AlphaFold会改变什么？

第一部分：蛋白质的结构及其解析方法

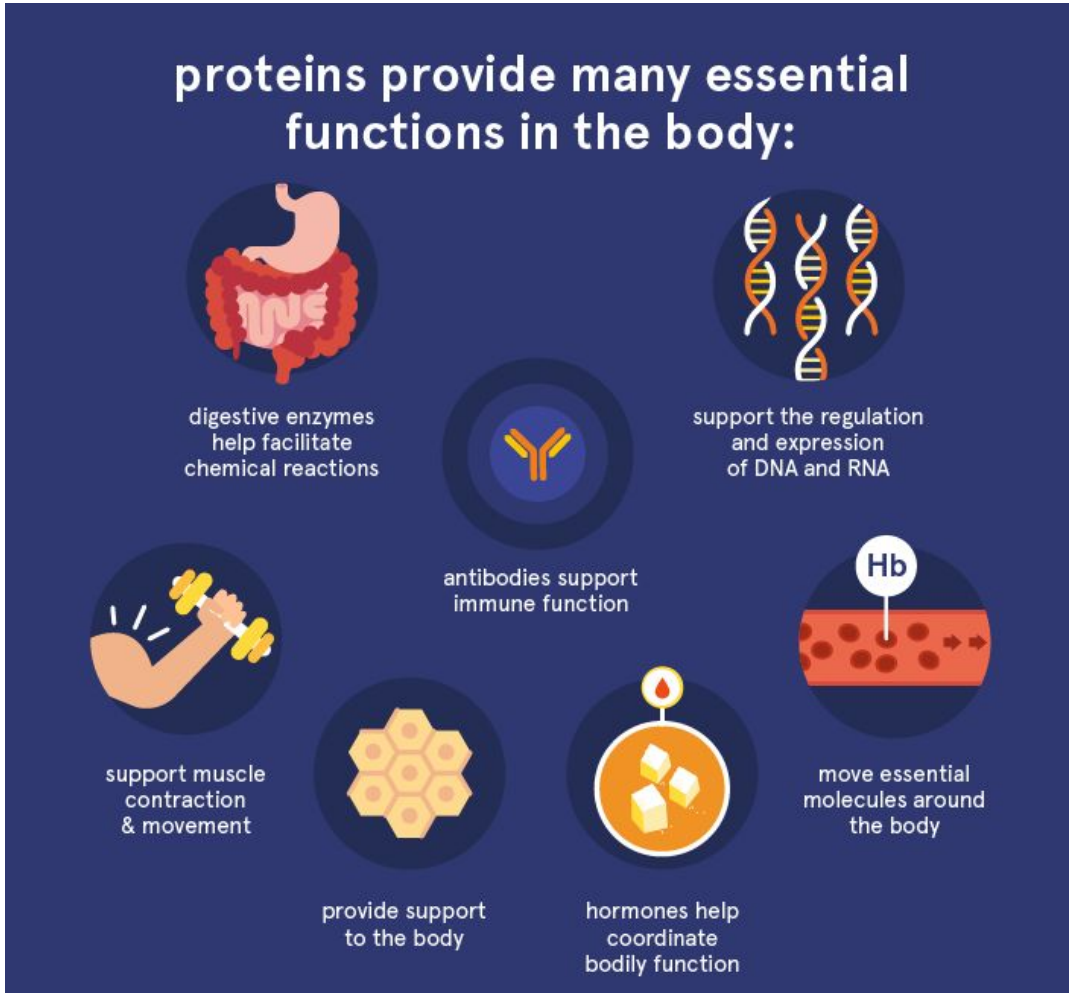
谭大治

dazhi@reverielabs.com

内容摘要

- 什么是蛋白质的结构？
- 为什么要研究解析蛋白质的结构？
- 如何解析蛋白质的结构？
  - 实验方法
  - 计算方法

蛋白质是几乎所有生命活动的执行者

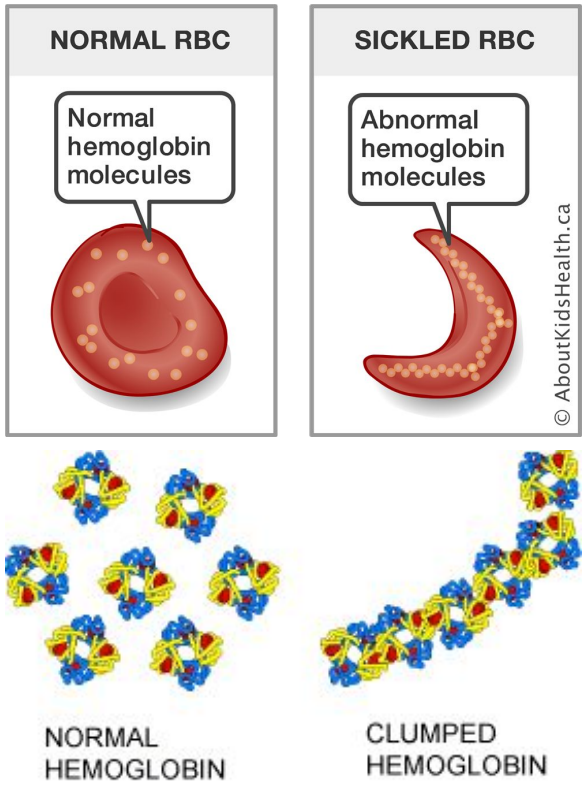


一些人体中蛋白质的例子：

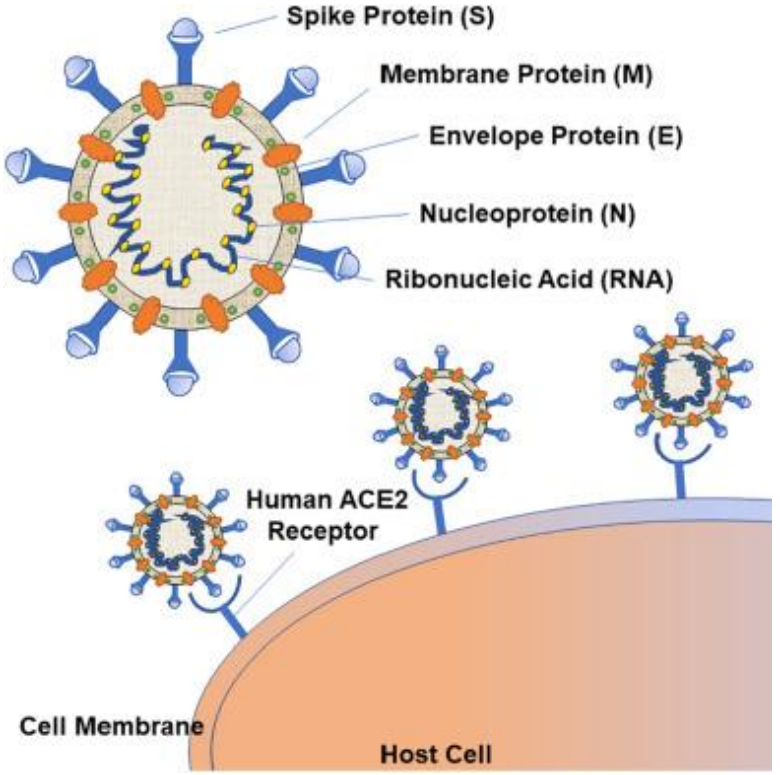
- 各种酶类(催化生化反应的胃蛋白酶，乙醛脱氢酶等)
- 抗体、干扰素(免疫反应)
- 肌红蛋白(肌肉运动)
- 血红蛋白(氧气输送)
- 组蛋白(遗传物质DNA的包裹组装)
- 角蛋白(毛发的主要组成部分)
- 胰岛素(新陈代谢的调节)
- ……

绝大多数人类疾病与蛋白质相关

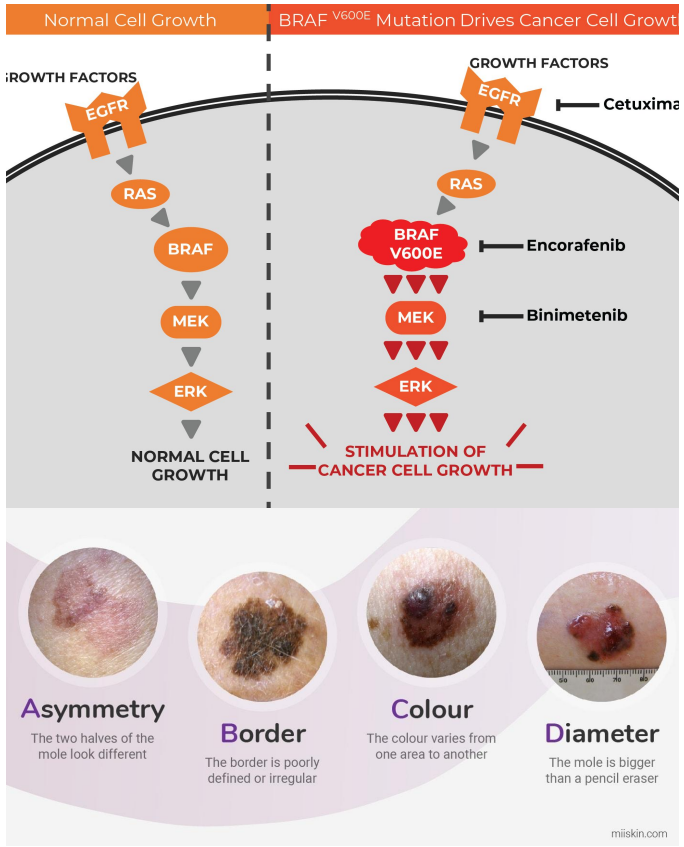
镰刀形细胞贫血症



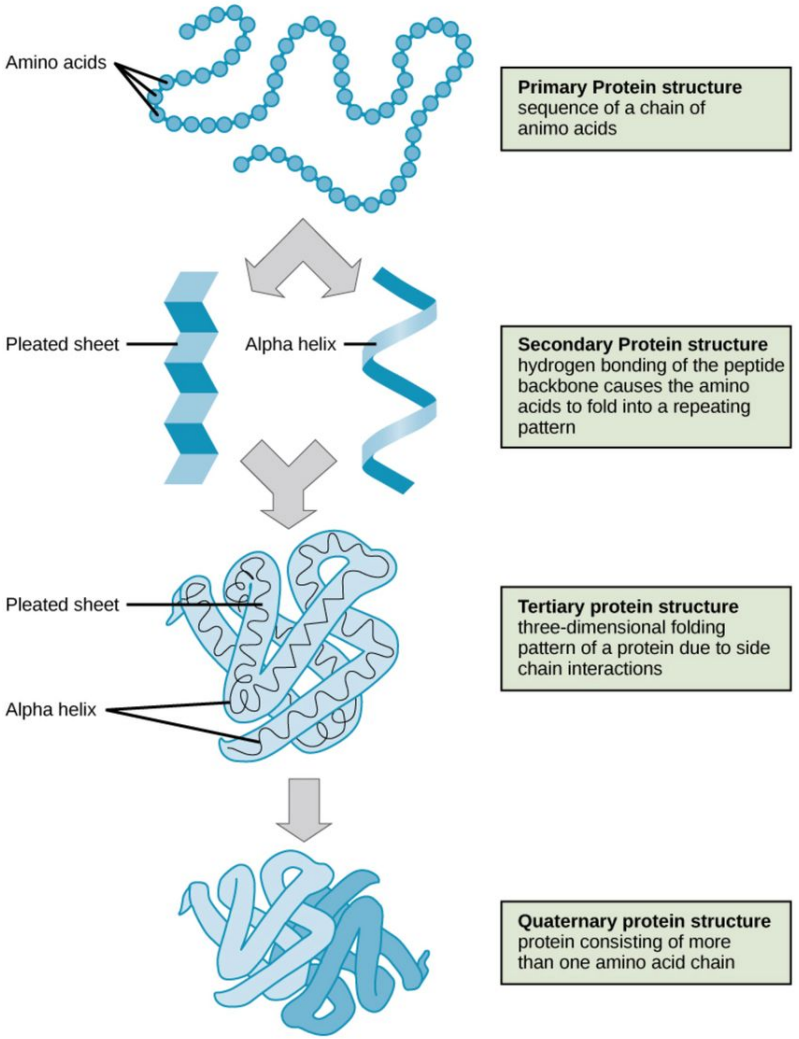
COVID-19



BRAF激酶变异导致的癌症



蛋白质多样的功能来源于其复杂的三维结构



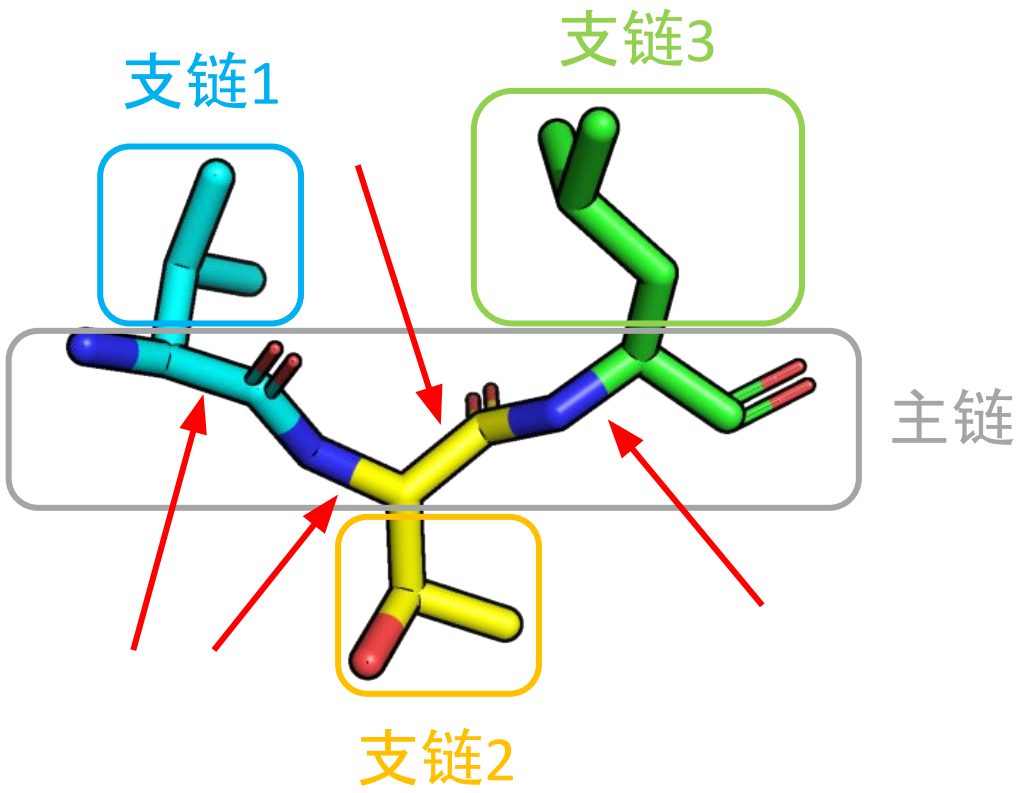
20种不同的氨基酸通过共价键连接形成线性多肽链，是为蛋白质一级结构。

多肽链局部通过非共价作用（主要是氢键）折叠形成例如α-螺旋或β片层的蛋白质二级结构。

二级结构再通过非共价作用（氢键、疏水作用、π电子堆积等）进一步整体折叠成蛋白质三级结构。

多个折叠好的多肽链再相互作用（或者与DNA或RNA作用）形成更加复杂的四级结构。

蛋白质结构有多复杂？



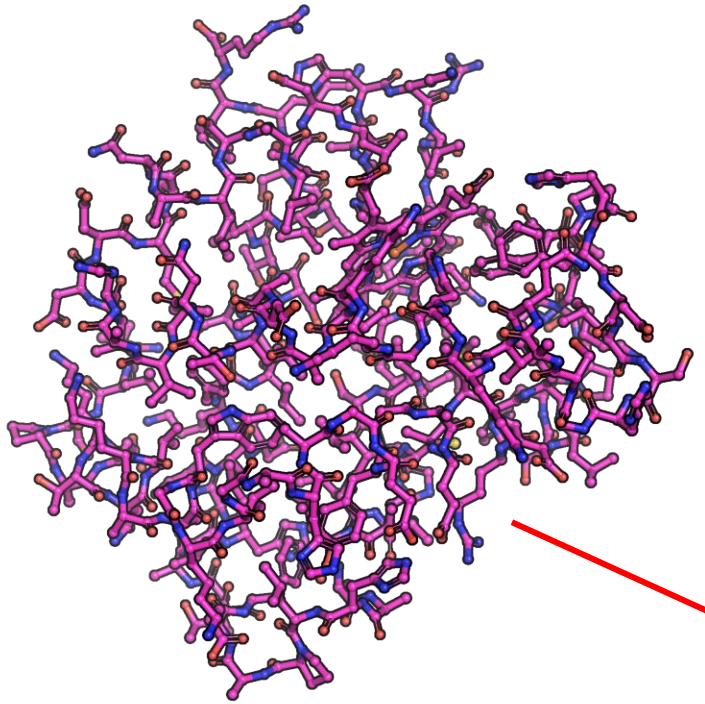
- 每一个氨基酸残基的支链有可能有1-55种不同的三维构型。平均有10种可能构型。
- 每两个氨基酸残基之间主链的连接也有可能有多种可能的构型。我们假设也有10种可能构型。
- 一个蛋白质分子可能含有几十至几千个氨基酸残基。一个典型的蛋白质分子约有300个氨基酸残基。

- 理论上该蛋白质可能的构型有$10^{300}*10^{299}=10^{599}$种。宇宙中原子的数目大约为$10^{80}$个。
- 当然，绝大多数的理论构型实际中都不可能存在。但即便如此，现实中蛋白质可以形成的结构仍然是一个天文数字。

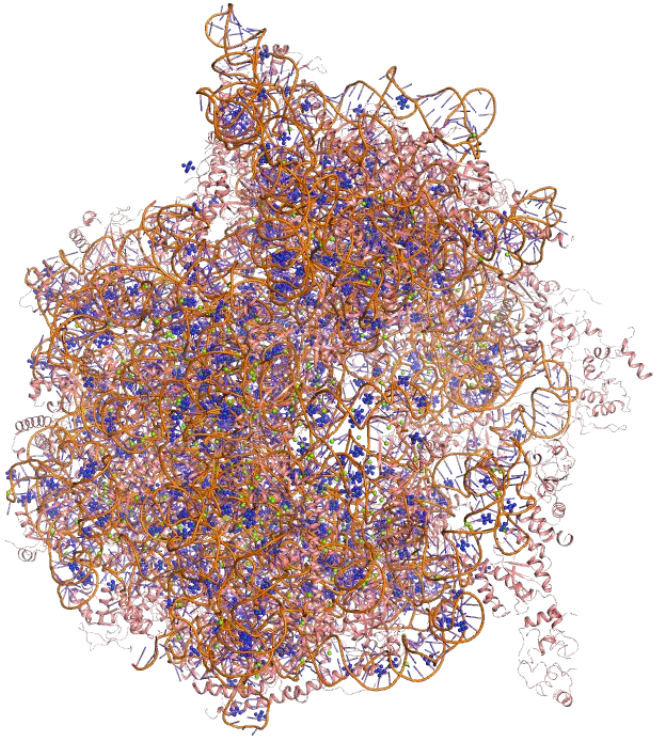
蛋白质结构有多复杂？

人血红蛋白单链结构
“卡通”视图

人血红蛋白单链结构
球棒视图

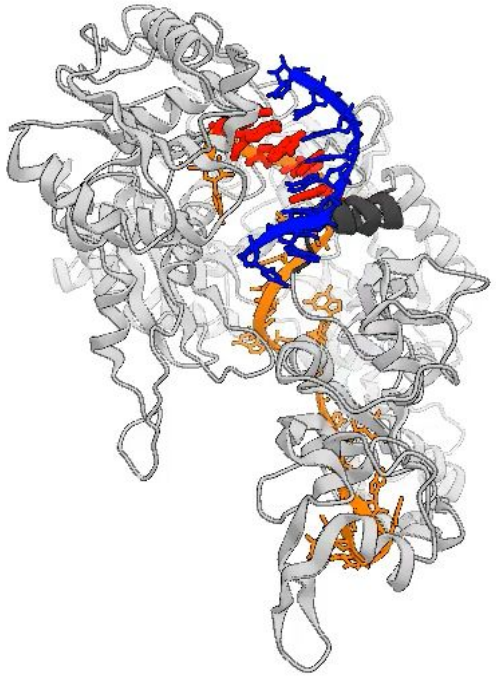


酿酒酵母(*S. cerevisiae*)核糖体结构
卡通视图



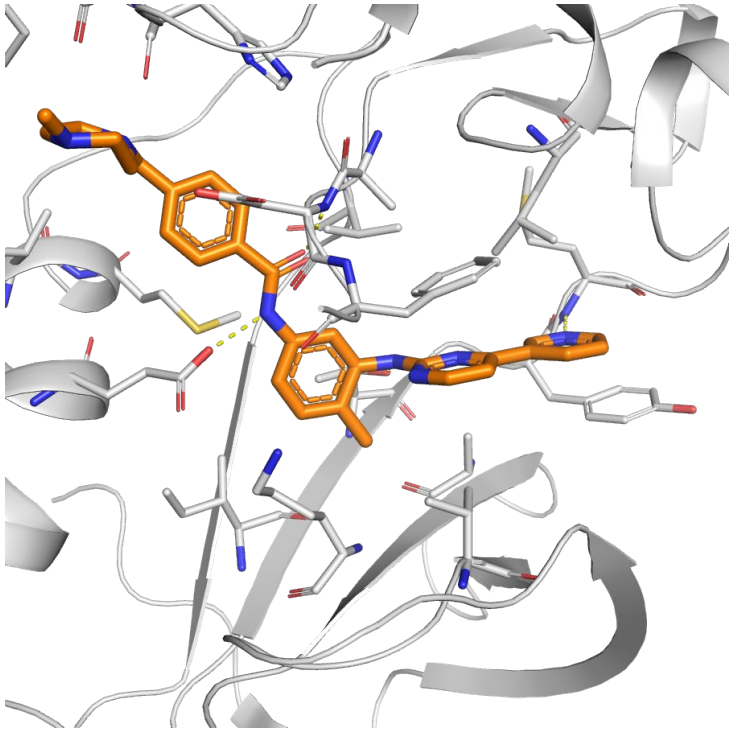
研究蛋白质结构的意义

从分子尺度理解生命活动的机理



指导我们设计新的疗法与药物

PD-1与其特异性抗体的复合体结构



Gleevec与Src-Abl激酶的结构

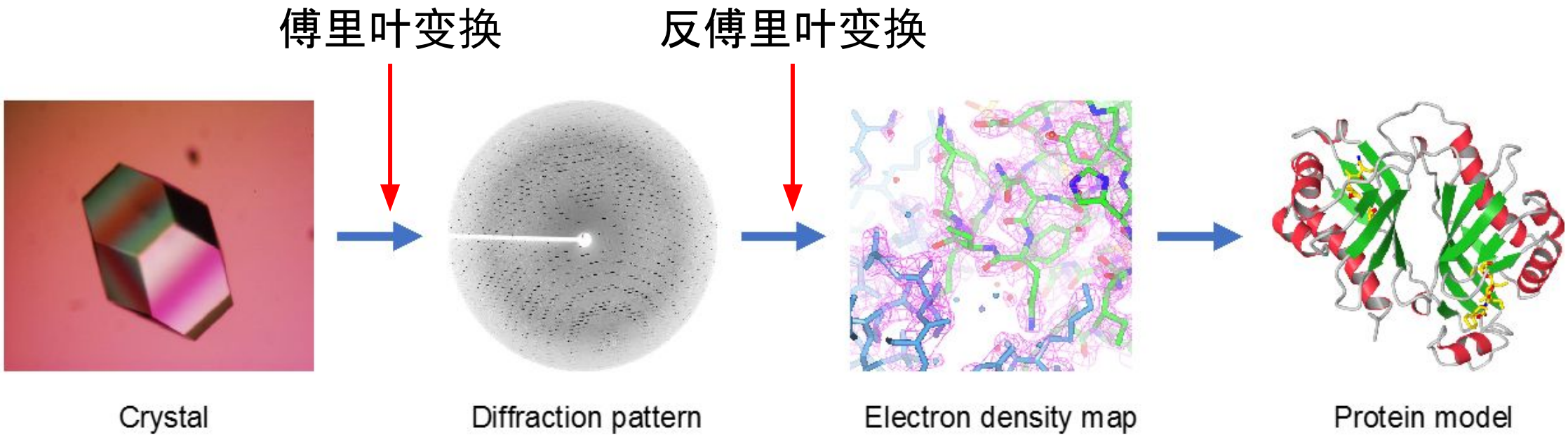
解析蛋白质结构的实验方法

- 大多数蛋白质的线度都在数百埃（1埃=0.1纳米）以下。光学显微镜无法直接观测到单个蛋白质分子。

- 目前主流的解析蛋白质结构的实验方法：

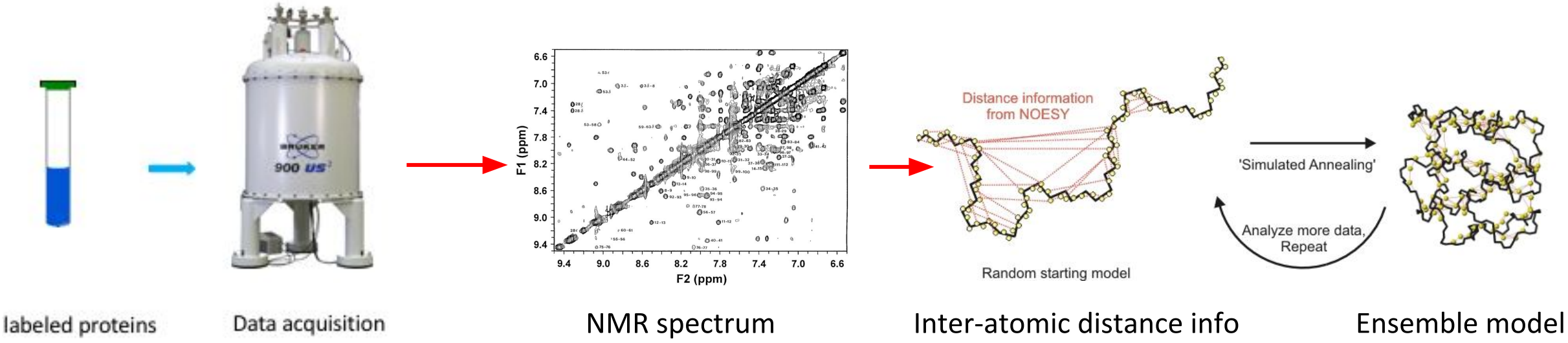
  - X光晶体学 (X-ray Crystallography): 肇始于1950年代。最传统也是产出最高的方法。仍为最主流(?)的结构解析方式。
  - 核磁共振(NMR): 1980年代开始进入实际应用
  - 冷冻电镜(Cryo EM): 1990年代起步，2010年后开始爆发式发展。有取代晶体学成为最主流的结构生物学方法。

X光晶体学



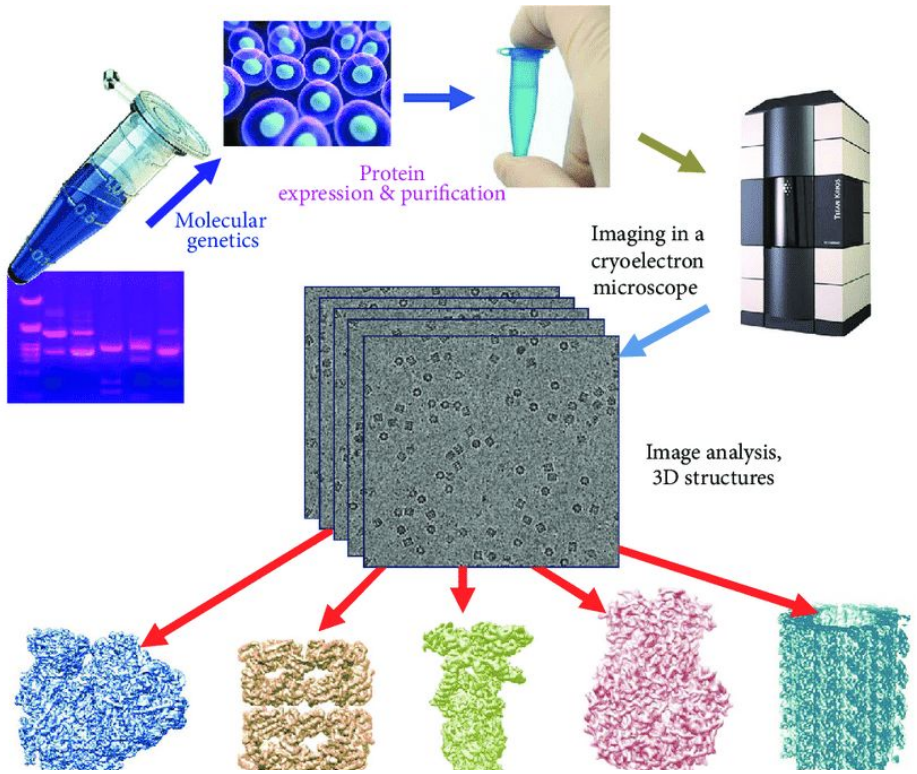
自1962年以来，有超过10次诺贝尔奖被授予生物大分子晶体学相关研究，包括1997, 2003, 2006, 2009, 2012和2020年诺贝尔化学奖

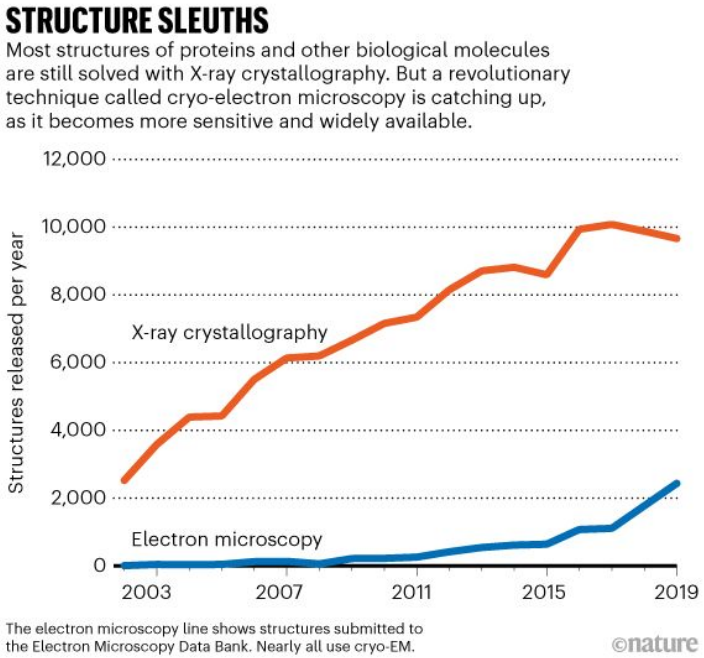
核磁共振

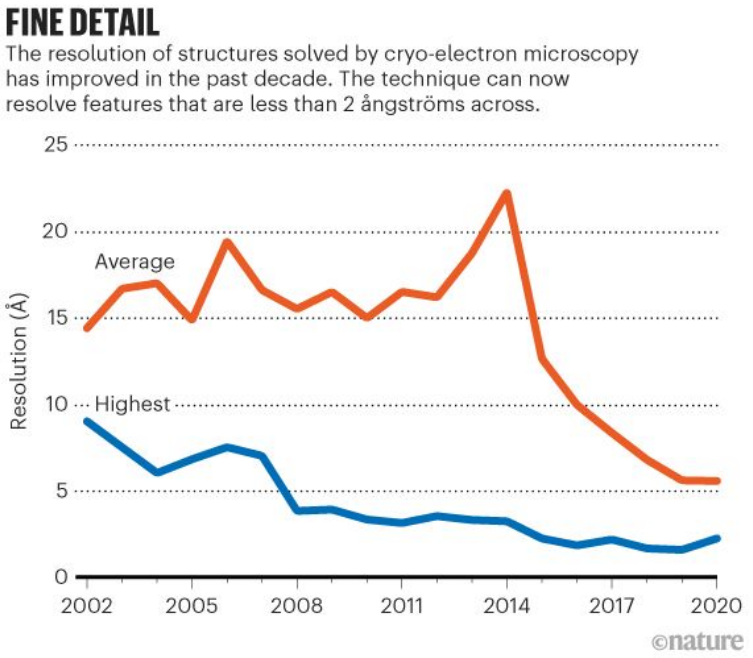


2002年诺贝尔化学奖被授予Kurt Wüthrich以表彰其利用核磁共振解析蛋白质结构方面的贡献。

冷冻电镜







2017年诺贝尔化学奖被授予三位科学家以表彰其利用冷冻电镜解析高分辨率蛋白质结构方面的贡献。

各种实验方法的利弊

方法	利	弊
晶体学	• 非常成熟且廉价的技术 • 可以解析各种大小的蛋白的结构 • 大多数结构可以获得非常高的分辨率	• **蛋白需要能够结晶!!!（玄学）** • 晶体中的蛋白结构可能与生理条件下的结构有偏差 • 难以获知蛋白结构动态的信息
核磁共振	• 目标蛋白无需能够结晶 • 可以探究溶液中蛋白的动态结构	• 一般只适用于较小的蛋白 • 所需要蛋白量较大 • 需要对蛋白进行同位素标记 • 所需实验设备比较稀缺
冷冻电镜	• 目标蛋白无需能够结晶 • 非常善于解析分子量极大的复杂复合体的结构 • 也相对容易获知蛋白结构动态的信息	• 相当大比例的结构的分辨率仍不尽人意 • 较难解析较小的蛋白质的结构

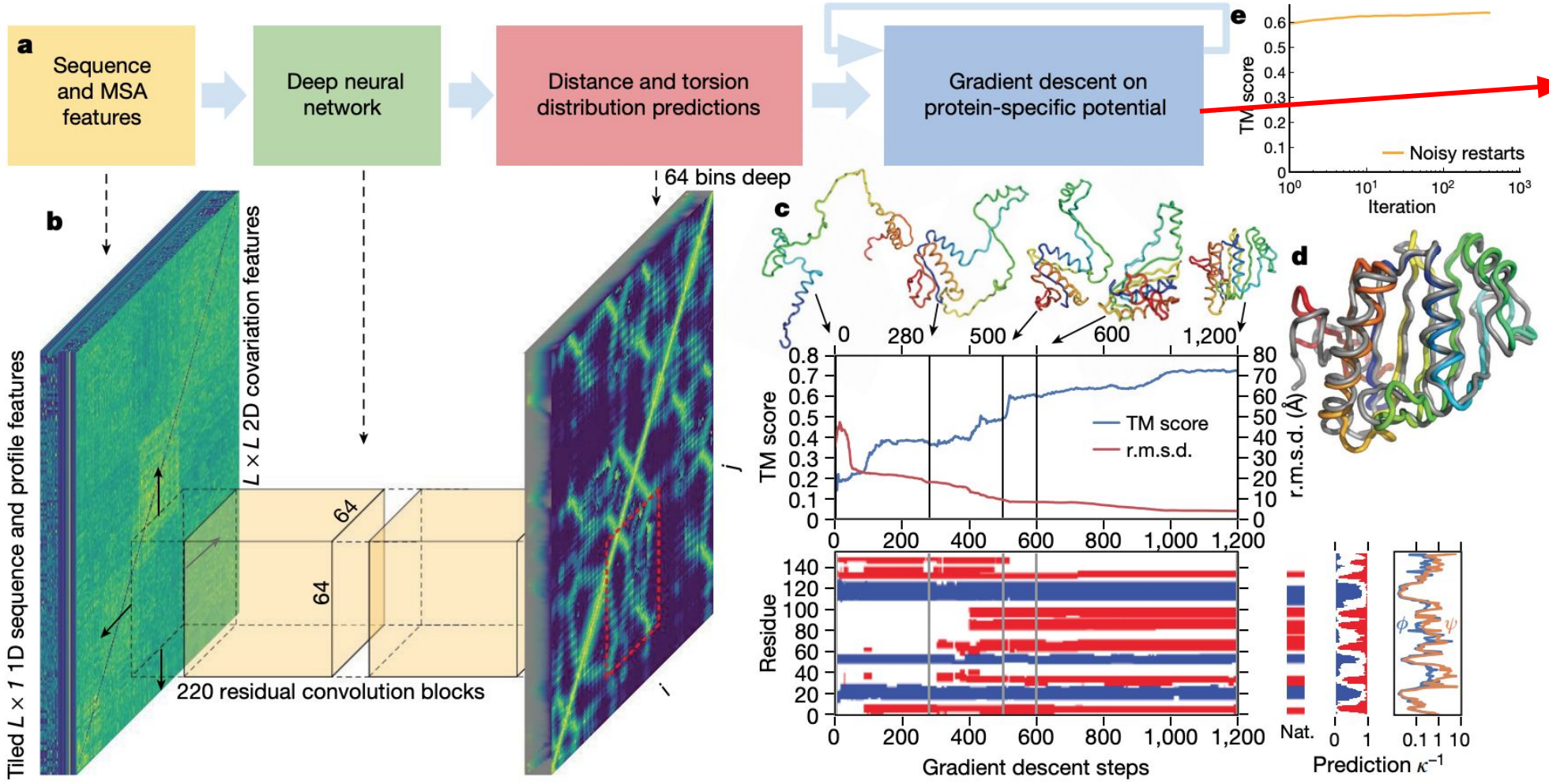
所有实验方法的共同限制条件：目标蛋白需要能够被体外表达并纯化，且其结构在样本制备过程中需要保持相对稳定。

预测蛋白质结构的计算方法

- 其实之前提到的实验方法也需要大量的计算来辅助后期的模型搭建

- 蛋白质结构预测的计算方法通过计算来取代以上图表中的“实验数据”：
  - 基于物理学的方法：如粗粒模型模拟及分子动态模拟等
  - 基于已有结构的经验方法：AlphaFold2
  - 混合方法：Rosetta，AlphaFold1

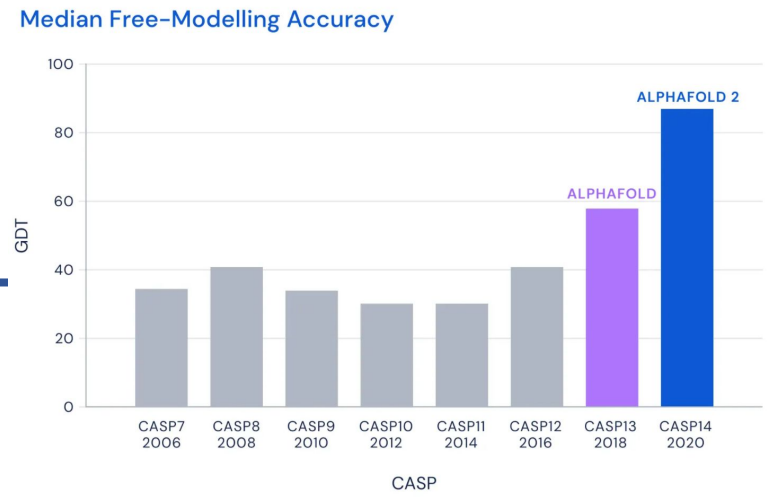
AlphaFold1

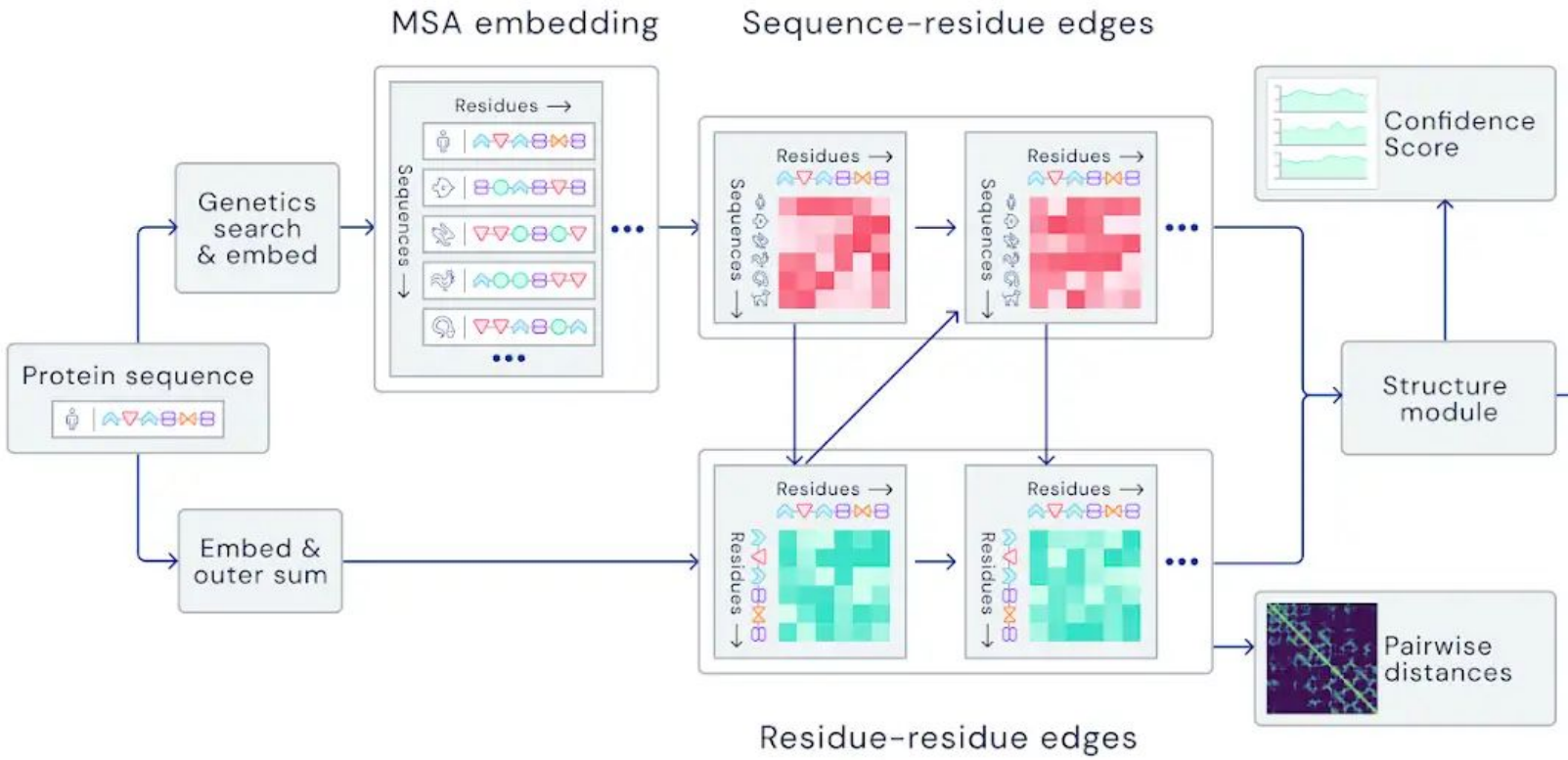


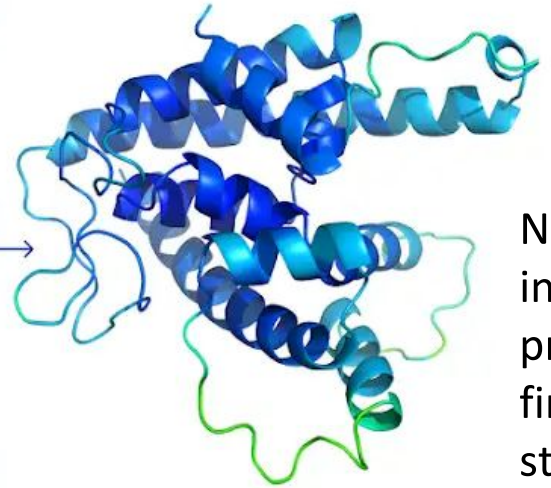
Combination of ML-based potential from AlphaFold and physics-based potential from Rosetta.

Senior et al., Nature (2020)

AlphaFold2







3D structure

No physical information is provided until the final refinement step

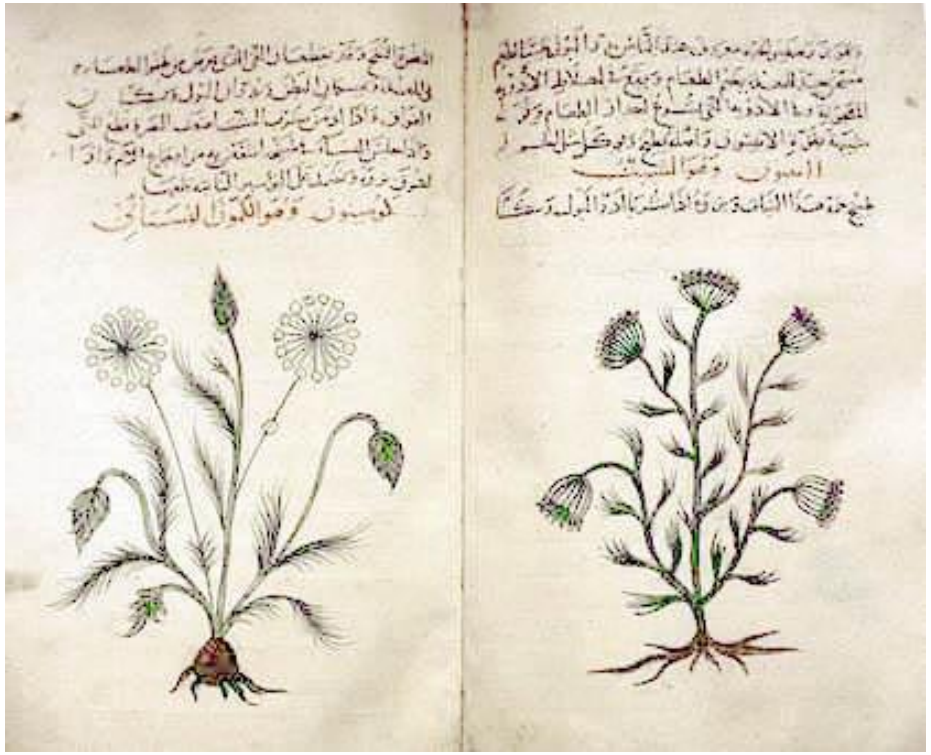
XtalPi

AlphaFold会改变什么

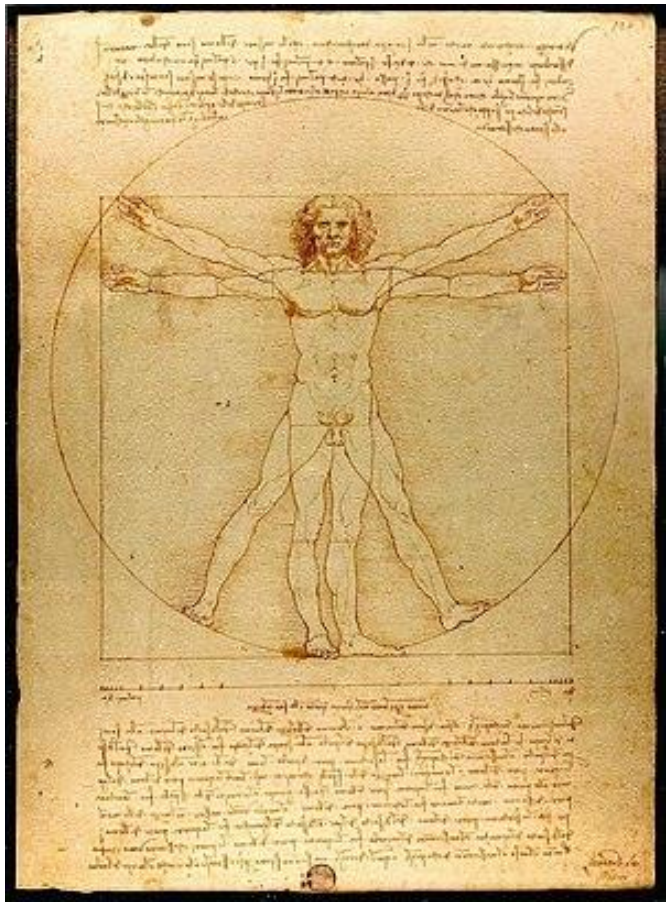
—— 药物研发及AI应用简介

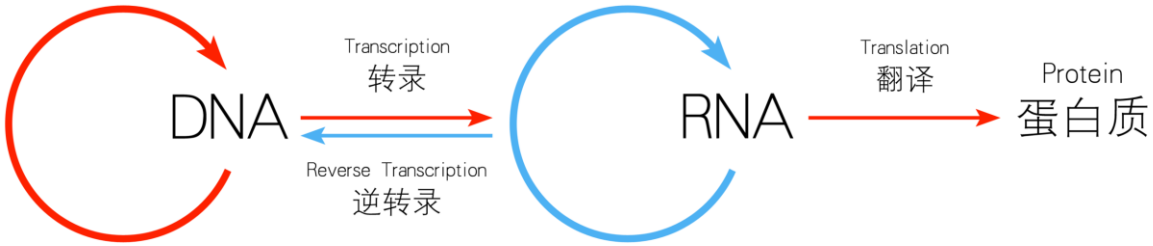
晶泰科技联合创始人 赖力鹏

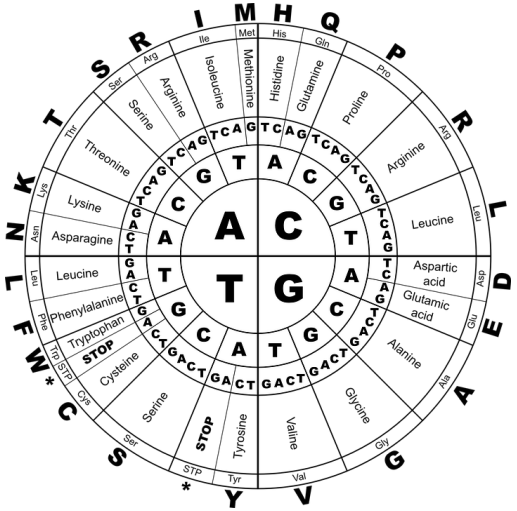
健康是人类一直以来的追求

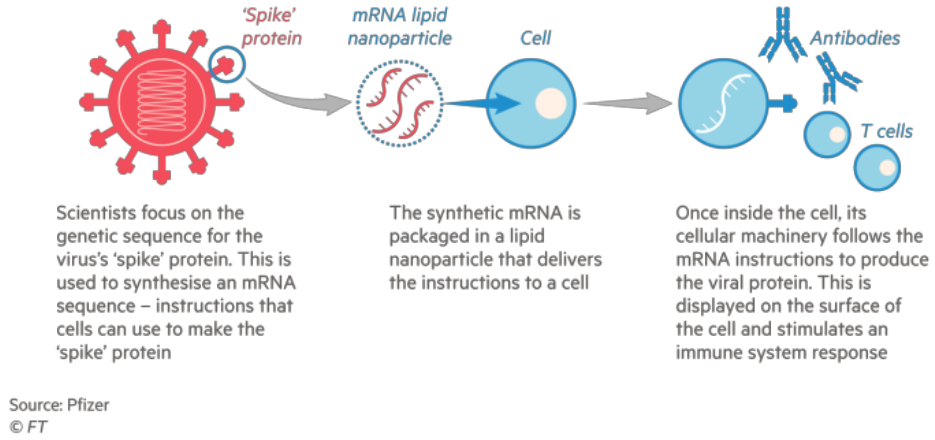


健康是人类一直以来的追求





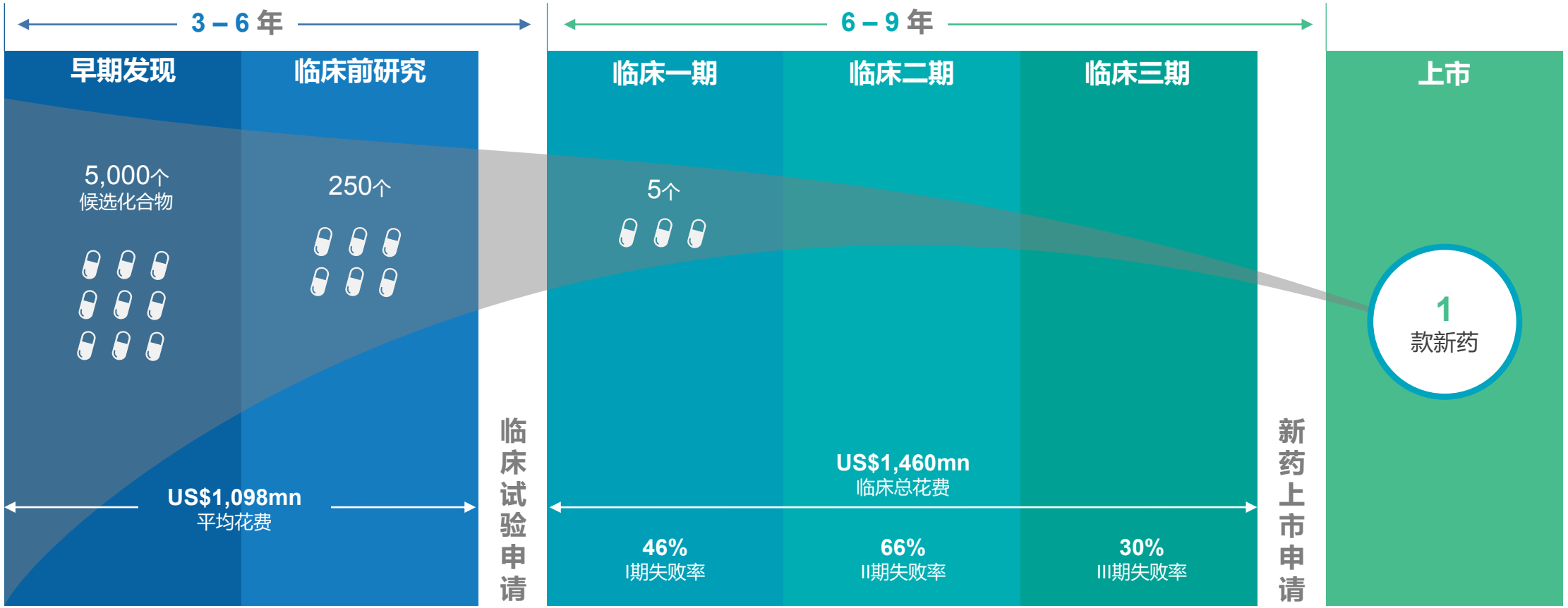




“他就想活命，他有什么罪”

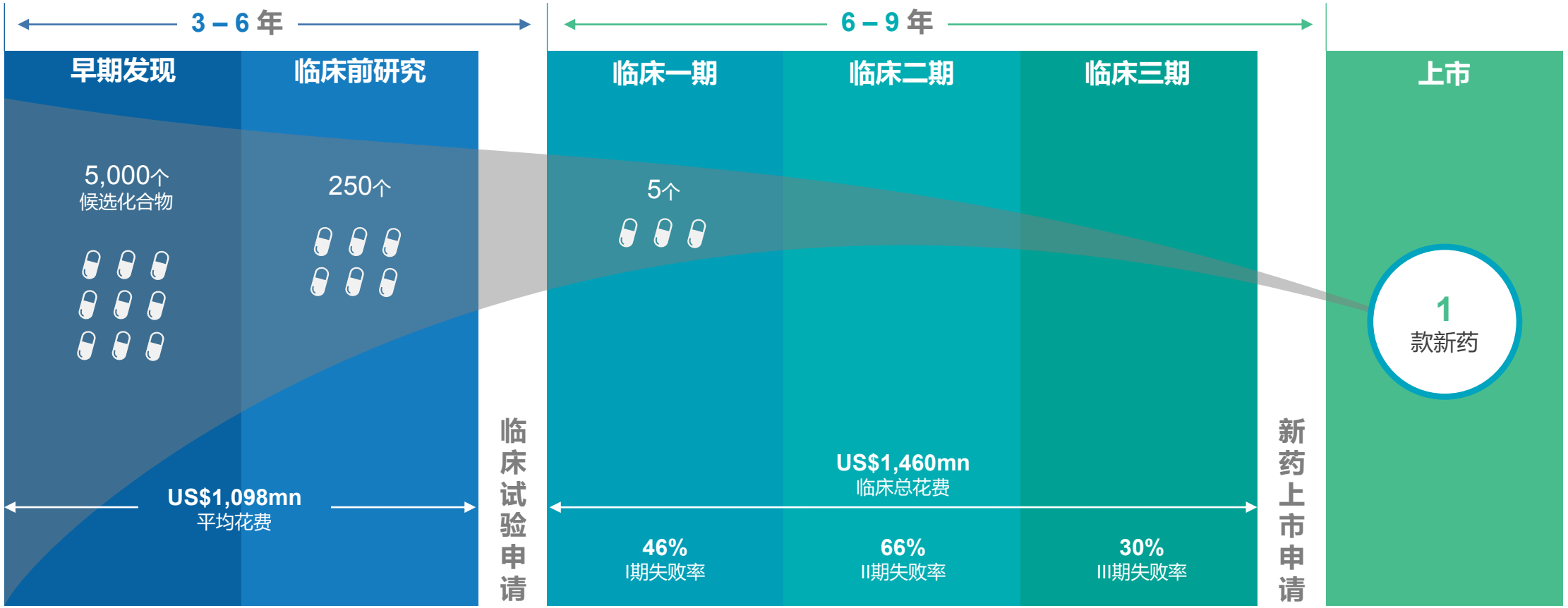
晶泰的技术将降低药物研发成本和患者用药成本，提高成功率

药物筛选：在浩瀚可能性中的搜索



数据来源：MarketResearch, Deloitte Consulting, 方达医药招股书

传统新药研发面临投入高，耗时长，风险高的痛点



1 投入高

26亿美元/新药

新药平均研发费用

2 耗时长

9-15年

新药研发总时长

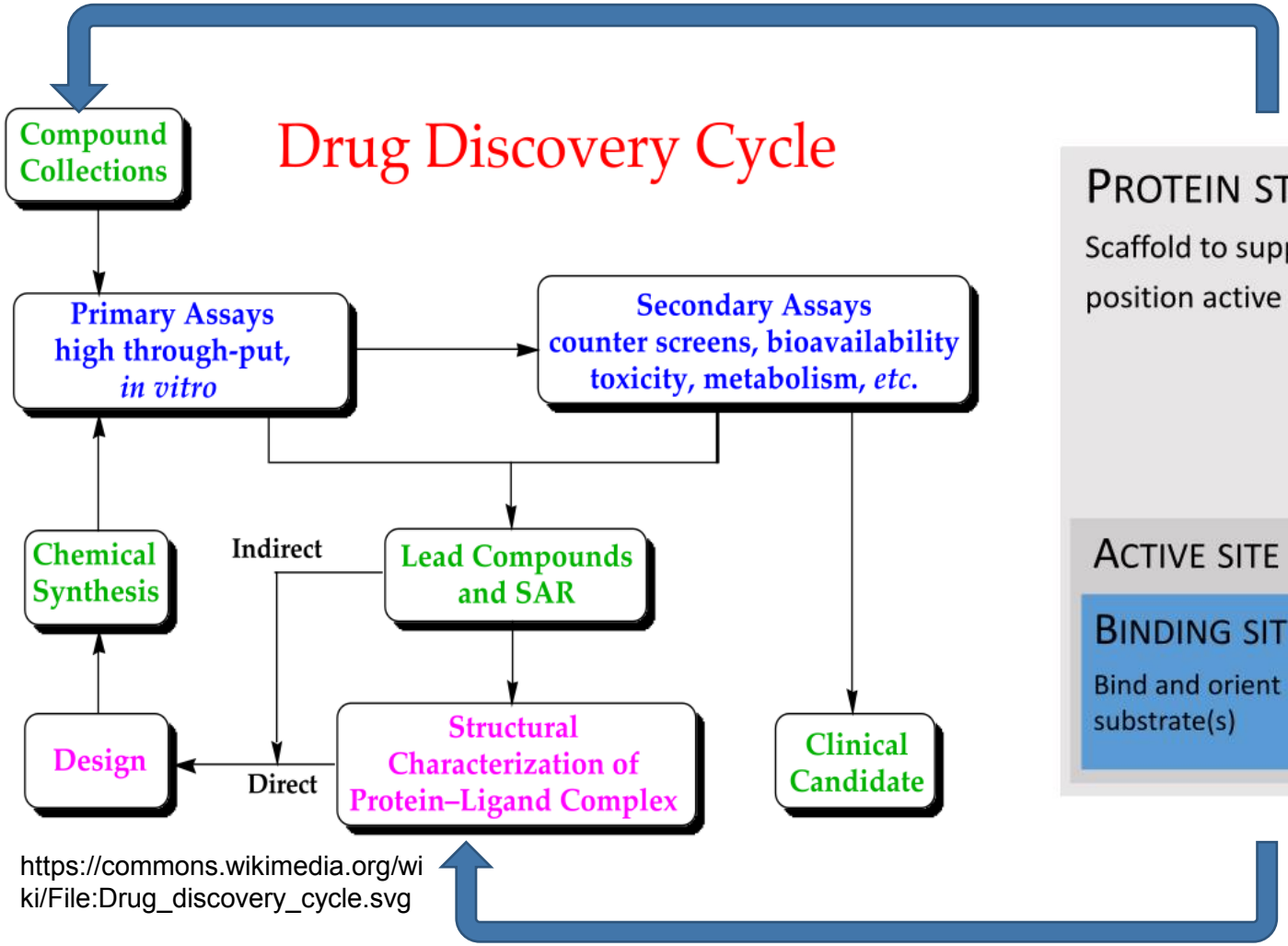
3 风险高

~10%

临床阶段成功率

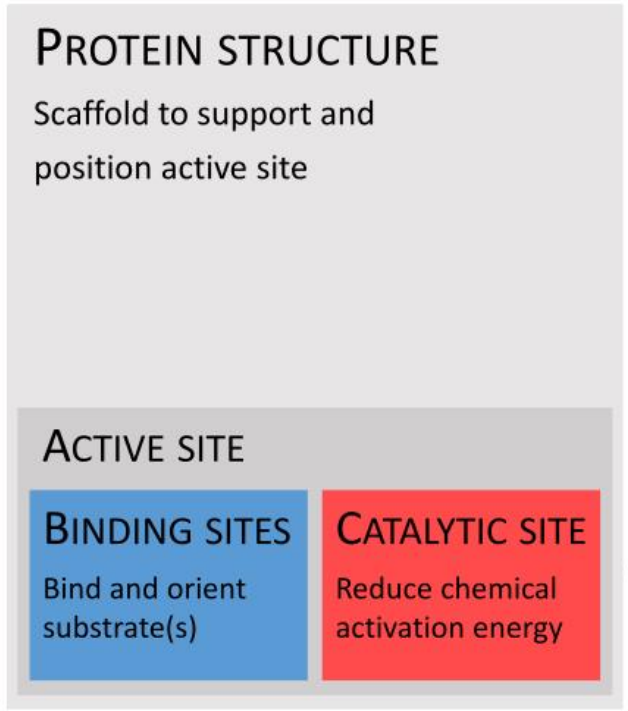
数据来源：MarketResearch, Deloitte Consulting, 方达医药招股书

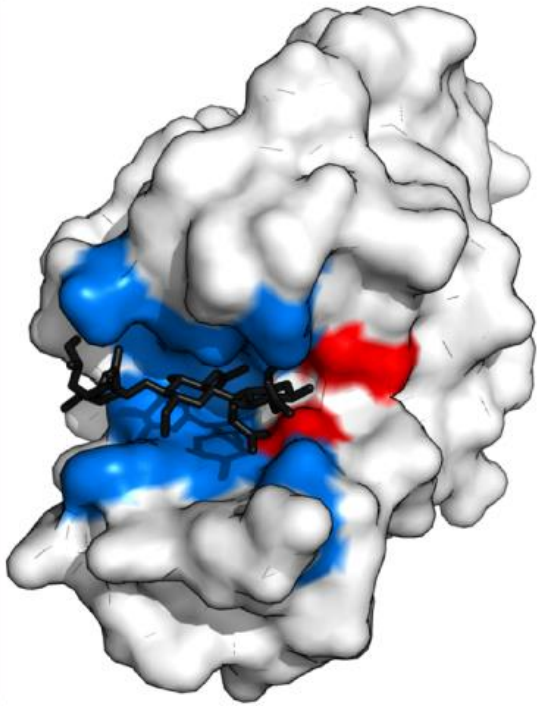
药物发现的过程



https://commons.wikimedia.org/wiki/File:Drug_discovery_cycle.svg

Structure Based Drug Design

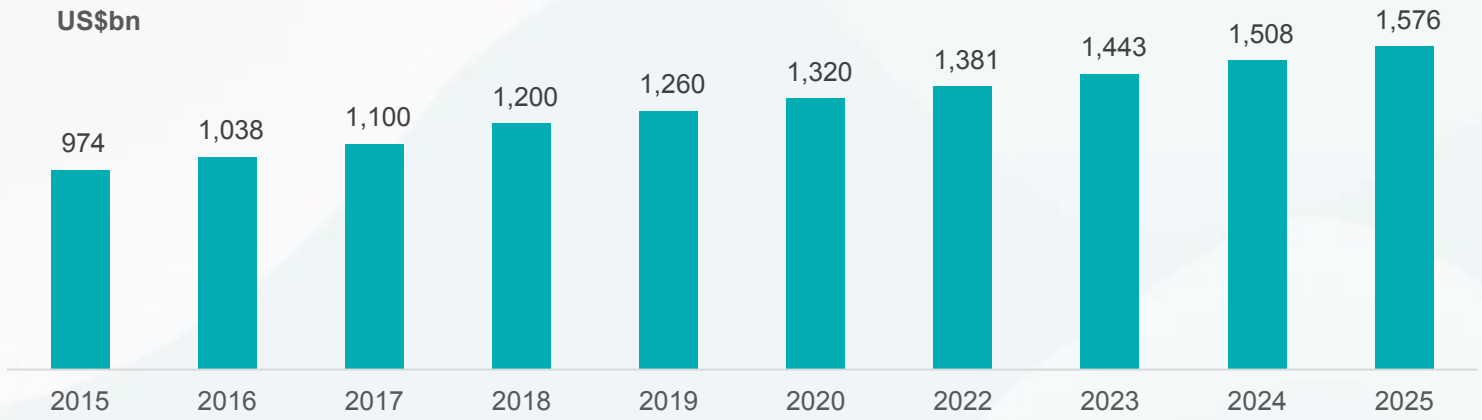




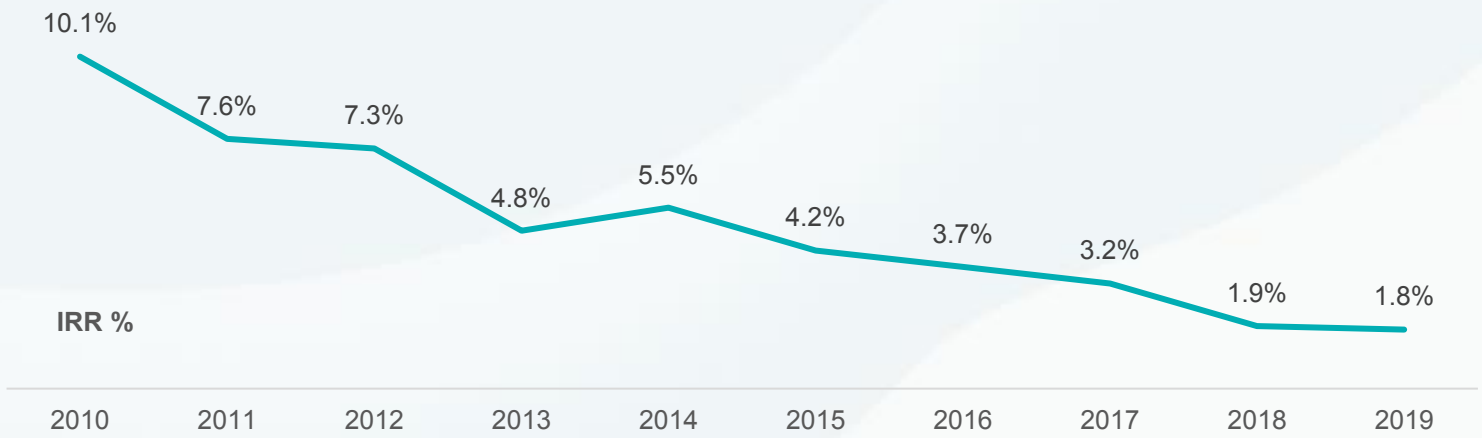
https://commons.wikimedia.org/wiki/File:Enzyme_structure.svg

万亿美元药物市场面临回报率持续下滑挑战，且未满足市场空间巨大

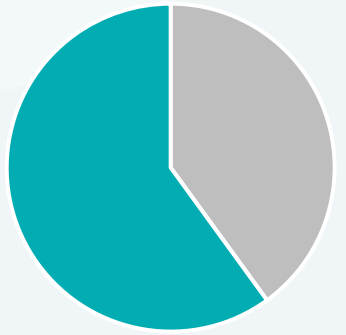
■ 2024年全球药物市场容量将突破1.5万亿美元



■ 药企研发回报率逐年下滑



数据来源：MarketResearch, Deloitte Consulting



60%
疾病无有效药物

50%-70%
患者对重磅药物无响应

AI的发展历史

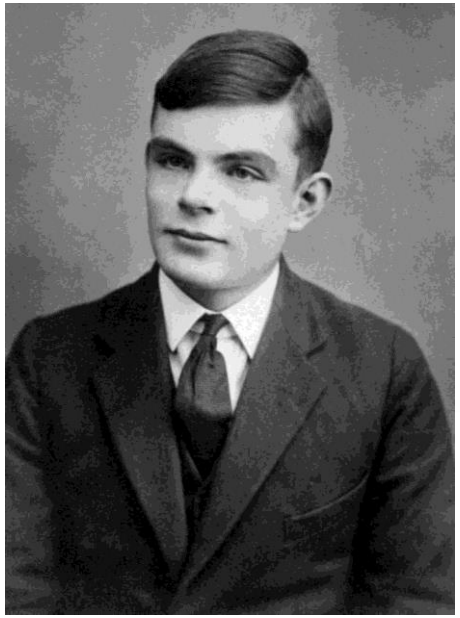
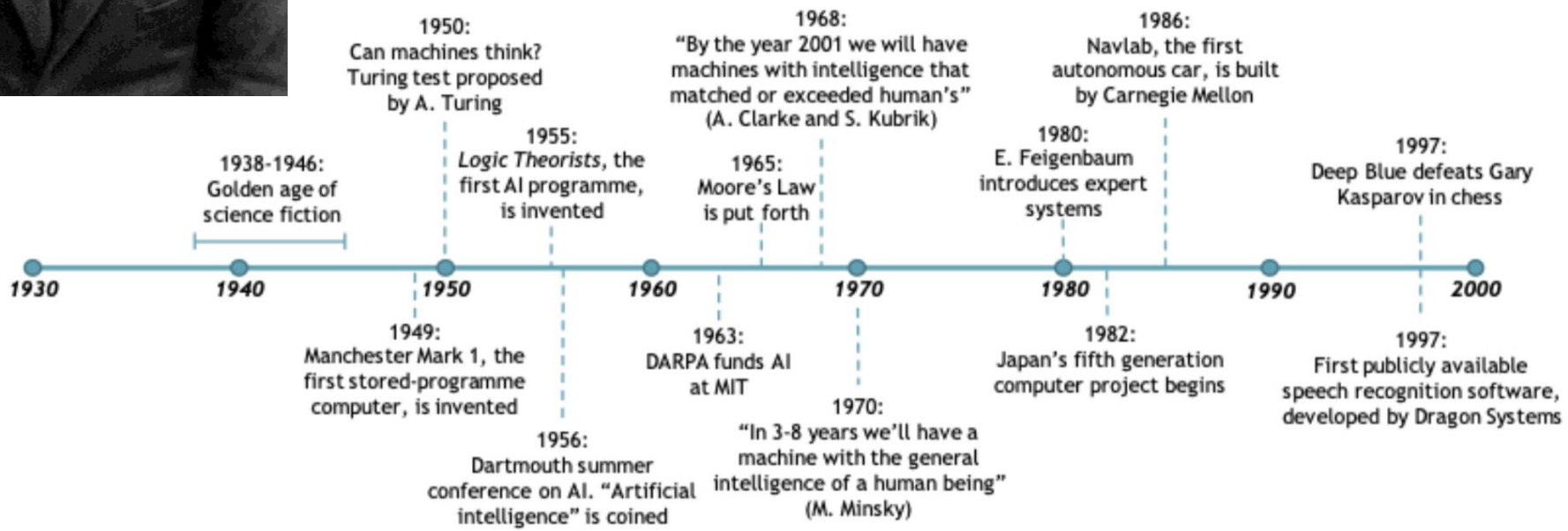
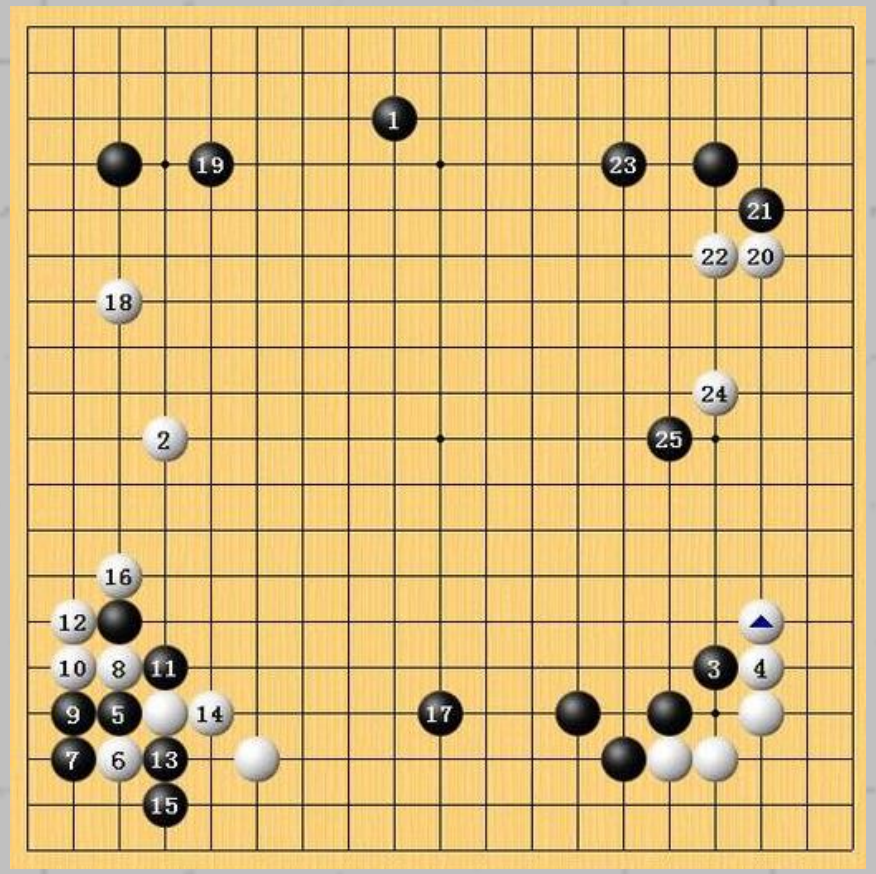
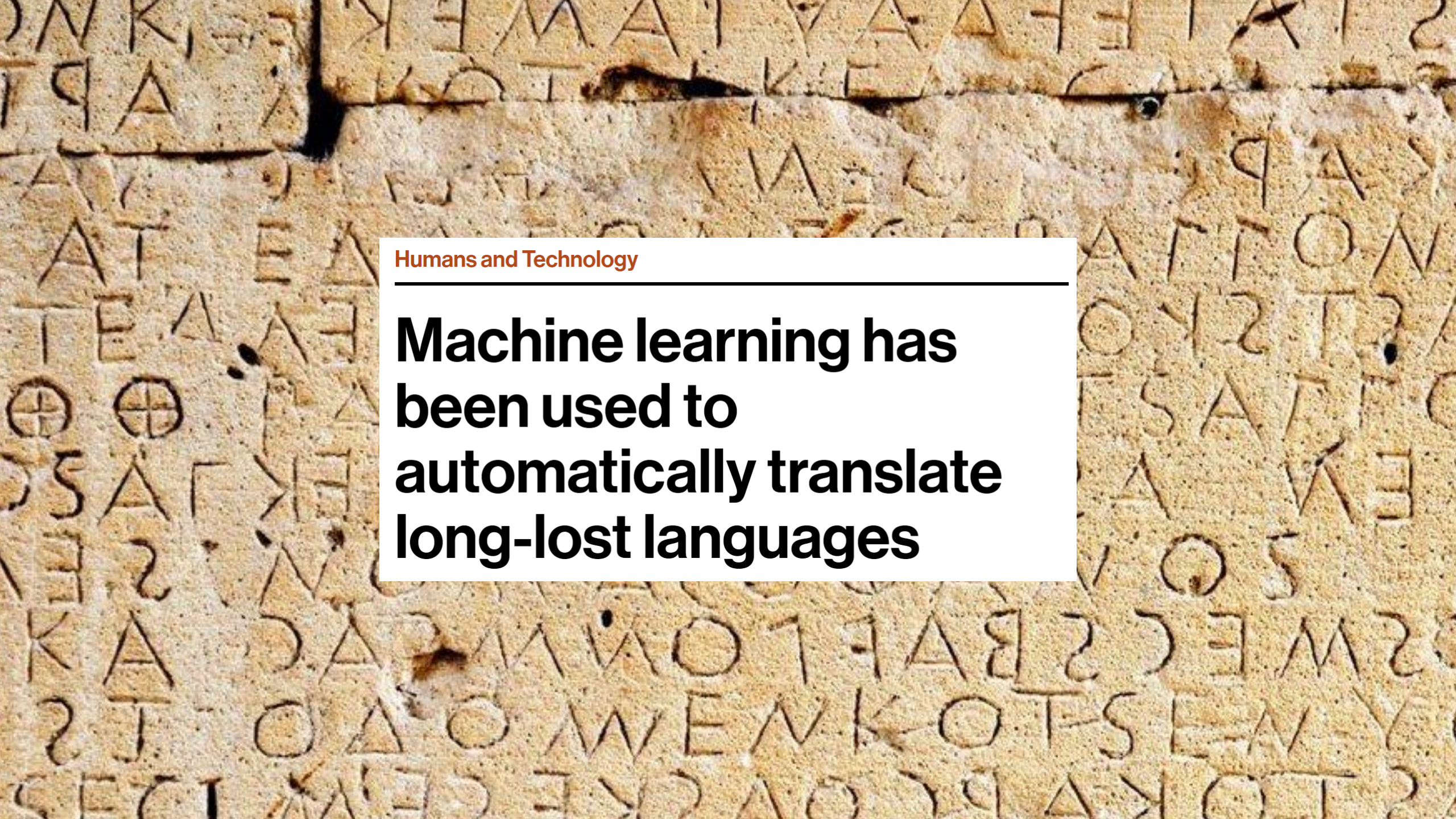


Figure 1.1. Timeline of early AI developments (1950s to 2000)



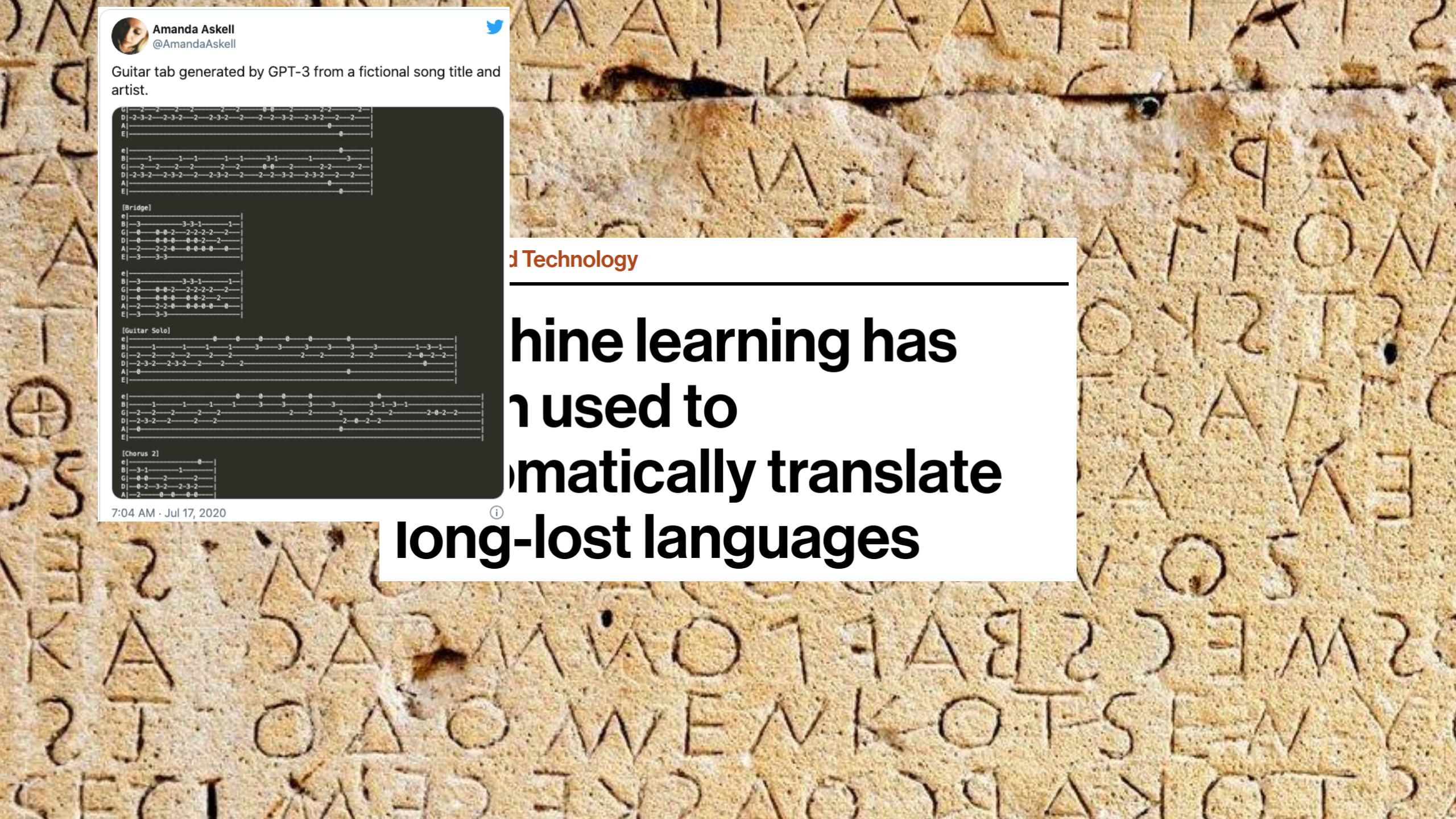
Source: Adapted from Anyoha (28 August 2017[4]), "The history of artificial intelligence", http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.


1
19
23
21
22
20
18
24
2
25
16
12
10
8
11
3
4
9
5
14
17
7
6
13
15



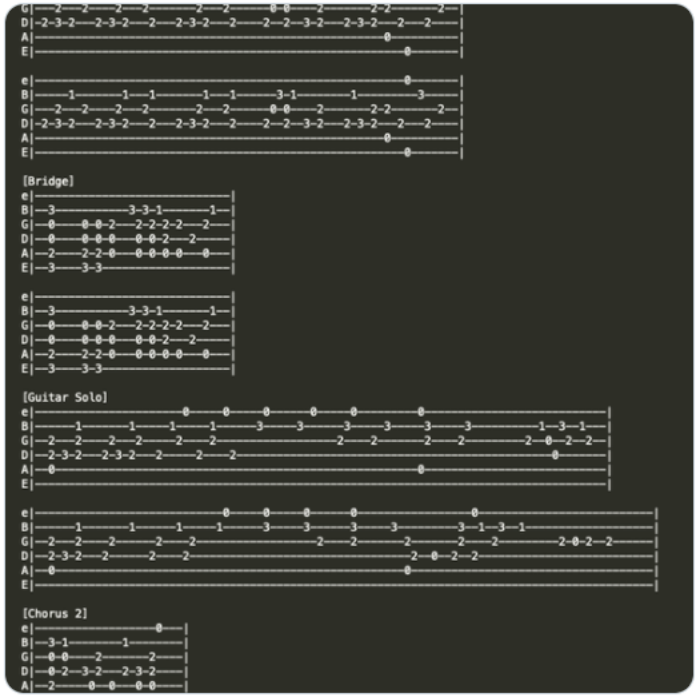
Humans and Technology

Machine learning has been used to automatically translate long-lost languages


Amanda Askell
@AmandaAskell
Guitar tab generated by GPT-3 from a fictional song title and artist.
[Bridge]
[Guitar Solo]
[Chorus 2]
7:04 AM · Jul 17, 2020
d Technology
hine learning has
n used to
matically translate
long-lost languages

Amanda Askell
@AmandaAskell

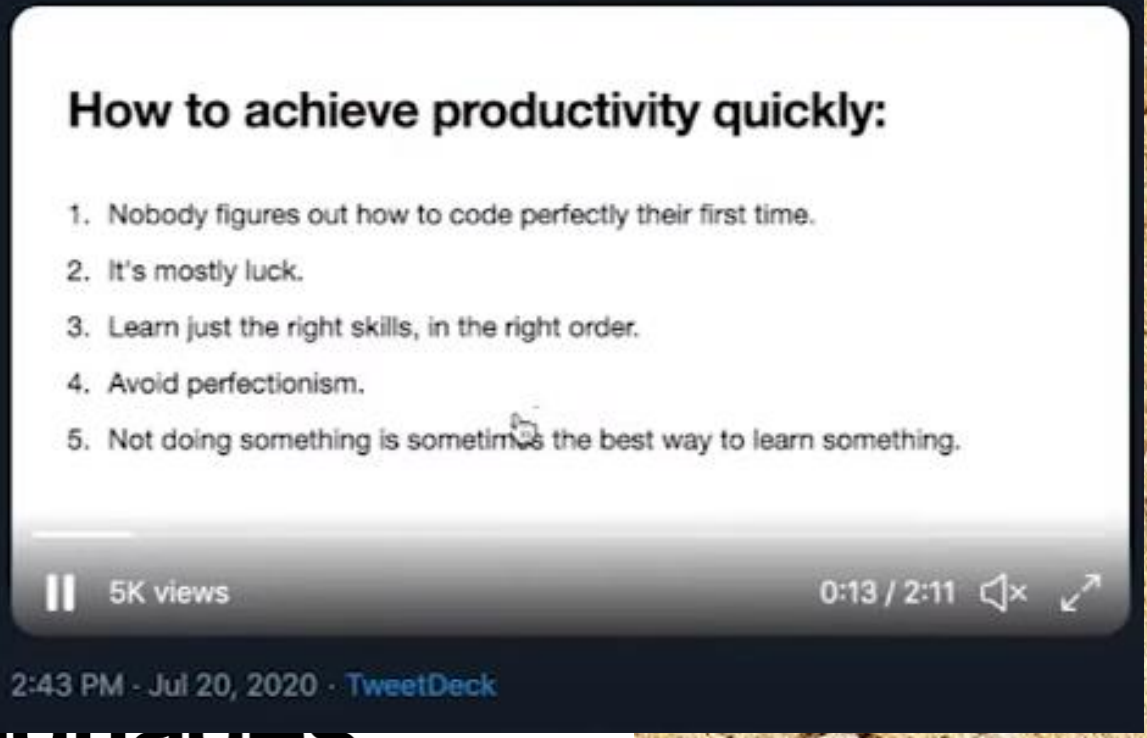
Guitar tab generated by GPT-3 from a fictional song title and artist.



7:04 AM · Jul 17, 2020

@bemmu

I asked GPT-3 to generate a presentation for me, and here's what it came up with.



2:43 PM · Jul 20, 2020 · TweetDeck

d Technology

hine lea
n used t
matical
long-lost languages

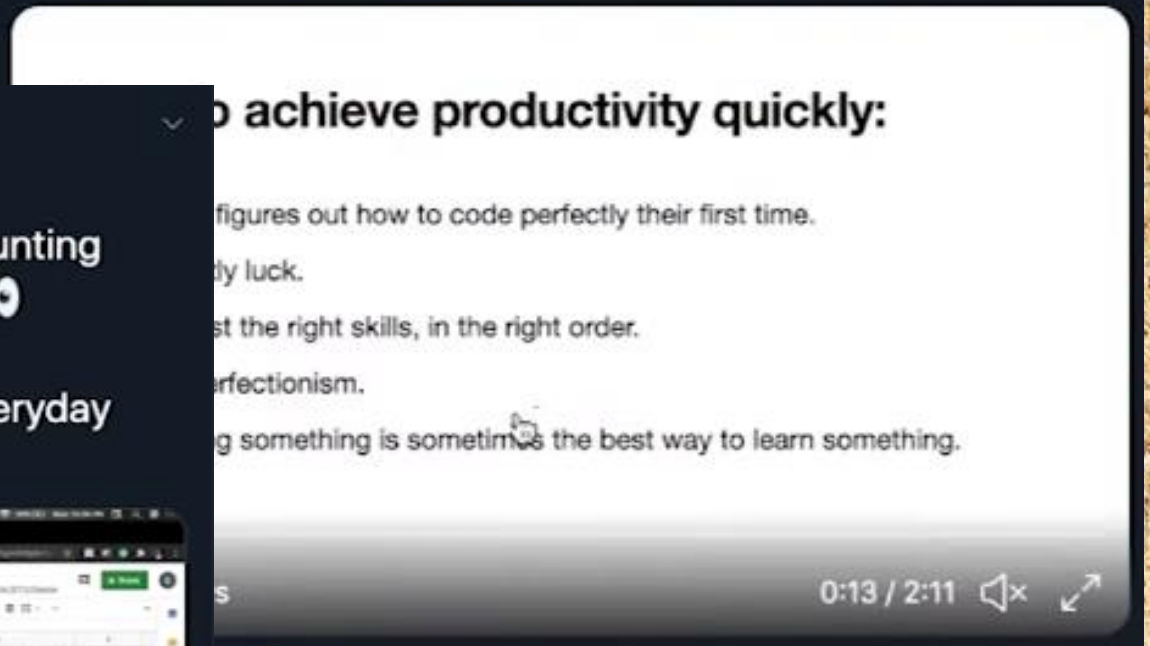
Amanda Askell
@AmandaAskell

Guitar tab generated by GPT-3 from a fictional song title and artist.

7:04 AM · Jul 17, 2020

@bemmu

I asked GPT-3 to generate a presentation for me, and here's what it came up with.

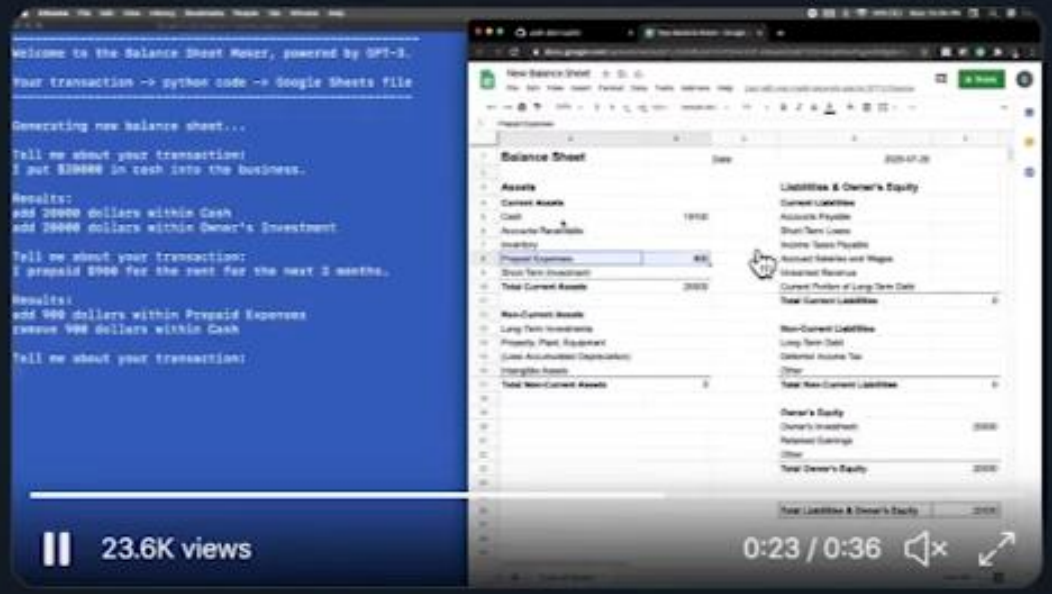


0, 2020 · TweetDeck

yash
@itsyashdani

Built a GPT-3 bot that lets people with no accounting knowledge generate financial statements 📈 👀

Here it is creating balance sheets by turning everyday language into Python code:



5:59 PM · Jul 21, 2020 · Twitter Web App

Amanda Askell
@AmandaAskell

Guitar tab generated by GPT-3 from a fictional song title and artist.

7:04 AM · Jul 17, 202

@bemmu

I asked GPT-3 to generate a presentation for me, and here's what it came up with.

o achieve productivity quickly:

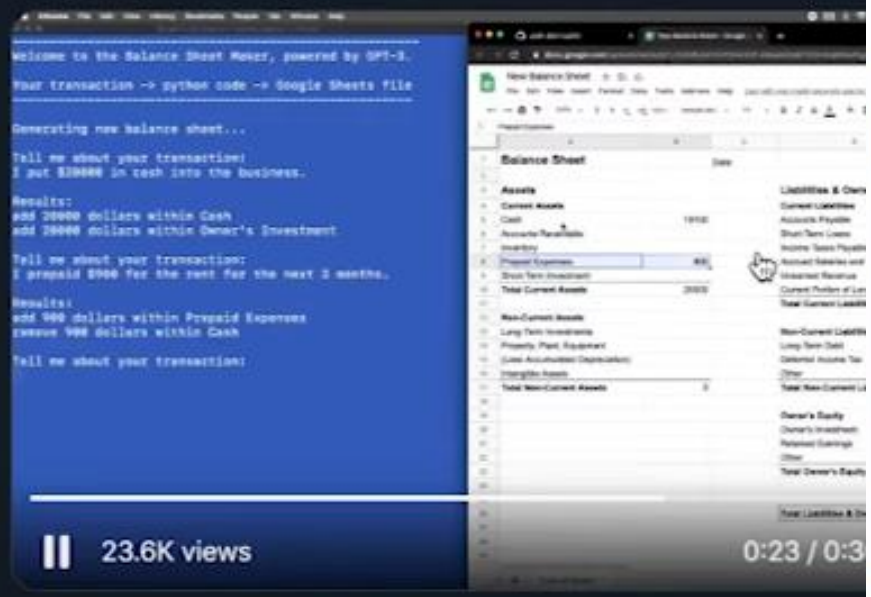
figures out how to code perfectly their first time.

dy luck.

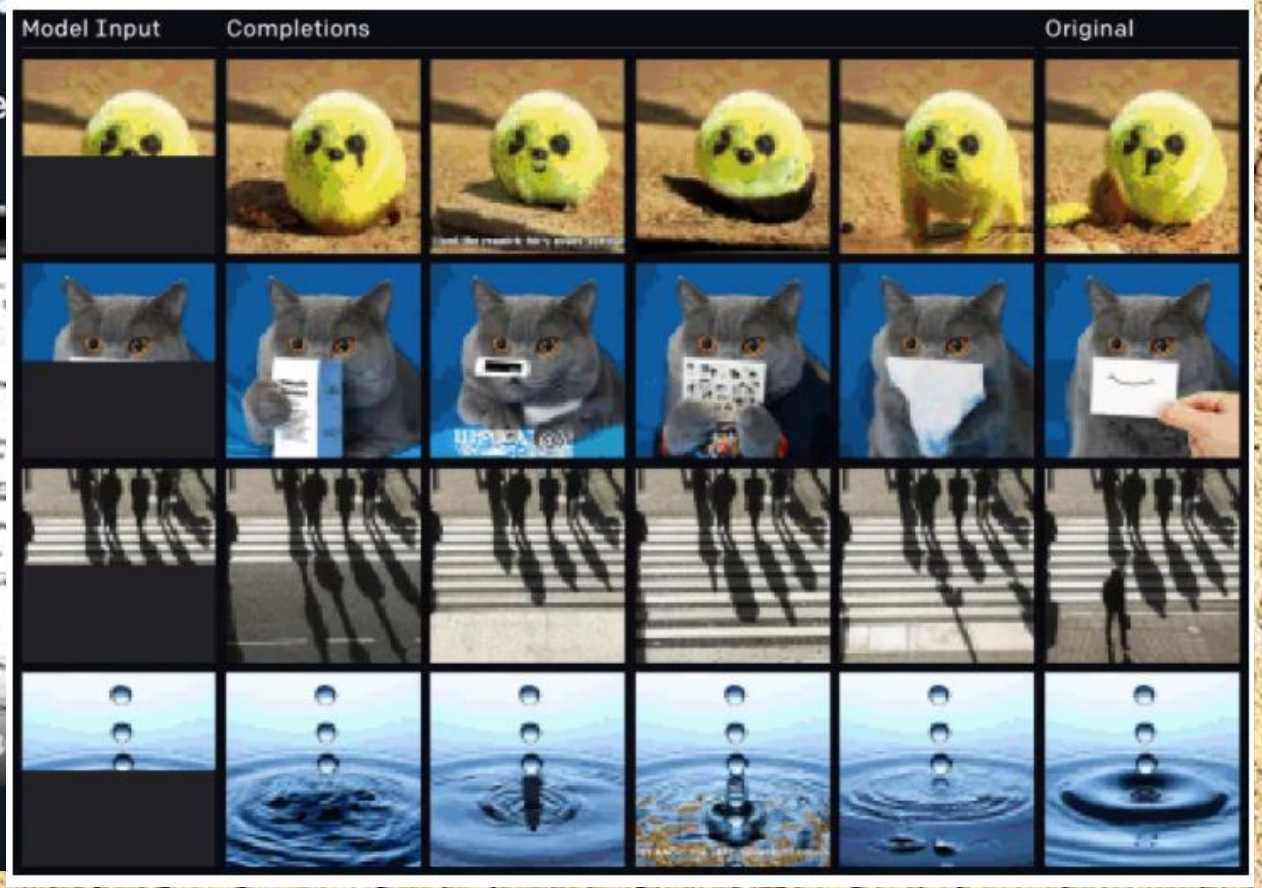
yash
@itsyashdani

Built a GPT-3 bot that lets people with no accounting knowledge generate financial statements 📈 👀

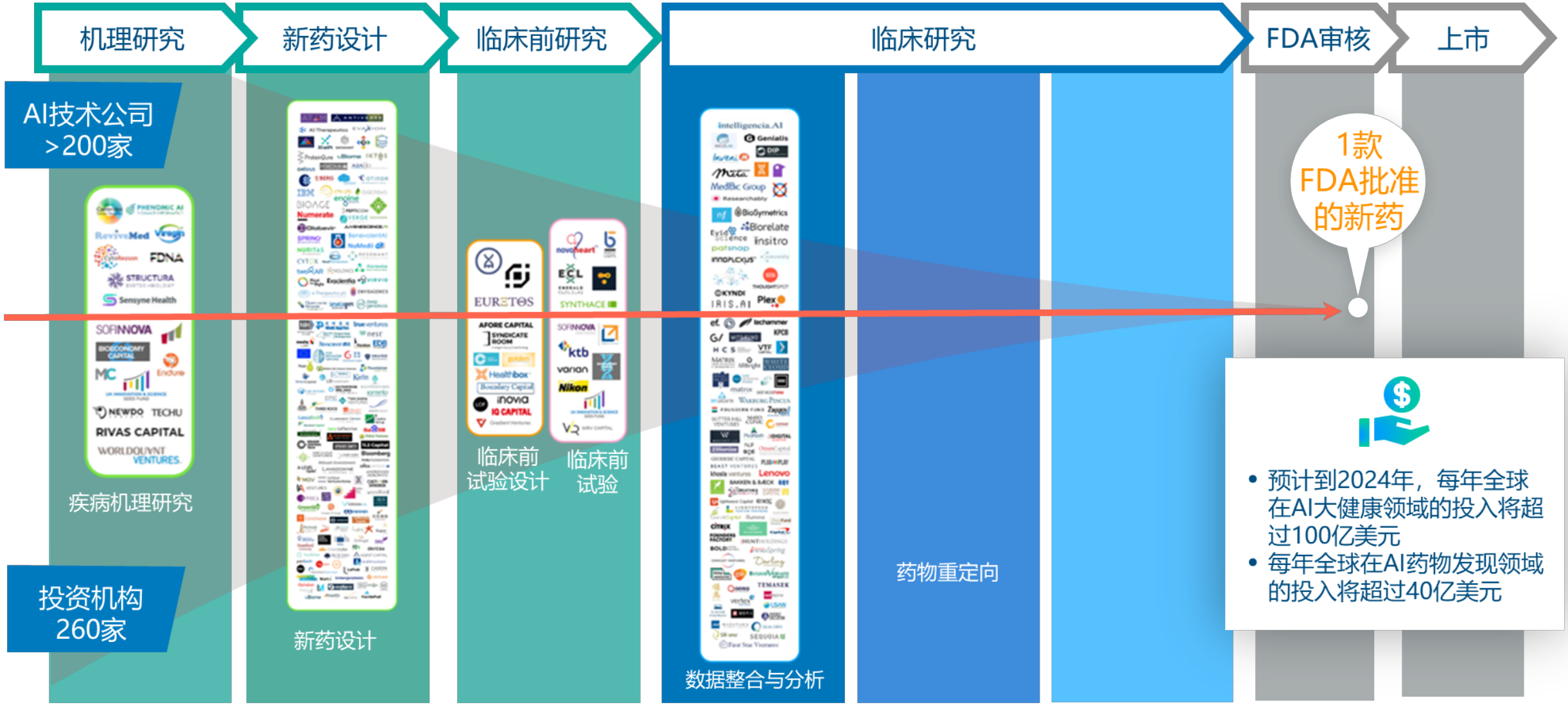
Here it is creating balance sheets by turning eve
language into Python code:



5:59 PM · Jul 21, 2020 · Twitter Web App



AI药物研发的快速发展



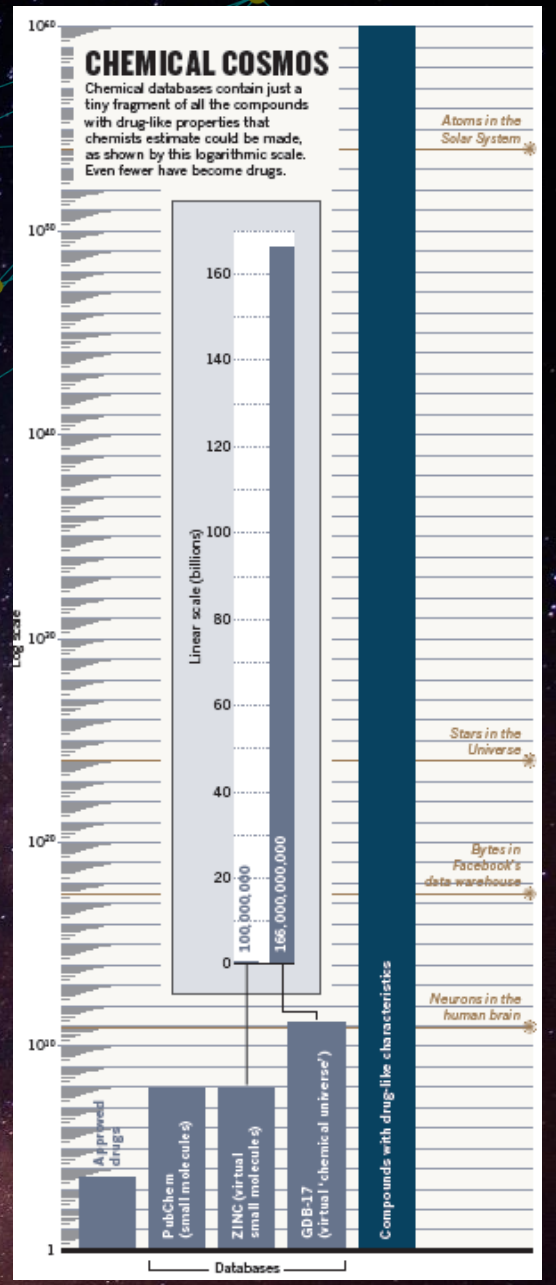
AI帮助我们更好地在可能的成药空间中进行搜索

- 创造虚拟空间
- 设定搜索边界

AI Applied to Chemistry and Drug R&D

The Drug-Maker's Guide to the Galaxy

— Nature, 2017



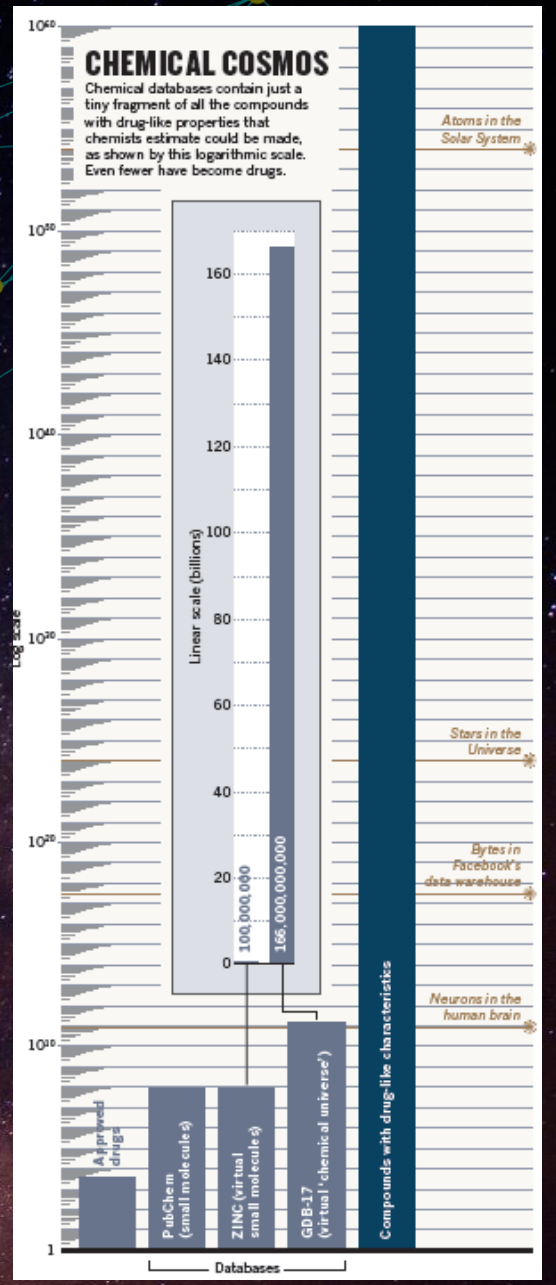
x^n

DIQMTQSPSSLSA
SVGDRVTITCRAS
QGIRNYLAWYQQK
PGKAPKLLIYAAST
LQSGVPSRFSGSG
SGTDFTLTISSLQP
EDVATYYCQRYNR
APYTFGQGTKVEI
KRTVAAPSVFIFPP
SDEQLKSGTASVV
CLLNNFYPREAKV
QWKVDNALQSGN
SQESVTEQDSKDS
TYSLSSTLT
LSKADYEKHKVYA
CEVTHQGLSSPVT
KSFNRGEC

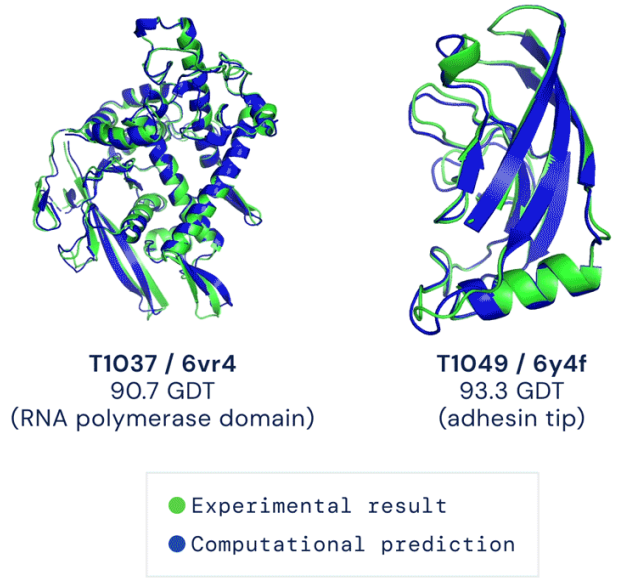
AI Applied to Chemistry and Drug R&D

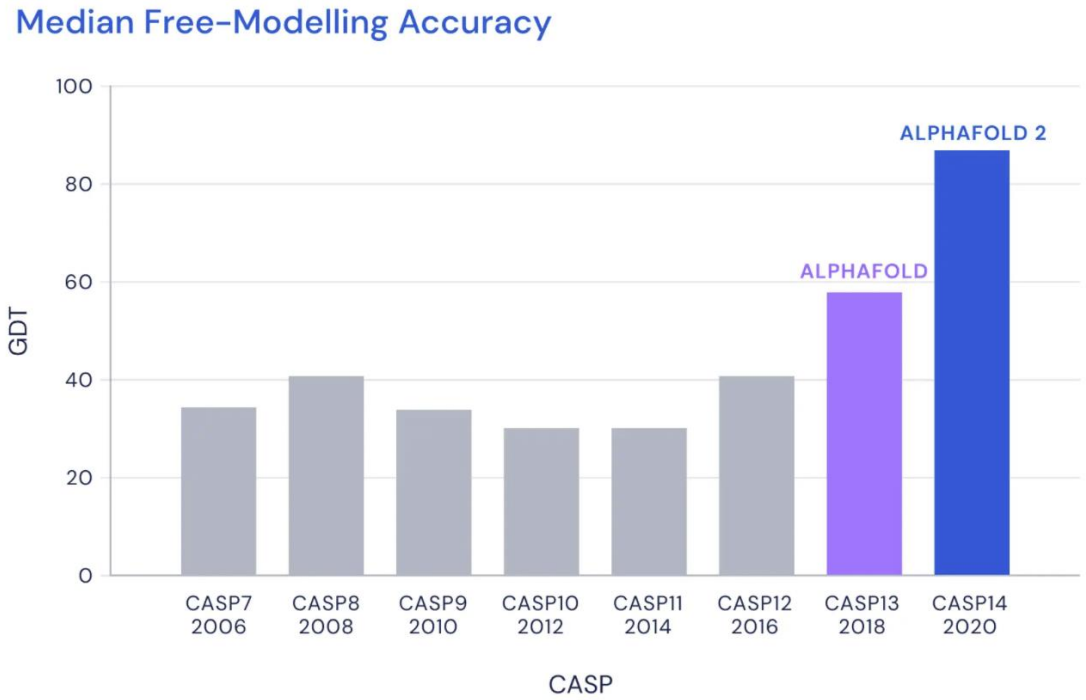
The Drug-Maker's Guide to the Galaxy

— Nature, 2017



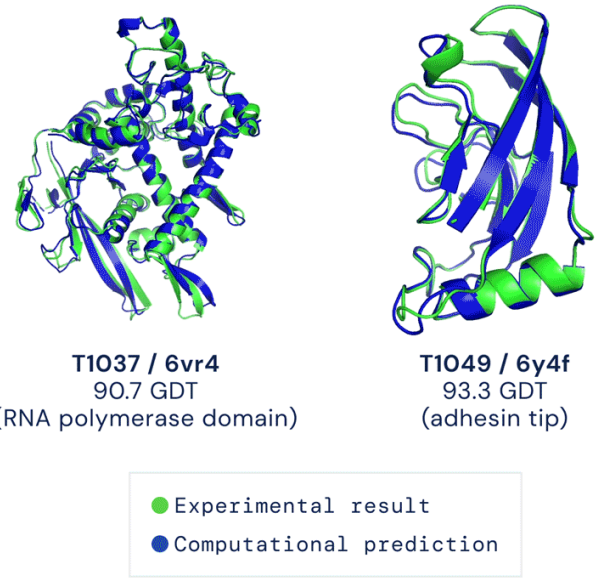
AlphaFold2

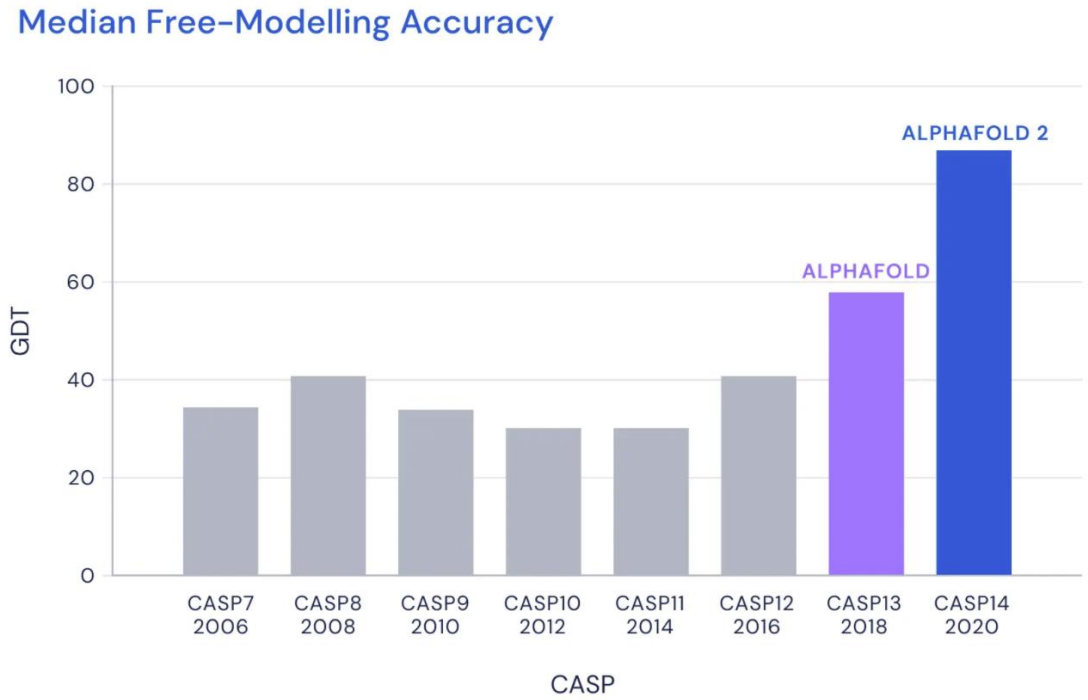




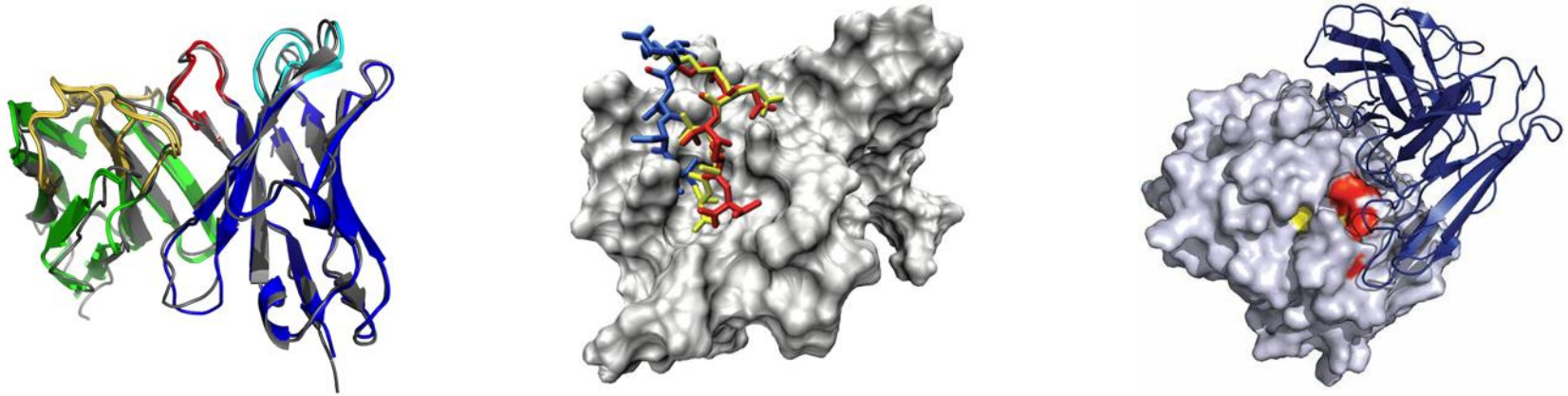
https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

AlphaFold2





https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology



GALILEI LINCEO FILOSOFO E
MATEMATICO DEL SER^MO GRAN
DVCA DI TOSCA
GALILEO