# Refinement and Validation

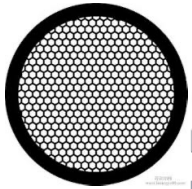
**Pavel Afonine**


**October 28th, 2025**
**UTMB, Galveston, TX**

# Solving structure by electron cryo-microscopy

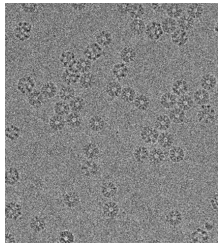
Pure sample

Grid

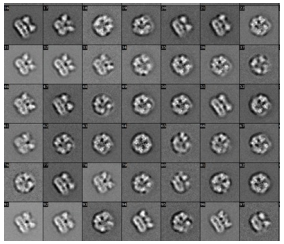


Electron microscope

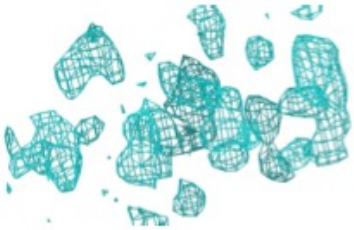
Electron microscope



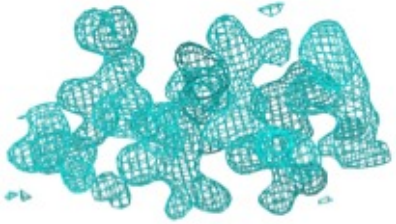
Pick, align & average particles



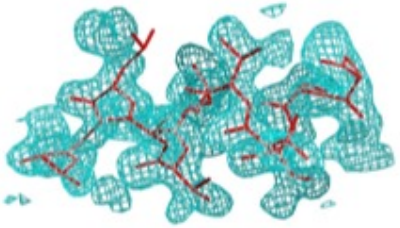
Initial map



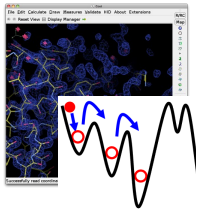
Improved map



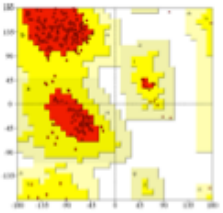
Build model



Refine model

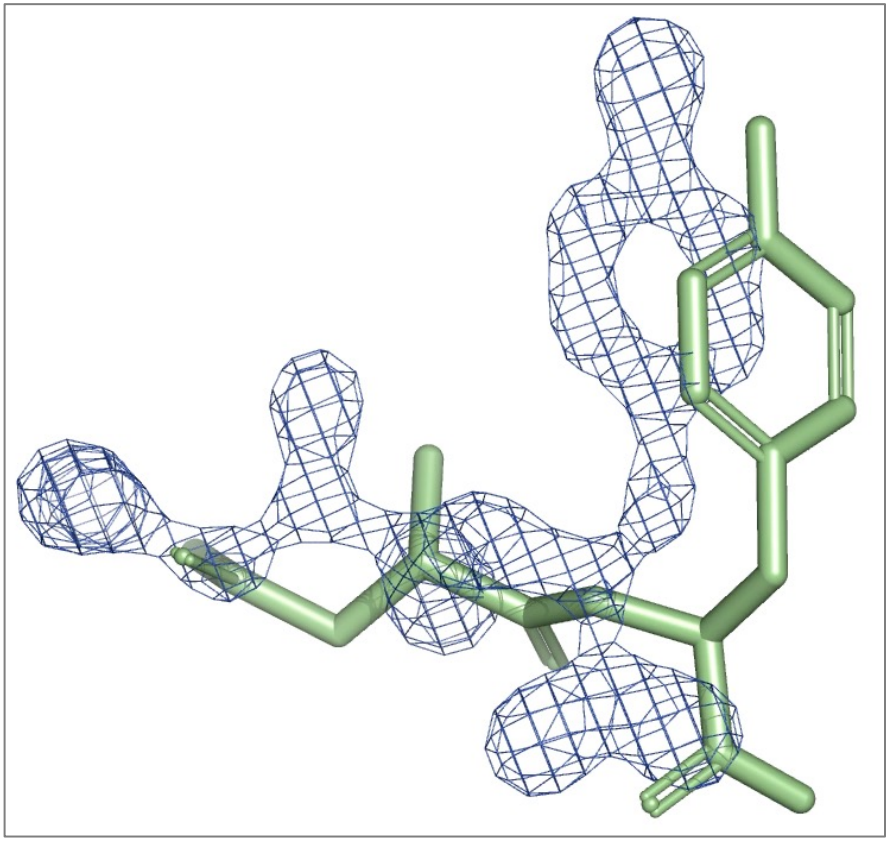


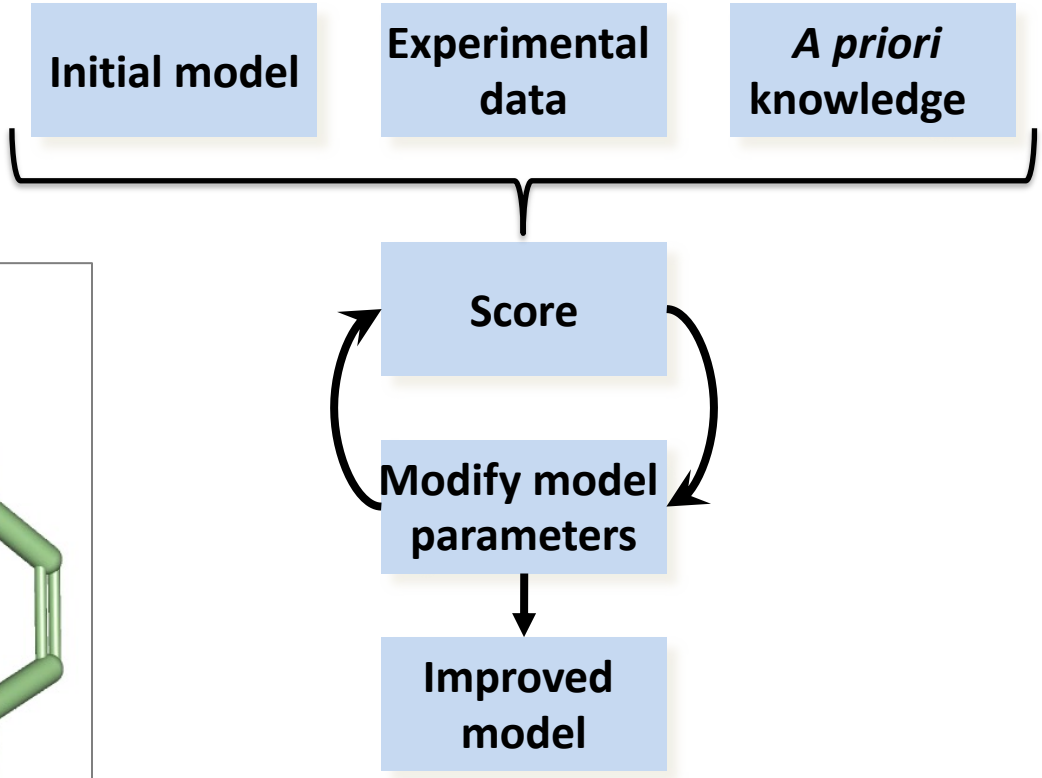
Validate model



Publish & deposit

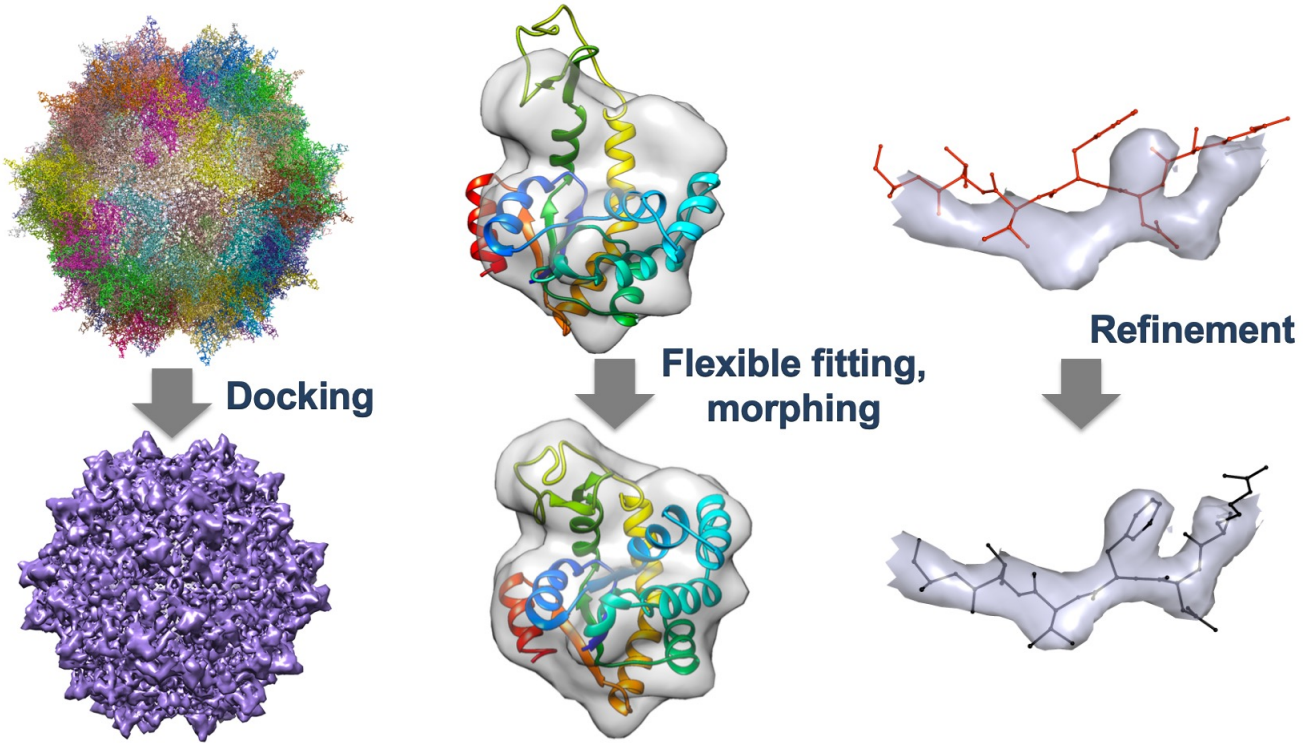

# Model refinement






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 fine atomic model
  - Refinement does only small changes to the model
  - *Convergence radius of refinement* ~ 1Å

# Refining structure in the past

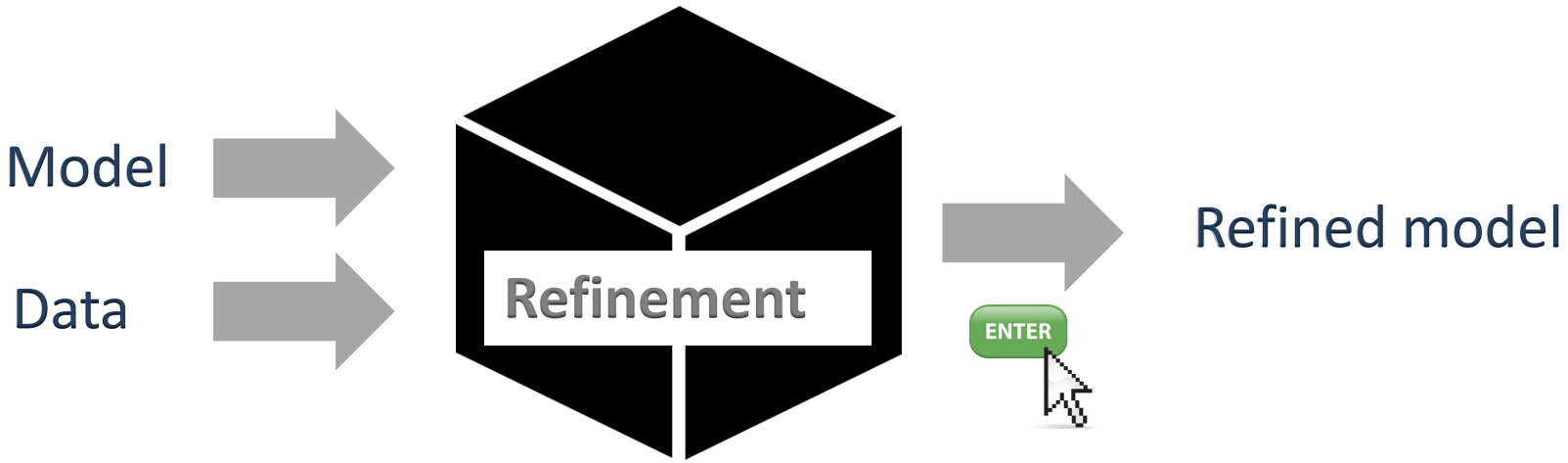
- Familiar with many software packages (often with “orthogonal” philosophies)
- Mutually incompatible file formats for common data exchange
- Coding experience was a must (typically using arcane languages FORTRAN or C)
- No GUIs. Command line expertise (Unix)
- Reading thick books (no Google, YouTube or Chat bots!)
- Limited online forums
- Don’t expect your questions answered quickly by email
- Slow computers (with sometimes limited access)

# Solving structure in the past

- From many months to years
  - Spend days on graphics (manual atomic model building)
  - Run computations overnight
  - In the morning hope all worked, nothing crashed

Solving my first structure back in 1997

# Model refinement: black box

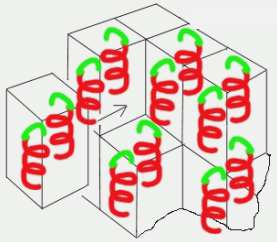


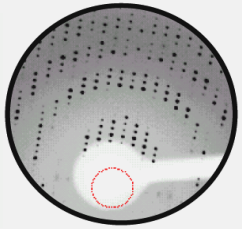

- Does it always work?

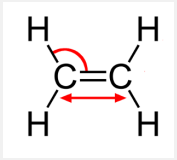
- Is it always as easy as *poor model in, better model out*?

# Refinement

## Crystallography








Initial model

Experimental data

*A priori* knowledge

Score

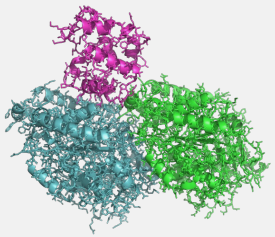
Modify model parameters

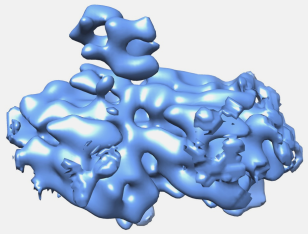
Improved model

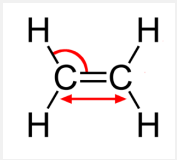
phenix.refine

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

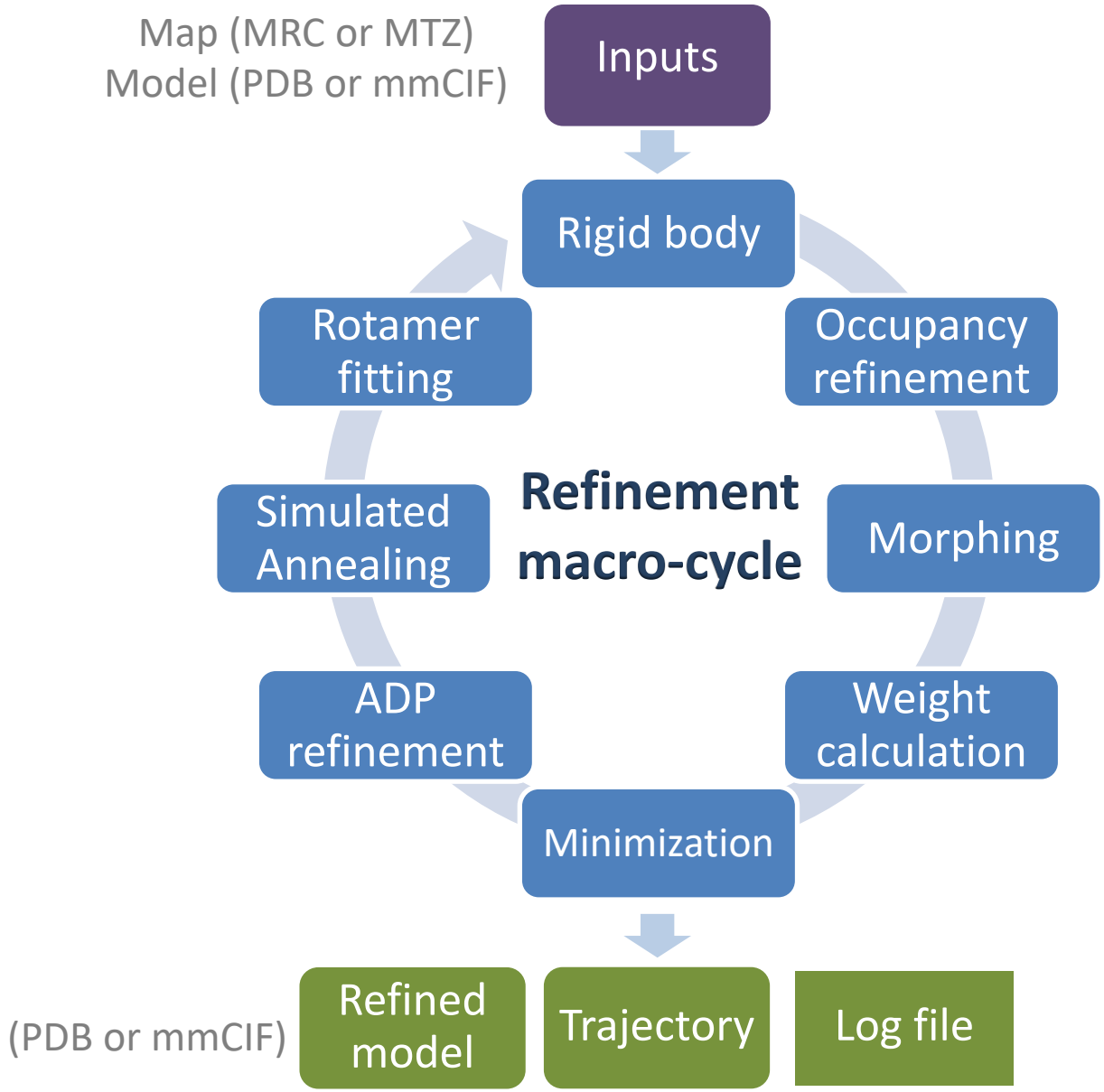
## Crystallographic refinement

- 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

# Refinement protocol



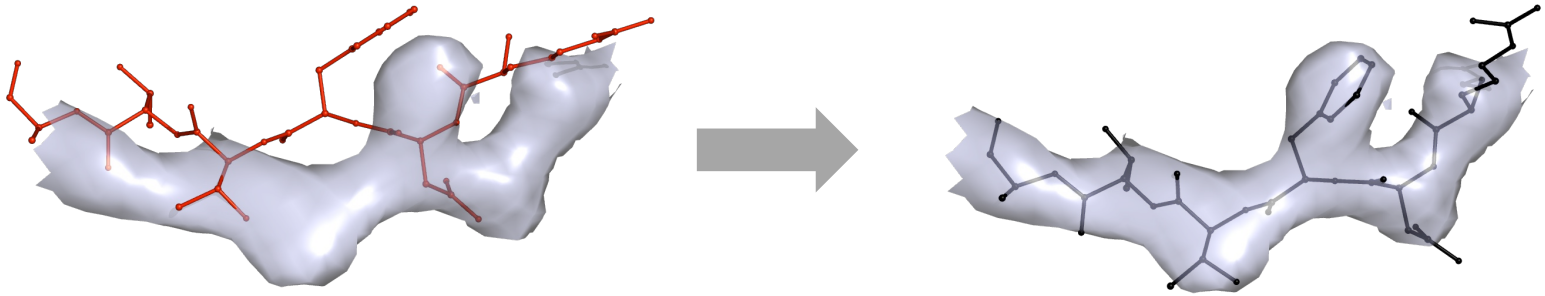
# Atomic model refinement: *phenix.real_space_refine*

## Direct refinement of atomic models against the map

Acta Cryst D STRUCTURAL BIOLOGY
ISSN 2059-7983

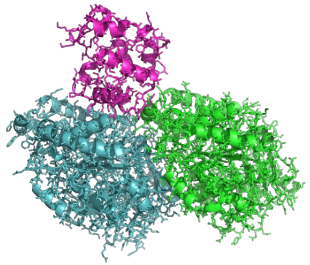
**Real-space refinement in *PHENIX* for cryo-EM and crystallography**

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]

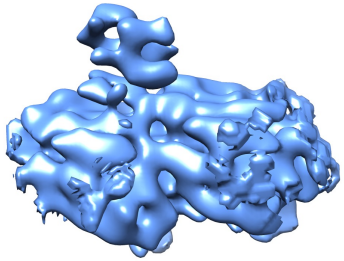


# Validation

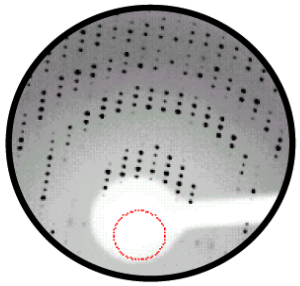
## Model



## Data



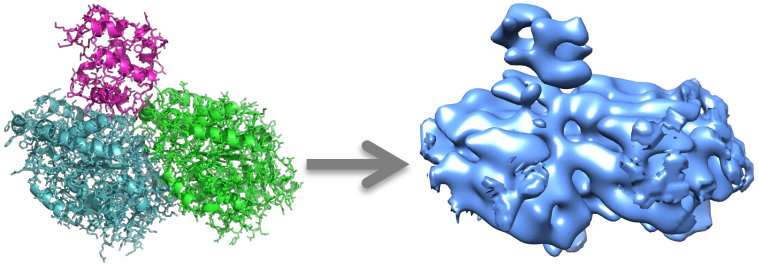
or



Cryo-EM

Diffraction

## Model to data fit



Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

# Validation: *why to do?*

- Can help to:
  - save (a lot of) time
  - produce better models
  - set correct expectations

- Subjectivity:
  - lot's of manual steps that depend on skills, pressure and ethics

- Software isn't perfect

- Databases are not perfect

# Validation: why to do?

- **Quality filters:**
  - You
  - Software you use
  - Your boss
  - Reviewers (of your paper)
  - PDB deposition (software and people)
  - Community

- **Unnoticed (intentionally or not) problems**
  - Likely discovered anyway, sooner or later

- **PDB is a public data bank of structures and experimental data**
  - PDB does not police the quality of depositions
    - PDB may point out some issues, but it is on you to fix them
  - You can (relatively easily) end up depositing a rubbish structure

# Validation: why to do?

## Retraction: Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution

Michael A Hanson & Raymond C Stevens

*Nat. Struct. Biol.* 7, 687–692 (2000); retracted 6 July 2009

In this paper, we described both the three-dimensional crystal structure of a botulinum toxin catalytic domain separated from the holotoxin (BoNT/B-LC, PDB 1F82) and a structure of the toxin catalytic domain in complex with a peptide (Sb2–BoNT/B-LC, PDB 1F83). The complex was later refined and deposited in the Protein Data Bank (PDB 3G94). The apo structure (PDB 1F82) remains valid. However, because of the lack of clear and continuous electron density for the peptide in the complex structure, the paper is being retracted. We apologize for any confusion this may have caused.

- H.M. Krishna Murthy (University of Alabama) – Protein Fabrication scandal
  - 12 falsified structures and 10 related papers
  - 1BEF, 1CMW, 1DF9, 2QID, 1G40, 1G44, 1L6L, 2OU1, 1RID, 1Y8E, 2A01, and 2HR0
  - Murthy's falsified data ended up affecting 449 papers at that time

Proper validation can prevent this!

# Validation

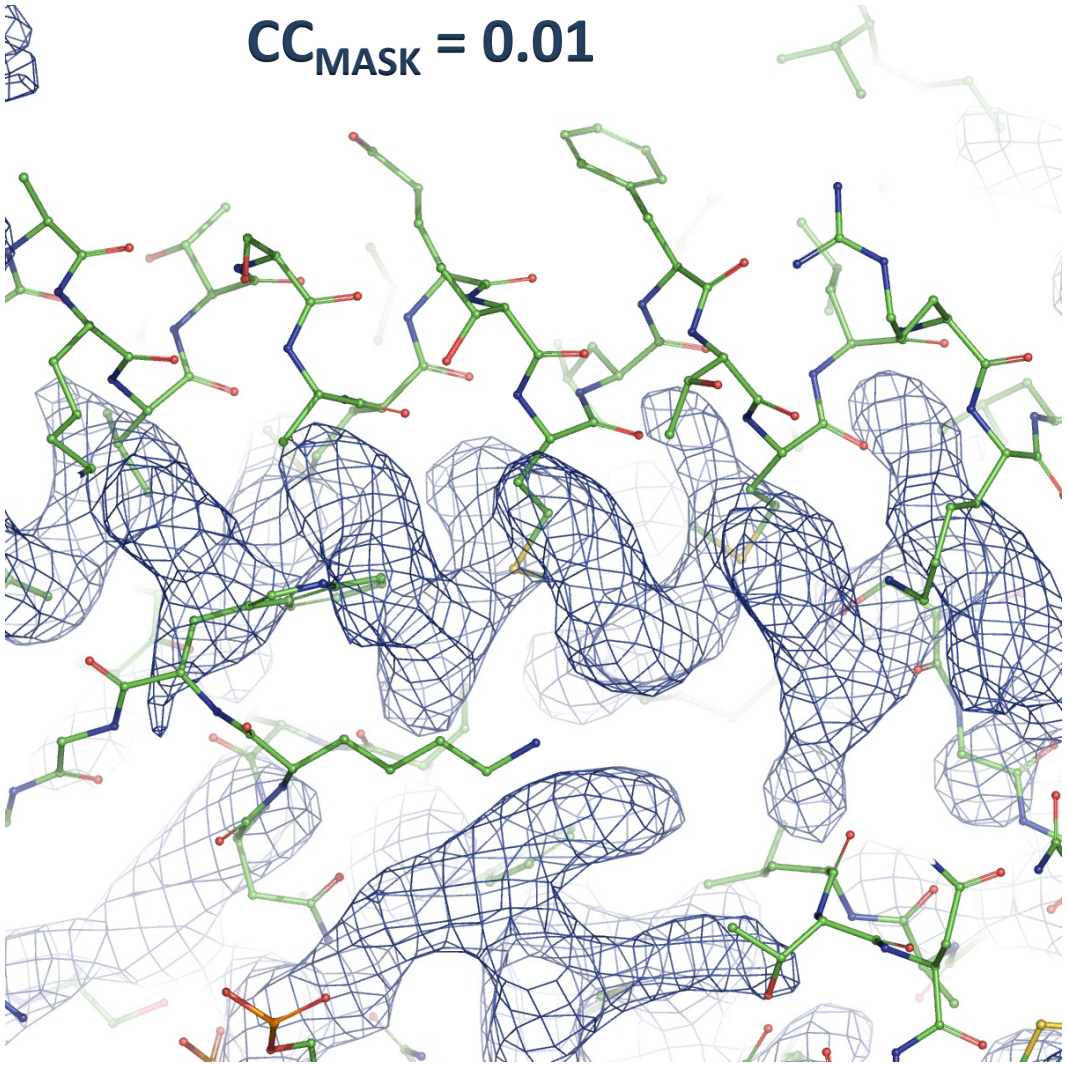
Despite all efforts to popularize the validation in recent years, poorly scoring models are still getting into databases now

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 |

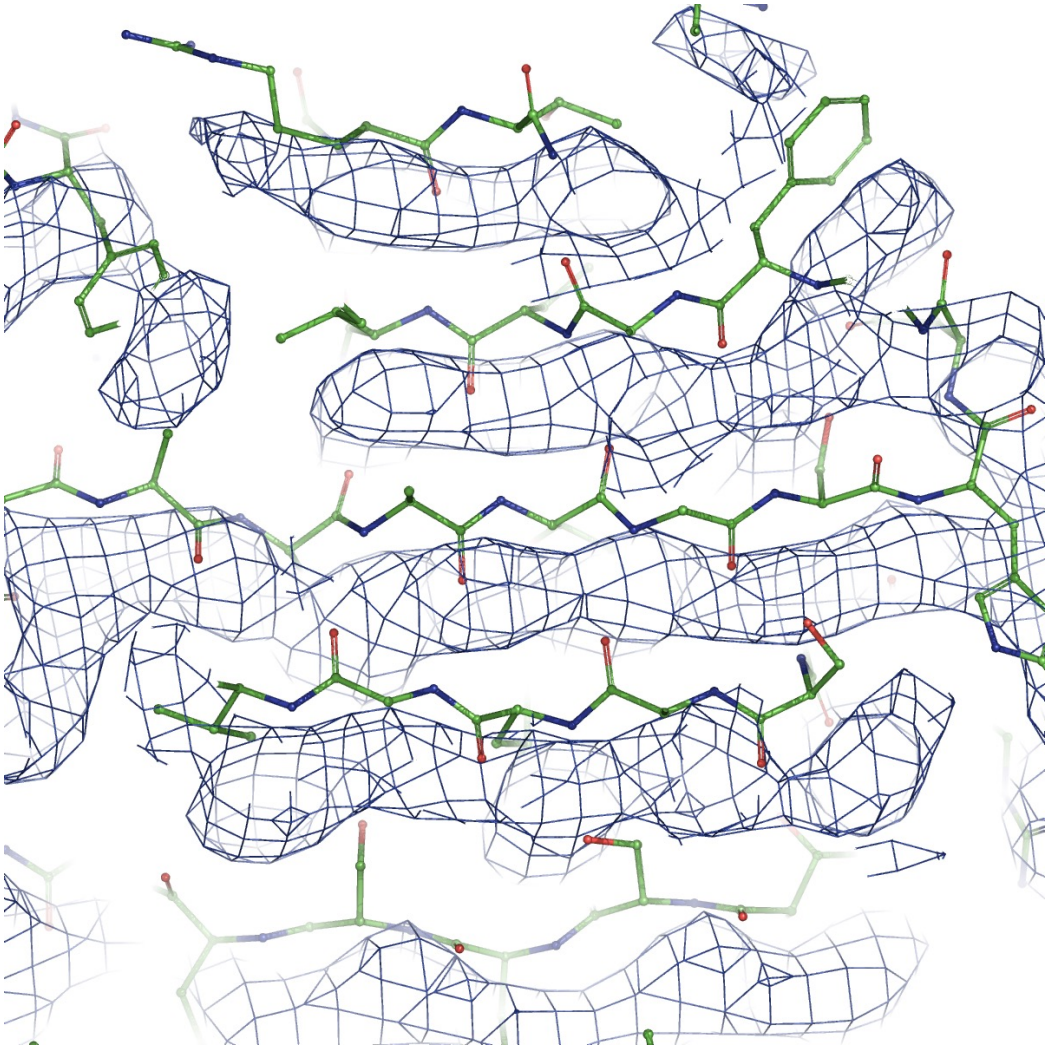


# Model does not fit the map

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**

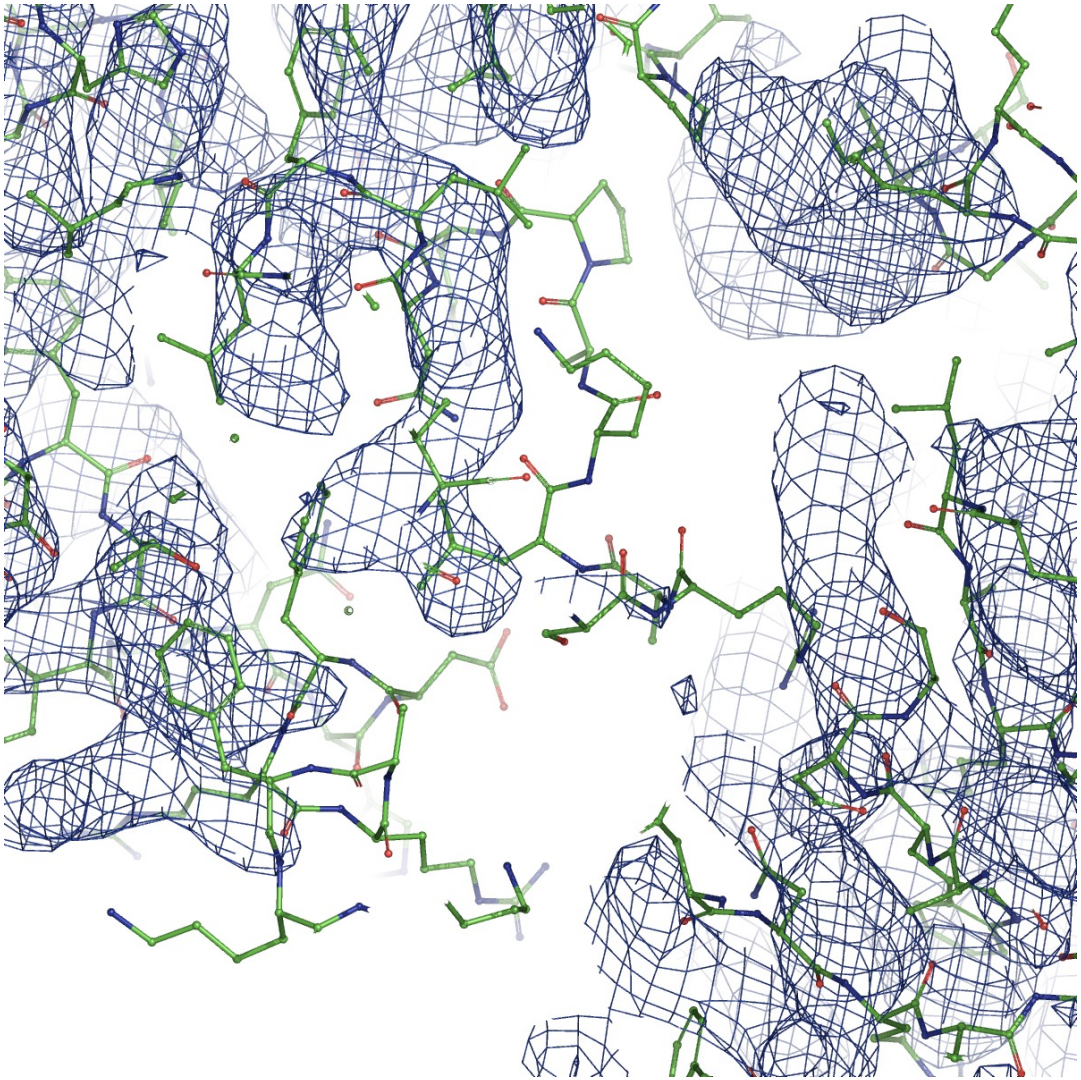


# Model does not fit the map

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 |

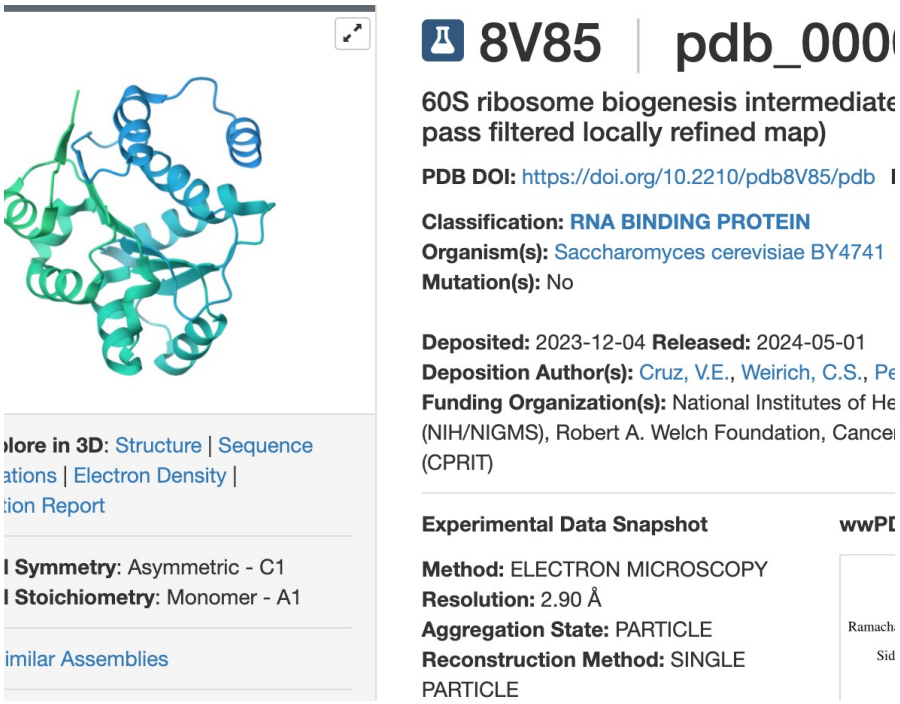
$CC_{MASK}$ = 0.1



# Model does not fit the map

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

$CC_{MASK}$ = 0.15



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)

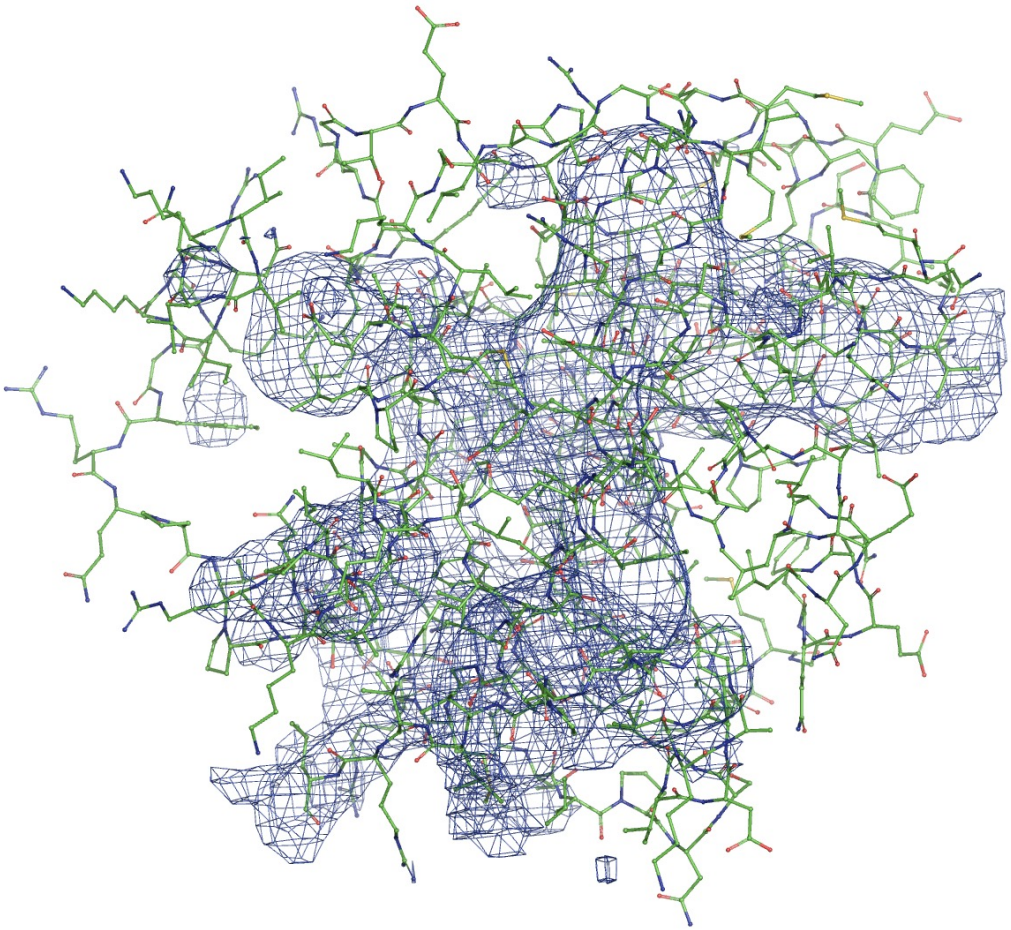
**Experimental Data Snapshot** **wwPD**

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

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

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

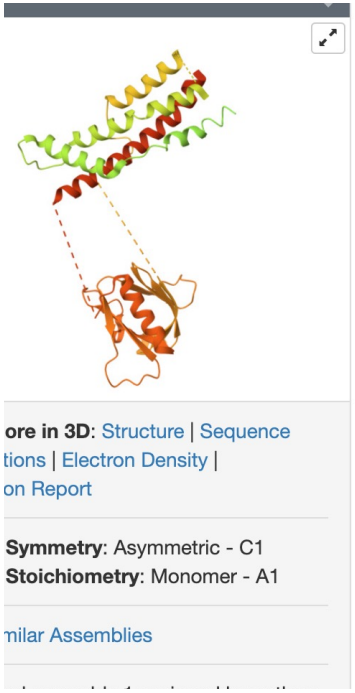
imilar Assemblies



# 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

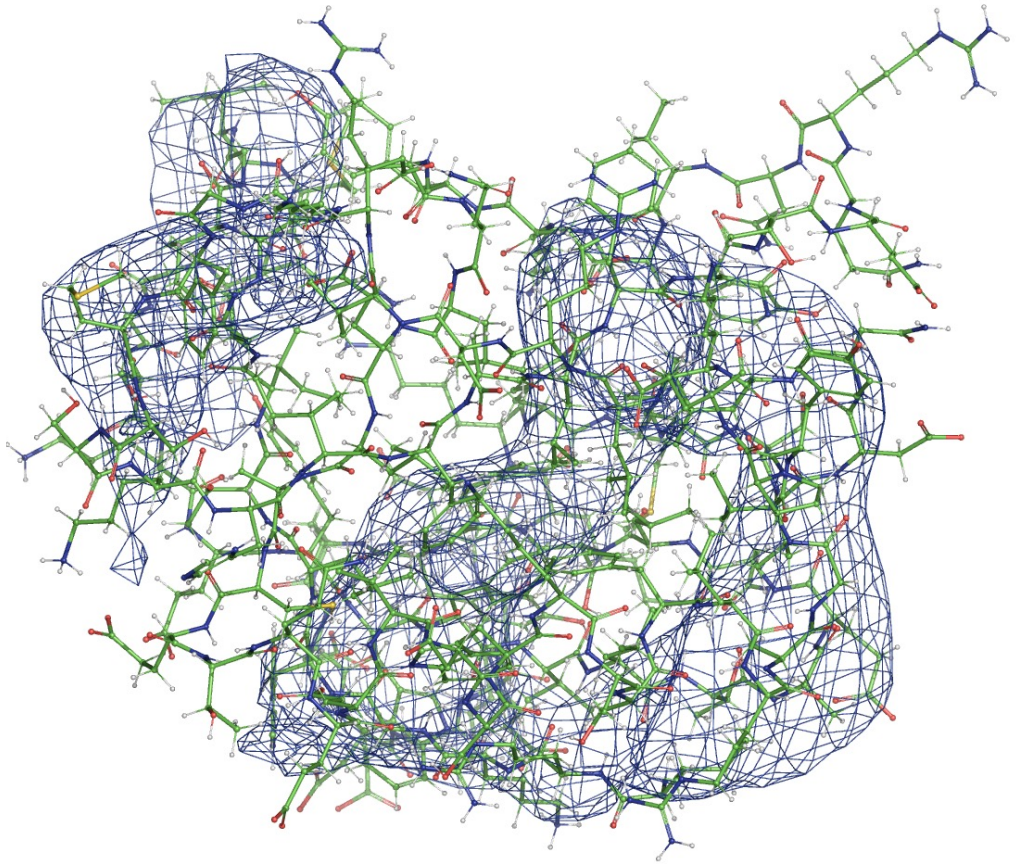
**Classification:** HYDROLASE
**Organism(s):** Homo sapiens
**Expression System:** Escherichia coli 'BL21-Go
**Mutation(s):** Yes

**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

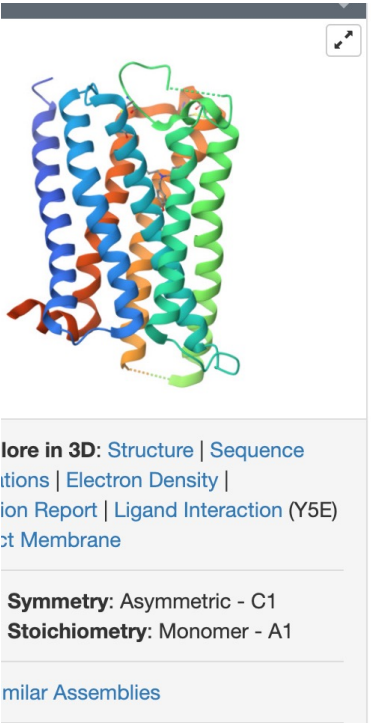
Ran



# 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

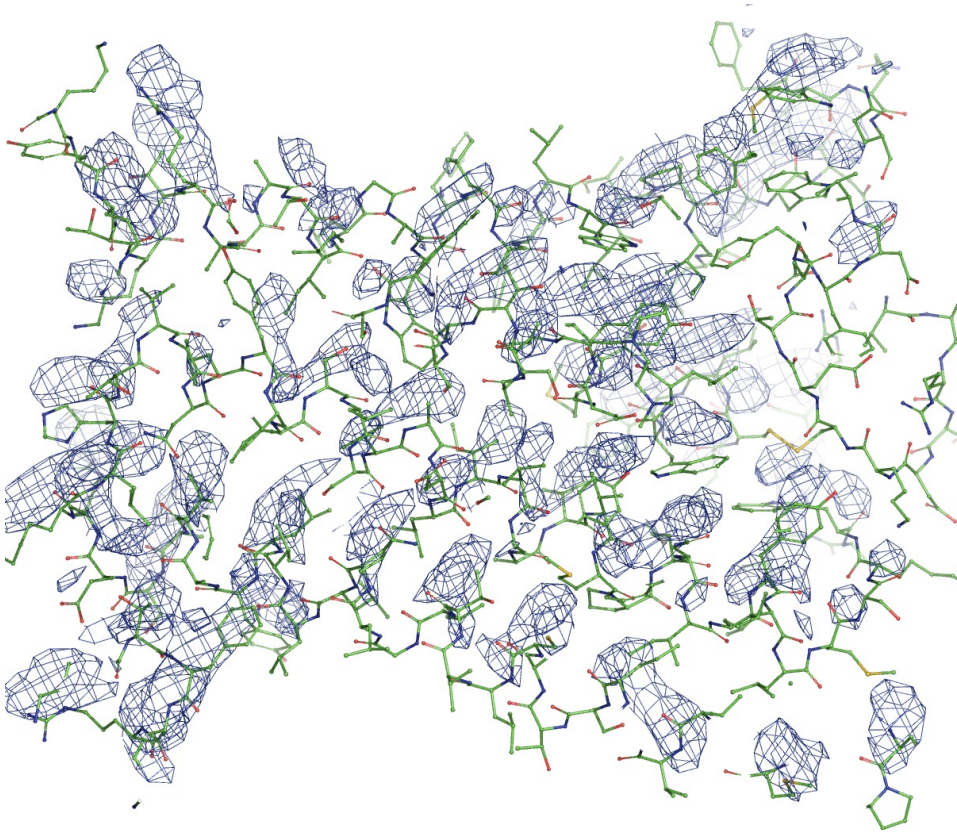
**Deposited:** 2023-11-20 **Released:** 2024-01-17
**Deposition Author(s):** Wang, D.D., Guo, Q.
**Funding Organization(s):** National Natural Science Found

**Experimental Data Snapshot**

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

**wwPDB Vali**

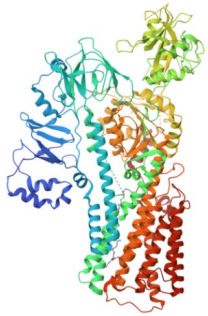
Clash
Ramachandran ou
Sidechain ou



# 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**

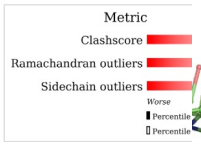
**Classification:** TRANSPORT PROTEIN
**Organism(s):** Homo sapiens
**Expression System:** Homo sapiens
**Mutation(s):** No
**Membrane Protein:** Yes PDBTM MemProtMD

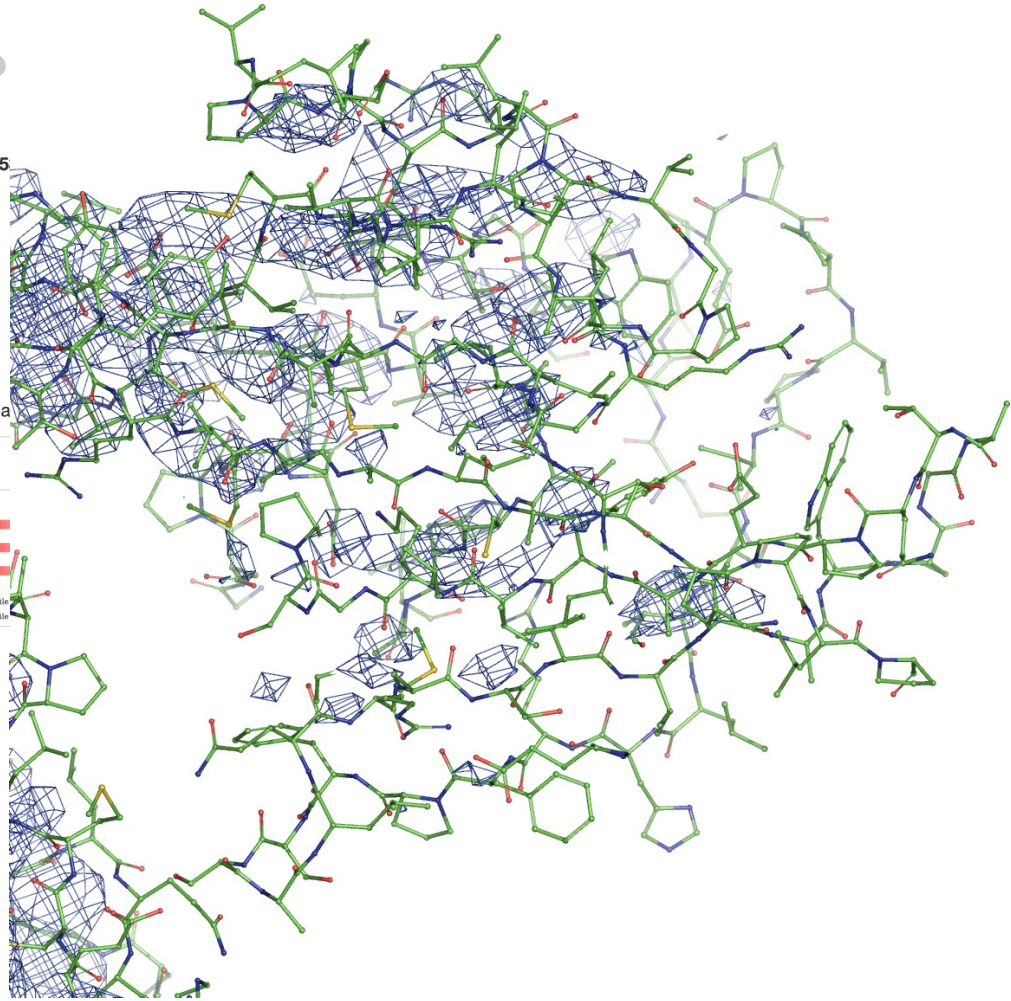
**Deposited:** 2023-02-15 **Released:** 2023-12-20
**Deposition Author(s):** Liu, Z.M., Mu, J.Q., Xue, C.Y.
**Funding Organization(s):** National Science Foundation (NSF, China

**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

nature communications

Explore content ~ About the journal ~ Publish with us ~

nature > nature communications > articles > article

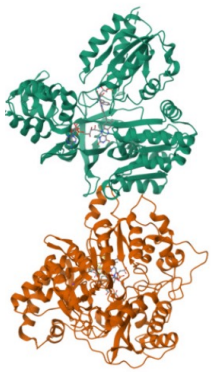
Article | Open access | Published: 26 March 2025

## Structure of dimerized assimilatory NADPH-dependent sulfite reductase reveals the minimal interface for diflavin reductase binding

Behrouz Ghazi Esfahani, Nidhi Walia, Kasahun Neselu, Yashika Garg, Mahira Aragon, Isabel Askenasy, Hui Alex Wei, Joshua H. Mendez & M. Elizabeth Stroupe

*Nature Communications* **16**, Article number: 2955 (2025) | Cite this article

**1343** Accesses | **1** Altmetric | Metrics



in 3D: Structure | Sequence
s | Electron Density |
Report |
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

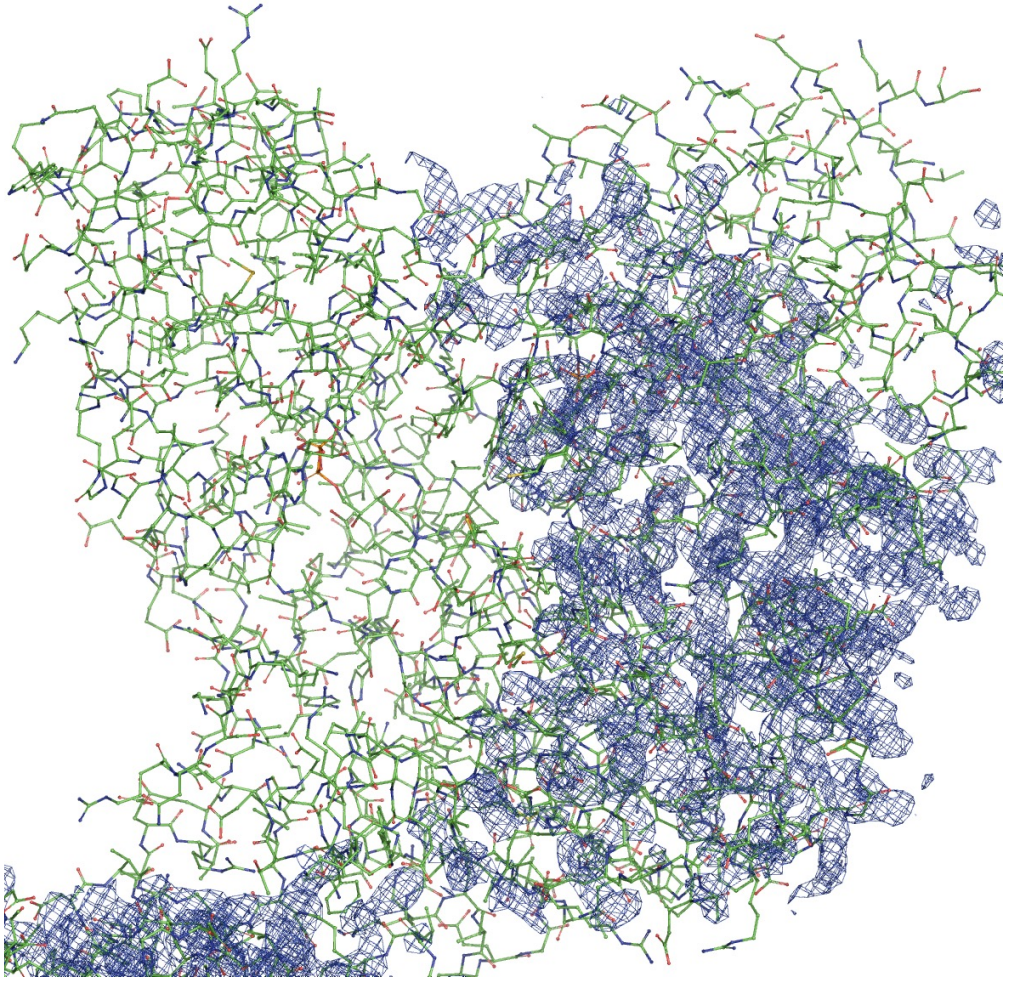
**Deposited:** 2024-06-13 **Released:** 2025-02-12
**Deposition Author(s):** Ghazi Esfahani, B., Walia
Mendez, J.H., Stroupe, M.E.
**Funding Organization(s):** National Science Fou

**Experimental Data Snapshot** ww

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

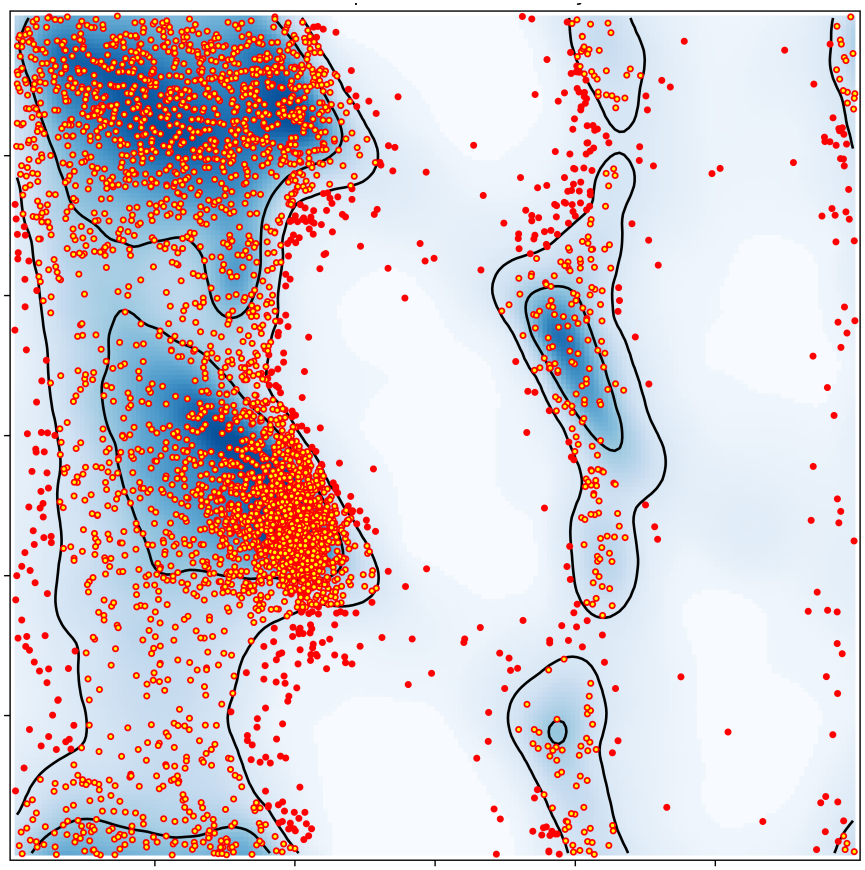
Ran

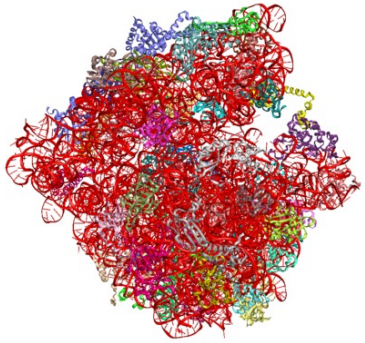
$CC_{MASK}$ = 0.0



# Poor model geometry

**(2019) Nature 570: 400-404 | PDB: 6o9j | EMDB: 0661 | 3.9Å**





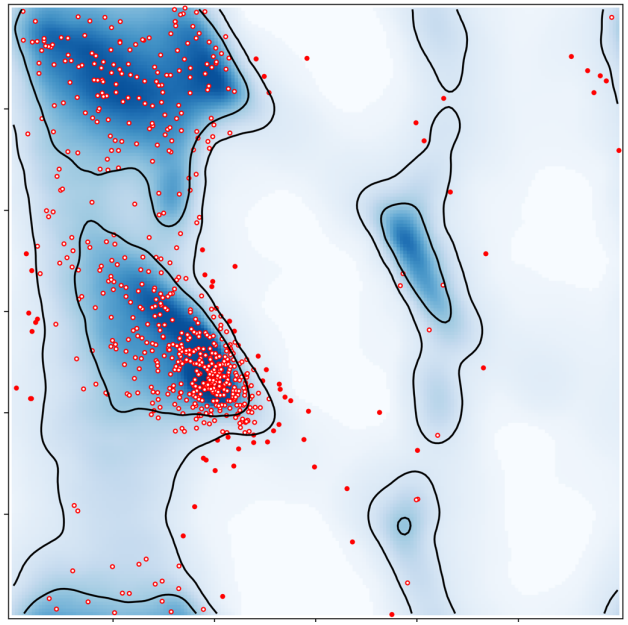
| Metric | 6o9j | Expected |
|---|---|---|
| Clashscore | 70 | Less than 10 |
| Ramachandran favored, % | 59 | More than 98 |
| Ramachandran outliers, % | 15 | 0 |
| Rotamer outliers, % | 23 | 0 |
| $C_\beta$ deviations, % | 0.5 | 0 |

# Poor model geometry

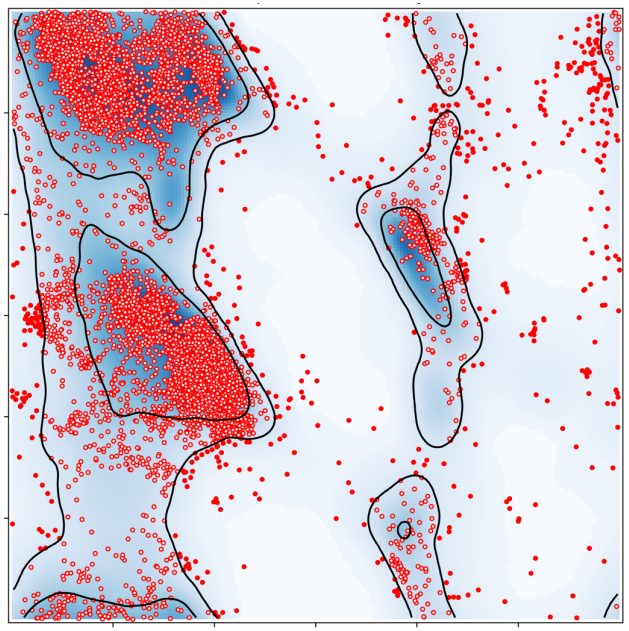
| | 6kio | 6kiq | 6o9j | 7ase |
|---|---|---|---|---|
| Bond/angle | 0.04/3.4 | 0.04/3.7 | 0.01/1.3 | 0.02/2.2 |
| Clashscore | 11 | 12 | 55 | 9 |
| Rotamer outl., % | 8 | 15 | 23 | 3 |
| Cb deviations, % | 5 | 16 | 0.5 | 1.4 |
| Ramachandran, % | | | | |
| favored | 74 | 70 | 59 | 79 |
| outliers | 7 | 11 | 15 | 7 |
| Resolution (Å) | **3.9** | **3.6** | **3.9** | **3.3** |
| Published in | Nature Comm. | Nature Comm. | Nature | Cell |
| Year | 2020 | 2020 | 2019 | 2020 |

# Poor model geometry

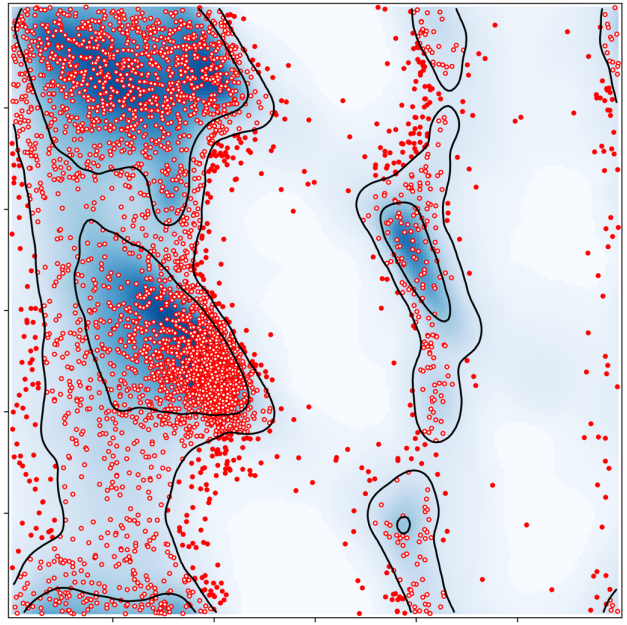
6kiq



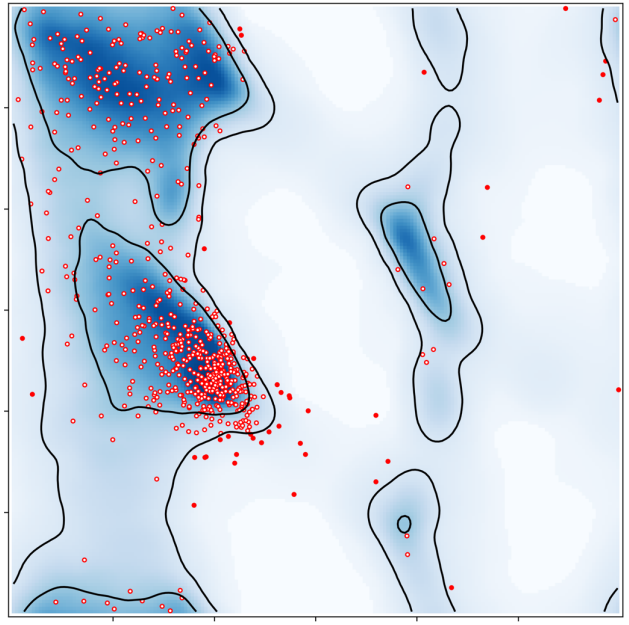
7ase



6o9j



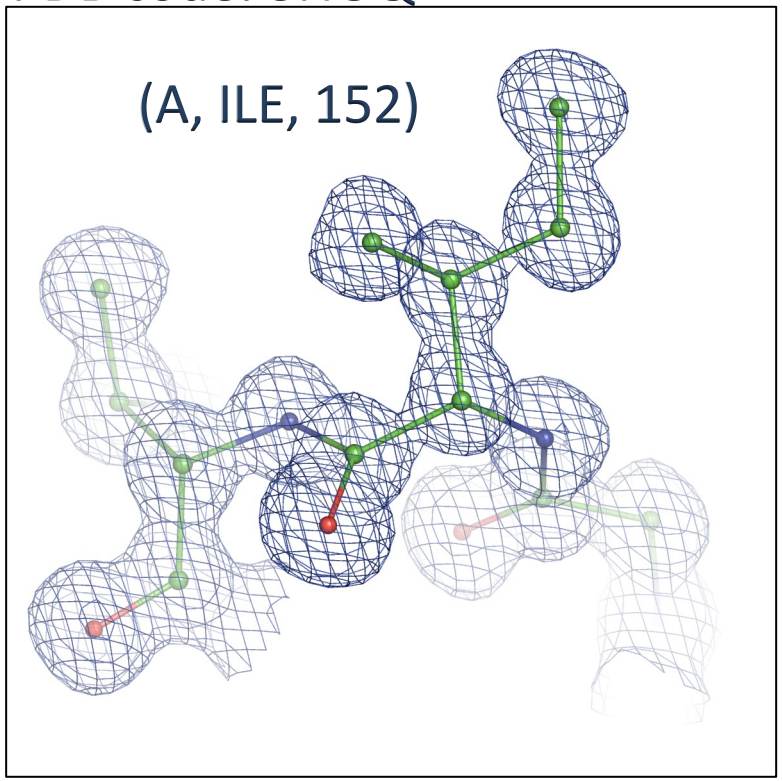
6kio

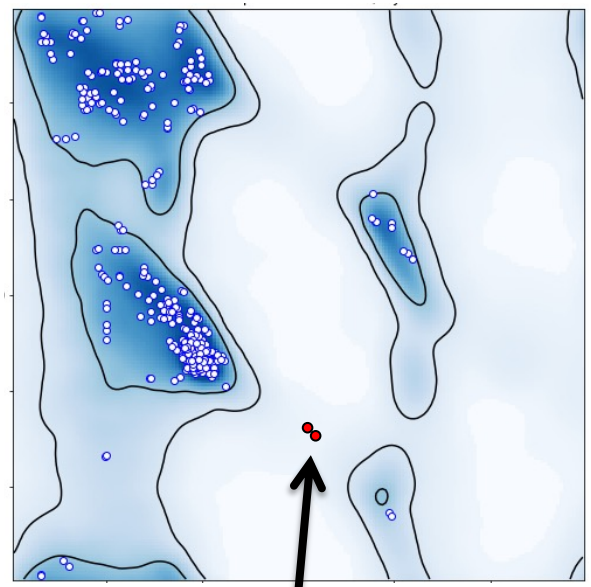


# Validation: *outliers are not always wrong*

- A Ramachandran plot outlier ≠ wrong

PDB code: 3NOQ



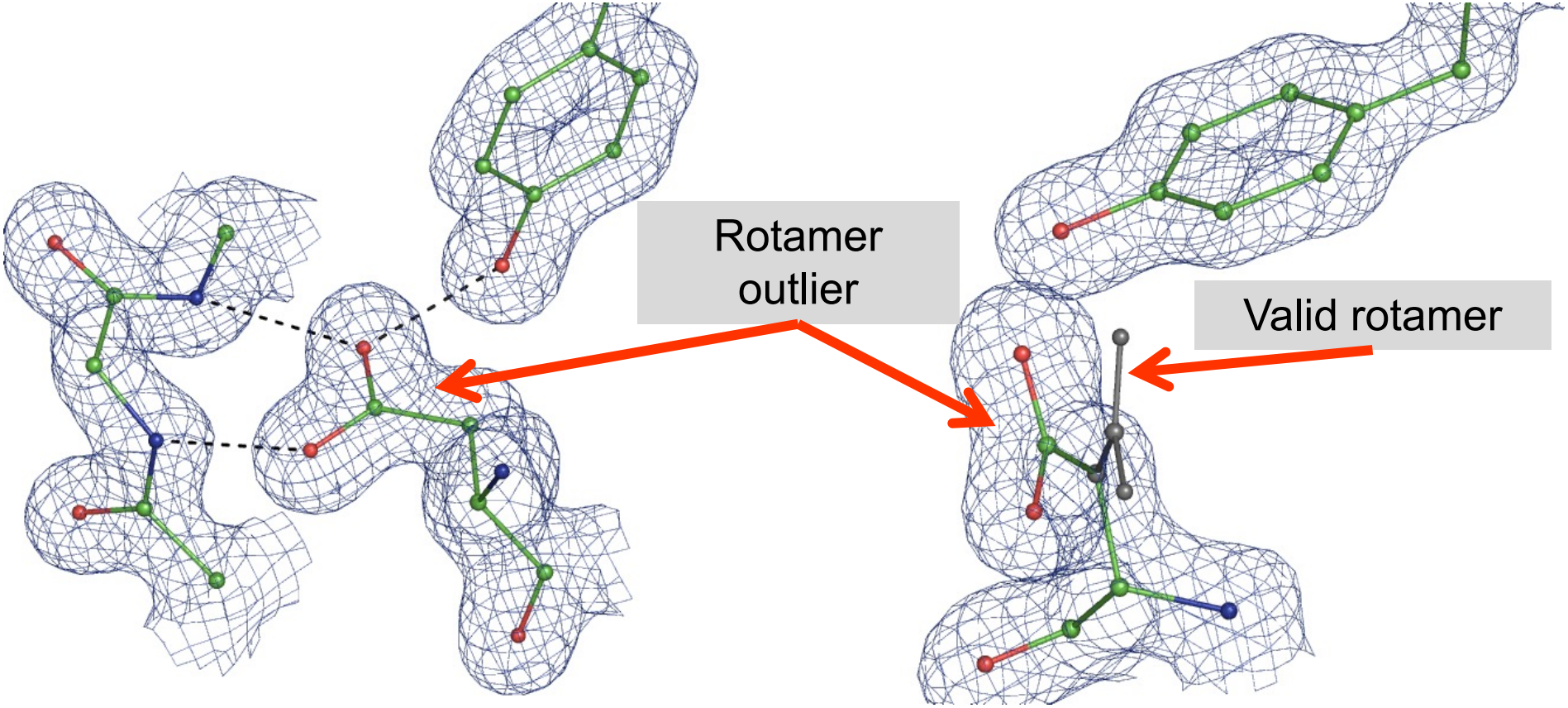
Outliers:

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

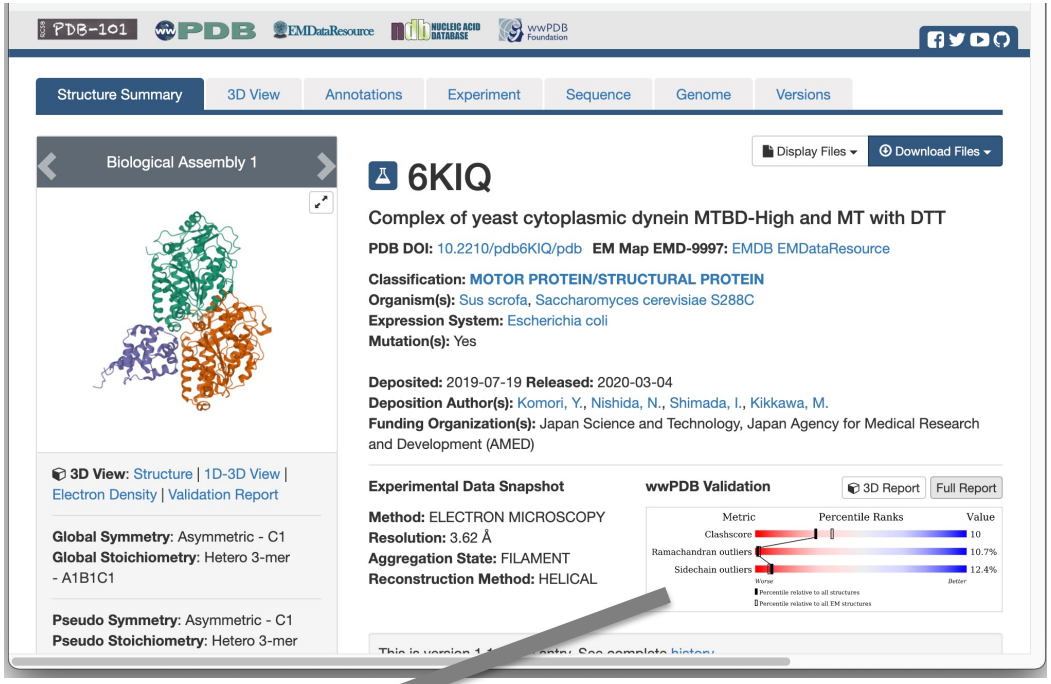
- All outliers need to be explained (supported by the data)

# Validation: *outliers are not always wrong*

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

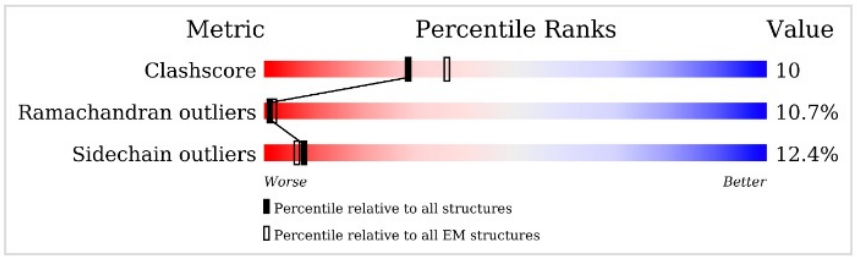


# Validation reports (RCSB)






**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

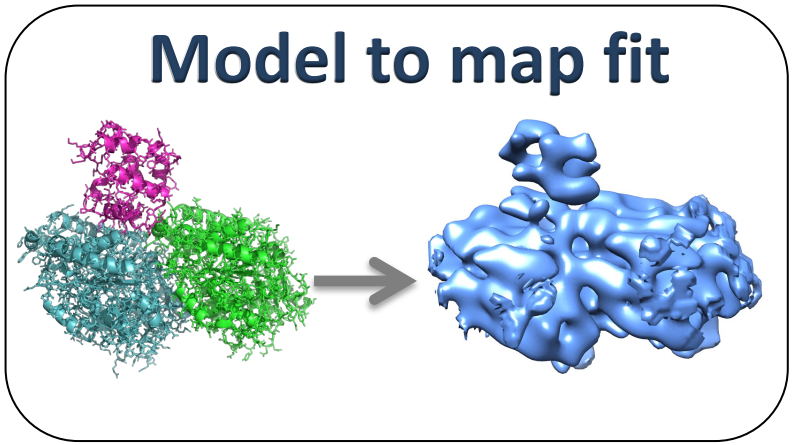
# Q-Score

**Q-score**:

- Measures the resolvability of atoms in a cryo-EM map, using an atomic model fitted to or built into the map
- It does not (at least directly) measure model-to-map fit

- 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: $CC_{MASK}$



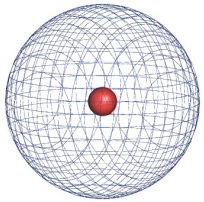

$$CC_{MASK} = \frac{\sum \rho_{obs}\, \rho_{calc}}{(\sum \rho_{obs}^{2}\ \sum \rho_{calc}^{2})^{1/2}}$$

$\rho_{obs}$ = experimental map
$\rho_{calc}$ = 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

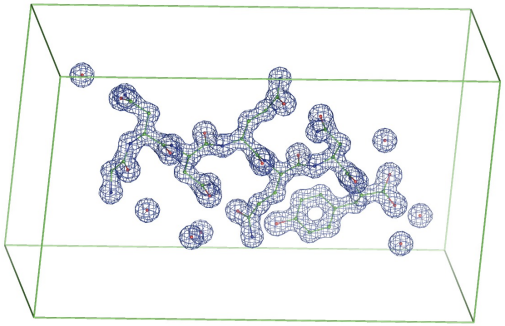
| Metric | Expected value |
|---|---|
| $CC_{MASK}$ | Poor: < 0.3<br>So-so: 0.3-0.6<br>Good: > 0.6 |

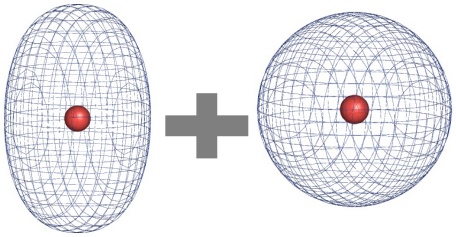
# Model map: Independent Atom Model (IAM)



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

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

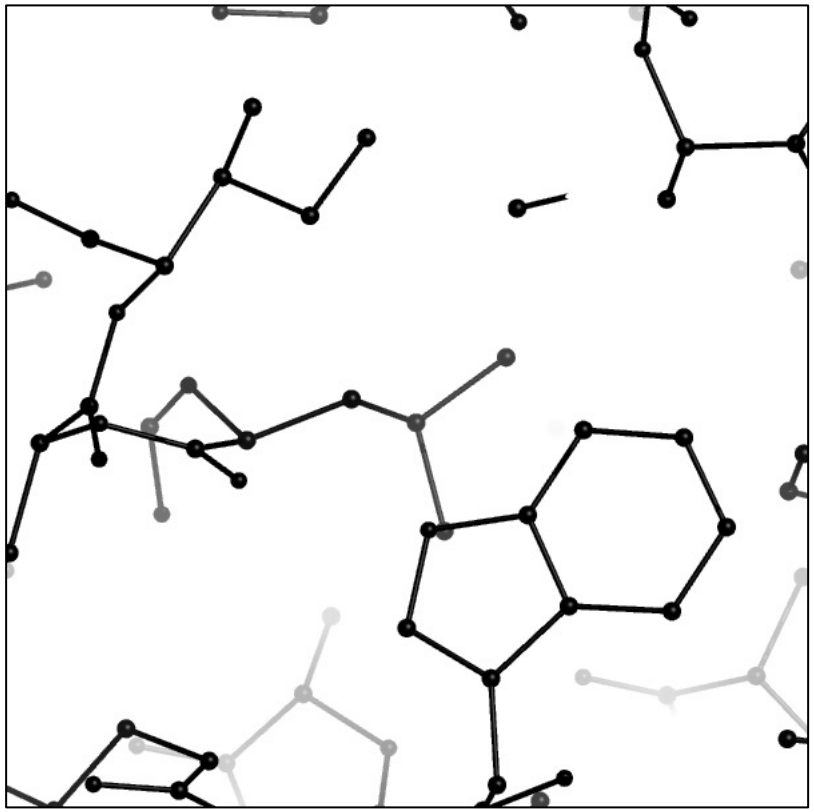




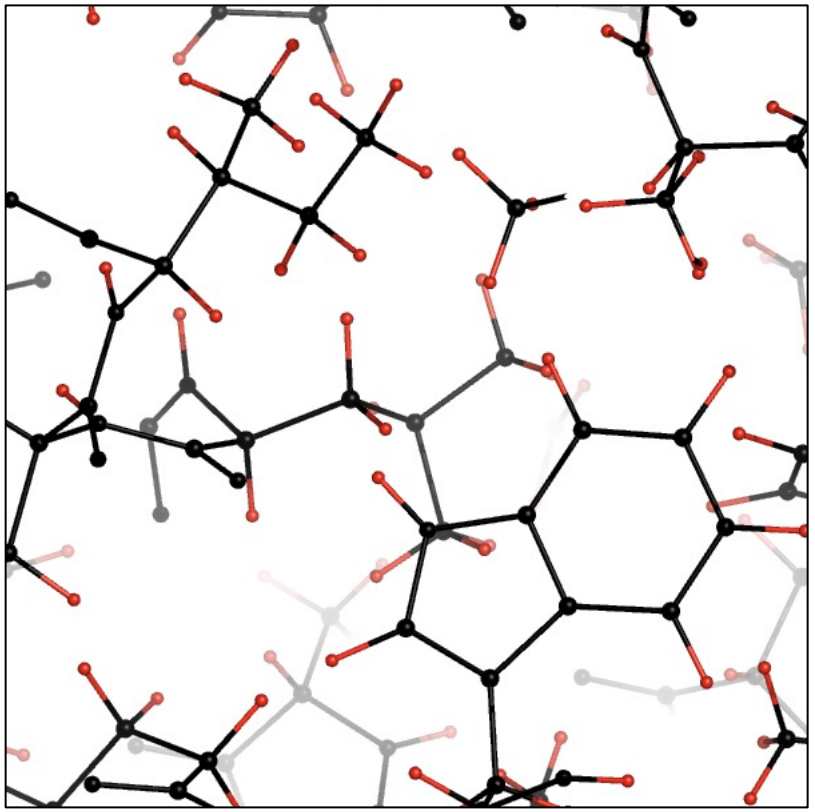
$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

# Hydrogen atoms

- H: half of the atoms in a protein molecule
- Make most interatomic contacts
  - Using H in refinement helps prevent or eliminate clashes



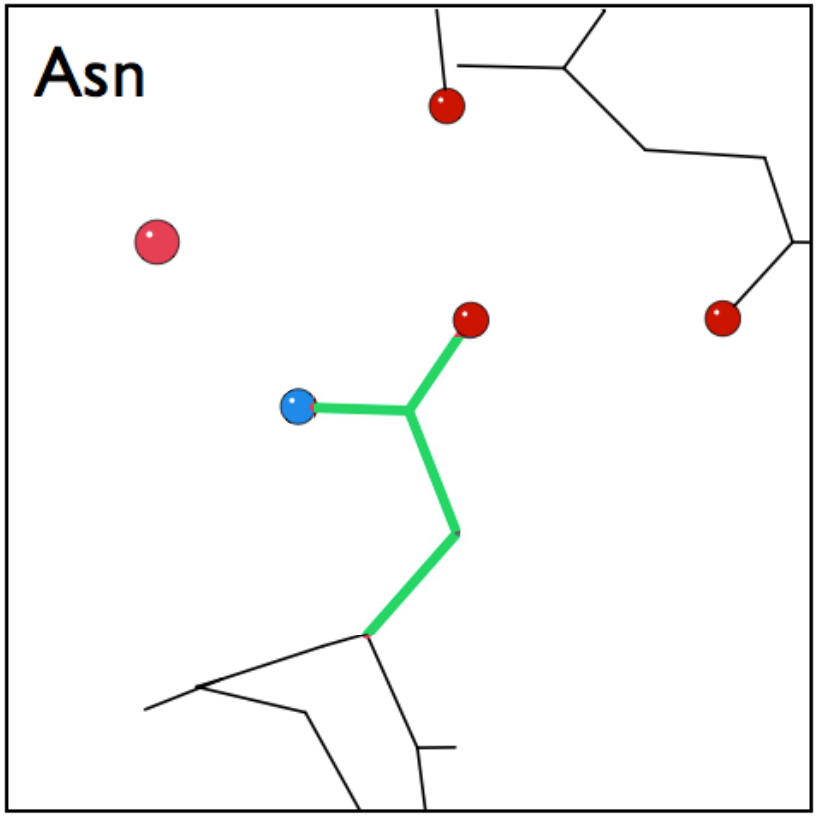
No H atoms

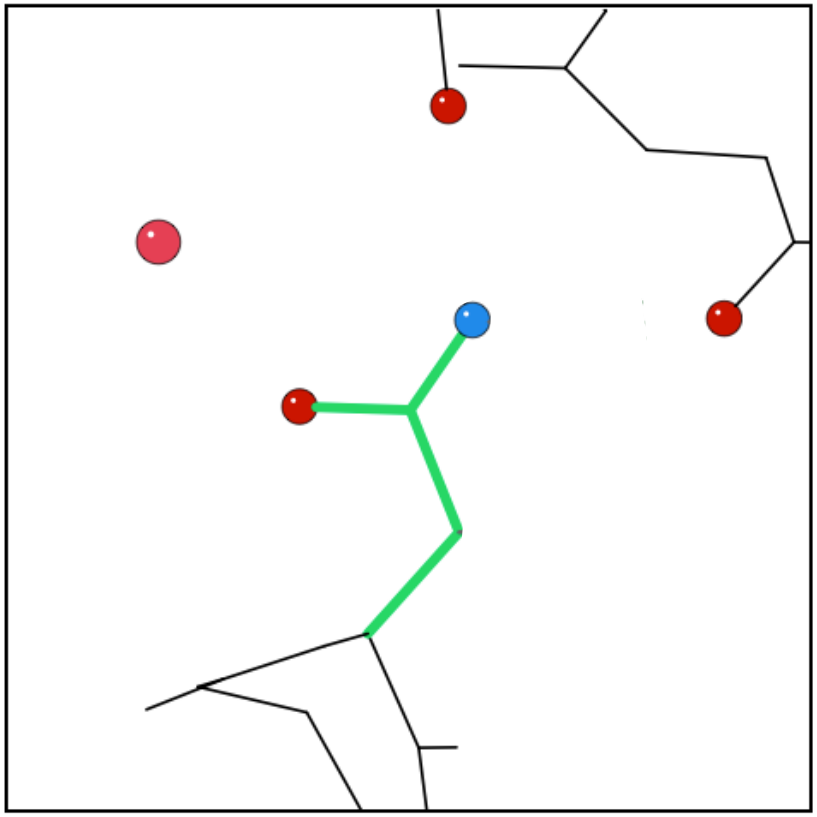


H atoms added

# Model validation: clashes

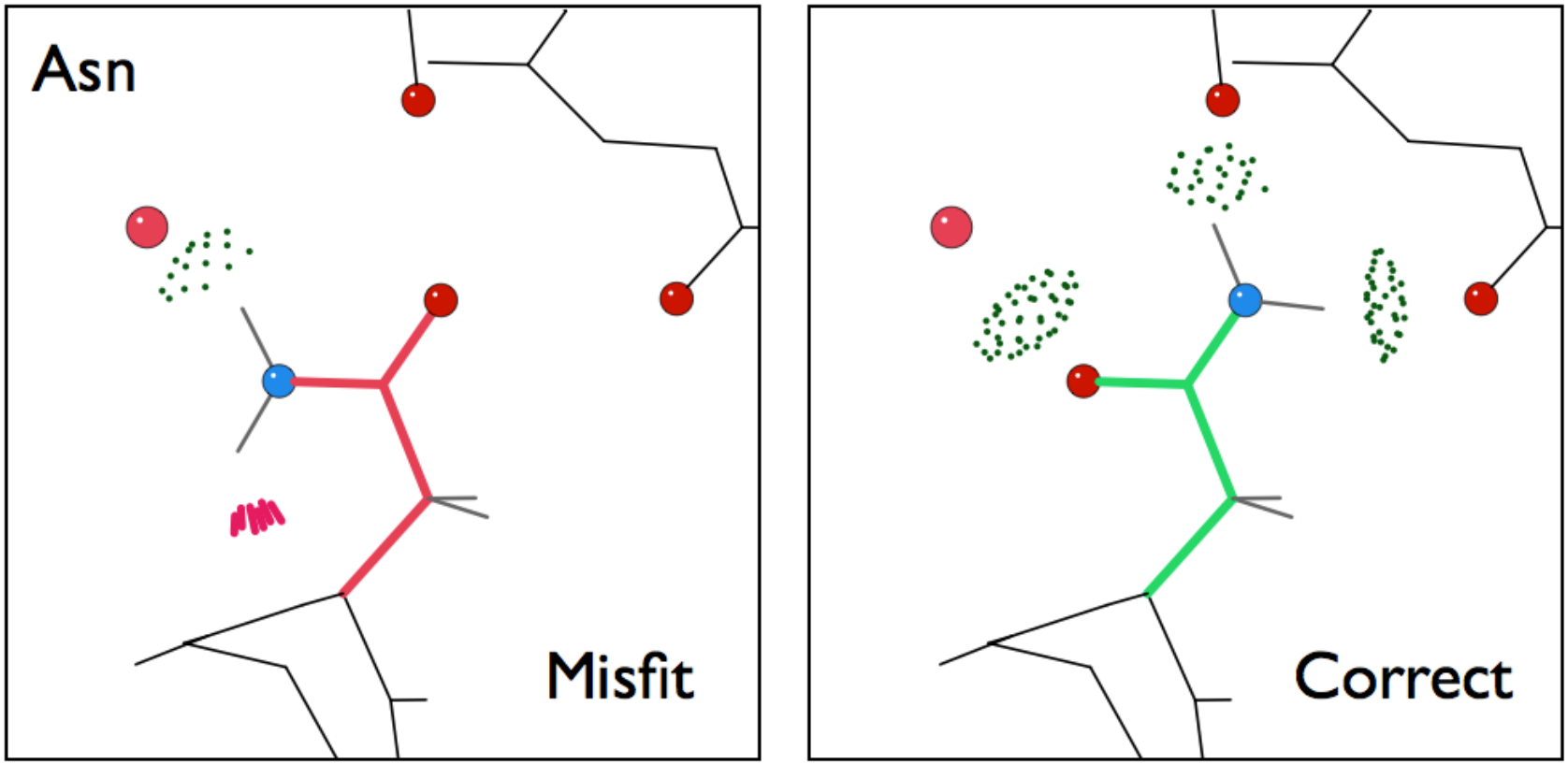
- N/Q/H flips (asparagine/glutamine/histidine)
  - Based on clash analysis
  - Requires H present



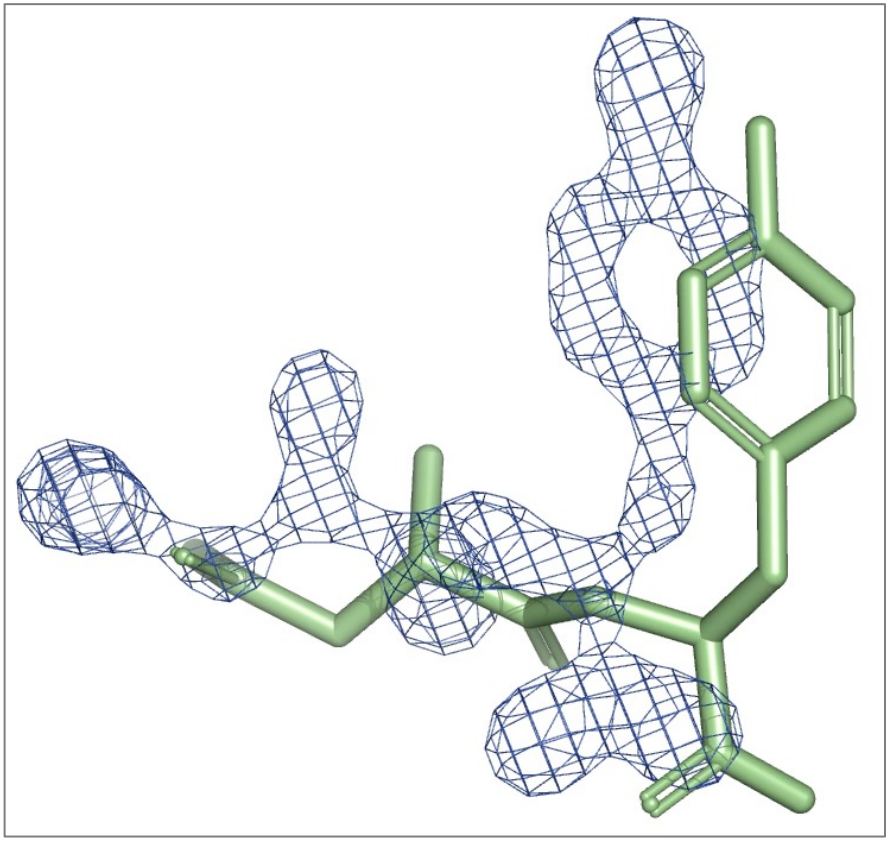
# Model validation: clashes

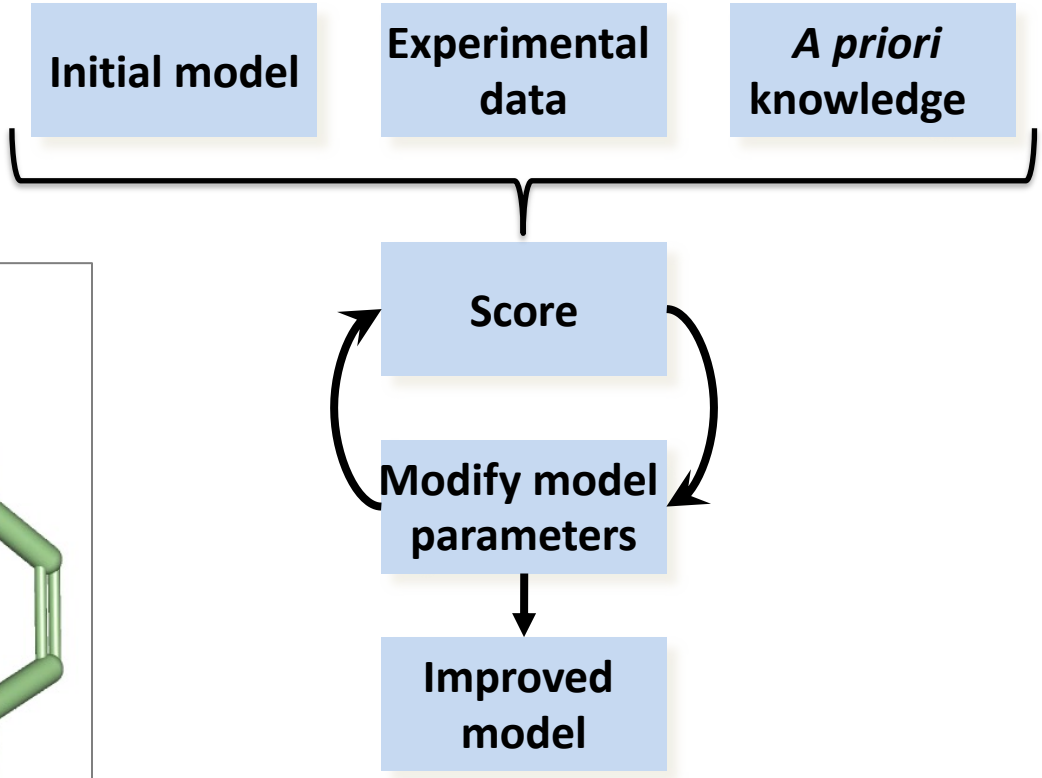
- N/Q/H flips
  - Based on clash analysis
  - Requires H present



# Validation and Refinement ”conflict”

# Model refinement



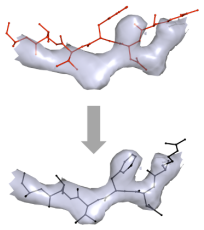


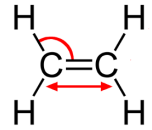

Optimization process of fitting atomic model parameters to experimental data

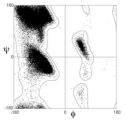
# Restraints and data resolution

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

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

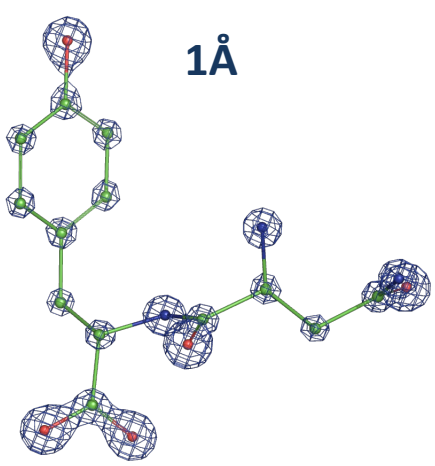


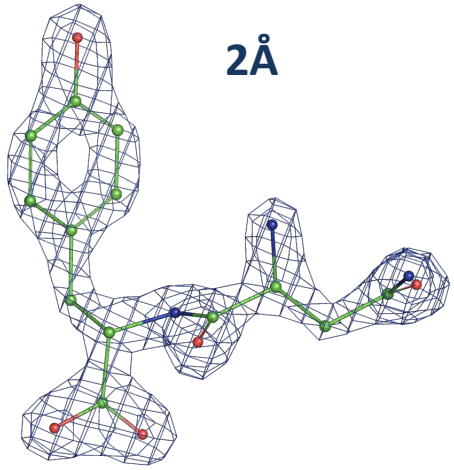


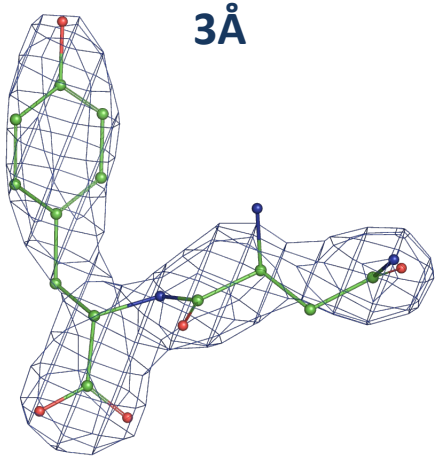



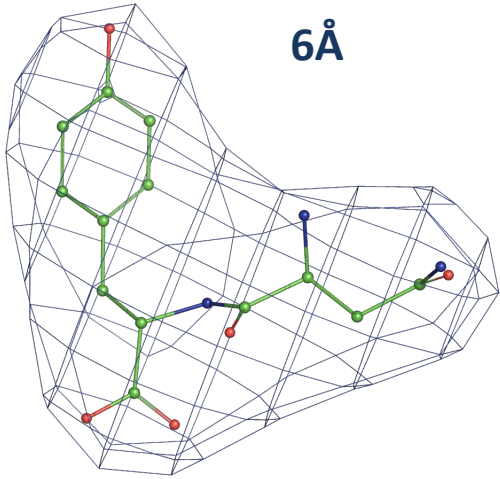
Bonds, angles, planes, torsions, chirality, non-bonded repulsion

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



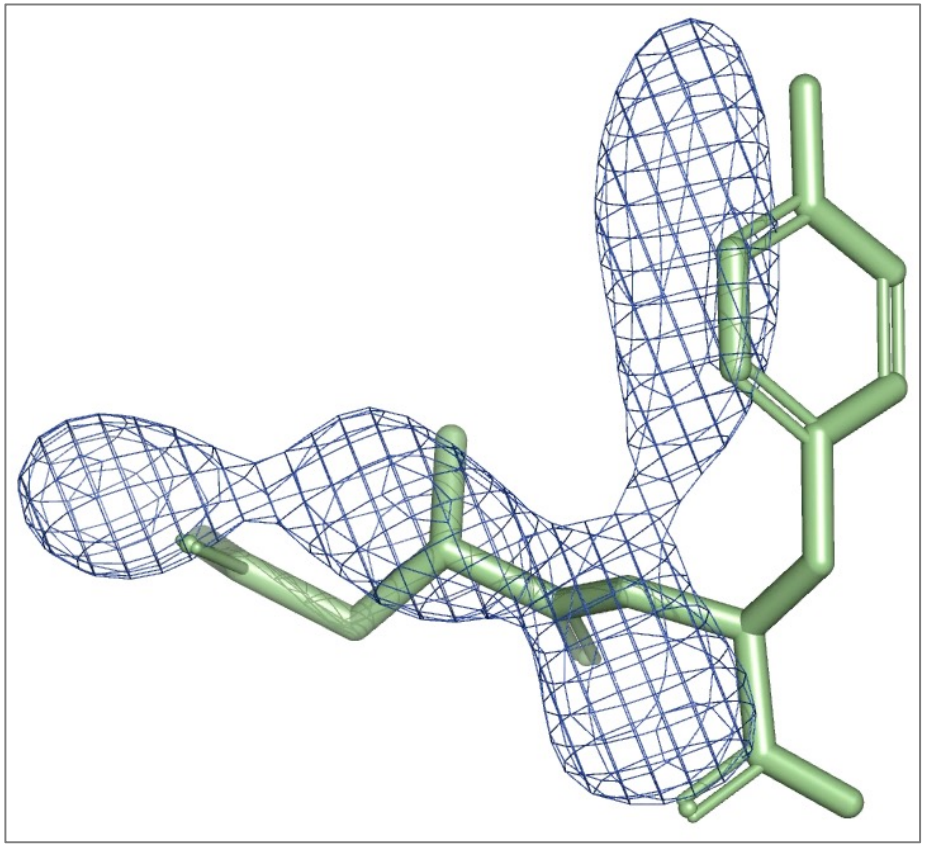



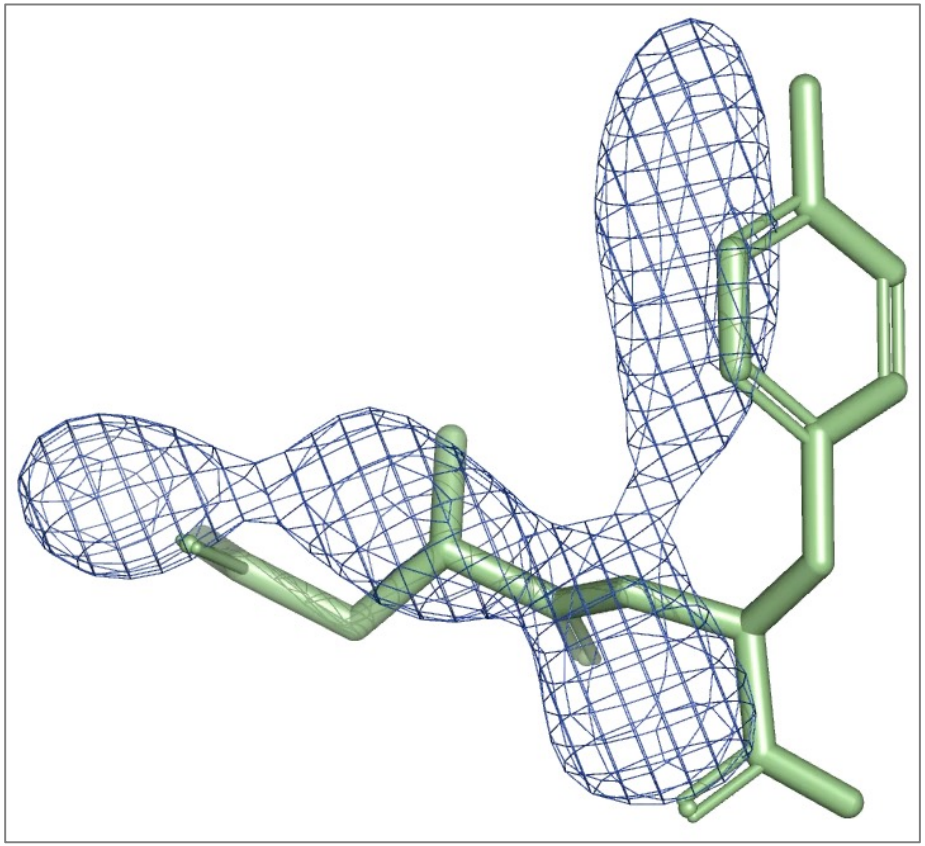

High Resolution Low

# Model refinement with *vs* no restraints

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



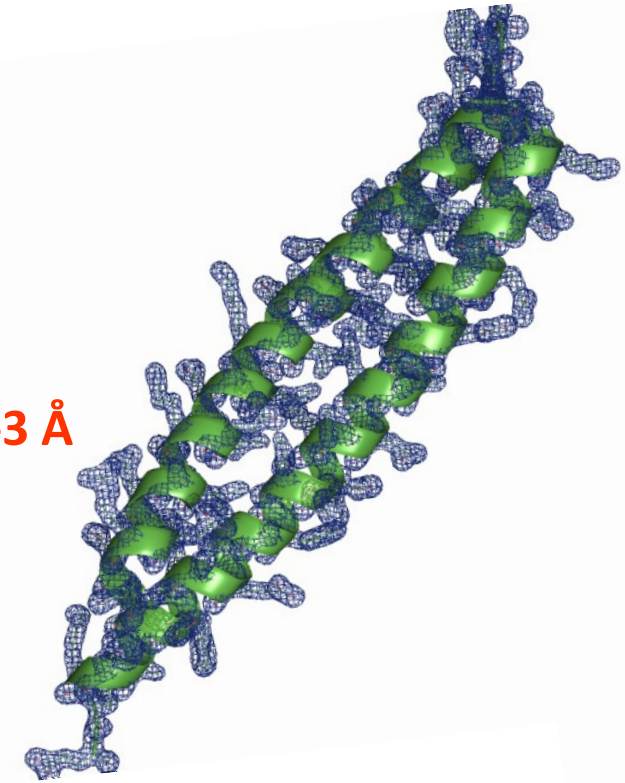
**Using** restraints



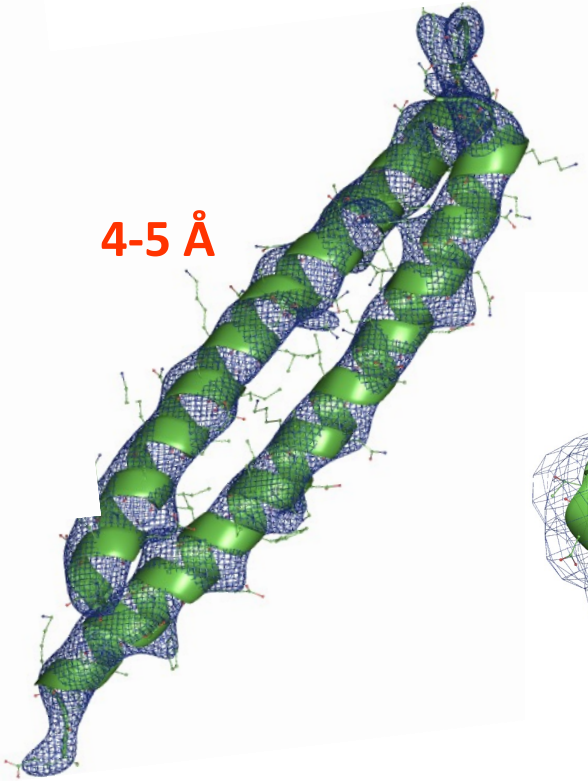
**Not using** restraints

# Restraints: low resolution

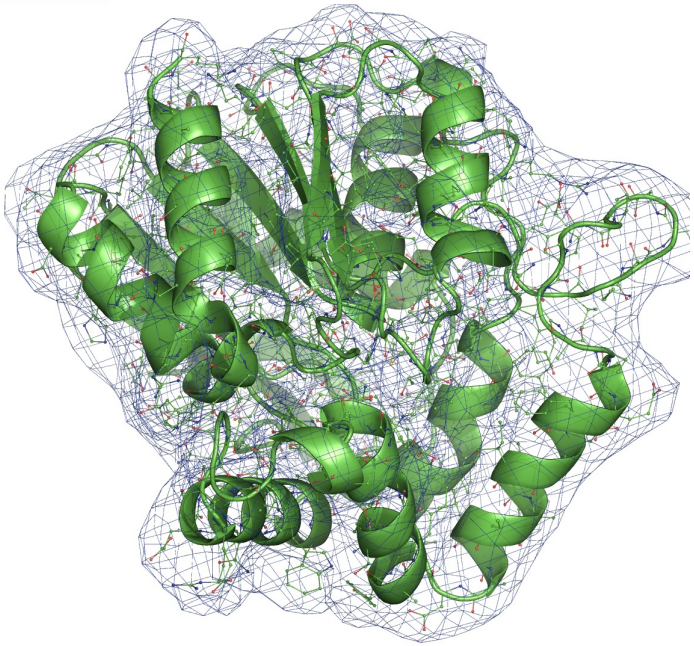
2-3 Å



4-5 Å

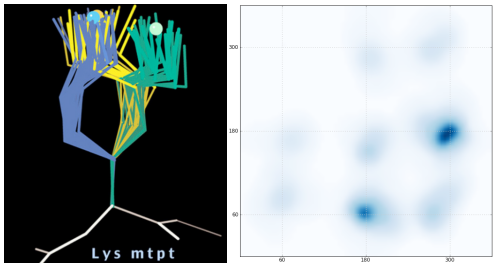


6Å-lower

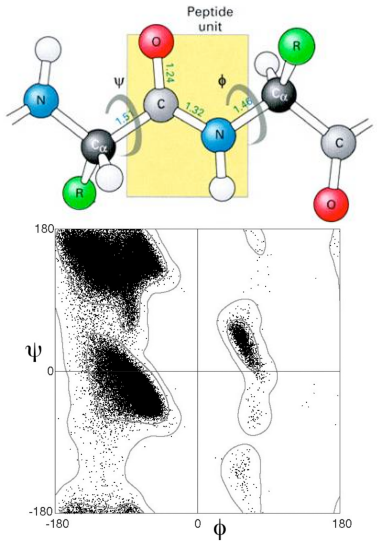


# More restraints for low resolution

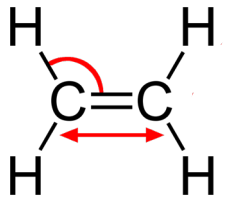
Side chain distributions



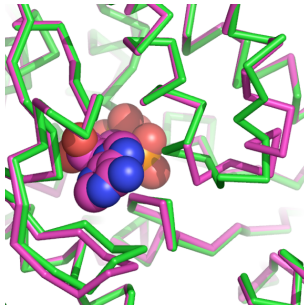

Main chain distributions



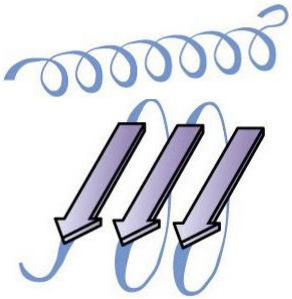

Covalent geometry




Internal symmetry (NCS)



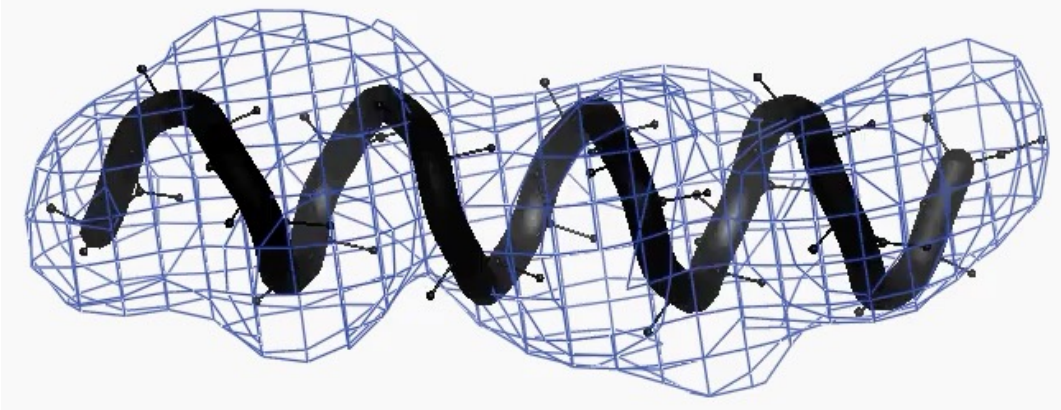
Similar (homologous) structures (reference model restraints)



Secondary structure

# Model refinement with insufficient restraints

- Refinement of a perfect α-helix into low-res map
  - Using simplistic (standard) restraints on covalent geometry
    - Model geometry deteriorates as result of refinement

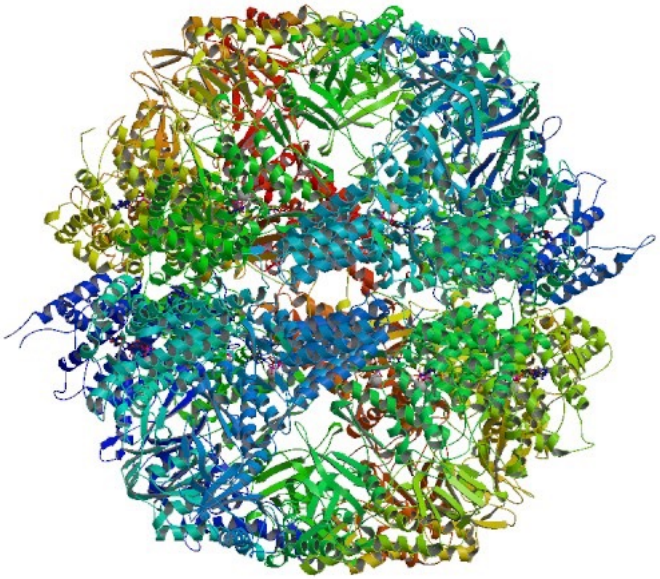


# Validation and Refinement "conflict"

- Validation metrics progressively become refinement goals
    - Ramachandran plot restraints
    - Cβ deviation restraints
    - Secondary structure restraints
    - Restraints on χ angles of amino-acid side-chain rotamers
- As result, validation becomes less capable of catching problems

# Example

## PNAS, 2019 116 (39) 19513-19522

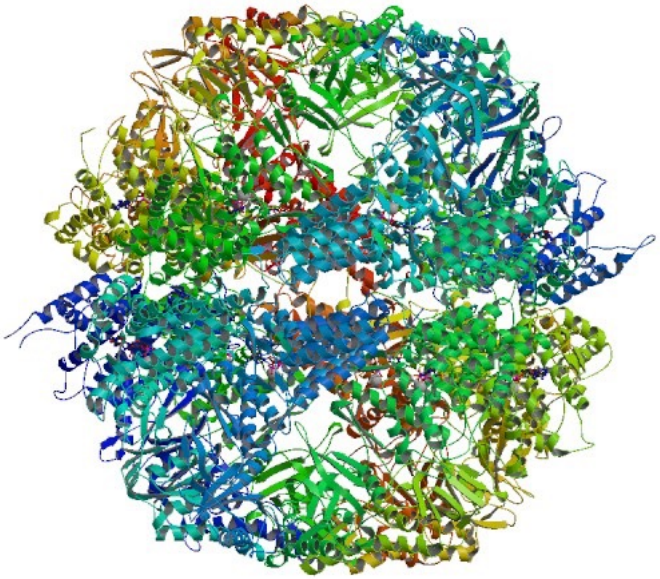


| Metric / PDB code | | 6KS6 |
|---|---|---|
| Clashscore | | 8.8 |
| Rama. (%) | favored | 96.4 |
| | outliers | 0.2 |
| Rotamer outliers (%) | | 0 |
| $C_\beta$ deviations | | 0 |
| RMSD | Bond (Å) | 0.002 |
| | Angle (°) | 490 |
| Resolution (Å) | | 3.0 |

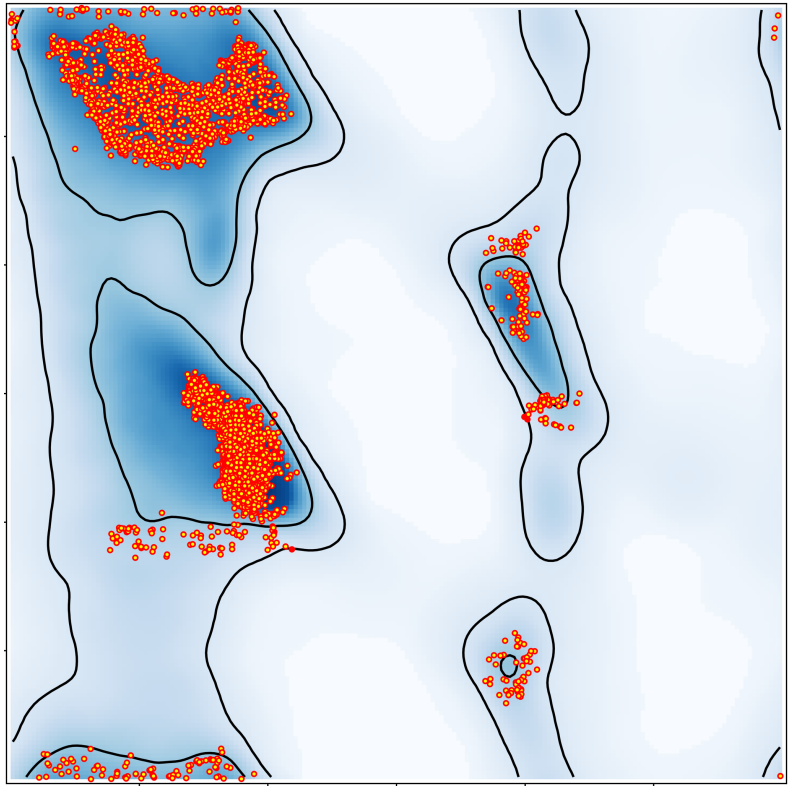
Perfect statistics! All looks just great!

# Example

## PNAS, 2019 116 (39) 19513-19522



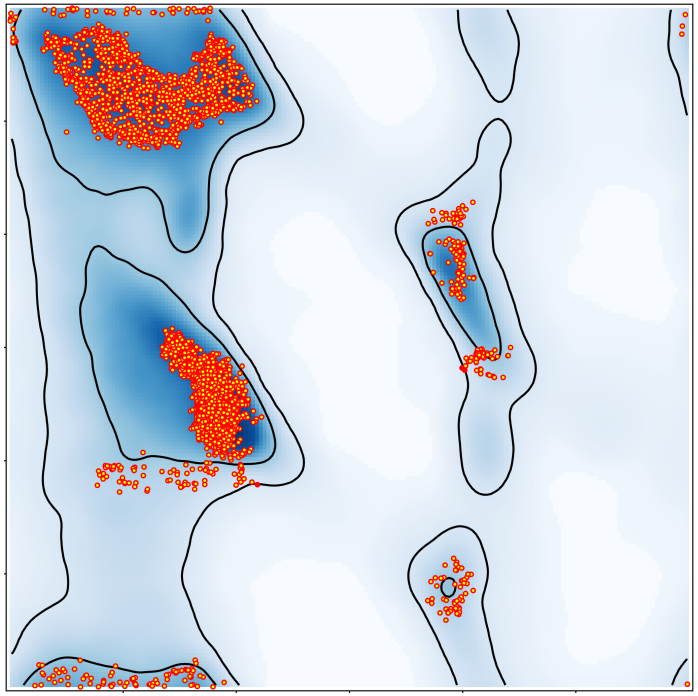
The plot looks very wrong!



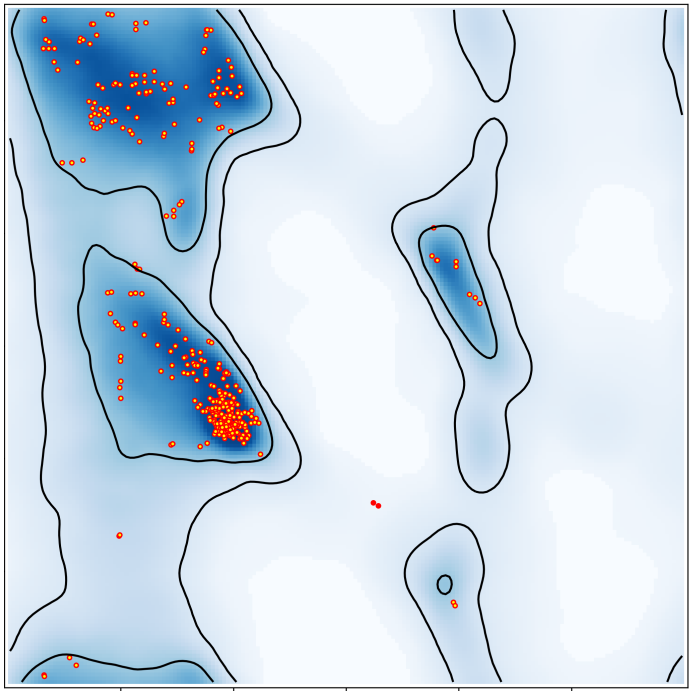
**1. How we know the plot looks wrong?**

**2. How did that happen?**

**Q:** How we know the plot looks wrong?



**A:** Because 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

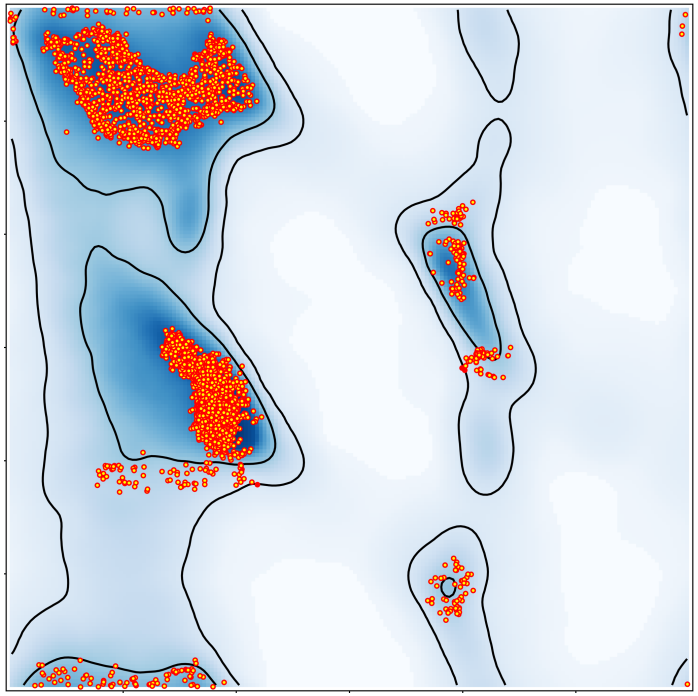
**A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry**

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]

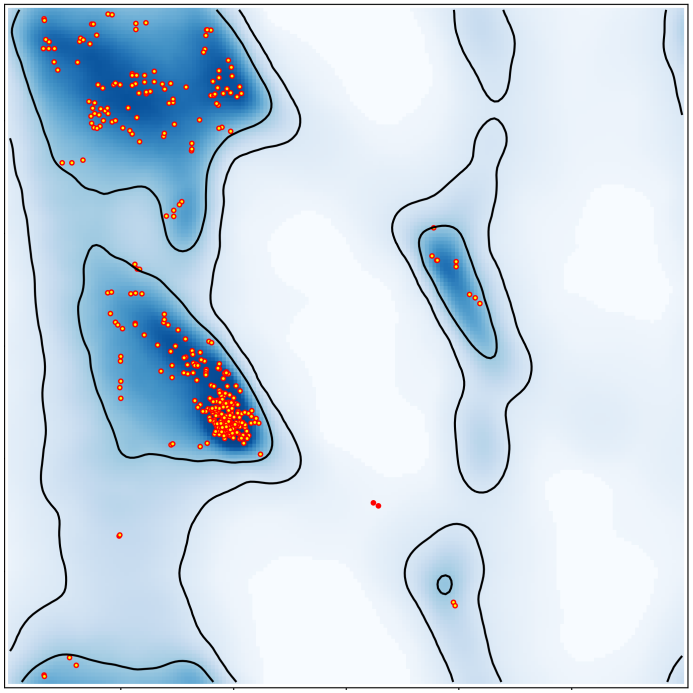
# Ramachandran plot Z-score

**Q:** How we know the plot looks wrong?

**A:** Because we know how good plot looks like!



RamaZ = -4.1 (Poor)

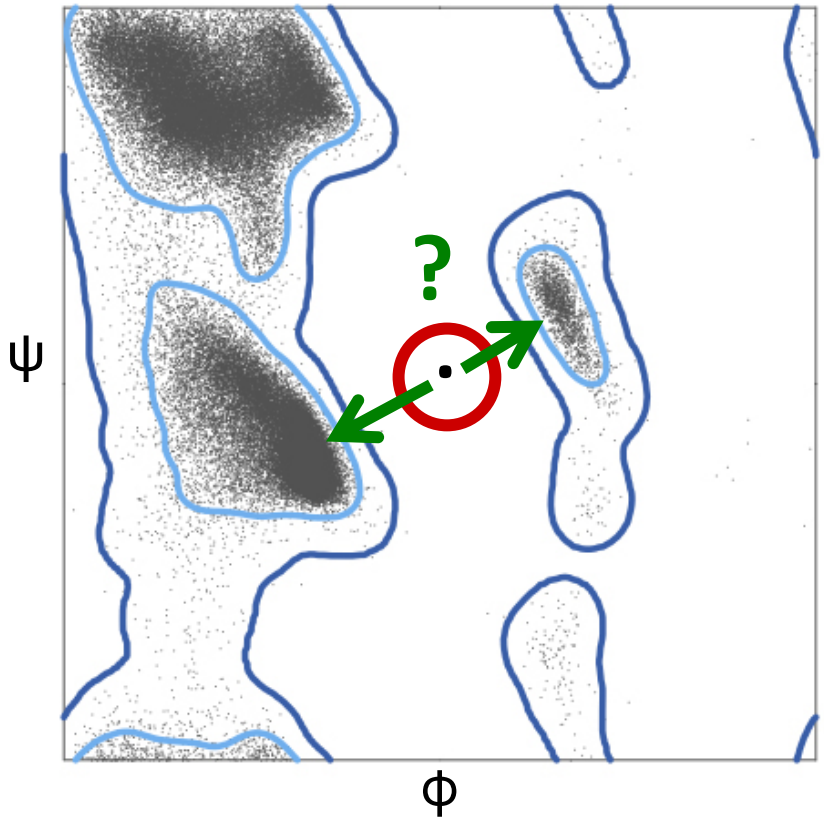


RamaZ = -1.9 (Good)

Poor: |Z| > 3  Suspicious: 2 < |Z| < 3  Good: |Z| < 2

# How did that happen?

$$E = \Sigma\, w * (\varphi_{model} - \varphi_{target})^2 + \Sigma\, w * (\psi_{model} - \psi_{target})^2$$

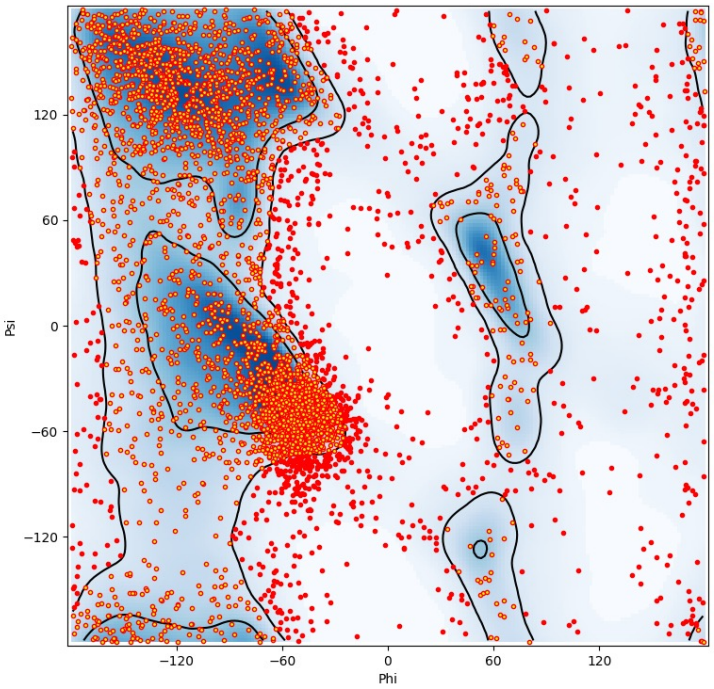



- Setting up Ramachandran plot, secondary structure, etc, restraints can be ambiguous and is error prone!

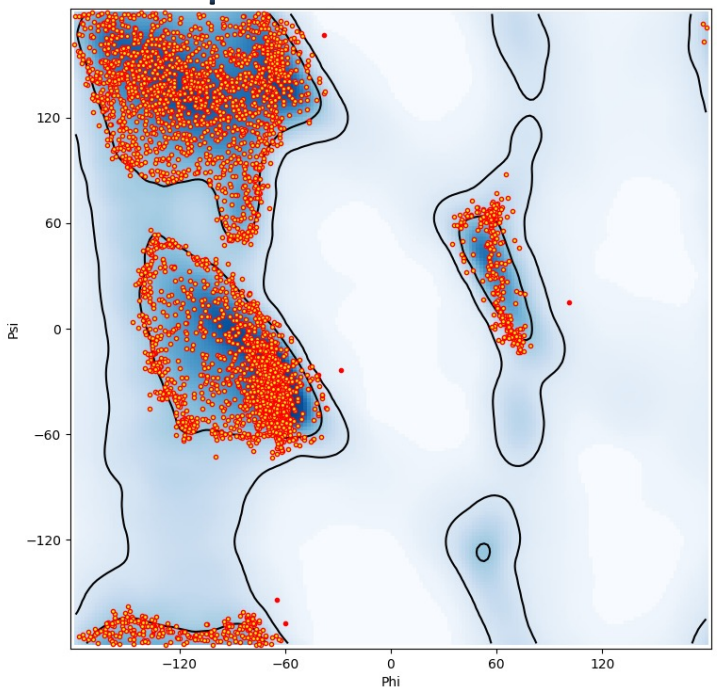
# How did that happen?

**PDB code: 5a9z**

Original



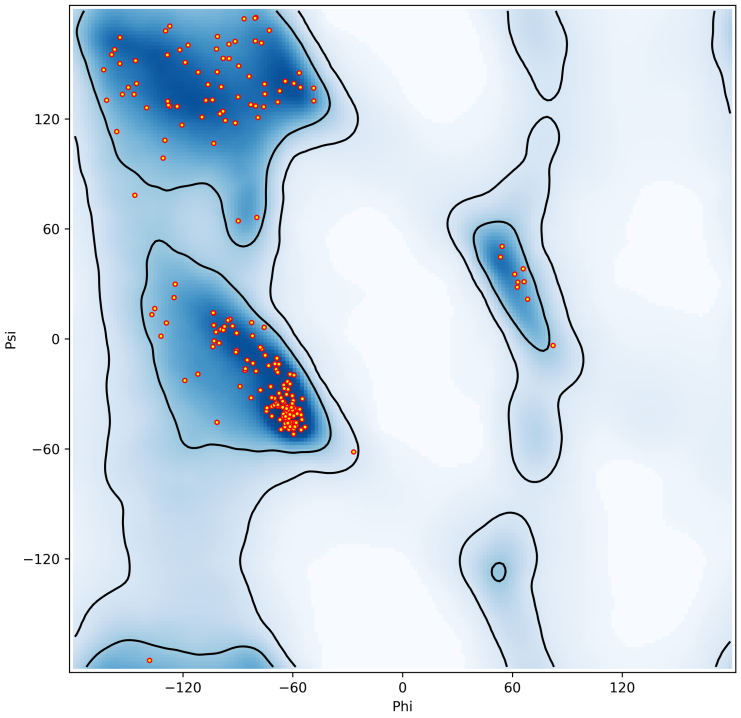

Refined with Ramachandran plot restraints



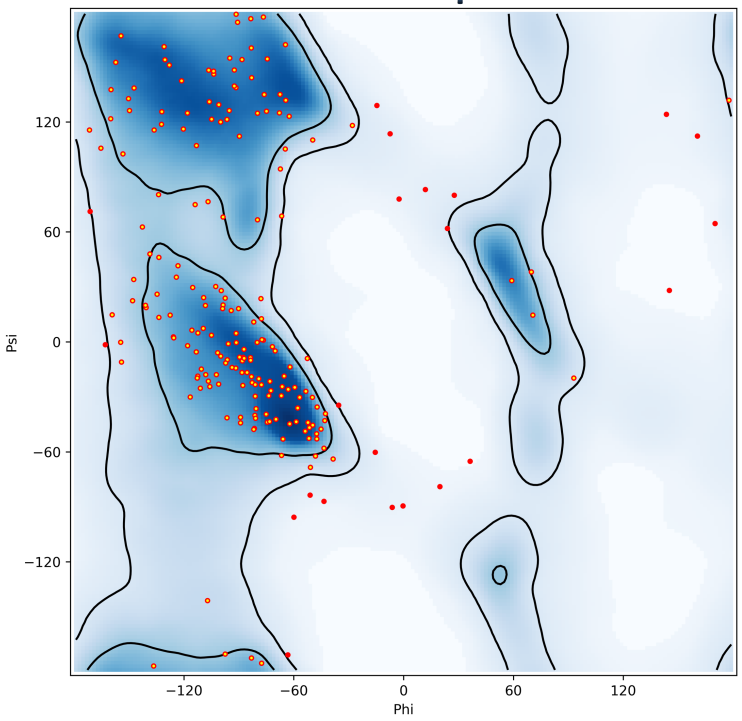

**Don’t use Ramachandran plot restraints to remove outliers!**

# Ramachandran plot restraints

Before refinement




After refinement
(No Ramachandran plot restraints)




Use Ramachandran plot restraints to prevent outliers from occurring!

# Restraints and limitations

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

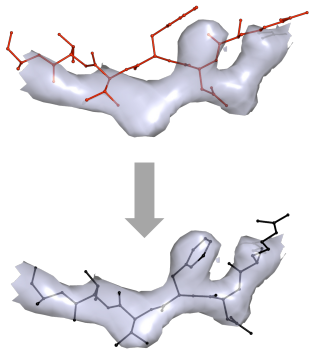
$$T_{RESTRAINTS} = T_{BOND} + T_{ANGLE} + T_{DIHEDRAL} + T_{PLANE} + T_{REPULSION} + T_{CHIRALITY}$$

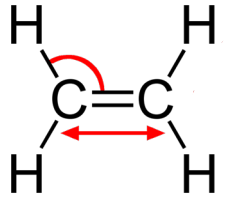
- Restraints are too limited:
  - No attraction terms (electrostatics, etc)
  - Not using information about protein structure (secondary structure, rotamers)
  - Limited to tabulated entities in the libraries (e.g., Monomer Library, GeoStd)

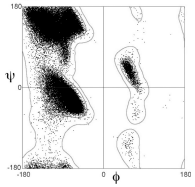
# Restraints from QM

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

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***

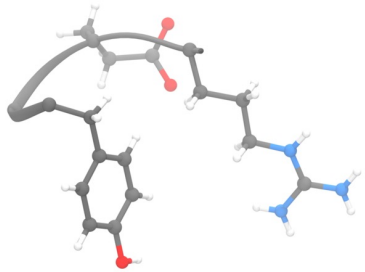
# History of progress

Impossible

2010

## QM Calculations

Impossible for proteins. Limited to small molecules only

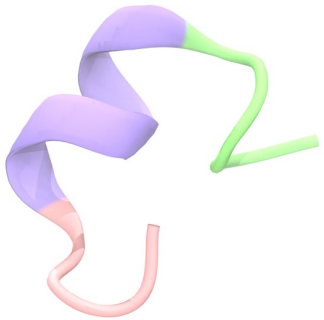


Limited

2012

## GPU Accelerated QM

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

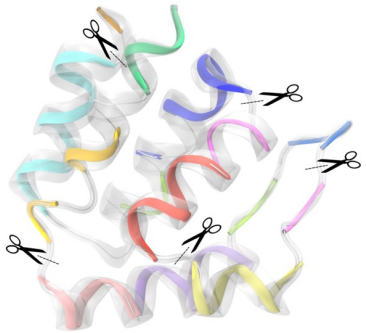


Possible

2017

## Q|R with Fragmentation

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

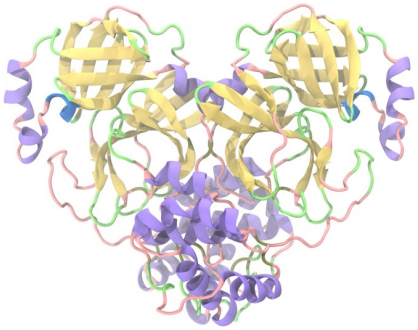


Practical

2024

## ML Potentials

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



# The Team

## Carnegie Mellon University


Olexandr Isayev


Roman Zubatyuk


Hatice Gökcan

## Uni. of Florida


Adrian Roitberg

**Blend of expertise and background**

Crystallography methods + Software development + QM expertise + ML / AI expertise

## Pending AI


Marl Waller


Holger Kruse

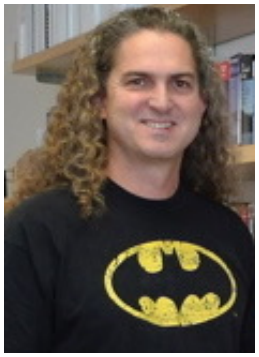
## Uni of Wrocław


Malgo Biczysko

## LBNL


Billy Poon


Pavel Afonine


Nigel Moriarty

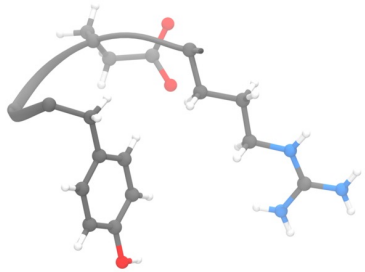
# History of progress

## Impossible

**2010**

### QM Calculations

Impossible for proteins. Limited to small molecules only

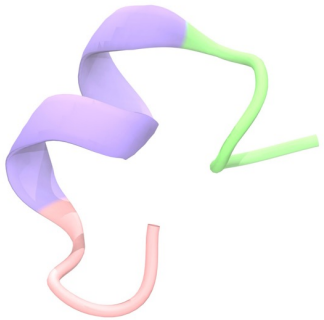


## Limited

**2012**

### GPU Accelerated QM

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

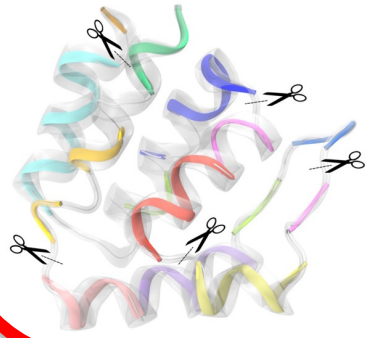


## Possible

**2017**

### Q|R with Fragmentation

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

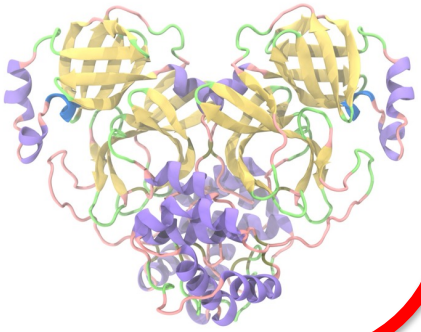


## Practical

**2024**

### ML Potentials

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



# 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

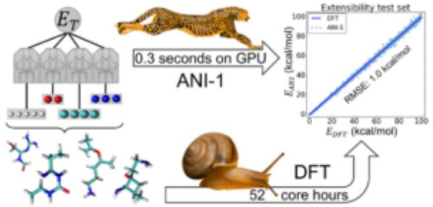
Large Dataset

DFT calculations

ML model

Calculation time:

About a week on one of big national computing resources




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

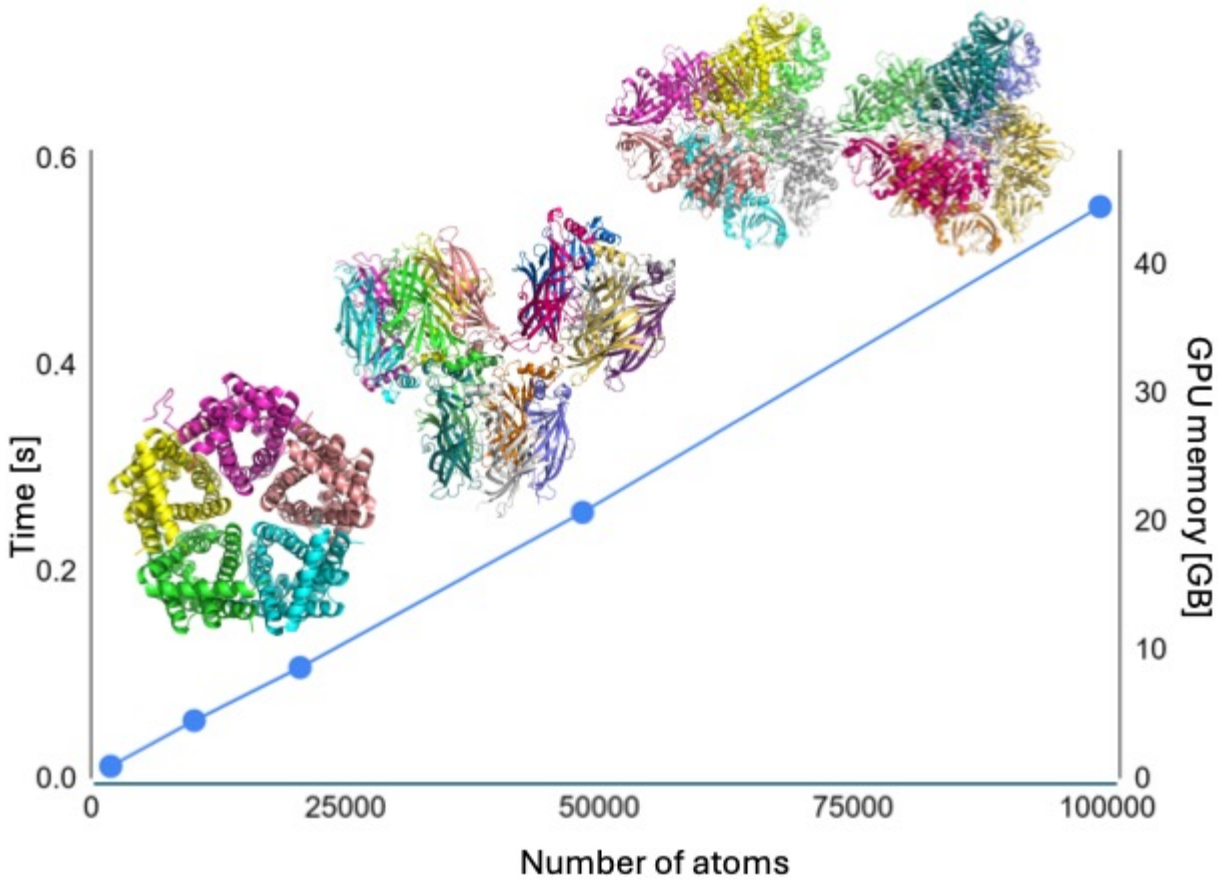
**ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost**

Journal Article

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

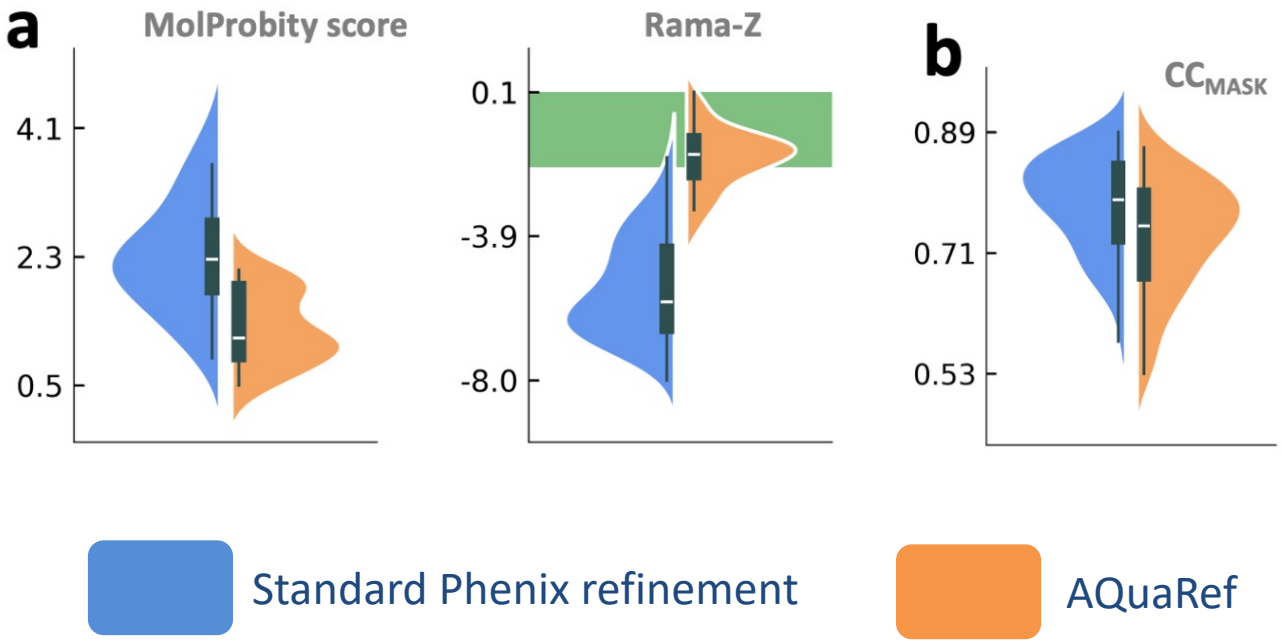
Abstract | **Links** | BibTeX | Tags: ANI, Machine learning potential

# Time & Memory Scaling: single energy calculation



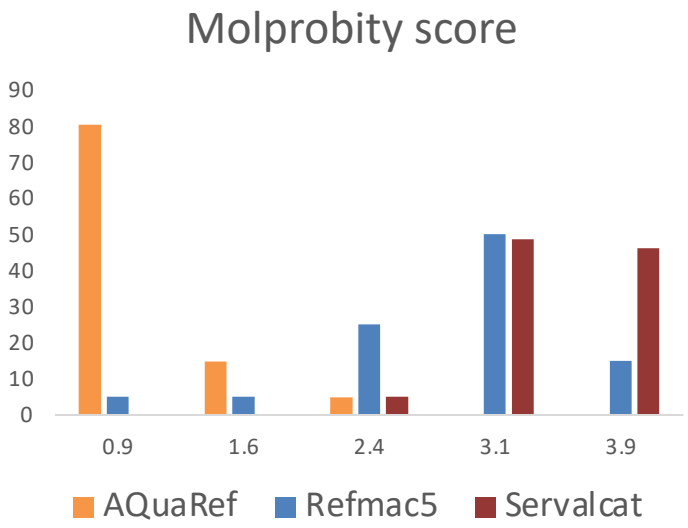
# AQuaRef *vs* standard *Phenix* refinement

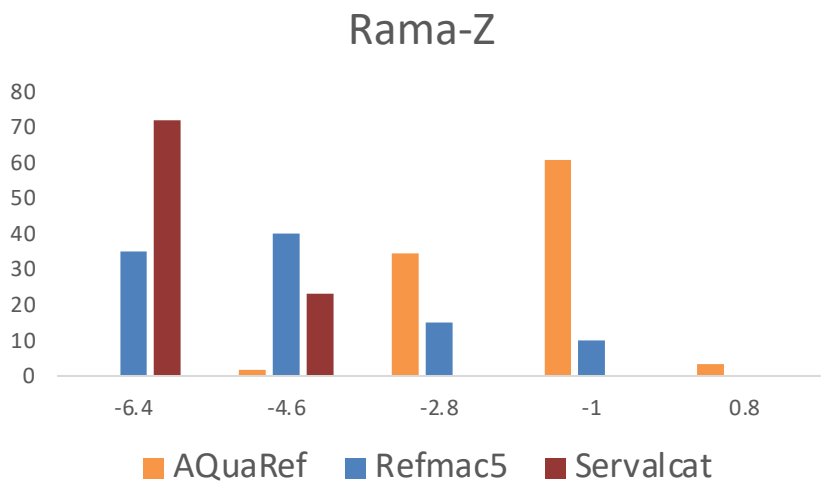
40 cryo-EM low resolution models (3Å or worse)



# Testing: AQuaRef *vs* others

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



# Availability

phenix-online.org/download/phenix/nightly/?version=dev-5430

## Installers for Phenix version dev-5430 (development version)

For a summary of possible errors and/or bugs, please see the build and testing logs:

ci, mac-intel-osx-x86_64, intel-linux-2.6-x86_64-centos6

Version documentation
Summary of changes

**Linux**

Kernel 2.6 (64-bit; CentOS 6) [ **download command-line installer** ]

CentOS 6 is the earliest linux distribution supported. This installer will work with newer linux distributions.

**macOS**

OS X 10.9+ (64-bit Intel) [ **download point-and-click installer** ]

OS X 10.9+ (64-bit Intel) [ **download command-line installer** ]

The graphical installer will install to /Applications; the command-line installer may be relocated with the --prefix argument.
On macOS 12 and later, you will need to run "xattr -c " to avoid errors about untrusted files
For Phenix 1.20.1 and before, the Phenix GUI will crash on macOS Sonoma (14). Please download this fix to fix your Phenix installation. Extract the tarball and run "sudo ./fix_wxpython.sh /Applications/phenix-1.20.1-4487". You will need administrative access to modify files in /Applications.

**Windows** (partially supported)

For full functionality, you should run the Python 2.7 linux installer on Windows using WSL. Instructions for setting up Phenix with WSL can be found here.

A native Windows build is only available for Python 3, but you will encounter crashes. See below for the link to the Windows installer.

The next official release of Phenix will be in Python 3 and Windows will have a working installer again.

**Python 3.9 (Alpha) - CUDA 11 and CUDA 12 support only for quantum refinement on Linux, general Phenix GUI is still in testing and does not have full functionality**

Linux (CUDA 12) [ **download command-line installer** ]

Linux (CUDA 11) [ **download command-line installer** ]

**Python 3.9 (Alpha) - For testing and development, do not expect full functionality in the Phenix GUI**

Linux [ **download command-line installer for development builds and modifications (contains "modules" directory)** ]

Linux [ **download command-line installer** ]

macOS (Intel) [ **download graphical installer** ]

macOS (Intel) [ **download command-line installer** ]

macOS (Apple Silicon) [ **download graphical installer** ]

macOS (Apple Silicon) [ **download command-line installer** ]

Windows [ **download graphical installer** ]

AQuaRef

nature communications

Article

https://doi.org/10.1038/s41467-025-64313-1

# AQuaRef: machine learning accelerated quantum refinement of protein structures

Received: 29 July 2024

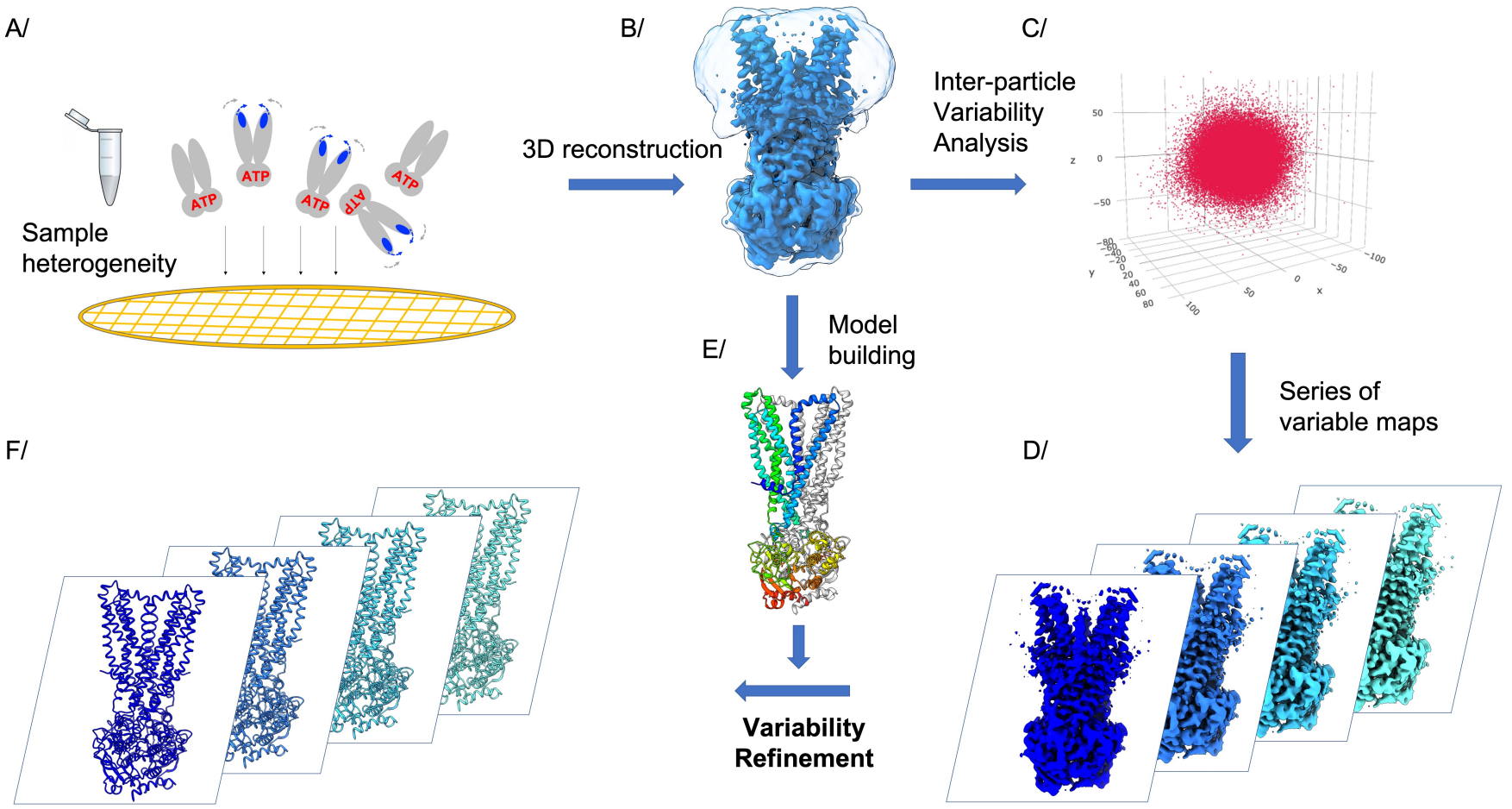
Accepted: 12 September 2025

Published online: 17 October 2025

Roman Zubatyuk[1,7], Malgorzata Biczysko[2,7], Kavindri Ranasinghe[3,7], Nigel W. Moriarty[4], Hatice Gokcan[1], Holger Kruse[5], Billy K. Poon[4], Paul D. Adams[4,6], Mark P. Waller[5], Adrian E. Roitberg[3], Olexandr Isayev[1] ✉ & Pavel V. Afonine[4] ✉

# Phenix is feedback & need driven

# Variability refinement: treasuring conformational changes



# Feedback & need driven – Example

## Vincent's post on phenix mailing list (phenixbb)

From vincent Chaptal <vincent.chaptal@ibcp.fr> | Reply | Reply List | Forward | Archive | Junk | Delete | More

To PHENIX user mailing list <phenixbb@phenix-online.org> 1/11/22, 04:48

Subject **[phenixbb] refinement of an ensemble of structures -> cryoEM variability**

Hi Phenix-ers,

I thought to ask for something that I believe you have already implemented, but I'm not sure of the best tool to use.

I have a cryoEM map where I refine my "high resolution" structure. I also have the 3D variability of this map that shows several maps varying around the consensus high-res map. I want to refine an ensemble (20) of structures, one for every 20 maps around the consensus map.
Is there a tool in phenix to do this?

I could refine individually the high-res structure into each map incrementally; since every map differs a little from the original one, Real-space-refinement could move the structure a little at a time. Then I could combine all the PDBs in an ensemble?
A tool to refine variability would be very useful. Input could be a PDB and an ensemble of maps, and output would be all the PDBs combined?

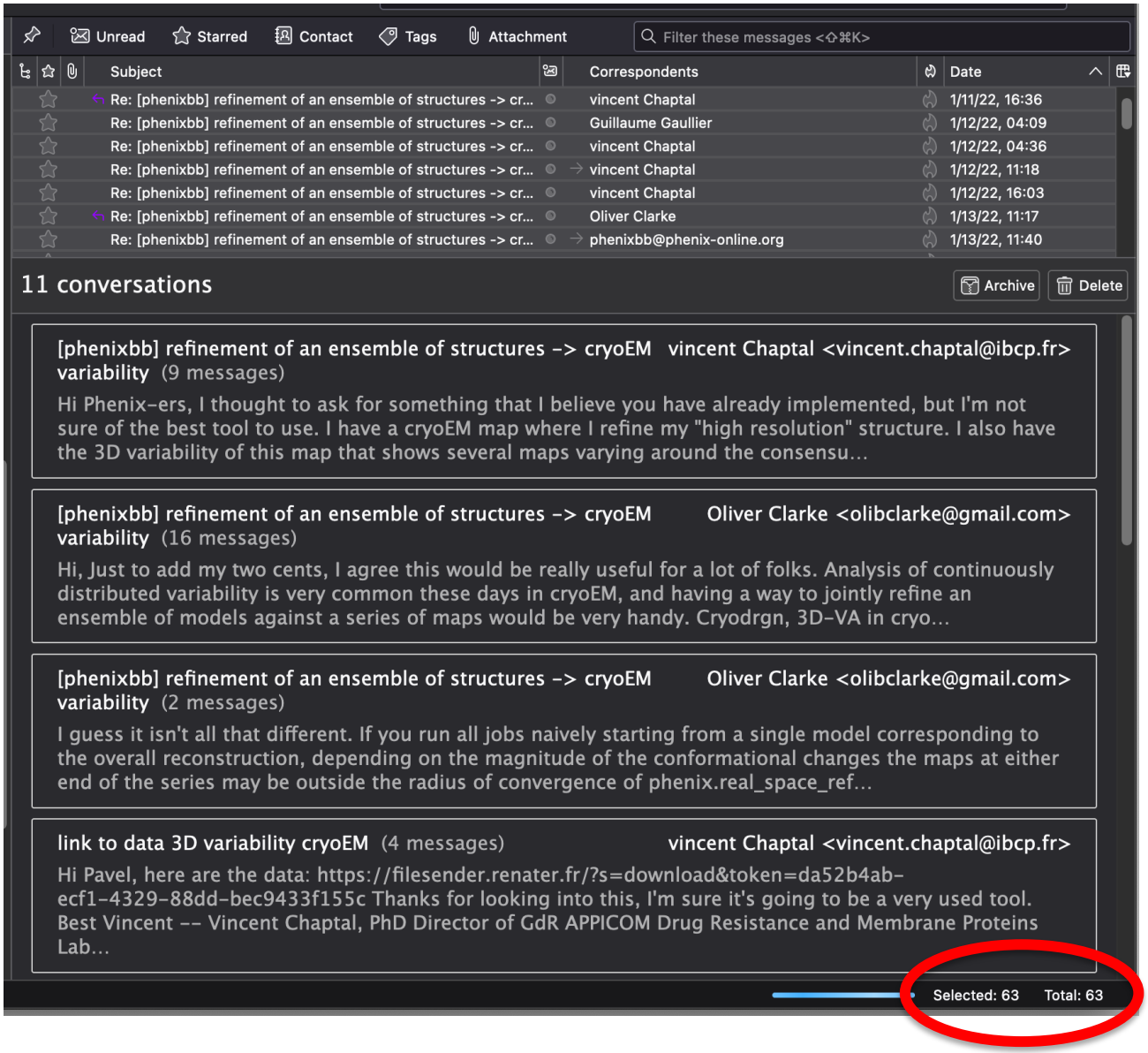
Thank you.

All the best
Vincent

--
Vincent Chaptal, PhD
Director of GdR APPICOM
Drug Resistance and Membrane Proteins Lab

MMSB -UMR5086
7 passage du Vercors
69007 LYON
FRANCE
+33 4 37 65 29 01
http://www.appicom.cnrs.fr
http://mmsb.cnrs.fr/en/

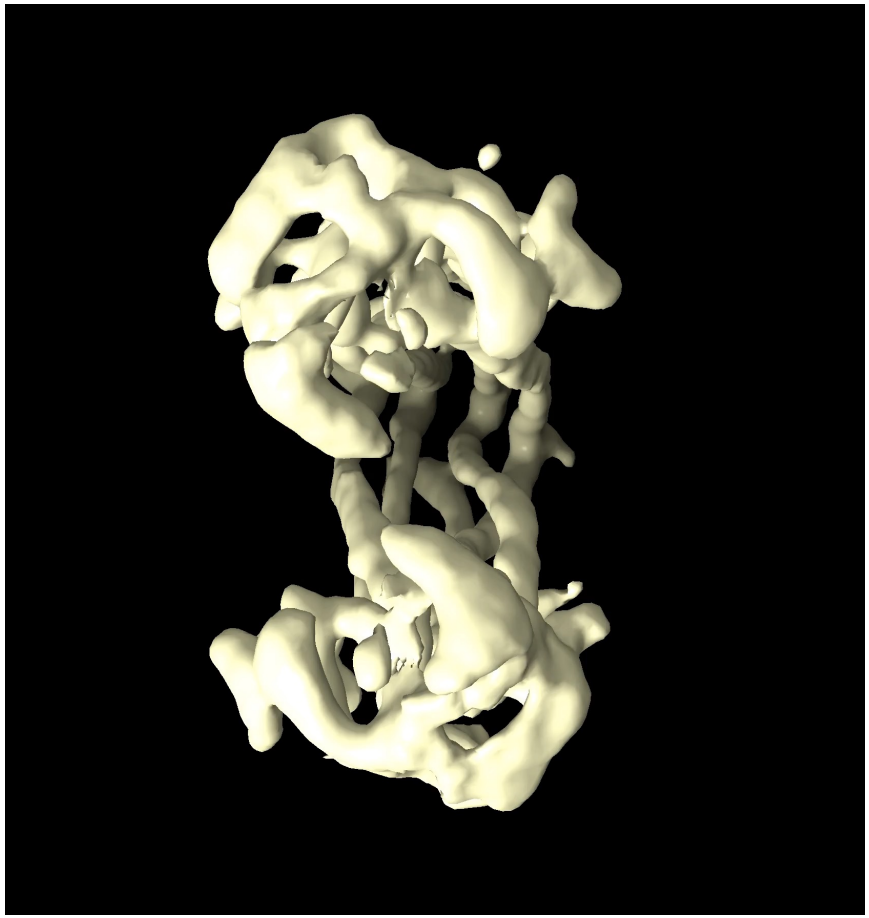
# Feedback & need driven – Example

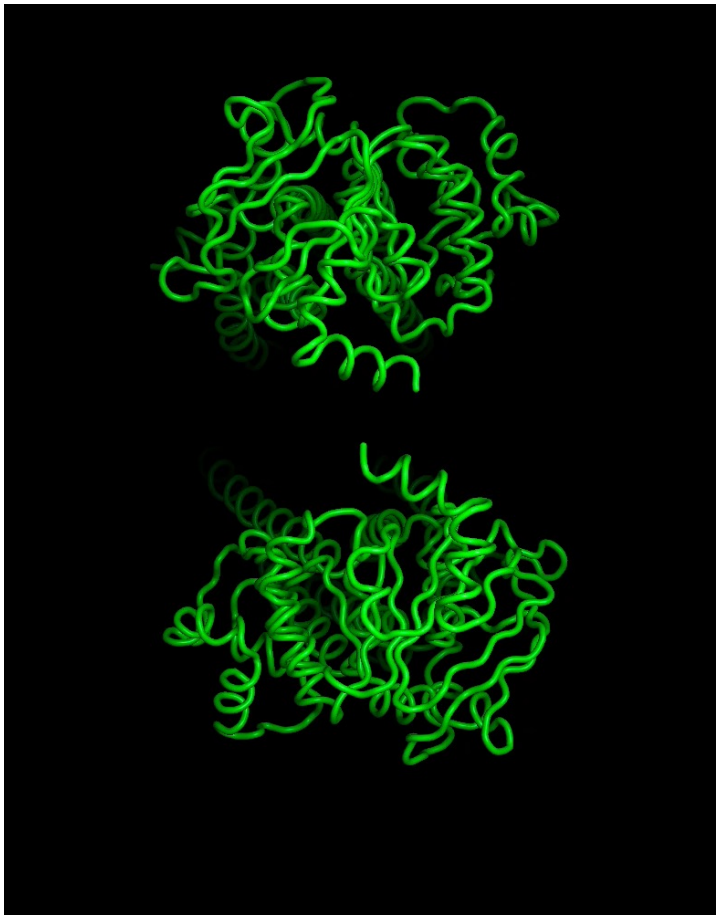



**63 emails later....**

# Feedback & need driven – Varref (Variability Refinement)

**… 63 emails later,** we came up with a tool to effectively model ensemble of maps with ensemble of atomic models in a fully automated manner…





# Feedback & need driven – Varref (Variability Refinement)

… and wrapped that into a publication and a user-accessible tool

Contents lists available at ScienceDirect

BBA - Biomembranes

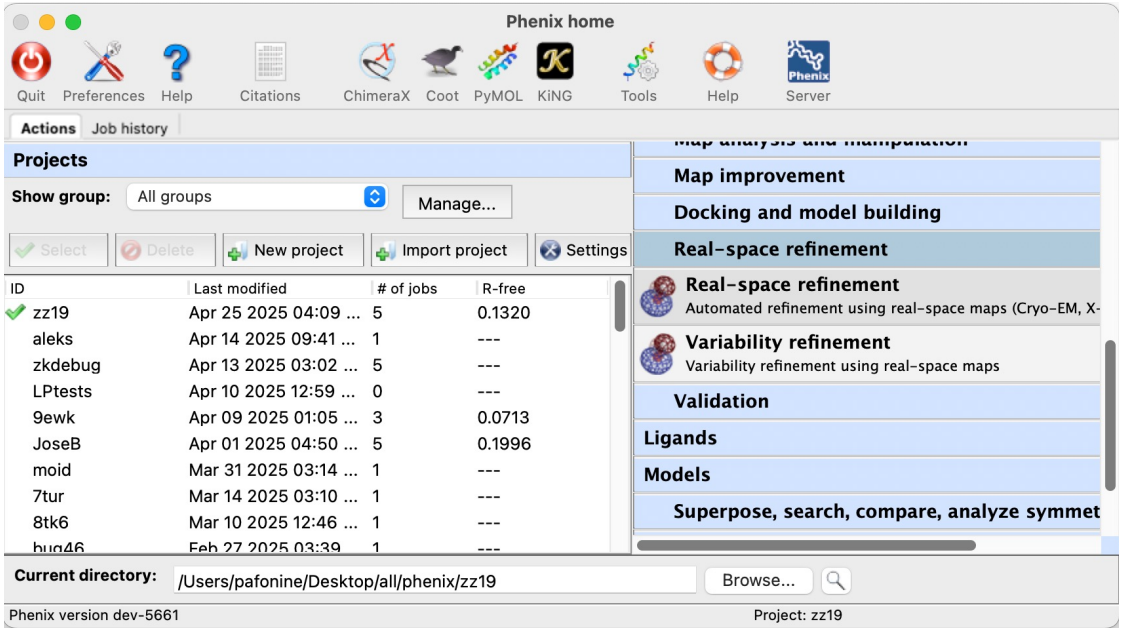
journal homepage: www.elsevier.com/locate/bbamem

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



# Feedback & need driven – Varref (Variability Refinement)

nature communications

Article

https://doi.org/10.1038/s41467-024-55439-9

## Conformational diversity in class C GPCR positive allosteric modulation

Giuseppe Cannone[1,7], Ludovic Berto[2,7], Fanny Malhaire[2], Gavin Ferguson[2], Aurelien Fouillen[2], Stéphanie Balor[3], Joan Font-Ingles[4], Amadeu Llebaria[4], Cyril Goudet[2], Abhay Kotecha[5], Vinothkumar K.R.[6] & Guillaume Lebon[2]

Received: 23 November 2023

Accepted: 10 December 2024

Published online: 13 January 2025

nature communications

Article

https://doi.org/10.1038/s41467-025-56849-z

## Rhodamine6G and Hœchst33342 narrow BmrA conformational spectrum for a more efficient use of ATP

A. Gobet[1,7], L. Moissonnier[2,7], E. Zarkadas[3], S. Magnard[2], E. Bettler[4], J. Martin[5], R. Terreux[4], G. Schoehn[6], C. Orelle[2], JM Jault[2], P. Falson[2] & V. Chaptal[2]

Received: 5 April 2024

Accepted: 3 February 2025

Published online: 18 February 2025

www.nature.com/scienti

scientific reports

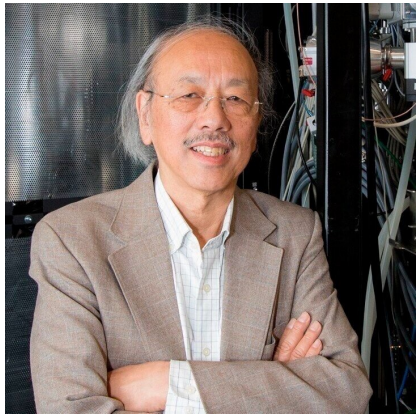
OPEN

## Conformational landscape of soluble α-klotho revealed by cryogenic electron microscopy

Nicholas J. Schnicker[1,2,5], Zhen Xu[1,5], Mohammad Amir[3,5], Lokesh Gakhar[1,4] & Chou-Long Huang[3]

# Feedback & need driven – Real Space Refinement

Around 2015: cryoEM resolution revolution about to happen!



- More and more high-resolution cryoEM maps, lack of model building and refinement tools!
- Birth of *phenix.real_space_refine* – refinement program in Phenix to refine atomic models into cryoEM maps

PNAS

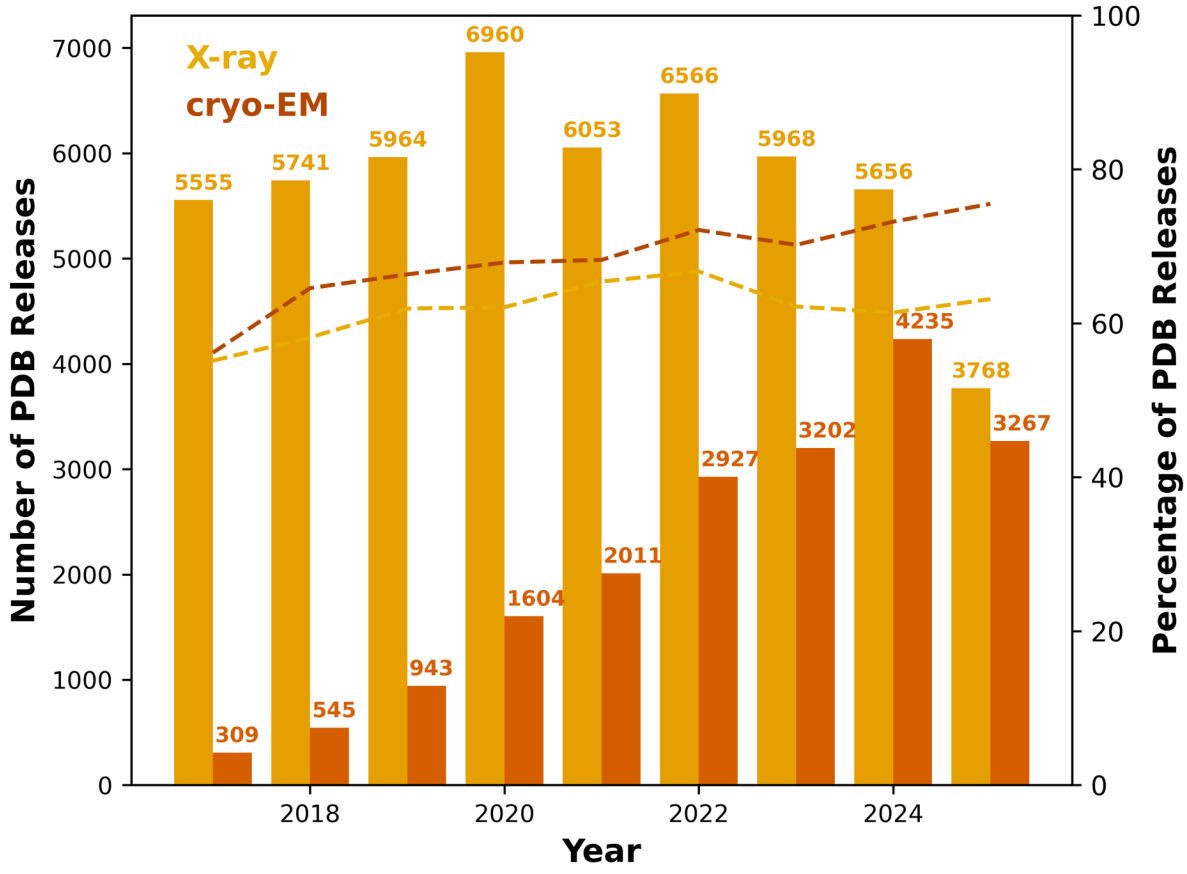
## Accurate model annotation of a near-atomic resolution cryo-EM map

**Corey F. Hryc[a,1], Dong-Hua Chen[b,1,2], Pavel V. Afonine[c], Joanita Jakana[b], Zhao Wang[b], Cameron Haase-Pettingell[d], Wen Jiang[e], Paul D. Adams[c], Jonathan A. King[d], Michael F. Schmid[a,b], and Wah Chiu[a,b,3]**

[a]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030; [b]National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030; [c]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [d]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [e]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907

Contributed by Wah Chiu, February 2, 2017 (sent for review December 7, 2016; reviewed by Terje Dokland and Jack E. Johnson)

# 10 years later..



# Pixel size calibration (refinement)

Reply | Reply List | Forward | Archive | Junk | Delete | More

From Oliver Clarke

To phenixbb@phenix-online.org 6/3/25, 08:19

Subject **[phenixbb] Refine pixel size for EM map?**

Tags Important

List-ID Phenix user mailing list <phenixbb.phenix-online.org>

Hi,

I remember asking about this some time ago, but I am not sure if it was ever implemented.

Is there a way to refine just pixel size for an EM map in phenix.real_space_refine (or another phenix utility)? Even just automatically rigid-body fitting a crystal structure and calculating the pixel size with max correlation would be handy. We do this at the moment using Chimera/ChimeraX, but an automated way to do it in phenix would be handy.

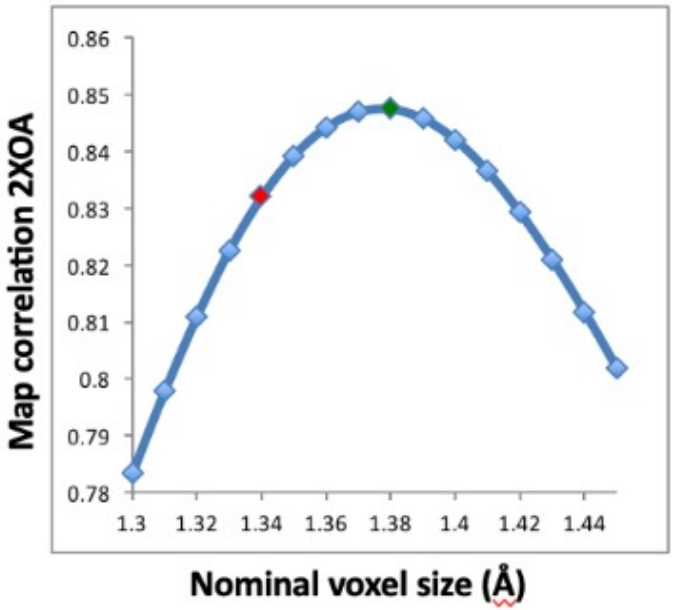
In the absence of a crystal structure, I wonder if directly refining pixel size during real space refinement would be possible/reasonable at higher resolution (<2Å)?

Cheers
Oli

# Pixel size calibration (refinement)

**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).



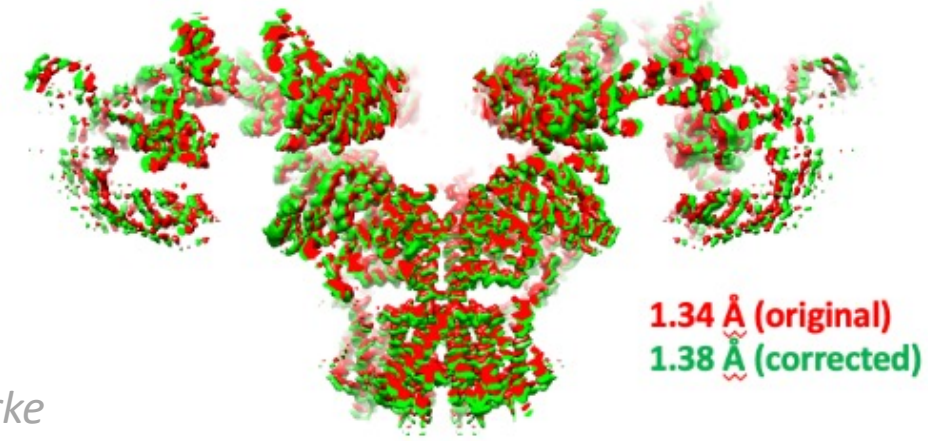
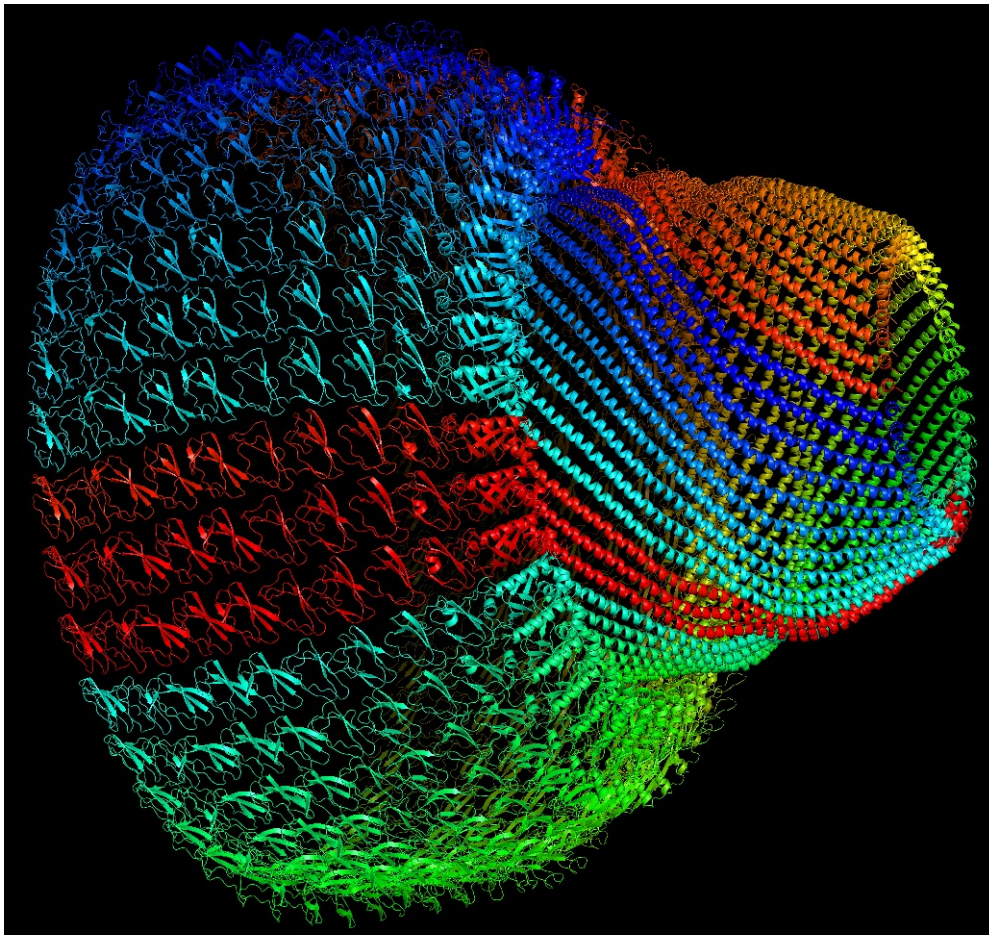




*Figure: Oliver Clarke*

# MagRef: Pixel size calibration (refinement)

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

PDB: 4hl8 (vault ribonucleoprotein particle), 3.5 Å, 250,000 atoms



**`phenix.magref`** takes about 4 minutes on a 6 years old Mac laptop (~100 of rigid-body refinements)