接盒子尺寸为2.2 nm×2.2 nm×2.2 nm. 使用打分函数的随机全局优化将配体对接到CDK2受体结合口袋中. 根据对接评分和分子构象选择最好的对接情况. 对接后的部分化合物列于表S1(见本文支持信息)，化合物的详细信息可根据ID在PubChem数据库(https://pubchem.ncbi.nlm.nih.gov/)检索到.

1.6 类药性分析

类药性表示一个理想药物应具备的特征，是成为药物所表现出来的理化性质(如分子量、脂水分配系数)以及结构特征(如环结构、可旋转键数目和氢键供体/受体数目等)的综合反映. 当类药性化合物所表现出来的结构和理化性质与体内药动学参数具有较好的相关性时，化合物才可以作为先导化合物. 使用SwissADME[25]和ADMETlab[26]在线网站对筛选后的4种化合物的类药性进行分析.

1.7 分子动力学模拟与结合自由能计算

1.7.1 *初始结构的获取与处理* 本文使用的CDK2晶体蛋白(PDB code：3PXZ)来自RCSB(http://www.rcsb.org)，删除结晶水分子和其它辅助结晶的复合物. 4个分子(Z1766368563，Z363564868，Z2701273053和Z1891240670)的几何参数均使用Gaussian 16程序[27]，在6-31+G(*d*)基组和B3LYP[28]方法下计算得到.

1.7.2 *分子动力学模拟* 将处理好的晶体结构使用PROPKA3.1程序[29]进行质子化处理，pH=7.0时确定了蛋白质氨基酸残基的状态. 使用AmberTools21[30]生成小分子的力场文件. 将处理过后的蛋白-配体使用GROMACS程序中的Amber力场[31]和三点水模型(TIP3P)[32]进行溶剂化，加入4个氯离子平衡系统的电荷. 首先，经过1000步能量最小化消除原子间不合理的接触，采用最快下降法使系统能量最小. 为使水分子和离子均匀分布在系统周围，在310 K的温度下，使用V-rescale耦合算法进行了100 ps的等容等温(NVT)模拟，在相同温度和一个大气压下使用Parrinello-Rahman耦合算法进行200 ps的等温等压(NPT)模拟. 整个过程中添加周期性边界条件，使用LINCS算法[33]约束键长. 范德华相互作用截断半径为1.2 nm，长程静电相互作用使用PME[34]方法. 最后在MD系综进行100 ns的分子动力学模拟. 分子动力学模拟采用GROMACS 2018.8程序[35]完成，使用VMD[36]，Pymol[37]和Discover Studio[38]软件对模拟结果进行处理及可视化分析.

1.7.3 *结合自由能计算* 为了比较4种化合物对CDK2的亲和力，采用MM-PBSA[39]方法计算结合自由能，包括溶剂化能、非溶剂化能和熵的贡献，从每个配体-受体系统中的最后10 ns提取500个快照，计算了4个体系的结合自由能. 计算公式如下：

$$\begin{aligned}\Delta G_{\text{bind}} &= \Delta G_{\text{complex}} - \Delta G_{\text{receptor}} - \Delta G_{\text{ligand}} \\ &= \Delta E_{\text{MM}} + \Delta G_{\text{solv}} - T\Delta S \\ &= \Delta E_{\text{ele}} + \Delta E_{\text{vdW}} + \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}} - T\Delta S\end{aligned} \tag{5}$$

式中：ΔG_{bind}(kJ/mol)为蛋白与配体小分子之间的结合自由能；$\Delta G_{\text{complex}}$，$\Delta G_{\text{receptor}}$和$\Delta G_{\text{ligand}}$(kJ/mol)分别为蛋白-配体复合物、蛋白和配体的自由能；ΔE_{MM}(kJ/mol)为真空平均分子力学能；ΔG_{solv}(kJ/mol)为溶剂化自由能；$-T\Delta S$(kJ/mol)为配体结合构象熵；ΔE_{ele}和ΔE_{vdW}(kJ/mol)分别为静电和范德华相互作用；ΔG_{polar}和$\Delta G_{\text{nonpolar}}$(kJ/mol)分别为极性溶剂化能和非极性溶剂化能.

2 结果与讨论

2.1 化学空间可视化

在机器学习模型中结构多样性的化合物通常会覆盖较大化学空间，并且使得机器学习模型具有普遍的适用性. 在建立模型之前，分析了CDK2数据集中化合物的多样性. 图2(A)为基于ECFP6指纹的CDK2训练集与验证集的空间分布，图2(B)为在ECFP6指纹的基础上加了6个理化性质描述符的空间

分布. 由图2可以看出，CDK2数据分布广泛，说明此数据集中的化合物结构有一定的多样性；并且训练集与验证集的分布几乎一致，验证集中化合物大致分布在训练集内部，表明验证集评估模型是可靠的. 图2结果表明，CDK2数据中分子的空间分布广泛，有助于提高训练集的多样性，克服模型过拟合和欠拟合的问题，从而提高了模型的可靠性和泛化性.

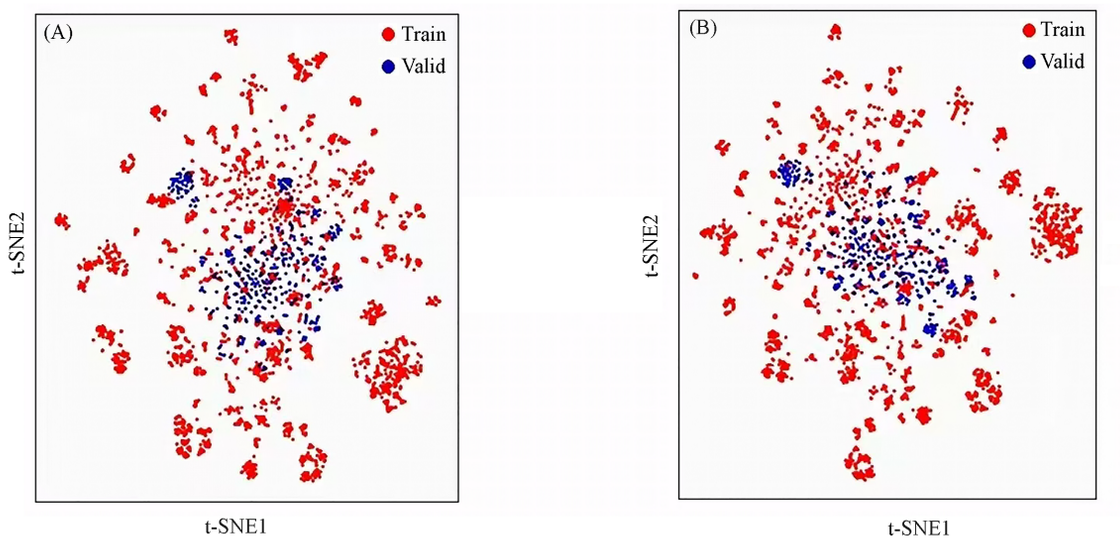

Fig. 2　Visualized spatial distribution of CDK2 training sets and validation sets by t-SNE

(A) Based on ECFP6 fingerprints; (B) based on ECFP6 fingerprint plus 6 physicochemical property descriptors.

2.2　分子骨架多样性分析

使用RDKit分别分析CDK2数据集中活性与非活性化合物的骨架多样性. RDKit提供了根据分子二维结构将分子分解为Murcko骨架和碳骨架的标准. Murcko骨架是除去所有R基团，但保留了环系统的接头；碳骨架是将Murcko骨架中的杂原子更改为碳原子，所有键的类型转换为单键. 结果显示，在CDK2数据集中，活性抑制剂数据中产生了863个独特的Murcko骨架和456个独特的碳骨架；1398个Murcko骨架和863个碳骨架存在于CDK2非活性的数据中. 其中，在Murcko骨架中，大约95%的骨架不超过10个分子；同时大约90%的碳骨架不超过10个分子. 整理了CDK2活性与非活性排在前五位的骨架(表S2，见本文支持信息)，从中能够观察到活性分子与非活性分子的骨架有所不同，表明抑制剂活性的差异可能与骨架相关. 数据集中骨架的统计结果反映了CDK2数据集中化合物的结构具有多样性. 骨架活性与非活性的差异可以为药物的优化和设计提供指导.

2.3　机器学习预测模型的性能评估

使用了两种传统机器学习模型和3种深度学习模型. 所有模型均在Python环境下实现，并且经过了多次验证，以避免得到随机的结果. CDK2模型的验证集与测试集评估结果分别列于表1和表2.

从表1与表2可以看出，验证集总体分数略微高于测试集或与测试集持平，这是因为在训练过程中，验证集参与了模型的超参数选择，而测试集模拟未知数据的分布可能与训练集有差别. 从表2可见，基于ECFP6指纹的RF和XGBoost模型在CDK2数据集表现很出色，测试集AUC分别是93.80%和93.56%. 与之相比的MACCS指纹的RF和XGBoost模型在预测抑制剂的能力上略显不足. 使用ECFP6指纹建立的RF模型中SP(准确识别非活性化合物分数)值为91.77%，表明RF模型在预测非活性化合物时具有较高的正确性. 基于ECFP6指纹的XGBoost模型预测活性和非活性化合物的能力良好并且预测两者的能力均衡. 在图模型中，Attentive FP预测效果最好，测试集AUC值为92.34%，因为模型考虑了原子和化学键的信息，并且通过注意力机制提高了重要原子的权重. GAT模型的SE值为94.58%，表明该模型相较于其它模型能够更准确地预测活性抑制剂；但是该模型在预测非活性抑制剂时效果差，SP值仅为56.40%.

Table 1　Performance of CDK2 inhibitor machine learning algorithm on validation set

Model	Feature	AUC(%)	ACC(%)	MCC(%)	SE(%)	SP(%)
RF	MACCS	91.73	84.55	69.16	82.53	86.57
	ECFP6	93.80	86.80	74.23	82.37	91.22
XGBoost	MACCS	86.00	77.14	55.02	84.17	70.16
	ECFP6	94.14	88.20	76.65	88.58	90.81
GAT	Graph	90.53	73.56	55.16	92.17	57.63
GCN	Graph	91.53	83.10	64.17	75.36	87.42
Attentive FP	Graph	93.11	86.99	71.13	88.13	86.99
Mean value		91.55	82.91	66.50	84.76	81.54
Variance		6.58	26.17	65.30	25.75	138.63

Table 2　Performance of CDK2 inhibitor machine learning algorithm in the test set

Model	Feature	AUC(%)	ACC(%)	MCC(%)	SE(%)	SP(%)
RF	MACCS	90.85	83.14	66.21	81.88	84.39
	ECFP6	93.80	86.00	72.55	79.76	91.77
XGBoost	MACCS	91.31	84.20	68.15	84.65	83.75
	ECFP6	93.56	87.93	73.37	88.28	89.58
GAT	Graph	90.67	75.49	52.85	94.58	56.40
GCN	Graph	91.00	81.30	63.72	73.75	88.72
Attentive FP	Graph	92.34	85.96	70.73	87.63	86.45
Mean value		91.93	83.43	66.80	84.36	83.01
Variance		1.48	14.50	42.56	38.64	124.98

2.4　机器学习模型解释

Shapley additive explanations(SHAP)是一个解释机器学习模型结果的工具[40,41]. 使用SHAP对基于ECFP6指纹的RF与XGBoost模型结果进行解释. 在两个模型中，当化合物中存在1476位、1582位、74位和1292指纹片段时，会提高化合物为活性抑制剂的可能性[图3(A)]；当化合物结构中存在1480位、80位、352位以及1573指纹片段时，可能会降低化合物抑制CDK2的活性[图3(B)]. 具体的指纹片段结构列于表S3(见本文支持信息).

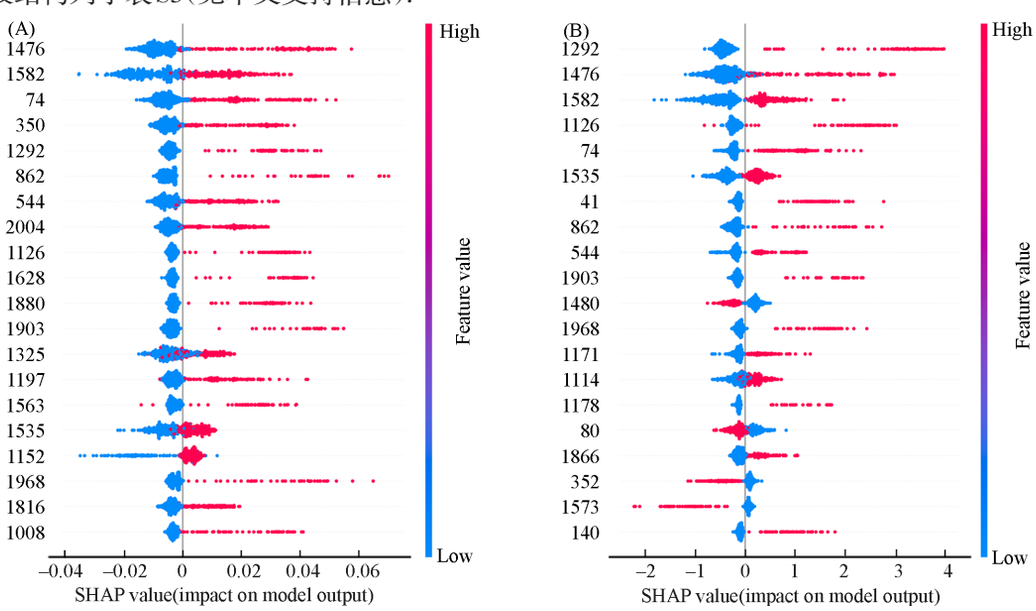

Fig. 3　Visualized feature importance of RF model(A) and XGBoost model(B) based on ECFP6 fingerprint in CDK2 dataset

The higher the feature value is, the higher the feature importance is.

除了对模型指纹进行解释外，单个分子中原子的权重图也有助于对一些活性与非活性结构的理解. 在活性抑制剂[图4(A)]中，分子结构中氮元素颜色较深，尤其位于中心的氨基以及苯环上的氮，说明模型在训练时对这些原子高度关注，该结构对CDK2抑制剂的活性起到了重要作用. 而在非活性

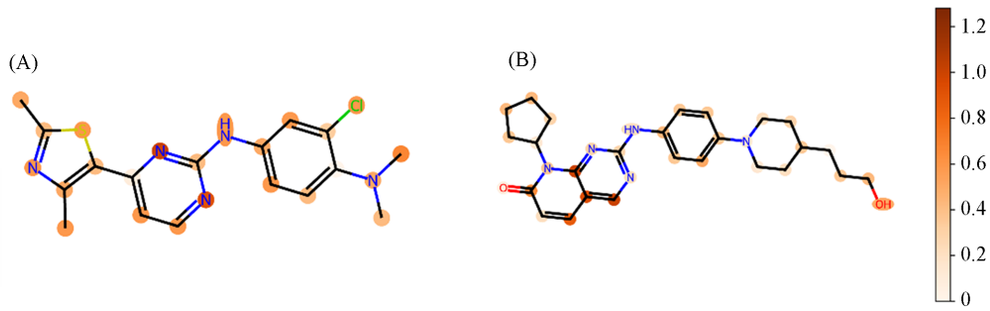

Fig. 4　Atomic heat diagrams of the active inhibitor(A) and inactive inhibitor(B)

(A) The darker atomic color indicates a positive correlation with the activity of the inhibitor; (B) the darker primary color indicate a negative correlation with inhibitor activity

抑制剂[图4(B)]中，被关注的是一些羰基碳、苯环和链上的一些碳原子、羟基等，这些结构被模型认为是非活性结构. 上述使用原子权重的结果与分子指纹模型解释的结果一致.

2.4.1　*机器学习模型的药物筛选*　研究建立了5个机器学习模型，使用评估指标对已建立的模型进行评价. XGBoost模型在预测活性和非活性化合物时都有很好的效果. RF与Attentive FP效果良好，但是在预测活性或者非活性数据时稍有不足. 考虑到各个指标的利弊，最终选择使用XGBoost模型预测CDK2潜在抑制剂. 此次筛选所用的数据来源于Enamine HLL-50 (https：//enamine. net/)，共包含5万个新型未知活性化合物. 首先，去掉与数据集相同的分子，删除盐和络合物；随后，使用XGBoost模型进行预测，得到了1152个潜在活性化合物(https：//github. com/yulinliu111/DATA).

2.4.2　*分子对接药物筛选*　将机器学习模型初步筛选的1152个CDK2潜在抑制剂进行分子对接，发现1152个化合物与CDK2受体的对接评分均在−29. 3以下[图5(A)]，说明建立的模型在预测CDK2活性

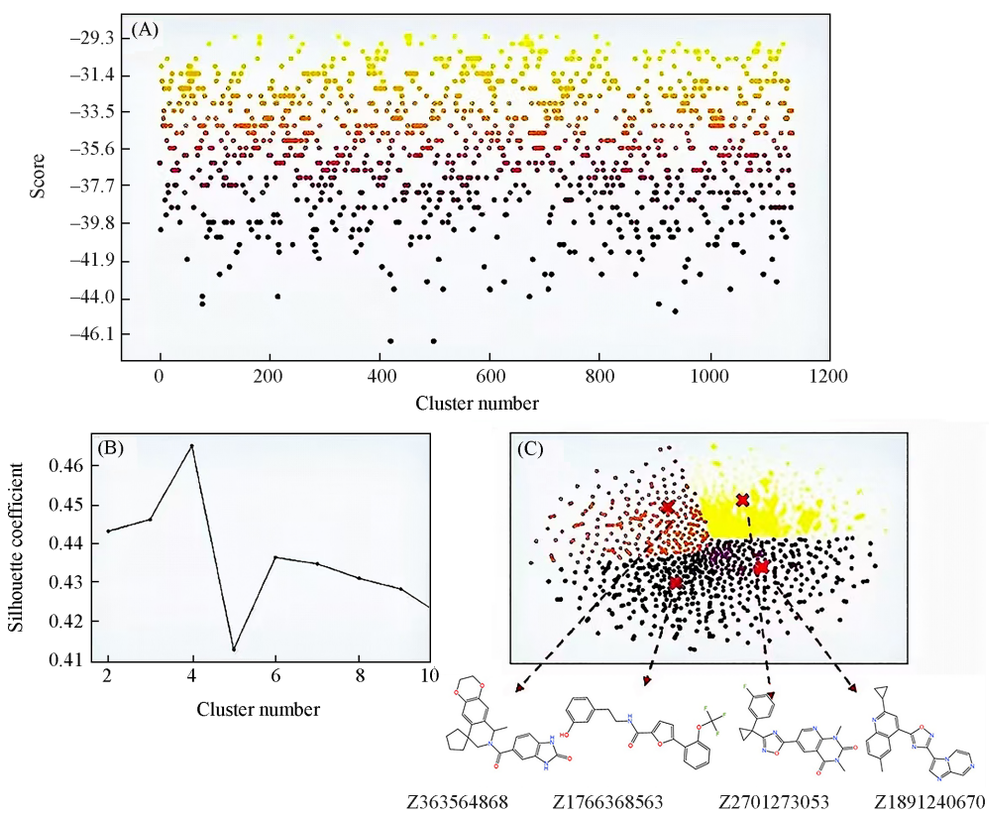

Fig. 5　Docking scores of 1152 screened molecules(A), the number of clusters determined based on the silhouette coefficient index(B), cluster diagrams of compounds and molecular structures of the four chosen compounds(C)

(B) The closer the value of silhouette coefficient is to 1, the better the clustering effect is.

抑制剂方面具有良好的效果. 考虑到筛选的化合物结构应具有多样性, 使用K-Means算法进行聚类, 利用轮廓系数聚类指标来判断最优聚类个数. 轮廓系数取值范围为[-1, 1], 取值越接近1, 说明聚类效果越好; 取值越接近-1则效果越差. 从图5(B)可见, 当聚类类别为4时, 轮廓系数最大. 因此, 将1152个化合物分为4类, 每种颜色代表不同的类别, 相同颜色的化合物具有相似的结构[图5(C)]. 从每个类别中找出对接评分较高的化合物作为筛选的CDK2潜在抑制剂, 并进行后续的类药性分析.

通过SHAP对模型预测出的化合物的活性与非活性结构进行解释, 通过检查是否存在已知数据集中得出的活性或非活性结构, 来判断模型是否是随机预测活性化合物. 从图S1(A)(见本文支持信息)可见, 模型将一个化合物预测为活性抑制剂的原因是: 该化合物的活性指纹片段(如1292, 156和1272等)的权重(红色)超过了非活性的指纹片段(蓝色)的权重值(2.23). 从图S1(B)可见, 当一个化合物中含有活性片段(红色), 但是化合物中的非活性指纹片段权重(蓝色)远超过活性指纹片段时, 模型会认为此化合物是非活性抑制剂.

2.5 类药性分析

表3列出了使用SwissADME和ADMETlab在线网站对筛选后的4种化合物的类药性分析指标(其中, MW是分子量, HBD/HBA代表氢键供体/受体数量, TPSA代表拓扑极性表面积, GI absorption代表肠胃吸收程度, QED是定量评估类药性分数, lg*P*是脂/水分配系数).

Table 3 Chemoid-like analysis of the four chosen compounds

Compound	MW	HBD	HBA	TPSA/nm^2	GI absorption	QED	lg*P*
Z1766368563	391.3	2	7	0.717	High	0.648	4.487
Z363564868	419.4	2	4	0.872	High	0.630	3.360
Z1891240670	393.4	0	7	0.958	High	0.528	2.415
Z2701273053	368.4	0	6	0.820	High	0.475	5.003

通过表3可以看出, 4种分子均具有较高的QED分数和高的胃肠道吸收, HBA数量在0~10之间, HBD在0~5之间, TPSA在0~140之间, lg*P*在0~5之间. 可见, 选择的4种化合物具有良好的类药性, 可以作为先导化物进行后续的优化和改造.

2.6 分子动力学与结合自由能分析

将每种配体-受体复合物体系在显式水溶液中进行100 ns的MD模拟. 为了验证各配体在受体中的稳定性和MD模拟结果的合理性, 在100 ns的MD轨迹上检测了骨架原子的均方根偏差(RMSD). 由图6可见, 各体系在初始波动后逐渐达到动力学平衡状态, 波动范围在0.2~0.3 nm之间. 图6(B)~(D)

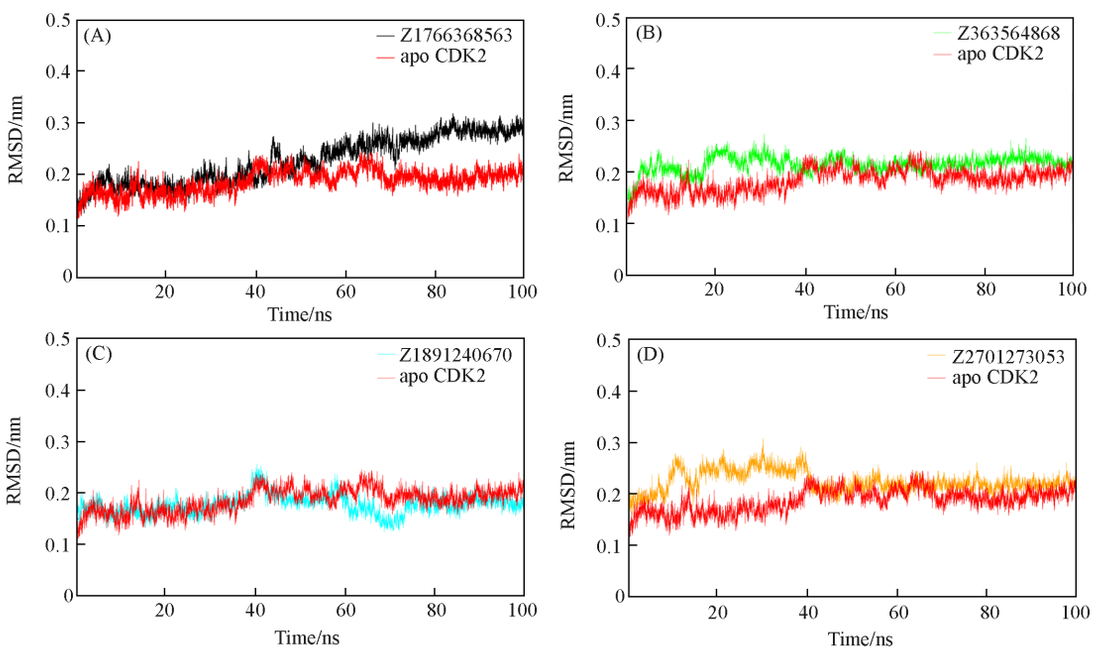

Fig. 6 RMSD for systems Z1766368563-CDK2(A), Z363564868-CDK2(B), Z1891240670-CDK2(C) and Z2701273053-CDK2(D)

在较短的时间内就达到了动力学平衡状态，而图6(A)在最后20 ns才逐渐达到平衡，且平衡状态时RMSD值最大，约为0.3 nm，表明Z1766368563配体的初始构象与CDK2结合模式不稳定，在最后20 ns才达到稳定状态.

采用MM-PBSA方法计算了4个体系的结合自由能(表4和图S2，见本文支持信息). 结果表明，蛋白质与配体之间的库仑力与溶剂化作用相反，表明真空(ΔE_{ele})和溶剂(ΔG_{PB})中静电作用贡献之和不利于配体与蛋白质结合. 然而，范德华能量(ΔE_{vdW})和非极性溶剂化能量(ΔG_{SA})之和的贡献对每种抑制剂是有利的. Z2701273053配体在CDK2受体中的结合自由能最低(ΔG_{bind}为−132.60 kJ/mol)，其中，主要的能量贡献是非极性相互作用，且范德华相互作用占主导地位. Z1891240670配体在CDK2受体中结合能较大，为−64.63 kJ/mol，由于体系的极性相互作用能较大，导致配体与蛋白之间的结合自由能降低. Z363564868配体与CDK2受体结合自由能为−81.42 kJ/mol，该配体的熵贡献较大，表明配体在与受体结合时构象变化较小，配体的构象适合受体口袋，因此，配体的结合自由能较低.

Table 4　Binding free energy of the four complex systems

System	ΔE_{ele}/ (kJ·mol^{-1})	ΔE_{vdW}/ (kJ·mol^{-1})	ΔG_{PB}/ (kJ·mol^{-1})	ΔG_{SA}/ (kJ·mol^{-1})	ΔG_e/ (kJ·mol^{-1})	ΔG_v/ (kJ·mol^{-1})	$T\Delta S$/ (kJ·mol^{-1})	ΔG_{bind}/ (kJ·mol^{-1})
Z1766368563-CDK2	−29.73	−200.10	136.64	−26.58	107.06	−226.68	−25.31	−94.31
Z1891240670-CDK2	−16.50	−197.78	145.86	−21.99	129.36	−219.77	−22.70	−64.63
Z363564868-CDK2	−49.71	−221.21	172.86	−27.24	123.15	−248.45	−40.61	−81.42
Z2701273053-CDK2	−39.14	−247.72	160.61	−25.15	121.47	−272.86	−18.79	−132.60

为了深入了解4种配体与激酶的选择性，将总结合能进一步分解为配体与残基之间的相互作用能，以评价单个残基对配体与蛋白质活性位点结合的紧密程度. 表S4(见本文支持信息)总结了单个残基贡献的相互作用能，相关残基位置如图S3(见本文支持信息)所示. 通过单个残基能量分解数据分析，Z1766368563配体的选择性主要由Ile10，Val18，Phe80，Asp145贡献；Z1891240670配体的选择性主要来自Ile10，Val18，Asp86，Leu134的贡献；Z363564868配体的选择性主要来自Ile10，Val18，Phe80，Asp86，Lys89，Leu134，Asp145的贡献；Z2701273053配体的选择性主要来自Lys33，Phe80，Leu134，Asp145的贡献.

通过监测配体与CDK2的氢键结合能力可以给出其相应蛋白亲和力的直接信息. 使用VMD监测所有可能出现的氢键的占有率. 由表5可见，虽然Z1766368563配体在模拟过程中与Val64，Lys33和Asp145残基形成了氢键，但是氢键的占有率较低(12%～19%)，表明生成的氢键并不稳定. Z2701273053配体与Asp145残基形成较稳定的氢键，占有率为74.61%. Z363564868配体与Asp86残基形成了稳定的氢键，占有率高达94.30%. Z1891240670配体与Glu12残基形成了较稳定的氢键，占有率为67.23%. 通过氢键分析，推测配体在CDK2中的高活性可能与Asp145残基相互作用有一定的关系.

Table 5　Hydrogen bond occupancy in simulation time

Compound	HBD	HBA	Occupancy(%)
Z1766368563	Lys33-Side-NZ	Mol299-Side-O2	18.99
	Asp145-Main-N	Mol299-Side-O2	17.59
	Mol299-Side-O3	Val64-Main-O	12.44
Z2701273053	Asp145-Main-N	Mol299-Side-O1	74.61
	Lys33-Side-NZ	Mol299-Side-O2	28.97
	Ile52-Main-CA	Mol299-Side-F	21.44
Z363564868	Mol299-Side-N3	Asp86-Side-OD1	94.30
	Mol299-Side-N3	Asp86-Side-CG	32.38
	Glu12-Main-N	Mol299-Side-O1	11.34
Z1891240670	Glu12-Main-N	Mol299-Side-N2	67.23

图S4(见本文支持信息)示出了4种配体与CDK2口袋氨基酸的相互作用. 图S4(A)示出了Z1766368563配体在CDK2受体中的作用模式, Asp145与配体中的呋喃环和羰基的氧元素形成了氢键, Phe80与呋喃结构和苯环形成了π-π键, 这与残基能量分解中Asp145与Phe80中范德华能量占比高一致.

由图S4(B)可见, Z363564868配体与Asp86和Glu12形成了氢键相互作用, 与Lys89形成了阳离子-π相互作用. Z363564868配体与CDK2口袋处的疏水性残基形成了丰富的疏水相互作用. 结合残基能量分解结果可知, 口袋处的关键氨基酸均对结合自由能提供了一定的贡献. 由图S4(C)可见, Glu12和Asp86与Z1891240670配体的噁二唑环分别形成了氢键和阴离子-π相互作用, Glu12与配体稠环结构额外形成了阴离子-π相互作用, 与残基能量分解结果中Glu12范德华能量贡献大的现象一致. 推测Z2701273053配体在4个化合物中具有最低的结合自由能的原因是具有噁二唑环与吡啶结构相连的结构(图S5, 见本文支持信息). 通过MD模拟后的相互作用分析[图S4(D)], 这个指纹片段与Lys33形成了氢键相互作用, 与Phe80和Asp145形成疏水相互作用. 并且通过残基能量分解得出这3种残基的范德华相互作用能(ΔG_v)之和占所有关键残基的ΔG_v之和的64.74%(表S4), 表明此结构为Z2701273053中的关键子结构. 这一结果与模型预测Z2701273053的1631位指纹片段为活性片段相符.

3 结　论

使用传统机器学习和图模型对CDK2建立二分类模型. 结果表明, 模型平均预测水平最优是基于ECFP6的XGBoost, 验证集与测试集的AUC都达到了90%以上. RF和图模型(Attentive FP)也有良好的预测效果. 模型建立完成后, 使用SHAP来解释分子指纹在RF和XGBoost模型的重要性, 并且利用可解释性的图模型在原子层面上对分子的活性做出解释, 来发现提高CDK2抑制剂活性的结构片段(如1476, 1292和1582指纹). 从分析结果中得到的分子指纹解释与原子权重解释一致. 在模型对已有活性化合物做出合理的解释后, 使用XGBoost模型在Enamine数据库中虚拟筛选出1152个化合物. 将筛选出的化合物与CDK2受体进行分子对接. 为了避免从打分排名中选取的化合物结构单一, 将1152个化合物通过指纹片段进行聚类. 根据聚类指标, 找到聚类的最优类别数(4类), 根据对接评分在每种类别中挑选出一个代表性化合物. 使用分子动力学模拟对4种配体进行测试, 结果表明, 4种配体在CDK2受体中结合稳定, 并在结合口袋处形成了丰富的范德华相互作用. 本文研究结果有助于CDK2抑制剂后续的改造和优化, 得到活性更好的CDK2抑制剂.

支持信息见http://www.cjcu.jlu.edu.cn/CN/10.7503/cjcu20240442.

参　考　文　献

[1] Swaffer M. P., Jones A. W., Flynn H. R., Snijders A. P., Nurse P., *Cell*, **2016**, *167*(7), 1750—1761
[2] Arellano M., Moreno S., *Int. J. Biochem. Cell. Biol.*, **1997**, *29*(4), 559—573
[3] Matsuura I., Denissova N. G., Wang G., He D., Long J., Liu F., *Nature*, **2004**, *430*(6996), 226—231
[4] Matsumoto Y., Hayashi K., Nishida E., *Curr. Biol.*, **1999**, *9*(8), 429—432
[5] Pagano M., Pepperkok R., Lukas J., Baldin V., Ansorge W., Bartek J., Draetta G., *J. Cell. Biol.*, **1993**, *121*(1), 101—111
[6] Hu S., Danilov A. V., Godek K., Orr B., Tafe L. J., Rodriguez-Canales J., Behrens C., Mino B., Moran C. A., Memoli V. A., Mustachio L. M., *Cancer Res.*, **2015**, *75*(10), 2029—2038
[7] Tetsu O., Mccormick F., *Cancer Cell*, **2003**, *3*(3), 233—245
[8] Faber A. C., Chiles T. C., *Cell Cycle*, **2007**, *6*(23), 2982—2989
[9] Karst A. M., Jones P. M., Vena N., Ligon A. H., Liu J. F., Hirsch M. S., Etemadmoghadam D., Bowtell D. D., Drapkin R., *Cancer Res.*, **2014**, *74*(4), 1141—1152
[10] Keck J. M., Summers M. K., Tedesco D., Ekholm-Reed S., Chuang L. C., Jackson P. K., Reed S. I., *J. Cell. Biol.*, **2007**, *178*(3), 371—385
[11] Sonntag R., Giebeler N., Nevzorova Y. A., Bangen J. M., Fahrenkamp D., Lambertz D., Haas U., Hu W., Gassler N., Cubero F. J., Müller-Newen G., *Proc. Natl. Acad. Sci.*, **2018**, *115*(37), 9282—9287

[12] Tadesse S., Caldon E. C., Tilley W., Wang S., *J. Med. Chem.*, **2019**, *62*(9), 4233—4251
[13] Tadesse S., Anshabo A. T., Portman N., Lim E., Tilley W., Caldon C. E., Wang S., *Drug Discov. Today*, **2020**, *25*(2), 406—413
[14] Beale G., Haagensen E. J., Thomas H. D., Wang L. Z., Revill C. H., Payne S. L., Golding B. T., Hardcastle I. R., Newell D. R., Griffin R. J., Cano C., *Br. J. Cancer*, **2016**, *115*(6), 682—690
[15] Cheng C. K., Gustafson W. C., Charron E., Houseman B. T., Zunder E., Goga A., Gray N. S., Pollok B., Oakes S. A., James C. D., Shokat K. M., *Proc. Natl. Acad. Sci.*, **2012**, *109*(31), 12722—12727
[16] Tripathi S. K., Muttineni R., Singh S. K., *J. Theor. Biol.*, **2013**, *334*, 87—100
[17] Rogers D., Hahn M., *J. Chem. Inf. Model*, **2010**, *50*(5), 742—754
[18] Belkina A. C., Ciccolella C. O., Anno R., Halpert R., Spidlen J., Snyder-Cappione J. E., *Nat. Commun.*, **2019**, *10*(1), 5415
[19] Kobak D., Berens P., *Nat. Commun.*, **2019**, *10*(1), 5416
[20] Su X., Bai M., *PLoS One*, **2020**, *15*(8), e0238000
[21] Sheridan R. P., Wang W. M., Liaw A., Ma J., Gifford E. M., *J. Chem. Inf. Model.*, **2016**, *56*(12), 2353—2360
[22] Lundberg S. M., Erion G., Chen H., DeGrave A., Prutkin J. M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S. I., *Nat. Mach. Intell.*, **2020**, *2*(1), 56—67
[23] Rodríguez-Pérez R., Bajorath J., *J. Med. Chem.*, **2020**, *63*(16), 8761—8777
[24] Trott O., Olson A. J., *J. Comput. Chem.*, **2010**, *31*(2), 455—461
[25] Daina A., Michielin O., Zoete V., *Sci. Rep.*, **2017**, *7*(1), 42717
[26] Xiong G., Wu Z., Yi J., Fu L., Yang Z., Hsieh C., Yin M., Zeng X., Wu C., Lu A., Chen X., *Nucleic Acids Res.*, **2021**, *49*(W1), W5—W14
[27] Frisch M. J., Trucks G. W., Schlegel H. B., Scuseria G. E., Robb M. A., Cheeseman J. R., Scalmani G., Barone V., Petersson G. A., Nakatsuji H., Li X., Caricato M., Marenich A. V., Bloino J., Janesko B. G., Gomperts R., Mennucci B., Hratchian H. P., Ortiz J. V., Izmaylov A. F., Sonnenberg J. L., Williams-Young D., Ding F., Lipparini F., Egidi F., Goings J., Peng B., Petrone A., Henderson T., Ranasinghe D., Zakrzewski V. G., Gao J., Rega N., Zheng G., Liang W., Hada M., Ehara M., Toyota K., Fukuda R., Hasegawa J., Ishida M., Nakajima T., Honda Y., Kitao O., Nakai H., Vreven T., Throssell K., Montgomery J. A., Peralta J. E., Ogliaro F., Bearpark M. J., Heyd J. J., Brothers E. N., Kudin K. N., Staroverov V. N., Keith T. A., Kobayashi R., Normand J., Raghavachari K., Rendell A. P., Burant J. C., Iyengar S. S., Tomasi J., Cossi M., Millam J. M., Klene M., Adamo C., Cammi R., Ochterski J. W., Martin R. L., Morokuma K., Farkas O., Foresman J. B., Fox D. J., *Gaussian* 16, *Revision C*.01, Gaussian Inc., Wallingford CT, **2016**
[28] Tirado-Rives J., Jorgensen W. L., *J. Chem. Theory. Comput.*, **2008**, *4*(2), 297—306
[29] Olsson M. H., Søndergaard C. R., Rostkowski M., Jensen J. H., *J. Chem. Theory. Comput.*, **2011**, *7*(2), 525—537
[30] Case D. A., Cheatham III T. E., Darden T., Gohlke H., Luo R., Merz Jr K. M., Onufriev A., Simmerling C., Wang B., Woods R. J., *J. Comput. Chem.*, **2005**, *26*(16), 1668—1688
[31] He X., Man V. H., Yang W., Lee T. S., Wang J., *J. Chem. Phys.*, **2020**, *153*(11), 114502
[32] Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W., Klein M. L., *J. Chem. Phys.*, **1983**, *79*(2), 926—935
[33] Hess B., Bekker H., Berendsen H. J., Fraaije J. G., *J. Comput. Chem.*, **1997**, *18*(12), 1463—1472
[34] Darden T. A., York D. M., Pedersen L. G., *J. Chem. Phys.*, **1993**, *98*(12), 10089—10092
[35] Kutzner C., Páll S., Fechner M., Esztermann A., de Groot B. L., Grubmüller H., *J. Comput. Chem.*, **2019**, *40*(27), 2418—2431
[36] Humphrey W., Dalke A., Schulten K., *J. Mol. Graph.*, **1996**, *14*(1), 33—38
[37] DeLano W. L, *Protein Crystallogr.*, **2002**, *40*(1), 82—92
[38] Dai Y., Wang Q., Zhang X., Jia S., Zheng H., Feng D., Yu P., *Eur. J. Med. Chem.*, **2010**, *45*(12), 5612—5620
[39] Sun H., Duan L., Chen F., Liu H., Wang Z., Pan P., Zhu F., Zhang J. Z., Hou T., *Phys. Chem. Chem. Phys.*, **2018**, *20*(21), 14450—14460
[40] Lundberg S. M., Erion G., Chen H., DeGrave A., Prutkin J. M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S. I., *Nat. Mach. Intell.*, **2020**, *2*(1), 56—67
[41] Rodríguez-Pérez R., Bajorath J., *J. Med. Chem.*, **2020**, *63*(16), 8761—8777

(Ed.: Y, K, M)