Model Refinement

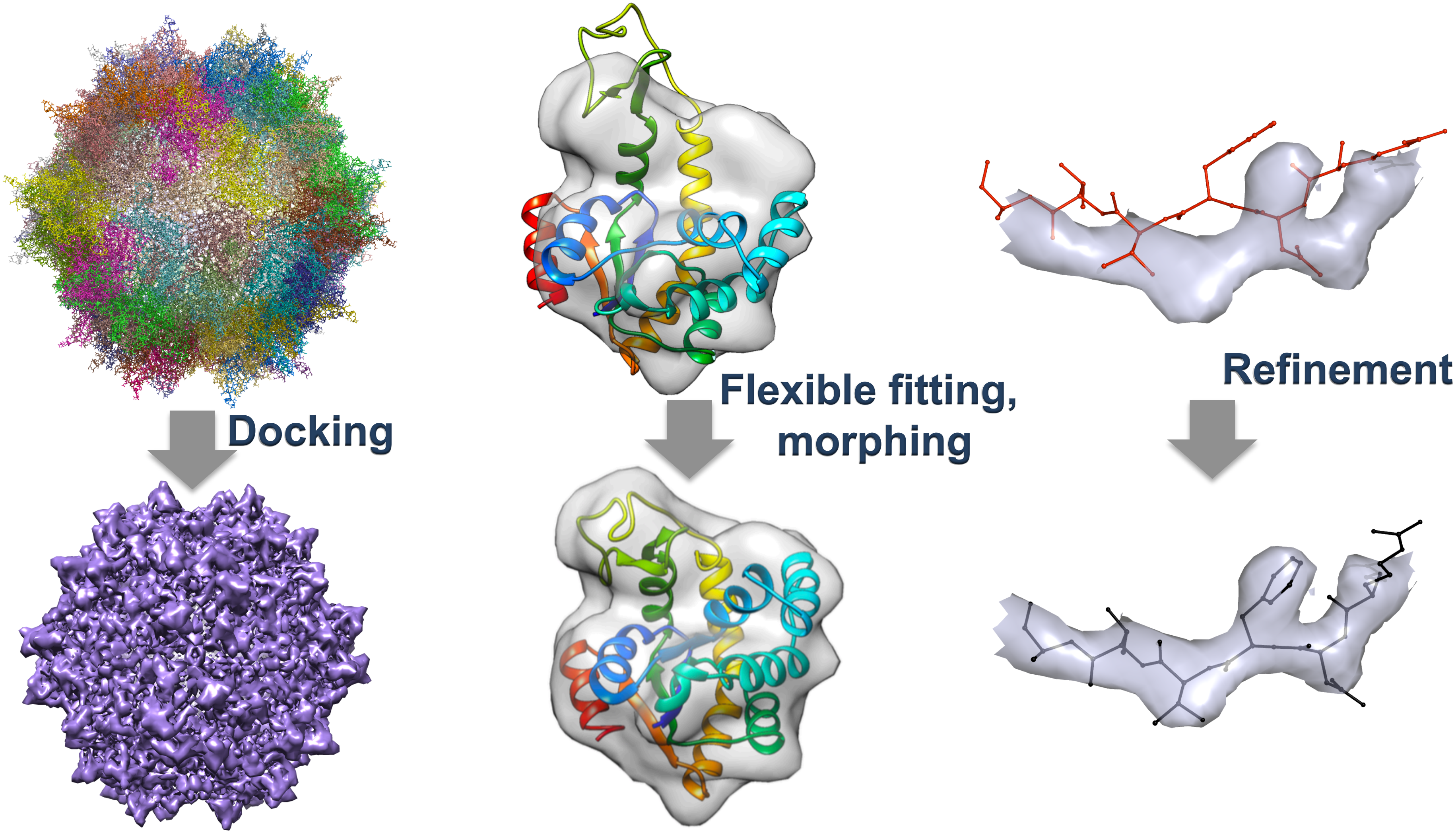
Oleg Sobolev

Phenix team

Lawrence Berkeley National Lab, California, USA

August 22, 2023
IUCr 2023, Melbourne

Model refinement vs other model fitting tools



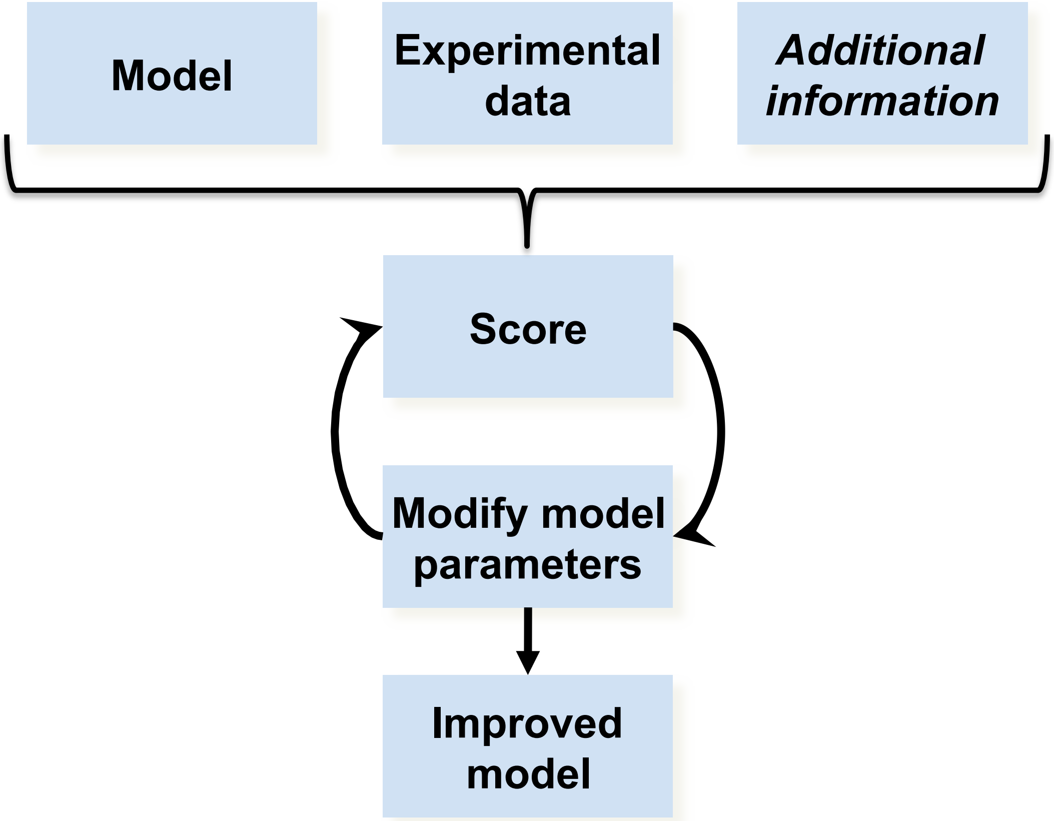
Docking

Flexible fitting, morphing

Refinement

- All the above move model to achieve better fit. The difference is: by how much

Model refinement



Model

Experimental data

Additional information

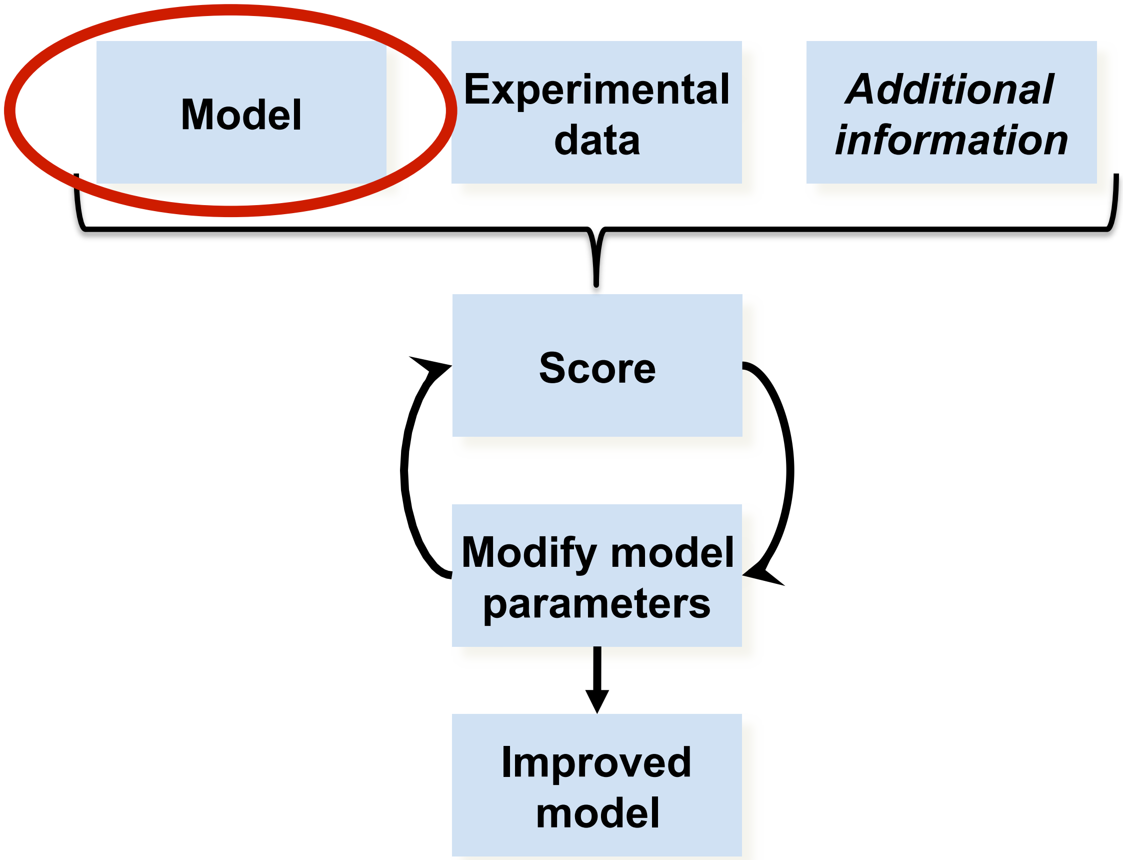
Score

Modify model parameters

Improved model

Refinement – optimization process of fitting model parameters to experimental data

Refinement: model

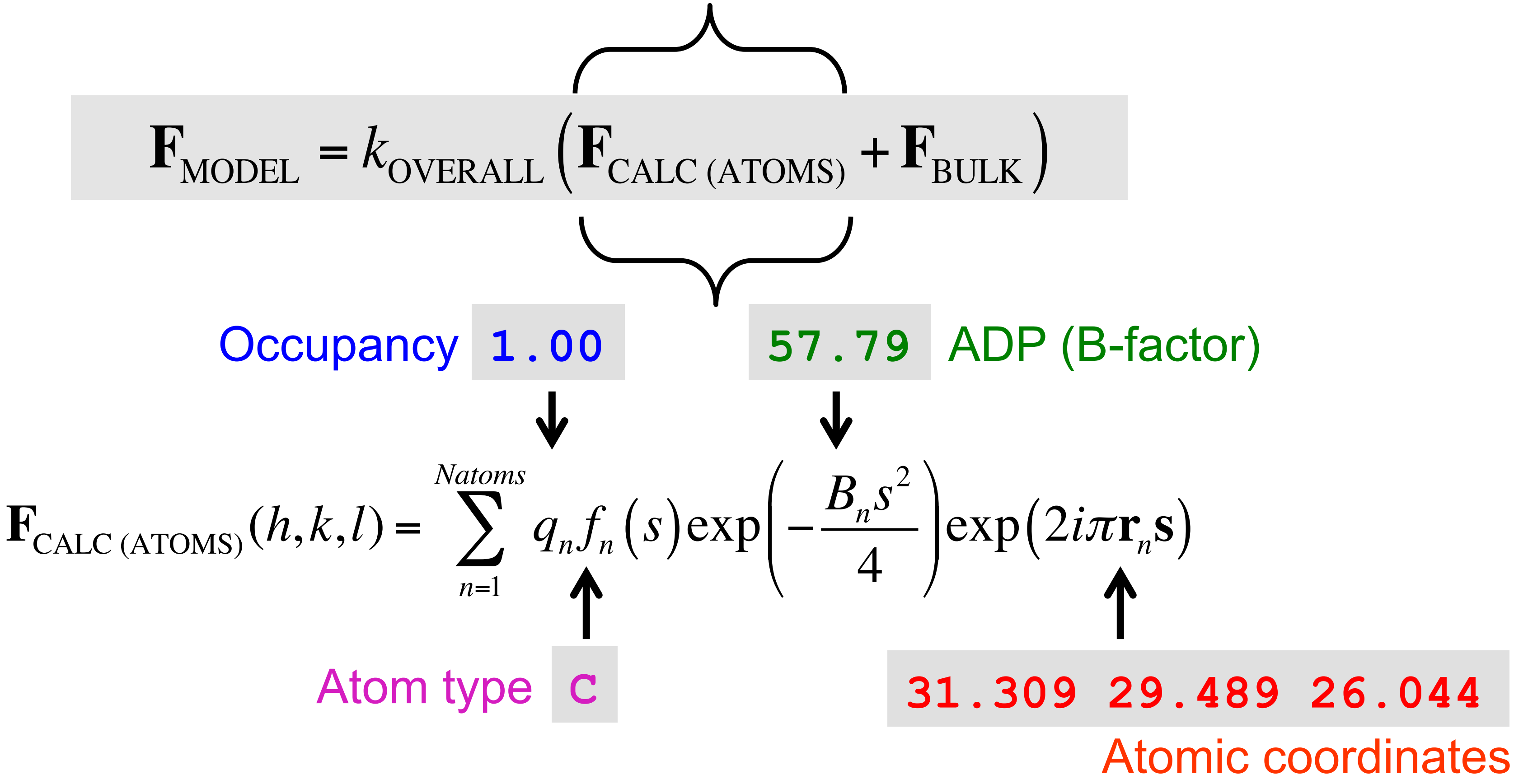


Atomic model parameters

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)



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

Occupancy `1.00`

`57.79` ADP (B-factor)

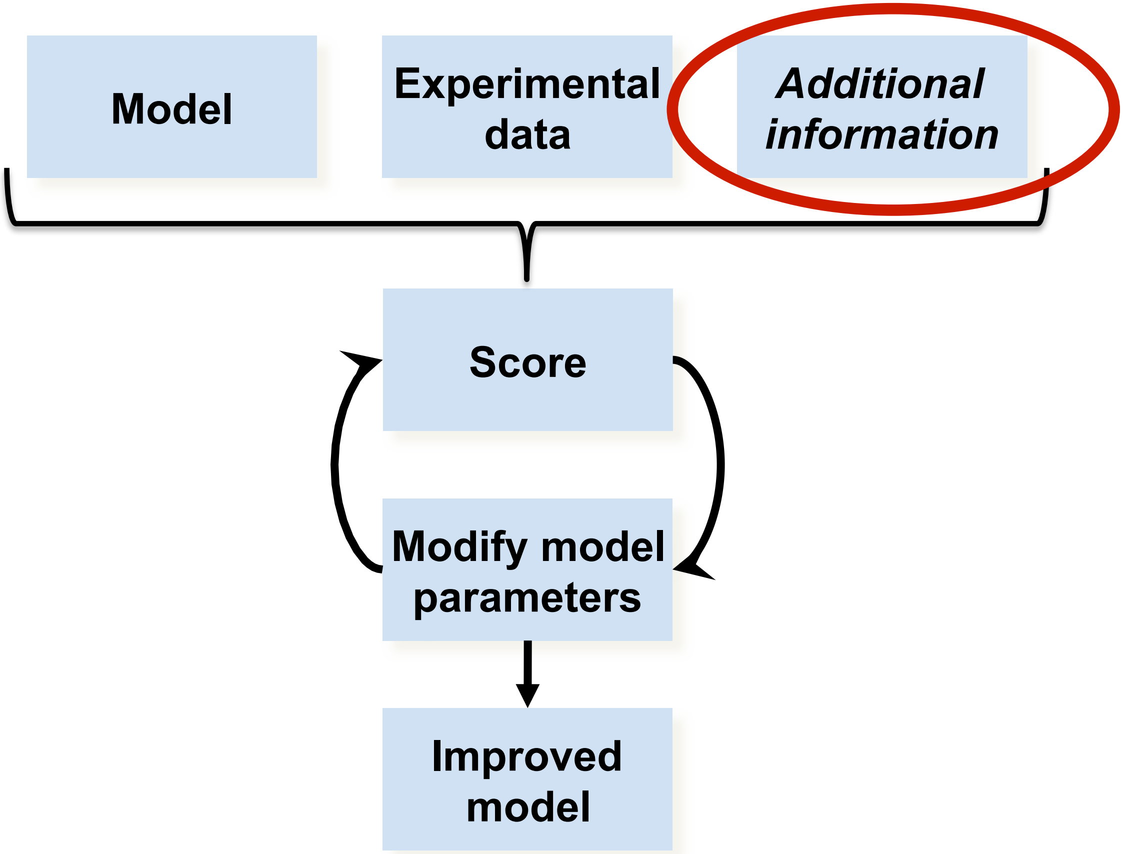
$$\mathbf{F}_{\mathrm{CALC\ (ATOMS)}}(h,k,l) = \sum_{n=1}^{Natoms} q_n f_n(s)\exp\left(-\frac{B_n s^2}{4}\right)\exp\left(2i\pi\mathbf{r}_n\mathbf{s}\right)$$

Atom type `C`

`31.309 29.489 26.044`

Atomic coordinates

Additional information (restraints, constraints)



Model

Experimental data

Additional information

Score

Modify model parameters

Improved model

Restraints for coordinate refinement

- The weight *w* balances data and restraints

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

- Too much restraints: model may not adequately describe the data

- Too much data: model may not obey prior knowledge about model geometry

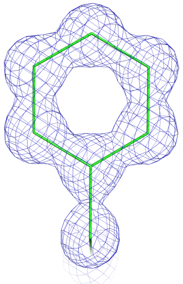
- Using optimal weight is very important
  - Programs know how to calculate it optimally
  - Sometimes programs fail to calculate it optimally
    - You need to be able to recognize this situation

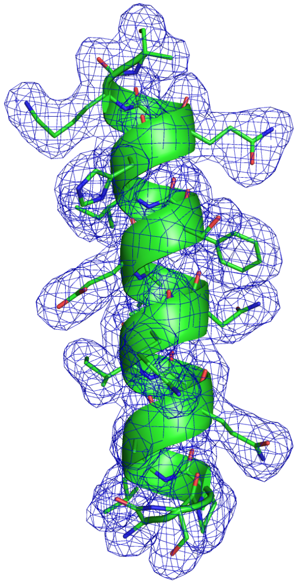
Restraints in structure refinement

- Refinement target is usually a weighted sum of experimental data and *a priori* chemical knowledge terms

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

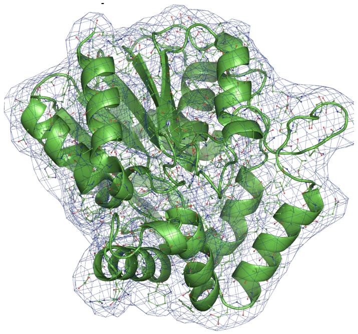
- At ultra-high resolution (<1Å) an unrestrained refinement sometimes may be possible.





- At ‘typical’resolutions (1-3Å) *standard* restraints are necessary:
  covalent bond, angles, etc

- At lower resolution (lower than 3Å) more restraints needed:
  NCS, Secondary Structure, Ramachandran, …



Restraints for coordinate refinement

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

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

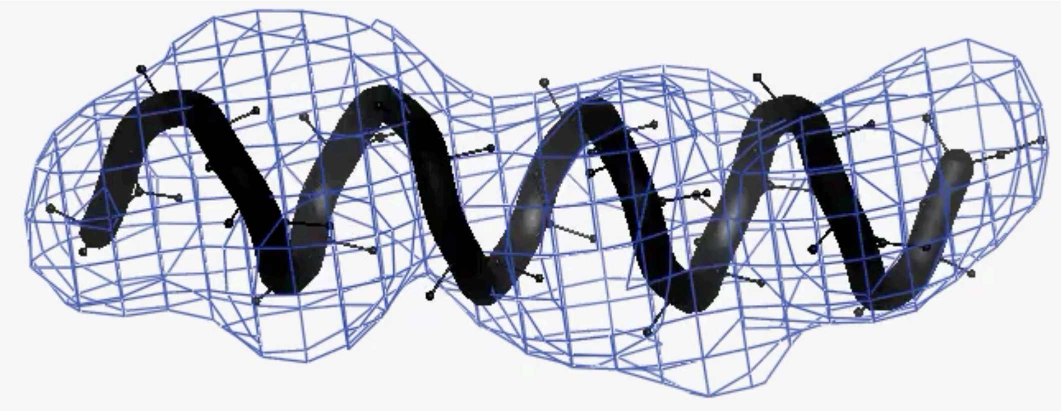
$$\mathrm{T_{BOND}} = \Sigma_{\text{all bonded pairs}}\ \mathrm{w}\ (\mathrm{d_{ideal}} - \mathrm{d_{model}})^2$$

From libraries (CCP4 Monomer Library or GeoStd in Phenix)

Calculated from actual atomic model

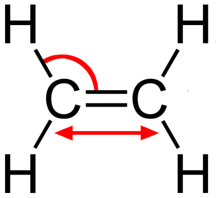
Importance of additional restraints

- 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

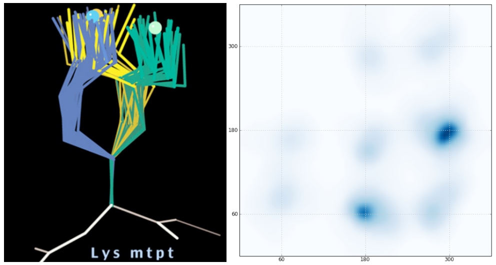


Restraints for low resolution

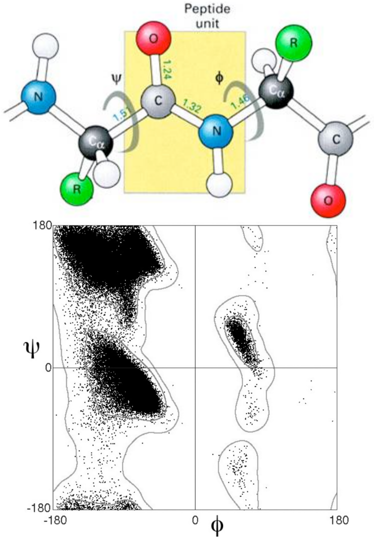
Covalent geometry



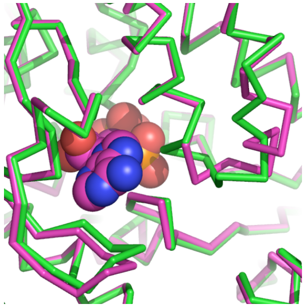
Side chain distributions



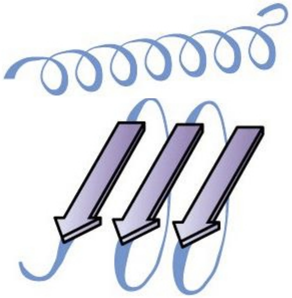
Main chain distributions (Ramachandran)



Internal symmetry (NCS)



Similar (homologous) structures (reference model restraints)

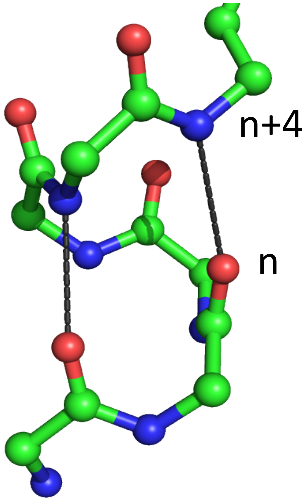


Secondary structure

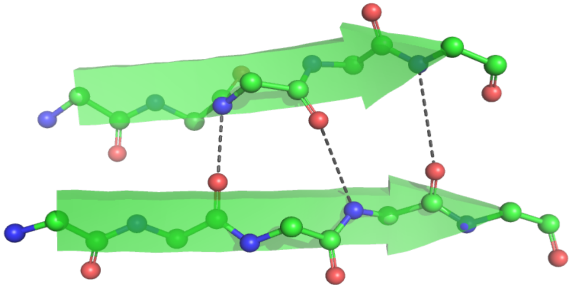
Secondary structure restraints

Proteins

Helices



Sheets

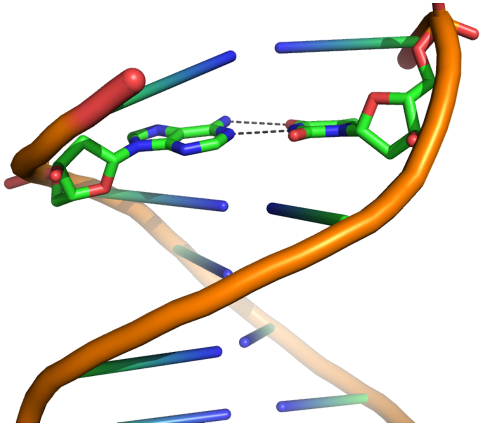


Restraints used

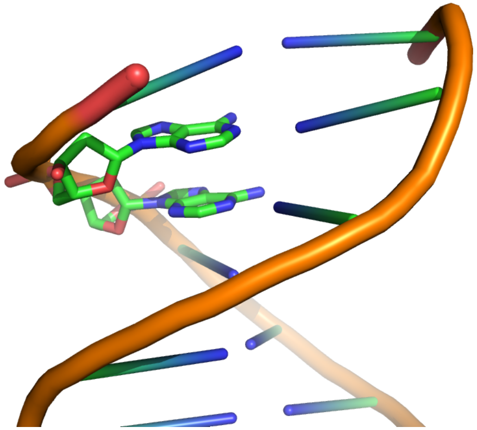
H-bond lengths
H-bond angles

Nucleic Acids

Base pairs



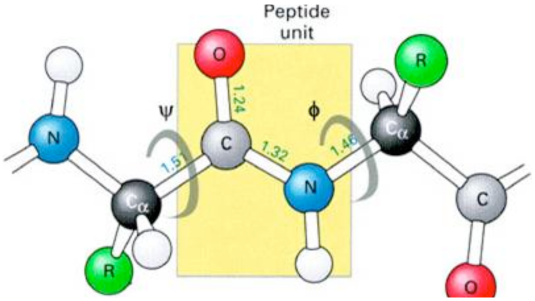
Stacking pairs



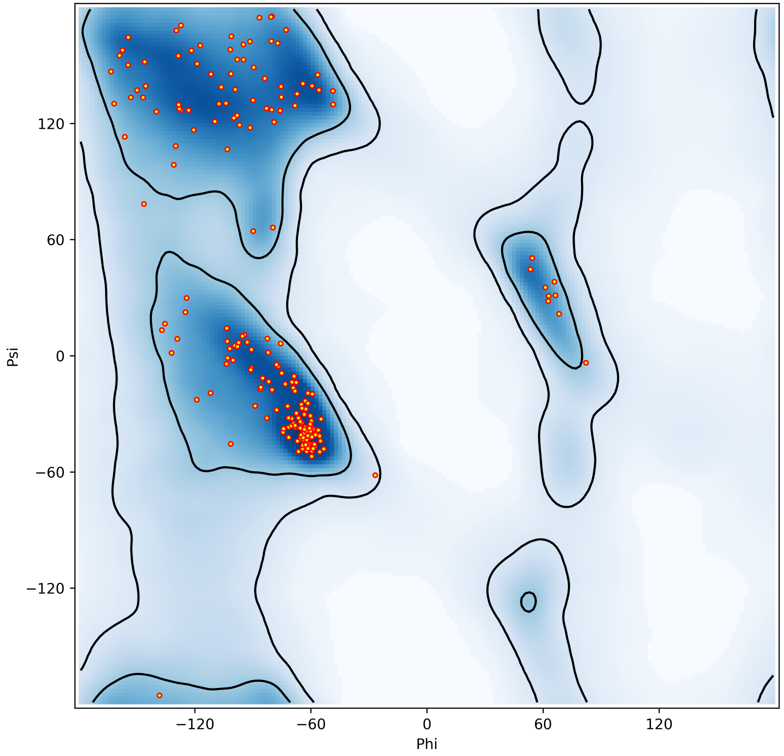
H-bond lengths
H-bond angles
Planarity
Parallelity

Ramachandran plot restraints

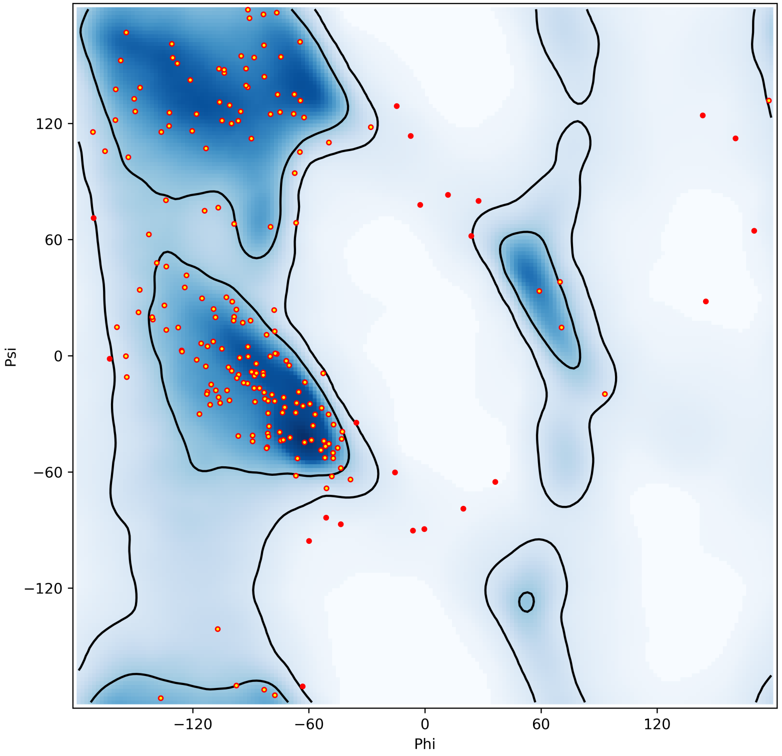
- Ramachandran plot restraints
  - Use to stop outliers from occurring



Before refinement

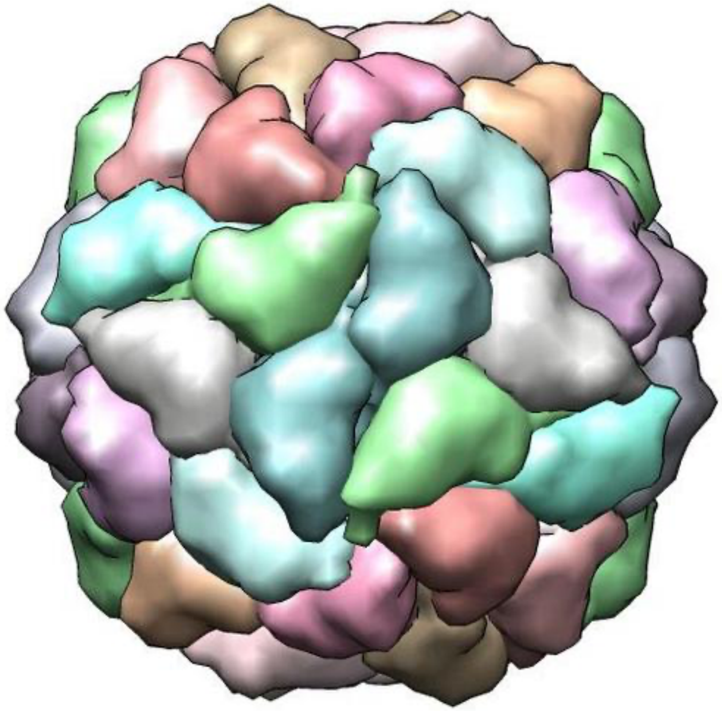


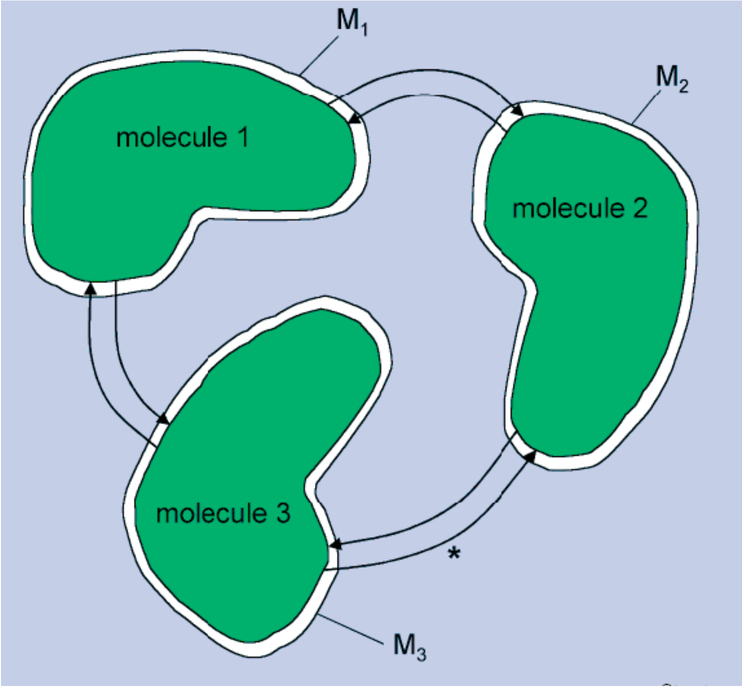
After refinement



Good idea to use Ramachandran plot restraints!

NCS (internal symmetry): constraints vs restraints



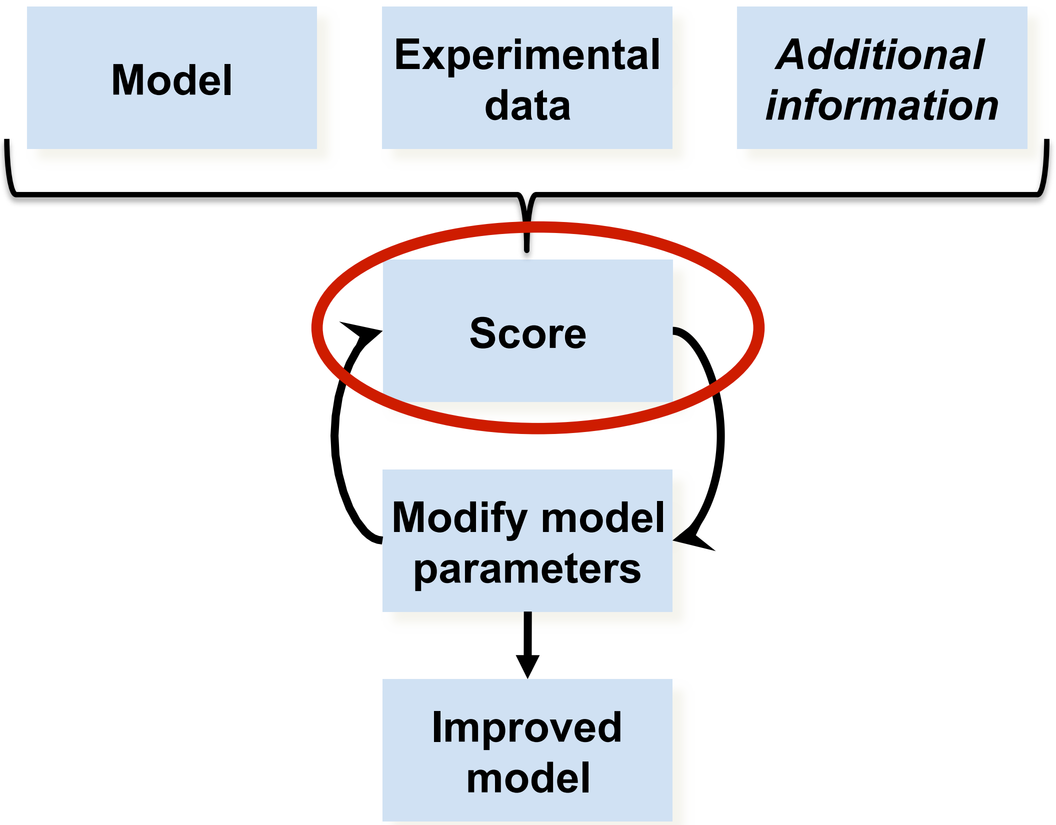


Source: Internet

- **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

Refinement target function (score)



Model

Experimental data

Additional information

Score

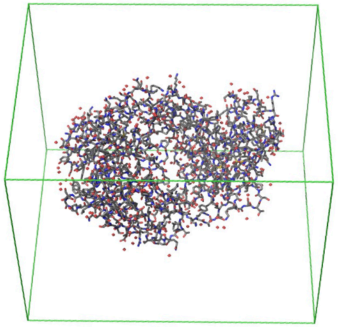
Modify model parameters

Improved model

Refinement target function (score)

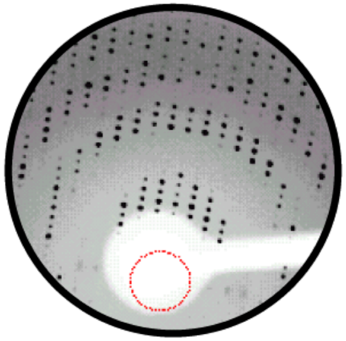
$$T = T_{\text{DATA}}(F_{\text{OBS}}, F_{\text{MODEL}}) + wT_{\text{RESTRAINTS}}$$

Model



$\mathbf{F}_{\text{MODEL}}$

Data



$\mathbf{F}_{\text{OBS}}$

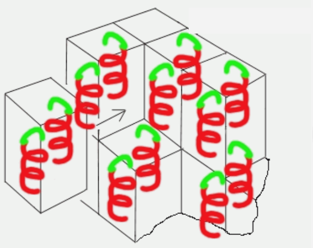
$$\boldsymbol{T_{DATA}} = \sum_{hkl}(F_{obs} - F_{model})^2$$

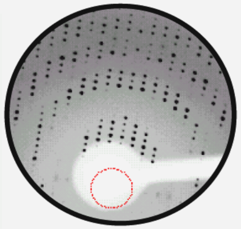
$$\boldsymbol{T_{DATA}} = \sum_{hkl}\frac{||F_{obs}| - |F_{model}||}{|F_{obs}|}$$

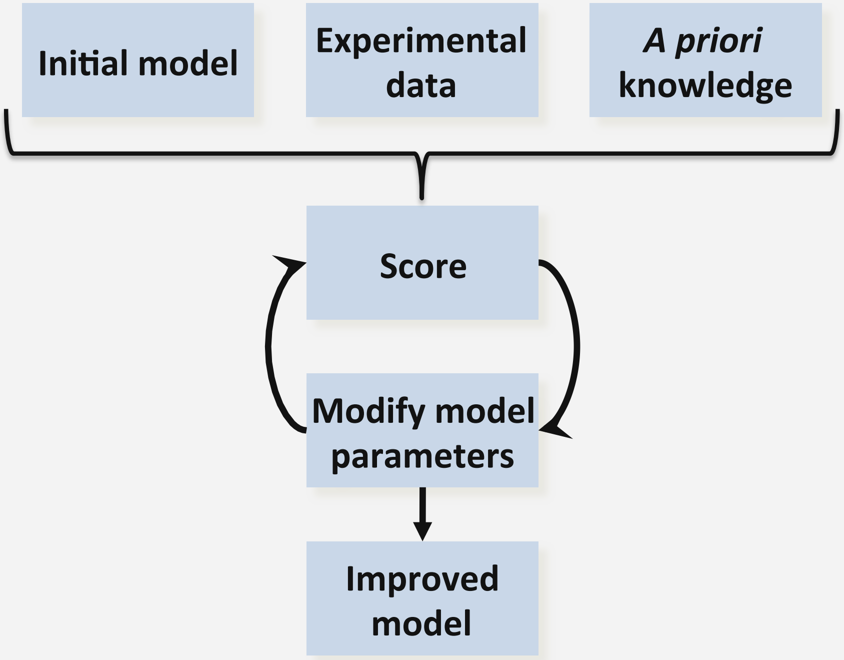
$\boldsymbol{T_{DATA}}$ =
Maximum-Likelihood score

Refinement

Crystallography







Initial model

Experimental data

A priori knowledge

Score

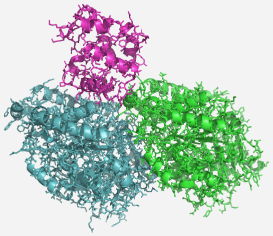
Modify model parameters

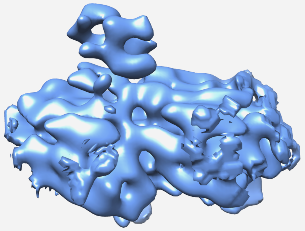
Improved model

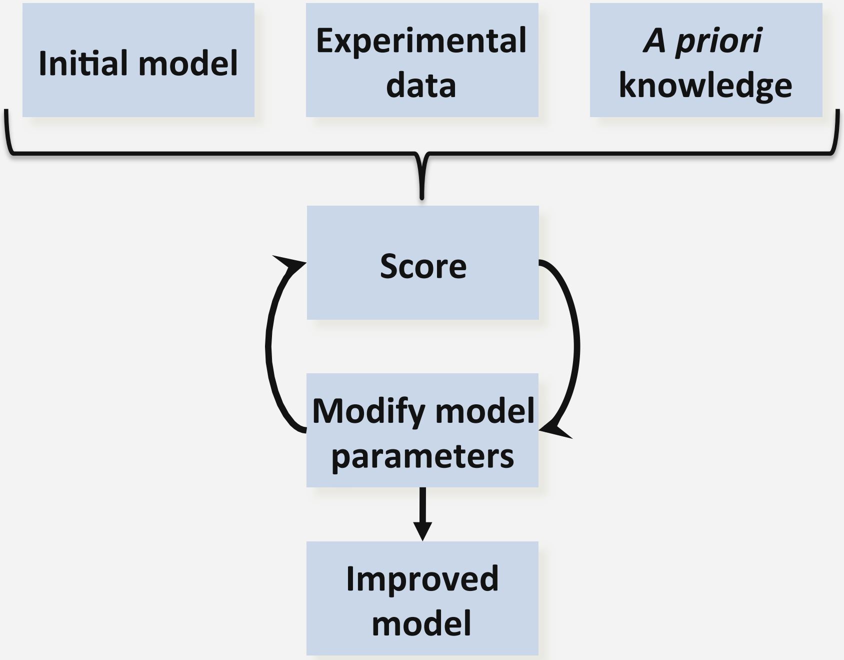
phenix.refine

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

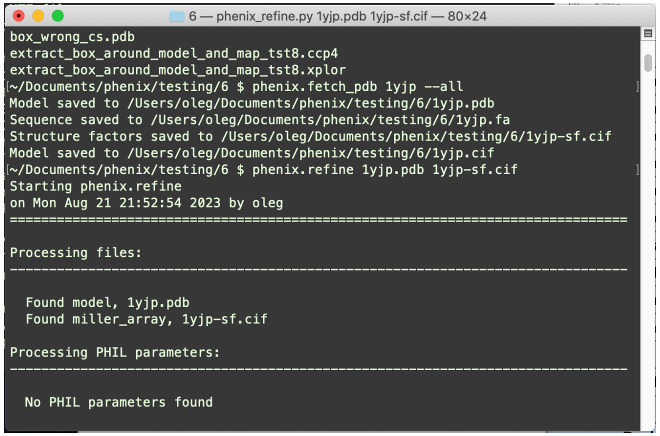
Refinement: command line

- Real-space (cryo-EM)

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

phenix.real_space_refine model.pdb map_coeffs.mtz

phenix.real_space_refine model.pdb map_coeffs.mtz ligans.cif

phenix.real_space_refine parameters.eff
```

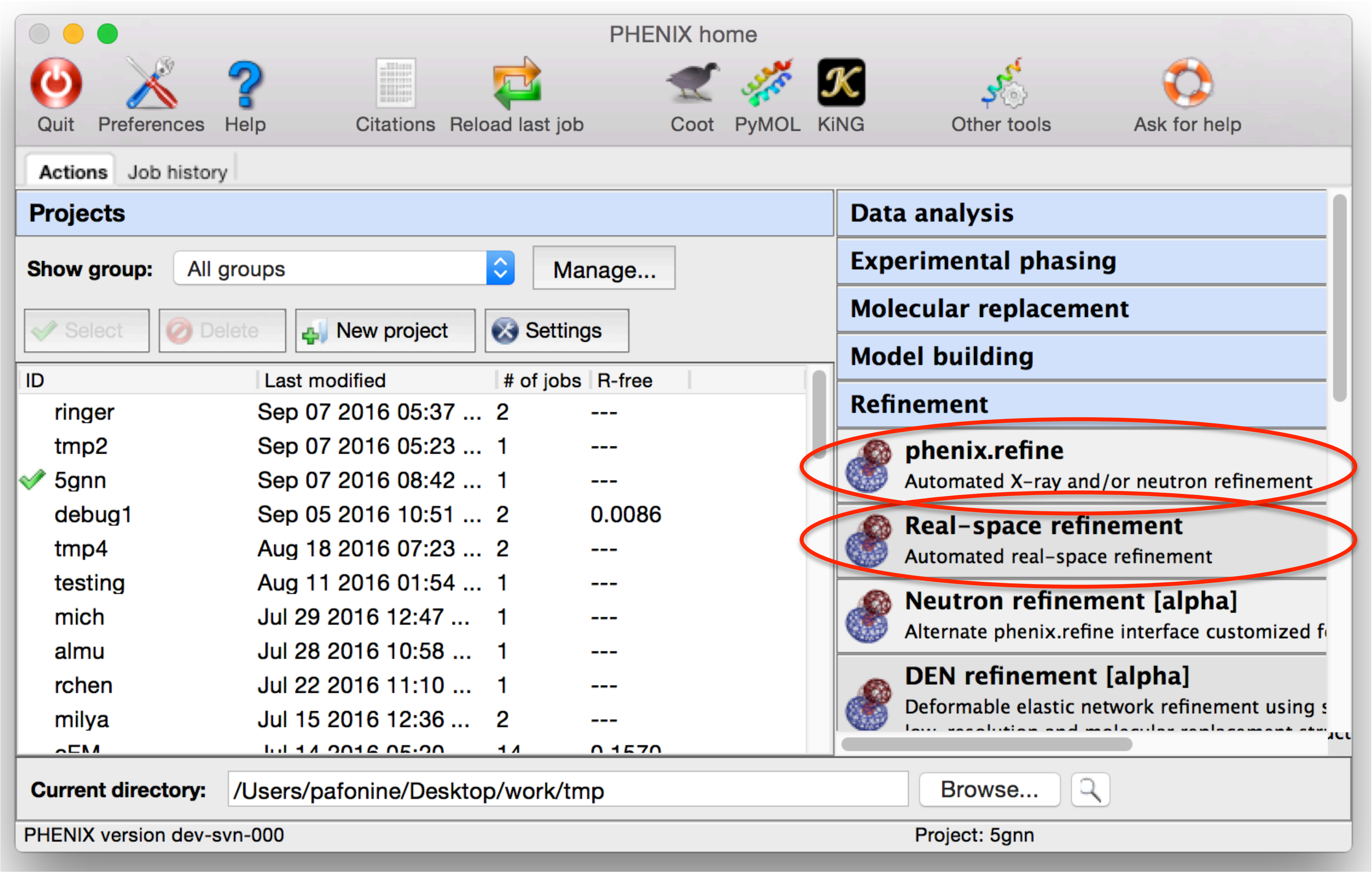
- Reciprocal-space (crystallography)

```
phenix.refine model.pdb data.mtz

phenix.refine model.pdb data.mtz ligans.cif

phenix.refine parameters.eff
```



```
box_wrong_cs.pdb
extract_box_around_model_and_map_tst8.ccp4
extract_box_around_model_and_map_tst8.xplor
~/Documents/phenix/testing/6 $ phenix.fetch_pdb 1yjp --all
Model saved to /Users/oleg/Documents/phenix/testing/6/1yjp.pdb
Sequence saved to /Users/oleg/Documents/phenix/testing/6/1yjp.fa
Structure factors saved to /Users/oleg/Documents/phenix/testing/6/1yjp-sf.cif
Model saved to /Users/oleg/Documents/phenix/testing/6/1yjp.cif
~/Documents/phenix/testing/6 $ phenix.refine 1yjp.pdb 1yjp-sf.cif
Starting phenix.refine
on Mon Aug 21 21:52:54 2023 by oleg
===============================================================================

Processing files:
-------------------------------------------------------------------------------


  Found model, 1yjp.pdb
  Found miller_array, 1yjp-sf.cif

Processing PHIL parameters:
-------------------------------------------------------------------------------


  No PHIL parameters found
```

Refinement tools in *Phenix*



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	---
cEM	Jul 14 2016 05:20	14	0.1570

phenix.refine
Automated X-ray and/or neutron refinement

Real-space refinement
Automated real-space refinement

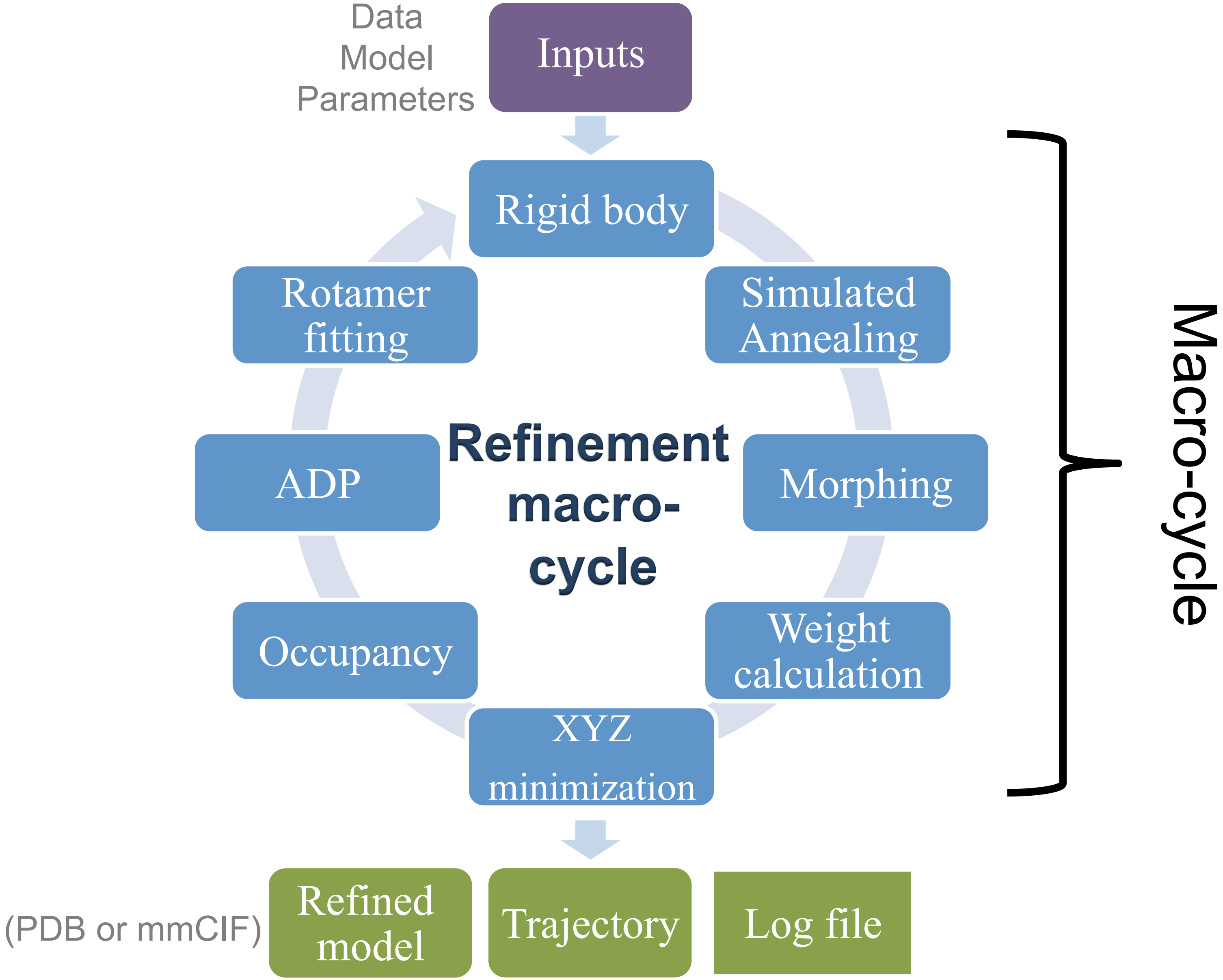
Neutron refinement [alpha]
Alternate phenix.refine interface customized f

DEN refinement [alpha]
Deformable elastic network refinement using s
low-resolution and molecular replacement struct

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

PHENIX version dev-svn-000 Project: 5gnn

Refinement protocol



Understanding inputs and outputs

- Real-space inputs
  - Atomic model (PDB, mmCIF)
  - Map (real map: MRC or Fourier map: MTZ)
  - Ligand restraints (”ligand CIF”)
  - Parameter files (as command line arguments or a file)

- Reciprocal-space inputs
  - Atomic model (PDB, mmCIF)
  - Reflection data (typically MTZ but most other formats are OK)
  - Ligand restraints (”ligand CIF”)
  - Parameter files (as command line arguments or a file)

Understanding inputs and outputs

- Real-space outputs
  - Atomic model (PDB, mmCIF)
  - .log file
  - .eff file – summary of all input parameters
  - .geo file (optionally)

- Reciprocal-space outputs
  - Atomic model (PDB, mmCIF)
  - .log file
  - .eff file – summary of all input parameters
  - MTZ file with copy of input data and 2Fo-Fc and Fo-Fc maps
  - .geo file (optionally)

.geo file contains description of all the geometry restraints used in refinement

Understanding inputs and outputs

- MTZ outputted by `phenix.refine` contains

1. Verbatim copy of input data considered for use

2. Data that was actually used in refinement

3. Total model structure factors $\mathbf{F}_{model}$

4. Fourier maps
   - $\texttt{2mF}_{\texttt{obs}}\texttt{-DF}_{\texttt{model}}$ ‘filled’
   - $\texttt{2mF}_{\texttt{obs}}\texttt{-DF}_{\texttt{model}}$
   - $\texttt{mF}_{\texttt{obs}}\texttt{-DF}_{\texttt{model}}$
   - Anomalous difference map (if anomalous data)

Refinement: practical considerations

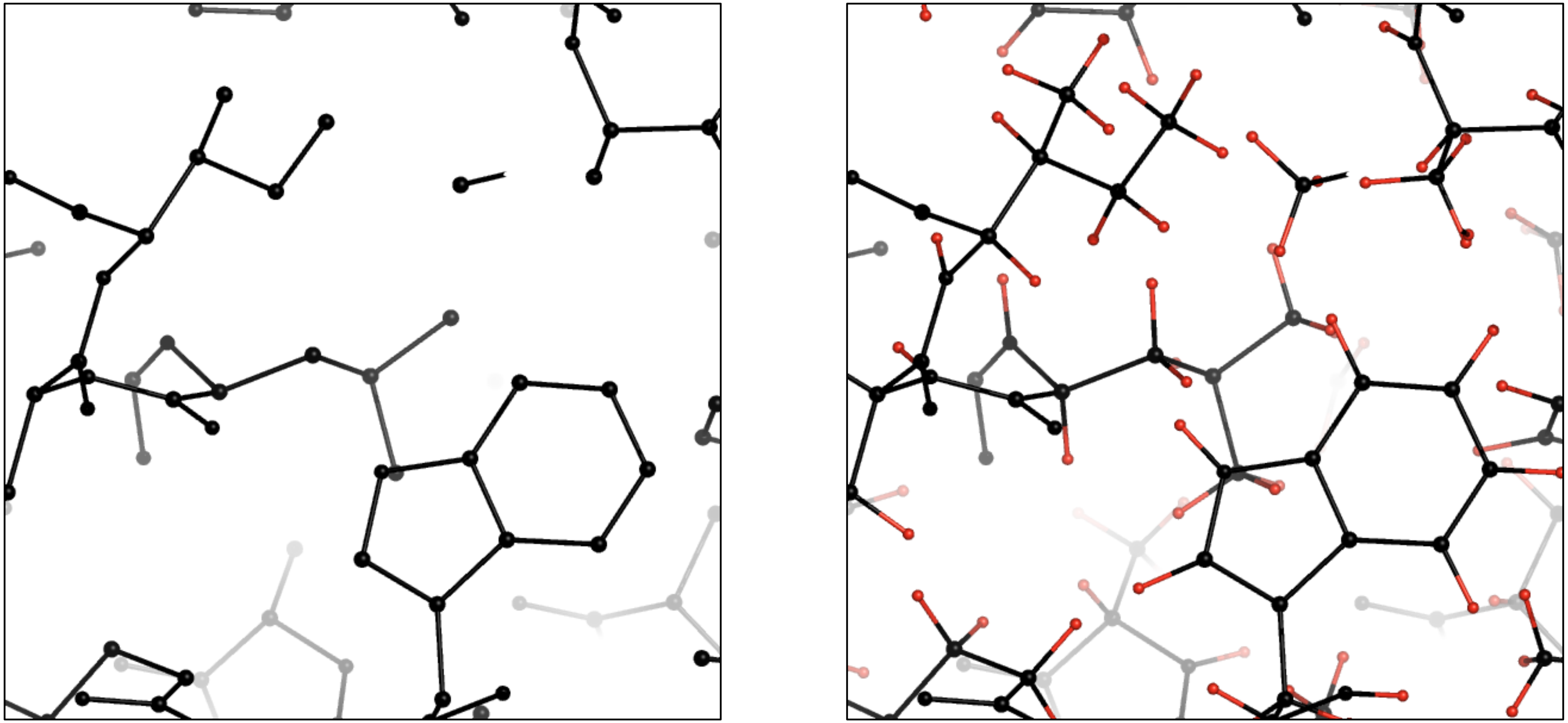
Aggressive optimization methods

- Simulated annealing (SA)
- Model morphing

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

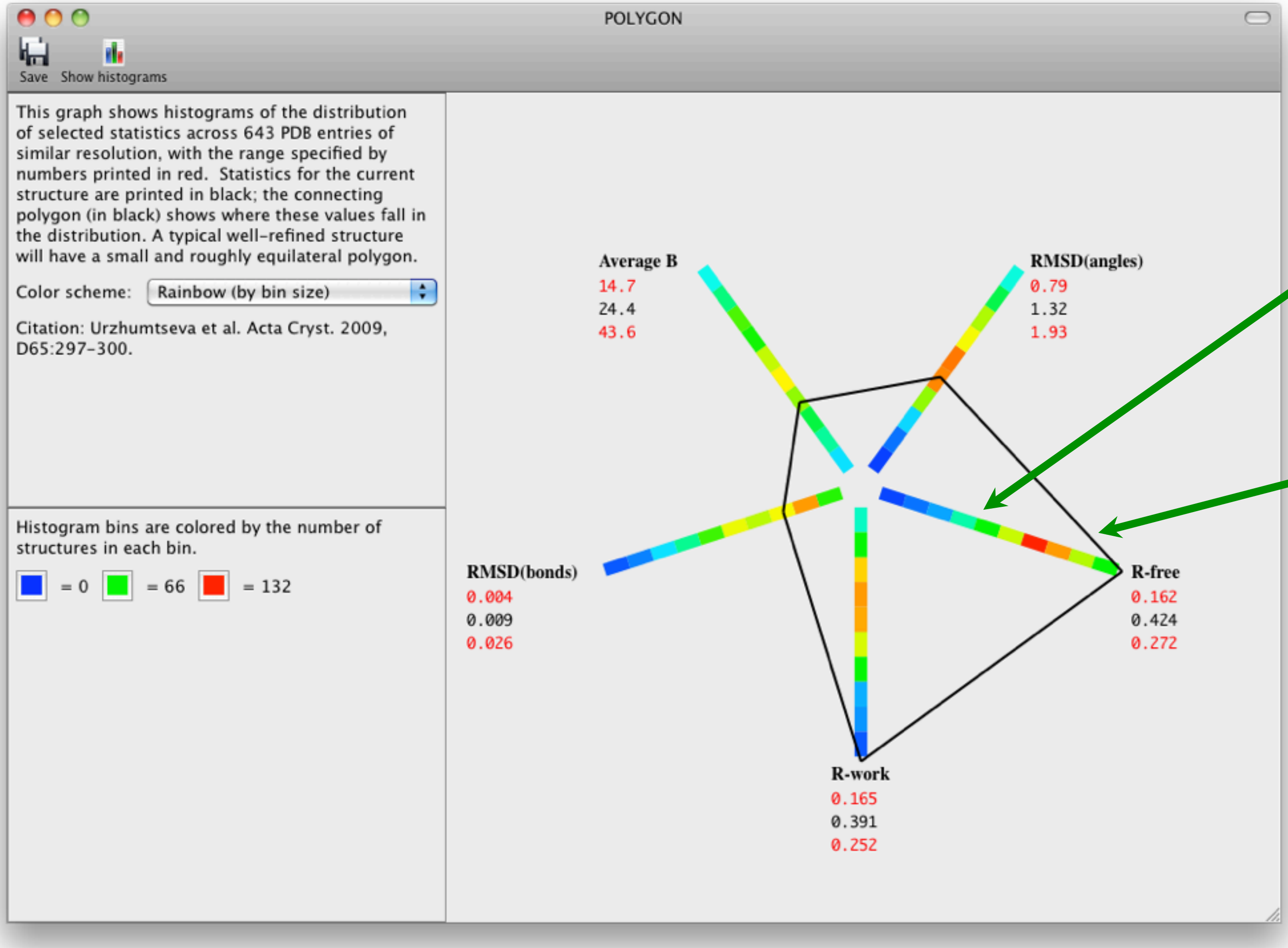
Use Hydrogen atoms

- 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

Know when to stop refinement



This graph shows histograms of the distribution of selected statistics across 643 PDB entries of similar resolution, with the range specified by numbers printed in red. Statistics for the current structure are printed in black; the connecting polygon (in black) shows where these values fall in the distribution. A typical well-refined structure will have a small and roughly equilateral polygon.

Color scheme: Rainbow (by bin size)

Citation: Urzhumtseva et al. Acta Cryst. 2009, D65:297-300.

Histogram bins are colored by the number of structures in each bin.

🟦 = 0 🟩 = 66 🟥 = 132

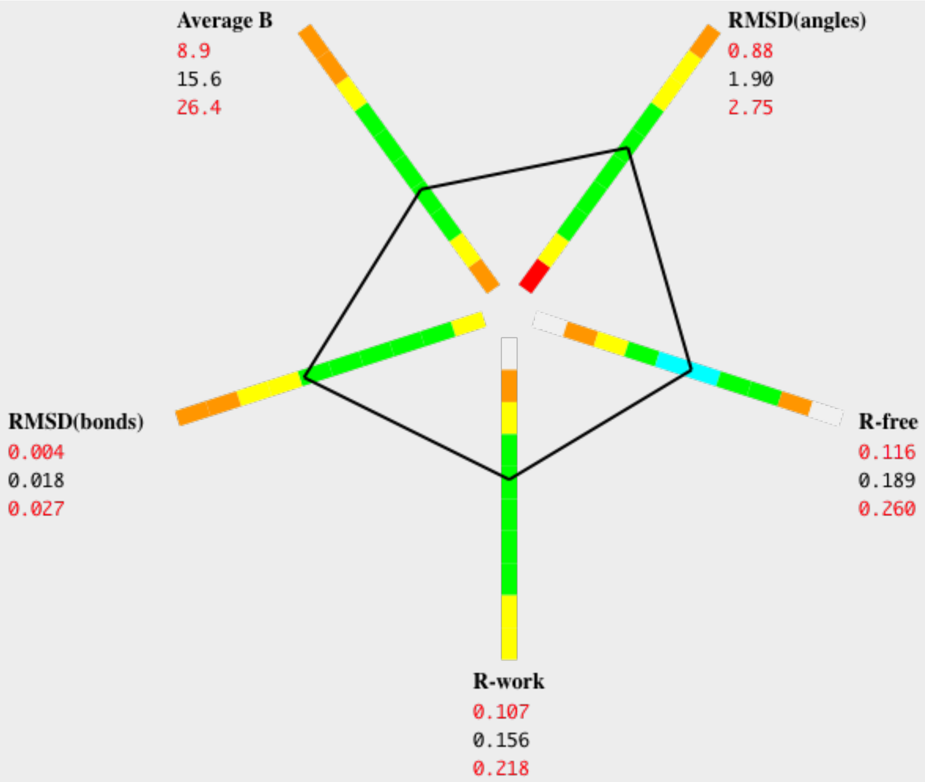
Colored bars are histograms showing distribution of values for structures at similar resolution

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

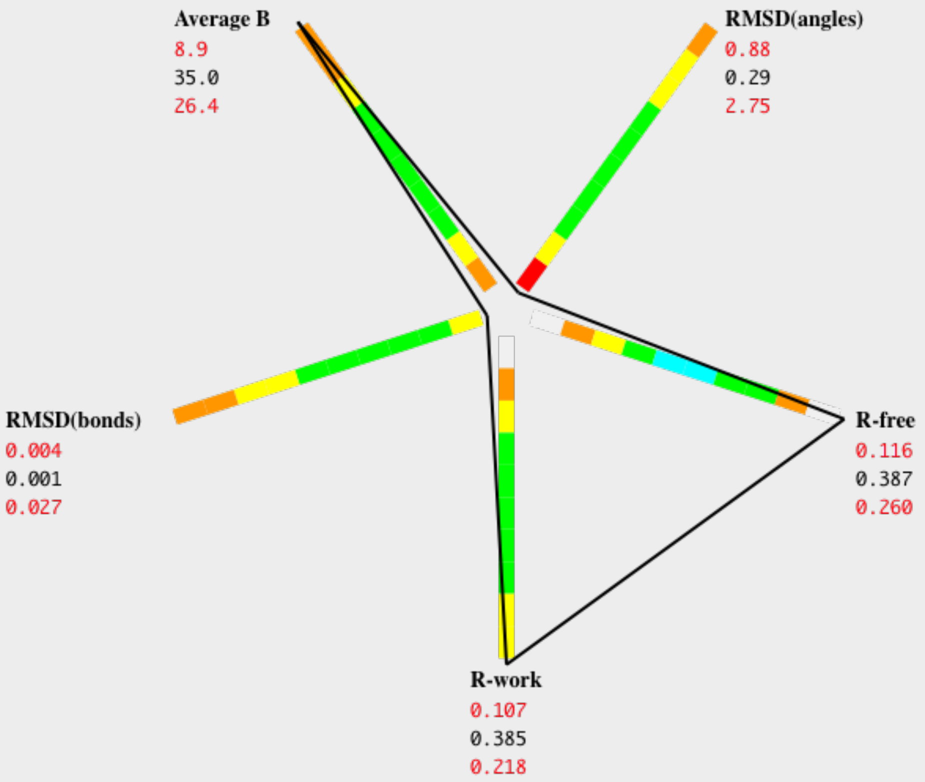
Crystallographic model quality at a glance.
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst*. **D**65, 297-300 (2009)

Know when to stop refinement

Likely overall good model



Clearly there are problems



Low resolution (3Å or worse)

- Use:
  - Ramachandran plot restraints
  - Secondary structure restraints
  - Reference model restraints (if quality homology model is available)
  - NCS (restraints or constraints)

NCS (Non-crystallographic symmetry)

- Constraints vs restraints
  - Constraints:
    - 4-5 Å or worse
    - Highly symmetric molecules
  - Restraints:
    - 2-4 Å

- Torsion vs Cartesian NCS
  - Torsion is preferable in most cases

- Symmetry related copies:
  - Can be found automatically as part of refinement
  - Can be specified manually
  - Automatic determination relies on model quality
    - Always check automatically detected NCS copies

Secondary structure (SS) restraints

- Always use at 3Å and worse
- Better than 3Å: use if needed
- Require SS annotation
- SS annotation must be accurate
  - Errors in SS annotation may propagate into refined model

- Secondary structure (SS) annotation
  - SS information
    - HELIX/SHEET records in PDB file or equivalent in mmCIF
    - *Phenix* generated parameter files
  - Tools to create SS annotation
    - Command line (*phenix.secondary_structure_restraints*)
    - *Phenix* GUI
  - Quality of SS annotation:
    - Depends on quality of input model (GIGO)
    - No software can annotate SS fully reliably and correctly
    - Manual validation and editing almost always required

Ramachandran plot restraints

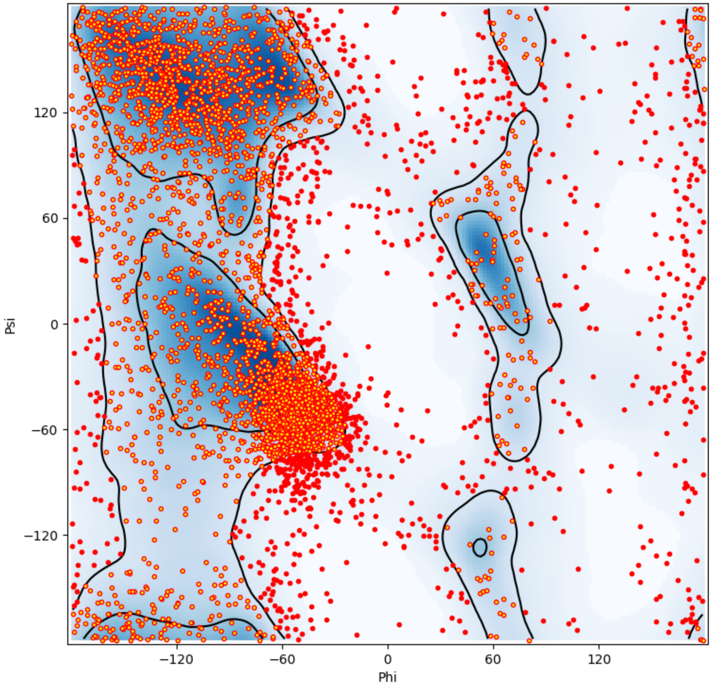
- Likely need at about 3Å and worse
- Better than 3Å: use if needed (preserve good initial model from deterioration)
- Check Ramachandran plot regularly
- Don’t use to fix outliers. Fix outliers first (manually), then use Ramachandran plot restraints to stop re-occurring outliers.

Ramachandran plot restraints

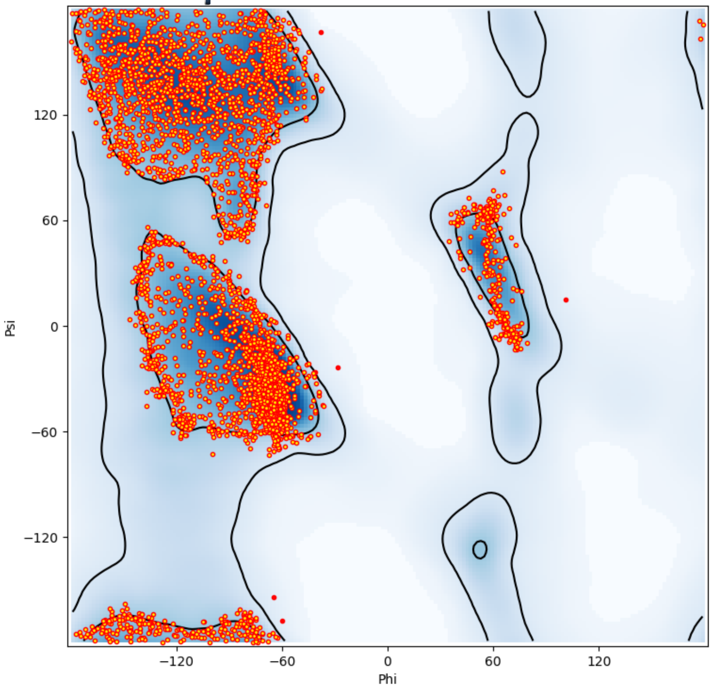
- Ramachandran plot restraints
  - Don’t use to fix outliers. Fix outliers first, then use Ramachandran plot restraints to prevent re-occurring outliers.

PDB code: 5a9z

Original

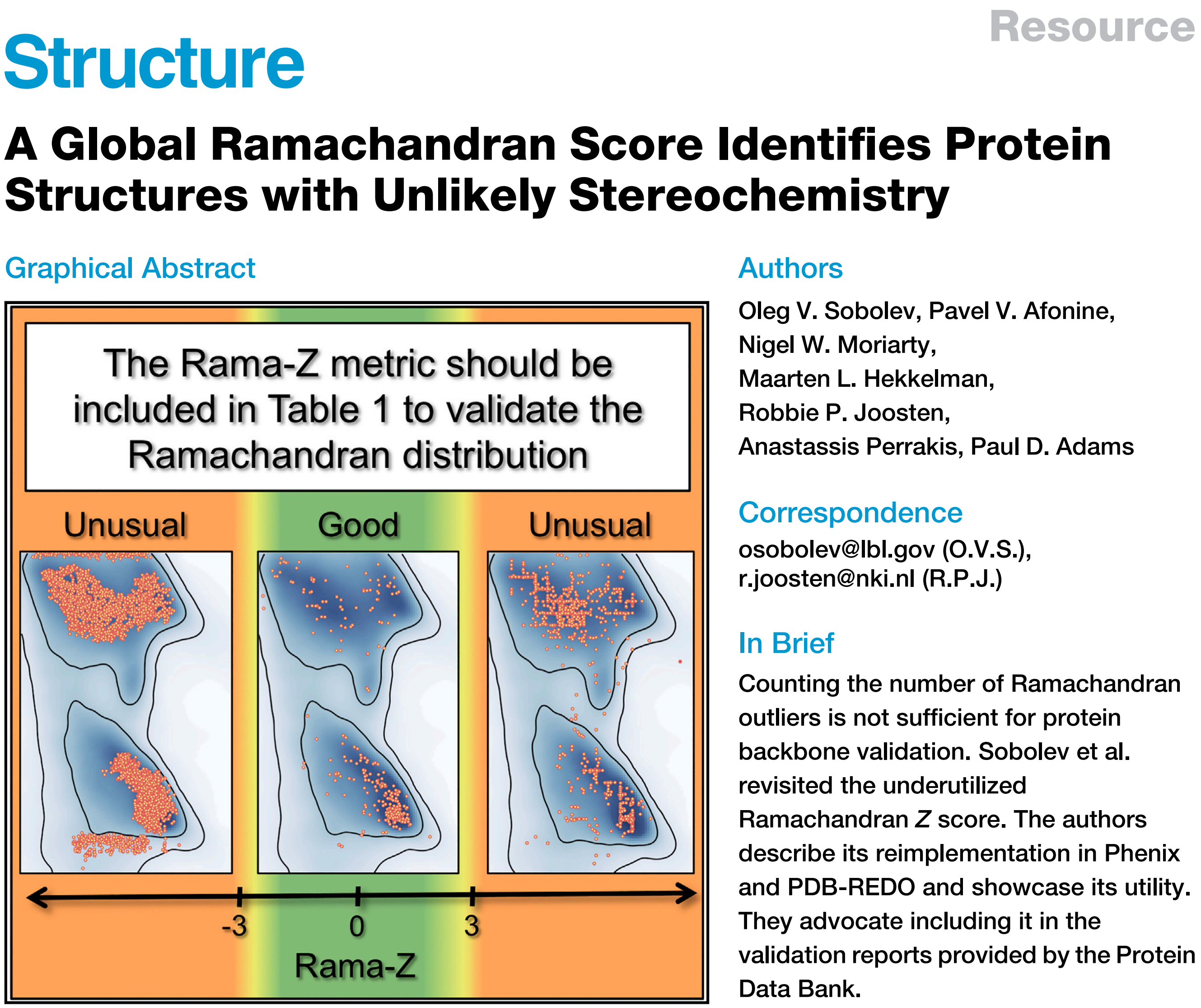


Refined with Ramachandran plot restraints



Bad idea to use Ramachandran plot restraints in this case. Fix outliers first!

Rama-Z score



Resource

Structure

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

Graphical Abstract

The Rama-Z metric should be included in Table 1 to validate the Ramachandran distribution

Authors

Oleg V. Sobolev, Pavel V. Afonine, Nigel W. Moriarty, Maarten L. Hekkelman, Robbie P. Joosten, Anastassis Perrakis, Paul D. Adams

Correspondence

osobolev@lbl.gov (O.V.S.), r.joosten@nki.nl (R.P.J.)

In Brief

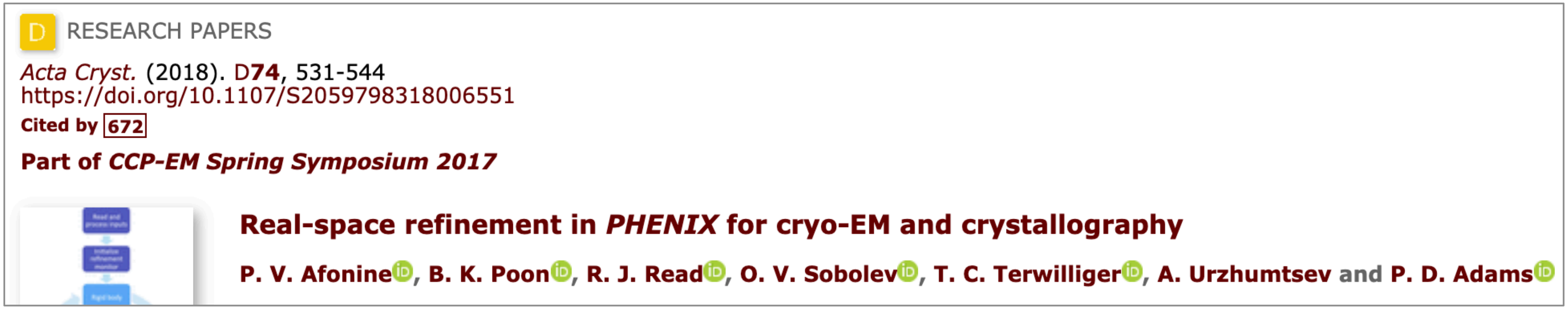
Counting the number of Ramachandran outliers is not sufficient for protein backbone validation. Sobolev et al. revisited the underutilized Ramachandran *Z* score. The authors describe its reimplementation in Phenix and PDB-REDO and showcase its utility. They advocate including it in the validation reports provided by the Protein Data Bank.

Talk tomorrow (Aug 23) 1:30 pm, room 209 Session A 20

Refinement: practical considerations

- Final stages
  - Make the model as complete as possible
  - Build alternative conformations
  - Use Hydrogen atoms (and keep them in the final model!)
  - Add ordered solvent components
- Remember: the better the model, the better the map
  - You may see and model your ligands better!

Reading



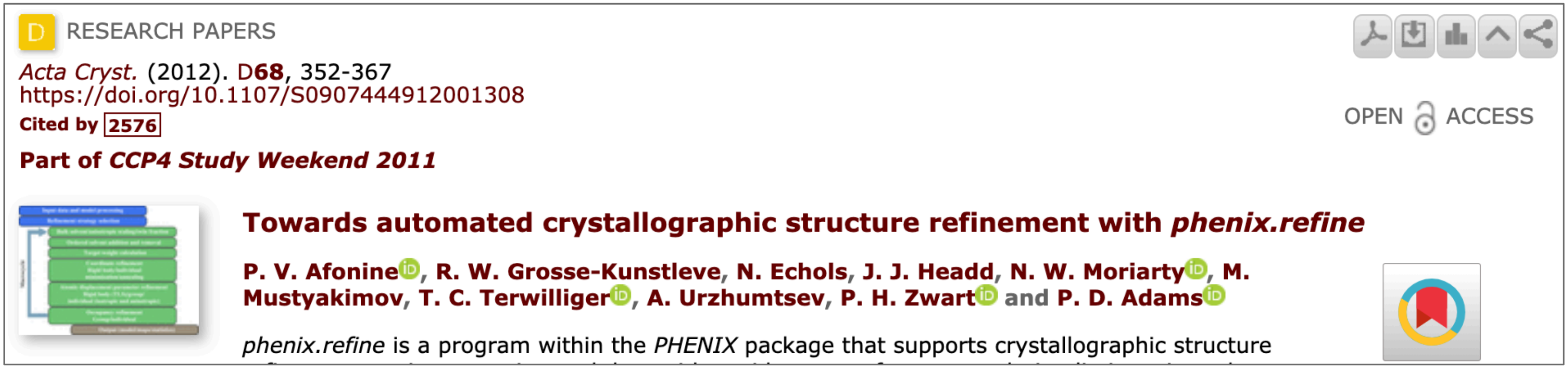
D RESEARCH PAPERS

Acta Cryst. (2018). D**74**, 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, B. K. Poon, R. J. Read, O. V. Sobolev, T. C. Terwilliger, A. Urzhumtsev and P. D. Adams



D RESEARCH PAPERS

Acta Cryst. (2012). D**68**, 352-367
https://doi.org/10.1107/S0907444912001308
Cited by 2576

OPEN ACCESS

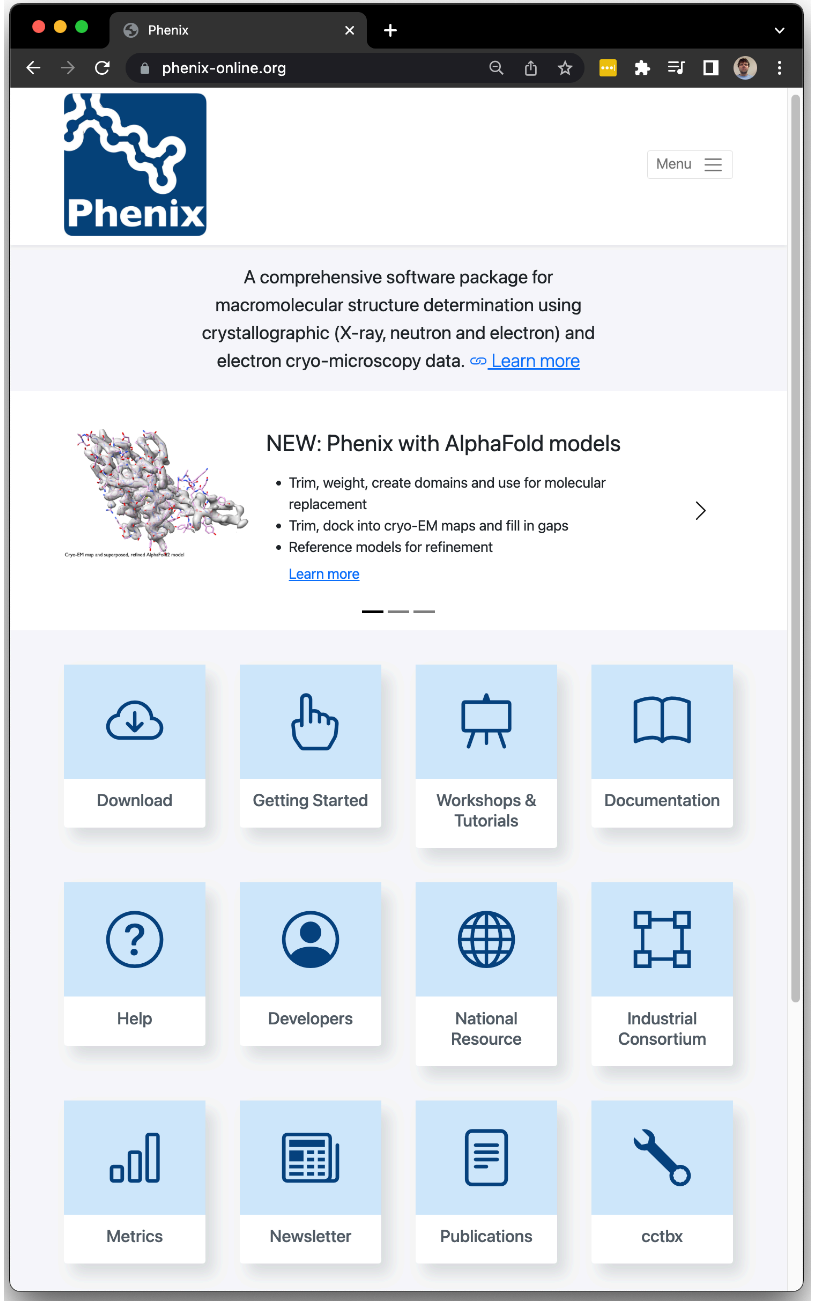
Part of *CCP4 Study Weekend 2011*

Towards automated crystallographic structure refinement with *phenix.refine*

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

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

Phenix resources



A comprehensive software package for macromolecular structure determination using crystallographic (X-ray, neutron and electron) and electron cryo-microscopy data. Learn more

NEW: Phenix with AlphaFold models

- Trim, weight, create domains and use for molecular replacement
- Trim, dock into cryo-EM maps and fill in gaps
- Reference models for refinement

Learn more

Phenix paper

Video tutorials

Documentation

Relevant papers

Bi-annual newsletters

Slides from workshops

User support

- **Feedback, questions, help**

Mailing list (anyone signed up):	phenixbb@phenix-online.org
Bug reports (developers only):	bugs@phenix-online.org
Ask for help (developers only):	help@phenix-online.org

- **Reporting a bug or asking for help:**
  - We can’t help you if you don’t help us to understand your problem
  - Make sure the problem still exist using the latest *Phenix* version
  - Send us all inputs (files, non-default parameters) and tell us steps that lead to the problem
  - All data sent to us is kept confidentially

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Dorothee Liebschner, Nigel Moriarty, Billy Poon, Christopher Schlicksup, Oleg Sobolev

BERKELEY LAB

Los Alamos National Laboratory
New Mexico Consortium

Tom Terwilliger, Li-Wei Hung

New Mexico CONSORTIUM

Los Alamos NATIONAL LABORATORY

UTHealth

Matt Baker, Corey Hyrc

UTHealth® The University of Texas Health Science Center at Houston

University of Cambridge

Randy Read, Airlie McCoy, Tristan Croll, Claudia Millán Nebot, Rob Oeffner

UNIVERSITY OF CAMBRIDGE

Duke University

Jane & David Richardson, Christopher Williams, Vincent Chen

Eruditio et Religio

An NIH/NIGMS funded Program Project

Liebschner D, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Cryst. 2019 **D75**:861–877