**A Survey/DrugAI**

药物开发过程主要包括：药物靶点确定，先导化合物发现和优化，候选药物确定，临床前研究和临床研究；现代药物开发的目的是通过机器学习方法加快中间步骤，从而节省开发成本——利用机器学习模型来预测早期分子特性。

人类生物学与分子病理学——确定与疾病有明确联系的靶点蛋白；

新药设计集中体现在药理活性（功能）是否可以从分子结构中推断出来，以及这种结构的哪些元素是相关的——先导化合物；

新药开发的瓶颈集中在靶点蛋白和药物先导化合物的确定，这涉及更深层次的生物学与病理学。而确定了靶点蛋白和药物先导化合物其中一方便能够利用基于特征相似性或是结构相似性的药物-靶标互作预测方法来加速药物开发；

目前的药物开发都是以药物重定位为目的构建预测模型。药物重定位是一种用于发现老药或在研药物超出原始批准的适应证，扩大其适用范围和用途的策略。与新药开发相比，该策略失败风险更低，并且可以利用现成的大规模药物-靶标网络知识图谱通过相似性进行药物-靶向蛋白预测。

药物-蛋白质相互作用也可以预测多重药理学，即药物分子与多种受体相互作用产生脱靶副作用的药物筛选。

---------------------------------------------------------------------------------------------------------------------------------

关于蛋白质化合物相互作用(CPI)模型的构建

建立预测模型的前提是有相应的有效数据库，即包括药物和靶向蛋白的相互作用对或者是可以量化的化合物与相关靶向蛋白之的相互作用强度（也可以理解为产生有效相互作用的可能性）的数据库。

有了相应的数据基础，之后便是决定采用何种方式描述数据，即数据表示方式，包括化合物、靶向蛋白及其相互作用对的表示。而关于这种带有3D结构的大分子物质的完全量化在目前很难实现，至少在计算机上的难以批量应用，所以目前的数据表示主要集中在1D组成成分信息与2D结构信息以及一些通过映射得到的3D结构信息。

组成成分信息包括蛋白质的氨基酸序列以及化合物分子的原子组成序列。由于化合物分子的规模较小并且功能与结构信息紧密相关，所以可以更多的量化其结构信息，目前广泛使用的化合物SMILES序列（简化分子输入线性输入系统）其中就包含了2D分子图的排列信息。基于相似性的预测都是建立在1D组成成分信息相似的基础上，包括整体序列的相似性以及局部序列的相似性。而关于结构相似性的预测，介于生物大分子及化合物的复杂结构，大多是利用局部子结构的相似性，也就是提取关键子结构（比如药物分子的功能团或药效团）信息。

除了浅层的组成及结构信息，利用1D组成成分信息与2D结构信息还可以得到更深层次的信息来描述数据，比如一些物理化学相关的性质、蛋白质的进化信息等等。目前更流行的是直接利用深度神经网络提取更深层次的潜在信息，这种从表层信息中提取的特征往往比人们预先定义的特征更有效，但缺点是缺少可解释性。

---------------------------------------------------------------------------------------------------------------------------------

图网络的应用

大多数深度学习模型不能直接从原始输入数据中学习结构信息，传统的 CNN 或其他深度学习模型难以直接处理分子图等非结构化数据。因此，将深度学习方法推广到非欧几里得领域的神经网络已成为处理药物相关任务的最新方法。在大多情况下，生物学数据被构建为非欧几里德域中的生物网络，可以通过两种方式进行建模: 基于分子结构的建模和基于生物网络的建模；

基于分子结构：使用原子或有效的化学子结构作为节点，使用键作为边，然后构建分子图，一般应用于分子性质预测和从头设计；

基于生物网络：各种实体被用作节点，如基因、疾病、RNA、分子、蛋白 等。节点之间的边缘意味着在一对实体之间存在已知的关联；比如疾病关联网络、药物—靶标互作等；

GNN 将非欧几里德数据作为图进行处理，旨在通过表示学习即图嵌入，在保持当前空间某些几何特征的前提下，将数据转化为一个低维的、更具有鉴别力的特征空间。通过节点之间的消息传递来捕获图的依赖性，同时考虑到输入数据的规模、异构性和深度拓扑信息。

GNNs在生物信息学中的典型应用包含三个主要主题：疾病关联预测、药物开发和发现以及医学成像；

药物开发：

蛋白质结构与功能预测：包括分子结构分析与蛋白质功能预测。最准确的结构预测可以从分子水平上充分阐明蛋白质作用的生物学机制。

分子结构分析：近年来，一些研究表明，GNN 在通过简化的图形表示中学习蛋白质的有效结构方面具有强大的能力；

蛋白质功能预测：有两种类型的方法: 基于蛋白质结构和基于 PPI 网络；蛋白质序列中含有的大量无效片段可能会影响对其功能的判断。利用 GNN 整合蛋白质关系网络的特征，是解决蛋白质序列和功能差异问题的方法之一；

蛋白质-蛋白质相互作用预测：能够间接地为药物开发提供理论支持

配体-蛋白(药物-靶点)相互作用预测：药物靶点与疾病或生物分子的病理状态相关，因此药物及其靶点的识别是新药开发的核心问题。药物靶标相互作用预测本质上是配体与蛋白质相互作用的预测问题；

知识图谱

分子性质预测：是药物开发的基础和重要组成部分；

新分子设计：药物设计的最终目标是发现具有理想化学性质的分子。

药物反应预测：将基因组学数据和药物信息结合起来进行药物反应预测；包括药物相互作用（消极的），药物-副靶标作用（有害的）；

---------------------------------------------------------------------------------------------------------------------------------

蛋白质三级结构预测

由于蛋白质序列从氨基酸中快速增长，用实验技术获得蛋白质的三维结构变得更加困难。 因此，蛋白质结构预测(PSP)仍然被认为是更重要的。[蛋白质结构预测的目标是使用计算方法从其氨基酸序列](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/peptide-sequence" \o "从 ScienceDirect 的 AI 生成的主题页面了解更多关于氨基酸序列的信息)开始确定蛋白质分子中每个原子的空间位置。

主要步骤：输入蛋白质序列→生成MSA同源性序列→精化MSA序列作为网络输入某种网络→输出距离/接触图→生成蛋白质结构

---------------------------------------------------------------------------------------------------------------------------------

名词：

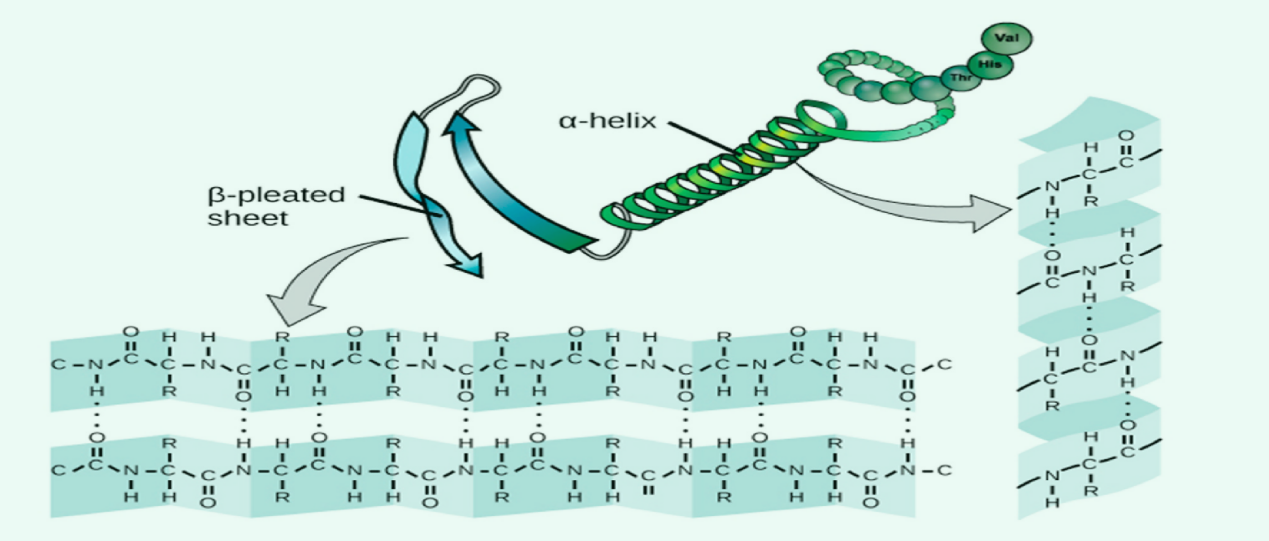
定量构效关系 (QSAR)指利用数理统计方法建立特定性质与分子理化性质参数或结构参数之间关系的方法；可作为分子描述符的函数；

药代动力学AMDET（absorption, distribution, metabolism, excretion, toxicity）：吸收、分布、代谢、排泄、毒性；

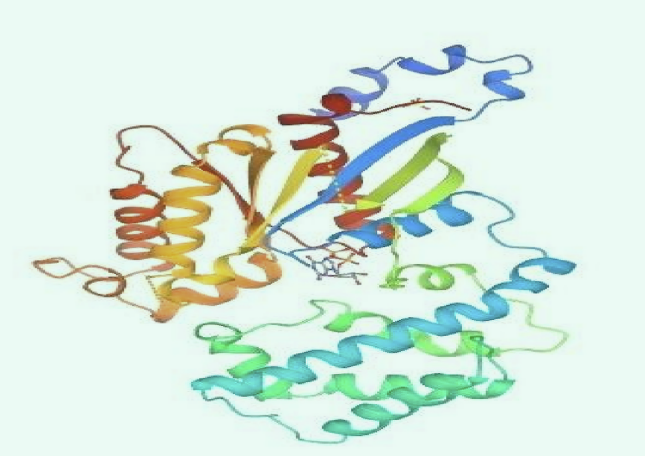
蛋白质一级结构：在每种蛋白质中，多肽链中氨基酸的排列顺序



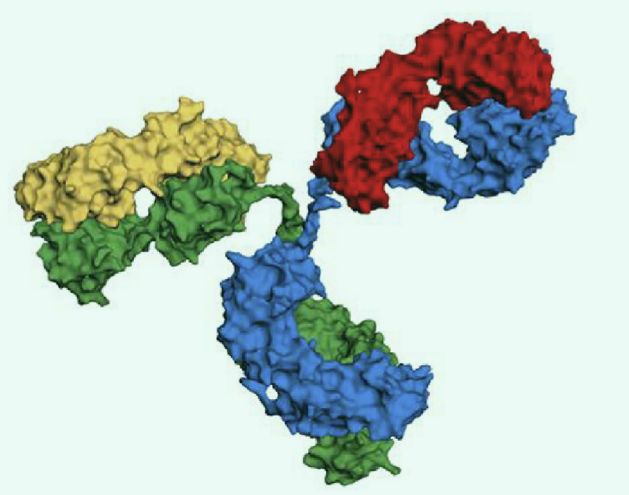
蛋白质二级结构：蛋白质多肽链本身的折叠和盘绕的方式



蛋白质三级结构：指一条多[肽链](https://baike.baidu.com/item/%E8%82%BD%E9%93%BE/8625112?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)在[二级结构](https://baike.baidu.com/item/%E4%BA%8C%E7%BA%A7%E7%BB%93%E6%9E%84/9644031?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)或者[超二级结构](https://baike.baidu.com/item/%E8%B6%85%E4%BA%8C%E7%BA%A7%E7%BB%93%E6%9E%84/1479765?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)甚至[结构域](https://baike.baidu.com/item/%E7%BB%93%E6%9E%84%E5%9F%9F/1236004?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)的基础上，进一步盘绕，折叠，依靠[次级键](https://baike.baidu.com/item/%E6%AC%A1%E7%BA%A7%E9%94%AE/10413614?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)的维系固定所形成的特定[空间结构](https://baike.baidu.com/item/%E7%A9%BA%E9%97%B4%E7%BB%93%E6%9E%84/3077578?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%B8%89%E7%BA%A7%E7%BB%93%E6%9E%84/_blank)



蛋白质四级结构：蛋白质分子中各个亚基的空间排布及亚基接触部位的布局和相互作用,称为蛋白质四级结构:体内许多功能性蛋白质含两条或以上多肽链。



---------------------------------------------------------------------------------------------------------------------------------

**实验室：**

Yangzhang

https://zhanggroup.org/

Jinbo Xu

<https://home.ttic.edu/~jinbo/> ; <http://raptorx.uchicago.edu/>

Yaoqi Zhou

<http://zhouyq-lab.szbl.ac.cn/> ; <https://sparks-lab.org/group/>

Dong Xu

<http://digbio.missouri.edu/>

David Baker

<https://www.bakerlab.org/> ; <https://www.ipd.uw.edu/>

Peer Bork

<http://www.bork.embl.de/j/>

Jianyi Yang

<http://yanglab.qd.sdu.edu.cn/>

待补充

---------------------------------------------------------------------------------------------------------------------------------

**研究现状：**

药物—靶标作用：

基于 ADMET 预测器和 ALGOPS 程序的神经网络已被用于预测各种化合物的亲脂性和溶解度；

SimBoost 利用回归树来预测 DTBA，并同时考虑基于特征和基于相似性的交互；

DeepDTA 接受 SMILES 形式的药物数据及蛋白质氨基酸序列用于蛋白质和药物结构的一维表示；所得到的表示比预定义的蛋白质和药物特征（相似性矩阵）表现更好；

WideDTA 是 CVNN DL 方法，相比于DeepDTA，在输入端加入了更多的预定义特征。将配体 SMILES (LS)、氨基酸序列、配体最大共同子结构（LMCS） 和蛋白质结构域和基序作为输入数据用于评估结合亲和力；

DeepAffinity 是一种可解释的 DL 模型，它同时使用 RNN 和 CNN 以及未标记和标记数据。它在结构和物理化学特性中考虑了 SMILES 格式的化合物和蛋白质序列；利用CNN对大量无标记数据进行无监督学习或获取化合物和蛋白质的潜在表征，。并将药物和靶蛋白的特征组合作为输入数据来预测药物靶标相互作用 (DTI)，即两者之间的相互作用强度；对于药物和靶点，分别使用 SMILES 表示和蛋白质序列组成 (PSC) 进行说明；

DeepcSeqSite 用于预测蛋白质-配体结合残基，这是一种基于深度卷积神经网络(CNN)的基于序列的方法。通过对几个卷积层进行叠加，从输入中获得层次特征。属于任何选定配体类别的结合残基在训练集中被分类为阳性样品，而其余的被标记为阴性样品。用于蛋白质-配体结合残基预测的七种类型的特征: 位置特异性评分矩阵，相对溶剂可及性，二级结构，二面体角，保守性评分，残基类型和位置嵌入，这些纯粹来源于蛋白质序列；

DeepSurf 一种基于深度学习架构的结合位点预测方法 。它利用了基于表面的表示(实现3D 体素化网格)以及最先进的深度学习架构来预测蛋白质上潜在的药物位点。在确定输入特征之后，将这些网格导入到3D CNN 中，由此得到的每个表面点的亲和力得分通过聚类来构建结合位点；

DeepDTAF 的预测模型，其中只使用一维序列数据生成局部和全局特征。(输入表示中排除了蛋白质、配体及其复合物的三维结构)；局部特征包括蛋白质-分子结合口袋的序列及结构特征；

GraphDTA将药物化合物视为原子间相互作用的图形，使用开源化学信息学软件 RDKit 提取原子特征。图中的节点是一个多维二元特征向量，表示五条信息: 原子符号，相邻原子的数量，相邻氢的数量，原子的隐含值以及原子是否在芳香结构中；

PADME是一个基于 DL 的平台，它利用前馈神经网络来预测药物靶标相互作用 (DTI)。它将药物和靶蛋白的特征组合作为输入数据，并预测两者之间的相互作用强度。对于药物和靶点，分别使用 SMILES 表示和蛋白质序列组成 (PSC) 进行说明；

DeepTox的 ML 算法通过识别分子化学描述符中的静态和动态特征（例如分子量 (MW) 和范德华体积），根据预定义的 2500 个毒物特征有效地预测分子的毒性；

**Thesis**

Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

Nguyen, Thin, et al. "GraphDTA: Predicting drug–target binding affinity with graph neural networks." Bioinformatics 37.8 (2021): 1140-1147.

Karimi, Mostafa, et al. "DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks." Bioinformatics 35.18 (2019): 3329-3338.

**CODE**

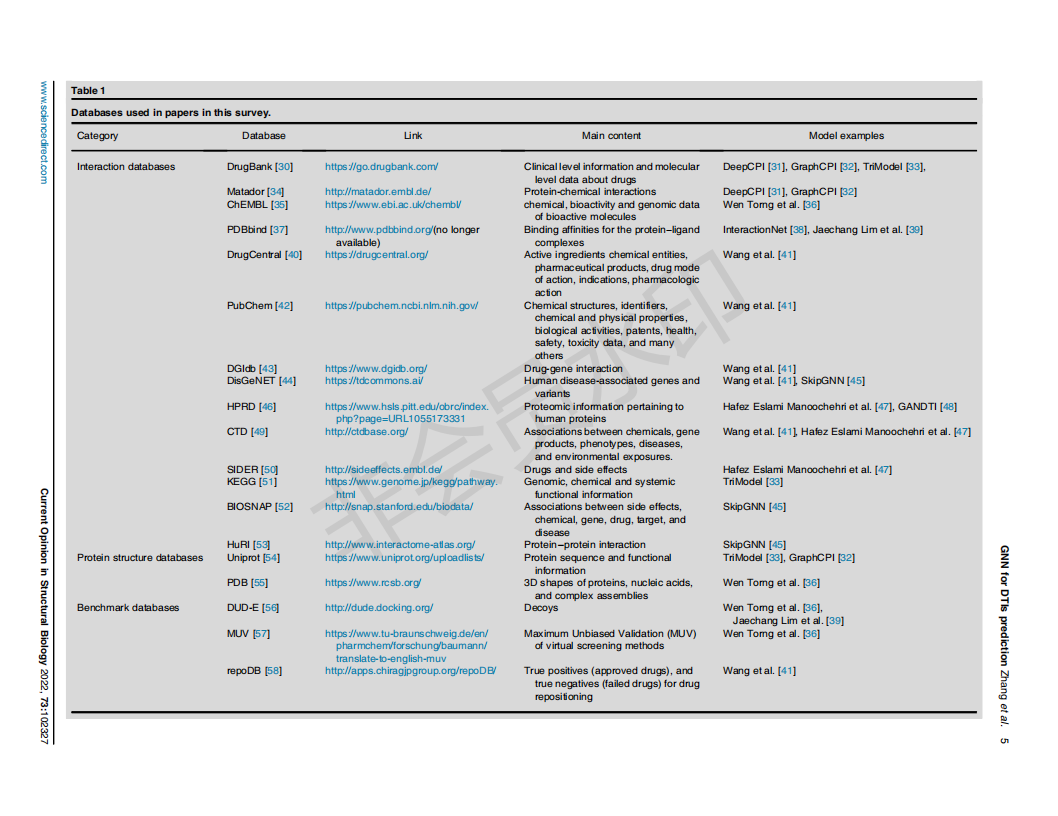
Deep DTA：https://github.com/hkmztrk/DeepDTA

**挑战：**

使用自然语言模型来处理序列会得到更多的特征；

可解释性的预测蛋白质结构与功能

**数据库**



CATH <http://cathdb.info/> 它提供了蛋白质结构域的进化关系。

SWISS-PROT https://www.expasy.org/resources/uniprotkb-swiss-prot 蛋白质序列数据库

**竞赛**

CAMEO

<https://cameo3d.org/>

CASP

<https://predictioncenter.org/>

RNA-Puzzles

<http://www.rnapuzzles.org/>

D3R

https://drugdesigndata.org/

领域了解

零基础 Pymol 作图入门

<https://www.bilibili.com/video/BV1fa4y1s7e6/?spm_id_from=333.337.search-card.all.click&vd_source=e1b0bf3aa5230c9f48f4ced8e2ec22ce>

秒杀AlphaFold！Science：用AI一秒设计自然界全新蛋白质

<https://mp.weixin.qq.com/s/ecLpD6hxTxOJeG59PBRAuw>

专题 | 蛋白质和多肽相互作用

<https://mp.weixin.qq.com/s/1XhUTxPJIpT-28h-fuWieQ>

用扩散模型进行蛋白质结构生成，已开源

<https://mp.weixin.qq.com/s/fpWgNRr38lOV_KkN1eAiSQ>

基于人工智能的蛋白质复合体三维空间结构预测：AlphaFold-Multimer原理介绍

<https://www.bilibili.com/video/BV1xB4y1i7x3/?spm_id_from=333.337.search-card.all.click>

综述：基于进化和物理启发建模的计算蛋白设计

<https://mp.weixin.qq.com/s/t3kmNp1D3yTutyr3tIoAKQ>

Cell专刊｜AI在蛋白结构、精准医疗、抗体疗法[综述]等的应用与未来预测

<https://mp.weixin.qq.com/s/D3sQyEU_qrSLiNDQF2BRkg>

【ISMB2022教程】图表示学习的精准医疗，哈佛大学

<https://mp.weixin.qq.com/s/bYHdKgpZV-Y272LAWBqt3g>

AlphaFold能不能用？MIT新研究：比随机猜测好不了多少，仍需继续改进

<https://mp.weixin.qq.com/s/idUBqIQChoQKlncxgavyOA>

许东的「AI生命科学」进化史：从90年代泡沫繁荣，到深度学习复兴

<https://mp.weixin.qq.com/s/g1EEZmYA-Is5ulZd3Ndz7w>

基于人工智能(AI)的蛋白结构预测工具合集

<https://mp.weixin.qq.com/s/7fb4tx0sKXE26IX1nXw4qg>

KDD 2022 | 编程指南：生命科学中的图神经网络

<https://mp.weixin.qq.com/s/qIf_ruPh4S3Ux45ZgOavuQ>

Nature Methods | 密西根大学张阳教授团队研究进展：蛋白质、核酸及其复合物通用结构比对算法US-align

<https://mp.weixin.qq.com/s/lXUYs2co-FWNFmZP5WWlHA>全球人工智能(AI)在药物发现中的应用市场--行业趋势和2029年预测

<https://mp.weixin.qq.com/s/lo9P8Ke2xGlHoksr-VuByw>

工具