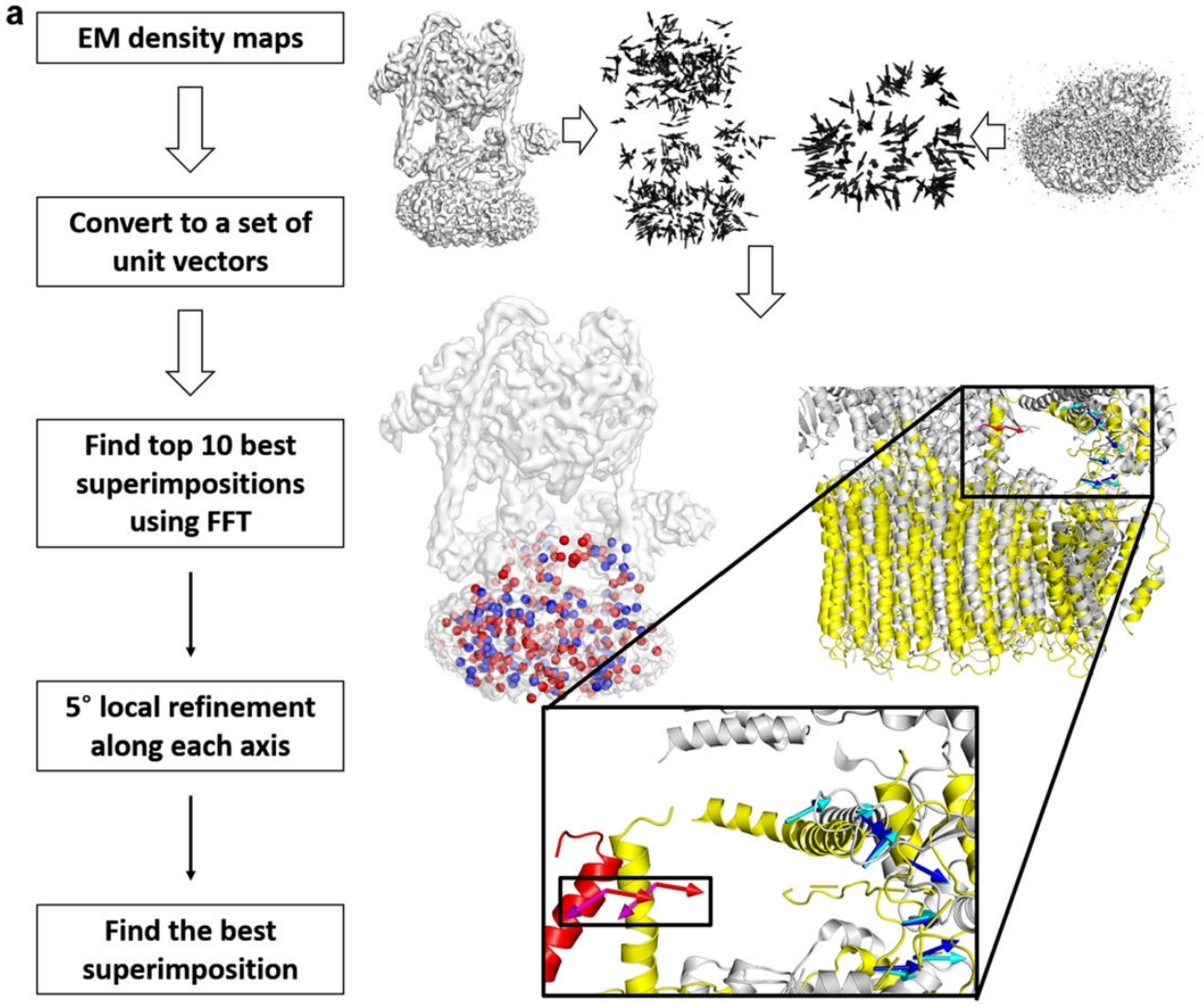
DiffModeler: large macromolecular structure modeling for cryo-EM maps using a diffusion model

2024, Nature Methods

Introduction

- 当分辨率高于5A时，可以利用深度学习来检测图谱中的原子位置，从而直接得到蛋白质和核酸的主链。
- 对于中等分辨率范围内的地图（5-10A），从头建模通常是不可行的，即使使用深度学习技术，氨基酸残基和原子的识别仍然难以实现。通常是与已有PDB中的原子模型或预测出的原子模型进行拟合。
  - 现有方法：Phenix, Flex-EM, Assembline, MultiFit, Chimera, MarkovFit, **VESPER**
  - 缺点：分辨率很低或亚基数量很多时拟合效果不好
- DiffModeler: 5A-10A、
- VESPER：VEctor-based local SPace ElectRon density map alignment
  - Pipeline



  - Mean-Shift算法：

$$y_i = \frac{\sum_{n=1}^{N} k(x_i - x_n)\Phi(x_n)(x_n - x_i)}{\sum_{n'=1}^{N} k(x_i - x_{n'})\Phi(x_{n'})}$$

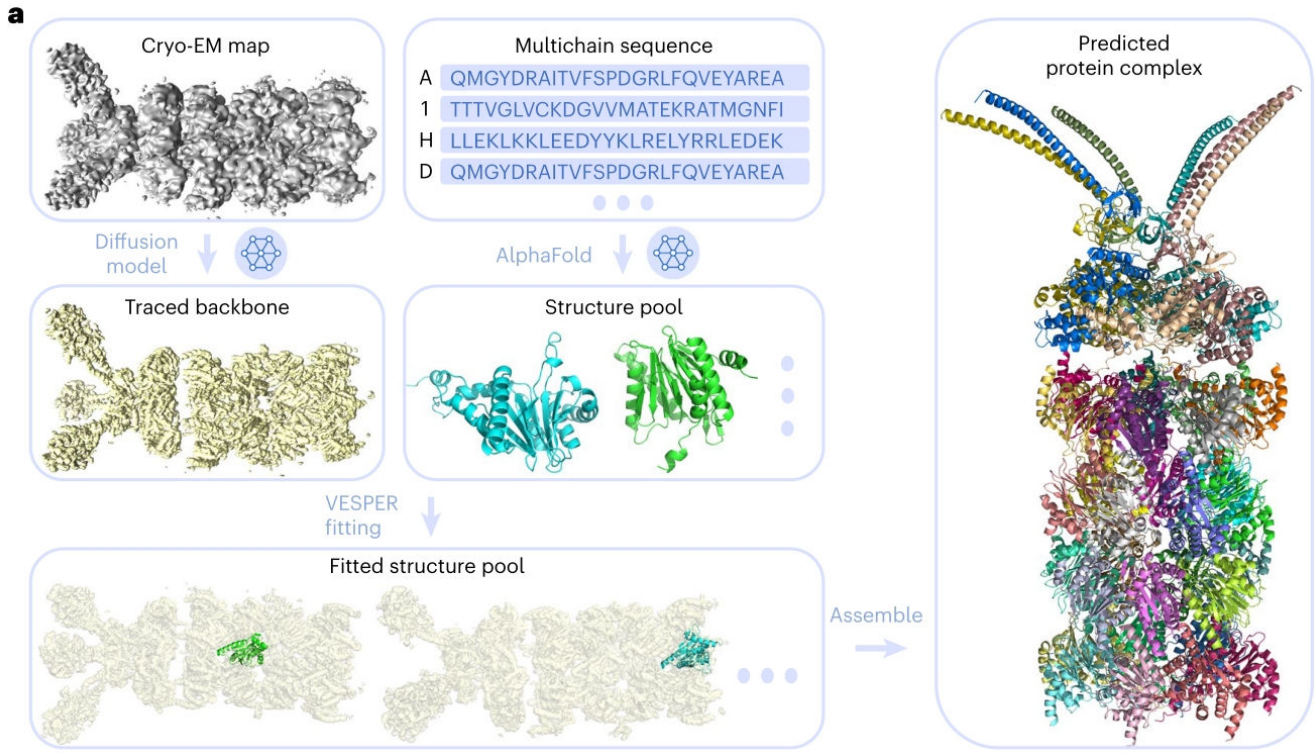
其中，x_i 是网格点， $\Phi(x_n)$ 是网格点 x_n 的密度值, $k(p)$ 是高斯核函数(σ是超参数):

$$k(x) = \exp(-\frac{1.5x^2}{\sigma^2})$$

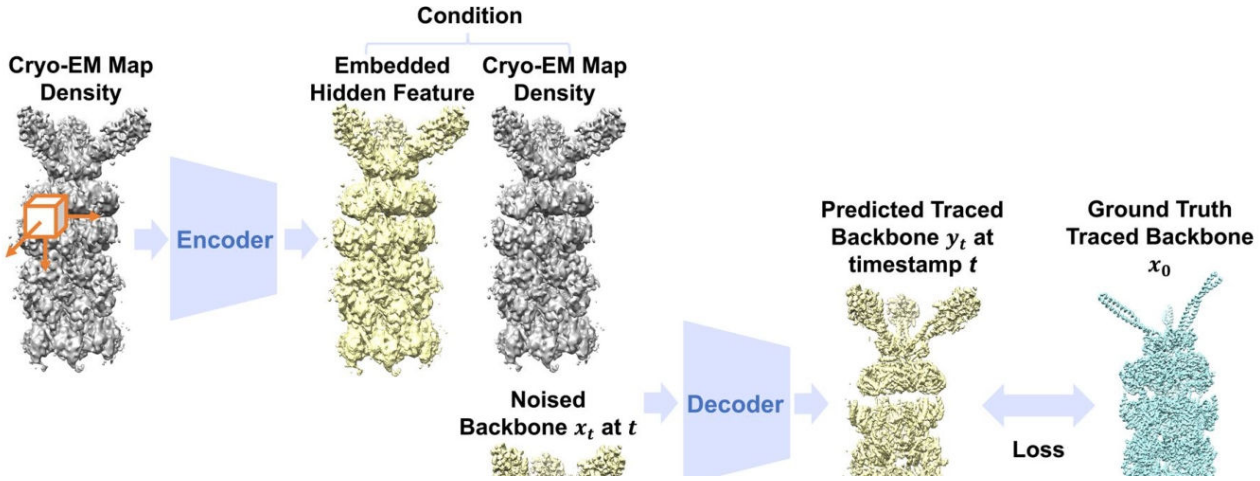
- DOT score

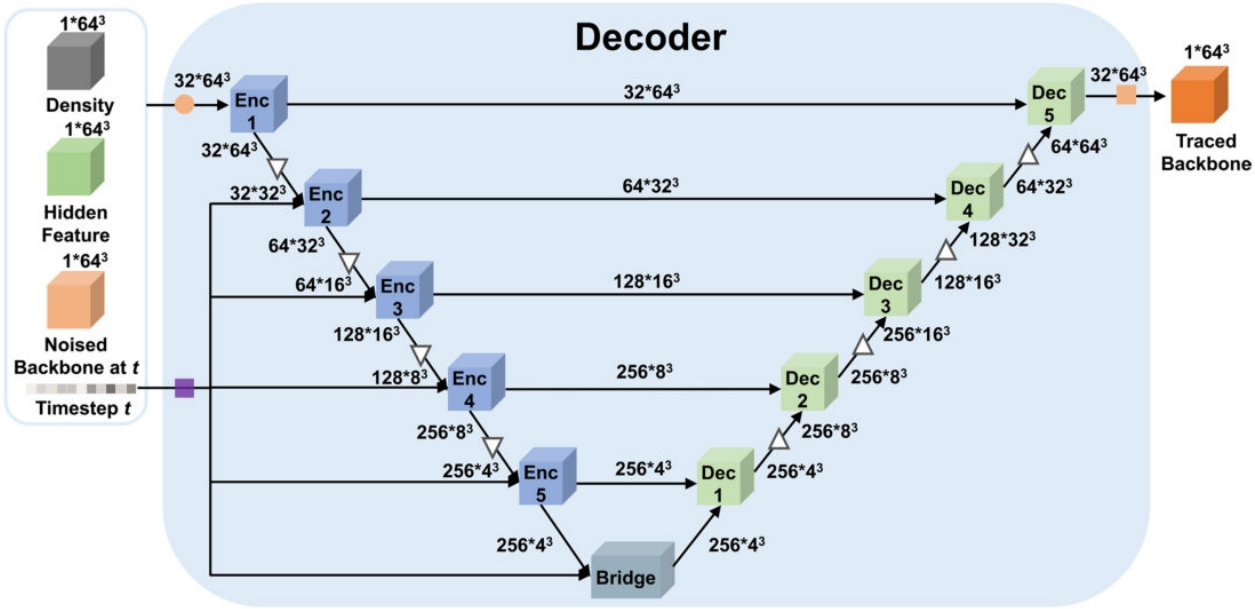
Method

- Overview



- Backbone tracing via the diffusion model
  - 训练集：从EMDB中选取5A-10A且与对应PDB中的原子模型拟合较好的密度图。将其在每个pixel处裁剪出一个64^3的方块，将方块中的点按是否属于主链(距离主链小于2A)来设置0和1.
  - Encoder-Decoder结构

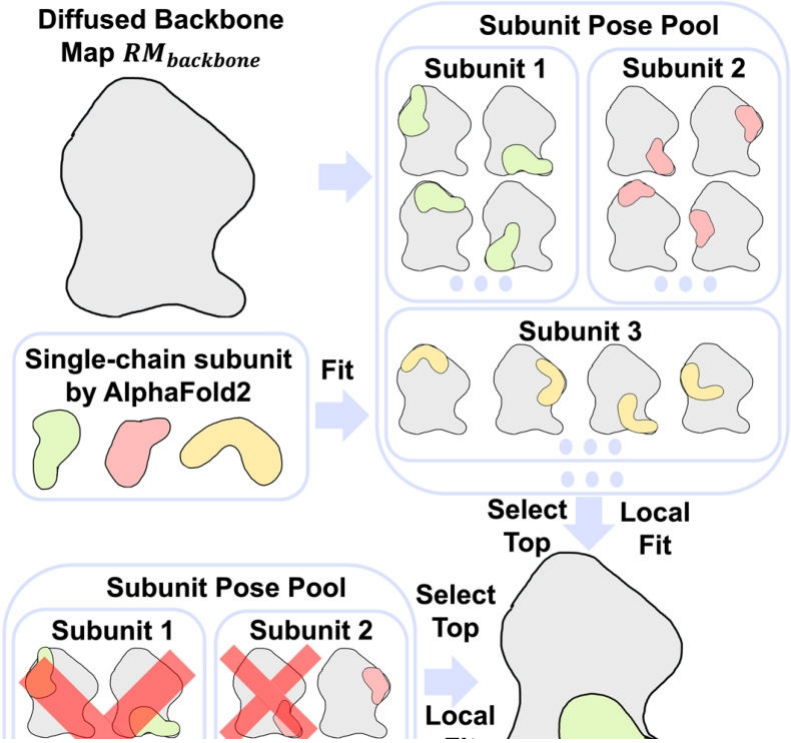


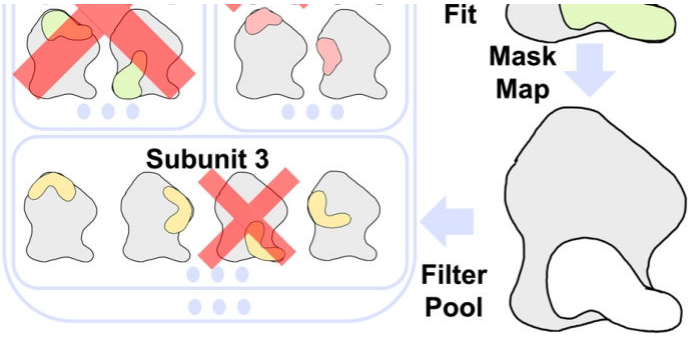


◦ 损失函数:

$$\begin{cases} L_{\text{Dice}} = 1 - \frac{2\times\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} {p_i}^2 + \sum_{i=1}^{N} {g_i}^2 + \varepsilon} \\ L = \frac{1}{B}\sum_{k=1}^{B} L_{\text{Dice}}(k) \end{cases}$$

- Structure prediction by AF2
- Structure model fitting with VESPER
- Protein complex modeling by a greedy assembling algorithm





- Fitting quality estimation in DiffModeler