Hans Journal of Medicinal Chemistry 药物化学, 2025, 13(3), 229-238
Published Online August 2025 in Hans. https://www.hanspub.org/journal/hjmce
https://doi.org/10.12677/hjmce.2025.133024

Hans汉斯

基于机器学习的抗胰岛β细胞凋亡化合物分类模型研究

岳　悦，张　娜

北京工业大学化学与生命科学学院，北京

收稿日期：2025年3月4日；录用日期：2025年4月2日；发布日期：2025年8月12日

摘　要

糖尿病是一种以胰岛素抵抗和胰岛细胞功能衰竭为特征的慢性代谢性疾病，现有的胰岛素注射和口服降糖药治疗方法均无法从根本上逆转糖尿病的原发性病理机制，故针对胰岛β细胞受损干预，寻找抑制β细胞凋亡的药物将对糖尿病的治疗具有重大意义。基于已报道的103个具有抗胰岛β细胞凋亡活性化合物，结合6种分子指纹和7种机器学习方法对上述化合物构建了分类模型，并运用10倍交叉验证和测试集对模型性能进行评估。结果表明，基于PubChem指纹的随机森林算法所构建的模型表现最佳(AUC = 0.992、CA = 0.96和MCC = 0.901)。同时结合信息增益和子结构频率分析，识别出具有抗胰岛β细胞凋亡活性的9个特征子结构，如含氮芳香杂环、胺类和吡啶等高活性片段，为抗胰岛β细胞凋亡化合物的开发提供理论参考和指导。

关键词

糖尿病，抗胰岛β细胞凋亡，分类模型

Machine Learning-Based Classification Model of Anti-Pancreatic β-Cell Apoptosis Compounds

Yue Yue, Na Zhang

School of Chemistry and Life Sciences, Beijing University of Technology, Beijing

Received: Mar. 4th, 2025; accepted: Apr. 2nd, 2025; published: Aug. 12th, 2025

Abstract

Diabetes is a chronic metabolic disease characterized by insulin resistance and pancreatic β-cell

文章引用: 岳悦, 张娜. 基于机器学习的抗胰岛 β 细胞凋亡化合物分类模型研究[J]. 药物化学, 2025, 13(3): 229-238.
DOI: 10.12677/hjmce.2025.133024

dysfunction. Current treatments, including insulin injections and oral hypoglycemic drugs, cannot fundamentally reverse the primary pathological mechanisms of diabetes. Therefore, targeting interventions that protect pancreatic *β*-cells and identifying drugs that inhibit *β*-cell apoptosis are of significant importance for diabetes treatment. Based on 103 reported compounds with anti-*β*-cell apoptosis activity, a classification model was constructed using six molecular fingerprints and seven machine learning methods. The model's performance was evaluated using 10-fold cross-validation and a test set. The results show that the model built using the Random Forest algorithm based on PubChem fingerprints performed the best (AUC = 0.992, CA = 0.96, and MCC = 0.901). Additionally, through information gain and substructure frequency analysis, nine characteristic substructures with anti-*β*-cell apoptosis activity were identified, including nitrogen-containing aromatic heterocycles, amines, and pyridines. These findings provide theoretical reference and guidance for the development of anti-*β*-cell apoptosis compounds.

Keywords

Diabetes, Anti-Pancreatic *β*-Cell Apoptosis, Classification Model

1. 引言

糖尿病可分为 1 型糖尿病(T1DM)、2 型糖尿病(T2DM)、妊娠期糖尿病或其他类型的糖尿病。T2DM 是最常见的糖尿病类型，占全球糖尿病总数的 90%以上，是以胰岛素抵抗和胰岛细胞功能衰竭为特征的慢性代谢性疾病[1]。目前针对 T2DM 的治疗策略为口服降糖药和胰岛素注射[2]，但无法从根源解决胰岛 *β* 细胞功能缺陷。因此，将胰岛 *β* 细胞作为研究对象，寻找可能减少胰岛 *β* 细胞死亡或促进 *β* 细胞增殖的治疗方法在 T2DM 治疗方面拥有良好前景。

胰岛 *β* 细胞的凋亡受多种因素影响，如内质网氧化应激[3]、细胞炎性因子和糖脂毒性[4]等，开发设计能够预防或逆转胰岛 *β* 细胞死亡的小分子成为糖尿病治疗药物的研究热点[5]-[8]。多种多酚类似物，如没食子酸[9]、姜黄素[10]、香豆素和黄酮类化合物，以及一系列氮杂环衍生物，已被证实具有抑制胰岛 *β* 细胞凋亡活性。香豆素类代表性化合物白屈菜素[11]的体外试验结果显示，浓度为 40 μM 的白屈菜素可将胰岛 *β* 细胞活性提高到 80%。黄酮类化合物包括芹菜素、桑黄素、黄芩苷、山柰酚和杨梅素，均被证实对胰岛 *β* 细胞具有显著的保护作用，从多种通路抑制胰岛 *β* 细胞凋亡[12]。除上述天然产物外，研究人员通过高通量筛选(High-Throughput Screening, HTS)和构效关系(Structure-Activity Relationship, SAR)方法，筛选出一系列具有抑制胰岛 *β* 细胞凋亡活性氮杂环衍生物，如螺环氮杂环丁烷-*β*-咔啉化合物[13]、7-氮杂环衍生物[14]和 2,4-二氨基喹唑啉化合物[14]等。其中一种 2,4-二氨基喹唑啉衍生物拥有最高的生物活性，EC_{50} = 0.56 μM。

目前关于具有抗胰岛 *β* 细胞凋亡活性化合物的研究，主要集中在新型骨架化合物的发现和结构优化方面，且该类药物的活性评价大多进行于细胞层面，需要耗费大量的时间与人力。因此，开发设计相关构效关系模型，来总结化学结构和生物活性之间的规律将对后续药物发现与优化具有指导性作用。定性分类研究(Qualitative Classification, QC)可通过机器学习(Machine Learning, ML)找出与抑制剂化合物抑制活性密切相关的特征子结构，为新型抑制剂的结构设计提供理论依据[15]。

2. 材料与方法

2.1. 数据集收集

综合考虑化合物分子骨架的多样性及其活性覆盖范围的广度，从已发表的文献中筛选出 103 个具有抗胰岛 β 细胞凋亡活性的化合物[13] [14] [16]-[18]，代表性化学骨架如图 1 所示。使用截止值 $EC_{50}=5$ μM 作为阈值，将低于阈值的化合物判定为高活性化合物(P)，反之则判定为低活性化合物(N)。按照 3:1 的比例随机分为训练集和测试集构建分类模型。数据集包含 76 个 P 类化合物和 27 个 N 类化合物。训练集中含 78 个化合物，其中含 58 个 P 类和 20 个 N 类；测试集中含 25 个化合物，其中含 18 个 P 类和 7 个 N 类(表 1)。活性化合物在各组中分布为：训练集 =74.3%，测试集 =72%，分布大致平衡，适用于评估模型预测性能。

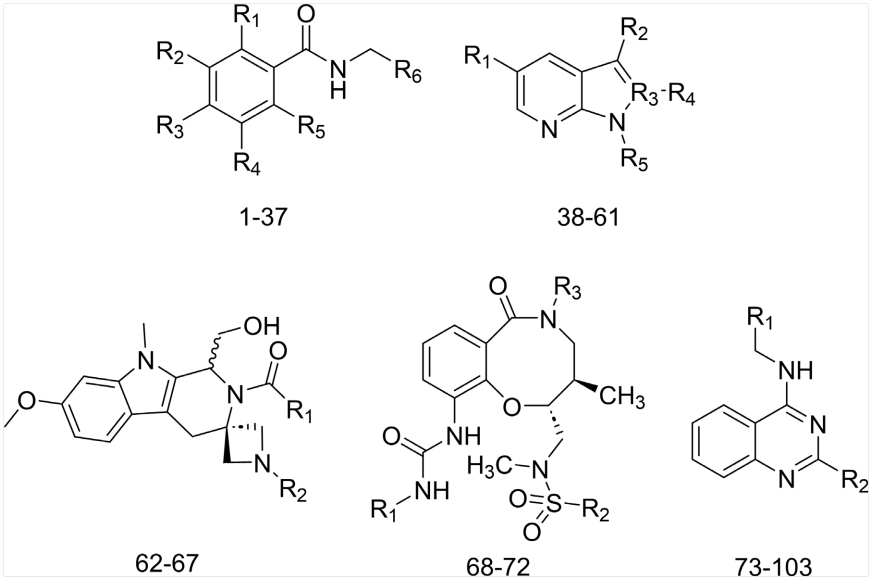


Figure 1. Molecular skeleton structures of anti-pancreatic β-cell apoptosis compounds heterocyclic inhibitors of data set
图 1. 抗胰岛 β 细胞凋亡化合物数据集分子骨架结构

Table 1. The distribution of compounds in the dataset
表 1. 数据集化合物分布情况

Data Set	Potent Compounds	Non-Potent Compounds	Total
Train set	58	20	78
Test set	18	7	25
Total	76	27	103

2.2. 分子指纹与机器学习方法

分子指纹属于化合物的化学结构特征，可显示化合物分子的结构或构象特征。通过采用二进制值(0 表示无，1 表示有)或计数来表征分子的二维或三维特征，将化学结构转换成计算机数据格式[19]。使用 Padel Descriptors [20]计算了 6 种分子指纹，即 Molecular Access System (MACCS)、Extended (Ext)、Estate (Est)、PubChem (Pub)、Substructure (Sub)、Graph-Only (Graph)。

本研究使用了 7 种机器学习算法，分别是决策树(Tree)、逻辑回归(LR)、随机森林(RF)、朴素贝叶斯

(NB)、k-最邻近(kNN)、人工神经网络(NN)和支持向量机(SVM) [21]。所有算法均在 Orange Canvas 3.11 软件(https://orange.biolab.si/)中生成。

2.3. 模型性能评价

采用十倍交叉验证和测试集评估分类模型的准确性，并用灵敏性(Sensitivity, SE)、特异性(Specificity, SP)和预测准确率(Classification Accuracy, CA)等参数对模型进行评估[22]，分别表示阳性化合物被正确识别为阳性的准确率、阴性化合物被正确识别为阴性的准确率和正确预测所有抑制剂百分比。其中正阳性(TP)表示阳性化合物被正确识别为阳性的数量；真阴性(TN)表示阴性化合物被正确识别为阴性的数量；假阳性(FP)表示阴性化合物被错误识别为阳性的数量；假阴性(FN)表示阳性化合物被错误识别为阴性的数量。

通过马修相关系数(Matthews Correlation Coefficient, MCC)对模型进行更平衡地评价，综合考虑分类模型的真阳性、假阳性、真阴性和假阴性。其取值范围从−1 到+1，+1 未出现任何误分类，0 表示模型预测与随机猜测相当，−1 则表示完全错误的预测。

此外，根据接收机性能特征(Receiver Operating Characteristic curve, ROC)曲线下面积(Area Under the Curve, AUC)评估分类模型准确性。ROC 曲线以 FP 率为横坐标，TP 率为纵坐标绘制而成。如果 AUC 值为 1，则说明分类器性能最优；如果 AUC 值为 0.5，则说明该分类器无分辨能力[23]。

2.4. 子结构片段分析

特征子结构是指与化合物活性密切相关的基团。通过综合评估信息增益值(Information Gain, IG)和子片段频率贡献来确认化合物的特权子结构[24]。若某子结构片段在 P 类化合物中频繁出现，则认为该片段有助于提高化合物的抗胰岛β细胞凋亡活性，是高活性化合物的特征子结构。

3. 结果与讨论

3.1. 数据集分析

数据集中 103 个具有抗胰岛β细胞凋亡化合物的 pEC_{50} ($pEC_{50} = -lgEC_{50}$)范围在 4.6 至 7.5 间(图 2(a))。分子热图中，红色(1)和绿色(0)分别表示分子的最高和最低多样性。图 2(b)中多数分子分布在红色区域，说明数据集分子结构具有化学多样性，基于该数据集训练所构建的分类模型具有较强的泛化能力。图 2(c)表征了数据集化合物分子的化学空间分布，数据集 60%的特征描述符方差是由前三个最主要的成分解释的，且训练集和测试集化合物所覆盖化学空间分布广泛且重合，说明训练集和测试集具有高度结构多样性和化学空间相似性。由 Lipinsiki 五规则分子描述符所构建的雷达图(图 2(d))无明显偏好。综上所述，该数据集适合开发稳健且具有广泛预测能力的分类模型。

3.2. 十倍交叉验证结果

基于训练集 78 个化合物，使用 6 种分子指纹和 7 种机器学习算法构建了 42 个分类模型(图 3)。大多数模型的 CA 和 AUC 值均大于 0.8；SE、SP 值分别在 0.74~0.96 和 0.6~0.9 之间。ExtFP 和 Graph 指纹的表现明显优于其他类型的指纹，尤其是在复杂算法(RF 和 NN)中。而 MACCS 和 SubFP 指纹较为简单，适合与简单的算法(LR)结合。从整体来看，SE 值普遍高于 SP 值，表明所有模型对 P 类化合物具有良好的预测能力，可能是由于数据集中 P 类化合物占比高(74.3%)，模型对高活性化合物预测更敏感。可通过引入 MCC 参数更好地衡量模型在各个类别(P 类和 N 类)上的整体表现，从而促使模型在提高对 P 类化合物预测能力的同时，平衡对 N 类化合物的预测，避免过度偏向主类。

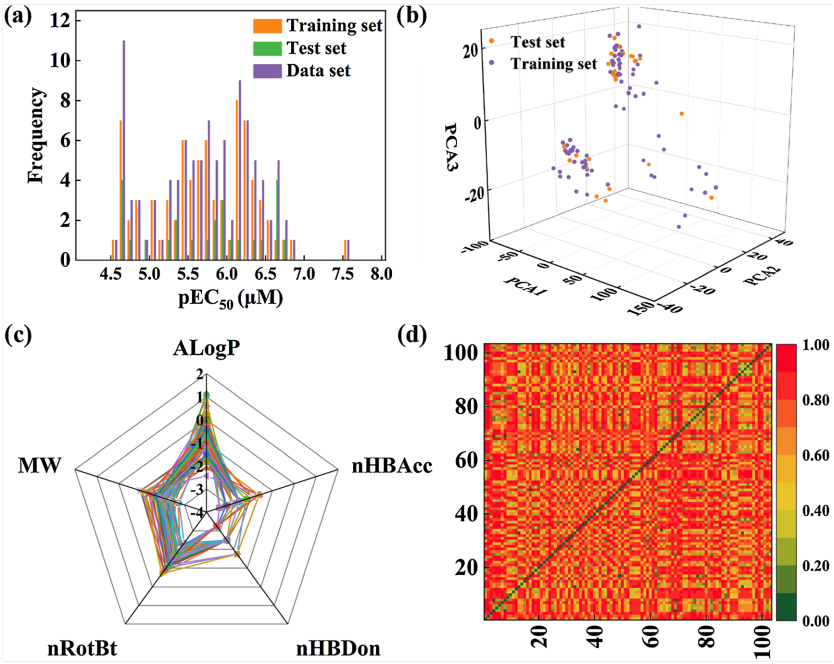


Figure 2. The pEC_{50} value distribution plot of the dataset (a), principal component analysis (PCA) plot (b), Lipinski's rule-of-five descriptor radar plot (c), and molecular similarity heat map based on Euclidean distance (d)
图 2. 数据集化合物 pEC_{50} 值分布图(a)，主成分分析图(b)，Lipinski 五规则标准描述符雷达图(c)以及基于欧几里得距离的分子相似性热图(d)

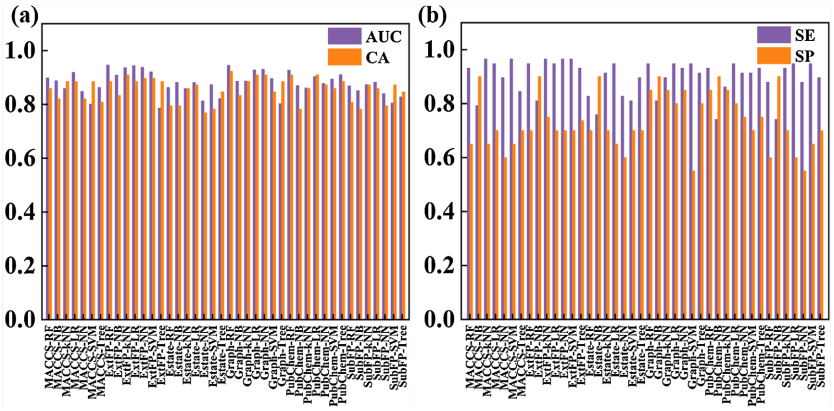


Figure 3. (a) AUC-CA and (b) SE-SP histogram of 42 models by ten fold cross validation
图 3. 10 倍交叉验证法 42 个模型(a)AUC-CA 及(b)SE-SP 柱状图

Table 2. Results of top ten classification models
表 2. 前 10 分类模型结果

Data set	model	AUC	CA	MCC	TN	TP	FN	FP	SE	SP
Training set	ExtFP-RF	0.946	0.885	0.686	14	55	3	6	0.948	0.700
	Graph-RF	0.945	0.923	0.798	17	55	3	3	0.948	0.850
	ExtFP-LR	0.944	0.885	0.686	14	55	3	6	0.948	0.700
	ExtFP-NN	0.938	0.897	0.720	14	56	2	6	0.966	0.700
	ExtFP-kNN	0.937	0.91	0.757	15	56	2	5	0.966	0.750

续表

	Graph-NN	0.93	0.91	0.769	17	54	4	3	0.931	0.850
	Graph-LR	0.928	0.91	0.761	16	55	3	4	0.948	0.800
	PubChem-RF	0.927	0.91	0.769	17	54	4	3	0.931	0.850
	ExtFP-SVM	0.921	0.897	0.720	14	56	2	6	0.966	0.700
	MACCS-LR	0.919	0.885	0.686	14	55	3	6	0.948	0.700
	ExtFP-RF	1	0.96	0.901	6	18	0	1	1.000	0.857
	Graph-RF	0.976	0.96	0.901	6	18	0	1	1.000	0.857
	ExtFP-LR	0.992	0.96	0.901	6	18	0	1	1.000	0.857
	ExtFP-NN	0.968	0.96	0.901	6	18	0	1	1.000	0.857
	ExtFP-kNN	1	0.96	0.901	6	18	0	1	1.000	0.857
Test set	Graph-NN	0.968	0.96	0.802	6	17	1	1	0.944	0.857
	Graph-LR	0.96	0.92	0.802	6	17	1	1	0.944	0.857
	PubChem-RF	0.992	0.96	0.901	6	18	0	1	1.000	0.857
	ExtFP-SVM	0.984	0.92	0.802	5	18	0	2	1.000	0.714
	MACCS-LR	0.96	0.96	0.901	6	18	0	1	1.000	0.857

3.3. 前十模型性能

排名前十的模型是 Ext-RF、Graph-RF、Ext-LR、Ext-NN、Ext-kNN、Graph-NN、Graph-LR、PubChem-RF、Ext-SVM 和 MACCS-LR。整体来看，前十模型在 AUC、准确率和 MCC 三个指标上均表现出了较为均衡的性能。AUC 值越高，表示模型的区分能力越强，能够更好地将两类分开。Graph-RF 模型在 AUC(0.945)和准确率(0.923)上表现突出，表明该模型在区分正负类别的能力和整体预测准确性方面具有较强的优势。其 MCC 值(0.798)进一步反映了该模型在处理不平衡数据集时，能够较好地平衡对少数类样本的预测，这对于解决类别不平衡问题尤为重要。ExtFP 系列模型(如 ExtFP-RF、ExtFP-LR)尽管 AUC 较高(0.946、0.944)，但 MCC 较低(0.686)，表明这些模型在少数类样本的识别上可能存在一定的偏差。总体而言，Graph-RF、Graph-NN 和 PubChem-RF 模型在综合评估指标上均表现出较为优异的能力，特别适用于类别不平衡或少数类样本重要性的应用场景。PubChem-RF 计算复杂度适中，ExtFP-kNN 计算复杂度较低但存储和检索效率受限，而 Graph-RF 在处理复杂分子结构时计算成本较高。且 Graph-RF 适用于复杂拓扑分子，ExtFP-kNN 适用于特征空间稳定的数据。PubChem-RF 则兼具高预测性能和良好可解释性，因此最适用于本研究任务。

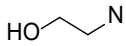
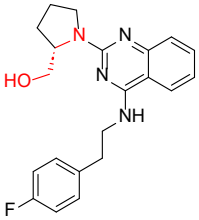
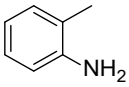
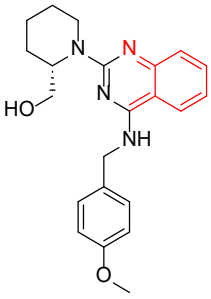
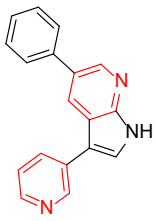
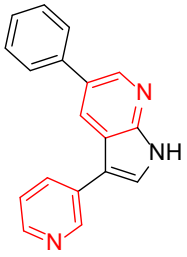
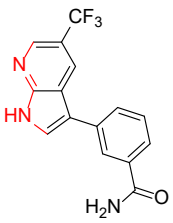
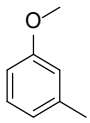
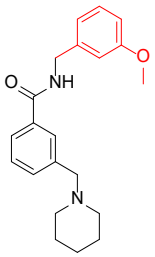
3.4. 测试集表现

PubChem-RF 模型在 AUC (0.992)、准确率(0.96)和 MCC (0.901)上均表现优异，展现了出色的整体分类能力和少数类预测性能，尤其在类别不平衡的情况下表现稳健，符合实际应用中的高准确性需求(表 2)。此外，ExtFP-RF、ExtFP-LR 和 ExtFP-kNN 同样具备较高的 AUC 和准确率，且 MCC 值相同(0.901)，在少数类识别上表现良好。而 Graph-RF、Graph-NN 和 ExtFP-SVM 的 AUC 和 MCC 略低，表明它们在少数类型预测上存在一定局限。因此，PubChem-RF 模型的综合表现使其成为最优选择，尤其适用于类别不平衡的应用场景。

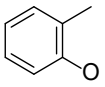
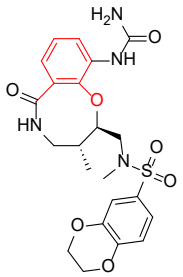
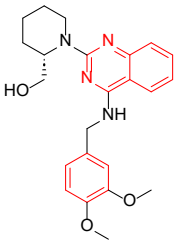
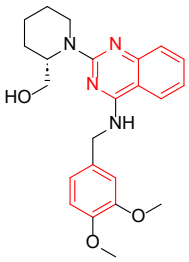
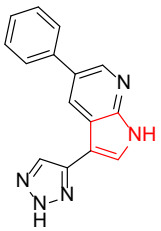
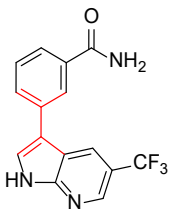
3.5. 特征子结构的筛选与分析

本研究利用 PubChem 指纹图谱，综合信息增益和子结构频率分析方法识别化合物中与保护胰岛 β 细胞免凋亡活性相关的子结构特征，IG 值和子结构片段频率值越大，说明该分子指纹所代表的结构片段则有助于提高化合物的抗胰岛 β 细胞凋亡活性。

Table 3. Substructural features of anti-β-cell apoptosis activity based on Pubchem fingerprints
表 3. 基于 PubChem 指纹的抗 β 细胞凋亡活性特征子结构

No.	Privileged substructures	General substructures	Representative compounds	IG	FP	FN
FP659	O-C-N-[#1]			0.228	1.34 (31)	0 (0)
FP758	Cc1c(N)cccc1					
FP821	CC1C(N)CCCC1			0.180	1.34 (26)	0 (0)
FP187	≥2 saturated or aromatic nitrogen-containing ring size 6			0.093	1.34 (15)	0 (0)
FP636	C-N-C-N-C			0.273	1.31 (41)	0.09 (1)
FP698	O-C-C-C-C-C-C-C					
FP598	C-C:C-O-C			0.035	1.25 (13)	0.18 (1)

续表

FP756	Cc1c(O)cccc1			0.180	1.15 (6)	0.56 (1)
FP259	≥3 aromatic rings			0.236	1.28 (42)	0.18 (2)
FP145	≥1 saturated or aromatic nitrogen-containing ring size 5			0.113	1.26 (30)	0.24 (2)
FP712	C-C(C)-C(C)-C			0.070	1.24 (24)	0.30 (2)

本研究得到 9 个与抗胰岛 β 细胞凋亡活性相关的子结构(表 3)。大多数化合物子结构含多个芳香环，如含有 PubChemFP259 的化合物 40-63 和 73-102 都是高活性化合物，这些化合物都含有三个以上芳香环。含氧原子的链状结构通常作为链接子或侧链出现在大骨架化合物中，含氮芳香环则作为骨架在化合物中呈现。与仅以苯环作为骨架的苯甲酰胺衍生物(化合物 1~37，EC_{50} 均值为 8 μM)相比，苯环上含有氮原子或含氮取代基(PubChemFP758、PubChemFP821 和 PubChemFP636)作为骨架的杂环并联化合物(化合物 38~61、73~103)有着更高的生物活性，EC_{50} 均值为 1.5 μM。其中，酚基(PubChemFP756)作为高活性特征子结构被识别出来，其存在于 6 个高活性化合物(化合物 87~90、94 和 95)中，EC_{50} 范围为 0.17~1.88 μM。筛选出的子结构在高活性抗 β 细胞凋亡化合物中具有较高出现频率，能够作为药效特征子结构，用于预测和设计新型化合物。

4. 结论

本研究以 103 个具有抗胰岛 β 细胞凋亡活性化合物为数据集，构建了基于 7 种机器学习方法及 6 种分子指纹方法的分类模型，并对模型进行了基于交叉验证和测试集的性能评估。其中 PubChem-RF 模型(AUC = 0.992、CA = 0.96 和 MCC = 0.901)在测试集中表现最佳。结合信息增益和子片段频率贡献识别出

9 个抗胰岛 β 细胞凋亡药效特征子结构，如含氮芳香杂环、胺类、吡啶等高活性片段，对新药的设计和开发具有重要意义，为抗胰岛 β 细胞凋亡化合物的开发提供了理论参考和指导。未来研究可进一步优化模型，探索更多算法和融合策略，结合实验验证提高模型的可靠性，并探讨其在抗糖尿病药物开发中的应用前景。

基金项目

北京工业大学教育教学课题(ER2024RCA02)。

参考文献

[1] Zheng, Y., Ley, S.H. and Hu, F.B. (2017) Global Aetiology and Epidemiology of Type 2 Diabetes Mellitus and Its Complications. *Nature Reviews Endocrinology*, **14**, 88-98. https://doi.org/10.1038/nrendo.2017.151

[2] Clapham, J.C. (2019) Sixty Years of Drug Discovery for Type 2 Diabetes: Where Are We Now? In: Stocker, C., Ed., *Methods in Molecular Biology*, Springer, 1-30. https://doi.org/10.1007/978-1-4939-9882-1_1

[3] Böni-Schnetzler, M. and Meier, D.T. (2019) Islet Inflammation in Type 2 Diabetes. *Seminars in Immunopathology*, **41**, 501-513. https://doi.org/10.1007/s00281-019-00745-4

[4] Cnop, M., Welsh, N., Jonas, J., Jörns, A., Lenzen, S. and Eizirik, D.L. (2005) Mechanisms of Pancreatic β-Cell Death in Type 1 and Type 2 Diabetes. *Diabetes*, **54**, S97-S107. https://doi.org/10.2337/diabetes.54.suppl_2.s97

[5] Wang, N., Yi, W.J., Tan, L., Zhang, J.H., Xu, J., Chen, Y., *et al*. (2017) Apigenin Attenuates Streptozotocin-Induced Pancreatic β Cell Damage by Its Protective Effects on Cellular Antioxidant Defense. *In Vitro Cellular & Developmental Biology-Animal*, **53**, 554-563. https://doi.org/10.1007/s11626-017-0135-4

[6] Vanitha, P., Senthilkumar, S., Dornadula, S., Anandhakumar, S., Rajaguru, P. and Ramkumar, K.M. (2017) Morin Activates the Nrf2-ARE Pathway and Reduces Oxidative Stress-Induced DNA Damage in Pancreatic Beta Cells. *European Journal of Pharmacology*, **801**, 9-18. https://doi.org/10.1016/j.ejphar.2017.02.026

[7] Wang, N., Zhang, J., Qin, M., Yi, W., Yu, S., Chen, Y., *et al*. (2017) Amelioration of Streptozotocin-Induced Pancreatic β Cell Damage by Morin: Involvement of the AMPK-FOXO3-Catalase Signaling Pathway. *International Journal of Molecular Medicine*, **41**, 1409-1418. https://doi.org/10.3892/ijmm.2017.3357

[8] Roy, S., Metya, S.K., Sannigrahi, S., Rahaman, N. and Ahmed, F. (2013) Treatment with Ferulic Acid to Rats with Streptozotocin-Induced Diabetes: Effects on Oxidative Stress, Pro-Inflammatory Cytokines, and Apoptosis in the Pancreatic β Cell. *Endocrine*, **44**, 369-379. https://doi.org/10.1007/s12020-012-9868-8

[9] Sameermahmood, Z., Raji, L., Saravanan, T., Vaidya, A., Mohan, V. and Balasubramanyam, M. (2010) Gallic Acid Protects Rinm5f β-Cells from Glucolipotoxicity by Its Antiapoptotic and Insulin-Secretagogue Actions. *Phytotherapy Research*, **24**, S83-S94. https://doi.org/10.1002/ptr.2926

[10] Hao, F., Kang, J., Cao, Y., *et al*. (2015) Curcumin Attenuates Palmitate-Induced Apoptosis in MIN6 Pancreatic β-Cells through PI3K/Akt/FoxO1 and Mitochondrial Survival Pathways. *Apoptosis*, **20**, 1420-1432. https://doi.org/10.1007/s10495-015-1150-0

[11] Vinayagam, R. and Xu, B. (2017) 7, 8-Dihydroxycoumarin (Daphnetin) Protects INS-1 Pancreatic *β*-Cells against Streptozotocin-Induced Apoptosis. *Phytomedicine*, **24**, 119-126. https://doi.org/10.1016/j.phymed.2016.11.023

[12] 王鹏程, 曹泽彧, 许冶良, 等. 以 2 型糖尿病胰岛 β 细胞为靶点的天然产物研究进展[J]. 中草药, 2019, 50(18): 4502-4510.

[13] Chou, D.H., Duvall, J.R., Gerard, B., Liu, H., Pandya, B.A., Suh, B., *et al*. (2011) Synthesis of a Novel Suppressor of β-Cell Apoptosis via Diversity-Oriented Synthesis. *ACS Medicinal Chemistry Letters*, **2**, 698-702. https://doi.org/10.1021/ml200120m

[14] Huang, Z., Tremblay, M.S., Wu, T.Y.-H., Ding, Q., Hao, X., Baaten, J., *et al*. (2019) Discovery of 5-(3,4-Difluorophenyl)-3-(Pyrazol-4-Yl)-7-Azaindole (GNF3809) for β-Cell Survival in Type 1 Diabetes. *ACS Omega*, **4**, 3571-3581. https://doi.org/10.1021/acsomega.8b03050

[15] Kong, W., Wang, W. and An, J. (2020) Prediction of 5-Hydroxytryptamine Transporter Inhibitors Based on Machine Learning. *Computational Biology and Chemistry*, **87**, Article 107303. https://doi.org/10.1016/j.compbiolchem.2020.107303

[16] Small, J.C., Joblin-Mills, A., Carbone, K., Kost-Alimova, M., Ayukawa, K., Khodier, C., *et al*. (2022) Phenotypic Screening for Small Molecules That Protect β-Cells from Glucolipotoxicity. *ACS Chemical Biology*, **17**, 1131-1142. https://doi.org/10.1021/acschembio.2c00052

[17] Duan, H., Li, Y., Arora, D., Xu, D., Lim, H. and Wang, W. (2017) Discovery of a Benzamide Derivative That Protects Pancreatic β-Cells against Endoplasmic Reticulum Stress. *Journal of Medicinal Chemistry*, **60**, 6191-6204. https://doi.org/10.1021/acs.jmedchem.7b00435

[18] Duan, H., Lee, J.W., Moon, S.W., Arora, D., Li, Y., Lim, H., *et al*. (2016) Discovery, Synthesis, and Evaluation of 2,4-Diaminoquinazolines as a Novel Class of Pancreatic β-Cell-Protective Agents against Endoplasmic Reticulum (ER) Stress. *Journal of Medicinal Chemistry*, **59**, 7783-7800. https://doi.org/10.1021/acs.jmedchem.6b00041

[19] Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S. and Pujadas, G. (2015) Molecular Fingerprint Similarity Search in Virtual Screening. *Methods*, **71**, 58-63. https://doi.org/10.1016/j.ymeth.2014.08.005

[20] Yap, C.W. (2010) PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *Journal of Computational Chemistry*, **32**, 1466-1474. https://doi.org/10.1002/jcc.21707

[21] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A. (2019) Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC *Medical Informatics and Decision Making*, **19**, Article No. 281. https://doi.org/10.1186/s12911-019-1004-8

[22] Fan, T., Sun, G., Zhao, L., Cui, X. and Zhong, R. (2018) QSAR and Classification Study on Prediction of Acute Oral Toxicity of *N*-Nitroso Compounds. *International Journal of Molecular Sciences*, **19**, Article 3015. https://doi.org/10.3390/ijms19103015

[23] Pérez-Garrido, A., Helguera, A.M., Borges, F., Cordeiro, M.N.D.S., Rivero, V. and Escudero, A.G. (2011) Two New Parameters Based on Distances in a Receiver Operating Characteristic Chart for the Selection of Classification Models. *Journal of Chemical Information and Modeling*, **51**, 2746-2759. https://doi.org/10.1021/ci2003076

[24] Chen, Z., Zhang, L., Zhang, P., Guo, H., Zhang, R., Li, L., *et al*. (2023) Prediction of Cytochrome P450 Inhibition Using a Deep Learning Approach and Substructure Pattern Recognition. *Journal of Chemical Information and Modeling*, **64**, 2528-2538. https://doi.org/10.1021/acs.jcim.3c01396