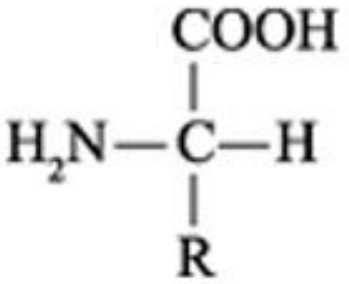
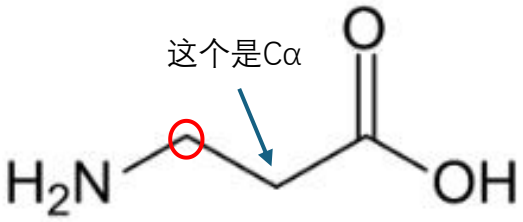
Preliminary knowledge

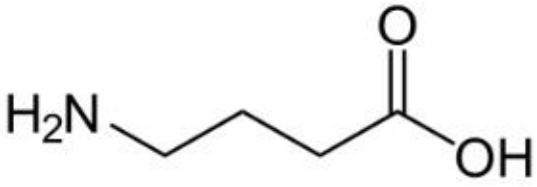
- 氨基酸是组成蛋白质的基本单位，其分子中含有一个羧基（-COOH）和一个氨基（$-NH_2$）。
- 依据氨基连在碳链上的不同位置，可将氨基酸分为α-，β-，γ-等氨基酸。
- 生物界中构成天然蛋白质的氨基酸被称作**天然氨基酸**（**基本氨基酸**），天然氨基酸共有20种，天然氨基酸全部是α-氨基酸（脯氨酸是**α-亚氨基酸**）
- 唯一常见的天然存在的β-氨基酸是β-丙氨酸， β-氨基酸组成的β-肽被认为有潜力作为防止产生抗生素耐药性的方法
- γ-氨基酸：最常见的是γ-氨基丁酸，重要的神经递质。它符合氨基酸的化学定义，但是在生物学角度上通常并不这么归类。



α-氨基酸的基本结构



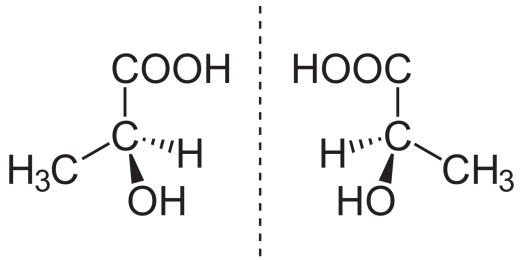
β-丙氨酸
可见氨基和羧基并不连接在同一个碳原子上



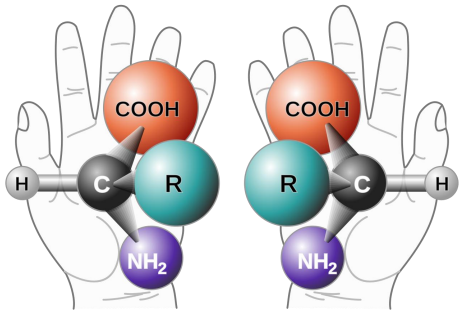
γ-氨基丁酸

Preliminary knowledge

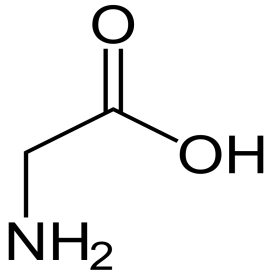
- 旋光异构：同分异构体的一种。指互为镜像（mirror images）的分子。不对称碳/手型碳（asymmetric carbon，C*）和四种不同的原子或原子基团连结，不对称碳为对称中心（chiral centers）。
- 除甘氨酸外，其余氨基酸的α-碳原子均为手性碳原子，存在L-型和D-型两种旋光异构体。组成人体蛋白质的氨基酸均为L-型。
- 氨基酸通过肽键相连而成的化合物称为**肽**（peptide），肽链中的氨基酸分子因脱水缩合而残缺不全，称为**氨基酸残基**。多肽链有两个末端，有自由氨基的一端称为氨基末端（简称"N-端"），有自由羧基的一端称为羧基末端（简称"C-端"），N-端在左边，C-端在右边。



乳酸的旋光异构



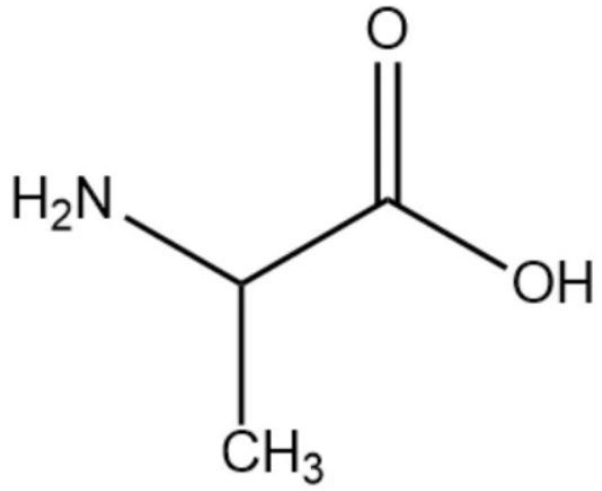
α-氨基酸的手性



甘氨酸
因为R基是H，所以并没有手性

Preliminary knowledge

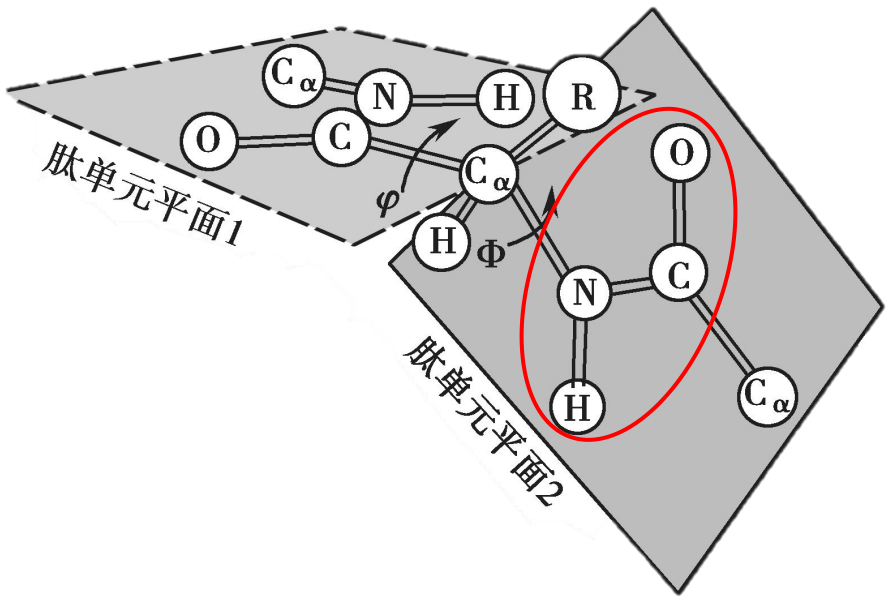
- 利用以上性质，就可以用一些方法表示氨基酸（以pdb数据为例子）。
- B因子：表示该原子的热运动程度

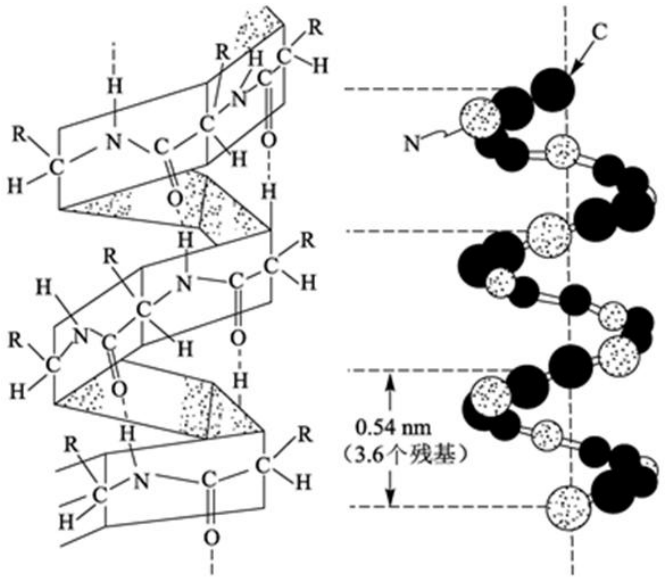


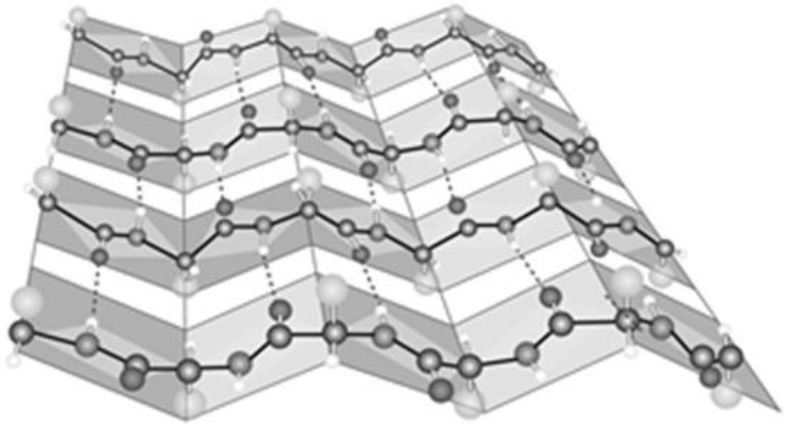
序号	原子名称	残基名称	链标识符	残基序号	X 坐标	Y 坐标	Z 坐标	占有率	B 因子	元素符号
1	N	ALA	A	2	17.406	18.438	57.995	1.00	74.03	N
2	CA	ALA	A	2	18.345	17.403	57.577	1.00	68.42	C
3	C	ALA	A	2	18.325	17.235	56.060	1.00	60.96	C
4	O	ALA	A	2	17.346	16.740	55.499	1.00	42.05	O
5	CB	ALA	A	2	18.022	16.085	58.263	1.00	57.36	C
6	N	PRO	A	3	19.413	17.639	55.398	1.00	67.14	N
7	CA	PRO	A	3	19.459	17.584	53.931	1.00	59.38	C

Preliminary knowledge

- 蛋白质的一级结构：氨基酸序列。维持蛋白质一级结构稳定的主要化学键是肽键、二硫键等
- 肽单元是由参与肽键组成的C、O、N、H四个原子和与它们相邻的两个氨基酸的α-碳原子（Cα1、Cα2）共同构成的刚性平面。
- 蛋白质的二级结构：蛋白质分子中某一段肽链的局部空间结构。
- α-螺旋：是指多肽链中肽单元通过α-碳原子的相对旋转，围绕中心轴做有规律盘绕而形成的一种紧密螺旋结构
- β-折叠：多肽链主链中一种比较伸展、呈锯齿状的二级结构形式

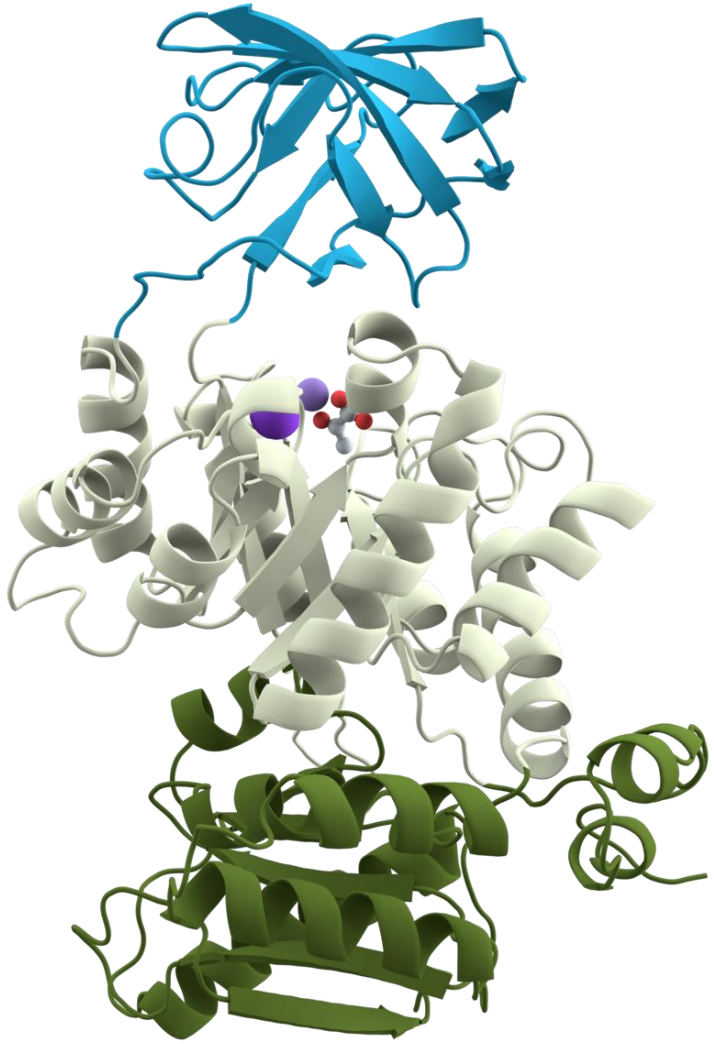






Preliminary knowledge

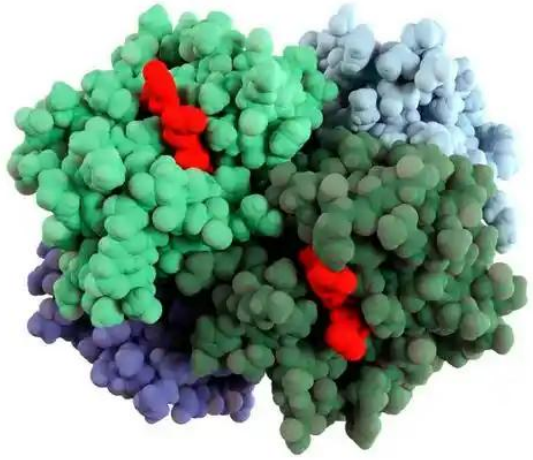
- β-转角：在蛋白质分子中，肽链形成180°回折，在回折角处的构象即为β-转角。β-转角通常由4个连续的氨基酸残基构成。
- 无规卷曲：肽链其余部分还没有确定规律性的局部空间构象称为无规卷曲。可构成酶活性部位和其他蛋白质特异的功能部位。
- 超二级结构：超二级结构是指若干相邻的二级结构组合在一起形成的有规则的、稳定的、在空间上能辨别的二级结构组合体，如α-螺旋组合（αα）、β-折叠组合（βββ）和α-螺旋β-折叠组合（βαβ）等。
- 结构域：结构域是指在二级结构或超二级结构的基础上，多肽链形成的相对独立的紧密球状实体。



丙酮酸激酶的三个不同结构域

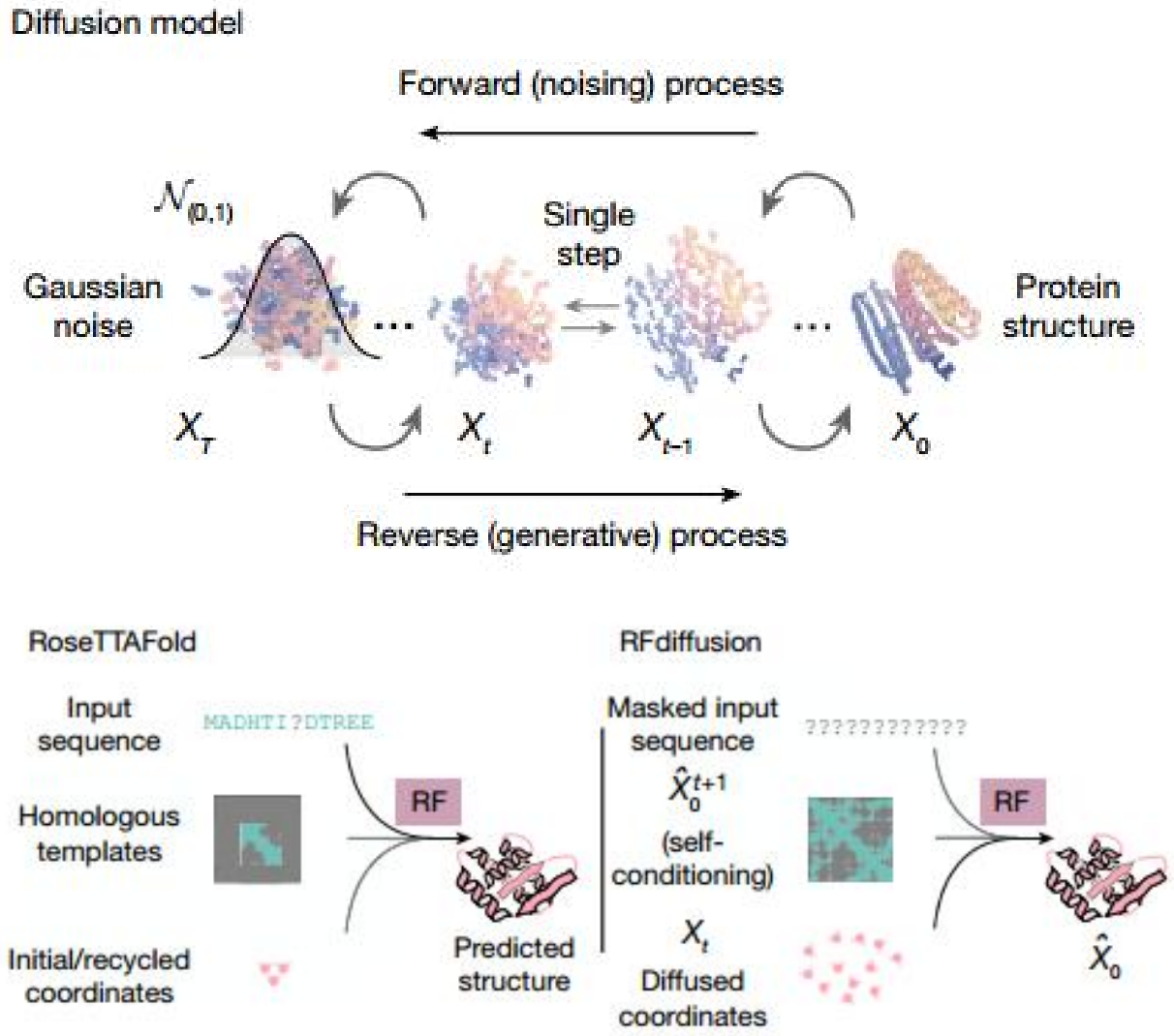
Preliminary knowledge

- 三级结构：二级结构的基础上，多肽链进一步盘曲、折叠而形成的三维空间结构，包括主链和侧链中所有原子的空间排布。
- 蛋白质三级结构的形成和稳定主要靠多肽链R基团之间相互作用而形成的次级键来维持，如氢键、疏水键、离子键和范德华力。
- 由一条多肽链构成的蛋白质**只有**具有三级结构时才能发挥其生物活性。
- **亚基**：具有完整三级结构的多肽链。亚基也是蛋白质的最小共价单位。蛋白质的大型装配体（如病毒）往往只使用少量类型的蛋白亚基作为构建单元。一些天然存在的蛋白质具有相对少量的亚基，因此被描述为寡聚体。
- 四级结构是指蛋白质分子中各个亚基之间的空间排布及相互作用。



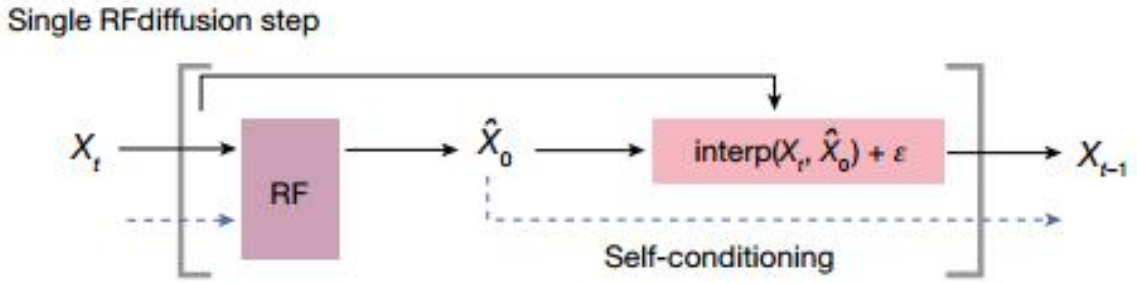
血红蛋白由四个亚基组成，其中每个亚基由一条肽链和一个血红素分子构成

RFdiffusion的扩散框架

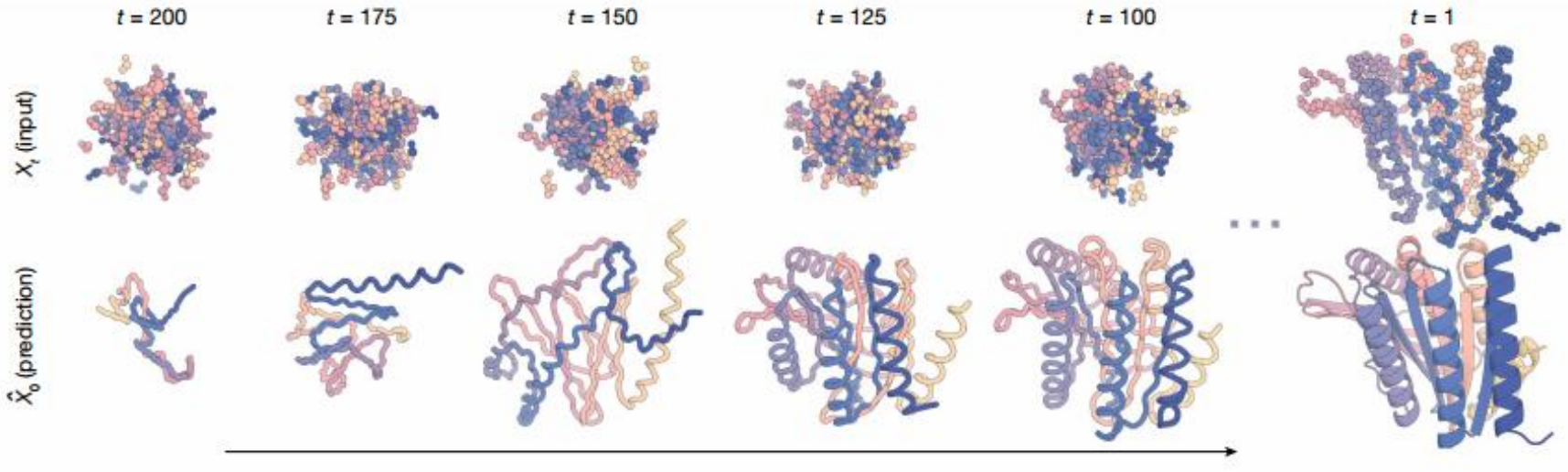


- RoseTTAFold拥有序列、结构模板以及初始坐标这三个信息的输入
- 对于骨架结构的生成看成是每个氨基酸局部坐标系相对于全局坐标系而言的旋转矩阵R和平移向量Z的生成
- 前文提到，只需要确定Cα就可以确定氨基酸的其他原子的相对位置，所以平移向量Z相当于就是Cα原子在全局坐标系下的坐标。
- 在全局坐标系发生旋转和平移时，预测的结果也应当发生相同的旋转和平移。
- 在旋转矩阵加噪时，这里将生成建模扩展到黎曼流形上，并将扩散过程定义为在流形上的布朗运动。其原因其是高斯分布不明确。

Self-Condition

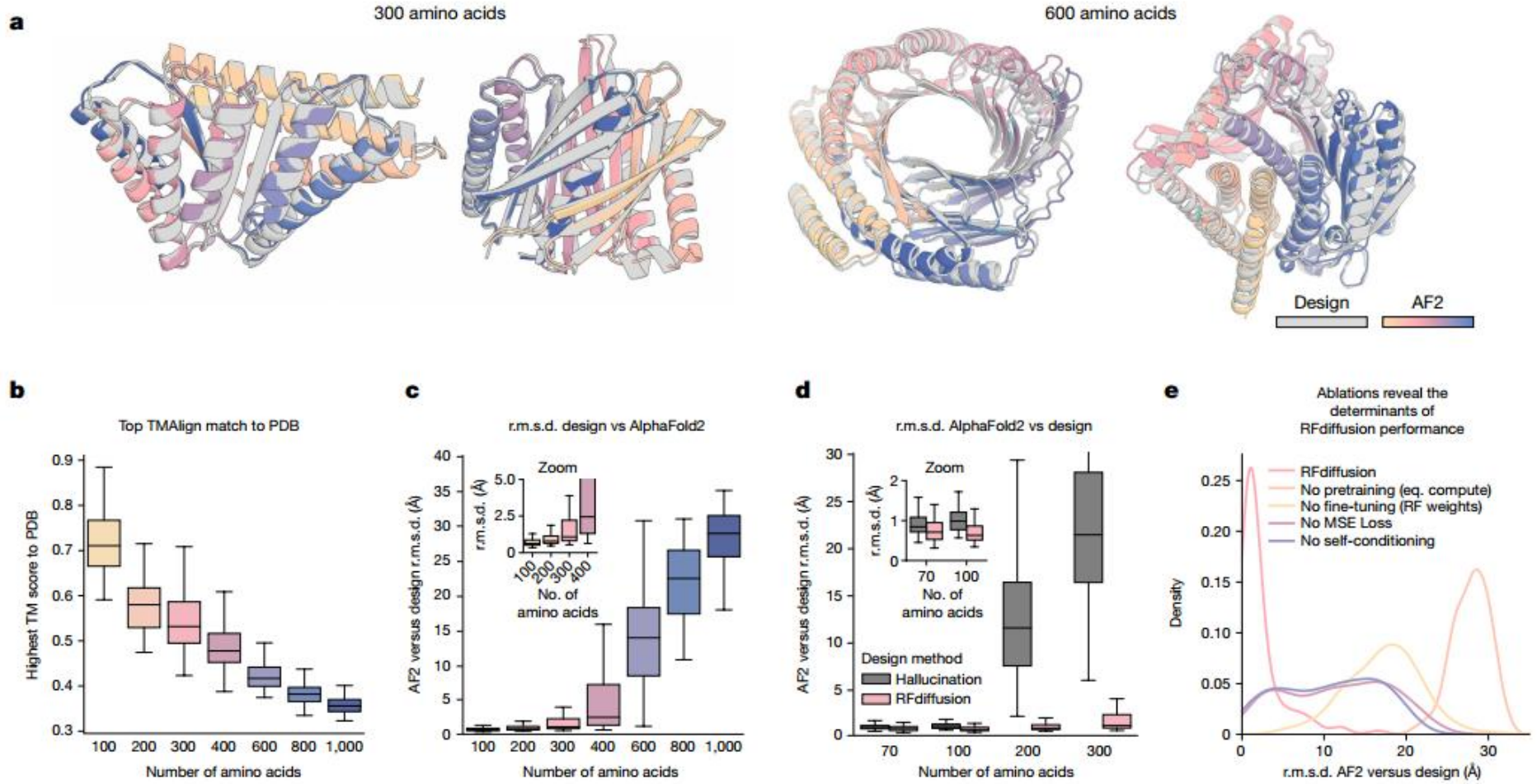


$$
\begin{aligned}
x_{t+1} &= \text{ForwardNoise}(x_0, t+1) \\
\hat{x}_{0,\text{prev.}} &= \text{RFDiffusion}_\theta(x_t, \vec{0}, t) \\
\hat{x}_{0,\text{prev.}} &= \text{StopGradient}(\hat{x}_{0,\text{prev.}}) \\
x_t &= \text{ReverseStep}(x_{t+1}, \hat{x}_{0,\text{prev.}}, t)
\end{aligned}
$$



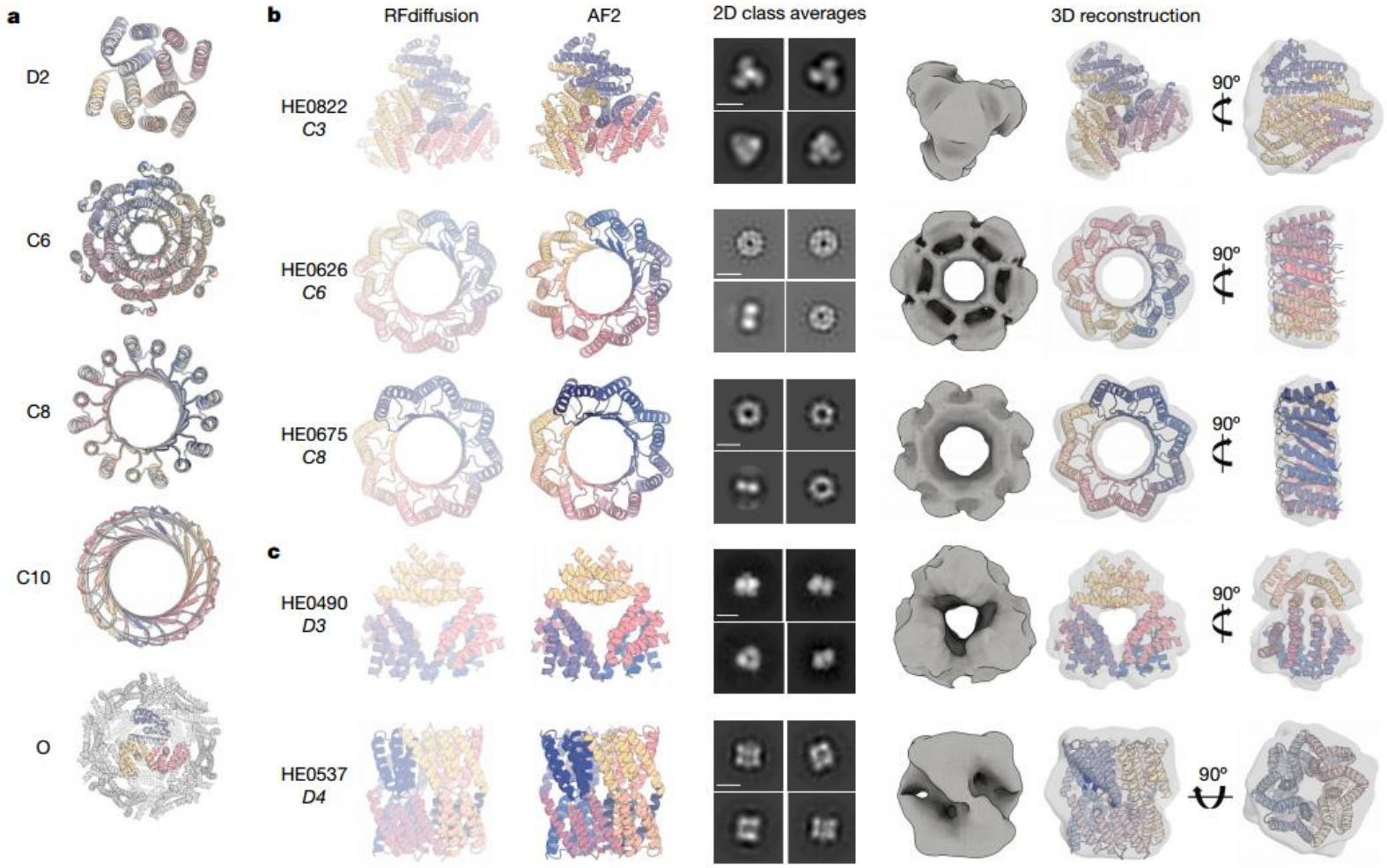
Unconditional protein monomer generation

作者按照蛋白质序列长度进行分组进行生成，并**使用ProteinMPNN对输出的结构进行设计**，使用alphafold2单序列预测模型来评价生成效果。
但值得注意的是，在100个氨基酸序列长度区间中，RF Diffusion生成的蛋白结构非常类似于PDB中已有的结构，在200-400的区间内逐步展示出一定的多样性生成能力。



high-order symmetric oligomers generation

高阶对称寡聚体（对称自组装颗粒蛋白）已经被证明是对疫苗、药物纳米载体的重要形式。由于RF Diffusion可以显式地对初始坐标输入直接进行操作，因此可以很轻松地在其中加入对称性的概念，在本次测试中，RF Diffusion已经实现了直接生成整个完整的颗粒八面体和二十面体。



点群对称性可以用旋转矩阵的有限集合来表示，我们可以用围绕z轴旋转(360/K)°的旋转矩阵集合来表示K阶循环对称组。

即对于一个多聚体

$X = [x^1 \ldots x^k]$

对于其中的某一个单体

$x^k=([z_1^k,\ldots,z_M^k],[r_1^k,\ldots,r_M^k])$

只要重建出其中一个，就可以重建出完整的多聚体。

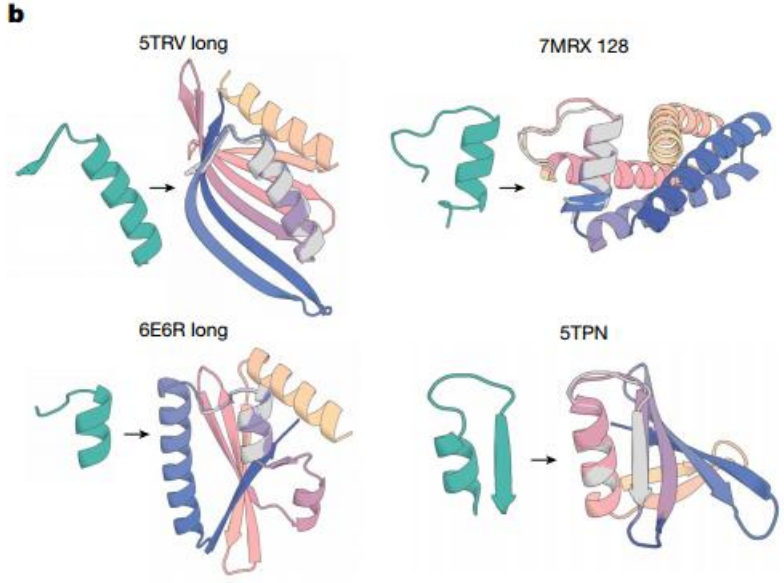
Functional-motif scaffolding

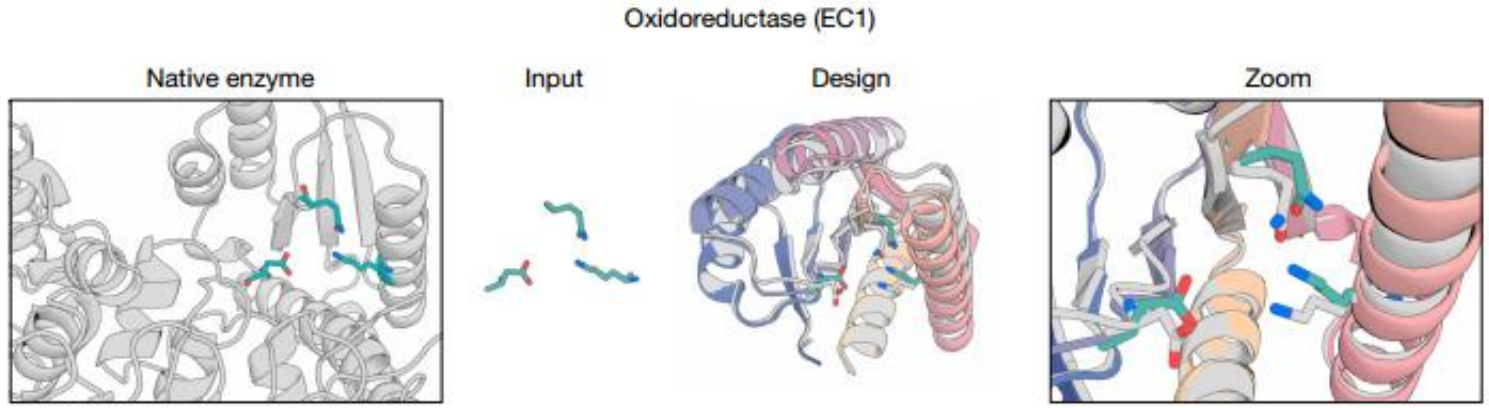
已知某段连续或者不连续的功能位点的氨基酸信息以及结构信息，创建一个稳定的支架来支持功能位点。

训练方式：

- 从pdb数据库中抽取原始数据X
- 将数据分为Xn，Xm（假设Xn为骨架）
- 给骨架坐标上加噪。
- 计算*loss*

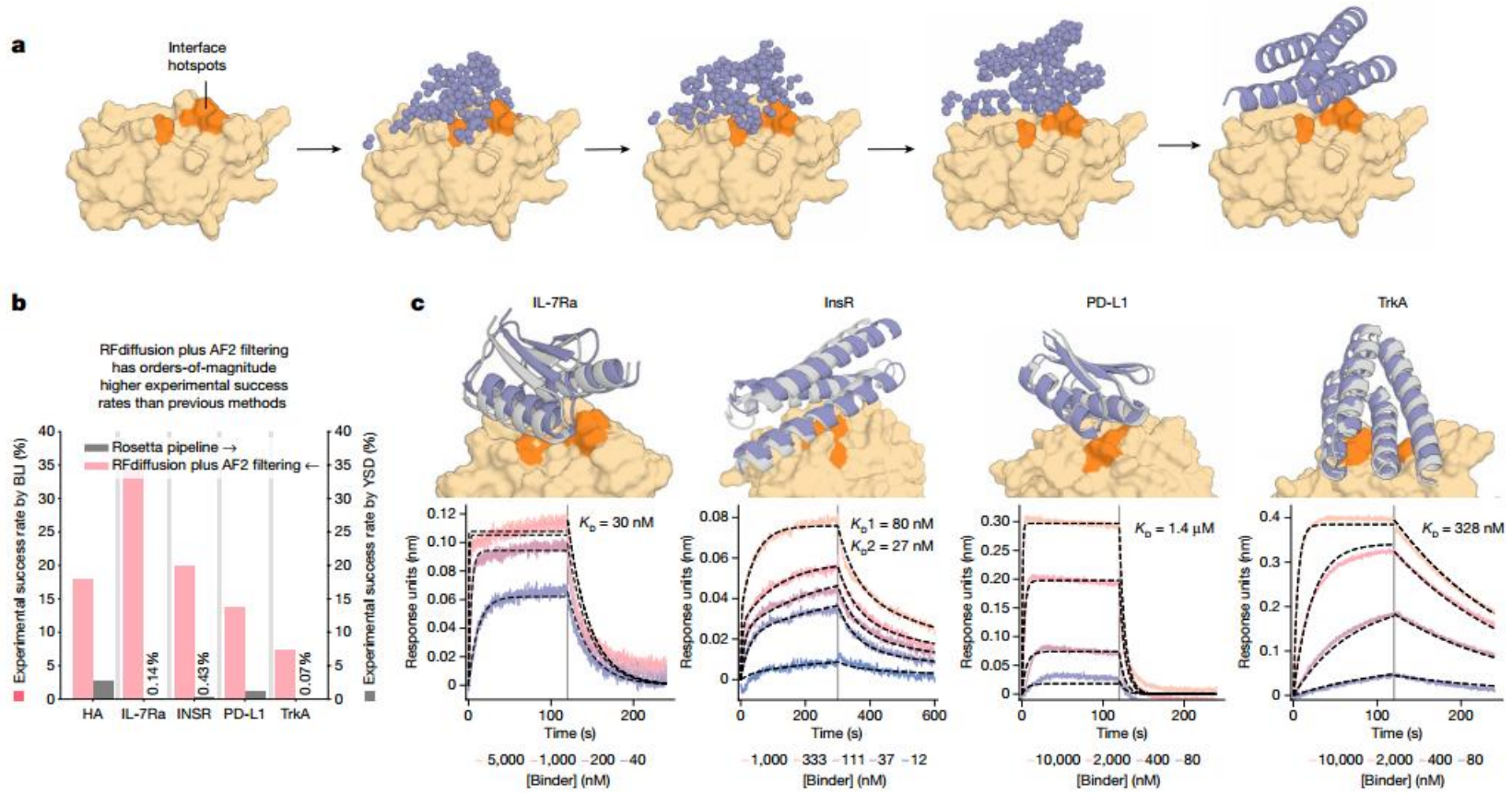
为了鼓励功能位点上的结构信息不发生改变，所以在预测时，每一步去噪中用的功能位点的输入坐标信息即为原始坐标信息，并且还会引入功能位点的侧链扭转角以及氨基酸序列信息。并同时计算在功能位点和骨架上的预测损失。





Design of protein-binding proteins

设计靶向结合蛋白的高亲和力结合物是一个重要的蛋白质设计任务。为了解决这个任务，*RF Diffusion*提供了*ppi*版本，通过输入一组接触界面的残基编号，让模型自动生成对应结合到该表位的*binder*结构，不仅如此，还可以输入邻近的二级结构信息来控制*binder*的拓扑结构。



当数据为复合物时，只对其中一条链进行加噪，另一条链保持固定。

在训练时，50%的时间会引入加噪对应链的二级结构信息。剩下50%的时间会引入块邻信息。