论文介绍：Determining structures in a native environment using single-particle cryoelectron microscopy images

- 通过将非目标蛋白的密度视为非高斯噪声，提出了一种新的目标函数，大幅提高了在单张microscoph中识别目标蛋白的效率。
- 提出了一种排序函数，减小了model bias，并在后续的refine的过程中提高了分辨率。

背景介绍

冷冻电子断层扫描（Cryo-electron tomography）虽然能够实现接近纳米分辨率的结构解析，但其过程需要采集倾斜系列图像，tilt series通常包含30多张在不同倾斜角度下拍摄的图像，采集过程通常需要30分钟，效率较低。

在insitu SPA中，密度可能会被来自周围蛋白的密度覆盖。这些覆盖的密度可以被视为低频噪声。然而，这些低频部分具有较高的信噪比，对于确定蛋白复合物的初始位置和方向至关重要。

在目标蛋白被覆盖的情况下，通过单颗粒结果或子断层平均提取蛋白复合物的初始中心和方向信息后，无需subtract覆盖密度直接使用local refinement即可重构。

核心公式 weighting function

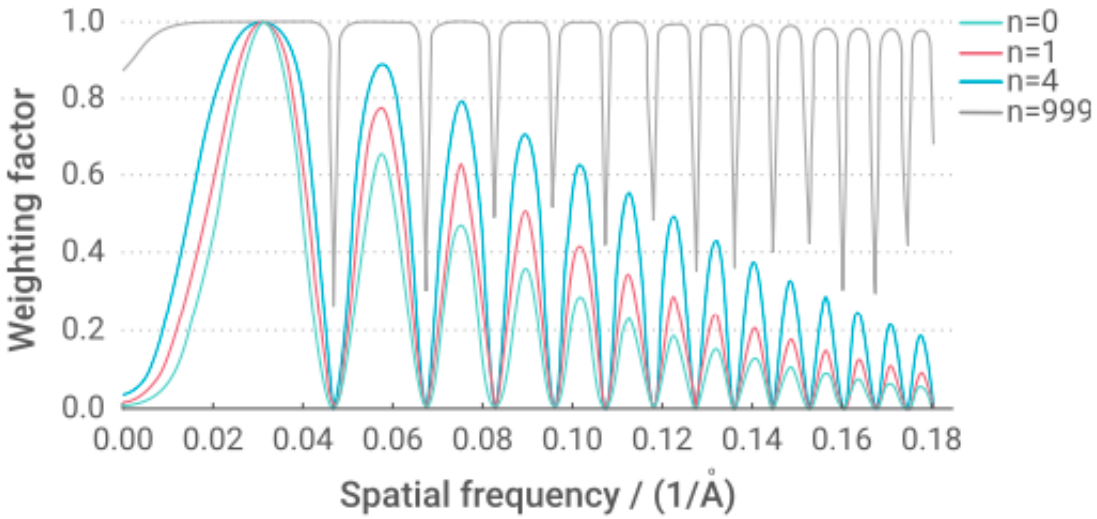
将重叠的蛋白质密度视为CTF调制的噪声，并将重叠蛋白质与目标蛋白的比例近似为一个常数n。

$$W(k) = \frac{CTF(k) \cdot FSC(k)}{\frac{1}{SSNR(k)} + n \cdot CTF^2(k)}$$

其中 CTF 是颗粒的衬度传递函数；SSNR 是目标蛋白质密度与由功率谱评估的射频噪声的比率；FSC 是 三维模板的傅里叶壳相关；n 是重叠蛋白质强度与目标蛋白质密度的比率。n 的增加表明来自重叠蛋白质密度的噪声增多，归一化的高频权重随 n 的增加而增加。

【推导见supplementary material 2】

weighting function的形状：



颗粒检测 particle detection

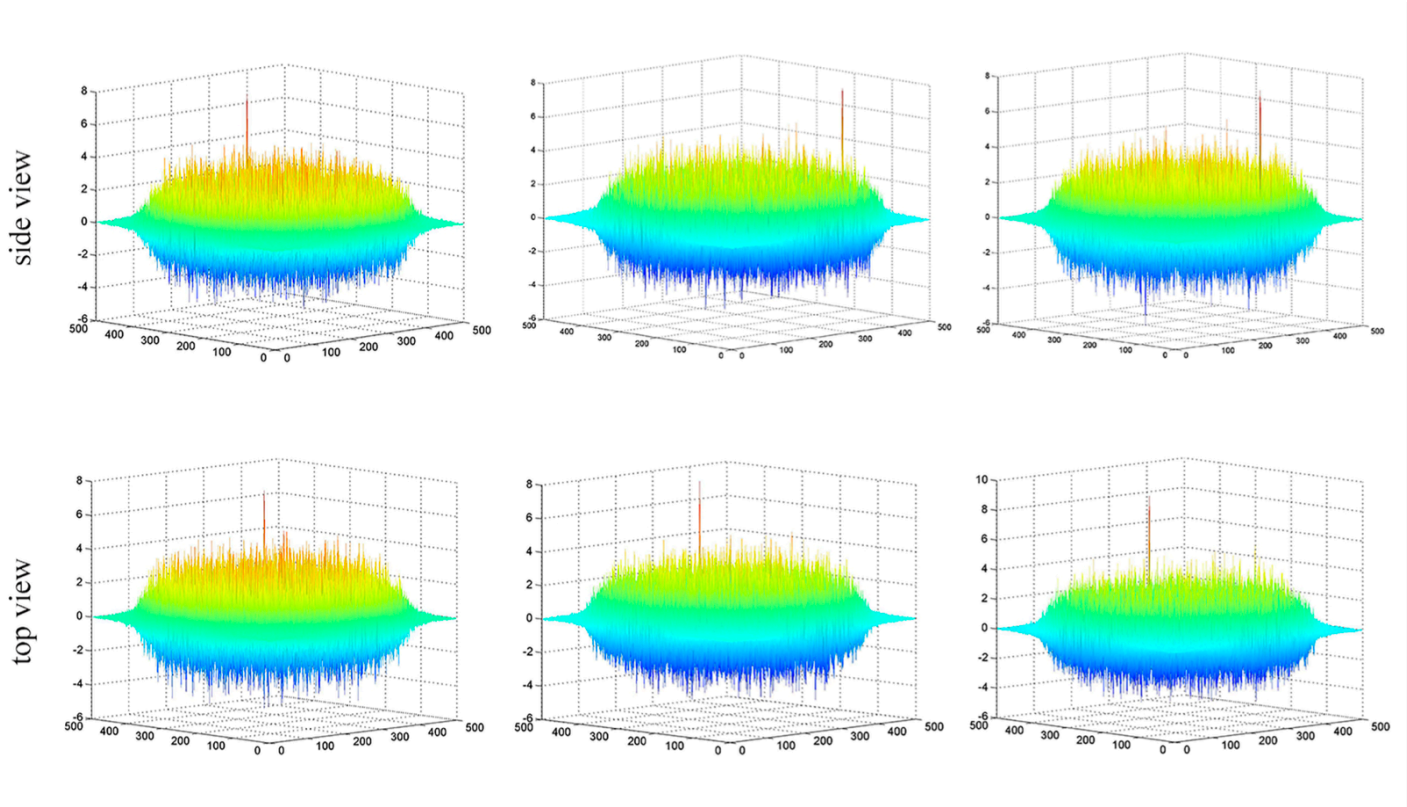
在三种病毒数据集中测试，这些蛋白复合物的密度与其他蛋白质和病毒基因组的密度重叠，模拟了复杂的成像环境。**每个颗粒图像的旋转和平移参数由对这些病毒的高分辨率SPA结构确定，故一定准确**。初始模型的投影图列表通过EMAN间隔5度来获得。

目标蛋白质的particle位置确定是基于 Weighted Cross Correlation Grams，输出是 CCG 峰值与标准差的比值，用于决定是否保留某一位置。加权交叉相关定义为：

$$\mathrm{cc} = \sum_{k} \mathrm{W}(k) \cdot \mathrm{X}(k) \cdot \mathrm{M}^{*}(k)$$

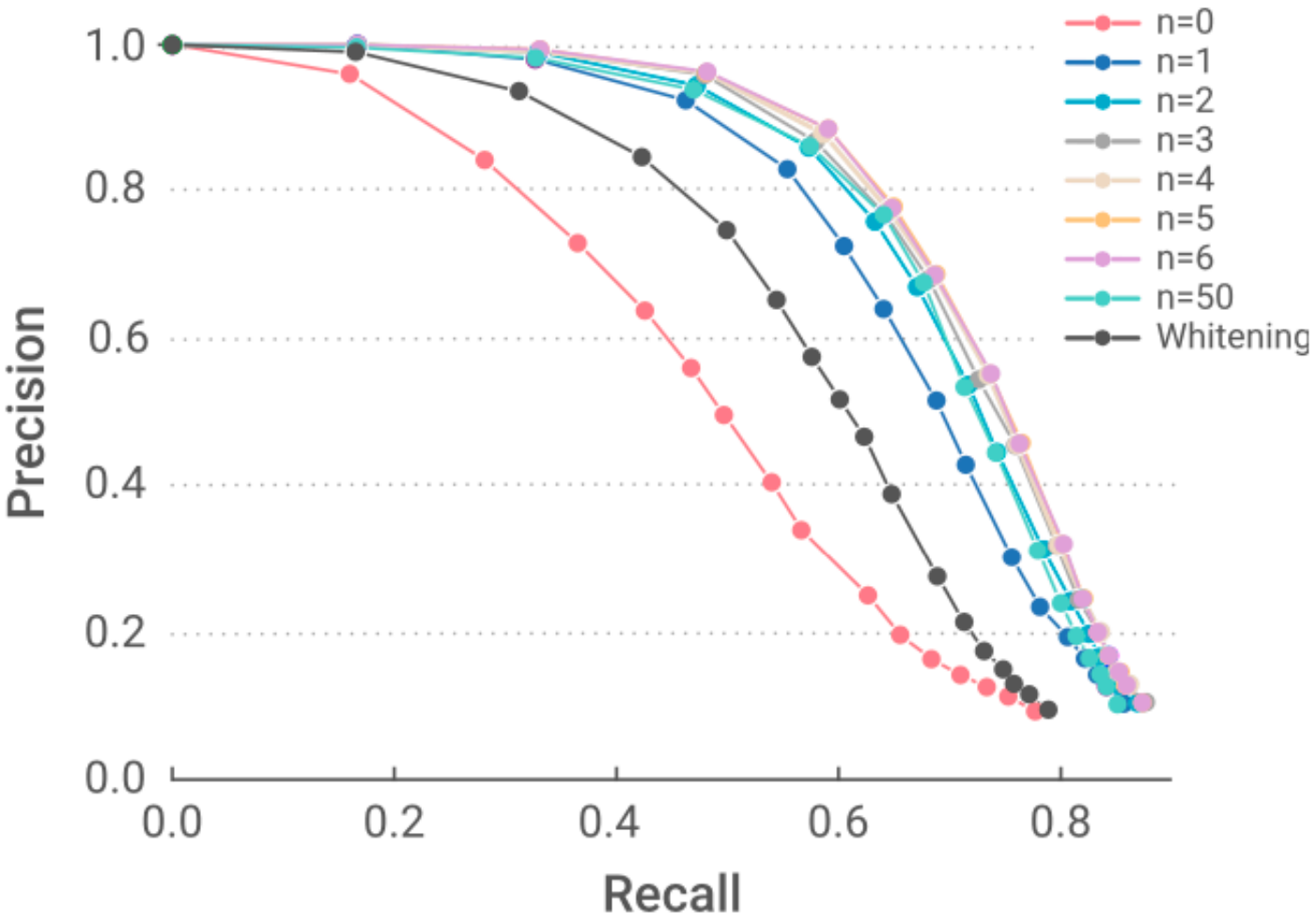
其中，k 是空间频率，$W(k)$ 是加权函数，$M^{*}(k)$ 是三维模板投影的傅里叶变换的共轭复数值，$X(k)$ 是傅里叶形式的原始颗粒图像。

投影图和颗粒计算的CCGs (Cross Correlation Grams)如图所示：

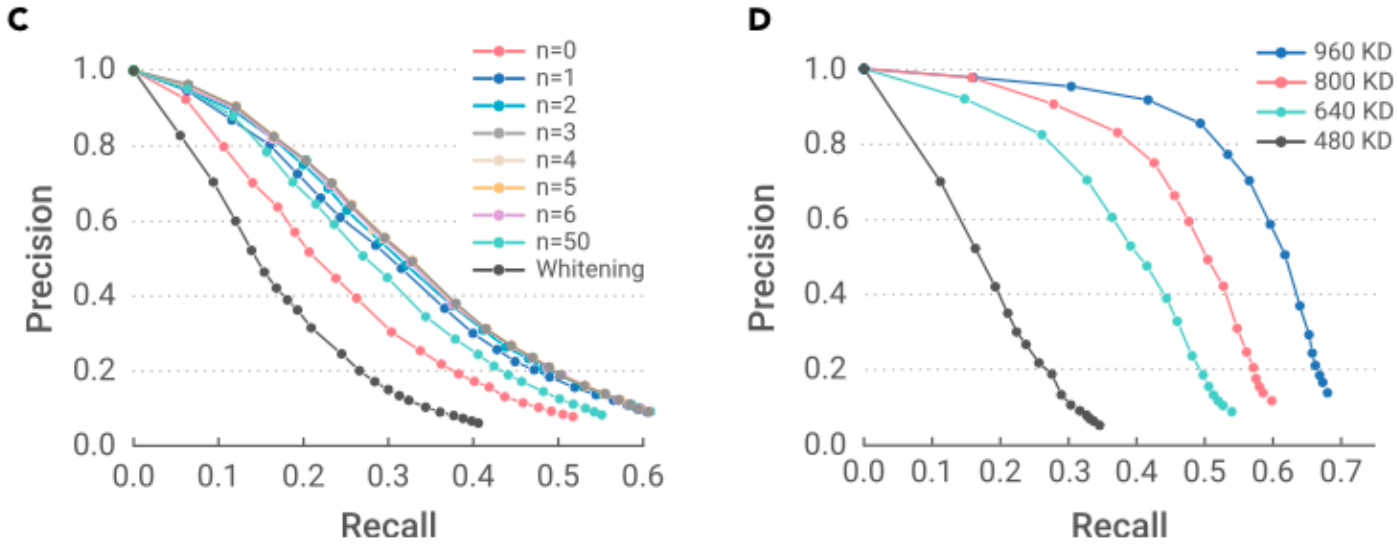


如果平移参数误差超过5 pixels或者旋转参数误差超过6度，认定为false positive。

准召率曲线如图，n = 3-6时最高。

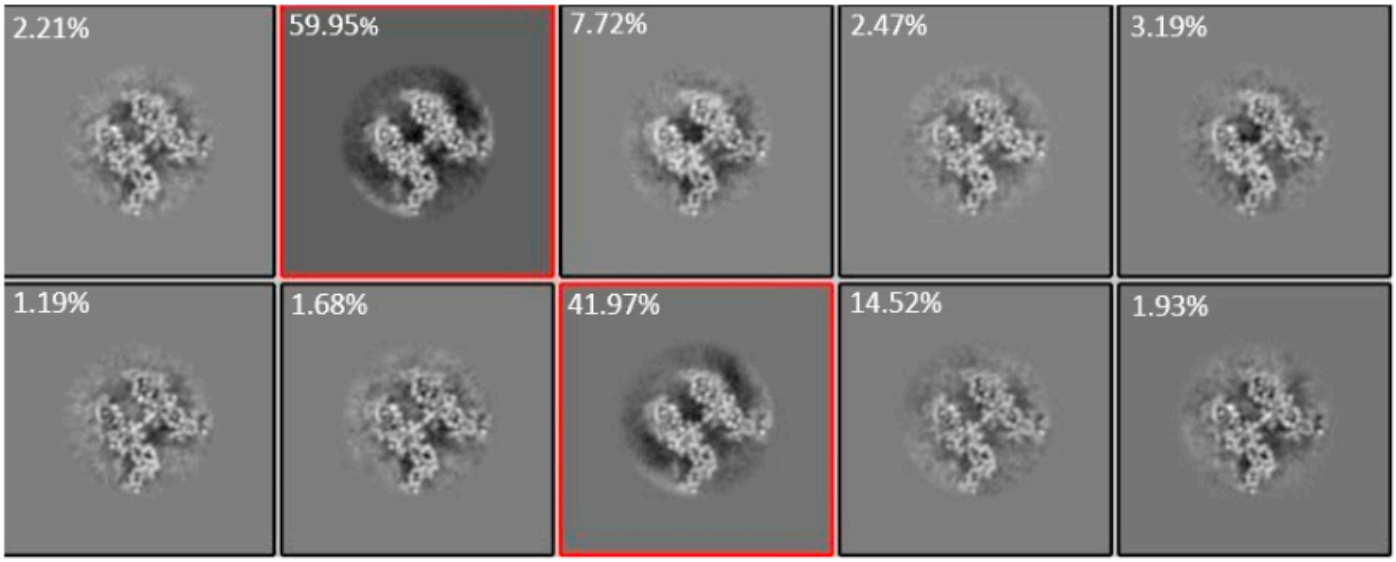


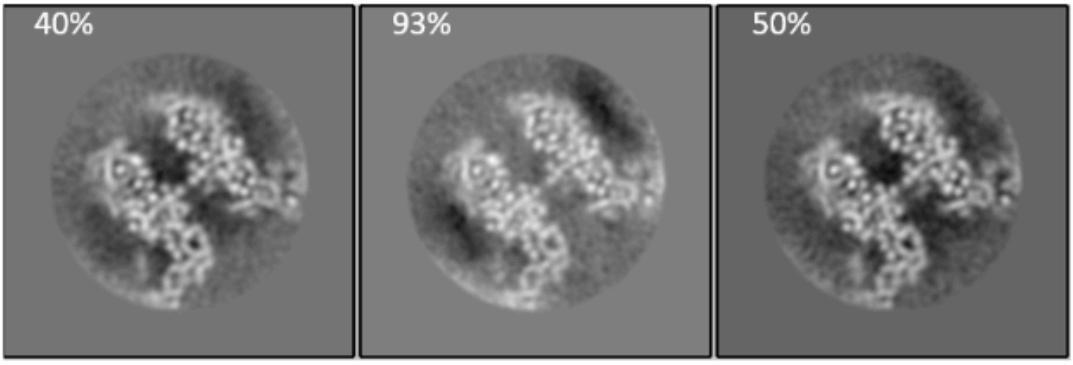
结果表明，当蛋白复合物的尺寸减小时，正确结果与总检测结果的比例迅速降低。



设置n为4，并使用从100Å 到8 Å 的频率范围进行颗粒检测。根据CCG函数的得分计算并排序了目标蛋白的可能位置和取向。在将具有相似取向（7°以内）且邻近位置（10像素以内）的颗粒合并为一个颗粒后，从每个病毒颗粒中选取了前500个得分最高的蛋白进行进一步的数据处理，其中约88%的结果是假阳性。在RELION中执行了3D分类，跳过了对齐，使用了由我们的选择函数提供的

位置和取向信息。正确率如图，10类里有2类包含最少的假阳性颗粒，仍然有50%左右的假阳性颗粒，可能也可以做进一步的分类筛选：





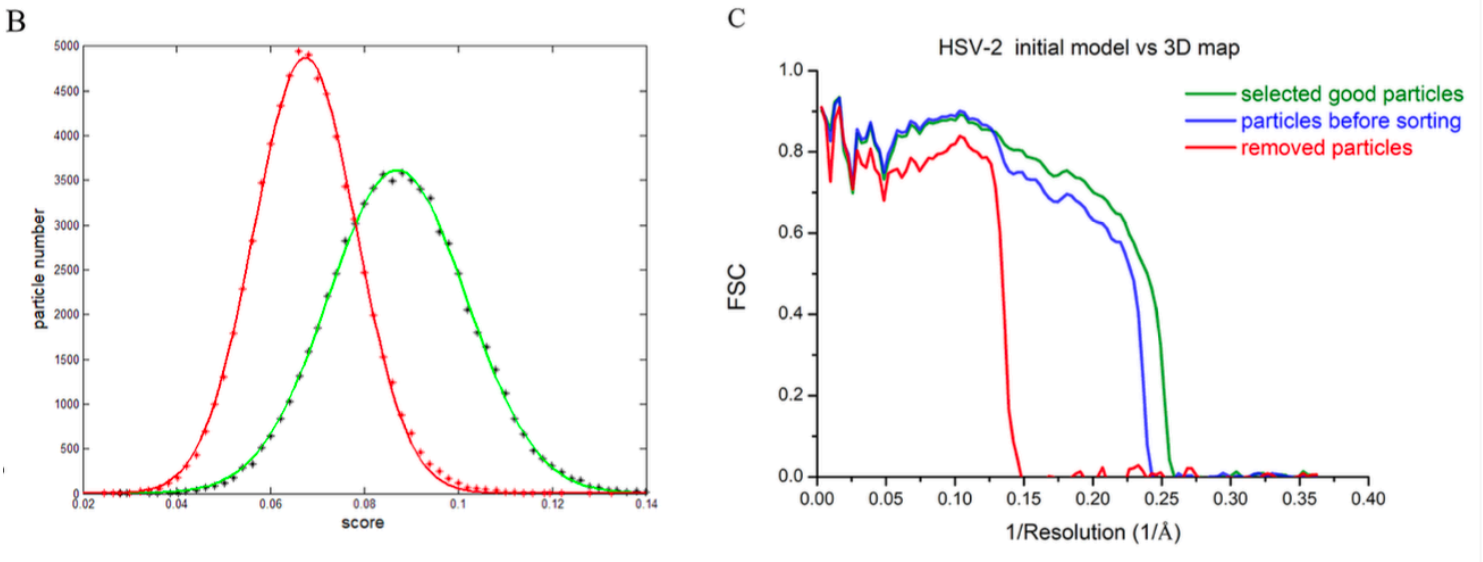
图中分别为：对原始挑选的阳性颗粒进行的三维分类，共分为十个类别，每个类别中正确率显示在左上角。两个选定的类别通过红色方框标出。对选定颗粒进行的三维分类，共分为三个类别，每个类别中正确率显示在左上角。

particle sorting

为了进一步降低假阳性结果的比例，计算了refine后有角度信息的raw颗粒和模板的phase residue分数，对颗粒进行了排序，分数定义如下：

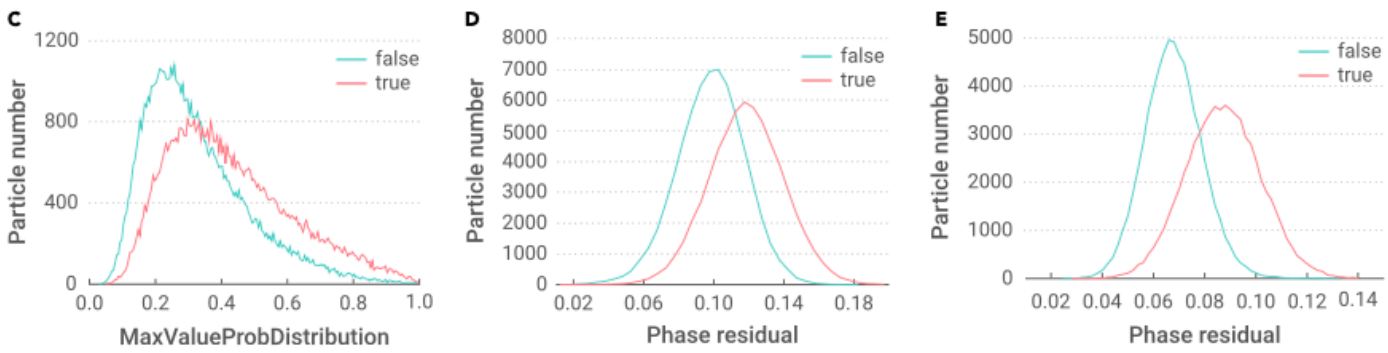
$$score = FSC \cdot CTF \cdot \cos(\Delta\phi)$$

分数中，频率范围从 8Å 到 5Å，排序把正确颗粒与假阳性颗粒通过两个类似高斯分布的峰分开，如图所示：



但是当排序基于从20到8Å的频率范围的particle picking的得分时，分离能力显著下降。所以可以认为，不使用particle picking的频率范围进行排序可以减少reference bias。

另外，在RELION中测试了用于排序颗粒的“MaxValueProbDistribution”的最大似然得分函数，然而，正确结果几乎无法与假阳性结果区分开来。根据“NrOfSignificantSamples”参数进行排序的结果与MaxValueProbDistribution的结果相似。因此，使用与优化中不同的得分进行排序也可能有助于减少reference bias。



图中为：
(C) 使用RELION中的MaxValueProbDistribution项的好坏颗粒分布。
(D) 使用频率范围从20 Å到8 Å的相位残差的好坏颗粒分布。
(E) 使用频率范围从8 Å到5 Å的相位残差的好坏颗粒分布。

Workflow of isSPA Method

(1) 通过用两个指数函数（一个用于噪声，另一个用于信号）拟合功率谱，估计信噪比（SSNR）参数；
(2) 对于颗粒检测，首先使用已知的SSNR参数通过优化的加权函数对原始图像进行滤波，然后将滤

波后的图像和三维模型的投影输入程序；
(3) 程序输出大量位置和方向数据，根据它们的平移和旋转对相邻的颗粒进行合并；
(4) 根据相关系数（cc）值对颗粒进行排序，并估计的每一部分中的颗粒数量；
(5) 使用脚本提取潜在颗粒，并利用不同的冷冻电镜软件包（如RELION等）对这些颗粒进行局部优化（local refinement）和三维分类；
(6) 使用排序算法以去除假阳性颗粒
(7) 对剩余的颗粒进行局部优化（local refinement），最终完成重构。