doi: 10.13241/j.cnki.pmb.2024.04.002

基于机器学习的环氧合酶 -2 抑制剂分类模型的构建 *

萧耿苗 穆云萍 千爱君 李芳红 赵子建△

（广东工业大学生物医药学院 广东 广州 510006）

摘要 目的：构建环氧合酶 -2（Cyclooxygenase-2，COX-2）抑制剂分类模型，用以筛选和优化 COX-2 抑制剂。**方法：**基于八种机器学习算法构建模型，比较不同模型的预测性能，筛选出最优模型后利用 Y 随机验证法对其进行测试，最后运用 SHAP（Shapley Additive eXplanation）算法对最优模型进行可解释性分析。**结果：**八种不同模型的性能比较结果显示，基于随机森林算法建立的模型最优，其预测准确率、平衡准确率、马修斯相关系数、特征曲线下面积和 F1 分数（分别为 0.893、0.825、0.673、0.909 和 0.933）最高；Y 随机验证结果表明最优模型的预测结果并非偶然；此外，通过 SHAP 算法挖掘出 20 个最有可能影响 COX-2 抑制剂活性的结构片段。**结论：**本研究为新型 COX-2 抑制剂的开发提供理论依据，可供本领域其他研究人员对先导化合物进行优化或设计更好的 COX-2 抑制剂。

关键词：COX-2 抑制剂；机器学习；可解释性；重要结构片段

中图分类号：R-33；Q55；Q811；R914 **文献标识码：**A **文章编号：**1673-6273（2024）04-606-06

Construction of a Classification Model for Cyclooxygenase-2 Inhibitors based on Machine Learning*

XIAO Geng-miao, MU Yun-ping, QIAN Ai-jun, LI Fang-hong, ZHAO Zi-jian△

(School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China)

ABSTRACT Objective: This study aims to develop a classification model for cyclooxygenase-2 (COX-2) inhibitors for the purpose of screening and optimizing COX-2 inhibitors. **Methods:** Eight machine learning algorithms were used to construct models, and their predictive performance was compared to identify the best model. The optimal model was tested by using Y-scrambling validation method, finally the interpretability analysis of the optimal model was performed by using Shapley Additive eXplanation (SHAP) algorithm. **Results:** Among the eight different models compared, the Random Forest algorithm exhibited the best performance. With the highest accuracy, balanced accuracy, Matthew's correlation coefficient, area under the ROC curve, and F1 scores (0.893, 0.825, 0.673, 0.909 and 0.933, respectively), it comes out on top. Validation with Y-scrambling showed that the predictions of the optimal model were not coincidence. Moreover, the SHAP algorithm was used to mine 20 structural fragments that could affect COX-2 inhibitor activity. **Conclusions:** In this study, we developed a theoretical basis for developing COX-2 inhibitors, which is useful to other researchers in this field when optimizing lead compounds and designing new COX-2 inhibitors.

Key words: COX-2 inhibitors; Machine learning; Interpretation; Important structural fragments

Chinese Library Classification (CLC): R-33; Q55; Q811; R914 Document code: A

Article ID: 1673-6273(2024)04-606-06

前言

环氧合酶 -2（Cyclooxygenase-2，COX-2）是机体合成前列腺素的关键限速酶之一，可催化花生四烯酸转化为前列腺素[1]。在细胞中，两个紧密相连的 COX-2 单体组成同源二聚体，并通过四螺旋膜锚结构域固定在内质网膜和细胞核膜上[2]。生理状态下 COX-2 表达量较低，但人体在受到创伤或感染时会产生大量 COX-2，引发炎症并产生疼痛信号[3]。COX-2 抑制剂适用于减轻由多种疾病引起的轻度至中度疼痛或炎症，以及运动损伤引起的急性疼痛[4]。尽管已研制出多种 COX-2 抑制剂，如塞来昔布、罗非昔布、伐地考昔、依托考昔和罗美昔布等，但真正可安全使用的药物却很少[3,4]。罗非昔布和伐地考昔因存在增加心血管疾病的风险，已分别于 2004 年和 2005 年被停用。截至目前，依托考昔和罗美昔布尚未获得美国 FDA 批准上市，可用的药物仅有塞来昔布[3-5]。因此，需要多途径探索研发安全可用的新型 COX-2 抑制剂，以缓解临床上药物缺乏的困境。

利用传统生物学实验筛选 COX-2 抑制剂，研发周期长，且成本巨大，不利于新药的开发。当前，机器学习算法在新药研发中的应用越来越受到人们的关注，可显著减少筛选实验数量，并提供可用信息，加快新药开发进程。然而，没有任何一种机器

* 基金项目：国家重点研发计划项目（2018YFA0800603）；广东省 " 珠江人才计划 " 项目（2016ZT06Y432）；
广东省重点领域研发计划项目（2019B020201015）

作者简介：萧耿苗（1994-），男，博士研究生，主要研究方向：内分泌，E-mail: 2111706036@mail2.gdut.edu.cn

△ 通讯作者：赵子建（1962-），男，博士生导师，教授，主要研究方向：生殖、肥胖、代谢性疾病和创新药物研发，E-mail: azzhao@gdut.edu.cn

（收稿日期：2023-08-18 接受日期：2023-09-13）

学习算法能在所有任务中都取得最佳的预测性能。使用多种机器学习算法构建模型，并从中选出性能最优的模型，是当前新药开发的重要策略[6,7]。在浅层机器学习算法中，使用朴素贝叶斯（naïve Bayesian，NB）[8]、k 近邻（k-nearest neighbor，KNN）[9]、随机森林（random forests，RF）[10]和支持向量机（support vector machines，SVM）[11] 算法构建分类模型的频率明显高于其他算法，取得最优性能的模型也往往是基于这四种算法构建的[12]。在多个二分类任务中，使用深度神经网络（Deep Neural Networks，DNN）[13]、图注意力网络（Graph Attention Network，GAT）[14]、消息传递神经网络（Message Passing Neural Network，MPNN）[15]和基于注意力机制的图神经网络（Attentive FP）[16]四种深度学习算法构建的模型，同样能够取得优异性能[6,17]。已有研究大多聚焦于对 COX-2 抑制剂模型的开发，而对影响抑制剂活性的结构片段的研究却鲜有报道[18-20]。基于此，本研究利用公共数据库的测试数据构建八种 COX-2 抑制剂分类模型，对不同模型的预测性能进行比较，筛选出最优模型，利用 Y 随机验证法对其展开进一步测试；此外，利用 SHAP 算法对最优模型进行可解释性分析，进而挖掘出 COX-2 抑制剂的优势和劣势片段，以期为本领域科研人员开发或设计出活性更好的 COX-2 抑制剂提供依据。

1 材料与方法

1.1 材料

本研究从公开数据库 ChEMBL[21]中获取以人源 COX-2 为靶点的测试数据。采用以下四个步骤对数据进行优化：（1）仅保留活性测试类型的数据；（2）仅保留 COX-2 抑制剂的测试数据；（3）保留具有明确 IC50、Ki 测定值的化合物，并移除重复测量的数据；（4）去除分子量大于 1000 Da 的化合物。最终得到由 3289 个小分子化合物组成的数据集。测定值小于 10 μM 的化合物归类为有抑制活性（标签记为 "1"），其余归类为没有抑制活性（标签记为 "0"）[17]。数据集按 8∶1 的比例划分为训练集和测试集。

1.2 模型的构建和评估

1.2.1 模型的构建 使用八种机器学习算法构建模型，包括四种浅层机器学习算法：NB、KNN、RF 和 SVM，以及四种深度学习算法：DNN、GAT、MPNN 和 Attentive FP。通过使用 Scikit-learn 算法库构建 NB、KNN、RF 和 SVM 模型，使用 DeepChem 深度学习框架构建 DNN、GAT、MPNN 和 Attentive FP 模型。此外，采用网格搜索优化每个模型的超参数。

1.2.2 模型性能的评估 在药物发现领域，常用于评估分类模型性能的指标包括预测准确率（Accuracy，ACC）、平衡准确率（Balanced accuracy，BA）、马修斯相关系数（Matthew's correlation coefficient，MCC）、受试者工作特征曲线下面积（Area under the curve，AUC）和 F1 分数（F1 score）[17]。指标值越大，模型性能越好。在本研究中，利用测试集评估模型的预测性能，ACC、BA、MCC、AUC 和 F1 分数最高的模型为最优模型。

1.2.3 Y 随机验证 本研究使用 Y 随机验证对最优模型做进一步测试。该方法通过随机打乱原训练集的标签，使用与最优模型相同的超参数训练模型并在原测试集上评估其性能；或者随机打乱原测试集的标签，利用打乱的测试集对最优模型进行评估。如果发现 AUC 和 MCC 分数明显下降，则可以确信最优模型不是偶然的[22,23]。在本研究中，训练集和测试集的标签各随机打乱 500 次。

1.3 模型的可解释性

SHAP 算法[24]是近年来开发的一种基于博弈论的机器学习模型解释工具，能够分析模型做出相应预测的判断依据。本研究利用 SHAP 算法对最优模型进行可解释性分析，通过将所有化合物的结构片段视为贡献者，计算每个结构片段的贡献值（SHAP 值），从而解释模型如何判断小分子化合物是否具有 COX-2 抑制活性。SHAP 值越大的结构片段越有可能影响小分子化合物的抑制活性。

2 结果

2.1 模型性能的比较

八种模型的预测性能如表 1 所示。在四种浅层机器学习模型中，RF 模型的各项指标最高，ACC、BA、MCC、AUC 和 F1 分数分别为 0.893、0.825、0.673、0.909 和 0.933。KNN 模型具有和 RF 模型相同的 ACC 和 F1 分数，但其 BA、MCC 和 AUC 分数低于 RF 模型。在四种深度学习模型中，Attentive FP 模型的 ACC 和 MCC 分数最高（分别为 0.89 和 0.806），DNN 模型的 BA 分数最高（为 0.821），GAT 模型的 AUC 和 F1 分数最高（分别为 0.902 和 0.927）。深度学习能够通过构建多个隐层，从而提高模型的性能[25,26]，但在本研究中深度学习模型并没有表现出更高的预测性能，四种深度学习模型的各项指标均低于 RF模型。综上所述，基于随机森林算法构建的 RF 模型为最优模型。

表 1 八种模型在测试集上的性能

Table 1 Performance of 8 models in the test set

	Models	ACC	BA	MCC	AUC	F1
Shallow learning	NB	0.749	0.742	0.416	0.812	0.825
	SVM	0.883	0.799	0.634	0.88	0.927
	KNN	0.893	0.82	0.67	0.902	0.933
	RF	0.893	0.825	0.673	0.909	0.933
Deep learning	MPNN	0.866	0.740	0.566	0.859	0.653
	GAT	0.882	0.783	0.629	0.902	0.927
	DNN	0.885	0.821	0.655	0.899	0.801
	Attentive FP	0.890	0.806	0.660	0.886	0.764

2.2 模型的测试

本研究使用 Y 随机验证法对最优模型做进一步评估，结果如图 1 所示。打乱训练集后模型的 MCC 分数下降至 0.005，AUC 分数下降至 0.495，其分类准确性仅相当于随机分类[27]。打乱测试集后，RF 模型的 MCC 分数下降至 0.003，AUC 分数下降至 0.502，表明基于随机森林算法建立的 RF 模型具有可靠性，其预测结果不是偶然的。

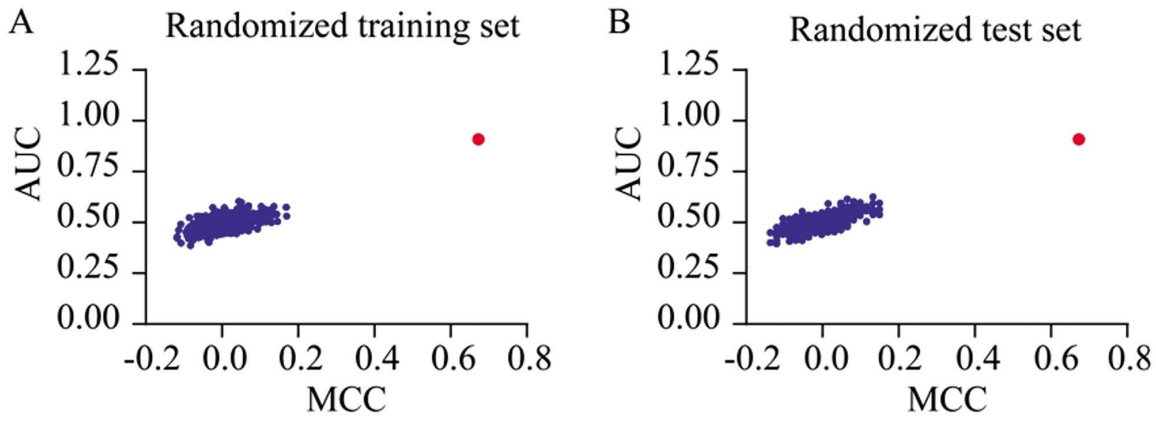

图 1 Y 随机验证

Fig.1 Y-scrambling validation

Note: Blue dots represent the performance of the model after disrupting the training or test set; Red dot represents the performance of the RF model.

2.3 模型可解释性分析

机器学习的 " 黑盒效应 " 导致模型难以解释，一旦预测结果出现错误，无法分析原因并进行修正，不利于更深入的研究[28]。本研究运用 SHAP 算法解释性能最优的 RF 模型，前 20 个重要结构片段如图 2 所示。SHAP 值集中在正值部分表示该结构为优势片段，集中在负值则为劣势片段 [17]。结果显示，Morgen_319、816、715、216、45、659、173、136、378、828、350、896、801、233、833、361、471 这 17 个结构片段的红点集中在正值部分，表明具有这些结构片段的小分子化合物更可能是 COX-2 抑制剂；Morgen_168、175、389 的红点集中在负值部分，表明携带这三个结构片段的小分子化合物可能不具备 COX-2 抑制作用。

由图 3 可知，优势片段中大多含有磺酰胺基团及其衍生基团（如 Morgen_319、715、173、350、833），这可能与磺酰基能够和 COX-2 结合口袋内的精氨酸 Arg513 形成氢键相互作用有关[3]；含氟原子或氯原子的结构片段也是 COX-2 抑制剂的优势片段，如 Morgen_816、216、46、659；此外，氮原子也出现在多个优势片段中。以当前 COX-2 抑制剂为例，塞来昔布、罗非昔布、伐地考昔和依托考昔均不包含 Morgen_168、175、389 这三个劣势结构片段，但包含多个优势结构片段，尤其是磺酰胺基团及其衍生基团，见图 4A-D。根据 RF 模型，这四种药物均被预测为 COX-2 抑制剂，与实验结果一致。

3 讨论

本研究基于机器学习算法构建了八种 COX-2 抑制剂分类模型，其中利用随机森林算法构建的 RF 模型为最优模型且具有可靠性。通过 SHAP 算法对 RF 模型进行可解释性分析，挖掘出最重要的 20 个结构片段，其中优势片段大多为磺酰胺基团及其衍生基团，以及含有氟原子、氯原子或氮原子的结构片段，为新型 COX-2 抑制剂的研发提供支持。

深度学习模型在众多领域中表现出优异的性能，但需要庞大的数据支撑才能充分发挥其性能[29-31]。在本研究中，深度学习模型性能稍差的原因可能是数据量太少，无法训练出优异性能的模型。近年来多个药物发现领域的研究也显示，基于随机森林算法构建的分类模型性能不弱于深度学习算法，且计算成本

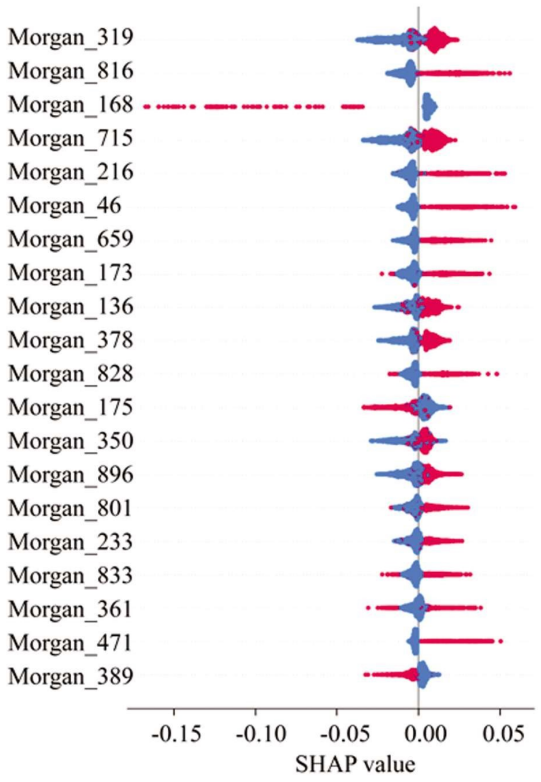

图 2 模型特征的 SHAP 树状图

Fig.2 SHAP tree diagram of model features

Note: Each row represents a structural fragment, each dot represents a sample, and the position of the dot represents the SHAP value. Red dots indicate that the sample contains this structural fragment, blue dots indicate that it does not.

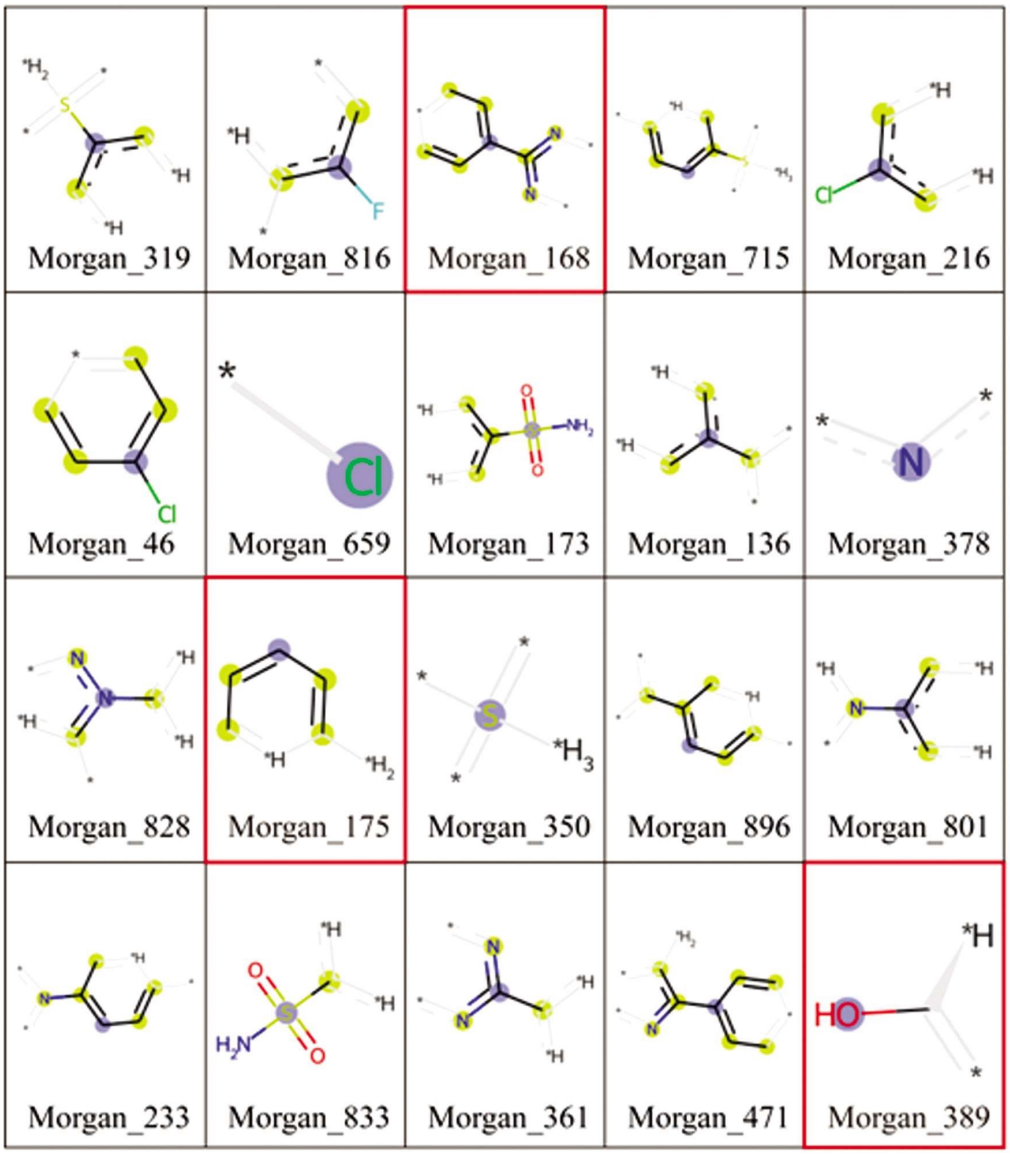

图 3 重要结构片段的二维结构图

Fig.3 Two-dimensional structure diagrams of important structural fragments

Note: Black boxes represent advantageous fragments; Red boxes represent disadvantageous fragments.

更低[6,17]。本文是首个利用 SHAP 算法挖掘 COX-2 抑制剂重要结构片段的研究，可用于优化化合物结构。目前，相关研究已证实磺酰胺和氯原子显著影响小分子化合物的 COX-2 抑制活性[32,33]，当小分子化合物的磺酰胺被羧基取代后，其 COX-2 抑制活性消失。此外，有研究报道在 COX-2 抑制剂中引入氯原子，有助于提高其抑制活性[32]，本研究也表明以 Morgen_216、46、659 三种结构片段的形式引入氯原子，更有可能提高 COX-2 抑制活性。

综上所述，本研究构建了一个准确可靠的 COX-2 抑制剂分类模型，并挖掘出 20 个重要的结构片段，不仅有助于筛选潜在的 COX-2 抑制剂，也可用于优化 COX-2 抑制剂结构，通过去除劣势结构片段或在合适的位置添加优势结构片段，得到抑制效果更好的小分子化合物。值得注意的是，不同模型的适用范围、最佳运行条件、以及对数据量的要求，在实际使用时均影响其预测结果，也是判断模型优劣的关键因素。因此，多算法筛选构建最优模型将助力新药研发。

参 考 文 献(References)

[1] Mitchell J A, Kirkby N S, Ahmetaj-Shala B, et al. Cyclooxygenases and the cardiovascular system[J]. Pharmacol Ther, 2021, 217: 107624.

[2] Nina M, Berneche S, Roux B. Anchoring of a monotopic membrane protein: the binding of prostaglandin H2 synthase-1 to the surface of a phospholipid bilayer[J]. Eur Biophys J, 2000, 29(6): 439-454.

[3] Marnett L J. The COXIB experience: a look in the rearview mirror[J]. Annu Rev Pharmacol Toxicol, 2009, 49: 265-290.

[4] Stiller C O, Hjemdahl P. Lessons from 20 years with COX-2 inhibitors: Importance of dose-response considerations and fair play in comparative trials[J]. J Intern Med, 2022, 292(4): 557-574.

[5] Mcgettigan P, Henry D. Cardiovascular risk with non-steroidal anti-inflammatory drugs: systematic review of population-based controlled observational studies[J]. PLoS Med, 2011, 8(9): e1001098.

[6] Jiang D, Wu Z, Hsieh C Y, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models [J]. J Cheminform, 2021, 13(1): 12.

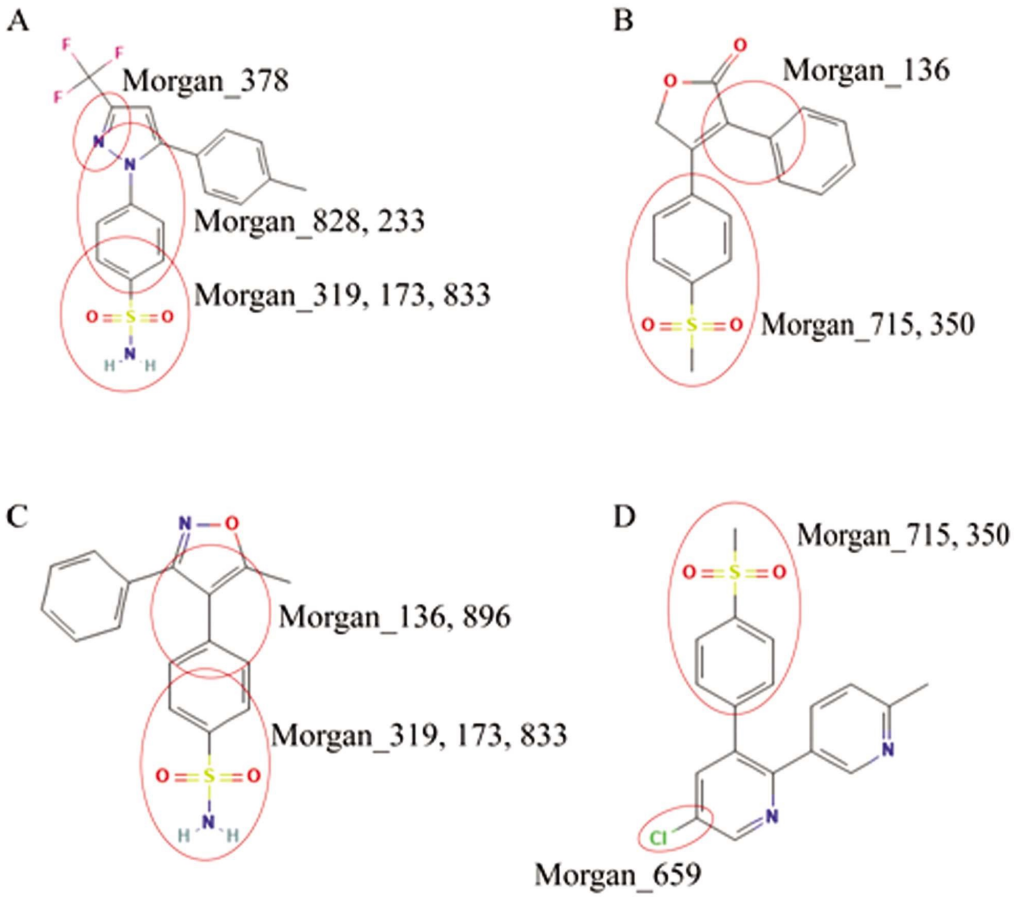

图 4 COX-2 抑制剂的二维结构图

Fig.4 Two-dimensional structure diagrams of COX-2 inhibitors

Note: A represents celecoxib; B represents rofecoxib; C represents vadicoxib; D represents etoricoxib; Red boxes indicate important structural fragments in the COX-2 inhibitor.

[7] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery[J]. Comput Struct Biotechnol J, 2021, 19: 4538-4558.

[8] Burstrom G, Edstrom E, Elmi-Terander A. Foundations of Bayesian Learning in Clinical Neuroscience [J]. Acta Neurochir Suppl, 2022, 134: 75-78.

[9] Zainuddin A Z A, Mansor W, Lee K Y, et al. Comparison of Extreme Learning Machine and K-Nearest Neighbour Performance in Classifying EEG Signal of Normal, Poor and Capable Dyslexic Children [J]. Annu Int Conf IEEE Eng Med Biol Soc, 2019, 2019: 4513-4516.

[10] Sipper M, Moore J H. Conservation machine learning: a case study of random forests[J]. Sci Rep, 2021, 11(1): 3629.

[11] Hao P Y, Chiang J H, Chen Y D. Possibilistic classification by support vector networks[J]. Neural Netw, 2022, 149: 40-56.

[12] Uddin S, Khan A, Hossain M E, et al. Comparing different supervised machine learning algorithms for disease prediction [J]. BMC Med Inform Decis Mak, 2019, 19(1): 281.

[13] Vogt M. Using deep neural networks to explore chemical space[J]. Expert Opin Drug Discov, 2022, 17(3): 297-304.

[14] Lv Q, Chen G, Yang Z, et al. Meta Learning With Graph Attention Networks for Low-Data Drug Discovery[J]. IEEE Trans Neural Netw Learn Syst, 2023[Epub ahead of print].

[15] Tang M, Li B, Chen H. Application of message passing neural networks for molecular property prediction[J]. Curr Opin Struct Biol, 2023, 81: 102616.

[16] Xiong Z, Wang D, Liu X, et al. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism[J]. J Med Chem, 2020, 63(16): 8749-8760.

[17] He S, Zhao D, Ling Y, et al. Machine Learning Enables Accurate and Rapid Prediction of Active Molecules Against Breast Cancer Cells[J]. Front Pharmacol, 2021, 12: 796534.

[18] Moussa N, Hassan A, Gharaghani S. Pharmacophore model, docking, QSAR, and molecular dynamics simulation studies of substituted cyclic imides and herbal medicines as COX-2 inhibitors [J]. Heliyon, 2021, 7(4): e06605.

[19] 李秉轲, 刘杰, 刘军, 等. COX-2 抑制剂的定量构效关系研究[J]. 广东化工, 2018, 45(15): 62-63+53.

[20] 艾上杰. 基于机器学习的 COX-2 抑制剂虚拟筛选方法研究[D]. 海南大学, 2018.

[21] Gaulton A, Bellis L J, Bento A P, et al. ChEMBL: a large-scale bioactivity database for drug discovery [J]. Nucleic Acids Res, 2012, 40(Database issue): D1100-7.

[22] Saini R, Agarwal S M. EGFRisopred: a machine learning-based classification model for identifying isoform-specific inhibitors against EGFR and HER2[J]. Mol Divers, 2022, 26(3): 1531-1543.

[23] Danishuddin, Kumar A, Mobeen F, et al. Development of Ligand and Structure-based classification models to design novel inhibitors against antibiotic hydrolyzing enzymes: Integration of web server[J]. Journal of Biomolecular Structure & Dynamics, 2018, 36 (11): 2966-2975.

[24] Bifarin O O. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification[J]. PLoS One, 2023, 18(5): e0284315.

[25] Sarker I H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions[J]. SN Comput Sci, 2021, 2(6): 420.

[26] Alzubaidi L, Zhang J L, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of Big Data, 2021, 8(1): 53.

[27] Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation [J]. Caspian J Intern Med, 2013, 4(2): 627-35.

[28] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods [J]. Entropy (Basel), 2020, 23(1): 18.

[29] Choudhary K, Decost B, Chen C, et al. Recent advances and applications of deep learning methods in materials science [J]. Npj Computational Materials, 2022, 8(1).

[30] Mousavi S M, Beroza G C. Deep-learning seismology [J]. Science, 2022, 377(6607): eabm4470.

[31] Krentzel D, Shorte S L, Zimmer C. Deep learning in image-based phenotypic drug discovery[J]. Trends Cell Biol, 2023, 33(7): 538-554.

[32] Li J J, Anderson G D, Burton E G, et al. 1,2-Diarylcyclopentenes as selective cyclooxygenase-2 inhibitors and orally active anti-inflammatory agents[J]. J Med Chem, 1995, 38(22): 4570-4578.

[33] Güngör T, Ozleyen A, Yılmaz Y B, et al. New nimesulide derivatives with amide/sulfonamide moieties: Selective COX-2 inhibition and antitumor effects[J]. Eur J Med Chem, 2021, 221: 113566.

（上接第 605 页）

[21] Erburu M, Muñoz-Cobo I, Diaz-Perdigon T, et al. SIRT2 inhibition modulate glutamate and serotonin systems in the prefrontal cortex and induces antidepressant-like action[J]. Neuropharmacology, 2017, 117: 195-208.

[22] Buzoglu HD, Burus A, Bayazıt Y, et al. Stem Cell and Oxidative Stress-Inflammation Cycle[J]. Curr Stem Cell Res Ther, 2023, 18(5): 641-652.

[23] Dos Santos JM, Rodrigues Lacerda AC, et al. Oxidative Stress Biomarkers and Quality of Life Are Contributing Factors of Muscle Pain and Lean Body Mass in Patients with Fibromyalgia [J]. Biology (Basel), 2022, 11(6): 935.

[24] Bruehl S, Milne G, Schildcrout J, et al. Perioperative oxidative stress predicts subsequent pain-related outcomes in the 6 months after total knee arthroplasty[J]. Pain, 2023, 164(1): 111-118.

[25] Ye S, Mahmood DFD, Ma F, et al. Urothelial Oxidative Stress and ERK Activation Mediate HMGB1-Induced Bladder Pain [J]. Cells, 2023, 12(10): 1440.

[26] Zhao M, Zhang X, Tao X, et al. Sirt2 in the Spinal Cord Regulates Chronic Neuropathic Pain Through Nrf2-Mediated Oxidative Stress Pathway in Rats[J]. Front Pharmacol, 2021, 12: 646477.