Optimizing image processing

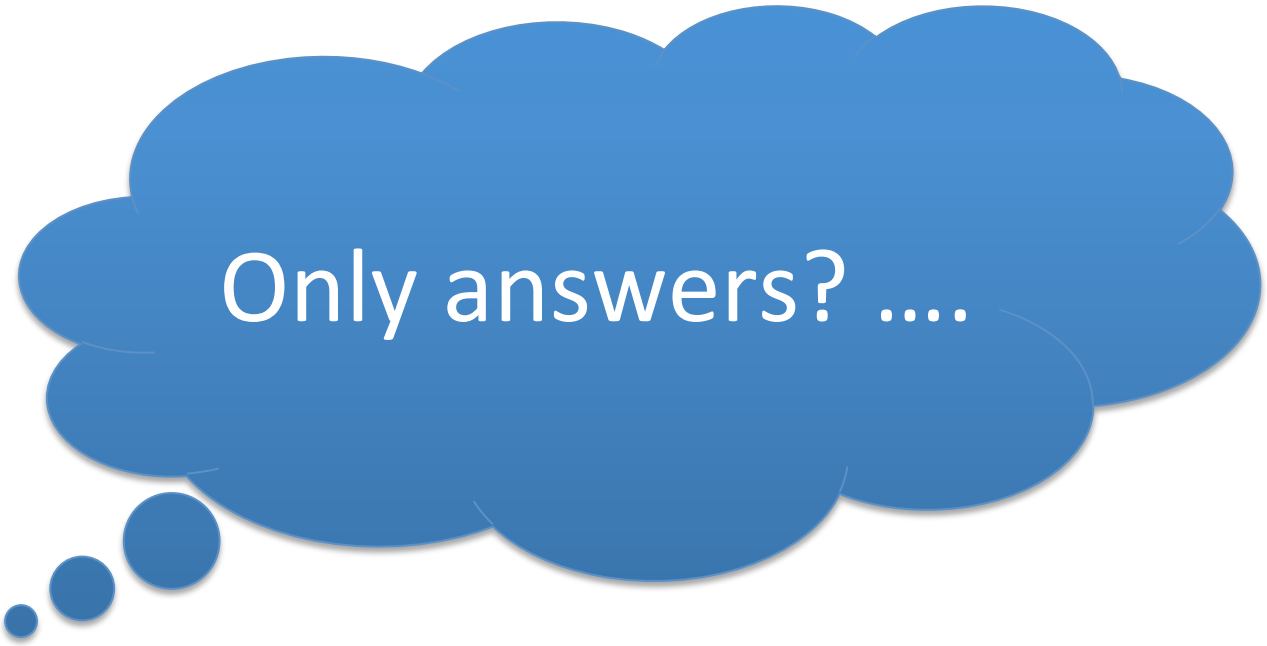
Sjors Scheres
MRC Laboratory of Molecular Biology

What if one would have many noisy 2D projections of assumedly identical 3D objects in unknown orientations, and one would want to know that 3D structure?

And how strictly adhering to theory helps...

Questions?

Conventional and ML 3D (projection matching) refinement (and the differences between them), Bayesian extension, how to avoid overfitting, how to avoid model bias, multi- reference refinement (classification).



Cryo-EM inconveniences

- Electrons damage biological mat rial
  - Low dose: large am s of
- Incomplete, ill-posed inverse problem
- particles fall on the grid
  - Unkn n orientations & classes

Inverse problems

The forward model

$$X_i = \mathrm{CTF}_i \mathbf{P}_\phi V_k + N_i$$

Given V and CTF, we can simulate X very well

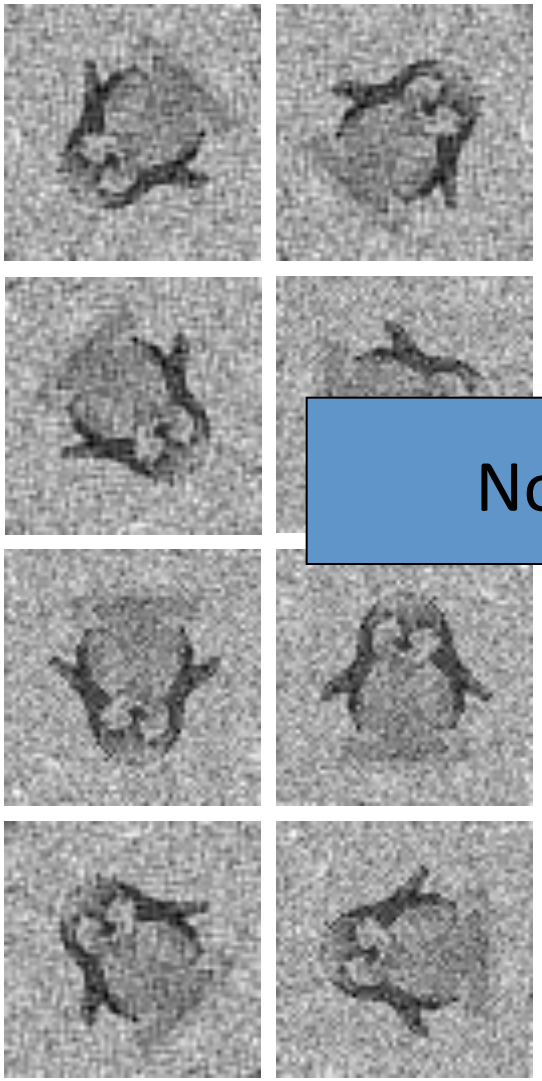
But the other way
around is more difficult!

Incompleteness

Incomplete data problems

- Part of the data was not observed experimentally
  - Orientations
  - Class assignments
- Difficult to solve!
  - Iterative methods?
- Complete data problem would be very easy to solve
- (Another famous one: the phase problem in XRD)

Incomplete data problems

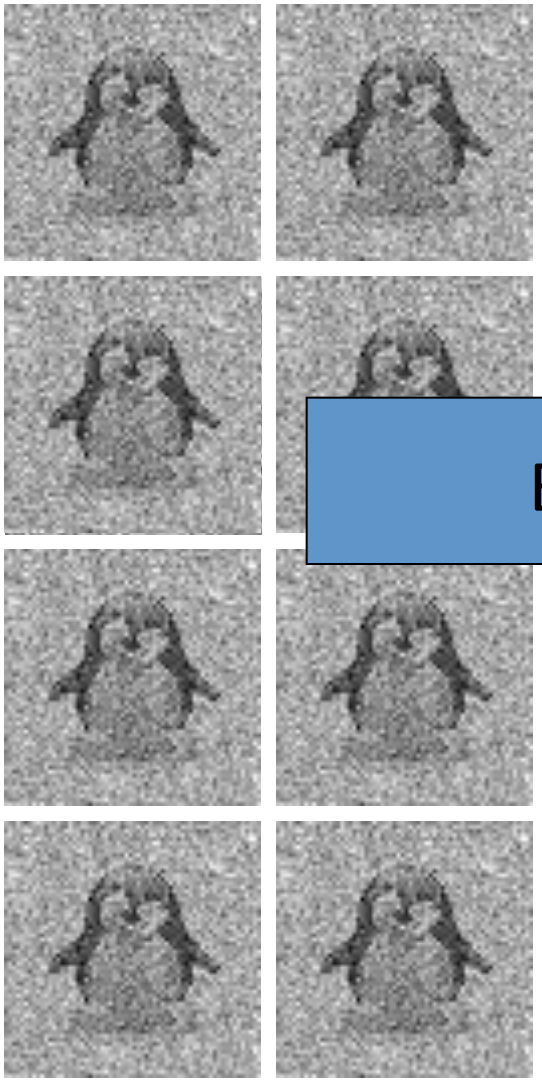


Not easy

Observed data (*X*): images

Missing data (Y): orientations

Complete data problems



white Gaussian noise

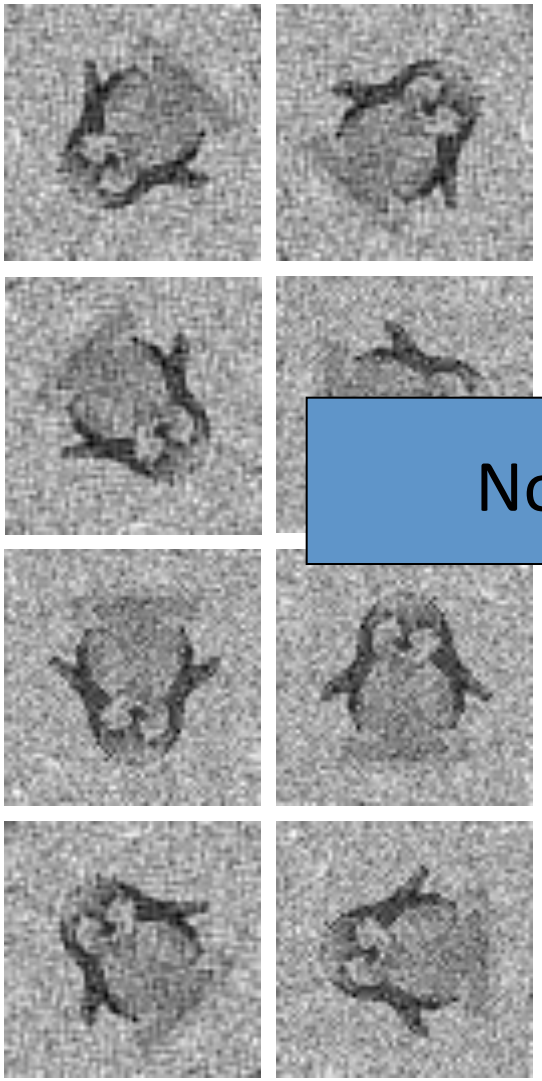
$$L(\Theta) = P(X \mid \Theta)$$

Easy!

$$A^{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Observed data (*X*): images

Incomplete data problems



Not easy

Observed data (*X*): images

Missing data (Y): orientations

Incomplete data problems

- Option 1: add *Y* to the model → Maximum cross-correlation / least-squares

$$L(Y,\Theta) = P(X \mid Y,\Theta)$$

- Option 2: marginalize over *Y* → Maximum Likelihood

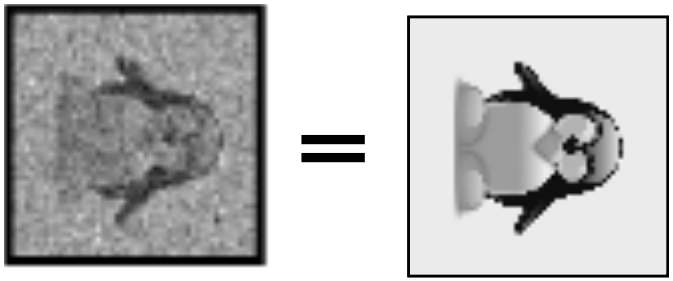
$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y,\Theta) P(Y \mid \Theta) d\phi$$

Probability of X, regardless Y

The maxCC approach

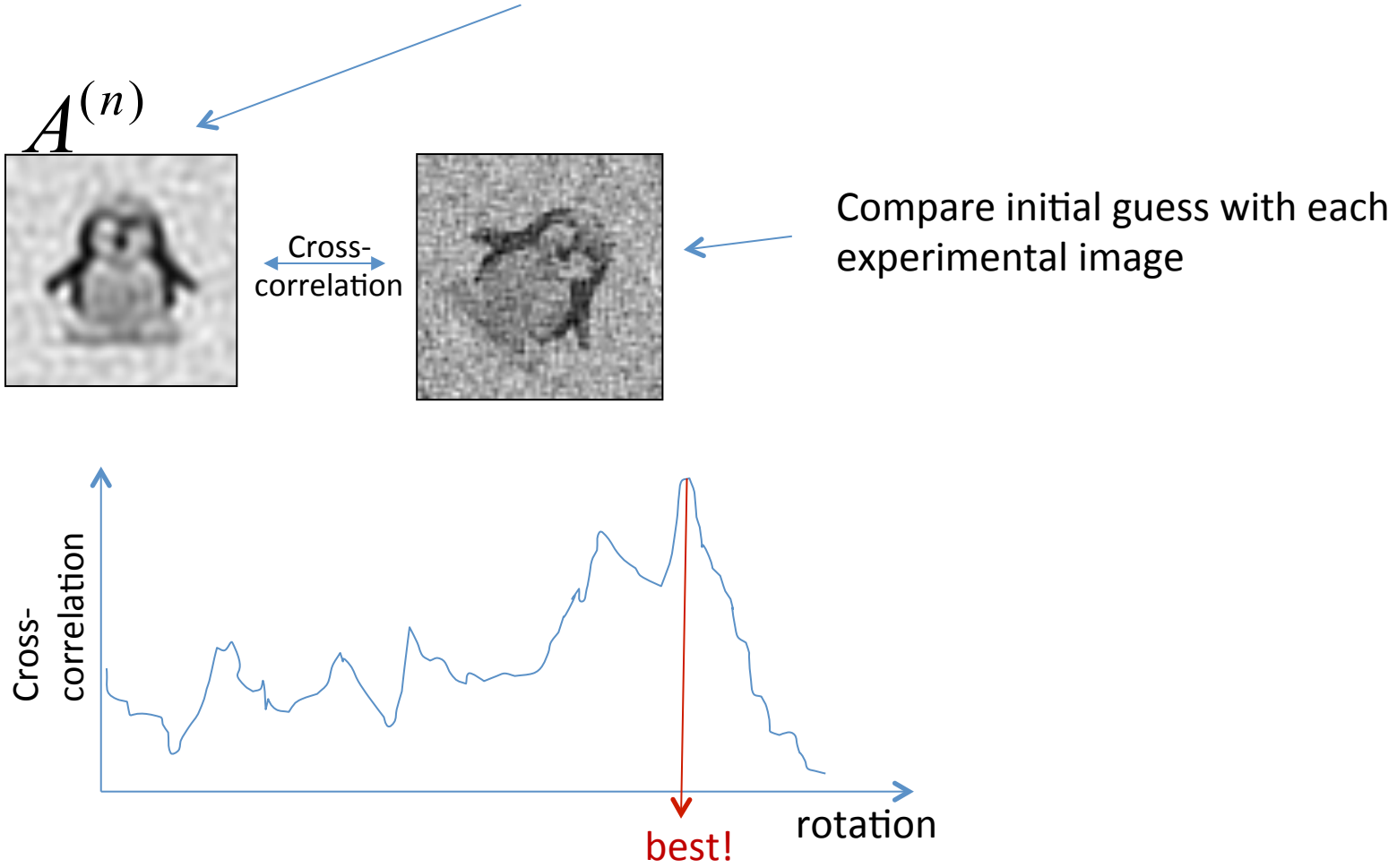
Statistical data model

$$X_i = P_\varphi V_k$$

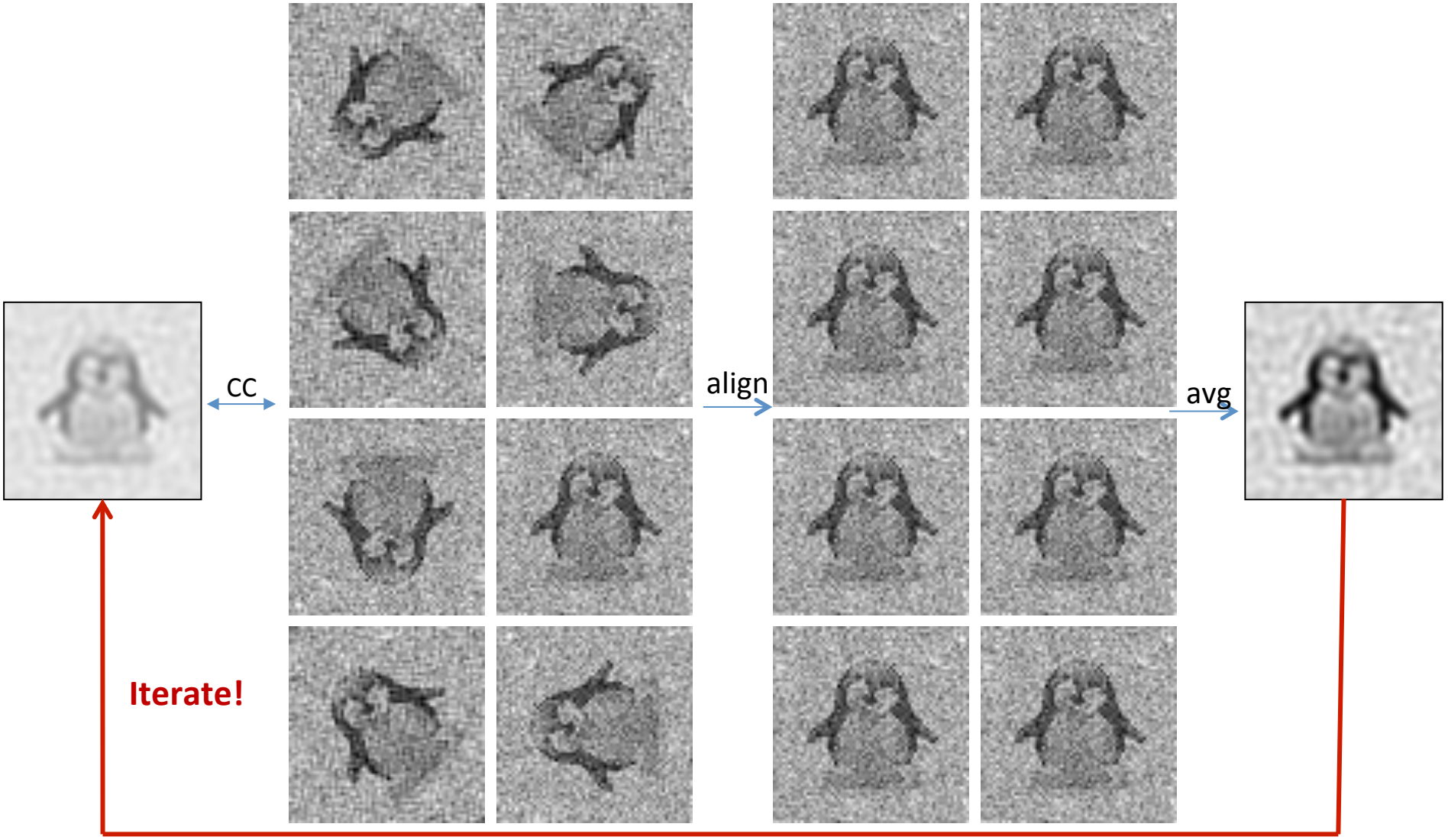


Reference-based alignment

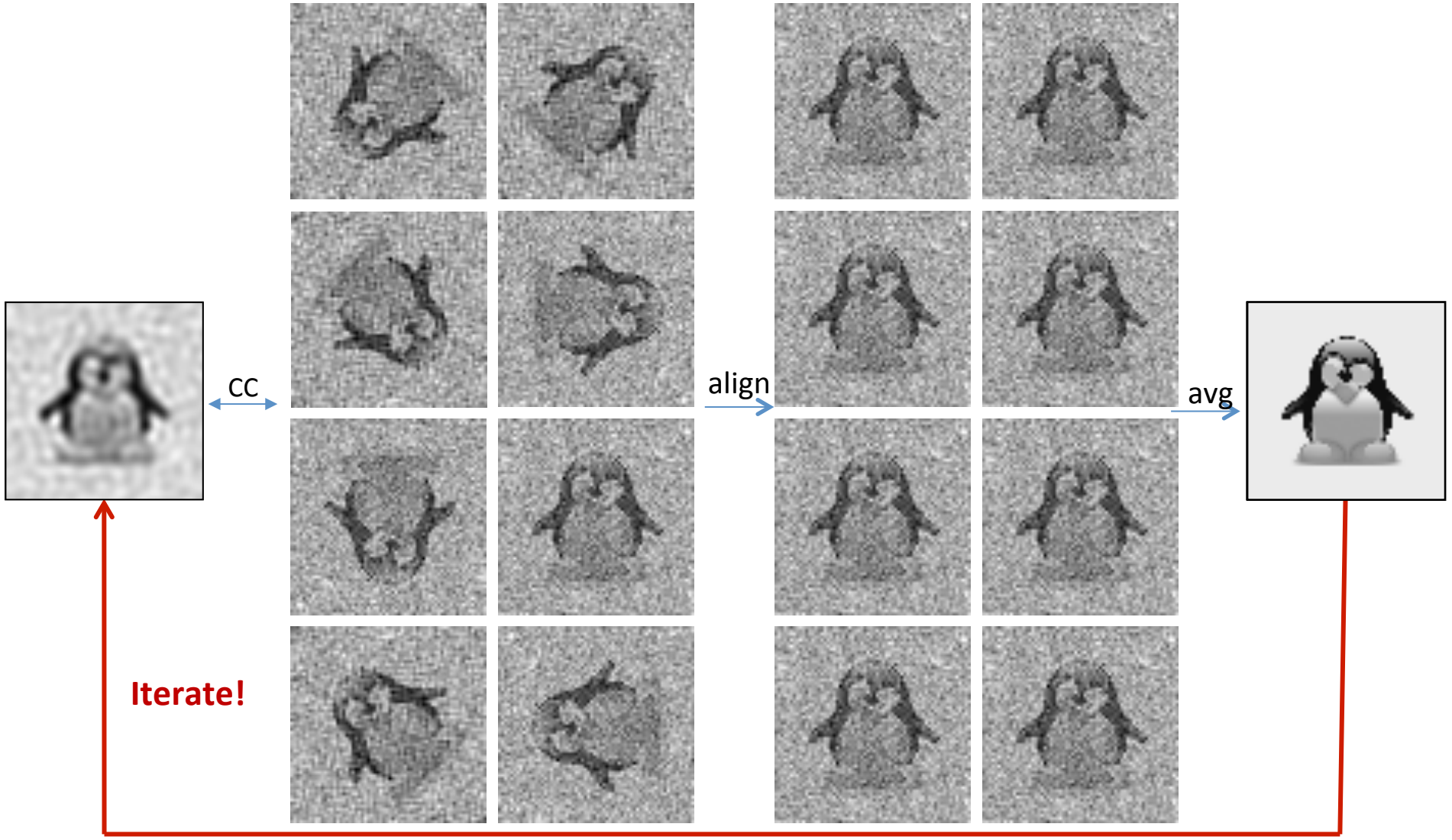
- Starts from some initial guess about the structure



Align and average



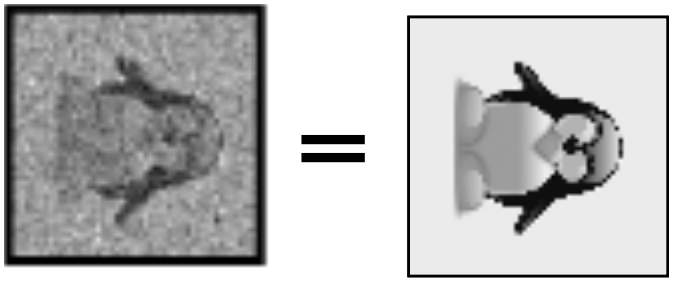
Align and average



The ML approach

Statistical data model

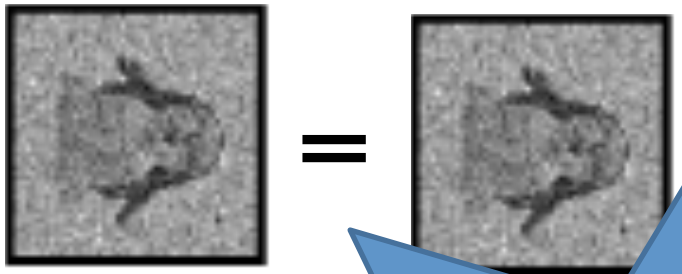
$$X_i = P_\varphi V_k$$



Statistical data model

$$X_i = P_\varphi V_k + N_i$$

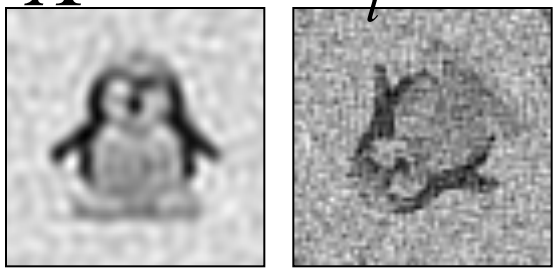
white / coloured Gaussian noise



Statistical description of the noise

Maximum likelihood

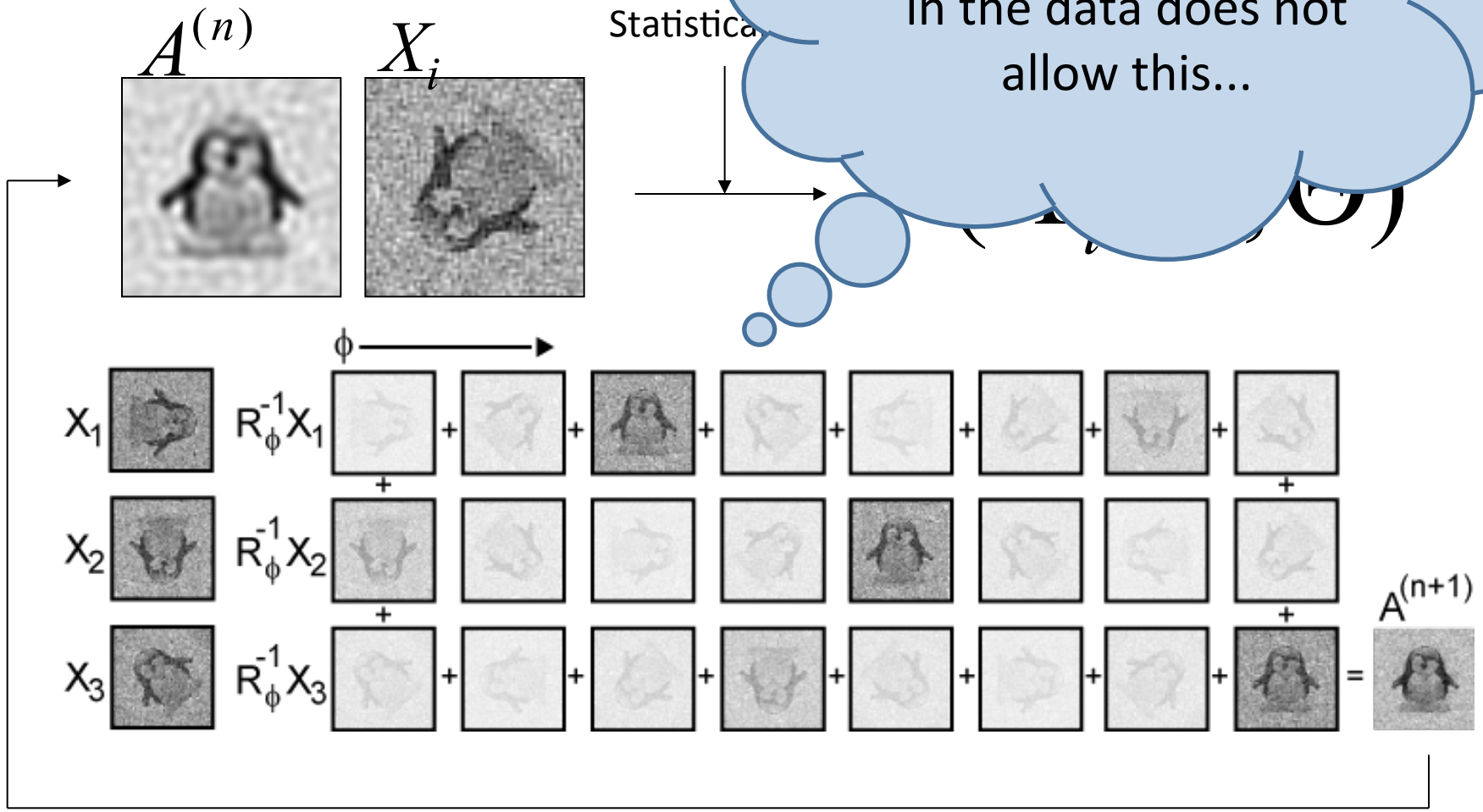
$A^{(n)}$ X_i



Statistical model

$$P(X_i \mid \phi, \Theta)$$

Maximum li


Do not assign discrete orientations if the noise in the data does not allow this...
$A^{(n)}$
X_i
Statistica
ϕ
X_1
$R_\phi^{-1}X_1$
X_2
$R_\phi^{-1}X_2$
X_3
$R_\phi^{-1}X_3$
$A^{(n+1)}$

Incomplete data problems

- Option 1: add *Y* to the model

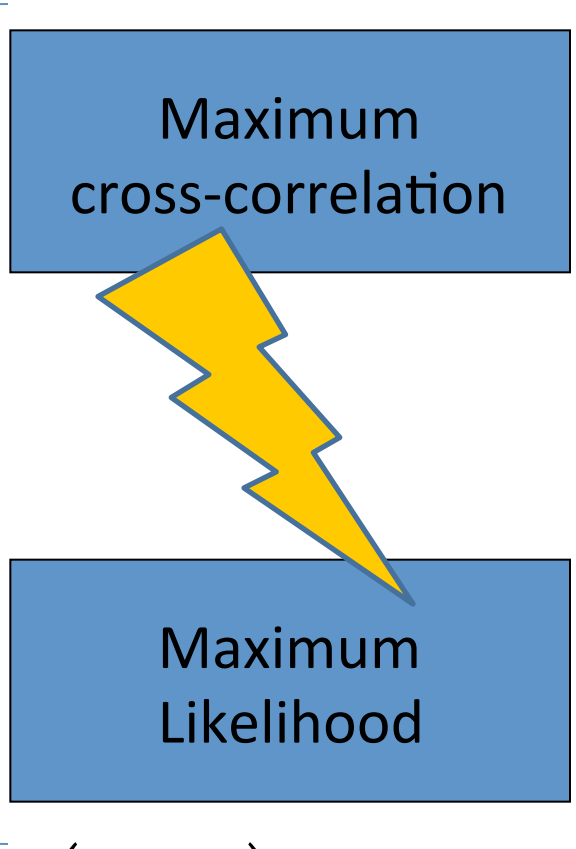
$$L(Y,\Theta) = P(X \mid Y,\Theta)$$

In the limit of **noiseless data** the Two techniques are equivalent!

- Option 2: marginalize over *Y*

$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y,\Theta)P(Y \mid \Theta)d\phi$$

Probability of X, regardless Y



Read more? See *Methods in Enzymology*, **482** (2010)

maxCC projection matching

- Compare X_i with $\mathrm{CTF}_i \mathbf{P}_\phi V$ for all ϕ, and select optimal ϕ^* based on some similarity measure (e.g. CC)

- Reconstruction:

iterate

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \mathbf{P}_{\phi^*}^{\mathrm{T}} \mathrm{CTF}_i X_i}{\sum_{i=1}^{N} \mathbf{P}_{\phi^*}^{\mathrm{T}} \mathrm{CTF}_i^2}$$

- Least-squares solution to *V* (?)

Maximum likelihood refinement

- Calculate a probability $P(X_i | \phi, \Theta)$ for all ϕ, based on an explicit noise model (e.g Gaussian)

- Probability-weighted angular assign[illegible]nt:

iterate

- Max[illegible] V

Theory says this is the best one can do *(in the limit of infinitely large data sets)*

Remaining issues

1. What to use as initial guess?
   - Local optimizer:
   - Wrong initial model -> wrong answer!
   - **Model bias!**

Model bias

- common-lines models are difficult
  - 2D projections are OK
  - Their combination in 3D is not
- Better (?)
  - RCT, sub-tomogram averaging, homologous structure
- EMAN(2) better than projection matching
  - But also not guaranteed…

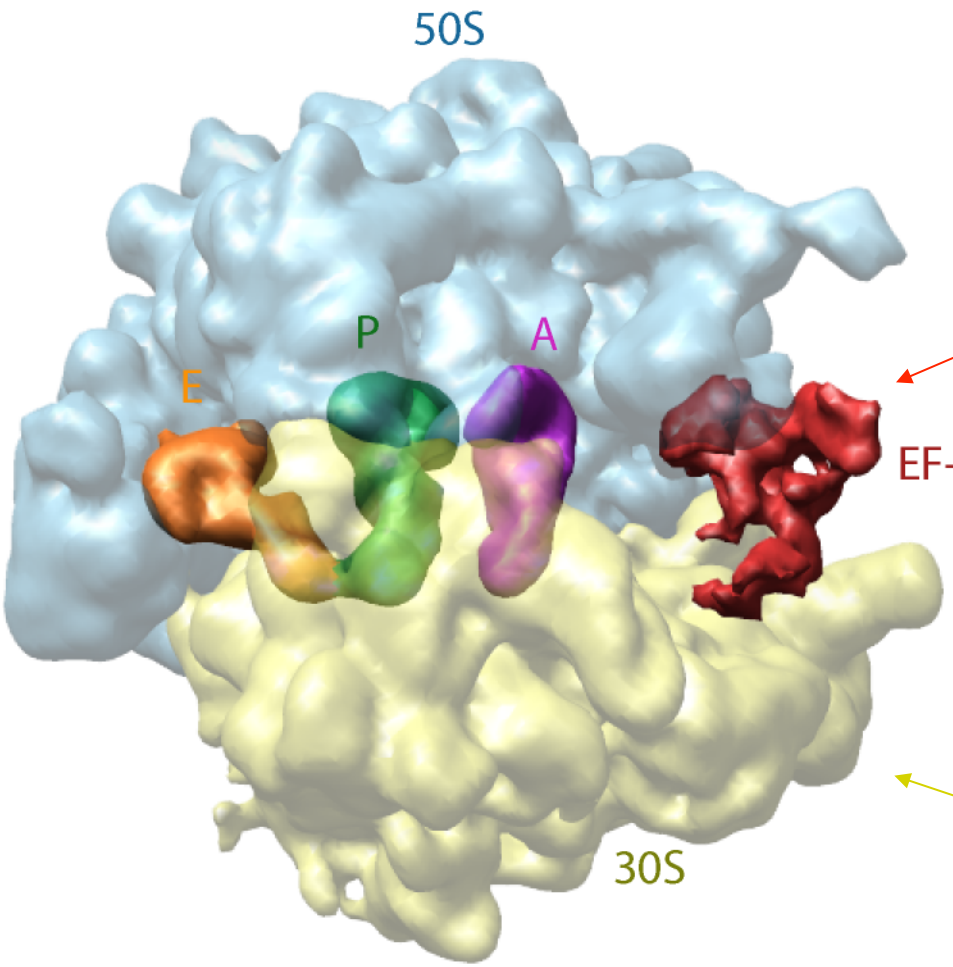
Remaining issues

1. What to use as initial guess?
   - Wrong initial guess may lead to wrong answers!
   - **Model bias!**

2. What if multiple structures are present?
   - Cannot align against 1 reference
   - Alignment + classification problem

Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution



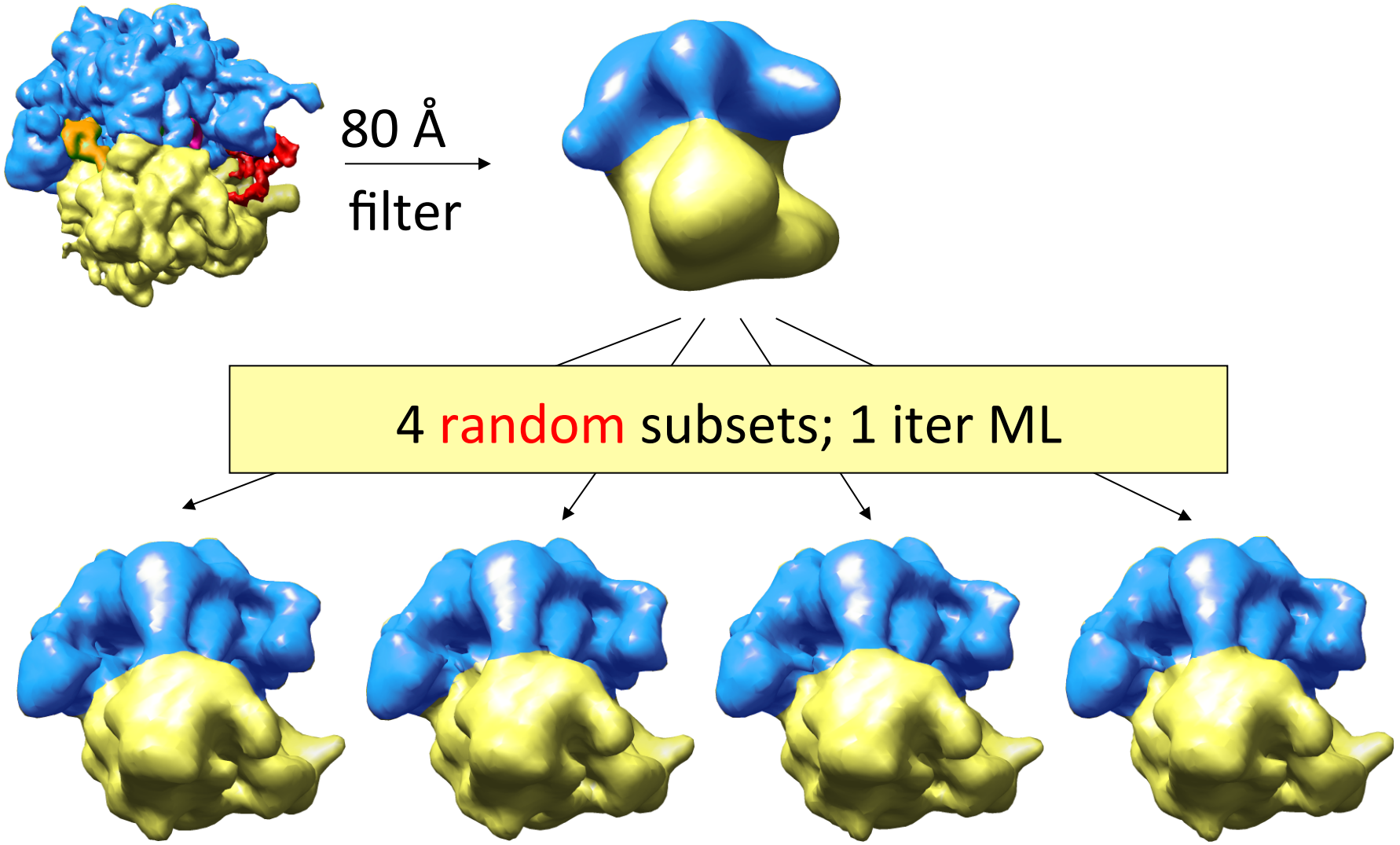
fragmented

(depicted at a lower threshold)

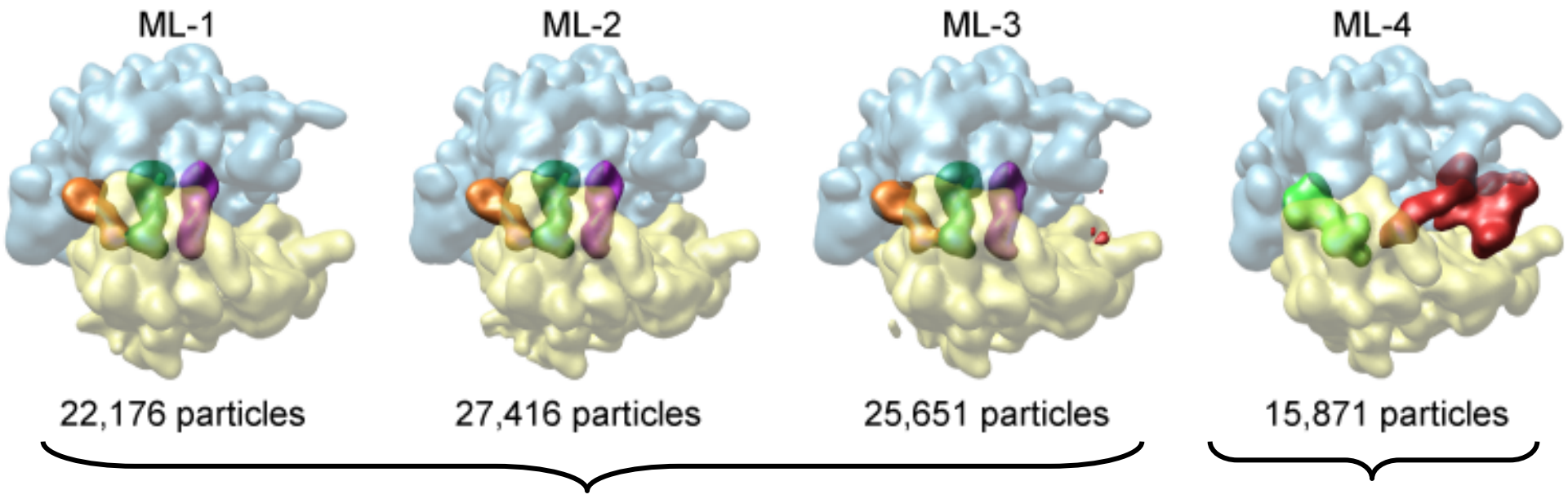
blurred

In collaboration with Haixiao Gao & Joachim Frank

Seed generation



ML-derived classes



no ratcheting; no EF-G; 3 tRNAs
differences: overall rotations

ratcheting,
EF-G, 1 tRNA

(Results coincided with a supervised classification)

BUT….

- 3D-classification is not a cure for bad data….
- Works best for few well-defined states
- Not all variability can be resolved
  - Continuous heterogeneity -> compromises
  - Many states may be tricky (expensive at least)
    - Supervised classification may be an alternative:
    - Fischer et al, Nature, 2010 (>20 states, 2M particles)
  - Ultimately a signal-to-noise ratio issue

Remaining issues

1. What to use as initial guess?
    1. Wrong initial guess may lead to wrong answers!
    2. **Model bias!**

2. What if multiple structures are present?
    1. Cannot align against 1 reference
    2. Alignment + classification problem

3. What if I do not infinite amounts of data...

Ill-posedness

The bad news

- The experimental data alone is not enough to determine a unique solution! (*ill-posed*)
  - Noise tends to accumulate in the reconstruction

The bad news

- The experimental data alone is not enough to determine a unique solution!
  - Noise tends to accumulate in the reconstruction
  - Overfitting
  - Over-estimation of resolution
  - Incorrect interpretations

The good news

- By incorporating external information, a different problem may be solved for which a unique solution does exist!

- Regularization

- Conventional approaches
  - Wiener filtering
  - Low-pass filtering

2D Wiener filter

- Assume noise is independent
  - with spectral power $\sigma^2(\upsilon)$
- Assume signal is independent
  - with spectral power $\tau^2(\upsilon)$
- **Minimise noise in 2D average**: (optimal filter)

$$A_{\mathrm{j}}^{(n+1)} = \frac{\sum_{i=1}^{N} \mathbf{R}_{\phi^*}^{\mathrm{T}} \frac{\tau^2(\upsilon)}{\sigma^2(\upsilon)} \mathrm{CTF}_{ij} X_{ij}}{\sum_{i=1}^{N} \mathbf{R}_{\phi^*}^{\mathrm{T}} \frac{\tau^2(\upsilon)}{\sigma^2(\upsilon)} \mathrm{CTF}_{\mathrm{ij}}^2 + 1}$$

Damp A for those Fourier components where all CTFs are zero or τ^2/σ^2 is small

Correct CTF AND low-pass filter!

3D Wiener filter

CHAPTER ONE

Fundamentals of Three-Dimensional Reconstruction from Projections

Pawel A. Penczek

Reconstruction methods based on backprojection and direct Fourier inversion methods require the implementation of a form of Wiener filter, which schematically is written as (see Chapter 2):

$$D = \frac{\sum_n \mathrm{CTF}_n \mathrm{SSNR}_n G_n}{\sum_n \mathrm{CTF}_n^2 \mathrm{SSNR}_n + 1}. \qquad (1.27)$$

The summation in the numerator can be realized as a backprojection of the Fourier transforms of $(n-1)$D projections multiplied by their respective CTFs and SSNRs, so the result is nD. However, it is far from obvious how the summation in the denominator can be realized such that the result would have the intended meaning after the division is performed.

Meth. Enzym. (2010)

3D Wiener filter

- Same assumptions
- Plus (often):

$$\frac{\tau^2(\upsilon)}{\sigma^2(\upsilon)} = \mathrm{SSNR}(\upsilon) = 1/C$$

BUT THIS IS NOT TRUE!!!!

Low-pass filtering effect is lost!

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \mathbf{P}_{\phi^*}^{\mathrm{T}} \mathrm{CTF}_i X_i}{\sum_{i=1}^{N} \mathbf{P}_{\phi^*}^{\mathrm{T}} \mathrm{CTF}_i^2 + C}$$

"Wiener constant"

“Arbitrary” low-pass filters

- Many different ones exist
  - choose shapes, effective resolution, width, etc.
- User expertise is required!

A Bayesian view on regularization

$$P(\Theta \mid X) = \frac{P(X \mid \Theta)P(\Theta)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

Maximum A Posteriori estimation

Likelihood

- Assume noise is Gaussian and independent
  - in Fourier space
  - with spectral power $\sigma^2(\upsilon)$: *coloured noise*

$$P(X_i \mid k, \phi, \Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}} \exp\left(\frac{\left\| X_{ij} - \mathrm{CTF}_{ij} (\mathbf{P}_\phi V_k)_j \right\|^2}{-2\sigma_{ij}^2} \right)$$

Prior

- Assume signal is Gaussian and independent
  - in Fourier space
  - Limit power $\tau^2(\upsilon)$: *smoothness in real space!*

$$P(\Theta) = \prod_l \frac{1}{2\pi\tau_{kl}} \exp\left\{ \frac{\|V_{kl}\|^2}{-2\tau_{kl}^2} \right\}$$

Expectation maximization

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^{2(n)}} X_i d\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^{2(n)}} d\phi + \frac{1}{\tau^{2(n)}}}$$

→ Wiener filter for 3D reconstruction

$$\sigma_i^{2(n+1)} = \frac{1}{2} \int_{\phi} \Gamma_{i\phi}^{(n)} \left\| X_i - \mathrm{CTF}_i \mathbf{P}_{\phi} V^{(n)} \right\|^2 d\phi$$

→ Estimate resolution-dependent power of noise from the data

$$\tau^{2(n+1)} = \frac{1}{2} \left\| V^{(n)} \right\|^2$$

→ Estimate resolution-dependent power of signal from the data

$$\Gamma_{i\phi}^{(n)} = \frac{P(X_i \mid \phi, \Theta^{(n)}) P(\phi \mid \Theta^{(n)})}{\int_{\phi'} P(X_i \mid \phi', \Theta^{(n)}) P(\phi' \mid \Theta^{(n)}) d\phi'}$$

3D Wiener filter

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^{2(n)}} X_i d\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^{2(n)}} d\phi + \frac{1}{\tau^{2(n)}}}$$

- Calculates SSNR(υ) (as a 3D function)
- Handles uneven orientational distribution
- Handles anisotropic CTFs & CTF e[illegible]s
- Corrects CTF & low-pass[illegible]
- *Optimal linear filter*

WITHOUT ARBITRARINESS!

Recapitulating...

- Inverse problem: needs iterating
- Incomplete problem: needs marginalizing
- Ill-posed problem: needs regularizing

- Bayesian approach:
  - Does all 3 things in optimizing a single function!
  - "Learns" optimal parameters from the data
  - No *ad-hoc* parameters to tune by the user

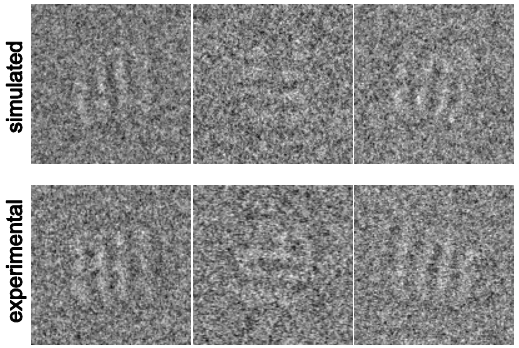
Preventing overfitting

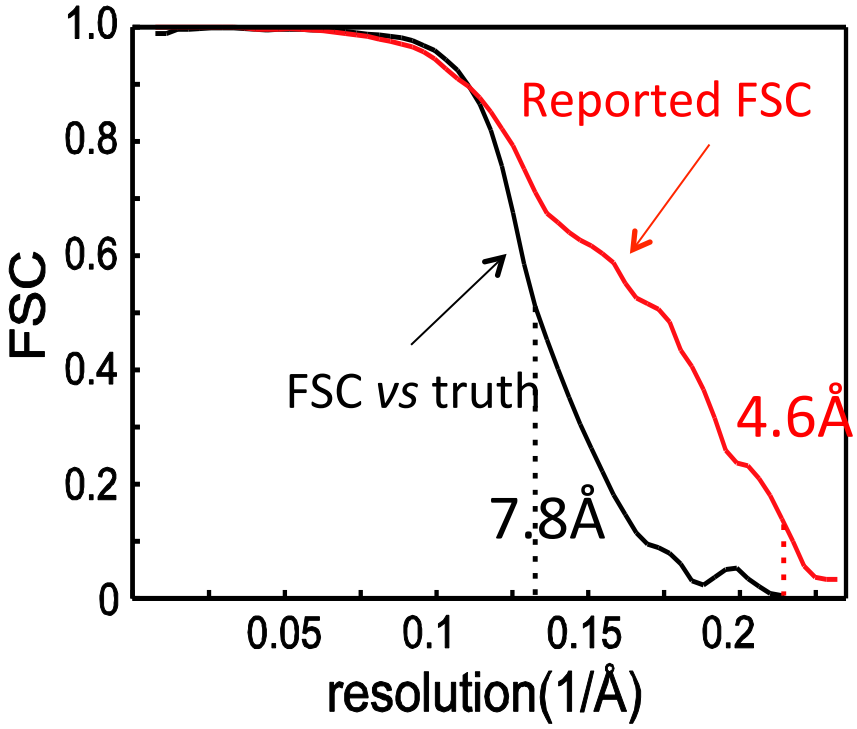
A little detour...

Scheres & Chen (2012) *Nature Methods*

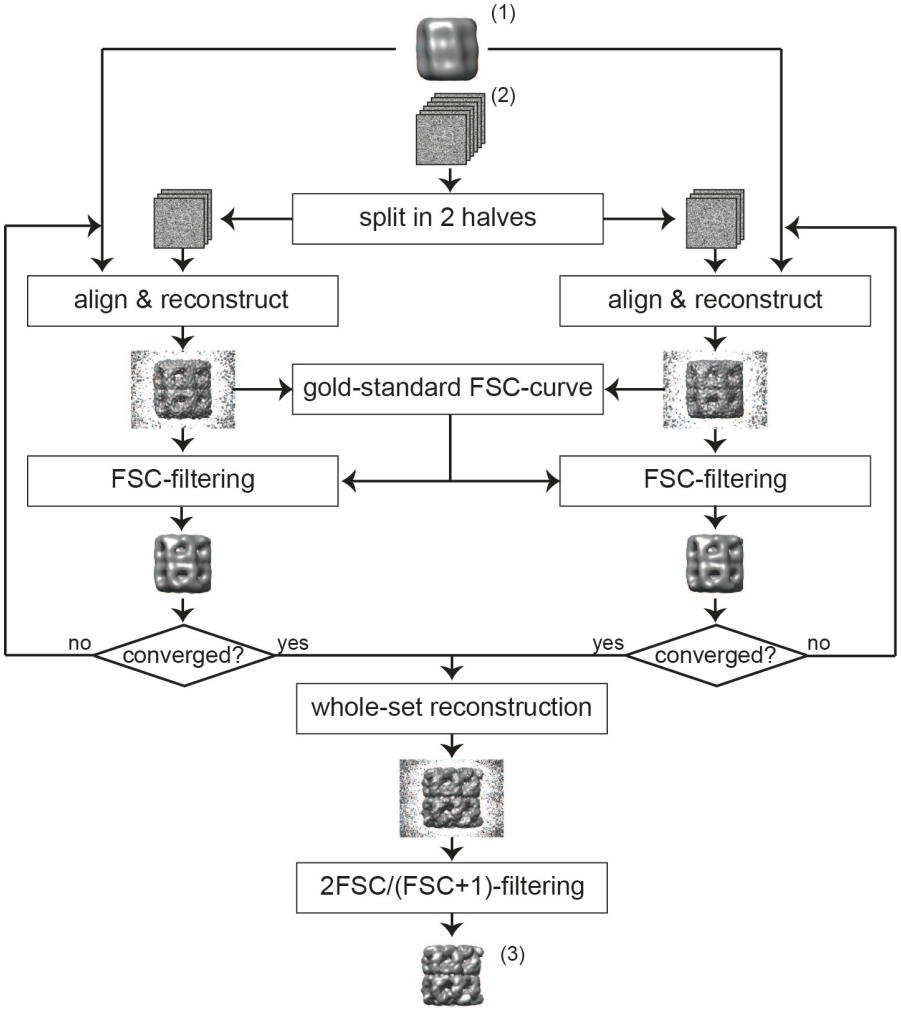
The pitfalls of undetected overfitting

- 20k simulated GroEL particles
- Conventional projection matching



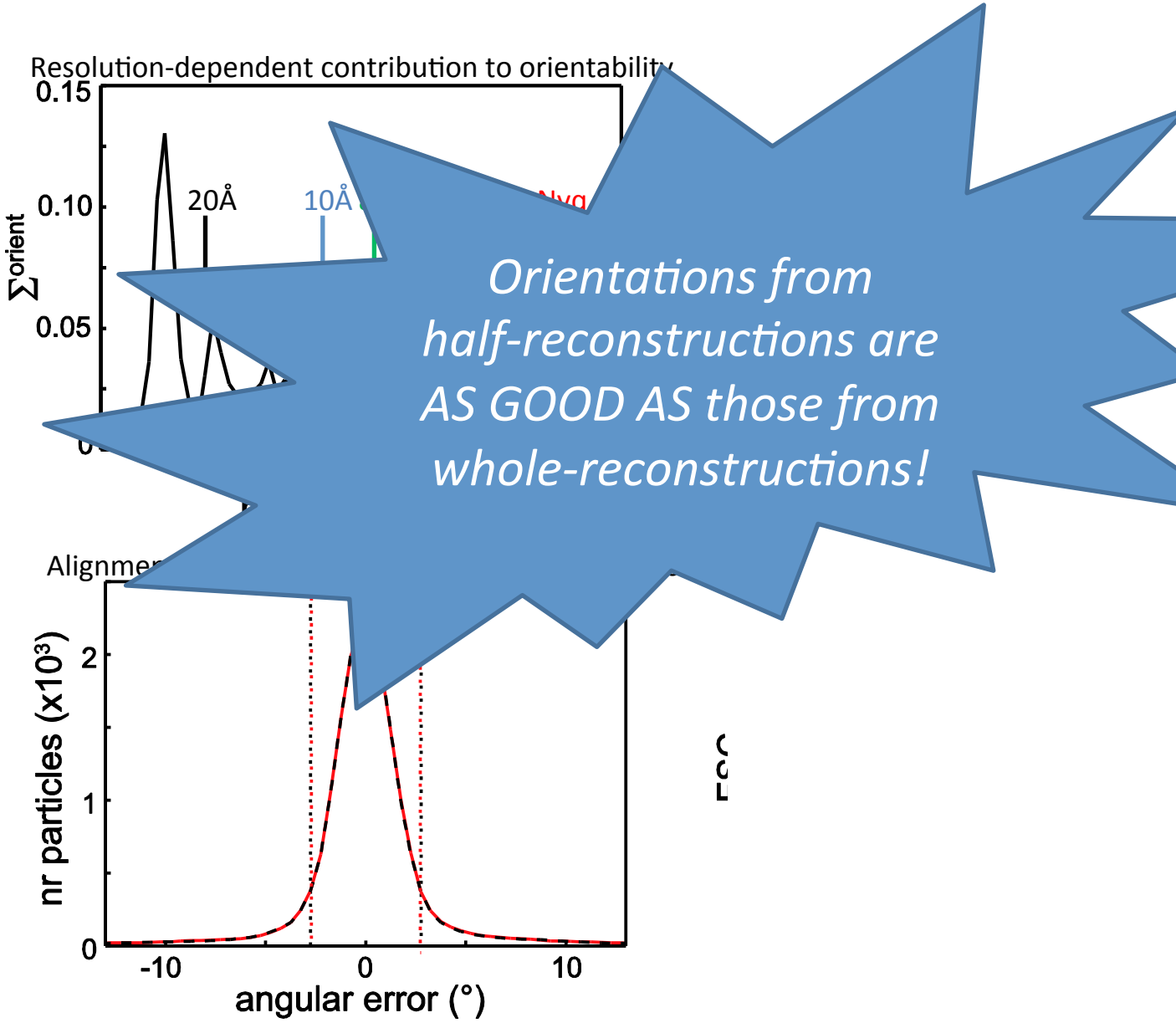


Overfitting-free refinement



easy to script into many packages...

Only lower resolution data drive alignment



Experimental data

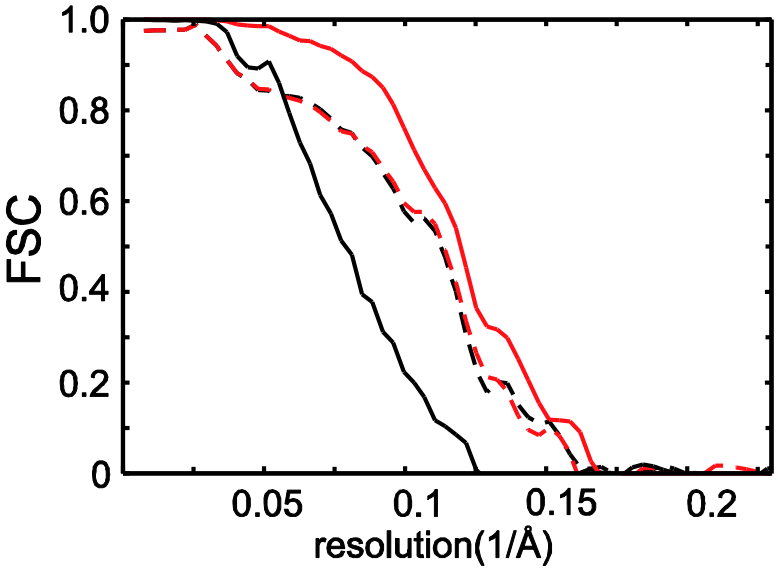
- 5,053 GroEL particles*
- 50,330 β-galactosidase particles
- 5,403 hepatitis B capsid particles**

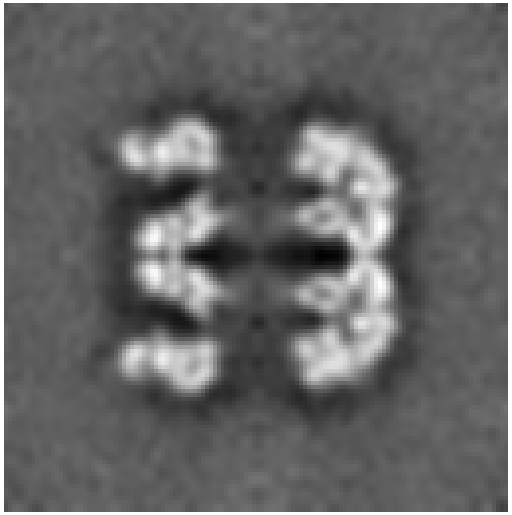
- **High-resolution crystal structures!**

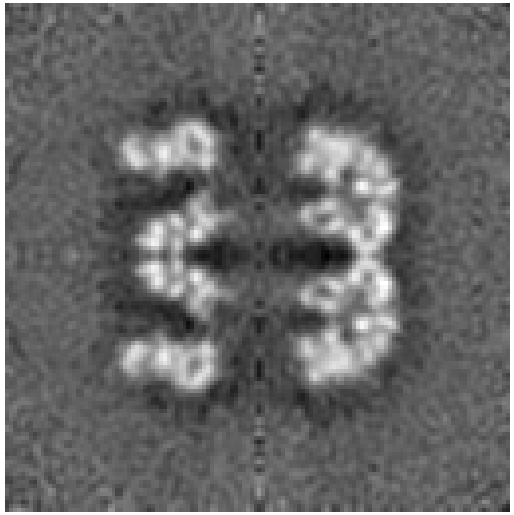
* *kindly provided by NCMI/Steven Ludtke*

** *kindly provided by Tony Crowther*

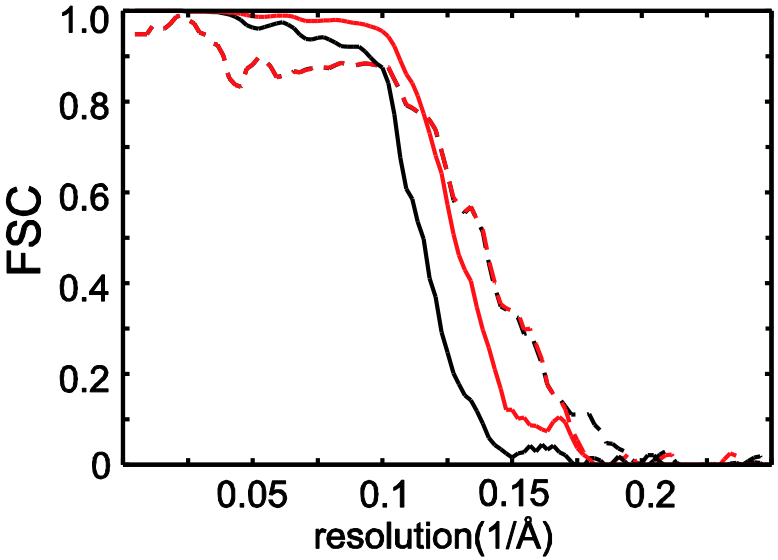
GroEL

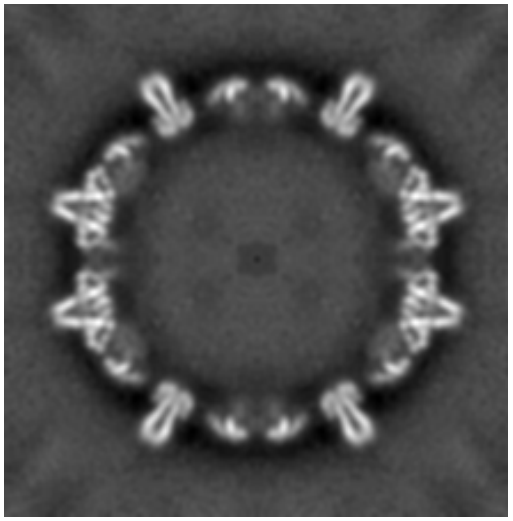


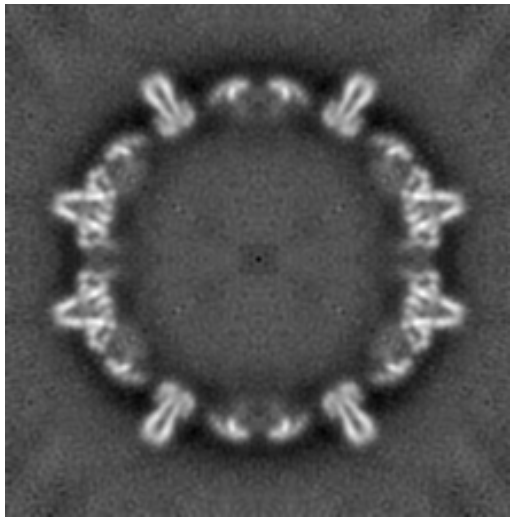




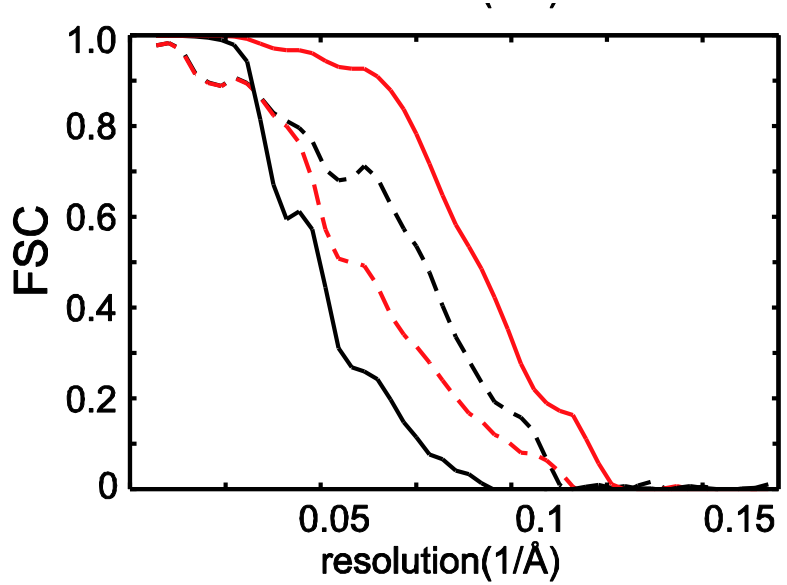
Hepatitis B capsid

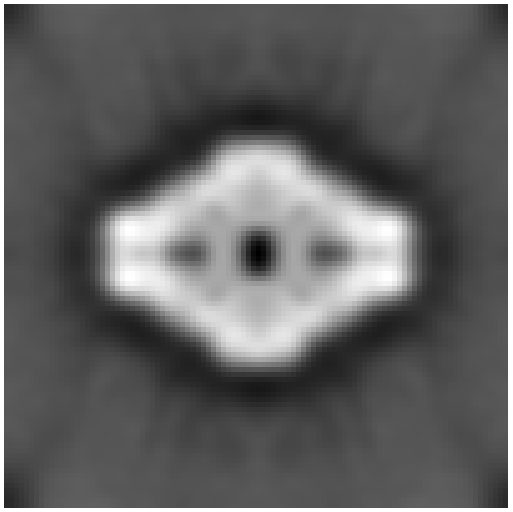


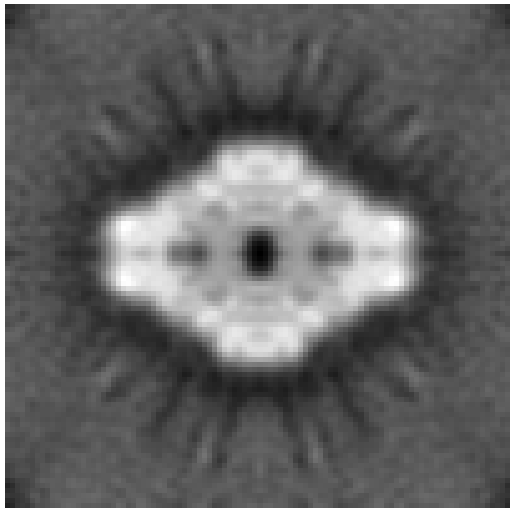




β-galactosidase







Conclusions

- Overfitting may be avoided without loss of resolution
  - Gold-standard FSCs between 2 independent models

- In the absence of overfitting
  - Higher-resolutions may be obtained
  - Maps are clean and easy to interpret, fit, etc.
  - FSC=0.143 is a reliable resolution estimate

Back to the statistical approach

Gold-standard FSC in the Bayesian approach

- Refine two models independently
- At each iteration: calculate $\tau^2(\upsilon)$ based on FSC_{gold}

REgularised **LI**kelihood **O**ptimisatio**N**

http://www2.mrc-lmb.cam.ac.uk/relion

RELION

Page | Discussion | Read | Edit | View history | Go | Search

Running RELION

Using the GUI [edit]

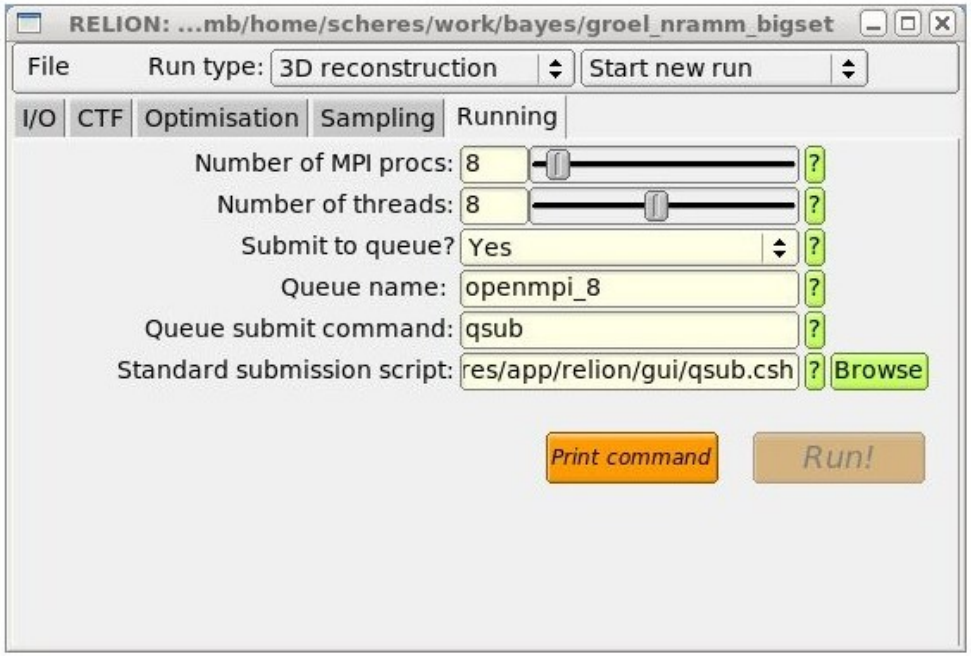
Navigation

Main page
Community portal

Toolbox

What links here
Related changes
Upload file
Special pages
Printable version
Permanent link

The graphical user-interface (GUI) has been designed to provide an intuitive interaction with the user. **It is strongly recommended to always run the GUI from the same directory for a given data set.** The GUI may be launched from the command line by typing the command `relion`. A screenshot is given below.



RELION may be used to perform different tasks (run types). The following run-types may be selected from the drop-down menu at the top of the GUI:

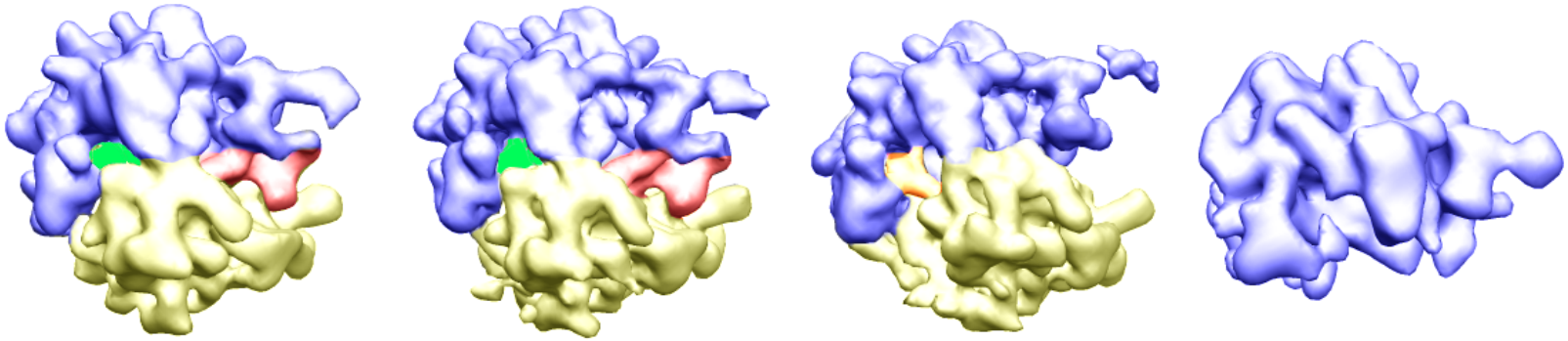
- 2D averaging: calculate reference-free 2D class averages
- 3D reconstruction: perform 3D (multi/single-reference) refinements

Some results

Tom Walz: test new programs on old data!

Classify structural variability

- Standard data set *(i.e. used by many groups...)*
  - 10,000 70S ribosomes (50% +EFG; 50% -EFG)
  - MAP-refinement K=4



24%	28%	42%	6%
26Å	19Å	19Å	30Å

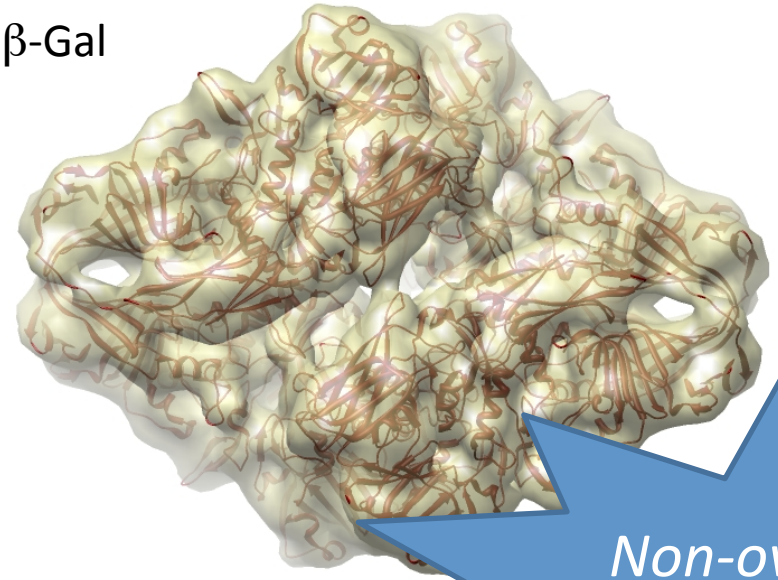
8 hrs on 64 CPUs

3D auto-refine results

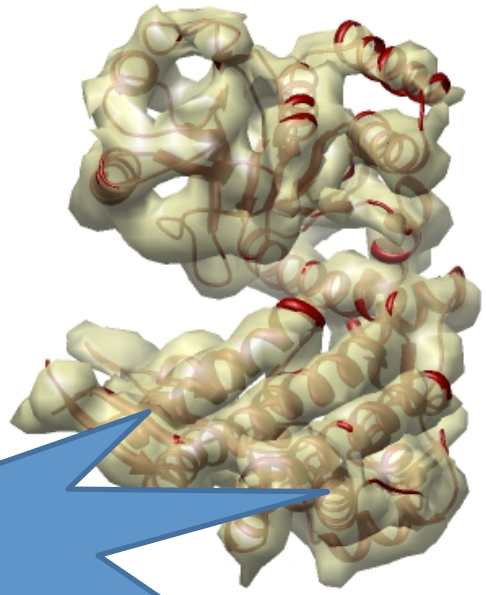
	β-galactosidase	groEL	hepatitis B	rotavirus
Sample characteristics				
Size (MDa)	0.45	0.8	4	60
Symmetry	D2	D7	I	I
Microscopy settings				
Microscope	FEI Polara G2	Jeol 3000SFF	Hitachi HF2000	FEI Tecnai F30
Voltage (kV)	80	300	200	300
Defocus range (μm)	1.2-2.7	1.9-3.2	1.0-2.0	1.2-2.9
Detector	Kodak SO163	Kodak SO163	Kodak SO163	Kodak SO163
Data characteristics				
Image size (pixel^2)	100 × 100	128 × 128	220 × 220	400 × 400[a]
Pixel size (Å)	2.93	2.12	2.00	2.40
Nr. particles	50,330	5,053	5,403	3,700
RELION parameters				
Particle mask diameter (Å)	200	205	400	785
Initial low-pass filter (Å)	60	60	50	40
Initial angular sampling (°)	7.5	7.5	3.7	3.7
Local searches from (°)	1.8	1.8	0.5	0.5
Initial offset range (pixel)	6	6	6	6
Initial offset step (pixel)	1	1	1	1
RELION results				
Wall-clock time (hr)	13.6	2.0	8.2	41.5
Reported resolution (Å)	9.8	8.2	7.3	5.6
Resolution vs X-ray (Å)	10.1	8.4	7.3	4.4[b]
Previous results				
Refinement program	XMIPP[c]	EMAN2[d]	MRC	FREALIGN[e]
Reported resolution (Å)	13.9	8.4	7.4	≈6
Resolution vs X-ray (Å)	12.7	8.7	7.5	4.4[b]

3D auto-refine results

β-Gal

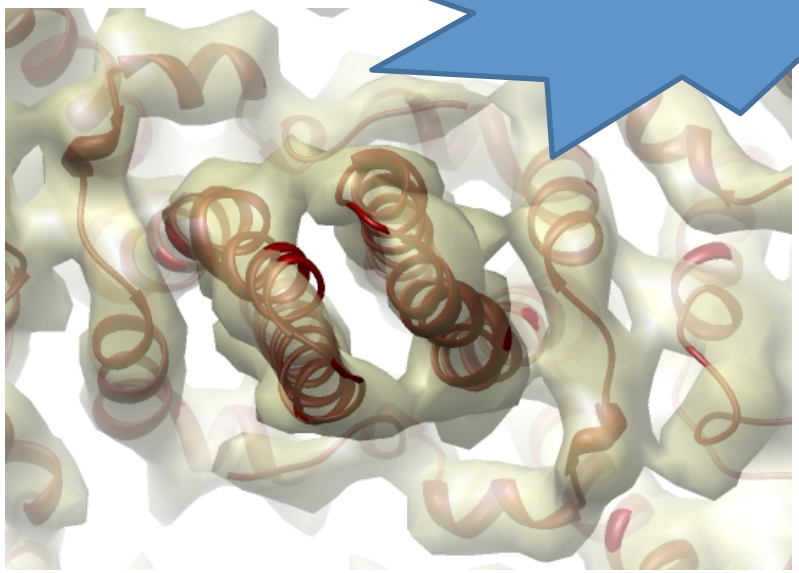


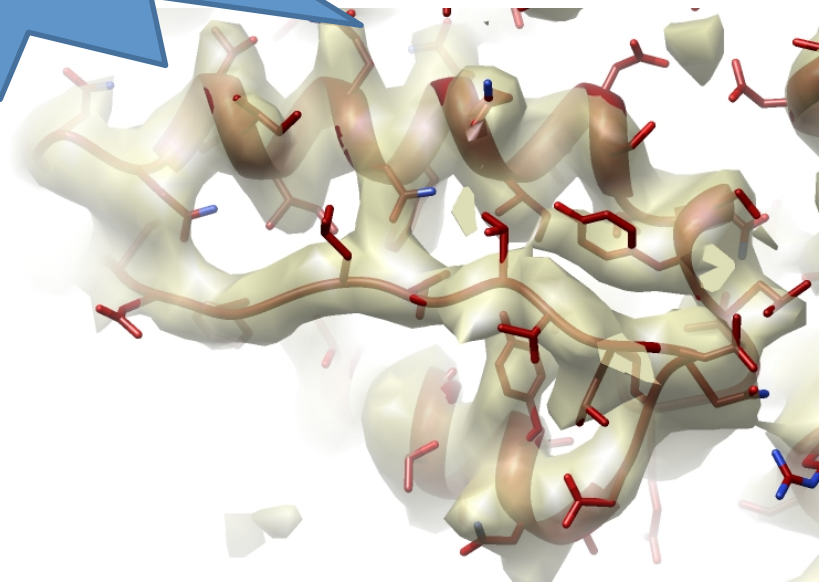
GroEL



Non-overfitted maps are clean!

hepB





More exciting RELION results

- DNA-origami object @ 11.5 Å resolution
  - See poster (Xiao-chen Bai)

Conclusions

- 3D-EM reconstruction is ill-posed, incomplete inverse problem
  - Needs: regularization, marginalization and iteration
- Initial model generation & classification remain problematic in some projects
- Overfittting may be avoided w/o loss of reconstruction quality
  - Use gold-standard FSCs, or high-res limited refinement!
- Bayesian framework provides a firm theoretical basis for 3D-EM
  - Learns optimal parameters from the data
  - Very little user input -> objective and easy-to-use
  - Excellent quality reconstructions

Acknowledgements

- HepB data
  - Tony Crowther
  - Greg McMullan
- GroEL data
  - Steven Ludtke
- 70S Ribosome data
  - Haixiao Gao
  - Joachim Frank
- β-galactosidase data
  - Shaoxia Chen
  - Richard Henderson
- Rotavirus data
  - James Chen
  - Niko Grigorieff
- **80S ribosome data**
  - **Xiaochen Bai**
  - Israel Sanchez
  - Venki Ramakrishnan
- Computing
  - Jake Grimmett
  - Toby Darling
- Some code in RELION
  - Xmipp (Carazo *et al.*)
  - Bsoft (Heymann *et al.*)
- Discussions
  - LMB colleagues
- Funding

MRC | Laboratory of Molecular Biology