冷冻电镜中的计算方法：图像处理与三维重构

张 法
zf@ncic.ac.cn
中国科学院计算技术研究所

2010年11月11日

Outline

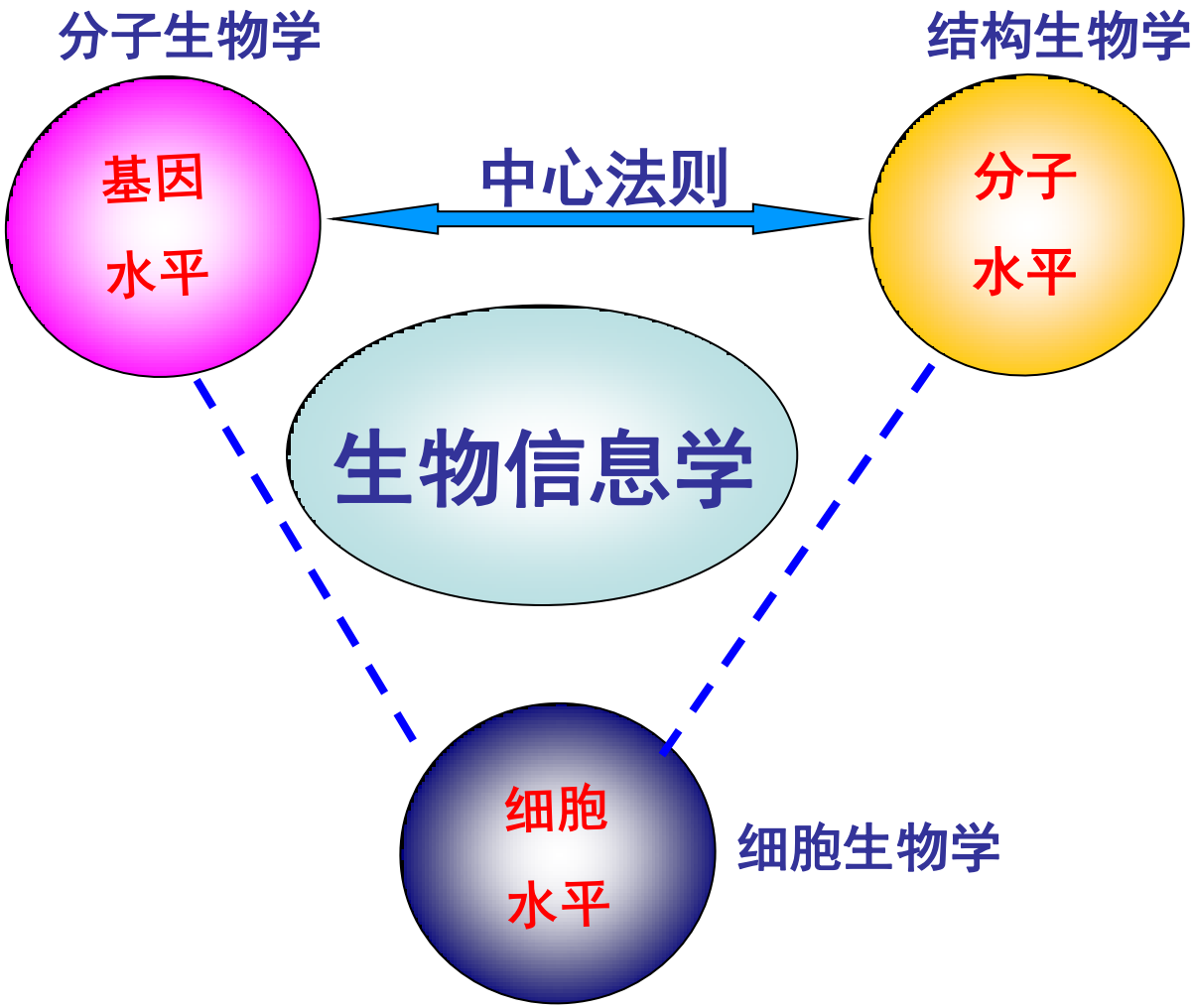
一、冷冻电镜三维重构(CryoEM)简介

二、CryoEM面临的科学问题

三、科研进展

四、研究趋势

生命科学研究的三个层次



蛋白质功能取决于其空间结构

背景简介

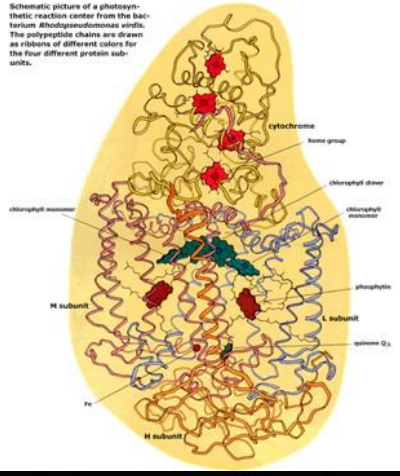
序列

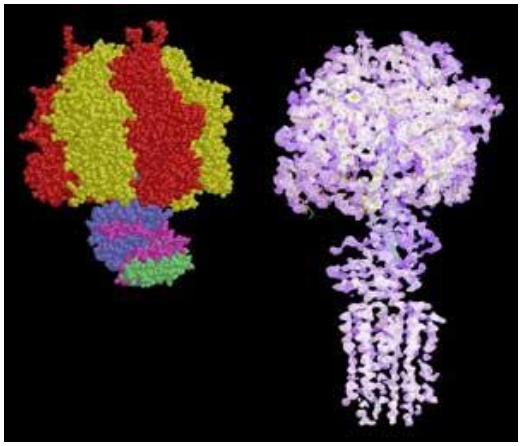
↓

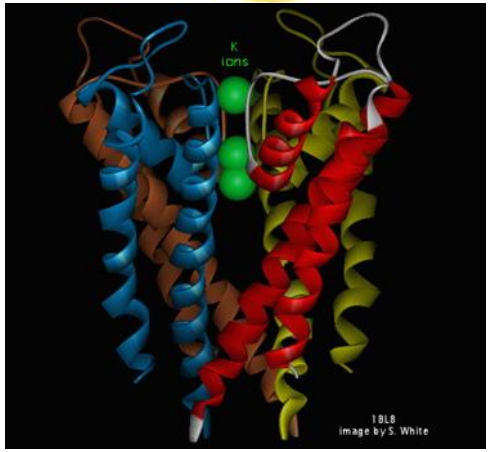
结构

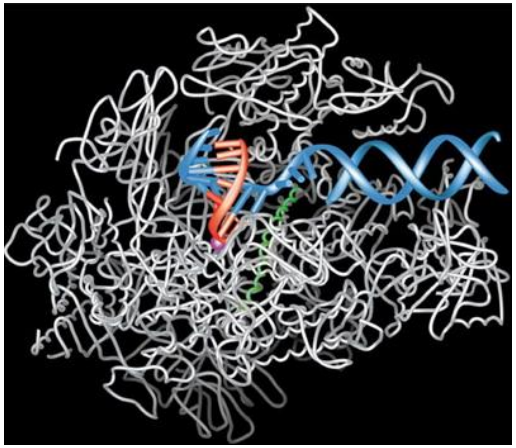
↓

功能

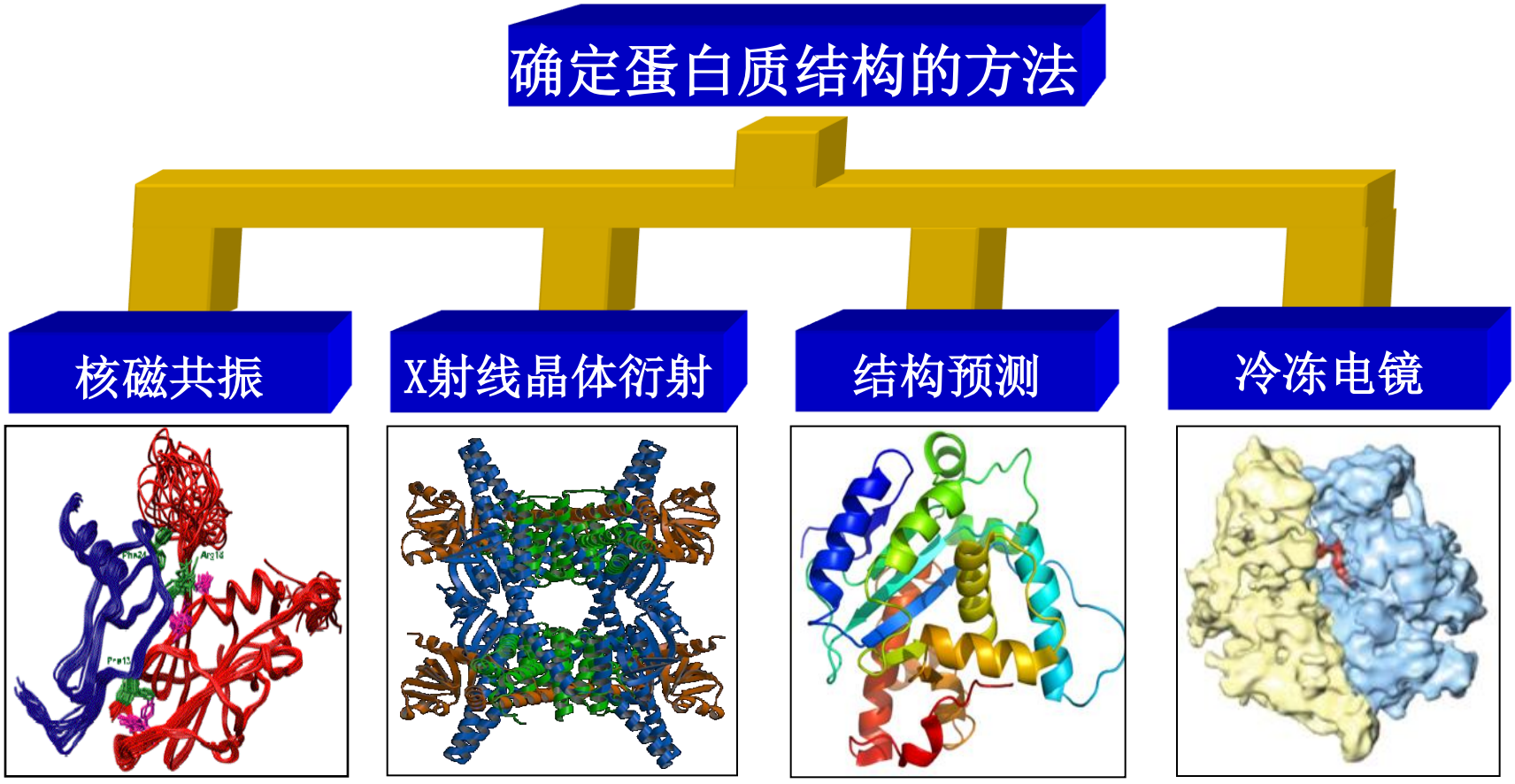






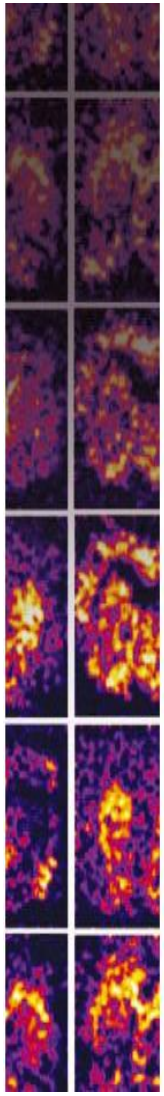


确定蛋白质结构的方法


确定蛋白质结构的方法
核磁共振
X射线晶体衍射
结构预测
冷冻电镜
小分子量蛋白
（细胞趋化因子与抑制分子的复合体）
蛋白质复合体
（SARS病毒非结构蛋白复合体）
小分子量蛋白
膜蛋白和超大分子复合体
酵母核糖体

冷冻电镜的优势

- 研究生物大分子结构的强有力手段。
- 冷冻电镜重构技术在未来结构生物学中起到了一个联系的纽带
- 基于CryoEM的多种技术的融合，被Science评选为2002年世界十大科技进展之一。



insight review articles

From words to literature in structural proteomics

Andrej Sali*, Robert Glaeser†, Thomas Earnest‡ & Wolfgang Baumeister§

*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California, San Francisco, California 94143, USA
†Department of Molecular and Cell Biology, Stanley/Donner ASU, University of California, Berkeley, California 94720, USA
‡Berkeley Center for Structural Biology, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
§Department of Structural Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18 a, 82152 Martinsried, Germany
(e-mail: baumeist@biochem.mpg.de)

Technical advances on several frontiers have expanded the applicability of existing methods in structural biology and helped close the resolution gaps between them. As a result, we are now poised to integrate structural information gathered at multiple levels of the biological hierarchy — from atoms to cells — into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

The structures of individual macromolecules are often uninformative about function if taken out of context. Just as words must be assembled into sentences, paragraphs, chapters and books to make sense, vital cellular functions are domains are important units that are shuffled, duplicated, and fused into larger proteins. Although the universe of distinct amino acid sequences is essentially unlimited, the number of different folding patterns for the domains is not. Extrapolation based on the existing databases of protein

Sali et.al. *Nature* (2003)

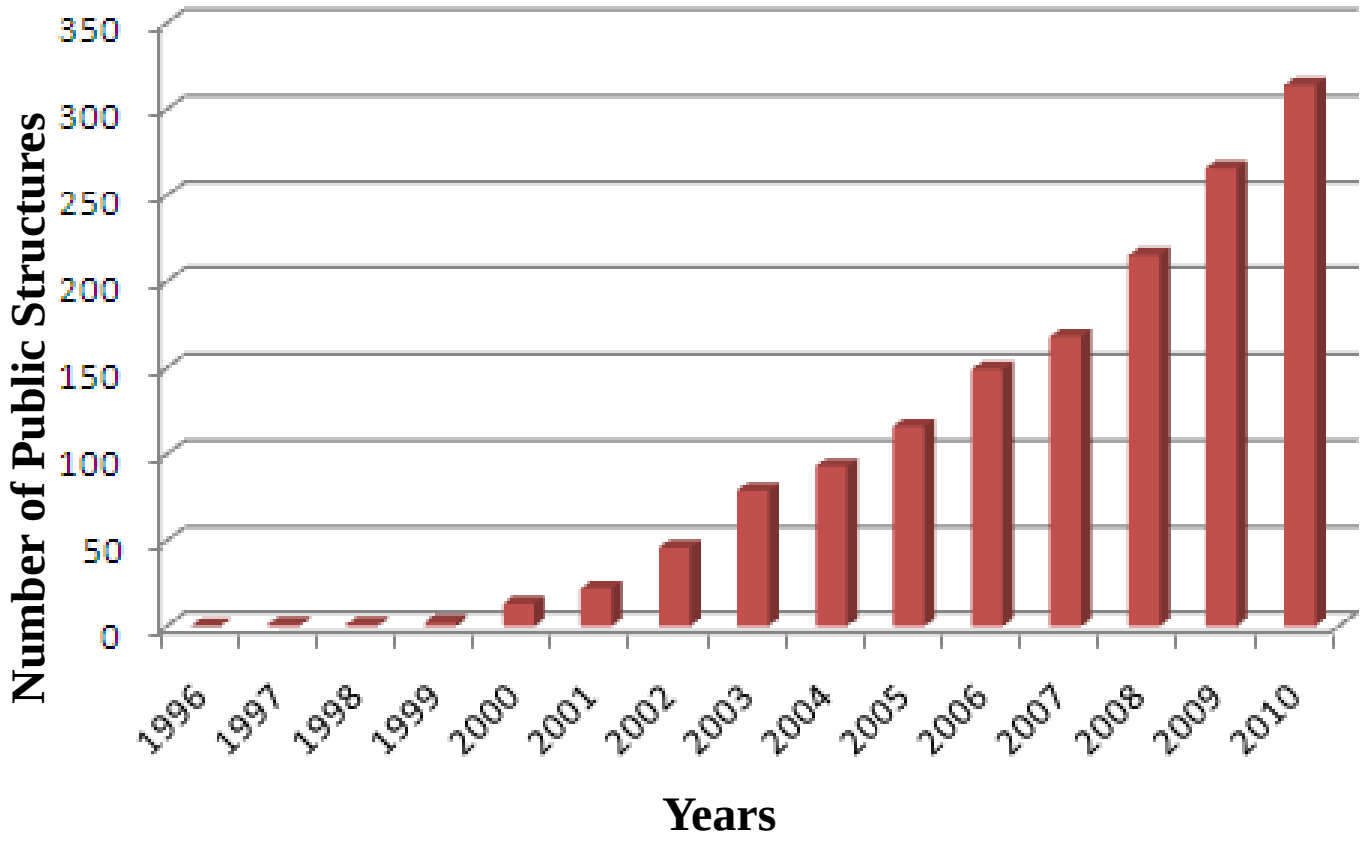
Trends in Macromolecular CryoEM

背景简介

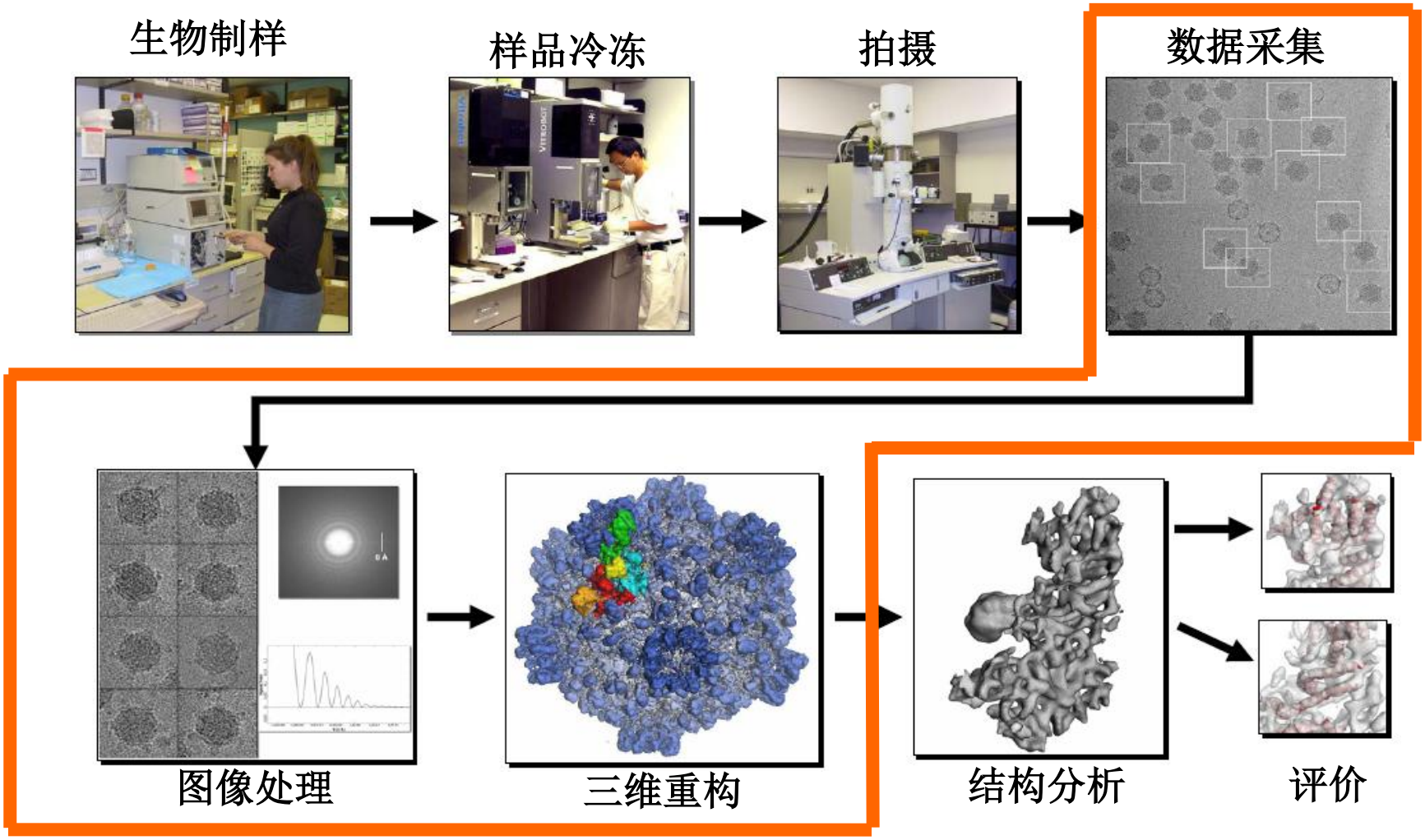
EMDB

EM DATA BANK

http://www.emdatabank.org/



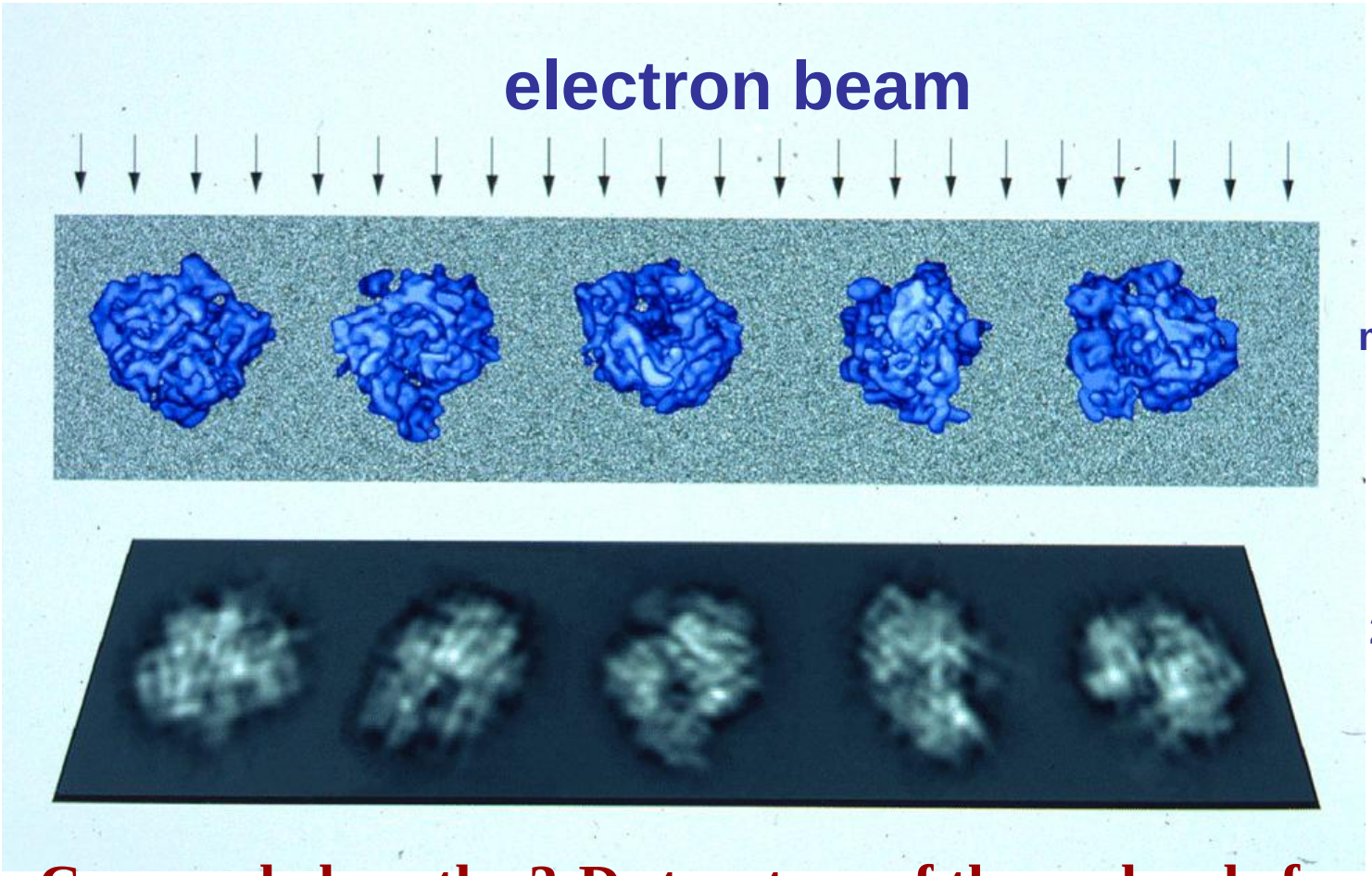
冷冻电镜三维重构处理流程



Chiu et.al. *JEOL News* (2006)

冷冻电镜三维重构问题描述

背景简介



3D macromolecule

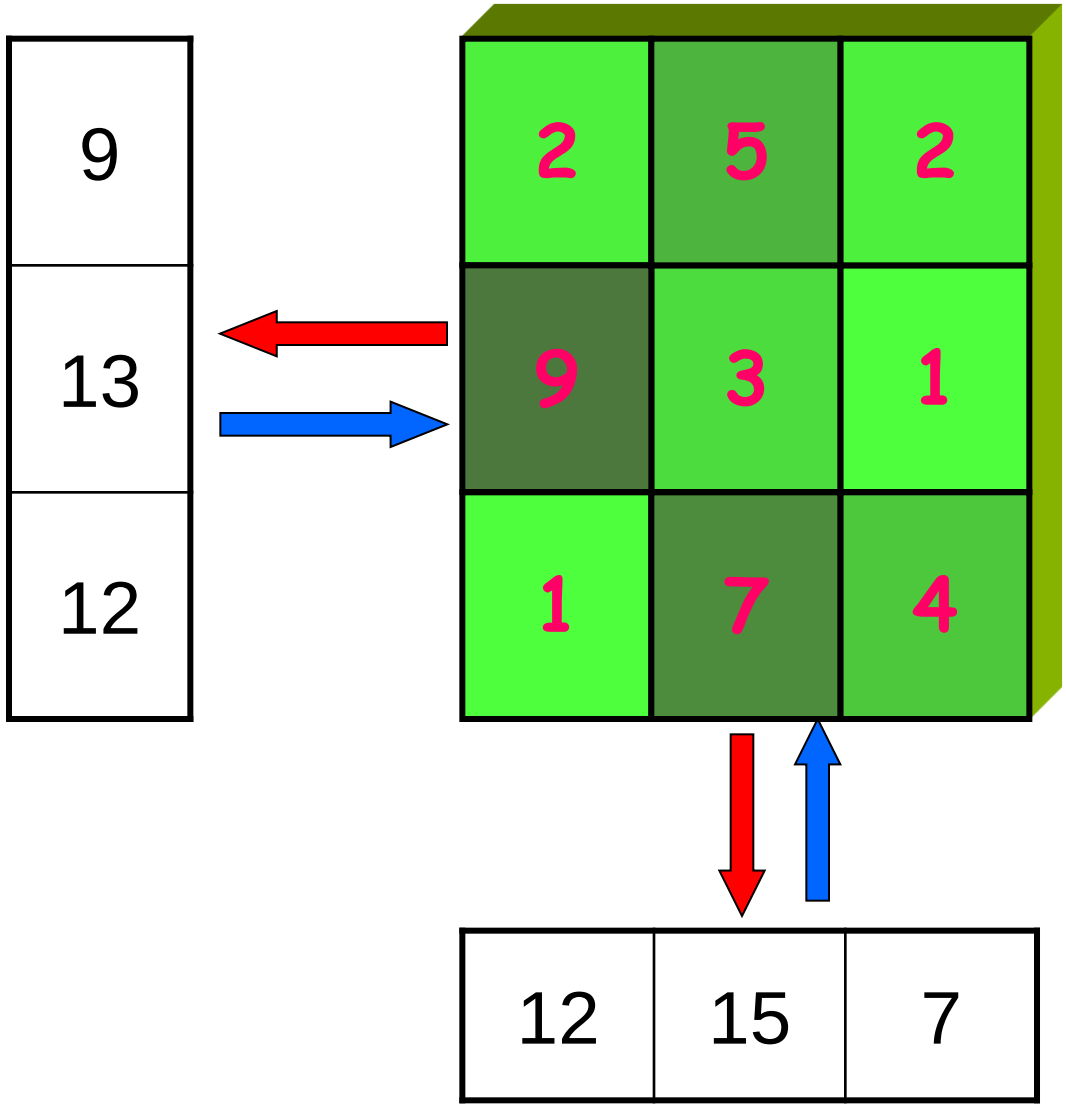
2D projections

(CryoEM)

Can we deduce the 3-D structure of the molecule from a set of 2-D projection images with unknown relative orientations?

冷冻电镜三维重构问题描述

背景简介



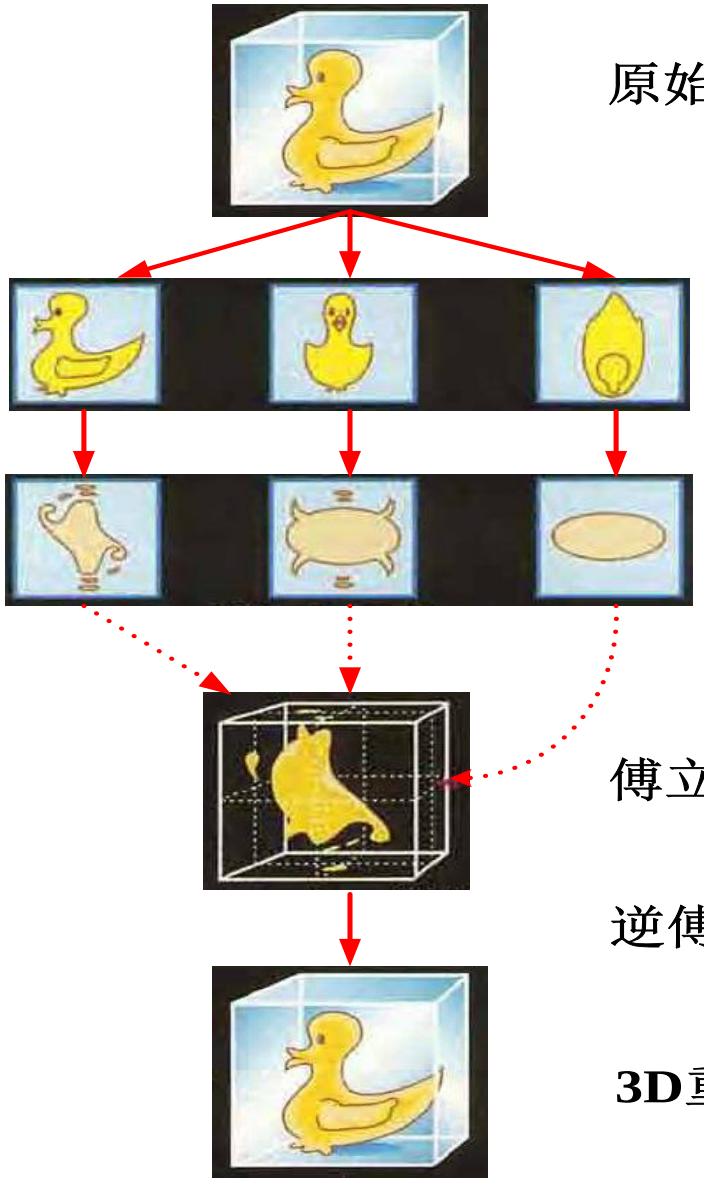
冷冻电镜三维重构问题形式描述

- **Data:** $g_i \in \Re^{n^2}, i = 1,2,...m;$
- **Unknown Parameters:**
  - **Density:** $f \in \Re^{n^3}$
  - **Rotations:** $(\varphi_i, \theta_i, \phi_i), i = 1,2,..., m$
  - **Translations:** $(s_{x_i}, s_{y_i}), i = 1,2,...m;$
- **Objective:**

$$\min_{\phi_i, \theta_i, \varphi_i, s_{x_i}, s_{y_i}, f} \sum_{i=1}^{m} \|r_i\|^2 = \sum_{i=1}^{m} \left\| P\left(\phi_i, \theta_i, \varphi_i, s_{x_i}, s_{y_i}\right) f - g_i \right\|^2$$

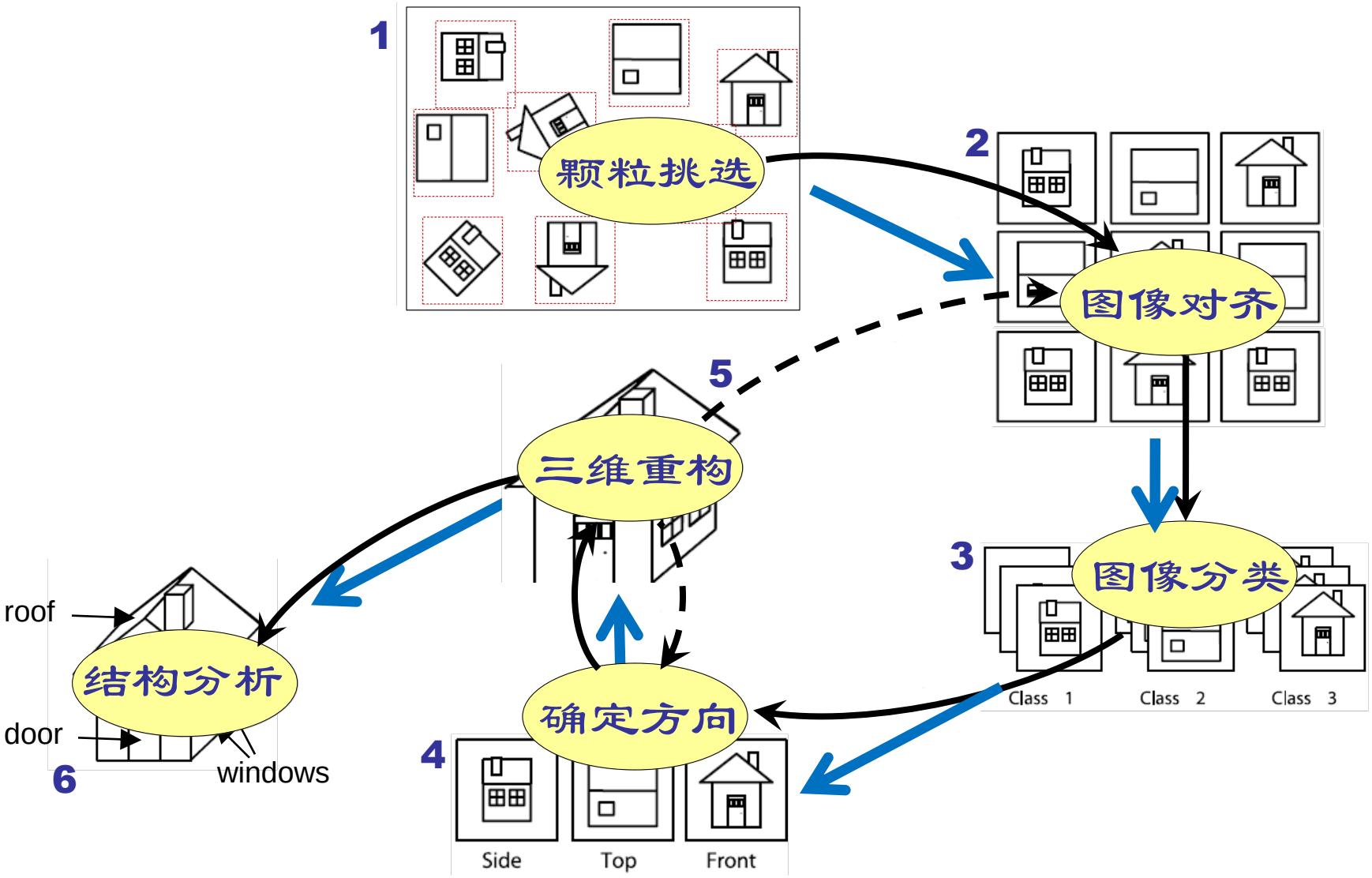
C. Yang et.al. *JSB* (2005)

冷冻电镜三维重构基本思想

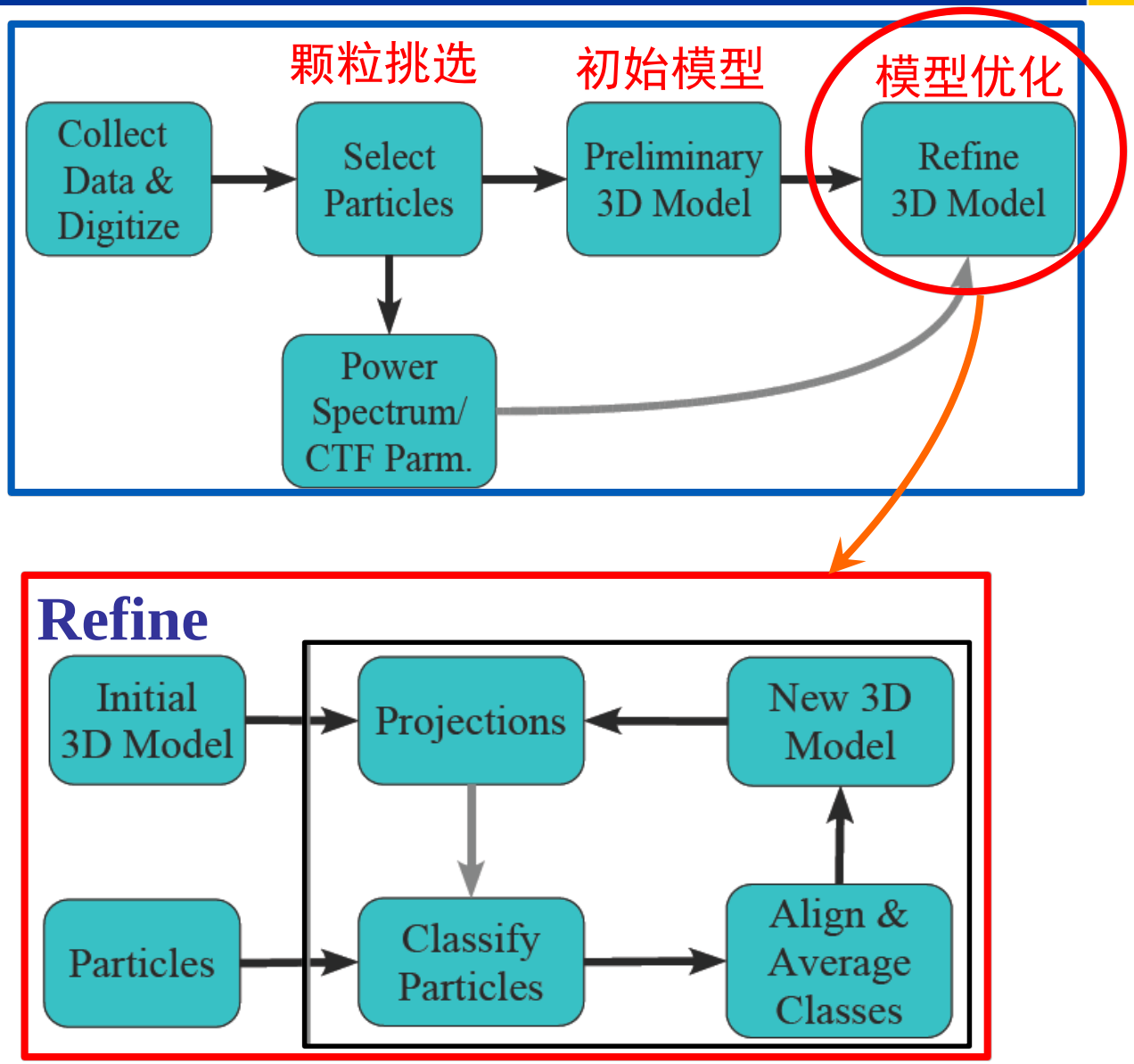


基本思想：相同生物大分子某方向的投影，在实空间中经过调整，叠加平均，提高信噪比，使共同部分的结构信息得到加强，最后对各种不同方向的投影在三维空间中进行重构，从而获得其三维结构信息。

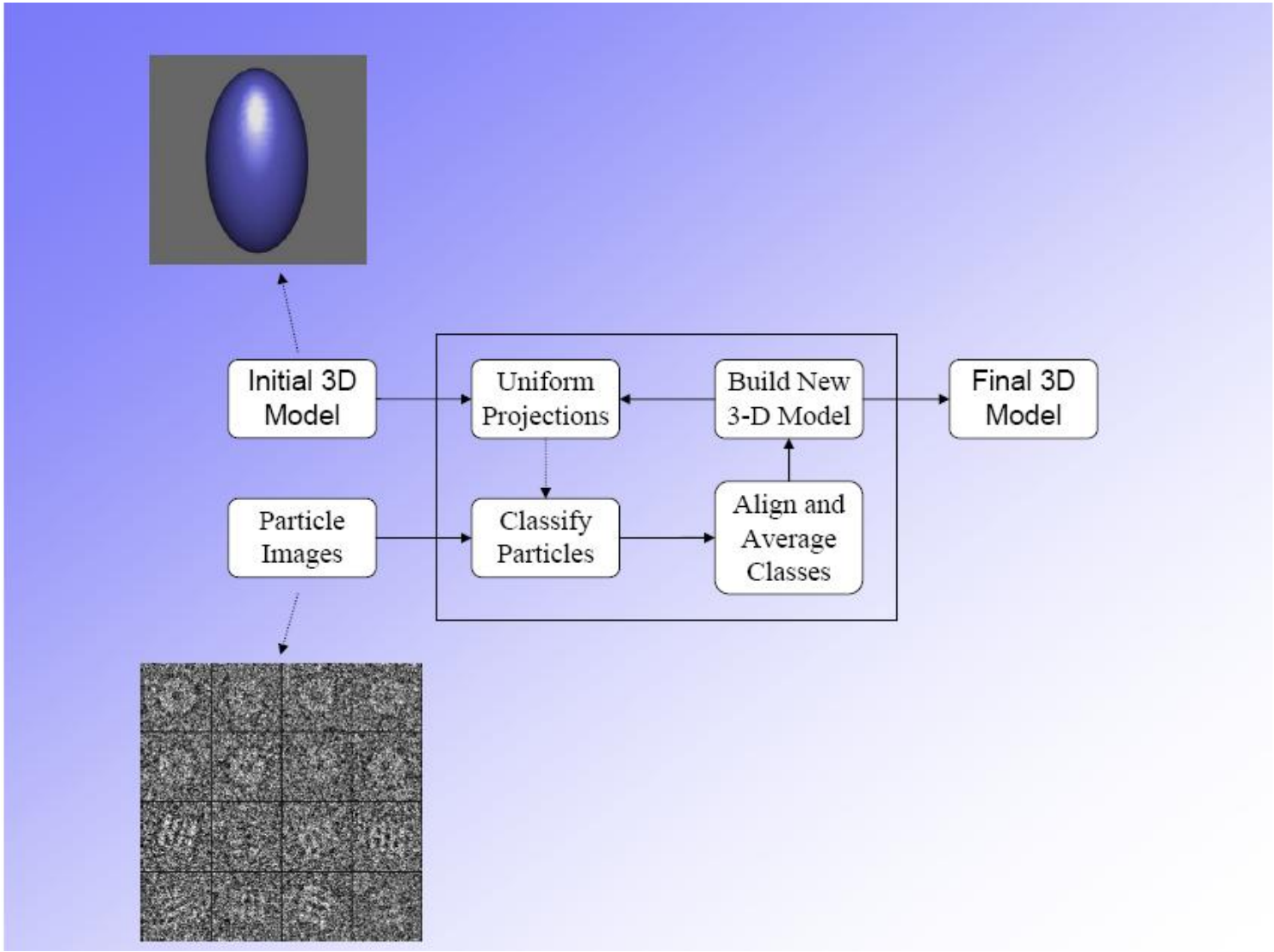
冷冻电镜三维重构基本思想



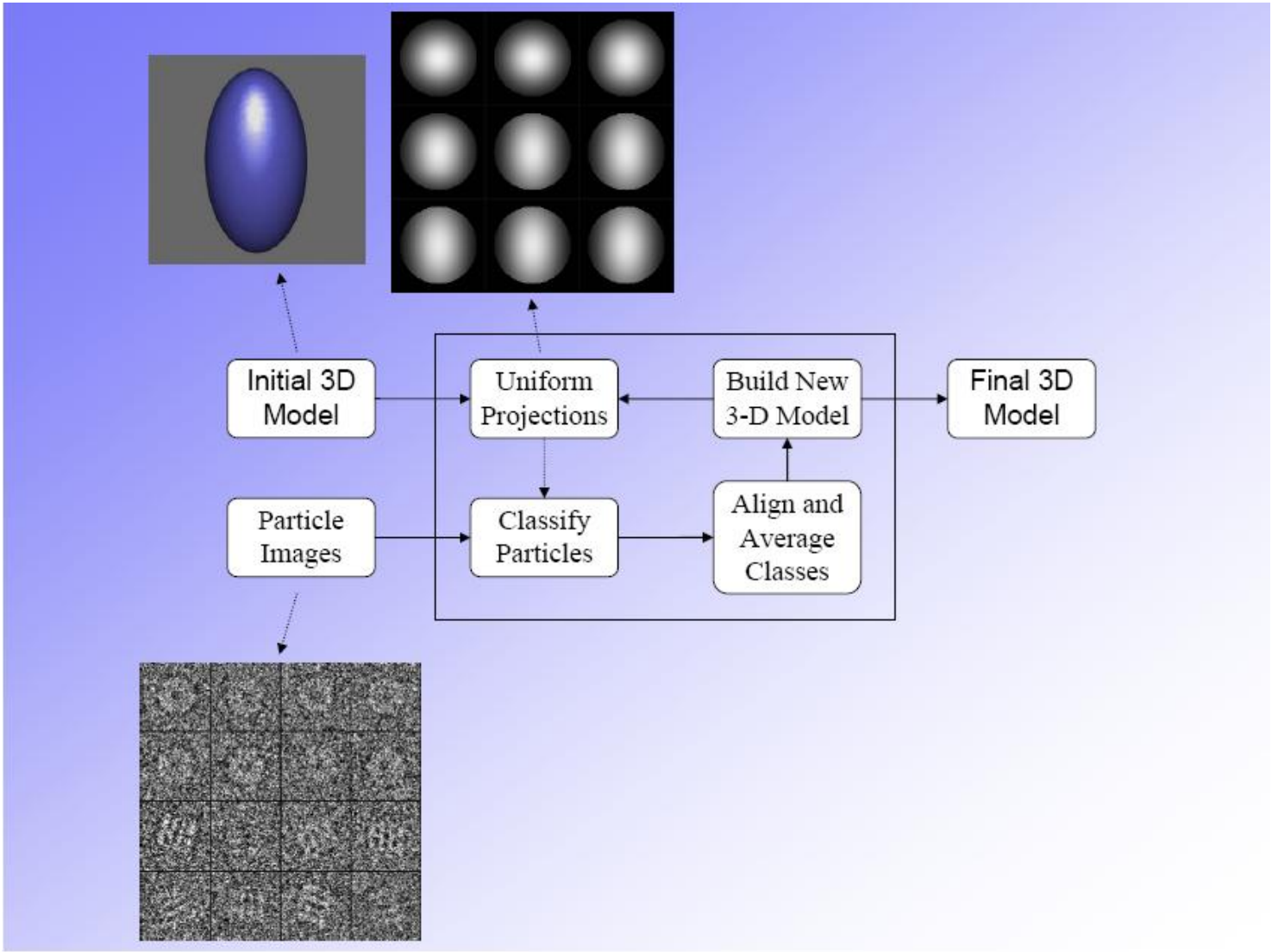
冷冻电镜三维重构的基本步骤



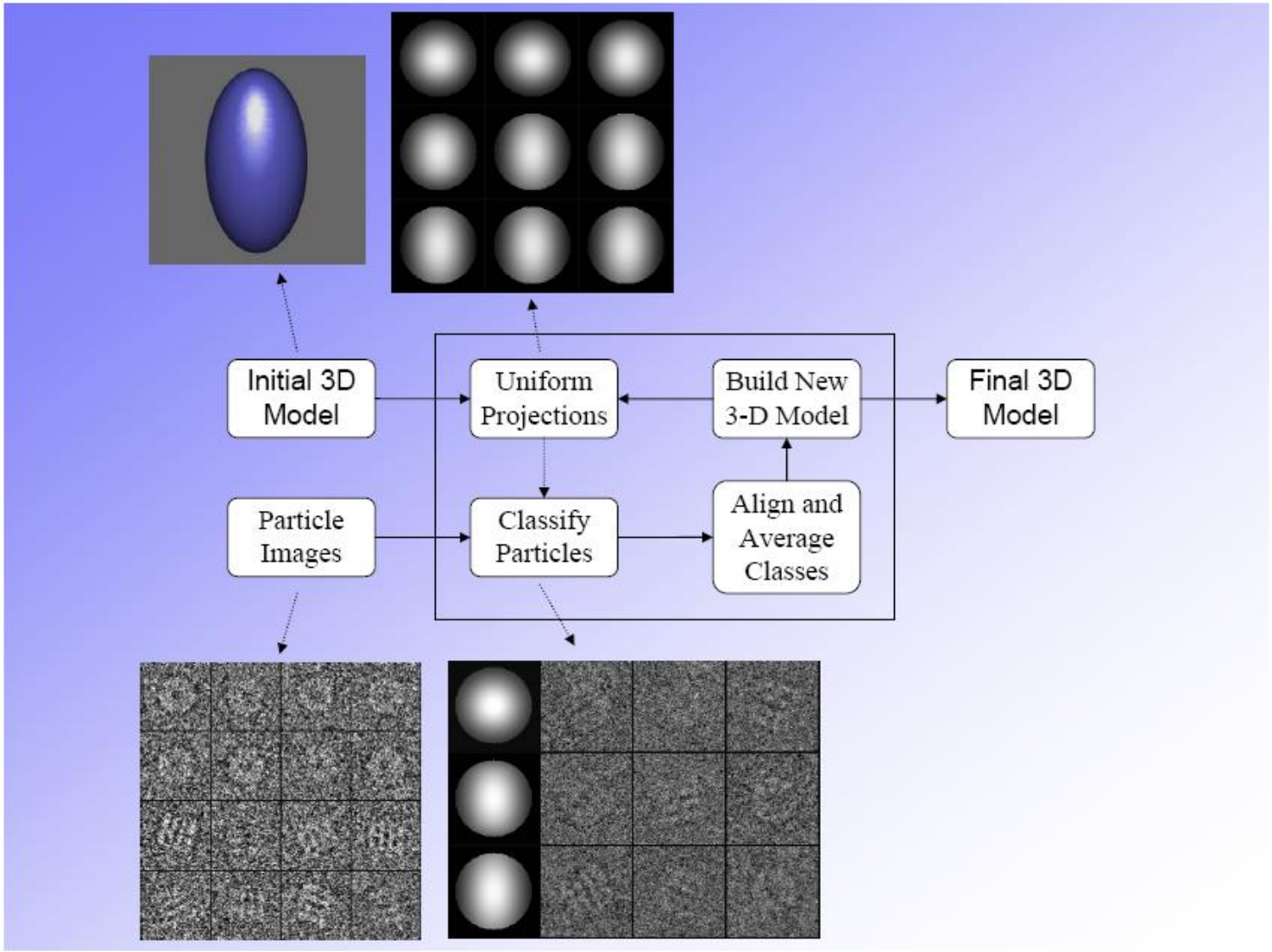
冷冻电镜三维重构的基本步骤



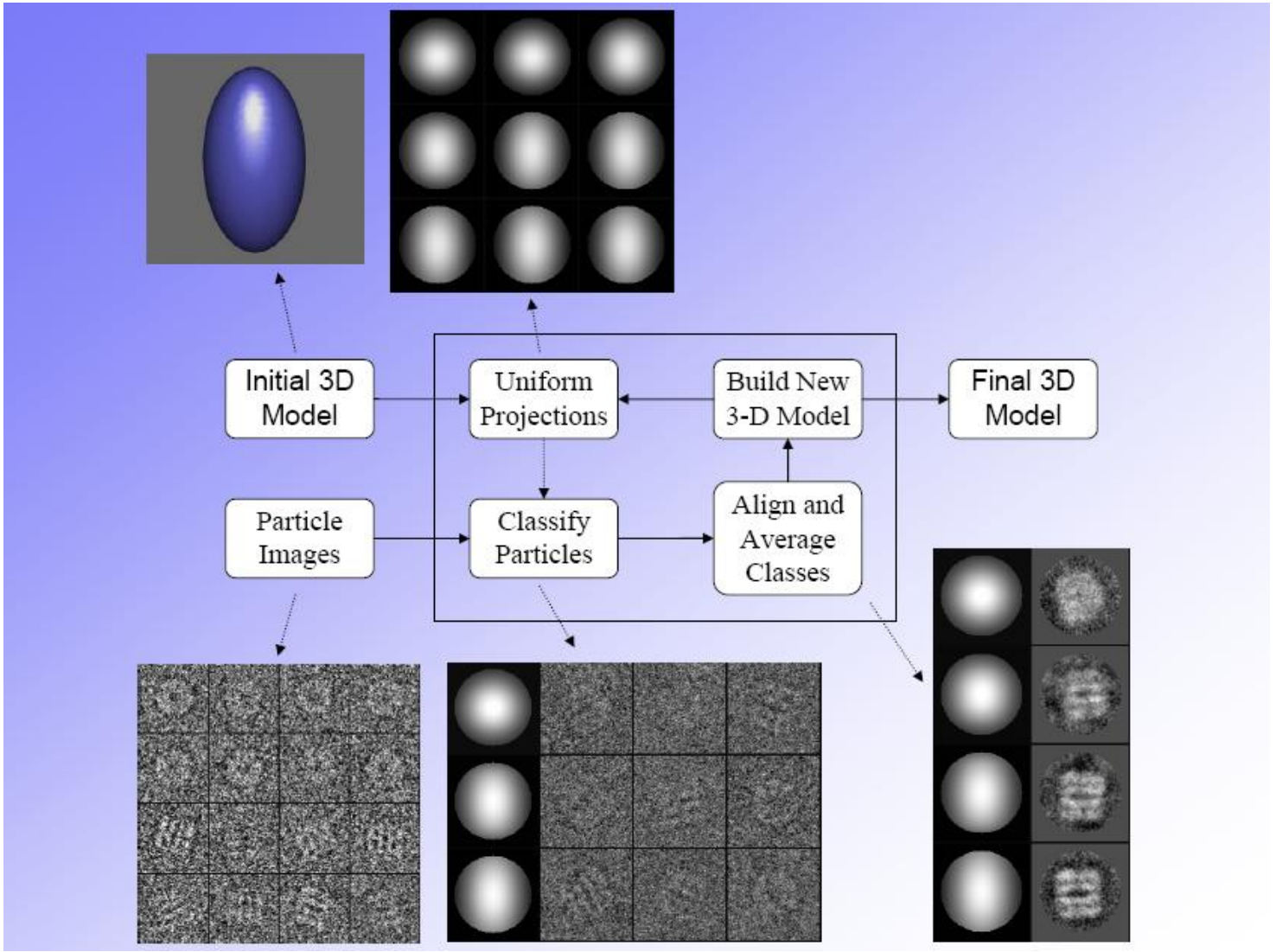
冷冻电镜三维重构的基本步骤



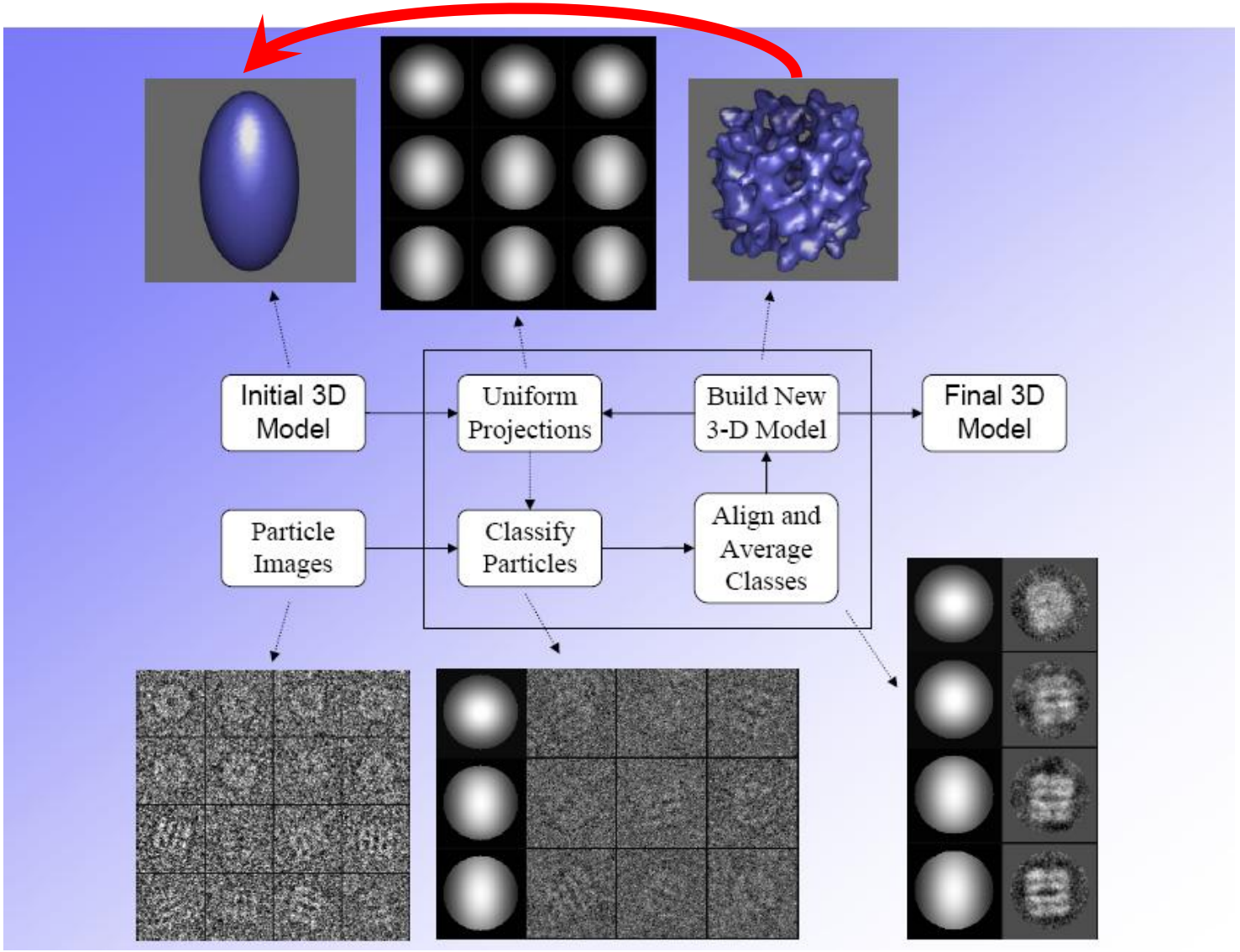
冷冻电镜三维重构的基本步骤



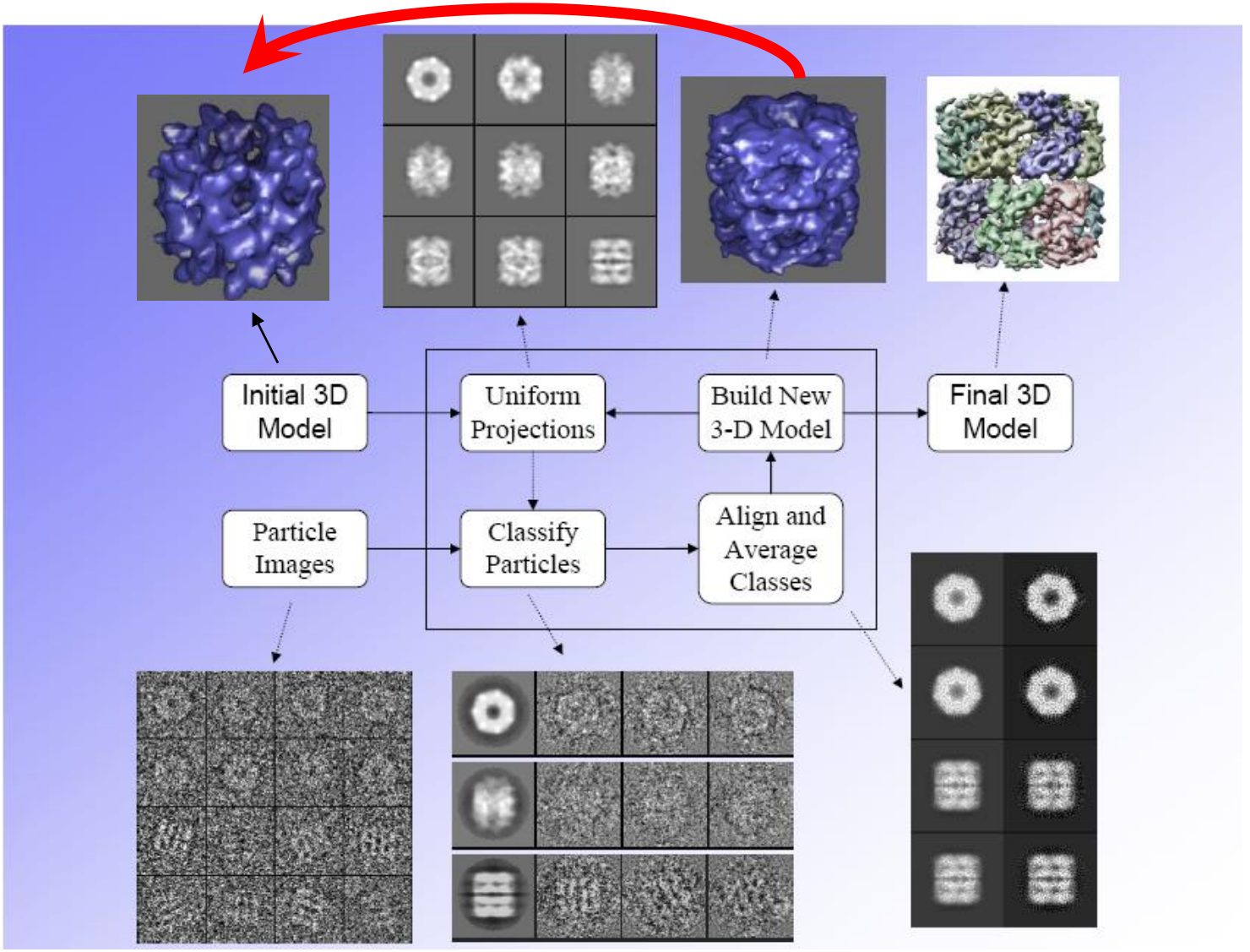
冷冻电镜三维重构的基本步骤



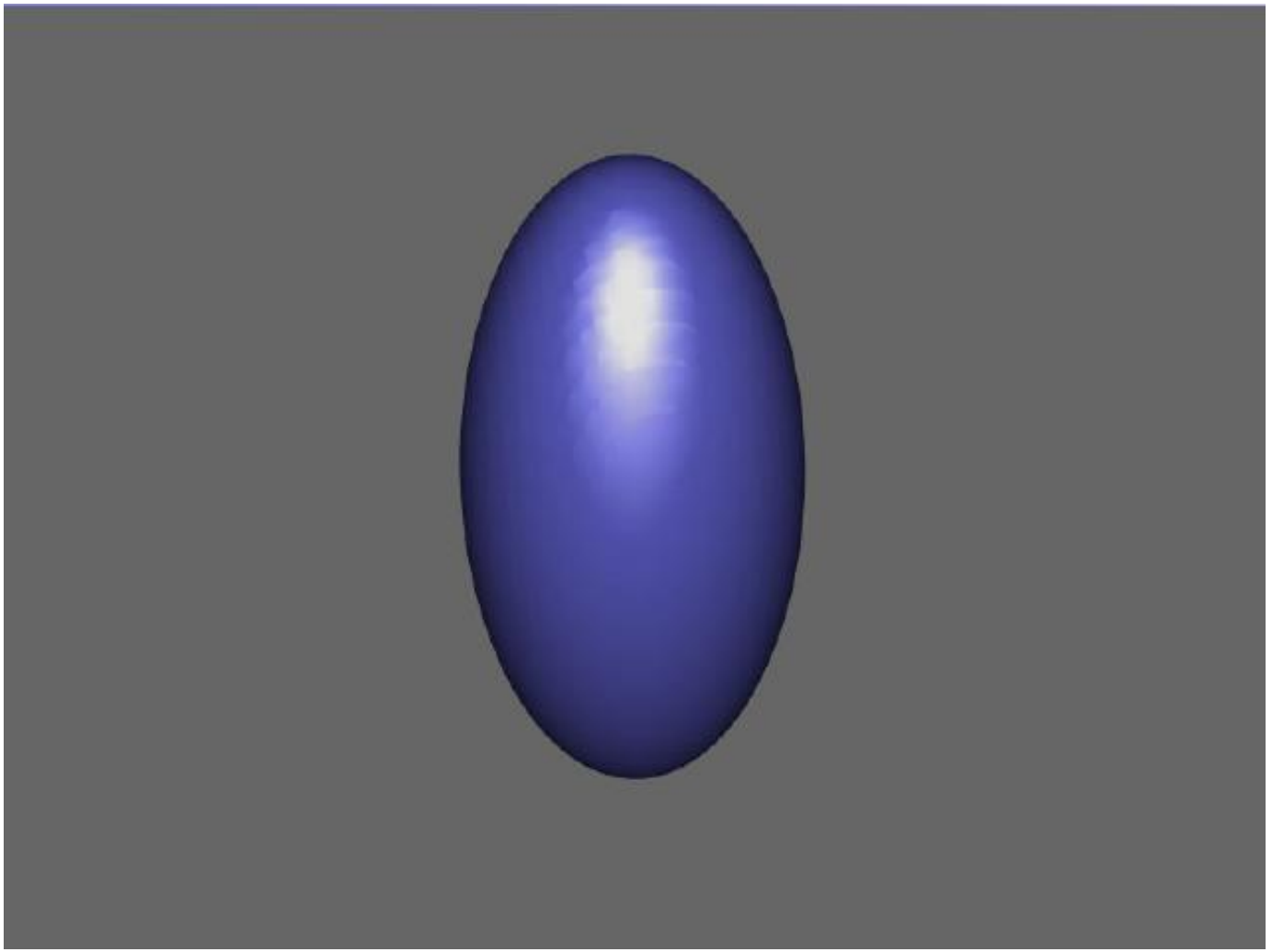
冷冻电镜三维重构的基本步骤



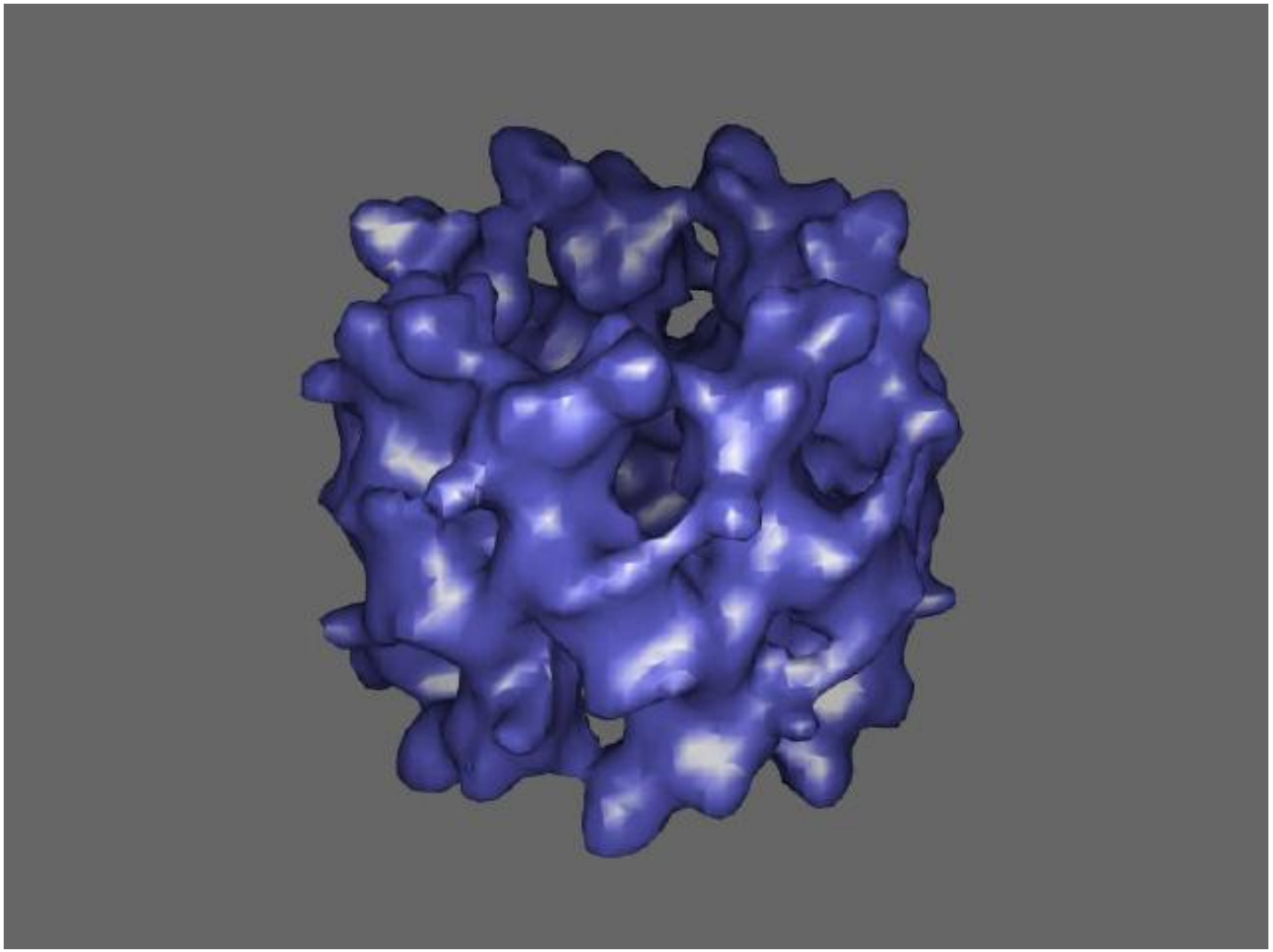
冷冻电镜三维重构的基本步骤



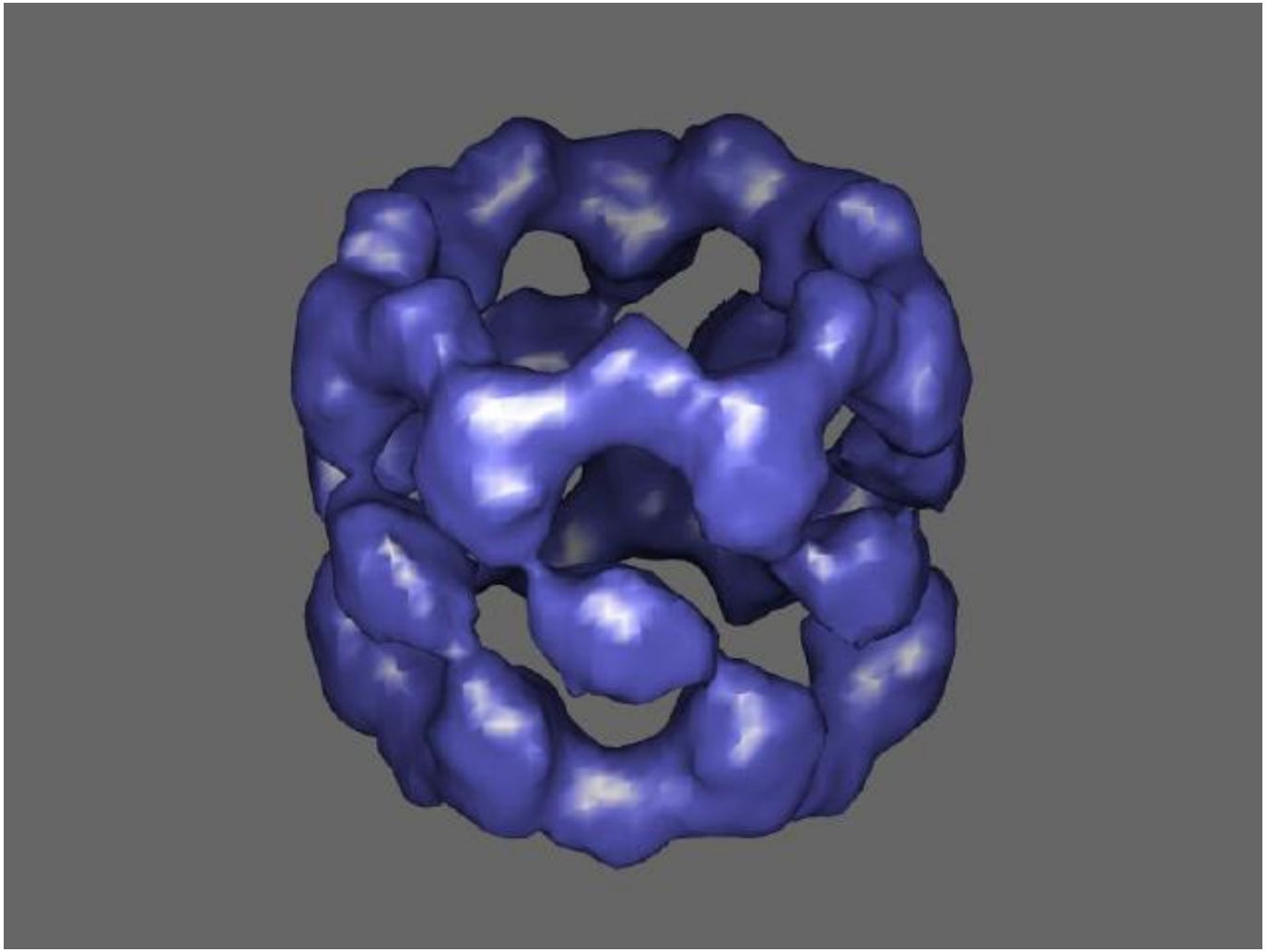
冷冻电镜三维重构结果：初始模型



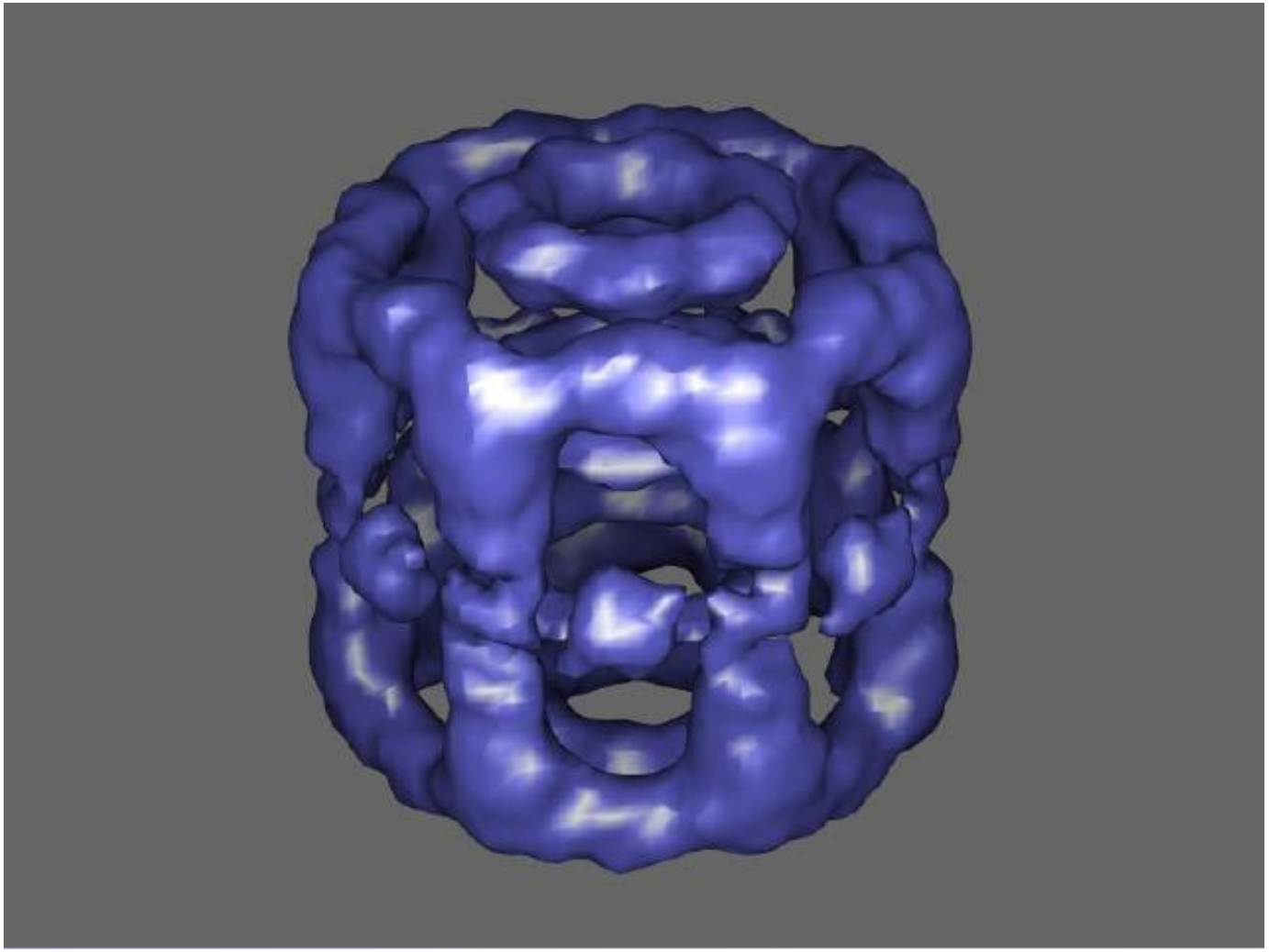
冷冻电镜三维重构结果：第1次迭代



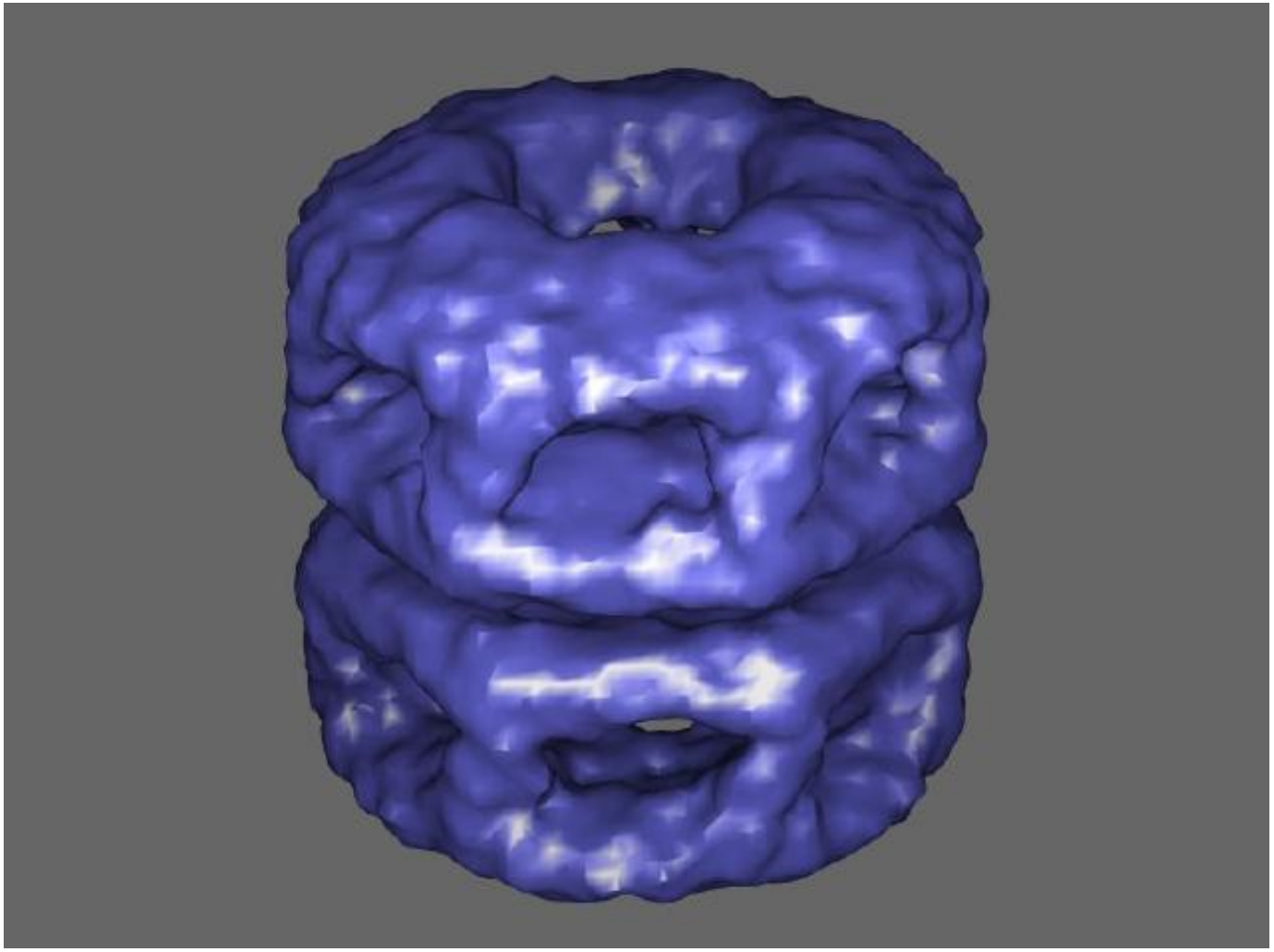
冷冻电镜三维重构结果：第2次迭代



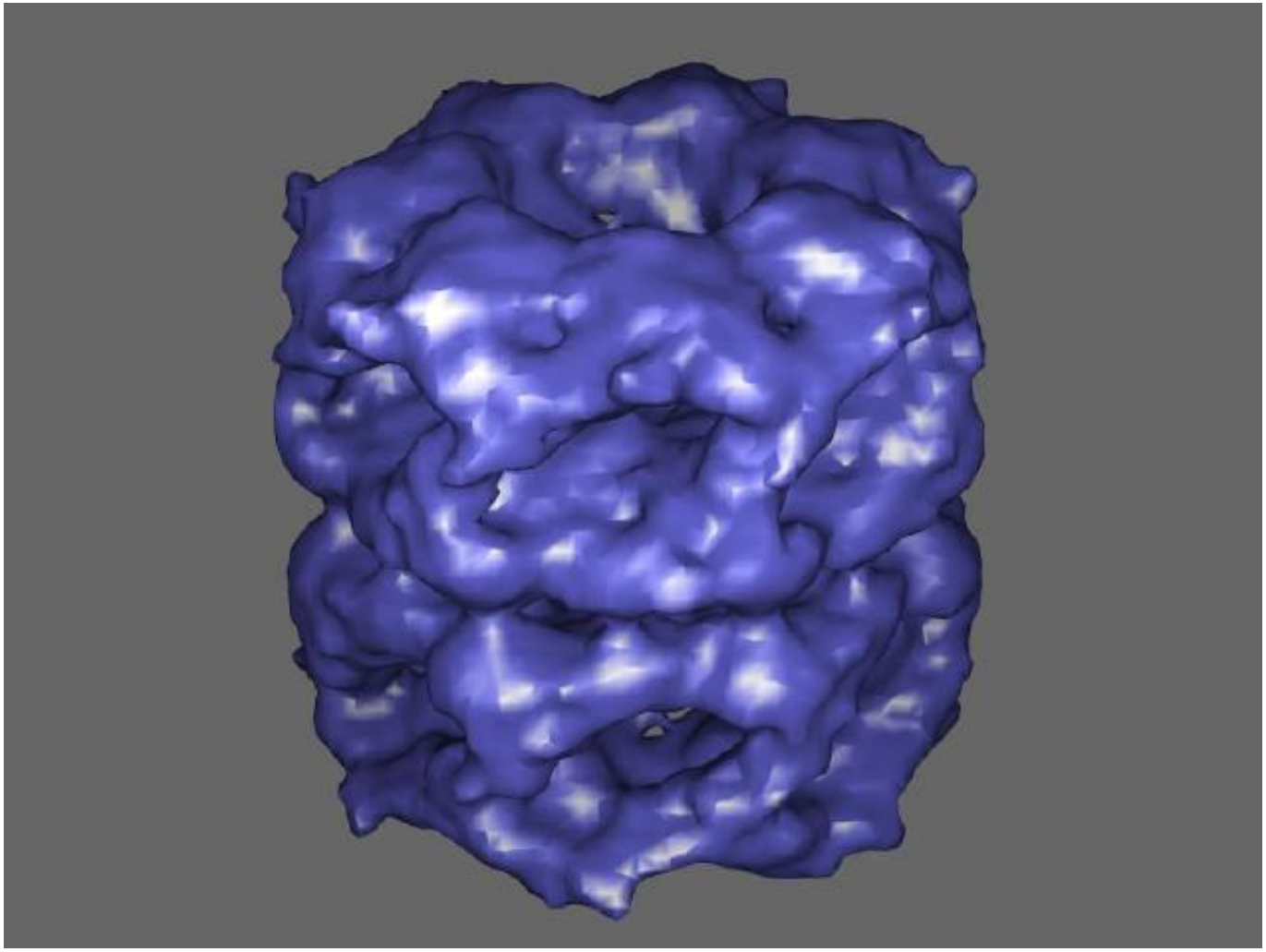
冷冻电镜三维重构结果：第3次迭代



冷冻电镜三维重构结果：第4次迭代



冷冻电镜三维重构结果：第5次迭代



冷冻电镜三维重构面临的科学问题

一、颗粒图像信噪比极低

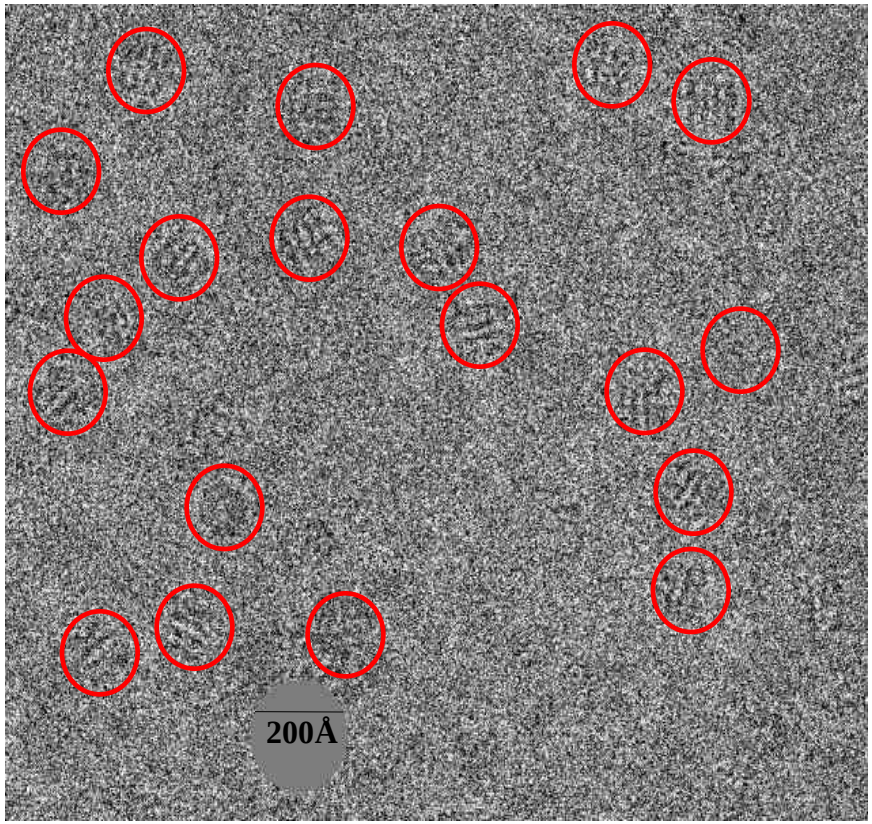
二、三维重构计算时间极其漫长

三、急需新的重构算法

颗粒图像信噪比极低

科学问题

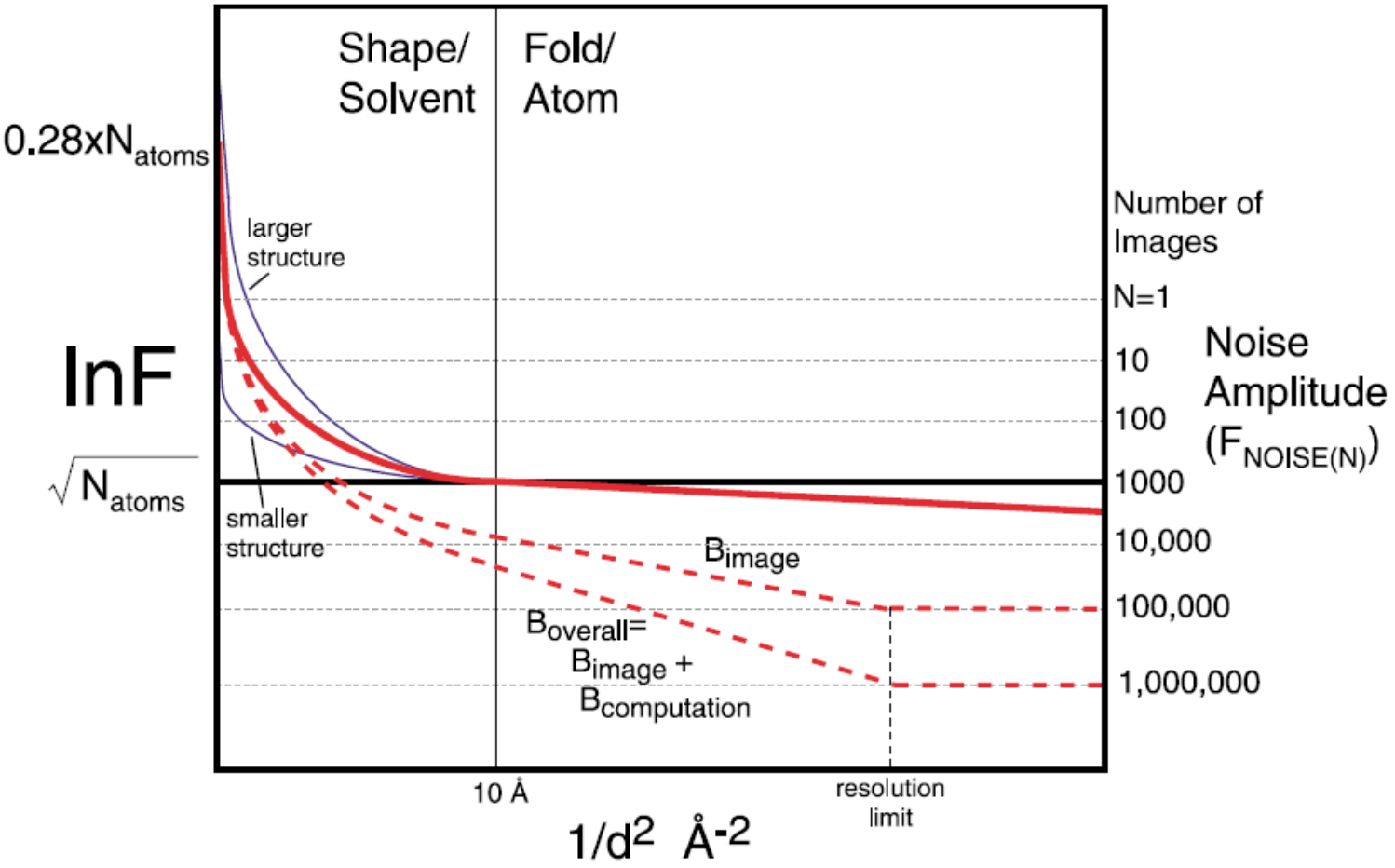
- 颗粒图像信噪比极低
  - 电子成像技术
  - 样品不均一
- 需采集大量图像数据
- 如何快速、准确进行蛋白颗粒的挑选？
  - 手工挑选
  - 半自动化



颗粒图像信噪比极低

科学问题

要达到1Å精度，需要100万张分子颗粒图片

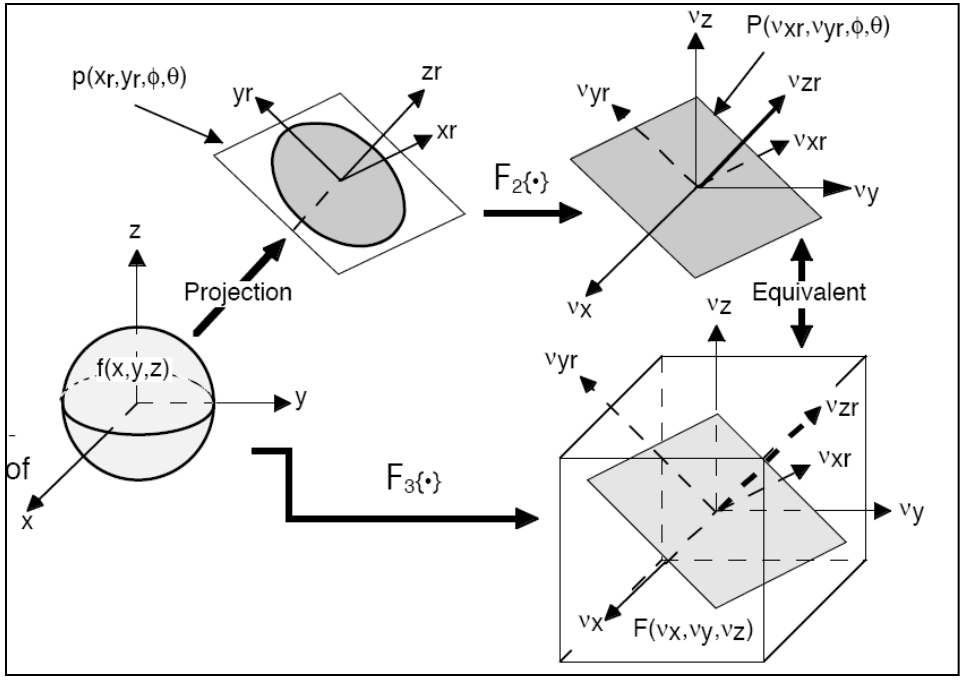


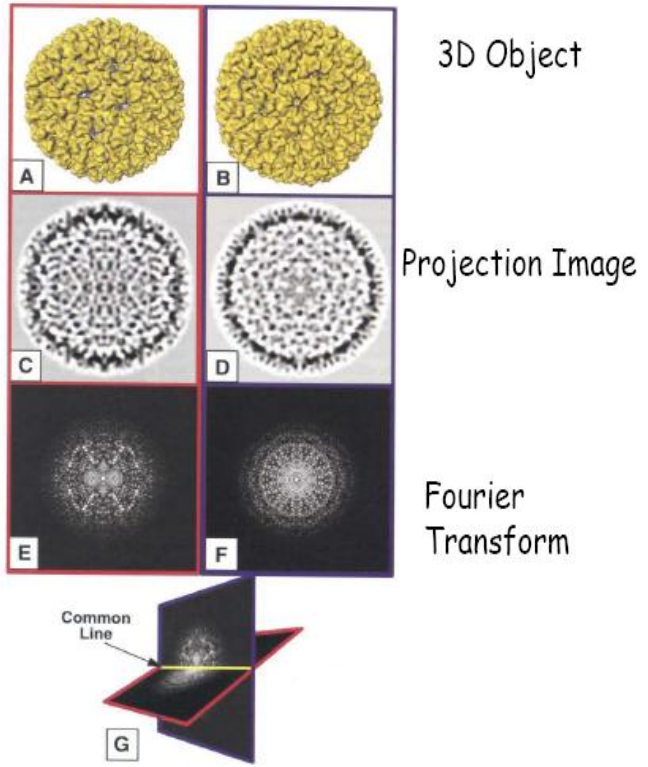
三维重构时间极其漫长

科学问题

数万~数十万颗粒图像

- Classification
- Average
- Fourier Transform
- Refinement
- Alignment
- Cross Common Line
- Inverse Fourier Transform





三维重构时间极其漫长

科学问题

Near-atomic Resolution Structures by Single Particle Cryo-EM						
Complex	Publication	Selected Particles	Software	CPU hrs	Effective resolution (Å)	Modeling method
CPV	Nature (453) 15 May 2008	17,123	IMIRS	~ 10^3	3.8	Fourier-Bessel & 球谐函数
ε 15 phage	Nature (451) 28 Feb. 2008	36,259	EMAN	~10^6	4.5	Direct Fourier inversion
GroEL	Structure (16) March 2008	20,401	EMAN	N/A	4.2	Direct Fourier inversion
Rotavirus	PNAS (105) 12 Feb. 2008	18,125	FREALIGN	~10^5	3.8	Direct Fourier inversion

数据来自 “Current Opinion in Structural Biology” 2008, 18:218-228

急需新的重构算法

科学问题

- 现有算法绝大部分基于Fourier-Bessel模型，70年代初提出的。
- 基于现有的算法，重构精度很难再提高
- 如何进一步提高三维重构的精度？

我们的工作

一、颗粒图像的挑选

——颗粒图像识别软件Picker

二、三维重构的高性能计算

——重构并行软件ParaEMAN

三、重构结果精度优化

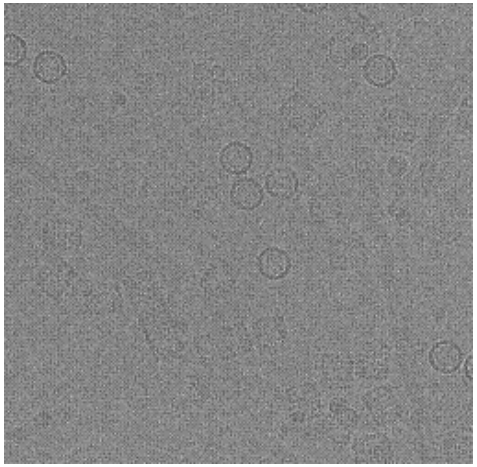
——原有算法的完善

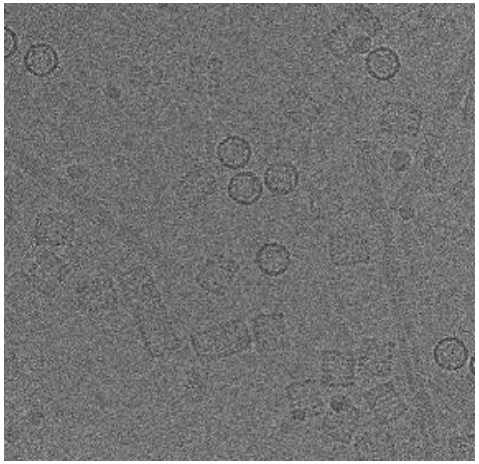
——球面坐标系下的三维重构算法

颗粒图像挑选 — 降噪处理

工作进展

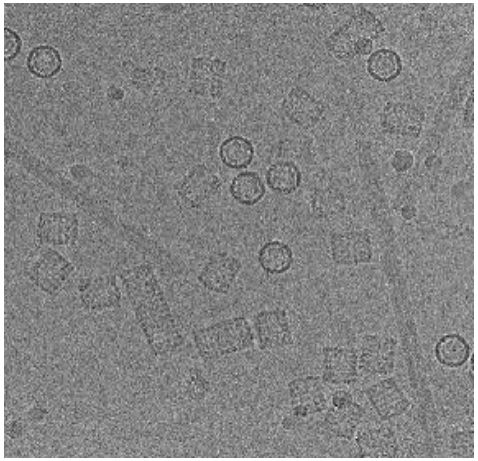
原始图片

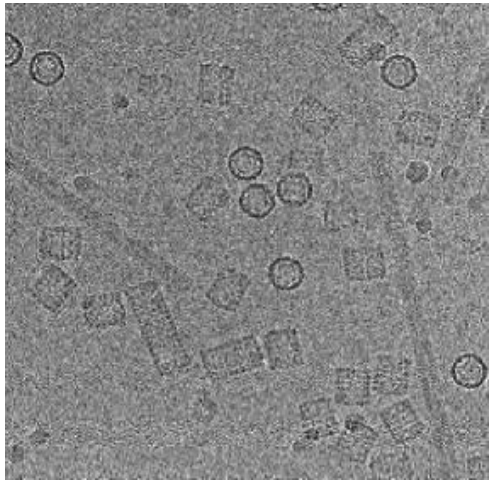




一次降噪

二次降噪





三次降噪

原始电镜图片01nov26b.001.003.001.001 ,来自SCRIPPS

颗粒图像挑选 — Picker

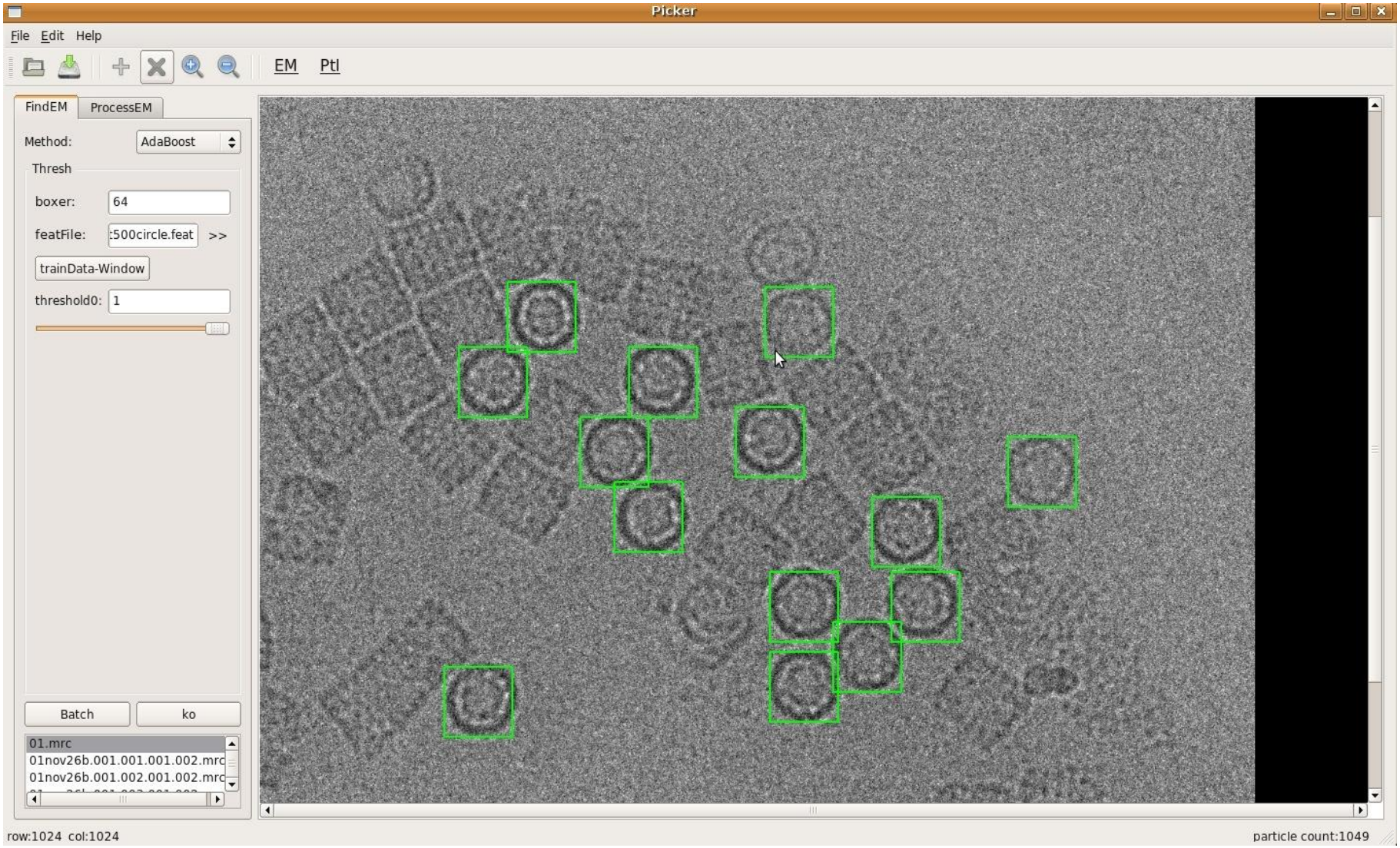
工作进展

- 模板匹配
  - ✓ 直方图信息熵的方法
  - ✓ 相关性匹配
- 特征学习
  - ✓ Adaboost方法
  - ✓ 最小距离分类器
  - ✓ 贝叶斯分类器
  - ✓ 矩不变量/旋转不变量

Algorithm	FPR (%)	FNR (%)	DR (%)
Adaboost （726&786）	3.1	10.7	89.3
Improved adaboost （826&803）	7.5	4.4	95.6
Bayesian classifier （729&778）	5.5	12	88
Min-distance （758&760）	8.2	11	89
Cross –correlation （854&897）	1.67	6.47	93.53

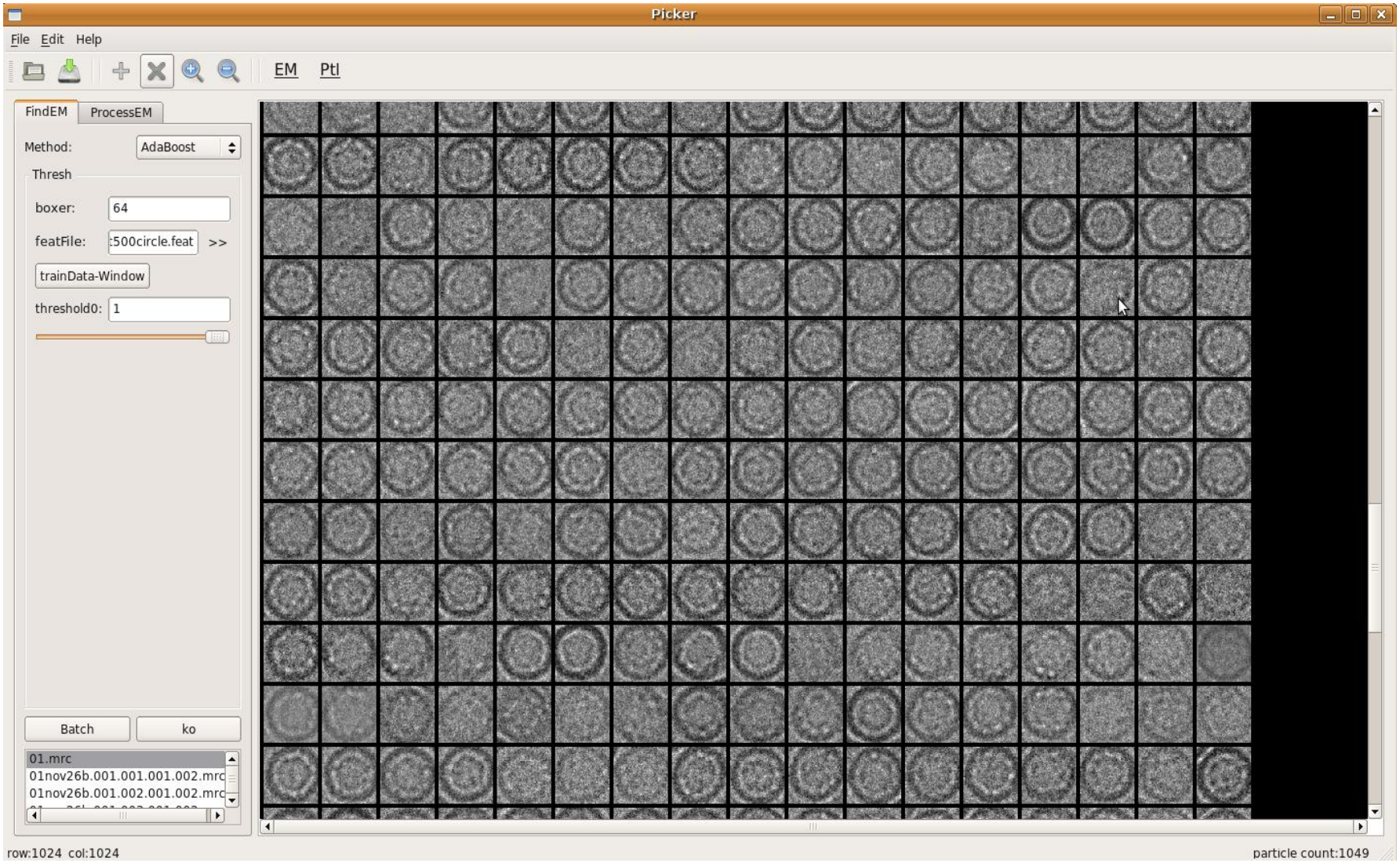
颗粒图像挑选

工作进展


Picker
File Edit Help
EM Ptl
FindEM ProcessEM
Method: AdaBoost
Thresh
boxer: 64
featFile: :500circle.feat >>
trainData-Window
threshold0: 1
Batch ko
01.mrc
01nov26b.001.001.001.002.mrc
01nov26b.001.002.001.002.mrc
row:1024 col:1024
particle count:1049

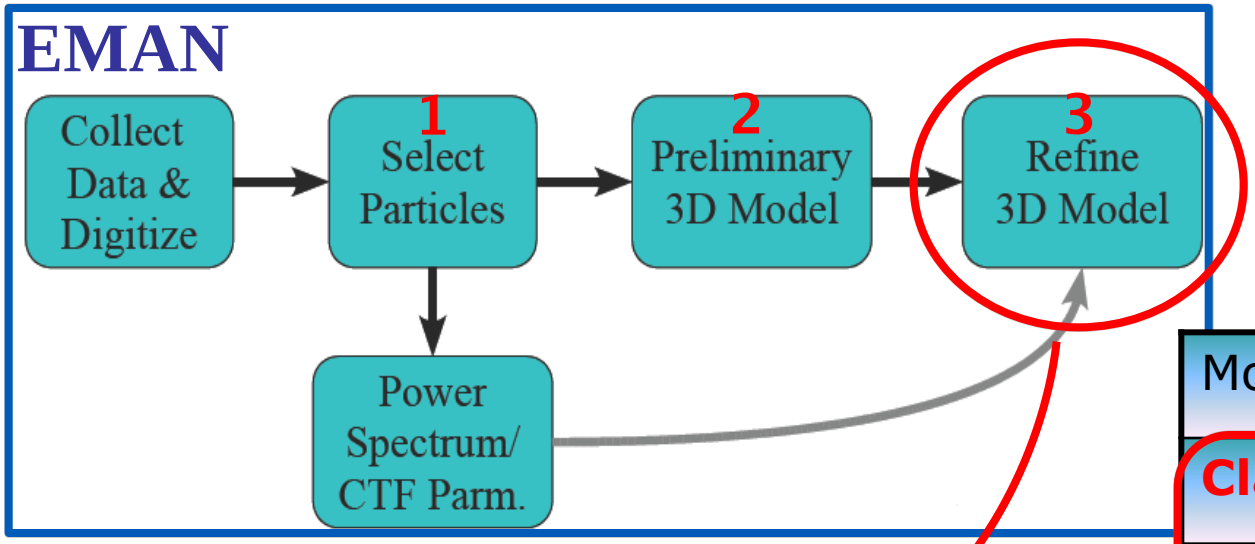
颗粒图像挑选

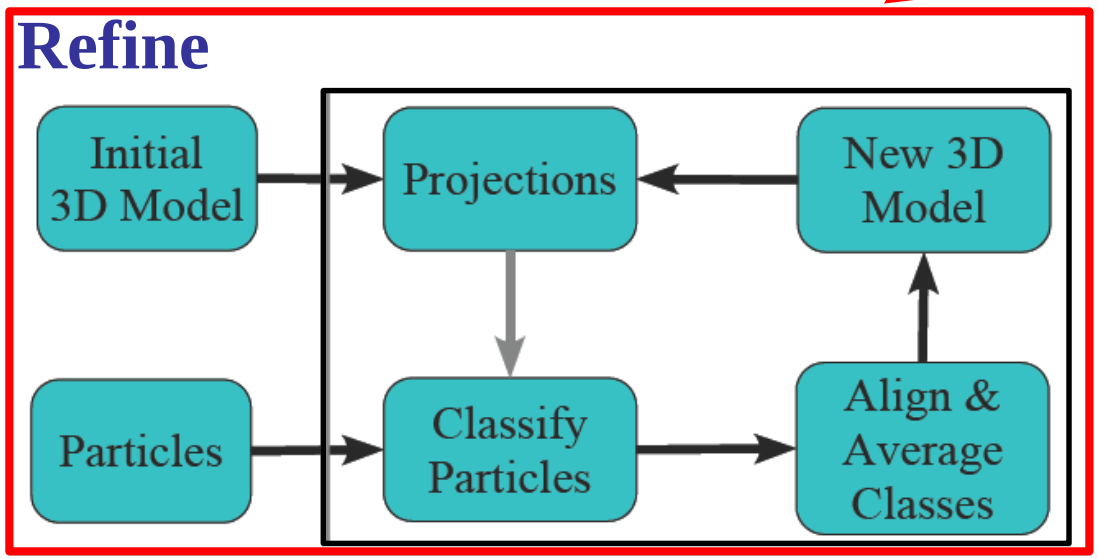
工作进展


Picker
File Edit Help
EM Ptl
FindEM ProcessEM
Method: AdaBoost
Thresh
boxer: 64
featFile: 500circle.feat >>
trainData-Window
threshold0: 1
Batch ko
01.mrc
01nov26b.001.001.001.002.mrc
01nov26b.001.002.001.002.mrc
row:1024 col:1024
particle count:1049

三维重构高性能计算 — 时间分析

工作进展



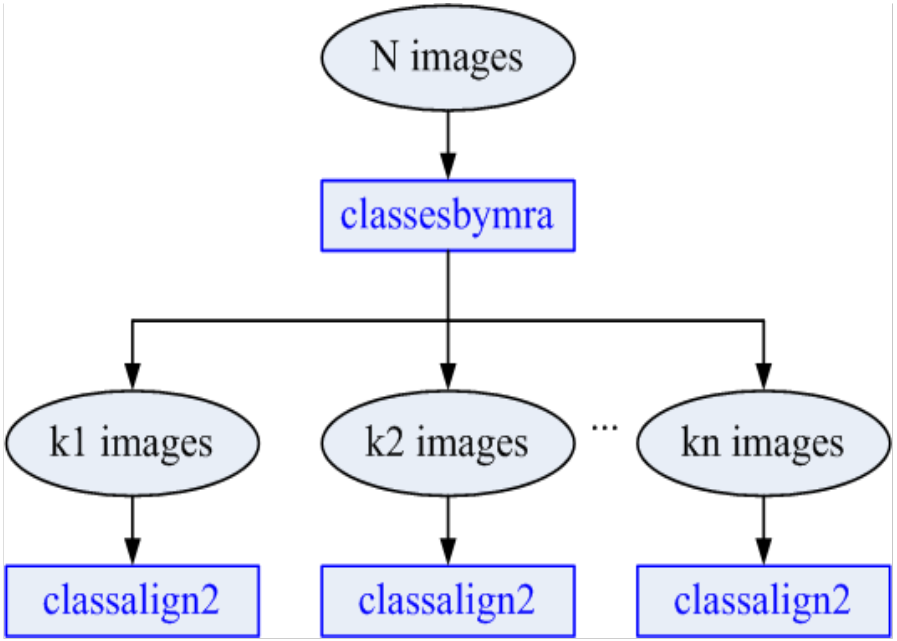


96.97%

Module	Ratio
Classalign2	**50.59%**
Classesbymra2	**24.86%**
Classesbymra1	**21.52%**
Make3d	1.61%
Project3d	0.65%
Projtree	0.03%
Proc3d	0.03%

三维重构高性能计算 — 核心问题

工作进展



n组图像数目差异很大

- **目标1：** N个图像的分类处理时间最短
- **目标2：** 尽可能使各处理器分配的任务平均

三维重构高性能计算 — 核心问题

工作进展

问题描述（Makespan Minimization)

- s：J→P （J = {1,2…n}, P={1,2…m})
- The finishing time of processor *i*

$$F_i(s) = \sum_{s(j)=i} p_j \qquad i = 1,2...m$$

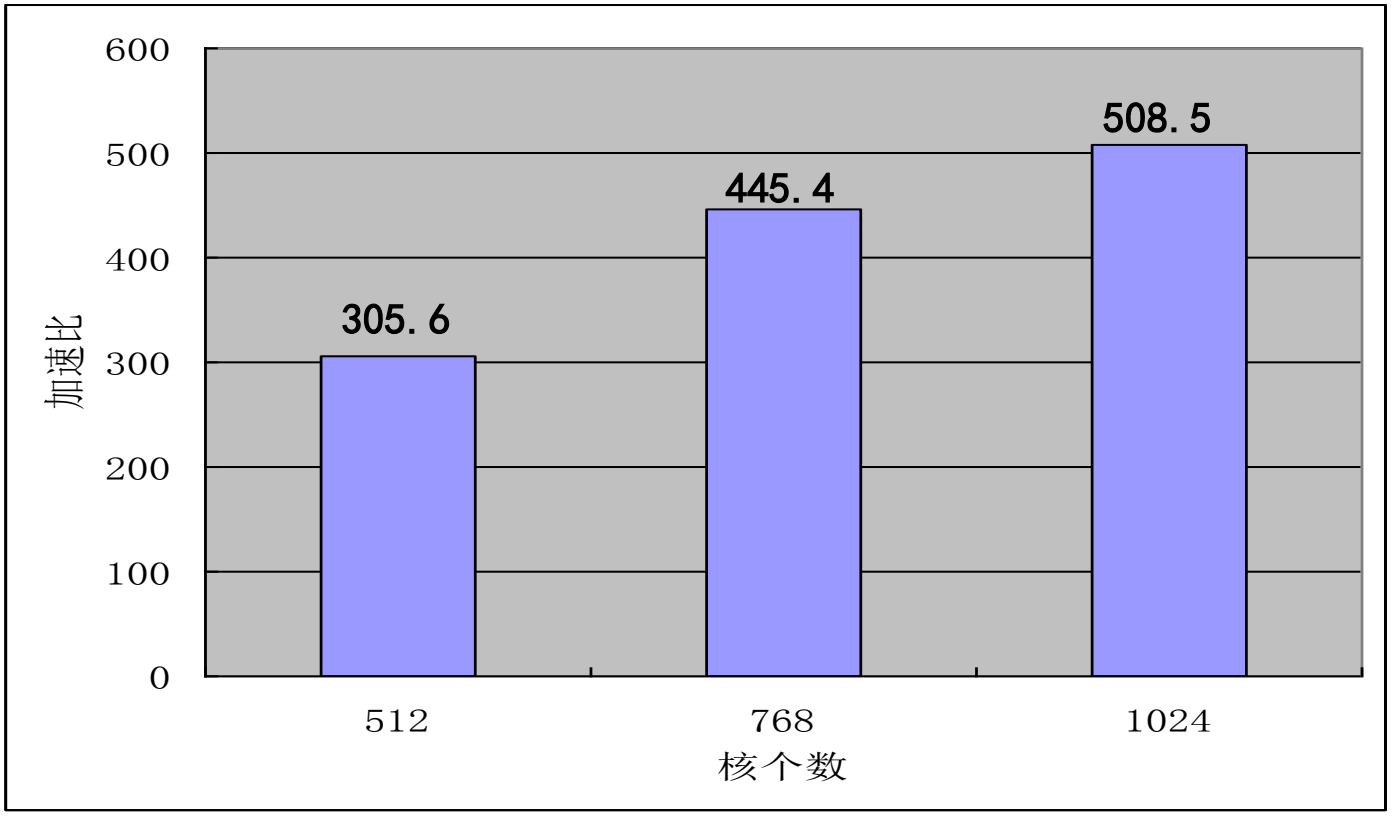
- The makespan is: $F(s) = \max_{1 \le i \le m} \{F_i(s)\}$
- Strongly NP-hard problem

我们提出了一种自适应的调度算法

三维重构高性能计算

工作进展

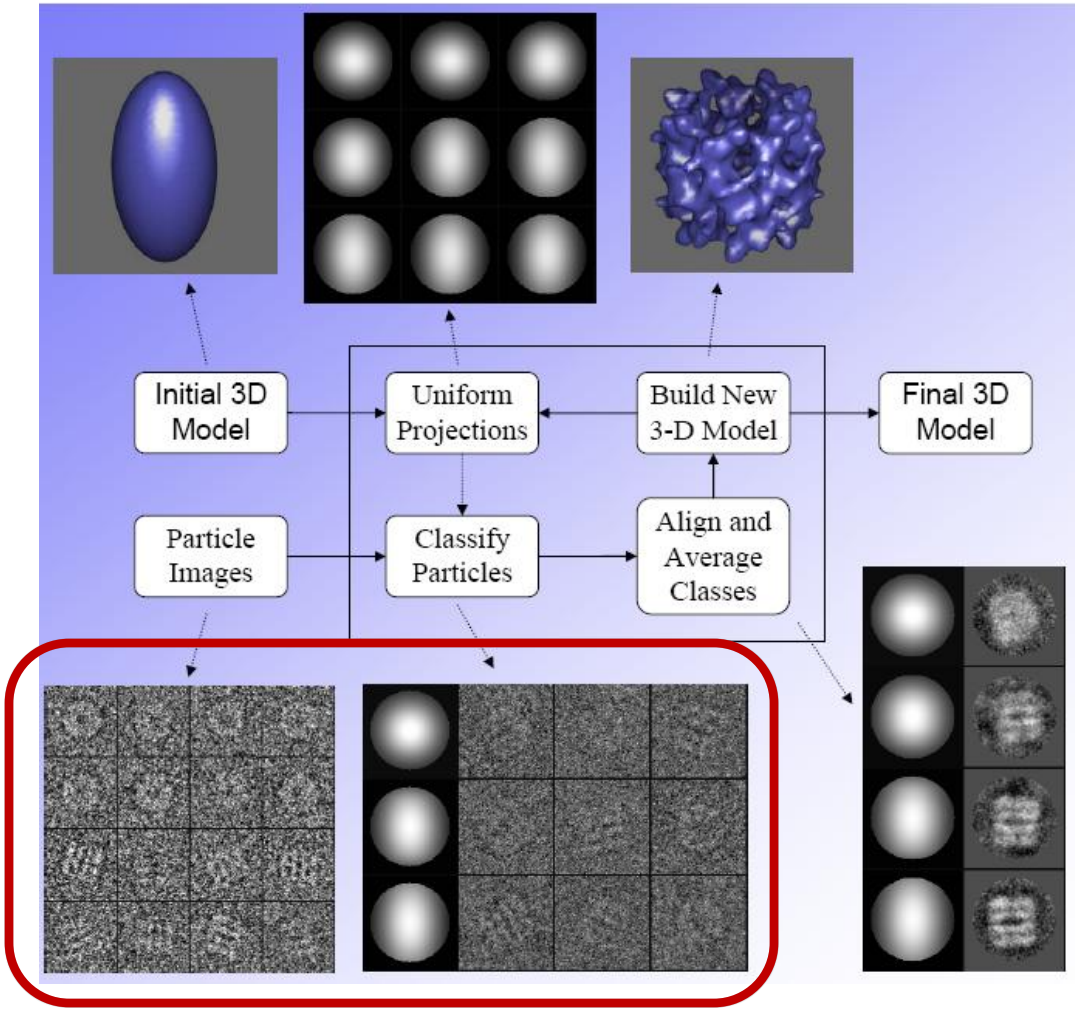
ParaEMAN曙光5000上的加速比



三维重构结果精度优化

工作进展

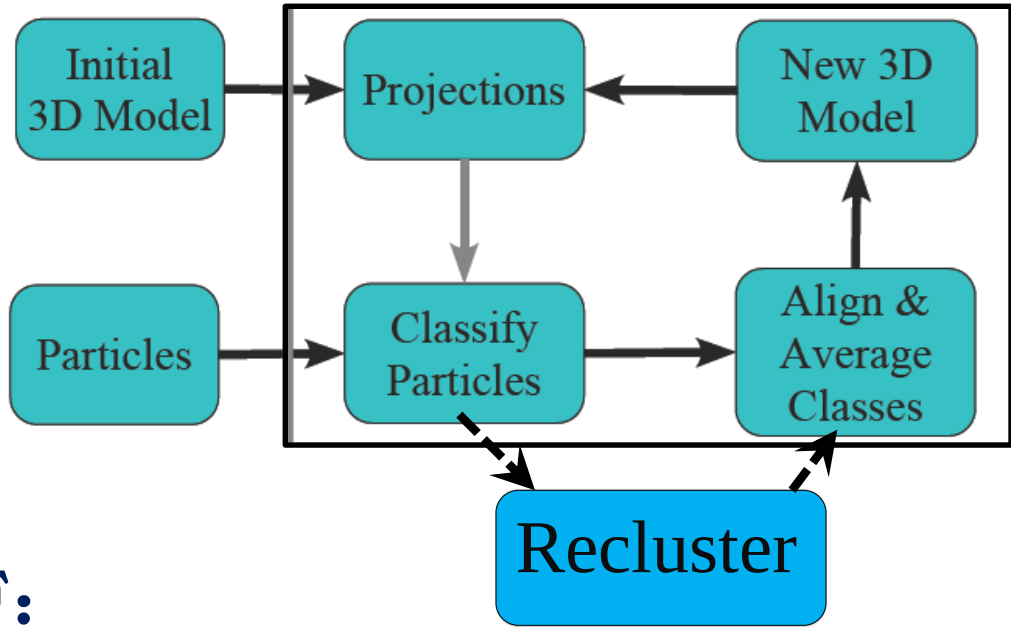
- 颗粒图像分类结果决定了重构的精度



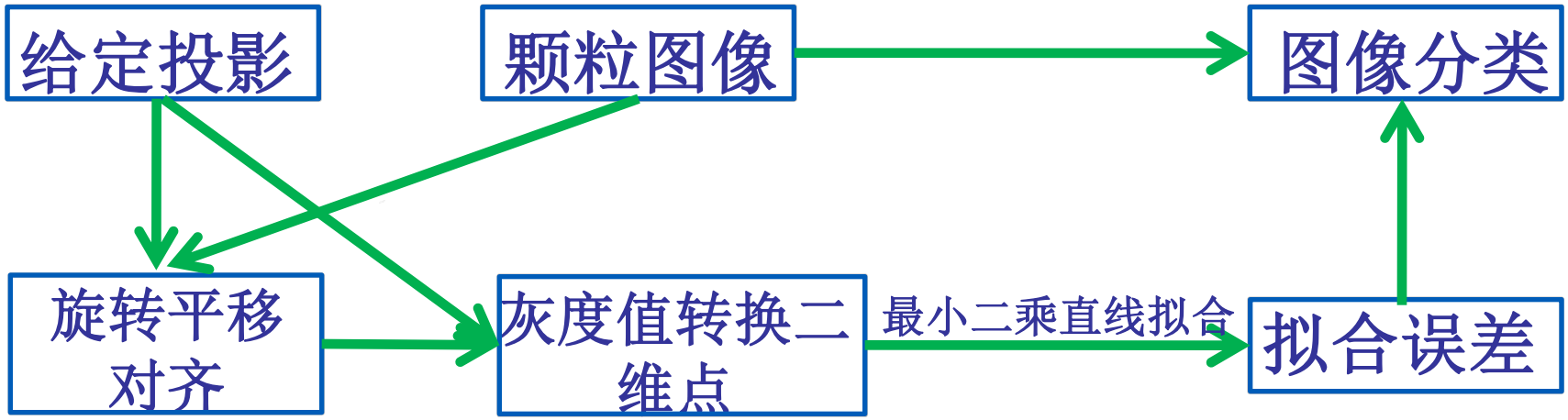
三维重构结果精度优化

工作进展

- 对原有算法进行了优化



颗粒图像重聚类框架PRF：

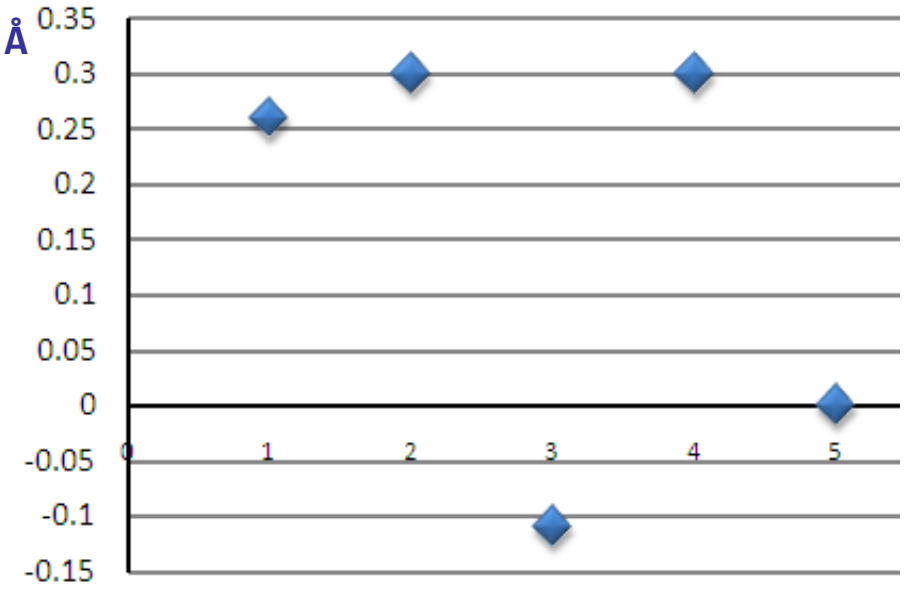


三维重构结果精度优化

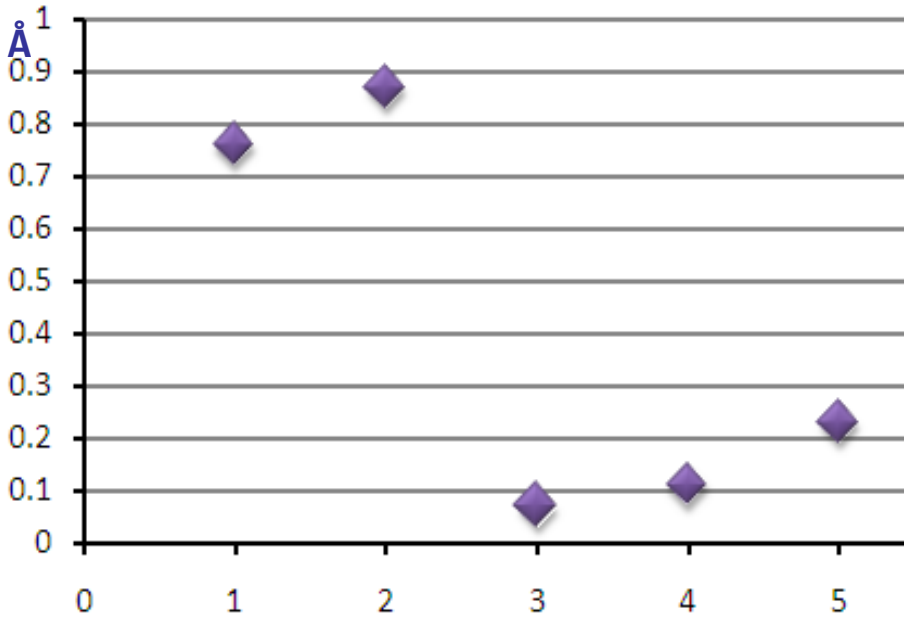
工作进展

乙肝病毒　refine 1 hard=15 sym=icos mask=80　pad=374　proc=3
classkeep=0.9 ang=1 classiter=5　refine xfiles=2.5 ctfc=7

Beta1.1　refine 1 mask=60 sym=c9 proc=3 hard=40 classkeep=0.9
shrink=2 pad=200 amask=55,0.8,30 classiter=5 ang=8



乙肝病毒数据



Beta1.1数据

基于球谐函数的三维重构算法

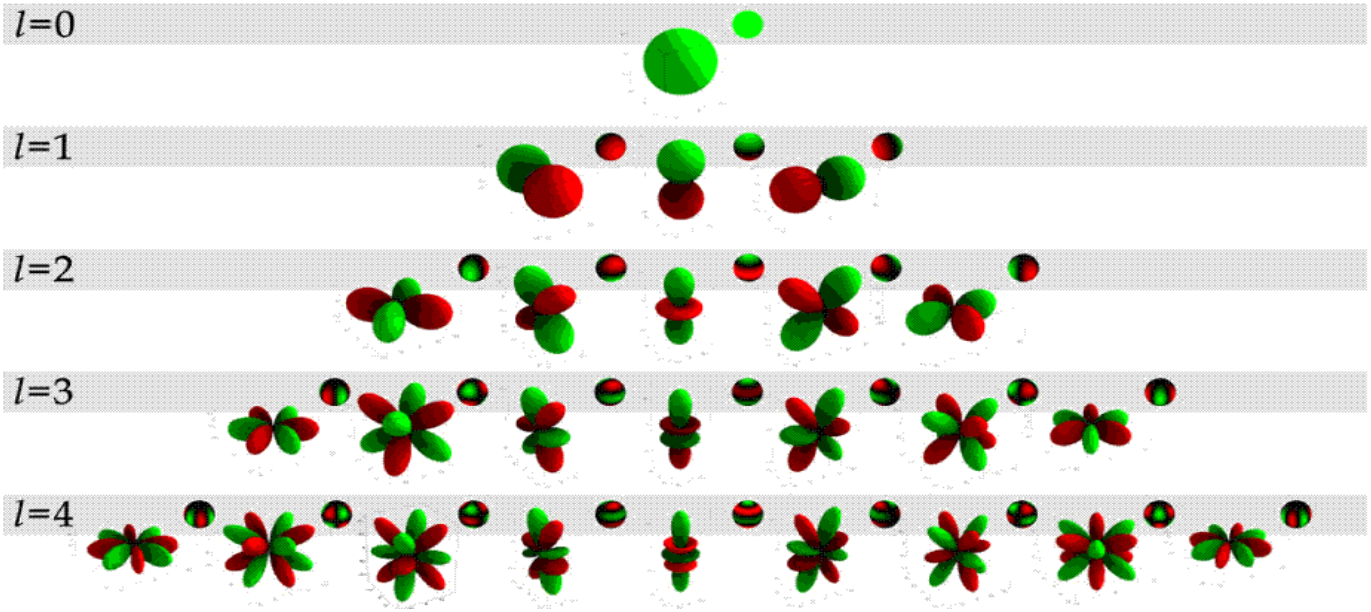
工作进展

理论基础

- 任何空间几何体表面都可以表示成球谐函数各分量的线性组合

$$R(\theta_i,\phi_i)=\sum_{j=0}^{(M+1)^2} c_j Y_j(\theta_i,\phi_i)$$

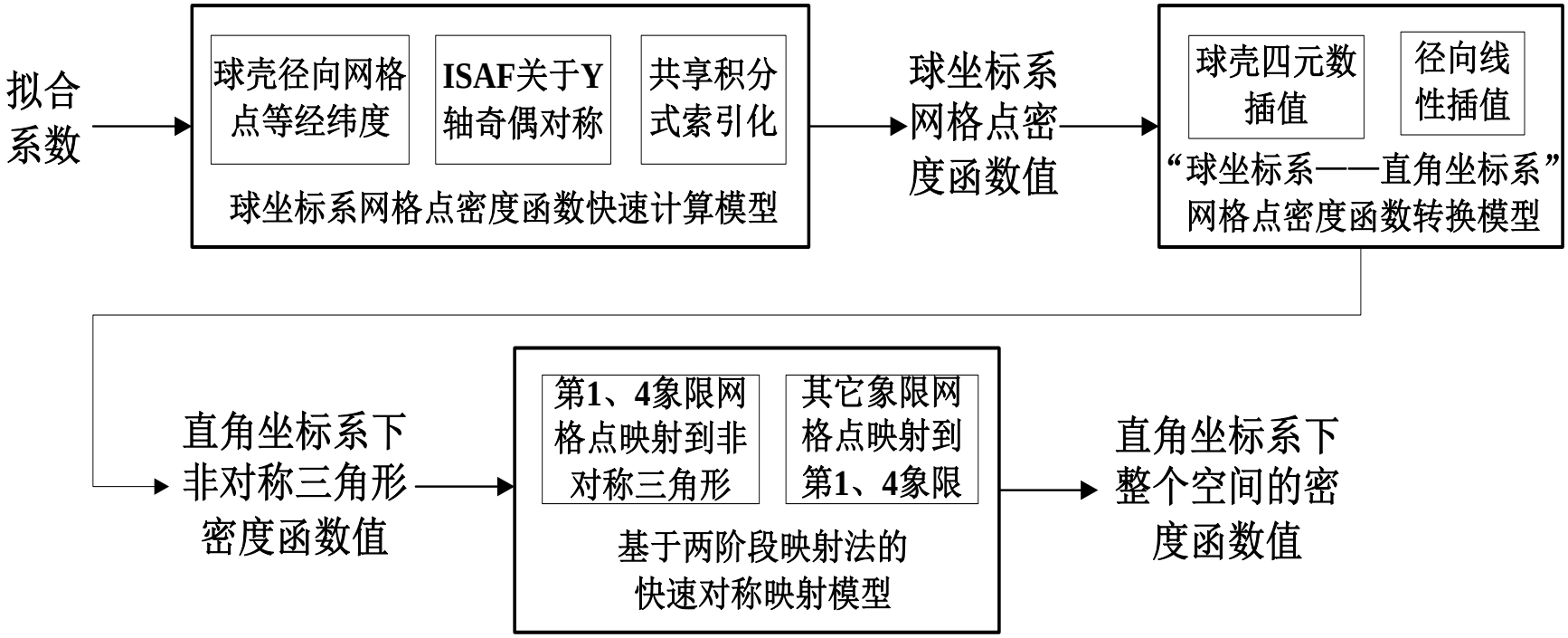
- 0到4阶球谐函数的图形



基于球谐函数的三维重构算法

工作进展

快速ISAF三维重构模型

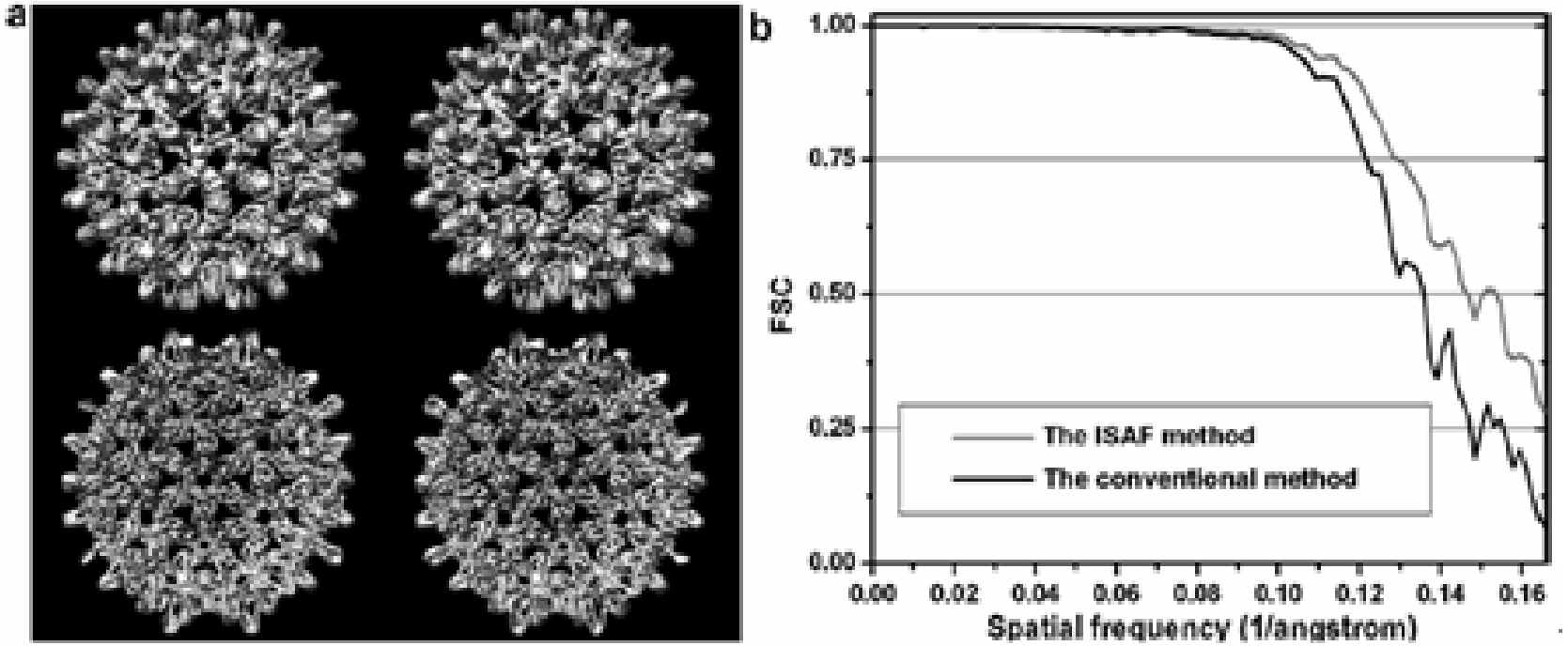


- 开发完成了ISAF三维重构系统ISAFResys

基于球谐函数的三维重构算法

工作进展

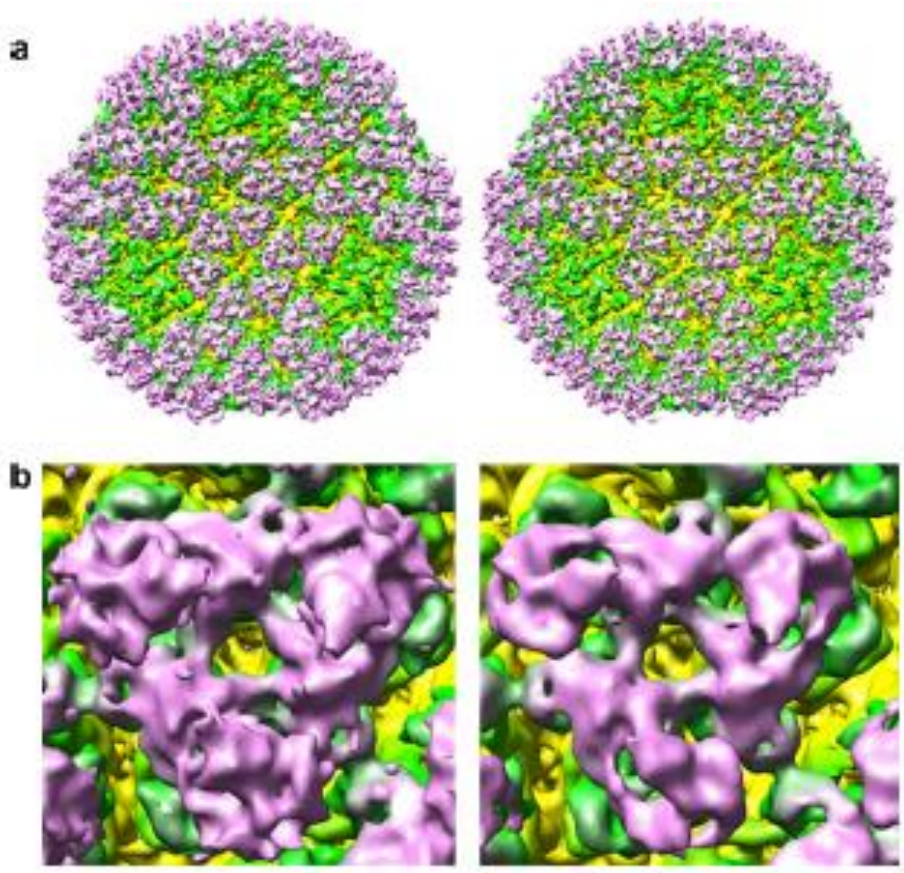
HBV乙肝病毒(HBV)重构

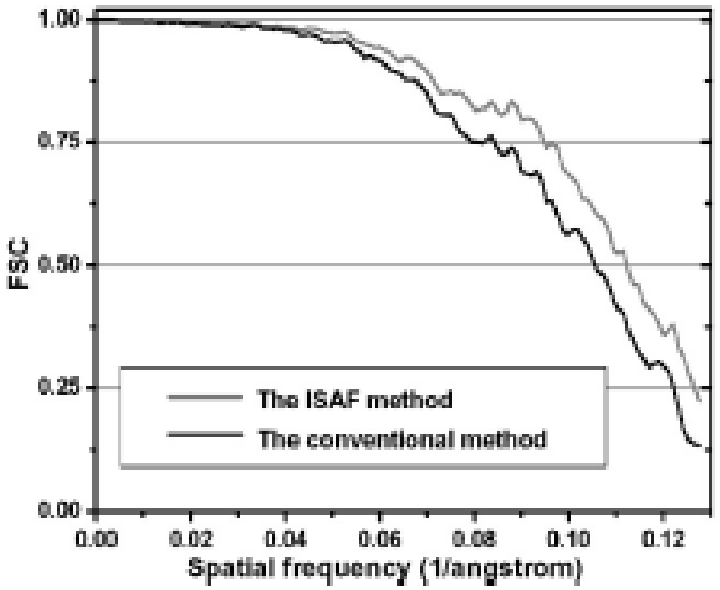


基于球谐函数的三维重构算法

工作进展

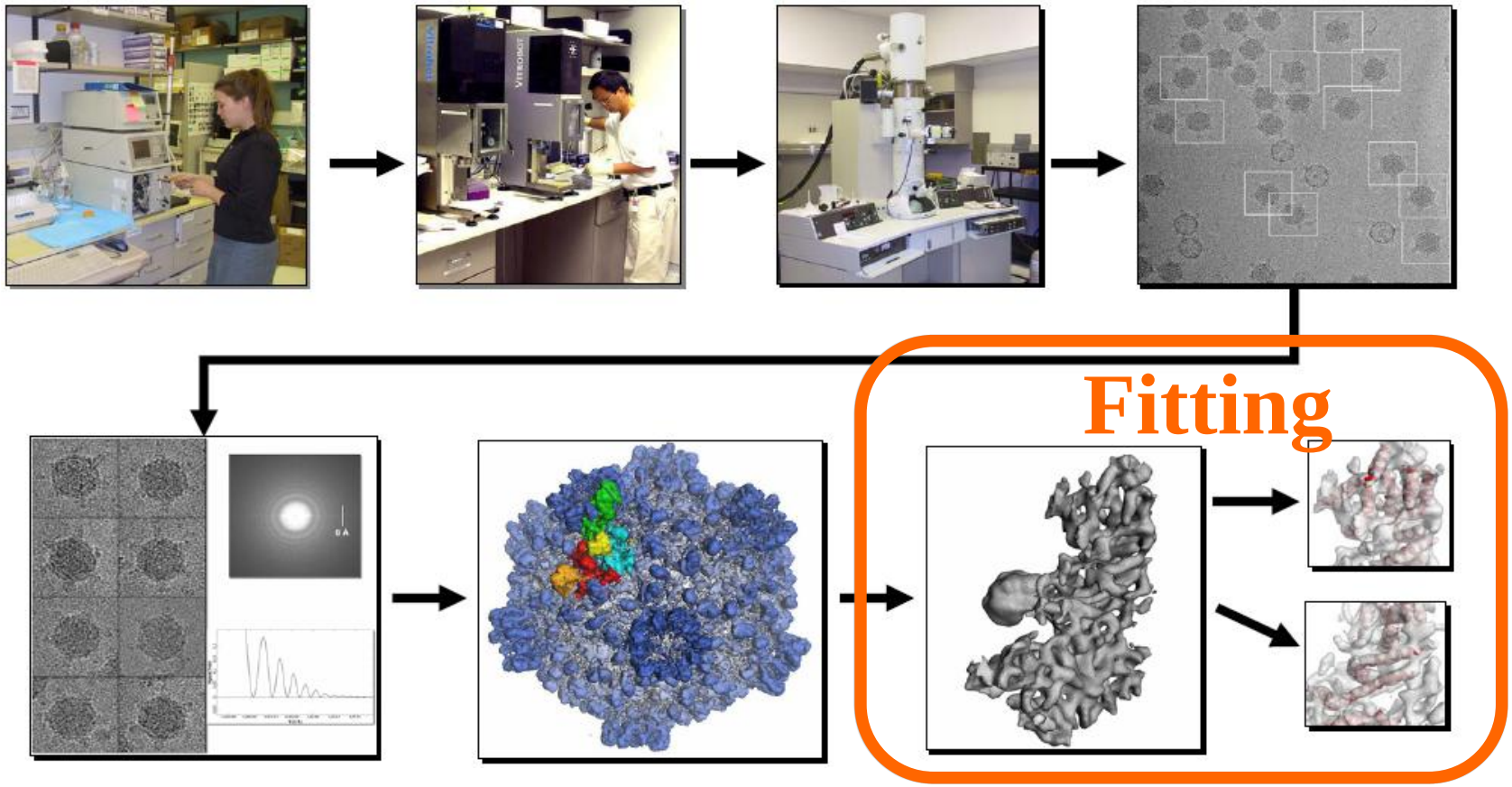
鲤鱼呼吸肠道病毒(GCVR)重构





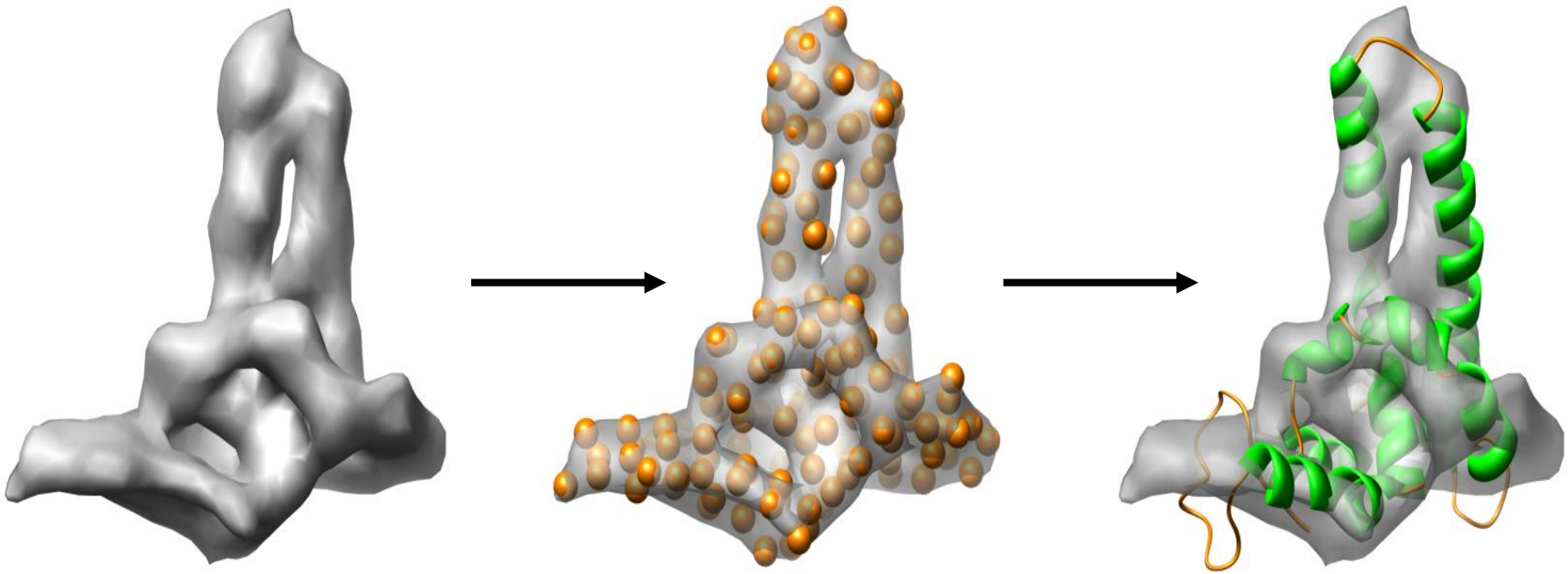
研究趋势

- I've got a structure, what next?



研究趋势

- **I’ve got a structure, what next?**



Modeling + docking

致谢

孙飞研究员 （生物物理所）
张燕博士（生物物理所）

Prof. Steven Ludtke (NCMI)
Prof. Albert Lawrence (UCSD)

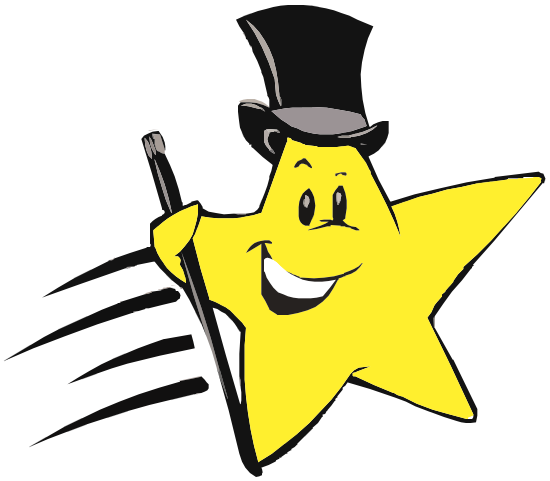
课题组成员：

陈翔博士 助理研究员
樊莉亚博士　邵书伟
王功明　万晓华
储琪　王林
贾琰　侯晨颖

杨启斌教授 （湘潭大学）
刘洪荣博士 （湘潭大学）

中国科学院
CHINESE ACADEMY OF SCIENCES

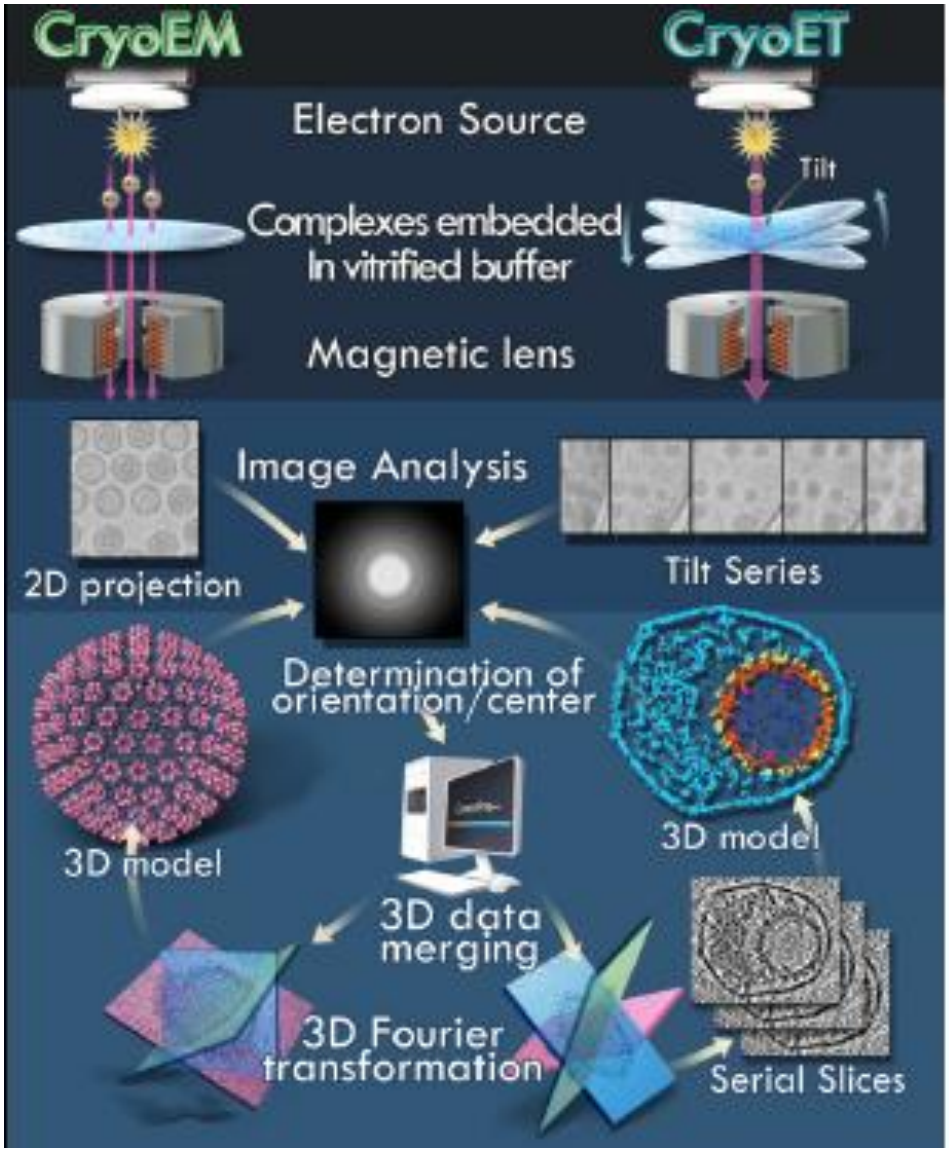
NSFC



THANKS

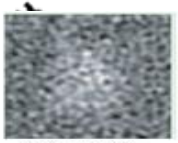
冷冻电镜三维重构

背景简介



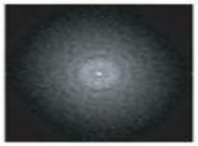
图像获取
- 颗粒挑选
- 噪声剔除
- 归一化
- 欠采样处理

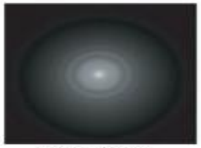

电子显微镜图片


颗粒图像

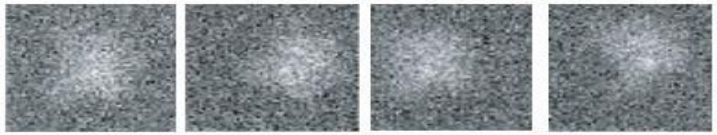

降噪处理

CTF修正
- 功率谱密度(PSD)估计
- 1-D/2-D 理论CTF模型
- CTF-相位修正


PSD估计


CTF模型

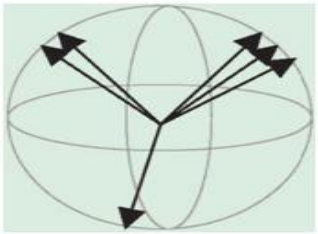
2-D分析
- 旋转和平移调整
- 旋转光谱


蛋白颗粒图像旋转和平移调整

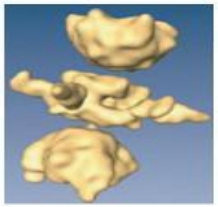
图像分类
- 特征提取
- 硬聚类和模糊聚类
- 维数消减
- 自组织映射


图像的分类和自组织映射

角度测定
- 中央截面定理
- 并行处理
- Radon空间
- 小波空间
- Fourier空间


确定每个颗粒图像的投影方向

3-D重构
- 直接傅里叶求逆
- 加权反投影
- 迭代优化
- 并行加速


三维重构模型

国际最新进展

工作简介

Algorithms/Criteria for Particle Selection			
分类	典型算法	优点	缺点
基于模板匹配	Bern's algorithm Ludtke's algorithm Penczek's algorithm Roseman's algorithm Sigworth's algorithm	手工挑选模板 借助3D参照结构 实现简单	如何保证模板无遗漏 结果假阳性高
基于特征	Bajai's algorithm Hall's algorithm Mallick's algorithm Volkmann's algorithm Zhu's algorithm	少量样本集 利用几何与统计信息 自动化程度高	特征提取困难 低对比度图像效果差

数据来自 "Nature Methods" 2008, 5(7): 651- 658
"Journal of Structural Biology" 2004, 145

颗粒图像挑选

工作进展

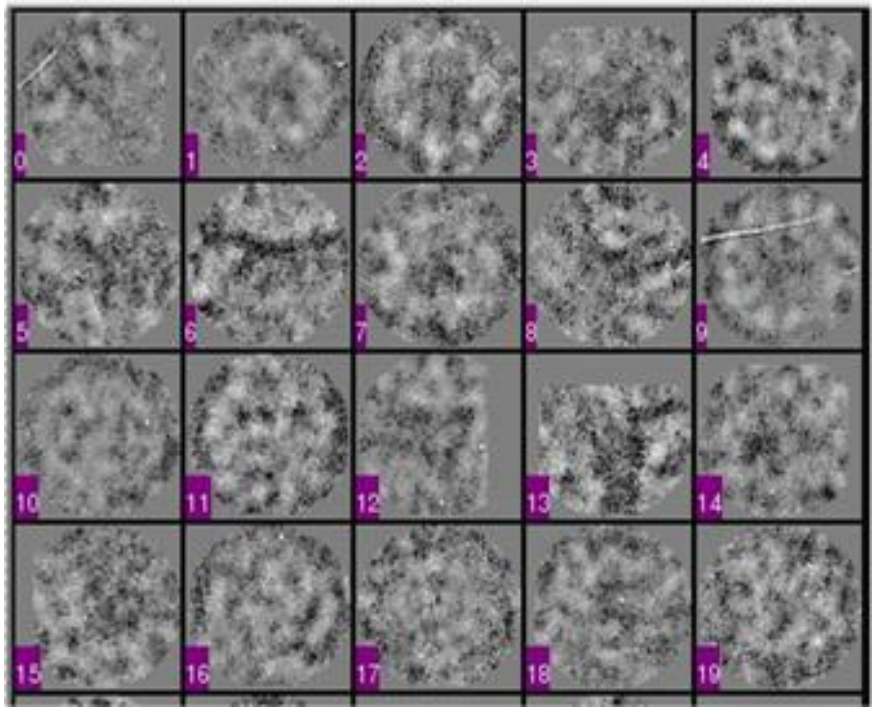
“Journal of Structural Biology” 145 （1-2）的统计结果

Test / Truth		Bajaj	Bern	Mouche (Manual)	Haas (Manual)	Hall	Ludtke	Mallick	Penczek	Roseman	Sigworth	Volkmann	Zhu
Bajaj (1269)			33.9 11.5	24.7 8.3	31.0 7.0	42.2 24.3	51.9 21.0	28.0 9.8	52.9 25.2	17.4 14.0	37.4 5.1	38.5 9.2	24.0 11.4
Bern (948)		11.5 33.9		16.2 23.8	21.5 21.0	36.3 37.7	43.1 30.3	17.7 23.1	48.4 38.8	10.3 30.3	26.4 16.7	29.9 22.8	17.1 28.0
Mouche (1042)		8.3 24.7	23.8 16.2		11.7 2.3	27.4 22.0	43.4 23.7	14.2 11.7	46.8 30.7	2.4 16.6	23.2 4.5	27.4 12.2	9.7 13.7
Haas (944)		7.0 31.0	21.0 21.5	2.3 11.7		26.2 28.2	41.1 28.4	12.2 18.4	44.0 33.9	1.5 23.9	18.4 8.4	22.9 15.7	8.8 21.3
Hall (969)		24.3 42.2	37.7 36.3	22.0 27.4	28.2 26.2		52.0 39.9	30.1 33.2	55.9 46.6	19.3 35.8	35.3 25.2	39.3 31.7	25.7 33.7
Ludtke (775)		21.0 51.9	30.3 43.4	23.7 43.4	28.4 41.1	39.9 52.0		23.0 41.2	48.3 50.0	20.3 49.4	27.1 32.7	32.3 39.1	23.5 45.4
Mallick (1015)		9.8 28.0	23.1 17.7	11.7 14.2	18.4 12.2	33.2 30.1	41.2 23.0		46.7 32.5	7.0 22.6	25.8 10.3	30.1 17.9	14.5 20.5
Penczek (799)		25.2 52.9	38.8 48.4	30.7 46.8	33.9 44.0	46.6 55.9	50.0 48.3	32.5 46.7		23.7 50.0	38.4 41.3	39.7 44.0	30.2 49.1
Roseman (1219)		14.0 17.4	30.3 10.3	16.6 2.4	23.9 1.5	35.8 19.3	49.4 20.3	22.6 7.0	50.0 23.7		33.1 2.7	34.9 7.8	17.5 7.8
Sigworth (838)		5.1 37.4	16.7 26.4	4.5 23.2	8.4 18.4	25.2 35.3	32.7 27.1	10.3 25.8	41.3 38.4	2.7 33.1		12.3 14.6	6.8 28.1
Volkmann (861)		9.2 38.5	22.8 29.9	12.2 27.4	15.7 22.9	31.7 39.3	39.1 32.3	17.9 30.1	44.0 39.7	7.8 34.9	14.6 12.3		11.5 30.0
Zhu (1109)		11.4 24.0	28.0 17.1	13.7 9.7	21.3 8.8	33.7 25.7	45.4 23.5	20.5 14.5	49.1 30.2	7.8 17.5	28.1 6.8	30.0 11.5	
Median/Mean	FNR	11.4/13.3	28.0/27.9	16.2/16.2	21.5/22.0	43.4/44.5	33.7/34.4	20.5/20.8	48.3/47.9	7.8/10.9	27.1/28.0	30.1/30.7	17.1/17.2
	FPR	33.9/34.7	21.5/25.3	23.2/21.7	18.4/18.7	27.1/28.9	30.1/33.6	23.1/23.8	33.9/35.4	30.3/29.8	10.3/15.1	15.7/20.6	28.0/26.3
Standard Deviation	FNR	7.0	7.1	8.6	8.0	6.0	6.7	7.3	4.1	7.9	7.7	8.0	7.8
	FPR	11.3	12.8	14.2	14.4	8.6	11.9	13.0	8.2	12.4	12.7	12.4	13.2

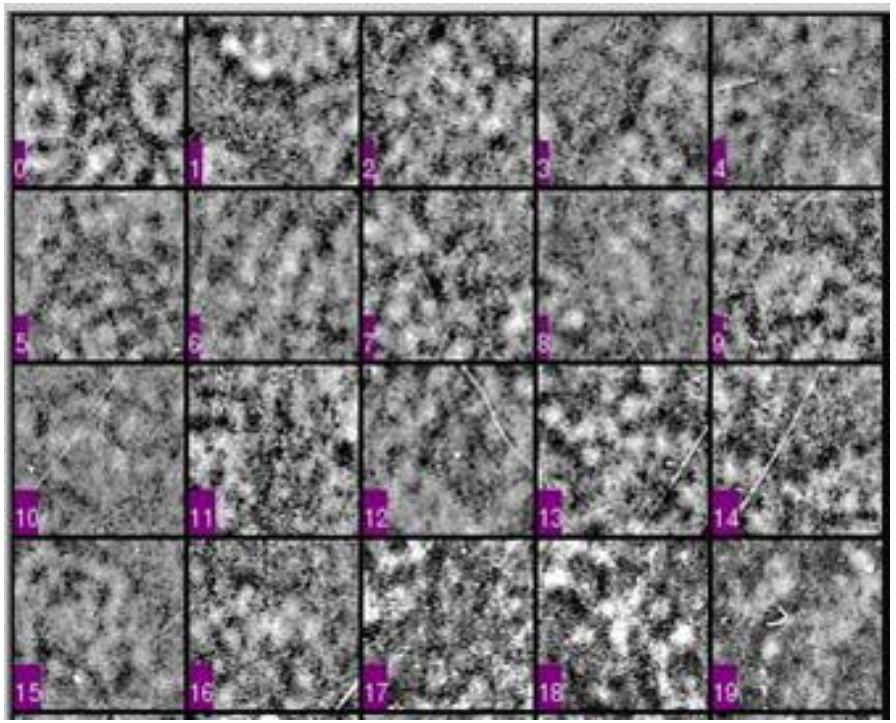
颗粒图像信噪比极低

科学问题

Good

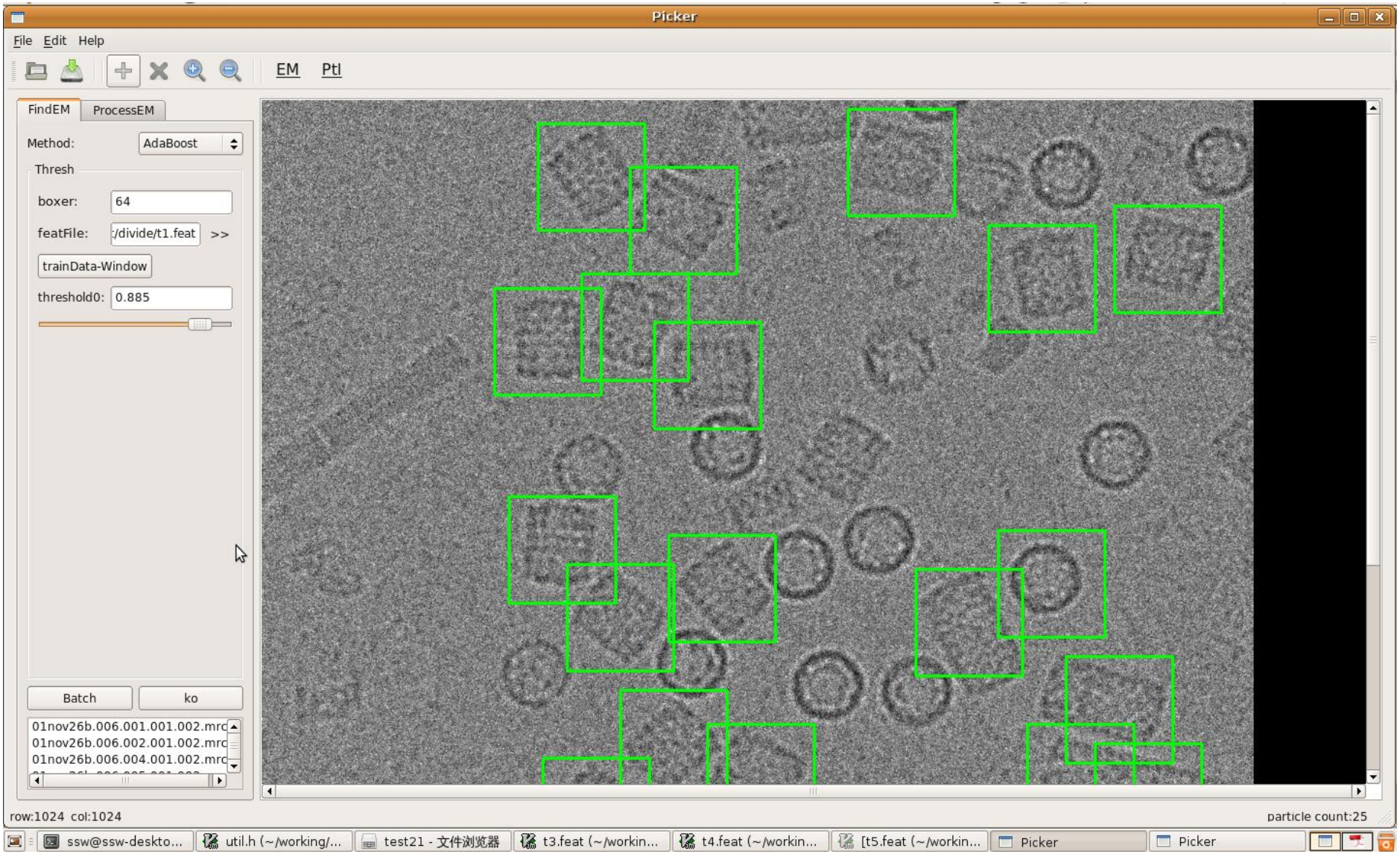


Bad



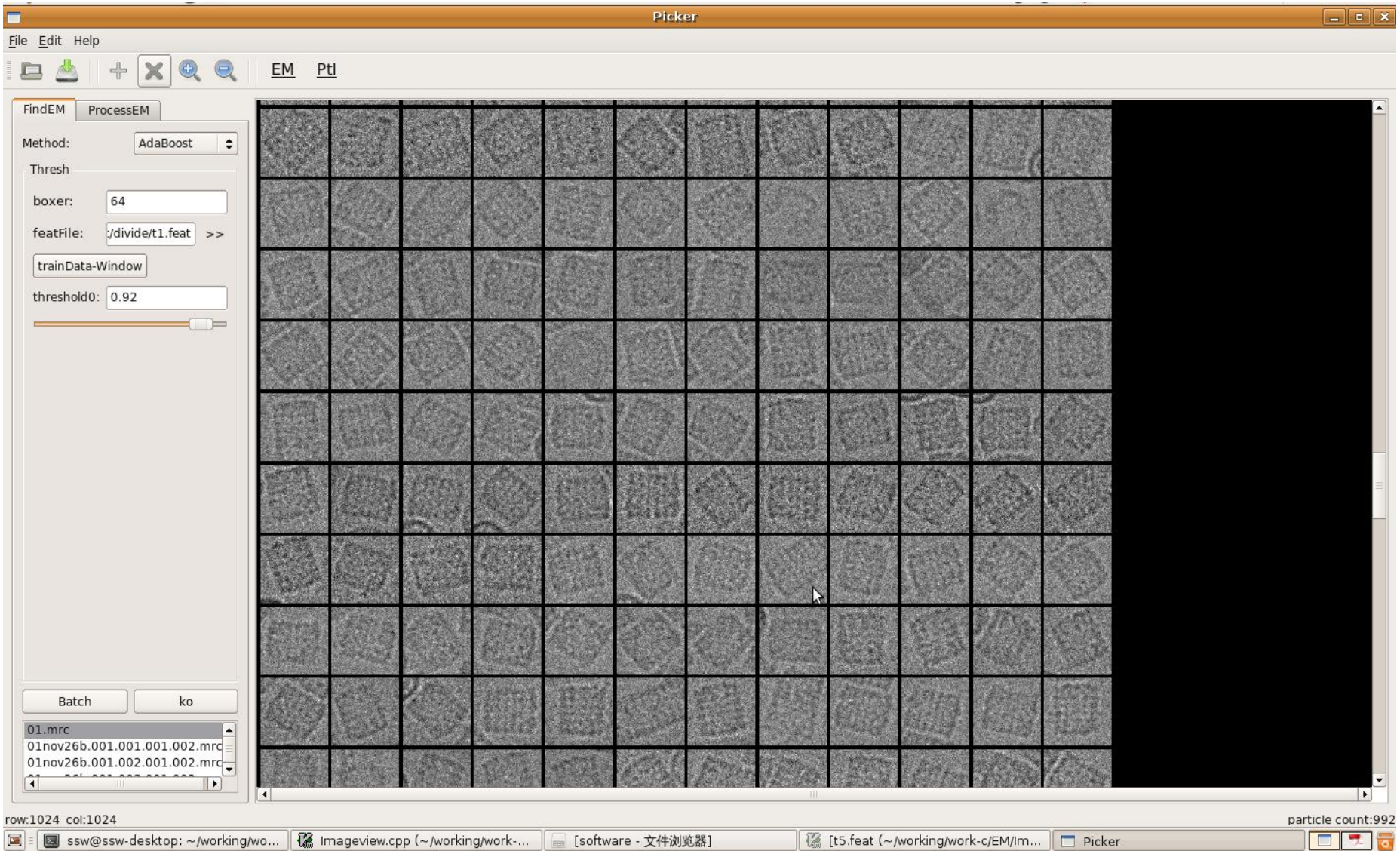
颗粒图像挑选

工作进展


Picker
File Edit Help
EM Ptl
FindEM ProcessEM
Method: AdaBoost
Thresh
boxer: 64
featFile: /divide/t1.feat >>
trainData-Window
threshold0: 0.885
Batch ko
01nov26b.006.001.001.002.mrc
01nov26b.006.002.001.002.mrc
01nov26b.006.004.001.002.mrc
row:1024 col:1024
particle count:25
ssw@ssw-deskto...
util.h (~/working/...
test21 - 文件浏览器
t3.feat (~/workin...
t4.feat (~/workin...
[t5.feat (~/workin...
Picker
Picker

颗粒图像挑选

工作进展


Picker
File Edit Help
EM Ptl
FindEM
ProcessEM
Method:
AdaBoost
Thresh
boxer:
64
featFile:
:/divide/t1.feat
>>
trainData-Window
threshold0:
0.92
Batch
ko
01.mrc
01nov26b.001.001.001.002.mrc
01nov26b.001.002.001.002.mrc
row:1024 col:1024
particle count:992
ssw@ssw-desktop: ~/working/wo...
Imageview.cpp (~/working/work-...
[software - 文件浏览器]
[t5.feat (~/working/work-c/EM/Im...
Picker

自适应动态调度算法-- SADS

并行加速

自适应动态调度算法 -- SADS

1. Determining the processing time;
2. Minimizing the makespan;
3. Solving the sub-set sum problem (Integer Non-linear Programming)

针对EMAN的两个简化：

1. Target value has an upper bound polynomial in the input size.

$$p_i = \frac{1}{i}\sum_{j \in s_i} p_j \le \sum_{j \in s_i} p_j \le \sum_{j=1}^{n} p_j = \sum_{j=1}^{n} (ak_j + b) = aN + b$$

2. Processing time p_j can be rounded to an integer.

2、自适应动态调度算法-- SADS 并行加速

SADS

1. Set initial values, $a^{(0)} = 1$, $b^{(0)} = 0$, , and $p_i^{(0)} = k_i^{(0)}$, i=1,2…n

2. For k=0 to ITER – 1 do
   - 2.1 Distribute n tasks to m processors by means of estimated values of $p_i^{(k)}$
   - 2.2 Run classalign2 for each task and record the actual processing time $t_i^{(k)}$.
   - 2.3 Update values of a and b by means of the least square method.

$$a^{(k+1)} = \frac{n\sum_{i=1}^{n} k_i^{(k)} t_i^{(k)} - \sum_{i=1}^{n} k_i^{(k)} \sum_{i=1}^{n} t_i^{(k)}}{n\sum_{i=1}^{n} (k_i^{(k)})^2 - (\sum_{i=1}^{n} k_i^{(k)})^2} \quad b^{(k+1)} = \frac{\sum_{i=1}^{n} t_i^{(k)} \sum_{i=1}^{n} (k_i^{(k)})^2 - \sum_{i=1}^{n} k_i^{(k)} \sum_{i=1}^{n} k_i^{(k)} t_i^{(k)}}{n\sum_{i=1}^{n} (k_i^{(k)})^2 - (\sum_{i=1}^{n} k_i^{(k)})^2}$$

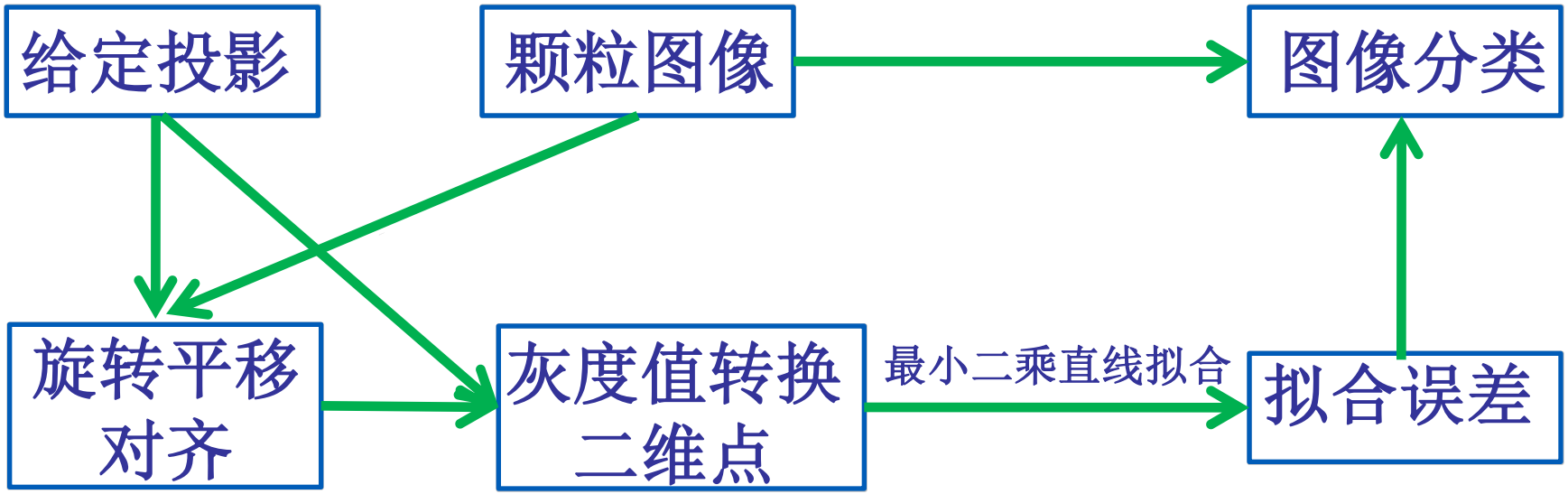
   - 2.4 Estimate processing time of the next round:

$$p_j^{(k+1)} = a^{(K+1)} K_j^{(k+1)} + b^{(k+1)}, \qquad j = 1,2...n$$

End

3、算法优化

- 重构结构的优化
  - 提出一种颗粒图像再分类方法



基于学习(adaboost)方法

- 学习流程
  - For t=1,...,T
    1. 归一化权重，使得w_t为一个概率分布 $w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
    2. 对每个特征j，训练一个弱分类器h_j
       并计算其带权重的错误率

$$\varepsilon_j = \sum_{i=1}^{n} w_{t,i} \left| h_j(x_i) - y_i \right|$$

    3. 选择误差最小的弱分类器h_t
    4. 更新每个样本的权重 $w_{t,i} = w_{t,i}\beta_t^{1-e_i}, \beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$

其中：x_i被正确分类，e_i=0，否则e_i=1

- 正确分类则权重下降，否则不变；错误率越小，β 也越小