药物表示学习综述

S191000860 俞琳荟

摘要

药物的化学成分是药物生物活性的物质基础，它通过键合相应的受体发挥各种生理功能，因此学习到药物的表征信息对于研究药物与受体的关系意义重大[1]。与传统的机器学习方法不同，利用深度学习的方法提取药物的特征，将药物投射到低维的表示空间，是从“原始”的药物去获得的而不是手动输入分子描述符，是一种自动学习获取药物表征的方法，减少了大量的人力与物力，避开了繁琐的人工输入表征的过程。本文将现有的药物表示学习方法划分为 5 类:（1）基于 SMILES 字符串的药物表征学习（2）基于分子指纹的药物表征学习（3）基于 2D 分子图片的药物表征学习（4）基于分子拓扑图的药物表征学习（5）基于 3D 结构的药物表征学习

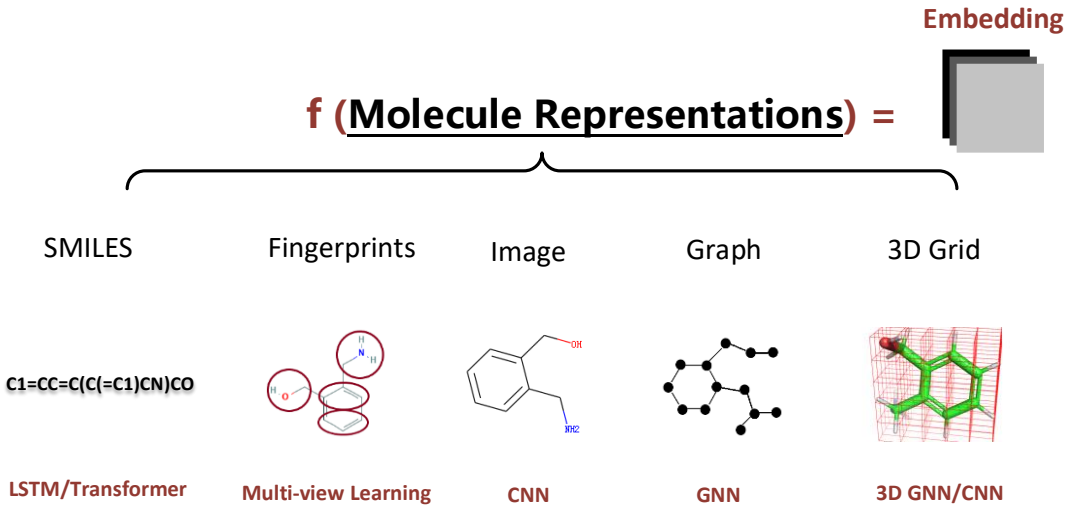

图 1 药物表示学习分类及学习方法

介绍

新药的发现和开发是一个漫长、昂贵、具有挑战性且效率低下的过程，平均需要 10 到 15 年的时间，研制过程中任何的失败都将造成巨大的财务损失[2]。然而，失败其实并不罕见，很多药物可能在进入临床二三期的阶段宣告失败，导致前期投入付之一炬。随着信息技术的不断发展，人们对信息的收集逐步重视起来，化合物信息、药物信息和临床医疗的数据集越来越多。这些数据的数量和复杂性带来了新的可能性，例如研究个性化医疗方案，与此同时也带来了新的挑战。计算机速度和功能的增强使大数据分析更加可行，然而

我们不可能仅通过经典的统计方法和人工评估来处理这些大量数据并找到合适的模式。

在过去的 10 年中，深度学习方法在语音识别、计算机视觉、自然语言处理和数据挖掘等不同的领域均取得了显著的成功。药物化学领域的研究人员也尝试将新兴的深度学习方法应用于药物开发过程，降低药物开发的成本，缩短药物开发的周期[3]。深度学习方法有别于传统的机器学习方法，它能够在大量标注数据的监督下自动学习数据的表示，从而实现数据特征的自动提取，绕开传统机器学习方法的特征工程。深度学习方法的核心在于使用神经网络模块自动从分子结构或保留大量原始结构信息的描述符中直接学习分子的低维稠密表示向量。

在 2005 年，Merwirth 等人提出直接从药物的分子图中学习对分子拓扑结构高度敏感的自适应的“描述符”[4]，而不是使用预定义的描述符，这一工作可以视作药物表示学习的“先驱”。目前而言，对于药物表示学习这一研究工作，总的大体方向可分为两个“基于文本的药物表示学习”与“基于图像的药物表示学习”。

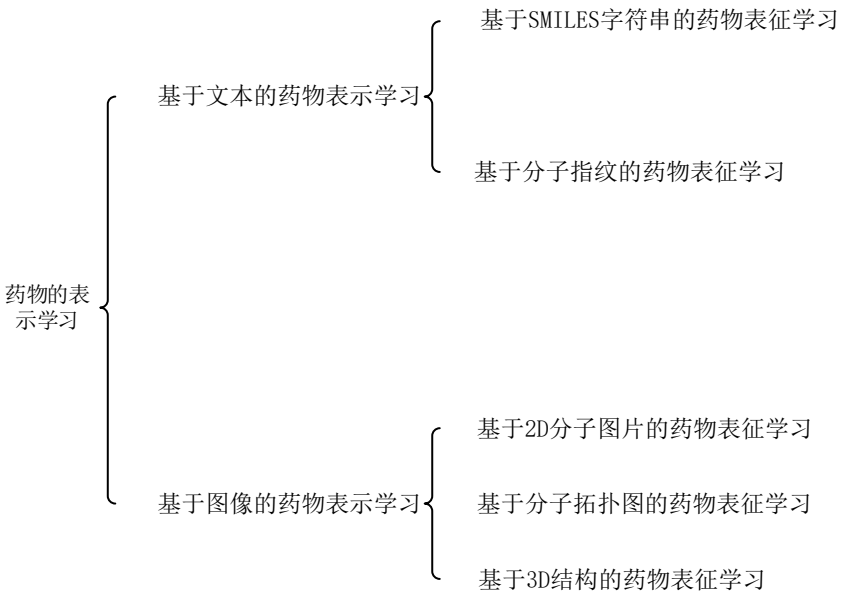

图 2 药物表示学习分类

基于文本的表示学习

基于文本的药物表示学习是基于药物用字符串表示来实现的，可以细分为“基于 SMILES 字符串的药物的表征学习”和“基于分子指纹的药物表征学习”。

（1）基于 SMILES 字符串的药物的表征学习

由于大多数药物包含少于 100 个重原子，所以具有相对较小的结构空间，因此药物分

子可以用其线性符号很好地表示。SMILES(简化分子线性输入规范)，是一种用 ASCII 字符串明确描述分子结构的规范,主要用了一些短的字符串无歧义地来表征化学分子的结构信息，这样就将药物分子转化为了文本形式[5]。基于 SMILES 表达式的药物的表示学习方法大多用于解决药物开发过程中的药物从头设计问题。所谓的药物从头设计问题,指的是高通量筛选或虚拟筛选得到的候选小分子化合物虽然对靶标具有理想的生物活性,但其水溶性、毒性等相关性质却并不令人满意,需要在保证生物活性基本不变的情况下重新合成具有期望生物活性和理想性质的化合物分子[6]。药物的从头设计问题可以类比于文本生成问题。最近的研究证明，当前的深度学习技术可以根据其线性表示准确预测结构特性[7]。近期，中山大学杨跃东教授团队[8]利用 BiLSTM 学习药物分子上下文之间的联系，进一步研究了药物的结构特点，并且加入了自注意力机制对药物的虚拟筛选增加了可解释性。

（2）基于分子指纹的药物表征学习

分子相似性是指两个分子或化合物在结构上的相似程度[9]。在相似度计算中，最常用的抽象就是分子指纹，是将分子表示为一串易于进行数值化比较的比特位序列。生成分子指纹的方式多种多样，有基于子结构的、基于药效团的、基于蛋白-配体相互作用的等。在化学分析中，基于相似性的物化性质分析、相似相溶原理等都是以“具有相似结构的化合物一般也具有相似的物理化学性质”为基础[10]。分子结构转换为分子指纹，这种抽象的表示使得处理和比较分子间的化学结构更有计算效率，两个分子之间的相似性可以用简单的相似性度量。DeepCPI 就是利用分子指纹加上深度学习“end-to-end”的方法进行化合物相似性度量的一个模型[11]，通过它预测化合物与蛋白质是否具有相互作用的可能，并取得了不错的结果。药物发现和虚拟筛选中研究非常广泛且相对成熟的定量构效关系，都离不开分子相似性的概念。

基于图像的药物表征学习

总所周知，化学分子的一个重要表现形式就是分子图，因此在图表征学习研究炙手可热的当下，我们也考虑利用利用分子图来学习药物的表征。

（1）基于 2D 分子图片的药物表征学习

ADMET （药物的吸收，分配，代谢，排泄和毒性）药物动力学方法是当代药物设计和药物筛选中十分重要的方法，通过已知药物的 ADMET 属性预测新药未知的 ADMET 属性将极大的减少药物研发过程中临床试验的投入[12]。卷积神经网络（CNN）是深度学习中最具代表

性的体系结构之一，在许多领域尤其是图像分类和目标检测中被广泛采用。在过去的几年中，CNN 在药物发现领域引起了越来越多的关注。Python 的第三方软件包 RDKit 是生物信息学常用的一款软件，它可以根据化合物的 SMILES 字符串画出分子的棍棒图并存为图片格式。基于此 Shi 等人设计了基于分子二维图像的 CNN 方法[13]，并用该方法建立 ADMET 属性的预测模型。这也是利用的分子相似性的特点，拥有相同 ADMET 属性的药物分子在结构图片上往往是相似的。

（2）基于分子拓扑图的药物表征学习

药物的分子图可以看作是典型的拓扑图——原子即节点、化学键即边，这样就可以利用图表征学习的方法处理药物分子图。从药物分子图的一个原子节点出发，通过归纳与它以化学键相连的原子节点的信息后更新自身信息，最终学习到的向量即为药物的表征[15]。近期提出的 MR-GNN 模型就是利用了药物分子图的拓扑结构特征预测 DDI（药物-药物相互作用）问题[15]，研究人员将药物分子的结构化实体表示为图，然后使用图卷积操作从每个单独的图提取特征，最后总结每个图的局部特征，并提取成对图之间的交互特征以探求到药物对之间产生相互作用的特征。

（3）基于 3D 结构的药物表征学习

化合物与蛋白质能否相互作用、如何相互作用的，这都与化合物和蛋白质能否对接在一起有很大的关系，只有能对接上的才能进一步产生相互作用。化合物与蛋白质的关系类似“钥匙”和“锁”的关系，有的钥匙可以开好几把锁，有的锁可以有很多把钥匙，有的钥匙暂时没有找到它可以开启的锁，而有的锁可能连锁孔都没有，实际上我们体内很多蛋白质暂时都未发现可以“开启”它们的化合物。由于 3D 结构复杂难以分析，因此即使 3D 结构包含了药物结构最多的信息，利用药物的 3D 结构去学习药物表征的方法目前不多。主流思路是将 3D 结构离散到网格中，然后将结构信息转为 2D 的位置矩阵后，再利用图表征学习的模型去学习药物的表征。Wallach 等人开创了用 3D 结构预测的先河[21]，他们记录下结合体在 3D 网格里周围的复杂结构，然后将 3D 卷积神经网络（CNN）应用于分类任务。然而他们的方法中 3D 矩形网格表示包含大量冗余网格点，这些网格点对应于没有原子存在的空隙空间，从而导致计算效率低下。于是，Lim 等人[21]提出了一种新的深度学习方法预测药物-目标相互作用的图形神经网络，该模型引入距离感知图注意力算法来区分不同类型的分子间相互作用，并从蛋白质-配体结合位姿的三维结构信息中直接提取分子间相互作用的图形特征。因此，该模型可以学习准确预测药物-靶相互作用的关键特征，而不只是记住配体分子的某些模式，所以，该模型对于药物-受体分子对接的虚拟有重要意义。

挑战

我们可以看到深度学习在推动药物表示学习方面已取得了喜人的进展，但是仍然存在很多挑战需要我们去解决。

首先是对新描述符的生成以及对化学系统正确表示的补充标准的进一步研究对于不久的将来至关重要。我们现在很多的药物描述符生成的标准都是统一的，但其实针对不同类别的药物，应该要根据其不同的特点去研究不同的标准，要区别对待而不是一概而论，这样才能保留不同药物的不同属性。

其次，深度学习的一个很大特点就是需要大量的数据集去训练才能得出更好的效果，然而，由于我们的数据集都来源于临床，而生物医学又很复杂，我们目前有记录的数据其实是很有限的，也就是说可用于训练的样本相较于其他领域来说其实并不是很多。这就导致了我们需要进一步研究如何优化模型，使其在训练集较小的学习上也可以做的很好[18]。

最后，针对药物的 3D 结构还需要跟进，我们目前对于 3D 结构的利用受限于它结构信息复杂的特点，虽然 2D 结构在目前的研究中已取得不错的成绩，但忽略药物的三维结构信息有可能会造成信息损失[19]，这一点亟待解决。

结论

传统的机器学习技术在提取或工程设计中依赖于深加工的功能。例如，为了在图像分类中获得不错的结果，必须应用几种预处理程序，例如过滤器，边缘检测等。深度的最大优点是，如果有足够（有时为数百万）的训练样本可用，则可以从数据中自动学习到很多特征。传统的机器学习方法通过分解问题，而深度学习则采用端到端的方法来解决问题，无需将编码训练和解码训练分解为两个单独的步骤[20]。尽管深度学习相对于其他传统机器学习方法有很大的优势，但是我们不能断言深度学习就一定比机器学习好。例如，当问题涉及化合物或目标蛋白输入描述符的组合集时，深度学习的结果与机器学习没有显着差异[21]。因此，对于药物表示学习的某些方面，我们可以将深度学习与传统的机器学习方法结合起来以共同取得新进展。

参考文献

[1] 曹东升. 化学生物信息学新方法及其在医药研究中的应用[D].中南大学,2013.

[2] Lo Y C, Rensi S E, Torng W, et al. Machine learning in chemoinformatics and drug discovery[J]. Drug discovery today, 2018, 23(8): 1538-1546.

[3] Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery[J]. Drug discovery today, 2018, 23(6): 1241-1250.

[4] Merkwirth, Christian & Lengauer, Thomas. (2005). Automatic Generation of Complementary Descriptors with Molecular Graph Networks. Journal of chemical information and modeling. 45. 1159-68. 10.1021/ci049613b.

[5] 孔德毅. 分子相似性网络中关键化合物发现算法研究[D].兰州大学,2015.

[6] 陈鑫,刘喜恩,吴及.药物表示学习研究进展[J/OL].清华大学学报(自然科学版):1-10[2019-12-30].https://doi.org/10.16511/j.cnki.qhdxxb.2019.21.038.

[7] Shuangjia Zheng, Xin Yan, Yuedong Yang, and Jun Xu. Identifying structure-property relationships through smiles syntax analysis with self-attention mechanism. Journal of chemical information and modeling, 2018.

[8] Zheng S, Li Y, Chen S, et al. Predicting Drug Protein Interaction using Quasi-Visual Question Answering System[J]. bioRxiv, 2019: 588178.

[9] 郑明月. 数据驱动的药物设计方法学及应用研究[C]. 中国化学会.中国化学会第 14 届全国计算(机)化学学术会议暨分子模拟国际论坛会议手册.中国化学会:中国化学会,2017:46.

[10] 唐玉焕,林克江,尤启冬.基于 2D 分子指纹的分子相似性方法在虚拟筛选中的应用[J].中国药科大学学报,2009,40(02):178-184.

[11] Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. Bioinformatics, 2018, 35(2): 309-318.

[12] 邓芳芳. 药物分子的计算机辅助理论模拟及分子设计[D].兰州大学,2014.

[13] Shi T, Yang Y, Huang S, et al. Molecular image-based convolutional neural network for the prediction of ADMET properties[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 194: 103853.

[14] Li J, Cai D, He X. Learning graph-level representation for drug discovery[J]. arXiv preprint arXiv:1709.03741, 2017.

[15] Xu, Nuo & Wang, Pinghui & Chen, Long & Tao, Jing & Zhao, Junzhou. (2019). MR-GNN: Multi-

Resolution and Dual Graph Neural Network for Predicting Structured Entity Interactions. 3968-3974. 10.24963/ijcai.2019/551.

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

[17] Lim, Jaechang & Ryu, Seongok & Park, Kyubyong & Choe, Yo & Ham, Jiyeon & Kim, Woo. (2019). Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. Journal of Chemical Information and Modeling. 59. 10.1021/acs.jcim.9b00387.

[18] Alcaro S, Bolognesi M L, García-Sosa A T, et al. Multi-target-directed ligands (MTDL) as challenging research tools in drug discovery: From design to pharmacological evaluation[J]. Frontiers in chemistry, 2019, 7: 71.

[19] Kelm J M, Lal-Nag M, Sittampalam G S, et al. Translational in vitro research: Integrating 3D drug discovery and development processes into the drug development pipeline[J]. Drug discovery today, 2019, 24(1): 26-30.

[20] Manchanda S, Anand A. Representation learning of drug and disease terms for drug repositioning[C]//2017 3rd IEEE International Conference on Cybernetics (CYBCONF). IEEE, 2017: 1-6.

[21] Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects[J]. Drug discovery today, 2019.