•专家论坛•

深度学习在药物设计与发现中的应用

李 伟[2], 杨金才[1], 黄 牛[1,3*]

(1. 北京生命科学研究所, 北京 102206; 2. 瑞璞鑫 (苏州) 生物科技有限公司, 江苏 苏州 215123;
3. 清华大学生物医学交叉研究院, 北京 102206)

摘要: 在新药创制的药物设计与发现所采用的多种技术中, 深度学习仍处于初级阶段, 但近年来以其独有的特点, 开始应用于虚拟化合物库的生成, 化合物活性、代谢和毒性的预测, 以及有机合成反应预测等多个方面。与传统的机器学习方法相比, 深度学习的预测能力无明显优势, 但其无需人工归纳总结数据特征, 而是具有学习能力, 自动提取特征。与基于第一性原理的计算化学相比, 深度学习虽然因为对标注明晰的大数据集的依赖, 存在泛化能力的不足, 但其以原子为中心进行卷积的表征开始助力计算化学。深度学习作为新兴技术发展迅速, 不依赖于大量标注数据的非监督学习等方法在逐渐完善, 有望能更好地助力新药研发。

关键词: 新药研发; 深度学习; 机器学习; 计算化学; 全新药物设计

中图分类号: R916 **文献标识码**: A **文章编号**: 0513-4870(2019)05-0761-07

Deep learning in drug design and discovery

LI Wei[2], YANG Jin-cai[1], HUANG Niu[1,3*]

(*1. National Institute of Biological Sciences, Beijing 102206, China; 2. RPXDs (Suzhou) Biotechnology Co., Ltd., Suzhou 215123, China; 3. Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing 102206, China*)

Abstract: Among various technologies used in drug design and discovery, deep learning is still in its infancy. Recently, deep learning approaches have been rapidly developed and applied to address various problems in drug discovery, including generation of virtual compound library, prediction of compound activity, metabolism and toxicity, and prediction of organic synthesis routes. Compared with the traditional machine learning methods, the prediction power of deep learning did not show significant improvement. However, proactively learning and automatically feature extraction bring advantages for deep learning approaches. Compared to first principle-based computational chemistry methods, deep learning can not be generalized because it depends on large-scale and high-quality annotated data sets. But its molecular representation with single-atom atomic environment vectors could be useful for computational chemists. As an emerging technology, deep learning, especially the unsupervised learning method that does not rely on large datasets with labels, is gradually improving. It is expected that someday deep learning method will become practical for drug discovery.

Key words: drug discovery; deep learning; machine learning; computational chemistry; *de-novo* design

1 新药研发寻求新技术

从20世纪90年代末到21世纪初, 全球新药研发成功率持续下滑, 进入临床Ⅰ期的化合物90%都以失败告终。2007至2010年间每年仅20个左右的新药 (指新分子实体, 下同) 获得FDA批准上市[1-3]。面对每个新药平均20亿美元的开发费用和10年以上的研发周期[4,5], 各大制药公司寻求的突破口之一是开始从大众病转向专科病药物开发[2], 近年来新药研发成功率开始回升, 2017年和2018年FDA批准的新药各达46个和59个[6]; 另一个突破口则是寻找用于新药研发的

收稿日期: 2019-03-20; 修回日期: 2019-03-27.
*通讯作者 Tel: 86-10-80720645, Fax: 86-10-80720813,
E-mail: huangniu@nibs.ac.cn
DOI: 10.16438/j.0513-4870.2019-0189

新技术, 如高通量筛选[7]、DNA 编码化合物库[8]、计算机辅助药物设计[9]和人工智能[10]等。得益于迅猛增长的计算能力和大数据集的兴起, 深度学习作为人工智能里机器学习的分支, 开始应用于生物医药领域[11-13], 其中先导化合物的发现与优化是进展最快的领域之一。

2 深度学习

深度学习 (deep learning) 是基于机器学习中的人工神经网络 (artificial neural network, ANN) 发展而来。McCulloch 等于 1943 年最早提出采用数学模型模拟人类大脑的概念, 在此基础上, Rosenblatt 于 1958 年首次发表了人工神经网络用于模式识别的算法, 而 20 世纪 70 年代初, Baskin 等[14]便已将其应用于药物设计。早期人工神经网络只包含三层, 输入层、输出层以及中间的隐含层。单层隐含层表达能力有限, 只能应用于识别手写数字等简单任务中, 多层隐含层由于梯度消失等问题难以训练, 人工神经网络进展停滞[15]。近年来, 随着新方法的出现和计算能力的提升, 如图形处理器 (graphics processing units, GPU) 的应用[16]等, 促使人工神经网络向深度学习发展。

2006 年, Hinton 和 LeCun 等[17,18]在人工神经网络的基础上, 提出深度学习的算法, 用于数据的维度降低和特征提取。深度学习包含大量隐含层, 因此深度学习能够处理大量未经人工深度加工的数据, 通过不断学习从而自动提取特征来进行表征、模拟和分类。深度学习的参数量巨大, 如果没有方法的改进, 即使使用 GPU 也难以训练。首先是使用 ReLU 激活函数, 使梯度反向传播中不会随网络层数的增加而减小, 解决了梯度消失问题[15]; 其次是 Mini-Batch 随机梯度下降法每次基于随机选取的小批量样本更新参数, 在收敛稳定性和收敛速度间取得平衡, 节约计算量, 同时解决大数据集无法同时加载到内存的问题; 最后是一系列正则化 (regularization) 方法解决过度拟合的问题, 包括 Dropout、L1 和 L2 范数、数据增强等, 增加模型的泛化能力, 其中, Hinton 等[19]提出的 Dropout 方法是神经网络特有的正则化方法, 在深度学习的训练中, 随机丢掉隐含层部分节点的输出, 从而避免前后神经元相互之间过于共适应, 在视觉、语言辨识和文件分类等多种测试场景中, 都达到显著减低过拟合风险的效果[20]。

3 大数据

深度学习需要大量数据作为训练集, 大数据使构建大规模高质量训练集成为可能。Deng 等[21]搭建的海量数据库 ImageNet 是标志性事件。就先导化合物的发现与优化而言, 海量化合物数据是关键的大数据集[22]。ChEMBL 数据库[23]持续从已发表文献、公开的数据库和临床试验数据等多个渠道收集化合物信息, 截止 ChEMBL_24, 数据库中有 180 多万个不同结构的化合物, 1 500 多万个生物学活性数据点, 由 100 多万种不同生物学测定方法产生, 覆盖 12 000 多个生物学靶标, 收集了近 7 万篇文献。除专注于小分子化合物的数据库, RCSB 数据库拥有接近 13 万个生物大分子晶体结构, 其中蛋白质晶体结构数目达 12 万个, 包含酶、G 蛋白偶联受体和离子通道等[24], 其中蛋白和小分子的复合物结构可供深度学习对小分子和蛋白的分子识别模式进行训练。除普适性的公众数据库, 还有多个专为深度学习而开发搭建的数据库。基于深度学习在图像处理上的强大能力, Cell Image Library 提供上万的化合物处理细胞后不同图像和形态学数据[25], 以供寻找新的药物作用新机制。除此之外, 工业界也逐渐提供内部的实验数据。阿斯利康的研发人员搭建了 Chemistry Connect, 包含 4 500 万个结构不同的化合物及对应的生物学或物化性质的数据, 其中 1 亿 5 千万个数据点都是来自阿斯利康内部的数据[26]。

不过现阶段用于深度学习的数据仍有不少问题。首先, 已有的可供学习的新药数据有限, 目前只有 1 600 个左右被 FDA 批准的新药[27]。其次, 由于各大制药公司对内部数据的产权保护, 公司之间以及与学术界的资源和数据共享, 仍处于起步阶段[28]。再次, 不同测试方法产生的数据需要进一步的归纳分类。最后, 来源于公开发表的文献和专利的数据质量参差不齐, 存在一定的错误信息和结论[29], 可能会误导深度学习。

4 药物设计与发现的实践

深度学习擅长的是对已有知识的挖掘, 在海量的数据中寻找已有知识的关联性。BenevolentAI 公司宣称基于其深度学习技术发现强生公司开发治疗注意缺陷多动障碍而失败的化合物 Bavisant 对帕金森症患者的日间极度嗜睡症 (EDS) 可能有疗效, 因此开展了 Phase IIb 的验证性临床试验[30]。除了基于文献专利等已公开数据的分析外, 来自杨森等公司和学校的研究人员利用传统的高通量筛选针对糖皮质激素受体的细胞模型筛选了 50 万个化合物, 获得化合物的细胞表型图像数据, 生成基于图像的分子指纹, 同时结合化合物之前在 500 多种不同靶标的筛选模型中测定的生物学活性作为训练集, 利用深度神经网络 (deep neural networks, DNN) 进行训练, 然后可以根据化合物在糖皮质激素受体的细胞表型图像数据, 来预测化合物对其他不相关靶标的生物学活性数据, 从而发现新的活性化合物[31]。这意味着单个高通量细胞表型图像筛选模型可以取代多个耗时耗力构建的特定靶标和通路的筛选模型, 显著降低人力和时间成本。Insilico Medicine

公司利用678种不同药物分别处理3种不同细胞系所获得的转录组数据, 与12种不同治疗领域相关联, 用深度神经网络进行训练, 可以根据化合物的转录组数据来预测其可能的治疗领域, 从而发现新的先导化合物或实现药物用途的重定向[32]。

5　传统机器学习与深度学习

定量构效关系 (quantitative structure-activity relationship, QSAR) 从Hansch等[33]开始, 已有50多年的历史, 机器学习方法不断发展, 实用价值得到药物化学家的一定认可[34]。但是QSAR受限于数据集的偏向性、参数选择不当、过拟合和模型缺乏解释等原因, 缺乏预测和指导能力[35,36]。随着深度学习的兴起, 基于上世纪90年代末就已运用于QSAR活性预测的神经网络发展而来的深度神经网络开始在药物研发领域发挥作用[37,38]。默沙东研究组和多伦多大学合作开发的深度学习方法DNN, 与随机森林 (random forest, RF) 相比, 能更好地预测默沙东公司内部药物研发实践中多样性的化合物活性数据[39]; 基于深度卷积神经网络 (deep convolutional neural networks, DCN) 的Chemception在活性预测方面也表现较好, 无需提供传统QSAR所需的分子描述符, 仅基于化合物本身自动提取相关特征[40]。Korotcov等[41]将深度神经网络与其他多种机器学习方法在药物研发的多个方面进行系统的比较, 所采用的数据集包括了细胞筛选的活性数据 (如美洲锥虫病抑制剂筛选和恶性疟原虫抑制剂筛选)、单个蛋白的活性数据 (如hERG的抑制)、化合物物理化学性质 (如溶解度) 等, 综合评分上深度神经网络的表现优于支持向量机 (support vector machines, SVM), 而支持向量机又优于其他机器学习方法, 如线性回归分析 (logistic linear regression, LLR) 和随机森林等。相反的是, Russo等[42]对比了深度神经网络与其他多种机器学习方法在预测化合物与雌激素受体结合能力的表现, 发现深度神经网络在非训练集的化合物活性预测能力上与朴素贝叶斯 (Naive Bayes)、决策树 (decision tree)、支持向量机和随机森林等多个传统的机器学习方法相比, 并无明显优势。Rodríguez-Pérez 等[43]根据高通量筛选的海量数据 (53个不同的靶标, 十万以上不同的化合物) 构建了小分子–靶标的活性谱, 根据活性数据标注了活性和非活性化合物, 然后在深度学习和多种传统的机器学习上进行测试, 总体而言, 深度学习的预测准确度并没有优于随机森林和支持向量机等机器学习。基于目前的研究程度, 虽然深度学习在图像识别上有一定优势, 但在其他多个方面, 譬如化合物活性预测等, 无法断言深度学习一定优于机器学习, 更多的还是依赖于数据集的广度和深度, 而不是算法本身[37]。而且, 因为深度学习是自动从原始数据中提取特征, 某些特征可能并无明确的物理化学含义, 所以虽然能够避免人工偏向性以及减少人力成本, 但如果无法理解深度学习提取的特征所表征的含义, 就难以在药物研发过程中作出理性可靠的决策。

6　计算化学与深度学习

除了机器学习, 计算机技术运用于药物设计与发现中的还有基于第一性原理 (first principle) 的计算化学。药物与靶标的结合, 其本质是一个自由能驱动的物理学过程。近年来, 计算能力的提高和新算法的发展, 对基于物理学原理的计算化学领域的发展有较大的推进作用, 譬如自由能微扰 (free energy perturbation, FEP) 在某些生物体系能精确到1 kcal·mol^{-1}, 接近试验测量误差[44]。因为深度学习本身也是基于计算, 所以数据驱动的深度学习与基于物理学原理的计算化学方法或计算机辅助药物分子设计也有相互交叉。就技术成熟度曲线 (hype cycle) 而言, 当前可用技术如计算化学已经历低谷, 成熟可用。深度学习刚经过巅峰, 渐渐在滑落去泡沫化低谷, 而深度学习用于药物设计还在促动期, 其与当前可用的其他技术相比, 优势和劣势需要进一步评估[45]。

基于靶标结构的药物设计需要依赖打分函数来预测小分子与蛋白的相互作用强弱。传统的打分函数主要包括基于力场 (force field-based)[46]、基于经验性函数 (empirical-based)[47]和基于知识 (knowledge-based)[48]。深度学习开始学习三维结构进行活性预测。目前不同模型的差异主要来自两个方面: 一是训练集, 二是表征方法。对于训练集, 分为实验获得的和通过传统分子对接预测的蛋白–配体复合体结构, 前者以PDBBind数据库[49]为代表, 数据更可靠, 但是数据量小; 后者可以产生大量结构数据用于训练, 但依赖于分子对接的准确度, 训练集中混入错误的配体结合构象会影响深度学习的表现。表征方法大体分为两类: 一类是三维格点, 即把蛋白–配体复合体当作三维图像, 用计算机视觉中的三维卷积模型识别蛋白–配体结合模式; 另一类是以原子为中心进行卷积, 把邻接原子环境编码为原子环境向量来表示蛋白–配体结合模式。前者有现成的计算机视觉领域的软件包可以借用, 方便做早期探索。后者基于原子生成特征, 更适合表征原子–原子相互作用。以三维格点为表征的研究已较为广泛, 而以原子为中心进行卷积的表征近年来逐渐开展。2007年, Behler和Parrinello[50]开创性利用高斯径向函数和高斯角函数作为基组编码原子位置信息, 引入神经网络来表征量化计算密度泛函理论 (density functional theory, DFT) 势能面, 比DFT的计算要快上好几个数

量级; 2017年, ANI-1更进一步, 使用修改后的2007年Behler和Parrinello的对称函数建立单原子环境矢量(atomic environment vectors, AEV)来表征分子, 所以深度学习可基于DFT量化计算结果进行训练, 学习产生的ANI-1可以计算比训练集所含体系更大的体系, 而且和DFT的量化计算准确程度基本一致, 但是速度要快得多, 有助于快速评估有机小分子的能量[51]。正如DUD[52]已成为评估分子对接方法的常用测试集, MoleculeNet也把目前已有的数据集、表征方式、网络模型都放到Github DeepChem上, 以供整个深度学习领域作为基准(Benchmark)测试使用[53]。

7 全新药物设计

20世纪90年代, 基于计算机的全新药物设计(*De-novo* drug design)已有相关的文献报道[54], 包括人工神经网络的应用[55]。全新药物设计已成为新药研发的重要手段[56], 但因为受限于分子生长和连接方式、成药性、合成难易及计算资源的问题, 全新药物设计能直接成功的案例并不多。药物设计包括对药物分子结构的识别, 对药物分子合成路线的分析以及构效构性关系分析[57]。深度学习在多个方面的应用都已有相关文献报道。

7.1 新分子的生成及活性预测 循环神经网络(recurrent neural network, RNN)能够接受序列数据作为输入特征, 之前用于自然语言处理领域, 现在应用于生成新化合物结构。Segler与AstraZeneca的研究人员合作, 以常见的SMILES字符串格式表征化合物, RNN首先通过类似学习语言的方式学习大量的SMILES文本是如何表征分子, 由此拟合出的模型可以生成全新的SMILES字符串, 即全新的分子且无偏向性, 适用于虚拟筛选等各种用途; 其次再基于迁移学习(transfer learning), 将之前训练出的模型用某个针对性靶标的小分子数据集进行再度训练, 而且这个小分子数据集无需大量的数据; 然后对模型微调, 在针对两种病原菌的全新药物设计中, 产生的分子与真实世界中药物化学家设计的化合物能有部分重合[58]。Schneider等搭建与上述原理类似的模型, 同时结合之前开发的靶标预测工具—SPiDER[59], 针对RXR和PPAR两个蛋白靶标, 实地合成了由深度学习自动设计的5个排名靠前的全新分子, 发现其中4个化合物在细胞活性显示纳摩尔到低微摩尔的水平[60]。Olivecrona等[61]进一步引入强化学习(reinforcement learning)用于调整前述的RNN模型来保持整体一致性, 避免生成的分子逐渐偏离方向和重复生成相同分子, 并且具有预先指定的性质, 适用于药物设计中的骨架替换以及定向化合物库的生成。匹配分子对(matched molecular pairs, MMP)分析是药物设计中重要的方法[62], 化合物同一骨架不同基团的取代时常会导致活性、溶解度、生物利用度和毒性等性质发生显著变化[63-65]。Turk等[66]将MMP与深度学习相结合, 用逆合成分析的规则来判断和提取ChEMBL数据库中匹配分子对作为数据集, 利用DNN训练匹配分子对活性发生变化的模式, 从而根据所得的模型产生新的分子, 并预测其活性变化。

7.2 代谢及毒理预测 定量构性关系早已运用于预测化合物的毒性[67]。而基于经验规则也广泛运用, 如ToxAlerts数据库收集大量已被文献报道的引起不良反应的分子或分子片段, 提示类似的分子可能引起相同或相似不良反应[68]。深度学习在预测化合物代谢及毒理学性质方面也陆续展开研究与应用。Fernandez等[69]基于深度学习发展的预测化合物不良反应的方法, 基于化合物的二维结构图, 在Tox21的数据集上对不良反应预测的准确度比肩目前其他机器学习方法[70]。由于药物毒性常来自于药物在体内的活性代谢物, 如药物诱导的肝脏毒性[71]等, Hughes等[72]基于深度学习开发了第一个可用于预测化合物是否会经过一步或两步形成醌从而潜在导致毒性的模型, 以及开发了用于预测化合物是否形成环氧化物从而潜在导致毒性的模型[73], 都整合在XenoSite模块中(http://swami.wustl.edu/xenosite), 可用于快速筛选化合物潜在毒性。同时, 深度学习方法也用于预测化合物的其他性质, 如Lusci等[74]通过改进化合物的表征方式并结合深度学习中的递归神经网络(recursive neural network, RNN), 开发了用于预测化合物溶解度的AquaSol, 其准确度与已有的计算方法相当。

7.3 化合物合成路线预测 目前合成路线的设计依赖于化学家的经验和知识, 最常用的策略是E. J. Corey提出的逆向合成[75]。通过学习海量的专利及文献中的化学反应来预测化学反应, IBM公司将深度学习在语言分析中的算法转移到对化学反应的解构上, 从而把预测化学反应的问题转变为语言翻译的问题[76]。Segler等[77]采用深度学习结合蒙特卡洛算法, 通过学习大量已经被多次验证过的化学反应后, 可采用逆向合成的策略来设计合成路线, 挑选出合适的起始原料。更重要的是, 该方法比目前常用的计算机辅助合成路线设计方法要高效, 且在双盲测试中得到专业人员的肯定。

8 深度学习的展望

药物设计作为一门综合学科, 需要考虑化合物生物学活性、药代动力学、药效动力学、毒理学和物化性质等各方面的因素, 而有些因素并无清晰的判断标准, 不同药物化学家对同一分子的评价也常常存在分歧[78], 这对于目前仍然需要明确数据标识的深度学习

而言是较大的障碍。目前的深度学习依赖于高质量有标识的大数据集, 这就要求数据点是清晰的, 同时是低成本的。深度学习作为分析数据提出假说的工具, 最适用的领域是缺乏假说、但同时又能以较低成本验证假说的领域, 因而验证体外活性和化学合成的预测可行性更大。虽然目前深度学习的进展主要体现在监督学习, 但是发展的方向是不依靠大量标注数据的非监督学习[18]。用于非监督学习的新网络架构逐渐兴起, 如深度自动编码器网络 (deep auto-encoder network, DEAN) 和生成对抗网络 (generative adversarial networks, GANs) 等, 已开始应用于药物设计与发现[79]。Kadurin 等[80]基于 GANs 开发的 druGAN 用于全新分子设计, 能处理大量数据集, 在生成分子所需特性上可调节性强, 以及更有效地对回归模型进行无监督预训练。Altae-Tran 等开发的迭代求精长短期记忆 (iterative refinement long short-term memory) 新构架, 用于迁移学习, 无需大量数据集进行训练, 在 Tox21 数据集[70]和 SIDER 数据集[81]上表现良好[82]。非监督学习的一大优势即是降低对标注清晰的数据量的要求, 是深度学习的发展方向。期待不断完善中的深度学习在真实世界的药物研发中发挥更加重要的作用。

References

[1] Smietana K, Siatkowski M, Møller M. Trends in clinical success rates [J]. Nat Rev Drug Discov, 2016, 15: 379-380.

[2] Mullard A. 2010 FDA drug approvals [J]. Nat Rev Drug Discov, 2011, 10: 82-85.

[3] Hughes B. 2007 FDA drug approvals: a year of flux [J]. Nat Rev Drug Discov, 2008, 7: 107-109.

[4] DiMasi JA, Grabowski HG, Hansen RW. The cost of drug development [J]. N Engl J Med, 2015, 372: 1972.

[5] Avorn J. The $2.6 billion pill--methodologic and policy considerations [J]. N Engl J Med, 2015, 372: 1877-1879.

[6] Mullard A. 2018 FDA drug approvals [J]. Nat Rev Drug Discov, 2019, 18: 85-89.

[7] Macarron R, Banks MN, Bojanic D, et al. Impact of high-throughput screening in biomedical research [J]. Nat Rev Drug Discov, 2011, 10: 188-195.

[8] Franzini RM, Neri D, Scheuermann J. DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries [J]. Acc Chem Res, 2014, 47: 1247-1255.

[9] Jorgensen WL. The many roles of computation in drug discovery [J]. Science, 2004, 303: 1813-1818.

[10] McCarthy J, Minsky M, Rochester N, et al. A proposal for the dartmouth summer research project on artificial intelligence [J]. AI Magazine, 2006, 27: 12-14.

[11] Mamoshina P, Vieira A, Putin E, et al. Applications of deep learning in biomedicine [J]. Mol Pharm, 2016, 13: 1445-1454.

[12] Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine [J]. Nat Biotechnol, 2018, 36: 829-838.

[13] Smalley E. AI-powered drug discovery captures pharma interest [J]. Nat Biotechnol, 2017, 35: 604-605.

[14] Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery [J]. Expert Opin Drug Discov, 2016, 11: 785-795.

[15] Xu YJ, Pei JF. Deep learning for chemoinformatics [J]. Big Data Res, 2017, 3: 45-66.

[16] Gawehn E, Hiss JA, Brown JB, et al. Advancing drug discovery *via* GPU-based deep learning [J]. Expert Opin Drug Discov, 2018, 13: 579-582.

[17] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313: 504-507.

[18] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521: 436-444.

[19] Srivastava N, Hinton G, Alex Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overtting [J]. J Machine Learning Res, 2014, 15: 1929-1958.

[20] Geoffrey E, Hinton NS, Alex Krizhevsky, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv: 1207.0580,2012.

[21] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, U S A, 2009: 248-255.

[22] Tetko IV, Engkvist O, Koch U, et al. BIGCHEM: challenges and opportunities for big data analysis in chemistry [J]. Mol Inf, 2016, 35: 615-621.

[23] Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017 [J]. Nucleic Acids Res, 2017, 45: D945-D954.

[24] Rose PW, Prlic A, Altunkaya A, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information [J]. Nucleic Acids Res, 2017, 45: D271-D281.

[25] Bray MA, Gustafsdottir SM, Rohban MH, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay [J]. Gigascience, 2017, 6: 1-5.

[26] Muresan S, Petrov P, Southan C, et al. Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data [J]. Drug Discov Today, 2011, 16: 1019-1030.

[27] Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets [J]. Nat Rev Drug Discov, 2017, 16: 19-34.

[28] Alteri E, Guizzaro L. Be open about drug failures to speed up research [J]. Nature, 2018, 563: 317-319.

[29] Ioannidis JP. Why most published research findings are false [J]. PLoS Med, 2005, 2: e124.

[30] https://clinicaltrials.gov/ct2/show/NCT03194217.

[31] Simm J, Klambauer G, Arany A, et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery [J]. Cell Chem Biol, 2018, 25: 611-618e613.

[32] Aliper A, Plis S, Artemov A, et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data [J]. Mol Pharm, 2016, 13: 2524-2530.

[33] Hansch C, Maloney PP, Fujita T, et al. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients [J]. Nature, 1962, 194: 178-180.

[34] Lombardo F, Desai PV, Arimoto R, et al. In silico absorption, distribution, metabolism, excretion, and pharmacokinetics (ADME-PK): utility and best practices. An industry perspective from the international consortium for innovation through quality in pharmaceutical development [J]. J Med Chem, 2017, 60: 9097-9113.

[35] Maggiora GM. On outliers and activity cliffs – why QSAR often disappoints [J]. J Chem Inf Model, 2006, 46: 1535.

[36] Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? [J]. J Med Chem, 2014, 57: 4977-5010.

[37] Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery [J]. Drug Discov Today, 2018, 23: 1241-1250.

[38] Zhang L, Tan J, Han D, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery [J]. Drug Discov Today, 2017, 22: 1680-1685.

[39] Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure-activity relationships [J]. J Chem Inf Model, 2015, 55: 263-274.

[40] Garrett BG, Siegel C, Vishnu A, et al. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models [J]. arXiv: 1706.06689,2017.

[41] Korotcov A, Tkachenko V, Russo DP, et al. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets [J]. Mol Pharm, 2017, 14: 4462-4475.

[42] Russo DP, Zorn KM, Clark AM, et al. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction [J]. Mol Pharm, 2018, 15: 4361-4370.

[43] Rodríguez-Pérez R, Miyao T, Jasial S, et al. Prediction of compound profiling matrices using machine learning [J]. ACS Omega, 2018, 3: 4713-4723.

[44] Wang L, Wu Y, Deng Y, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field [J]. J Am Chem Soc, 2015, 137: 2695-2703.

[45] Jordan AM. Artificial intelligence in drug design – the storm before the calm? [J]. ACS Med Chem Lett, 2018, 9: 1150-1152.

[46] Brooks BR, Brooks CL 3rd, Mackerell AD Jr, et al. CHARMM: the biomolecular simulation program [J]. J Comput Chem, 2009, 30: 1545-1614.

[47] Eldridge MD, Murray CW, Auton TR, et al. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes [J]. J Comput Aided Mol Des, 1997, 11: 425-445.

[48] Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach [J]. J Med Chem, 1999, 42: 791-804.

[49] Liu Z, Su M, Han L, et al. Forging the basis for developing protein-ligand interaction scoring functions [J]. Acc Chem Res, 2017, 50: 302-309.

[50] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces [J]. Phys Rev Lett, 2007, 98: 146401.

[51] Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost [J]. Chem Sci, 2017, 8: 3192-3203.

[52] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking [J]. J Med Chem, 2006, 49: 6789-6801.

[53] Wu ZQ, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning [J]. Chem Sci, 2018, 9: 513-530.

[54] Lewis RA. Automated site-directed drug design: approaches to the formation of 3D molecular graphs [J]. J Comput Aided Mol Des, 1990, 4: 205-210.

[55] Schneider G, Wrede P. Artificial neural networks for computer-based molecular design [J]. Prog Biophys Mol Biol, 1998, 70: 175-222.

[56] Schneider G, Fechner U. Computer-based *de novo* design of drug-like molecules [J]. Nat Rev Drug Discov, 2005, 4: 649-663.

[57] Schneider G. Automating drug discovery [J]. Nat Rev Drug Discov, 2018, 17: 97-113.

[58] Segler MHS, Kogej T, Tyrchan C, et al. Generating focused molecule libraries for drug discovery with recurrent neural networks [J]. ACS Cent Sci, 2018, 4: 120-131.

[59] Reker D, Rodrigues T, Schneider P, et al. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus [J]. Proc Natl Acad Sci U S A, 2014, 111: 4067-4072.

[60] Merk D, Friedrich L, Grisoni F, et al. *De novo* design of bioactive small molecules by artificial intelligence [J]. Mol Inf, 2018, 37: 1700153.

[61] Olivecrona M, Blaschke T, Engkvist O, et al. Molecular *de-novo* design through deep reinforcement learning [J]. J Cheminform, 2017, 9: 48.

[62] Griffen E, Leach AG, Robb GR, et al. Matched molecular pairs as a medicinal chemistry tool [J]. J Med Chem, 2011, 54: 7739-7750.

[63] Leach AG, Jones HD, Cosgrove DA, et al. Matched molecular

pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure [J]. J Med Chem, 2006, 49: 6672-6682.

[64] Hajduk PJ, Sauer DR. Statistical analysis of the effects of common chemical substituents on ligand potency [J]. J Med Chem, 2008, 51: 553-564.

[65] Wassermann AM, Bajorath J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets [J]. J Chem Inf Model, 2010, 50: 1248-1256.

[66] Turk S, Merget B, Rippmann F, et al. Coupling matched molecular pairs with machine learning for virtual compound optimization [J]. J Chem Inf Model, 2017, 57: 3079-3085.

[67] Ekins S. Progress in computational toxicology [J]. J Pharmacol Toxicol Methods, 2014, 69: 115-140.

[68] Sushko I, Salmina E, Potemkin VA, et al. ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions [J]. J Chem Inf Model, 2012, 52: 2310-2316.

[69] Fernandez M, Ban F, Woo G, et al. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images [J]. J Chem Inf Model, 2018, 58: 1533-1543.

[70] Andersen ME, Krewski D. Toxicity testing in the 21st century: bringing the vision to life [J]. Toxicol Sci, 2009, 107: 324-330.

[71] Fraser K, Bruckner DM, Dordick JS. Advancing predictive hepatotoxicity at the intersection of experimental, in silico, and artificial intelligence technologies [J]. Chem Res Toxicol, 2018, 31: 412-430.

[72] Hughes TB, Swamidass SJ. Deep learning to predict the formation of quinone species in drug metabolism [J]. Chem Res Toxicol, 2017, 30: 642-656.

[73] Hughes TB, Miller GP, Swamidass SJ. Modeling epoxidation of drug-like molecules with a deep machine learning network [J]. ACS Cent Sci, 2015, 1: 168-180.

[74] Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules [J]. J Chem Inf Model, 2013, 53: 1563-1575.

[75] Corey EJ, Long AK, Rubenstein SD. Computer-assisted analysis in organic synthesis [J]. Science, 1985, 228: 408-418.

[76] Schwaller P, Gaudin T, Lanyi D, et al. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models [J]. Chem Sci, 2018, 9: 6091-6098.

[77] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI [J]. Nature, 2018, 555: 604-610.

[78] Lajiness MS, Maggiora GM, Shanmugasundaram V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds [J]. J Med Chem, 2004, 47: 4891-4896.

[79] Jing YK, Bian YM, Hu ZH, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era [J]. AAPS J, 2018, 20: 58.

[80] Kadurin A, Nikolenko S, Khrabrov K, et al. druGAN: an advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties in silico [J]. Mol Pharm, 2017, 14: 3098-3104.

[81] Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects [J]. Nucleic Acids Res, 2016, 44: D1075-D1079.

[82] Altae-Tran H, Ramsundar B, Pappu AS, et al. Low data drug discovery with one-shot learning [J]. ACS Cent Sci, 2017, 3: 283-293.