蒋华良：AI 赋能创新药物研究——现状与未来丨理解未来科学讲座实录

未来科学论坛 2022-07-26 20:09 发表于北京

2022《理解未来》科学讲座 02 期——“AI+分子模拟与药物研发”，我们特别邀请到中国科学院院士、中国科学院上海药物研究所研究员蒋华良做专题分享《AI 赋能创新药物研究——现状与未来》，蒋教授在分享中介绍了国际创新药研发的进展与趋势，结合自己团队的研究案例展示了 AI 技术在新药研发中的应用。他还特别提出了 AI 应用于药物研发的十个挑战性问题，并鼓励青年学子参与这一领域的发展。

蒋华良：各位嘉宾、线上观众，大家好！非常高兴未来论坛理解科学讲座邀请我来讲 AI 关于药物研发方面的内容，也要感谢谢晓亮院士刚才对我本人的介绍，今天我主讲的题目是“AI 赋能创新药物研究——现状与未来”，主要分为四个部分介绍：药物研发的挑战和关键问题，AI 药物研发探索与实践，主要介绍我们一些案例，第三是 AI 在药物研发中的问题和展望，第四部分制药领域 AI 的十大挑战性问题。

首先介绍第一部分药物研发的挑战和关键问题：对于我国来说，新药研究已经不是问题，比如 2021 年我国药监局批准的创新药物已经达到 50 多个，创新药物对我国来说不是问题，关键是原创性药物就是 First in Class 药物的产出，我国的原创能力还非常薄弱。我给大家一个数据，统计 2012-2021 年世界上的 First in Class 药物，中文目

前还没有对应的名字，我们暂且称其原创药物。所谓的原创药物就是在一个疾病领域颠覆性的治疗方案，或者一个新作用机制、新的靶点产生的第一代在国际上首发的药物，比如 PD1 的第一代抗体 O 药和 K 药就是属于 First in Class 的药物。这个领域仍然是美国的强项，十年中美国产生 113 个原创药物，欧盟是 24 个，日本是 22 个，中国仅仅只有 4 个，所以原创药物对我国来说是一个迫切需要重视的领域。

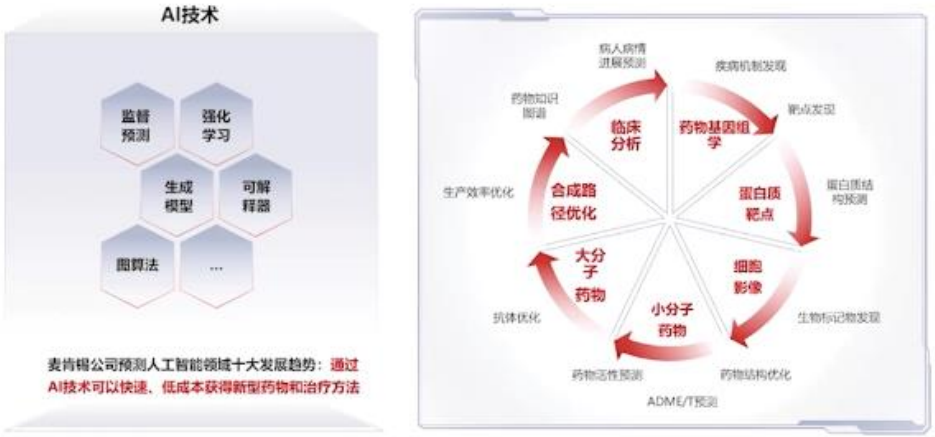
为什么药物研发特别是原创药物研发这么难呢？药物研发是一个系统工程，流程要从靶标发现到药物发现和临床前的研究、临床研究和审批，而从化学药来讲，我自己主要是从事化学药的，筛选成千上万个化合物，得到数百个候选，选择几个化合物进行临床前研究，再进入临床研究，即使到了临床研究阶段，大概 10 个候选药进入临床研究也就 1 个新药能够上市，所以这是一个非常漫长的过程。十五年前新药研发投入是 8 亿美元，五年前是 26 亿美元，现在已经超过 35 亿美元，一个靶点的发现需要十五年到二十年以后才能出现第一代药物，研究一个药物又要花十几年的时间，周期非常长，效率也非常低，失败率非常高，从万到百再到十，十再到一。这是一个漏斗型转化的过程，转化也是非常漫长的，所以投入大、风险大，产出也比较低，解决药物研发的投入问题，缩短周期、提高效率、加快转化，就是四个关键性的挑战问题。

第二部分我再讲一讲 AI 在药物研发中的探索和实践。刚才已经讲过，药物研发的周期长、转化慢、研发投入大，要想解决这样的一个问题，一百多年来药物科学工作者想尽各种办法，用上各种技术，

最近人们寄希望于 AI 加速解决刚才我讲的四个问题。

近年来药物领域发生了哪些技术性的变革？个人认为主要是信息技术和生物技术的融合对药物研发产生了深刻的影响，尤其是最近五年来人工智能技术大大促进药物研发领域的技术发展，也成为药物研发技术发展的一个热点。

人工智能可以加速新药研发的多个环节



实际上药物研发从临床开始到候选产物，然后进行临床前的研究，再进入临床研究，这样一个全链条的过程，AI 是全部可以用上的，实际上已经全部用上了。比如生命科学的基础研究中发现药物的靶点，AI 技术是可以用上的，生物标志物的发现过程中，比如细胞影像数据、测序数据，我们可以发现生物标志物，靶标发现、生物标志物发现以后进行小分子药物设计，AI 技术也可以用上。最近抗体类、生物技术类药物，疫苗的研发 AI 也可以用上，甚至在生产过程中化学药合成工艺的路径优化也可以用 AI 进行赋能。现在还有一个方向，就是一个药物做出来不容易，70%-80%的经费花在临床研究，预测一个

药物进入临床能不能成功，成功率是多少也是一个热点领域，如果成功率是 30%，建议公司可以不要去做这个药物的临床，如果成功率是 70%以上，那么做临床肯定有希望，这也是一个非常好的领域。

国际上各大研究机构和高校，特别是企业对 AI 在药物研发中的应用是非常重视的，去年已经有 40 家顶尖的制药公司，其中 3 家是中国的，包括 55 家 AI 初创公司和云服务、云计算的高校发表 AI 辅助药物研发的行动白皮书，充分显示药物研发的各个方面、各种战略科技力量对 AI 这样一个技术在药物研发中应用的重视。有些初创公司和大型制药公司合作用 AI 寻找新先导化合物，有些也已进入临床研究，截止去年为止，大概有 20 多家 AI Biotech 公司在纳斯达克等交易所上市。

我们的实验室还是根据发病机制、靶标发现、新的化合物优化一直延续到临床研究，发展用于药物研发的 AI 软件系统，我们跟高校、制药公司、IT 公司，比如华为、阿尔脉、先声、再鼎、药明康德等都有很好的合作。作为算法和软件的发展，一定要和应用结合才能得到更好的反馈，真正用于药物的研发。

介绍一下第一个算法，基于蛋白质序列进行预测，跟蛋白质进行结合的药物化合物算法，实际上 AlphaFold2 出来以前就有做这项工作，现在依然有重要的意义。传统的药物设计最好是有靶标结构，比如蛋白质结构、生物大分子结构，到现在为止也是这样，大部分靶标结构已经没有晶体结构，当时就发展基于序列的 TransformerCPI 模型，这样的算法也有得到很好的应用，涉及到很多激酶抑制剂或者 G 蛋

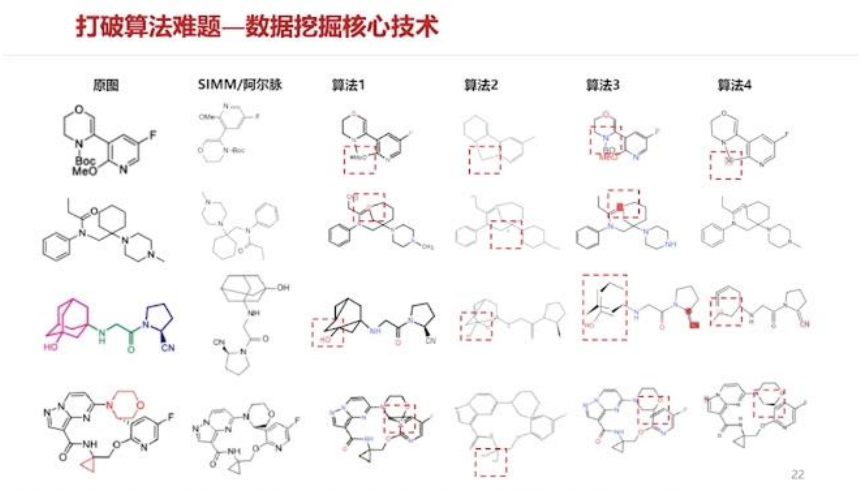
白偶联受体蛋白拮抗剂的设计，并与工业界合作，有些药物也已进入开发阶段。

目前有一个艾滋病药物叫做 Nelfinavir，我们发现一个宿主靶点，这对减缓重症病人死亡率有很好的作用，但机制是不知道的，通过我们的算法就找到一个宿主靶点。ENPP1 也是肿瘤方面的重要靶点，其抑制剂的深入研究具有重要临床价值。如果我们知道有些药物靶点不清楚就可以去寻找靶点，如果知道一个靶点，但没有化合物的话，我们也可以筛选这样的一些化合物，所以有一举两得的优势，现在有些和制药公司合作，也有发现原创性靶点的候选化合物。

第二个例子是离子通道，因为离子通道很多都是没有结构的，测定离子通道结构是非常难的，人类大概有 400 多个离子通道，80%以上是没有结构的，这个方法尤其适合离子通道，我们设计的一些非成瘾性镇痛的离子通道的抑制剂已经优化，可以进入开发的阶段。这些显示出 AI 的算法在药物设计上有些前瞻性的用途，相比传统基于结构的药物设计相比有着一定的优势。

通过这样的算法，我们的学生 2018 年参加比赛，叫做 Multi-Targeting Drug Design，这个比赛是非常苛刻的，同一个化合物要针对 2 个激酶有活性，针对一个 1 个激酶一定没有活性，另外的激酶是毒性靶点，还有一个是 tau 蛋白，是和阿尔茨海默症有关的蛋白，我们的学生在这两个比赛都拿到了冠军，提供线索和候选化合物，然后第三方购买，进行实验验证，所以这个方法从序列到靶标功能和活性化合物的设计，再到候选化合物的优化都会有一定的推进作用。

在 AI 药物研发领域数据是一个难题，大部分制药公司的数据都是不共享的，公共数据也是有限的，现有的 AI 算法都是数据驱动的，如果没有大量的数据，AI 的算法就做不好。我们也有搜集很多数据，包括商业化的数据、药物所自己产生的数据和文献上的数据，但是对化学药物来讲，文献上或者专利的数据有一个困难，就是在文献上和专利上的化学结构解读，翻译成可以进行 AI 计算或者搜集数据就是一个难题。



我们发展了一种药物分子结构的表示方法，这个方法有广宽的应用前景，但到实际应用还有一段距离，也是借助去年的这个比赛，即化学分子式的翻译比赛，就是把文献上的有些一百多年前的结构、模糊化的结构能够读出来，我们的学生参加比赛也拿到了冠军。比赛归比赛，但到实用需要经过好几次迭代，我们经过了 5 次迭代。图中这样的模糊结构需要识别出来我们可以进行算法编制的、模型构建的结构，我们和其它一系列的算法相比精确度是非常高的，几乎可以达到 99%以上的识别率，所以这是图像识别的算法进行改进完成。有了这样的结构，现在读取文献和专利中的文字不是一个困难，可以针对结

构把做过的临床数据，包括药效数据、安评数据、毒性数据、物理化学性质的数据统统搜集起来，变成一个很大的数据库，在此基础上就可以进行各种算法的模型构建，可以解决一部分数据难题，至少对化学药来讲，我们在文献中和专利中挖掘数据进行药物研发就是一个很大的促进作用。

我讲 2 个案例， 一个是我们怎么根据序列解决药物研发，没有结构的情况下怎么来做。还有一个数据突破的难题，我们跟华为合作，把这些算法都弄在华为的云上，因为我们研究所的算力不够，也建立了亿级水平的成药性化合物数据库 DrugSpaceX，根据现有的药物以及文献中报道的活性化合物为基础，构建成药性比较高的化合物牵库，供大家进行药物设计，传统的虚拟筛选也可以用我们这样的库进行药物的研发，科制药公司和科研单位均可以应用。

我们自己也有一个案例，就是 G 蛋白偶联受体，开发抗精神分裂症的药，我们第一代药物做了 15 年，积累了 3000 多个化合物的数据，在此数据上进行建模，仅仅花了半年的时间合成 5 个化合物，现在已经有 1 个化合物和药物化学家、药理学家合作进行临床前的研究，药代动力学性质和安全性都比较好，希望完成临床前研究之后可以申报临床，因此在原有的数据基础上开发第二代药物研发时，如果有 AI 的技术介入，比第一代药物研发要好得多、快得多，时间和研究成本也节约下来了。

iDEL（intelligent DNA Encoded Library）技术可以合成上亿级的化合物进行筛选，但有一个问题，就是库里噪音比较多，怎样进行筛

选呢？可能 1 亿个化合物只有几个有活性，怎么挑出来是一个问题，所以我们跟 DEL 专家合作建立 AI 赋能的工作，也是取得了比较好的进展。

我举一个案例， Pictet-Spengler 反应就是和药明康德合作，DEL 一个关键步骤就是挑试剂，如果反应效率不高，合成出来库的质量就不高，AI 模型介入之前，大部分反应的效率不到 30%，那么这样的库的质量就成问题，筛选出来活性化合物的成功率也不高。我们就是利用这样的数据作为基础建立模型，大部分试剂的产率大于 70%，所以这样的库质量就比较高，我们把 DEL 这样药物研发新的技术和 AI 结合起来，促进 DEL 这项技术的发展，也是在应用中得到了很好的结果。

第三部分，存在的问题和展望。

刚才讲过，虽然我们解决了结构识别的问题，可以到专利和文献中挑数据，但专利和文献中的数据肯定没有制药公司一百多年来积累下来的数据值钱，但制药公司给的数据是商业化价值很高，不会共享，现有数据的偏差性也是比较高的，数据量依然是比较少的，相比其它领域，我们药物研发的数据量是偏少的。总结一下，数据的高壁垒、高成本、高机密，数据量小、有偏差。怎么来解决这样的问题，这是 AI 在药物研发领域存在的第一个问题。

最近我们企图解决这样的问题，利用联邦学习，就是用原来 AI 应用于金融系统的算法移植到药物上来，各大制药公司和研究机构可以共享模型，但不共享数据。我们跟华为合作建立起个性化的联邦学

习模型，目前已经上线，大家可以使用，我们进行化合物水溶性的预测，包括抑制剂的模型。我们也已进行联邦学习的尝试，所有的制药公司参与者要用我们的模型把数据加入进去，提高这种模型的预测能力，经过不断的迭代，最后预测的效率或者准确率会越来越高。好处就是把公司商业数据的保密问题解决了，同时也共享了提供的数据产生的模型效率。

数据量小的话就是要发展主动学习的策略，我们原来是被动学习，就是基于大数据，现在是要发展一种算法的话就是利用主动学习的策略指导实验，增加实验数据，不断迭代，就是把我们自己的实验数据加入模型中，进行模型预测能力的提高，也是第二个可以解决的方向。

我们也有进行一些尝试，发表一些论文，不同的采样策略都有显著差异，如果找到一个合适的主动学习模型的话，部分体系上可以达到小样本建模优于全量数据建模的结果。目前这些刚刚开始，我们利用不确定性采样指导药物体内的暴露量，因为药物在体内的暴露量是药物成药性一个很关键的问题，也是和实验工作者合作开展这项研究，已经取得比较好的结果。

（二）模型开发方面的问题

AI 在药物研发方面的另一个问题就是模型开发，这是 AI 通用的，也是 2019 年《Nature》报道的，所有领域的 AI 都有存在这样的问题，就是有偏的数据集，不合适的数据集中 AI 无法学习真正有用的知识，包括隐藏变量的问题，AI 学习是隐藏变量对应错误的信息，也就得到错误的模型，错误建模的目标和不合理模型的评价指标无法真实反映 AI 算法的效果。这些都是陷阱，AI 在药物研发应用中依然是存在的。

未来的方向也是全链条，从靶标发现、化合物合成、化合物筛选、晶型预测、患者招募、临床试验的设计以及药物临床成功率的预测都是未来发展的方向，AI 在药物研发的全过程中仍然是大有可为的。

最后我想讲一讲制药领域 AI 十大挑战性问题，特别是供年轻人参考，这些都是我个人的想法，请大家批评指正。

药物设计有五十多年的历史，六十年代开始直到八十年代真正应用，传统的药物设计几乎没有一个化学药 CADD 方法来做。AI 现在处在什么状况呢？个人认为处于上个世纪八十年代传统药物设计的阶段，现在应用领域从理论突破还是比较少的，主要还是在药物发现的阶段。临床前的研究是有，但还是比较少，临床治疗阶段的研究主要是用于诊断，但用于药物真正的临床研究应用还是比较少的，这些是目前所处的状态。

AlphaFold2 是对蛋白质预测突破性、革命性的工作，解决了蛋白质折叠、三维结构预测的问题，也是对药物研发利好的消息，即

使是传统基于结构的药物设计，这也是有很大的促进作用。刚才我讲过，原来大部分靶标都是没有结构的，现在至少可以在AlphaFold2 预测的结构进行基于结构的药物设计，但是 AlphaFold2 没有解决蛋白质科学的所有问题，也是存在一定的挑战。

比如靶标蛋白动态变化的重要性就是动力学的问题，因为靶标蛋白在体内相互作用是一个动态的过程，不是一个静态的过程，AI 应该怎样介入？这是一个挑战，靶标不是单一的，而是网络体系，比如蛋白质-蛋白质相互作用，AlphaFold2 没有完全解决这一问题，国际上也在努力。

最后，我想讲一下，AI 是年轻人的领域，所以年轻人应该介入到这样的领域中，同时它也是多学科交叉的领域，至少是信息科学、生命科学、药物学、化学交叉的领域，属于年轻人的天下。像我这样比较老的只能知道大概的，要是让我去弄具体的算法，我是做不出来的，还是要靠学生和年轻人来做。

AI 药物研发面临十大挑战。

第一大类，药物研发和新型治疗方案主要有六个问题：一、靶标发现和实验验证的迭代技术，现在发表的文献很多，每个文章都说 Potential Target，但最后怎么确证，真正成为可药性靶点；二、难成药的孤儿受体，连化合物的线索都没有，在此基础上 AI 技术怎样介入到药物研发的过程？三、成药性预测，比如说很简单的水溶性，原来我们说药物设计出来的分子水溶性是大问题，包括其它成药性，包括安全性和毒性等等，与实验验证联合技术怎样介入；

四、临床药理和毒理预测技术；五、药物临床成功率的预测也是比较难的，因为现在有些医院的数据也是共享性有问题，不会拿出来共享建模。当然，我们可以从《新英格兰医学杂志》(NEJM)《柳叶刀》(The Lancet)等杂志搜集大量的数据进行建模；六、现在又出现一些新的治疗方案、治疗技术，比如基因治疗和干细胞治疗，AI 怎样赋能这样的新型治疗方案。

第二大类基础研究性问题包括：七、蛋白质-蛋白质的相互作用及蛋白质复合体三维结构预测，AlphaFold2 只解决单个蛋白的集合；八、配体-靶标结构三维结构的预测和结合强度预测，特别是亲和力的精准预测也是一种挑战，现在的虚拟筛选只是给出一个排序，没有给出一个定量化的数据；九、调控生物通路和网络的药物设计技术，AI 怎样介入这样的领域。

第三是底层的问题，未来 AI 技术，目前的人工智能主要是机器学习或者深度学习，如何在神经科学、脑科学发展的基础上来衍生出非深度学习的强人工智能的底层技术，也是一大挑战。这是我提出的十大挑战。

以上就是我的报告内容，致谢上海药物所的团队，特别是我的同事和学生的工作。另外我们和上海科技大学、华为、阿尔脉、先声、再鼎、药明康德都有很好的合作，也得到了自然科技基金委、科技部、中国科学院的大力支持。

谢谢大家！