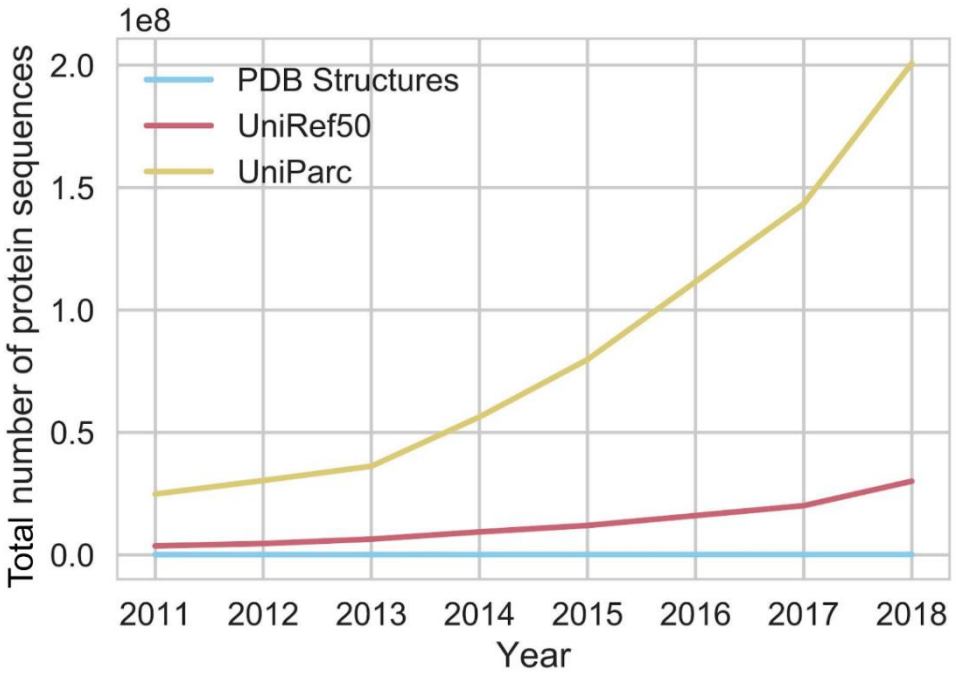
深度学习在蛋白质结构预测中的应用

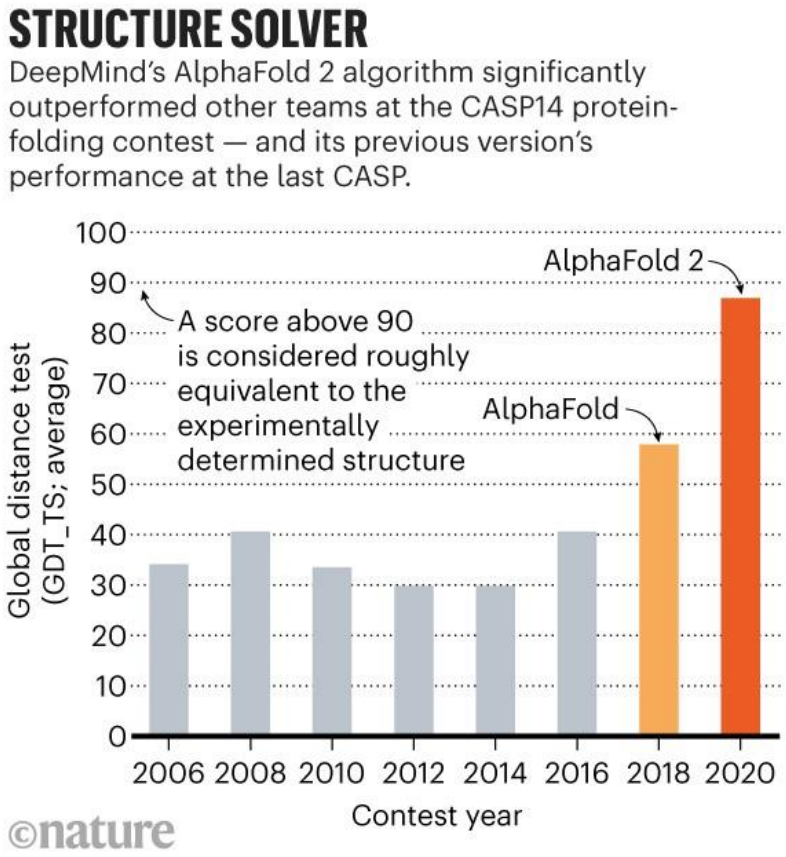
1. 研究背景

根据伯克利人工智能研究所在 2019 年 11 月的一项研究显示，UniParc 蛋白质数据库中约有 300,000,000 个氨基酸序列，PDB 蛋白质数据库中约有 160,000 个蛋白质结构。也就是说，蛋白质序列数据库的数据积累速度非常快，但是已知结构的蛋白质相对很少。



传统的解析蛋白质结构的方法主要有 **X 射线晶体衍射(XRD)、核磁共振(NMR)、冷冻电镜(EM)**。其中近 90%检测得到的蛋白质结构来自于 X 射线晶体衍射。通过实验方法确定蛋白质结构的过程非常复杂，代价较高，因此**实验测定的蛋白质结构比已知的蛋白质序列要少得多**。

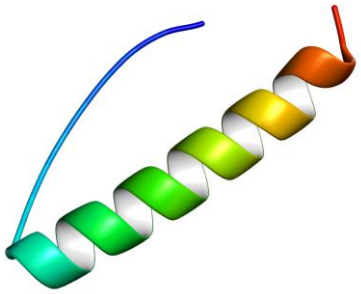
- 1994 年，计算生物学家约翰·莫尔特（JohnMoult）发起全球蛋白质结构预测竞赛(CASP)，让 AI 加入到蛋白质三维结构的研究中。但在此之后的 20 多年中，各个 AI 实验室在这项比赛中始终缺乏实质性突破。
- 2018 年，DeepMind 加入 CASP13 比赛，由 DeepMind 开发的蛋白质三维结构预测算法“AlphaFold”在 98 名参赛队伍中排名第一。
- 2020 年，在 CASP14 中，DeepMind 的“AlphaFold2”再次以绝对的优势排名第一。在受检验的近 100 个目标蛋白中，AlphaFold2 对 2/3 的目标蛋白的氨基酸序列给出的预测结构与实验手段获得的结构相差无几。



2. 数据库

PDB：一个专门收录蛋白质及核酸三维结构资料的数据集，由美国 Brookhaven 国家实验室于 1971 年创建。该数据库通过 X 射线单晶衍射、核磁共振、电子衍射等方法确定蛋白质、多糖、核酸、病毒等生物大分子的三维结构。数据集的信息主要包括生物大分子的原子坐标、参考文献、1 级和 2 级结构信息，晶体结构因数，NMR 实验数据等。

例：**PDB ID 为 1C26 的蛋白质**包含一个有 32 个残基的 α 螺旋链，其三维结构及 PDB 数据库中 1C26 蛋白质坐标部分的前 25 行数据如下图所示：



```
ATOM      1  N   GLY A 325       9.163 -34.215  11.815  1.00 38.89           N
ATOM      2  CA  GLY A 325       9.101 -35.492  12.566  1.00 44.53           C
ATOM      3  C   GLY A 325       8.510 -36.592  11.708  1.00 38.94           C
ATOM      4  O   GLY A 325       9.103 -36.938  10.697  1.00 49.11           O
ATOM      5  N   GLU A 326       7.347 -37.113  12.104  1.00 36.53           N
ATOM      6  CA  GLU A 326       6.643 -38.161  11.376  1.00 34.07           C
ATOM      7  C   GLU A 326       6.197 -37.651  10.000  1.00 28.31           C
ATOM      8  O   GLU A 326       6.167 -36.442   9.746  1.00 20.08           O
ATOM      9  CB  GLU A 326       5.429 -38.649  12.165  1.00 41.40           C
ATOM     10  CG  GLU A 326       5.761 -39.080  13.595  1.00 73.86           C
ATOM     11  CD  GLU A 326       4.586 -39.731  14.353  1.00 84.91           C
ATOM     12  OE1 GLU A 326       3.422 -39.285  14.188  1.00 92.90           O
ATOM     13  OE2 GLU A 326       4.836 -40.687  15.134  1.00 90.37           O
ATOM     14  N   TYR A 327       5.863 -38.593   9.123  1.00 27.69           N
ATOM     15  CA  TYR A 327       5.438 -38.312   7.757  1.00 22.96           C
ATOM     16  C   TYR A 327       3.946 -38.583   7.619  1.00 19.61           C
ATOM     17  O   TYR A 327       3.396 -39.486   8.271  1.00 18.87           O
ATOM     18  CB  TYR A 327       6.254 -39.164   6.771  1.00 24.50           C
ATOM     19  CG  TYR A 327       7.698 -39.040   7.102  1.00 36.69           C
ATOM     20  CD1 TYR A 327       8.211 -39.664   8.253  1.00 39.92           C
ATOM     21  CD2 TYR A 327       8.494 -38.114   6.453  1.00 42.44           C
ATOM     22  CE1 TYR A 327       9.451 -39.348   8.761  1.00 37.84           C
ATOM     23  CE2 TYR A 327       9.751 -37.790   6.958  1.00 45.56           C
ATOM     24  CZ  TYR A 327      10.210 -38.405   8.120  1.00 41.38           C
ATOM     25  OH  TYR A 327      11.393 -38.016   8.686  1.00 51.24           O
```

第 1 列表示类型：原子，第 2 列表示原子序号，第 3 列表示原子名称(例如，N 表示氮原子，CA 表示 Cα 碳原子,OE1 表示名为 OE1 的氧原子，OXT 表示氨基酸序列中最后 1 个氧原子)，第 4 列表示原子所在氨基酸残基的名称，第 5 列表示原子所属的链，第 6 列表示氨基酸残基的序列号，第 7-9 列分别表示原子的 x,y,z 轴坐标，第 10 列表示原子占有率，第 11 列表示温度因子，第 12 列表示原子的元素符号。

UniProt：一个包含蛋白质序列、功能信息、研究论文索引的蛋白质数据库， 也是目前为止收录蛋白质序列目录最广泛、功能注释最全面的一个数据库。目前，UniProt 由主要由以下子库构成：

数据库名	全名	用途
UniProtKB/Swiss-Prot	Protein knowledgebas (review)	高质量的、手工注释的、非冗余的数据库
UniProtKB/TrEMBL	Protein knowledgebase (unreview)	自动翻译蛋白质序列，预测序列，未验证的数据库
UniParc	Sequence	非冗余蛋白质序列数据库
UniRef	Sequence clusters	聚类序列减小数据库，加快搜索的速度
Proteomes	Protein sets from fully sequenced genomes	为全测序基因组物种提供蛋白质组信息

UniParc：一个综合性的非冗余蛋白质数据库，**包含了所有主要的、公开的数据库的蛋白质序列**。该数据库只含有蛋白质的序列信息，而没有注释数据。 由于一种蛋白质可能在不同的数据库中存在，且可能在同一数据库中有多个版本，为了去冗余，UniParc 对每条唯一的序列只存一次。氨基酸序列只要相同，就被合并为一条，每条序列提供稳定的、唯一的编号 UPI。

BFD：一个蛋白质序列比对数据集，包含 65983866 个蛋白质家族的 MSA 信息和 2204359010 个蛋白质序列的 HMM(隐藏马尔可夫模型)信息。

Uniclust30：包含 1.3 亿条氨基酸序列，使用 hhsuite 中的 HHblits 工具进行快速 MSA 检索，总计 13 个文件 86G 数据量。

Uniref90：包含 1.3 亿条氨基酸序列，是 MSA 检索所需要的库，使用 JackerHammer 工具进行检索，总计 1 个文件 58G 数据量。

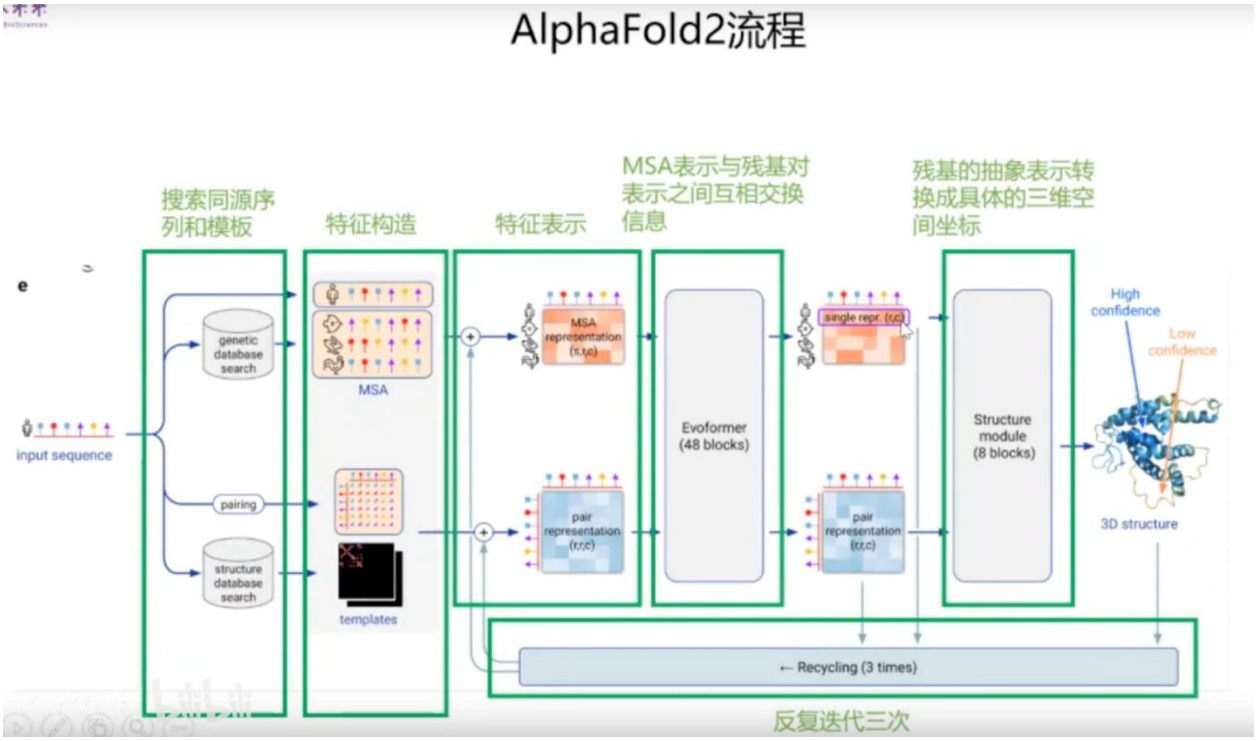
3. 近期工作

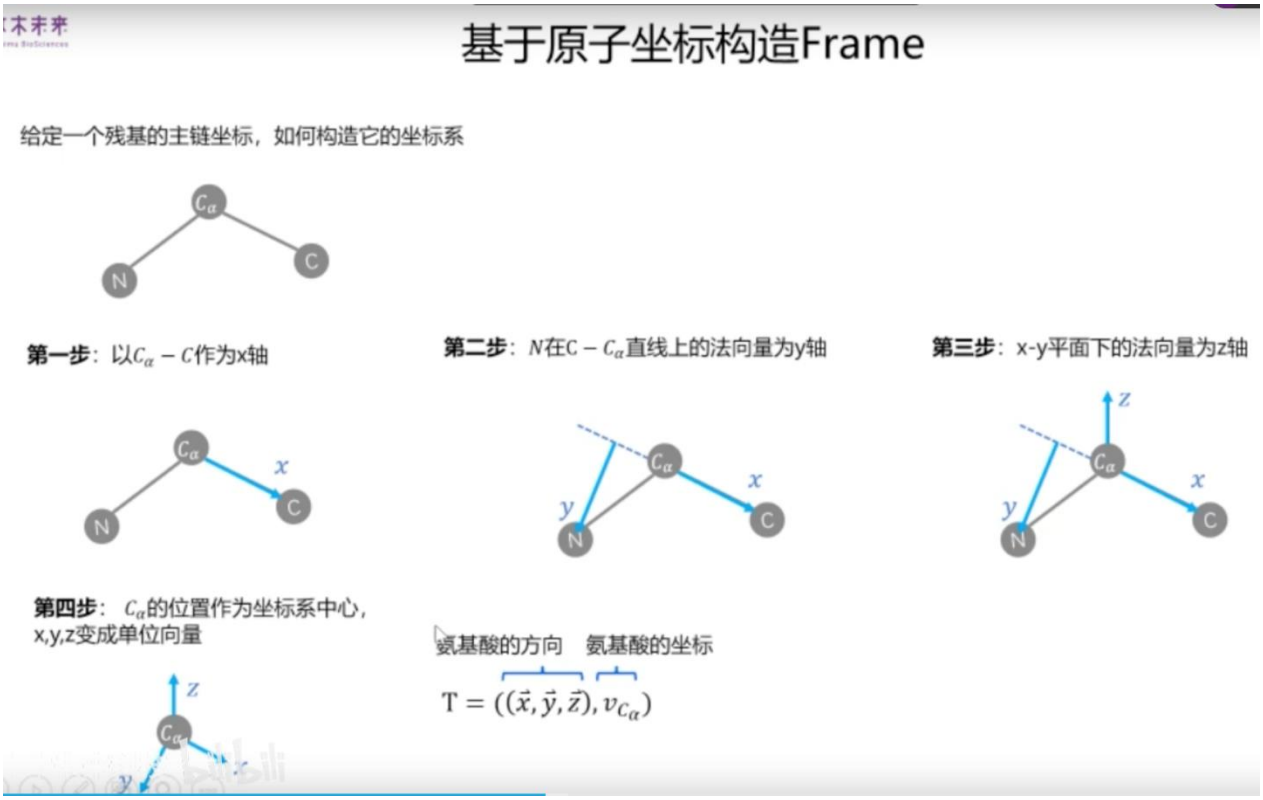
1. *Highly accurate protein structure prediction with AlphaFold. (2021)*

目标：根据氨基酸序列，用端到端的方法预测蛋白质的三维结构。
数据：
搜索同源模板序列：PDB
构造多序列比对(MSA)特征：Uniref90(58GB)，MGnify(64GB)，Uniclust30(86GB)，BFD(1.7TB)
训练模型：PDB

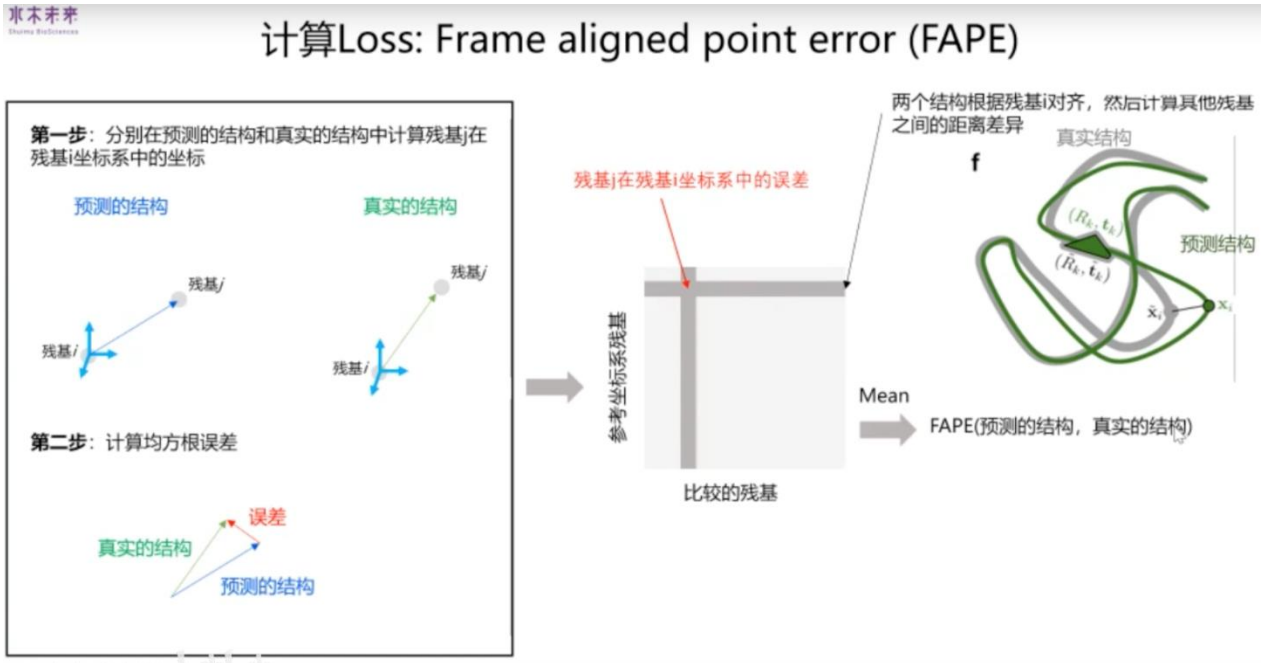
算法的输入和输出：
输入：氨基酸序列
输出：氨基酸序列中每一个原子的空间坐标，每一个残基的 pLDDT 指标(范围是 0-100，得分越高预测结果越好)，残基对距离误差、TM-得分等

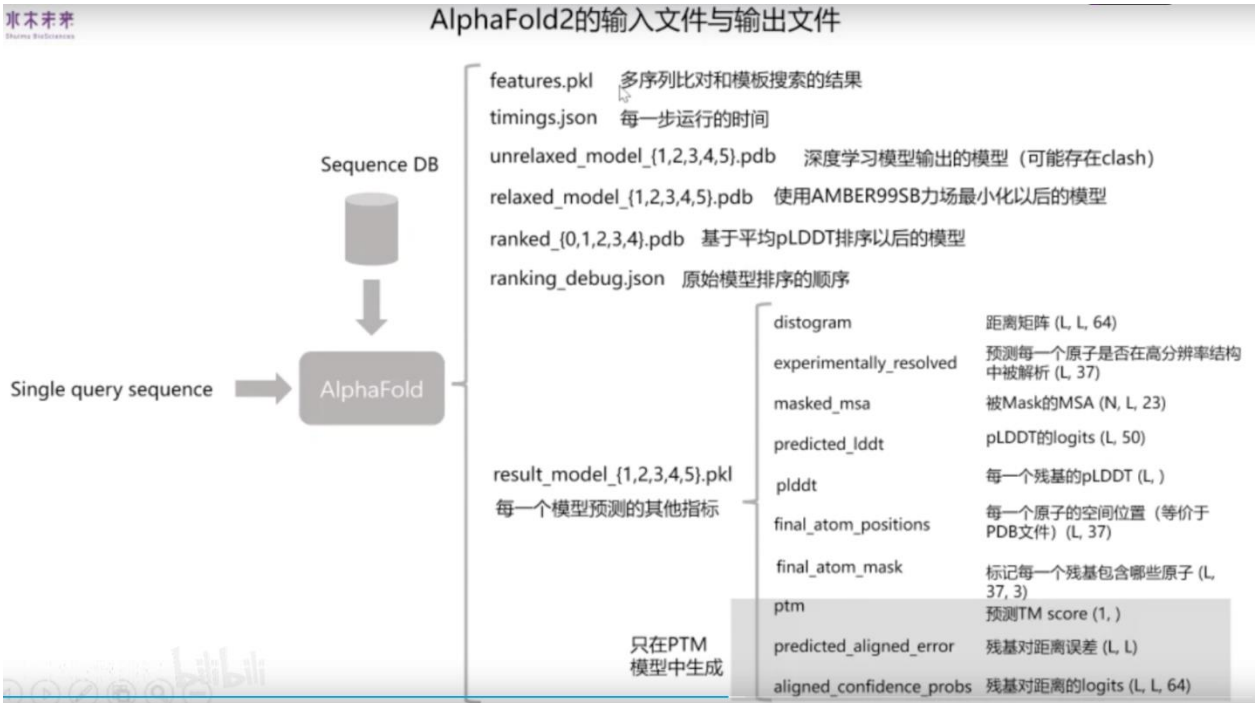
方法：
用深度学习和多序列比对(MSA)来预测蛋白质结构。应用同源建模法，原理是**相似的氨基酸序列对应着相似的蛋白质结构**。输入序列，在数据库种找到与输入序列同源的已知结构的氨基酸序列作为模板序列;对输入序列与模板序列进行序列比对，生成 MSA 特征；再根据生成的 MSA 特征，用同源建模软件预测氨基酸序列对应的蛋白质结构类型。该方法适用于输入序列与模板序列之间的一致度超过 30%的氨基酸序列。
参考链接：https://www.deepmind.com/blog/AlphaFold-a-solution-to-a-50-year-old-grand-challenge-in-biology





给定一个氨基酸残基，若已知其中 C_α,C,N 原子的空间位置，则以 C_α为坐标原点，C_α指向 C 的方向为 X 轴正方向；过 N 原子所在点作 X 轴的垂线为 Y 轴方向，且 y 轴正向指向 N 原子；以 X-Y 平面的法向量为 Z 轴，建立三维局部坐标系。





AlphaFold2 缺点：

(1)AlphaFold2 使用深度学习和编码在多序列比对(MSA)来预测蛋白质结构，对2/3 的蛋白预测达到到实验精度，还有 1/3 并未做到。

(2)AlphaFold2 对于找不到(显著)同源关系的蛋白质和快速进化的蛋白质，不能为其生成多序列比对(MSA)。

(3)模型输入仅限制于单链,不能输入多链。

(4)输出结果适用于三维晶体结构，对于无法结晶的蛋白质，不一定适合这种方法。

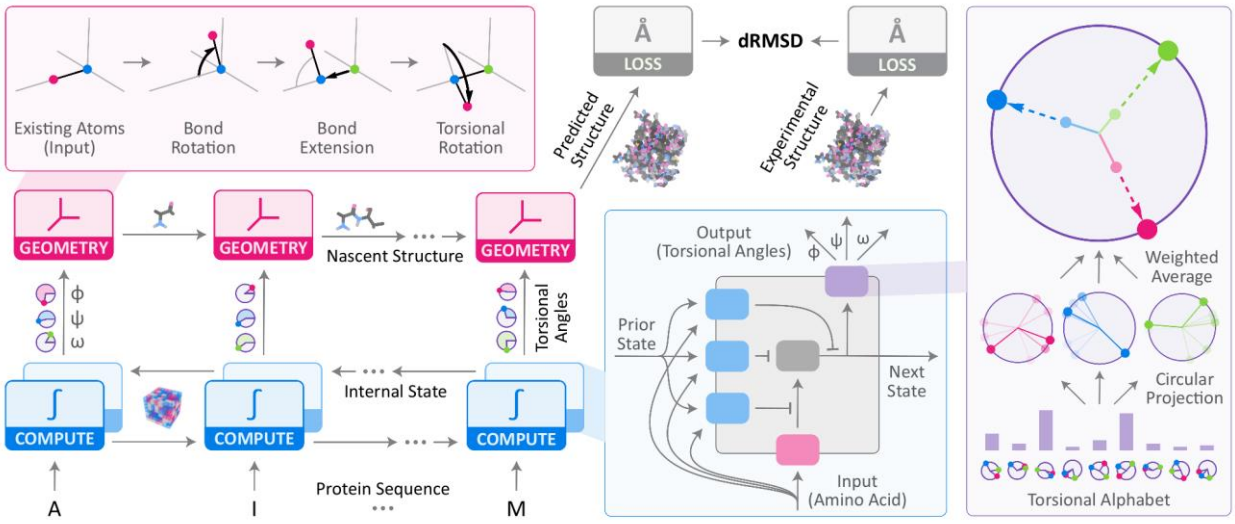
2.End-to-End Differentiable Learning of Protein Structure(2019)

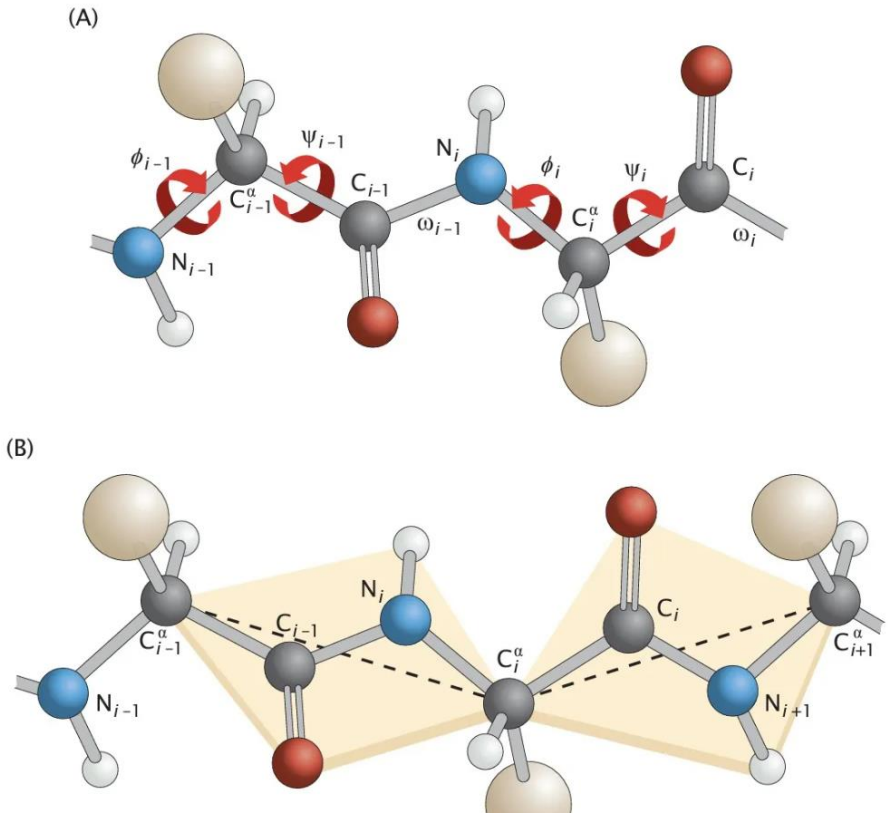
目标：根据氨基酸序列预测蛋白质的三维结构。
数据：
ProteinNet 7 - 12 数据集、CATH 结构分类数据库

RGN 算法的输入和输出：
输入：氨基酸序列
输出：扭转角、氨基酸序列的每个原子的坐标、dRMSD 得分(基于距离的均方误差得分)

RGN 方法：
(1)输入氨基酸序列，对于每一个残基，利用位置特异性评分矩阵（PSSM），应用 LSTM（长短期记忆),预测主链中各个键的 3 个扭转角。
(2)进入几何模块，将扭转角转化为三维笛卡尔坐标。
(3)计算预测的三维结构和实验测得的三维结构的差距(dRMSD)

RGN(Recurrent Geometric Networks)网络框架





肽平面示意图。由于 N-Cα 键和 Cα-C 键可以自由旋转，一个肽平面可相对于相邻肽平面进行旋转。绕 N-Cα 键的旋转角被称为**扭转角 φ**，绕 Cα-C 的旋转角被称为**扭转角 ψ**，绕 C-N 肽键的旋转角则被称为**扭转角 ω**。

3. *Single-sequence protein structure prediction using language models from deep learning.(2022)*

目标：根据氨基酸序列预测蛋白质的三维结构。

数据：
训练 Aminobert 语言模型： UniParc 序列数据库(包含约 2.5 亿个天然蛋白质序列)
已知结构的孤儿蛋白质序列： 77 个，取自 UniRef30,PDB70,MGnify
人工合成的蛋白质序列： 149 个，计算机软件合成
训练 RGN2：ProteinNet12、一个源于 ASTRAL SCOPe 的小数据集

RGN2 算法的输入和输出：
输入：氨基酸序列
输出：键角、扭转角、dRMSD 得分(基于距离的均方误差得分)、GDT_TS (总体距离测试总得分)等

RGN2：提出了端到端的循环几何网络计算模型 RGN2，使用 AminoBERT 蛋白质语

言模型预测没有显著同源序列的蛋白质(如孤儿蛋白)的结构，并预测人工合成的蛋白质的结构。评估了新模型 RGN2 在 77 个孤儿蛋白和 149 个人工合成的蛋白质的测试集上的准确性，计算时间更短、性能更好。

(1)从 UniParc 序列数据库中输入蛋白质氨基酸序列，用 12 层的 Transformer 重建氨基酸序列，训练 AminoBERT 蛋白质语言模型；

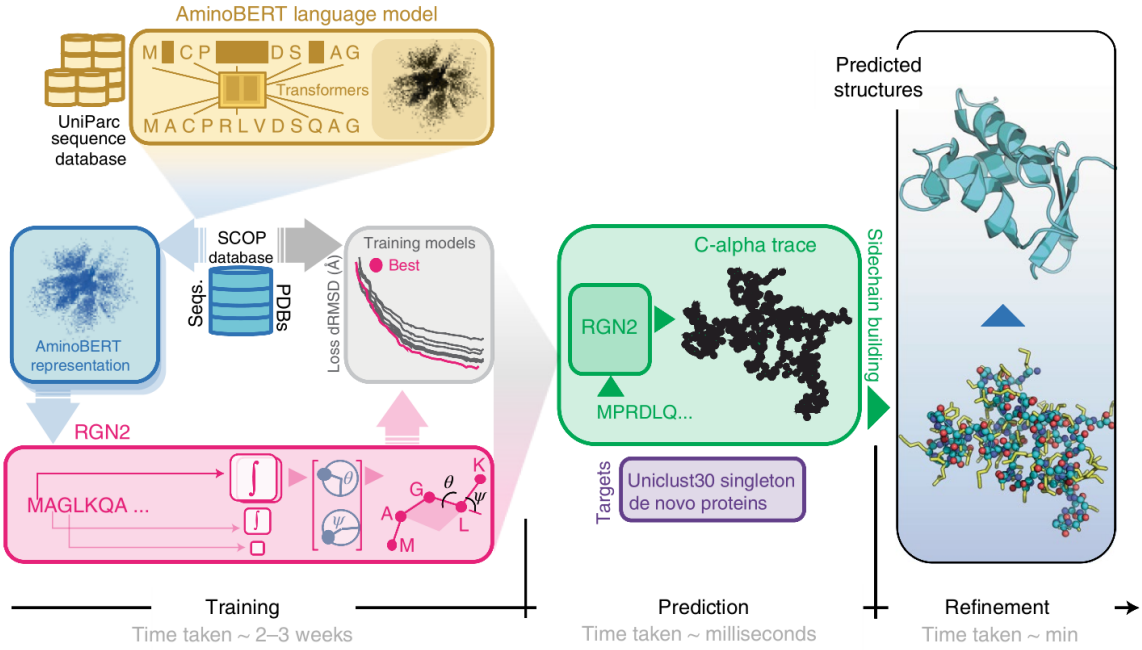
(2) 从 PDB 序列数据库中输入氨基酸序列，重建氨基酸序列，关于每个残基用 Frenet-Serret 公式在 C_α 原子处建立新的**切向量-副法向量-法向量三维坐标系**，通过转换矩阵来得到整个序列的骨架结构，计算预测的结构和实验得到的结构的 dRMSD 得分，以训练 RGN2 模型；

(3)构建侧链和氢键空间网络，获得输入蛋白质序列的空间结构；

(4)使用 AF2Rank 对空间结构进行细化。

AminoBERT 训练过程

AminoBERT 是一个 **12 层的 Transformer**，每一层包含 12 个注意头。它从 UniParc 序列数据库中获得的约 2.6 亿个天然蛋白质序列中提取蛋白质序列。输入蛋白质氨基酸序列，对序列中 2-8 个相邻的残基做掩码处理，用 AminoBERT 蛋白质语言模型重建氨基酸序列。**Batchsize = 3072, epochs = 13,** 每一层均用 0.1 的概率进行 Dropout，使用 GELU 激活函数，在 512-core TPU pod 上进行了大约 1 周的训练。



模型效果：

(1)对于选取的孤儿蛋白的结构预测准确率，RGN2 模型优于 AF2(AlphaFold2)；

(2)对于具有单螺旋及弯曲结构的蛋白质，RGN2 模型优于 AF2,RF(RoseTTAFold)；

(3)在较长螺旋蛋白质上的预测表现上，RGN2 模型优于 AF2；

模型创新点：

蛋白质几何结构基于 Frenet－Serret 公式确定，保持了蛋白质的平移和旋转不变性。

Frenet－Serret 公式(弗莱纳公式)用来描述欧几里得空间 R 中的粒子在连续可微曲线上的运动，给出了曲线的单位切向量 T，单位法向量 N，单位副法向量 B 之间的关系。弗莱纳坐标系具体定义：
T ：单位切向量，方向指向粒子运动的方向；
N ：单位法向量 T 对弧长参数的微分单位化得到的向量；
B ： T 和 N 的外积。

给定一个氨基酸序列以及整体三维坐标系，得到每个 C_α 原子的空间坐标向量，在C_α-C_α上定义单位切向量，副法向量、法向量：

$$t_i = \frac{r_{i+1} - r_i}{|r_{i+1} - r_i|}$$

$$b_i = \frac{t_{i-1} \times t_i}{|t_{i-1} \times t_i|}$$

$$n_i = b_i \times t_i$$

定义旋转矩阵

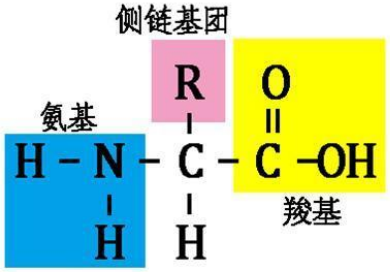
$$\begin{pmatrix} n_{i+1} \\ b_{i+1} \\ t_{i+1} \end{pmatrix} = \mathcal{R}_{i+1,i} \begin{pmatrix} n_i \\ b_i \\ t_i \end{pmatrix}$$

固定C_α-C_α键长为 3.8 埃，将蛋白质股价的第一个C_α原子应为在坐标原点，重建C_α原子坐标：

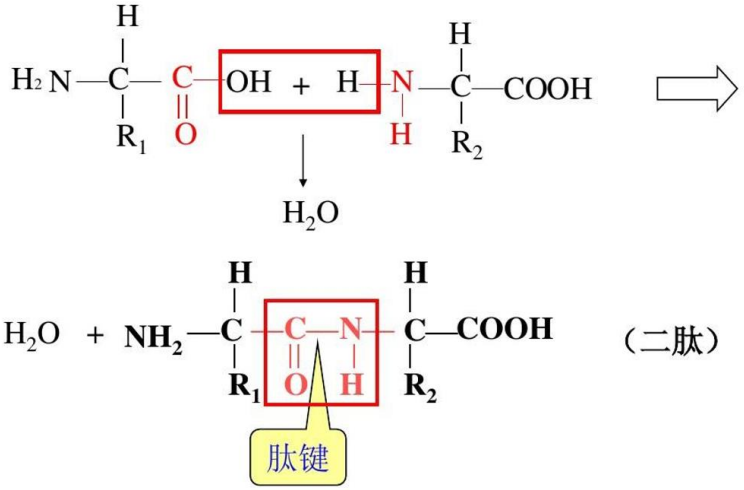
$$\begin{pmatrix} n_{i+1} \\ b_{i+1} \\ t_{i+1} \\ r_{i+1} \end{pmatrix} = \begin{pmatrix} & & & 0 \\ & \mathcal{R}_{i+1,i} & & 0 \\ & & & 0 \\ 0 & 0 & 3.8 & 1 \end{pmatrix} \begin{pmatrix} n_i \\ b_i \\ t_i \\ r_i \end{pmatrix}$$

4. 基础知识

氨基酸：称含有氨基和羧基的有机化合物为氨基酸。氨基连在α-碳上的氨基酸为 α-氨基酸，组成蛋白质的氨基酸大部分为 α-氨基酸。



氨基酸脱水缩合形成多肽：一个氨基酸分子的羟基-COOH 和另一个氨基酸分子的氨基-NH2 相连接，同时脱去一分子的水。称由多(二)个氨基酸分子脱水缩合而成的含有肽键的化合物为多(二)肽。



氨基酸残基：指不完整的氨基酸。氨基酸脱水缩合形成多肽时，部分基团失去一分子水，称由肽键连接的氨基酸失水后的剩余部分为氨基酸残基。

氨基酸序列：氨基酸相互连接形成多肽的顺序，即蛋白质的一级结构。

蛋白质：多肽通常呈链状结构，叫做肽链。肽链盘曲、折叠（有的蛋白质分子由几条肽链通过某些化学键相连），形成具有一定空间结构的蛋白质分子。

蛋白质的四级结构：

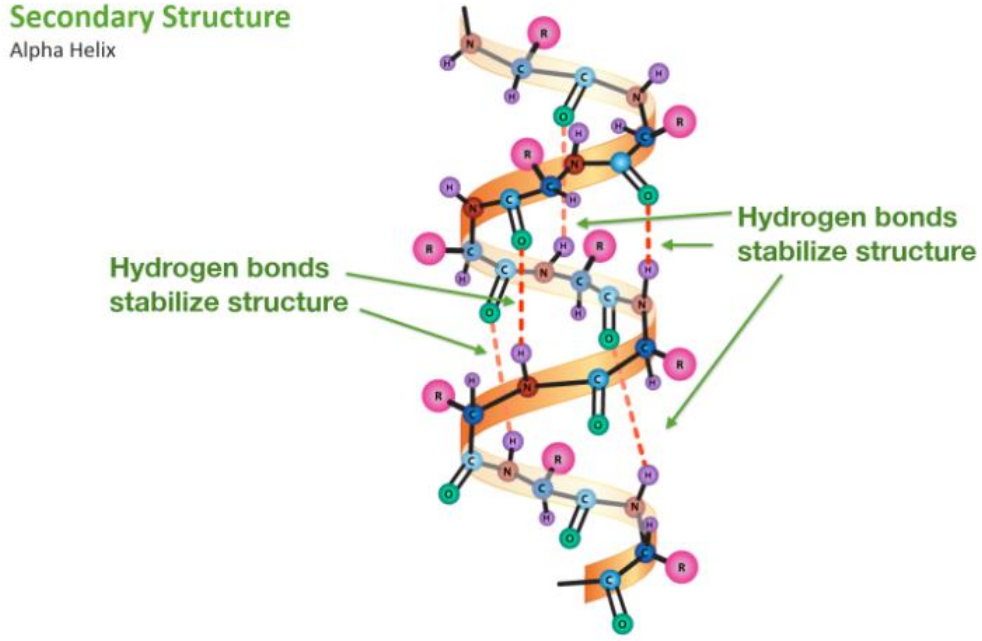
一级结构：独特的氨基酸序列，由遗传物质决定。

例：胰岛素 A 链的一级结构：

Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-

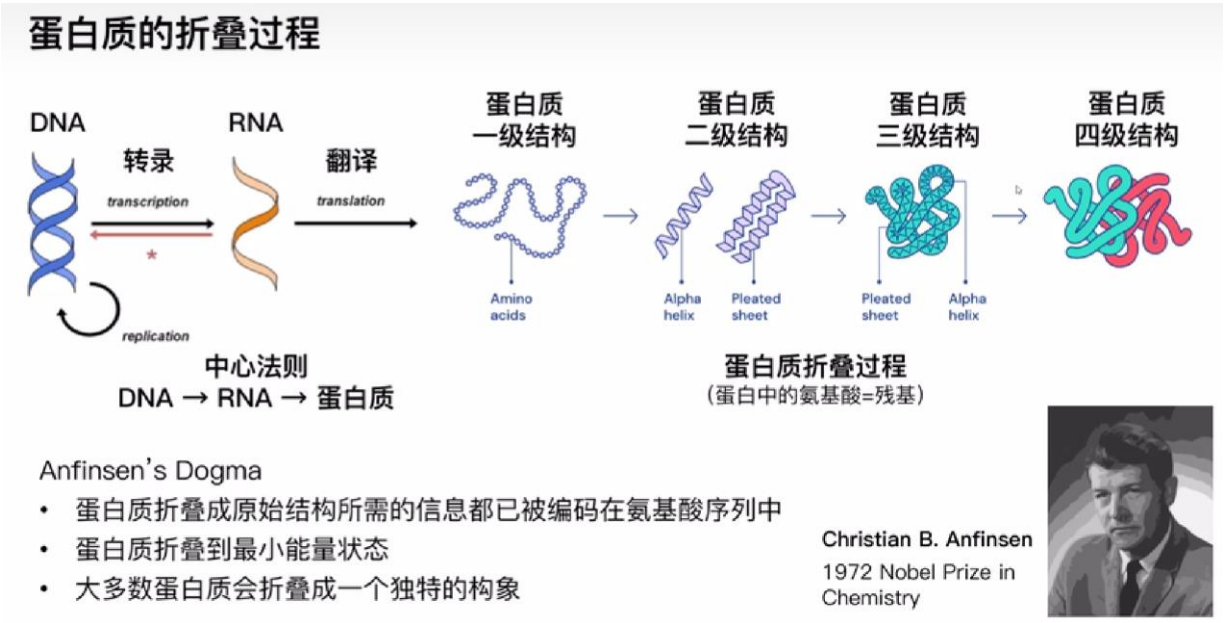
Asn-Tyr-Cys-Asn

二级结构：多肽链的主链骨架（不包括 R 基）在空间上有规律的折叠和盘绕形成的特定的构象。二级结构的主要形式包括 α-螺旋、β-折叠、β-转角、Ω 环和无规卷曲。



三级结构：多肽链在二级结构的基础上，进一步盘绕、卷曲和折叠，形成三维结构。

四级结构：具有两条和两条以上多肽链的寡聚蛋白质或多聚蛋白质才会有四级结构。四级结构包括亚基的种类、数目、空间排布以及亚基之间的相互作用。



20 种常见氨基酸：

中文名称	英文名称	三字母缩写
甘氨酸	Glycine	Gly
丙氨酸	Alanine	Ala
缬氨酸	Valine	Val
亮氨酸	Leucine	Leu
异亮氨酸	Isoleucine	Ile
脯氨酸	Proline	Pro
苯丙氨酸	Phenylalanine	Phe
酪氨酸	Tyrosine	Tyr
色氨酸	Tryptophan	Trp
丝氨酸	Serine	Ser
苏氨酸	Threonine	Thr
半胱氨酸	Cystine	Cys
蛋氨酸	Methionine	Met
天冬酰胺	Asparagine	Asn
谷氨酰胺	Glutarnine	Gln
天冬氨酸	Asparticacid	Asp
谷氨酸	Glutamicacid	Glu
赖氨酸	Lysine	Lys
精氨酸	Arginine	Arg
组氨酸	Histidine	His

孤儿基因：孤儿基因指找不到同源关系，也没有明确的演化历史的基因。生物体内大部分基因来自于几个大基因家族，家族中相似的基因拥有共同的祖先，可以追溯到千百万年前，然后不断变异演化，形成庞大的基因家族。人体内的基因片断中，有超过三分之一的部分既找不到同源基因，也没有发现它们的演化史——看上去既没有“父母”，也没有任何“亲属”，就像是不知道从哪里冒出来的“孤儿”。孤儿基因可以代表真核生物基因组编码基因的 10–20%。他们的功能仍然鲜为人知。有证据表明，绝大多数是转录和编码蛋白质，但它们的功能相关性仍有待确定。

孤儿蛋白：孤儿蛋白指找不到(显著)同源关系的蛋白质。

参考文献：

[1]Chowdhury, R. , Bouatta, N. , Biswas, S. , Rochereau, C. , Church, G. M. , & Sorger, P. K. , et al. (2021). Single-sequence protein structure prediction using language models from deep learning. Cold Spring Harbor Laboratory.

[2]AlQuraishi, M. End-to-end differentiable learning of protein structure. Cell Syst. 8, 292－301 (2019).

[3]Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature596, 583－589 (2021).

[4] AlphaFold2 架构视频解读
https://www.bilibili.com/video/BV1444y1h7Tf/?vd_source=6b561fa986b1dd6838229bc0e765d550

[5]https://zhuanlan.zhihu.com/p/574229937

[6]https://www.nature.com/articles/s41586-021-03819-2

[7]Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706－710 (2020).

[8]AlQuraishi, M. End-to-end differentiable learning of protein structure. Cell Syst. 8, 292－301 (2019).

[9]https://www.tsu.tw/edu/11609.html

[10]https://www.nature.com/articles/s41587-022-01432-w

[11]使用语言模型和深度学习的单序列蛋白质结构预测
https://zhuanlan.zhihu.com/p/574229937

[12]1c26-p53 四聚功能域的晶体结构-日本蛋白质结构数据库
https://pdbj.org/mine/structural_details/1c26?lang=zh-CN

[13]3D PFV: 1C26 https://www.rcsb.org/3d-sequence/1C26?assemblyId=1

[14]https://blog.csdn.net/weixin_30501857/article/details/96485481

[15]http://www.bioengx.com/pdb-file/

[16] “蛋白质结构预测”问题描述
http://bitjoy.net/2019/05/25/%E8%9B%8B%E7%99%BD%E8%B4%A8%E7%BB%93%E6%9E%84%E9%A2%84%E6%B5%8B%E9%97%AE%E9%A2%98%E6%8F%8F%E8%BF%B0/