

• 专家论坛 •



刘 琦 同济大学生物信息系教授, 博士生导师, 入选同济大学青年“百人计划”。IEEE 会员, 美国计算机协会 (ACM) 会员, 中国计算机协会 (CCF) 高级会员 (生物信息学专业组委员), 中国人工智能学会会员 (生物信息学与人工生命专委会委员), 美中药物开发协会 (SAPA) 会员。入选上海市启明星人才计划、浦江人才计划。第七届吴文俊人工智能自然科学奖三等奖获得者。致力于计算机和生命科学的智能计算和机器学习的交叉研究, 目前基于人工智能和机器学习方法从事药物信息学、肿瘤免疫治疗以及基因编辑的小 RNA 设计等方向的研究工作。在 *Nature* 旗下刊物 *Clin Pharmacol Ther*、*Nat Commun*、*Genome Biol*、*Cell* 旗下刊物 *Trends Biotechnol*、*Trends Pharmacol Sci*, 主流的生物信息学及药物信息学期刊如 *Bioinformatics*、*J Cheminformatics*、*Chem Inf Model*、*J Comput Chem*、*Brief Bioinform*、*NAR*, 以及机器学习领域国际顶级期刊和会议如 TKDE、SDM、ICDM 等发表系列论文 70 余篇。开发相应的生物数据挖掘平台 20 余项。

DOI:10.16781/j.0258-879x.2018.08.0869

人工智能与药物研发

刘 琦*

同济大学附属第十人民医院, 同济大学生命科学与技术学院生物信息学系, 上海 200092

[摘要] 新药研发是一个系统工程, 周期长、成功率低。传统药物研发在于发现疾病相关的有效靶点, 借助各种技术进行小分子 (或大分子) 的筛选与设计。人工智能技术在医学多个领域已取得显著进展, 其在新药研发领域能整合大量高通量组学数据、网络药理学数据和图像等高维表型数据, 进行有效靶点的筛选和药物设计, 节省药物研发成本, 缩短药物研发时间。本文探讨了在新一代人工智能技术驱动下的药物发现过程, 旨在为新药研发提供参考。

[关键词] 人工智能; 新药; 药物研发; 靶点; 表型

[中图分类号] R 9 **[文献标志码]** A **[文章编号]** 0258-879X(2018)08-0869-04

Artificial intelligence and drug discovery

LIU Qi*

Department of Bioinformatics, School of Life Science and Technology, the Tenth People's Hospital, Tongji University, Shanghai 200092, China

[Abstract] The discovery of new drugs is a systematic project with long cycle and low success rate. The traditional drug discovery is to find effective targets related to diseases, and then to screen and design effective small molecules (or large molecules) using various technologies. Artificial intelligence technology has made significant progress in the medical field. In the field of new drug discovery, artificial intelligence technology can integrate a large number of high-dimensional phenotype data, including high-throughput omics data, network pharmacology data and images, so as to effectively screen therapeutic targets and design drugs, saving the costs of drug discovery and shortening the time required for drug discovery. In this article, we explored the drug discovery process driven by new generation of artificial intelligence technology, hoping to provide a reference for the development of new drugs.

[Key words] artificial intelligence; new drug; drug discovery; target; phenotype

[Acad J Sec Mil Med Univ, 2018, 39(8): 869-872]

[收稿日期] 2018-06-21 **[接受日期]** 2018-08-01

[作者简介] 刘 琦, 博士, 教授, 博士生导师.

*通信作者 (Corresponding author). Tel: 021-66540109, E-mail: qiliu@tongji.edu.cn

新药研发是一个系统工程,从靶点的发现、验证,到先导化合物的发现和优化,再到候选化合物的挑选和开发,最后进入临床研究,其周期长、成功率低。传统药物研发的起点在于发现疾病相关的有效靶点,进而明确后续的药物研发路径,可借助各种计算机辅助虚拟筛选技术、高通量组学技术,综合计算化学、物理学和结构生物学的相关知识进行小分子(或大分子)的筛选与设计。传统药物靶点的发现本质是由分子生物学和系统生物学两个学科共同驱动的。疾病的靶点、突变基因和功能蛋白的鉴定是与药物研发同等复杂的过程。因此,可以认为药物靶点是联系分子生物学与药物研发的桥梁,但其本身归属于分子生物学范畴。本文关注在新一代人工智能(artificial intelligence)技术驱动下的药物发现过程,这是一个涉及多学科交叉的研究领域。

1 药物研发的转变:从以靶点为核心到综合表型筛选

人工智能技术是否可用于新靶点的发现呢?首先,人工智能相关技术已大量应用于疾病的靶点预测、高通量数据分析和系统生物学的建模过程中。然而,人工智能技术并不是新靶点发现的核心手段,因为靶点发现属分子生物学研究范畴,但不能否认人工智能技术在靶点发现中的重要作用。*Cell* 及其子刊同期发表的针对癌症基因图谱(The Cancer Genome Atlas, TCGA)的 Pan-cancer 分析的系列报道就是一个典型案例,这些研究使用了大量高通量组学数据的分析技术和人工智能相关建模方法,挖掘和分析肿瘤相关的突变位点^[1-2],提示人工智能技术已融合到复杂疾病的靶标分析中。

此外,还可利用人工智能技术挖掘海量文献进而发现新靶点。自然语言处理(natural language processing, NLP)技术是人工智能技术的一个分支,人工智能技术在靶点挖掘领域中的应用有若干方向,面向海量文献的挖掘是人工智能技术比较容易落地的一个应用场景。如前所述,靶点发现归属于分子生物学研究范畴,仅分析文献数据并不一定能发现可靠的靶点;此外,应将靶点发现与后续的药物设计分开研究,相较于靶点发现,人工智能技术目前在药物设计层面有更好的应用和落地。

当前药物研发的范式正从单一的以靶点为

核心(target-centric)的模式向整合表型筛选(phenotypic-screening)的模式转换。传统的药物研发以靶点为核心,但复杂疾病是一个综合系统,单一或若干个靶点突变可能并非该疾病发生、发展的根本原因。如由于肿瘤系统不断进化且具有高度异质性,肿瘤的靶向药物会产生抗性。大多数复杂疾病的发病机制并不清晰,靶点也不明确,如阿尔茨海默病(Alzheimer disease, AD)至今也没有较为明确的靶点。在这种情况下,通过整合表型筛选进行药物研发越来越受到学术界和工业界的关注^[3]。表型筛选是指在不明确疾病靶点和相关作用机制(mechanism of action, MoA)的情况下,基于疾病的表型数据(phenotype data)进行药物的筛选和设计。即发现某种小分子可逆转疾病的表型,那该分子针对该疾病可能具有潜在成药性,可行后续验证;同时,小分子的确定也有助于疾病靶点的筛选,如通过反向对接技术(inver-dock)和网络药理学(network pharmacology)的方法预测小分子的结合靶点。2011—2015年,诺华公司的新药研发统计数据表明,其采用表型筛选策略进行药物筛选的增长率远高于以靶点为核心的筛选模式。

利用化合物的细胞图像数据进行药物筛选的 Imagenome 概念本质是表型筛选的一种体现,这里的表型即为图像。在新药研发领域,可以基于细胞系的转录组表型进行药物重定位,其典型代表为对博德研究所开发的 Connective Map 数据库的挖掘和应用^[4]。2011年美国 *Sci Transl Med* 杂志发表了美国加州大学旧金山分校计算健康信息学研究所所长 Atul Butte 的工作,即成功应用表型筛选的方法筛选出一种抗消化性溃疡药西咪替丁,并进行了体外和体内验证,该药物可用于治疗非小细胞肺癌^[5]。综上所述,疾病的靶点发现并非是药物研发的唯一起点。

北京大学来鲁华教授也在 JACS 中撰文指出,药物研发的发展趋势之一是在传统基于结构的药物设计基础上开展系统生物学的研发模式^[6]。未来的药物研发是一个综合多学科、多手段的系统工程,需要整合大量高通量组学数据、网络药理学数据乃至图像等高维表型数据,面对这些数据,人工智能技术将大有作为。

2 大规模标记样本：是人工智能技术的阿喀琉斯之踵？

人工智能模型在药物研发领域的应用将同样面对数据预处理、模型选择、结果评估等诸多问题，本文重点讨论在药物研发领域人工智能模型对于标记样本的依赖问题。当前的人工智能模型，特别是深度学习（deep learning）模型往往需要大量标记样本进行训练，对标记样本的需求很高。在生物医学和药物研发的应用场景下，标记样本的获取依赖于领域专家知识和实验验证，成本较高。深度学习在生物医学领域最先落地的应用场景是病理切片的图像读取，究其原因深度学习的逐层学习模式天然适合对图像这种低层特征（low-level feature）的样本进行表征学习。通过逐层网络学习，深度学习技术可自动学习图像的高层特征（high-level feature），在一定程度上避免了人工进行特征工程（feature engineering）的烦琐。同时，ImageNet 等大规模标注图像数据库也为深度学习模型的快速发展提供了可靠的训练数据来源。

深度学习模型可以应用于病理图像分析，门槛相对较低，但目前病理图像的深度学习远未达到可以替代病理医师的水平，其主要原因在于：

（1）高质量可靠的病理图像人工标记样本仍然缺乏，模型的泛化能力差；（2）各公司或论文发表的预测结果缺乏统一的独立测试数据集以及有效的评价标准；（3）在模型的方法学设计上，还需要更深入考虑病理图像的特点，并且引入先验的领域知识。

药物研发领域同样存在样本标记问题，但与病理图像领域有很大差别：（1）药物研发领域积累的数据多样化，既有高通量组学数据，又有各种表型数据（如 Image）和文本数据，整合分析多源高维的异质数据可以弥补单一数据源小样本层面存在的问题；（2）人工智能技术的发展趋势正从传统的大样本训练向小样本学习及反馈学习的模式转变。Yann LeCun 形象地比喻利用非监督学习对大量无标记样本进行分析是人工智能“这块蛋糕的主体”，而强化学习是“蛋糕中美味的樱桃”。

中国科学院上海生命科学研究院陈洛南教授团队建立了一套完整有效的基于高维数据小样本（甚至是单样本）的疾病靶点标志物筛选方法^[7]；

近年来发展的弱监督学习（如迁移学习、多任务学习、半监督学习等）、小样本学习（one/few-shot learning）、零样本学习（zero-shot learning）也逐渐应用于药物研发领域。本团队基于迁移学习、多任务学习、半监督学习等小样本/弱监督学习方法在药物虚拟筛选和组合用药预测领域进行了一定的探索和尝试^[8-9]。斯坦福大学 Altae-Tran 等^[10]近期也尝试使用小样本学习来进行小样本药物发现（low data drug discovery）；而强化学习已被应用于小分子的生成设计^[11]；上述工作均是药物研发领域中面向小样本进行的有益尝试。

人脑对于客观事物的理解并不一定需要大样本训练，很多时候基于简单的类比即可进行学习。DeepMind 公司最近在 *Nat Neurosci* 上发表论文探讨大脑如何在少量的经验下进行学习，即“元学习”（meta-learning）或“学习如何学习”（learning to learn），而对于元学习模式的理解是达到通用智能的重要途径之一^[12]。2018 年 3 月底，*Nature* 发表了基于人工智能进行药物逆合成路线设计的工作，完全借鉴了 AlphaGo 的思想，而 AlphaGo 之后的 AlphaGo zero 版本则实现了不需要依赖训练样本的对弈模式，在药物研发领域这种模式有望发生根本性的改变^[13]。

总而言之，对大规模标记样本的依赖并不能说是药物研发中应用人工智能技术的阿喀琉斯之踵，其原因如下：（1）药物研发是一个大量多源多层面数据共存的场景，多源数据的整合分析非常重要，同时也可以弥补单一数据源的小样本问题；（2）小样本学习的发展是人工智能技术发展的重要方向，相信未来新的学习范式能在药物研发领域落地。

3 药物研发+人工智能：从数据到治疗还有多远？

药物研发的每个阶段都有多种可用的方法和技术，各自优缺点并存，而人工智能技术可应用于药物研发的各个层面（这里特指靶点筛选，小分子筛选、设计、合成，成药性评估等验证前阶段。大分子药物设计较为复杂，不在此次讨论范畴）。传统的计算机辅助药物设计技术更偏向于以靶点和结构信息为核心的计算机辅助设计，如传统的定量构效关系（quantitative structure-activity relationship, QSAR）模型的构建、基于结构的虚

拟筛选 (virtual screening) 等, 而人工智能技术已逐渐跳出以靶点为核心的传统研发模式, 面对海量、多源、异质性的数据, 其在药物研发中的应用是以数据为核心驱动的。

综上所述, 人工智能技术几乎涵盖药物研发实验前的所有步骤。药明康德曾提出药物研发的VIC模式, 即“VC (风险投资) + IP (知识产权) + CRO (研发外包)”, 其也可以拥有另外一种解读, 即“Virtual (虚拟) + IP (知识产权) + Capital (资本)”。未来的药物研发可在资本的介入和知识产权的保护下, 在实验验证前最大程度虚拟化、人工智能化, 由计算机评估药物成药的各个指标, 最大程度降低失败率, 通过选取最可能成药的小分子进入后续实验和临床验证, 节省药物研发成本, 缩短药物研发时间。

编者按: 本文内容来自于刘琦教授于2017年5月在知识分子微信公众号发表的相关讨论, 稍作修改, 特收录于本期“人工智能与医学”专刊。

参考文献

- [1] BAILEY M H, TOKHEIM C, PORTA-PARDO E, SENGUPTA S, BERTRAND D, WEERASINGHE A, et al. Comprehensive characterization of cancer driver genes and mutations[J]. *Cell*, 2018, 173: 371-385.
- [2] WAY G P, SANCHEZ-VEGA F, LA K, ARMENIA J, CHATILA W K, LUNA A, et al. Machine learning detects pan-cancer Ras pathway activation in the Cancer Genome Atlas[J]. *Cell Rep*, 2018, 23: 172-180.
- [3] HAASEN D, SCHOPFER U, ANTCZAK C, GUY C, FUCHS F, SELZER P. How phenotypic screening influenced drug discovery: lessons from five years of practice[J]. *Assay Drug Dev Technol*, 2017, 15: 239-246.
- [4] LAMB J, CRAWFORD E D, PECK D, MODELL J W, BLAT I C, WROBEL M J, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease[J]. *Science*, 2006, 313: 1929-1935.
- [5] SIROTA M, DUDLEY J T, KIM J, CHIANG A P, MORGAN A A, SWEET-CORDERO A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data[J/OL]. *Sci Transl Med*, 2011, 3: 96ra77. doi: 10.1126/scitranslmed.3001318.
- [6] PEI J, YIN N, MA X, LAI L. Systems biology brings new dimensions for structure-based drug design[J]. *J Am Chem Soc*, 2014, 136: 11556-11565.
- [7] LIU R, YU X, LIU X, XU D, AIHARA K, CHEN L. Identifying critical transitions of complex diseases based on a single sample[J]. *Bioinformatics*, 2014, 30: 1579-1586.
- [8] SUN Y, SHENG Z, MA C, TANG K, ZHU R, WU Z, et al. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer[J/OL]. *Nat Commun*, 2015, 6: 8481. doi: 10.1038/ncomms9481.
- [9] LIU Q, ZHOU H, LIU L, CHEN X, ZHU R, CAO Z. Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: an in-silico study[J/OL]. *BMC Bioinformatics*, 2011, 12: 294. doi: 10.1186/1471-2105-12-294.
- [10] ALTAE-TRAN H, RAMSUNDAR B, PAPPU A S, PANDE V. Low data drug discovery with one-shot learning[J]. *ACS Cent Sci*, 2017, 3: 283-293.
- [11] OLIVECRONA M, BLASCHKE T, ENKVIST O, CHEN H. Molecular de-novo design through deep reinforcement learning[J/OL]. *J Cheminform*, 2017, 9: 48. doi: 10.1186/s13321-017-0235-x.
- [12] WANG J X, KURTH-NELSON Z, KUMARAN D, TIRUMALA D, SOYER H, LEIBO J Z, et al. Prefrontal cortex as a meta-reinforcement learning system[J]. *Nat Neurosci*, 2018, 21: 860-868.
- [13] SEGLER M H S, PREUSS M, WALLER M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. *Nature*, 2018, 555: 604-610.

[本文编辑] 曾奇峰