New tools for automated refinement and model completion

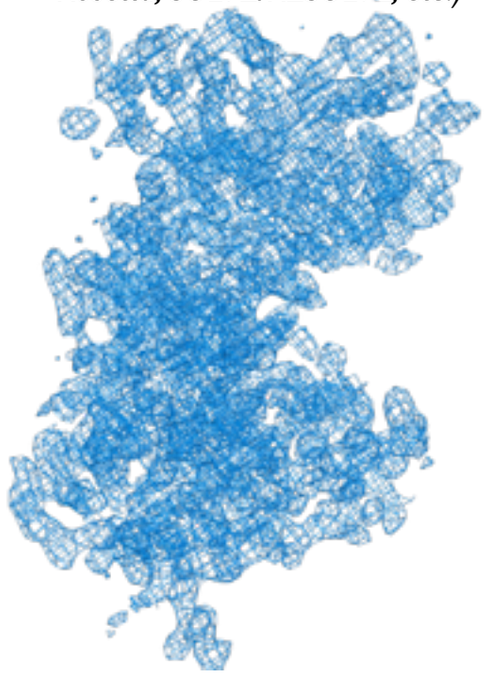
Nathaniel Echols

(Frank DiMaio, Tom Terwilliger, Nigel Moriarty, Pavel Afonine, Paul Adams, et al.)

The structure determination workflow

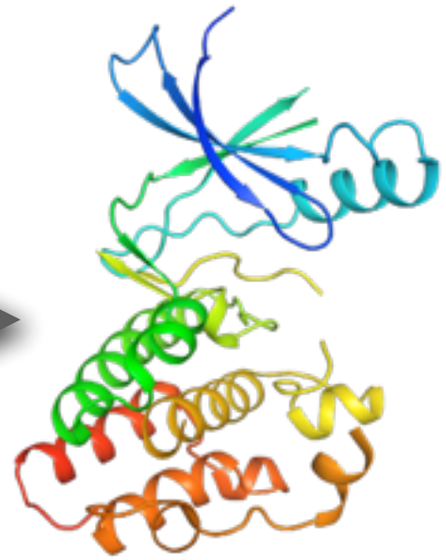
Phasing

(Phaser, AutoSol, MRage, AutoMR, MR-Rosetta, SOLVE/RESOLVE, etc.)



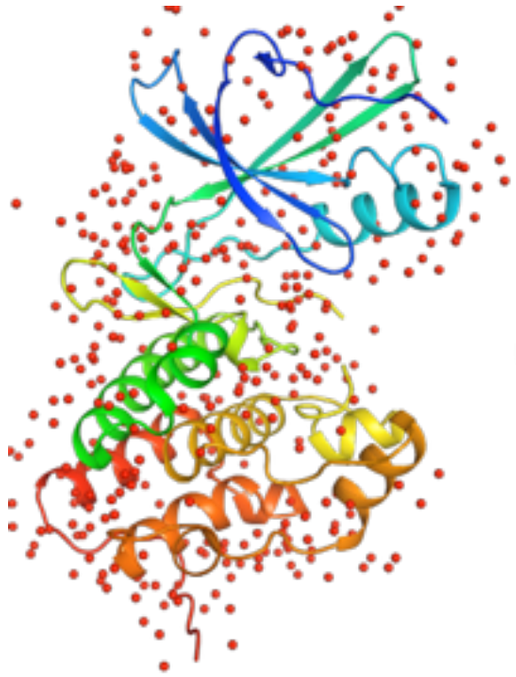
Initial model-building

(AutoBuild, AutoSol, phenix.find_helices_strands)



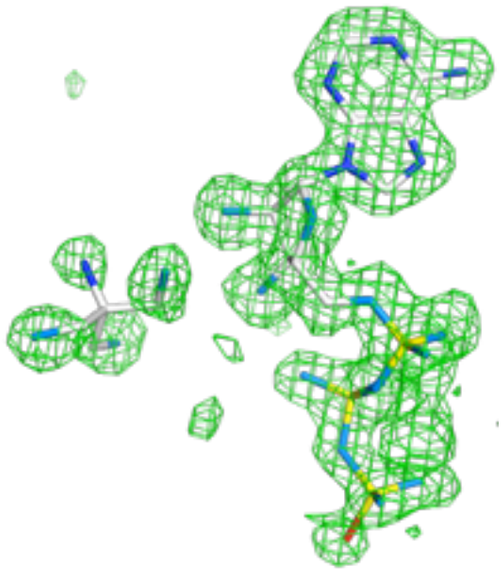
Loops and waters

(phenix.fit_loops, phenix.refine)



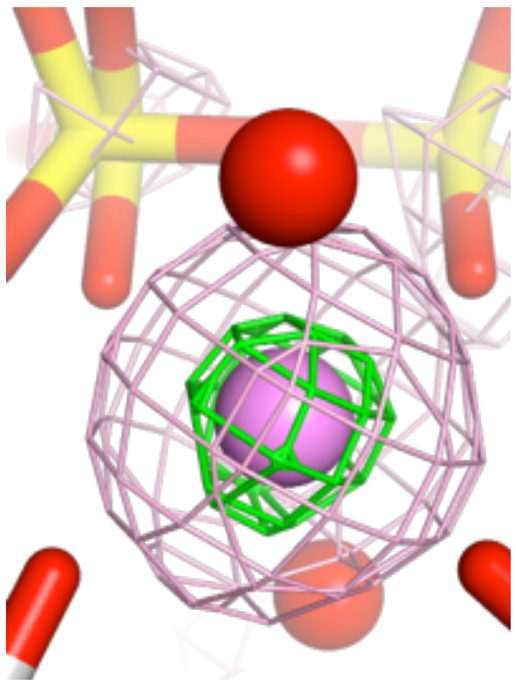
Ligands

(LigandFit, ligand identification, ligand pipeline, eLBOW)



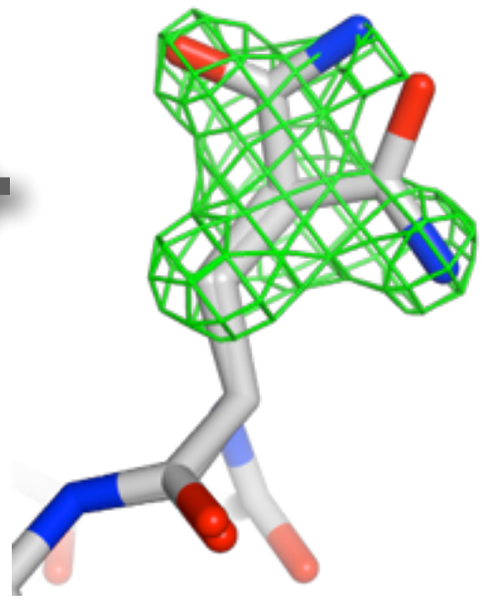
Ions

(phenix.refine, Phaser)



Alternate conformations

(in development)



PDB deposition*

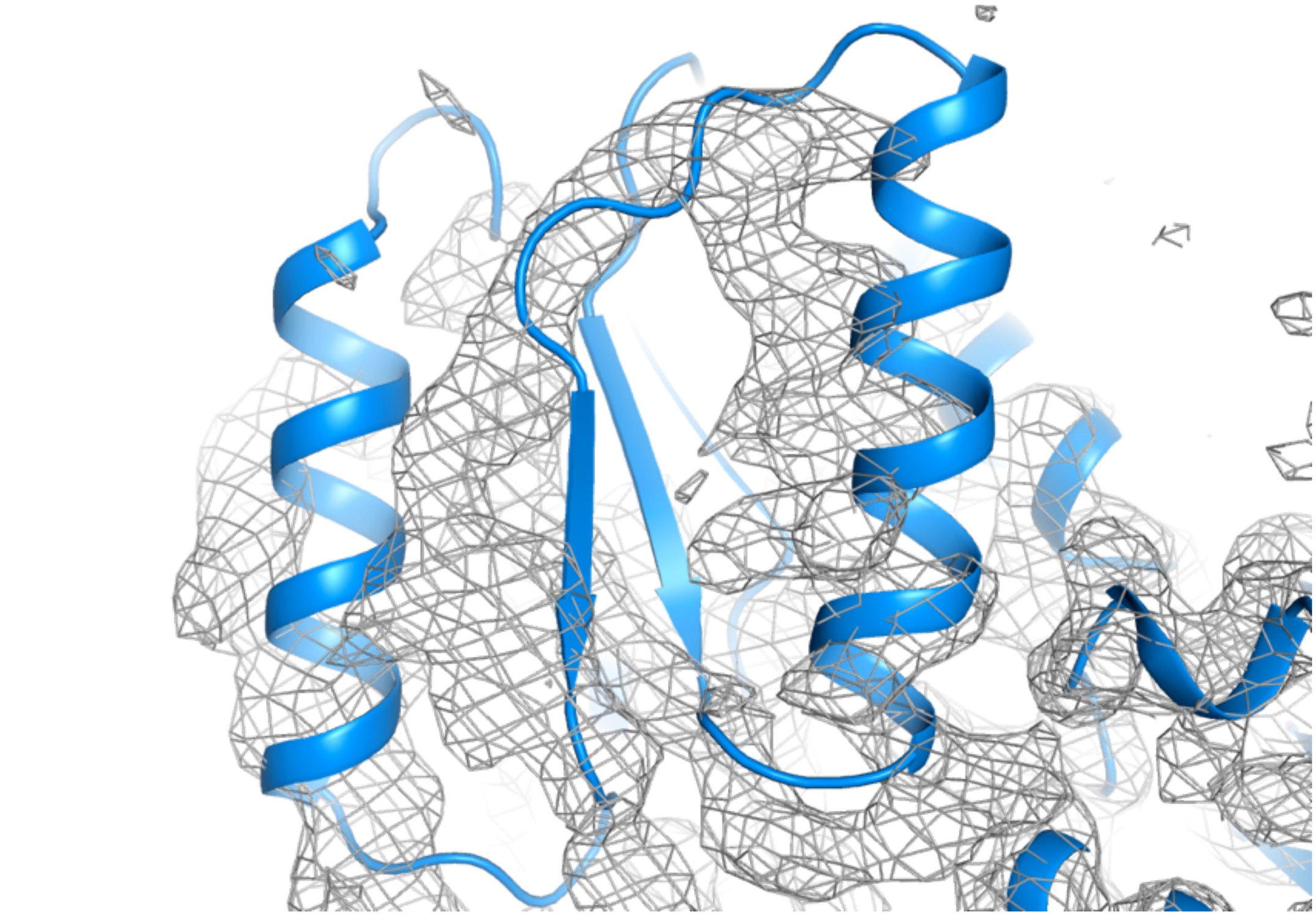
(in development)

Publication*

(phenix.table_one, TBA)

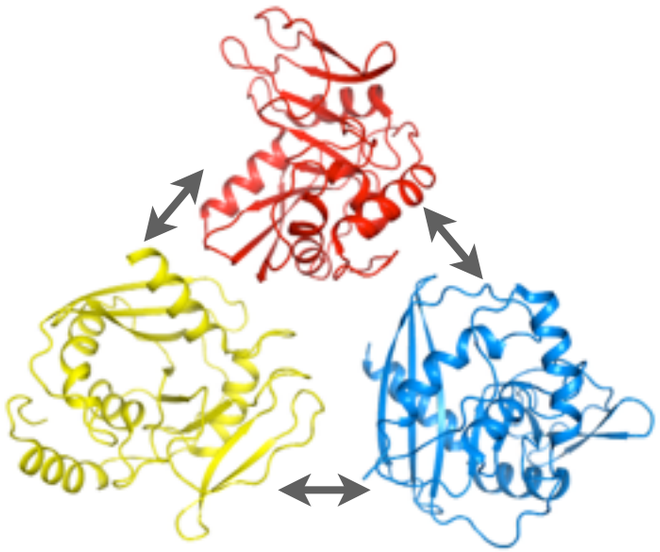
* or whatever industrial crystallographers do instead

The problem of low resolution refinement



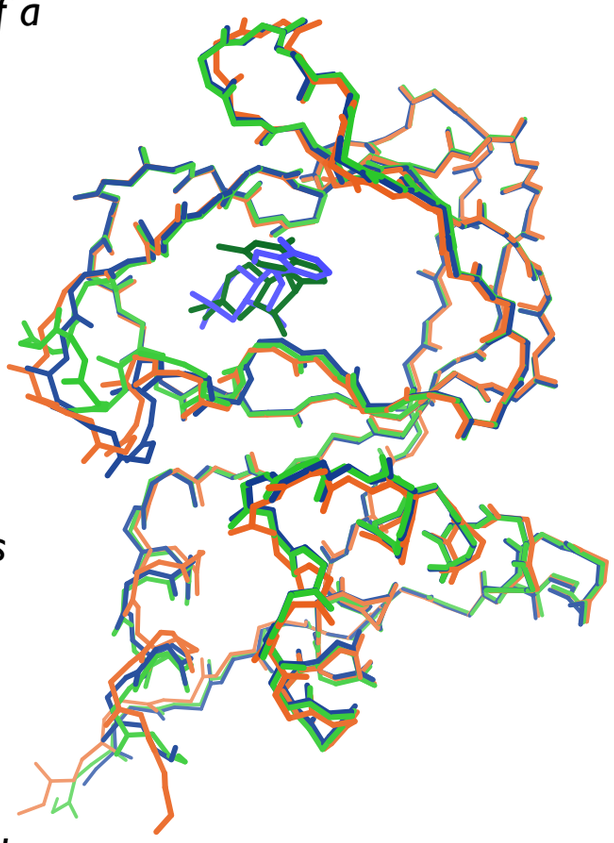
How can we optimize a model and maintain good geometry when it is far from correct?

Restraints for medium-to-low-resolution structures



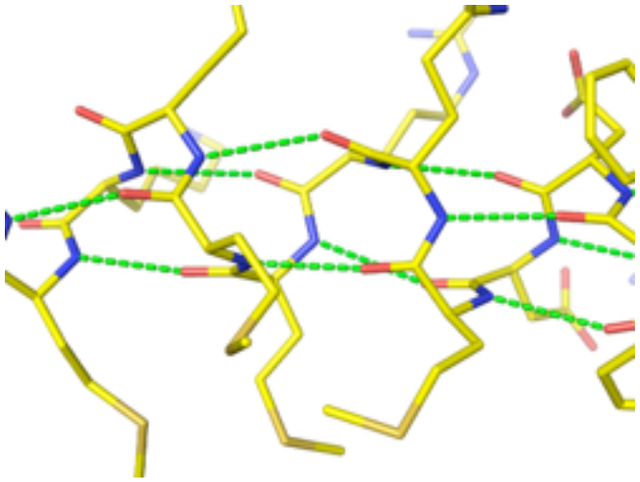
Non-crystallographic symmetry (NCS)

When the asymmetric unit (ASU) contains more than one copy of a molecule, the conformations of each chain can be restrained together. Can be parameterized either globally (keeping the structures rigid); or locally (torsion angles, in Phenix); local NCS restraints are preferable in most cases, since some deviation from ideal symmetry is common.



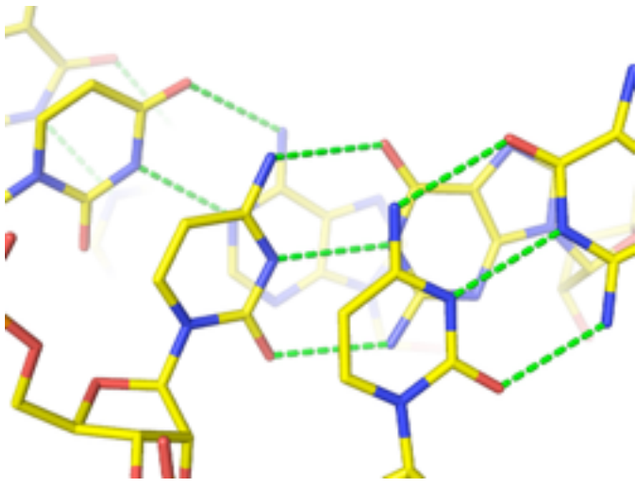
Reference model

If a high-quality, high-resolution structure (or homology model) of part or all of the model is available, the local conformation of the refined model can be restrained to this. Our implementation uses torsion angles, and shares many features with the torsion NCS restraints. Rotamer correction is also performed.



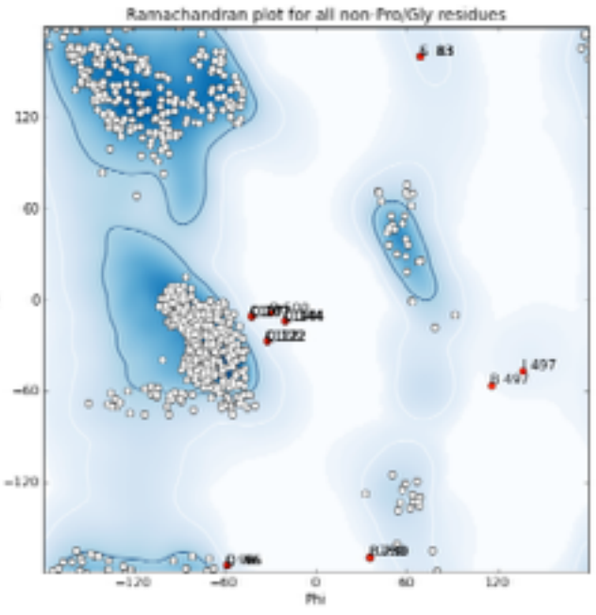
Secondary structure

Distances between hydrogen-bonding atoms in protein helices and sheets or nucleic acid base pairs can be restrained. This helps keep regular structure from unravelling during refinement, but the impact on R-factors is very small.



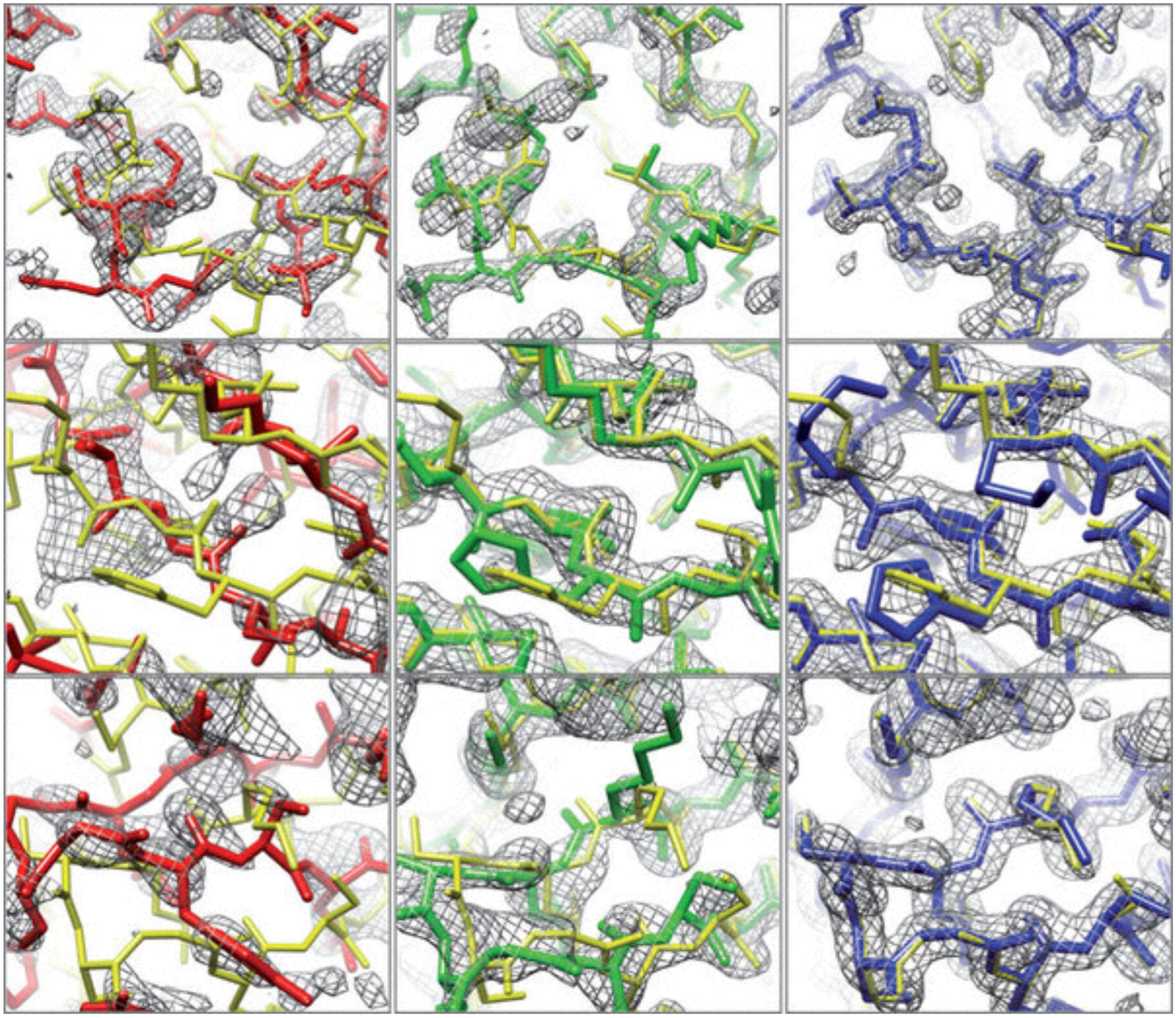
Ramachandran angles
(only under special circumstances)

For desperate cases at very low resolution, it may be necessary to restrain the protein backbone to stay within allowed regions of the Ramachandran plot. These restraints should be used only as a last resort, and not just to improve validation statistics.



Further reading: Headd et al. , Smart et al., Nicholls et al. in Acta Crystallgraphica D. April 2012 issue

MR-Rosetta: coarse refinement coupled with Phaser



DiMaio et al. (2011) Nature 473:540-3.

Combining crystallography and modeling

(with Frank DiMaio and David Baker, U. Washington)

Rosetta:

- Realistic energy function
- Repacking to remove steric clashes and build rotameric sidechains
- Torsion-angle minimization
- Real-space target
- Fragment-based rebuilding (optional, not used here)

phenix.refine[1]:

- X-ray target functions (ML, MLHL, LS-twin)
- Bulk solvent correction
- B-factor refinement (including TLS)
- Map calculation[2]
- Density modification (RESOLVE)

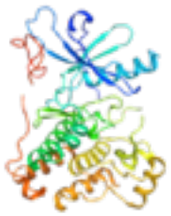
The Python and C++ architecture of these programs makes it (relatively) easy to share functionality

1. Afonine et al. (2012) *Acta Cryst.* **D68**:352-367.

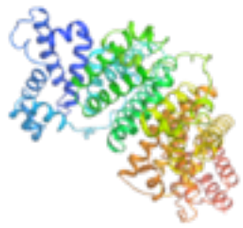
2. With R-free flagged reflections omitted

A realistic test set of poor MR solutions

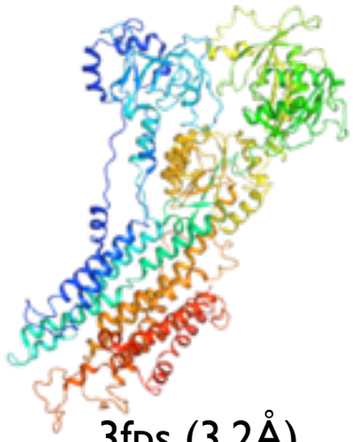
Starting RMSD to deposited structures: 1.5 - 6.0Å



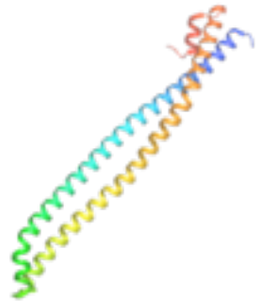
2j5f (3.0Å)



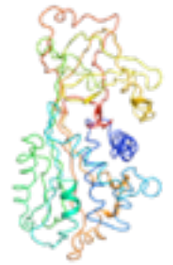
1bke (3.15Å)



3fps (3.2Å)

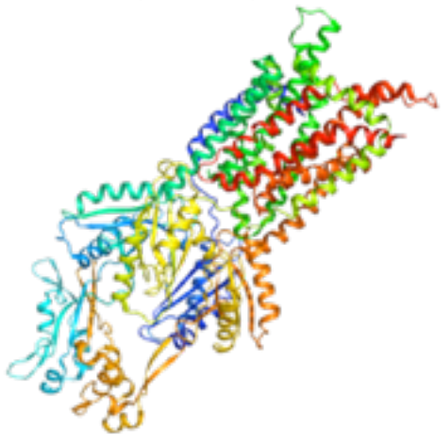


3mtt (3.3Å)

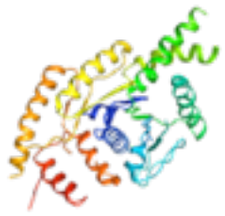


1kct (3.46Å)

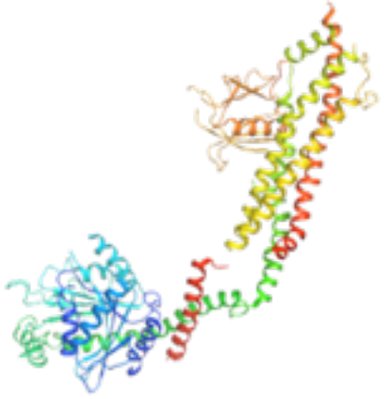
3pwy (3.5Å)



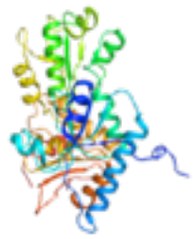
3k07 (3.52Å)



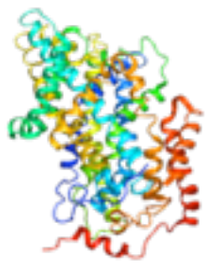
3idq (3.7Å)



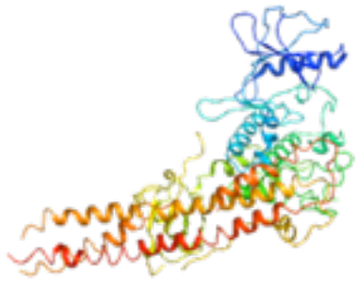
3snh (3.7Å)



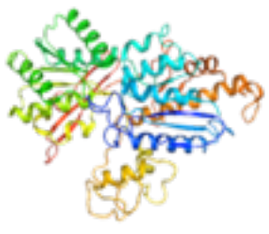
2vaf (3.8Å)



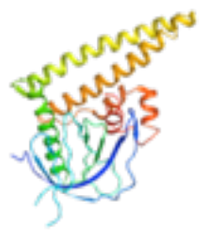
2x79 (3.8Å)



3rzf (4.0Å)



1isr (4.0Å)

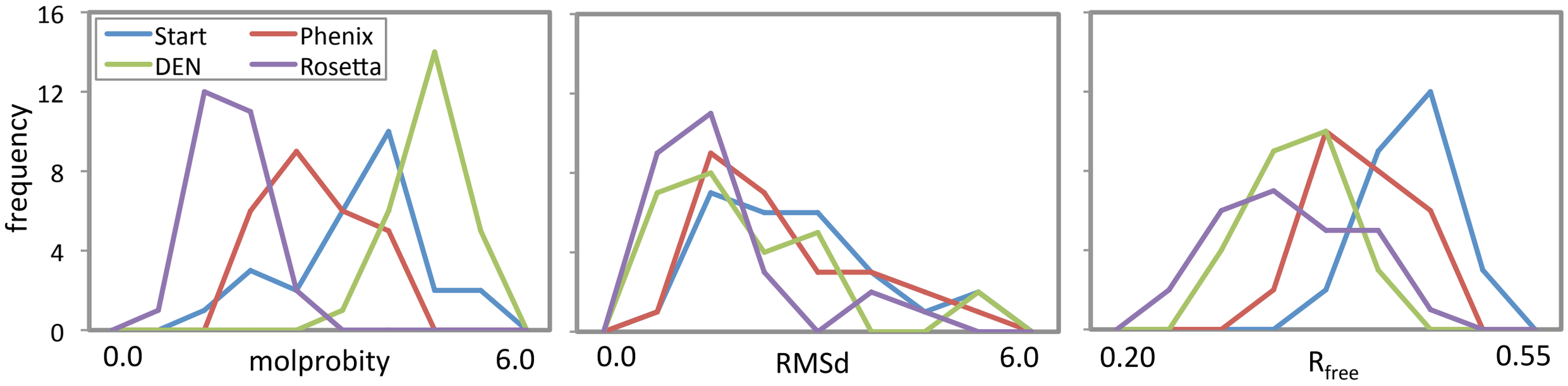


3a8n (4.5Å)

underlined structures are membrane proteins
italic structures used homologous proteins for MR

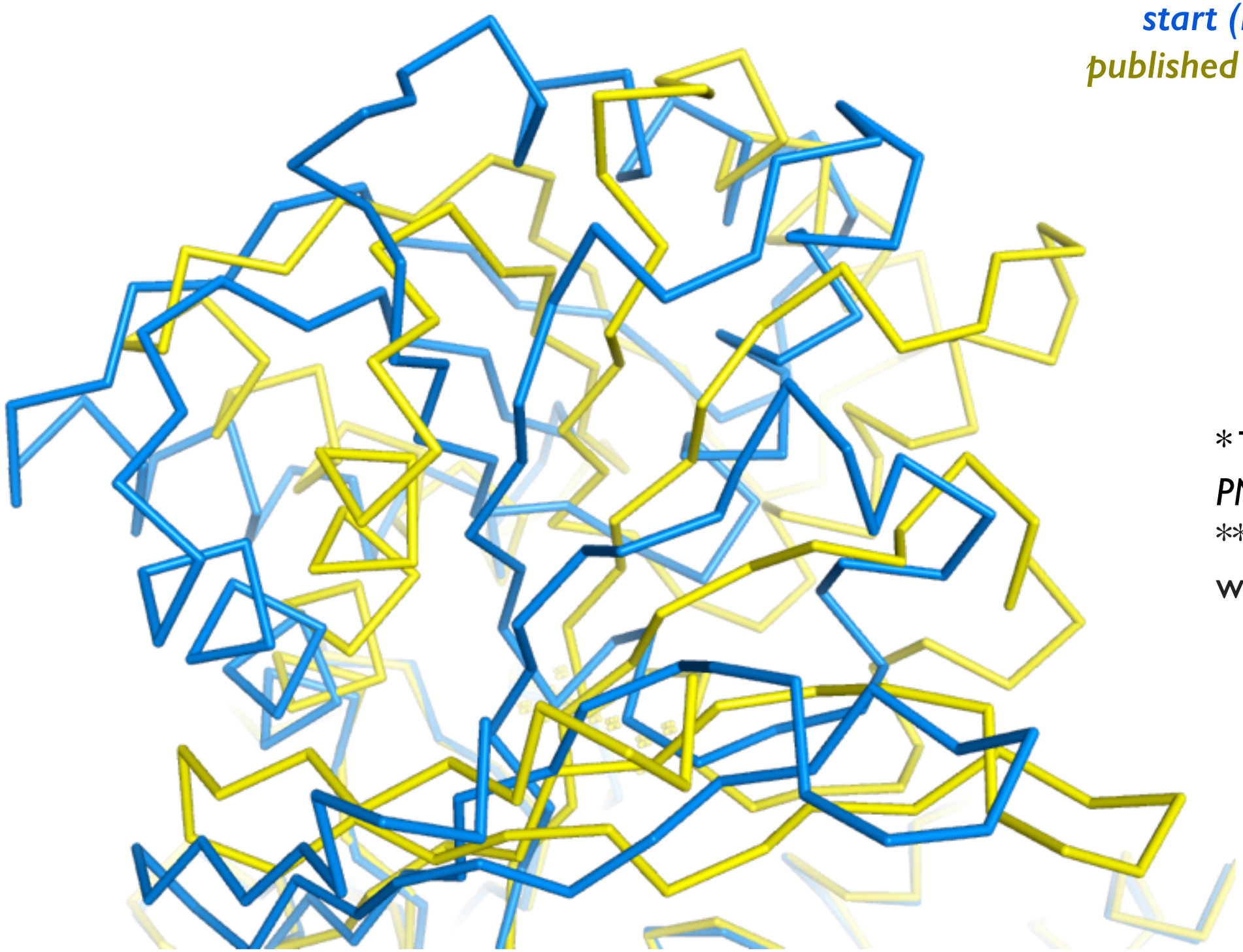
Rosetta+Phenix versus Phenix: overview

- Rosetta+Phenix always produces a higher-quality model
- Rosetta+Phenix usually refines to a better RMSD than our conservative phenix.refine strategy
- Comparable to CNS/DEN refinement but with better geometry, and performs better on a few particularly difficult cases
- Refmac "jelly body" refinement is comparable to CNS/DEN



A large conformational change (1isr*)

Starting model is
PDB ID 1ewt

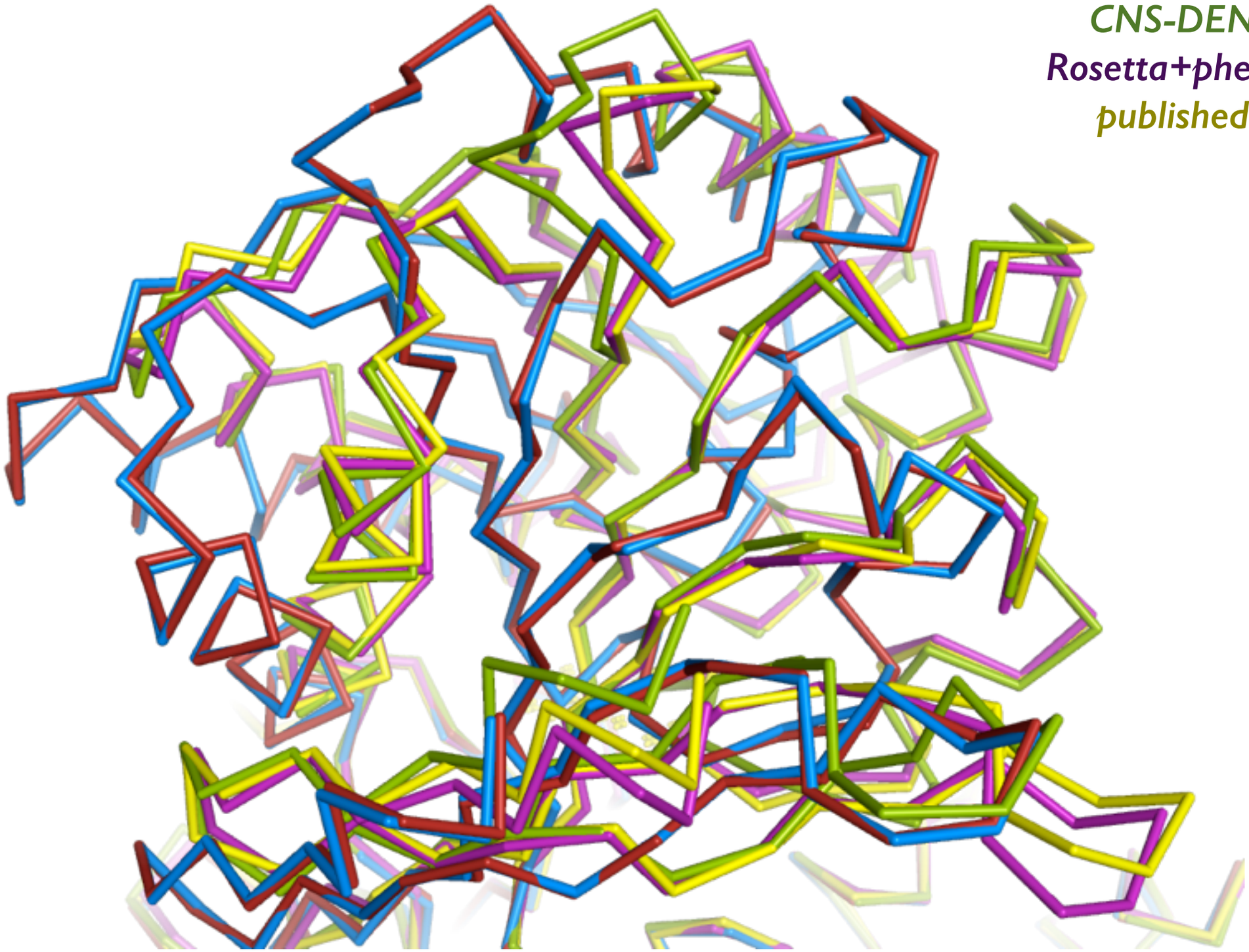


start (r_free=0.45)
*published (r_free=0.20**)*

* Tsuchiya et al. (2002)
PNAS **99**: 2660-2665
** after re-refinement with Phenix

A large conformational change (1isr)

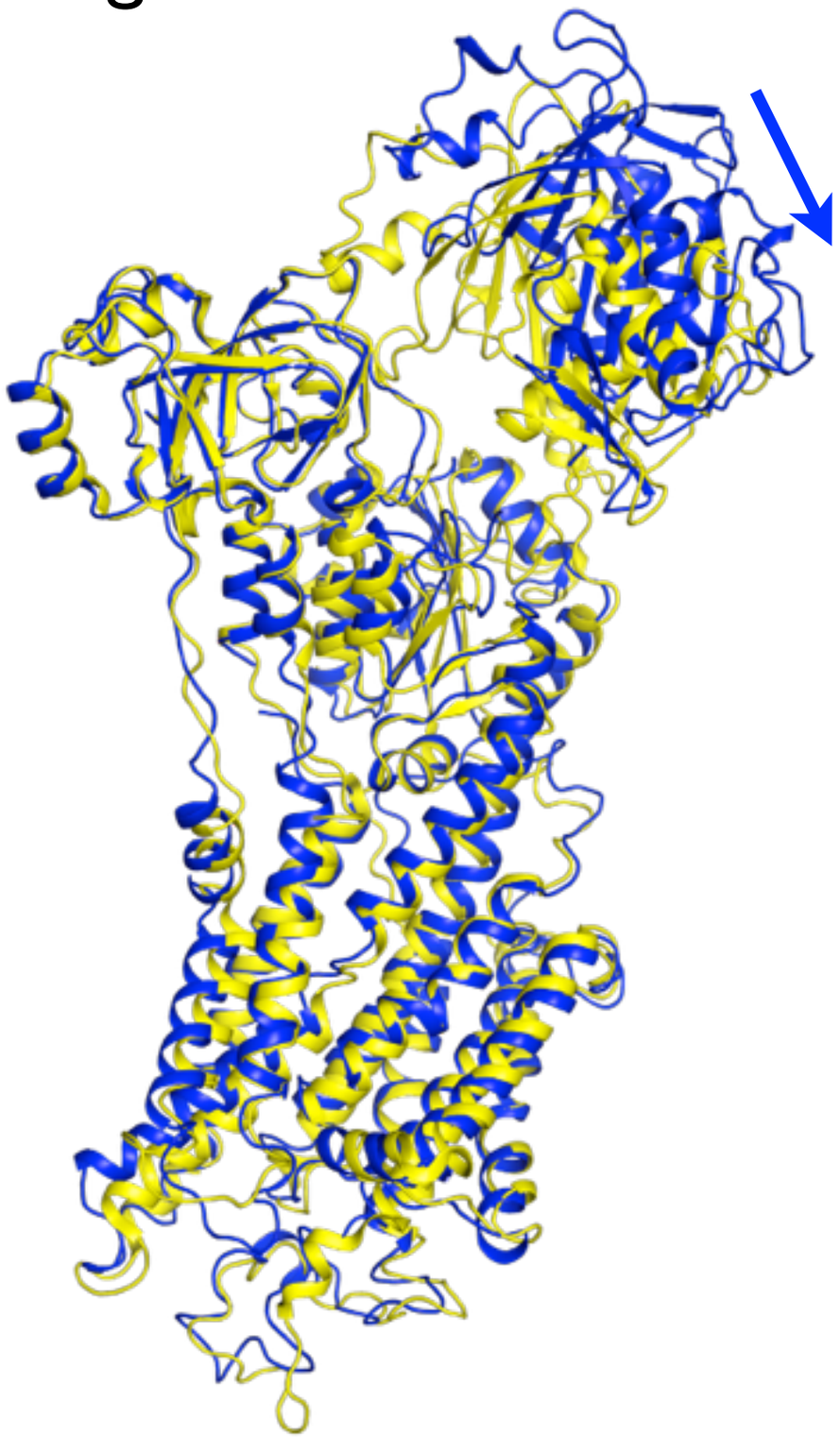
Starting model is
PDB ID 1ewt

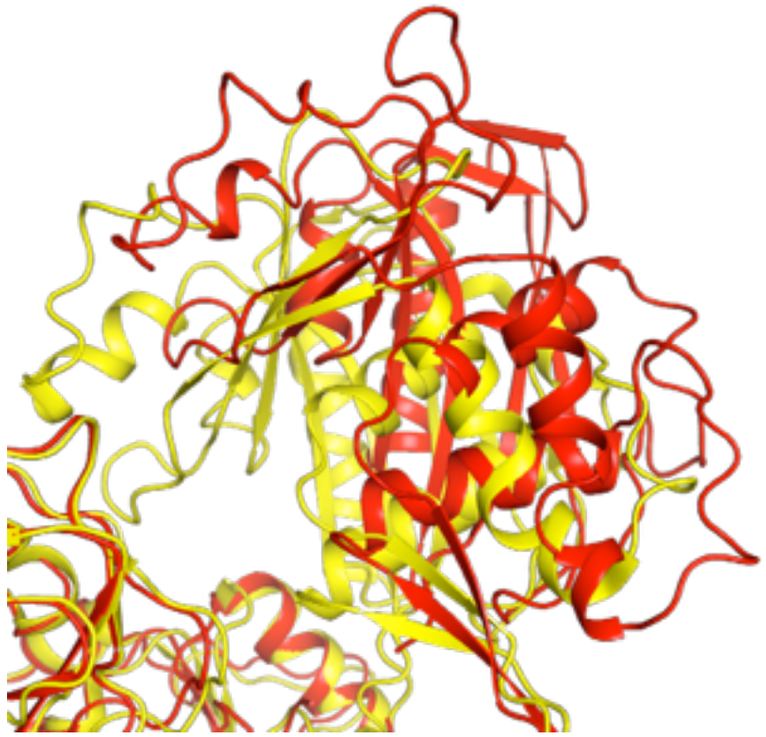


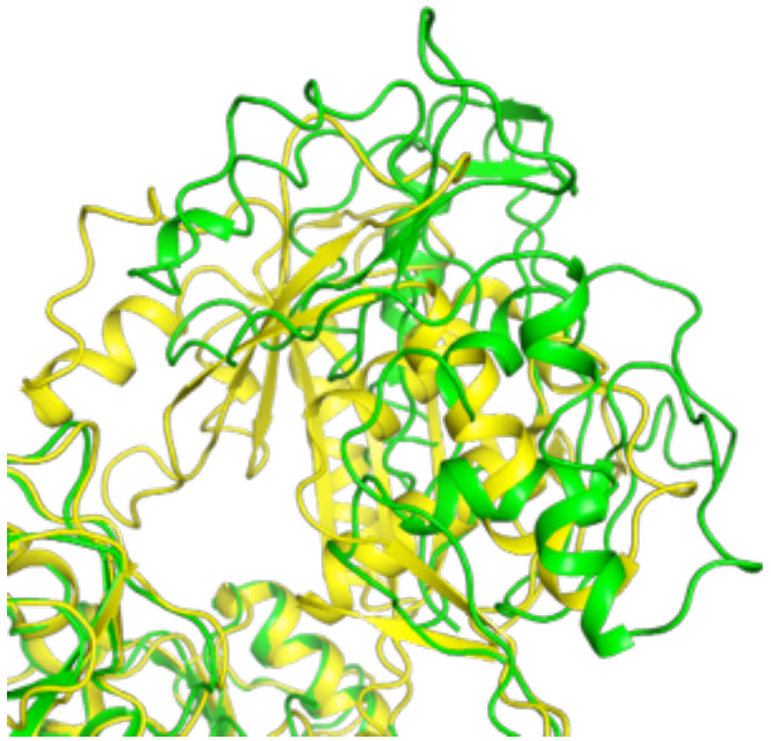
start (r_free=0.45)
phenix (r_free=0.40)
CNS-DEN (r_free=0.30)
Rosetta+phenix (r_free=0.23)
published (r_free=0.20)

Calcium pump ATPase (3fps, starting from 2zbg)

- Large conformational change required (starting model in blue)



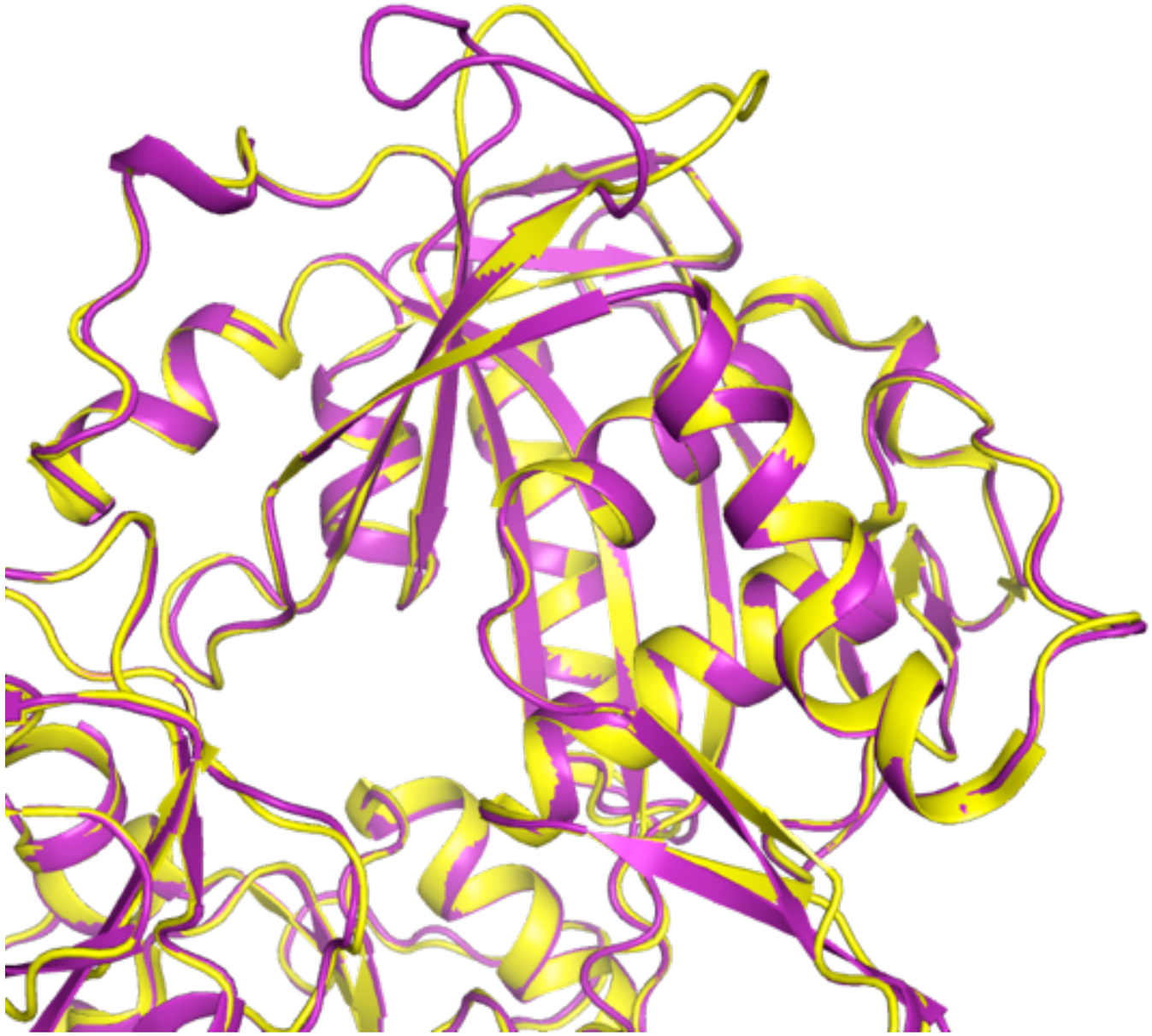


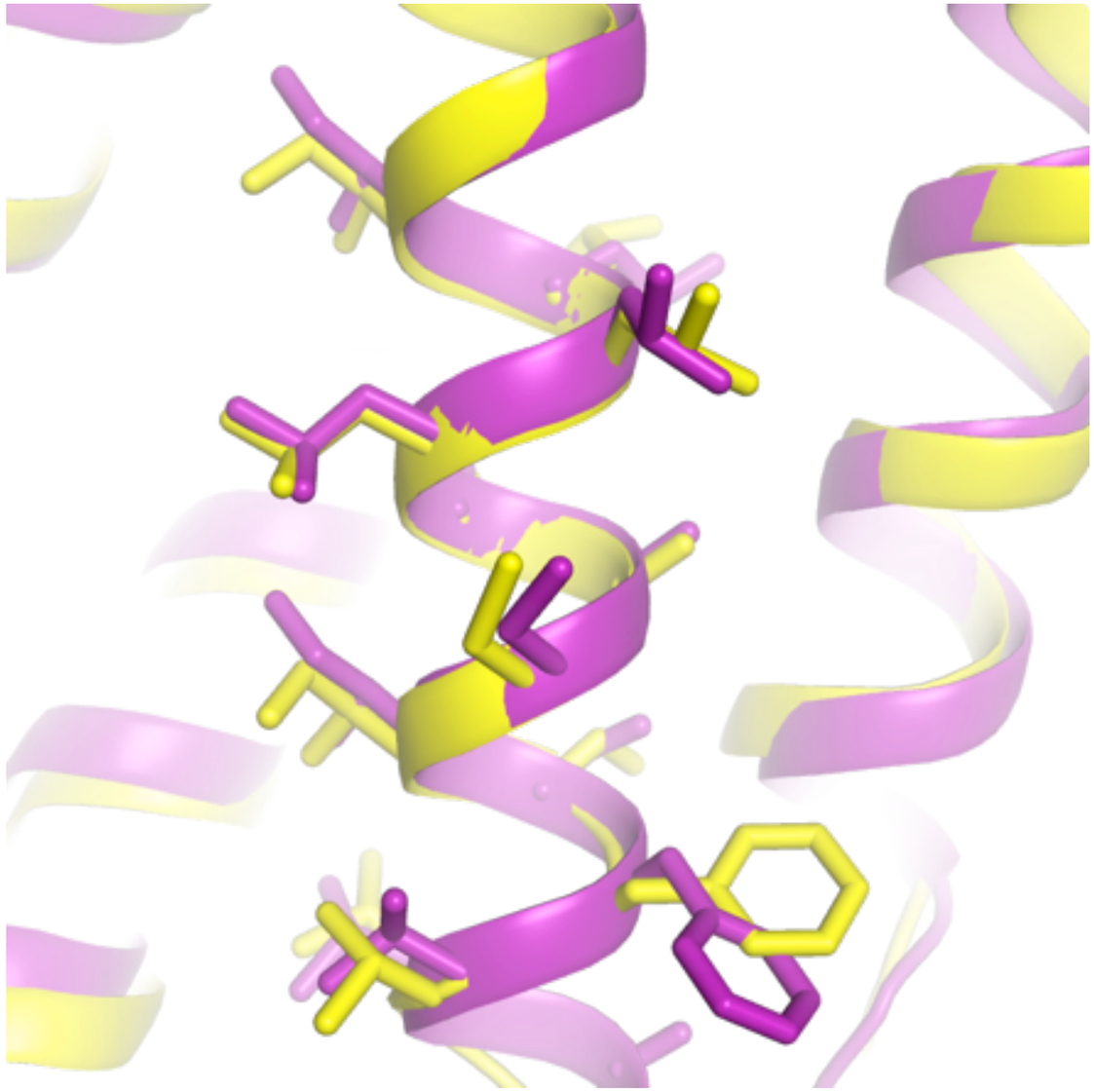


Neither Phenix alone (red model) nor CNS-DEN (green model) is able to refine the large domain movement

Calcium pump ATPase (3fps, starting from 2zbg)

- Rosetta+Phenix model (purple) very close to published (yellow)





Performance on structures near convergence

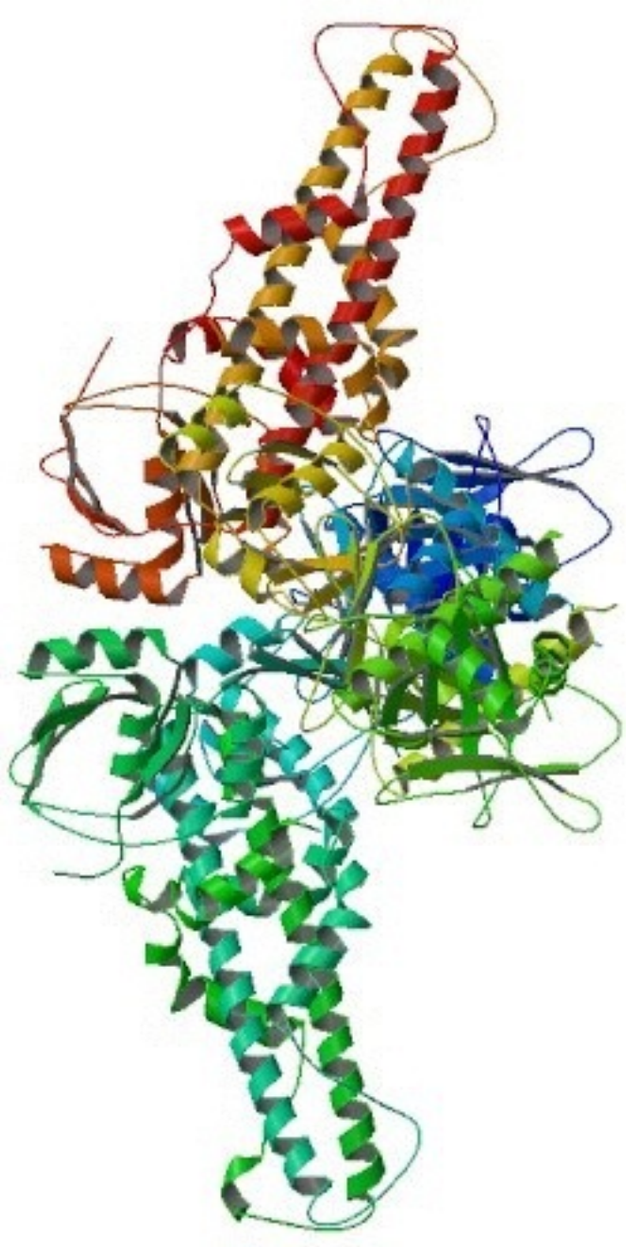
3qkr, 3.4Å [Williams et al. (2011) NSMB 18:423-431]

	published (REFMAC)	phenix (current)	rosetta +phenix	rosetta+phenix, then phenix.refine
Rama. outliers	0.25%	0.25%	0.25%	0.5%
Rota. outliers	14.1%	11.9%	0.28%	0.55%
Clashscore	15	2.36	3.38	1.18
MPscore	**2.9**	**2.23**	**1.38**	**1.16**
rms(bonds)	0.008	0.003	0.013	0.002
rms(angles)	1.25	0.66	1.89	0.56
R-work	0.211	0.199	0.202	0.193
R-free	0.274	0.242	0.258	0.256

phenix.refine improvements over the years

Older versions tended to improve R-factors at the expense of chemistry; recent versions are much more conservative

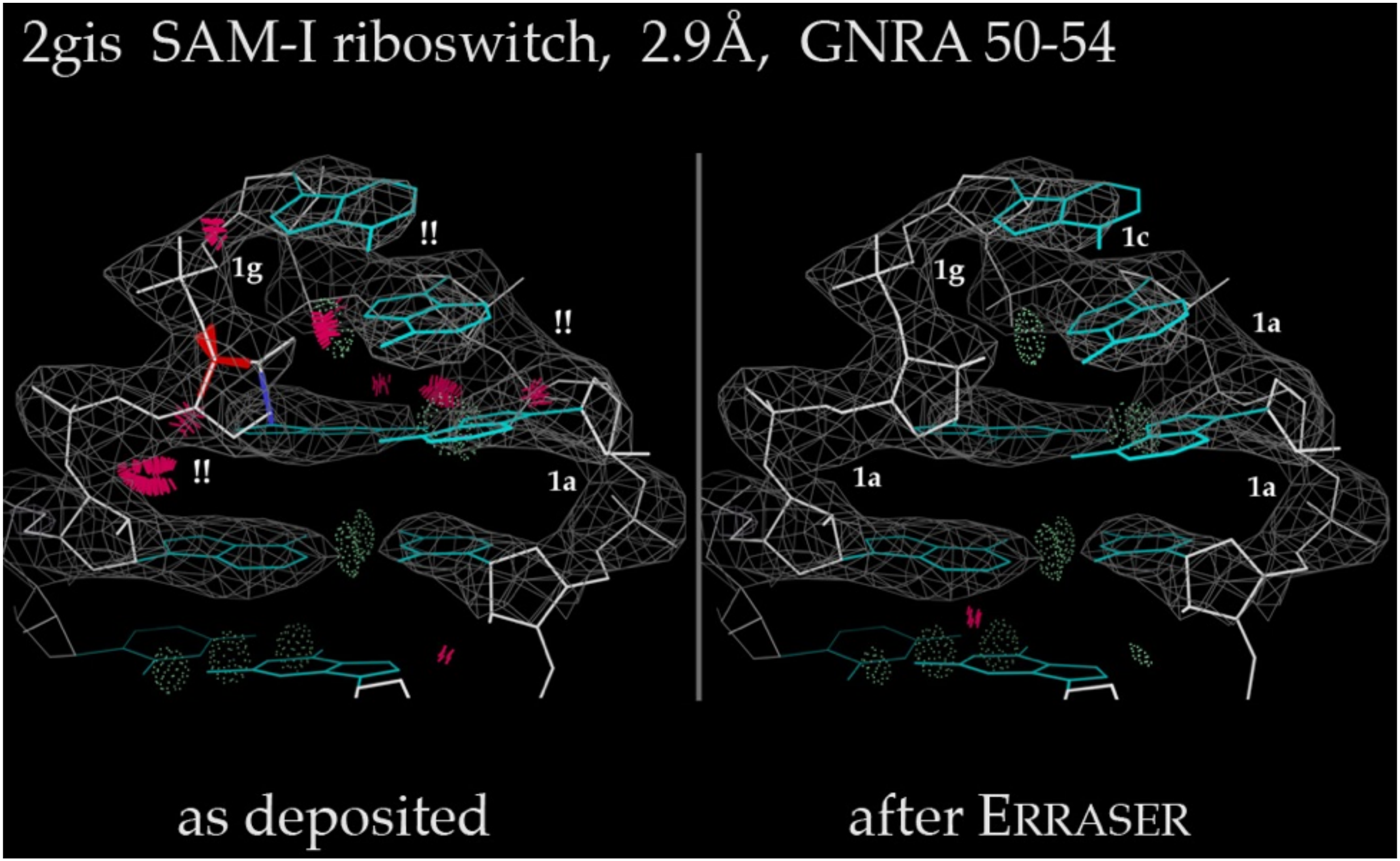
XLF-XRCC4 (3.97Å)	published (phenix 1.6.1)	phenix (current)	rosetta+phenix, then phenix.refine
Rama. outliers	6.82%	3.4%	2.5%
Rota. outliers	15.65%	12.0%	0.47%
Clashscore	58.4	12.1	3.6
MPscore	**3.92**	**3.04**	**1.66**
R-work	0.358	0.314	0.312
R-free	0.369	0.361	0.360



You can't always expect an improvement in R-factors — some structures are just difficult!

ERRASER: Rosetta rebuilding for RNA

Fang-Chieh Chou, Rhiju Das - *Chou et al. (2013) Nature Methods 10:74-76.*



(image courtesy of Jane Richardson)

Part II: accelerating high-throughput ligand-binding studies

phenix.ligand_pipeline

Process input files

Import data and R-free flags

Xtriage

(If restraints are already available for the target ligand, eLBOW will simply pass these on unchanged)

eLBOW

Phaser

(plus Sculptor, phenix.pdbtools, and phenix.find_alt_orig_sym_mate)

mmtbx.prune_model

phenix.refine (rigid-body and restrained)

AutoBuild

Coot

(Intermediate Coot steps are optional, used only in interactive mode)

(Several changes from default settings: conformers=20, delta_phi_ligand=20, n_group_search=5)

LigandFit

Coot

mmtbx.extend_sidechains

phenix.refine

(This step is now identical to the protocol used for re-refinement of PDB structures)

validation & maps

Coot

An academic study: AKR1C3*

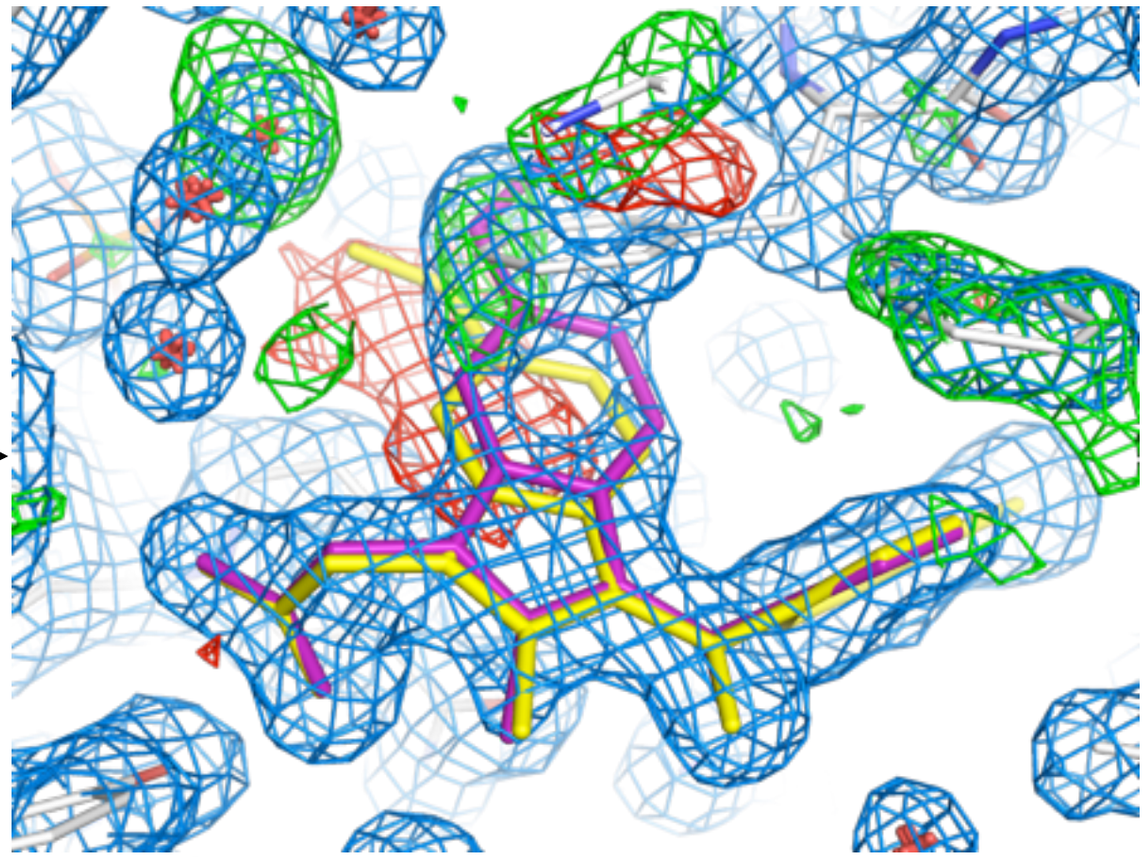
- Using 1s1p as search model (as specified in publication)

ID	d_min	fit	r_work	r_free	rw_pub	rf_pub
3r43	2.0	yes	0.1621	0.1812	0.185	0.226
3r6i	1.95	yes	0.1613	0.1820	0.175	0.206
3r7m*	2.10	yes	0.1642	0.2013	0.182	0.232
3r8g	1.80	yes	0.1555	0.1748	0.174	0.213
3r8h	1.90	yes	0.1522	0.1760	0.164	0.204
3r94	2.01	mostly	0.1613	0.2046	0.175	0.221
3ufy	1.90	yes	0.1653	0.1862	0.173	0.194
3ug8	1.73	partial	0.1626	0.1837	0.173	0.200

Every structure refines to a lower R-free than the published values!

3r94: ligand correctly positioned but one ring needs to flip

3ug8: ligand correctly positioned but needs to be adjusted due to conflicting Phe residue →

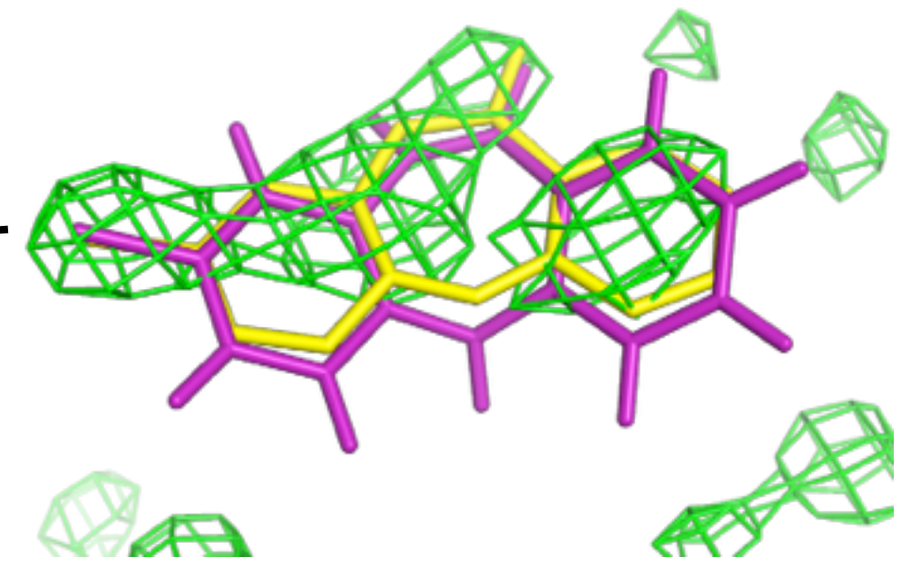


* Flanagan et al. (2012) *PLoS ONE* 7:e43965 [online Aug. 28th]

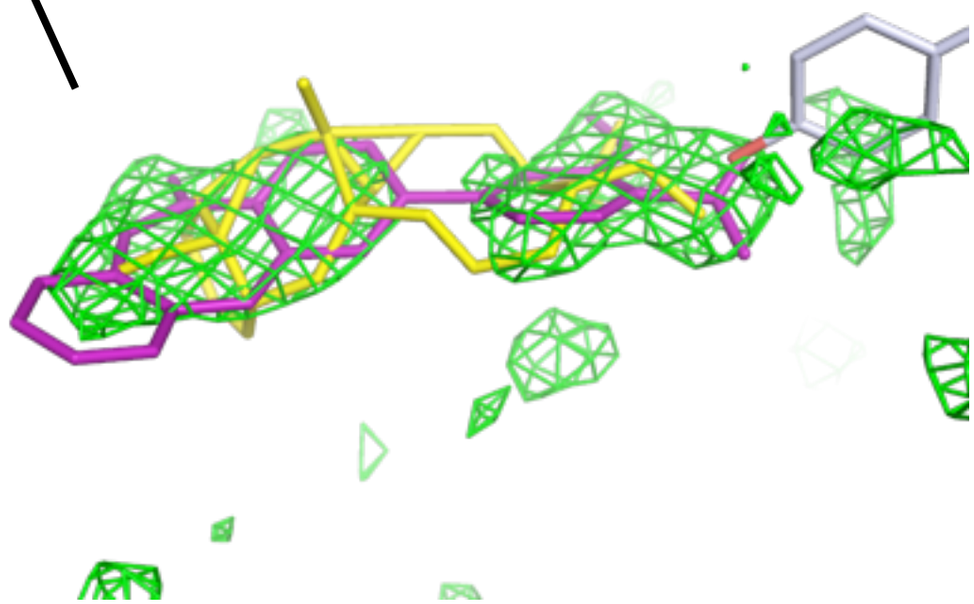
A high-throughput academic study (CHK1)*

ID	d_min	fit	r_work	r_free	rw_pub	rf_pub
4fsm	2.2	yes	0.1984	0.2237	0.175	0.198
4fsn	2.1	yes	0.1834	0.2205	0.176	0.217
4fsq	2.4	yes	0.1923	0.2348	0.175	0.216
4fsr	2.49	yes	0.1882	0.2133	0.175	0.198
4fst	1.9	yes	0.1699	0.1982	0.166	0.185
4fsu	2.09	yes	0.1791	0.1968	0.173	0.181
4fsw	**2.3**	**no**	**0.2186**	**0.2577**	**0.184**	**0.212**
4fsy	2.29	yes	0.1917	0.2282	0.181	0.226
4fsz	2.3	yes	0.1870	0.2249	0.173	0.220
4ft0	**2.3**	**no**	**0.2038**	**0.2404**	**0.172**	**0.203**
4ft3	2.5	yes	0.1864	0.2421	0.168	0.213
4ft5	2.39	yes	0.1789	0.2302	0.178	0.223
4ft7	2.2	yes	0.1710	0.1972	0.171	0.195
4ft9	2.2	yes	0.1802	0.2226	0.160	0.206
4fta	2.4	yes	0.1822	0.2326	0.170	0.211
4ftc	2.0	yes	0.1721	0.2085	0.169	0.194
4fti	2.19	yes	0.1688	0.2016	0.174	0.194
4ftj	2.2	yes	0.1848	0.2058	0.180	0.205
4ftk	2.3	yes	0.1829	0.2164	0.167	0.206
4ftl	2.5	yes	0.2034	0.2307	0.178	0.226
4ftm	1.9	yes	0.1719	0.2045	0.172	0.201
4ftn	2.02	yes	0.1855	0.2220	0.173	0.213
4fto	2.08	yes	0.1864	0.2153	0.181	0.195
4ftq	**2.0**	**yes**	**0.1847**	**0.2184**	**0.184**	**0.220**
4ftr	**2.25**	**no**	**0.2160**	**0.2544**	**0.172**	**0.214**
4ftt	**2.29**	**no**	**0.2127**	**0.2629**	**0.171**	**0.215**
4ftu	2.1	yes	0.1797	0.2143	0.190	0.210

4fsw: weak density, inhibitor needs to bend in middle



4ft0: weak density, incorrect ligand geometry, and a stray sidechain

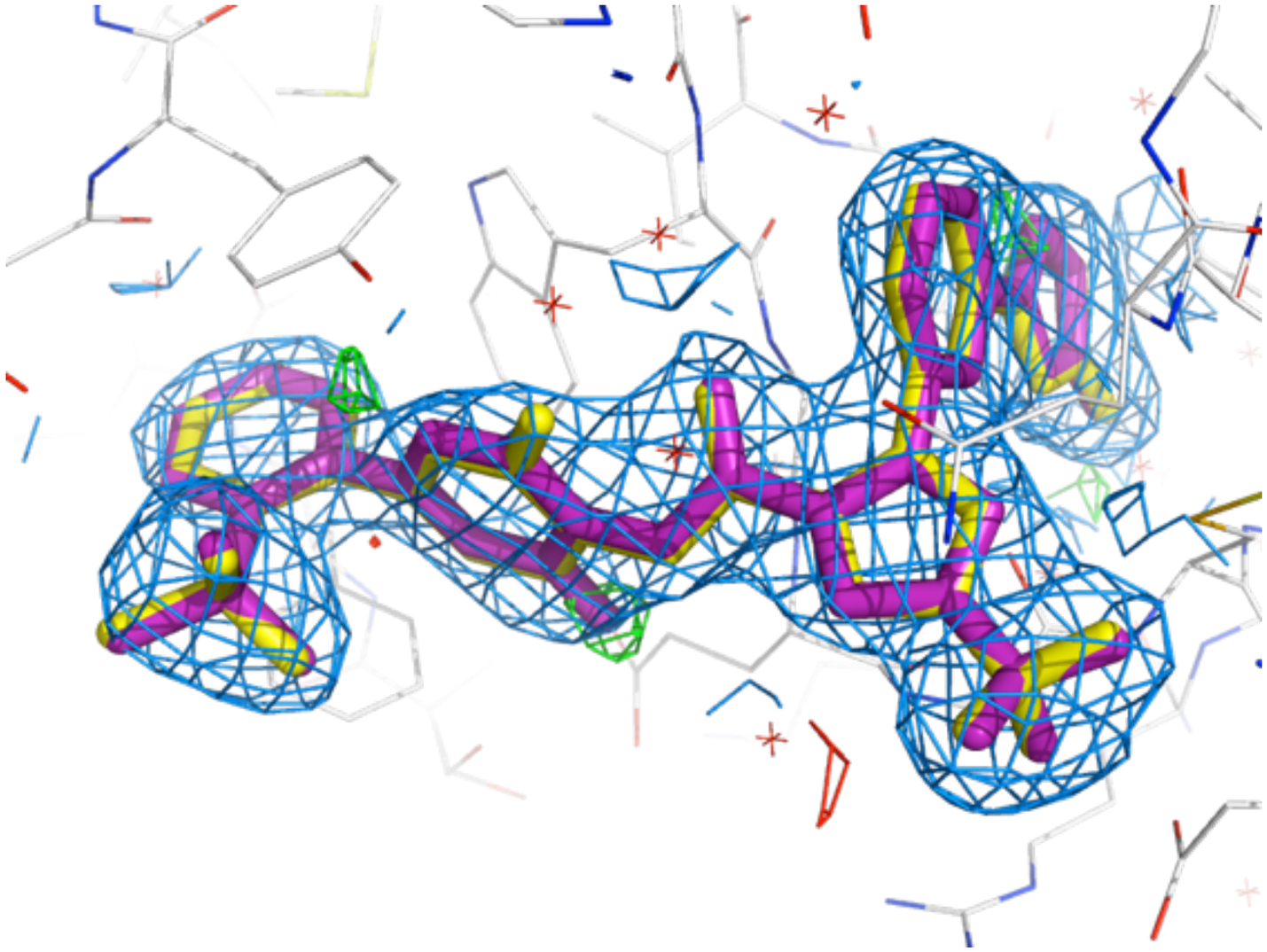


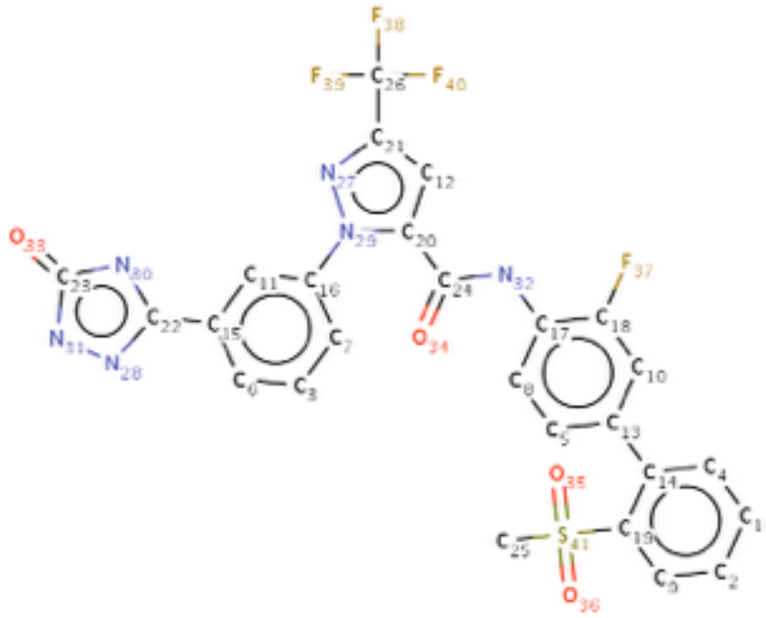
* *from Stuckey lab, U. Michigan - currently unpublished*

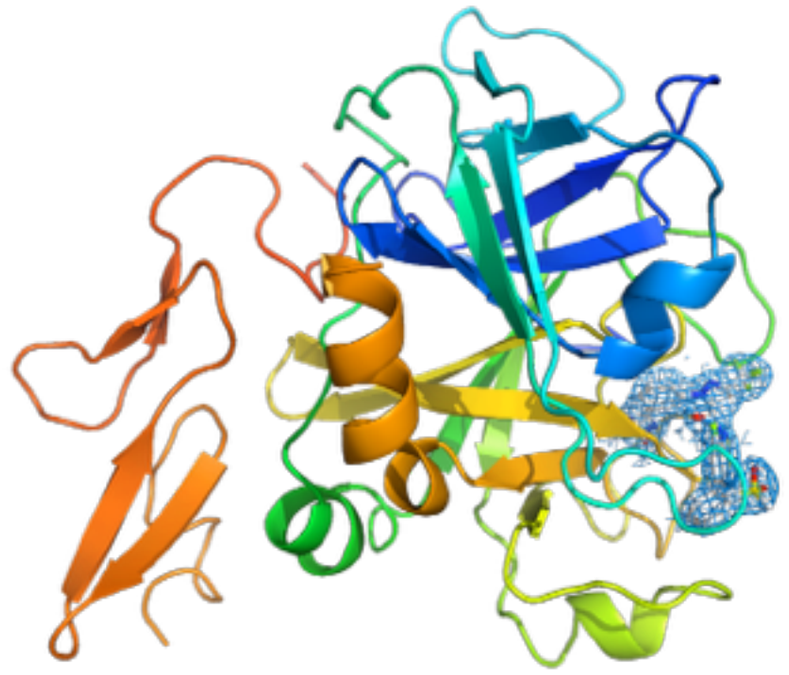
A representative industrial structure

- 3kqb: Factor Xa with inhibitor (BMS)

Resolution = 2.25Å; MR search model: 3ffg
Published R-work/R-free = 0.189/0.221 (purple sticks)
Pipeline R-work/R-free = 0.1771/0.2077 (yellow sticks)





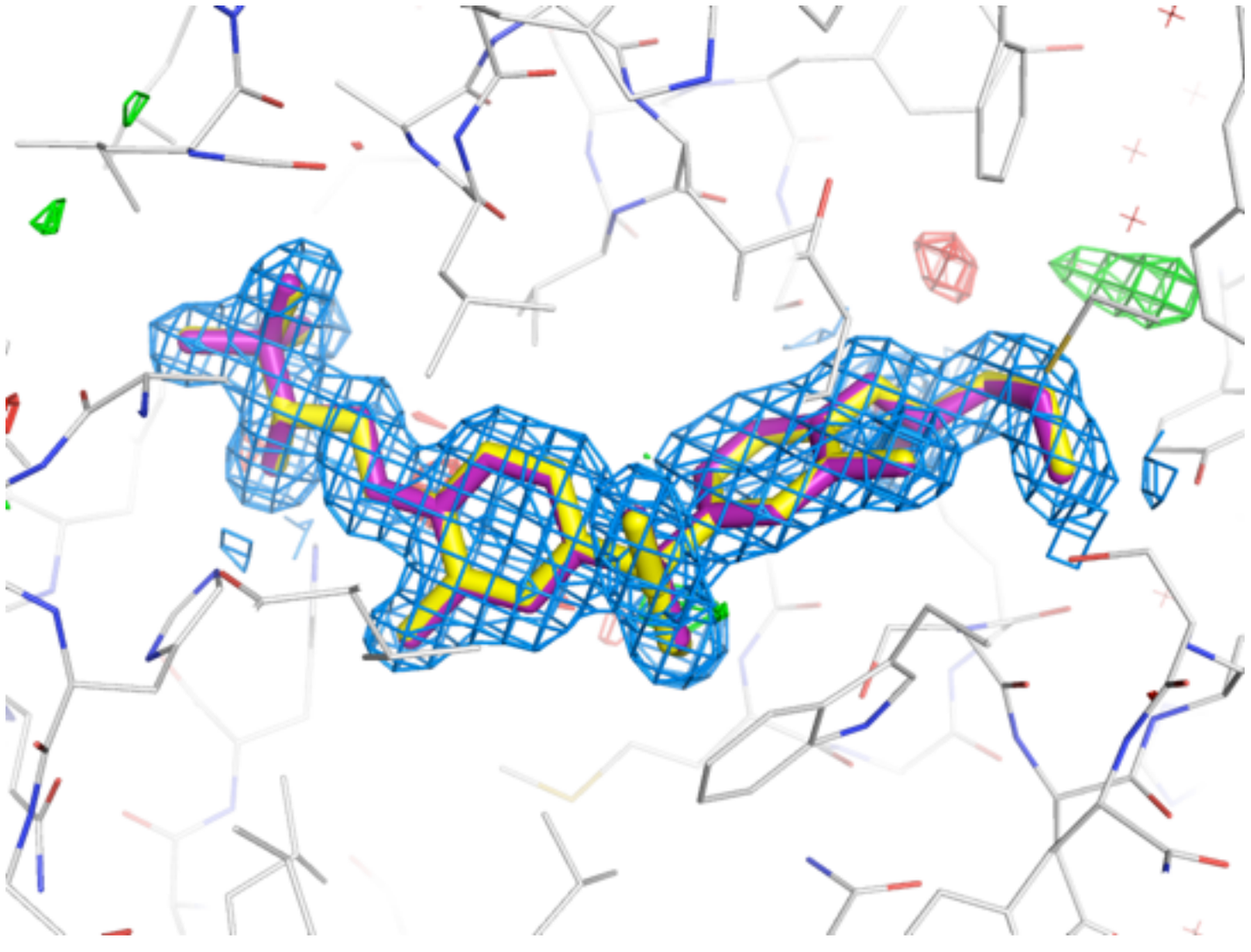


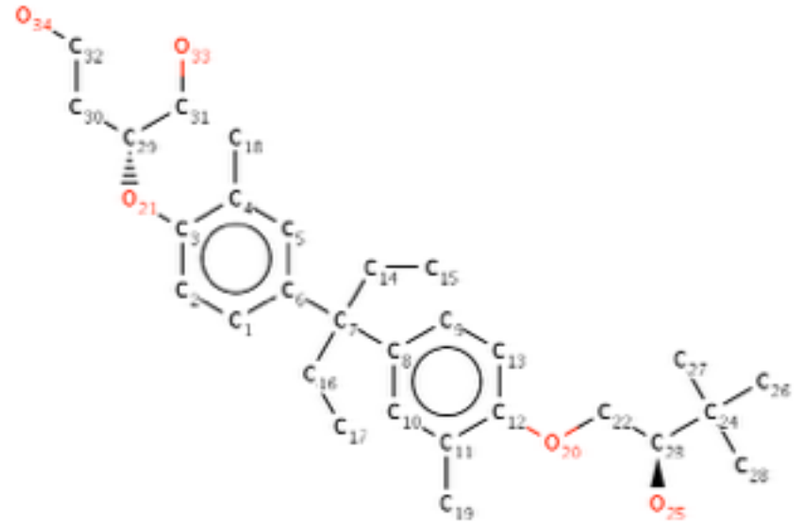
Also very easy - lower R-free than published structure

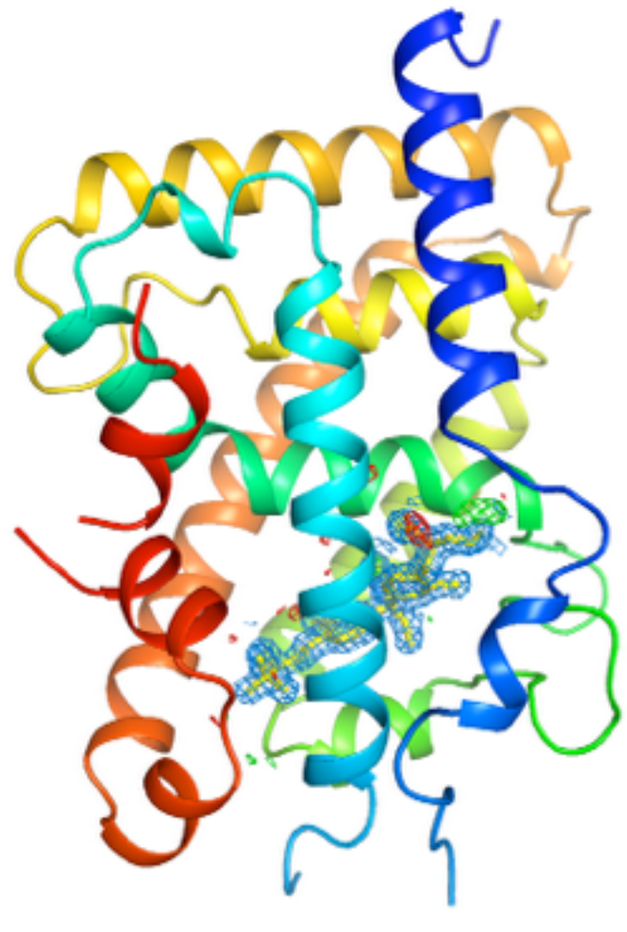
A representative academic structure

- 3aun: Vitamin D receptor (academic)

Resolution = 1.81Å; MR search model: 2zfx
Published R-work/R-free = 0.198/0.237 (purple sticks)
Pipeline R-work/R-free = 0.1712/0.2054 (yellow sticks)



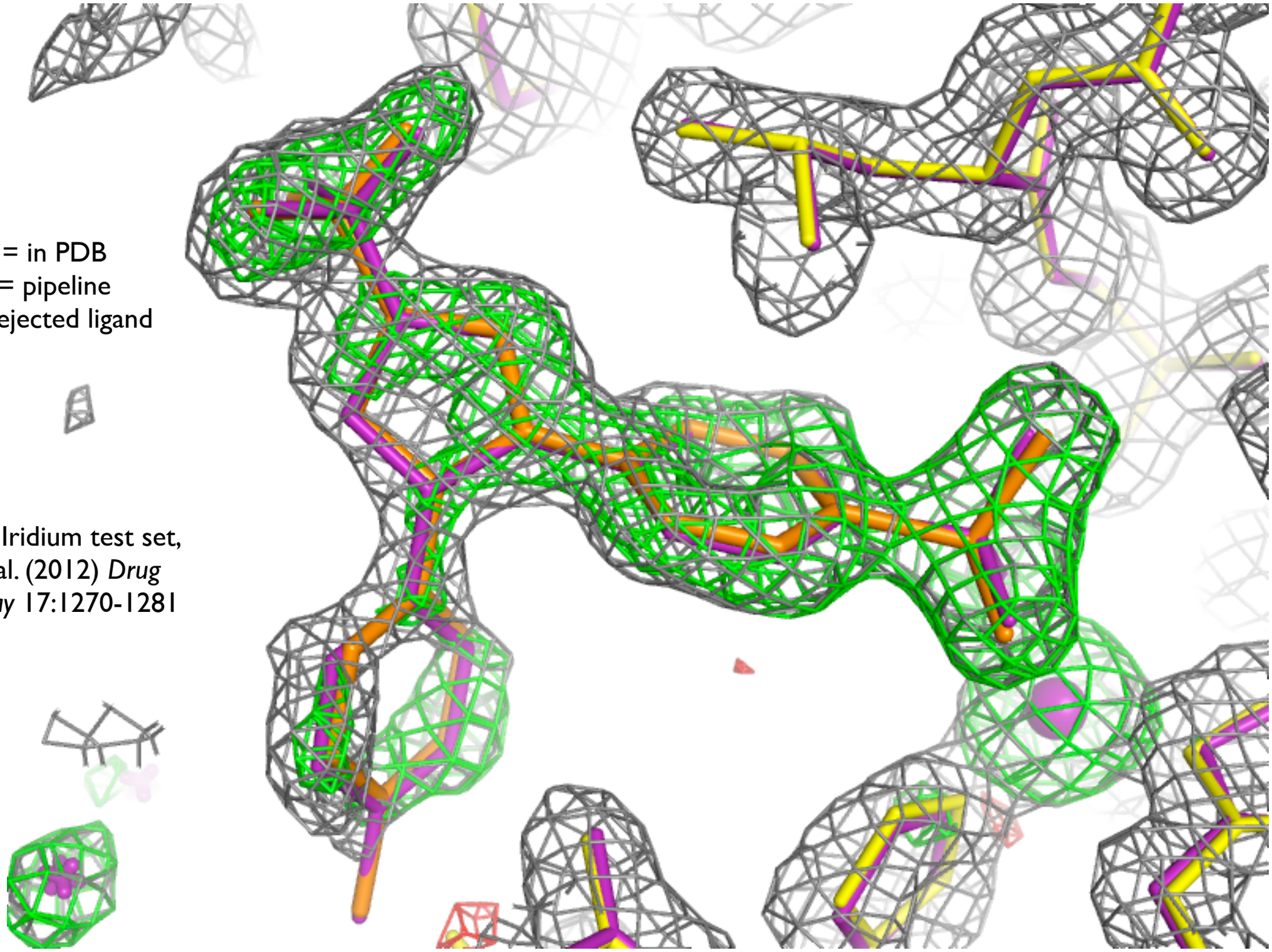




Significantly better than published structure

1oq5* (1.5 Å): poor initial CC

purple = in PDB
yellow = pipeline
orange = rejected ligand

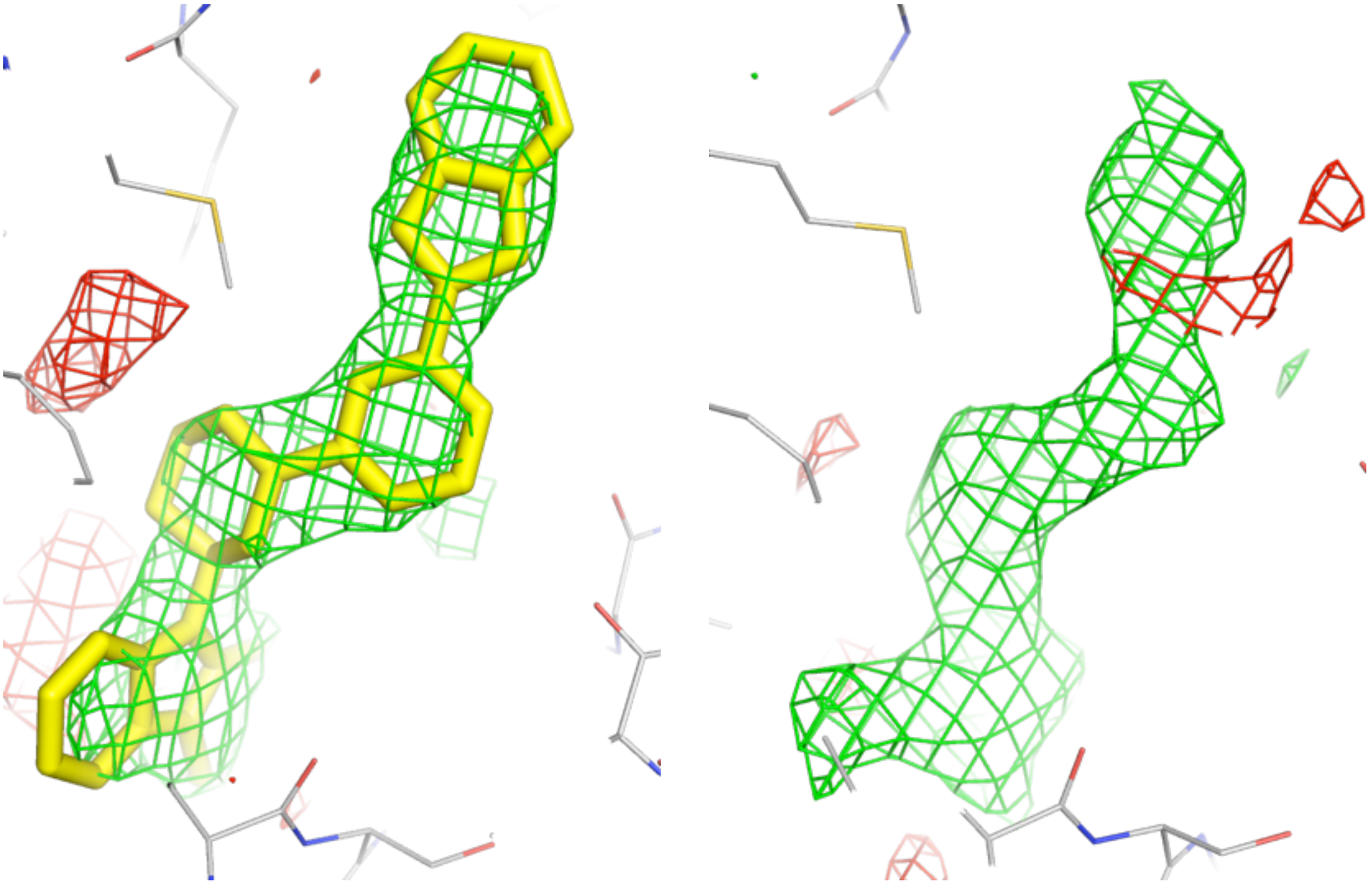


* taken from Iridium test set, Warren et al. (2012) *Drug Discovery Today* 17:1270-1281

Re-running with min_ligand_cc_keep=0.6 was successful.

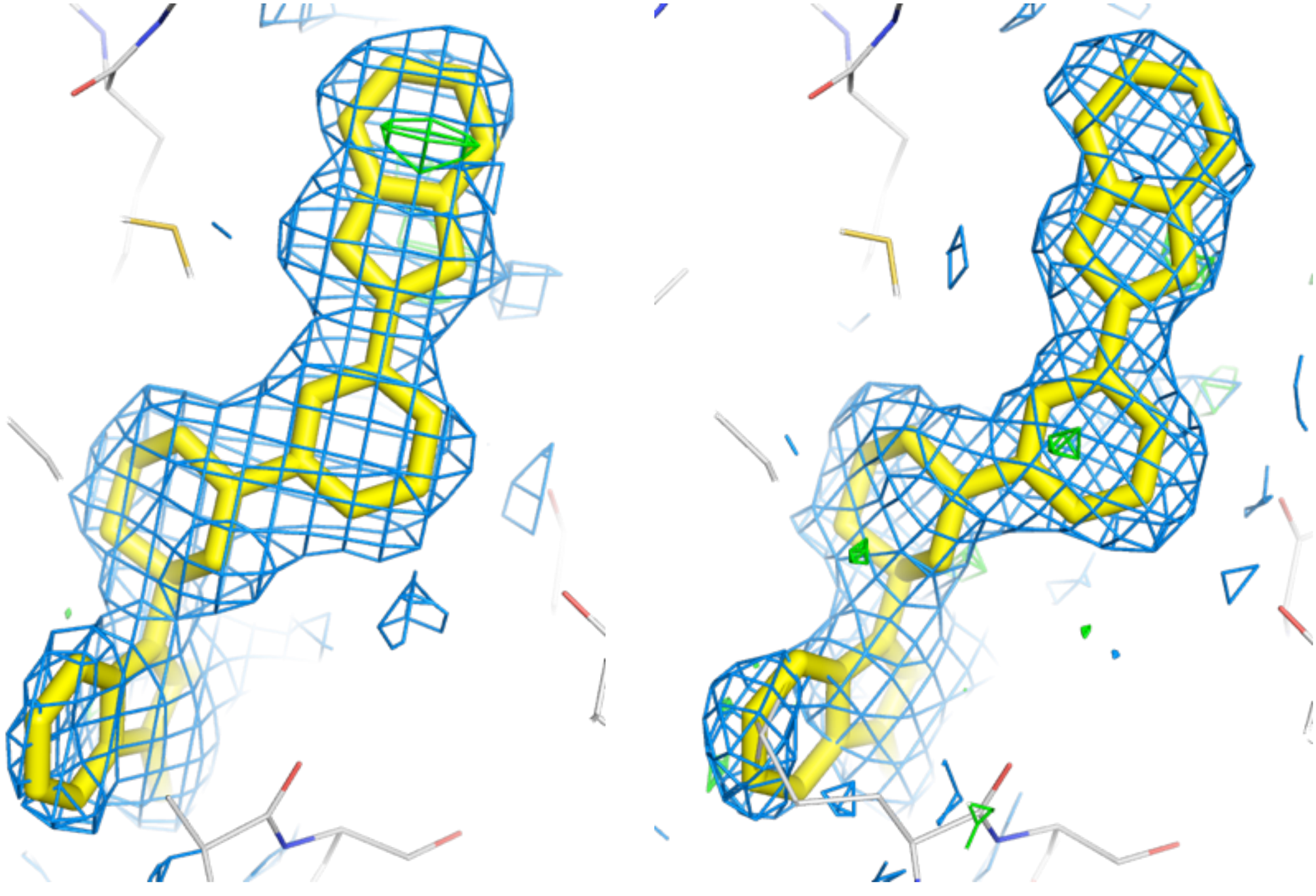
Taking advantage of NCS

- 3qj9 (Amgen): one copy correctly placed by LigandFit, but another ends up in density for missing protein residues



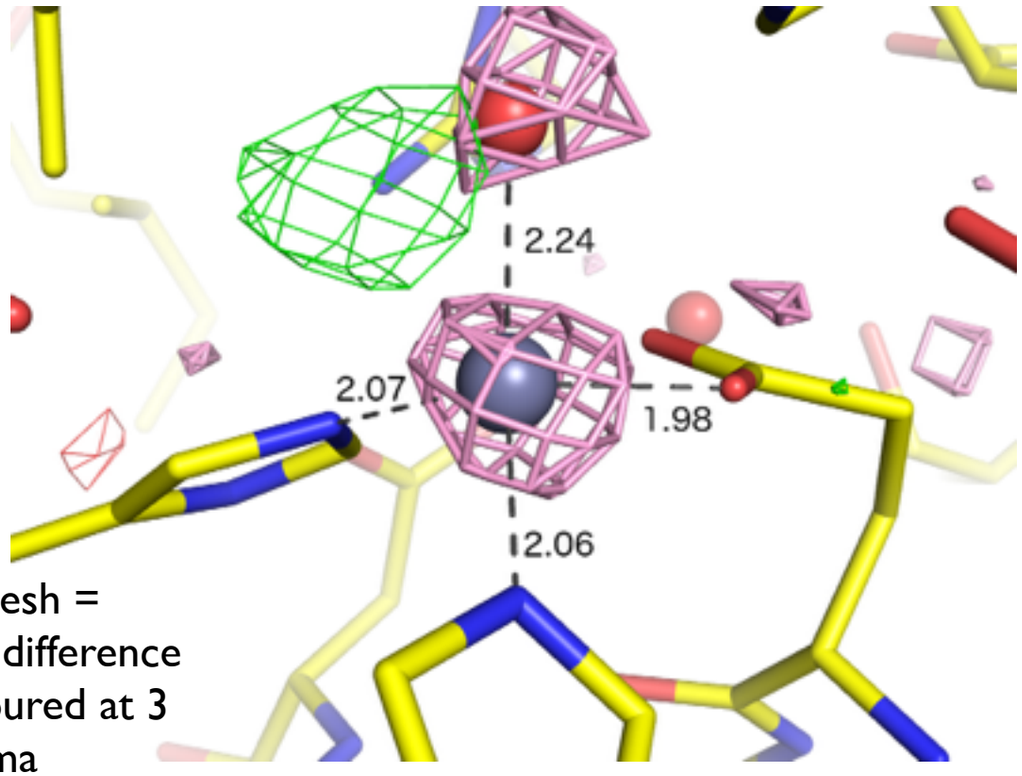
Taking advantage of NCS

- 3qj9: applying NCS operator to the good ligand (CC=0.873) results in good fit to 2mFo-DFc map

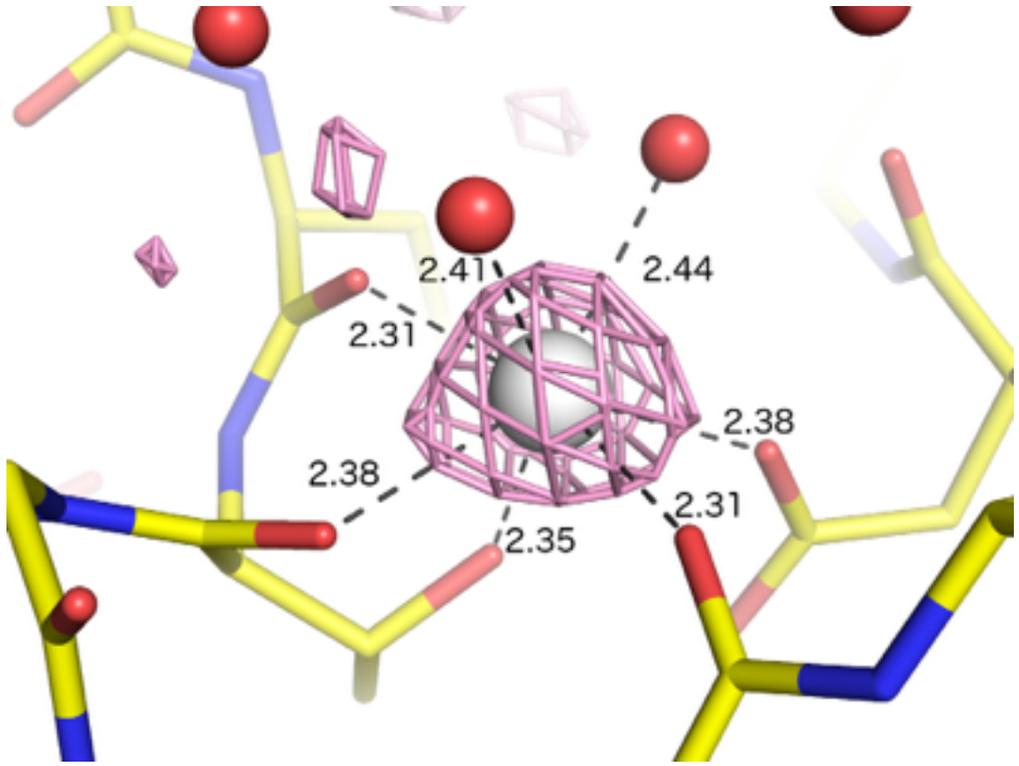


Ion placement as an extension of water picking

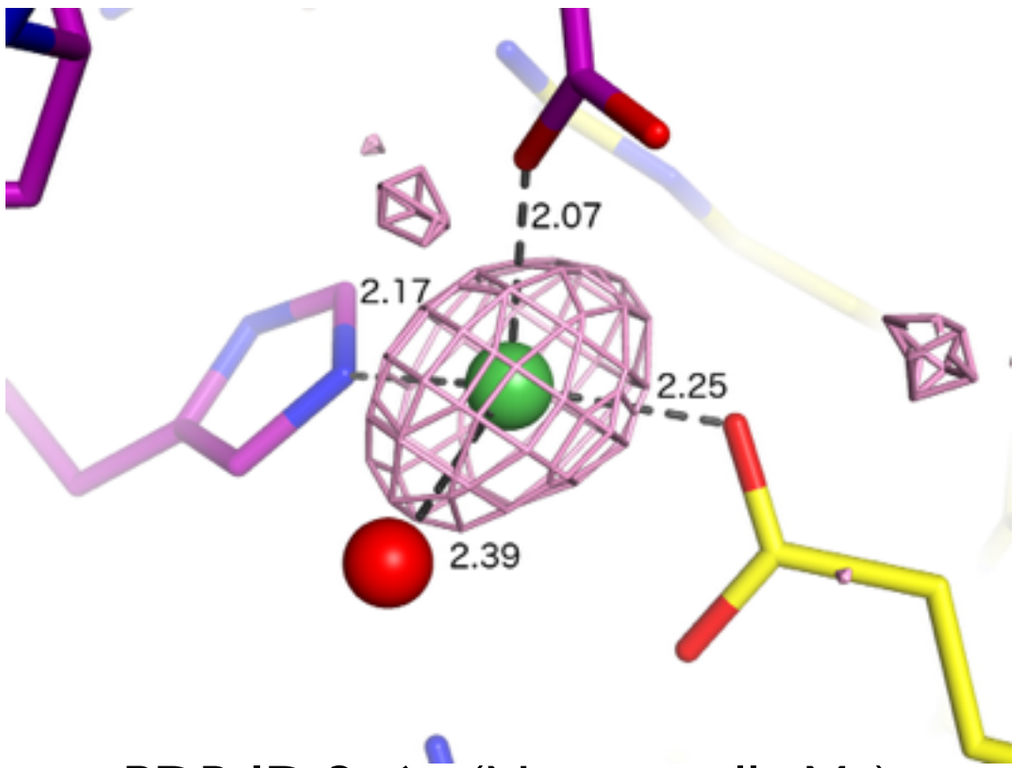
pink mesh = anomalous difference map contoured at 3 sigma



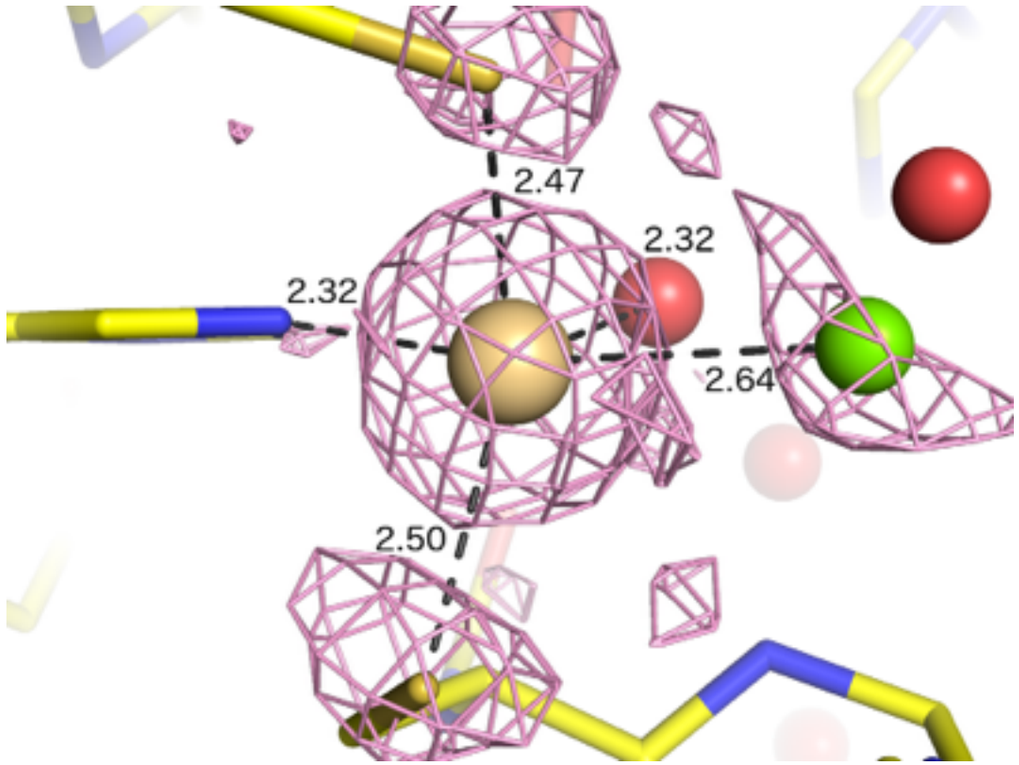
PDB ID 2hwz (Zn)



PDB ID 2hwz (Ca)

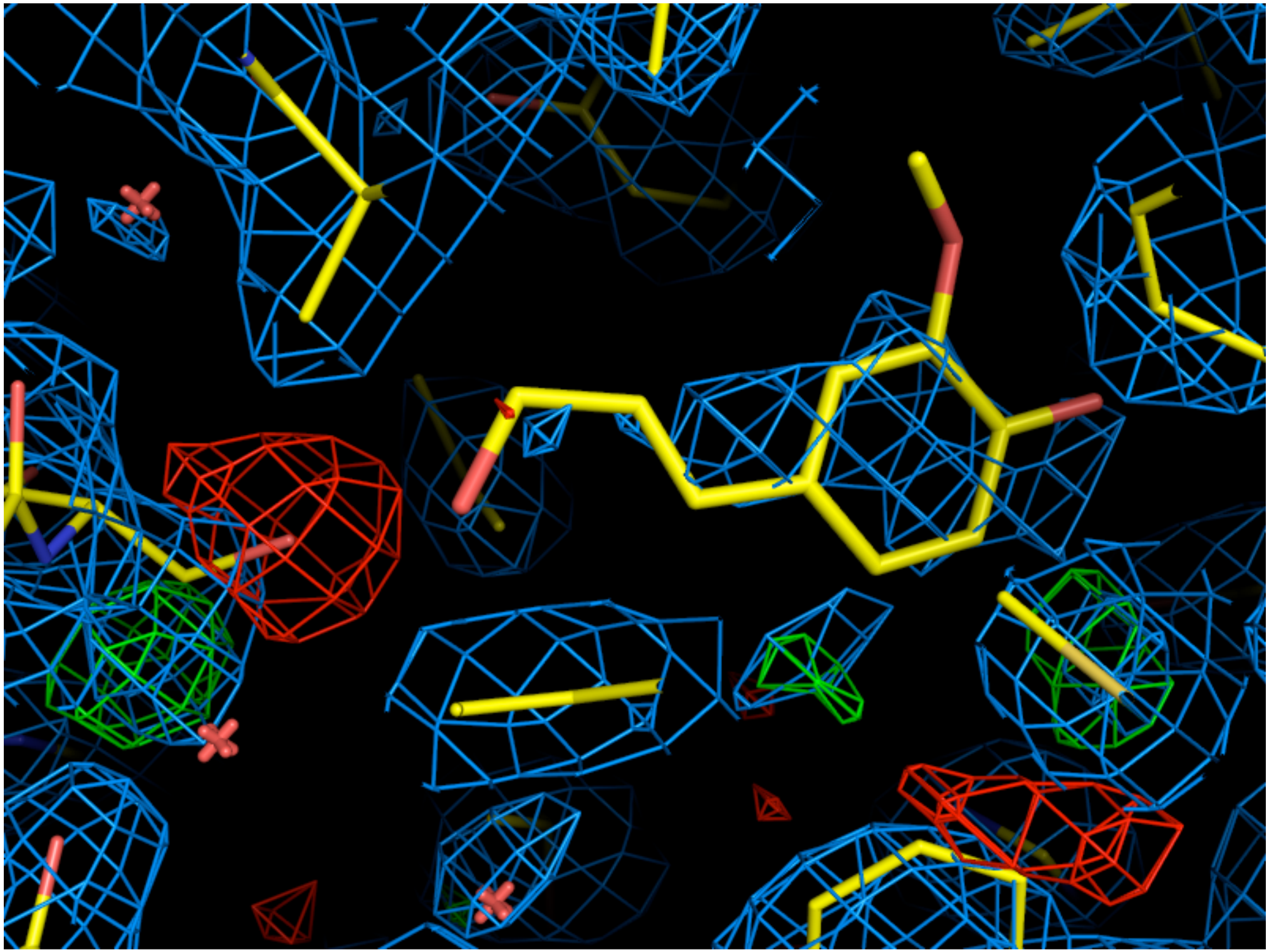


PDB ID 3n6q (Ni, originally Mg)



PDB ID 3bob (Cd, Cl)

The importance of validation



See also Pozharski et al. (2013) Acta Cryst D69:150-167.

Acknowledgments

University of Washington

Frank DiMaio, David Baker

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Youval Dar, Nat Echols, Richard Gildea, Ralf Grosse-Kunstleve, Jeff Headd, Nigel Moriarty, Nader Morshed, Ian Rees, Nicholas Sauter, Oleg Sobolev, Peter Zwart

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Duke University

Jane & David Richardson, Vincent Chen, Swati Jain, Gary Kapral, Chris Williams, Bryan Arendall, Bradley Hintze

Funding: NIH, Phenix Industrial Consortium