Robust and efficient likelihood-based docking of models into cryo-EM reconstructions

UNIVERSITY OF CAMBRIDGE

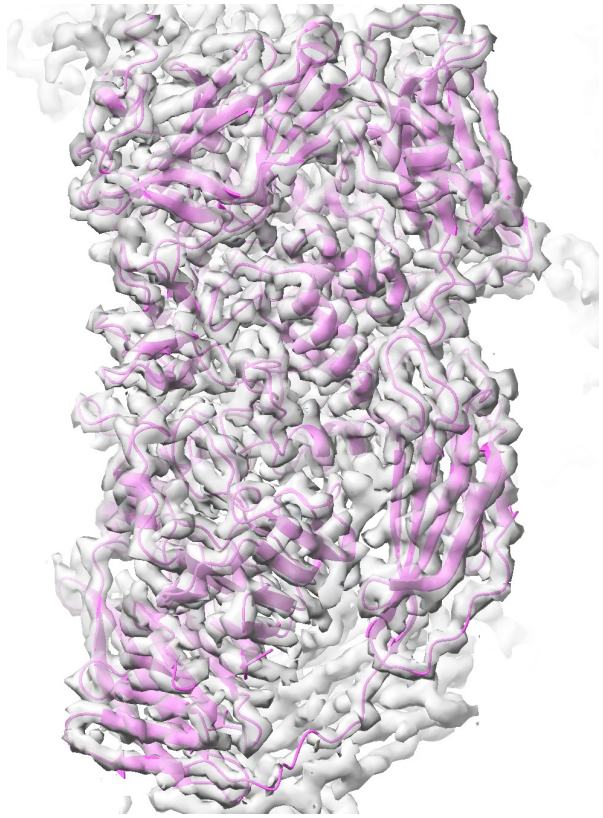
Randy J Read
Department of Haematology

CIMR
Molecules
Mechanisms
Medicine

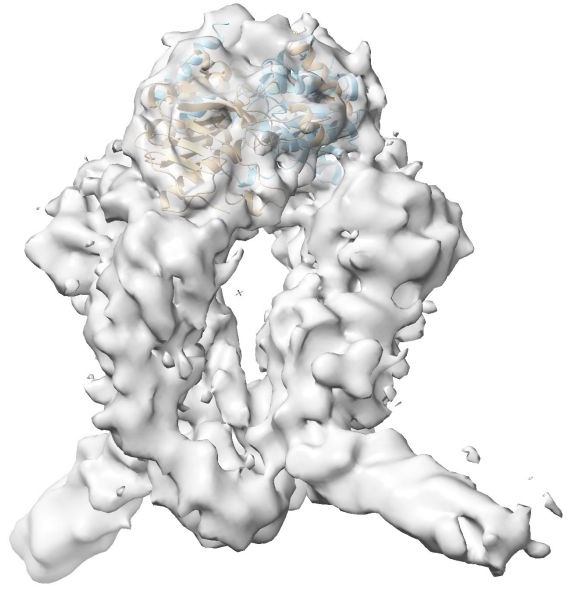
The docking problem in cryo-EM

- 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?

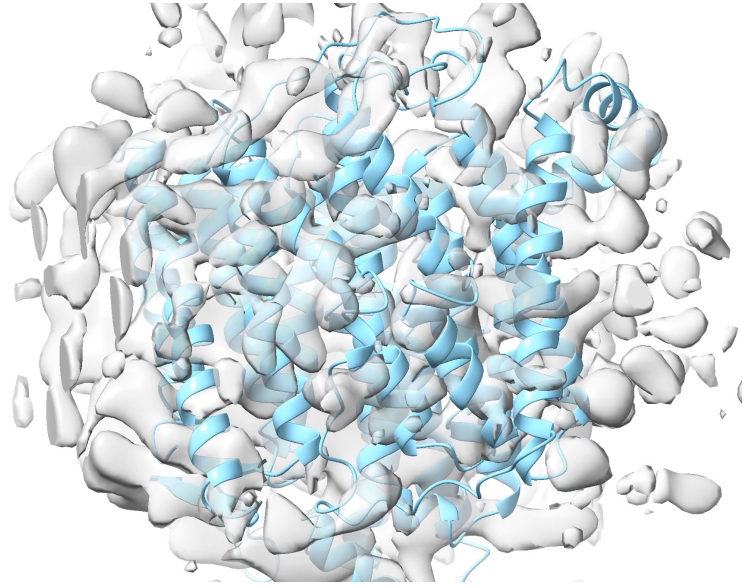
Which docking cases are important?



β-galactosidase
2.2 Å



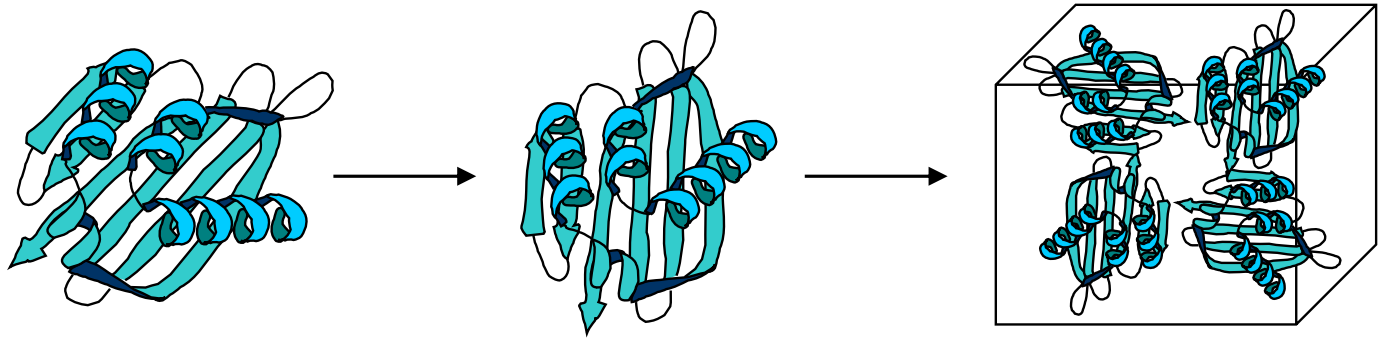
C-terminal domain of MutS
6.9 Å



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

Solving crystal structures by molecular replacement

- Rotate and translate atomic model
- Score the rotations and translations using likelihood in *Phaser*
  - accounts for errors in the data and in the model



- Some lessons can be applied to docking in cryo-EM
  - reconstructions are carried out in Fourier space *but they include phase information*

Advantages of likelihood for MR and docking

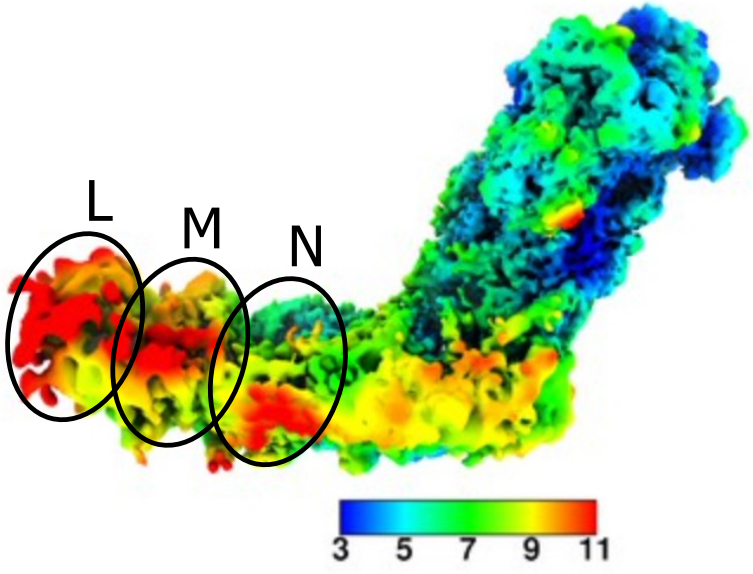
- Measures the consistency of the model (atomic model plus rigid-body rotation/translation parameters) with the data
  - probabilistic: accounts for errors in both model and data
- Likelihood is an absolute score
  - compare alternative hypotheses
  - judge confidence in solution
- Achievable score can be predicted from quality of model and data
  - optimise choices of strategy

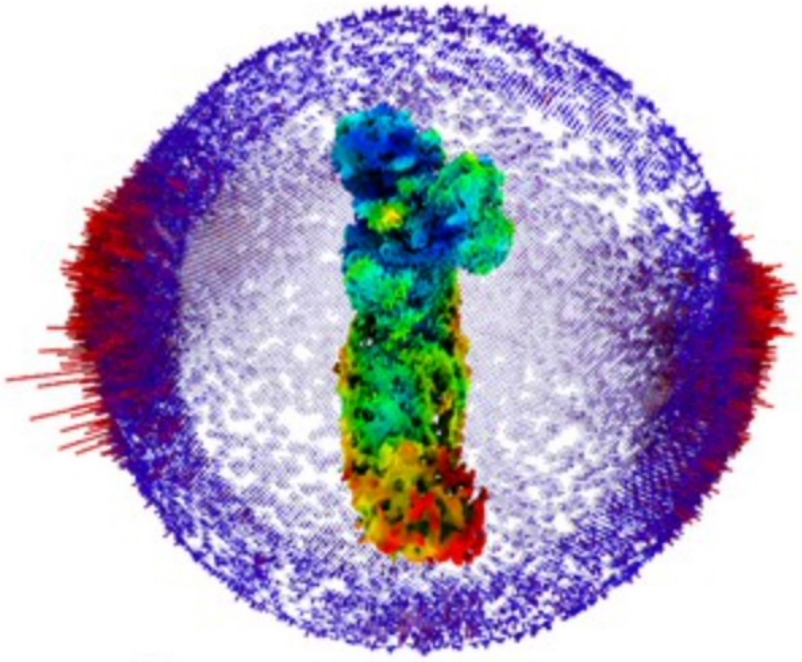
Likelihood: signal and noise in cryo-EM data

- Individual particle images are very noisy
  - average data from many particles to reduce noise
- Signal reduced by lack of reproducibility of the sample
  - different conformations, radiation damage
- Signal and noise strength are analysed by comparing half-maps
  - differs from one part of the map to another
  - 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)

The log-likelihood gain (LLG)

- Likelihood is probability of data set given model
- Log-likelihood gain: difference between logarithm of likelihood for tested model and an uninformative model
  - score of 60 or more: usually correct

- Related to how much information in the data is explained by the model

The expected log-likelihood-gain (eLLG)

- Inspired by MR: eLLG is used to devise optimal strategies
  - predict LLG that will be obtained given the quality and resolution of the data
- Rotation eLLG: much lower than for translation search
  - rotation is the hard step!
  - rotation LLG and eLLG can be increased by putting the relevant density in a smaller box: inversely proportional to box volume
    - this does require phase information!

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 ← rotation eLLG
    - 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 ← translation eLLG
- 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

Practical aspects of running em_placement

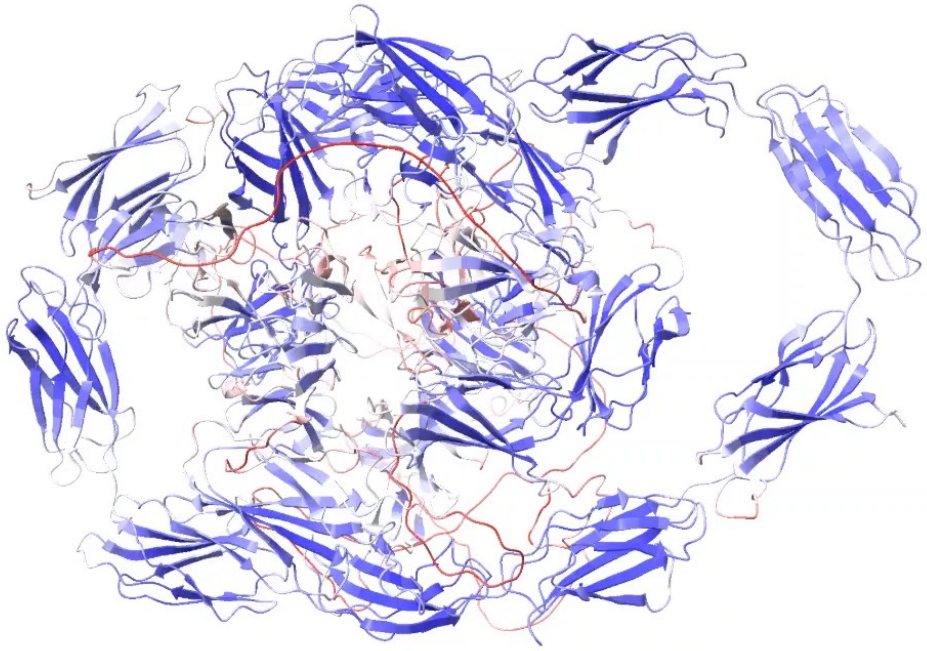
- Requires two half-maps
  - assess signal (correlation) and noise (difference)
  - also used to define ordered volume in the reconstruction
- Requires sequence information to define the total content of the reconstruction
  - set threshold for ordered volume determination
  - how much of the ordered volume is in a spherical subvolume?
    - is there enough to contain my search model?
- Resolution limit can be estimated, but useful to provide it
- Models should be edited appropriately

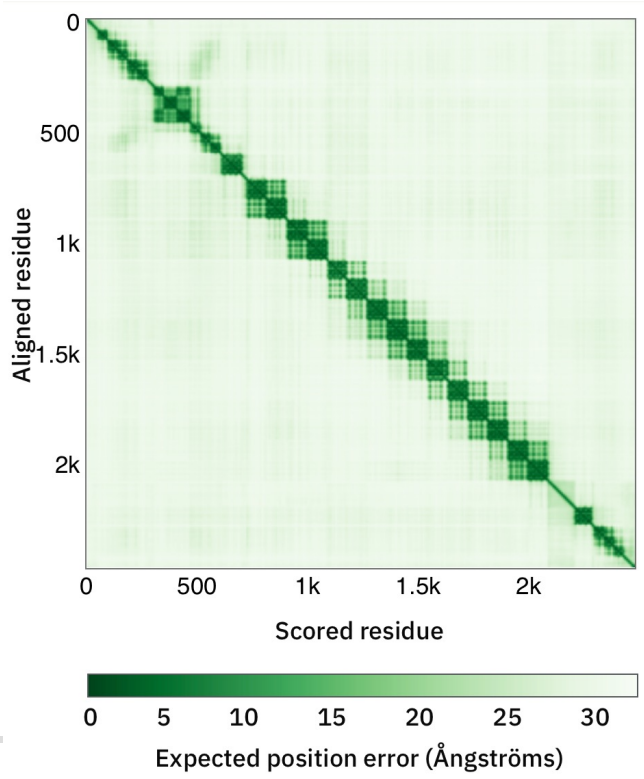
Process_predicted_model

- Docked models from AlphaFold (or other machine-learning methods) are an excellent starting point for model-building
- AlphaFold models have regions of high and low confidence
  - trim off low-confidence regions (pLDDT < 70)
  - turn pLDDT into sensible corresponding B-factors!
- Relative orientations of domains may be poorly predicted
  - PAE matrix is very useful

AlphaFold model of human fibronectin

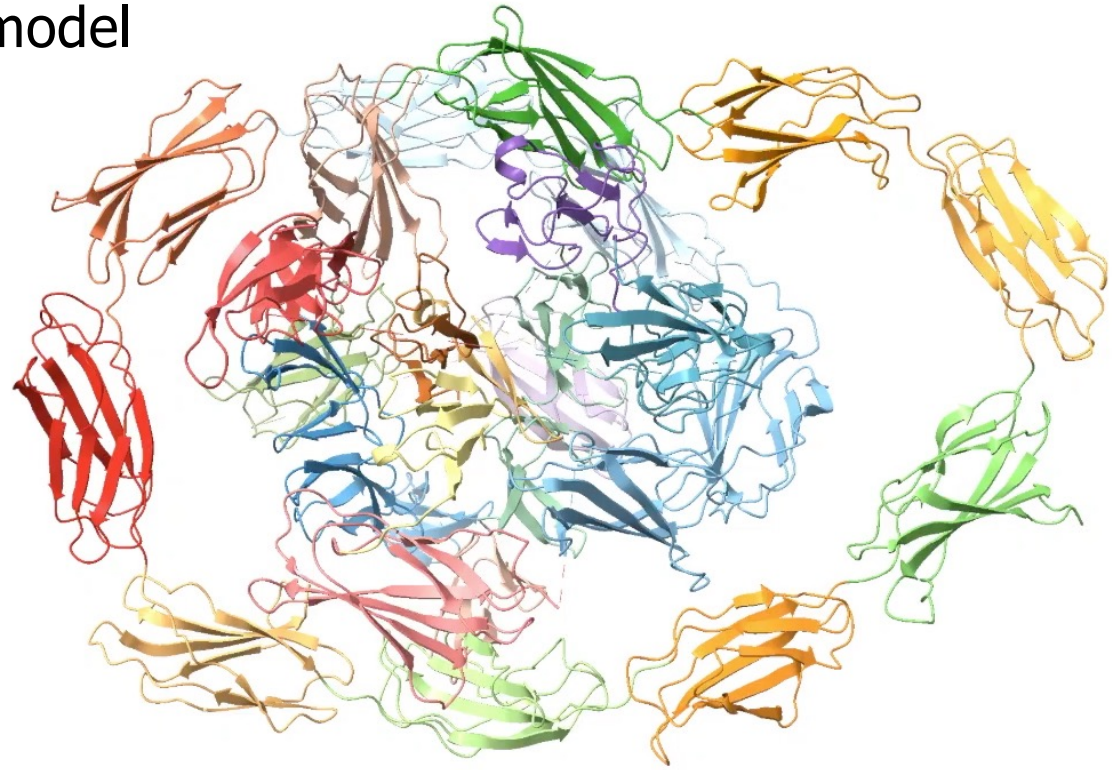
- Fibronectin repeats often have different relative orientations
- Large segments (in red) poorly predicted (or disordered)





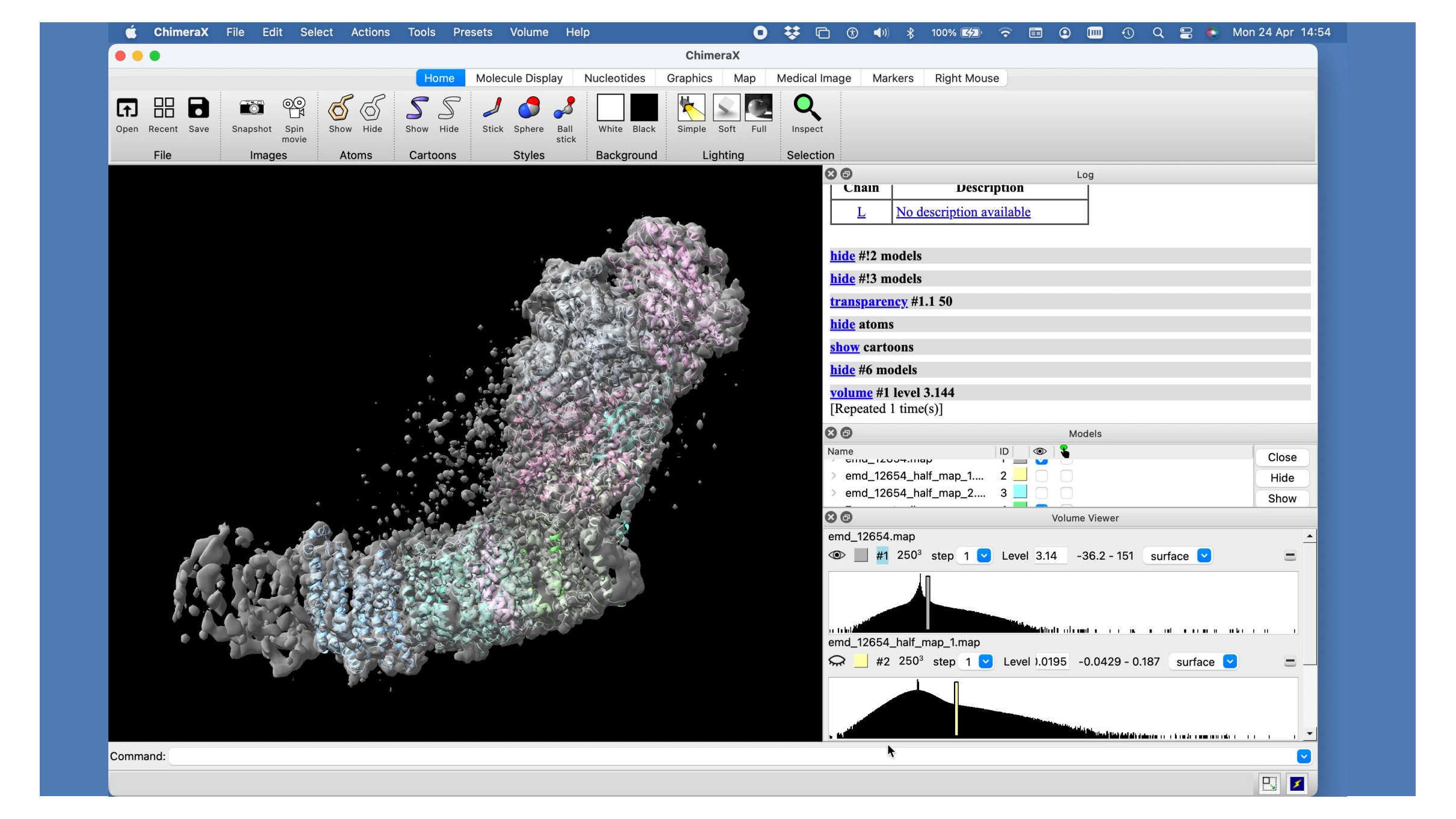
Fibronectin parsed into domains

- Community clustering of PAE matrix (Tristan Croll)
  - phenix.process_predicted_model
    - cctbx library
  - CCP4
  - ChimeraX



Searching in a defined sphere: *emplace_local*

- More sensitive (and much faster) if you know approximately where a molecule should go
- Half-maps are strongly preferred but optional
  - assume that your docking location contains enough ordered density
  - if half-maps aren't provided, resolution must be specified and anisotropy in the signal and noise are neglected
- Easiest to run from new ChimeraX plugin
  - see YouTube tutorials by Dorothee Liebschner
    - https://www.youtube.com/c/phenixtutorials
    - Phenix/ChimeraX playlist


ChimeraX File Edit Select Actions Tools Presets Volume Help
Mon 24 Apr 14:54
ChimeraX
Home Molecule Display Nucleotides Graphics Map Medical Image Markers Right Mouse
Open Recent Save File
Snapshot Spin movie Images
Show Hide Atoms
Show Hide Cartoons
Stick Sphere Ball stick Styles
White Black Background
Simple Soft Full Lighting
Inspect Selection
Log
Chain Description
L No description available
hide #!2 models
hide #!3 models
transparency #1.1 50
hide atoms
show cartoons
hide #6 models
volume #1 level 3.144
[Repeated 1 time(s)]
Models
Name ID
emd_12654_half_map_1.... 2
emd_12654_half_map_2.... 3
Close
Hide
Show
Volume Viewer
emd_12654.map
#1 250³ step 1 Level 3.14 -36.2 - 151 surface
emd_12654_half_map_1.map
#2 250³ step 1 Level .0195 -0.0429 - 0.187 surface
Command:

Future plans

- Add rigid-body refinement of previously placed model, as a whole or in domains
- Account explicitly for point-group or even helical symmetry
- Search for multiple components
  - account for what has already been explained
  - avoid clashes
- Adapt search strategy to local map quality
- Move likelihood function closer to raw data?

Software availability

- Underlying algorithms in open-source CCTBX library

- *phasertng:* upcoming versions of Phenix, CCP4, CCP-EM
- *em_placement:* upcoming versions of Phenix, CCP-EM
- *emplace_local:* Phenix, ChimeraX

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