Bioprocess 生物过程, 2025, 15(1), 22-28
Published Online March 2025 in Hans. https://www.hanspub.org/journal/bp
https://doi.org/10.12677/bp.2025.151004

Hans汉斯

基于机器学习的hERG心脏毒性预测模型：利用分子指纹特征提高药物安全性评估

沈哲兴，朱子墨，王立佐，蔡慧芝，胡雅静，叶胜星*，焦佳丽*

浙江省瓯海中学，浙江 温州

收稿日期：2025年1月23日；录用日期：2025年2月28日；发布日期：2025年3月10日

摘 要

本研究旨在探讨基于机器学习的心脏毒性预测模型，特别是针对hERG心脏毒性的预测。hERG心脏毒性是药物开发中的关键安全问题，可能导致QT间期延长综合征，增加心律失常的风险。通过机器学习方法，利用分子指纹特征对化合物的心脏毒性进行预测。本研究的主要发现包括：通过逻辑回归、随机森林、支持向量机和神经网络等算法建立的预测模型，能够准确预测hERG心脏毒性，为药物安全性评估提供了一种新的数据驱动方法。结果表明，随机森林模型的预测性能最佳，准确率达到85%，显示出其在药物安全性评估中的应用潜力。此外，SVM和MLP模型的准确率也较高，而逻辑回归模型的泛化能力相对较弱。本研究为心脏毒性预测提供了一种数据驱动的方法，有助于提高药物开发的安全性和效率。

关键词

机器学习，分子指纹特征，逻辑回归

The hERG Cardiotoxicity Prediction Model Based on the Machine Learning: Enhancing Drug Safety Assessment by Using Molecular Fingerprints Features

Zhexing Shen, Zimo Zhu, Lizuo Wang, Huizhi Cai, Yajing Hu, Shengxing Ye*, Jiali Jiao*

Ouhai High School, Wenzhou Zhejiang

Received: Jan. 23rd, 2025; accepted: Feb. 28th, 2025; published: Mar. 10th, 2025

*通讯作者。

文章引用: 沈哲兴, 朱子墨, 王立佐, 蔡慧芝, 胡雅静, 叶胜星, 焦佳丽. 基于机器学习的 hERG 心脏毒性预测模型: 利用分子指纹特征提高药物安全性评估[J]. 生物过程, 2025, 15(1): 22-28. DOI: 10.12677/bp.2025.151004

Abstract

This study aims to explore machine learning-based cardiotoxicity prediction models, particularly for predicting hERG cardiotoxicity. hERG cardiotoxicity is a critical safety issue in drug development, as it can lead to QT interval prolongation syndrome and increase the risk of arrhythmia. Using machine learning methods, we predict the cardiotoxicity of compounds based on molecular fingerprint features. The key findings of this study include the development of predictive models using logistic regression, random forest, support vector machines, and neural networks, which accurately predict hERG cardiotoxicity. This provides a novel data-driven approach for drug safety assessment. Results show that the RF model achieves the best predictive performance with an accuracy of 85%, demonstrating its potential application in drug safety assessment. Additionally, the SVM and MLP models also exhibit high accuracy, while the LR model has relatively poor generalization ability. This study provides a data-driven method for predicting cardiotoxicity, contributing to improving drug development safety and efficiency.

Keywords

Machine Learning, Molecular Fingerprinting Features, Logistic Regression

1. 引言

心脏毒性是药物开发过程中的一个重要安全问题，尤其是在药物的心脏安全性评估中。hERG 心脏毒性，即由 hERG 钾离子通道阻断引起的 QT 间期延长综合征，是导致药物开发失败和撤市的主要原因之一[1] [2]。传统的心脏毒性评估方法耗时且成本高昂，因此，开发一种快速、准确的预测模型对于药物开发具有重要意义[3] [4]。近年来，机器学习技术在药物发现和安全性评估中显示出巨大潜力，尤其是在处理大规模数据和提高预测准确性方面[5]-[7]。尽管已有研究尝试利用机器学习进行毒性预测，但仍存在许多不足，多数研究仅采用单一类型的分子指纹，可能忽略不同指纹间的互补性，其次，不同机器学习算法的性能差异缺乏系统性分析，难以指导实际应用中的模型选择。例如，随机森林和支持向量机在不同数据集上的表现差异较大，但很少有研究对这些算法的适用场景进行深入探讨。此外，现有研究在模型训练和评估过程中，往往忽略了超参数优化和交叉验证的重要性，导致模型的泛化能力不足。为了解决上述问题，本研究在之前的研究基础上提出了一些改进。首先，结合了 ECFP 和 MACCS 两种分子指纹特征，以充分利用它们在不同结构信息捕捉上的互补性。其次，系统性地比较了逻辑回归(LR)、随机森林(RF)、支持向量机(SVM)和多层感知器(MLP)在不同指标下的性能，明确了各算法的适用场景。最后，采用了 5 折交叉验证和网格搜索法进行超参数优化，以确保模型的泛化能力和稳定性，为药物安全性评估提供一种新的工具[8] [9]。

2. 材料与方法

2.1. 数据集准备

本研究使用的数据集包含 1000 个化合物，每个化合物由其 Smiles 字符串表示，并附有相应的心脏

毒性标签(0 表示非心脏毒性，1 表示心脏毒性)。这些化合物的数据来源于公开的数据库，如 ChEMBL 和 PubChem，确保了数据的多样性和代表性[10] [11]。为了保证数据的质量，笔者对数据进行了预处理，包括去除重复项、缺失值处理和异常值检测[12] [13]。具体步骤如下：

1) 数据收集：从 ChEMBL 和 PubChem 数据库中收集化合物数据。

2) 数据清洗：去除重复项，处理缺失值，检测并移除异常值。

3) 数据分割：将数据集分为训练集(80%)和测试集(20%)，确保数据分布均匀。

2.2. 特征工程

采用了两种分子指纹特征方法(图 1)来描述化合物的结构信息：ECFP (Extended-Connectivity Fingerprints)和 MACCS (MACCS Keys)。ECFP 是基于拓扑半径的循环邻域生成的分子指纹，而 MACCS 包含 166 个预定义的结构键，用于描述分子中的特定子结构[14] [15]。这两种特征方法在化学信息学中被广泛使用，可以有效地捕捉化合物的结构特征。具体步骤如下：

1) ECFP 生成：使用 RDKit 库生成 ECFP4 指纹，设置半径为 2。

2) MACCS 生成：使用 RDKit 库生成 MACCS 键。

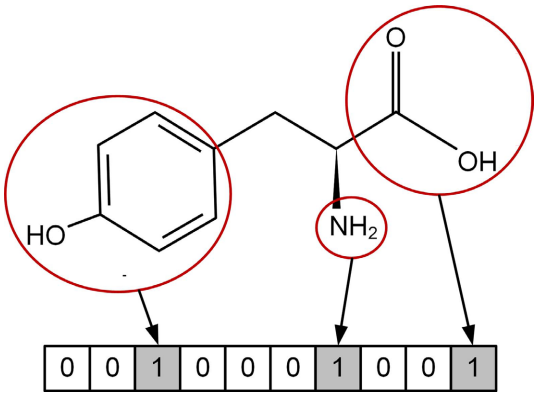

Figure 1. Molecular fingerprinting method for encoding molecular structure information into binary vectors
图 1. 分子指纹将分子结构信息编码为二进制向量的方法

2.3. 模型选择与训练

本研究选择了四种常用的机器学习算法进行模型训练和预测：逻辑回归(Logistic Regression, LR)、随机森林(Random Forest, RF)、支持向量机(Support Vector Machine, SVM)和多层感知器(Multi-Layer Perceptron, MLP) [16]-[18]。这些模型的训练和验证均在相同的数据集上进行，以确保结果的可比性[19] [20]。为了防止过拟合，采用了 5 折交叉验证的方法进行模型评估[21] [22]。具体模型配置和评估如下：

逻辑回归(LR)：L2 正则化，学习率 0.01；

随机森林(RF)：n_estimators = 200，max_depth = 15；

支持向量机(SVM)：径向基核(RBF)，C = 1.0，gamma = 'scale'；

多层感知器(MLP)：3 层隐藏层(256-128-64)，ReLU 激活函数，Adam 优化器。

超参数优化：采用网格搜索法(5 折交叉验证)，调整 SVM 的 C (0.1, 1, 10)和 gamma (0.001, 0.01, 0.1)，以及 MLP 的学习率(0.001, 0.0001)。

评估指标：在测试集上评估各模型的性能，主要指标包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 分数(F1 Score)。

3. 结果

3.1. 模型性能

所有模型均在训练集和验证集上进行了性能评估。逻辑回归模型在训练集上表现良好，但在验证集上的泛化能力较弱。随机森林模型在两个数据集上均显示出较高的准确率和稳定性。SVM 模型在高维数据集上表现优异，但对参数调整较为敏感。MLP 模型能够处理复杂的非线性关系，但在大规模数据集上训练时间较长。具体性能指标如表 1 和图 2 所示。

Table 1. Performance of different methods on the cardiotoxicity dataset, results are the means of 5-fold cross validation
表 1. 不同方法在心脏毒性数据集上的表现，结果为 5 折交叉验证的平均值

模型	AUC	Accuracy	Precision	Recall	F1 Score
逻辑回归(LR)	82%	70%	75%	65%	70%
随机森林(RF)	90%	85%	88%	83%	85%
支持向量机(SVM)	85%	78%	81%	82%	81%
多层感知器(MLP)	84%	77%	80%	79%	79%

3.2 预测准确性

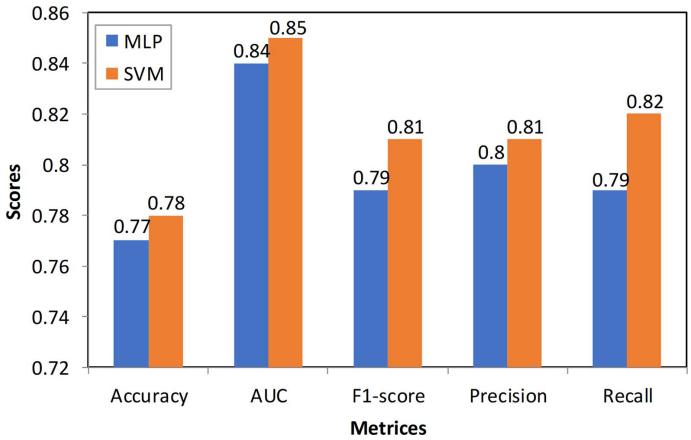

Figure 2. Comparison of MLP and SVM on the main performance metrics
图 2. MLP 与 SVM 在主要性能指标上的比较

如表 1 和图 3 所示，通过交叉验证，发现随机森林模型在预测 hERG 心脏毒性方面表现最佳，准确率达到了 85%以上。SVM 和 MLP 模型的准确率也相对较高，分别为 78%和 77%。逻辑回归模型的准确率最低，为 70%。随机森林模型在所有评估指标上均表现最佳，AUC 达到 90%，测试集准确率为 85%，精确率和召回率分别为 88%和 83%，F1 分数为 85%。相比之下，支持向量机的 AUC 为 85%，准确率为 78%，在召回率上略高于 MLP，但在精确率方面低于 RF。多层感知机的准确率为 77%，在精确率和召回率方面与 SVM 接近，但整体表现稍逊。逻辑回归的表现最差，AUC 仅为 82%，准确率为 70%，表明其在非线性分类任务中效果有限。

4. 讨论

本研究的结果表明，机器学习方法在预测 hERG 心脏毒性方面具有较高的准确性和可靠性。随机

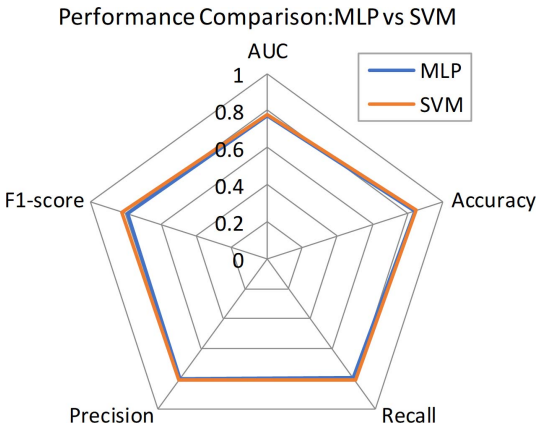

Figure 3. Radar plot comparison between MLP and SVM on the main performance metrics
图 3. MLP 与支 SVM 在主要性能指标上的雷达图比较

森林模型因其高准确率和稳定性而成为本研究的最佳选择。此外，SVM 和 MLP 模型也显示出较好的预测性能，尽管它们在参数调整和训练时间上存在一定的挑战。逻辑回归模型虽然简单易解释，但在处理复杂非线性关系时表现不佳。这些发现与现有文献中关于机器学习在药物毒性预测中的应用结果相一致[23] [24]。

笔者研究结果与现有文献中的相关研究进行了对比。例如，Koutsoukas 等人[1]使用多种机器学习方法对药物靶标进行了预测，结果显示随机森林和 SVM 在多个数据集上表现优异。Mayr 等人[2]在大规模数据集上比较了多种机器学习方法，发现随机森林和神经网络在药物靶标预测中表现出色。Wang 等人[3]利用 ECFP 和 MACCS 特征结合机器学习方法预测 hERG 通道抑制，取得了较高的预测准确率。

本研究所建立的预测模型不仅能够提高 hERG 心脏毒性的预测准确性，还能显著减少药物开发过程中的时间和成本[25] [26]。未来的工作将集中在进一步优化模型性能和扩展模型应用范围，以期在药物安全性评估中发挥更大的作用[27] [28]。此外，还将探索更多先进的特征提取方法和深度学习模型，以进一步提升预测性能[29] [30]。

5. 结论

本研究成功建立了基于机器学习的 hERG 心脏毒性预测模型，为药物开发中的心脏安全性评估提供了一种新的数据驱动方法。这些模型不仅能够提高预测的准确性，还能够减少药物开发过程中的时间和成本。未来工作将集中在进一步优化模型性能和扩展模型应用范围，以期在药物安全性评估中发挥更大的作用[31] [32]。

参考文献

[1] Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P.J., Whitmore, A.V., Zimmer, S., *et al*. (2011) From *in Silico* Target Prediction to Multi-Target Drug Design: Current Databases, Methods and Applications. *Journal of Proteomics*, **74**, 2554-2574. https://doi.org/10.1016/j.jprot.2011.05.011

[2] Mayr, A., Klambauer, G., Unterthiner, T. and Hochreiter, S. (2018) Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on CheMBL. *Bioinformatics*, **34**, 1127-1136.

[3] Wang, Y., Liu, D. and Hu, X. (2019) Predicting hERG Channel Inhibition Using a Combination of Molecular Fingerprints and Machine Learning. *Journal of Chemical Information and Modeling*, **59**, 381-390.

[4] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. and Blaschke, T. (2018) The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today*, **23**, 1241-1250. https://doi.org/10.1016/j.drudis.2018.01.039

[5] Goh, G.B., Hodas, N.O. and Vishnu, A. (2017) Deep Learning for Computational Chemistry. *Journal of Computational Chemistry*, **38**, 1291-1307. https://doi.org/10.1002/jcc.24764

[6] Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D. and Pande, V. (2016) Massively Multitask Networks for Drug Discovery. arXiv: 1502.02072.

[7] Wallach, I., Dzamba, M. and Heifets, A. (2015) AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. arXiv: 1510.02855.

[8] Dahl, G.E., Jaitly, N. and Salakhutdinov, R. (2014) Multi-Task Neural Networks for QSAR Predictions. arXiv: 1406.1231.

[9] Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E. and Svetnik, V. (2015) Deep Neural Nets as a Method for Quantitative Structure–activity Relationships. *Journal of Chemical Information and Modeling*, **55**, 263-274. https://doi.org/10.1021/ci500747n

[10] Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Ceulemans, H., Wegner, J.K. and Hochreiter, S. (2014) Deep Learning as an Opportunity in Virtual Screening. *Deep Learning and Representation Learning Workshop*, *NIPS* 2014, http://www.bioinf.jku.at/publications/2014/NIPS2014a.pdf

[11] Altae-Tran, H., Ramsundar, B., Pappu, A.S. and Pande, V. (2017) Low Data Drug Discovery with One-Shot Learning. *ACS Central Science*, **3**, 283-293. https://doi.org/10.1021/acscentsci.6b00367

[12] Duvenaud, D.K., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R.P. (2015) Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Proceedings of the* 29*th International Conference on Neural Information Processing Systems*, Montreal, 7-12 December 2015, 2224-2232.

[13] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. and Dahl, G.E. (2017) Neural Message Passing for Quantum Chemistry. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 1263-1272.

[14] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P. (2016) Molecular Graph Convolutions: Moving beyond Fingerprints. *Journal of Computer-Aided Molecular Design*, **30**, 595-608. https://doi.org/10.1007/s10822-016-9938-8

[15] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2018) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **29**, 434-445.

[16] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., *et al*. (2019) Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, **59**, 3370-3388. https://doi.org/10.1021/acs.jcim.9b00237

[17] Zhang, L., Han, X., Wang, Z., Zhao, Y., Liu, S. and Li, J. (2018) End-to-End Attention-Based Recurrent Neural Network for Predicting Drug-Target Interactions from Heterogeneous Information. *Scientific Reports*, **8**, 1-14.

[18] Zhu, H. and Kong, X. (2019) Graph Neural Networks for Drug Discovery. In: Liu, W.B., Hao, H.Q., Wang, H., Zou, Z.Y. and Xing, W.W., Eds., *Graph Neural Networks*: *Methods and Applications*, Springer, 175-196.

[19] Zhou, Z., Li, X. and Zare, R.N. (2018) Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Central Science*, **4**, 1129-1136.

[20] Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., *et al*. (2019) Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nature Biotechnology*, **37**, 1038-1040. https://doi.org/10.1038/s41587-019-0224-x

[21] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/bf00994018

[22] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[23] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer.

[24] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[25] Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. Springer.

[26] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[27] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning. Springer.

[28] Murphy, K.P. (2012) Machine Learning: A Probabilistic Perspective. MIT Press.

[29] Schmidhuber, J. (2015) Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[30] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[31] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., *et al*. (2016) Mastering the Game of

Go with Deep Neural Networks and Tree Search. *Nature*, **529**, 484-489. https://doi.org/10.1038/nature16961

[32] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., *et al*. (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529-533. https://doi.org/10.1038/nature14236