

Model Refinement

Pavel Afonine

Phenix team

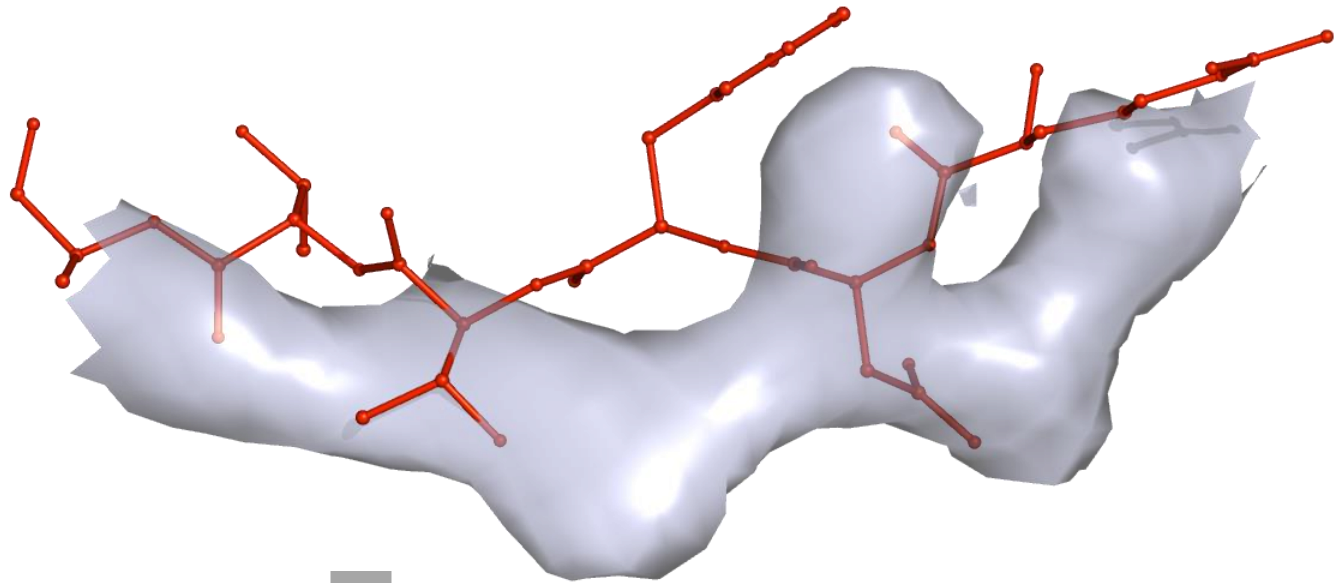
Lawrence Berkeley National Lab, California, USA

**May, 2025
MCCS, Madrid**

Model refinement in a nutshell

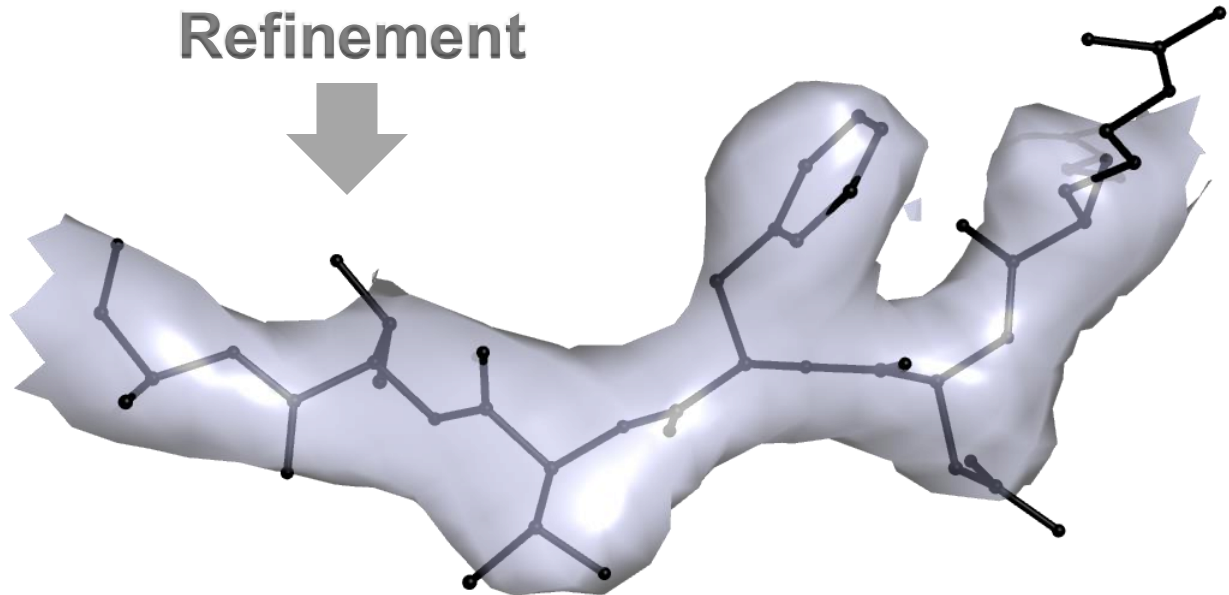
Model refinement

**Initial (poor)
model**

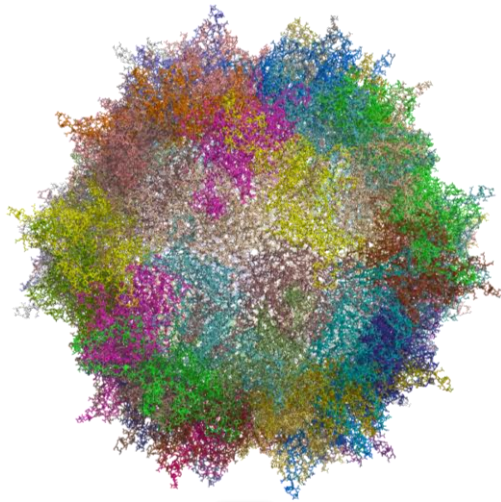


Refinement

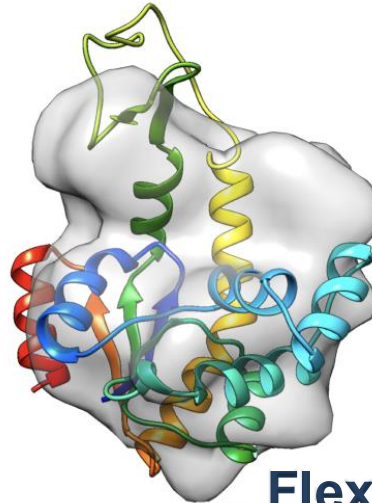
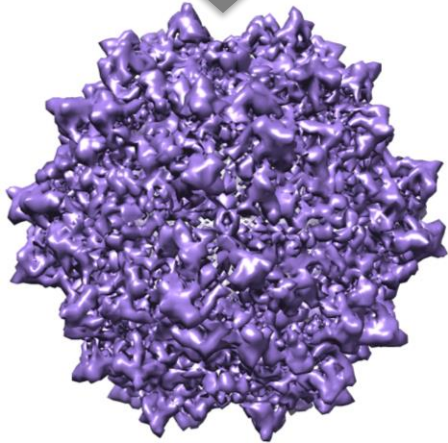
**Improved
(refined)
model**



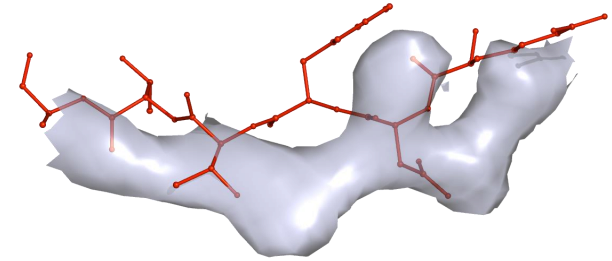
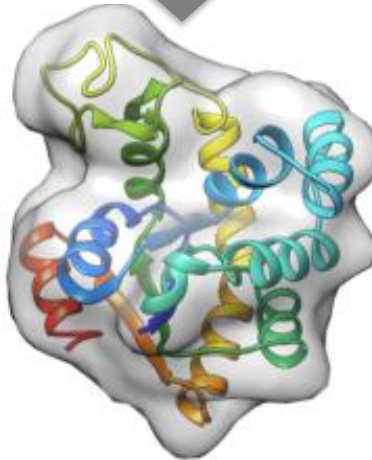
Model refinement vs other model fitting tools



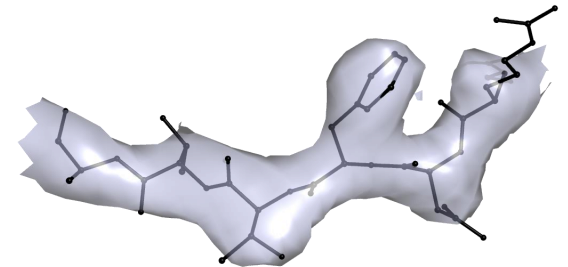
Docking



**Flexible fitting,
morphing**

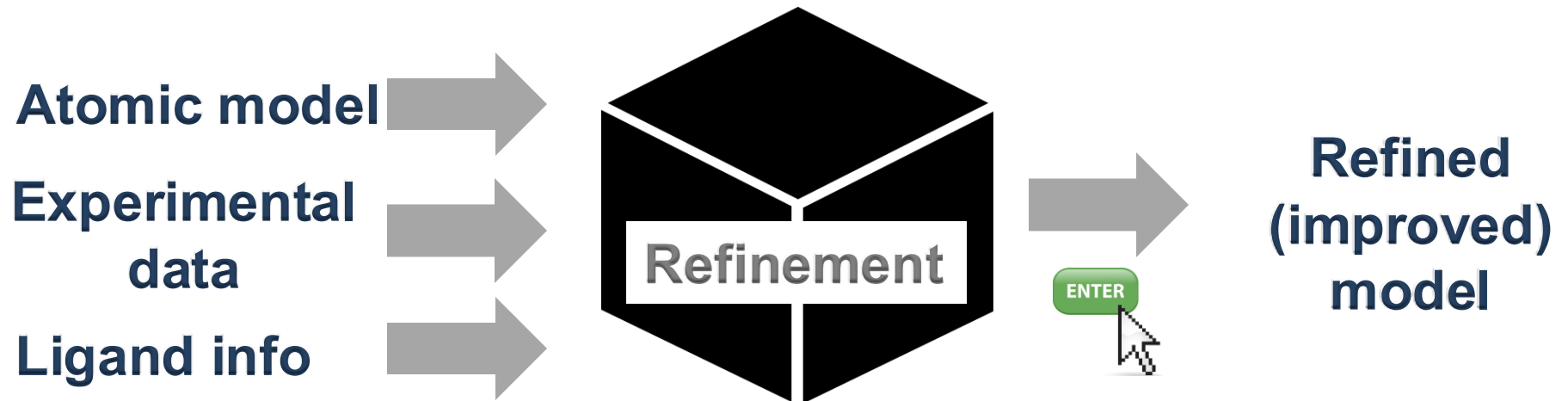


Refinement



- All the above move model to the map. The difference is: by how much

Structure refinement: black box



Structure refinement: black box

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

Structure refinement: black box

- **No.** Because:
 - Refinement parameterization isn't easy (next slide)
 - Default settings suit most common scenario:
 - $\sim 2 \text{ \AA}$ resolution data, model reasonably fits data
 - Less typical situations need customizations:
 - Low-high resolution data need special attention
 - Incomplete models need special attention
 - Specialized fine-tuning of complete models may be needed
 - AlphaFold predicted models may need special treatment

Structure refinement: lots of jargon

Reference model?

TLS?

Rotamer fixing?

AltLocs?

ADP?

Group B vs individual?

Local minima?

tNCS?

Clashes?

NCS?

IAS?

Weights?

CDL?

SA?

Grid search?



Minimization?

Rama plot restraints?

f' & f'' ?

Hydrogens?

Restraints?

Bulk-Solvent?

Rigid body?

Rama-Z?

Anisotropy?

NQH flips?

SS restraints?

Twinning?

Structure refinement: black box

- What to do when the 'black box' does not work?
 - Your decision-making is needed (and it is complicated!)

Structure refinement: decision-making variables

- **Crystal**
 - Mosaicity
 - Twinning
 - Solvent content
 - Symmetry
- **Data**
 - Resolution
 - Errors
 - Completeness
 - Processing
- **Model**
 - Stage
 - Source
 - Parameterization
 - Fit to data

How you know...

- ... refinement worked ?
- ... you did it correctly ?
- ... the model you got is good enough to publish ?

How you know...

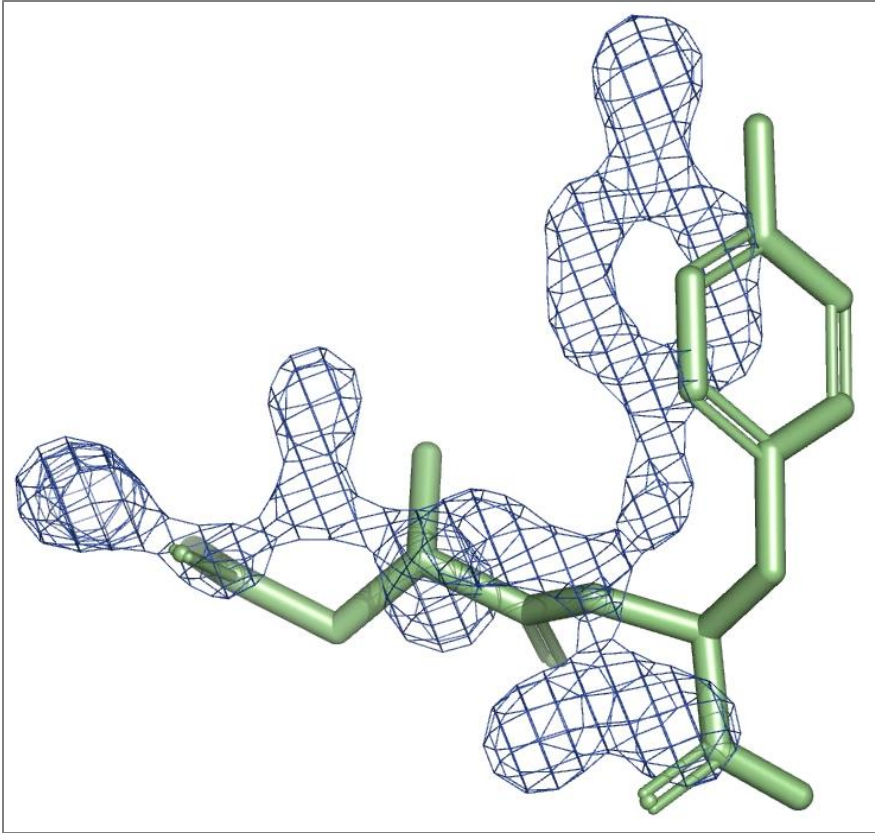
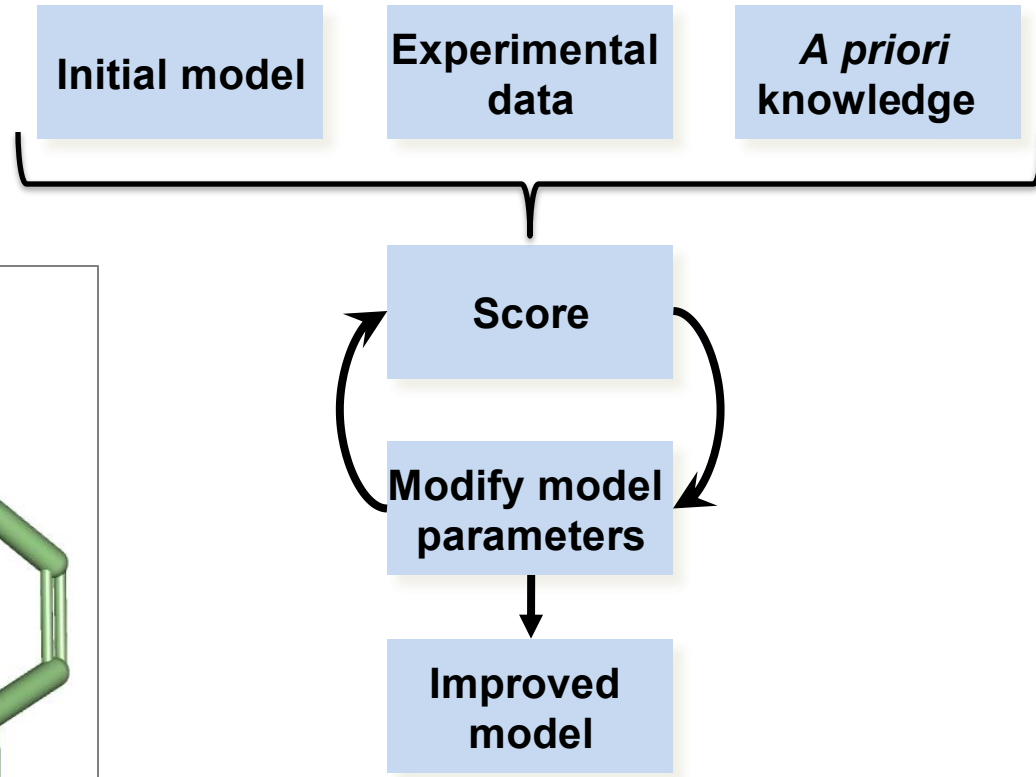
- ... refinement worked ?
 - ... you did it correctly ?
 - ... the model you got is good enough to publish ?
-
- Do validation! There are standard validation protocols that can answer these questions

Summary

- Refinement is reasonably automated but your decision-making is still often needed
- Knowing about model and data guides decision-making
- Use model parameterization that is appropriate for the data quality and stage of refinement
- Use optimization method that is appropriate for the refinement stage
- Do validation throughout the process

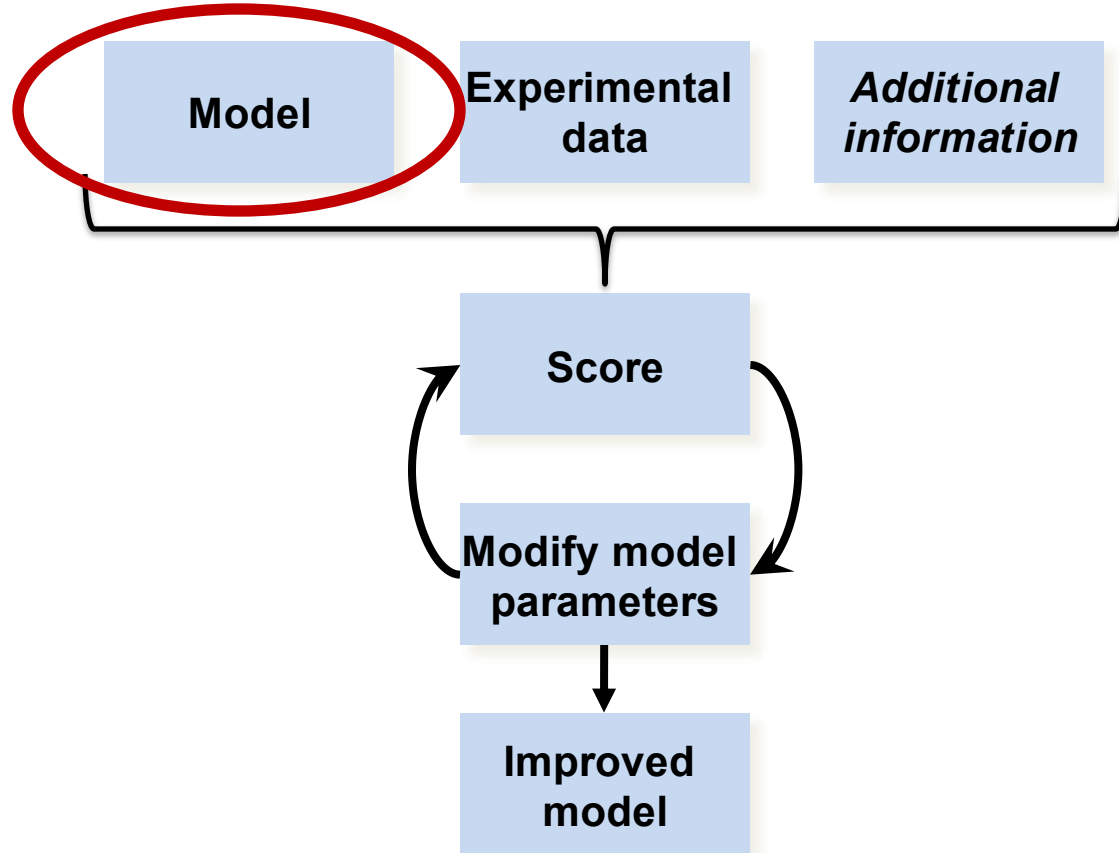
Refinement: a closer look

Model refinement



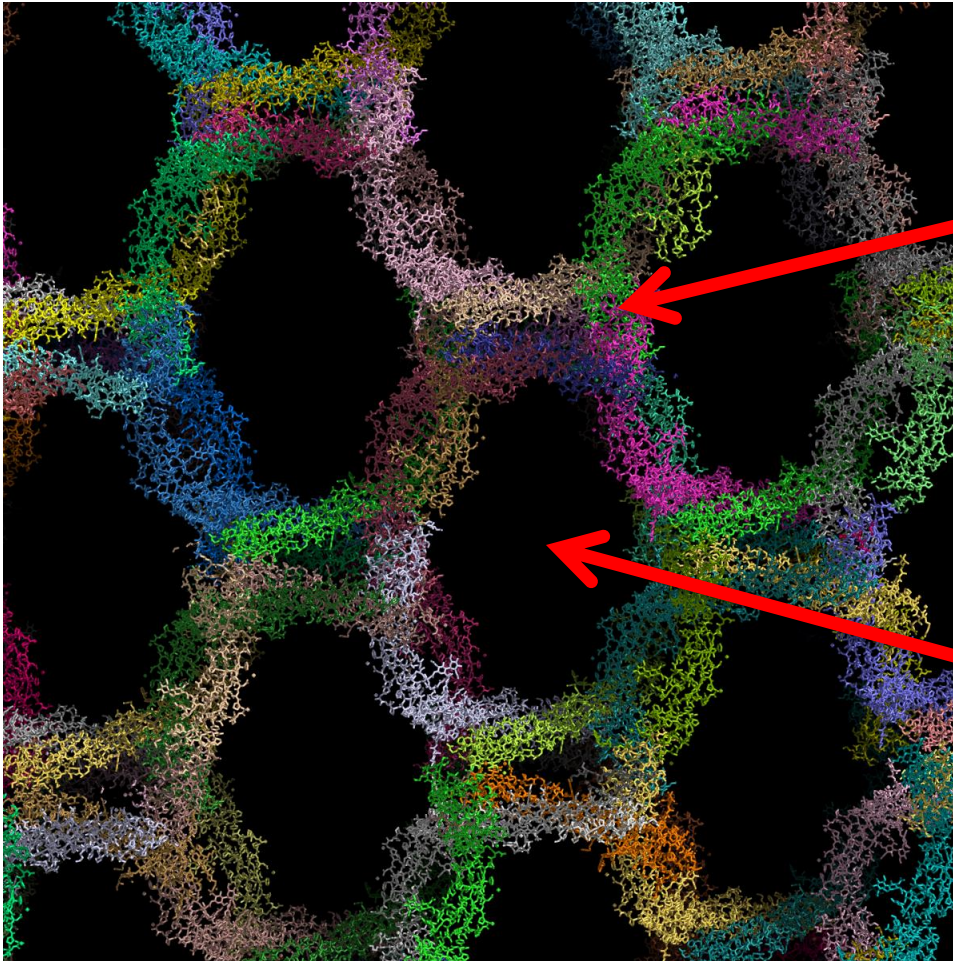
Optimization process of fitting atomic model parameters to experimental data

Refinement: model



Crystal model

PDB code: 1QUB

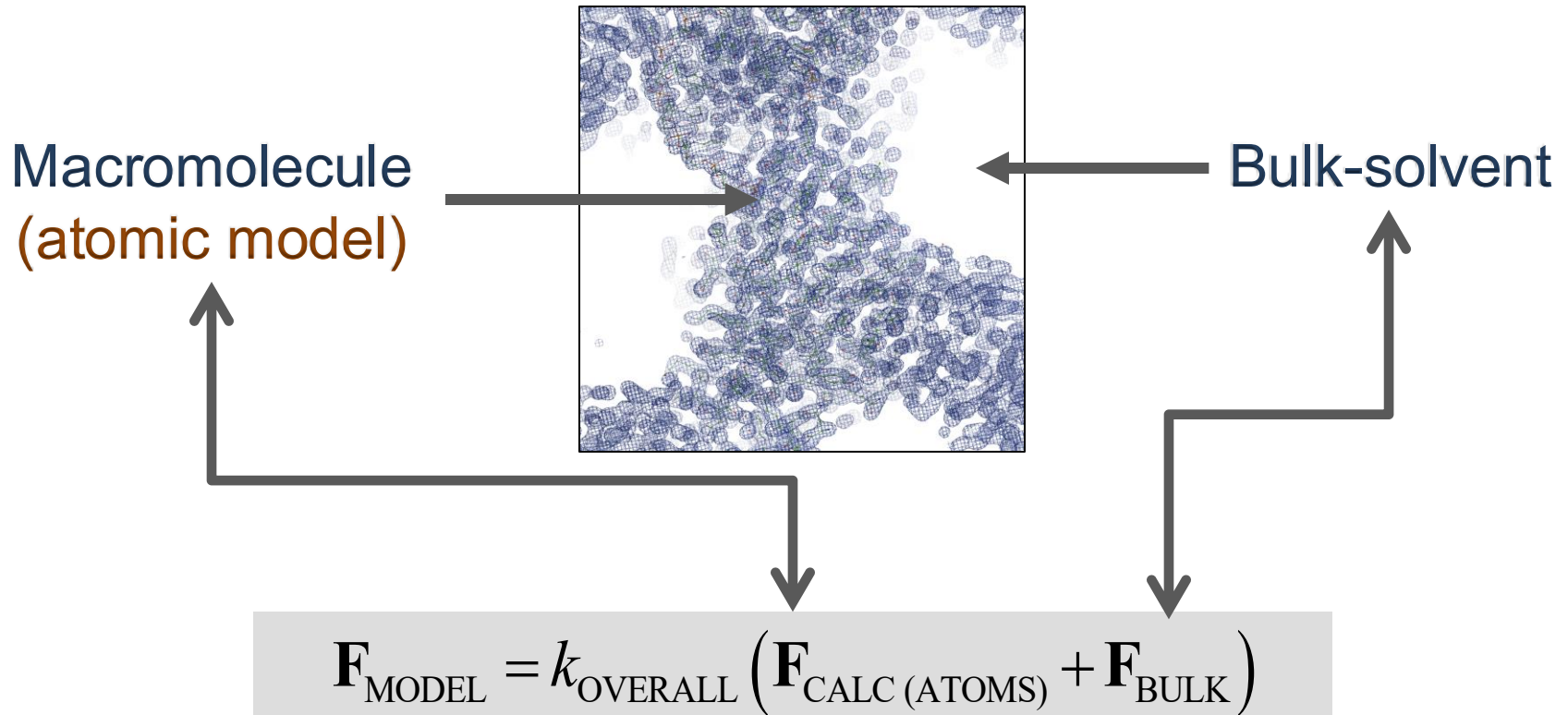


Macro-molecule

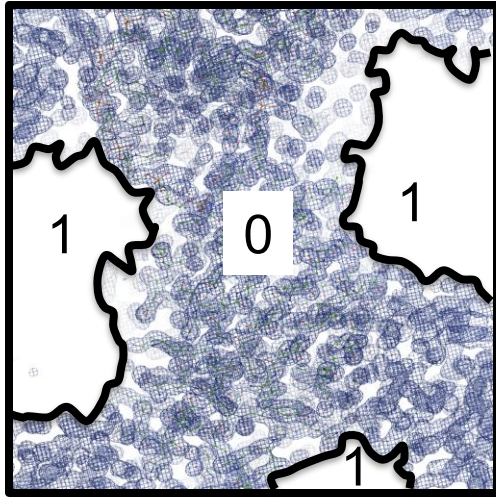
**Bulk-solvent:
~ 50% of unit cell
volume**

Crystal model: $\rho_{\text{crystal}} = \rho_{\text{atoms}} + \rho_{\text{bulk solvent}}$

Crystal model: structure factors



Bulk solvent: F_{BULK}



Steps to account for bulk-solvent:

1. Compute solvent mask, M:
0 – inside protein, 1 – outside

2. Structure factors from M:

$$F_{\text{MASK}} = \text{FT}(M)$$

3. Define solvent contribution F_{BULK} :

$$F_{\text{BULK}} = k_{\text{MASK}} * F_{\text{MASK}}$$

4. Combine with $F_{\text{CALC(ATOMS)}}$

Refine k_{MASK} by fitting $|F_{\text{MODEL}}|$ to F_{obs}

$$F_{\text{MODEL}} = k_{\text{OVERALL}} \left(F_{\text{CALC(ATOMS)}} + F_{\text{BULK}} \right)$$

Atomic model parameters

Position

Larger-scale disorder

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

Local mobility (harmonic vibrations)

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} \left(\mathbf{F}_{\text{CALC (ATOMS)}} + \mathbf{F}_{\text{BULK}} \right)$$

Occupancy

1.00

57.79

ADP (B-factor)

$$\mathbf{F}_{\text{CALC (ATOMS)}}(h, k, l) = \sum_{n=1}^{N_{\text{atoms}}} q_n f_n(s) \exp \left[-\frac{B_n s^2}{4} \right] \exp(2i\pi \mathbf{r}_n \cdot \mathbf{s})$$

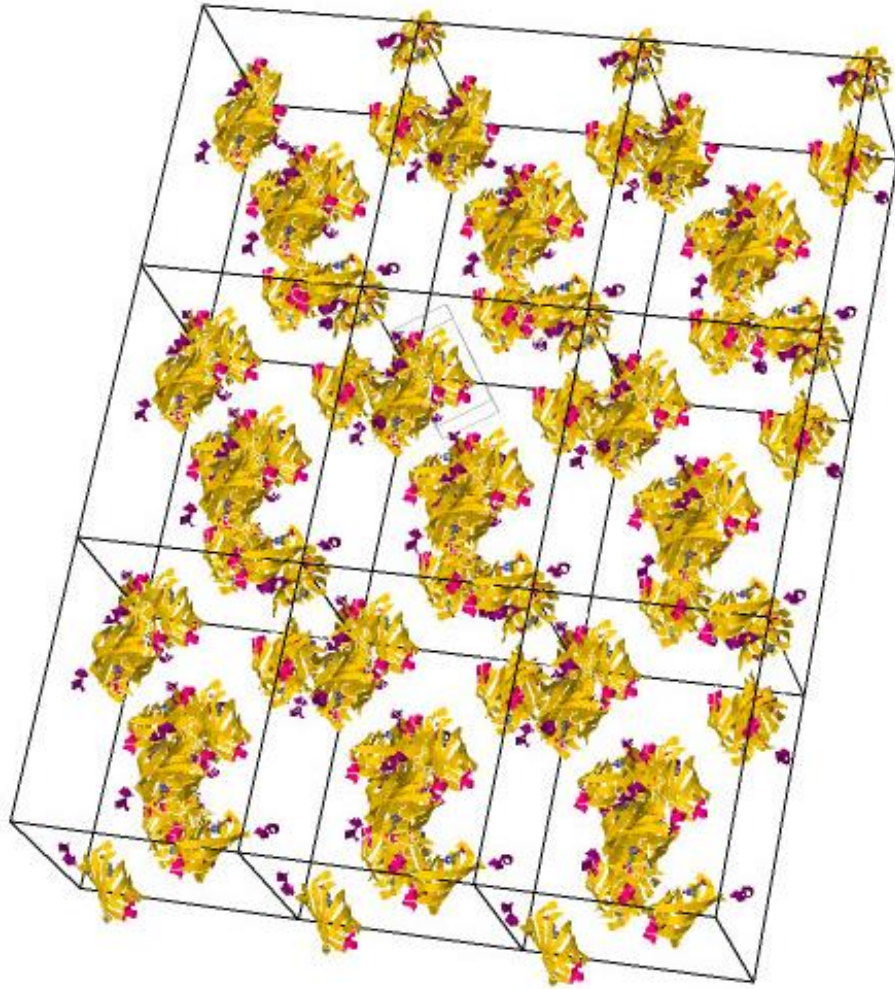
Atom type

C

31.309 29.489 26.044

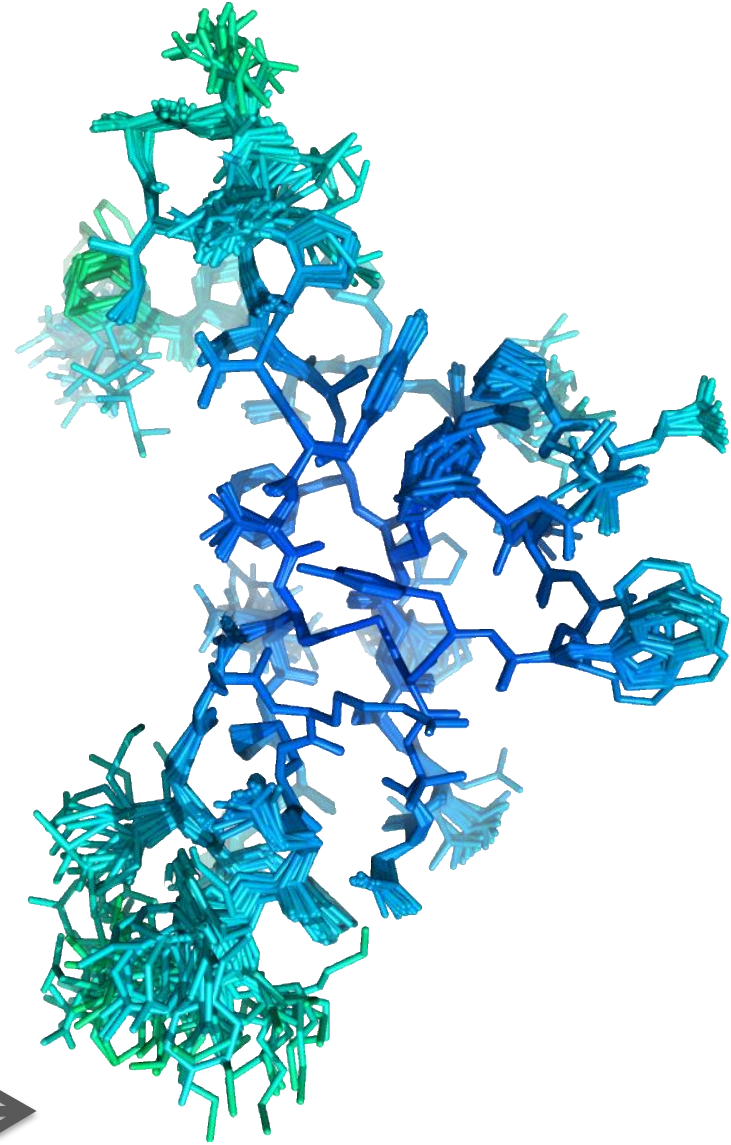
Atomic coordinates

Atomic model parameters: disorder

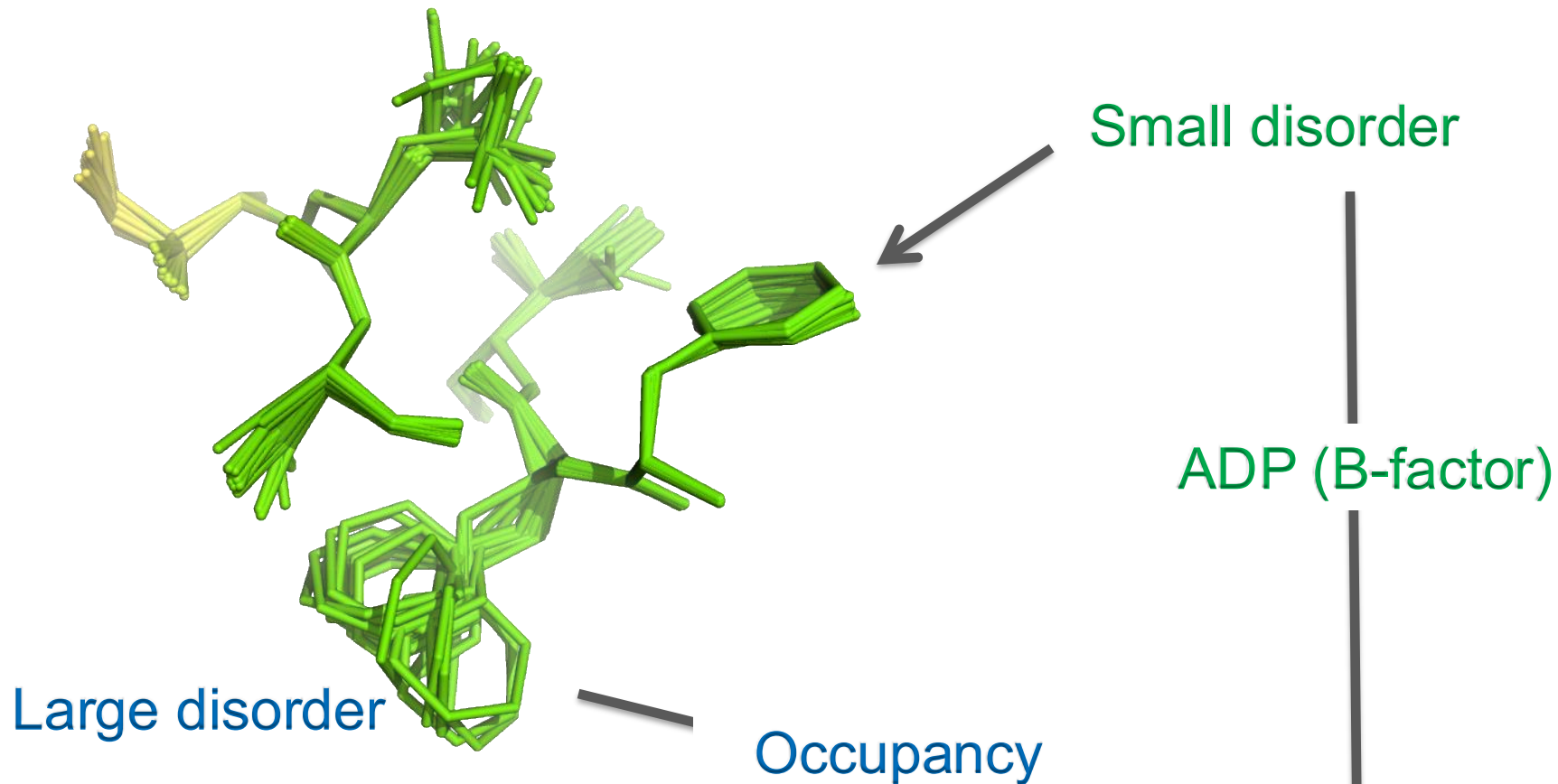


Crystal = many unit cells

Superpose all structures
from each unit cell

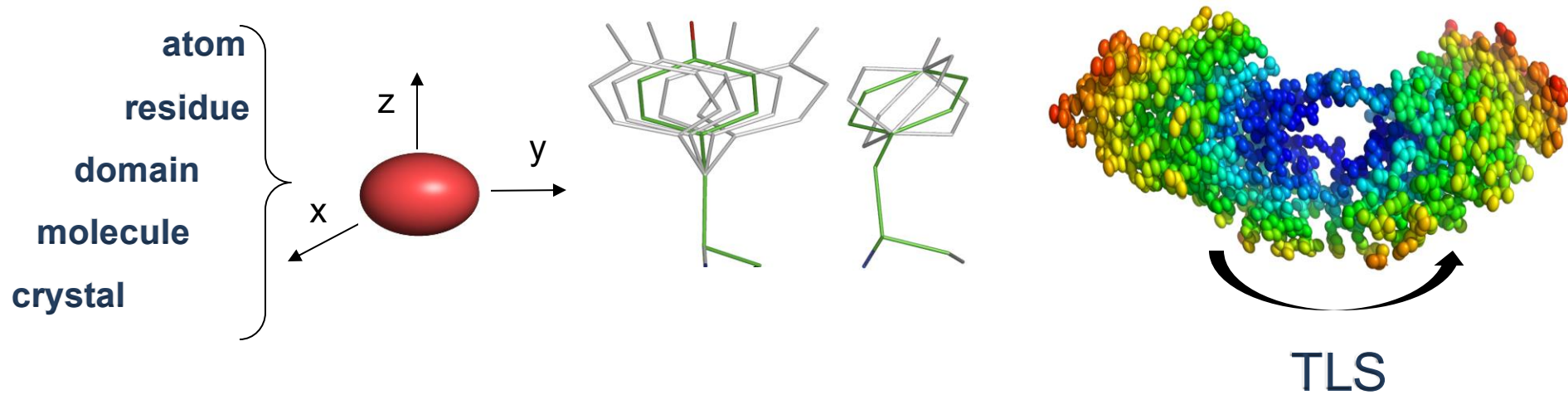


Atomic model parameters



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

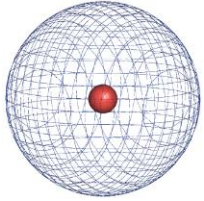
Atomic Displacement Parameters (ADP, B-factors)



B_{TOTAL} = sum of individual contributions

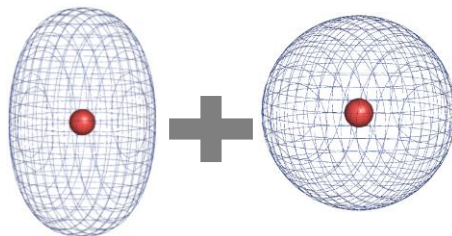
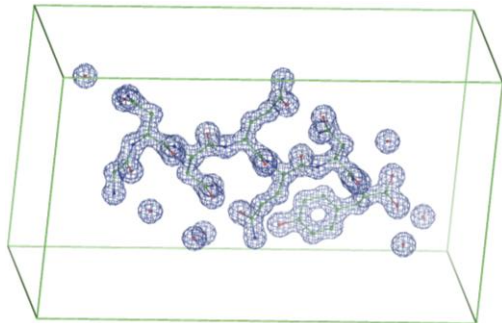
Model Parameters vs Refinement Parameters

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



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

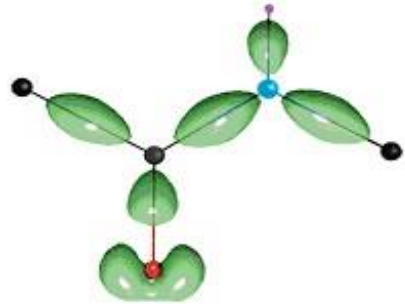
ATOM	25	CA	PRO	A	4	31.309	29.489	26.044	1.00	57.79	C
------	----	----	-----	---	---	--------	--------	--------	------	-------	---



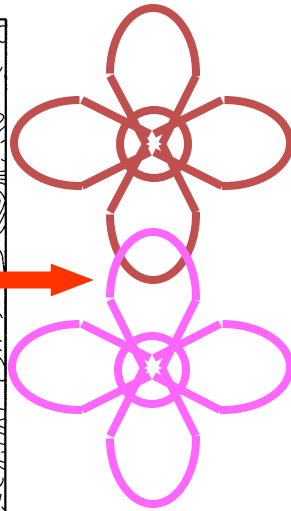
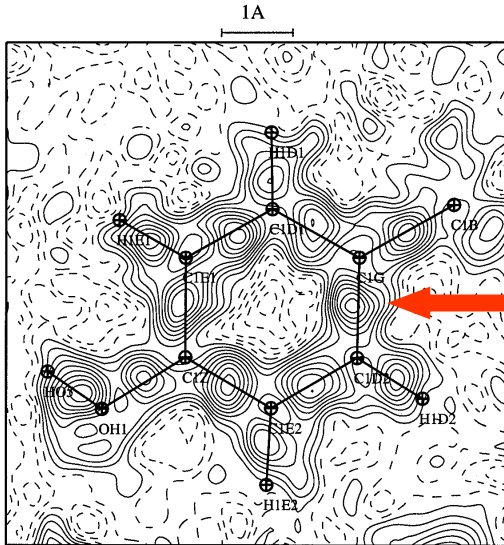
$$\rho_{MODEL}(\mathbf{r}) = \sum_{i=1}^{N_{atoms}} \rho_{atoms}(\mathbf{r})$$

Electron density distribution: multipolar model

- More accurate approximation assumes atoms are bonded: multipolar model



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



ρ_{ATOM} = core electrons

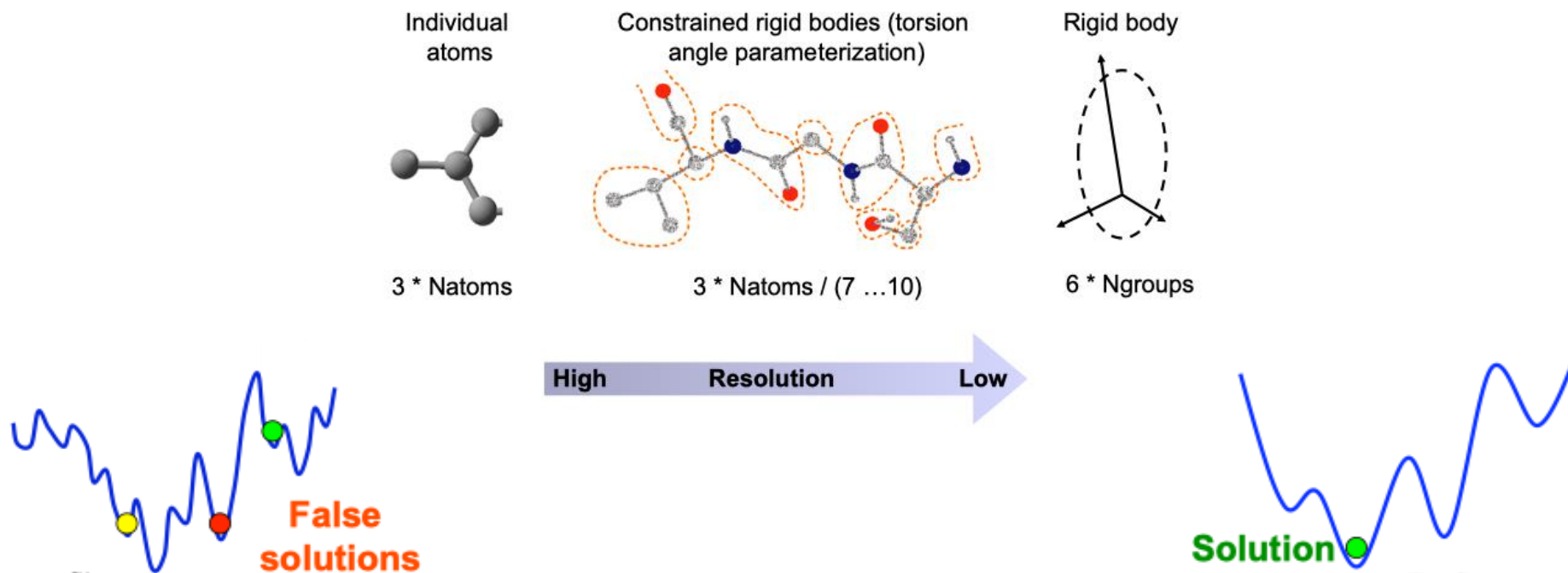
+ valence electrons

+ non-spherical part of the valence electron distribution

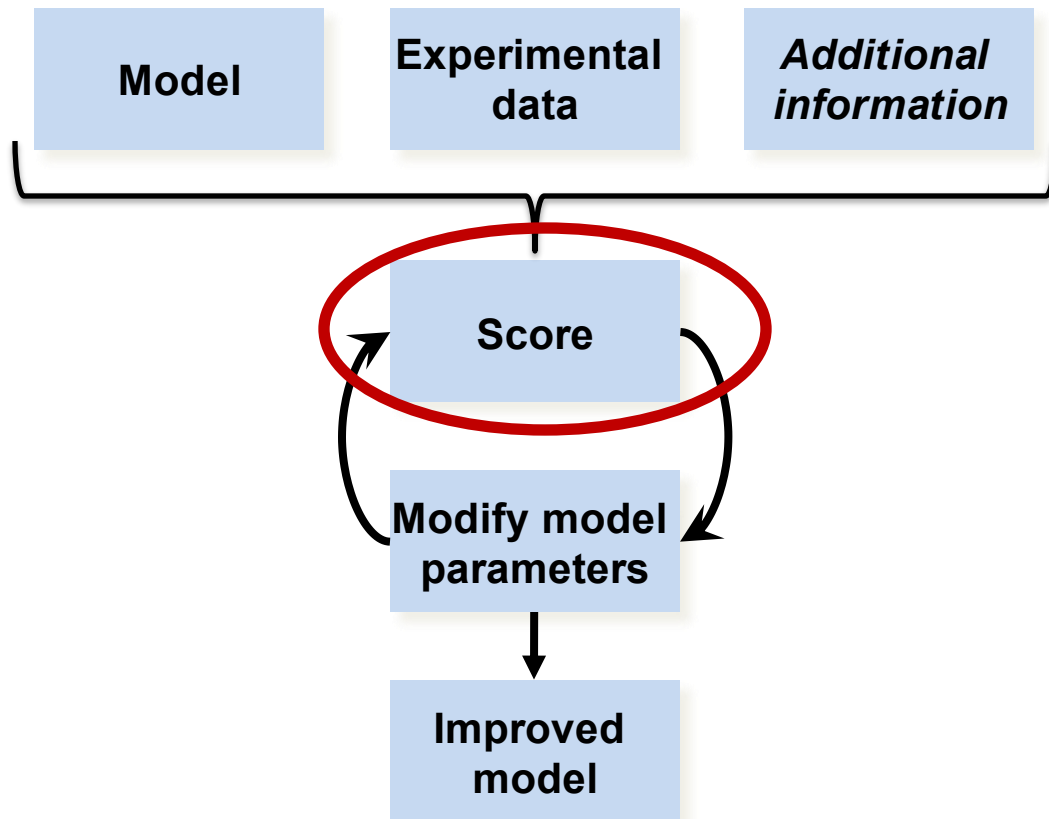
- Used at ultra-high resolution (better than 1 Å)

Model Parameters vs Refinement Parameters

- Refinable parameters can be actual atomic model parameters or can be functions of these parameters (meaning you refine atomic parameters indirectly)

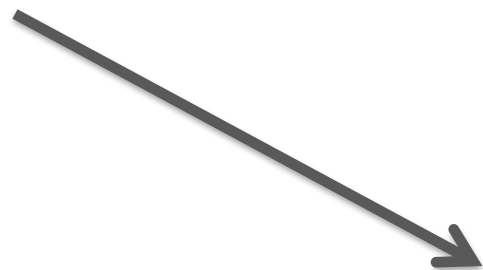


Refinement target function (score)

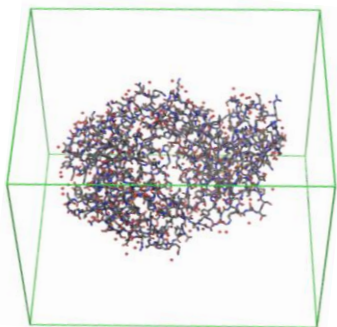


Refinement target function (score)

$$T = T_{\text{DATA}}(F_{\text{OBS}}, F_{\text{MODEL}}) + wT_{\text{RESTRAINTS}}$$

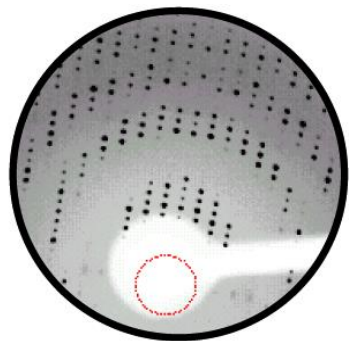


Model



→ **F**_{MODEL}

Data



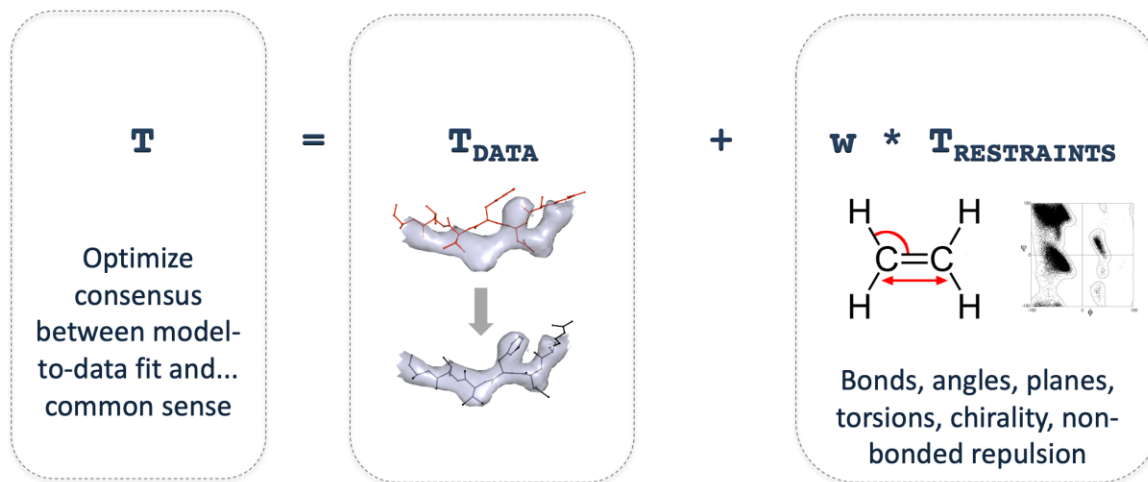
→ **F**_{OBS}

$$T_{\text{DATA}} = \sum_{hkl} (F_{\text{obs}} - F_{\text{model}})^2$$

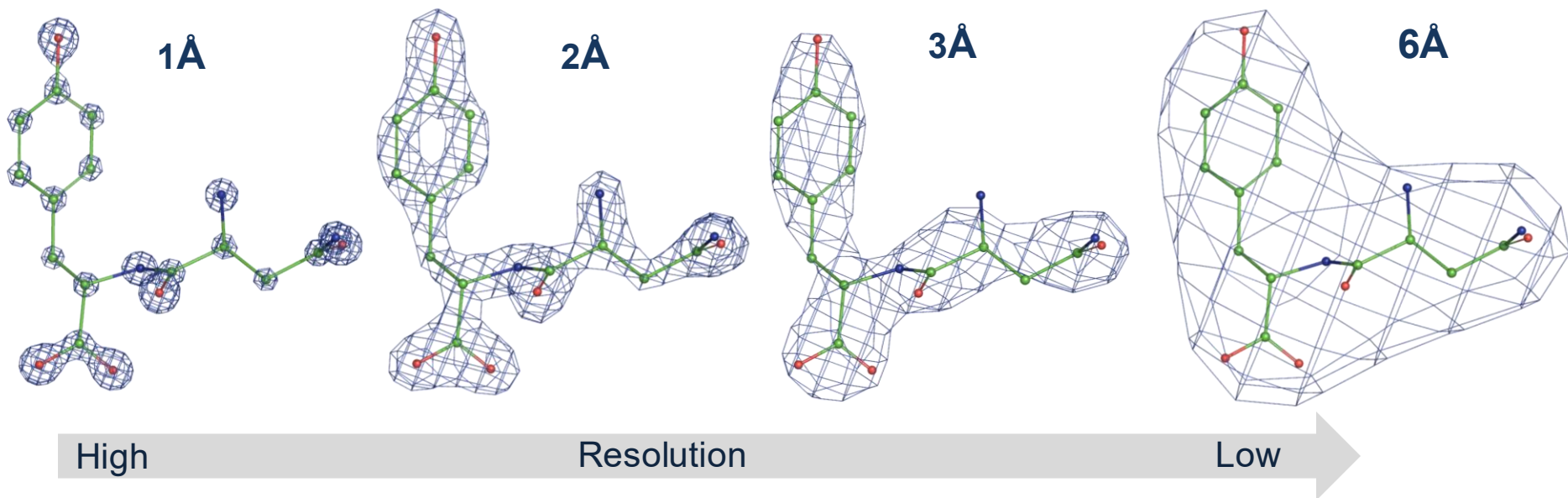
$$T_{\text{DATA}} = \sum_{hkl} \frac{||F_{\text{obs}}| - |F_{\text{model}}||}{|F_{\text{obs}}|}$$

T_{DATA} =
Maximum-Likelihood score

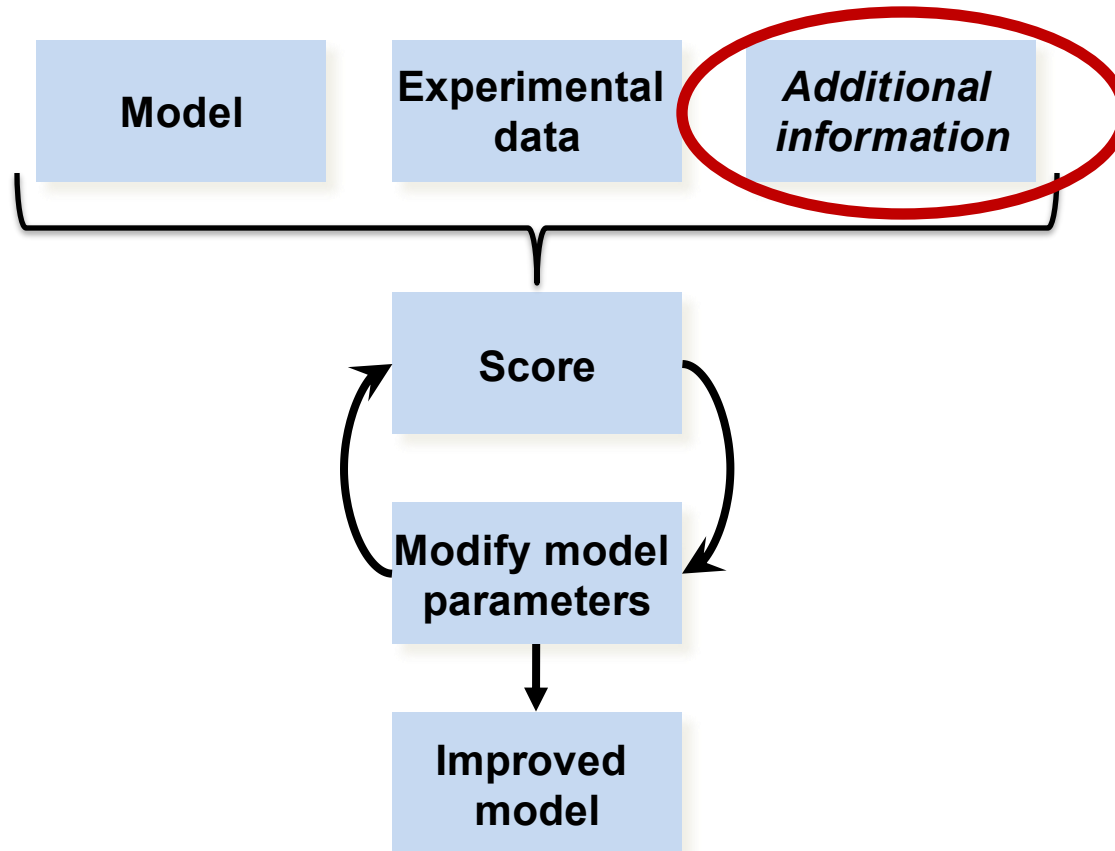
Refinement target and data resolution



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

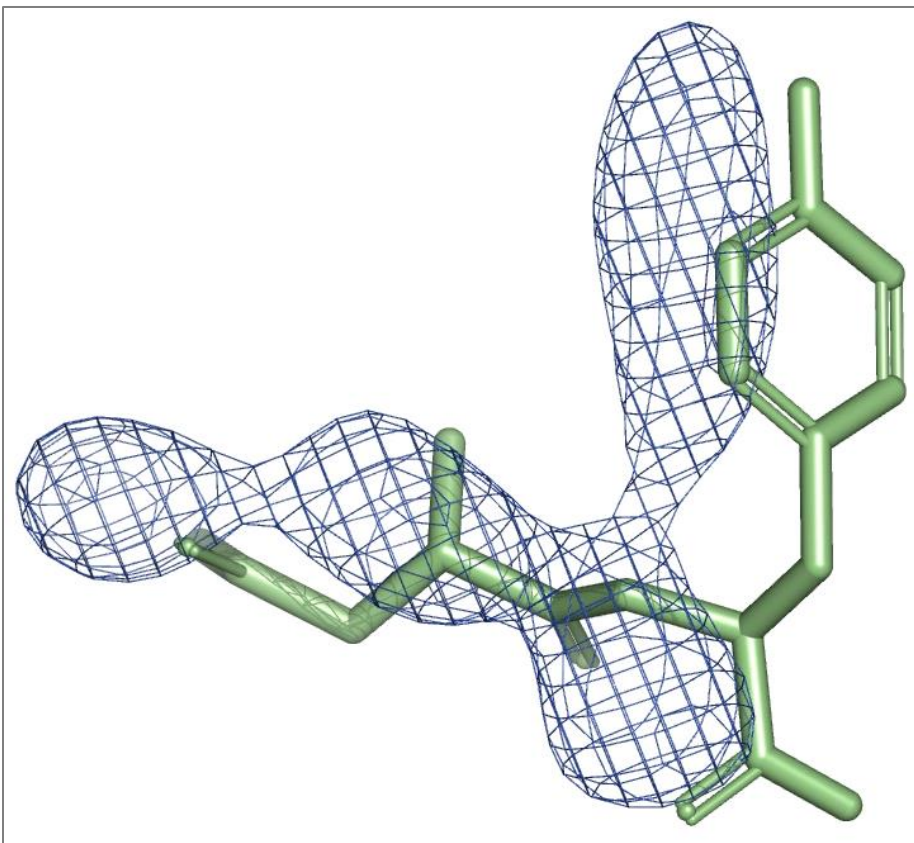


Additional information (restraints, constraints)

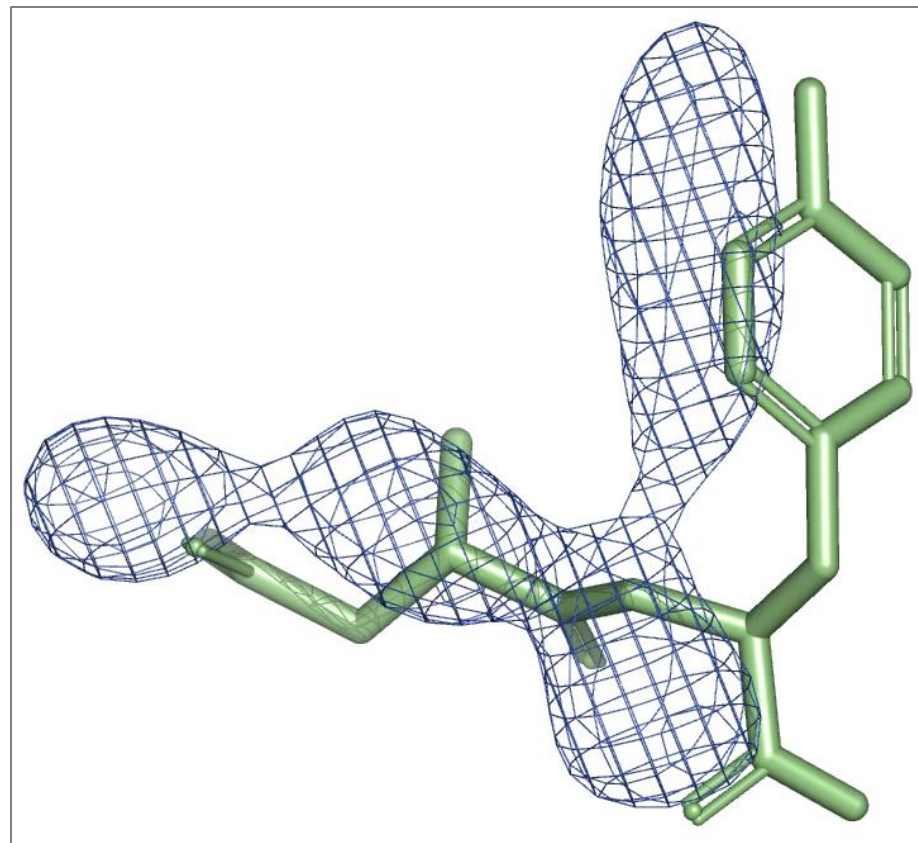


Model refinement with vs no restraints

$$\mathbf{T} = \mathbf{T}_{\text{DATA}} + \mathbf{w} * \mathbf{T}_{\text{RESTRAINTS}}$$



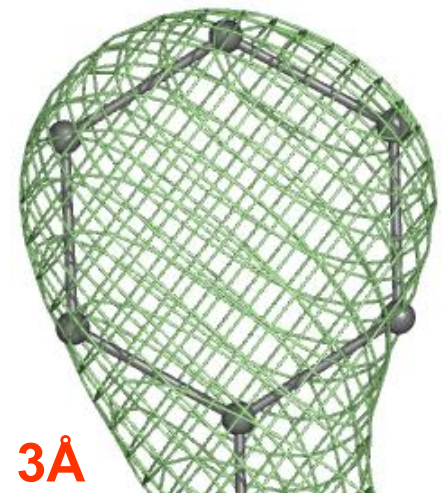
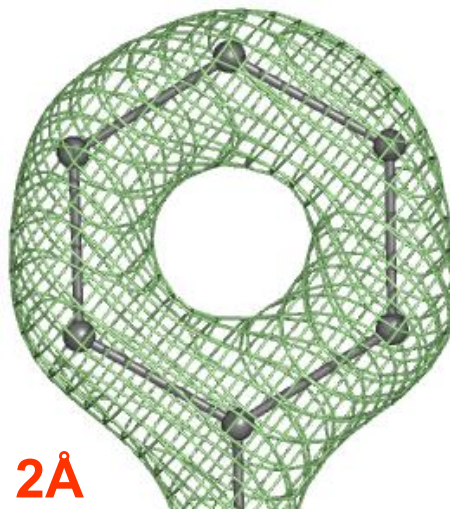
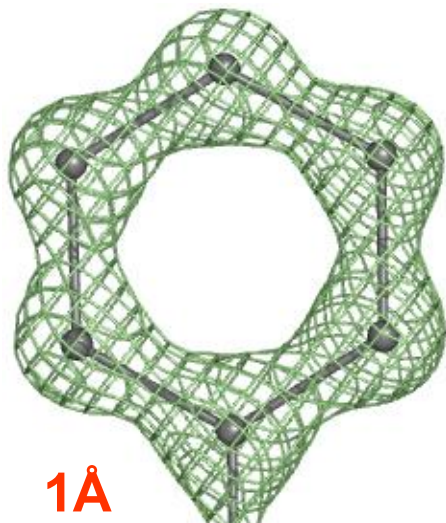
Using restraints



Not using restraints

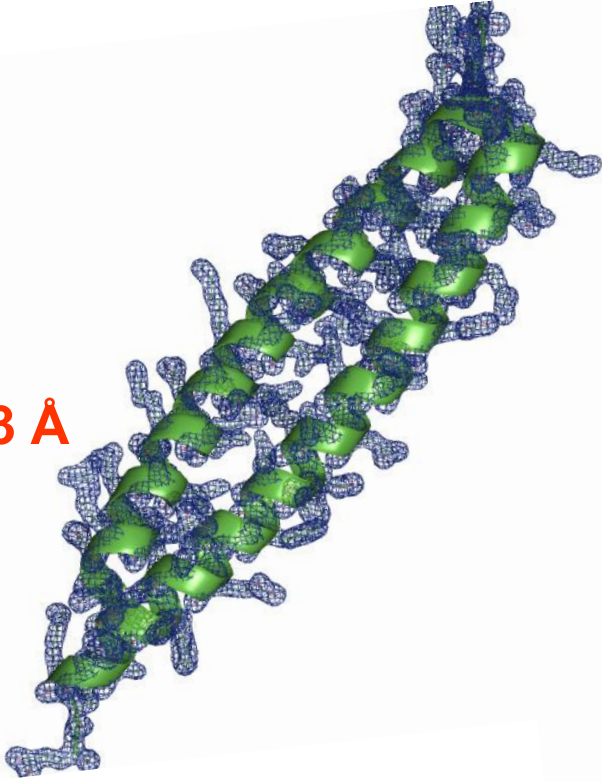
Restraints and constraints

- Why?
 - Experimental data are not perfect:
 - Finite resolution
 - Contains errors
 - Typically less than model parameters (overfitting)
 - Phases are approximate
- Effect of resolution

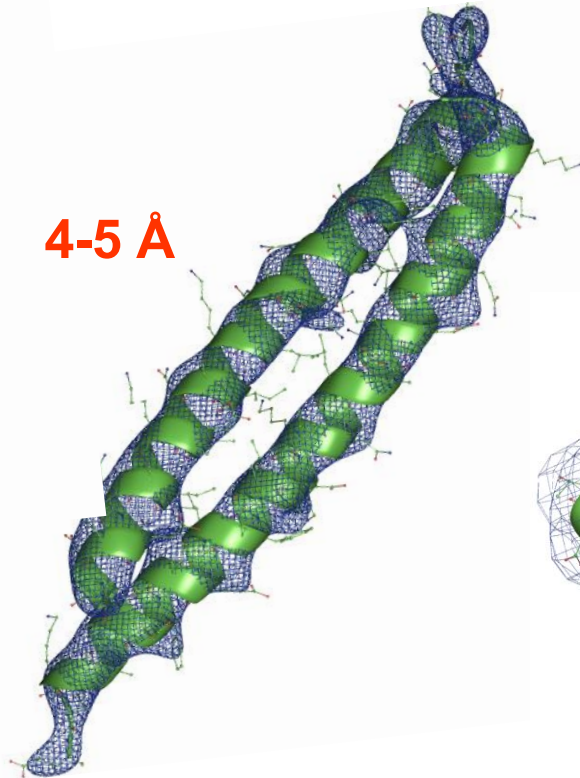


Restraints and constraints

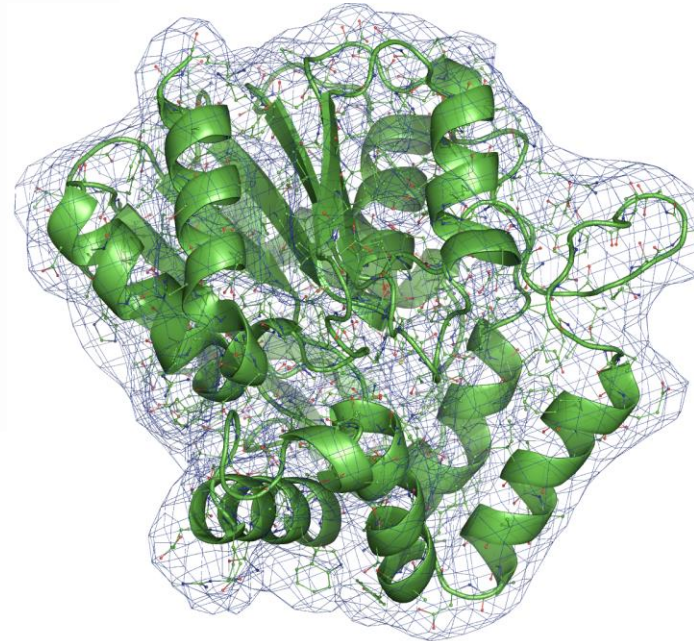
2-3 Å



4-5 Å

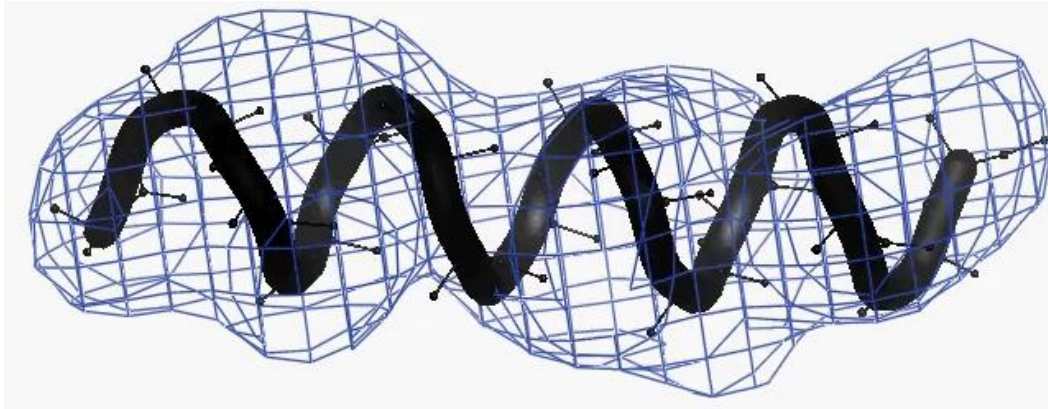


6Å-lower



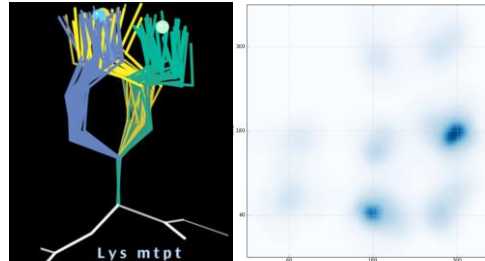
Importance of restraints

- Toy example: refinement of a perfect α -helix into low-res map
 - Standard restraints on covalent geometry isn't sufficient
 - Model geometry deteriorates as result of refinement

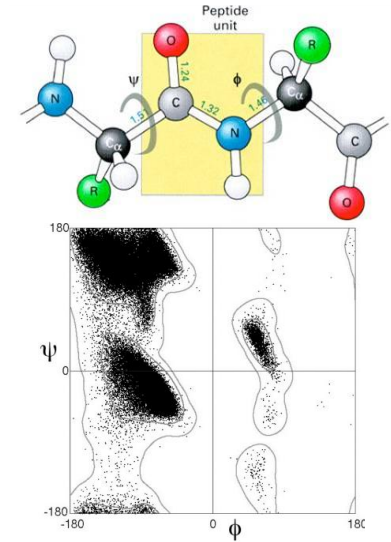


More restraints for low resolution

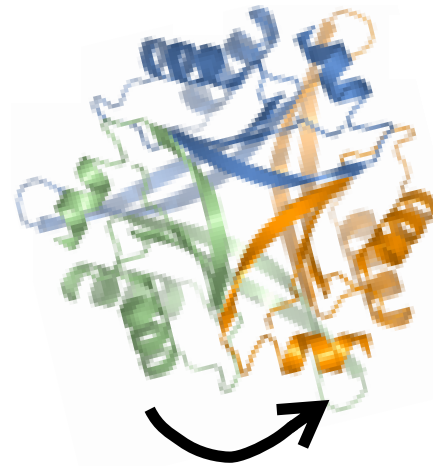
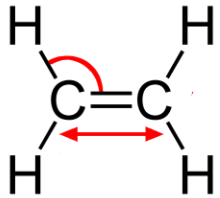
Side chain distributions



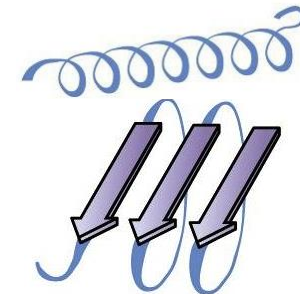
Main chain distributions



Covalent geometry

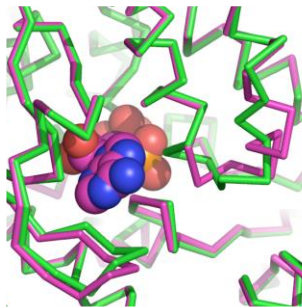


Internal symmetry
(NCS)



Secondary structure

Similar (homologous) structures
(reference model restraints)



Restraints for coordinate refinement

$$T = T_{\text{DATA}}(F_{\text{OBS}}, F_{\text{MODEL}}) + wT_{\text{RESTRAINTS}}$$

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

$$T_{\text{BOND}} = \sum_{\text{all bonded pairs}} w (d_{\text{ideal}} - d_{\text{model}})^2$$



From libraries (CCP4 Monomer
Library or GeoStd in Phenix)

Calculated from actual
atomic model

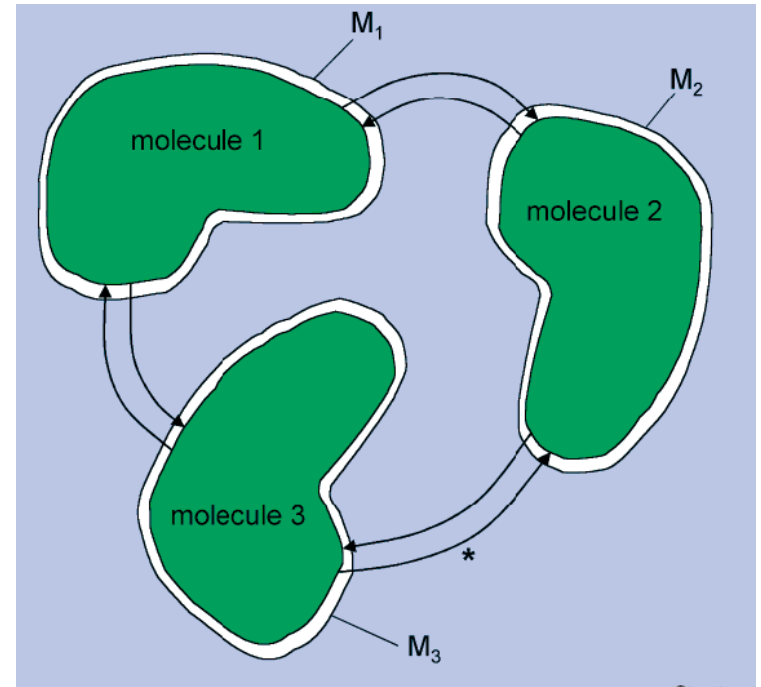
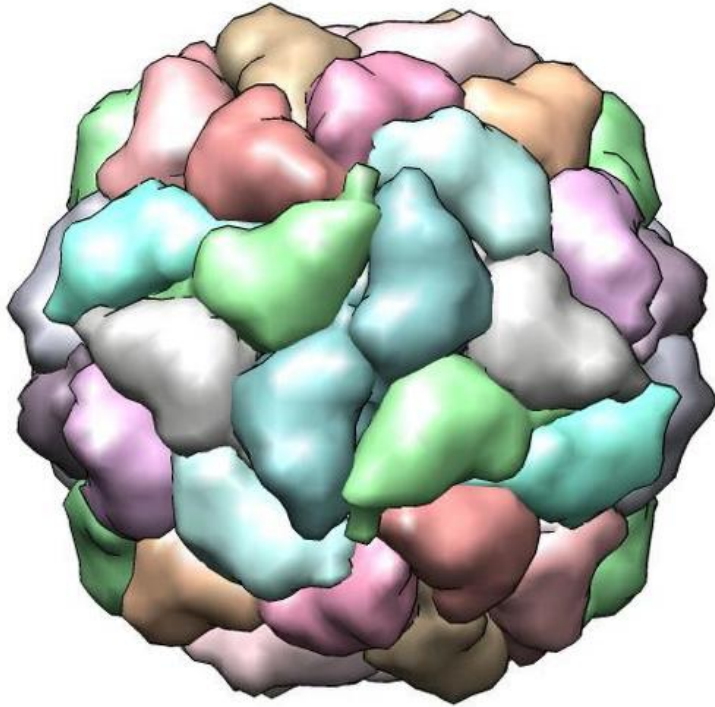
Restraints for coordinate refinement

- The weight w balances data and restraints

$$T = T_{\text{DATA}}(F_{\text{OBS}}, F_{\text{MODEL}}) + wT_{\text{RESTRAINTS}}$$

- Too much restraints: model may not adequately describe the data
- Too much data: model may not obey prior knowledge about model geometry
- Using optimal weight is very important
 - Programs know how to calculate it optimally
 - Sometimes programs fail to calculate it optimally
 - You need to be able to recognize this situation

NCS (internal symmetry): constraints vs restraints



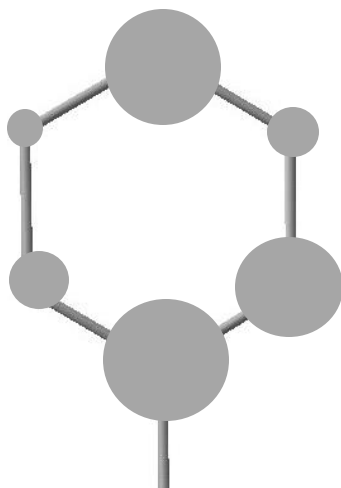
Source: Internet

- **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

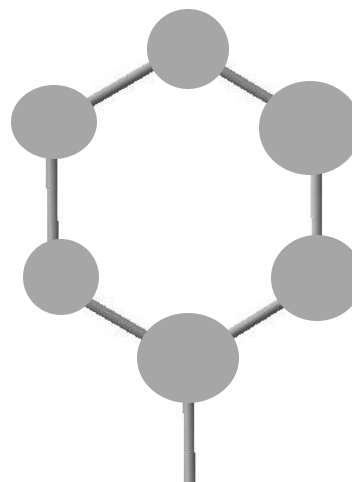
Restraints on Atomic Displacement Parameters

Isotropic

Unlikely

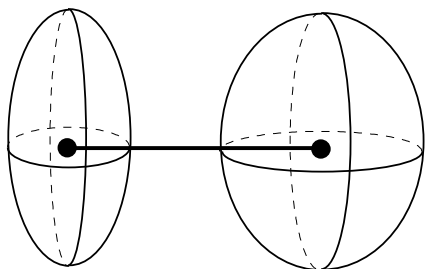


Reasonable

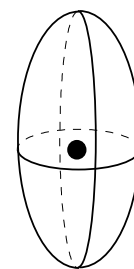
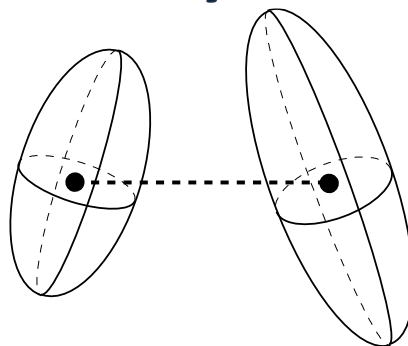


Restraints on Atomic Displacement Parameters

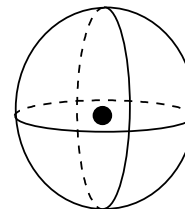
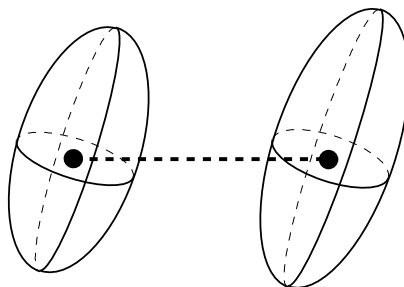
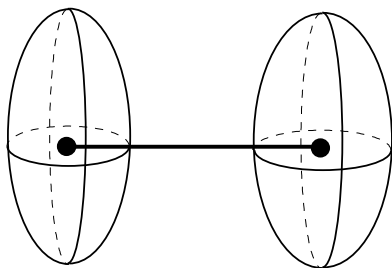
Anisotropic



Unlikely

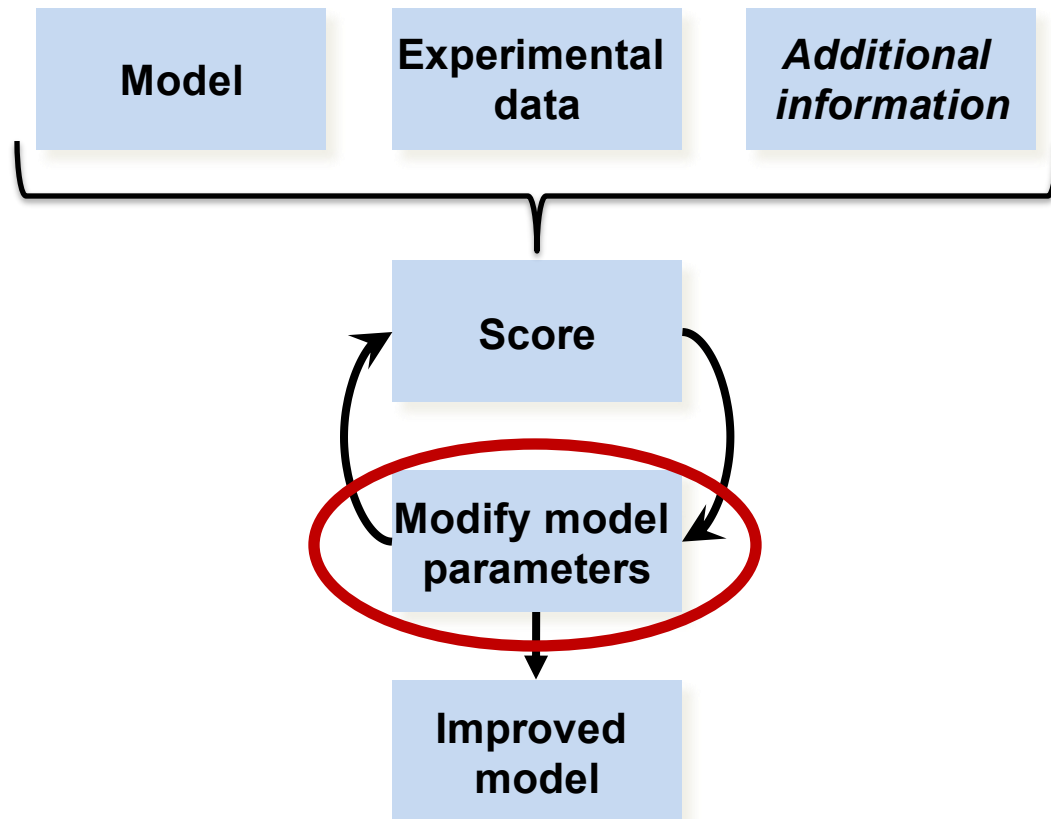


Reasonable



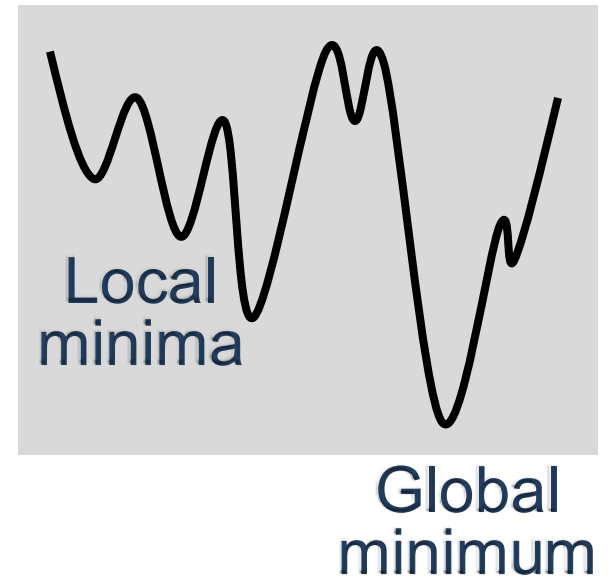
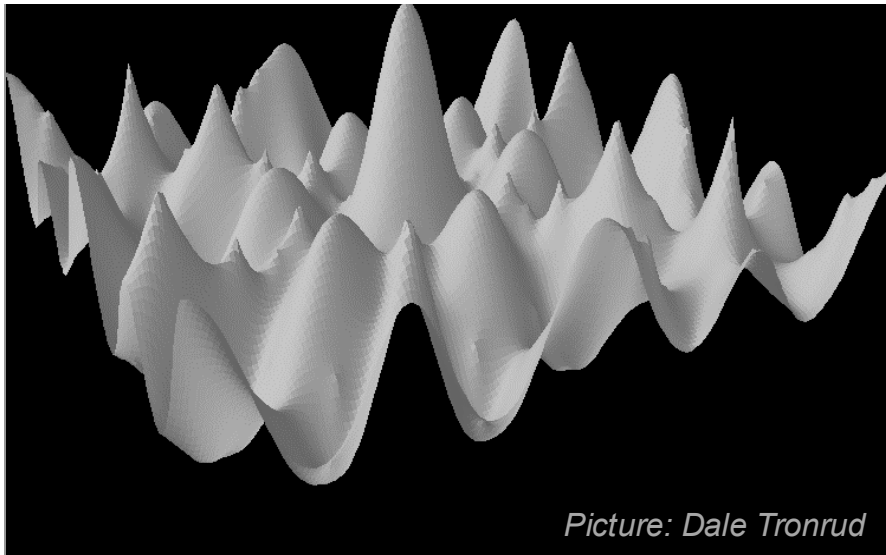
Pictures from Thomas Schneider

Refinement



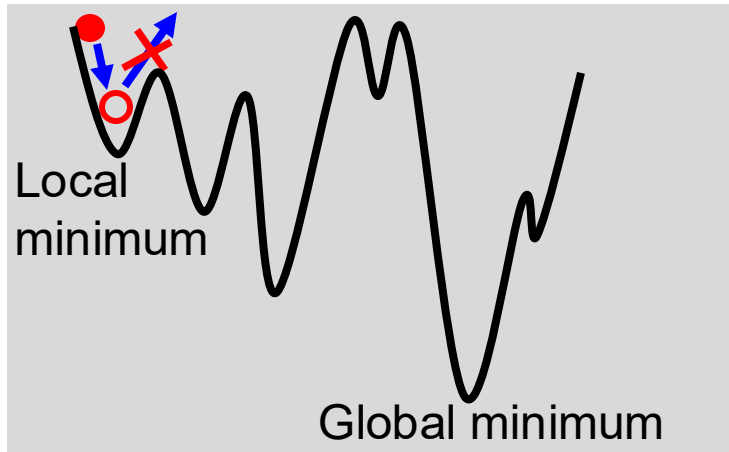
Complexity of refinement target

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

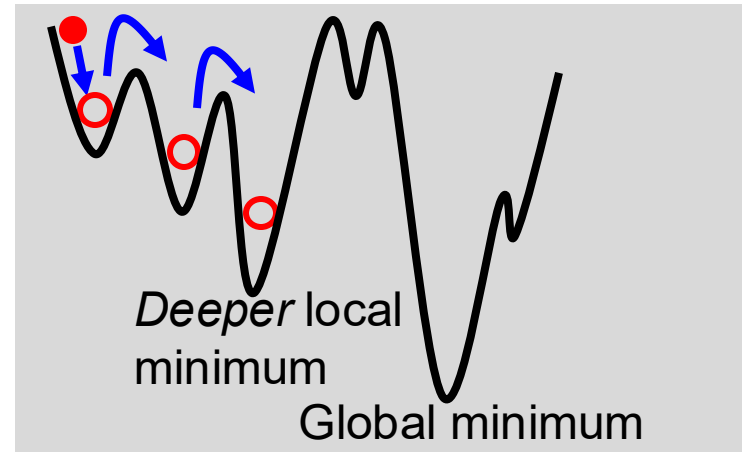


Refinement target optimization methods

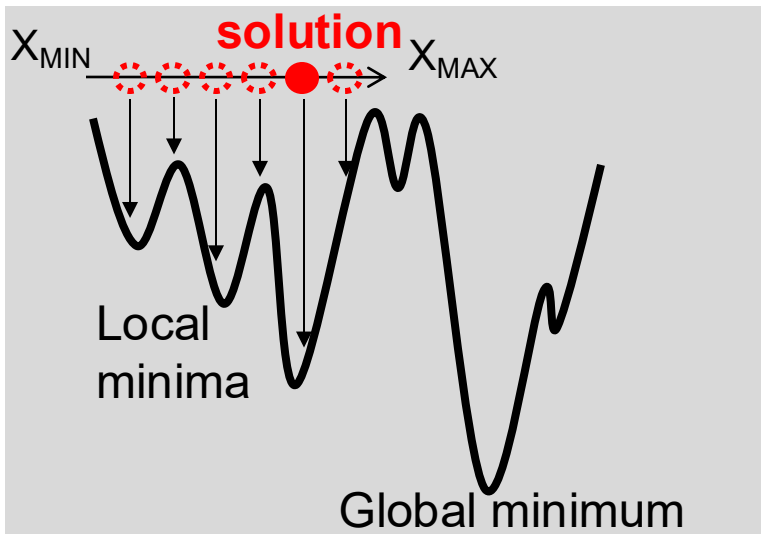
▪ Gradient minimization



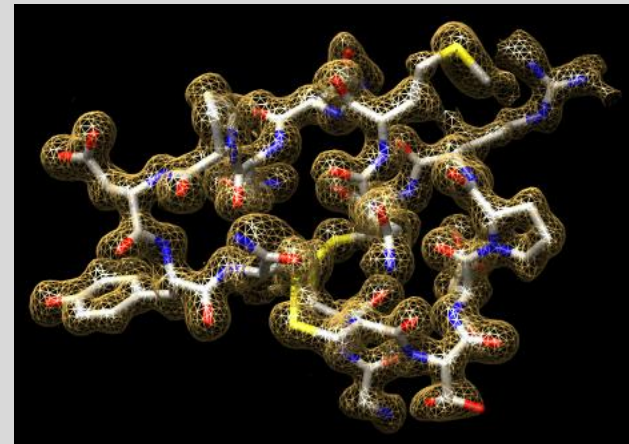
▪ Simulated annealing (SA)



▪ Grid search



▪ Hands & eyes (graphics)



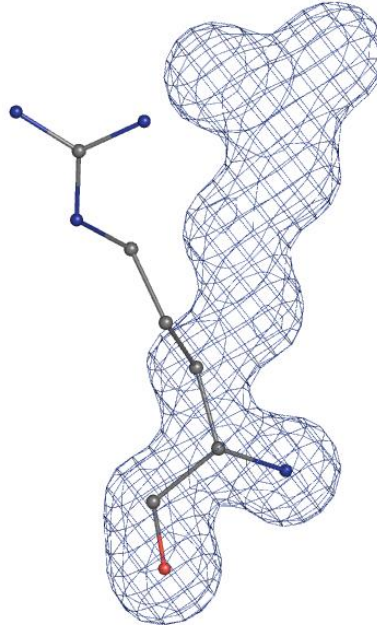
Refinement convergence

Minimization



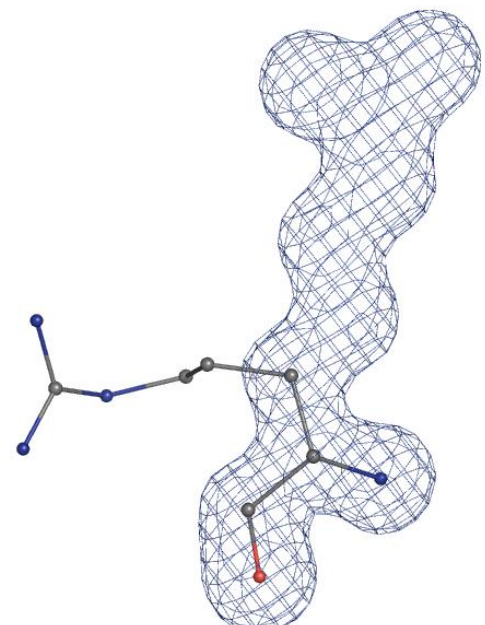
**Minimization or SA
can fix it**

Simulated Annealing



**Beyond
convergence radius
of minimization**

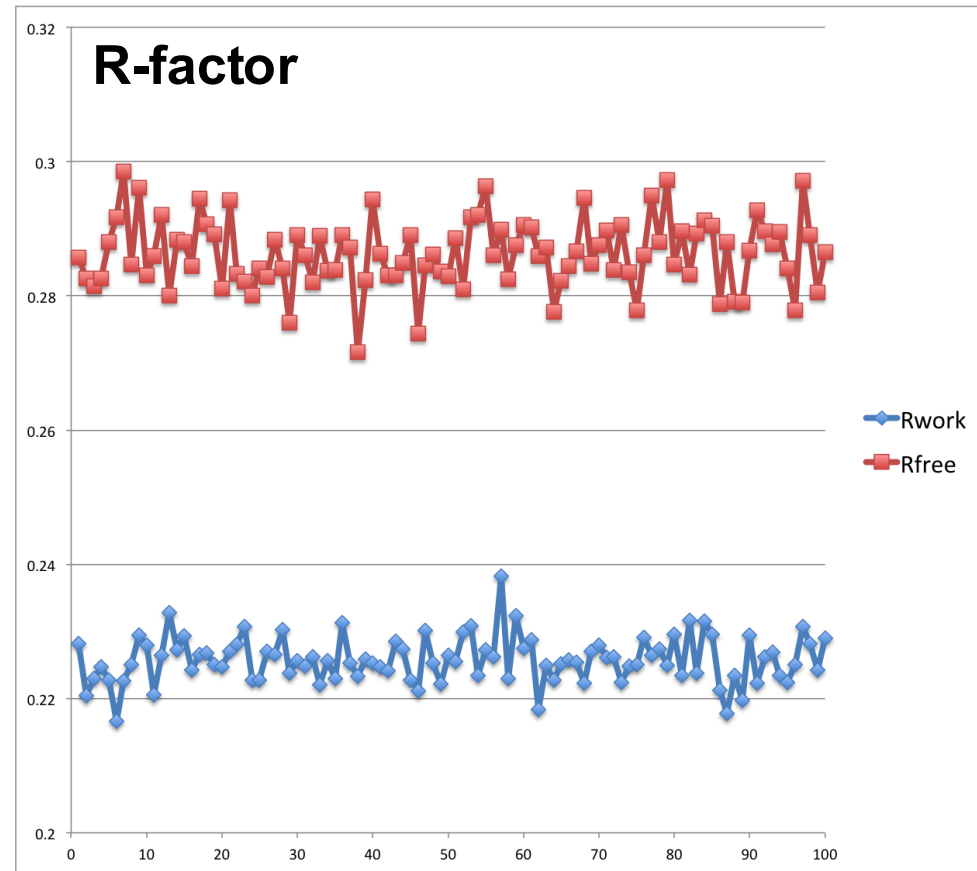
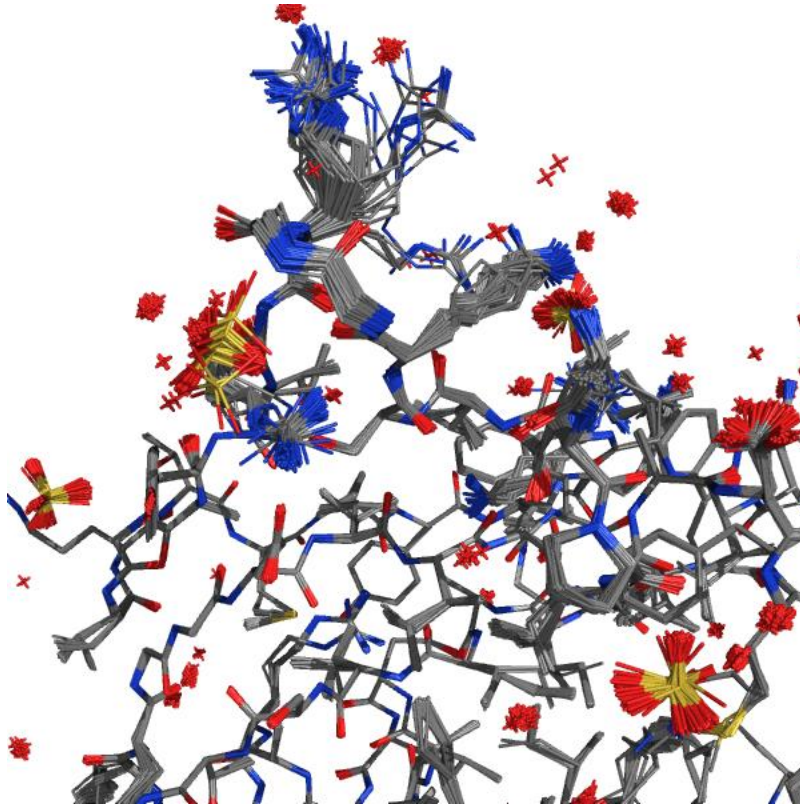
Real-space grid search



**Beyond convergence
radius of
minimization and SA**

Estimating and using uncertainty

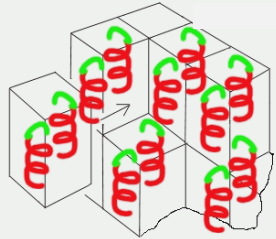
100 identical refinement runs each one starting with slightly perturbed model



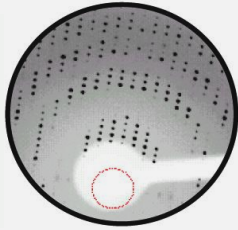
Refinement run

Refinement

Crystallography



Initial model



Experimental
data

A(priori knowledge

Score

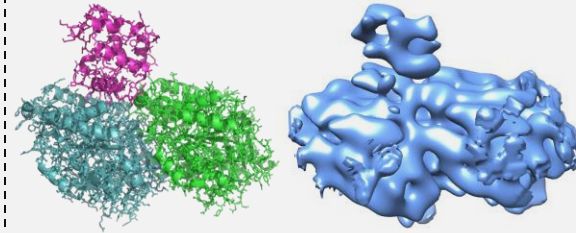
Modify model
parameters

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

Refinement tools in *Phenix*

PHENIX home

Quit Preferences Help Citations Reload last job Coot PyMOL KiNG Other tools Ask for help

Actions Job history

Projects

Show group: All groups Manage...

Select Delete New project Settings

ID	Last modified	# of jobs	R-free
ringer	Sep 07 2016 05:37 ...	2	---
tmp2	Sep 07 2016 05:23 ...	1	---
✓ 5gnn	Sep 07 2016 08:42 ...	1	---
debug1	Sep 05 2016 10:51 ...	2	0.0086
tmp4	Aug 18 2016 07:23 ...	2	---
testing	Aug 11 2016 01:54 ...	1	---
mich	Jul 29 2016 12:47 ...	1	---
almu	Jul 28 2016 10:58 ...	1	---
rchen	Jul 22 2016 11:10 ...	1	---
milya	Jul 15 2016 12:36 ...	2	---
SEM	Jul 14 2016 05:20 ...	14	0.1570





Data analysis

Experimental phasing

Molecular replacement

Model building

Refinement

-  **phenix.refine**
Automated X-ray and/or neutron refinement
-  **Real-space refinement**
Automated real-space refinement
-  **Neutron refinement [alpha]**
Alternate phenix.refine interface customized for neutron data
-  **DEN refinement [alpha]**
Deformable elastic network refinement using simulated annealing and molecular replacement strategies

Current directory: /Users/pafonine/Desktop/work/tmp Browse...

PHENIX version dev-svn-000 Project: 5gnn

Atomic model refinement: crystallography vs cryo-EM

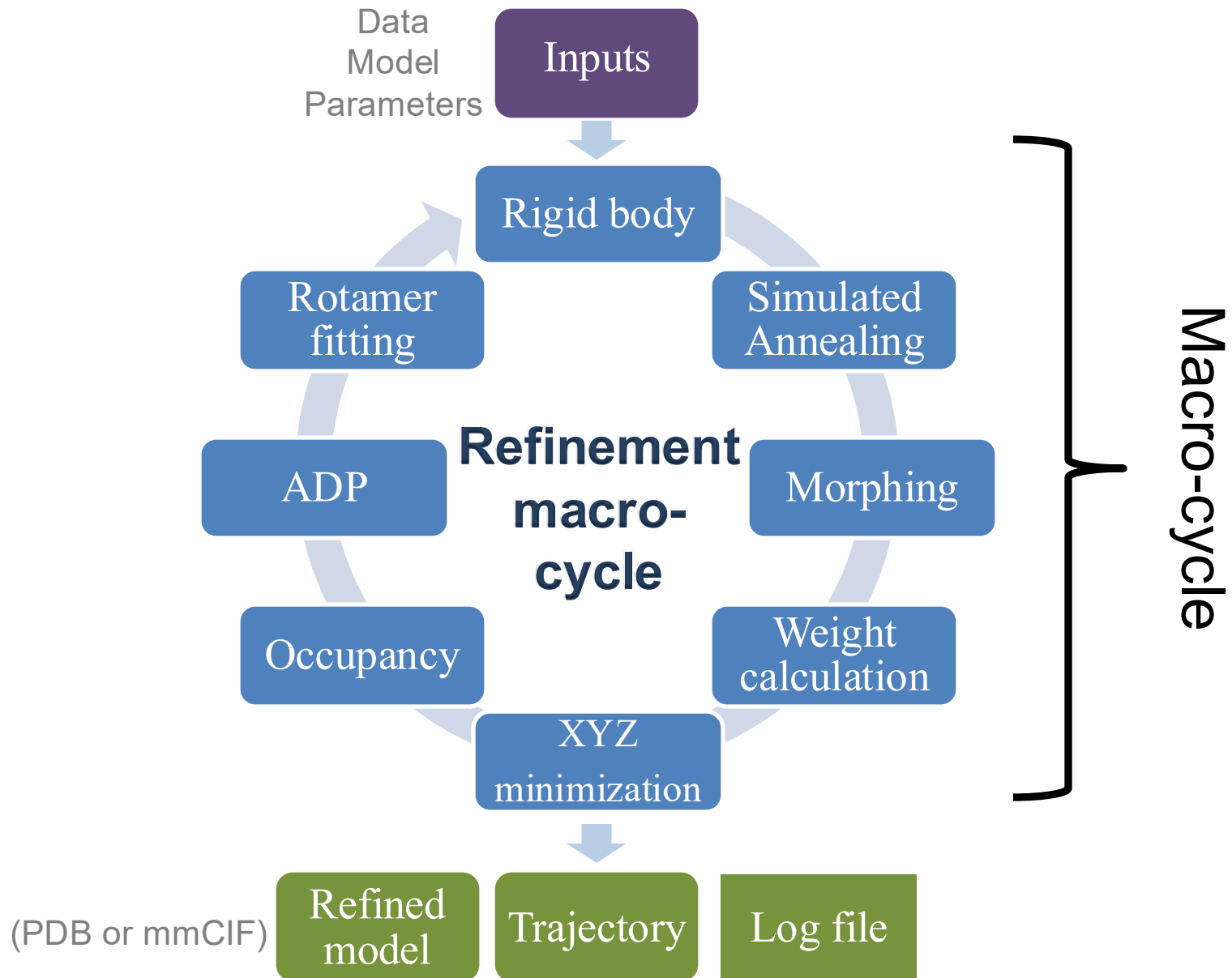
Crystallographic refinement

- 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



Reading



RESEARCH PAPERS

Acta Cryst. (2018). **D74**, 531-544
<https://doi.org/10.1107/S2059798318006551>

Cited by **672**

Part of *CCP-EM Spring Symposium 2017*



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

P. V. Afonine^{ID}, B. K. Poon^{ID}, R. J. Read^{ID}, O. V. Sobolev^{ID}, T. C. Terwilliger^{ID}, A. Urzhumtsev and P. D. Adams^{ID}



RESEARCH PAPERS

Acta Cryst. (2012). **D68**, 352-367
<https://doi.org/10.1107/S0907444912001308>

Cited by **2576**

Part of *CCP4 Study Weekend 2011*



Towards automated crystallographic structure refinement with *phenix.refine*

P. V. Afonine^{ID}, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty^{ID}, M. Mustyakimov, T. C. Terwilliger^{ID}, A. Urzhumtsev, P. H. Zwart^{ID} and P. D. Adams^{ID}

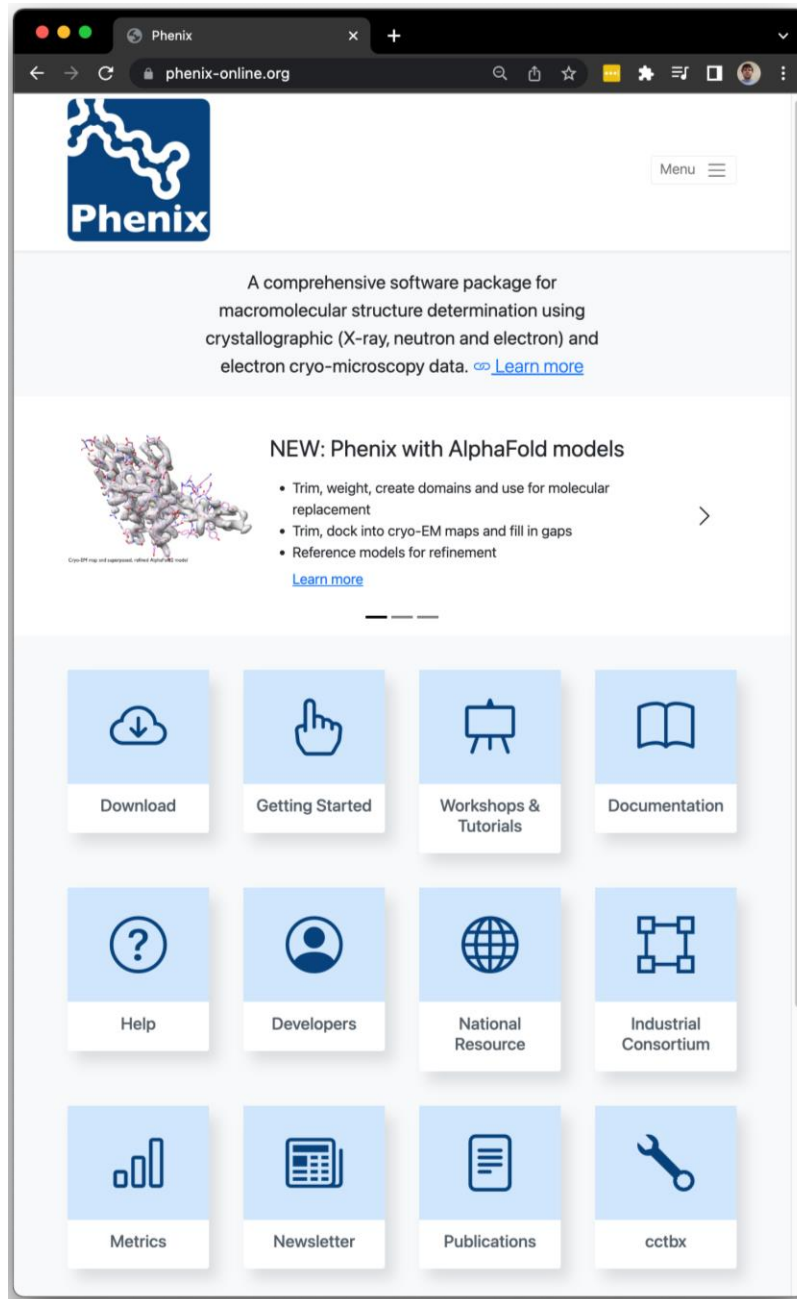
phenix.refine is a program within the *PHENIX* package that supports crystallographic structure



OPEN  ACCESS



Phenix resources



Phenix paper

Video tutorials

Documentation

Relevant papers

Bi-annual newsletters

Slides from workshops