Placing models with likelihood: molecular replacement and cryo-EM docking

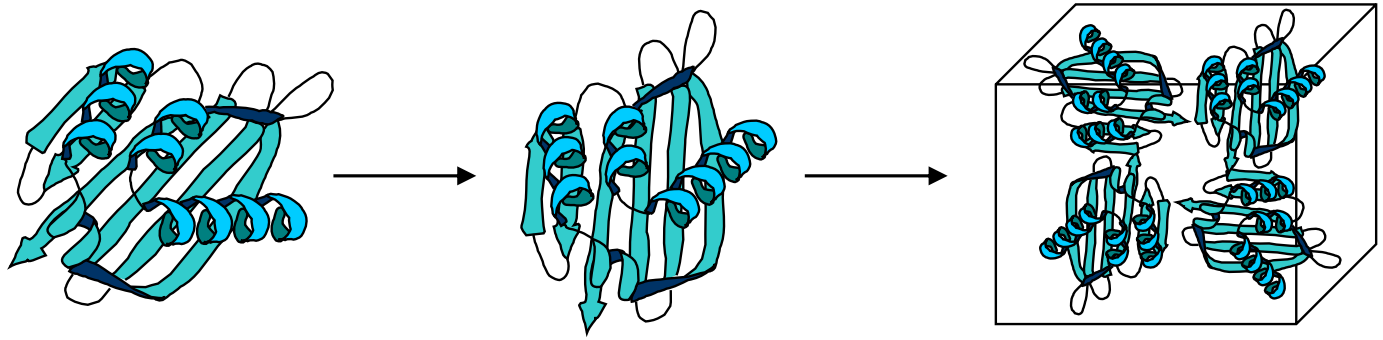
UNIVERSITY OF CAMBRIDGE

Randy J Read
Department of Haematology

CIMR
Molecules
Mechanisms
Medicine

Phasing by molecular replacement

- Phases can be calculated from atomic model
- Rotate and translate related structure
- Only one data set required!
- There is now almost always a good model!



What makes MR difficult?

- Incomplete model, or many copies
  - high non-crystallographic symmetry (NCS)
    - number of copies can be uncertain
  - part of complex
  - component(s) with no models, *e.g.* nucleic acid
- Poor data
  - low resolution
  - data pathologies (*e.g.* anisotropy, twinning, tNCS)
- Poor model
  - altered conformation
  - low-confidence AlphaFold model

Why likelihood?

- Accounts explicitly for effects of different sources of error
  - model error
  - measurement error
- More sensitive than other methods
  - especially for multiple copies or small fragments
- Exploits information from partial solutions
- Value of log-likelihood-gain (LLG) score gives good basis for automation: LLG > 60 usually means correct solution
  - expected value of LLG (eLLG) can be estimated in advance
  - choose among different possible solutions

How to attack a difficult MR problem

- Collect the best data possible
  - higher resolution helps
    - more signal with good models
    - more power for model completion algorithms
  - anomalous differences are very useful!
  - pathologies hinder progress
    - anisotropy reduces signal, makes maps harder to interpret
    - translational non-crystallographic symmetry (tNCS) must be accounted for
- Use eLLG to optimize strategy
- Prepare the best possible model

Models with estimated errors are far more useful!

- AlphaFold has been trained to predict the LDDT score used in CASP to assess the quality of each residue in a model
  - 100 = perfect
  - < 60-70 = poor
  - < 50 = possibly (probably?) intrinsically disordered
  - strong correlation with actual errors

trim from model

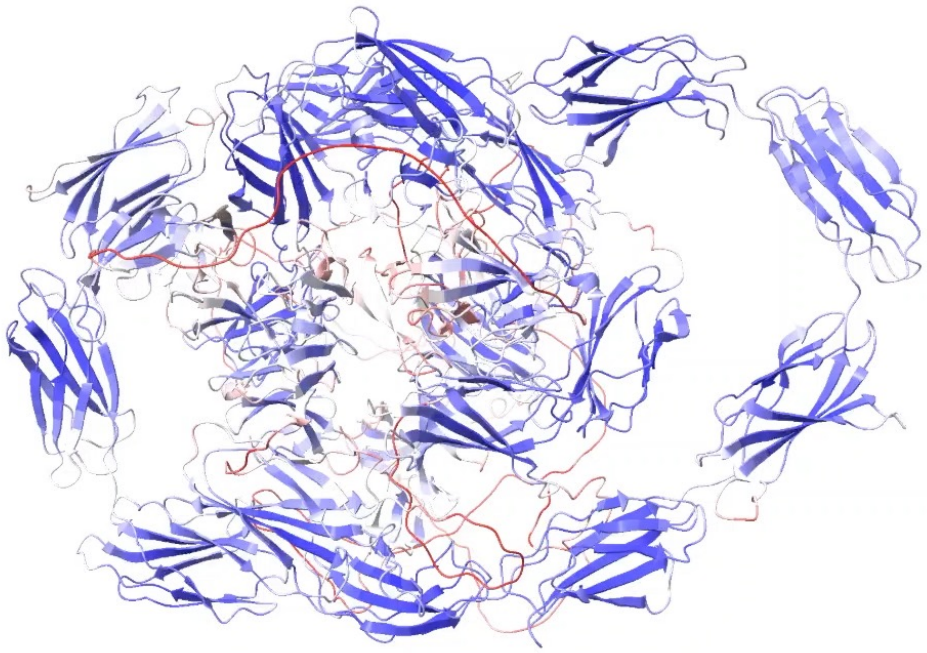
- AlphaFold computes a PAE (predicted aligned error) matrix
  - how certain are relative positions of residues in the structure
  - extremely useful for assessing confidence in domain orientations

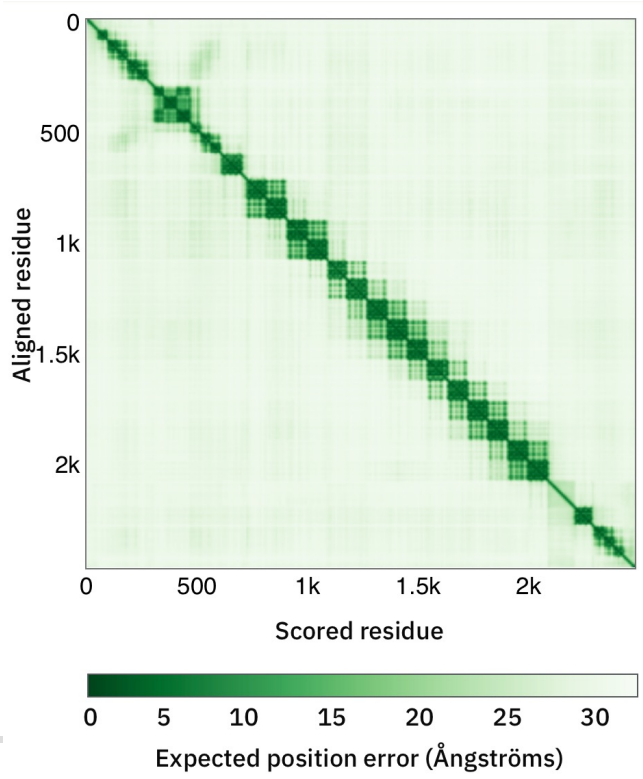
Using accuracy estimates

- Assign an overall estimated RMSD to each model
  - relative size and error taken into account in deciding search order
    - for AlphaFold models, size will be the major factor
- Change the relative weight of different parts of model
  - think of smearing out each atom over its possible positions
    - this is equivalent to adding a B-factor (Fourier transform of a Gaussian)
  - this is estimated from the pLDDT:
    - translate pLDDT into equivalent approximate RMSD, then to B-factor
- Use PAE (predicted aligned error) matrix to divide model into domains with uncertain relative orientation and position

Human fibronectin model

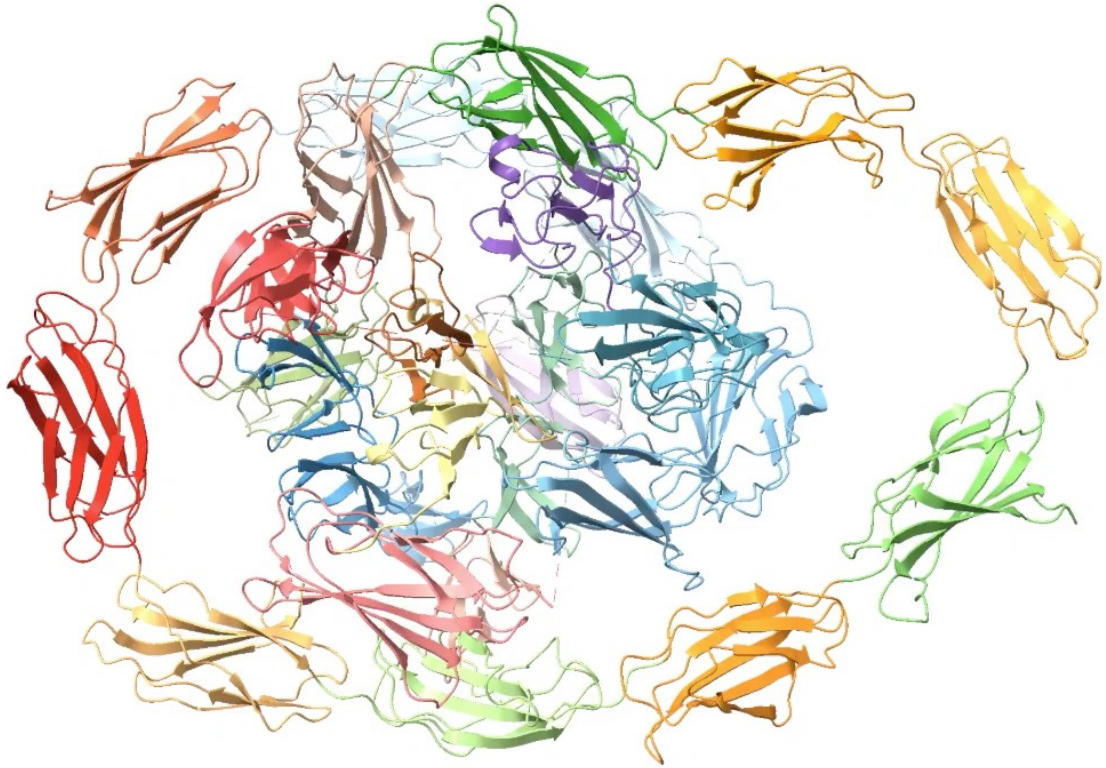
- Fibronectin repeats often have different relative orientations
- Large segments (in red) poorly predicted (or possibly disordered)





Fibronectin parsed into domains

- Community clustering of PAE matrix (Tristan Croll)



phenix.process_predicted_model

- Trim off low-confidence residues (pLDDT < 70 by default)
- Weight remaining structure by translating pLDDT to B-factor
- Divide into rigid domains
  - low-resolution “blob” analysis: Tom Terwilliger
  - PAE matrix parsing: Tristan Croll

Likelihood is sensitive…

- …to correct orientation and position of molecular replacement model
  - successful in solving structures with distant relatives, small fragments, or many copies in asymmetric unit
- …to violations of assumptions
  - data implicitly assumed to be isotropic
    - important to account for anisotropy
  - components may not be equally well-ordered
    - important to correct for differences in overall B-factors

Pathologies violating assumptions: translational NCS (tNCS)

- Found in about 8% of PDB entries

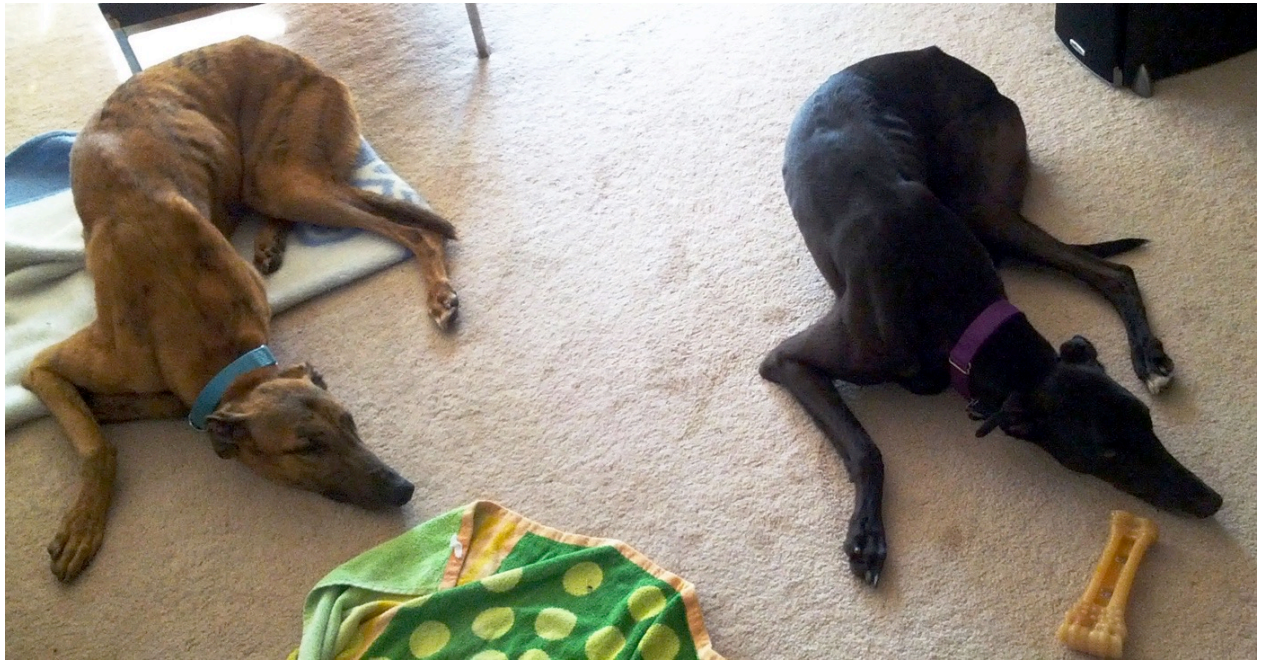
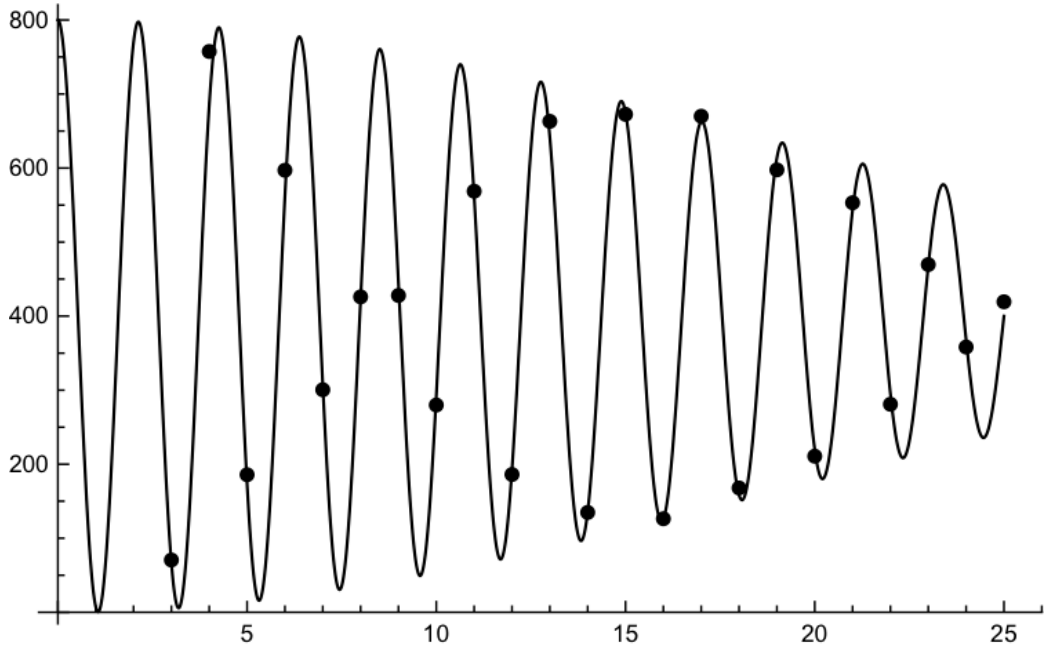


Photo courtesy of Laurie Betts

Accounting for translational NCS

- Model effect of translation combined with small rotation and random differences between copies



Hyp-1:
Sliwiak, Jaskolski, Dauter, McCoy, Read (2014)

Twinning

- Rotated diffraction pattern superimposed on itself
  - may mislead space group identification
    - consider subgroups of space group

SAD phasing in Phaser

- Likelihood for molecular replacement: probability of single structure factor measurement, given a model of the structure
- Likelihood for SAD: probability of Bijvoet pair of structure factor measurements, given a model of the anomalous substructure
  - generalisation of MR target

SAD log-likelihood gradient (LLG) map

- Compute derivative of log-likelihood with respect to heavy atom structure factor
- Fourier transform gives map of where likelihood target would like to see changes in anomalous scatterer model
- Very sensitive to minor sites
  - picks up sites identified as water molecules in refined structures determined by halide soaks

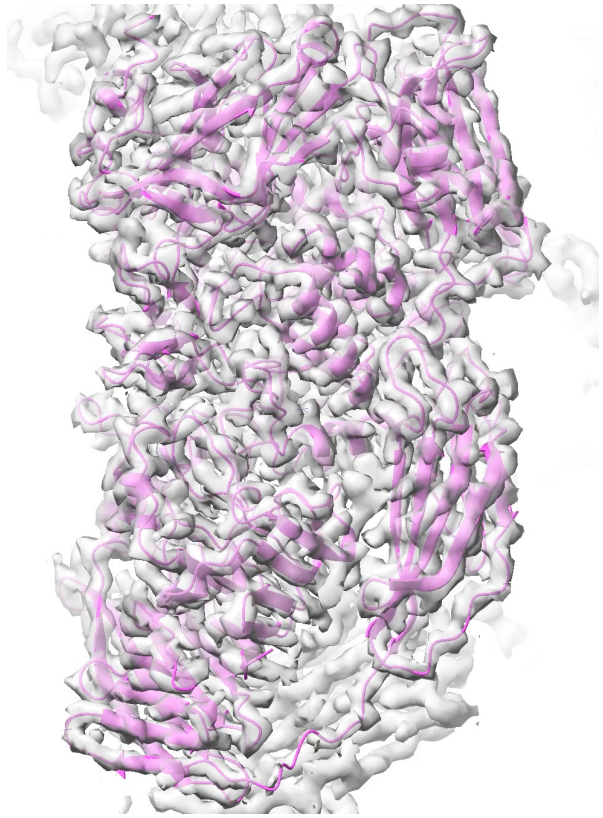
MR-SAD

- Use molecular replacement model as “substructure” with no anomalous scattering
- Find anomalous scatterer sites using SAD log-likelihood-gradient maps
  - in principle, different atom types give different scores in the log-likelihood-gradient maps
    - differ in relative contribution of real and imaginary scattering
- Used to improve phases and to help identify ambiguous atoms

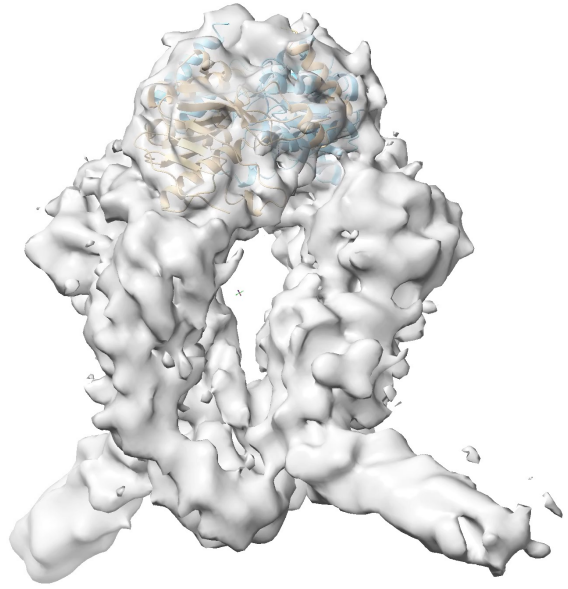
The docking problem in cryo-EM

- We have a map: how can we place an atomic model of a component in that map?
  - scoring problem
    - map correlations?
    - likelihood?
  - search problem: exploring rotations and translations
    - brute-force 6D search?
    - separate rotation and translation search?
  - decision problem
    - how confident can we be in the solution?

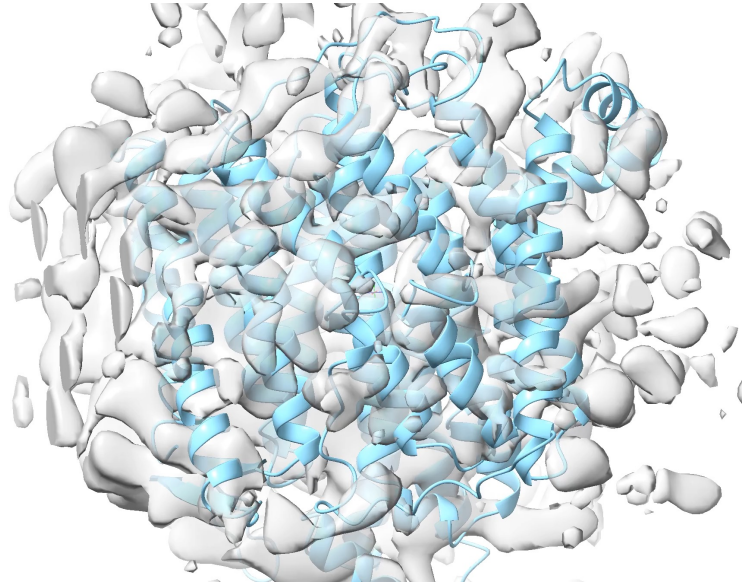
Which docking cases are important?



β-galactosidase
2.2 Å



C-terminal domain of MutS
6.9 Å



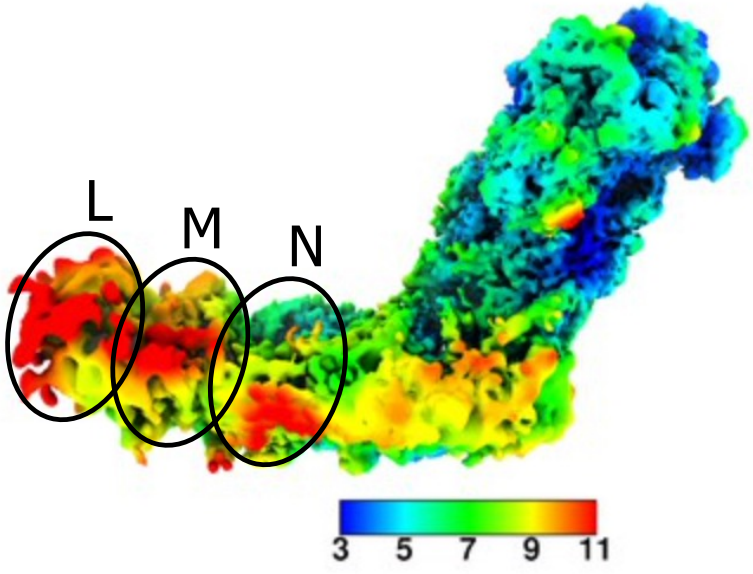
Chain L of *E. coli* complex I
3.8 - 11 Å

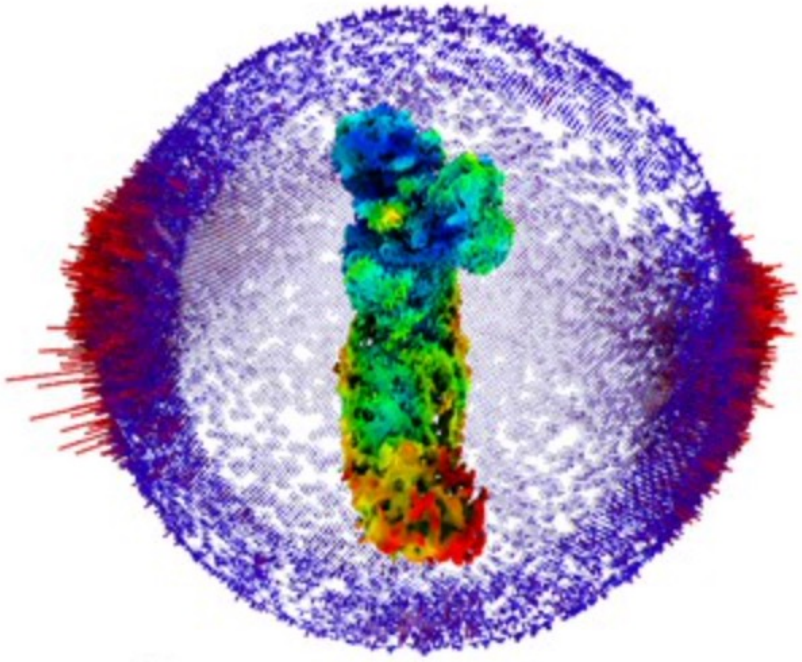
Likelihood: signal and noise in cryo-EM data

- Individual particle images are very noisy
  - average data from many particles
- Signal reduced by lack of reproducibility of the sample
  - different conformations, radiation damage
- Signal and noise strength are analysed by comparing half-maps
  - described in Read, Millán, McCoy & Terwilliger *Structural Biology (Acta Cryst D)*, 2023

Example: EMDB 12654: PDB 7nyu

- *E. coli* respiratory complex 1 in lipid nanodisc
  - Kolata & Efremov, eLife, 2021
  - resolution ranges from 3.8 to 11 Å





orientation preference

Docking a model to a cryo-EM map

- Break 6D search into two 3D searches for efficiency, as in MR
  - rotation search: equivalent to the crystallographic rotation function
  - translation search: the phased cryo-EM likelihood function can be evaluated exactly with a single FFT
- Details of strategy adapt to the quality of the data and the model, through the expected log-likelihood-gain (eLLG)

Overall docking strategy in *EM_placement*

- Evaluate signal and noise in entire reconstruction
  - will the rotation search probably succeed?
    - YES: run rotation search followed by translation search
    - NO: will rotation search for minimal sub-volume succeed?
      - YES: divide map into sub-volumes, carry on as before
      - NO: do brute-force rotation and translation search
- Implementation and test cases (1.7-8.5Å resolution, 5-50% complete model) described in Millán, McCoy, Terwilliger & Read *Structural Biology (Acta Cryst D)*, 2023

Searching in a defined sphere: *emplace_local*

- More sensitive (and much faster) if you know approximately where a molecule should go
- Easiest to run from new ChimeraX plugin
  - see YouTube tutorials by Dorothee Liebschner
    - https://www.youtube.com/c/phenixtutorials
    - Phenix/ChimeraX playlist

Acknowledgements

- Claudia Millán
- Airlie McCoy
- Tristan Croll

- Tom Terwilliger
- Dorothee Liebschner
- Billy Poon

Tom Burnley

- Eric Pettersen
- Tom Goddard

Cathy Lawson

Phenix

An NIH/NIGMS funded Program Project