Unipot 序列数据

结构 pdb

解析一个结构 也要几十万

Pdb大概有超过三十万结构的数据

Swissprot 函数

蛋白质语言的模型 词汇就是20个氨基酸

Bert 构造很多的完形填空

上下文

Msa的搜索，

Rowise 的attention

Column 的attention

Aflod evoformer的

从左到右边

Deffusion其实就是在去噪，用到的是全部的信息，用在大语言模型，预测每一个token， 可以由左到右也可以由右到左

图像是一个RGB，加的noise，加的是高斯的

Diffusion 的很快

大语言模型的缺点是比较慢，

谷歌的gemini 用diffusion来做大语言模型，其实是有

图神经网络

从左到右，从右到左

Atom graph 用

传统的gnn

Identity

技术思路

边跟边之间的关系

蛋白质边跟边的蛋白结构，model，边跟边二面角

以前在图像里面做自监督的

预训练后对

学习了能量函数，

这个能量函数

基于结构的模型 去解构

把抗原给小鼠，产生抗体，

在抗体里面，hit的

Lit里面挑一个lead，做抗体就是蛋白，本质上是对抗体进行优化，其实本质上是要提高蛋白的优化，换几个氨基酸，提高他们的亲和力，

AI想要解决这个问题，就是预测

预测一个函数，那就知道optimization是怎么做的

要去拟合这个能量函数其实是比较难的

从生物的角度怎么解决protein和proteino

新冠的抗体进一步的欧化，label data，但是很少，大概就是1-2万个，也做了预训练，核心是理解蛋白与蛋白之间怎么相互结合

但是预训练

做的也是csntruct

Nature的interface，random

两个nature

输入就是两个结构，

预训练，pdp数据，告诉我哪些数据可以bind，哪些是不能bind

做了一个实习那后，fine-tuning

生物真的做饱和突变库

最后做真实的方法和预测的方法之间的关联，

Case： Ab1372 optimization against JN.1开源模型

亲和力提高17倍

同样的思路，换了一个抗体，就又做了一个工作

抗体的设计 ，比较难的部分就是

酶的设计比抗体会更加容易， 酶的funtion会更加容易

在生化领域，AI非常重要，因为生化都是某个，AI喂给他数据，他就是function，但是今天神经网络就是建模型，fit

从AI的大模型 生成式模型 生产

酶比较多的是化学

做从头设计，会需要的