Refinement, Validation and more

Pavel Afonine

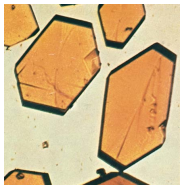
phenix-online.org lbl.gov

July 18th, 2025
ACA, Lombard, IL

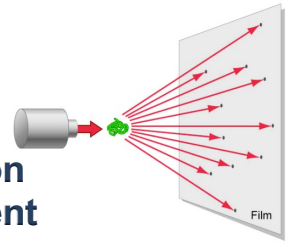
Solving structure - crystallography

Pure sample

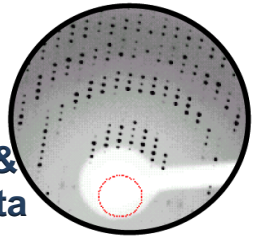
Crystals



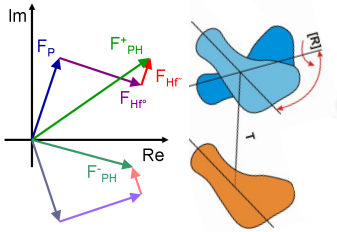
Diffraction experiment



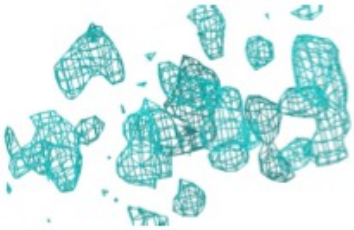
Process & check data



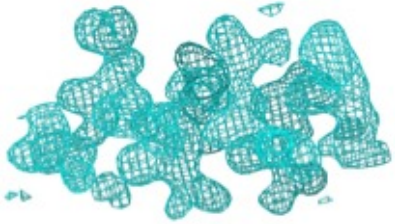
Solve phase problem



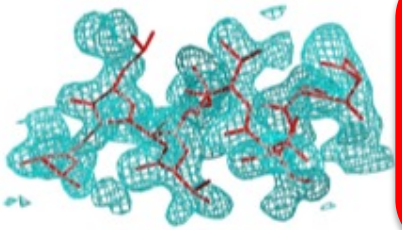
Initial map



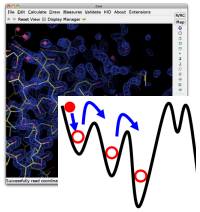
Improve d map



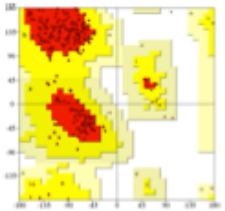
Build model



Refine model



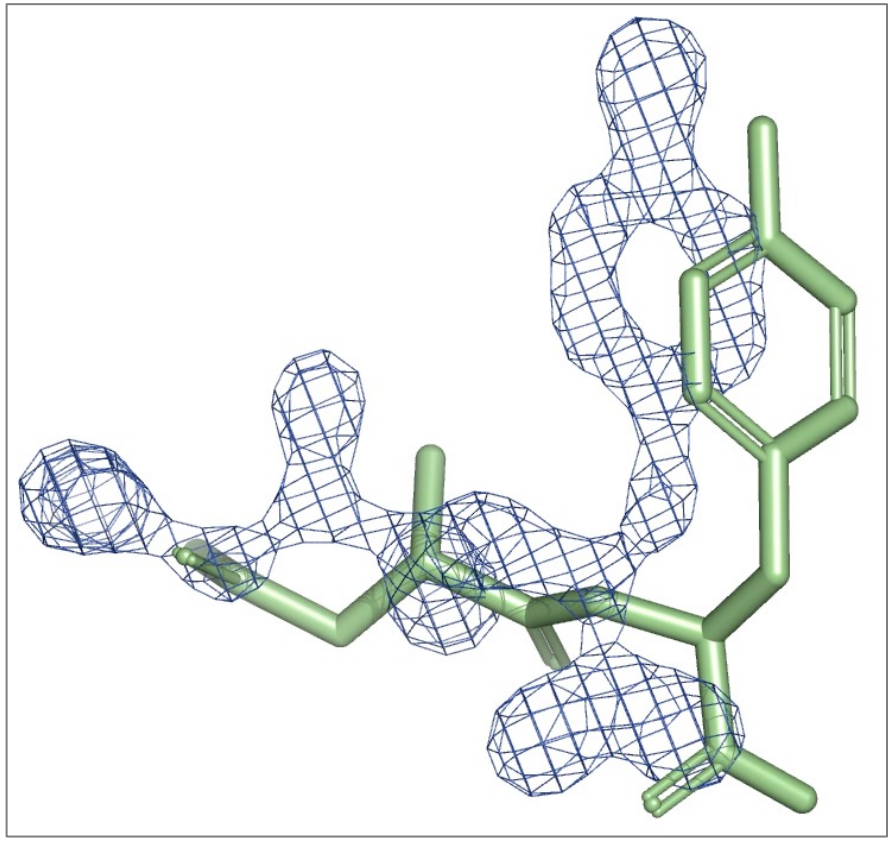
Validate model

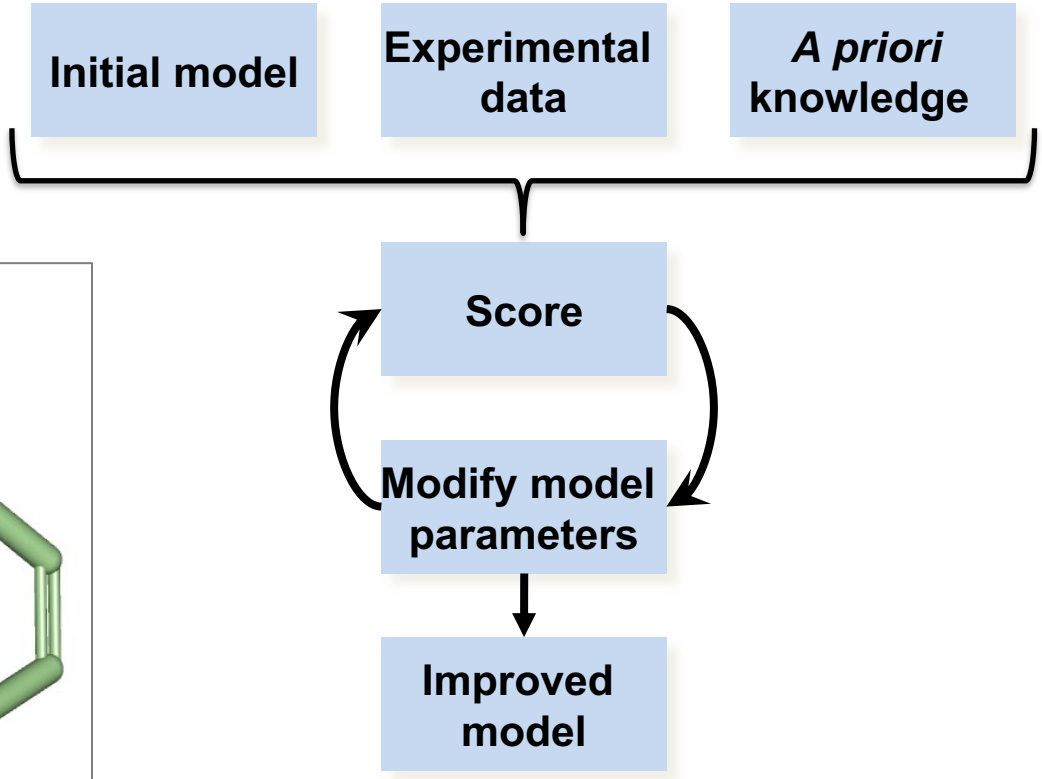


Publish & deposit

- Process is not as ‘linear’ as shown
- Each step has numerous sub-steps
- Crystals may not grow or exhibit pathologies
- Stuck solving phase problem

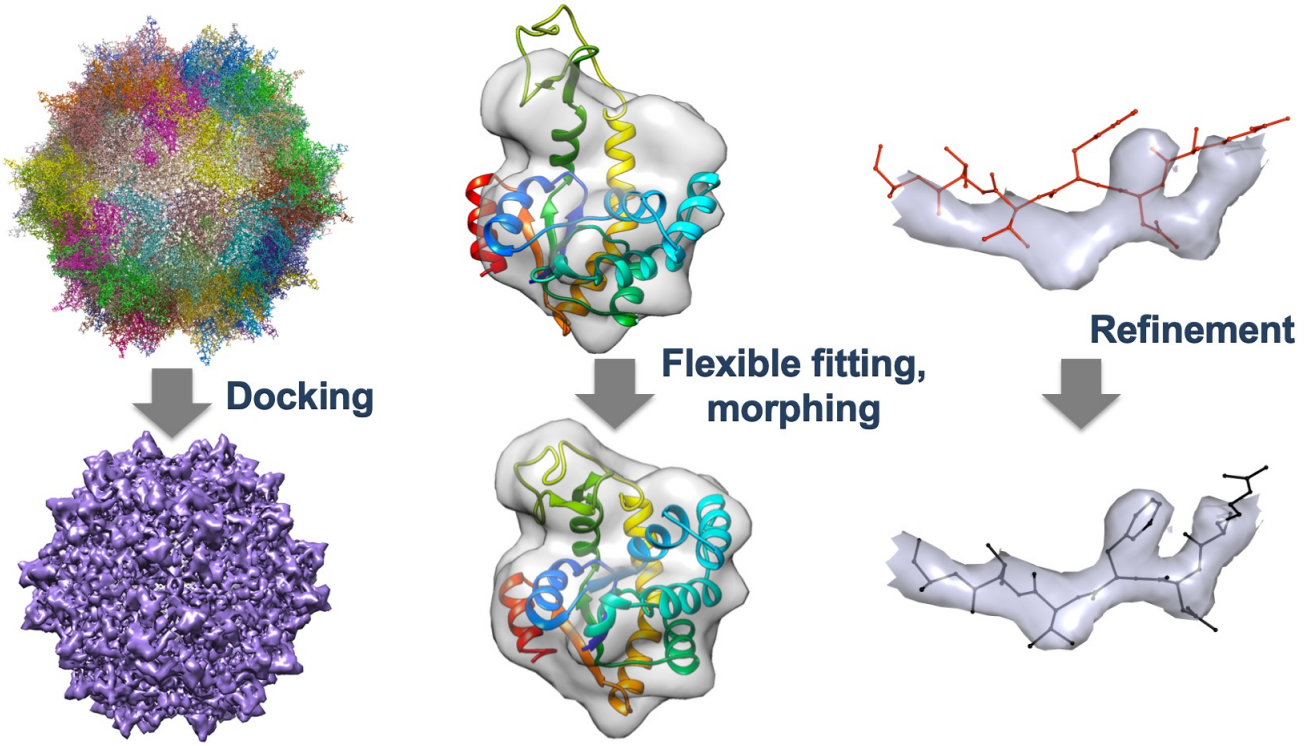
Model refinement





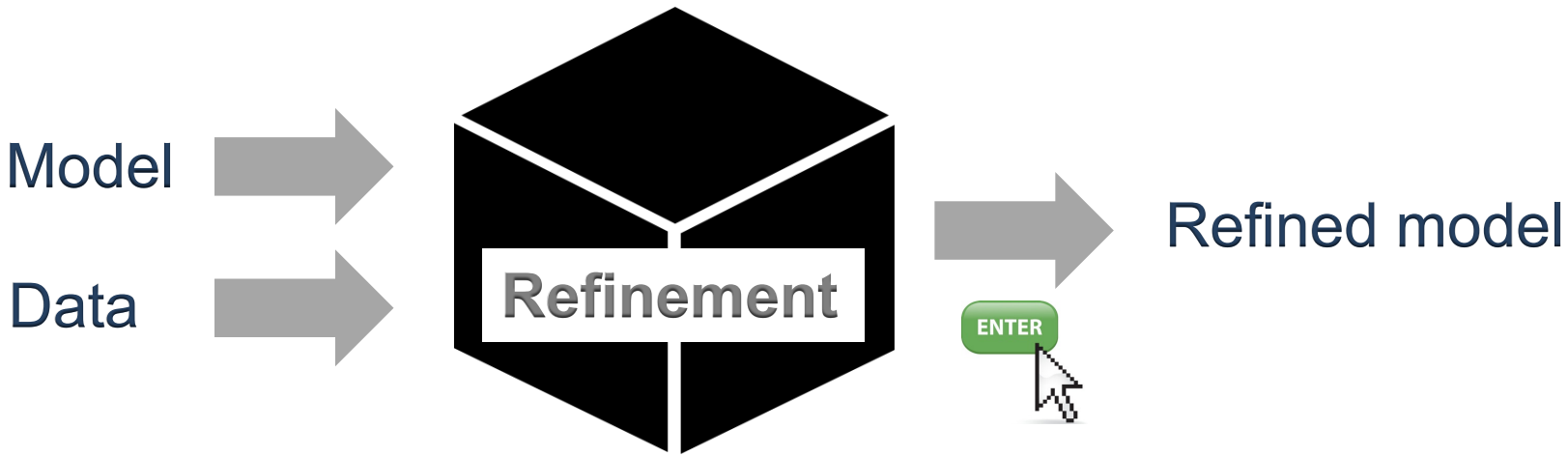
Optimization process of fitting atomic model parameters to experimental data

Not all model-to-data fitting is refinement

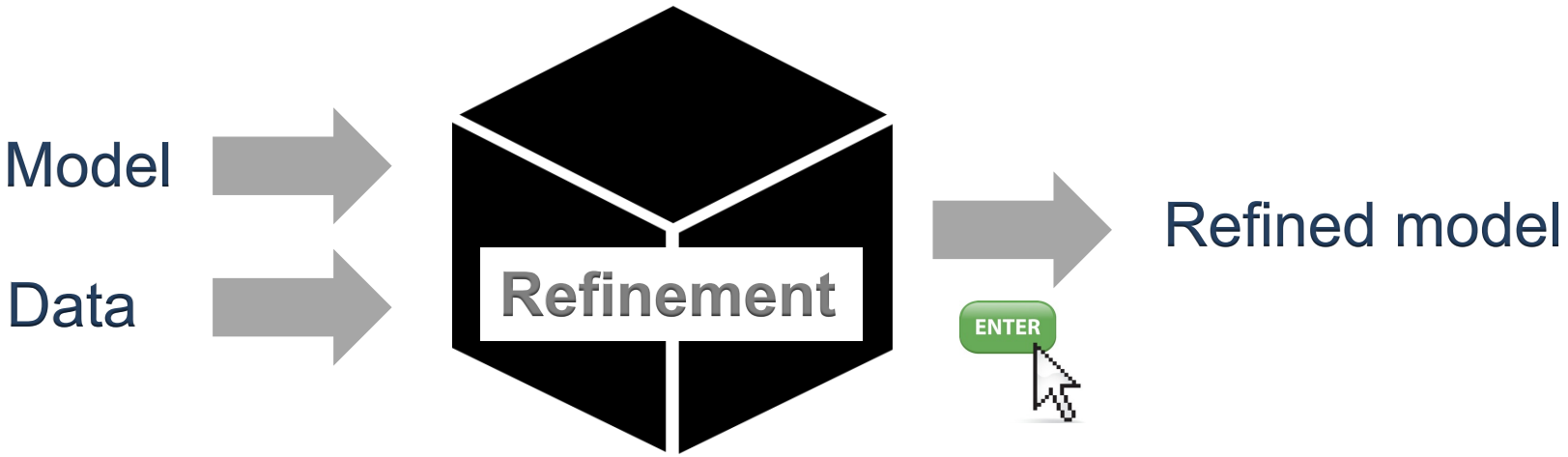


- Docking, flexible fitting, morphing are ***not*** refinement
- Refinement is to fine-tune an already ok atomic model
  - Refinement does only small changes to the model
  - *Convergence radius of refinement* ~ 1Å

Model refinement: black box



Model refinement: black box



- Does it always work?
- No.
  - Refinement parameterization isn't easy
  - Default settings suit most common scenario
  - Less typical situations need customizations

Model refinement: lot of stuff to know…

Reference model?

TLS?

Rotamer fixing?

AltLocs?

Group B vs individual?

Local minima?

ADP?

Clashes?

NCS?

IAS?

tNCS?

Grid search?

Weights?

CDL?

SA?

Minimization?

Rama plot restraints?

f’ & f”?

Hydrogens?

Restraints?

Bulk-Solvent?

Rigid body?

Rama-Z?

Anisotropy?

NQH flips?

SS restraints?

Twinning?

Model refinement: black box

- What to do when the ‘black box’ does not work?
  - Your decision-making is needed (and it is not always easy!)
- Validation helps to know you are on the right track

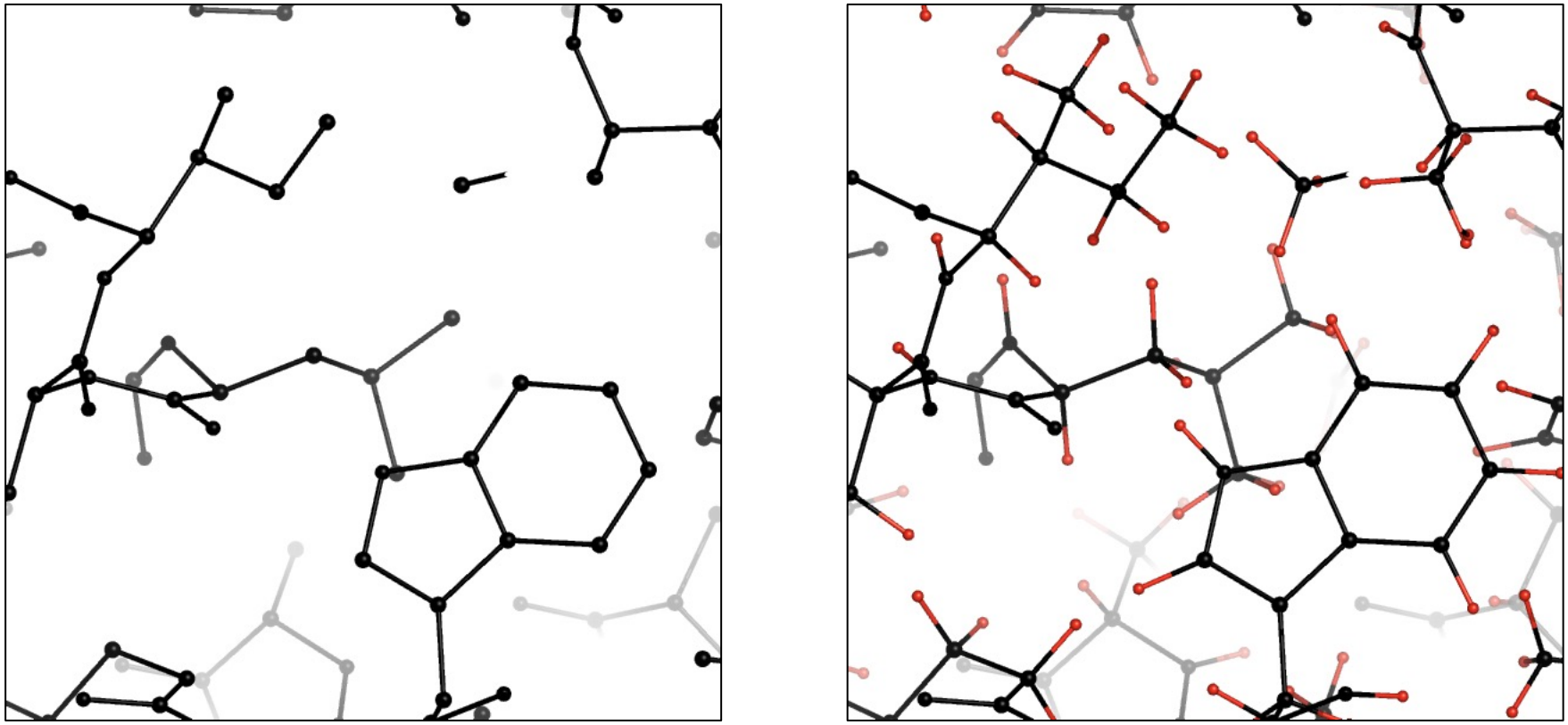
Model refinement: decision-making variables

- **Crystal**
  - Disorder
  - Twining, tNCS
  - Solvent content
  - Symmetry
- **Data**
  - Resolution
  - Errors
  - Completeness
  - Processing
- **Model**
  - Stage
  - Source
  - Parameterization
  - Fit to data

Model refinement: *random* topics

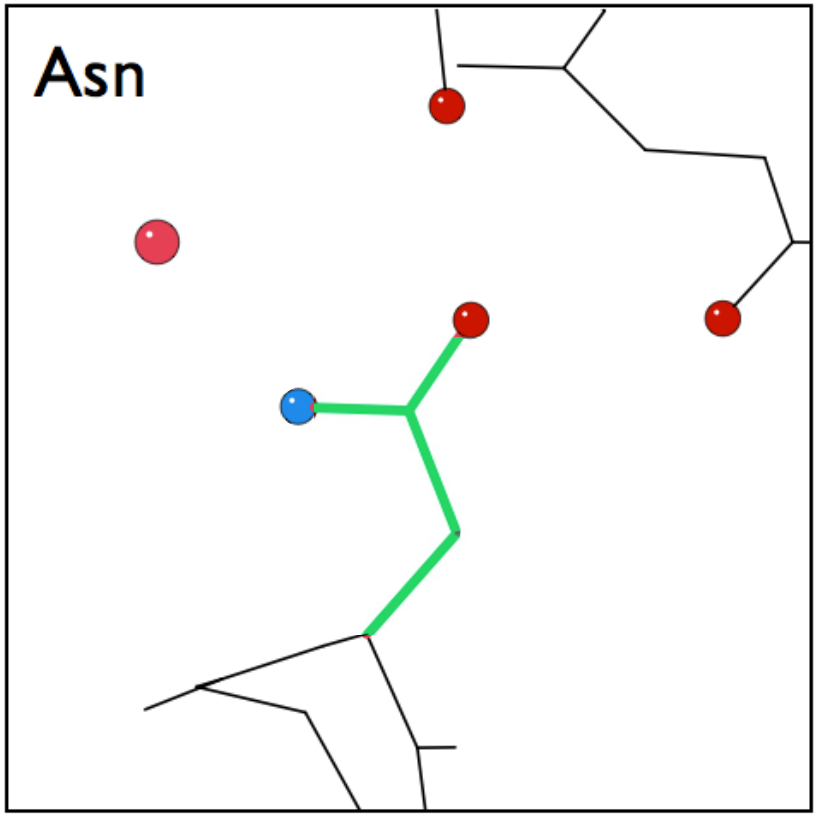
Hydrogens

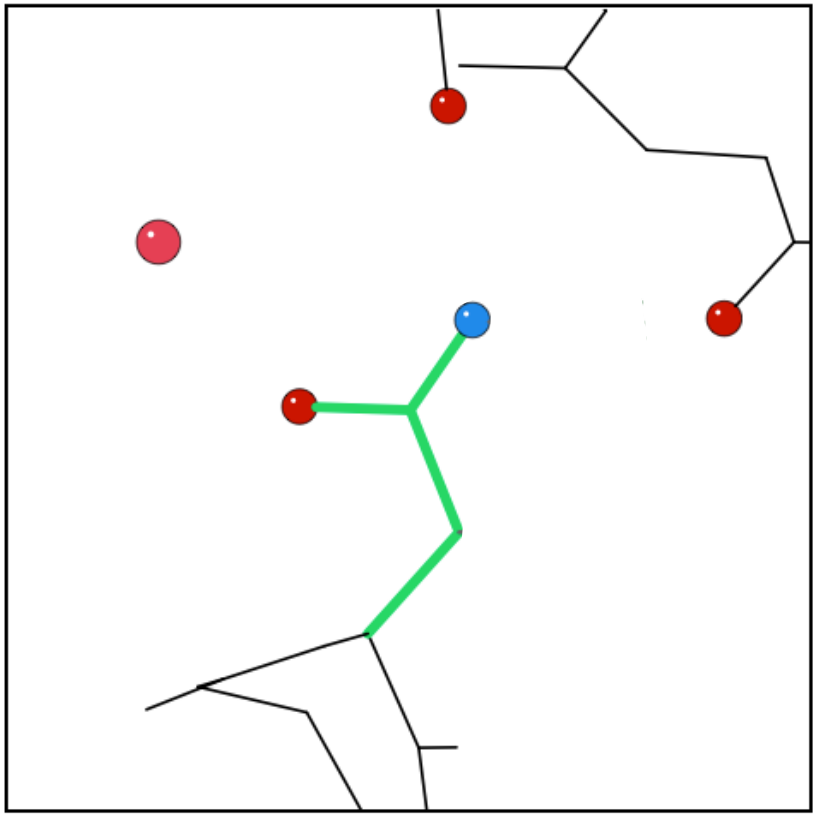
- Half of the atoms in a protein molecule
- Make most interatomic contacts
- Add to model towards the end, data resolution does not matter
- Once added, do not remove before the PDB deposition
- H do contribute to R-factors (expect 0.1-2% drop in R)



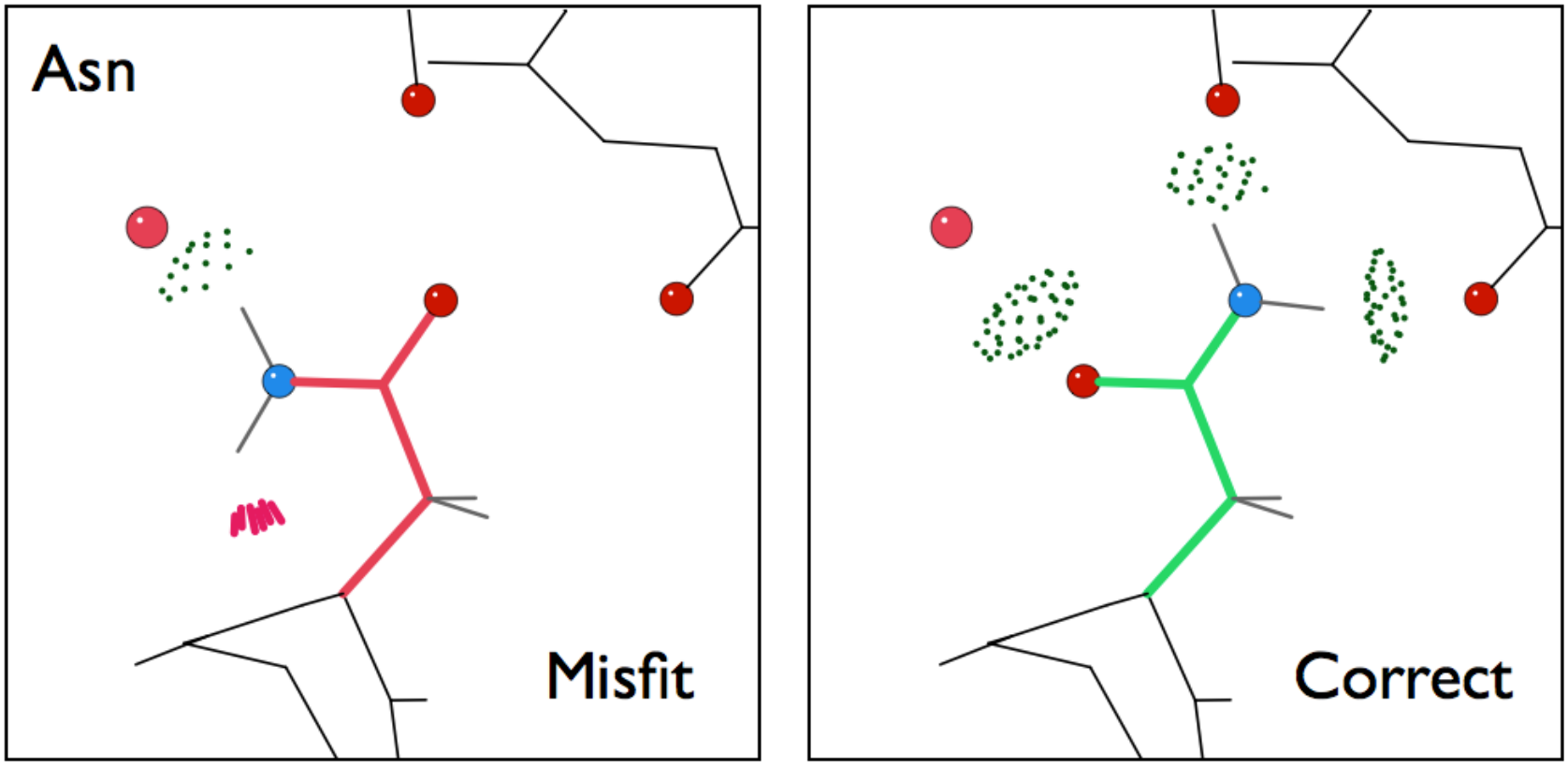
A structure without (left) and with (right) hydrogen atoms

- N/Q/H flips (asparagine/glutamine/histidine)
  - Based on clash analysis
  - Requires H present





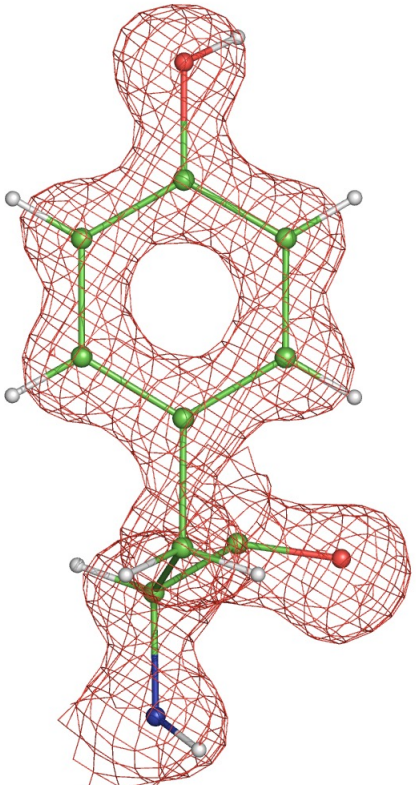
- N/Q/H flips
  - Based on clash analysis
  - Requires H present



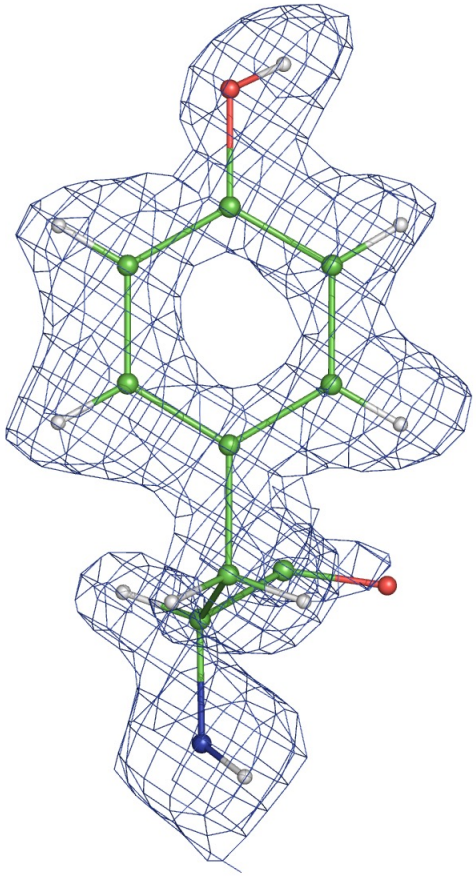
Hydrogens are best revealed by neutrons!

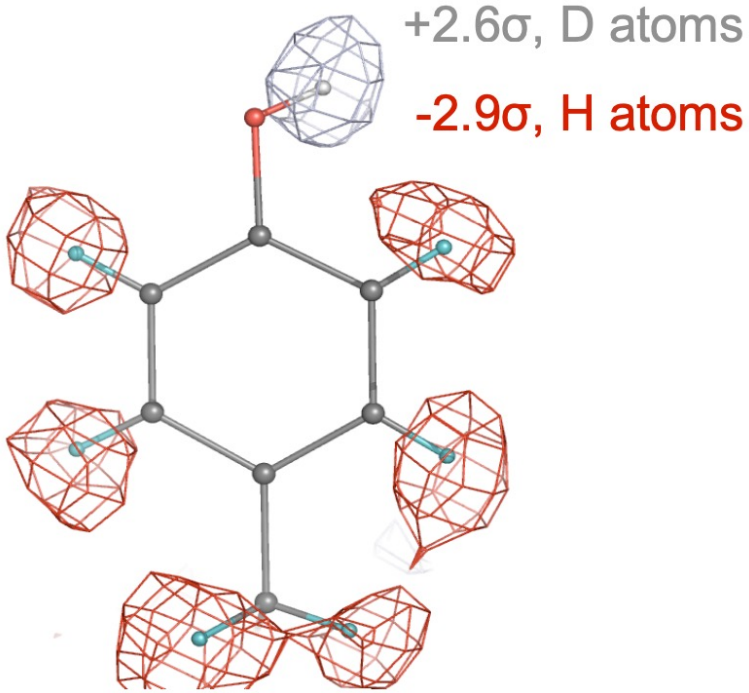
Nuclear density maps show H (D) at typical macromolecular resolutions (~2Å)

X-ray (1.1 Å)



Neutron (1.7 Å)

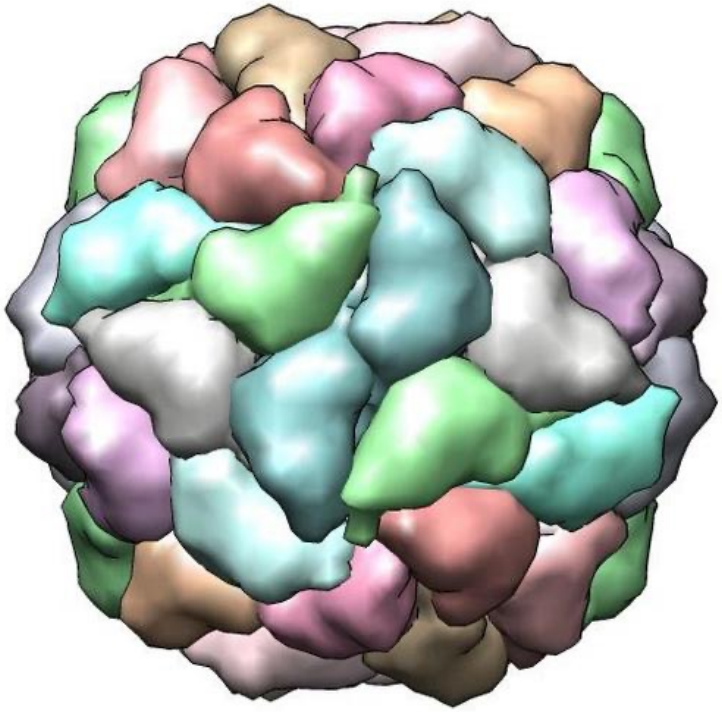


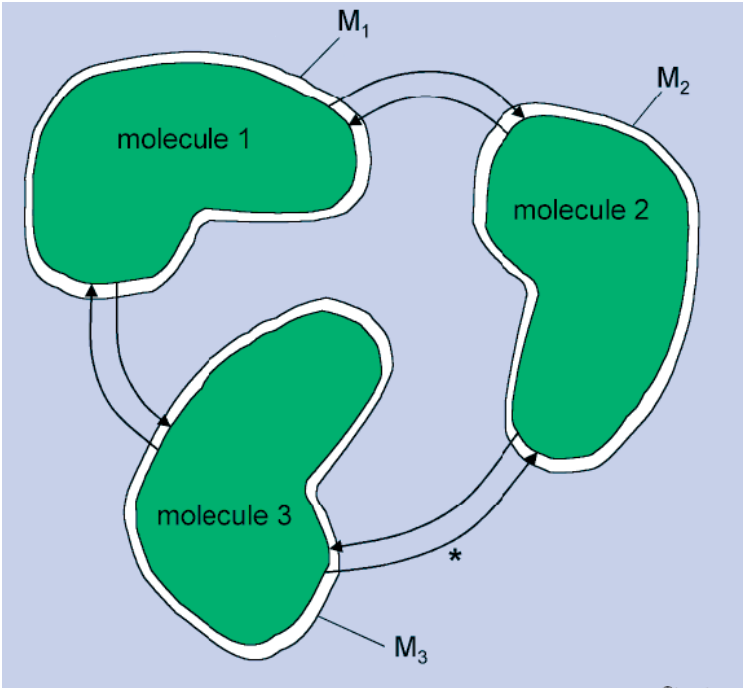


2mFo-DFc maps at 1.5σ (Rubredoxin, PDB code: 3KKY)

Constraints vs Restraints

NCS: constraints vs restraints

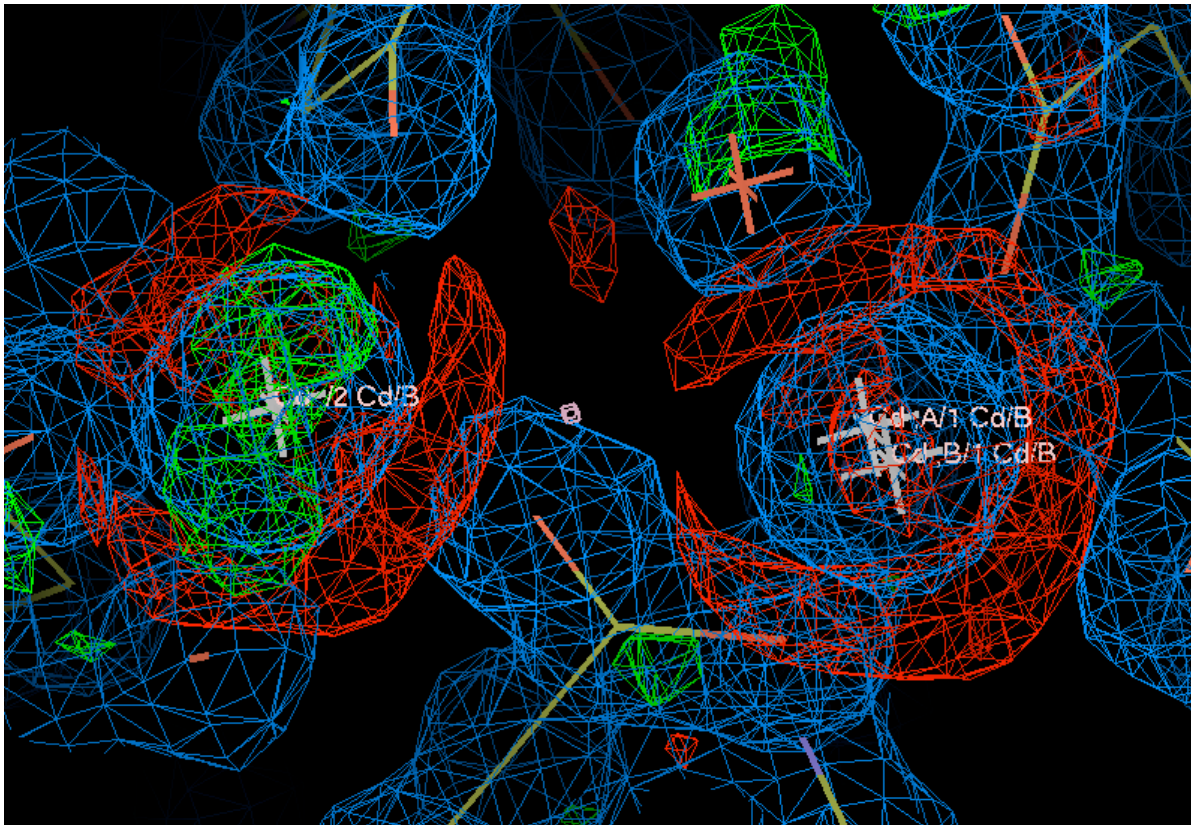




Source: Internet

- **Constraints**: molecules 1, 2 and 3 are required to be identical
- **Restraints**: molecules 1, 2 and 3 are required to be similar but not necessarily identical

Heavy atoms and map artifacts



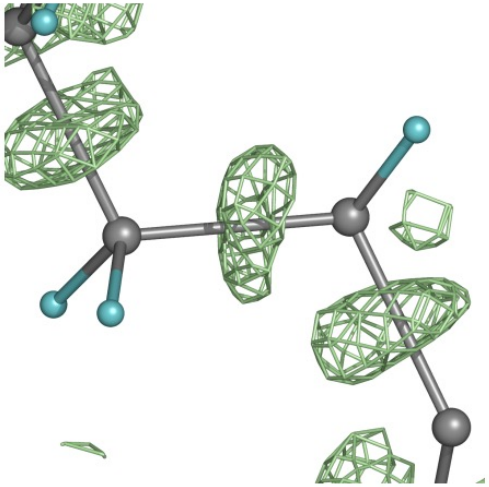
Reasons for +ve/-ve density:

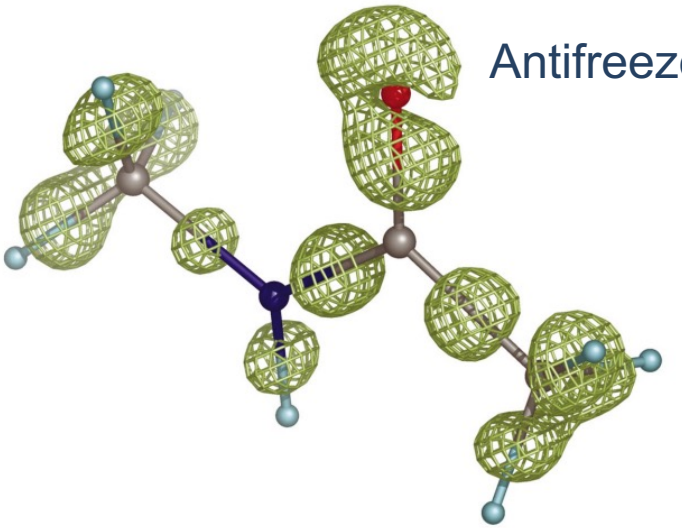
- Suboptimal xyz, occupancy, ADP (isotropic vs anisotropic), anomalous f’ & f’’, charge
- Refinement has not reached convergence
- Wrong atom

Map oddities at super-high resolution

Ultra-high resolution: 1Å or better

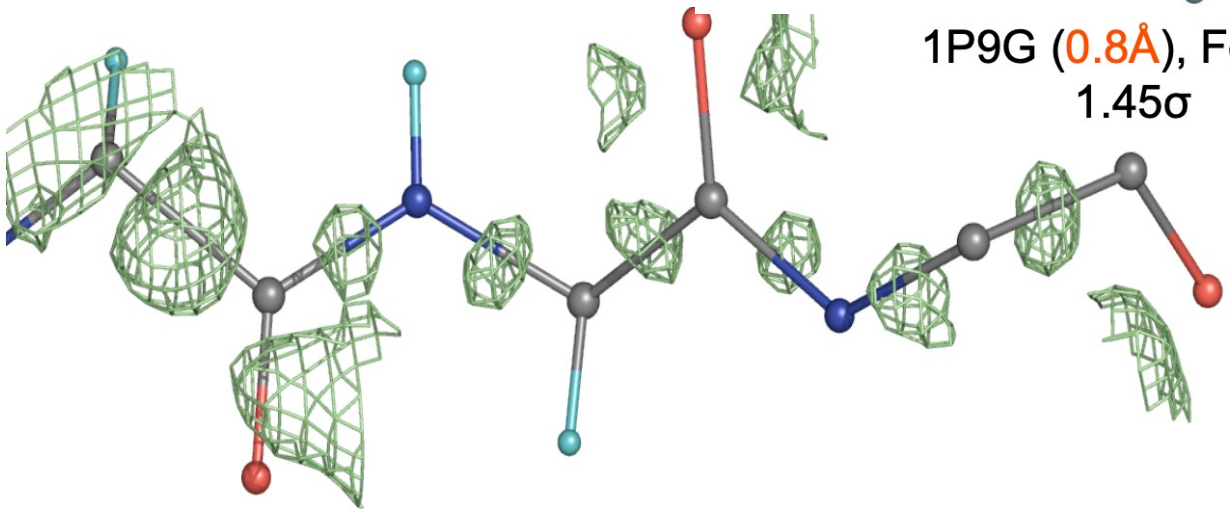
1UCS (0.62Å), Fo-Fc, 1.7σ





Antifreeze protein

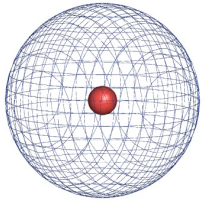
1P9G (0.8Å), Fo-Fc, 1.45σ



Multipolar refinement needs to be done. Will be available in Phenix later this year

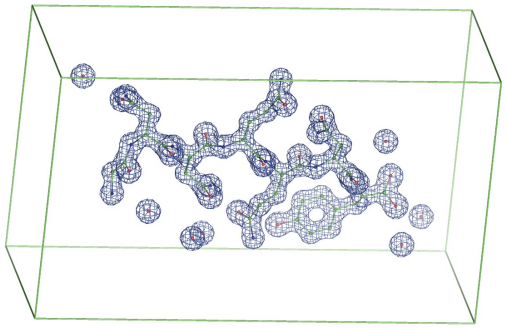
Atomic model

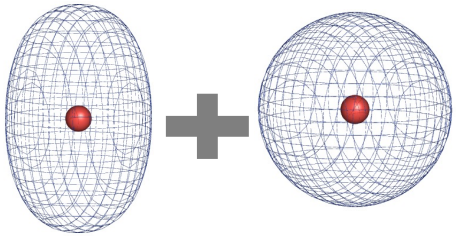
- Gaussian IAM (Independent Atom Model) – a physical model of ordered crystal structure



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

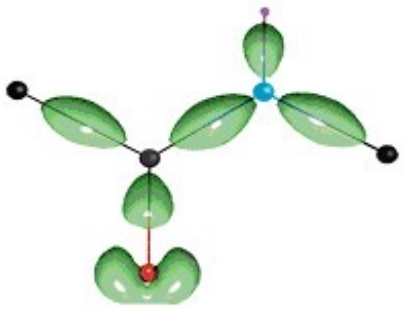
```
ATOM     25  CA  PRO A   4      31.309  29.489  26.044  1.00 57.79           C
```



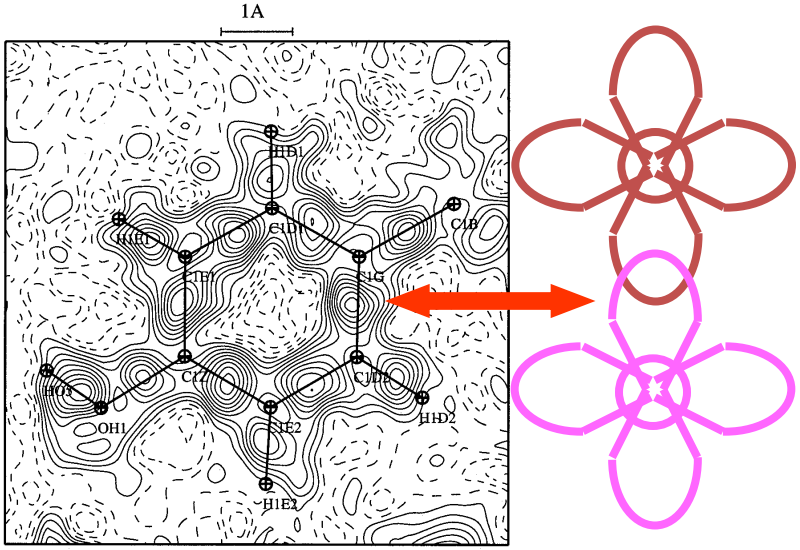


$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

More accurate approximation assumes atoms are bonded: multipolar model



$$\rho_{\text{atom}}(\mathbf{r}) = \rho_{core}(\mathbf{r}) + P_{val}\kappa^3\rho_{val}(\kappa\mathbf{r}) + \sum_{l=0}^{l_{\max}} \kappa'^3 R_l(\kappa'\mathbf{r}) \cdot \sum_{m=-l}^{l} P_{lm} y_{lm}(\theta,\varphi)$$



ρ_{ATOM} = core electrons

+ valence electrons

+ non-spherical part of the valence electron distribution

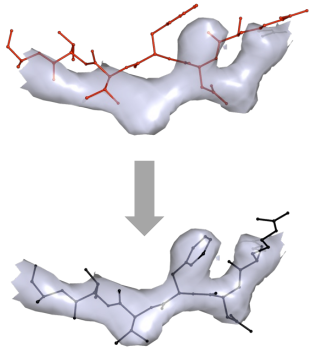
- Used at ultra-high resolution (better than 1Å)
- Coming soon in Phenix

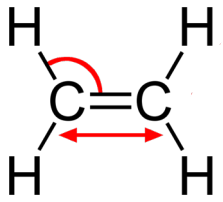
Refinement (target function) weight

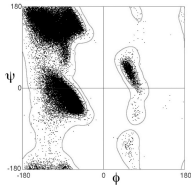
Refinement target (score)

T = T_{DATA} + $w * T_{RESTRAINTS}$

Optimize consensus between model-to-data fit and... common sense







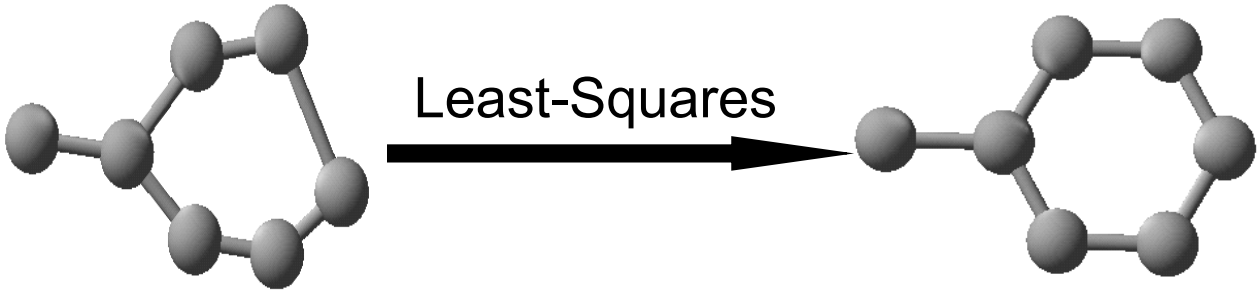
Bonds, angles, planes, torsions, chirality, non-bonded repulsion

Weight **w** doses helpful information such as restraints. The dose is very important!

Least-squares, Maximum-likelihood – what it's all about?

- **Complete model with errors**

Before refinement

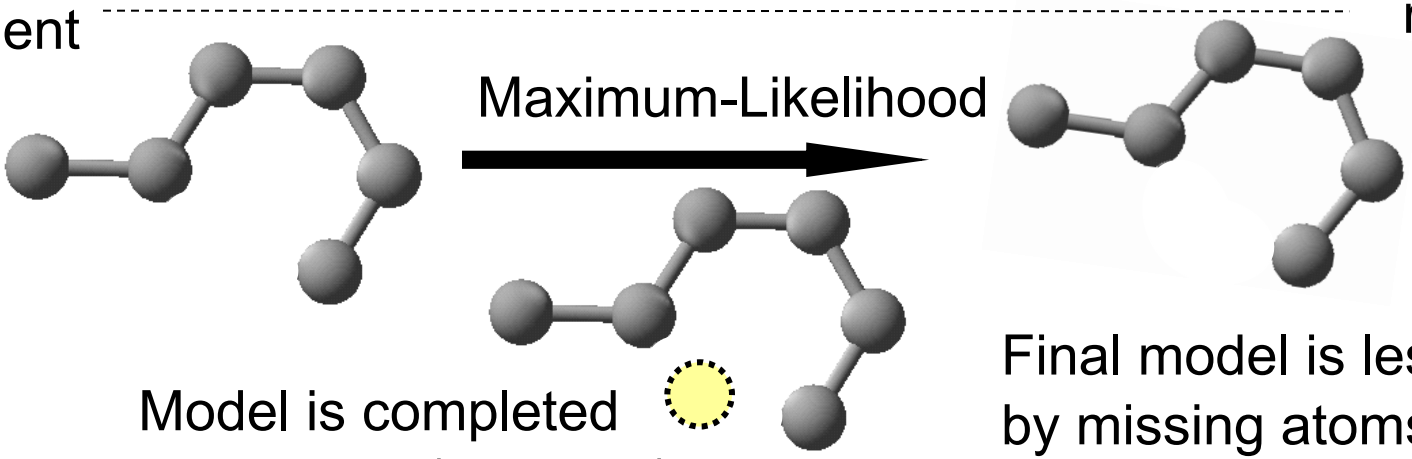


After refinement

- **Partial model, no errors**

Before refinement

After refinement



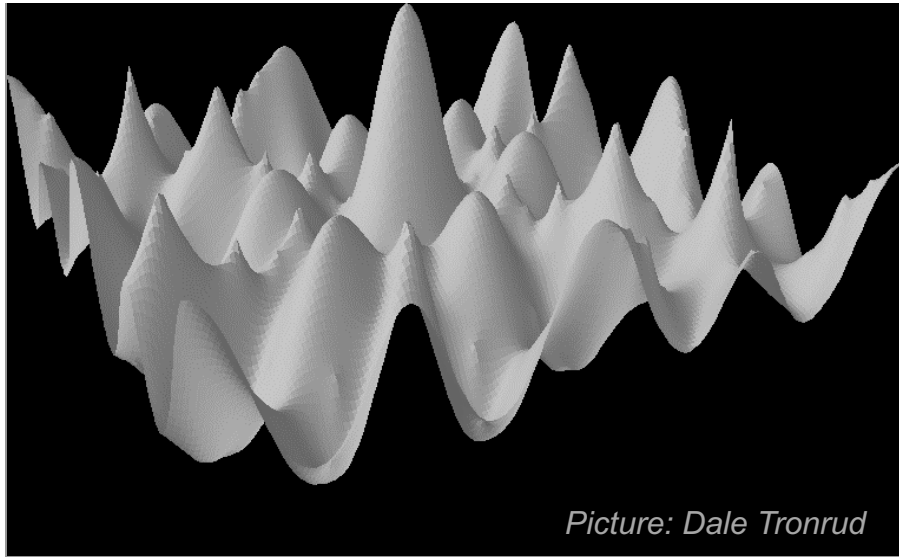
Model is completed statistically (implicitly)

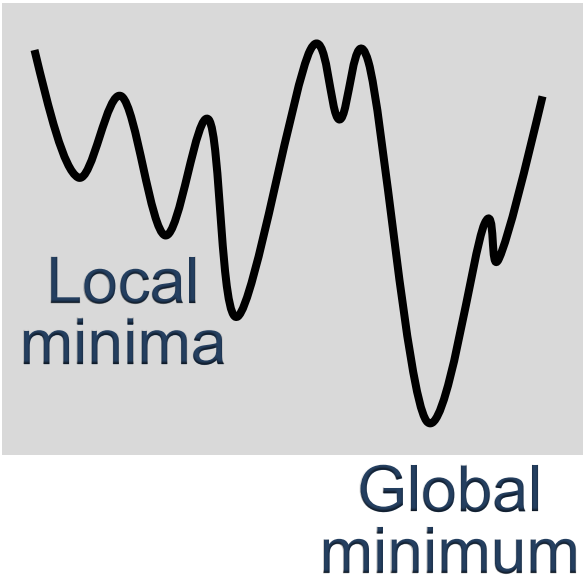
Final model is less affected by missing atoms

Refinement convergence and convergence radius

Complexity of refinement target

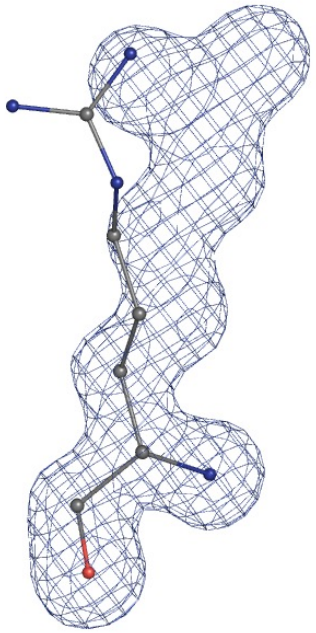
- Refinement target function (score) has very complex multi-dimensional profile





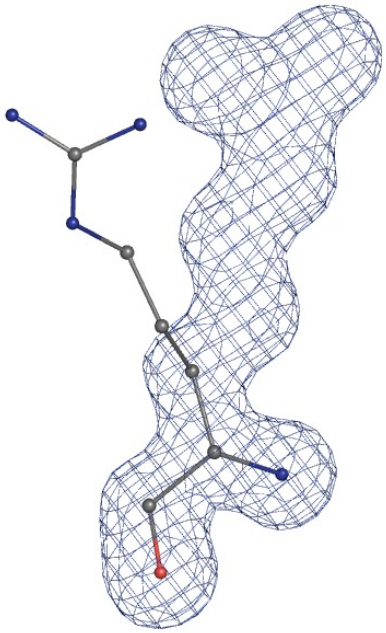
Refinement convergence

Minimization



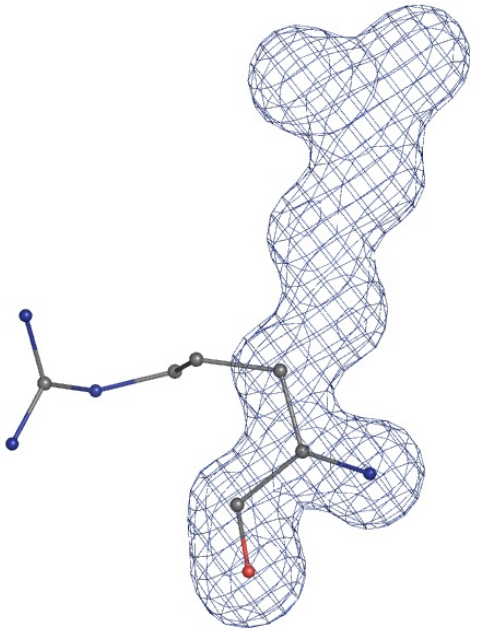
Minimization or SA can fix it

Simulated Annealing



Beyond convergence radius of minimization

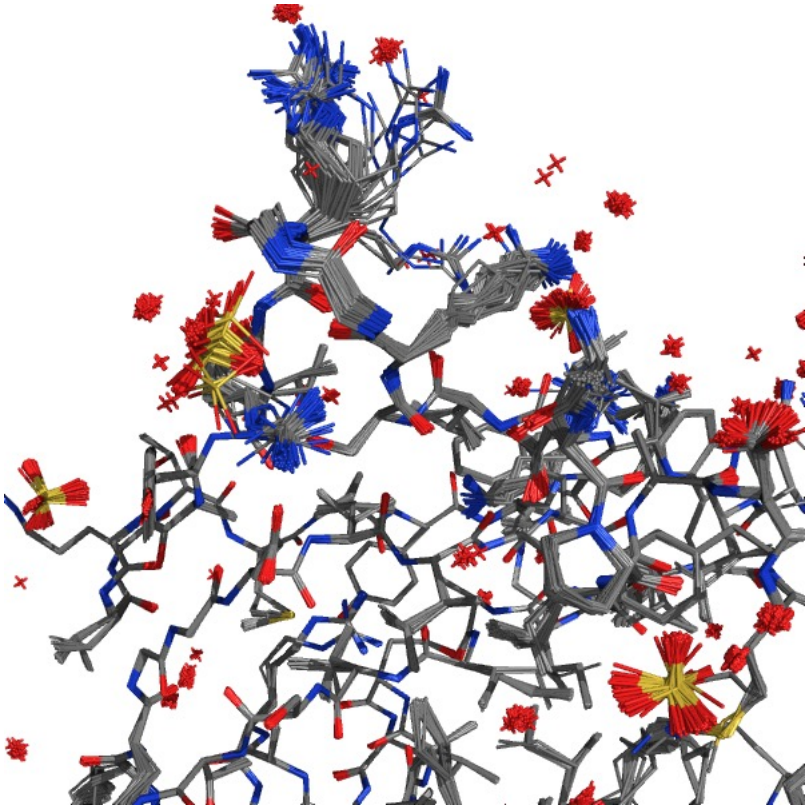
Real-space grid search

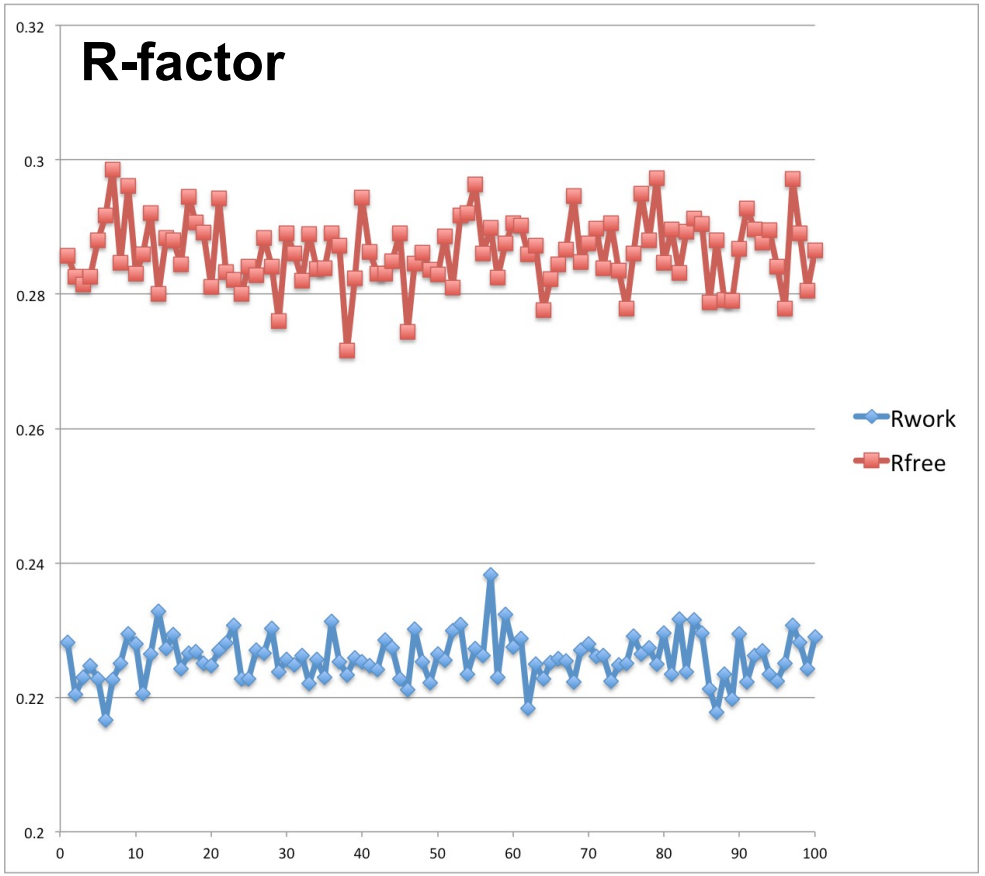


Beyond convergence radius of minimization and SA

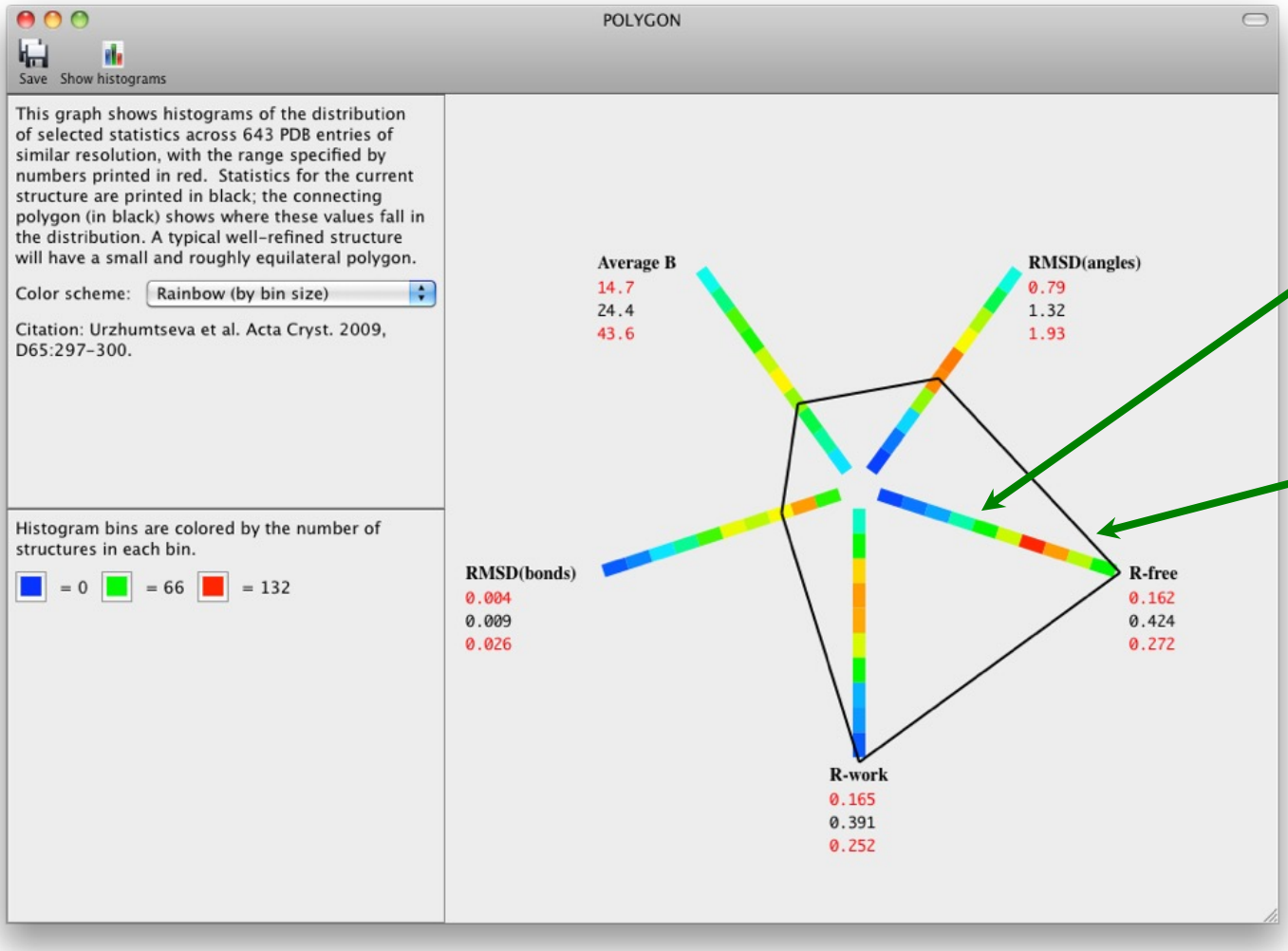
Uncertainty

100 identical refinement runs each one starting with slightly perturbed model





When you call it done?

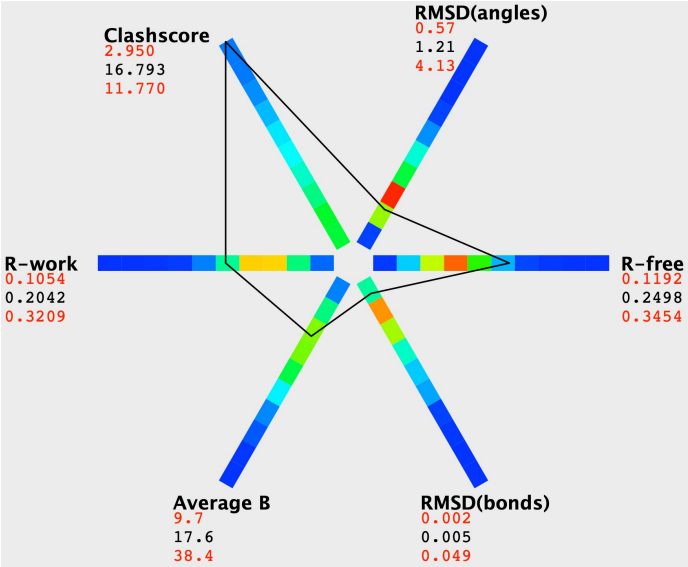


Colored bars are histograms showing distribution of values for structures at similar resolution

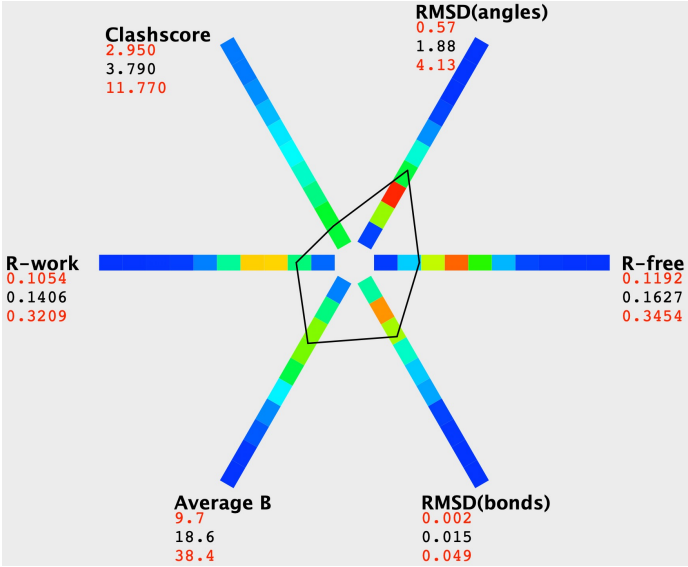
The black polygon shows where the statistics for the user's structure fall in each histogram

Crystallographic model quality at a glance.
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst*. **D**65, 297-300 (2009)

Clearly there are problems



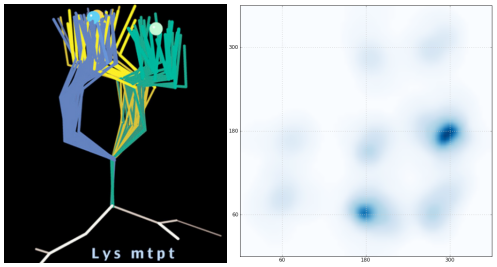
Likely overall good model



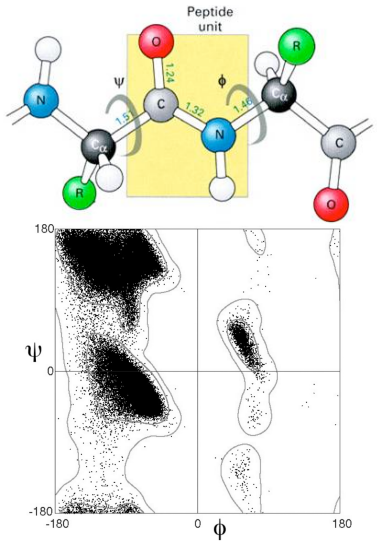
Low-resolution: things to consider

General idea: use all available information!

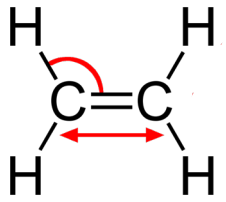
Side chain distributions



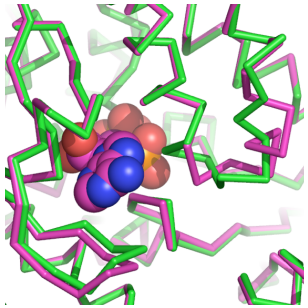
Main chain distributions



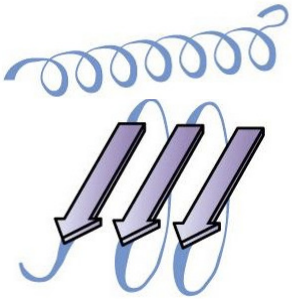
Covalent geometry



Internal symmetry (NCS)



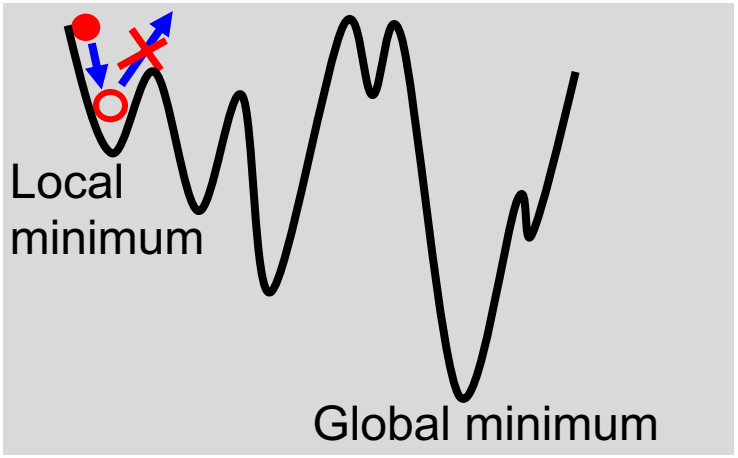
Similar (homologous) structures (reference model restraints)



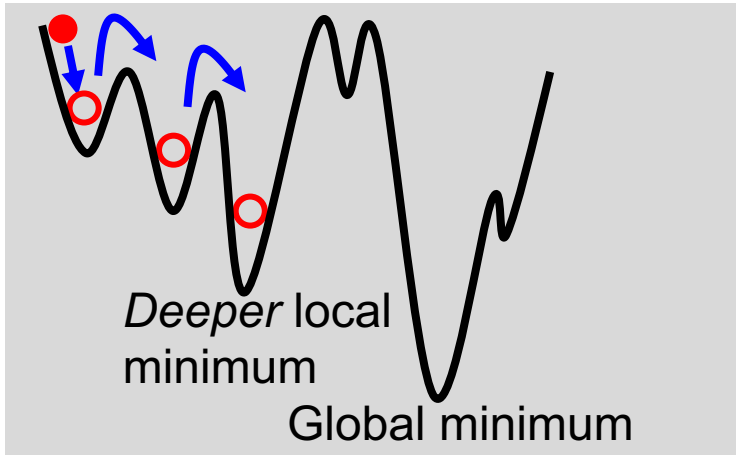
Secondary structure

Simulated Annealing (SA) ?

- **Gradient minimization**



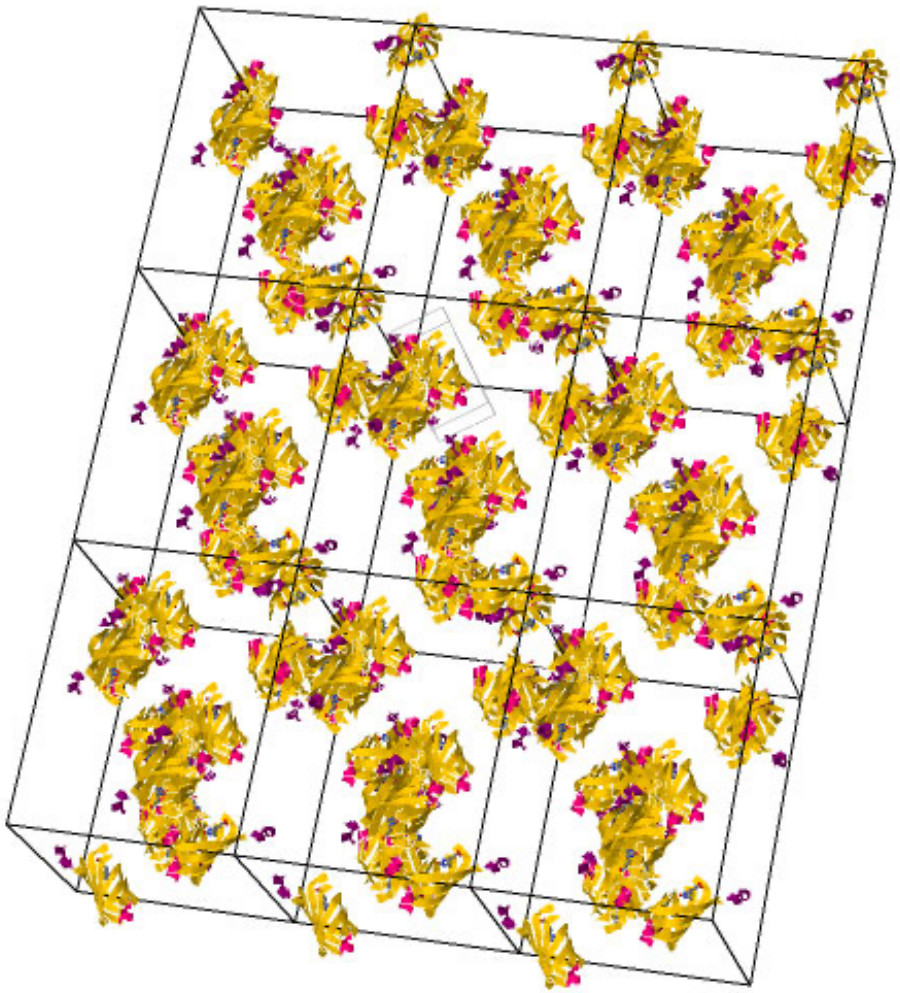
- **Simulated annealing (SA)**



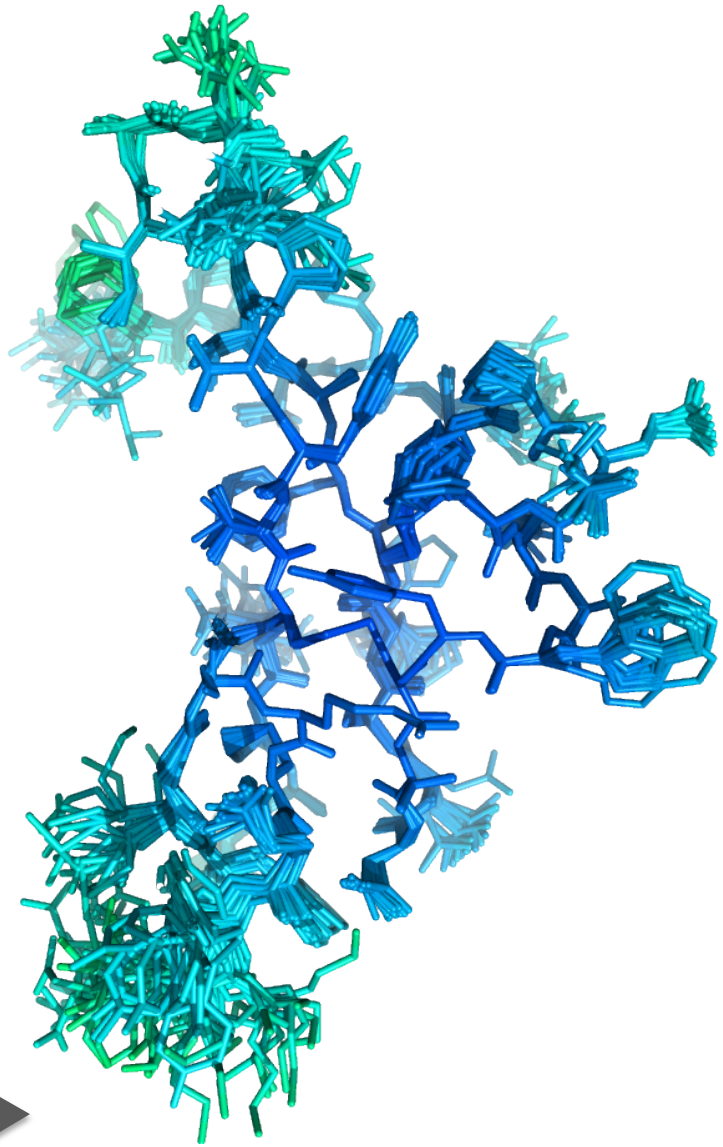
- Only use if model has gross errors (correction requires large movements)
- Do not use if model is relatively good and only needs small corrections

TLS

Disorder

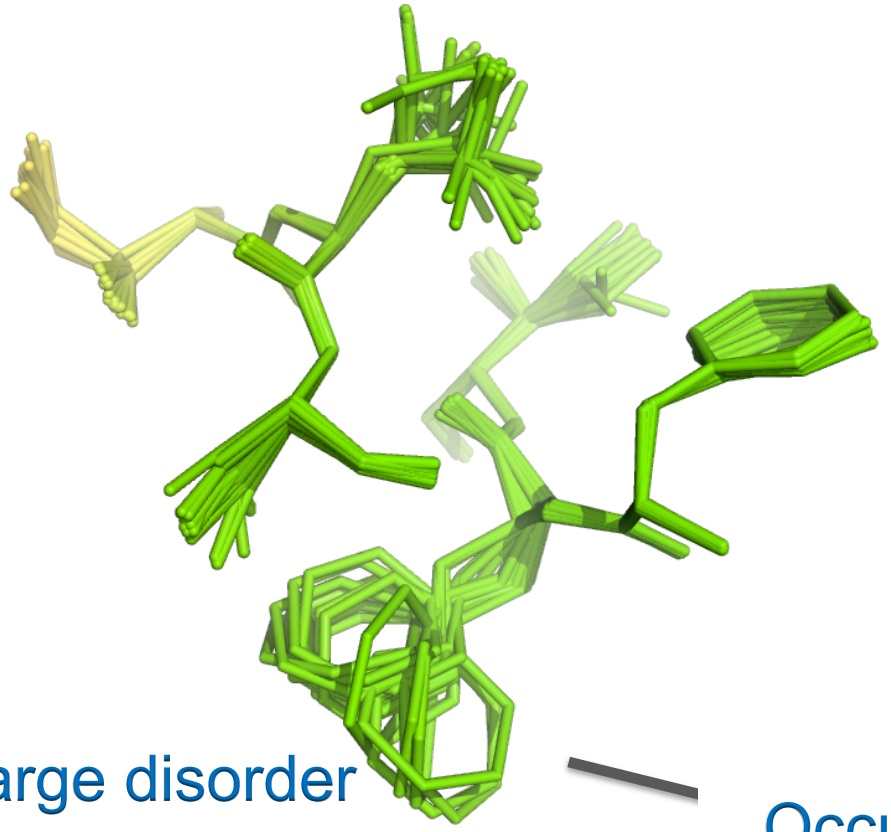


Crystal = many unit cells



Superpose all structures
from each unit cell

Disorder



Small disorder

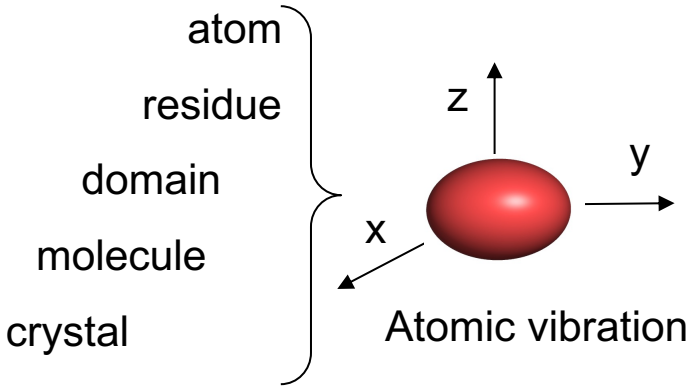
ADP (B-factor)

Large disorder

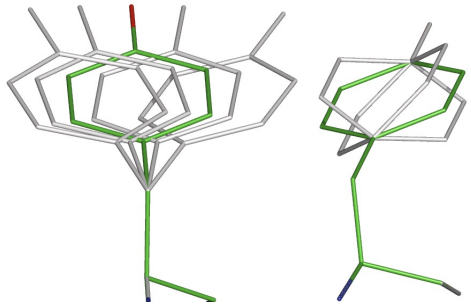
Occupancy

```
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
```

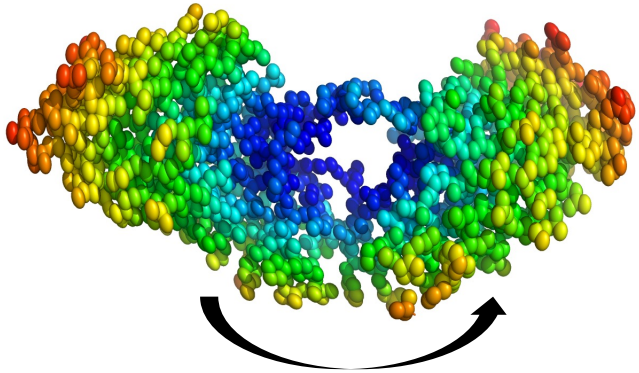
Atomic Displacement Parameters (ADP, B-factors)



Atomic vibration



Side chain libration



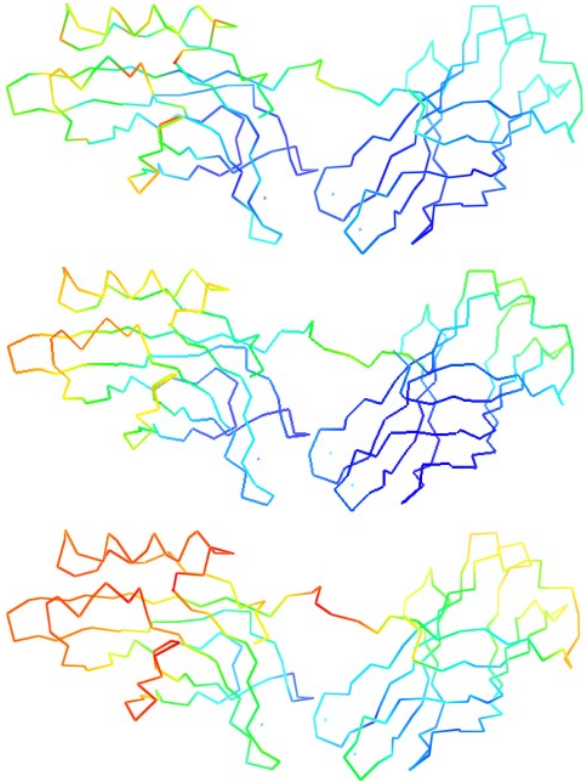
TLS: rigid-body domain vibrations

TLS: facts

- TLS assumes the atomic model consists of rigid domains that undergo anisotropic vibrations
- TLS uses 21 per rigid domain parameters to describe these vibrations
- TLS is not a way to reduce number of parameters
  - In fact, TLS adds more parameters!
- TLS offers more physically realistic model for atomic vibrations
- If correctly used may reduce R factors by up to 5%

TLS: example

Synaptotagmin refinement at 3.2 Å



Original refinement (PDB code: 1DQV)
R-free = ***34*** *%*
R = ***29*** *%*

PHENIX – Isotropic restrained ADP
R-free = ***28*** *%*
R = ***23*** *%*

PHENIX – TLS + Isotropic ADP
R-free = ***25*** *%*
R = ***20*** *%*

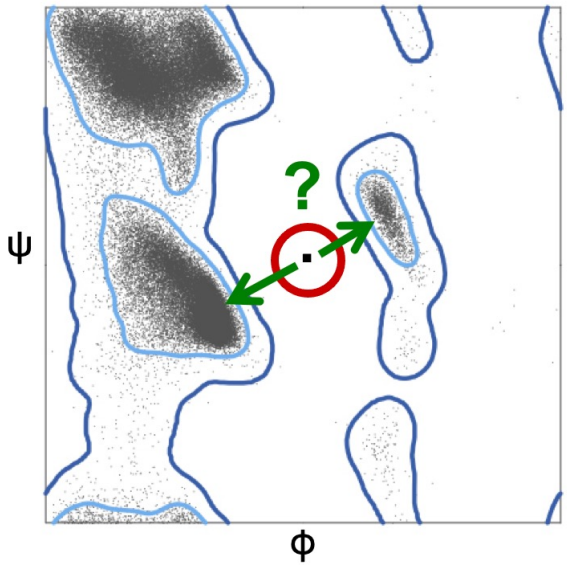
9% improvement in both Rwork and Rfree !

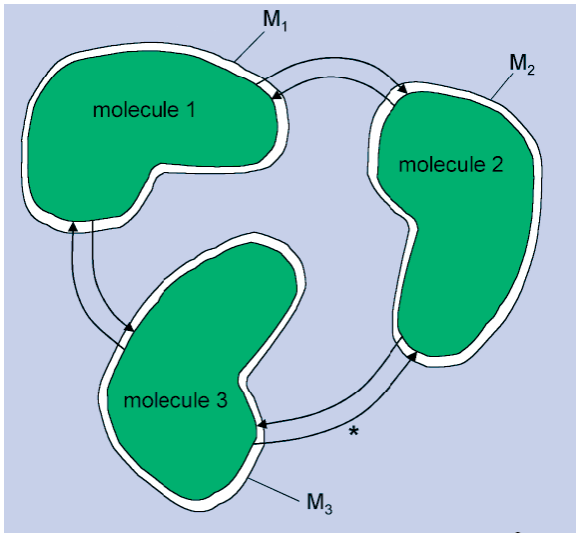
Refinement and validation conflict

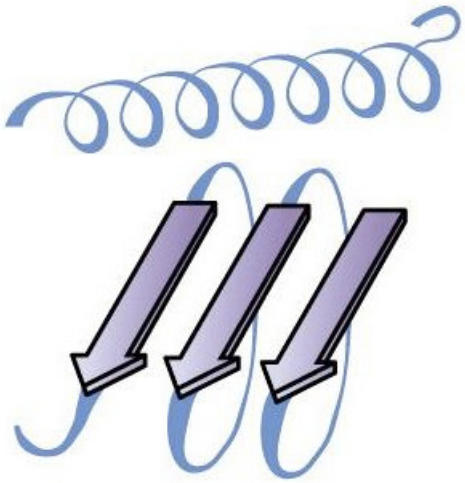
Refinement and validation conflict

- In low-resolution refinement we use extra restraints to compensate for lack of data:
    - Ramachandran plot restraints
    - Cβ deviation restraints
    - Secondary structure restraints
    - Restraints on χ angles of amino-acid side-chain rotamers
- These are standard validation tools… using them as restraints compromises their validation power
- Setting up extra restraints: manual work & very error-prone

Setting up extra restraints: manual work & very error-prone

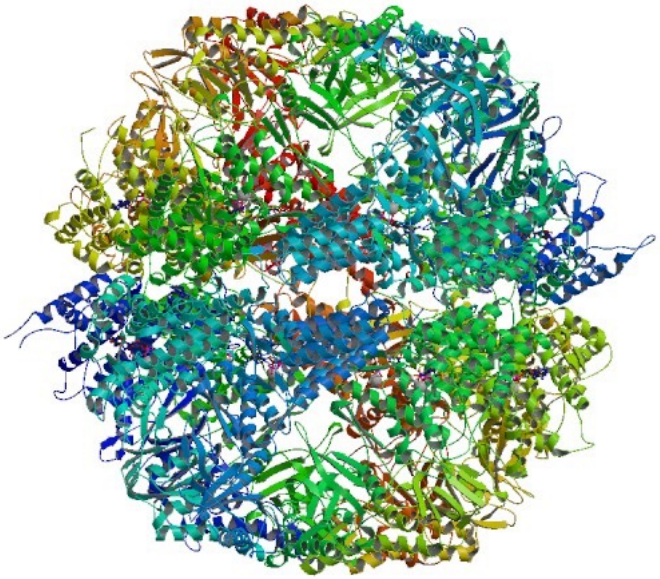






Model validation

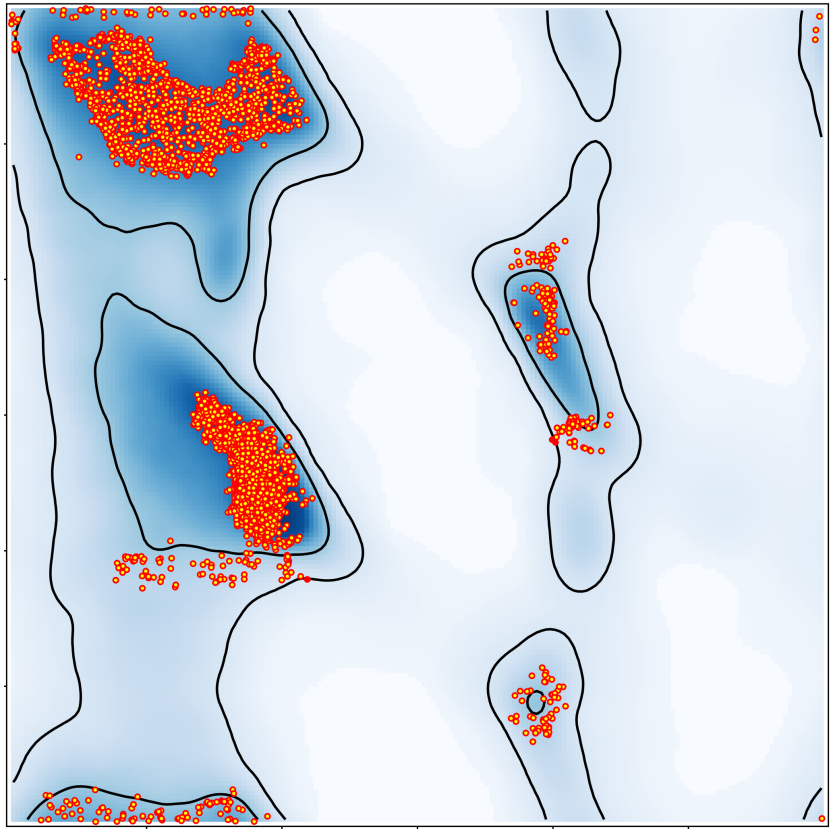
PNAS, 2019 116 (39) 19513-19522



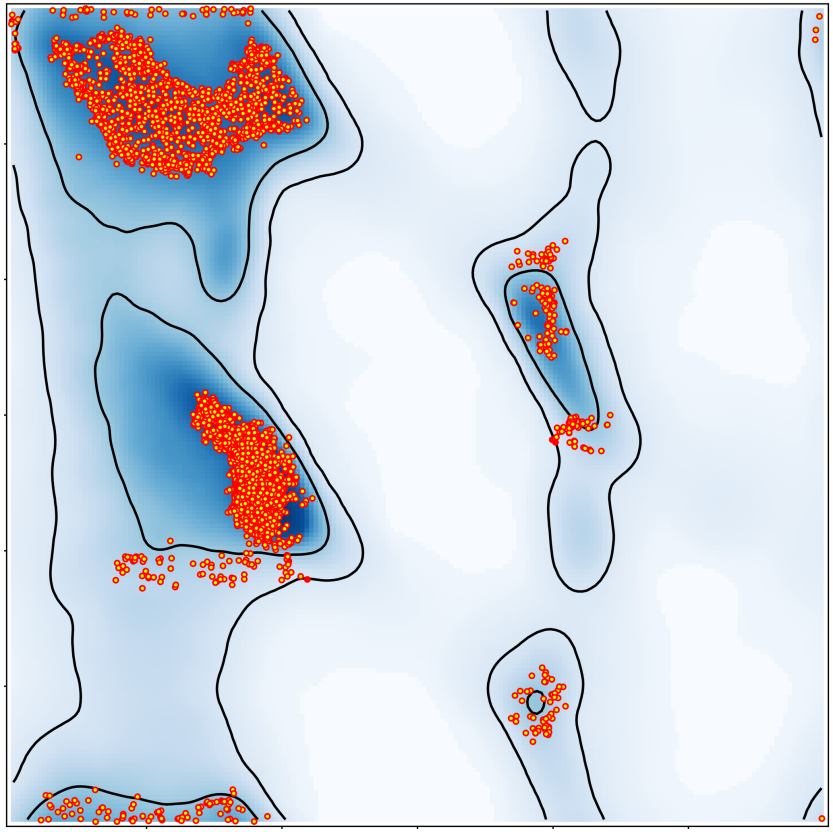
Metric / PDB code		6KS6
Clashscore		7.7
Rama. (%)	favored	96.4
	outliers	0.2
Rotamer outliers (%)		0
C_β deviations		0
RMSD	Bond (Å)	0.001
	Angle (°)	0.396
Resolution (Å)		3.0

Perfect statistics! All looks just great!

Model validation: *Ramachandran plot*



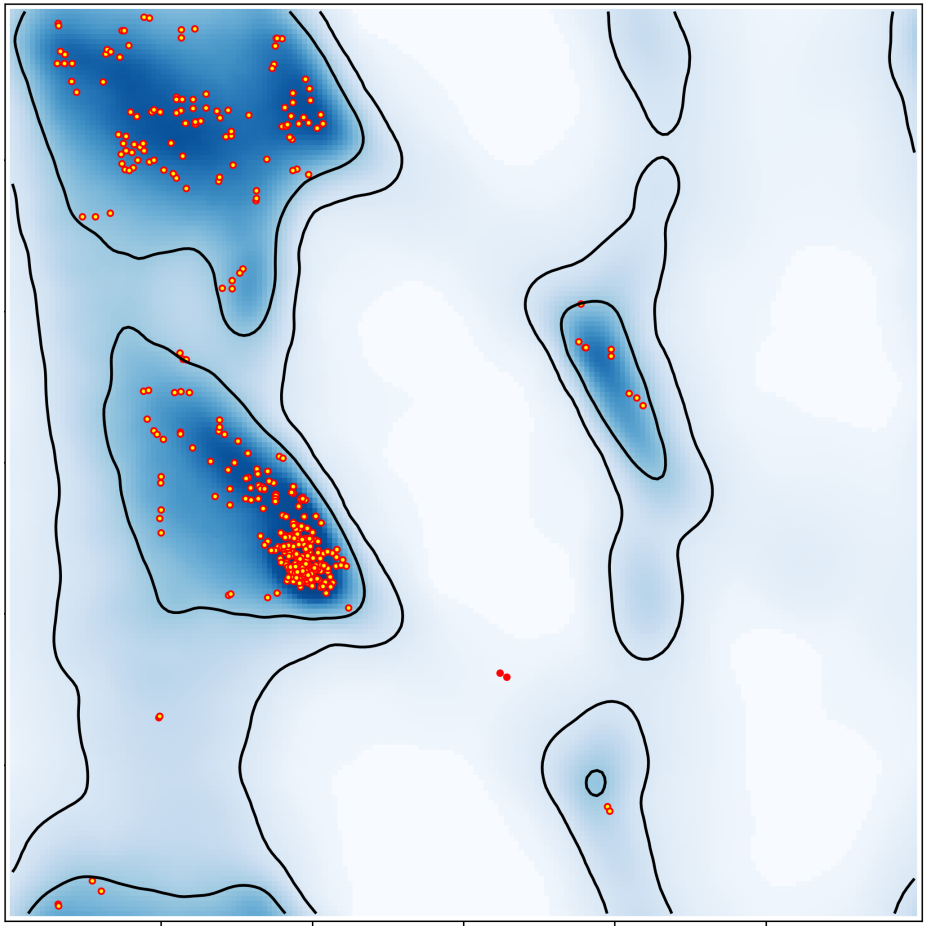
Model validation: *Ramachandran plot*



- Two questions:
  - **How we know the plot is poor?**
  - How did this happen?

Model validation: *Ramachandran plot*

We know how good plot looks like!



Ramachandran plot Z-score

CABIOS

Vol. 13 no. 4 1997
Pages 425–430

Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

- Good at spotting odd plots
- One number, simple criteria:
  - Poor: $|Z| > 3$ Suspicious: $2 < |Z| < 3$ Good: $|Z| < 2$

Structure

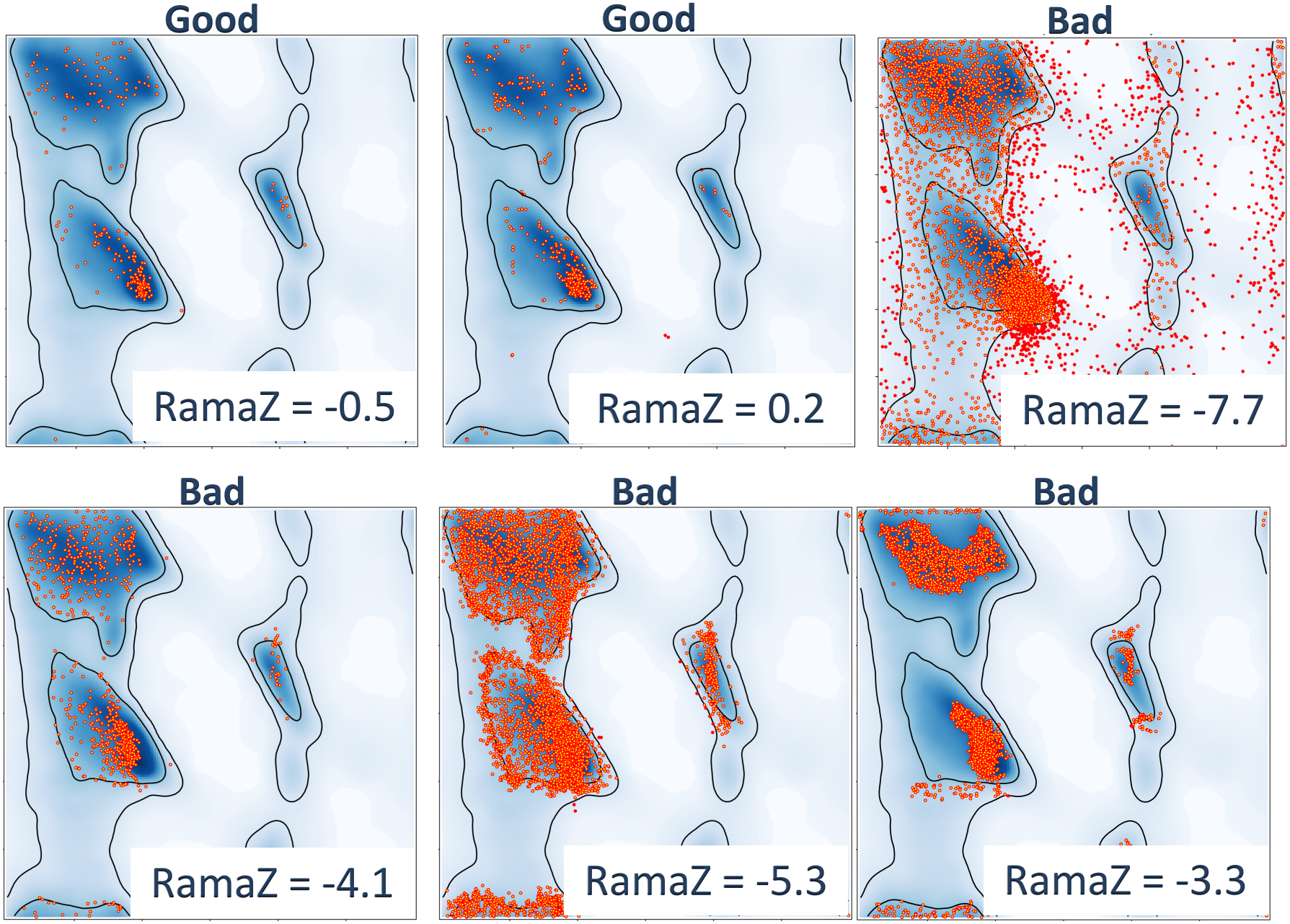
CellPress

Resource

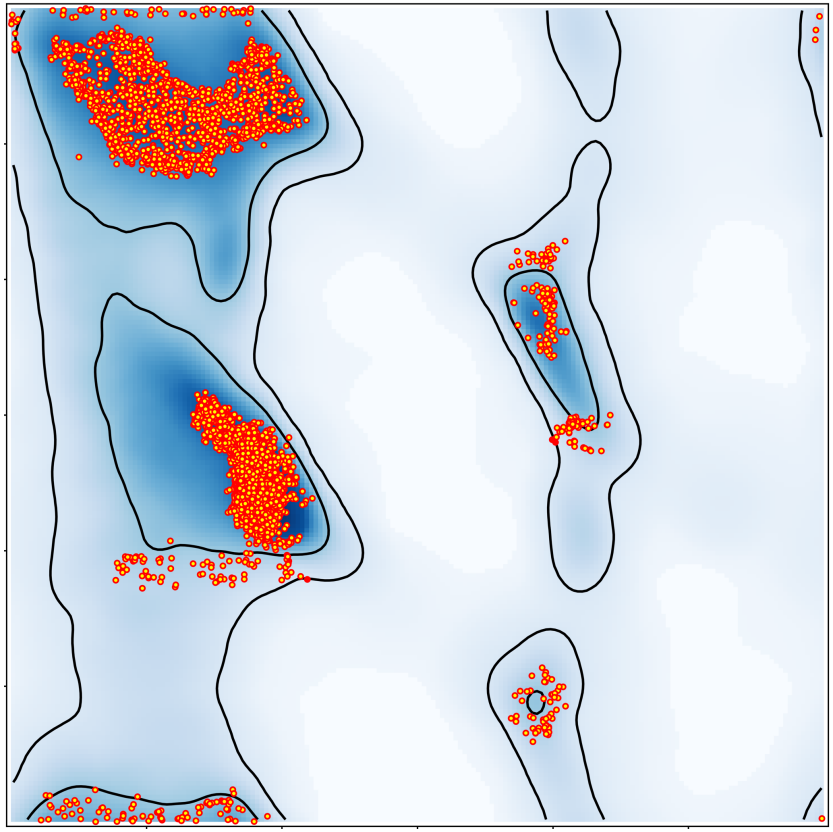
A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

Oleg V. Sobolev,[1,5,*] Pavel V. Afonine,[1] Nigel W. Moriarty,[1] Maarten L. Hekkelman,[2,3] Robbie P. Joosten,[2,3,*] Anastassis Perrakis,[2,3] and Paul D. Adams[1,4]

Model validation: *Ramachandran plot Z-score*



Model validation: *Ramachandran plot*



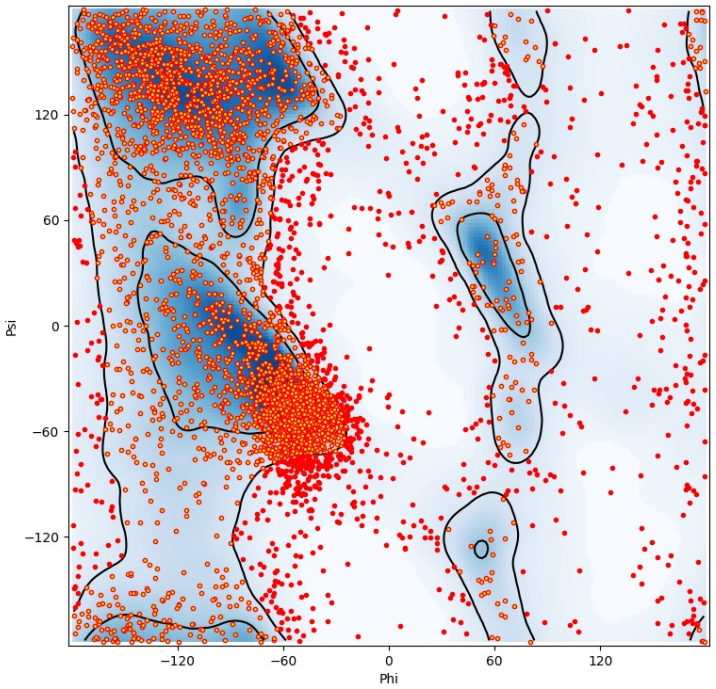
- Two questions:
  - How we know the plot is poor?
  - **How did this happen?**

Ramachandran plot restraints

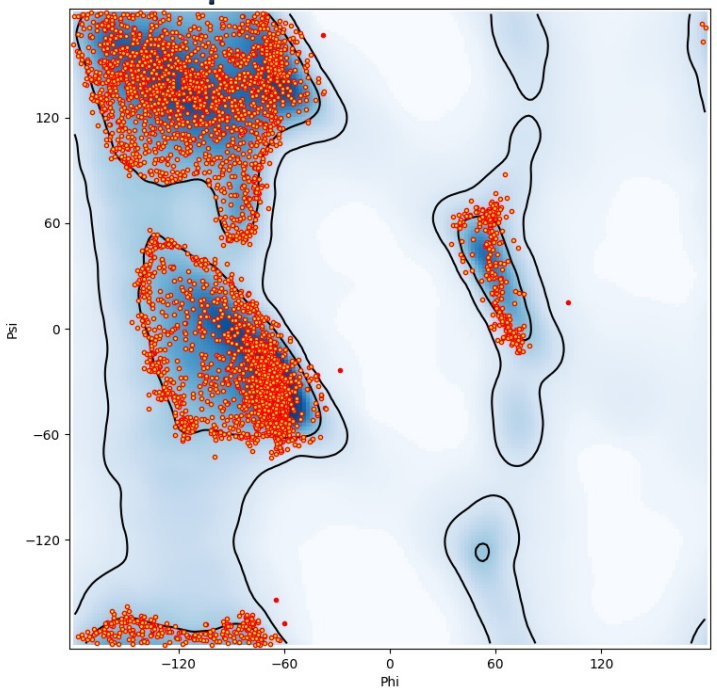
- Always use at low resolution
- **Do not use to fix existing outliers**

PDB code: 5a9z

Original



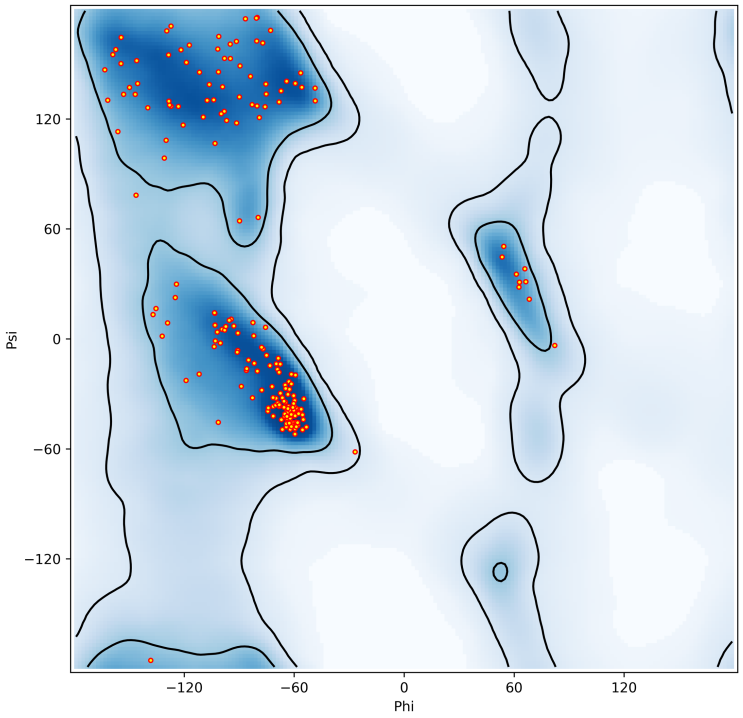
Refined with Ramachandran plot restraints



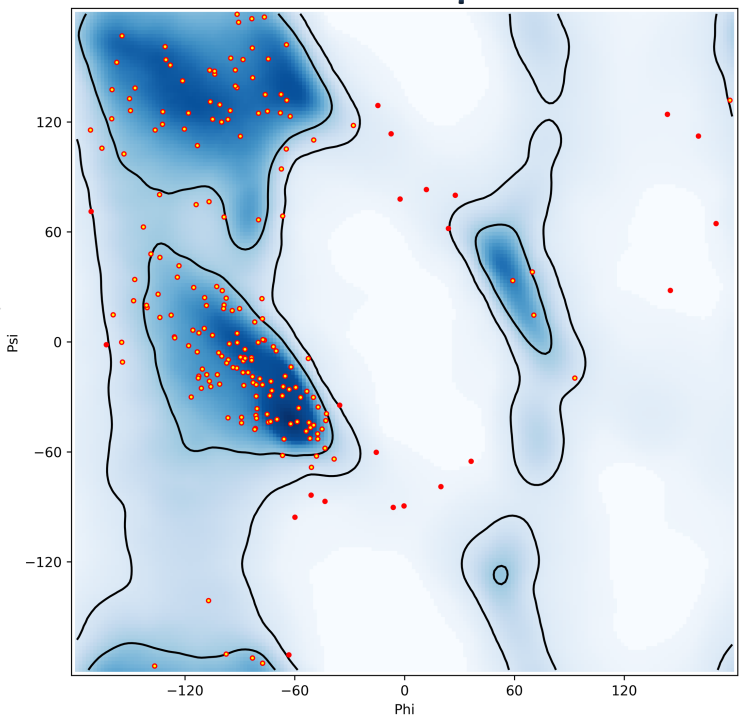
Ramachandran plot restraints

- Ramachandran plot restraints
  - **Use to stop outliers from occurring**

Before refinement



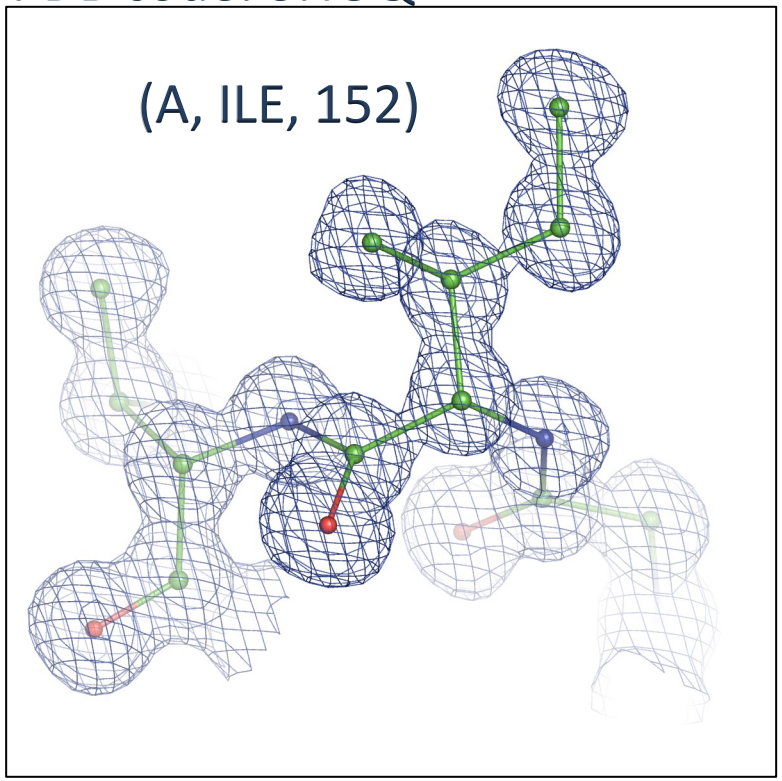
After refinement
(No Ramachandran plot restraints)

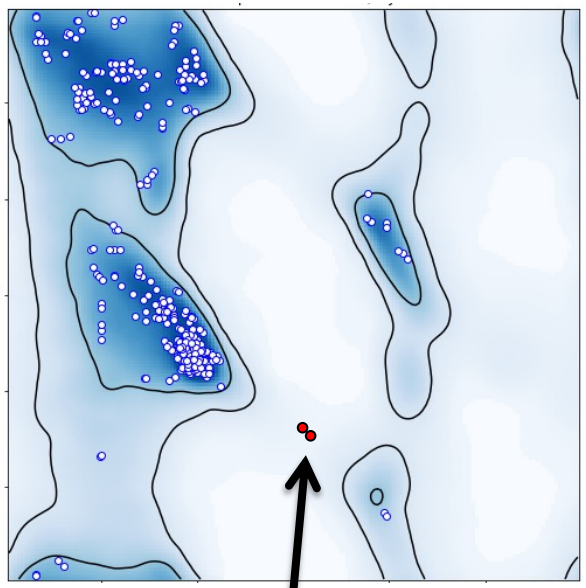


Validation: *outliers are not always wrong*

- A Ramachandran plot outlier ≠ wrong

PDB code: 3NOQ





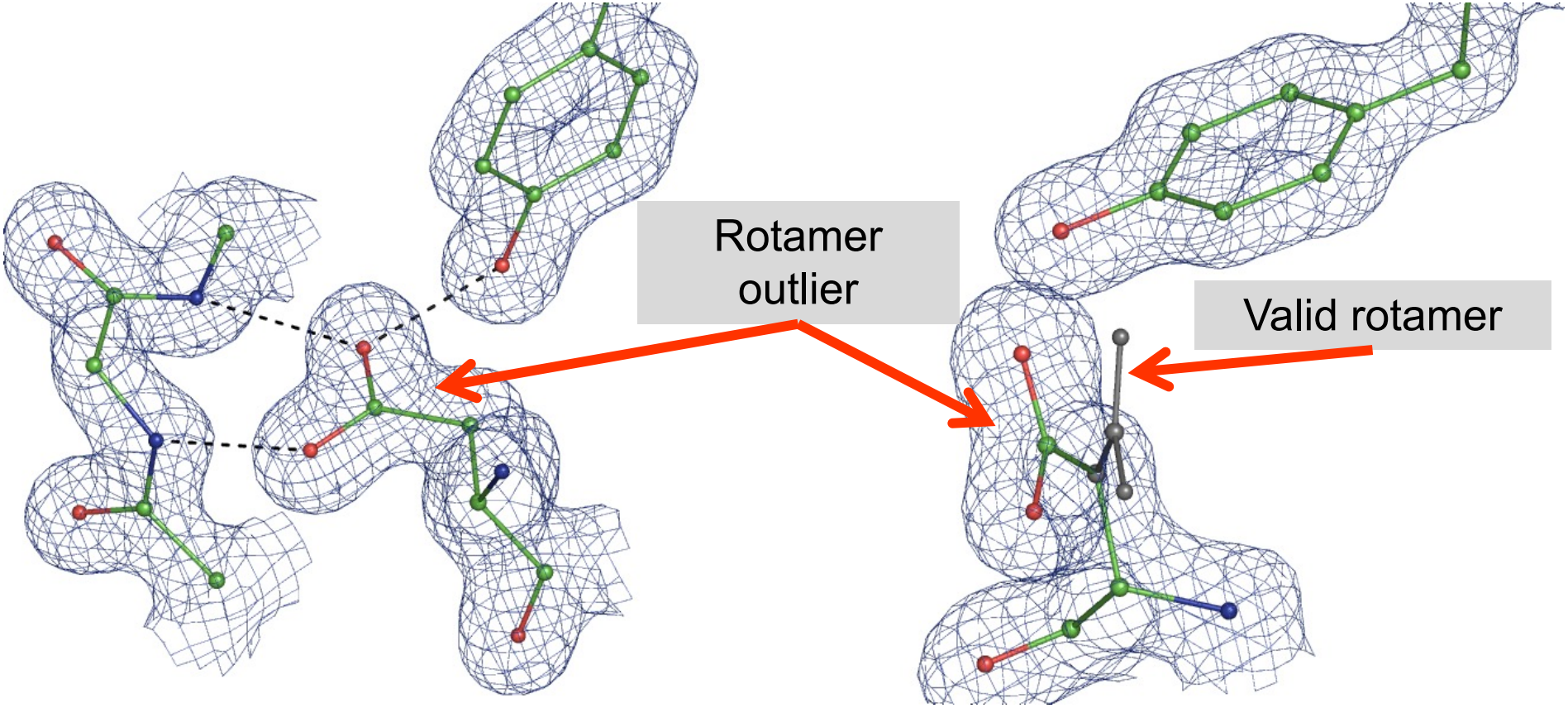
Outliers:

(A, ILE, 152), (B, ILE, 154)

- All outliers need to be explained (supported by the data)

Validation: *outliers are not always wrong*

- An outlier ≠ wrong
  - However, each outlier has to be explained



Refinement and validation conflict

Aren't we confused now?

Restraints and limitations

$$T = T_{DATA} + w * T_{RESTRAINTS}$$

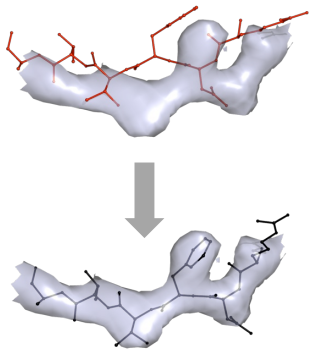
$$T_{RESTRAINTS} = T_{BOND} + T_{ANGLE} + T_{DIHEDRAL} + T_{PLANE} + T_{REPULSION} + T_{CHIRALITY}$$

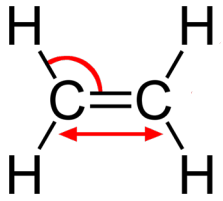
- Restraints are too limited:
  - No attraction terms (electrostatics, etc)
  - Not using information about protein structure (secondary structure, rotamers)
  - Limited to tabulated entities in the libraries (e.g., Monomer Library, GeoStd)

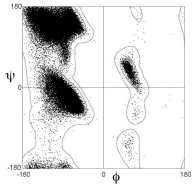
A better solution: restraints from QM

$T = T_{DATA} + w * T_{RESTRAINTS}$

Optimize consensus between model-to-data fit and... common sense







Bonds, angles, planes, torsions, chirality, non-bonded repulsion

Replace with energies/gradients from QM calculations

NEW: AQuaRef – QM based refinement in *Phenix*

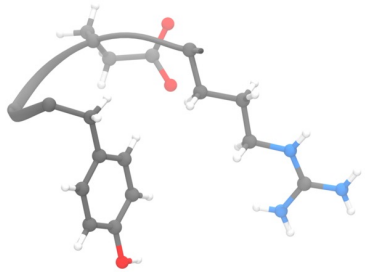
History of progress

Impossible

2010

QM Calculations

Impossible for proteins. Limited to small molecules only

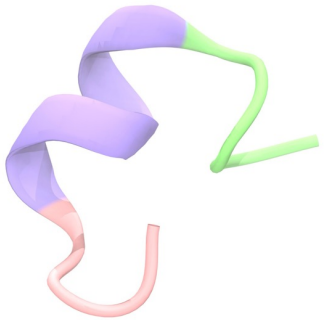


Limited

2012

GPU Accelerated QM

Limited to peptides and very small proteins (~hundreds of atoms)

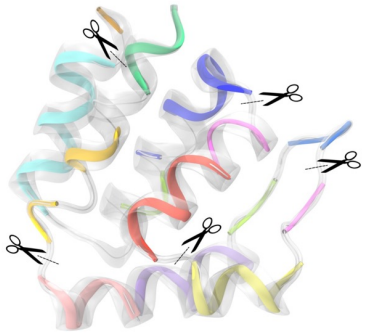


Possible

2017

Q|R with Fragmentation

QM-based protein refinement. Slow, resource-intensive, no inherent size limit

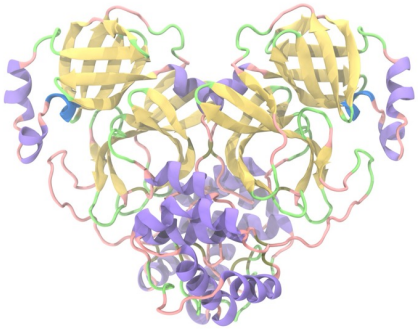


Practical

2024

ML Potentials

Fast, rivaling classical force fields, with QM-level accuracy and no fragmentation required

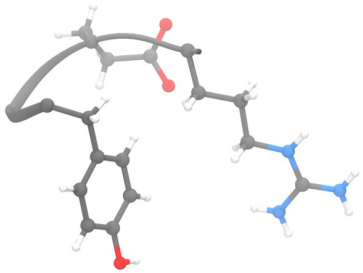


History of progress

Impossible
2010

QM Calculations

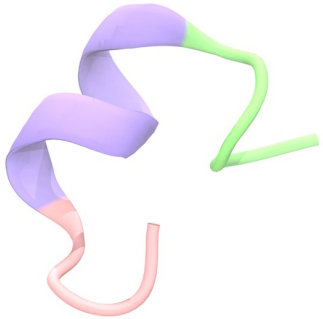
Impossible for proteins. Limited to small molecules only



Limited
2012

GPU Accelerated QM

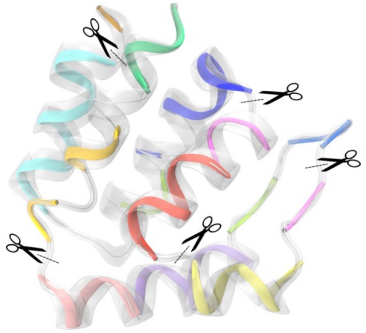
Limited to peptides and very small proteins (~hundreds of atoms)



Possible
2017

Q|R with Fragmentation

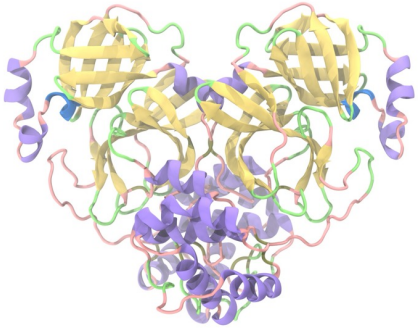
QM-based protein refinement. Slow, resource-intensive, no inherent size limit



Practical
2024

ML Potentials

Fast, rivaling classical force fields, with QM-level accuracy and no fragmentation required



Acta Cryst D STRUCTURAL BIOLOGY
ISSN 2059-7983

Solving the scalability issue in quantum-based refinement: Q|R#1

Min Zheng,[a,b] Nigel W. Moriarty,[c] Yanting Xu,[a] Jeffrey R. Reimers,[a,d] Pavel V. Afonine[a,c]* and Mark P. Waller[a]*

Acta Cryst D STRUCTURAL BIOLOGY
ISSN 2059-7983

Including crystallographic symmetry in quantum-based refinement: Q|R#2

Min Zheng,[a,b] Malgorzata Biczysko,[a] Yanting Xu,[a] Nigel W. Moriarty,[c] Holger Kruse,[d] Alexandre Urzhumtsev,[e,f] Mark P. Waller[g]* and Pavel V. Afonine[c]*

Acta Cryst D STRUCTURAL BIOLOGY
ISSN 2059-7983

Real-space quantum-based refinement for cryo-EM: Q|R#3

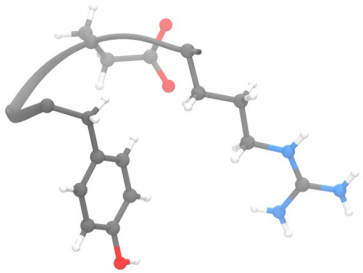
Lum Wang,[a] Holger Kruse,[b] Oleg V. Sobolev,[c] Nigel W. Moriarty,[c] Mark P. Waller,[d]* Pavel V. Afonine[c] and Malgorzata Biczysko[a]*

History of progress

Impossible — 2010

QM Calculations

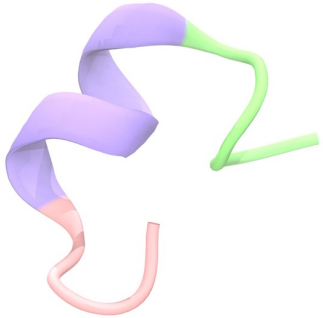
Impossible for proteins. Limited to small molecules only



Limited — 2012

GPU Accelerated QM

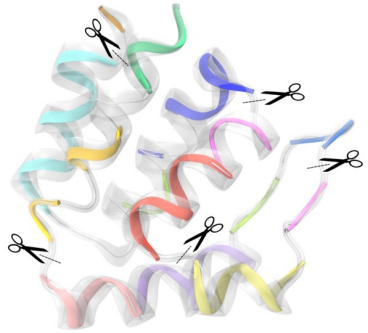
Limited to peptides and very small proteins (~hundreds of atoms)



Possible — 2017

Q|R with Fragmentation

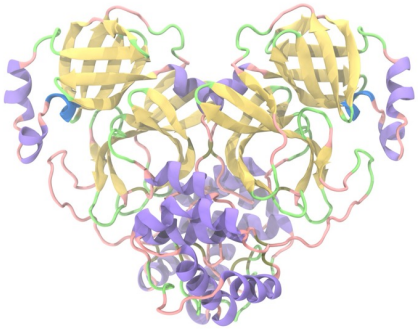
QM-based protein refinement. Slow, resource-intensive, no inherent size limit



Practical — 2024

ML Potentials

Fast, rivaling classical force fields, with QM-level accuracy and no fragmentation required



AQuaRef: Machine learning accelerated quantum refinement of protein structures

Roman Zubatyuk, Malgorzata Biczysko, Kavindri Ranasinghe, Nigel W. Moriarty, Hatice Gokcan, Holger Kruse, Billy K. Poon, Paul D. Adams, Mark P. Waller, Adrian E. Roitberg, Olexandr Isayev, Pavel V. Afonine

doi: https://doi.org/10.1101/2024.07.21.604493

Machine Learning potential (AIMNet2)

Standard amino-acids

- Generate all possible 1-, 2-, 3-, and 4-peptides (including S-S bridges)
- Torsion and non-equilibrium sampling

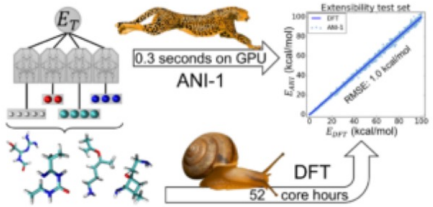
Large Dataset

DFT calculations

ML

Calculation time:

About a week on one of big national computing resources



Smith, Justin S.; Isayev, Olexandr; Roitberg, Adrian E.

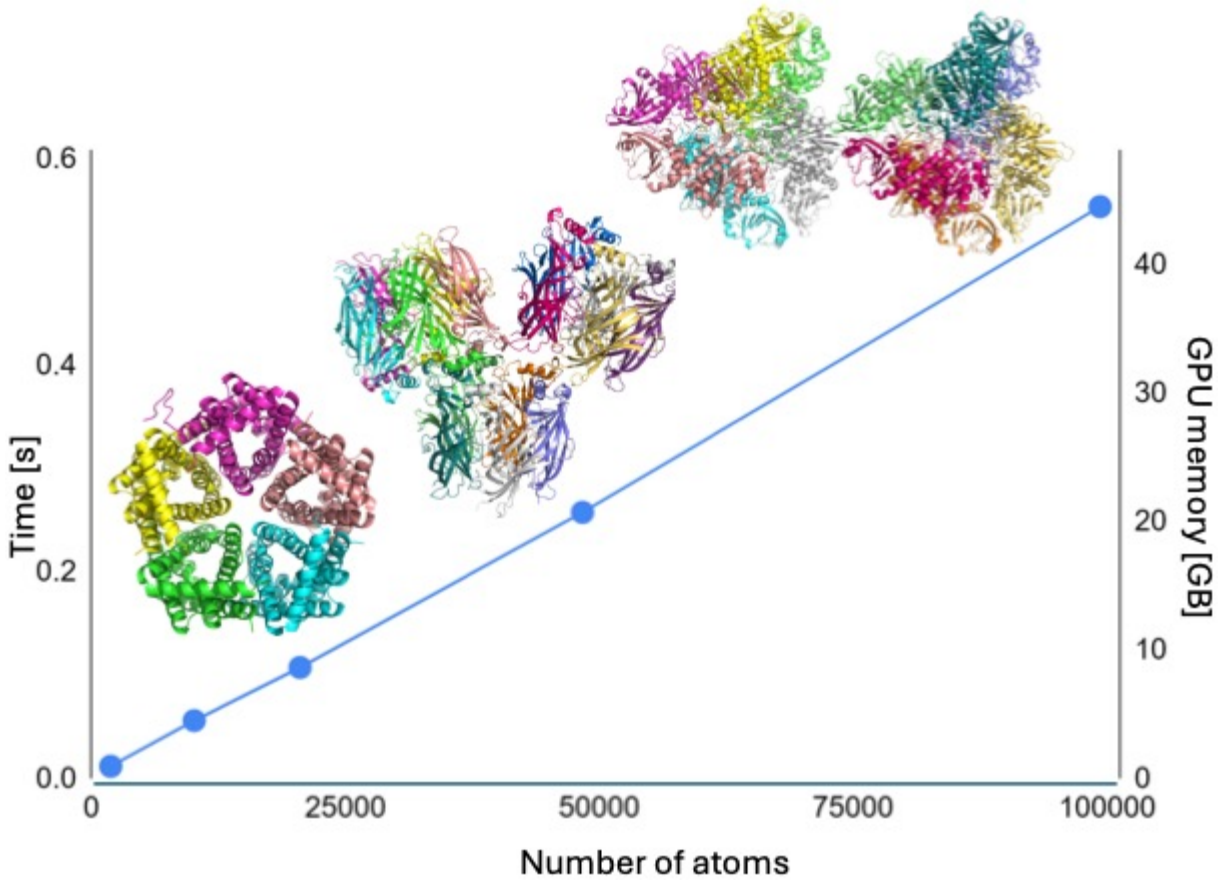
ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost

Journal Article

In: Chemical Science, iss. 8, pp. 3192-3203, 2017.

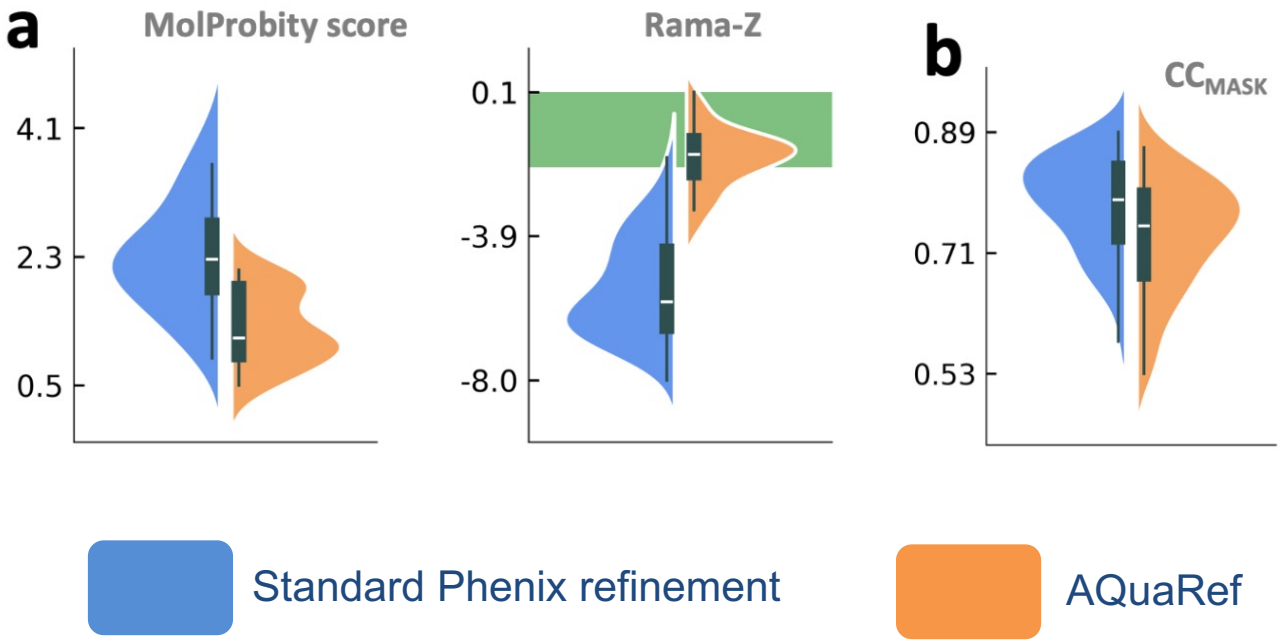
Abstract | **Links** | BibTeX | Tags: ANI, Machine learning potential

Time & Memory Scaling: single energy calculation



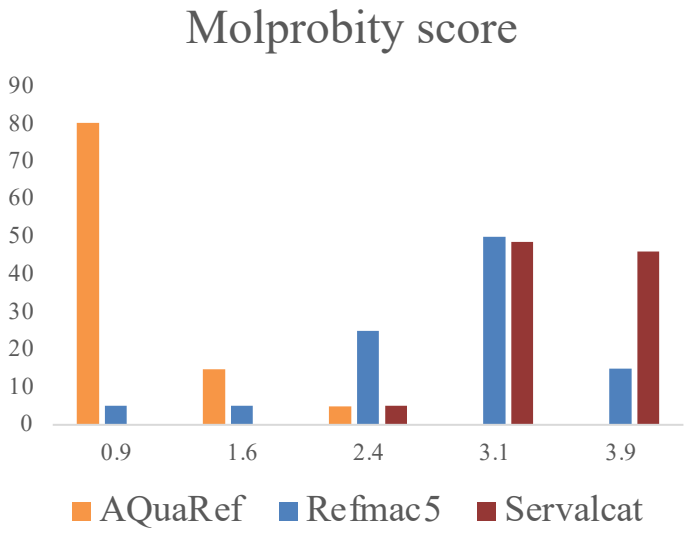
AQuaRef *vs* standard *Phenix* refinement

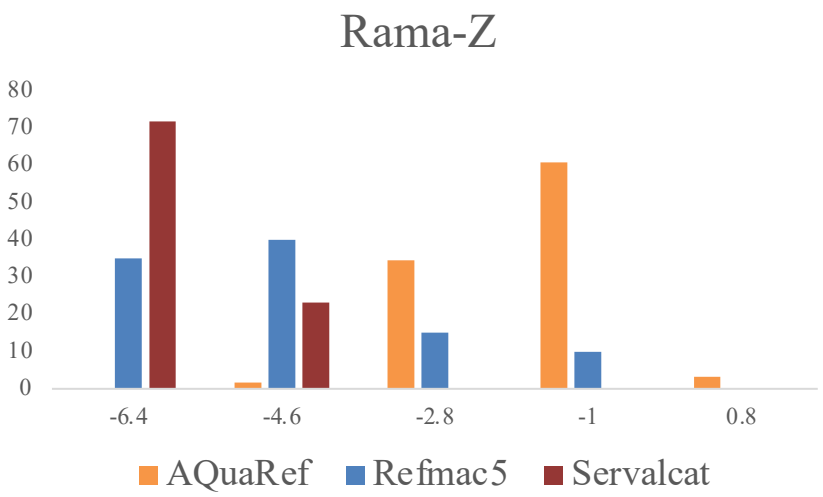
40 cryo-EM low resolution models (3Å or worse)



Testing: AQuaRef *vs* others

40 cryo-EM low resolution models (3Å or worse)





AQuaRef facts

- Refined models have superior geometry, same or better fit to the data
- Proteins only. Including other types of molecules is work in progress
- No alternative conformations
- Intolerant to nonsense: no severely distorted geometry, atom-complete models
- Same runtime as standard refinement
- GPU laptop or a workstation , CUDA 11/12 on Linux
- Available in nightly builds of Phenix

Validation (again!)

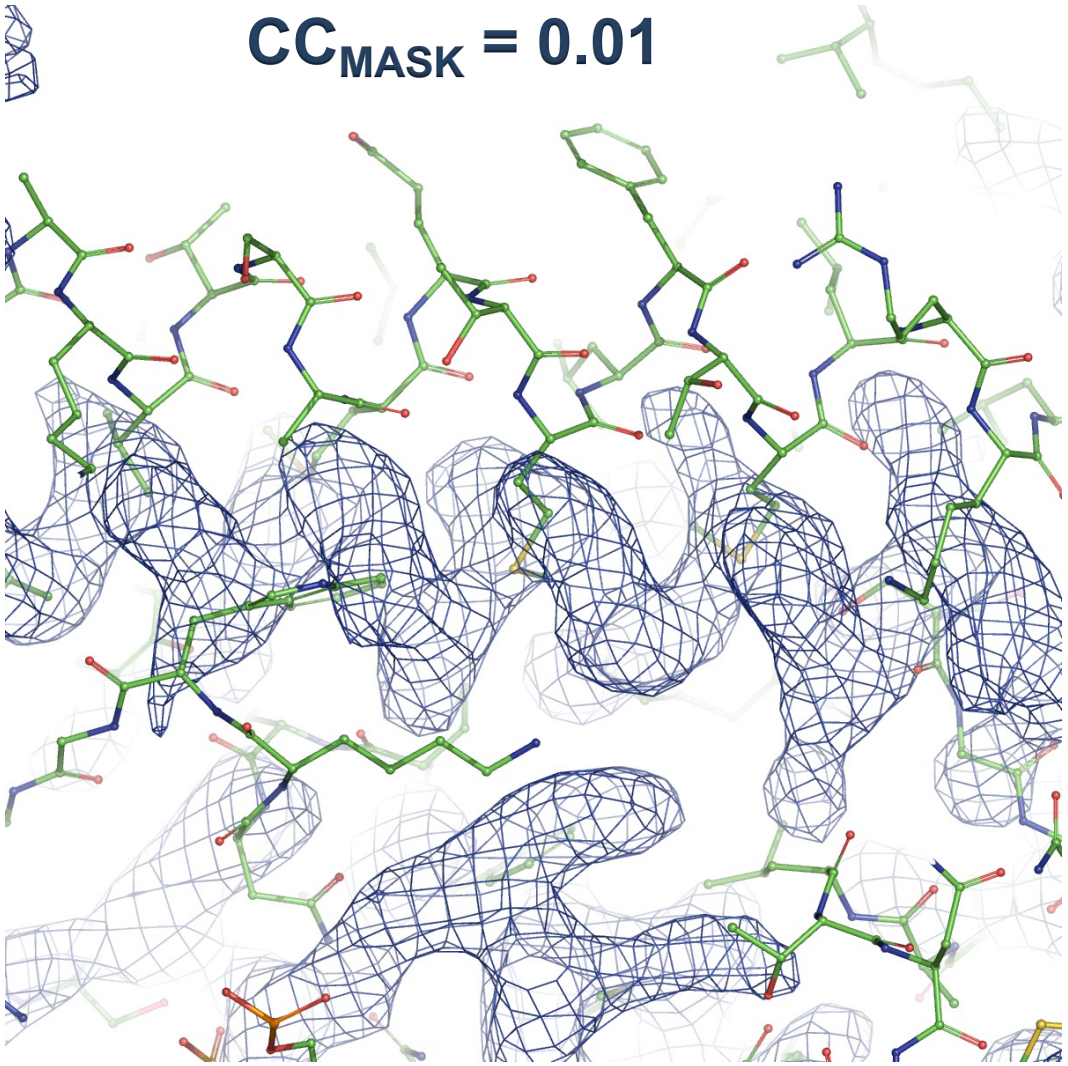
Despite efforts to promote and enforce validation, poorly scoring models are still being deposited into databases

Examples (recent years)

Model does not fit the map

PDB: 8gwb | EMDB: 34308 | 2.8 Å | Cell (2022) 185: 4347-4360

Chain	CC_{MASK}
A	0.01
B	0.02
C	0
D	0.01
I	0.04
J	0
F	0.12
E	0.08
G	0.1
M	0.16
A	0
F	-0.13
E	0.16
A	0.1
G	0.15
M	0.19

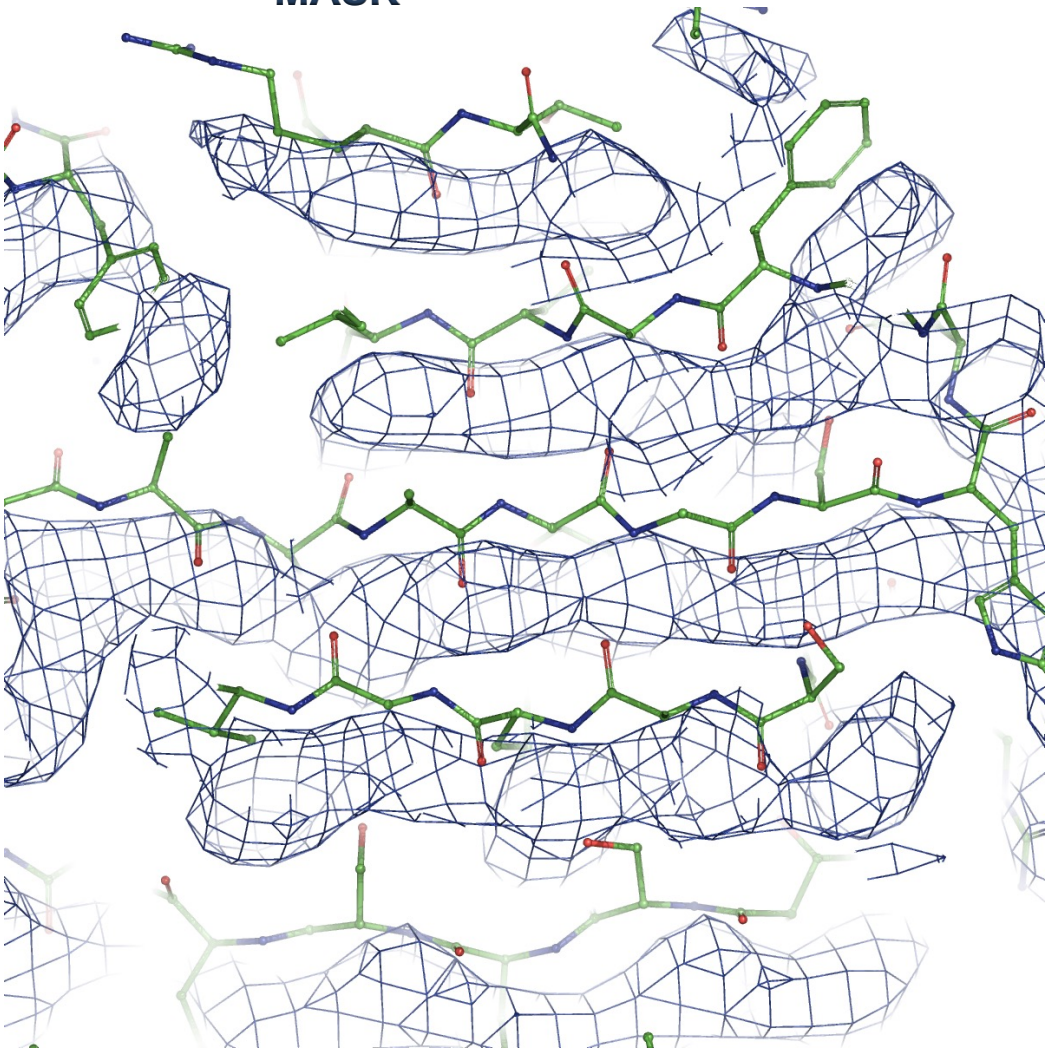


Model does not fit the map

PDB: 7xov | EMDB: 33360 | 3 Å | Cell Discov (2022) 8: 55-55

Chain	CC_{MASK}
A	0.04
B	-0.01
G	0.18
N	0.06
R	0.03
R	-0.02

CC_{MASK} = 0.02

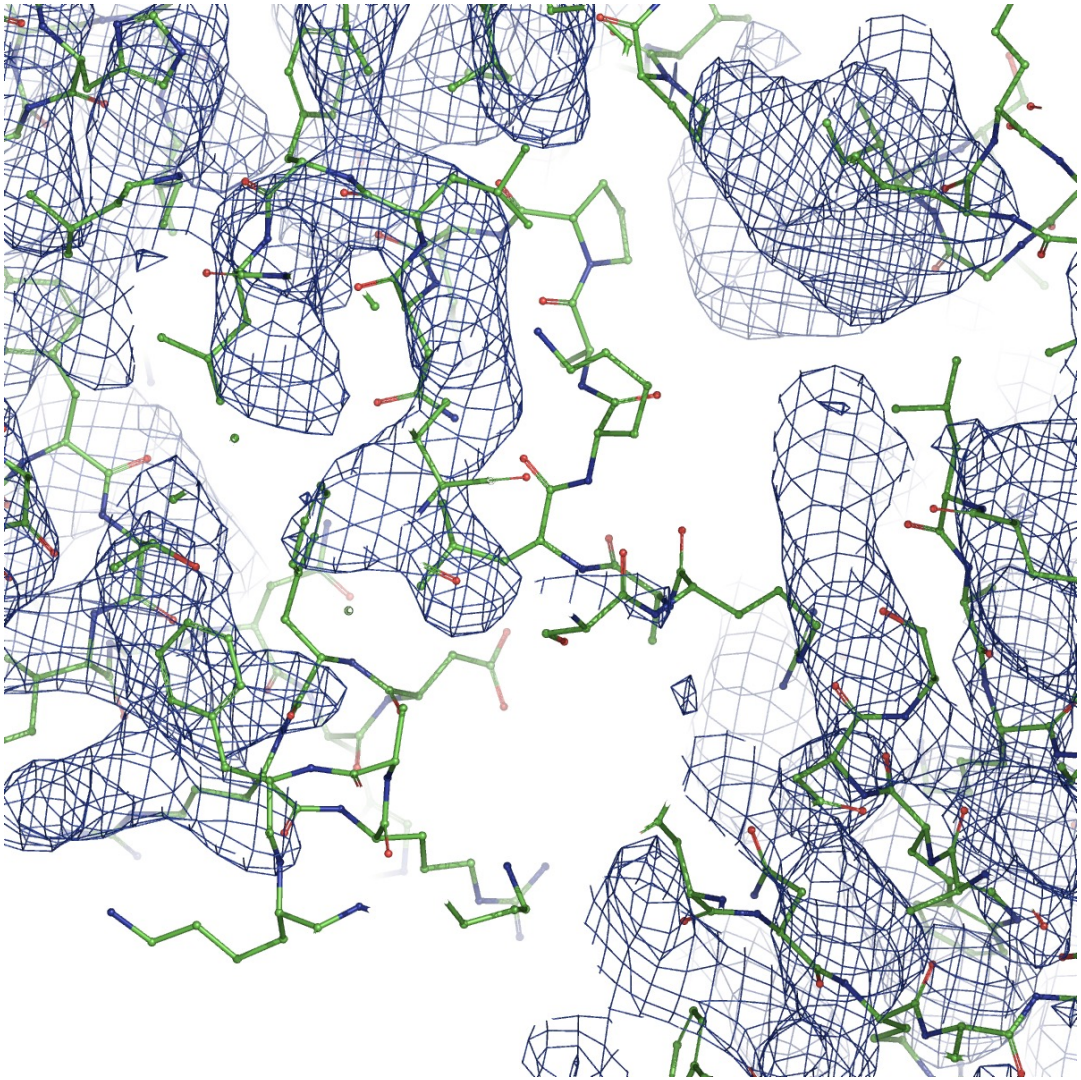


Model does not fit the map

PDB: 7w6p | EMDB: 32331 | 3.5 Å | Science (2022) 377: 7065-7065

Chain	CC_{MASK}
A	0.09
B	0.11
G	0.12
H	0.07
R	0.16
R	-0.08

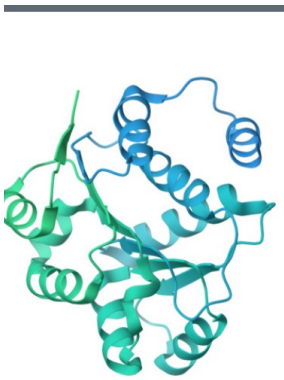
CC_{MASK} = 0.1



Model does not fit the map

PDB: 8V85 | EMDB: 43023 | 2.9 Å | Nat Commun (2024) 15: 3296-3296

CC_{MASK} = 0.15



lore in 3D: Structure | Sequence
ations | Electron Density |
tion Report

l Symmetry: Asymmetric - C1
l Stoichiometry: Monomer - A1

imilar Assemblies

8V85 | pdb_000(

60S ribosome biogenesis intermediate
pass filtered locally refined map)

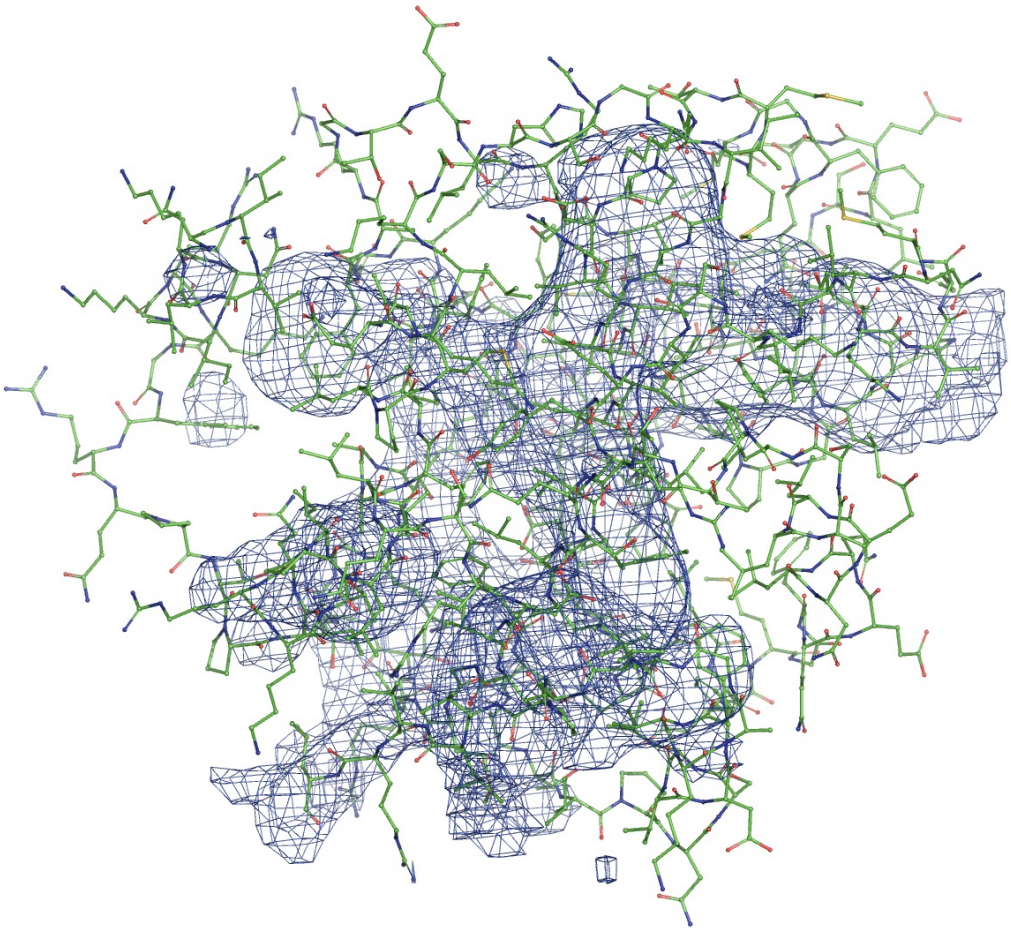
PDB DOI: https://doi.org/10.2210/pdb8V85/pdb

Classification: RNA BINDING PROTEIN
Organism(s): Saccharomyces cerevisiae BY4741
Mutation(s): No

Deposited: 2023-12-04 **Released:** 2024-05-01
Deposition Author(s): Cruz, V.E., Weirich, C.S., Pe
Funding Organization(s): National Institutes of He
(NIH/NIGMS), Robert A. Welch Foundation, Cancer
(CPRIT)

Experimental Data Snapshot wwPD

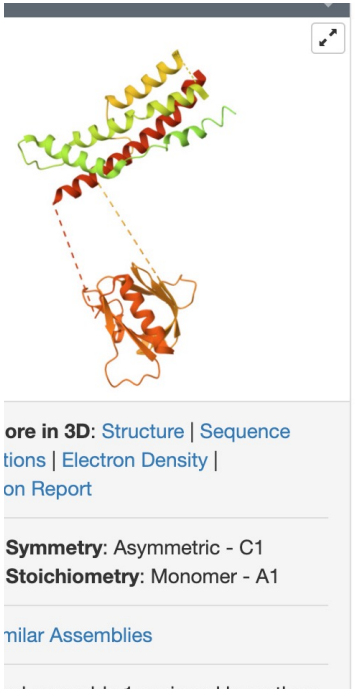
Method: ELECTRON MICROSCOPY
Resolution: 2.90 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE



Model does not fit the map

PDB: 8SZ7 | EMDB: 40902 | 2.8 Å | Dev Cell (2024) 59: 1783

CC_{MASK} = 0.19



ore in 3D: Structure | Sequence
tions | Electron Density |
on Report

Symmetry: Asymmetric - C1
Stoichiometry: Monomer - A1

milar Assemblies

8SZ7 | pdb_00

Cryo-EM of the GDP-bound human
membrane in the super constricted

PDB DOI: https://doi.org/10.2210/pdb8SZ7/pdb

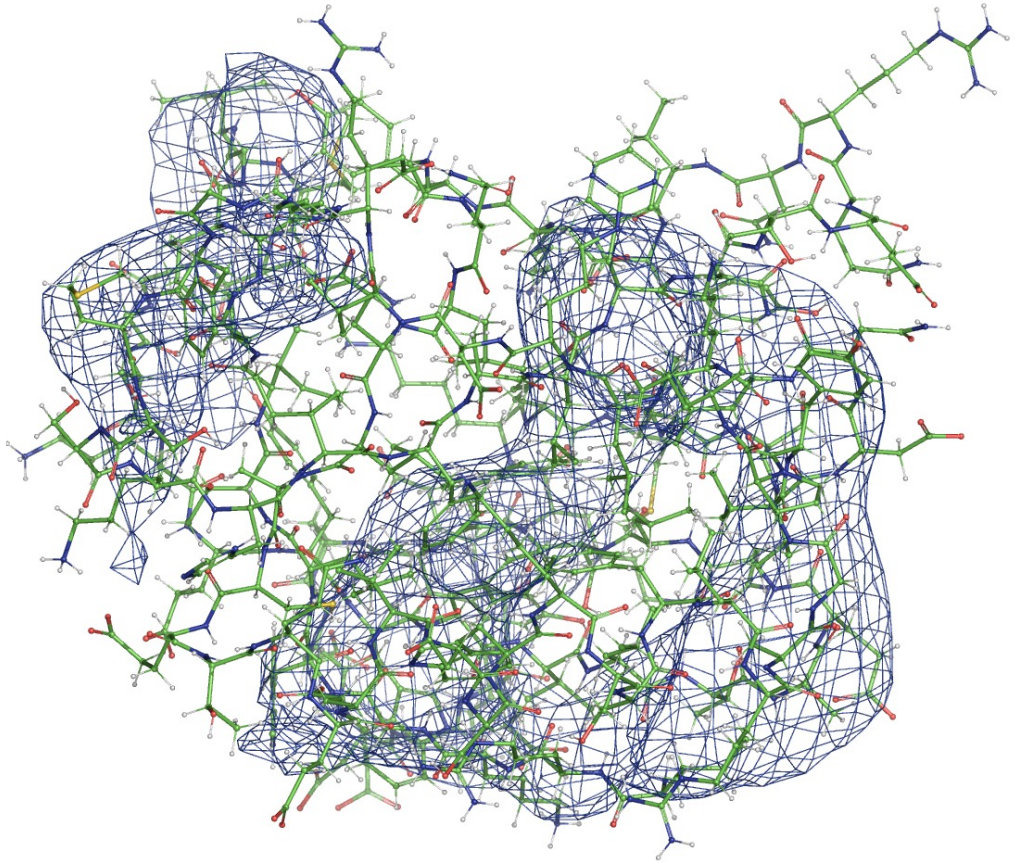
Classification: HYDROLASE
Organism(s): Homo sapiens
Expression System: Escherichia coli 'BL21-Go
Mutation(s): Yes

Deposited: 2023-05-27 **Released:** 2024-05-01
Deposition Author(s): Jimah, J.R., Canagarajah
Funding Organization(s): National Institutes of
and Kidney Disease (NIH/NIDDK), National Instit
Sciences (NIH/NIGMS)

Experimental Data Snapshot ww

Method: ELECTRON MICROSCOPY
Resolution: 2.84 Å
Aggregation State: FILAMENT
Reconstruction Method: HELICAL

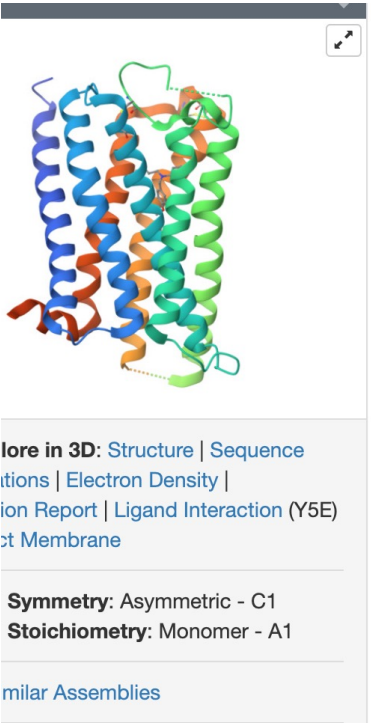
Ran



Model does not fit the map

PDB: 8x63 | EMDB: 38078 | 3.2 Å | Nat Commun (2024) 15: 84-84

CC_{MASK} = 0.13



lore in 3D: Structure | Sequence
tions | Electron Density |
ion Report | Ligand Interaction (Y5E)
ct Membrane

Symmetry: Asymmetric - C1
Stoichiometry: Monomer - A1

milar Assemblies

8X63 | **pdb_00008**

CryoEM structure of the histamine H1 recep
with mepyramine

PDB DOI: https://doi.org/10.2210/pdb8X63/pdb **EM Ma**

Classification: MEMBRANE PROTEIN
Organism(s): Homo sapiens, Escherichia coli
Expression System: Spodoptera frugiperda
Mutation(s): Yes
Membrane Protein: Yes PDBTM MemProtMD mpstruc

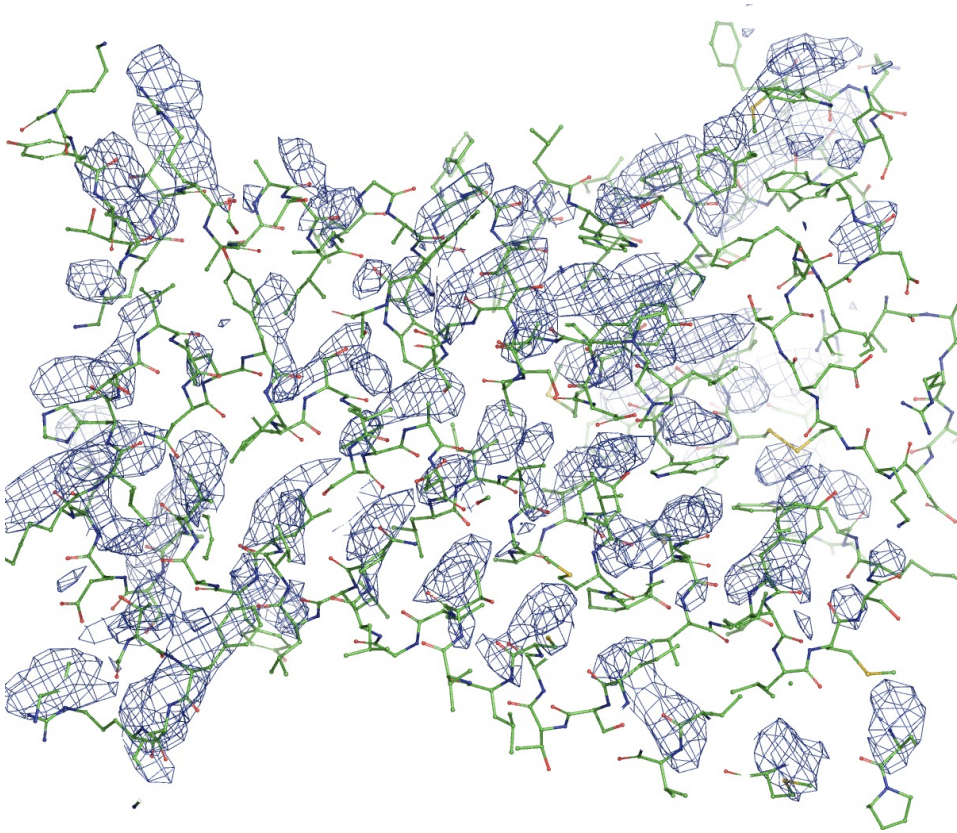
Deposited: 2023-11-20 **Released:** 2024-01-17
Deposition Author(s): Wang, D.D., Guo, Q.
Funding Organization(s): National Natural Science Found

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY
Resolution: 3.20 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE

wwPDB Vali

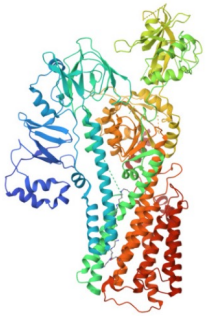
Clash
Ramachandran ou
Sidechain ou



Model does not fit the map

PDB: 8iEN | EMDB: 35387 | 3.25 Å | Nat Commun (2023) 14: 1978-1978

CC_{MASK} = 0.0



xplore in 3D: Structure | Sequence
tations | Electron Density |
ation Report |
nd Interaction (SPM) |
ct Membrane

al Symmetry: Asymmetric - C1
al Stoichiometry: Monomer - A1

8IEN | **pdb_00008ien**

Cryo-EM structure of ATP13A2 in the E2-Pi state

PDB DOI: https://doi.org/10.2210/pdb8IEN/pdb **EM Map EMD-35**

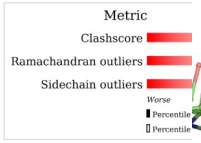
Classification: TRANSPORT PROTEIN
Organism(s): Homo sapiens
Expression System: Homo sapiens
Mutation(s): No
Membrane Protein: Yes PDBTM MemProtMD

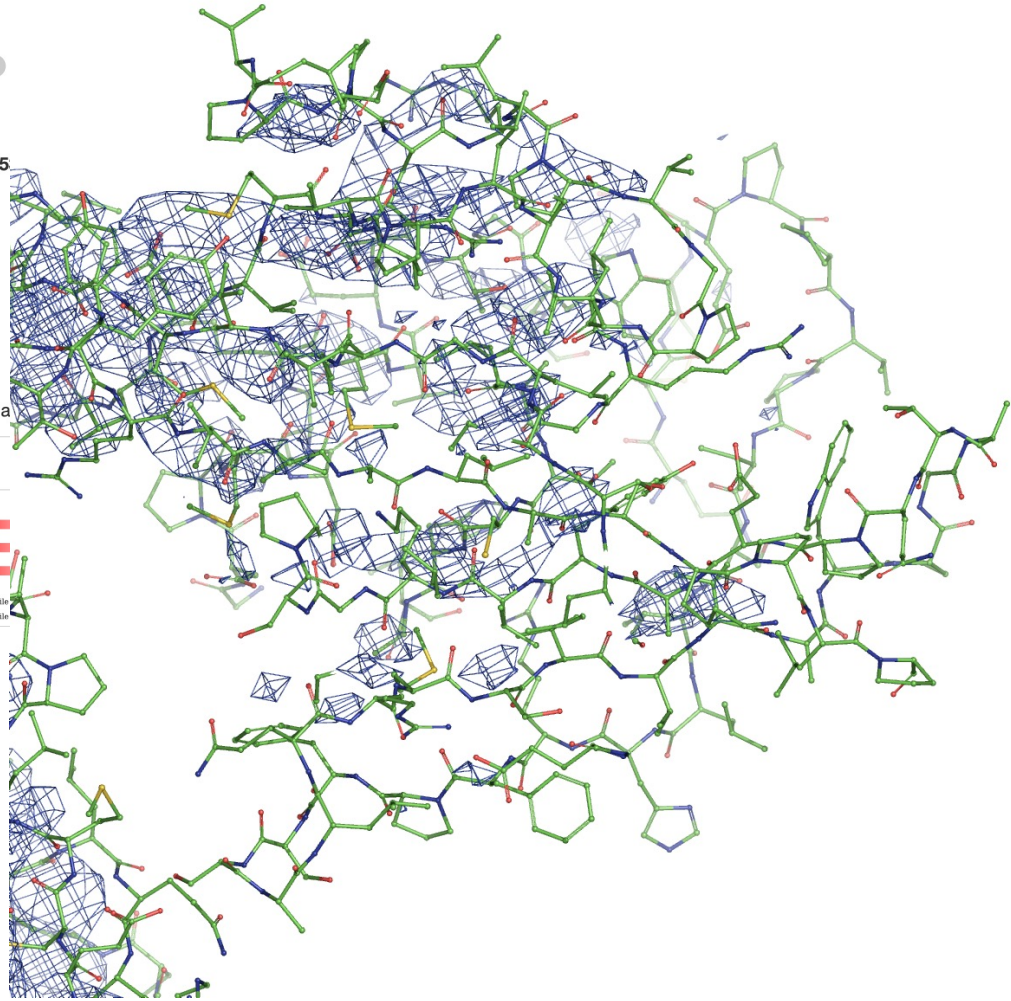
Deposited: 2023-02-15 **Released:** 2023-12-20
Deposition Author(s): Liu, Z.M., Mu, J.Q., Xue, C.Y.
Funding Organization(s): National Science Foundation (NSF, China

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY
Resolution: 3.25 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE

wwPDB Validation

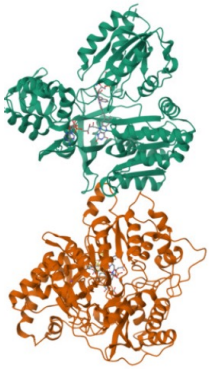




Model does not fit the map

PDB: 9c91 | EMDB: 45359 | 2.78 Å | Nat Commun (2025) 16: 2955

CC_{MASK} = 0.0



in 3D: Structure | Sequence
s | Electron Density |
eport |
raction (SRM)

metry: Asymmetric - C1
chiometry: Hetero 2-mer -

9C91 | pdb_00

Assimilatory NADPH-dependent sulf

PDB DOI: https://doi.org/10.2210/pdb9C91/pdb

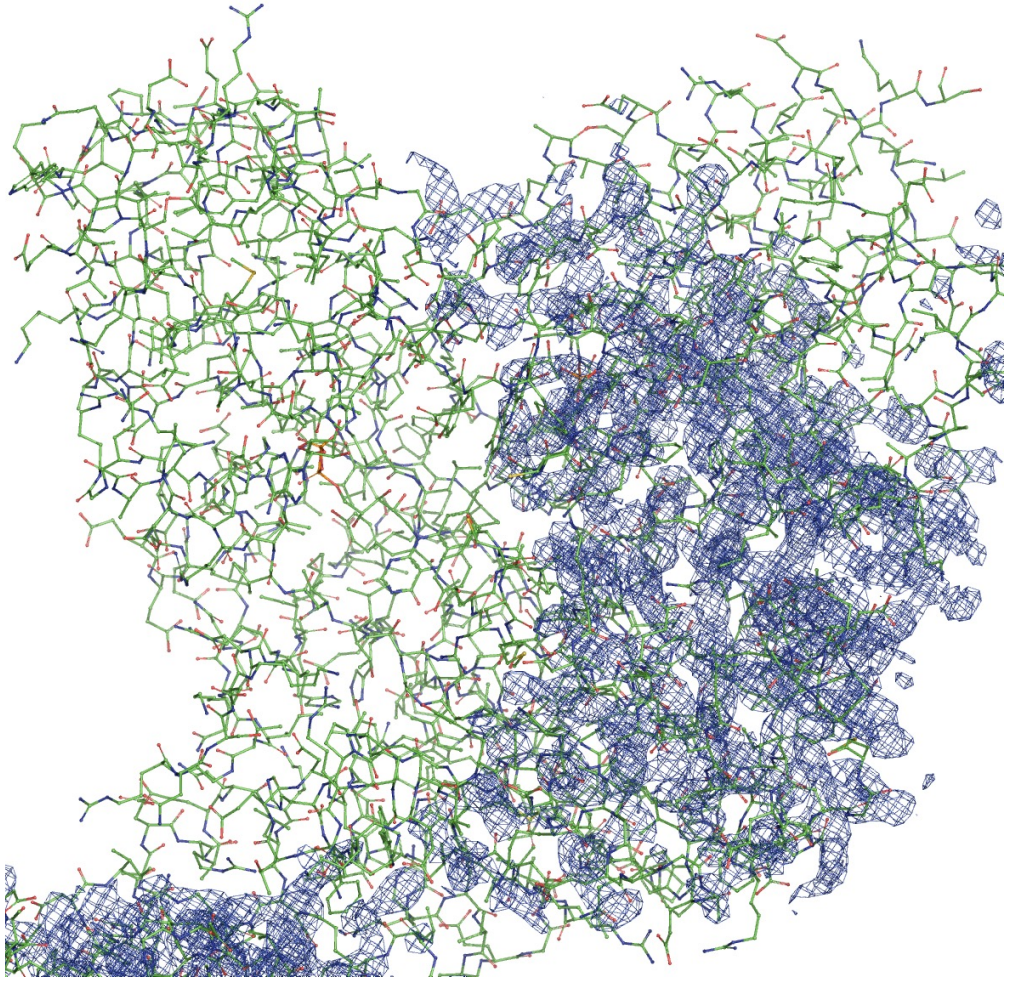
Classification: FLAVOPROTEIN
Organism(s): Escherichia coli
Expression System: Escherichia coli
Mutation(s): No

Deposited: 2024-06-13 Released: 2025-02-12
Deposition Author(s): Ghazi Esfahani, B., Walia
Mendez, J.H., Stroupe, M.E.
Funding Organization(s): National Science Fou

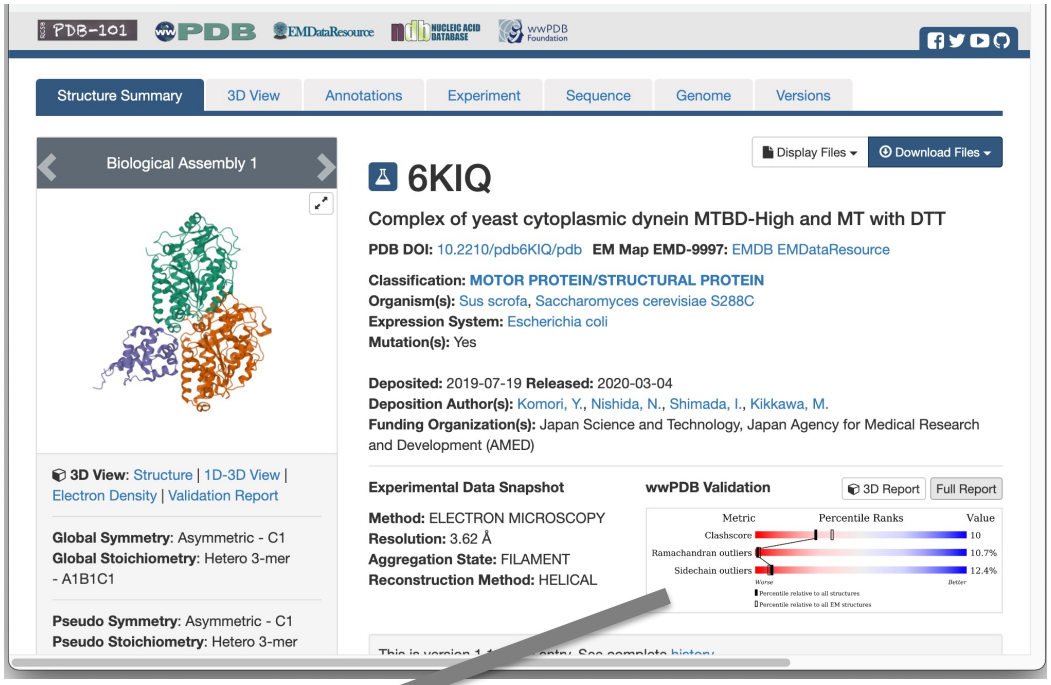
Experimental Data Snapshot ww

Method: ELECTRON MICROSCOPY
Resolution: 2.78 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE

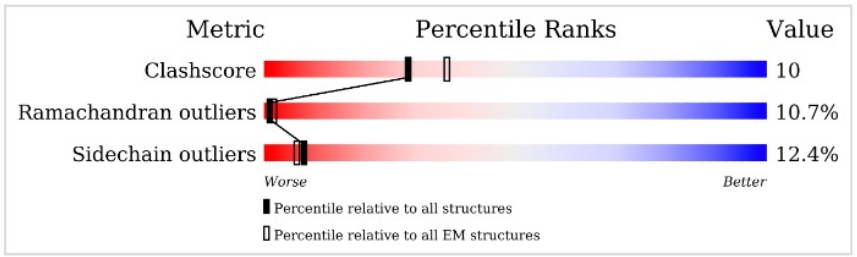
Ran



Validation reports (RCSB)



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

Lack of (useful) model-to-map fit statistics!

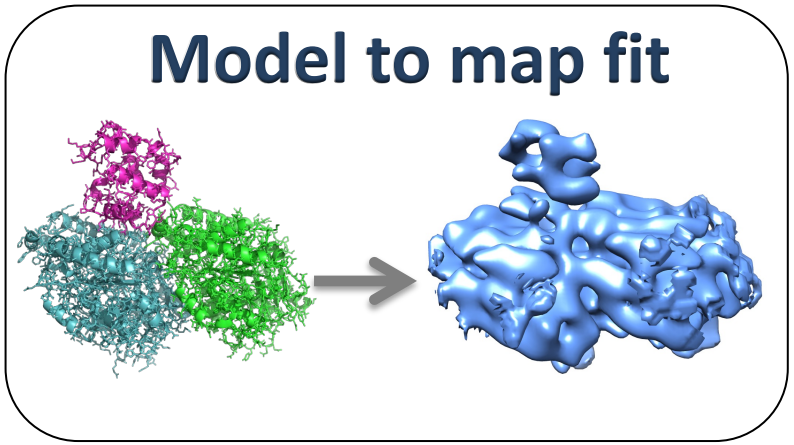
Atom inclusion

- **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
  - 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 is not used:
    - 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
  - Anisotropic atoms are ***not*** handled

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



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

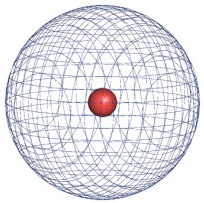
ρ_{obs} = experimental map
ρ_{calc} = model calculated map

- 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

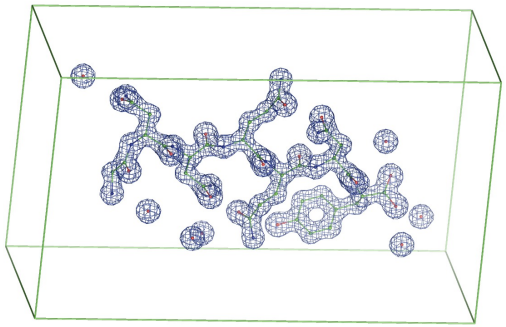
Model-to-map fit validation: CC_{MASK}

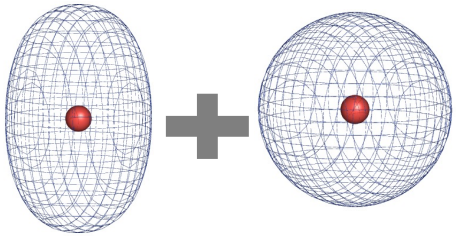
- Gaussian IAM (Independent Atom Model)



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

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



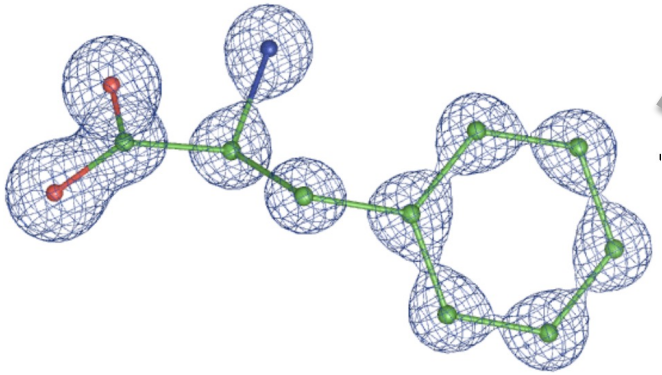


$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

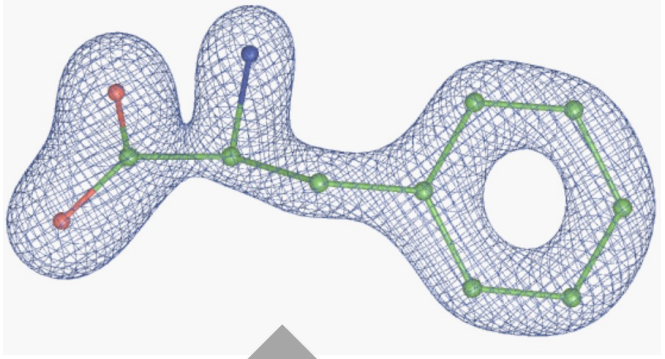
Model-to-map fit validation: CC_{MASK}

3Å model-calculated map

Exact model map



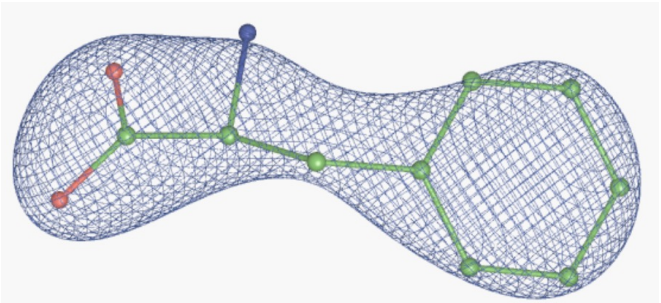
FT



CC_{MASK}

CC_{MASK}

3Å experimental map



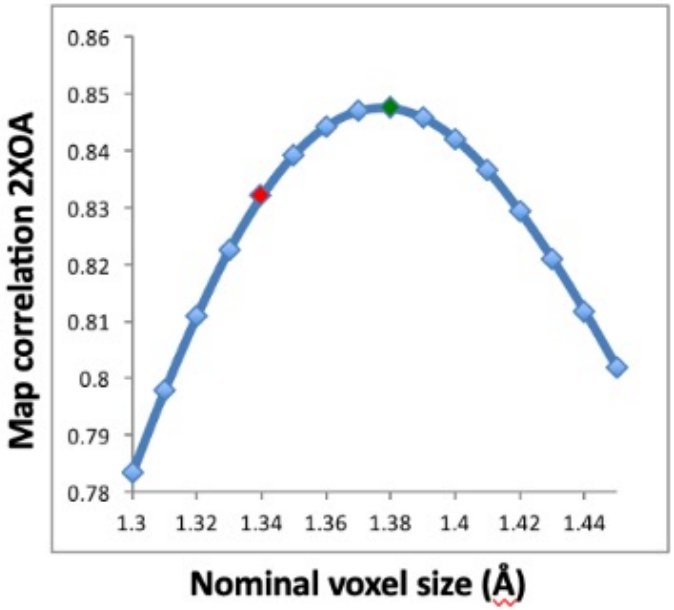
$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

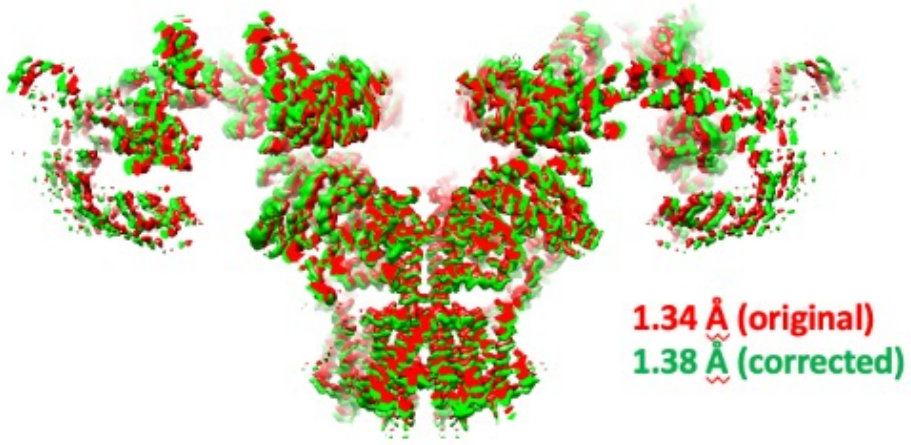
- FT exact model map
- Remove terms up to specified resolution
- FT back to real space to get a Fourier image = “Model map”

Voxel size (magnification) calibration

Using UCSF Chimera for voxel size calibration (of your map and others)

- Voxel size generally requires calibration against a crystal structure.
- Once calibrated, generally stable between samples/datasets at same magnification.
- Can calibrate by fitting in Chimera at range of nominal voxel sizes and measuring correlation.
- Incorrect voxel sizes are common in deposited maps - **be aware of this when comparing structures**. E.g. here there is a 3% difference – affects structural alignment, reported resolution (3.8 vs 3.9Å).






```
phenix.magref map.mrc model.pdb resolution=3.4
```