文 献 综 述

毕业设计题目：

基于卷积神经网络的药物-靶标相互作用预测

基于卷积神经网络的药物-靶标相互作用预测

LOW REN HONG

(21 计算机留学生（1）班　2021529620004)

1　引言

近年来，随着计算能力的迅速提升、大数据技术的普及以及生物医学数据的爆发性增长，药物研发这一复杂且耗时的工程得到了全新的技术支持[1]。然而，传统药物开发方法仍面临诸多挑战。为了研发出可靠且安全的药物，研究人员需要投入大量的人力和财力资源，而这些资源往往受到时间、成本以及技术能力的限制。药物开发周期漫长，从靶标发现到临床试验及最终获批，通常需要耗费数十年时间和数十亿美元的投入[2]。高失败率是药物研发的一大难题，尤其是在早期阶段，由于对药物与靶标相互作用的预测不准确，导致许多潜在药物在临床试验中被淘汰。

为了有效的解决这个问题，药物-靶标相互作用作为一个解决方案得到了广泛的关注。传统的实验筛选方法由于成本高昂且效率低下，已难以满足现代药物研发的需求。药物-靶标相互作用（Drug-Target Interaction, DTI）预测作为药物研发中的关键步骤，旨在通过预测药物与靶标之间的结合强度，有效缩小实验筛选的范围，从而加速药物开发进程。传统的 DTI 预测方法多基于统计学和机器学习，这些方法依赖于人工特征提取和设计，对数据质量和专业知识要求较高。然而，近年来深度学习技术的引入显著提高了预测的准确性和效率。特别是卷积神经网络（Convolutional Neural Network, CNN）凭借其强大的特征学习能力，能够从蛋白质序列和化合物分子表示中提取高层次特征，为 DTI 预测提供了更加全面和精准的解决方案。深度学习和机器学习在自然语言处理和计算机视觉等领域取得了卓越的表现。[3][4][5][6]

本研究旨在借助深度学习的技术，开发一个基于卷积神经网络的药物-靶标相互作用预测。

本文阐述了机器学习在药物-靶标相互作用预测中的应用，包括监督学习、无监督学习和强化学习。第二节介绍了机器学习的理论知识，基础介绍以及对于药物-靶标相互作用的算法选择。第三节，药物-靶标相互作用的数据来源，总结了药物靶标相互作用的评估指标。第 4 节，综述了国内外在药物-靶标相互作用预测领域的研究现状。最后，总结了目前的研究存在的问题和挑战，并展望了未来的研究方向。

2 机器学习

2.1 监督学习

监督学习是机器学习中通过标记数据来进行学习，通过输入和输出的关系做出准确的预测。模型通过学习样本，从而在新数据上进行准确的预测。适用于分类，回归任务等等。

在药物-靶标相互作用中，监督学习利用已知的药物-靶标相互作用数据集训练模型，以预测新的相互作用。通过监督学习，可以充分挖掘已有数据，提高预测的准确性。然而，监督学习依赖大量高质量标记数据，而药物-靶标相互作用数据集通常稀缺，这限制了其在 DTI 预测中的广泛应用。

2.2 无监督学习

无监督学习对比与监督学习，是没有标注数据，通过分析输入数据，提取有意义的信息。适用于聚类，降维等任务。[7]

在药物-靶标相互作用中，无监督学习用于探索性分析和特征提取。通过无监督学习，可以利用聚类的方法将相似的药物或靶标分组，从而推测出潜在的药物-靶标相互作用。降维技术可以帮组简化高维特征，提高计算效率增强可视化效果。由于缺乏明确的标签，预测的结果和解释性较弱，可以作为辅助工具。

2.3 强化学习

强化学习是通过与环境交互并根据奖励反馈优化决策策略的机器学习方法。强化学习的核心是智能体（Agent）通过采取动作（Action）与环境（Environment）交互，根据奖励（Reward）优化策略（Policy），以实现累计奖励最大化。

在药物-靶标相互作用中，强化学习可以用于药物分子的生成和优化。通过强化学习，可以设计智能体，通过与环境交互，生成具有特定生物活性的药物分子。强化学习的优势在于可以自动化生成药物分子，减少人工干预，提高效率。然而，强化学习的训练过程较为复杂，需要大量的计算资源和时间。

尽管强化学习在探索性任务和复杂决策问题中具有巨大潜力，但其在 DTI 预测中的应用仍面临挑战，如环境模拟的复杂性和奖励设计的难度。强化学习需要大量计算资源和时间进行训练，这限制了其在现实场景中的广泛使用。

3 药物-靶标相互作用预测

3.1 数据来源

药物靶标相互作用数据库是药物研发的重要资源，包括药物与靶标之间的相互作用信息，如药物的化学结构、靶标的序列信息、药物与靶标之间的结合强度等。药物靶标相互作用数据库的建立和维护对于药物研发具有重要意义，可以帮助研究人员快速获取药物与靶标之间的相互作用信息，加速药物研发进程。用于药物-靶标相互作用预测的数据集主要包括以下几类：ChEMBL,DrugBank,BindingDB，PubChem。ChEMBL[8]是生物活性化合物数据库，由欧洲分子生物学实验室（EMBL）的欧洲生物信息学研究所（EBI）维护和更新。主要提供化合物与靶标之间的活性数据，包括实验测量值（如 IC50、EC50、Kd）。DrugBank[9]最初由加拿大阿尔伯塔大学开发并维护，绝大多数药物靶标类型：G 蛋白偶联受体（GPCR）、酶（enzyme）、载体蛋白（transporters）和离子通道（ion channels）。BindingDB[10]是收录了小分子化合物与蛋白质靶标之间的结合数据。提供化合物与蛋白质靶标的实验亲和力数据（如 IC50、Kd、Ki、EC50）。PubChem[11]是美国国家生物技术信息中心（NCBI）维护的一个化学信息数据库，提供化合物结构、性质和生物活性信息，适合大规模筛选和基础研究。

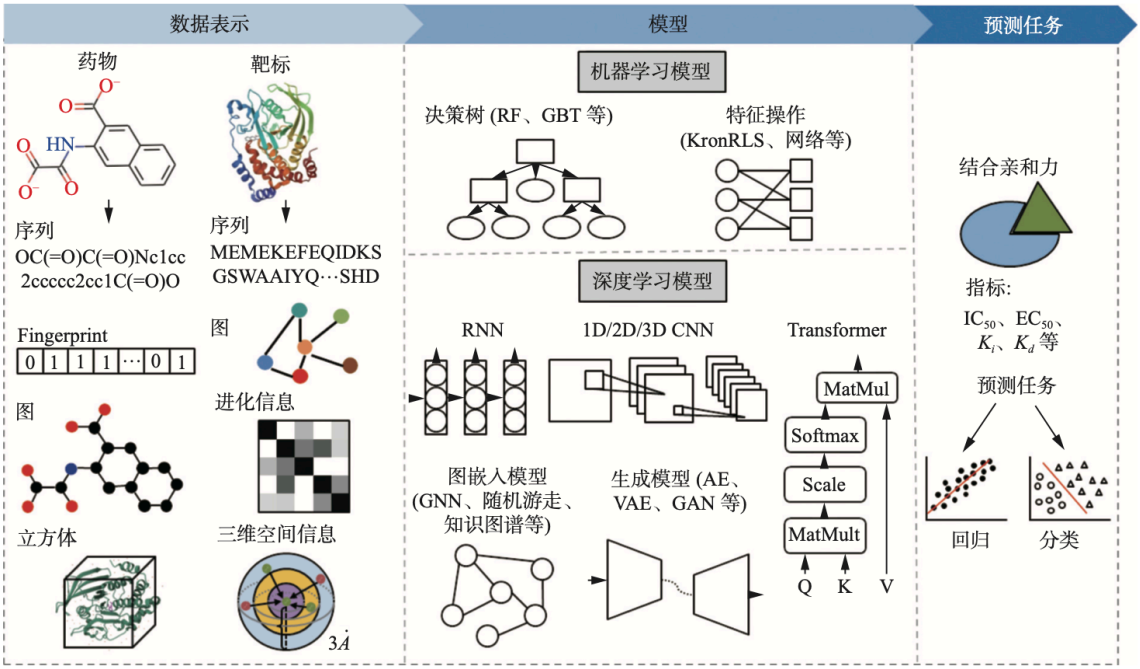

图 1 不同数据类型及其在机器学习与深度学习模型中的应用场景

3.2 评价指标

药物-靶标相互作用预测主要是二分类任务和回归任务。二分类任务需要判断药物是否存在交互，样本的预测结果分为 4 类，TP(预测值为 1，真实值为 1),FN(预测值为 0，真实值为 0),TN(预测值为 1，真实值为 0),FP(预测值为 0，真实值为 1)。常用的模型评价指标 TPR(真阳性率),FPR(伪阳性率),AC(准确率)，PR(精确率)，RC(召回率)，ROC,AUPR 等。回归任务需要预测药物与靶标之间的结合强度，可采用均方误差（MSE)、均方根误差（RMSE)、平均绝对误差（MAE）等指标。

4 国内外研究现状

4.1 国内研究现状

相较于国外，国内在药物-靶标相互作用 (DTI) 预测领域的研究起步较晚，但随着生物信息学、机器学习和深度学习技术的发展，近年来也取得了较快的进展。

首先是，DTINet[12] 是清华大学曾坚阳课题组提出的另一种重要模型，发表于 2017 年的《Nature Communications》。DTINet 提出了基于异构网络整合的创新方法，通过随机游走和特征学习算法，有效整合多种网络信息（如药物-疾病关联、蛋白质-蛋白质相似性等)，从而实现药物-靶标相互作用的高效预测。该模型不仅显著提高了预测准确性，还在药物重定位方面表现出潜力，也同时展示了异构网络在该领域的潜力。

DTINet 为国内基于异构网络的研究奠定了基础。随后，国内研究者开始引入卷积神经网络（CNN）等先进模型。2019 年，发表于《IEEE/ACM Transactions on Computational Biology and Bioinformatics》的一篇研究文章提出了一种基于卷积神经网络的预测方法**Hu2021**，使用药物描述符和蛋白质序列的特征表示，通过卷积神经网络模型进行学习与预测。这项研究在测试集中取得了优异的预测性能，进一步证明了深度学习在 DTI 预测领域的优势和潜力，其成果为后续结合注意力机制的 MCANet 提供了思路。

近年来随着注意力机制的发展，在 2023 年，东北林业大学研究团队提出了用于药物-靶标相互作用预测 MCANet[13] 。引入了基于共享权重的多头交叉注意力（Multi-headCrossAttention）模块，能够高效提取药物和蛋白质之间的双向交互特征，同时采用 PolyLoss 损失函数[14]有效缓解数据集中的过拟合和类别不平衡问题。在 DrugBank、Davis[15] 和 KIBA[16] 等公共数据集上。其中，PolyLoss 是一种专为小样本和不平衡数据设计的损失函数，能够动态调整各类别样本的贡献，从而提升模型的泛化能力。AUPRC（Area Under Precision-Recall Curve）是衡量模型在处理不平衡数据集时性能的重要指标。MCANet 在 DrugBank 数据集上的 AUPRC 提升至 0.89，表明其在小样本预测中

的可靠性。

2024 年山西农业大学与北京中医药大学合作提出了一种名为 DrugMAN[17] 的深度学习模型，该模型通过整合异构网络信息预测药物-靶标相互作用（DTI）。该模型在冷启动场景下表现出色，尤其是在药物和靶标均未出现在训练集中的情况下，模型优于 DTINet。

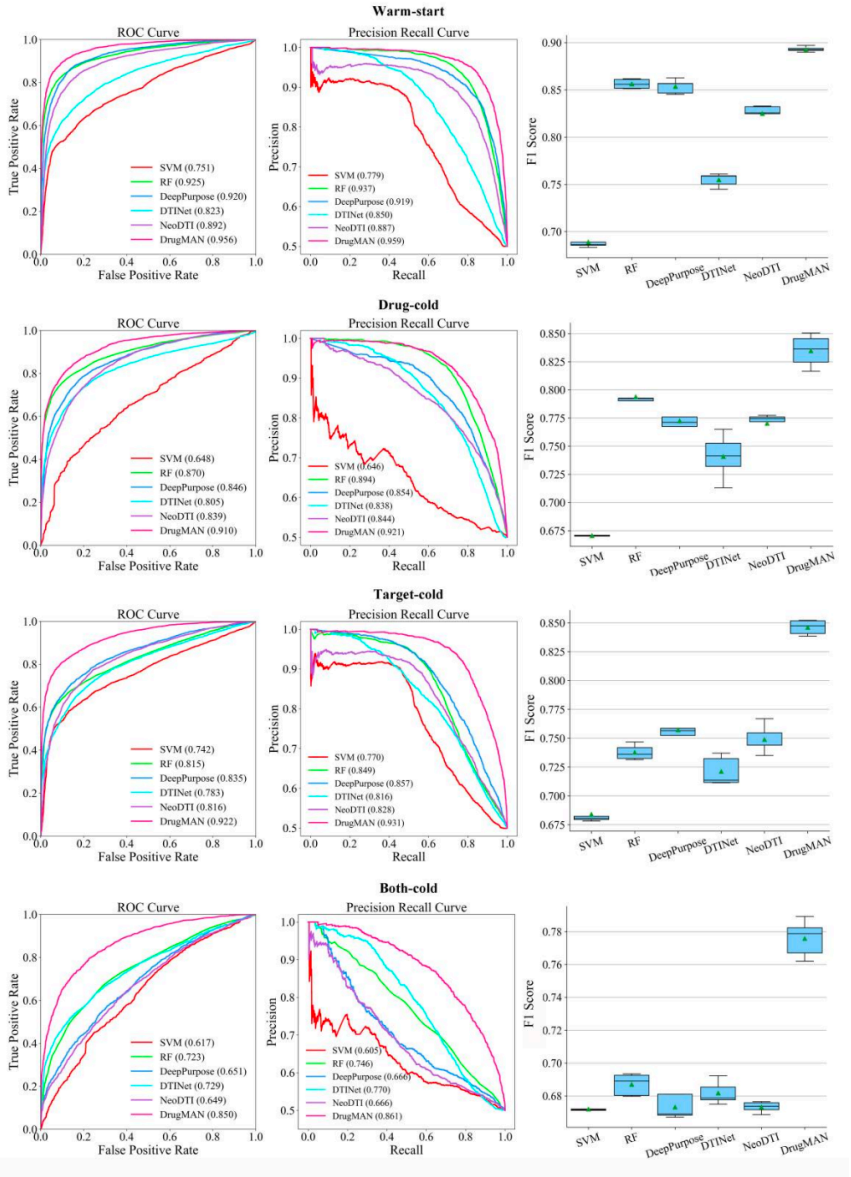

图 2 DrugMAN 与最新方法的比较

这些研究表明，国内在 DTI 预测领域正逐步从传统机器学习方法向基于深度学习和异构网络整合的创新方法转变。未来，国内研究者可以进一步探索深度学习与无监督学习结合，3D 结构特征提取等等，以提高预测的准确性和效率。

4.2 国外研究现状

药物-靶标相互作用（DTI）预测是药物发现中至关重要的一步，近年来，随着大数据和深度学习的快速发展，国外的研究也取得了显著的进展。

早期研究，DeepDTA[18]由 Öztürk et al. 在 2018 年提出的模型，旨在通过深度神经网络模型来处理药物分子的序列和靶标蛋白的序列数据，从而预测药物与靶标蛋白质的相互作用。分别采用了 KronRLS 回归算法**Pahikkala et al.,2014**和 SimBoost 方法**He et al,2017**作为基线方法。DeepDTA 是第一个使用深度学习架构（特别是卷积神经网络）来同时处理药物和蛋白质序列信息的模型，也是后续的深度学习模型提供了启发。

随着图神经网络（GNN）的快速发展，国外学者将其引入 DTI 预测领域。例如，GraphDTA[19] 由 Nguyen et al. 在 2021 年，利用分子图和蛋白质序列中的拓扑特征，通过 GNN 捕捉化学键和分子结构的信息，与 CNN 模型相比表现出色。

GraphDTA 展示了 GNN 在药物分子表示中的潜力，而 iNGNN[20] 由 Sun et al 在 2024 年，将 NGNN 和 AplhaFold2 生成的蛋白质 3D 结构，利用图表示增强序列信息的表达能力，并通过注意力机制捕捉药物与靶标之间的交互。特别是在测试集中包含从未见过的药物或靶标时，iNGNN-DTI 的表现显著优于传统模型，显示了其在冷启动场景下的鲁棒性。iNGNN-DTI 在冷启动场景下的表现如图 3 所示

Table 5. Model performance on the test set with unseen drugs and proteins.

Method	unseen drugs		unseen proteins	
	AUROC	AUPRC	AUROC	AUPRC
DeepDTA	0.736±0.0550	0.145±0.0300	0.770±0.0594	0.145±0.0300
Moltrans	0.744±0.0540	0.144±0.0370	0.778±0.0538	0.231±0.0534
ML-DTI	0.737±0.0700	0.169±0.0730	0.840±0.0357	0.259±0.0519
DGraphDTA	0.718±0.0045	0.169±0.0049	0.780±0.0478	0.164±0.0391
iNGNN-DTI	**0.763 ± 0.0490**	**0.224 ± 0.0640**	**0.867 ± 0.0357**	**0.296 ± 0.0534**

图 3 iNGNN 在冷启动下的表现

通过多种深度学习模型的不断优化，国外研究在提高预测精度、利用 3D 结构信息和实现模型可解释性方面均取得了重要进展。

5 总结与展望

本文综述了药物与靶标的现状以及药物靶标的相关研究。在提取药物和蛋白质方面，专家们通过监督学习，无监督学习，自监督学习等多个因素考虑，提出了方案和优化方法，为药物靶标相互作用提供丰富的理论和实践经验。在药物靶标相互作用预测方面，研究者们采用了多种算法，包括卷积神经网络、图神经网络、强化学习等，以解决药物与靶标之间的相互作用问题，提高了药物研发的效率。

目前的研究还存在一些问题和挑战。对于 3D 结构的蛋白质，如何更好地提取特征，以及如何更好地利用蛋白质的结构信息，仍然是一个挑战。药物-靶标相互作用预测的数据集通常稀缺，如何更好地利用有限的数据，提高预测的准确性，也是一个重要的问题。

无监督学习在药物-靶标相互作用预测中的应用仍然有待深入研究。因此，未来的研究应深入探索无监督学习与深度学习的结合，进一步提高预测模型的准确性与效率

随着大数据和深度学习的快速发展，未来研究可以从深度学习与无监督学习结合，3D 结构特征提取等等。研究者通过自监督学习和无监督预训练框架，利用大规模无标签数据增强特征表示能力，提高模型对新药物和未知靶标的预测性能。研究者还可借助 AlphaFold2 等工具生成蛋白质的高精度 3D 结构，并探索适合空间特性的新型建模算法，充分挖掘结构信息的潜力。

参考文献

[1] Abbasi K, Razzaghi P, Poso A, et al. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks[J/OL]. Bioinformatics, 2020, 36(17): 4633-4642. DOI: 10.1093/bioinformatics/btaa544.

[2] Adams C P, Brantner V V. Estimating the cost of new drug development: is it really 802 million dollars?[J/OL]. Health Affairs (Millwood), 2006, 25(2): 420-428. DOI: 10.1377/hlthaff.25.2.420.

[3] Dong S, Wang P, Abbas K. A survey on deep learning and its applications[J/OL]. Computer Science Review, 2021, 40: 100379. https://www.sciencedirect.com/science/article/pii/S1574013721000198. DOI: https://doi.org/10.1016/j.cosrev.2021.100379.

[4] Feng X, Jiang Y, Yang X, et al. Computer vision algorithms and hardware implementations: A survey[J/OL]. Integration, 2019, 69: 309-320. https://www.sciencedirect.com/science/article/pii/S0167926019301762. DOI: https://doi.org/10.1016/j.vlsi.2019.07.005.

[5] Otter D W, Medina J R, Kalita J K. A Survey of the Usages of Deep Learning for Natural Language Processing[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 604-624. DOI: 10.1109/TNNLS.2020.2979670.

[6] Khan A, Laghari A A, Awan S A. Machine Learning in Computer Vision: A Review[J/OL]. EAI Endorsed Transactions on Scalable Information Systems, 2021, 8(32): e4. https://publications.eai.eu/index.php/sis/article/view/2055. DOI: 10.4108/eai.21-4-2021.169418.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning[J/OL]. Nature, 2015, 521(7553): 436-444. https://doi.org/10.1038/nature14539. DOI: 10.1038/nature14539.

[8] Mendez D, Gaulton A, Bento A P, et al. ChEMBL: towards direct deposition of bioassay data [J/OL]. Nucleic Acids Research, 2019, 47(D1): D930-D940. https://doi.org/10.1093/nar/gky1075. DOI: 10.1093/nar/gky1075.

[9] Wishart D S, Feunang Y D, Guo A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J/OL]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082. https://doi.org/10.1093/nar/gkx1037. DOI: 10.1093/nar/gkx1037.

[10] Gilson M K, Liu T, Baitaluk M, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology[J/OL]. Nucleic Acids Research, 2016, 44(D1): D1045-D1053. https://doi.org/10.1093/nar/gkv1072. DOI: 10.1093/nar/gkv1072.

[11] Kim S, Thiessen P A, Bolton E E, et al. PubChem Substance and Compound databases[J/OL]. Nucleic Acids Research, 2016, 44(D1): D1202-D1213. https://doi.org/10.1093/nar/gkv951. DOI: 10.1093/nar/gkv951.

[12] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information[J/OL]. Nature Communications, 2017, 8. DOI: 10.1038/s41467-017-00680-8.

[13] Bian J, Zhang X, Zhang X, et al. MCANet: shared-weight-based MultiheadCrossAttention network for drug-target interaction prediction[J/OL]. Briefings in Bioinformatics, 2023, 24(2): bbad082. DOI: 10.1093/bib/bbad082.

[14] Leng Z, Tan M, Liu C, et al. PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions[EB/OL]. 2022. https://arxiv.org/abs/2204.12511. arXiv: 2204.12511 [cs.CV].

[15] Davis M I, Hunt J P, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity [J/OL]. Nature Biotechnology, 2011, 29(11): 1046-1051. https://doi.org/10.1038/nbt.1990. DOI: 10.1038/nbt.1990.

[16] Tang J, Szwajda A, Shakyawar S, et al. Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis[J/OL]. Journal of Chemical Information and Modeling, 2014, 54. DOI: 10.1021/ci400709d.

[17] Zhang Y, Wang Y, Wu C, et al. Drug-target interaction prediction by integrating heterogeneous information with mutual attention network[J/OL]. BMC Bioinformatics, 2024, 25(1): 361. DOI: 10.1186/s12859-024-05976-3.

[18] Öztürk H, Ozkirimli E, Özgür A. DeepDTA: Deep Drug-Target Binding Affinity Prediction [J/OL]. Bioinformatics, 2018, 34(17): i821-i829. https://doi.org/10.1093/bioinformatics/bty593. DOI: 10.1093/bioinformatics/bty593.

[19] Nguyen T, Le H, Quinn T P, et al. GraphDTA: Predicting Drug-Target Binding Affinity with Graph Neural Networks[J/OL]. Bioinformatics, 2021, 37(8): 1140-1147(2021-05-23). https://d

oi.org/10.1093/bioinformatics/btaa921. DOI: 10.1093/bioinformatics/btaa921.

[20] Sun Y, Li Y, Leung C K, et al. iNGNN-DTI: Prediction of Drug-Target Interaction with Interpretable Nested Graph Neural Network and Pretrained Molecule Models[J/OL]. Bioinformatics, 2024, 40(3): btae135(2024-03-04). https://doi.org/10.1093/bioinformatics/btae135. DOI: 10.1093/bioinformatics/btae135.