·综述·

基于“人工智能+药物基因组学”的药物不良反应预测方法进展

韩芳芳[1#]，刘静欣[1#]，柴克燕[2]，吴嘉瑞[2]，蔡永铭[1]

1. 广东药科大学医药信息工程学院（广州 510006）
2. 北京中医药大学中药学院（北京 102488）

【摘要】药物不良反应（ADR）是全球药物警戒关注的首要问题，个体遗传差异，尤其是药物基因组学（PGx）特征，是导致 ADR 发生的关键因素。近年来，人工智能（AI）技术为整合多组学数据、精准预测 ADR 提供了可能。本文系统梳理了基于 PGx 预测 ADR 的 AI 方法。首先整理了常用的 PGx 与 ADR 相关的多源异构数据集，然后重点列举了传统机器学习（如支持向量机、随机森林等）及深度学习（如卷积神经网络、图神经网络等）等 AI 模型在该领域的应用实例。这些模型通过挖掘基因变异、临床用药特征与 ADR 之间的复杂非线性关系，实现了 ADR 的智能预测。然而，该领域仍面临数据异质性、模型可解释性及临床转化障碍等挑战。文章最后展望了多模态数据融合、可解释 AI 等未来研究方向，旨在推动个体化安全用药和精准医疗事业的发展。

【关键词】药物不良反应；药物基因组学；人工智能；机器学习；多源异构；神经网络

【中图分类号】R968　　**【文献标识码】**A

Recent advances in adverse drug reaction prediction using artificial intelligence and pharmacogenomics

HAN Fangfang[1#], LIU Jingxin[1#], CHAI Keyan[2], WU Jiarui[2], CAI Yongming[1]

1. School of Medical Information and Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China
2. School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing 102488, China
Co-first authors: HAN Fangfang and LIU Jingxin
Corresponding authors: WU Jiarui, Email: exogamy@163.com; CAI Yongming, Email: cym@gdpu.edu.cn

【Abstract】Adverse drug reaction (ADR) represents a primary concern in global pharmacovigilance. Individual genetic variations, particularly pharmacogenomics (PGx) characteristics, are key factors contributing to the occurrence of ADR. In recent years, artificial intelligence (AI) technologies have enabled the integration of multi-omics data for accurate ADR prediction. This review summarizes AI methods for predicting ADR based on PGx. It begins by organizing commonly used multi-source heterogeneous datasets related to PGx and ADR, then

DOI: 10.12173/j.issn.1005-0698.202508154
[#] 共同第一作者
基金项目：广东省中医药管理局中医药科研项目（20251212）；广东省医学科学技术基金（C2025084）；广州市科技计划项目（2025A03J3712）
通信作者：吴嘉瑞，博士，教授，博士研究生导师，Email：exogamy@163.com
蔡永铭，博士，教授，硕士研究生导师，Email：cym@gdpu.edu.cn

highlights application examples of AI models—such as traditional machine learning (e.g., support vector machine, random forests) and deep learning (e.g., convolutional neural networks, graph neural networks)—in this field. These models enable intelligent prediction of ADR by uncovering complex non-linear relationships among genetic variations, clinical medication features, and ADR. However, the field still faces challenges, including data heterogeneity, model interpretability, and obstacles in clinical translation. Finally, the review outlines future research directions, such as multi-modal data fusion and explainable AI, aiming to advance the development of personalized medication safety and precision medicine.

【Keywords】Adverse drug reaction; Parmacogenomics; Artificial intelligence; Machine learning; Multi-source heterogeneous; Neural network

药物不良反应（adverse drug reaction，ADR）是指合格药物在正常用法用量下出现的与用药目的无关的有害反应[1]。根据我国国家药品不良反应监测中心发布《国家药品不良反应监测年度报告（2024 年）》[2]显示，2024 年我国药品不良反应监测网络收到了 259.7 万份 ADR 报告，覆盖全国 98.7% 的县级地区。文献[3]调研结果显示，全球 65 岁以上的住院老年患者中，平均 16% 经历了显著的 ADR。而我国 2024 年 ADR/ 药物不良事件（adverse drug event，ADE）报告中，65 岁及以上老年患者占比为 33.4%，与 45~64 岁人群 33.9% 的占比相当，14 岁以下儿童患者占比为 8.7%。按照怀疑药物类别统计，化学药物占 81.0%，中药占 12.1%，生物制品占 3.9%[2]。ADR 不仅可能发生于各年龄段的人群，而且不论哪种类别的药物都可能产生。虽然导致 ADR 发生的可能因素有很多，包括联合用药、生活方式、环境、年龄和饮食等，但研究[4]显示，约 80% 的药物疗效和安全性都与基因相关，因此基因多态性是造成人群中个体 ADR 差异的最重要原因。

为了研究基因与药物疗效和安全性之间的关联，实现个体化精准用药，药物基因组学（pharmacogenomics，PGx）应运而生。PGx 是研究个体基因组变异如何影响药物反应的一门科学，可在优化药物疗效同时降低 ADR 的风险[5]。随着 PGx 的快速发展，个体间的基因序列差异与 ADR 之间的相关性不断被证实[6]。临床药物遗传学实施联盟（Clinical Pharmacogenetics Implementation Consortium，CPIC）、美国食品药品管理局（Food and Drug Administration，FDA）等权威组织相继发布 PGx 相关指南，对近年来有关基因多态性预测和 ADR 证据强度等级认定的最新进展及其重要意义进行总结，以期为基因导向的个体化给药提供参考。另外，遗传药理学和药物基因组学知识库（PharmGKB）还收录了与 ADR 相关的基因数据信息共 1 544 条，共 315 个药物–基因对（其中 111 对是被评定为具有高证据等级的）。然而这只是目前发现的具有高证据等级的基因记录，只要基因突变存在，新药不断研发，与基因多态性相关的 ADR 就需要被不断发现、监测，从而为个体化精准用药提供指导参考。因此，需要基于当前已发现的与 PGx 相关的 ADR，挖掘其相互关系规律，对 ADR 进行预测，才能真正实现临床精准用药指导，从而有效降低 ADR 的发生率。

随着基因分型技术的广泛应用和成本的降低，可用于 ADR 预测学习的数据种类被不断丰富，数据量飞速增长，数据的复杂度也更高。与此同时，人工智能（artificial intelligence，AI）中的机器学习方法也在飞速发展，从依赖人工特征（如将不同平台测得的基因组、转录组、拷贝数、突变谱等原始数据拼接成“样本 × 特征”矩阵）的传统机器学习方法（如支持向量机、随机森林等可监督分类器），到可以从更多类型数据（如整合基因组、转录组、表观基因组、蛋白质组等多维度数据）自动学习更多统计学特征的深度学习方法的出现，只要给足可供学习的数据，深度学习方法便可以主动精确剖析数据特征，根据概率计算最佳结果，从根本上改变了数据分析的传统范式，ADR 预测方法也被飞速推进到“AI+”时代。基于此，本文将从 PGx 数据与机器学习方法相结合发展的角度，对 ADR 预测机器学习方法的现状进行综述，并对当前所存在的关键问题和未来发展趋势进行探讨。

1 用于ADR预测的PGx相关数据简介

对于机器学习方法而言，可供学习的训练数据是至关重要的。由于ADR预测涉及药物和生物体之间的相互作用过程，不仅包括药物本身的物理化学属性特征（分子结构、作用机制等），还包括与其对应的生物属性特征（基因、蛋白质靶点等），以及患者的临床信息（年龄、性别等）；同时，这些不同属性特征的表达方式也有多种模式，如分子序列、文字文本和图像等，因此PGx和ADR相关数据的复杂程度可见一斑。

表1列出了一些有代表性的数据库，可为相关领域研究者提供参考。例如DrugBank、DrugCentral、PubChem、SuperTarget、Open Targets、ChEMBL，整合了药物结构、靶点、作用机制和生物活性数据；PharmGKB、PharmVar、FDA Table of Pharmacogenomic Associations、dbSNP、DGV、HLA-ADRs等数据库包含了基因变异对药物代谢、疗效与毒性的影响；CTD、KEGG、LINCS L1000整合了化学、基因与疾病网络及表达扰动机制。在预测ADR时，这些数据库可用于生成表示药物的关键特征。FAERS、SIDER、HCUP分别收录了ADR案例、已知的用药与ADR关联信息，以及真实的患者诊疗记录；UK Biobank、TOPMed提供人群级基因组和表型关联，支撑遗传相关ADR验证；OMOP-CDM、MedDRA、MeSH统一了术语表达，为不同来源的数据、语义实现了标准化。表1中大部分数据为开源，但也有一些需要申请访问，使用时需要注意遵守相关数据集的要求和伦理规范。

表1 PGx与ADR临床相关数据集汇总

Table 1. Summary of PGx and ADR clinical datase

序号	数据库名称	简要描述	参考网址
1	DrugCentral[7]	药物-靶点相互作用+系统/器官分类的ADR	https://drugcentral.org/
2	DrugBank[8]	药物-靶点相互作用	https://go.drugbank.com/
3	dbSNP[9]	高频SNP/SNV突变（通用基因变异数据库）	https://www.ncbi.nlm.nih.gov/snp/
4	PharmGKB[10]	经实验验证的影响药物疗效和安全性的基因突变核心资源，提供基因-药物-临床指南注释	https://www.clinpgx.org/
5	PharmVar[11]	提供高质量的药理学基因变异数据，包括基因型-表型关联、等位基因功能注释	https://www.pharmvar.org/
6	DGV[12]	基因变异数据库，旨在为与基因组变异与表型数据相关联的研究提供一个有用的对照数据参考。由加拿大多伦多大学收集和维护	https://ngdc.cncb.ac.cn/databasecommons/database/id/283
7	HLA-ADRs[13]	提供与ADR相关的人白细胞抗原（human leukocyte antigen，*HLA*）等位基因频率和单倍型	https://ngdc.cncb.ac.cn/databasecommons/database/id/1892
8	CTD[14]	整合大量化学物质、基因、功能表型和疾病之间相互作用数据，为疾病相关环境暴露因素及药物潜在作用机制研究提供基础数据	https://ctdbase.org/
9	KEGG[15]	通过基因组测序和其他高通量实验技术生成的大规模分子数据集，了解生物系统（如细胞、生物体和生态系统）的高级功能和效用	https://www.genome.jp/kegg/
10	SuperTarget[16]	包含332 828种药物靶点相互作用的网络资源	https://ngdc.cncb.ac.cn/databasecommons/database/id/564
11	PubChem[17]	包含化学结构信息	https://pubchem.ncbi.nlm.nih.gov/
12	LINCS L1000[18]	包含真实基因表达谱和变化数据	https://ngdc.cncb.ac.cn/databasecommons/database/id/6422
13	Table of Pharmacogenomic Associations[19]	基因-药物关联数据表	https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations
14	TOPMed[5]	全球18万+个体的全基因组测序数据	https://topmed.nhlbi.nih.gov/
15	FAERS[20]	临床观测ADR数据库	https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database

续表1

序号	数据库名称	简要描述	参考网址
16	SIDER[21]	药物-ADR数据库，提供ADR信息数	http://sideeffects.embl.de/
17	OMOP-CDM（从EHR提取）[22]	临床数据库，一个开放的社区数据，旨在规范观察性数据的结构和内容，并实现能够产生可靠证据的有效分析	https://www.ohdsi.org/data-standardization/
18	MedDRA[23]	《国际医学用语词典》（*Medical Dictionary for Regulatory Activities*），临床验证的国际医学术语集	https://www.meddra.org/
19	MeSH[24]	医学主题词表，可提供ADR语义数据	https://www.nlm.nih.gov/mesh/meshhome.html
20	ChEMBL[25]	一个大规模、开放的FAIR数据库，包含具有药物样特性的生物活性分子	https://www.ebi.ac.uk/chembl/

注：EHR. 电子健康记录（electronic health record）。

2 基于PGx数据预测ADR的AI方法概述

作为一种 AI 解决方案，机器学习方法可以辅助从未修正过的数据或非结构化文本中提取有用信息 [26]，并跨越数据源之间的界限 [27]，因此近年来应用非常广泛。机器学习方法按发展进程主要分为传统机器学习和深度学习两个阶段，本章中将对这两种方法用于 ADR 预测进行概述。

基于机器学习的 ADR 预测模式通常包括数据搜集与清洗对接、多模态数据标准化表征、基于机器学习的特征提取与分类模型构建、模型性能验证 4 个主要部分，如图 1 所示。表 1 中展示了多种不同源数据，包括药物和基因属性及两者的相互作用关系、蛋白质靶点、ADR 类别等，不同数据集包含的信息并不一一对应，因此需要对数据进行搜集、清洗和对接，才能将药代动力学或药效动力学过程中有相互关联的 PGx 信息与 ADR 对应起来，形成一个完整地确定 ADR 是否发生的数据集。然而由于药物和基因都有多种属性，且表达方式也不相同，需要进行表征标准化处理。如药物分子结构可以有多种表示方法，包括通用名称、化学分子式、IUPAC 名称、CAS RN、Canonical SMILES、InChI、WLN 等 [28]。此外，还可以用二维和三维图形来表示分子空间结构，考虑更多空间信息的优势是可以进一步模拟建立药物分子与其他分子的相互作用关系 [29]。大多数药物分子与 ADR 类别的相关性均通过报告文本的形式记录，因此除了序列结构、图形两种模态数据外，还有自然语言文本数据，这些数据要融合起来学习，则需要进行多模态数据标准化表征处理。然后，所有样本标签和多模态标准化表征数据进行对应后，才能建立基于机器学习的分类模型，实现 ADR 的 AI 预测。最后，再对预测模型的性能进行参数评估和实例验证。

2.1 面向ADR预测的传统机器学习方法

传统机器学习方法通过人工指导特征提取并拼接后，形成带有 ADR 标注的特征向量，再设计分类器学习分类，使其具有预测 ADR 是否会发生的能力。通常实现流程如图 2 所示。

在传统机器学习预测 ADR 的研究领域，机器学习方法发挥了重要作用，具体研究方法如表 2 所示，基于该方法与矩阵降维（kernel matrix

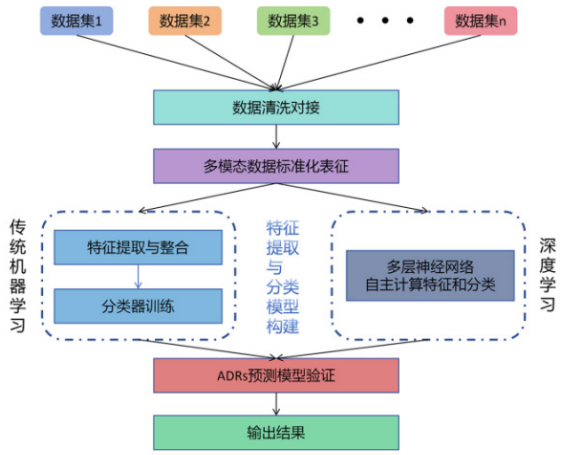

图1 基于机器学习的ADR预测通用模式

Figure 1. General flowchart for ADR prediction based on machine learning

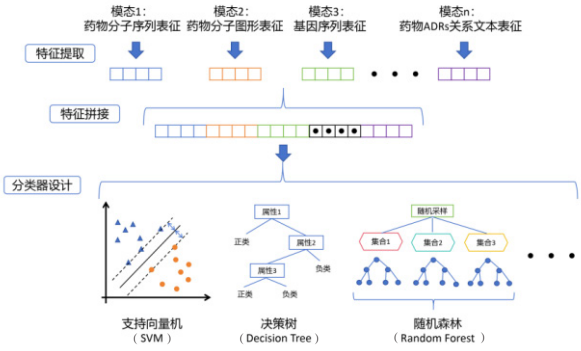

图2 面向ADR预测的传统机器学习方法通用流程

Figure 2. The general pipeline of traditional machine learning methods for ADR prediction

表2 ADR预测传统机器学习方法简述

Table 2. A brief review of traditional machine-learning methods for ADR prediction

序号	方法	实验数据来源	优势	局限
1	基于KMDR的传统机器学习算法，用于找到药物与ADR的关联[30]	DrugBank、Kegg、FAERS和SIDER	降维后模型简化了复杂矩阵运算，适用于大规模药物发现和临床安全评估	①模型性能依赖于降维参数p ②模型基于静态相似性，未整合动态变化因素或患者特定因素
2	QSP模型[31]	电子医疗记录、FAERS、NIH LINCS和hiPSC	①解决传统药理基因组学研究的样本大小不足和ADR多因素性质问题 ②整合多尺度数据，提供系统性描述机制的ADR表征 ③用hiPSC细胞系找人类个体变异，减少动物模型依赖	①hiPSC衍生细胞的成熟度不足，不能模拟成人表型 ②数据整合存在挑战，需要开发新算法，目前算法尚未成熟，无法准确预测 ③依赖大规模细胞库、数据库和基础设施的建立
3	改进的朴素贝叶斯网络模型，用于结合各节点（药物、基因、ADR）间的条件概率关系以计算基因引发ADR的概率，将无权重网络升级为复杂权重网络[32]	ADReCS、CTD、OMIM、LINCS L1000	①把分散的数据库统一整合成知识库，方便检索和分析 ②模型中用网络分析找出潜在基因靶点，有助于新药开发和个性化治疗	①数据有限，基因表达数据来源于体外细胞实验，未模拟人体药代动力学，可能偏差大 ②假阳性高，模型鲁棒性不高 ③未经大规模临床验证，潜在靶点需实验确认 ④技术限制，当前实验难系统预测常见ADR，模型性能依赖阈值
4	基于三元非负矩阵分解模型的中药配伍禁忌预测，通过功效性味属性及其关联关系实现预测；针对化学药物ADR预测，提出多任务多属性学习和依赖判别性特征选择两种模型，建立分子结构与ADR属性的关系[33]	文献、FAERS、SIDER、OMIM、CTD	①模型在真实和合成数据集上均做了性能测试，表现优异 ②通过特征选择能有效挖掘关键因素 ③多任务学习处理多属性数据，减少维度灾难	①中药数据源于文献，缺乏实时临床验证 ②模型参数多，需网格搜索优化，计算开销大，不适合实时预测 ③泛模型化能力有限，对罕见ADR或新药预测准确率较低 ④无法考虑动态因素，如患者个体差异
5	随机森林分类模型和投票决策[34]	GEO、cMap、SIDER	①整合了多数据源，提高生物学解释力 ②方法简单、效率高，适用于临床指导和药物开发	①数据集规模小，影响泛化能力 ②数据特征单一，仅用基因表达变化，未融合其他特征（如药物结构、靶点） ③限于二分类，算法简单，缺乏大规模验证
6	多核多任务学习模型（multi-kernel multi-task learning for adverse drug reaction，MKMT-ADR），贝叶斯个性化排序损失函数优化参数，实现个性化ADR预测[35]	FAERS、DrugBank、DisGeNET、Ensembl数据集	①将患者个体差异（年龄、性别等）和多模态数据进行融合，实现针对性预测 ②多任务共享信息提升了泛化能力，多核函数精准处理异质数据，降低过拟合	①包含多核+多任务，计算复杂，框架参数多，训练开销大 ②罕见ADR预测虽改善，但准确性仍较低 ③缺乏大规模临床验证，潜在生物学解释要进行进一步的实验支撑

dimension reduction，KMDR）的模型、量化系统药理学（quantitative systems pharmacology，QSP）模型、改进的朴素贝叶斯网络、三元非负矩阵分解、随机森林以及多核多任务学习模型等。这些方法整合了药物、基因和 ADR 数据，通过降维、多任务学习、网络分析或特征选择等技术挖掘潜在关联。它们简化运算，在处理数据时简单且高效，但也普遍面临数据质量依赖度高、数据整合算法不成熟、模型泛化能力有限等问题。

2.2 面向ADR预测的深度学习方法

由于传统机器学习方法需要依赖人工特征学习区分目标，对于多模态 PGx 数据的特征学习显然不是最高效的。深度学习方法的基础是神经网络（neural network）。而多层神经网络模拟神经元信号传递机制，无需受到人工特征的限制，对于结构复杂的多模态数据也无需人工特征指导计算，而是根据数据特性自动分析计算最佳区分特征，使得 ADR 预测更加高效，结果更加准确。几种常用的框架如图 3 所示。

在近些年预测 ADR 的研究领域，深度学习方法展现出显著优势[36]。具体研究方法如表 3 所示，各类方法普遍采用多模态数据融合策略：从西药的分子结构到中药复方的异质网络，再到多组学、知识图谱及遗传变异信息，均被有效整合。

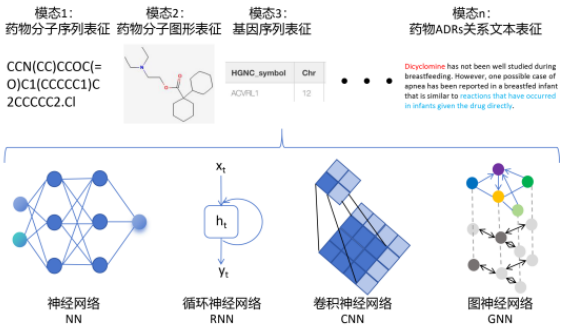

图3 面向ADR预测的深度学习方法常用框架

Figure 3. The general framework of deep learning methods for ADR prediction

在模型架构上，研究者创新性地运用了图神经网络、多头注意力机制、卷积神经网络等先进技术，不仅实现了高精度预测，还深入挖掘了药物与ADR之间的内在作用机制。这些方法在准确性和泛化性能上都优于传统机器学习方法，能够更好分析药物与ADR之间更深层次的特征关联，为提高ADR预测能力提供更加有力的技术支撑。

2.3 面向ADR预测的知识图谱与大语言模型方法

大语言模型（large language model，LLM）是深度学习技术的应用之一。其目标在于理解和生成人类语言，为了实现该目标，模型需要在大量文本数据上进行训练，以学习语言的各种模式和结构。面对庞大的ADR观测文字记录数据的分析需求，LLM无疑是可以大展身手的一项技术[43]。但ADR预测所依赖的并不完全是文本形式的数据，还包含了例如分子结构、PGx、蛋白质组学等图形和序列形式的数据，其相互之间存在一定的关联，此时常被用于描述数据关系的知识图谱（knowledge graph，KG）技术就可以弥补LLM对非文本数据分析的弱势。因此，面对药物研发中多模态数据并行处理的迫切需求，KG与LLM的融合分析近年来已成为该领域的重要研究方向。如表4所示，当前基于LLM辅助ADR分析的研究仍然以文本处理为主，深度融合KG及药物–基因–蛋白质靶点等多模态表征方法的方法尚不多见，是一个极具潜力的探索方向。

3 AI方法预测ADR研究的关键问题与未来趋势

3.1 数据质量与标准化问题

PGx数据涉及基因–基因与基因–环境相互作用以及全基因组关联研究，是一种庞大且异常复杂的组学数据；同时又与蛋白质组学、转录组、代谢组等多组学数据密切相关，其原始数据具有

表3 ADR预测深度学习方法简述

Table 3. A brief review of deep learning methods for ADR prediction

序号	方法	实验数据来源	优势	局限
1	西药模型：图神经网络表示西药分子结构，结合多层感知机提取分子指纹特征，实现ADR预测； 中药模型：基于双向编码器表征法（bidirectional Encoder Representations from Transformers，BERT）和异质图神经网络的中药复方ADR预测模型[37]	SIDER、ChEMBL、TCMSP、TCMID、HERB、STRING	①模型预测性能强，仅用分子结构特征以实现预测，适用于药物研发全流程 ②结合网络毒理学和异质图神经网络处理中药复杂性，捕获多模态关联	①模型忽略了动态相互作用 ②模型依赖关系的定义，容易错过隐含关联，需结合更多模态数据和大规模验证 ③不能覆盖罕见ADR或新药 ④模型参数多，训练需高资源，实时预测慢 ⑤模型的可解释性不足
2	基于转录组学的蛋白质组学预测器（transcriptomics-based Proteomics Predictor，TransPro）用图同构网络提取药物化学结构特征，多头注意力建模药物与细胞的交互，多层感知机实现组学特征转换，预测ADR[38]	LINCS L1000、DepMap、CCLE、FAERS、STRING	①与变分自编码器、生成对抗网络和传统多层感知机对比，TransPro在准确性和泛化性表现更优 ②采用分层设计来模拟信息流，有利于复杂疾病的研究	①对其他组织或罕见扰动的泛化能力有限 ②深度模型训练需图形处理器资源，处理大规模多组学数据时效率低 ③仅模拟静态信息流，忽略动态机制 ④缺乏大规模临床数据支撑
3	利用关系图卷积网络和关系图注意力网络，实现可解释的多模态图机器学习方法，从而预测指向与ADR预测相关的生物机制[39]	UniProt、GO、GEO、LINCS L1000、PubChem、ChEMBL、Broad Bioimage Benchmark Collection、SIDER、FAERS、STRING	①优于传统方法，支持早期临床试验决策 ②融合多种数据源（如影像和表达），能处理更全面生物机制，提高泛化能力	①罕见ADR覆盖不足 ②多模态图神经网络训练资源需求高，不适合实时预测 ③缺乏大规模临床数据支撑

续表3

序号	方法	实验数据来源	优势	局限
4	基于深度神经网络构建的自定义变体模型，采用4种不同输入组合，通过全连接多层感知机架构，解决维度压缩问题[40]	dbSNP、PharmGKB、MedDRA	①构建ADR-药物-靶点-突变网络预测模型 ②模型能分析大量化合物和基因多态性数据，适合高通量筛选	①数据不包含未表征的罕见变异 ②缺乏独立测试集或临床验证
5	提出用于同时预测多种毒性终点（肝/心/肾/神经毒性）的多任务深度学习方法[41]	组学数据、毒性数据	①多任务深度学习模型准确率较高，比传统实验方法表现优秀，能同时处理多种数据类 ②能分析复杂组学数据，标识生物标志物和新靶点，促进更安全的药物开发	①多组学数据复杂、维度多，需要高级计算工具，训练资源需求高 ②模型可解释性低，解释复杂生物交互机制仍是一个挑战 ③稀有或特异性药物反应的预测能力仍显不足
6	药物-基因-不良反应关联网络（Drug-Gene-ADR Association Network，DGANet），采用卷积神经网络提取药物和ADR的交叉特征，结合两个线性子网络处理多模态特征，通过多标签分类预测ADR[42]	CTD、SIDER、LINCS L1000、PubChem、MeSH	①提出化学结构和基因交互特征，和基因疾病相关性等新型基因组特征，提升多模态数据相关性 ②设计面向多源关联数据整合的深度学习模型	①没有纳入药物-基因交互和基因-疾病关联的多样性影响 ②模型可解释性仍不足，没有深入考虑个体因素，如性别、剂量、表型等

表4 ADR预测融合LLM与KG方法简述

Table 4. A brief review of LLM and KG methods for ADR prediction

序号	方法	实验数据来源	优势	局限
1	综述近10年有关生成式AI和LLM应用在药物相关伤害减少和警戒领域的论文[44]	PubMed，Embase，Web of Science，Scopus四个文献检索数据库，共3 988篇，最终纳入30篇	生成式AI和LLM被应用于3个关键方面：药物相互作用识别和预测、临床决策支持和药物警戒	没有研究前瞻性地测试这些模型的实验，表明当前研究仍需要进一步分析集成和现实世界的应用效果
2	从临床角度综述了ADE监测中LLM技术应用的现状[45]	文献来源于网络，包括同行评审的研究、会议论文、预印本、摘要、全文以及将LLM应用于临床数据的案例，共333篇，最终纳入39篇	研究分为面向人类用户的决策支持工具、免疫相关ADE监测、癌症相关和非癌症相关ADE监测以及个性化决策支持系统等不同方向，充分展示了LLM技术在药物警戒中的临床应用价值	特定领域的模型性能在不同领域表现会有很大差异、可解释性挑战、数据质量和隐私问题以及基础算力设施建设要求较高等
3	对将联邦学习和LLM融合用于ADR预测的论文进行综述[46]	PubMed、arXiv、IEEE和ACL Anthology中搜索获得的145篇文章	将多个机构、医院的LLM在不共享原始数据的前提下，联合训练出一个更精准的ADR预测模型，突破单一机构数据来源的局限，确保了所学习数据的广泛性	仍然存在可解释性不足，出现幻觉等问题
4	领域知识增强LLM模型[47]	英文推文中报告的ADR数据集和CSIRO ADE语料库（Cadec）数据集	通过统一的特征嵌入机制，检索外部值得信赖的知识资源并将其整合到框架中，系统地加强了ADR识别和医学术语标准化	该模型主要用于提升对语言表达文本数据中的ADR识别，不适用于对未知ADR预测的临床应用场景
5	构建一个基于元路径的单药ADR异构信息网络聚合嵌入模型，提取特征；构建一个组合药物与ADR异构网络图卷积网络，将预测药物对之间多种ADR的复杂任务转化为更易于管理的组合药物与单个ADR之间关系的预测[48]	斯坦福大学生物信息学数据库BioSNAP数据集	聚焦单药-ADR、组合药-ADR的预测研究	未对模型可解释性进行验证和提升，未涉及个性化医疗和多药物联合疗法的设计

多种来源的特征，不同公共数据集之间的数据质量和标准化差异巨大，为融合整理多组学数据带来阻碍[49]。

此外，ADR 数据通常来源于临床报告的文本数据，还有一些存在于自然语言环境中的关键信息和时序差异，如何更好地融合多组学和 ADR 文本异构数据，建立起药物的化学属性、生物作用与 ADR 的关联机制，供 AI 方法更好地分析计算，是一个亟需解决的重点问题。

3.2 多模态数据融合与深度学习模型可解释性挑战

AI 方法最擅长对标准化大数据进行高效分析，然而其对多模态数据融合分析的效能仍有待提升；另外，由于深度学习模型是建立在概率统计的基础上，对特征的定量计算与人类认知规律中对特征理解的方式有着本质差别，因此目前 AI 方法虽然通常展现出惊人的效果，却不能使人类判断其学习的特征正确与否，无法检验其运算过程的合理性与可推理性。

3.3 临床验证与转化障碍

目前 ADR 的 AI 预测方法仍然存在一些技术缺陷，例如对小数据、罕见 ADR 预测能力有限；算法的“幻觉”问题；模型更新与适应性等。由于这些技术缺陷涉及机器学习方法的科学基础，也一直是 AI 在各个领域应用中都存在的问题。这些问题的存在不可避免地要挑战临床应用的标准，可能需要各方协调平衡，才有可能真正实现技术的落地应用。

3.4 未来发展方向

（1）个性化精准用药指导：当前药物基因组学等多组学数据极大丰富了预测模型可依赖的数据基础，然而如何高效整合多组学数据，为每个患者构建一个全面的、动态的分子生物学画像，助力个性化精准用药辅助指导研究是临床应用发展的一个重要需求方向。

（2）预测模型全程可解释：受限于当前 AI 主流技术——深度学习方法的统计学数据分析特性，不可避免地存在模型特征不可解释和“幻觉”问题，如何利用当前能够搜集到的所有数据和生物学机制分析，提升 ADR 预测模型的可解释性并减少“幻觉”仍是该领域研究转化到临床应用实践中一个亟需突破的关键问题。

（3）PGx 等多组学数据的高效表征方法：ADR 预测所涉及的数据不只包括药物分子式和基因序列等一维数列，还包括二维、三维分子结构表达，另外还有自然语言表示的药物–基因–靶点–ADR 相互作用关系，这些不同信息本身具有多种数据结构和表征方式，再加上相互之间的作用关系类别多，如何设计有效的多模态特征及其相关性表征方法，让 AI 模型能够尽可能多地学习到与 ADR 相关的影响因素特性，是提升预测方法有效性的重要研究方向。

4 结语

随着基因分型技术的广泛应用和成本的降低，将 PGx 应用于临床实践被广泛认为是将基因组医学主流化的最初步骤之一，也是全球许多国家研究者关注的重点。因此，相对于 AI 技术出现之前，只能基于药物、疾病表型和 ADR 对应关系的粗放型预测方法，越来越多研究转向了对 PGx 及更多组学数据的探索。本文对基于 PGx 预测 ADR 的现有 AI 方法进行文献综述，以 PGx–ADR 关联数据集网络资源为基础，从数据整合方式和深度学习、KG 及 LLM 等先进技术相结合的角度对当前的研究现状进行总结，提出当前研究中存在的数据和技术问题，以及未来发展趋势，以期为相关领域研究者提供参考，促进研究成果向临床应用实践转化。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 俞镔 . 药物基因组学 : 指导常规用药的精准工具 [J]. 中国当代儿科杂志 , 2020, 22(11): 1143–1148. [Yu B. Pharmacogenomics: precision tool in routine prescription[J]. Chinese Journal of Contemporary Pediatrics, 2020, 22(11): 1143–1148.] DOI: 10.7499/j.issn.1008–8830.2006032.

2 国家药品不良反应监测中心 . 国家药品不良反应监测年度报告（2024 年）[R/OL]. (2025–04–07) [2025–08–20]. https://www.cdr–ADRs.org.cn/center_news/202504/t20250407_51076.html.

3 Jennings ELM, Murphy KD, Gallagher P, et al. In–hospital adverse drug reactions in older adults; prevalence, presentation and associated drugs–a systematic review and Meta–analysis[J]. Age Ageing, 2020, 49(6): 948–958. DOI: 10.1093/ageing/afaa188.

4 Cacabelos R, Cacabelos N, Carril JC. The role of pharmacogenomics in adverse drug reactions[J]. Expert Rev Clin Pharmacol, 2019, 12(5): 407–442. DOI: 10.1080/17512433.2019. 1597706.

5 Pirmohamed M. Pharmacogenomics: current status and future

perspectives[J]. Nat Rev Genet, 2023, 24: 350–362. DOI: 10.1038/s41576–022–00572–8.

6 李玉娇 , 初亚男 , 黄晓晖 , 等 . 基因多态性与药物不良反应发生风险的相关性及其临床证据 [J]. 药学进展 , 2021, 45(2): 100–111. [Li YJ, Chu YN, Huang XH, et al. Correlation between gene polymorphism and risk of adverse drug reaction and its clinical evidence[J]. Progress in Pharmaceutical Sciences, 2021, 45(2): 100–111.] https://d.wanfangdata.com.cn/periodical/CiBQZXJpb2RpY2FsQ0hJU29scjkyMDI1MTIyNDE1NDU1NRINeXhqejIwMjEwMjAwNBoIODFveXJlYzE%3D.

7 Avram S, Bologa CG, Holmes J, et al. DrugCentral 2021 supports drug discovery and repositioning[J]. Nucleic Acids Res, 2021, 49(D1): D1160–D1169. DOI: 10.1093/nar/gkaa997.

8 Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024[J]. Nucleic Acids Res, 2024, 52(D1): D1265–D1275. DOI: 10.1093/nar/gkad976.

9 Phan L, Zhang H, Wang Q, et al. The evolution of dbSNP: 25 years of impact in genomic research[J]. Nucleic Acids Res, 2025, 53(D1): D925–D931. DOI: 10.1093/nar/gkae977.

10 Gong L, Whirl–Carrillo M, Klein TE. PharmGKB, an integrated resource of pharmacogenomic knowledge[J]. Curr Protoc, 2021, 1(8): e226. DOI: 10.1002/cpz1.226.

11 Gaedigk A, Casey ST, Whirl–Carrillo M, et al. Pharmacogene variation consortium: a global resource and repository for pharmacogene variation[J]. Clin Pharmacol Ther, 2021, 110(3): 542–545. DOI: 10.1002/cpt.2321.

12 MacDonald JR, Ziman R, Yuen RKC, et al. The database of genomic variants: a curated collection of structural variation in the human genome[J]. Nucleic Acids Res, 2014, 42(D1): D986–D992. DOI: 10.1093/nar/gkt958.

13 Jeiziner C, Wernli U, Suter K, et al. HLA - associated adverse drug reactions–scoping review[J]. Clin Transl Sci, 2021, 14(5): 1648–1658. DOI: 10.1111/cts.13062.

14 Davis AP, Wiegers TC, Johnson RJ, et al. Comparative toxicogenomics database (CTD): update 2023[J]. Nucleic Acids Res, 2023, 51(D1): D1257–D1262. DOI: 10.1093/nar/gkac833.

15 Kanehisa M, Furumichi M, Sato Y, et al. KEGG: biological systems database as a model of the real world[J]. Nucleic Acids Res, 2025, 53(D1): D672–D677. DOI: 10.1093/nar/gkae909.

16 Hecker N, Ahmed J, Von Eichborn J, et al. SuperTarget goes quantitative: update on drug–target interactions[J]. Nucleic Acids Res, 2012, 40(D1): D1113–D1117. DOI: 10.1093/nar/gkr912.

17 Kim S, Chen J, Cheng T, et al. PubChem 2025 update[J]. Nucleic Acids Res, 2025, 53(D1): D1516–D1525. DOI: 10.1093/nar/gkae1059.

18 Wang Z, Clark NR, Ma'ayan A. Drug–induced adverse events prediction with the LINCS L1000 data[J]. Bioinformatics, 2016, 32(15): 2338–2345. DOI: 10.1093/bioinformatics/btw168.

19 Kim JA, Ceccarelli R, Lu CY. Pharmacogenomic biomarkers in US FDA–approved drug labels (2000–2020)[J]. J Pers Med, 2021, 11(3): 179. DOI: 10.3390/jpm11030179.

20 Moore TJ, Morrow RL, Dormuth CR, et al. US Food and Drug Administration safety advisories and reporting to the Adverse Event Reporting System (FAERS)[J]. Pharmaceut Med, 2020, 34(2): 135–140. DOI: 10.1007/s40290–020–00329–w.

21 Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects[J]. Nucleic Acids Res, 2016, 44(D1): D1075–D1079. DOI: 10.1093/nar/gkv1075.

22 Shin H, Lee S. An OMOP–CDM based pharmacovigilance data–processing pipeline (PDP) providing active surveillance for ADR signal detection from real–world data sources[J]. BMC Med Inform Decis Mak, 2021, 21(1): 159. DOI: 10.1186/s12911–021–01520–y.

23 Wu L, Ingle T, Liu Z, et al. Study of serious adverse drug reactions using FDA–approved drug labeling and MedDRA[J]. BMC Bioinformatics, 2019, 20(Suppl 2): 97. DOI: 10.1186/s12859–019–2628–5.

24 Lipscomb CE. Medical subject headings (MeSH)[J]. Bull Med Libr Assoc, 2000, 88(3): 265–266. https://pubmed.ncbi.nlm.nih.gov/10928714/.

25 Hunter FMI, Ioannidis H, Bento AP, et al. Drug and clinical candidate drug data in ChEMBL[J]. J Med Chem, 2025,68(19):19800–19827. DOI: 10.1021/acs.jmedchem.5c00920.

26 Thirumuruganathan S, Tang N, Ouzzani M, et al. Data curation with deep learning[C]. Denmark: The 23rd International Conference on Extending Database Technology, 2020: 277–286. DOI: 10.5441/002/edbt.2020.25.

27 Pan Y, Lei X, Zhang Y. Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factor, and disease networks: a comprehensive approach[J]. Med Res Rev, 2021, 42(1): 441–461. DOI: 10.1002/med.21847.

28 Wigh D, Goodman J, Lapkin A. A review of molecular representation in the age of machine learning[J]. Wiley Interdiscip Rev Comput Mol Sci, 2022, 12(5): e1603. DOI: 10.1002/wcms.1603.

29 Wei J, Chu X, Sun X, et al. Machine learning in materials science[J]. InfoMat, 2019, 1(3): 338–358. https://doi.org/10.1002/inf2.12028.

30 匡启帆 , 郭佳丽 , 李益洲 , 等 . 基于核矩阵降维算法对药物不良反应的预测 [J]. 中国科技论文 , 2017, 12(24): 2845–2849. [Kuang QF, Guo JL, Li YZ, et al. Prediction of drug adverse reactions based on kernel matrix dimension reduction method[J]. China Science Paper, 2017, 12(24): 2845–2849.] DOI: 10.3969/j.issn.2095–2783.2017.24.018.

31 Hasselt J and Iyengar R. Systems pharmacology–based identification of pharmacogenomic determinants of adverse drug reactions using human iPSC–derived cell lines[J]. Curr Opin Syst Biol, 2017, 4: 9–15. https://doi.org/10.1016/j.coisb.2017.05.006.

32 刘珂 . 基于生物大数据的基因与药物不良反应网络化关联分析与预测 [D]. 福建厦门 : 厦门大学 , 2018.

33 朱嘉静 . 基于机器学习的药物不良反应关键问题研究 [D]. 成都 : 电子科技大学 , 2020. DOI: 10.27005/d.cnki.gdzku.2020.004650.

34 杜瑶 . 基于多数据源与机器学习的药物副作用预测 [J]. 软件

导刊 , 2021, 20(5): 39–43. [Du Y. Prediction of drug side effects based on multiple data sources and machine learning [J]. Software Guide, 2021, 20(5): 39–43.] DOI: 10.11907/rjdk.201917.

35 刘佳宁 , 焦强 , 边太成 , 等 . 基于深度多核多任务学习的个性化药物不良反应预测 [J]. 计算机技术与发展 , 2024, 34(11): 148–156. [Liu JN, Jiao Q, Bian TC, et al. Personalized adverse drug reactions prediction based on deep multi–kernel multi–task learning[J]. Computer Technology and Development, 2024, 34(11): 148–156.] DOI: 10.20165/j.cnki.ISSN1673–629X.2024.0216.

36 尹璇 , 黄睿健 , 孔斯予 , 等 . 围手术期药物相关罕见不良反应发现的真实世界数据基础与方法学进展 [J]. 药物流行病学杂志 . 2025, 34(8): 917–925. [Yin X, Huang RJ, Kong SY, et al. Perioperative rare adverse reactions discovery: real–world data foundations and methodological advances[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(8): 917–925.] DOI: 10.12173/j.issn.1005–0698.202411007.

37 杨泽群 . 基于深度学习的药物不良反应预测方法的研究与实现 [D]. 南京 : 东南大学 , 2023. DOI: 10.27014/d.cnki.gdnau.2023.004124.

38 Wu Y, Liu Q, Xie L. Hierarchical multi–omics data integration and modeling predict cell–specific chemical proteomics and drug responses [J]. Cell Rep Methods, 2023, 3(4): 100452. DOI: 10.1016/j.crmeth.2023.100452.

39 Krix S, DeLong LN, Madan S, et al. MultiGML: multimodal graph machine learning for prediction of adverse drug events[J]. Heliyon, 2023, 9(9): e19441. DOI: 10.1016/j.heliyon.2023.e19441.

40 Dafniet B, Taboureau O. Prediction of adverse drug reactions due to genetic predisposition using deep neural networks[J]. Mol Inform, 2024, 43: e202400021. DOI: 10.1002/minf.202400021.

41 Son A, Park J, Kim W, et al. Recent advances in omics, computational models, and advanced screening methods for drug safety and efficacy[J]. Toxics, 2024, 12(11): 822. DOI: 10.3390/toxics12110822.

42 He M, Shi Y, Han F, et al. Prediction of adverse drug reactions based on pharmacogenomics combination features: a preliminary study[J]. Front Pharmacol, 2025, 16: 1448106. DOI: 10.3389/fphar.2025.1448106.

43 司书成 , 吴柳柳 , 王聪慧 , 等 . 大语言模型助力药物流行病学研究 [J]. 药物流行病学杂志 , 2025, 34(9): 1074–1083. [Si SC, Wu LL, Wang CH, et al. Large language models empowering pharmacoepidemiology research[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(9): 1074–1083.] DOI: 10.12173/j.issn.1005–0698.202504033.

44 Ong JCL, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication related harm[J]. NPJ Digit Med, 2025, 8(1): 182. DOI: 10.1038/s41746–025–01565–7.

45 Zitu MM, Owen D, Manne A, et al. Large language models for adverse drug events: a clinical perspective[J]. J Clin Med, 2025, 14(15): 5490. DOI: 10.3390/jcm14155490.

46 Guo D, Choo KKR. Applications of federated large language model for adverse drug reactions prediction: scoping review[J]. J Med Internet Res, 2025, 27: e68291. DOI: 10.2196/68291.

47 Zou H, Wang Y, Xiang K, et al. Knowledge–augmented large language model–based framework for adverse drug reactions analysis[J]. Appl Soft Comput, 2025, 185:114025. https://doi.org/10.1016/j.asoc.2025.114025.

48 Tian L, Wang Q, Zhou Z, et al. Predicting drug combination side effects based on a metapath–based heterogeneous graph neural network[J]. BMC Bioinformatics, 2025, 26(1): 16. DOI: 10.1186/s12859–024–06028–6.

49 Zack M, Stupichev DN, Moore AJ, et al. Artificial intelligence and multi–omics in pharmacogenomics: a new era of precision medicine[J]. Mayo Clin Proc Digit Health, 2025, 3(3): 100246. DOI: 10.1016/j.mcpdig.2025.100246.

收稿日期：2025 年 08 月 31 日 修回日期：2025 年 12 月 20 日
本文编辑：冼静怡 杨 燕