

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



UNIVERSITY OF
CAMBRIDGE

Randy J Read
Department of Haematology

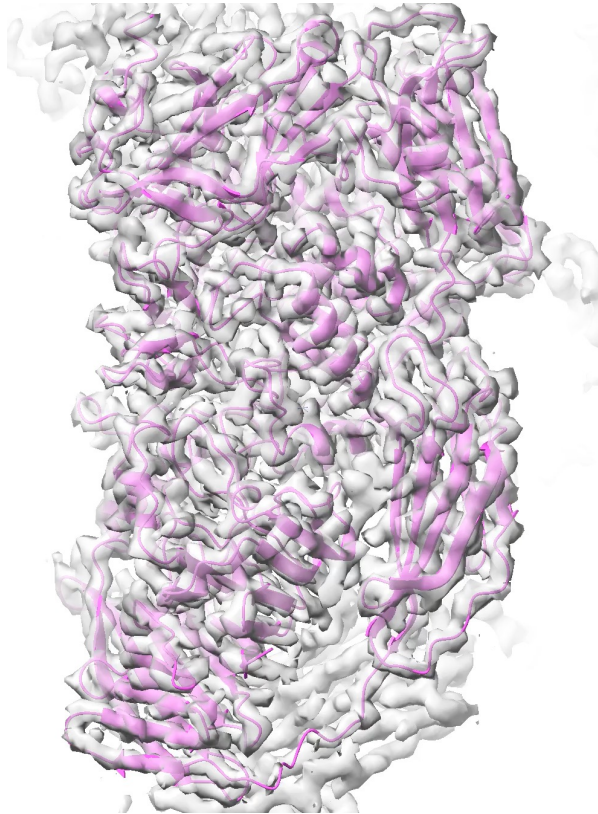


CIMR
Molecules
Mechanisms
Medicine

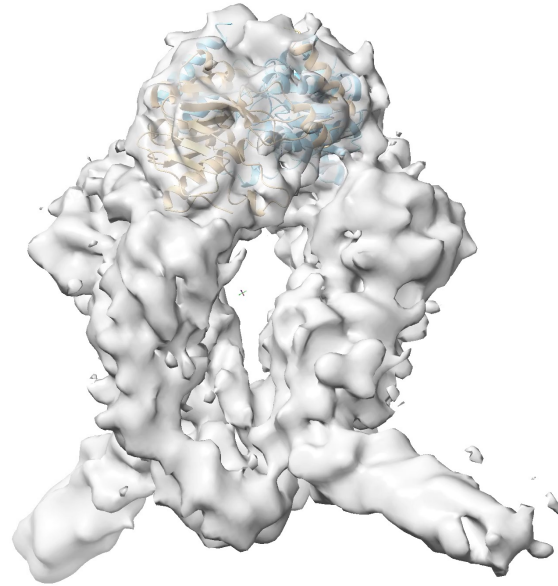
The docking problem in cryo-EM

- 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?
-

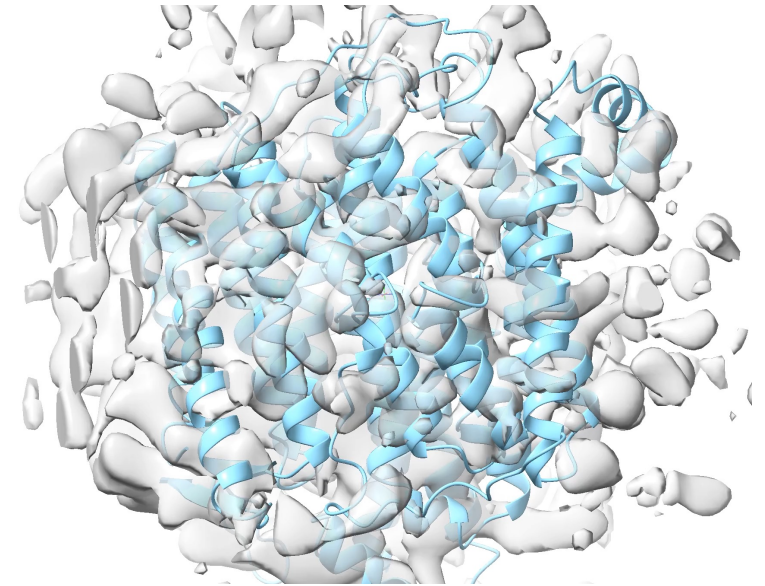
Which docking cases are important?



β -galactosidase
2.2 Å



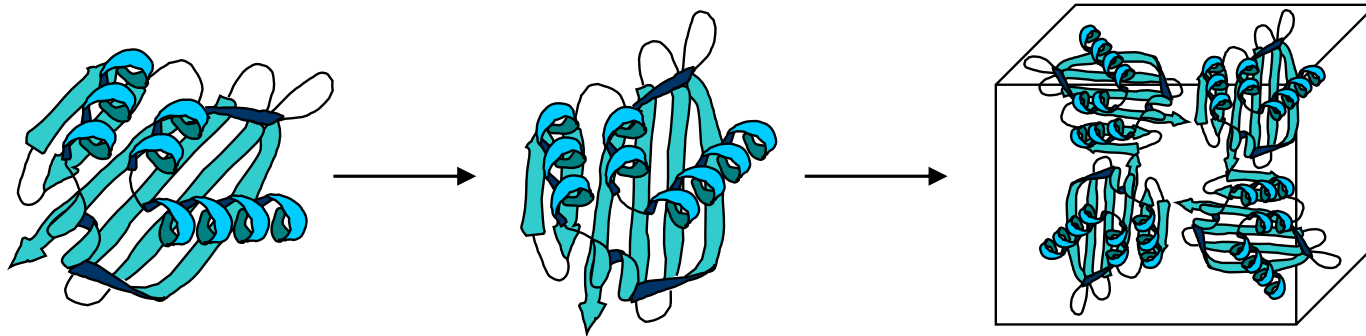
C-terminal domain of MutS
6.9 Å



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

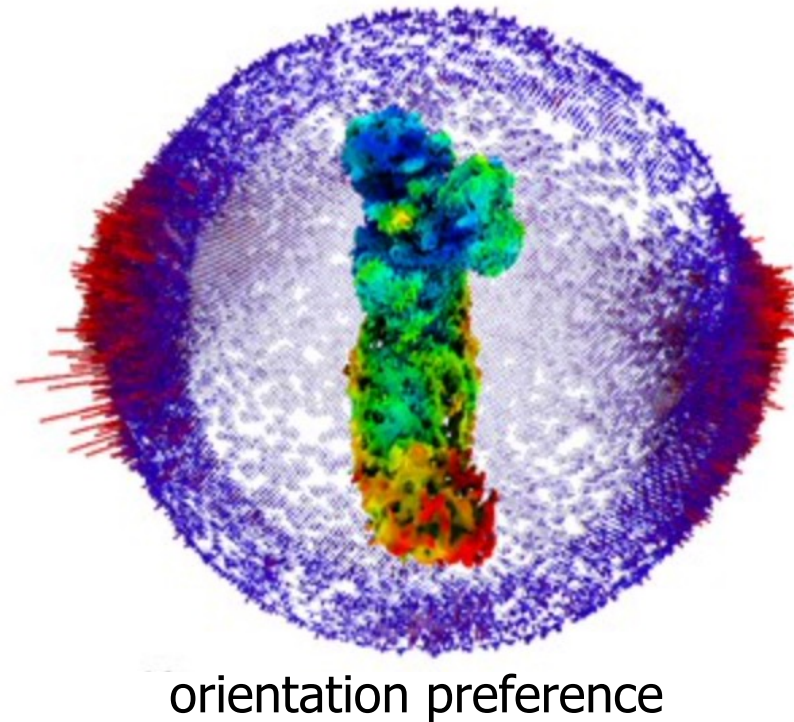
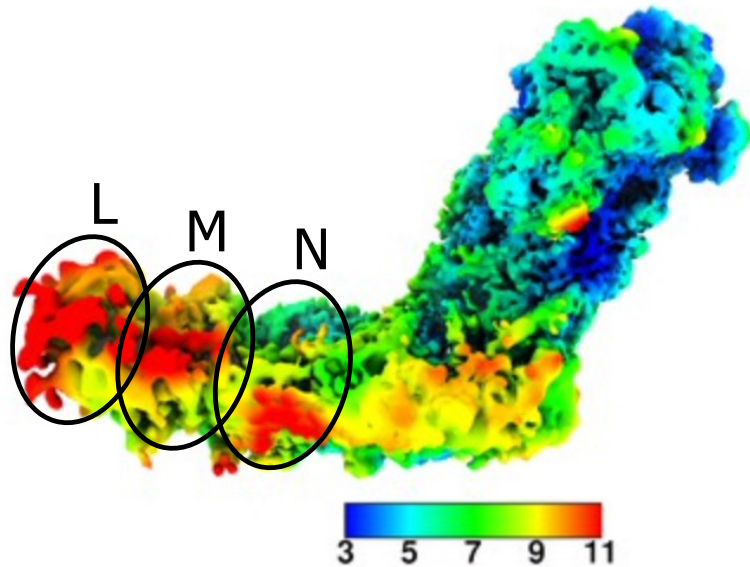
- Measures the consistency of the model (atomic model plus rigid-body rotation/translation parameters) with the data
 - probabilistic: accounts for errors in both model and data
 - Likelihood is an absolute score
 - compare alternative hypotheses
 - judge confidence in solution
 - Achievable score can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



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
-

CCP-EM EM_placement GUI

The screenshot displays the CCP-EM Doppio web interface. The top navigation bar includes the user 'randy@skate-2.local', the project path '~/phenix/cryodocking/Clemons_25375', and the application name 'CCP-EM Doppio'. The main interface is divided into a left sidebar and a right main panel.

Left Sidebar: Contains navigation tabs for 'PROJECT', 'JOBS', 'NODES', and 'NEW JOB'. Below these is a search bar 'Filter jobs by name or descripti' and an 'Expand all' button. A list of job categories is shown, including '2D Classification', '3D Classification', '3D Refinement', 'Atomic Model Build', and 'Atomic Model Fit'. Under 'Atomic Model Fit', three jobs are listed:

- em_placement** (EA icon): em_placement.atomic_model_fit - Fit model in whole map with em_placement
- emplace_local** (EA icon): emplace_local.atomic_model_fit - Locally fit model in map with em_placement
- Molrep** (MA icon): molrep.atomic_model_fit - Fit model in map with Molrep

Main Panel: Titled 'em_placement', it features three buttons: 'RUN' (green), 'JOB INFO' (green), and 'RESET OPTIONS' (red). Below these is a 'Job alias:' input field. The 'Main' configuration section includes:

- Model ***: 7spzA.pdb
- Sequence composition ***: 7spz.seq
- Input half map 1 ***: emd_25375_half_map_1.map
- Input half map 2 ***: emd_25375_half_map_2.map
- Best resolution ***: 8.46
- Point group symmetry**: C1
- Estimated RMSD ***: 1.2

The 'Queue submission options' section includes:

- Submit to queue?**: Radio buttons for 'Yes' and 'No' (selected).
- Queue name:** openmpi
- Queue submit command:** qsub

CCP-EM EM_placement GUI

CCP-EM Doppio – randy@skate-2.local

randy@skate-2.local Project: ~/phenix/cryodocking/Clemons_25375

PROJECT JOBS NODES NEW JOB

Filter jobs by name or descripti Expand all

- 2D Classification
- 3D Classification
- 3D Refinement
- Atomic Model Build
- Atomic Model Fit
 - em_placement
em_placement.atomic_model_fit – Fit model in whole map with em_placement
 - emplace_local
emplace_local.atomic_model_fit – Locally fit model in map with em_placement
 - Molrep
molrep.atomic_model_fit – Fit model in map with Molrep
- Atomic Model Refine
- Atomic Model Split
- Atomic Model Utilities

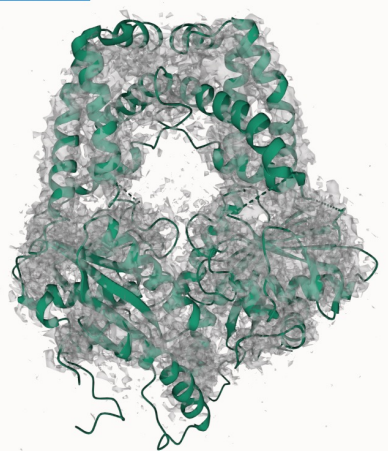
RESULTS LOGS I/O OPTIONS

Job 1 – EMplacement

```
-2      8      18      28      38      48      58      68      78      108
ATYMLPSLHDILDQHTYKWIFFGGKGVGKTTSSSFVSLMAETRPNEKFLLLSTDPAHNISDAFDQKFGKAPTQVSGIPNLYAMEVDDNDAAE
SKSEGDMEFGLNDLITCASSFKIDGTFPGDEMWSFINLIKLINEYSTVIFDTAPTGHTRFLELPETVNVKLEIFTRLKDNDMGGMLSMVMQT
MGLSQNDIFGLIDKTYPKIDVVKRISAEFRDPSLCTFVGVCIPEFLSLYETERLVQRLAVLDMDCCHAIVINFLVDANAATPCSMCRSRARMQN
```

Volume

Atomic Models



Structure

2 structures

Nothing Focused

Measurements

+ Add

Superposition

Chains Atoms

Quick Styles

Default Stylized Illustrative

Components 2 structures

Preset + Add

Polymer Cartoon

Volume

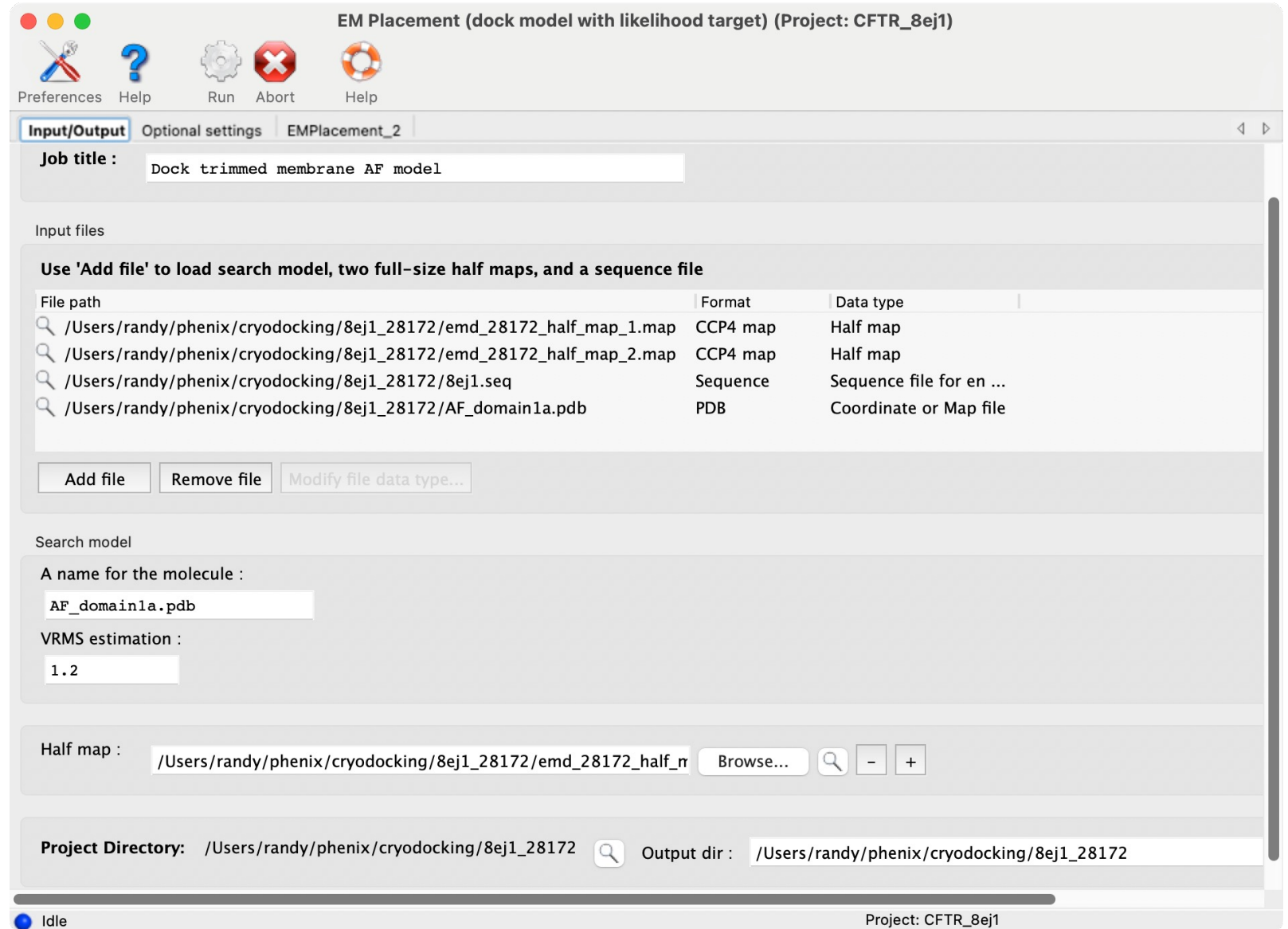
emd_25375_half_map_1.map

Isosurface 4.55 σ

Global map-fit scores

Solution_name	Model	Docking_region	mapLLG	Z-score	Mtz-ModelSol-CC
431744877-rmr-1.coord...	7spzA.pdb	focus1_691730648022_...	433.78	51.929	0.691
431744878-rmr-2.coord...	7spzA.pdb	focus2_691730648022_...	287.967	43.074	0.639

Phenix EM_placement GUI

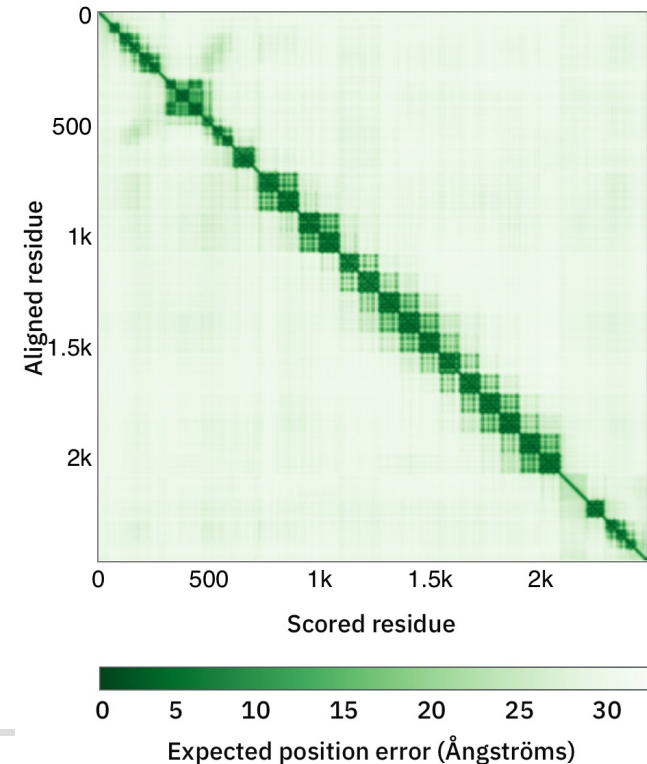
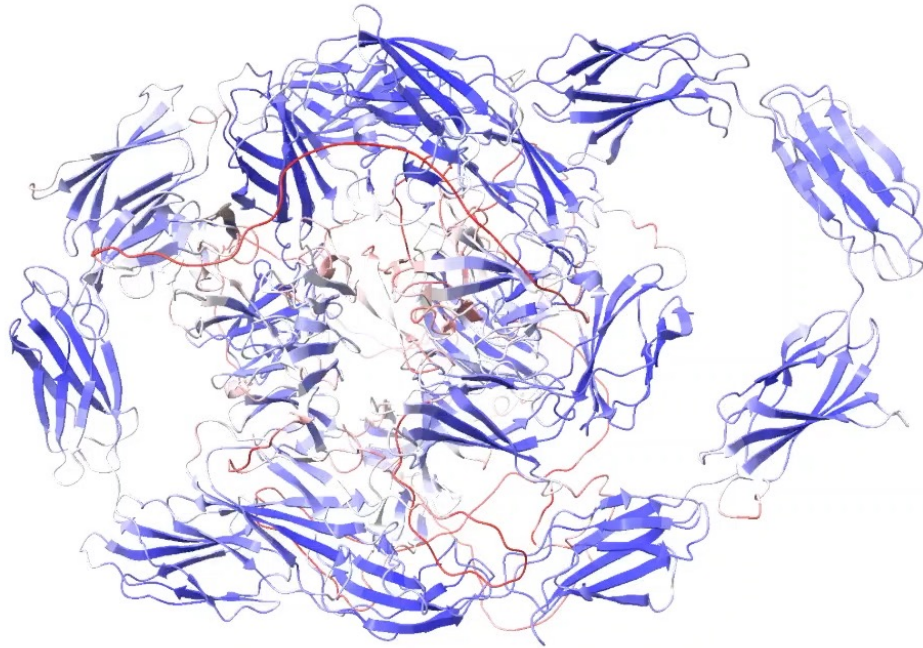


Process_predicted_model

- 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
-

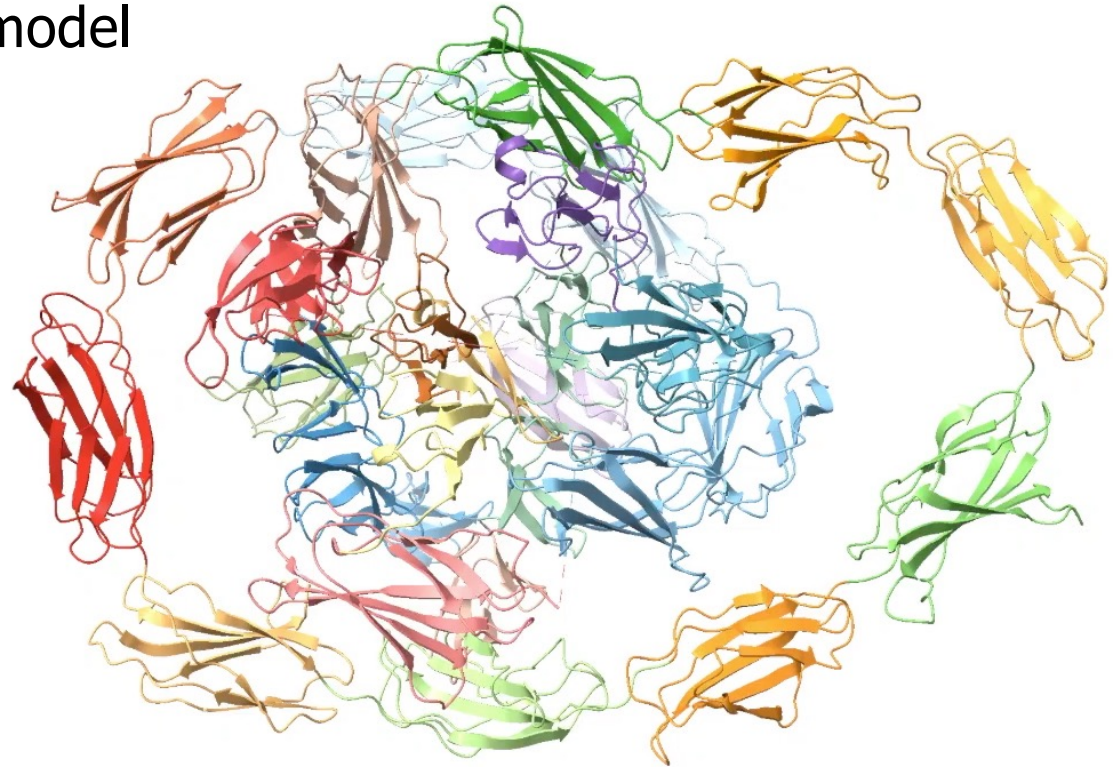
AlphaFold model of human fibronectin

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



Fibronectin parsed into domains

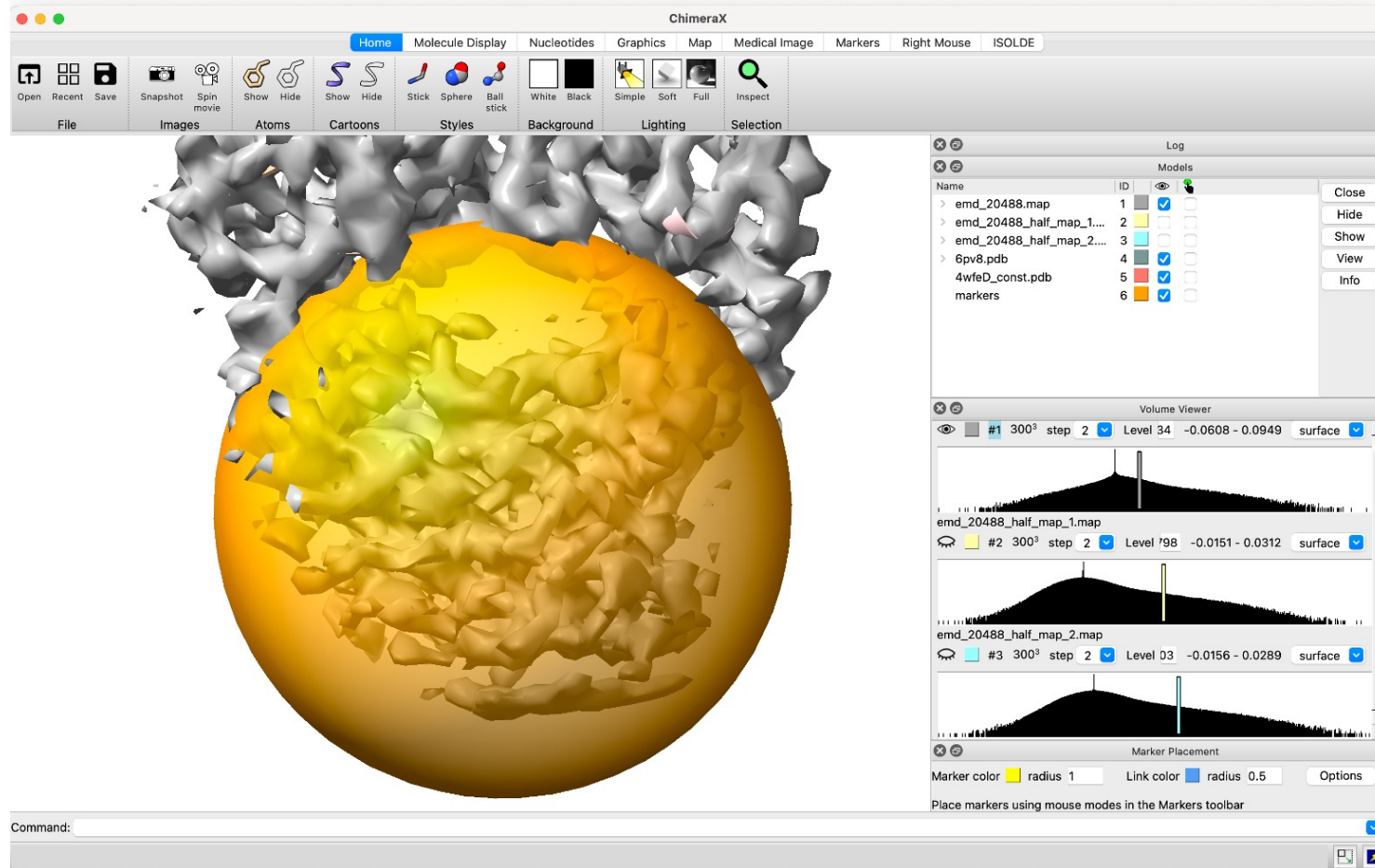
- Community clustering of PAE matrix (Tristan Croll)
 - phenix.process_predicted_model
 - cctbx library
 - CCP4
 - ChimeraX



Searching in a defined sphere: *emplace_local*

- More sensitive (and much faster) if you know approximately where a molecule should go
 - half-maps are recommended but not essential
 - Easiest to run from ChimeraX plugin
 - see YouTube tutorials by Dorothee Liebschner
 - <https://www.youtube.com/c/phenixtutorials>
 - Phenix/ChimeraX playlist
 - Read, Pettersen, McCoy, Croll, Terwilliger, Poon, Meng, Liebschner & Adams. *Structural Biology (Acta Cryst D)*, 2024
-

ChimeraX emplace_local GUI



New feature: accounting for placed components

- Most difficult cases: poorly-ordered component next to a well-ordered component in the map
 - docking tends to be confused by a large part of the signal being unexplained by the search model
 - Address this by masking out the explained part of the map?
 - no, this violates an assumption that noise is distributed pretty evenly over the map
 - Subtract contribution of placed components from map?
 - this works and is easier to implement
 - Add contribution of placed component to moving component
 - possibly better in theory, implementation in progress
-

Demos

- EM-placement in Doppio
 - Get3 targeting factor in closed conformation at 8.5 Å resolution
 - asymmetric dimer
 - EMD-25375
 - Use monomer from closed form as a model
 - `emplace_local` in ChimeraX
 - place 60 copies of monomer in icosahedral virus capsid
 - EMD-28022
 - 2.4 Å resolution
-

Future plans

- Account explicitly for point-group or even helical symmetry
 - Search automatically for multiple components
 - Adapt search strategy to local map quality
 - Move likelihood function closer to raw data?
-

Software availability

- Underlying algorithms in open-source CCTBX library
 - *em_placement*: Phenix, CCP-EM Doppio
 - *emplace_local*: Phenix, ChimeraX, CCP-EM Doppio
-

Acknowledgements

- Claudia Millán
- Airlie McCoy
- Tristan Croll

- Tom Terwilliger
- Dorothee Liebschner
- Billy Poon

- Eric Pettersen
- Tom Goddard

Tom Burnley

Cathy Lawson



Phenix
*An NIH/NIGMS funded
Program Project*
