PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024, 51(10): 2418~2429
www.pibb.ac.cn

基于冷冻电镜无倾转成像数据的新型蛋白质原位结构解析方法*

赵明洁　曹端方　章新政**
（中国科学院生物物理研究所生物大分子重点实验室，北京 100101）

摘要　与体外纯化的蛋白质复合物相比，细胞内处于工作状态的蛋白质复合物往往更为完整，并且其三维结构处于完全生理的构象，这对于理解蛋白质复合物在生命活动中发挥重要功能的结构基础尤为关键，也可以为药物设计等提供更精确的靶点信息。直接在细胞内解析蛋白质复合物的三维结构也被称为蛋白质的细胞原位结构解析，而冷冻电镜电子断层重构技术是原位结构解析的关键技术，但是电子断层重构技术的序列倾转数据采集通量低、数据处理较为繁琐，并且达到准原子分辨率对样品有特殊要求，这些问题成为了限制原位结构解析分辨率和实际应用的瓶颈。近年来，一种基于单张无倾转数据的蛋白质原位结构解析方法发展迅速，可以高通量地对细胞中的蛋白质复合物进行高分辨率结构解析，本综述将分析这类方法的原理，讨论这个方法相较于传统断层重构方法的优缺点，并对蛋白质的细胞原位结构解析进行展望。希望可以通过这篇综述，帮助结构生物学科研人员更好地选择合适的工具。

关键词　蛋白质原位重构，冷冻电镜原位结构解析，isSPA滤波，无倾转数据，高通量
中图分类号　Q615　　**DOI**：10.16476/j.pibb.2024.0282

近十年以来，冷冻电镜作为结构生物学的主要工具之一得到了迅速发展，包括直接电子相机[1-2]，能量过滤器等硬件的成熟应用，还有基于最大似然法的自动化三维重构方法的建立[3-4]，使得纯化蛋白质的单颗粒三维结构解析技术基本成熟，实现了第一次冷冻电镜的分辨率革命[5]，许多蛋白质三维重构分辨率已经到达2~3 Å，基本满足蛋白质结构和功能的研究需求，同时也可以为药物设计提供较为精确的靶点信息。另一方面，现阶段，基于人工智能（artificial intelligence，AI）的蛋白质结构预测软件发展迅速[6-7]，简单蛋白质的三维结构预测较为精确，并已经具有一定的实用价值。然而，蛋白质复合物的纯化会使得蛋白质离开生理环境，纯化条件对于许多复合物来说并不友好，常常导致蛋白质复合物亚基的丢失，或者复合物之间失去相互作用。并且，和细胞内处于工作状态的蛋白质复合物相比，纯化的蛋白质复合物通常停止工作，因此其多种构象状态难以代表生理的、实现其功能的动态过程。所以需要避免蛋白质的纯化，直接在细胞内对蛋白质复合物的三维结构开展解析。而基于AI的蛋白质结构预测软件的训练集为纯化的蛋白质三维结构，对于复合物的构成和组装以及动态多构象的预测在现阶段仍然难以实现，因此还是需要通过实验手段在细胞内直接对蛋白质复合物开展高分辨率的三维结构解析，这也被称为蛋白质的原位结构解析。

实现蛋白质原位结构解析的主要方法是冷冻电镜技术。传统的方案主要分为三步，首先运用聚焦离子束减薄方法将通过高压冷冻或者投入式冷冻得到的冷冻细胞减薄成~150 nm厚的薄片，减薄是为了使300 kV加速电压的电子可以高效地穿透样品，实现较高质量的冷冻电镜成像。第二步是在薄片上找到感兴趣的区域，对其采集序列倾转数据，并进行断层重构[8]。由于电子束对样品的辐照损伤，完成断层重构的总电子计量在200 $e^-/Å^2$左右，其基于信噪比的分辨率往往低于5 nm[9-10]。因此为了进一步推进三维重构分辨率，需要第三步，在断层重构中找出目标蛋白质复合物，提取出对应的成像数

* 国家自然科学基金（32150010）资助项目。
** 通讯联系人。
Tel：18513916820，E-mail：xzzhang@ibp.ac.cn
收稿日期：2024-06-30，接受日期：2024-08-17

据（sub-tomogram），对其进行对中（alignment）和平均操作，以进一步提升数据信噪比（signal-to-noise ratio，SNR）和分辨率，该过程也被称为亚单位平均[11-12]，可以将一些丰度较高、分子质量较大的蛋白质复合物分辨率迅速推进至亚纳米，甚至4 Å以内的准原子分辨率[13-14]。

上述流程自身存在一些问题，限制了蛋白质原位结构解析的全面开展，同时也影响了三维重构分辨率。首先，为了获得玻璃态冰，尺寸大于20 μm的细胞或者组织样品通常需要高压冷冻，尽管不同的课题组正在优化该流程[15-17]，但该流程依然复杂，且容易失败。而投入式冷冻对于微栅上的细胞或者亚细胞器的厚度要求比较严格，太厚或者太多的细胞外缓冲液均容易导致晶体冰的生成。另外，在切片过程中聚焦离子束对切片上下两个表面产生损伤[18-21]，严格的定量分析发现[21]，这会对每个表面~50 nm深度内造成无法忽略的损伤。即便通过将Ga^+离子源的加速电压从传统的30 kV降低至8 kV，该损伤深度依然达到~30 nm。因此，为了得到蛋白质复合物的高质量成像，30 kV的Ga^+离子束切片需要将厚度控制在120~150 nm之间，而8 kV的Ga^+离子束则可以进一步减薄到~90 nm。断层重构方法可以大约在5 nm分辨率下对冷冻细胞切片进行三维重构和可视化，观察细胞器、膜结构以及一些蛋白质复合物。然而，为了实现断层重构，需要采集序列倾转数据，即在一个数据点上，通过倾转样品台，对冷冻细胞切片上的目标位置进行多次成像。相较于纯化蛋白质仅需一张无倾转成像的单颗粒分析方法（~10 000张/24 h），序列倾转数据的采集通量（50套/24 h）仅为其1/200。现阶段基于电子束平移的采集方式可以将序列倾转数据的采集通量提升至~200套/24 h[22]，但是该方法的速度提升依赖于冷冻细胞切片中感兴趣区域的面积，例如，研究对象若为普遍存在于所有细胞质的核糖体，则提升明显，若为细胞中丰度很低的中心体，则基本无法提升通量。数据采集效率的低下，使得断层重构常常需要数十天甚至更多的数据采集时间才能将三维重构推进至准原子分辨率，这是单一课题组难以实现的[23]。另外，序列倾转数据需要将有限的电子剂量分配到数十张倾转成像中，导致每一张曝光电子剂量极低，成像中的样品漂移难以矫正，此外，断层重构的对中较为粗略，误差较大，这些会非常严重地限制亚单位平均后的三维重构分辨率。现阶段，亚单位平均可以借助多颗粒系统对中等方案[23-24]有效减少上述蛋白质漂移和对中造成的误差，极大地提升三维重构分辨率。然而，该方案需要高丰度的均一蛋白质实现多颗粒系统对中，而许多感兴趣蛋白质复合物所在区域并不具备该条件，这个要求很大地限制了该方法的实际应用。

针对断层重构中的数据采集通量极低，以及断层重构中误差较大的问题，近年来发展了一种基于无倾转冷冻电镜照片的蛋白质原位结构解析方法。该方法无需采集繁琐的序列倾转数据，无需断层重构，避免了传统基于断层重构的方法通量低、重构误差大的问题，可以实现高通量、高分辨率的蛋白质原位结构解析。这个方法的完整流程isSPA（*in situ* Single-Particle like Analysis）首次由Cheng等[25]在2020年提出，该方法考虑到原位数据生物大分子密度叠合的特点，去优化模版匹配算法，通过运用中等分辨率三维模型的模版匹配定位目标蛋白质，并通过交叉验证和排序算法去除假阳性数据，消除模型偏差问题[26-27]，实现对目标蛋白质复合物在细胞原位近原子分辨率的三维重构。之后，Lucas等[28]于2021年也发展了一种基于无倾转冷冻电镜照片的流程2DTM（2D template matching）。接下来将对其原理、适用场景、两种方法的原理差异以及较断层重构的优缺点等展开介绍，并对蛋白质原位结构解析的未来进行展望。

1　蛋白质在细胞原位或者纯化后的冷冻电镜成像的差异

对于纯化的蛋白质，一般通过投入式冷冻的制样方式将一定浓度的单一蛋白质包埋到厚度略大于蛋白质直径的玻璃态冰中。在制样过程中，通过控制蛋白质的浓度，使成像中蛋白质分布均匀，要求蛋白质足够密，但是每个蛋白质的成像相互独立，互不干扰，肉眼可见每一个蛋白质分子（图1a）。而在细胞中，蛋白质种类复杂，分布非常密集。将冷冻细胞减薄成~150 nm的切片后，该厚度远大于一般的蛋白质复合物尺寸，因此，对冷冻细胞切片进行成像，其和纯化蛋白质的成像数据相比，最大的差异是，几乎每一个蛋白质分子的成像都和其他生物大分子的成像叠合在一起，肉眼难以区分蛋白质（图1b）。

在冷冻电镜数据中，电子束对蛋白质样品的辐照损伤导致可用于成像的总电子剂量有限（单张照片一般低于60 $e^-/Å^2$）。类似于晚上拍摄曝光时间不

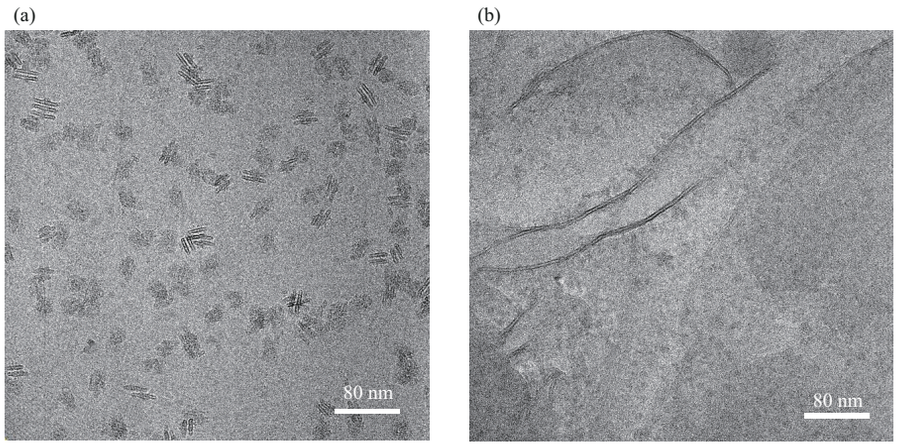


Fig. 1　Images of single particle data (a) and *in situ* data (b)

图1　单颗粒数据（a）和原位数据（b）的图像示例

（a）豌豆叶肉细胞提纯的$C_2S_2M_2$蛋白单颗粒数据；（b）豌豆叶肉细胞提纯叶绿体后的切片数据直接成像。

够长的照片，光子剂量不足导致照片中有大量噪音，从冷冻电镜获得的蛋白质成像中含有大量的噪音，使得数据的信噪比很低。然而，在基于直接电子探测相机电子计数算法得到的成像中[29]，由于曝光剂量不足导致的散粒噪音近似为白噪音，在功率谱中白噪音的振幅随着频率基本不变。图2a是一张冷冻电镜照片的功率谱，电子穿过样品时，被样品散射，这些散射信号在电镜中传递会被衬度传递函数（contrast transfer function，CTF）调制，因此散射信号的功率谱会随着频率震荡。成像中每一个像素的灰度都是由CTF调制的信号和白噪音两部分组成，因此功率谱也由两部分组成，振幅随着频率基本不变的白噪音以及振幅随着频率震荡的散射信号。如果认为散射信号大部分由蛋白质散射导致，那么数据的信噪比就是散射信号和白噪音之比。由此计算的信噪比（图2b）和一般认为冷冻电镜成像数据信噪比很低不一样，基于频率的信噪比在低频区域大于1，只有在中高频区域，数据的

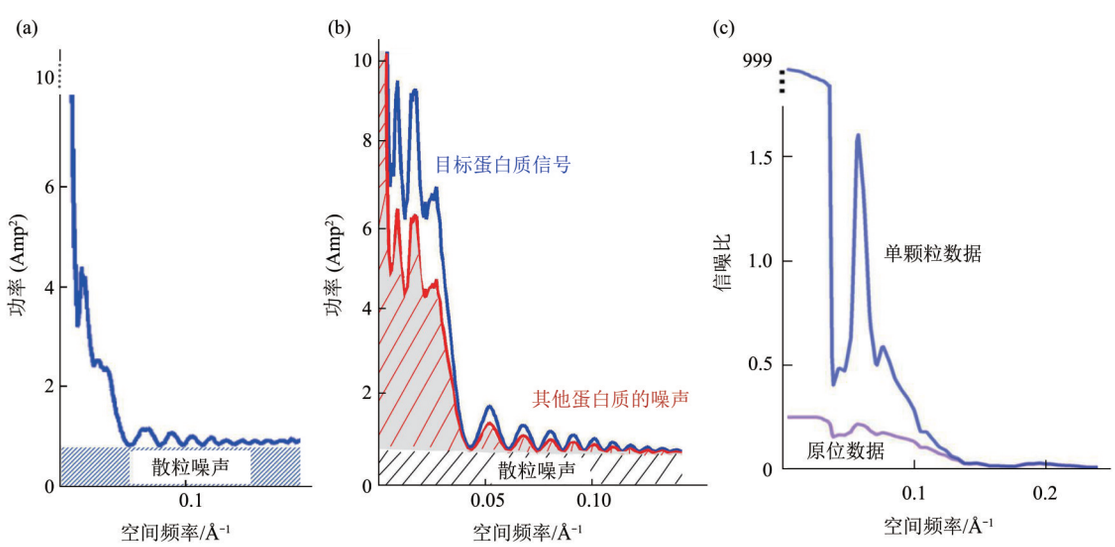


Fig. 2　Signal characteristics of cryo−EM single particle data and *in situ* data

图2　冷冻电镜单颗粒数据和原位数据信号特点

（a）采用电子计数算法得到的冷冻电镜单颗粒数据的径向平均功率谱，阴影部分是散粒噪声，其功率随频率几乎不变。信号经由衬度传递函数（contrast transfer function，CTF）调制后随频率震荡。（b）采用电子计数算法得到的冷冻电镜原位图像，其中的目标蛋白质和非目标蛋白质的径向平均功率谱。图中底部黑色阴影区域为散粒噪声，红色阴影区域为非目标蛋白质信号，在原位图像目标蛋白质的探测中，它作为一种蛋白质噪声，蓝色阴影区域合并上红色阴影区为目标蛋白质信号。（c）由图可知，单颗粒数据的信噪比（signal-to-noise ratio，SNR）在低频处很高；但在原位数据中，非目标蛋白质的叠合大约为目标蛋白质的3倍时，得到的目标蛋白质原位数据SNR。

信噪比才会远小于1。这是为什么对冷冻电镜数据进行低通滤波后，去除掉低信噪比的中高频信号，仅保留高信噪比的低频信号，就可以清晰地看清蛋白质的轮廓。低频区域信噪比高的原因是原子对电子的散射本身就是低频段散射强，而高频段散射弱。

然而在冷冻细胞切片的成像数据中，蛋白质成像相互叠合、相互干扰。对于期望重构的目标蛋白质复合物，在计算信噪比时，将与目标成像叠合的其他成像当做是噪音处理。因此，在原位冷冻电镜数据中，目标蛋白质的成像是信号，其他叠合蛋白质成像以及散粒噪音为噪音。与纯化蛋白质成像数据的信噪比相比（图2c），该蛋白质的原位数据在中低频的信噪比大幅下降，这也是为什么经过低通滤波后，原位数据中还是看不见单个蛋白质的原因。其本质原因是将叠合的非目标蛋白质当做噪音时，和散粒噪音不同，其低频信号振幅很强，会显著破坏目标蛋白质的低频信噪比。而非目标蛋白质的高频信号则很弱，远低于散粒噪音，因此叠合蛋白质成像时，其对目标蛋白质的中高频信号的信噪比几乎不影响。

2　传统的冷冻电镜数据处理方法依赖于高信噪比的低频信号

高信噪比的低频信号在冷冻电镜三维重构，包括单颗粒分析以及断层重构的亚单位平均中都非常重要。首先是颗粒挑选，无论是手动挑选，基于高斯球或者模版匹配的自动挑选，还是基于深度学习的自动挑选[30-33]，这些方法均依赖高信噪比的低频信号。颗粒挑选最终提供蛋白质颗粒的中心信息，而中心信息是基于迭代的二维/三维分类以及三维结构优化的起点，是必需的。在似然函数中，低频信号提供的蛋白质中心信息因为其准确率高，还被当做先验信息，用于计算符合先验的概率[34]。另外，当前的冷冻电镜三维重构包括单颗粒分析或者断层重构亚单位平均，它们均依赖于一个基于初始模型的迭代算法。迭代算法的基本工作原理是结构优化（refinement）的早期仅用低频信号参与计算，随着三维模型分辨率的缓步提升，慢慢地让更高频率的信号参与计算。在上述迭代算法的计算初期，单独使用低频信号可以得到正确结构的原因是低频信号具有高信噪比，可以准确地展示出蛋白质中心和轮廓信息。

因此，对于纯化的蛋白质，单颗粒冷冻制样的要求之一，就是蛋白质在非晶冰中有一定的分散度，使其在数据处理过程中（图3a），周围蛋白质的信号尽量不干扰目标蛋白质。这也是单颗粒分析这个名字的来由，即数据分析中只能包含一个蛋白质颗粒的成像信息。而在细胞切片的二维成像数据中，蛋白质成像相互叠合，无法做到互不干扰。因此，运用断层重构，将叠合的蛋白质密度在三维重构中分开，使其相互独立（图3b），恢复了低频数据的信噪比。在断层重构中，单个蛋白质的三维信号可以独立出来，用于颗粒挑选和迭代计算等。

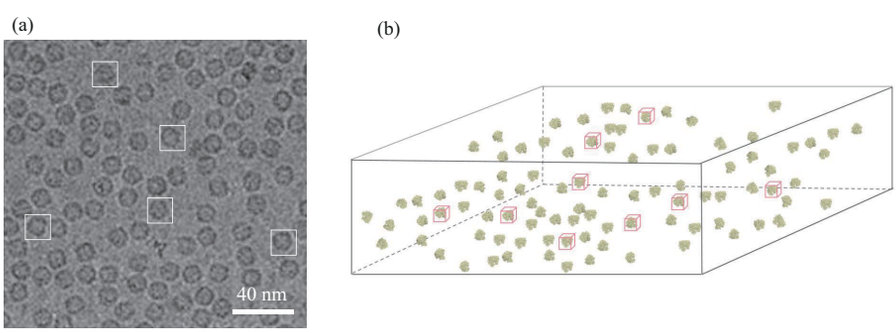


Fig. 3　Extract particles from images

图3　从图像中拾取颗粒

（a）从二维单颗粒图像中提取的有代表性的颗粒，颗粒之间不叠合（只提取4个颗粒作为示意）；（b）从三维重构后的体中提取颗粒，颗粒之间存在叠合，但从三维空间中可以把单个三维的蛋白质信号提取出来（仅提取个别颗粒作为示意）。

除了蛋白质信号相互独立的纯化蛋白质的二维成像数据和蛋白质信号相互干扰的细胞原位二维成像数据以外，有些生物样品（例如含有刺突蛋白的病毒颗粒[35-36]，嵌入在脂质体上的具有较明显可溶

结构域的膜蛋白[37-38]以及嵌入在线粒体膜上的线粒体复合物[39]等）的成像数据中，蛋白质具有部分叠合的性质，传统上是运用断层重构和亚单位平均对这些蛋白质进行高分辨率结构解析。而上述蛋白质有个特点，均为膜蛋白，并且具有较大的、特征明显的可溶性结构域。这些可溶性结构域的侧视图往往位于膜结构的一侧，可溶性结构域的成像相对较为独立。因此，可以基于这些具有较为独立成像信号的可溶性区域进行颗粒挑选并实现基于迭代的计算，而无需断层重构的参与。

当成像数据相互叠合时，例如细胞原位的成像，或者纯化蛋白质冷冻制样过于密集，绝大部分蛋白质叠合时，不仅无法用肉眼或者低频信号区分蛋白质，获得蛋白质的中心信息，而且从低频往高频逐渐收敛的二维分类和三维分类也丧失作用，这是由于仅仅由信噪比被破坏后的低频信号参与计算，会得到错误的结果。这里，我们在细胞冷冻切片的成像数据中已知一部分目标蛋白质（核糖体大亚基和C_2S_2捕光复合物）的中心信息，将这些目标蛋白质数据导出（extract），进行二维分类，可见二维分类失效（图4）。

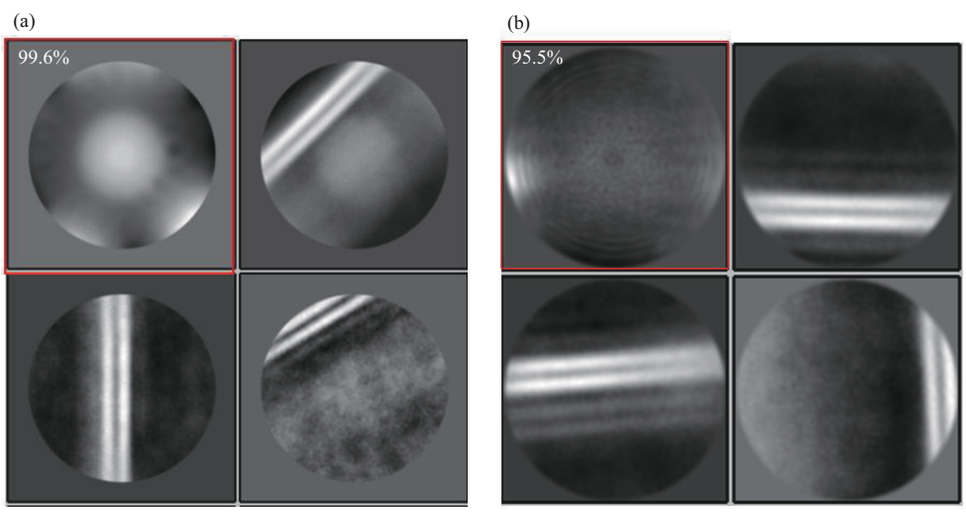


Fig. 4　2D Classification results of ribosome (a) and C_2S_2 (b)

图4　核糖体（a）和C_2S_2（b）的二维分类结果

（a，b）两图分别代表酿酒酵母细胞核糖体（a）和豌豆叶肉细胞叶绿体（b）的C_2S_2蛋白，在已知正确坐标后提取颗粒并进行二维分类的结果。两图左上角的数字代表二维分类后左上角那一类的颗粒分布比例。

3　在低频信号被破坏的细胞切片无倾转二维成像数据中直接进行三维重构

Cheng等[40-41]建议冷冻电镜采集细胞切片无倾转成像数据时，实际上需要倾转样品台一定角度，这是由于聚焦离子束制备的切片一般和微栅平面有10°左右的夹角，需要通过转动样品台以保证聚焦离子束切片和电子束垂直，这样可以保证冷冻电镜照片中对应切片的欠焦值变化最小，使得成像具有较强的索恩环（Thon ring），以便于欠焦值的测定和成像数据质量的判断。而Elferich等[42]和Lucas等[43]提出在样品台0°倾转的情况下，对具有一定倾斜角的切片直接采集数据的方案。

上面已经提到了在细胞切片的二维成像中，由于目标蛋白质复合物的成像与细胞中其他生物大分子的成像相互叠合，干扰了目标蛋白质复合物的低频信号，无法观察获得其中心和轮廓信息。除了观察以外，也可以运用基于关联函数的二维模版匹配方法来尝试定位蛋白质。如图5a所示，关联函数谱的峰值被噪音覆盖。这是由于通过成像数据直接计算关联函数谱时，冷冻电镜成像中低频信号所占比例较高，而传统的关联函数计算过多的依赖于低频信号，但低频信号被叠合密度破坏，这导致传统的关联函数谱对密度叠合非常敏感，使得其在细胞切片成像数据中对目标蛋白质的定位失效。

Cheng等[26]针对上述密度叠合的问题，发展了新型关联函数。他们认为，尽管中低频信号由于密度叠合导致信噪比下降明显，但是散粒噪音在中

高频信号中依然是最为主要的噪音来源，中高频信号随频率升高，信噪比迅速下降的特点无法忽略。因此，在关联函数的计算中需要严格控制信噪比极低的中高频信号的权重，因为过于依赖中高频信号，会出现严重的模型依赖问题（model bias problem）。因此，Cheng等[26]基于原位数据的特点，将叠合密度作为噪音，并兼顾散粒噪音，提出了基于信噪比谱（目标蛋白质作为信号，考虑叠合密度噪音和散粒噪音）的权重函数，作用于关联函数，用于定位叠合密度中的目标蛋白质（图5b）。当该关联函数计算得到较高的相关系数时（超过特定阈值），可以认为，在冷冻电镜照片的该位置上，存在一个目标蛋白质，取向和二维模版的取向一致。该方案在定位能力上有较为显著的提升，并且在一定程度上限制了定位完全错误的假阳性数据。

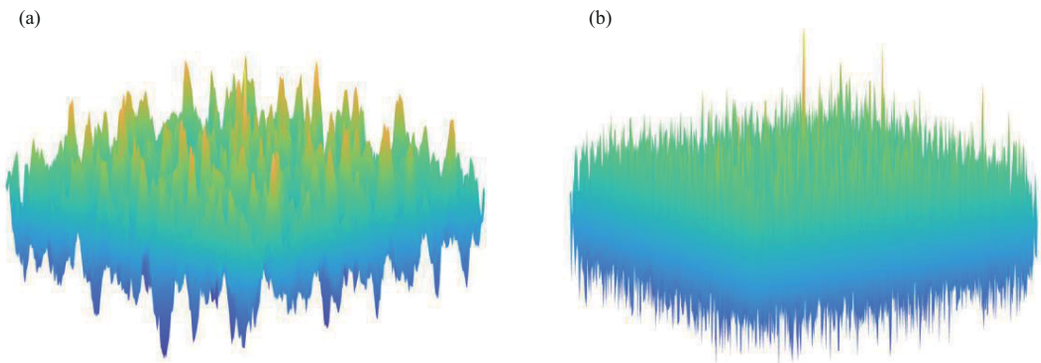


Fig. 5　CCG without isSPA weighting (a) and with isSPA weighting (b)

图5　未加isSPA滤波的原位图像与模版的互相关峰形图（a）与施加isSPA滤波后的原位图像与模版的互相关峰形图（b）

（a）互相关峰形图CCG中没有一个显著的高峰，是由于原位图像中大量噪声的影响，淹没了原本正确的峰。即通过该CCG计算无法确定出一个蛋白质取向定位；（b）施加isSPA滤波后的原位图像与模版再次做CCG，从图中可以明显看到一个高峰，高峰所在位置即目标蛋白质的取向及定位。CCG：互相关图（cross-correlation gram）；isSPA：原位单颗粒分析（*in situ* single particle like analysis）。

另一方面，Cheng等[25-27]在运用新关联函数定位了目标蛋白质后，即获得一个目标蛋白质数据集，其中不仅包括正确的目标蛋白质颗粒的成像，还含有错误的假阳性数据。可以看到，随着阈值的减小，获得定位的目标蛋白质比例（recall）增高，但同时假阳性数据的比例也迅速增高，从而目标蛋白质占该阈值以上定位总数的比例（precision）降低。无论是正确还是错误，目标蛋白质数据集中包含了每一个数据点的中心和投影角度信息。将这些信息输入到基于最大似然法的计算软件中，运用最大似然函数，对每一个数据点进行无对中（skip alignment）的三维分类。该三维分类过程可以认为是借助最大似然函数对目标蛋白质数据集中的每一个数据点进行打分，对数据点进行交叉验证，而交叉验证过程可以大幅减少数据集中的假阳性数据比例。需要指出的是，如果在三维分类中进行对中操作，如前所述，由于迭代计算的初期更多地依赖于低频信号，这使得基于最大似然函数的计算往往难以收敛到正确的解上，因此，需要避免对中操作。

通过交叉验证尽量获得正确率高的目标蛋白质数据集，之后可以在基于最大似然函数的三维重构软件中对该数据集进行三维重构的迭代优化，并且Cheng等[25, 27]在迭代优化的过程中又添加了一步排序过程（将在后面展开描述），进一步减少假阳性数据。和上述三维分类类似的，通常基于迭代的三维重构优化过程的初期更多地依赖低频信号，这在蛋白质的原位成像数据中需要避免，因此我们放入一个~10 Å的三维初始模型，通过设置，跳过全局搜索步骤，直接开始局域搜索进行三维重构的优化。而在原位数据中，蛋白质可以通过基于最大似然法的局域搜索进行三维重构优化的原理是，叠合在目标蛋白质上的生物大分子噪音的中高分辨率信号满足高斯分布。因此，中高频率段的叠合密度噪音和散粒噪音可以合并为一种高斯分布的噪音。

而Lucas等[28]发展了2DTM，他们基于之前发展的白化滤波（signal whitening）方案[44]，在传统定位函数仅运用低频信号的基础上，引入中高频信号参与关联函数的计算，并调低了低频信号在关

联函数中所占的权重。Lucas等[43]认为，当阈值足够高时，可以排除假阳性数据。这也是该课题组在后期发展2DTM时用的策略，利用高阈值减少假阳性数据。因为，2DTM是基于高阈值、低假阳性数据的方案，他们在2DTM的后继计算中[43]，没有交叉验证以及其他排除假阳性数据的手段。

4 定位目标蛋白质的三维模版的获得

上述通过模版匹配在细胞原位成像数据中定位目标蛋白质的这些方法需要一个中高分辨率的二维模版，然而所定位的目标蛋白质在成像时的取向几乎处于随机分布状态，因此需要目标蛋白质在各个投影方向的二维模版。通常需要获得一个目标蛋白质中高分辨率的三维重构，对其各个方向投影，生成数千个二维模版，然后进行模版匹配计算。需要指出的是，实际上通常并没有这个最佳的模版，作为最佳模版的三维重构正是目前缺乏的，并希望通过计算得到的。

由于原位结构解析通常是去探索纯化蛋白质无法获得的更为完整的复合物，因此往往复合物中的一部分可能通过纯化的方法，已经获得三维结构的解析。那这部分三维重构就可作为三维模版。例如，红藻的藻胆体（phycobilisome，PBS）已经用单颗粒方法解析至3.5 Å[45]，将该藻胆体的冷冻电镜三维重构作为模版，就可以在红藻细胞的冷冻切片中定位藻胆体。而红藻细胞中的藻胆体和光系统II以及光系统I形成了超级复合物，该定位过程实际上定位了藻胆体-光系统II-光系统I超级复合物，可以实现对其的准原子分辨率三维重构。

另一个常用的方法是对细胞切片数据先进行断层重构和亚单位平均。该过程一般可以较为便利地获得中等分辨率三维重构，可以用该三维重构作为模版，对目标蛋白质复合物进行定位，例如，在You等[41]的工作中，两类藻胆体超级蛋白质复合物的结构解析就是分别其运用了来自断层重构和亚单位平均获得的三维重构作为模版。随后，Zhang等[46]又讨论了以亚单位平均生成模版进行GisSPA颗粒探测的优势。

此外，三维结构非常类似的同源蛋白质复合物的三维重构也可以用作模版。这里相较于蛋白质的同源性，三维结构类似是更为重要的概念。其中三维结构的类似程度可以通过计算同源蛋白质三维结构之间的傅里叶球壳相关系数（Fourier shell coefficient，*FSC*）来测定。这里，三维结构非常类似定义为*FSC*在8 Å附近具有较为显著的值（>0.3）。

另外，三维模版也可以通过蛋白质的原子模型（PDB）生成三维重构密度图来获得。生成密度图的工具可以使用EMAN2软件包[47]中的pdb2mrc脚本计算得到的密度图再外加Bfactor来生成，也可以通过cisTEM软件[48]中内置的功能生成模拟三维模版来完成。Lucas等[43]就运用了PDB原子模型生成的大肠杆菌β半乳糖苷酶以及枯草芽孢杆菌核糖体大亚基的三维重构密度图，分别对其进行了蛋白质的定位分析。

三维模版的好坏会影响定位目标蛋白质的能力，好的三维模版和目标蛋白质的实际结构完全一致。三维模版好坏由两者之间的相似程度定义，同样可以通过计算*FSC*来判断。如复合物的一部分作为模版，若没有其他的构象变化，则定义分辨率的FSC曲线可用于判断在什么频率段，三维模版和实际结构几乎完全一致。在亚单位平均中，定义分辨率的FSC曲线，以及同源蛋白质之间的FSC起到类似的作用。而原子模型生成的三维重构密度图也需要和实际的三维重构密度图计算两者之间的*FSC*，可以看到该*FSC*比较特殊（图6），在0频附近就跌至0.6。这意味着目前的原子模型转换成三维重构密度的方法尚未成熟，会在中低频率段和真实的三维重构不符，产生较大的失真。这种类型的失真会显著降低定位目标蛋白质的能力，因此，该方法生成的模版并不是一类理想的模版。

在Cheng等[25-27]的目标蛋白质定位计算中，鼓励运用冷冻电镜三维重构作为模版，尽量避免PDB转换生成的模版，计算频率一般延伸至8 $Å^{-1}$左右。这是考虑到，一方面，加入在高频的信号会极大地增加计算量，另一方面，由于高频信号信噪比极低，在信噪比加权之后，对关联函数的贡献非常低。实际测试中也发现，进一步引入高频信号，几乎难以再提升定位能力。而在Lucas等[28]的目标蛋白质定位计算中，倾向于运用PDB生成高分辨率的模版，参与计算的频率段延伸至4 $Å^{-1}$及更高频率。由于高分辨率信息对欠焦量非常敏感，除了运用基于Thon ring测得的欠焦量以外，需要考虑目标蛋白质在细胞切片中的分布导致的单个蛋白质复合物欠焦量的涨落。另外，高分辨率信号参与计算提升了对模版的要求，需要模版和目标蛋白质的实际重构在4 Å附近有较高的相似度。

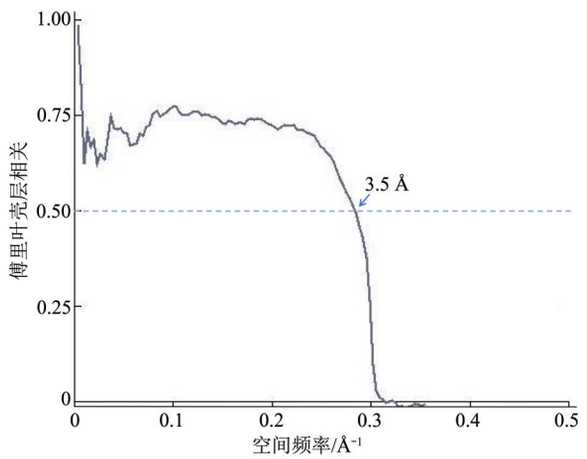


Fig. 6　The FSC curve between the map derived from the PDB and the map reconstructed from cryo−EM

图6　PDB转换来的密度图和冷冻电镜三维重构的密度图之间的傅里叶壳层相关（FSC）曲线

使用PDB 4v19模型和对应的EMD-2787的核糖体数据两者之间做FSC，使用的该模型为3.4 Å，因此*FSC*=0.5时，对应着分辨率3.5 Å。由此可见，在0频附近，*FSC*衰减到接近0.6。PDB：蛋白质数据库（protein data bank）；*FSC*：傅里叶壳层相关系数（fourier shell coefficient）。

5　三维重构中的模型偏差问题

上述方法提到，在冷冻电镜原位数据中，信噪比很低的中高分辨率信号参与模版匹配，随机噪音在偶然情况下会和模版产生较强的匹配，并且大量的噪音提升了偶然匹配的概率，导致最终定位到错误的生物大分子或者其他信号上。可以肯定的是，虽然是错误的定位，但是该位置的成像（假阳性数据）和模版是有较高匹配度的。当模版匹配运用到高至8 $Å^{-1}$的频率段时，假阳性数据在匹配频率段与模版相符，但是在更高的频率段，依然是随机噪音的形式存在。而对于定位到目标蛋白质上的数据（正确数据），不仅在匹配的频率段相符，在更高频率段也存在正确的成像信号。实际数据的计算发现，模型偏差问题主要发生在进行匹配计算的频率段，假阳性数据的存在使得三维重构该频率段的*FSC*值偏高。另外，假阳性数据会妨碍正确数据的最大似然方法对中，导致正确数据三维重构更高频率段的*FSC*偏低，进而降低三维重构分辨率。因此，假阳性数据的存在并不会导致三维重构更高频率段的*FSC*由于模型偏差问题而偏高，因此这部分*FSC*准确地定义了该频率段三维重构的信噪比，进而准确地定义了该重构的分辨率。由于这些频率段的信号来自于正确数据本身，而假阳性数据在这些频率段未经匹配、挑选，依然为随机噪音，和三维重构的相关性低，Cheng等[25, 27]发展了基于该频率段的排序算法，剔除假阳性数据。另一方面，如果模版匹配运用到高频率信号，频率范围接近甚至超过三维重构最终分辨率，那么假阳性数据的存在会使得*FSC*偏高，导致基于FSC的分辨率估算受模型偏差的影响而偏高。在这种情况下，很难直接判断出模型偏差对三维重构的影响到底有多严重。

另一个更为实用的判断模型偏差问题的标准是，当模版仅为复合物的一部分时，复合物中非模版区域的密度图质量是否和模版类似。例如，在You等[41]的工作中，模版为纯化PBS的单颗粒三维重构，解析得到的PBS-LHCII-LHCI超级复合物中，LHCII、LHCI以及PBS中新增的亚基，其三维重构密度图不受模型偏差的影响[41]。如果复合物是通过刚性的方式结合，一个好的三维重构要求这些亚基的分辨率或者密度图质量应该和PBS类似。因为细胞原位的蛋白质复合物往往和纯化的蛋白质复合物在构象或者亚基上有所差异，上述的判断方法应用性比较广泛。相反，如果模版对应的三维重构分辨率极高，而其他部分非常低，例如在Lucas等[28]的工作中，核糖体的大亚基作为模版，重构分辨率达到4 Å以内，而对应的小亚基部分则仅为15 Å左右，该论文的审稿意见指出，这个三维重构可能存在模型偏差问题。另一方面，如果三维重构分辨率在匹配频率段附近，而且三维重构没有出现模版以外的额外蛋白质亚基，Lucas等[43]展示，可以尝试在模版中删除一小部分密度（例如，氨基酸的一个侧链），当三维重构基本不受模型偏差问题影响时，模版中删除的密度会在三维重构中很好地恢复。

6　新方法的优缺点以及对未来的展望

断层重构本身是对细胞切片的一个三维重构，分辨率一般低于5 nm，可以清晰地观察到作为细胞器关键部分的各种膜结构的三维形貌，以及蛋白质复合物在细胞切片中的分布等。因此，断层重构的一个巨大优势在于对细胞切片进行相对较低分辨率的观察，这是原位研究的第一个目的。这种观察，一般期望观察范围尽可能越大越好，也就是在不影响成像质量的情况下，切片尽可能厚，一般200~300 nm的切片可以较好地满足观察的需求。

而蛋白质复合物的三维结构细节的获得需要更高的分辨率，一般需要4 Å以内（准原子分辨率），才可以搭建出蛋白质复合物的原子模型，这是原位研究的第二个目的，也就是蛋白质三维结构的原位解析。这个解析过程其实非常依赖切片厚度，切片厚度变薄可以大大降低结构解析的难度。因此，在聚焦离子束减薄样品的实验中，需要明确实验目的，若以观察为主要目的，可以将切片厚度维持在200~300 nm，而以蛋白质结构解析为目的，则尽量将切片厚度控制在100~150 nm，甚至更薄。值得注意的是，高质量的薄切片需要较低的离子加速电压[21]。

基于无倾转成像数据的原位结构解析方法因为没有断层重构，无法满足观察的目的，因此，仅仅能满足上述第二个目的，即对蛋白质复合物的三维结构开展解析。值得指出的是，这个方法也可以通过单复合物欠焦量的优化或者定位中欠焦量的搜索给出蛋白质复合物的三维定位信息，但是欠焦量的精度（*z*方向的精度）难以优于10 nm，远差于断层重构中复合物定位精度（可以达到1 nm以下）。因此，新方法在细胞观察和蛋白质的三维定位精度上较断层重构均有较大劣势。幸运的是，进行观察或者研究蛋白质在细胞中的定位无需大量数据，一般而言，十套左右的断层重构数据足以满足这两个要求。

新型原位结构解析方法的优点在于蛋白质原位结构的解析上，仅需采集无倾转数据，数据采集通量较断层重构提升近100倍。数据处理时，仅需对成像进行传统的平移矫正[49]，无需复杂、费时且对样品有较高要求的多颗粒系统对中方法。因此可以在较断层重构很短的时间内，获得大量的蛋白质成像数据，实现准原子分辨率的结构解析。两种技术方法的有机结合，可以既做到观察又做到结构解析。

经过了数十年的发展，基于断层重构的蛋白质原位结构解析方法在数据采集、数据质量和数据处理等方面都有了较大的进展。然而，获得一个近原子分辨率的蛋白质原位三维重构依然困难重重。现阶段，仅仅可以对细胞内丰度非常高或者局域丰度非常高亦或者具有很高对称性的蛋白质，例如核糖体、肌丝、鞭毛、二十面体病毒等展开高分辨率或者近原子分辨率的结构解析[13-14, 50-52]，而且蛋白质的分子质量一般需要约1 Mu或者更高。即便丰度高如核糖体，运用断层重构，在近原子分辨率研究其动态多构象，依然需要单个课题组无法轻易得到的巨大数据量，这使得现阶段动态多构象的原位研究由于数据的缺乏，平均分辨率往往限制在7 Å左右。而新型原位结构解析方法的开发，将数据采集通量提升近100倍，这将缓解原位结构解析对蛋白质丰度的要求。然而，真正低丰度的蛋白质复合物在冷冻细胞切片上的数目会很低，聚焦离子束减薄获得切片的效率也很低，因此，获得低丰度蛋白质复合物的原位成像数据的一个限制因素是切片的效率。另外，原位结构解析对目标蛋白质分子质量的要求是由于数据质量较低导致的，一方面聚焦离子束对冷冻细胞切片存在损伤，较大地降低了两个切面50 nm以内的蛋白质复合物的成像质量，另一方面，切片的厚度使得成像产生大量非弹性电子，降低数据的信噪比，而且叠合的生物大分子密度也会较大地影响目标蛋白质成像的对中。

综上，现阶段的原位结构解析技术在丰度较低（一个切片中少于100个蛋白质复合物）、分子质量稍小（<1 Mu）的蛋白质复合物上难以达到准原子分辨率，而真正的冷冻电镜第二个分辨率革命的发生，原位结构解析的时代来临，需要解决这个问题。而我们认为，高效率、高质量薄切片的获得将极大地降低对蛋白质复合物丰度和分子质量的需求。因此，发展一个可以获得高质量冷冻细胞薄切片的技术，将是蛋白质原位结构解析技术获得全面应用的关键。

参 考 文 献

[1] Faruqi A R, Henderson R, Pryddetch M, *et al*. Direct single electron detection with a CMOS detector for electron microscopy. Nucl Instrum Meth Phys Res Sect A Accel Spectrometers Detect Assoc Equip, 2005, **546**(1/2): 170-175

[2] McMullan G, Faruqi A R, Henderson R. Direct electron detectors. Methods Enzymol, 2016, **579**: 1-17

[3] Zivanov J, Nakane T, Forsberg B O, *et al*. New tools for automated high-resolution cryo-EM structure determination in RELION-3. Elife, 2018, **7**: e42166

[4] Punjani A, Rubinstein J L, Fleet D J, *et al*. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods, 2017, **14**(3): 290-296

[5] Kühlbrandt W. The resolution revolution. Science, 2014, **343**: 1443-1444

[6] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, **596**(7873): 583-589

[7] Abramson J, Adler J, Dunger J, *et al*. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024,

630(8016): 493-500

[8] Mastronarde D N, Held S R. Automated tilt series alignment and tomographic reconstruction in IMOD. J Struct Biol, 2017, **197**(2): 102-113

[9] Pawel A. Penczek J F. Resolution in electron tomography// Joachim F. Electron tomography: methods for three-dimensional visualization of structures in the cell. New York: Springer, 2006, 307-330

[10] Cardone G, Grünewald K, Steven A C. A resolution criterion for electron tomography based on cross-validation. J Struct Biol, 2005, **151**(2): 117-129

[11] Wan W, Briggs J A G. Cryo-electron tomography and subtomogram averaging. Methods Enzymol, 2016, **579**: 329-367

[12] Chen M, Bell J M, Shi X, *et al*. A complete data processing workflow for cryo-ET and subtomogram averaging. Nat Methods, 2019, **16**(11): 1161-1168

[13] Hoffmann P C, Kreysing J P, Khusainov I, *et al*. Structures of the eukaryotic ribosome and its translational states *in situ*. Nat Commun, 2022, **13**(1): 7435

[14] Xue L, Lenz S, Zimmermann-Kogadeeva M, *et al*. Visualizing translation dynamics at atomic detail inside a bacterial cell. Nature, 2022, **610**(7930): 205-211

[15] Wang S, Zhou H, Chen W, *et al*. CryoFIB milling large tissue samples for cryo-electron tomography. Sci Rep, 2023, **13**(1): 5879

[16] Wu Y, Qin C, Du W, *et al*. A practical multicellular sample preparation pipeline broadens the application of *in situ* cryo-electron tomography. J Struct Biol, 2023, **215**(3): 107971

[17] Kelley K, Raczkowski A M, Klykov O, *et al*. Waffle Method: a general and flexible approach for improving throughput in FIB-milling. Nat Commun, 2022, **13**(1): 1857

[18] Berger C, Dumoux M, Glen T, *et al*. Plasma FIB milling for the determination of structures *in situ*. Nat Commun, 2023, **14**(1): 629

[19] Martynowycz M W, Shiriaeva A, Clabbers M T B, *et al*. A robust approach for MicroED sample preparation of lipidic cubic phase embedded membrane protein crystals. Nat Commun, 2023, **14**(1): 1086

[20] Lucas B A, Grigorieff N. Quantification of gallium cryo-FIB milling damage in biological lamellae. Proc Natl Acad Sci USA, 2023, **120**(23): e2301852120

[21] Yang Q, Wu C, Zhu D, *et al*. The reduction of FIB damage on cryo-lamella by lowering energy of ion beam revealed by a quantitative analysis. Structure, 2023, **31**(10): 1275-1281.e4

[22] Eisenstein F, Yanagisawa H, Kashihara H, *et al*. Parallel cryo electron tomography on *in situ* lamellae. Nat Methods, 2023, **20**(1): 131-138

[23] Tegunov D, Xue L, Dienemann C, *et al*. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. Nat Methods, 2021, **18**(2): 186-193

[24] Himes B A, Zhang P. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nat Methods, 2018, **15**(11): 955-961

[25] Cheng J, Li B, Si L, *et al*. *In situ* structure determination using single particle cryo-electron microscopy images. bioRxiv, 2020. DOI:10.1101/2020.09.04.282509

[26] Cheng J, Zhang X. Optimizing weighting functions for cryo-electron microscopy. Biophys Rep, 2021, **7**(2): 152-158

[27] Cheng J, Li B, Si L, *et al*. Determining structures in a native environment using single-particle cryoelectron microscopy images. Innovation (Camb), 2021, **2**(4): 100166

[28] Lucas B A, Himes B A, Xue L, *et al*. Locating macromolecular assemblies in cells by 2D template matching with cisTEM. eLife, 2021, **10**: e68946

[29] Li X, Mooney P, Zheng S, *et al*. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods, 2013, **10**(6): 584-590

[30] Langlois R, Pallesen J, Frank J. Reference-free particle selection enhanced with semi-supervised machine learning for cryo-electron microscopy. J Struct Biol, 2011, **175**(3): 353-361

[31] Huang Z, Penczek P A. Application of template matching technique to particle detection in electron micrographs. J Struct Biol, 2004, **145**(1/2): 29-40

[32] Bepler T, Morin A, Rapp M, *et al*. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. Nat Methods, 2019, **16**(11): 1153-1160

[33] Wang F, Gong H, Liu G, *et al*. DeepPicker: a deep learning approach for fully automated particle picking in cryo-EM. J Struct Biol, 2016, **195**(3): 325-336

[34] Scheres S H W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol, 2012, **180**(3): 519-530

[35] Huang C Y, Draczkowski P, Wang Y S, *et al*. *In situ* structure and dynamics of an alphacoronavirus spike protein by cryo-ET and cryo-EM. Nat Commun, 2022, **13**(1): 4877

[36] Ke Z, Oton J, Qu K, *et al*. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. Nature, 2020, **588**(7838): 498-502

[37] Han Y, Zhou Z, Jin R, *et al*. Mechanical activation opens a lipid-lined pore in OSCA ion channels. Nature, 2024, **628**(8009): 910-918

[38] Yao X, Fan X, Yan N. Cryo-EM analysis of a membrane protein embedded in the liposome. Proc Natl Acad Sci USA, 2020, **117**(31): 18497-18503

[39] Zheng W, Chai P, Zhu J, *et al*. High-resolution *in situ* structures of mammalian respiratory super complexes. Nature, 2024, **631**(8019): 232-239

[40] Cheng J, Liu T, You X, *et al*. Determining protein structures in cellular lamella at pseudo-atomic resolution by GisSPA. Nat Commun, 2023, **14**(1): 1282

[41] You X, Zhang X, Cheng J, *et al*. *In situ* structure of the red algal phycobilisome-PSII-PSI-LHC mega complex. Nature, 2023, **616**(7955): 199-206

[42] Elferich J, Schiroli G, Scadden D T, *et al*. Defocus Corrected Large Area Cryo-EM (DeCo-LACE) for label-free detection of molecules across entire cell sections. Elife, 2022, **11**: e80980

[43] Lucas B A, Himes B A, Grigorieff N. Baited reconstruction with 2D template matching for high-resolution structure determination *in vitro* and *in vivo* without template bias. Elife, 2023, **12**: RP90486

[44] Rickgauer J P, Grigorieff N, Denk W. Single-protein detection in crowded molecular environments in cryo-EM images. Elife, 2017, **6**: e25648

[45] Ma J, You X, Sun S, *et al*. Structural basis of energy transfer in *Porphyridium purpureum* phycobilisome. Nature, 2020, **579**(7797): 146-151

[46] Zhang X, Xiao Y, You X, *et al*. *In situ* structural determination of cyanobacterial phycobilisome-PSII super complex by STAgSPA strategy. bioRxiv, 2023. DOI: 10.1101/2023.12.17.572042

[47] Tang G, Peng L, Baldwin P R, *et al*. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol, 2007, **157**(1): 38-46

[48] Grant T, Rohou A, Grigorieff N. *cis*TEM, user-friendly software for single-particle image processing. Elife, 2018, **7**: e35383

[49] Zheng S Q, Palovcak E, Armache J P, *et al*. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat Methods, 2017, **14**(4): 331-332

[50] Wang Z, Grange M, Wagner T, *et al*. The molecular basis for sarcomere organization in vertebrate skeletal muscle. Cell, 2021, **184**(8): 2135-2150.e13

[51] Chang Y, Zhang K, Carroll B L, *et al*. Molecular mechanism for rotational switching of the bacterial flagellar motor. Nat Struct Mol Biol, 2020, **27**(11): 1041-1047

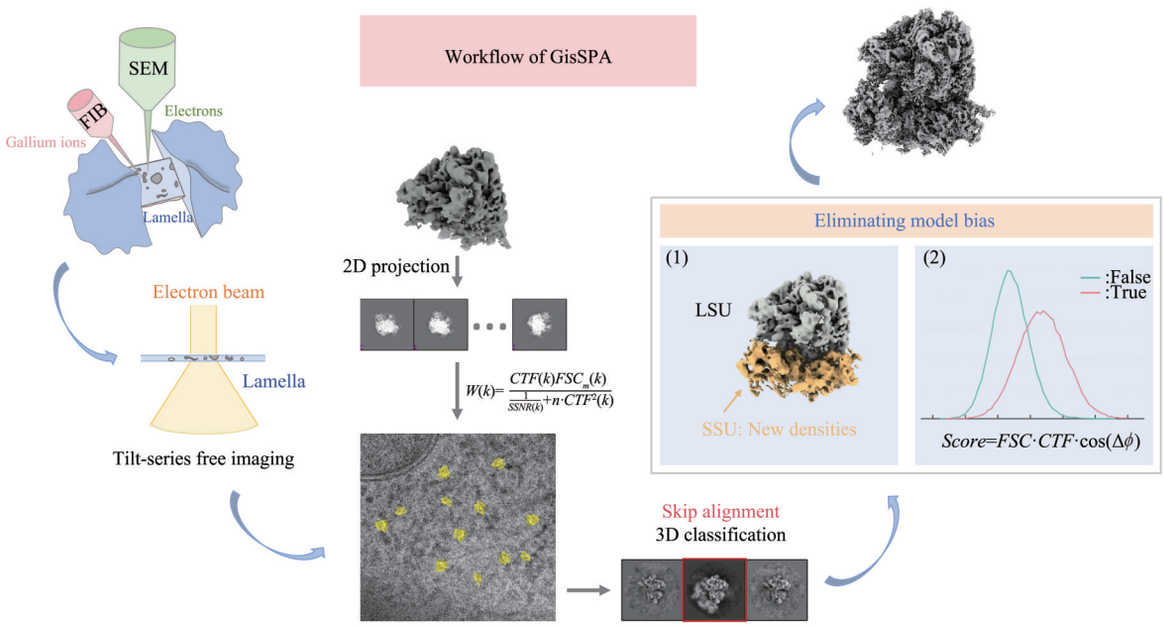
[52] Chmielewski D, Schmid M F, Simmons G, *et al*. Chikungunya virus assembly and budding visualized *in situ* using cryogenic electron tomography. Nat Microbiol, 2022, **7**(8): 1270-1279

GisSPA: a New Method for *in situ* Protein Structural Analysis Based on Cryo-EM Single-particle-like Non-tilting Imaging Data*

ZHAO Ming-Jie, CAO Duan-Fang, ZHANG Xin-Zheng**

(*National Laboratory of Biomacromolecules, Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China)

Graphical abstract



Abstract　Compared to *in vitro* purified protein complexes, protein complexes in a working state within cells are often more complete, and their three-dimensional structures are in a fully physiological conformation. This is crucial for understanding the structural basis of important functions that protein complexes play in life activities and can also provide more precise target information for drug design. The direct *in situ* structural analysis of protein complexes within cells is known as *in situ* structural analysis of proteins, and cryo-electron tomography (cryo-ET) is the key technology for *in situ* structural analysis. However, cryo-ET has limitations such as low data acquisition throughput for tilt series, cumbersome data processing, and special sample requirements to achieve near-atomic resolution. These issues have become bottlenecks limiting the resolution and practical application of *in situ* structural analysis. In recent years, a method based on the analysis of non-tilted images has developed rapidly, allowing high-throughput, high-resolution structural analysis of protein complexes within cells. This review will discuss the principles of this method, compare its advantages and disadvantages with traditional tomography, and provide an outlook on *in situ* structural analysis of proteins. It is hoped that this review will assist structural biologists in better choosing suitable tools.

Key words　*in situ* protein reconstruction, *in situ* structural analysis of cryoEM, isSPA (*in situ* single particle like analysis) filtering, non-tilting imaging data, high throughput

DOI：10.16476/j.pibb.2024.0282

* This work was supported by a grant from The National Natural Science Foundation of China (32150010).

** Corresponding author.

Tel：86-18513916820，E-mail：xzzhang@ibp.ac.cn

Received：June 30，2024　Accepted：August 17，2024