Model Refinement: cryo-EM

Pavel Afonine

phenix-online.org

lbl.gov

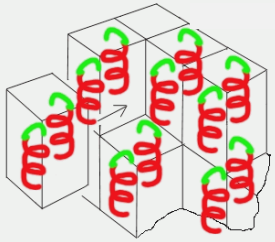
qrefine.com

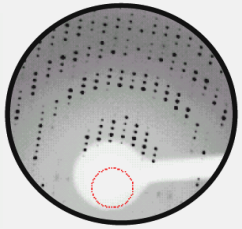
Missoula, MT

June 28th 2024

Refinement in Phenix

Crystallography





Initial model

Experimental data

A priori knowledge

Score

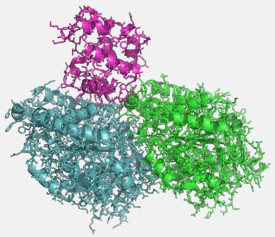
Modify model parameters

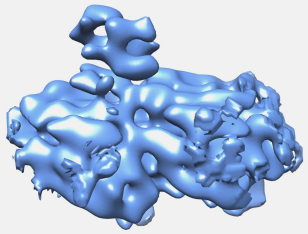
Improved model

phenix.refine

Available since 2005

Cryo-EM





Initial model

Experimental data

A priori knowledge

Score

Modify model parameters

Improved model

phenix.real_space_refine

Available since 2013

Atomic model refinement: crystallography vs cryo-EM

Crystallographic refinement

- Improving model improves map
  - (2mFo-DFc, Model phase), (mFo-DFc, Model phase)
  - Better model leads to better map
  - Better map leads to more model built
  - Improving model in one place lets build more model elsewhere in the unit cell
  - Refine all model parameters (XYZ, B) from start to end of structure solution
  - Build solvent (ordered water) early
- Experimental data never changed
- Data / restraints weight is global and time expensive to find best value
- Whole model needs to be refined

Cryo-EM refinement

- Changing model does not change map
  - Build solvent (water) last
  - Get as complete and accurate model as possible before refining B factors and occupancies
- Experimental data changes a lot during the process (filtering, boxing, using maps with implied symmetry or not, etc.)
  - What map to use in refinement?
  - Refined B factors depend on map used
- Data / restraints weight can be local and is always optimal
- Boxed parts of the model can be refined

Atomic model refinement: *phenix.real_space_refine*

Acta Cryst D STRUCTURAL BIOLOGY

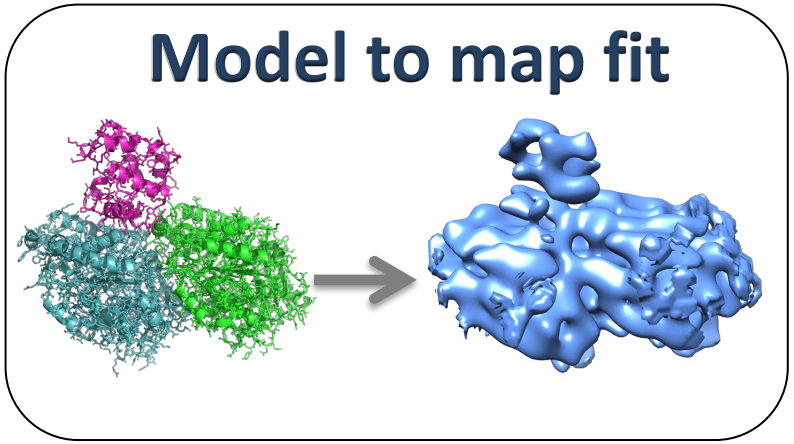
ISSN 2059-7983

Real-space refinement in *PHENIX* for cryo-EM and crystallography

Pavel V. Afonine,[a,b]* Billy K. Poon,[a] Randy J. Read,[c] Oleg V. Sobolev,[a] Thomas C. Terwilliger,[d,e] Alexandre Urzhumtsev[f,g] and Paul D. Adams[a,h]

How we evaluate refinement progress (model-to-map fit) or what's the analogue of crystallographic R-factor?

Model-to-map fit validation: CC_{MASK}



$$CC_{MASK} = \frac{\sum \rho_{obs}\, \rho_{calc}}{(\sum \rho_{obs}^{2}\ \sum \rho_{calc}^{2})^{1/2}}$$

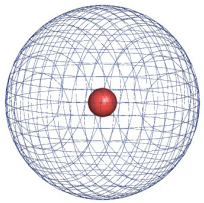
ρ_{obs} = experimental map
ρ_{calc} = model calculated map

- Easy interpretation: -1: anticorrelation, 0: no correlation, 1: perfect correlation
- Uses all atomic model parameters (XYZ, B-factors, occ, atom type)
- Not specific to map type (any map: x-ray, neutron, electron, cryo-EM, …)
- Can be calculated locally (per atom, residue, chain, molecule, whole box, …)
  - Local resolution can be trivially taken into account

Metric	Expected value
CC_{MASK}	Poor: < 0.3 So-so: 0.3-0.6 Good: > 0.6

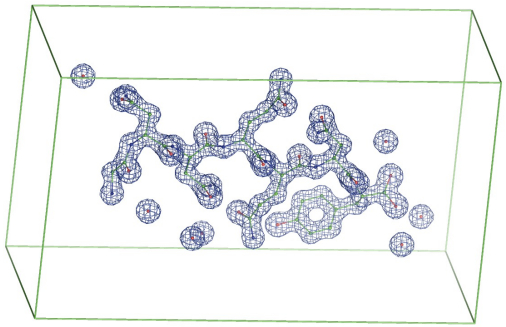
Model-to-map fit validation: CC_{MASK}

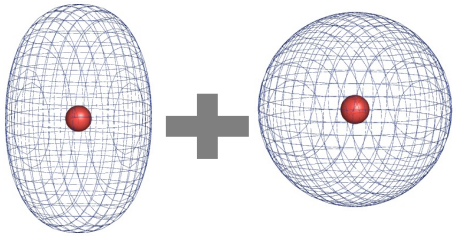
- Gaussian IAM (Independent Atom Model)



$$\rho_{atom}(\mathbf{r},\mathbf{r}_0,B,q) = q\sum_{k=1}^{5} a_k \left(\frac{4\pi}{b_k+B}\right)^{3/2} \exp\left(-\frac{4\pi^2|\mathbf{r}-\mathbf{r}_0|^2}{b_k+B}\right)$$

```
ATOM     25  CA  PRO A   4      31.309  29.489  26.044  1.00 57.79           C
```

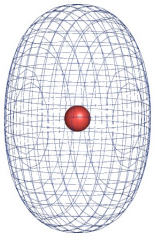




$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

Model map

- Gaussian IAM (Independent Atom Model)
- Anisotropic:

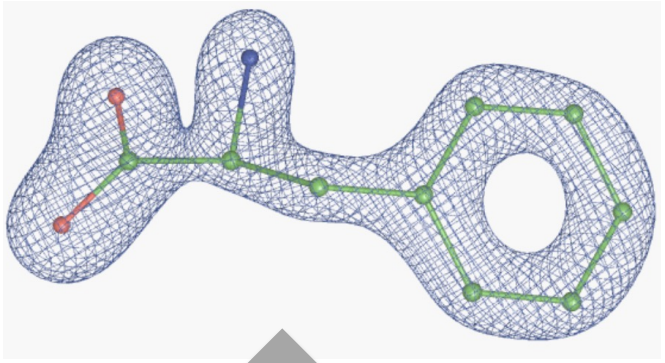


$$\rho_{atom}(\mathbf{r},\mathbf{U},q) = q\sum_{j=1}^{5}\frac{q\,a_j\,(4\pi)^{3/2}}{\left|8\pi^2\mathbf{U}_{\text{cart}} + b_j\mathbf{I}\right|^{1/2}}\exp\left(-4\pi^2(\mathbf{r}-\mathbf{r}_0)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left[8\pi^2\mathbf{U}_{\text{cart}} + b_j\mathbf{I}\right]^{-1}\mathbf{A}(\mathbf{r}-\mathbf{r}_0)\right)$$

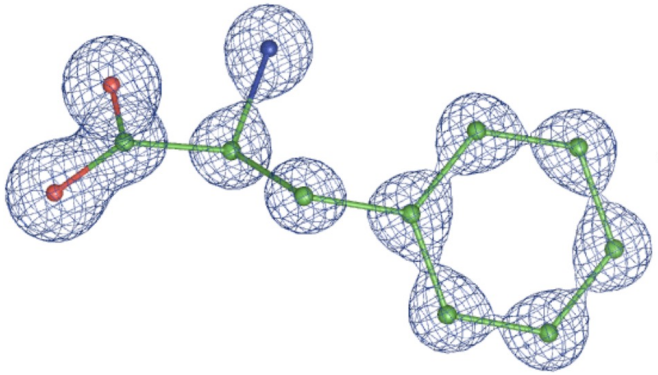
```
ATOM     25  CA  PRO A   4      31.309  29.489  26.044  1.00 57.79           C
ANISOU   25  CA  PRO A   4    8443   7405   6110   2093    -24    -80       C
```

Model-to-map fit validation: CC_{MASK}

3Å model-calculated map



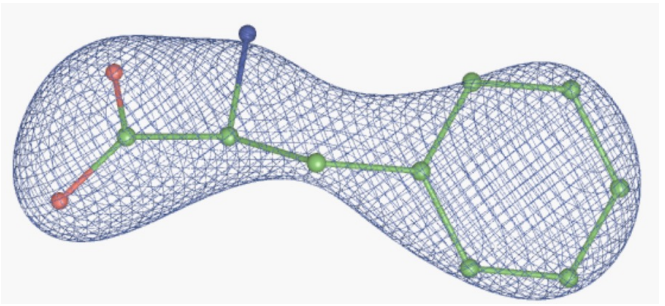
Exact model map



CC_{MASK}

CC_{MASK}

3Å experimental map



$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

- FT exact model map
- Remove terms up to specified resolution
- FT back to real space to get a Fourier image = “Model map”

Other popular model-to-map fit metrics and reasons why they are not as good as CCmask

Atom inclusion

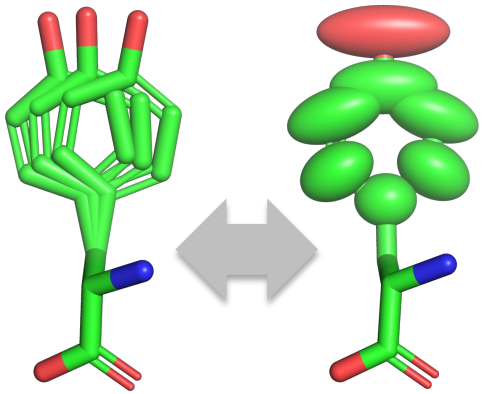
- **Atom inclusion**: fraction of atoms inside molecular envelope contoured at a given level
  - Contouring threshold: Arbitrarily? What is optimal level?
  - No use of atomic model parameters such as ADP, occupancy, atom type, …
  - Does not compare shape of density:
    - How SER placed into PHE density is going to score?
    - How water O placed into Mg peak will score?
  - Does not account for missing atoms
  - Does not use map type (x-ray, neutron, electron)
  - Partially occupied atoms (alternative conformations):
    - Chosen level for fully occupied atoms needs to be scaled by occupancy for partially occupied atoms

Q-Score

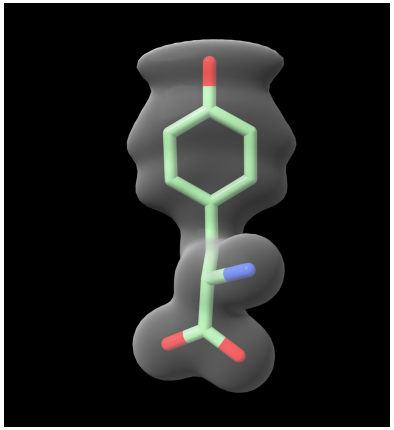
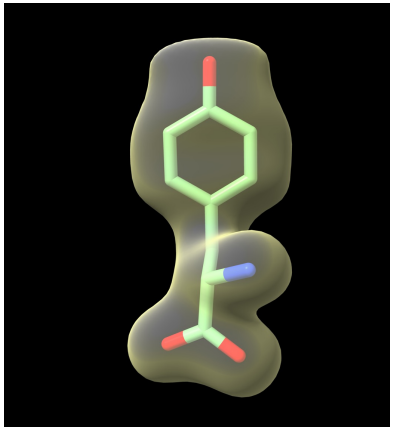
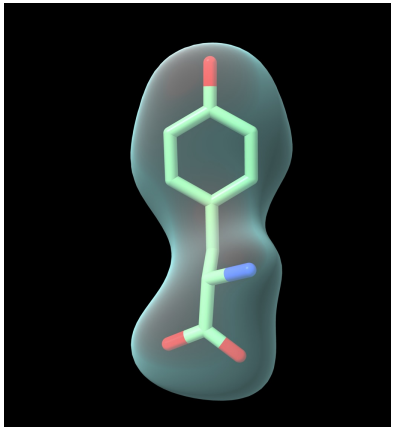
- **Q-score**: measure the resolvability of individual atoms in a cryo-EM map, using an atomic model fitted to or built into the map
  - No use of atomic model parameters such as ADP, occupancy, atom type, …
  - Shape of density:
    - How SER placed into PHE density is going to score?
    - How water O placed into Mg peak will score?
  - Does not account for missing atoms (it shouldn't given the definition)
  - Alternative conformations are ***not*** handled
  - How anisotropic atoms are ***not*** handled
  - Does not use map type (x-ray, neutron, electron)

Example: Q-Score for exact (model-generated) map

Model with anisotropic ADPs



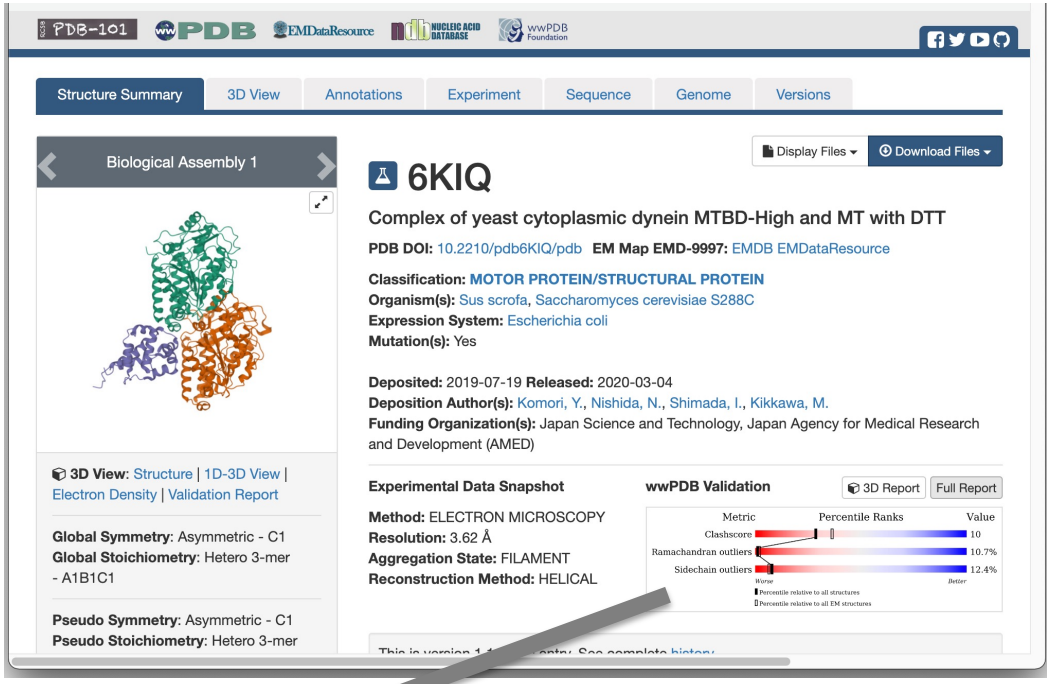
Model calculated maps (Fourier maps) at three resolutions

1 Å	2 Å	3 Å
		
0.97 (0.90)	0.93 (0.83)	0.84 (0.76)

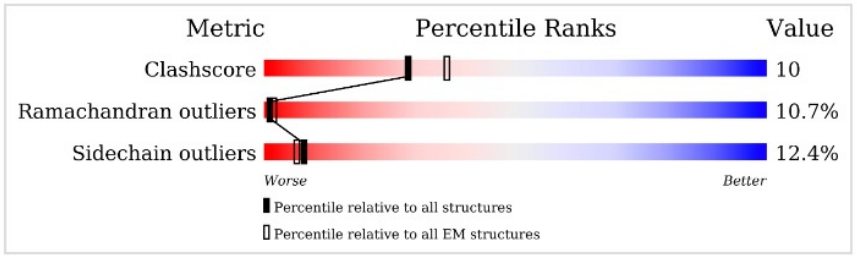
Overall and worst Q-Score (calculated in ChimeraX)

- Why Q-Score is not perfect (=1) given these are exact model-generated maps?
- Why it varies with the resolution?

Validation reports (RCSB): only Q-score and atom inclusion



wwPDB Validation 3D Report Full Report



9.5 Map-model fit summary

The table lists the average atom inclusion at the recommended contour level (0.125) a for the entire model and for each chain.

Chain	Atom inclusion	Q-score
All	0.9062	0.4550
M	0.5810	0.3210
a	0.9659	0.4790
b	0.9656	0.4730

Model-to-map fit statistics is insufficient and very well hidden!

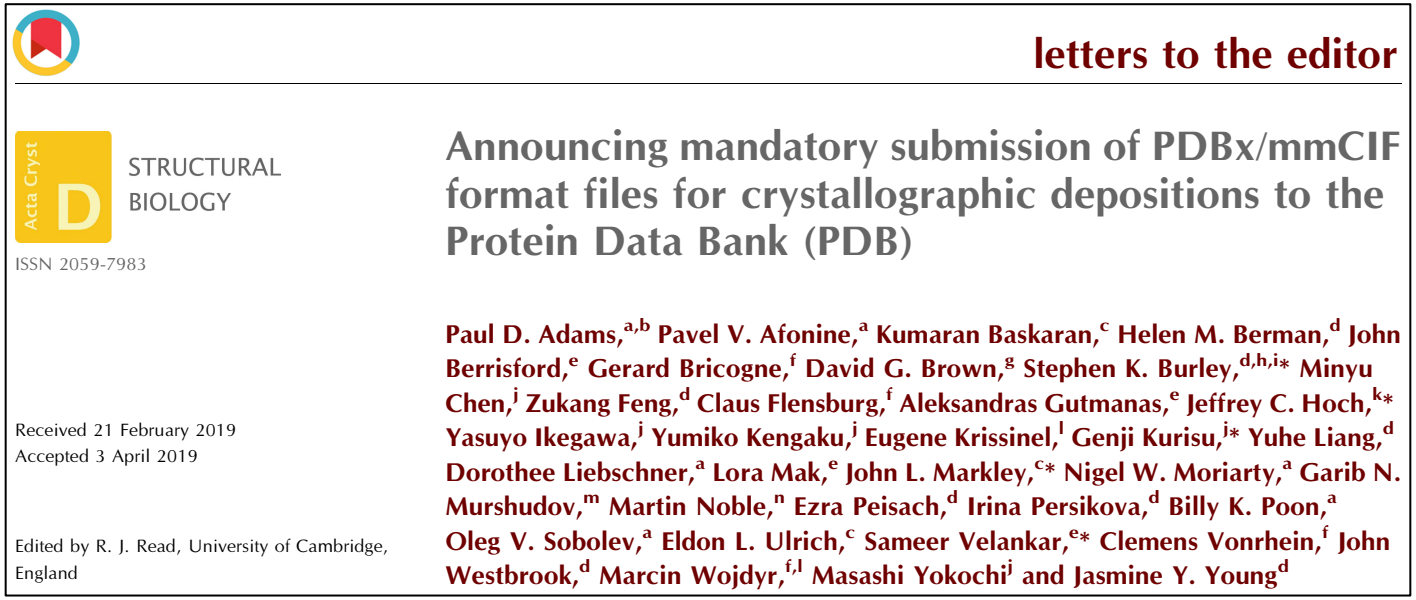
Refinement: practical considerations

- Final stages
  - Refine B-factors (Atomic Displacement Parameters)
    - Group B factor or individual
  - Refine occupancies
  - Use Hydrogen atoms (and keep them in the final model!)
  - Add water (phenix.douse: command line and GUI):

    Also available in ChimeraX

mmCIF

- mmCIF file format for atomic models
  - Mandatory use for crystallographic models since July 2019
    - PDB formatted files are not accepted any more
  - Some cryo-EM models may be too large to fit into PDB file format
  - *Phenix* provides full support for mmCIF I/O



letters to the editor

Acta Cryst D STRUCTURAL BIOLOGY

ISSN 2059-7983

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)

Paul D. Adams,[a,b] Pavel V. Afonine,[a] Kumaran Baskaran,[c] Helen M. Berman,[d] John Berrisford,[e] Gerard Bricogne,[f] David G. Brown,[g] Stephen K. Burley,[d,h,i]* Minyu Chen,[j] Zukang Feng,[d] Claus Flensburg,[f] Aleksandras Gutmanas,[e] Jeffrey C. Hoch,[k]* Yasuyo Ikegawa,[j] Yumiko Kengaku,[j] Eugene Krissinel,[l] Genji Kurisu,[j]* Yuhe Liang,[d] Dorothee Liebschner,[a] Lora Mak,[e] John L. Markley,[c]* Nigel W. Moriarty,[a] Garib N. Murshudov,[m] Martin Noble,[n] Ezra Peisach,[d] Irina Persikova,[d] Billy K. Poon,[a] Oleg V. Sobolev,[a] Eldon L. Ulrich,[c] Sameer Velankar,[e]* Clemens Vonrhein,[f] John Westbrook,[d] Marcin Wojdyr,[f,l] Masashi Yokochi[j] and Jasmine Y. Young[d]

Received 21 February 2019
Accepted 3 April 2019

Edited by R. J. Read, University of Cambridge, England

Variability refinement

Treasuring conformational changes

Contents lists available at ScienceDirect

BBA - Biomembranes

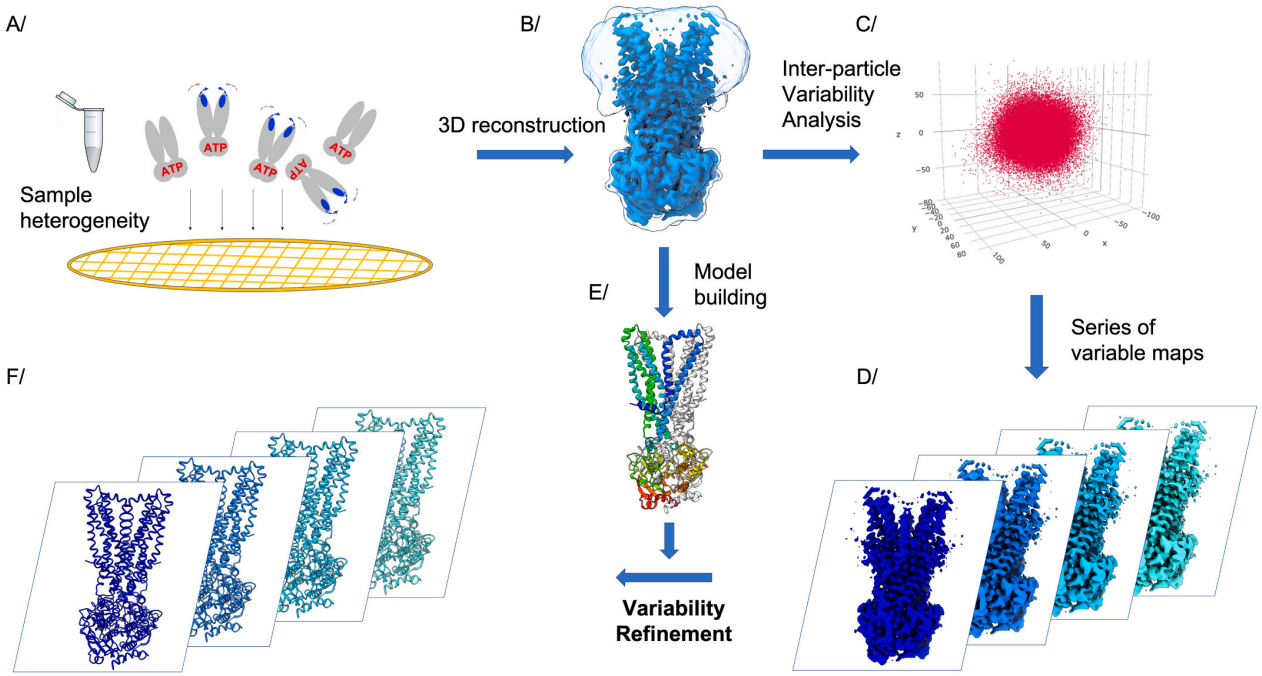
journal homepage: www.elsevier.com/locate/bbamem

Review

Conformational space exploration of cryo-EM structures by variability refinement

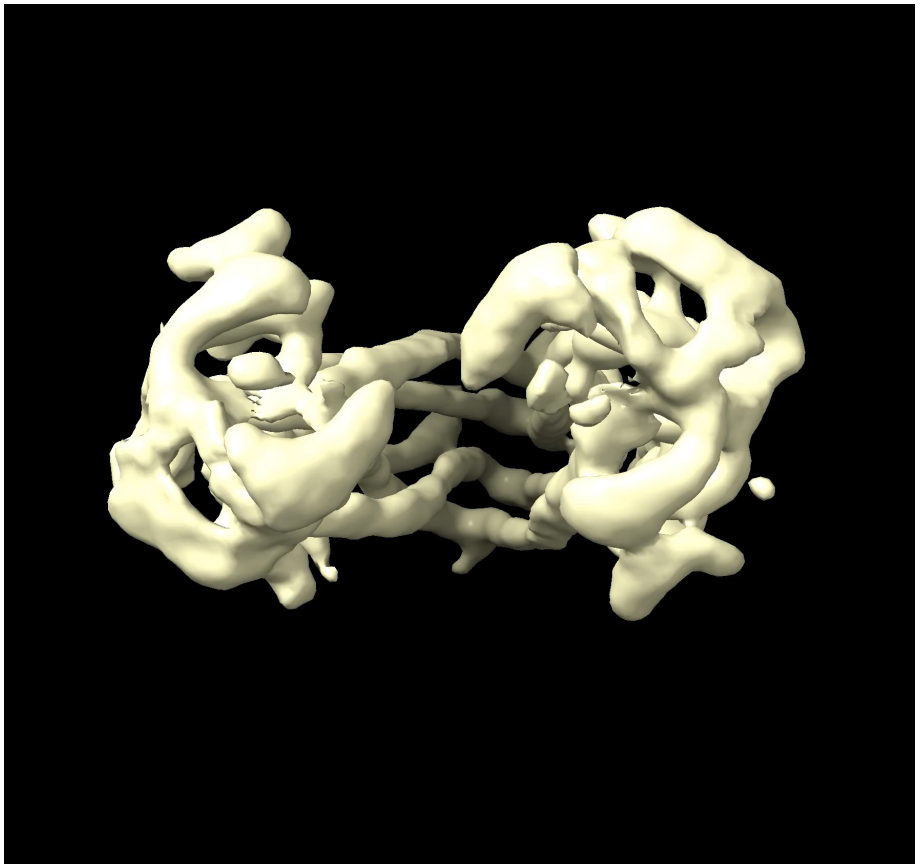
Pavel V. Afonine [a,*], Alexia Gobet [b], Loïck Moissonnier [b], Juliette Martin [b], Billy K. Poon [a], Vincent Chaptal [b,*]

[a] Molecular Biosciences and Integrated Bioimaging, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[b] Molecular Microbiology and Structural Biochemistry, UMR5086 CNRS University Lyon1, 7 passage du Vercors, 69007 Lyon, France



Maps

ABC transporter BmrA (unpublished!)



***phenix.varref* – Phenix tool to represent ensemble of maps with ensemble of atomic models**

phenix.varref

map1.mrc … mapN.mrc

model.pdb

resolution=3

nproc=100

models_per_map=100

Output: ensemble of refined models that represents all maps

Workflow

- Input model and maps
- Order maps by similarity using CC_{box}
- Identify the map that is closest to input model (by CC_{mask})
  - This is the starting point for the first refinement
  - Generate ensemble of 100 perturbed models (by MD)
  - Refine each model with *phenix.real_space_refine*
  - Combine all refined models to yield overall best fitting model
- Refine ensemble of refined models against the next closest map
  - Combined all refined models to yield overall best fitting model
- …and so on for all maps.
- Result:
  - N models corresponding to N maps
  - 100 models per map (can be used to estimate uncertainty)

Refined ensembles of models