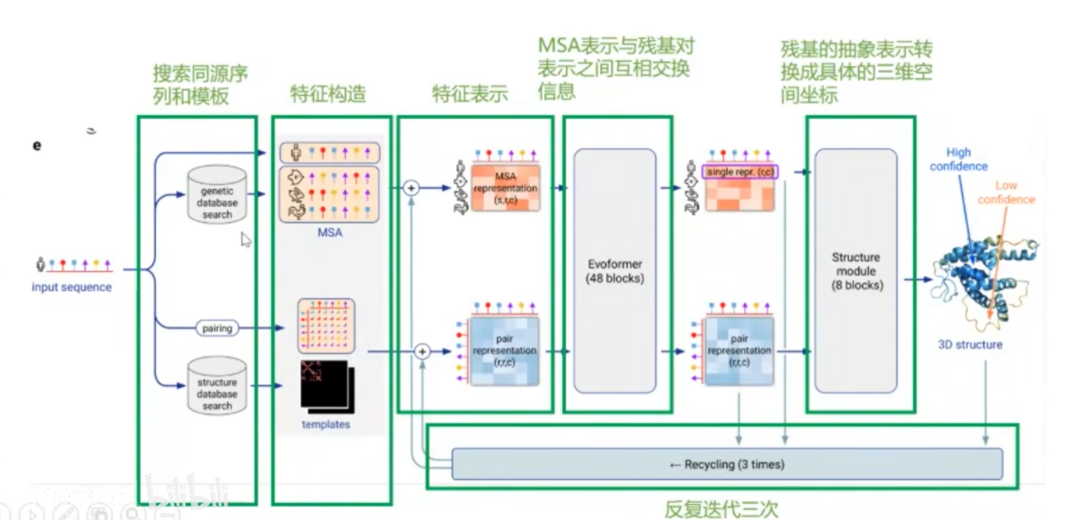
**Alphafold学习笔记**

1. **简介**

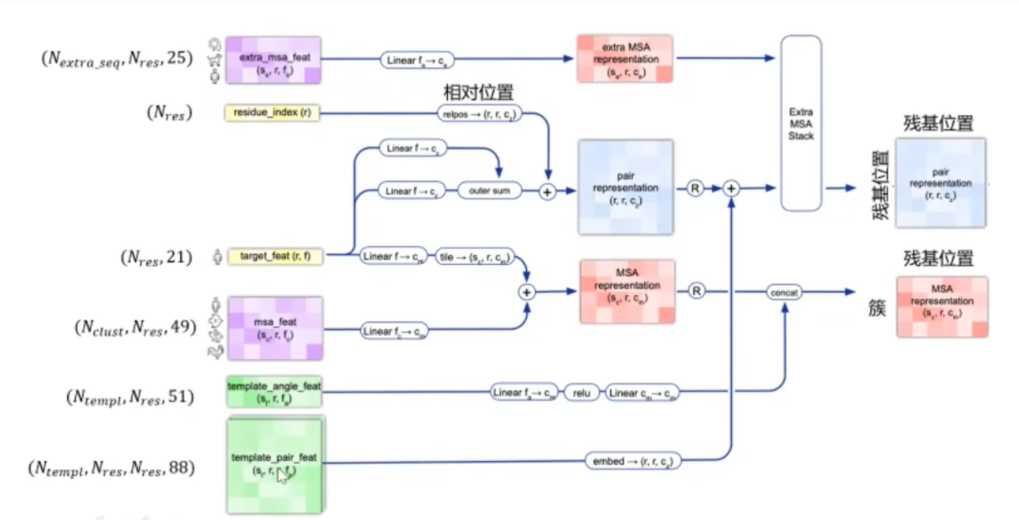
Alphafold是一款DeepMind开源的用于预测蛋白质、DNA、RNA、配体、离子等生物分子结构的工具。此外，最新的Alphafold3（[alphafoldserver.com](https://alphafoldserver.com/)）还可以对蛋白质和核酸的化学修饰进行建模。

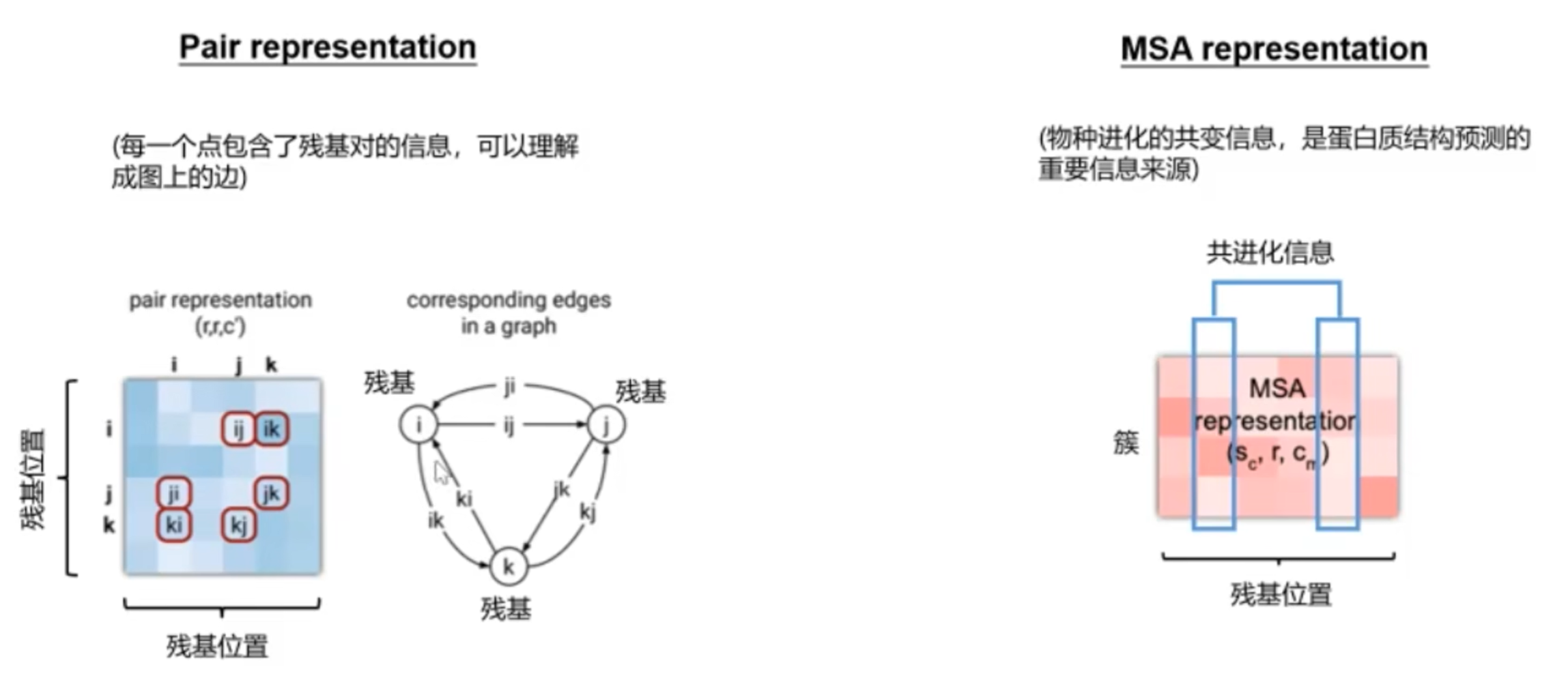
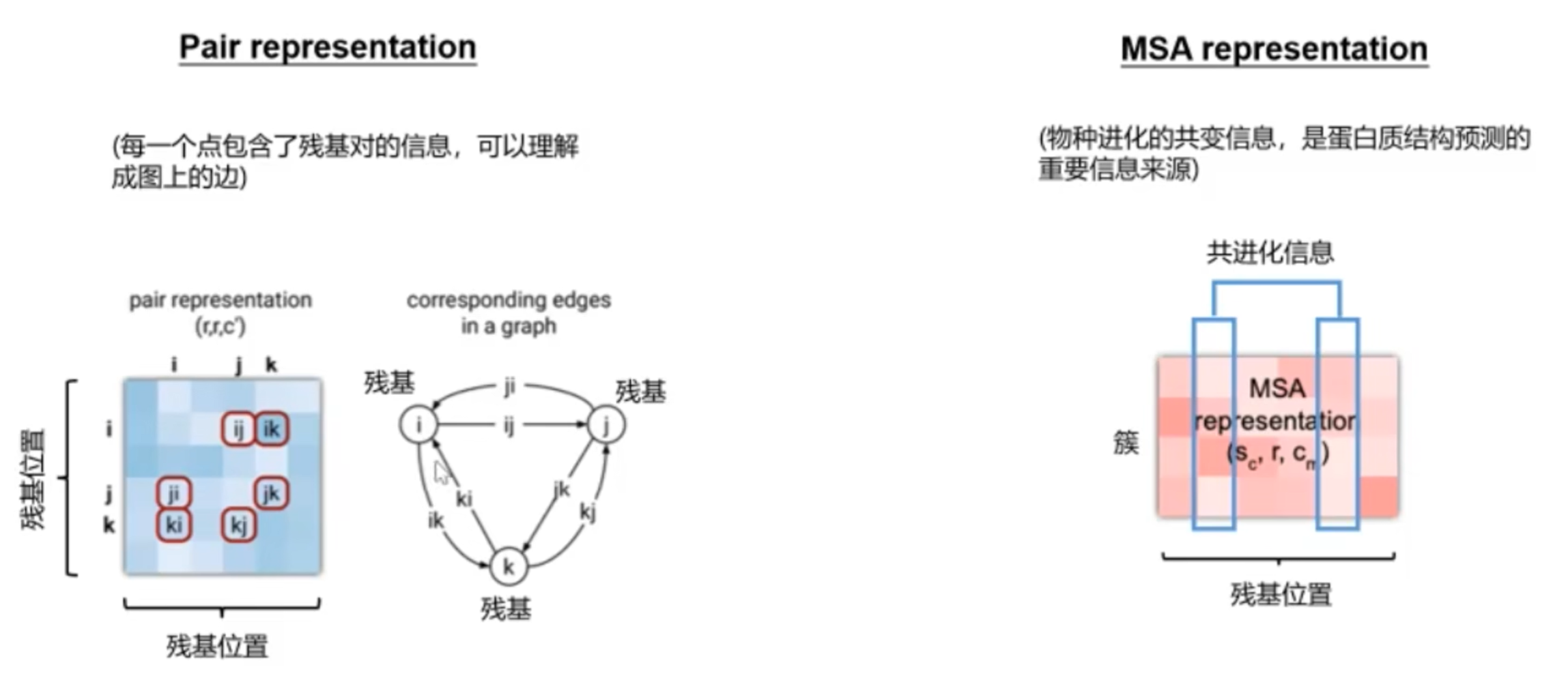
1. **算法原理**

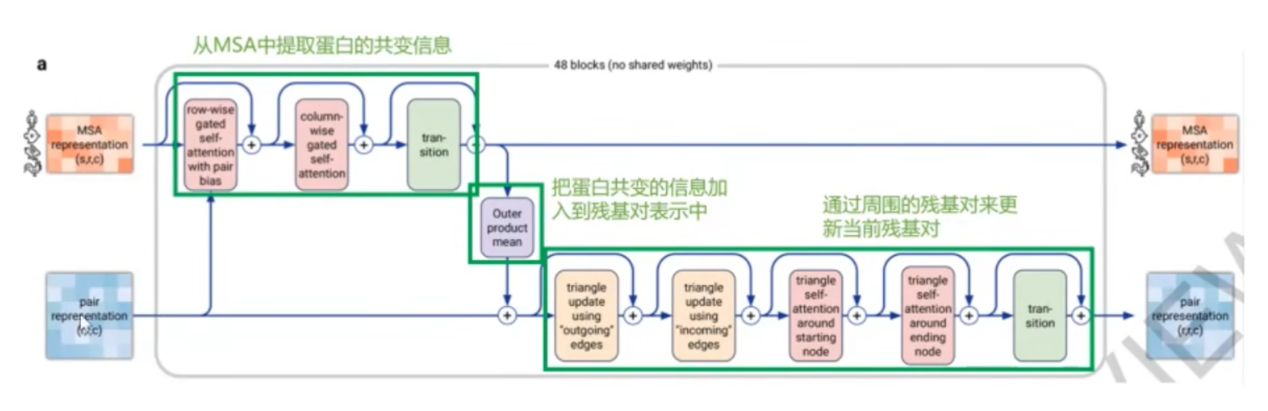
Alphafold2中要求使用的输入文件为目标氨基酸序列、MSA、模板，直接end to end的预测目标的三维结构，使用了Transformer进行预训练。通过从MSA中获得蛋白质的共进化信息来预测蛋白质的三级结构（蛋白质结构预测算法有PSSM（统计每一位点残基所占比例）、co-evolution（基于MSA提取共变矩阵）和Full MSA三种）。









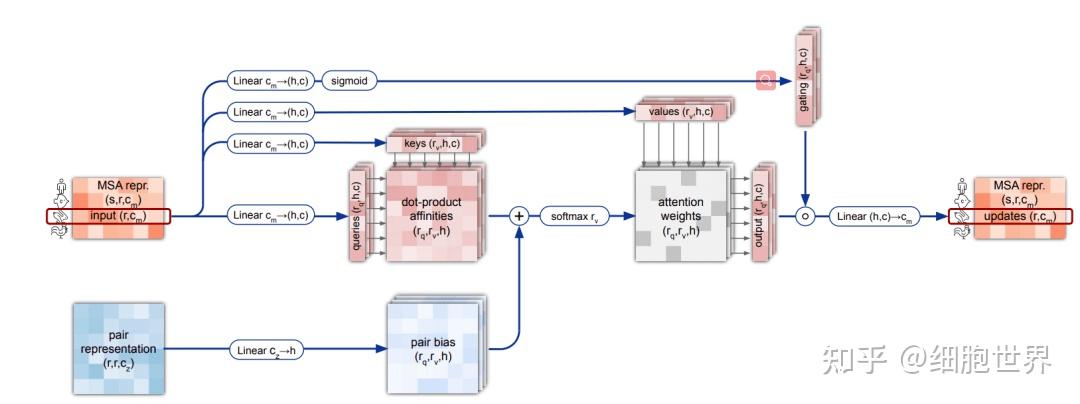
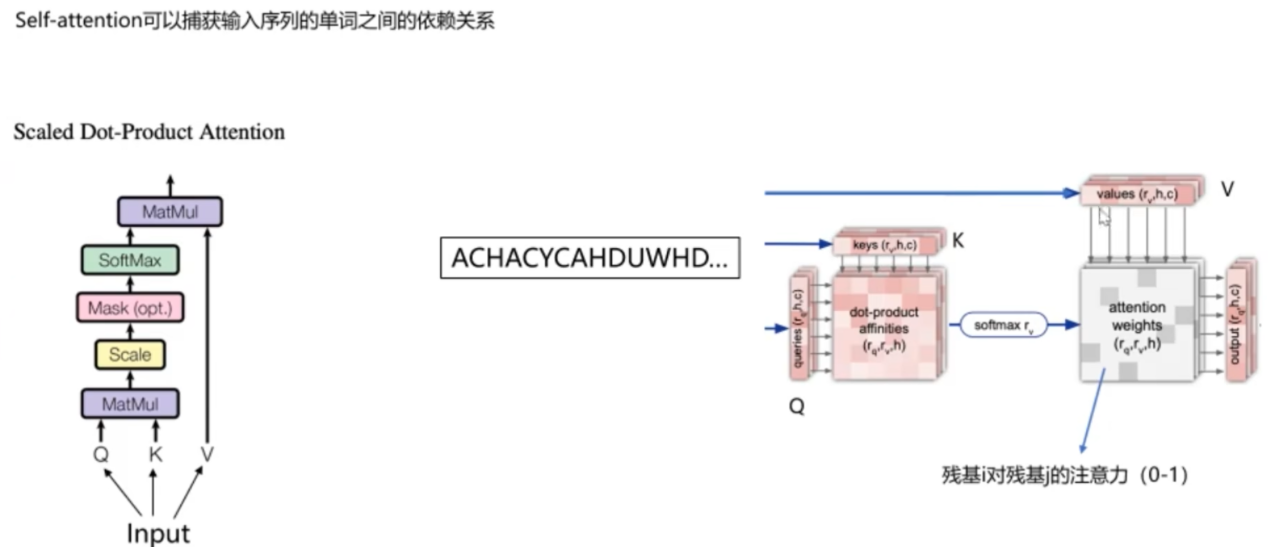


将物种进化的共变信息融入到单一物种是通过Evoformer结构实现的。

1. **背景**

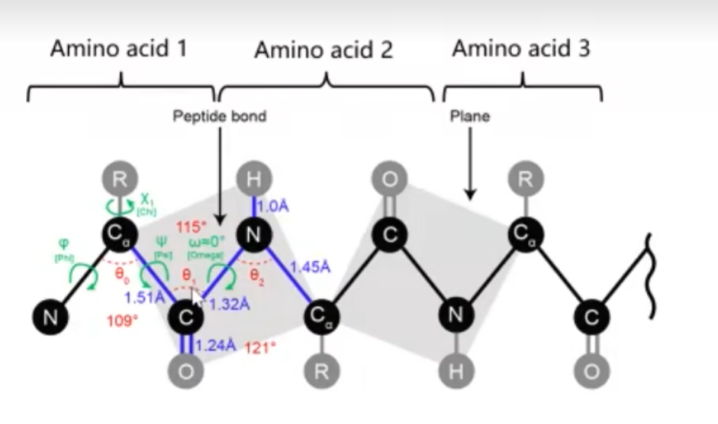
Transformer模型是一种基于自注意力机制（self-attention）的神经网络架构，其在包括机器翻译任务在内的自然语言处理（NLP）领域取得了显著的成果。ransformer模型完全依赖于注意力机制来捕捉输入和输出之间的全局依赖关系，从而避免了循环神经网络（RNN）中的重复计算问题。而Evoformer即进化former，进化魔改Transformer，用来计算MSA和pairwise features。

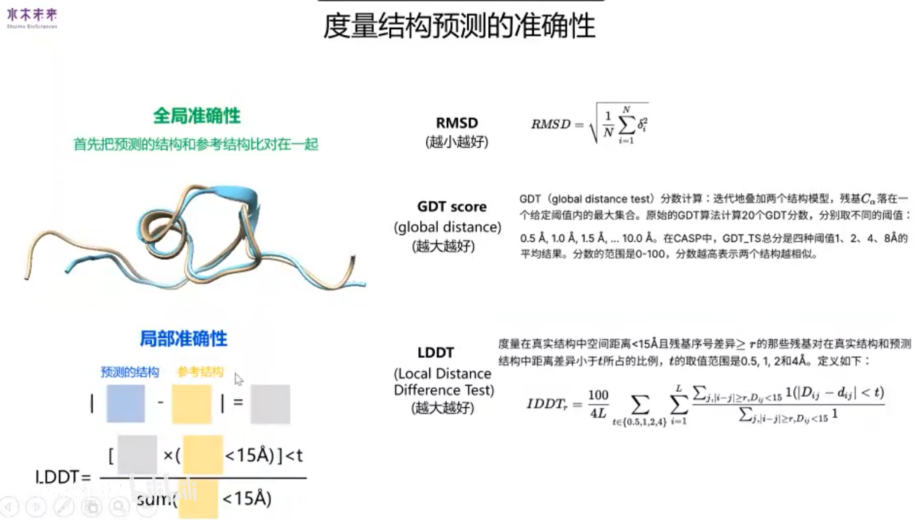
独热编码（One-Hot Encoding）:使用N位状态寄存器对N个状态进行编码，并且任何时刻只有一位是有效的（即设置为1）。该编码是针对具有明确分类值的数据进行预处理的有效方法。



残基-残基对的特征（是每个残基之间都有一个隐状态来描述他俩之间的关系，我们把每对儿残基之间的隐状态叫做pairwise features）。

backbone frames指的是构成蛋白质主链的框架结构，它们由蛋白质序列中每个氨基酸残基的N、Cα和C原子构成的三角形表示（一般在蛋白质结构中键长和键角时固定的，比如N、Cα和C原子形成的夹角一般固定在109°左右，键长在1.51Å左右）。这些三角形在空间中通过旋转和平移来确定蛋白质主链的三维结构。一般来说，Cα、C、O、N、H和Cα六个原子在一个平面内。





1. **提交的任务要求**
   1. 每天最多运行20个作业
   2. 每个作业的总大小受到结构中“token”的限制（最大为5000）

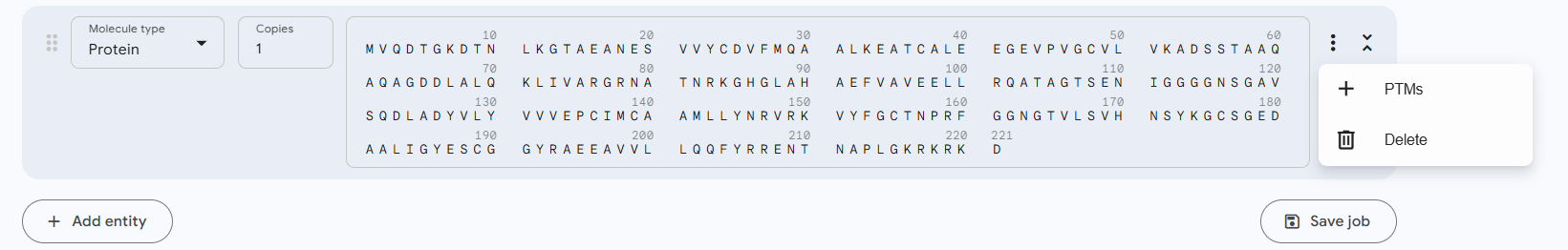
1个“token”标准：蛋白质——每个标准氨基酸残基；DNA/RNA——每个核苷酸残基；

配体——每个离子；修饰（不包括糖链）——每个修饰氨基酸残基或核苷酸的原子

具体查看：<https://alphafoldserver.com/faq>。

注意：每个蛋白质/核苷酸链必须**至少**包含**4个氨基酸或核苷酸**。

1. **提交步骤（以蛋白质序列为例）**
   1. **检查**是否有足够的**任务配额**，如果要运行的任务超出配额，可以先保存任务，待有配额后重新提交。
   2. **输入蛋白质序列**，可以选择添加蛋白翻译后修饰（post translational modifications，PTM）



* 1. 如果想添加其他类型的生物分子，可以点击“Add entity”。
  2. 添加完成后，点击“continue and preview job”。**修改**job name，选择**编辑或提交**任务。
  3. 任务提交完成后，等待一段时间便可查看运行结果。

1. **安装到服务器**

参考网址：https://github.com/google-deepmind/alphafold（[镜像访问](https://gitcode.com/gh_mirrors/alp/alphafold/overview?utm_source=csdn_github_accelerator)）

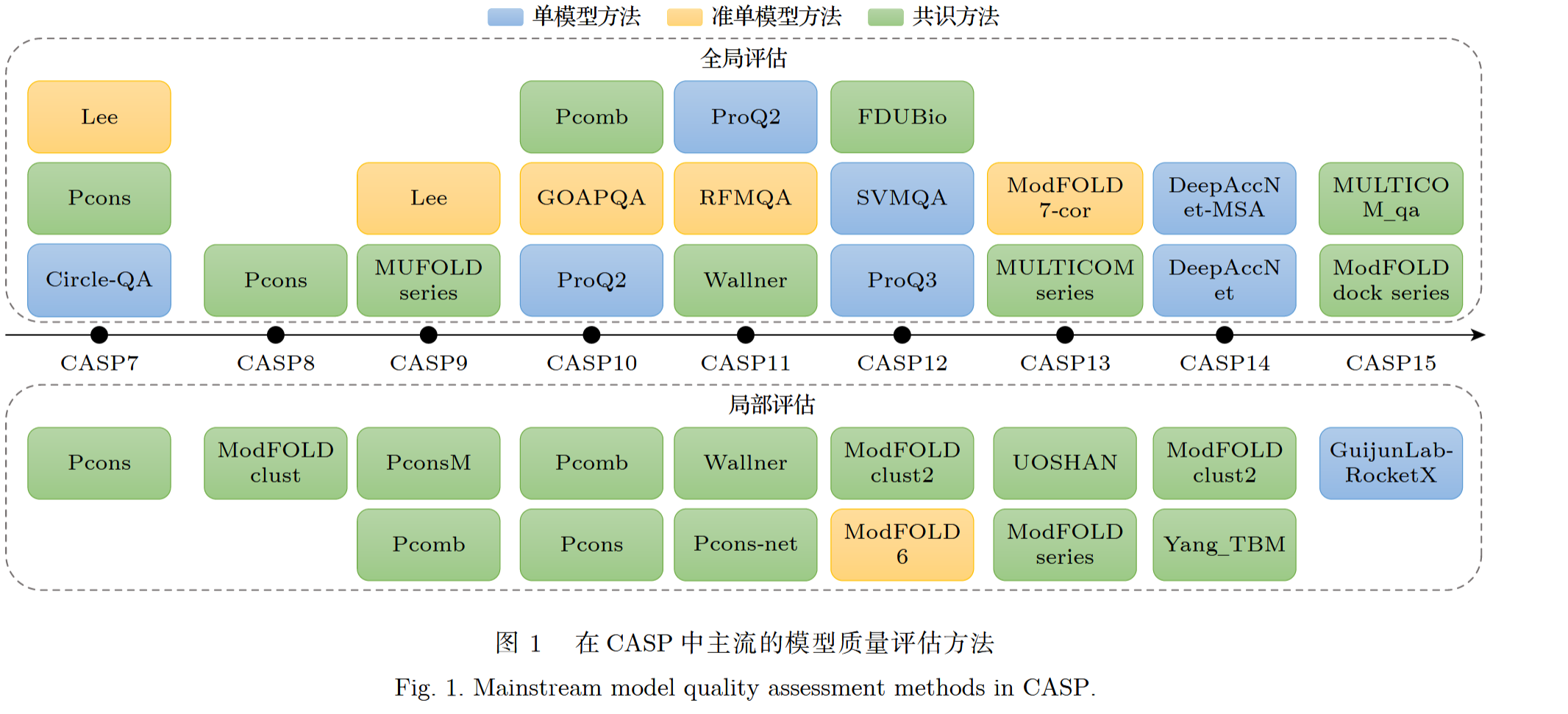
1. **使用Alphafold2在线预测**

若仅是预测个别蛋白质的结构，可以选择**谷歌的Colaboratory的在线版本**（[谷歌搜索Alphafold colab](https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb)）。

具体流程可以参考<https://blog.csdn.net/Eumenidus/article/details/123611569。>

1. **蛋白模型的准确性评估**

模型精度估计（estimation of model accuracy, EMA）：估计参考结构与预测模型在整体拓扑（全局结构）和残基级别（局部结构）相似的程度，并能够进一步实现模型单残基、连续残疾块的拓扑精修，常见的指标包括GDT-TS、TM-score、lDDT、CAD、SG。此外，对于复合物的评估除了全局/局部的评估外，还新增了接触界面精度估计，如DockQ和QS-score。



最近，在CASP和CAMEO比赛中采用的指标可以分为“单体结构质量评估指标”和“复合体结构质量评估指标”。

单体质量评估指标：

GDT-score (global distance test score)：计算预测模型结构中某种原子（如Cɑ）落在实验结构对应位置的某个阈值范围内所得到的最大原子数。通常，GDT-HA用到的阈值为0.5，1，2和4Å。GDT-TS用到的阈值为1，2，4和8Å。

TM-score：会对齐预测模型与参考结构之间所有残基对进行评估。

IDDT (local distance difference test)： 通过比较参考结构中一定范围内较近的、不属于同一残基的原子对之间的距离进行计算。如果模型中的距离与参考结构中的距离在一定的阈值范围内(如0.5，1，2和4Å)，则被认为是符合要求的距离。通过计算保留距离的比例，可以得到预测模型的lDDT。

PAE（预测对齐误差）：预测结构中两个标记之间的相对位置和方向的误差估计值。值越高表示预测误差越大，因此置信度越低。

pTM和ipTM分数：预测模板建模(pTM)分数和界面预测模板建模(ipTM) 分数均来自称为模板建模(TM)分数的度量。这可衡量整个结构的准确性。 pTM分数高于0.5表示复合物的整体预测折叠可能与真实结构相似。ipTM 衡量复合物内亚基预测相对位置的准确性。高于0.8的值代表有信心的高质量预测。

对于小结构或短链，PAE或pIDDT可能更代表预测质量。pIDDT在70以上，TM在0.8以上就可以认为准确性较高。