



COMPUTATIONAL CRYSTALLOGRAPHY INITIATIVE

Crystallographic Structure Validation

Pavel Afonine

Computation Crystallography Initiative
Physical Biosciences Division
Lawrence Berkeley National Laboratory, Berkeley CA, USA

Australasian Crystallography school
17th-24th July, 2010

PHYSICAL BIOSCIENCES DIVISION

Why validation?

- **Crystallography is not exact science** (Gerard J. Kleywegt):
 - Subjectivity in map interpretation:
 - we interpret the maps
 - some people more skilled than the other
 - you may be experienced but in rush or tired
 - we program the software that interprets the maps, builds the model
 - programs may contain bugs
 - results of automated protocols are not guaranteed to be 100% error-free
 - insufficient amount of data (typically at low resolution) creates multiple possibilities for interpretation
 - Subjectivity in refinement:
 - different model parameterization
 - different weights

A good model

A good model should be good...

...physically

- Packing, contacts

...chemically

- Bonds, angles, planarity, chirality, non-bonded (charge) interactions

...crystallographically

- R-factors, B-factors, density fit, bulk-solvent

...statistically

- No under-modeling (under-refinement) and no over-fitting (over-modeling)
- Model global quality figures should be in agreement with corresponding values found in similar structures

Model, data and model-to-data fit quality indicators

▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

What affects model quality

- Data quality (resolution, completeness, twinning)
 - crystal quality
 - data collection
- Experience of researcher
 - map interpretation is subjective (you interpret it)
 - refinement parameterization and strategy (too many options)
- Pressure to publish (paper, thesis, etc)
- “Good” R -factors (overfitting, NCS or twinning not considered when creating free- R flags)
- Post-refinement manipulations:
 - Final look before PDB deposition: I don’t like this water, let’s remove it (often statistics is not updated after such manipulation)
 - Removing “riding” hydrogen atoms naively thinking that they can be easily restored
 - Re-setting *high* B-factors, removing ANISOU records after TLS refinement.
- Misusing quality indicators (deciding about single water using R_{FREE})

Quality filters

- **Who checks your structure**

- Crystallographer (you)
- Software you use
- Your boss
- Reviewer (of your paper or thesis)
- PDB deposition (software and people)
- Community (those who eventually may come across of your structure or even use it for his/her research)

- **Ignored (or unnoticed) problems:**

- It will be discovered anyway, sooner or later
- Later you catch it – worse for you
- Better late than never

Model, data and model-to-data fit quality indicators

▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

R-factor

- R-factor formula

$$R = \frac{\sum_{\text{reflections}} |F_{\text{OBS}} - |F_{\text{MODEL}}||}{\sum_{\text{reflections}} F_{\text{OBS}}}$$

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} e^{-\mathbf{sU}_{\text{CRYSTAL}} \mathbf{s}^t} \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

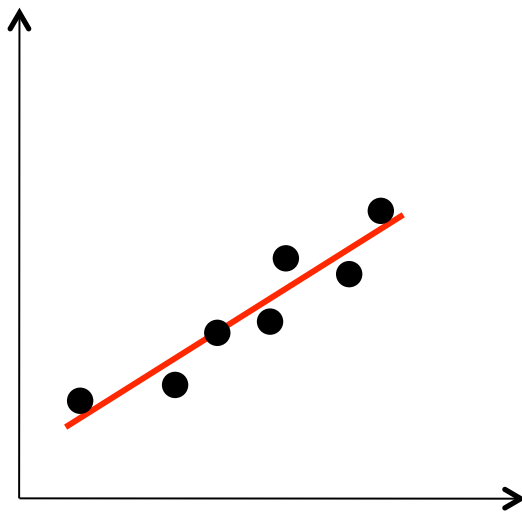
- R-factor values:
 - Expected value for a random model R~59%
 - You can see some model in 2mFo-DFc map, R~30%
 - You can see most of the model in 2mFo-DFc map, R<20%
 - Perfect model R~0%
- Sometimes the R-factor looks very good (you would expect a good model) but the model-to-map fit is terrible... Overfitting.

Overfitting (I)

Let's suppose:

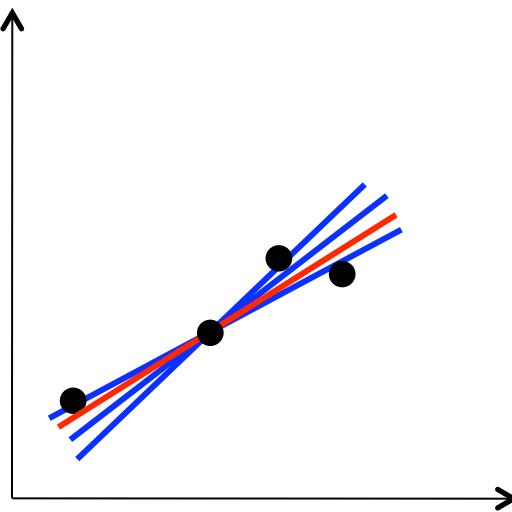
- (red, blue or green) is the model: $y = ax + b$ (2 parameters: a and b)
- is the data.

Lot's of data – one single correct model



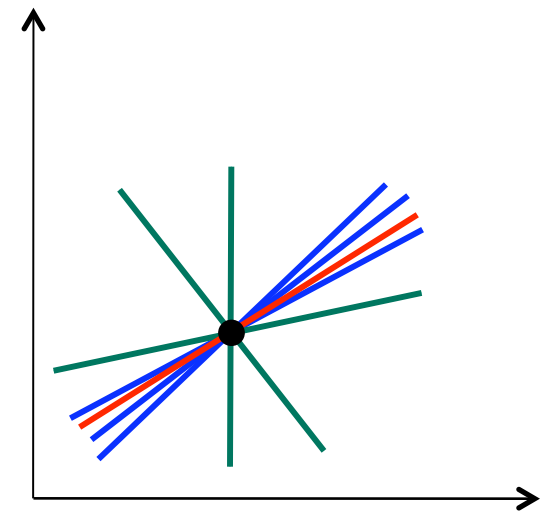
R -factor is good

Less data – more ambiguity, less certainty:
a bunch of models



R -factor may be good too

Little data – variety of models: from good to completely wrong

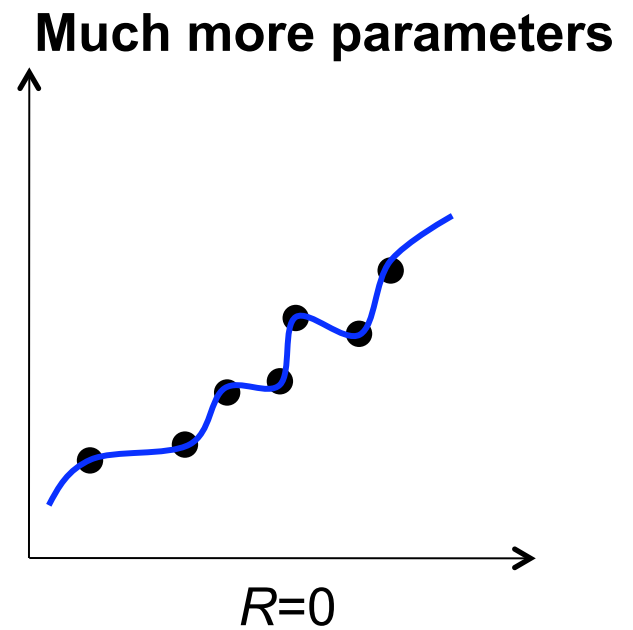
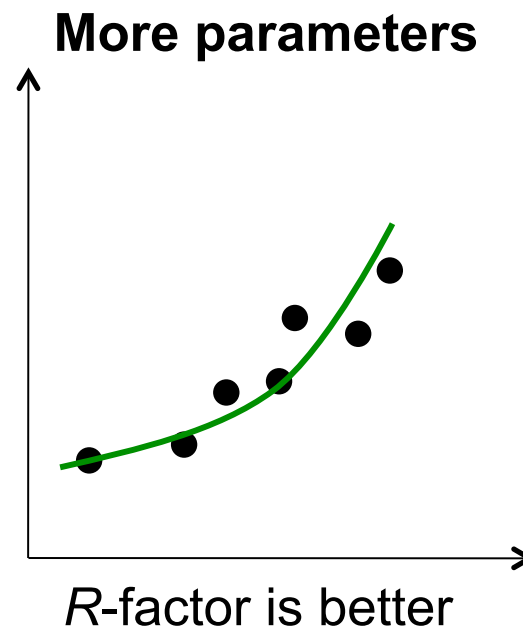
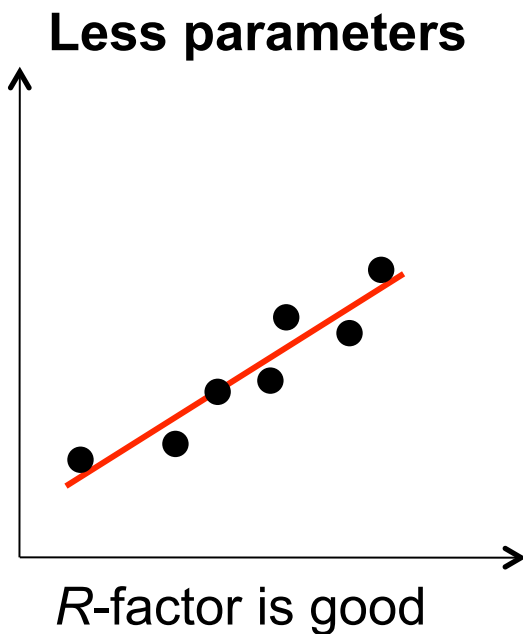


R -factor = 0 for all models
(including wrong ones)

Overfitting (II)

Let's suppose:

- model: $y = ax + b$ (2 parameters: a and b)
- data
- model described using more parameters: $y = ax^2 + bx + c$
- model described using even more parameters: $y = a_1x^n + a_2x^{n-1} + \dots$



Overfitting

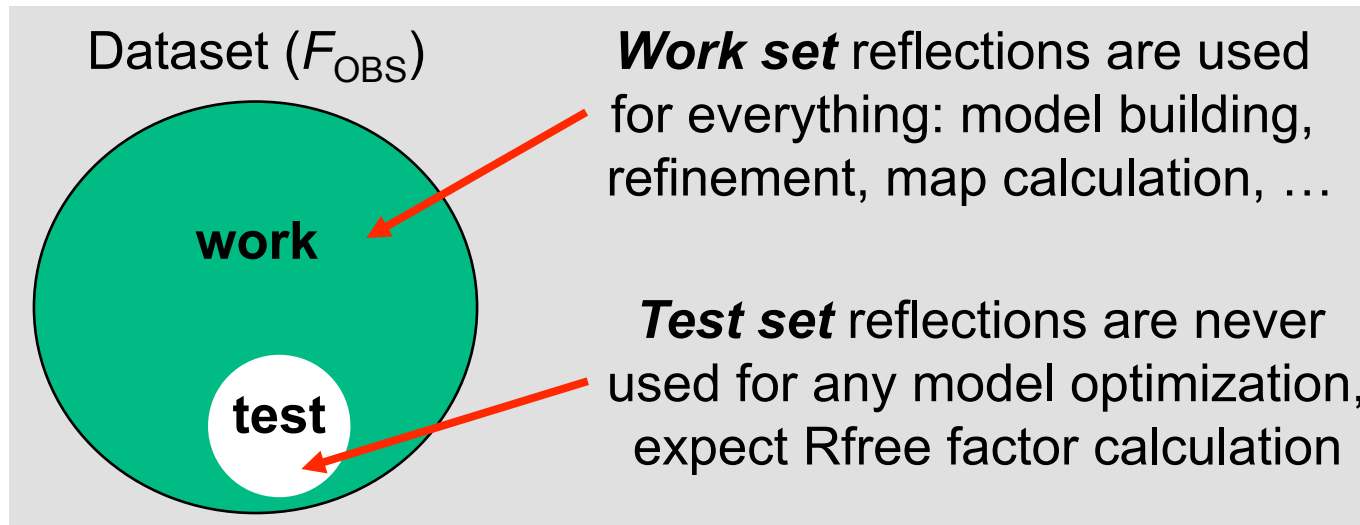
- **What leads to overfitting?**

- Insufficient amount of data (low resolution, poor completeness)
- Ignoring data (cutting by resolution, sigma, anisotropy correction)
- Inoptimal parameterization
- Excess of imagination
- Bad weights

Overfitting

▪ **Solution: cross-validation (R-free factor):**

- At the beginning of structure solution split the data into two sets: *test set* (~5-10% of randomly selected data), and *work set* (the rest).
- From this point on you look at two R-factors: R-work (computed using work set), and R-free (computed using test set)



- **Rationale:** the model that fits well ~90% of work set should fit well 10% of excluded data (test set). Since test set data does not participate in refinement, $R_{free} > R_{work}$. The gap $R_{free} - R_{work}$ depends on resolution and ranges from 5-7% (at medium to low resolution) to ~0.5Å (at ultra-high resolution)

How to tell if R-factor is good

- **Question:** “I got $R_{\text{WORK}}=18\%$ and $R_{\text{FREE}}=23\%$ after refinement, is it a good?”
 - A very common question
 - Answer depends on various factors
- **Answer:**
 - Yes, it’s likely a good result if the data resolution is around 2.5 Å.
 - No, it is very bad result, if the data resolution is 1.0 Å or higher.
- One can ask similar questions about other parameters, such as bond/angles RMSDs, average B-factors, etc...

Rwork and Rfree: typical values depend on resolution

- Say you are refining a structure at 1.0 Å resolution and the R-factors are:
 $R_{\text{WORK}} = 18\%$ and R_{FREE} is 23%.

– Are these values good? Is refinement completed?

- PDB statistics: histograms for R_{WORK} , R_{FREE} , $R_{\text{FREE}} - R_{\text{WORK}}$ for all similar structures:

R_{WORK} at 0.9-1.1Å

0.10 - 0.12:	68
0.12 - 0.14:	94
0.14 - 0.16:	73
0.16 - 0.18:	17 <<<
0.18 - 0.20:	12
0.20 - 0.21:	3
0.21 - 0.23:	5
0.23 - 0.25:	0
0.25 - 0.27:	0
0.27 - 0.29:	2

R_{FREE} at 0.9-1.1Å

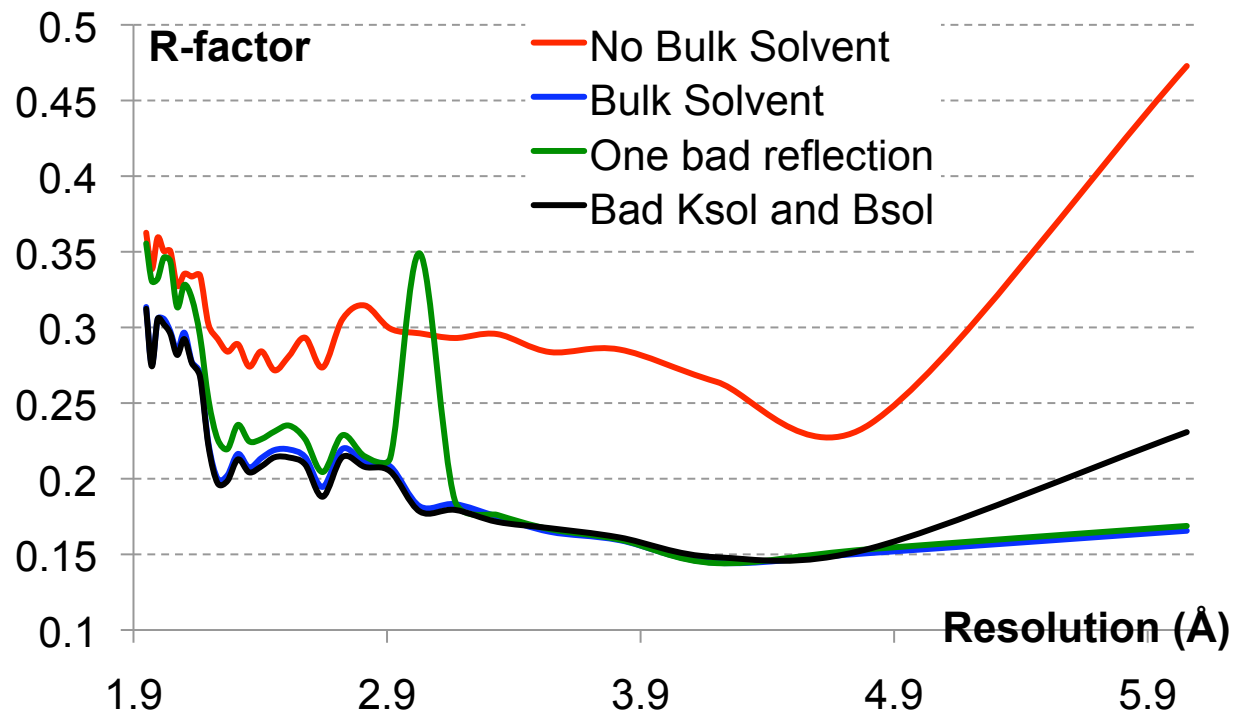
0.11 - 0.13:	16
0.13 - 0.15:	56
0.15 - 0.17:	97
0.17 - 0.18:	69
0.18 - 0.20:	14
0.20 - 0.22:	12
0.22 - 0.24:	3 <<<
0.24 - 0.26:	4
0.26 - 0.28:	1
0.28 - 0.30:	2

$R_{\text{FREE}} - R_{\text{WORK}}$ at 0.9-1.1Å

0.00 - 0.01:	8
0.01 - 0.01:	22
0.01 - 0.02:	56
0.02 - 0.03:	62
0.03 - 0.03:	58
0.03 - 0.04:	29
0.04 - 0.04:	14
0.04 - 0.05:	10 <<<
0.05 - 0.06:	6
0.06 - 0.06:	9

- **Answer:** the R-factors are not good, the structure needs some more work.

R-factor in resolution



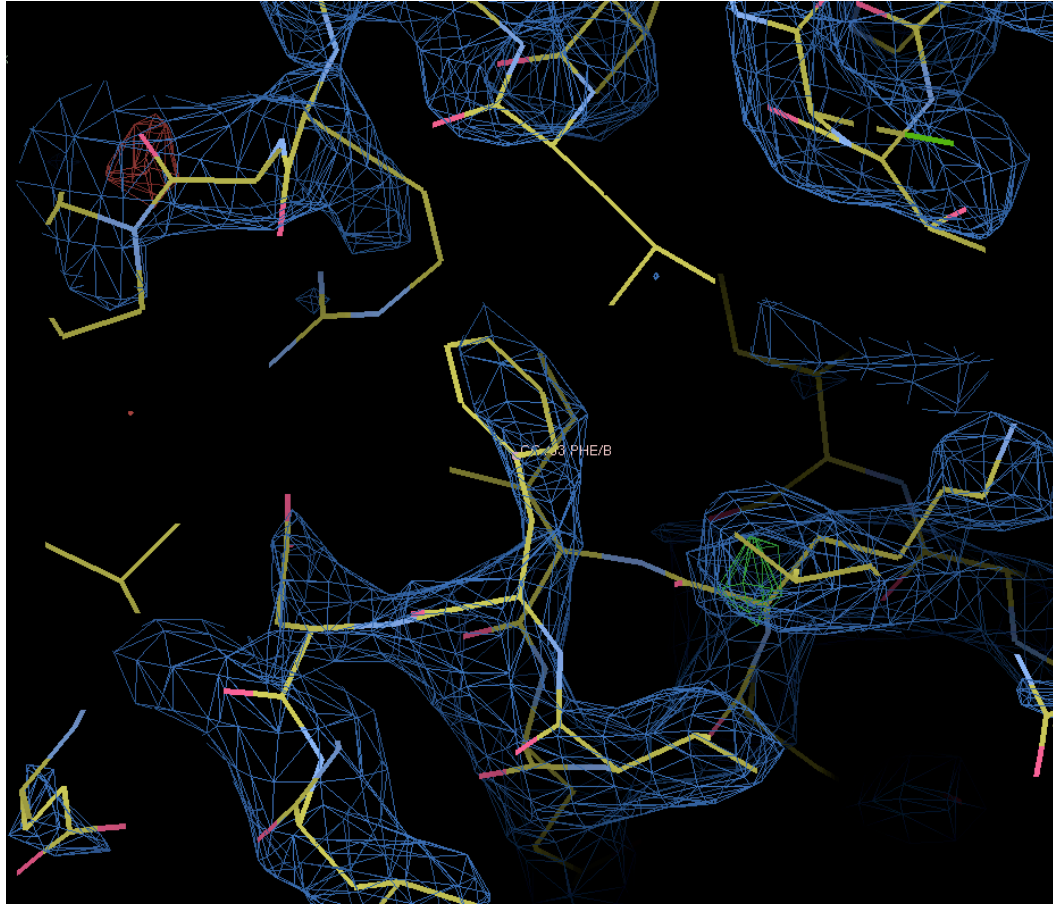
- Looking at R-factor in resolution bins helps to identify:
 - Poor or absence of bulk-solvent modeling (red and black lines at $>4.9\text{\AA}$)
 - Systematic problems with certain reflections (all lines at high resolution)
 - Artifacts (spike at around 3\AA resolution caused by nonsensical reflection amplitude value, green line):

```
INDE      3      5  -42 IOBS= 99999.999 SIGIOBS=      0.000
```

- Typically, one should expect an almost horizontal straight line, with some increase at high and low resolution ends

Good R-factors – bad map: twinning and free-R flags

- Data resolution: 2.8Å, $R_{\text{WORK}}=23.4\%$, $R_{\text{FREE}}=29.4\%$, poor map:



- Twinning was not accounted for when creating free-R set: R_{FREE} is biased
- After re-creating free-R set using lattice symmetry information and repeating refinement: $R_{\text{WORK}}=23.4\%$, $R_{\text{FREE}}=33.6\%$

Model, data and model-to-data fit quality indicators

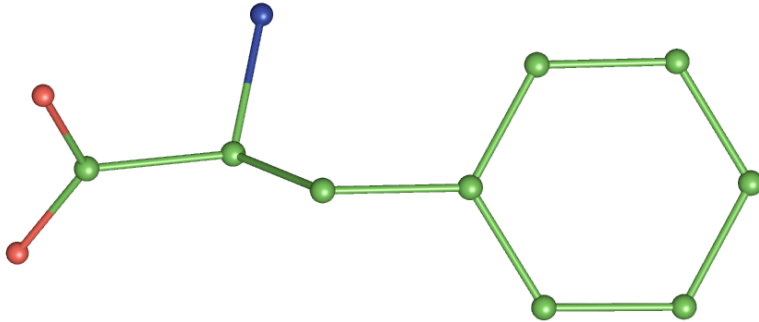
▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- **Geometry (stereochemistry):**
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- **Geometry and environment (rotamers, etc, main- side-chain conformations)**
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

Geometry: global figures



- *A priori* chemical knowledge is introduced (restraints) to keep the model chemically correct while fitting it to the experimental :

$$E_{\text{RESTRAINTS}} = E_{\text{BOND}} + E_{\text{ANGLE}} + E_{\text{DIHEDRAL}} + E_{\text{PLANARITY}} + E_{\text{NONBONDED}} + \dots$$

- Typically only rmsd for bonds and angles are reported along with R_{WORK} and R_{FREE}
- Typical values (resolutions $\sim 1.5\text{-}2\text{\AA}$): rmsd(bonds) $\sim 0.02\text{\AA}$, rmsd(angles) $\sim 2^\circ$
- These values can be smaller at lower resolution ($\sim 2.5\text{-}3\text{\AA}$), approaching 0 at $\sim 3\text{\AA}$ and lower resolution, and they can be larger at higher resolution ($\sim 1.5\text{\AA}$ and higher).

Geometry – histograms (I)

- Resolution 3.3Å:

$$R_{\text{WORK}} = 19.6\% \quad R_{\text{FREE}} = 24.5\% \quad \text{bonds} = 0.03\text{\AA} \quad \text{angles} = 4.6^\circ$$

- R-factors are great, geometry is terrible

Histogram of deviations from ideal values

Bonds				Angles		
0.000 – 0.035:	2645			0.000 – 9.313:	4208	
0.035 – 0.070:	19			9.313 – 18.626:	9	
0.070 – 0.106:	13			18.626 – 27.939:	3	
0.106 – 0.141:	5			27.939 – 37.252:	4	
0.141 – 0.176:	3			37.252 – 46.565:	0	
0.176 – 0.211:	0			46.565 – 55.878:	0	
0.211 – 0.246:	0			55.878 – 65.191:	2	
0.246 – 0.281:	0			65.191 – 74.504:	1	
0.281 – 0.317:	2			74.504 – 83.817:	0	
0.317 – 0.352:	18			83.817 – 93.130:	8	

- Problem with a few atoms, while the rest of ok
 - Incorrect ligand geometry

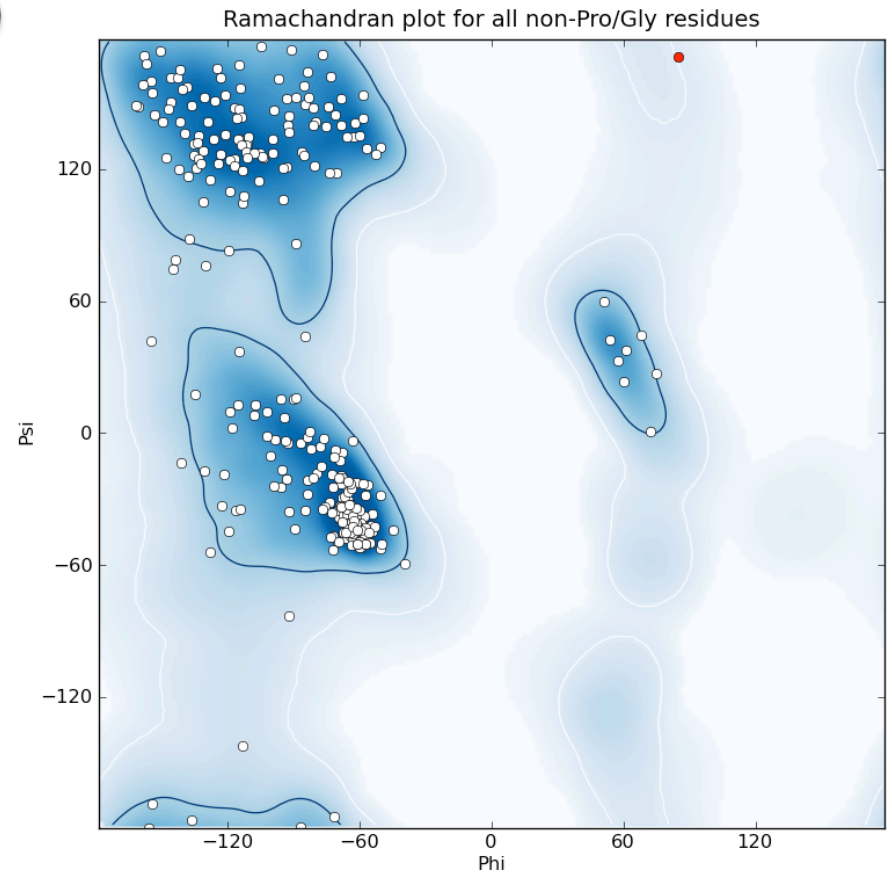
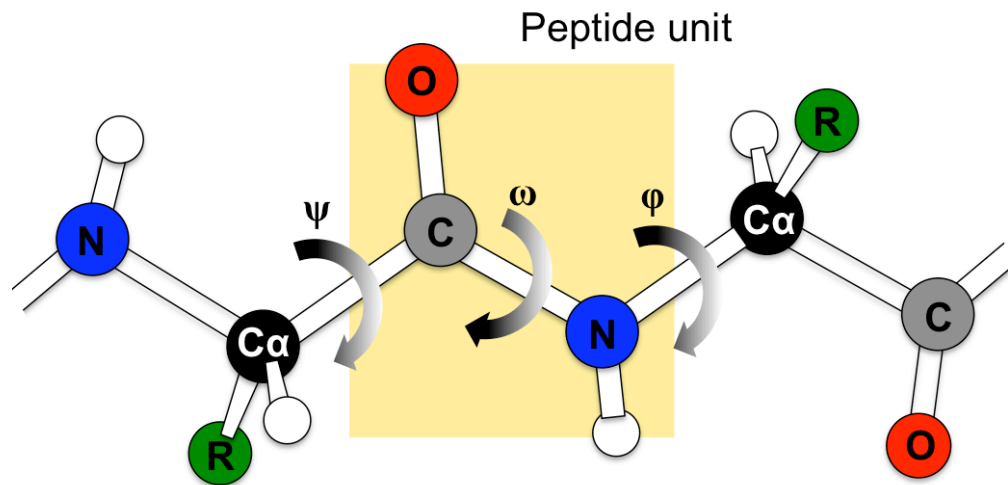
Geometry – histograms (II)

- After correcting the problem with the ligand: bonds = 0.01Å angles = 1.0°

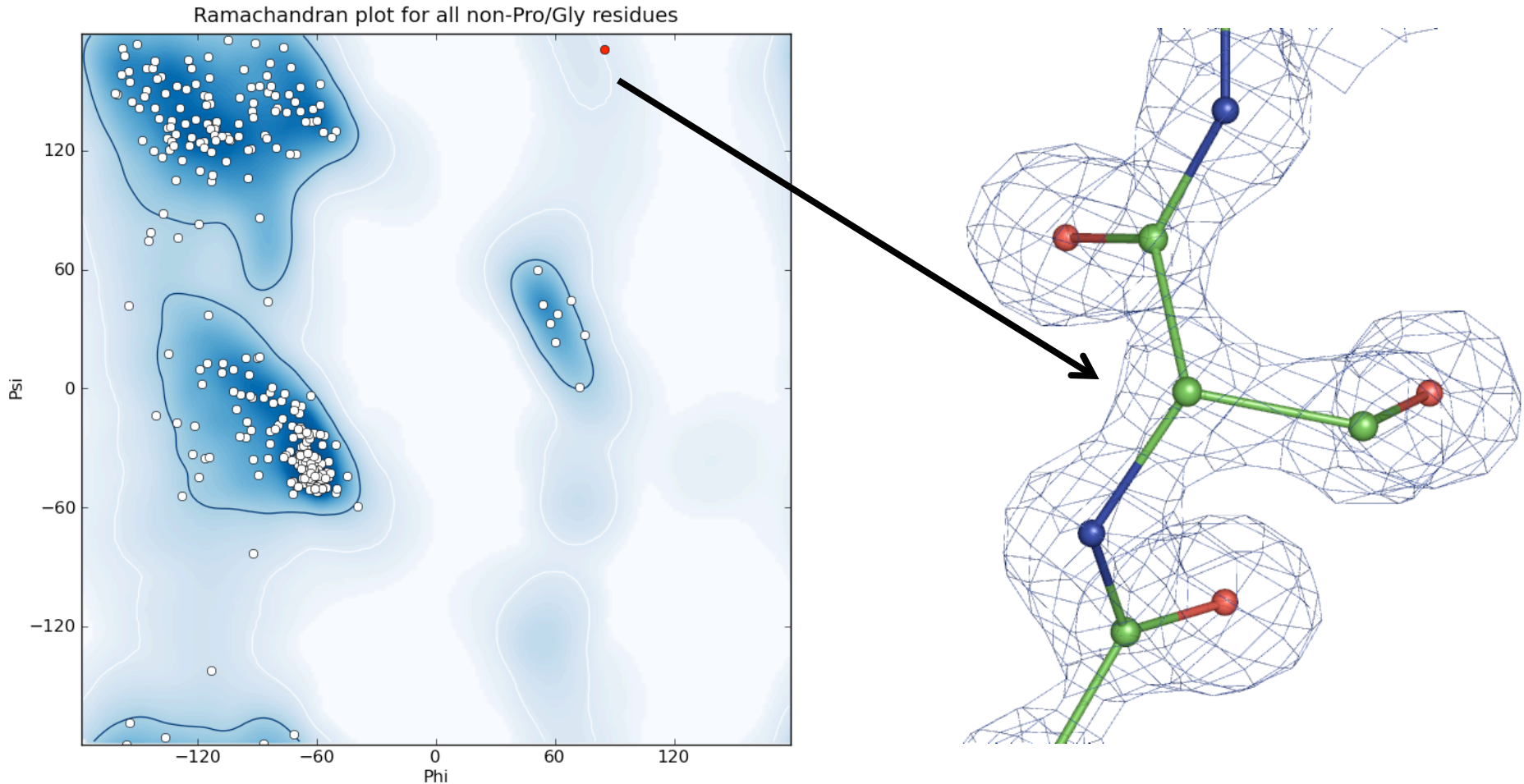
Histogram of deviations from ideal values

Bonds				Angles		
0.000 – 0.004:	1135		0.000 – 0.753:	2552		
0.004 – 0.008:	819		0.753 – 1.506:	1232		
0.008 – 0.012:	421		1.506 – 2.259:	266		
0.012 – 0.016:	179		2.259 – 3.012:	70		
0.016 – 0.020:	69		3.012 – 3.765:	28		
0.020 – 0.024:	35		3.765 – 4.518:	16		
0.024 – 0.028:	14		4.518 – 5.271:	8		
0.028 – 0.032:	5		5.271 – 6.024:	3		
0.032 – 0.036:	1		6.024 – 6.777:	3		
0.036 – 0.040:	1		6.777 – 7.530:	1		

Ramachandran plot



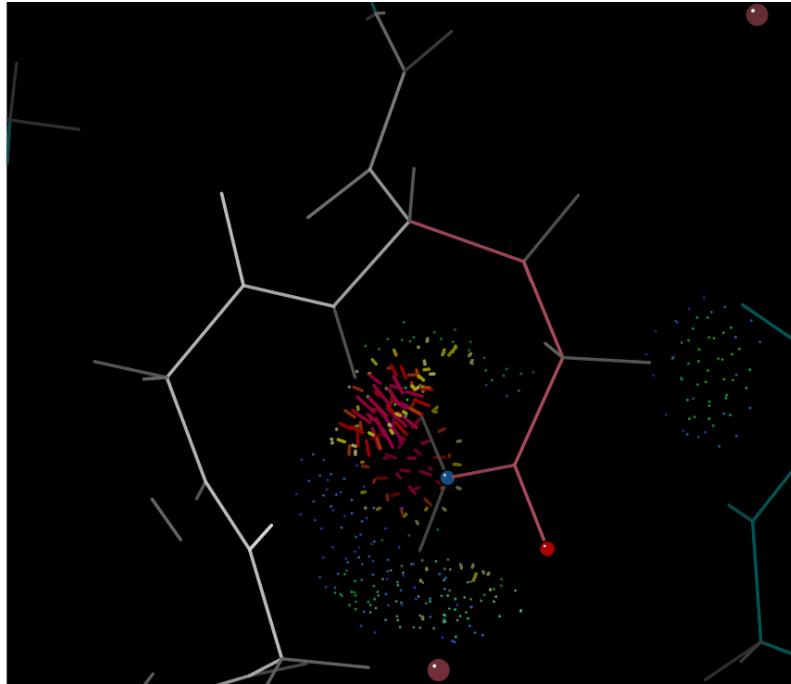
Ramachandran plot: outlier may be good



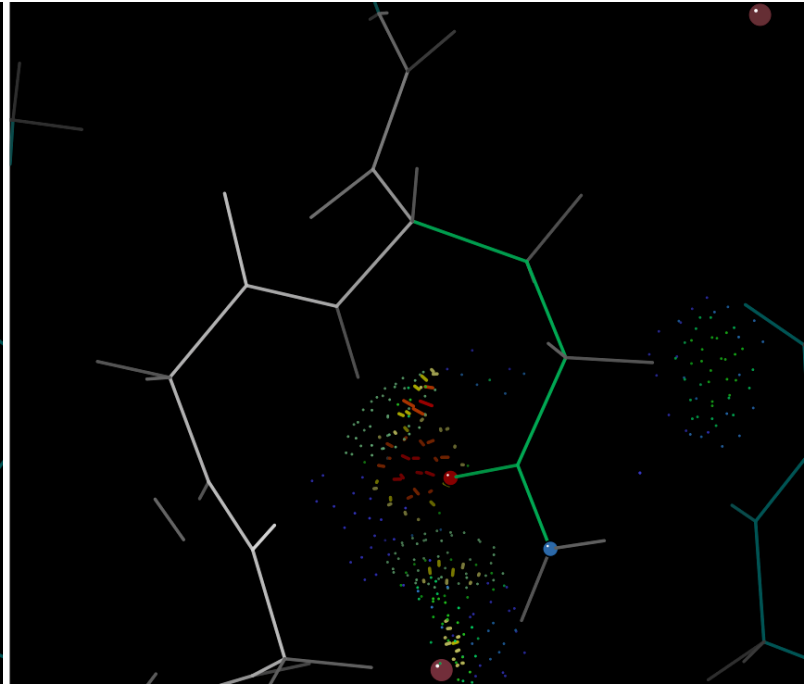
- Not everything flagged as outlier is actually wrong
 - Check the map
 - Make sure the map is not biased by the model
- Each outlier has to be explained

Steric clashes

Bad



Good

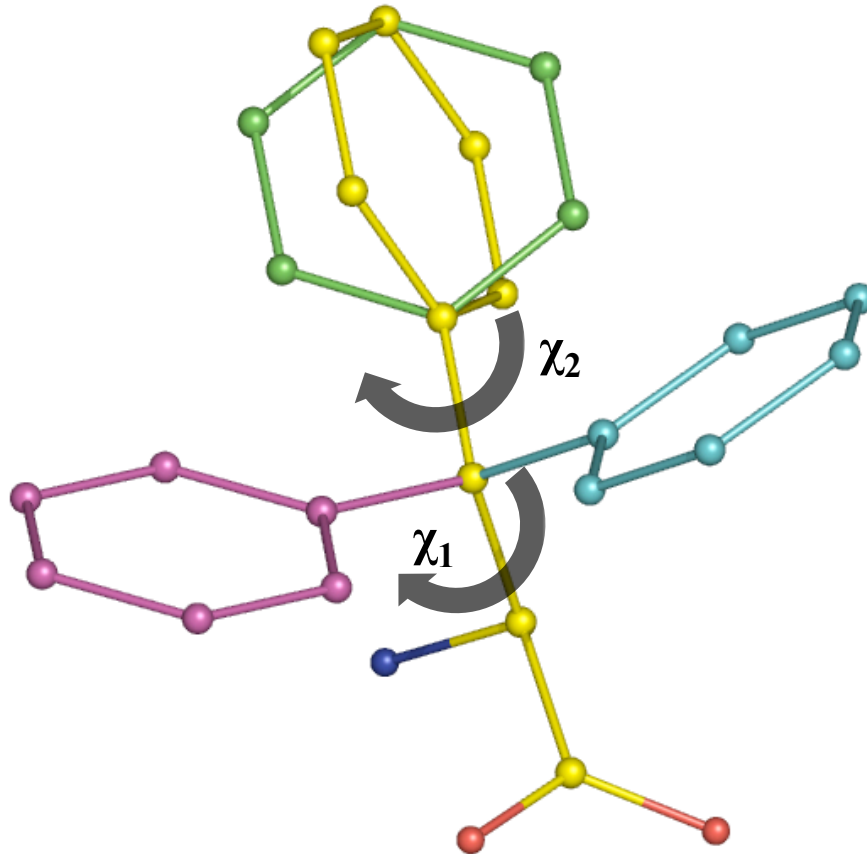


- **Overall clash score (number of bad overlaps per 1000 atoms)**
- **A clash: disallowed atom pair overlap ≥ 0.4 Å**

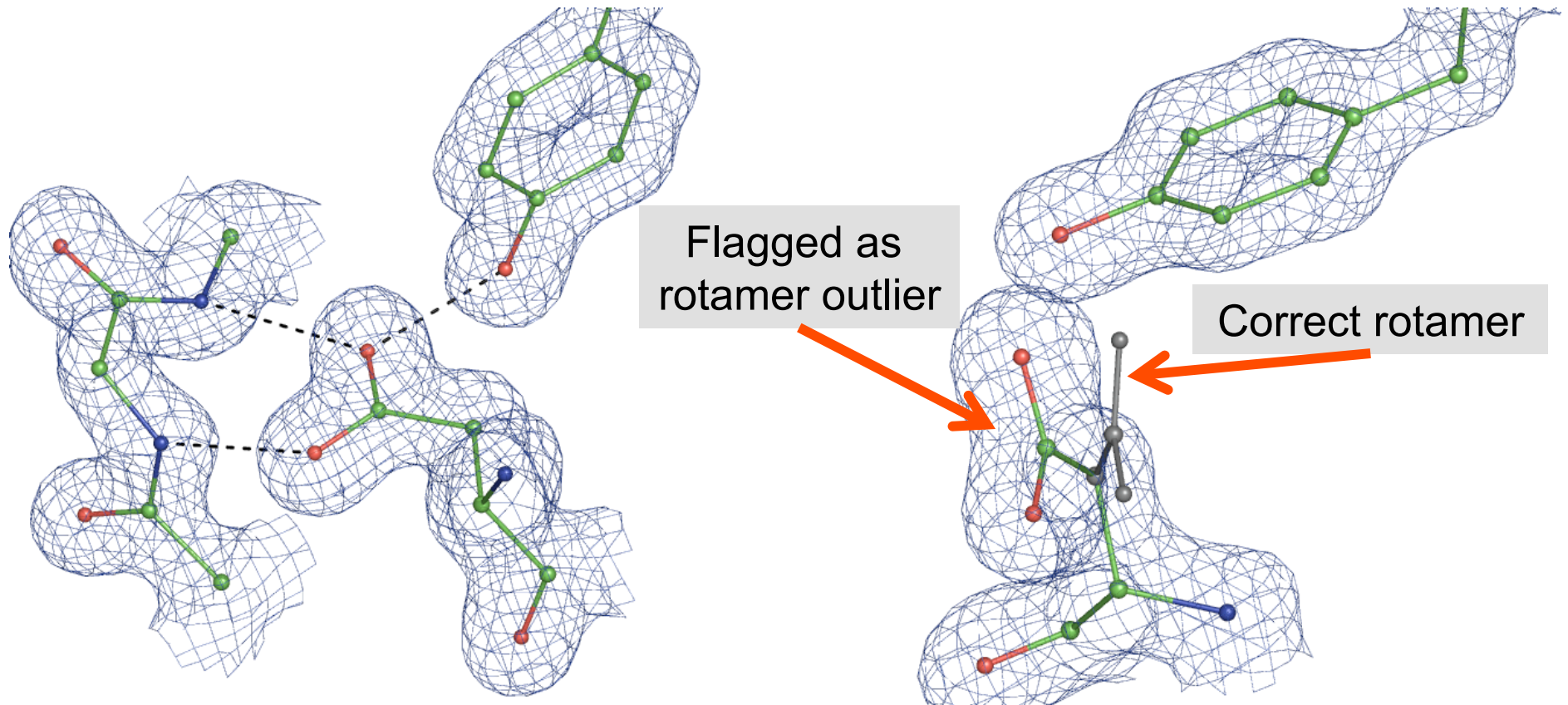
MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Davis et al, Nucleic Acids Research, 2007, Vol. 35

Rotamers

Rotamers: a set of conformers arising from restricted rotation about one single bond



Rotamers: outlier may be good



- Not everything flagged as outlier is actually wrong
 - Check the map
 - Make sure the map is not biased by the model
- Each outlier has to be explained

Model, data and model-to-data fit quality indicators

▪ Global:

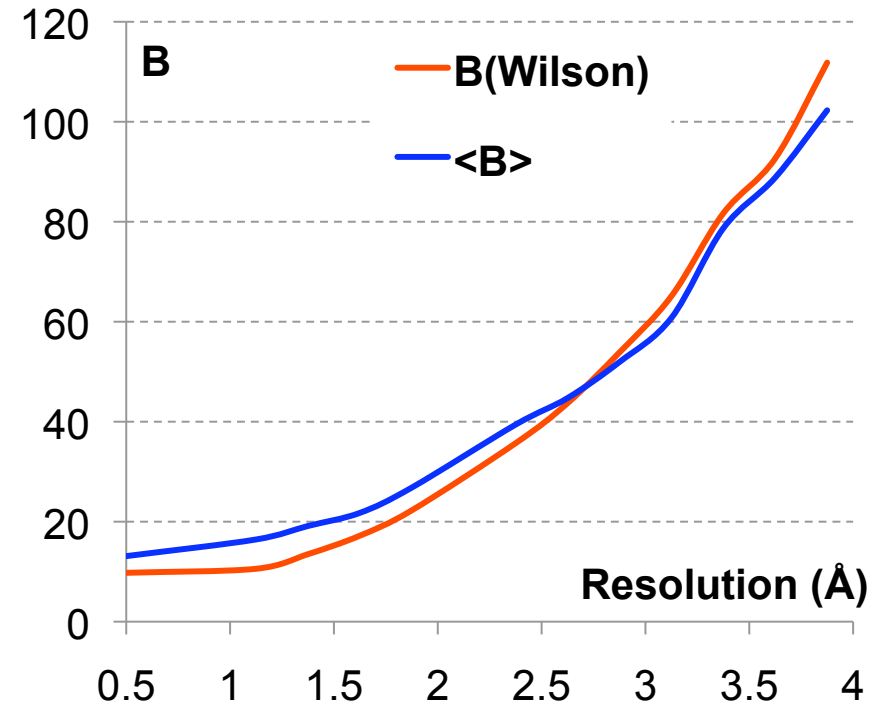
- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average *B*-factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (*B*-factors and their variations, occupancies).

Average B and Wilson B

Resolution	B (Wilson)	$\langle B \rangle$	Models
0.00-1.00	9.77	13.11	94
1.00-1.25	10.58	16.44	401
1.25-1.50	13.50	19.14	1050
1.50-1.75	17.20	21.76	3600
1.75-2.00	22.27	26.82	5510
2.25-2.50	35.70	39.42	3385
2.50-2.75	43.71	44.73	2844
2.75-3.00	53.86	51.94	1628
3.00-3.25	65.11	60.76	780
3.25-3.50	81.69	78.70	165
3.50-3.75	92.67	88.84	100
3.75-4.00	111.83	102.29	30



- Higher resolution – smaller B-factor
- Mean B does not differ too much from Wilson B
 - Wilson B is just an estimation (under some pretty unrealistic assumptions)
- There can be outliers

Model, data and model-to-data fit quality indicators

▪ Global:

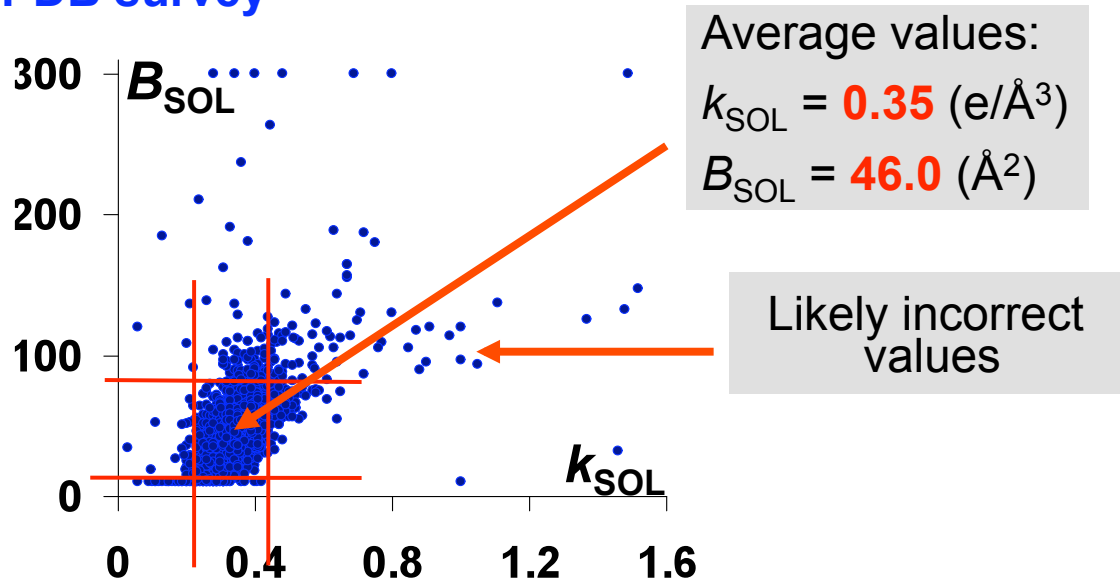
- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

Bulk-solvent parameters

PDB survey



- Wildly different k_{SOL} and/or B_{SOL} from average may indicate problem with the atomic or bulk-solvent model.
- B_{SOL} tends to be (too) large for incomplete typically low resolution structures.
- $k_{\text{SOL}}=0$ (no bulk-solvent) indicates either absence of low resolution data or severe problem with it.

Model, data and model-to-data fit quality indicators

▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- **Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions**
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

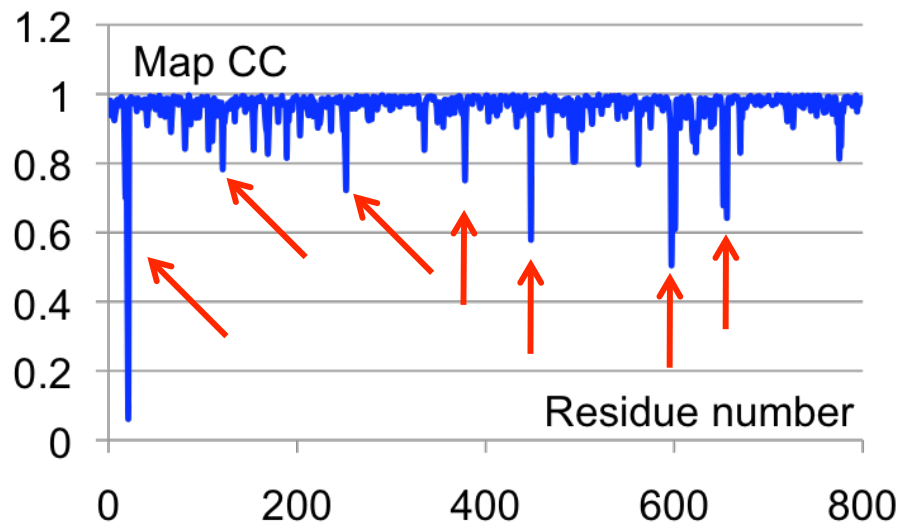
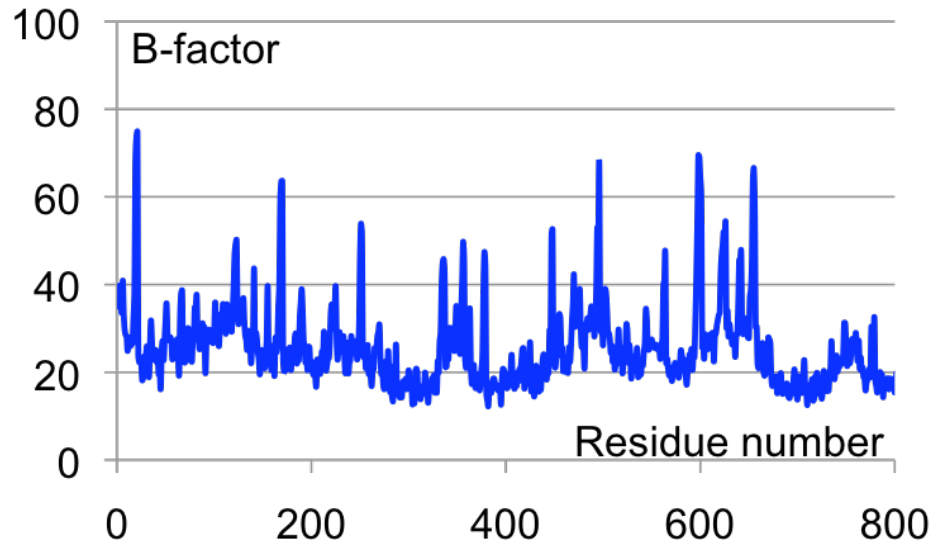
Real-space

$$CC = \frac{\sum_{\text{grid points}} |\rho_{\text{OBS}} - \langle \rho_{\text{OBS}} \rangle| \sum_{\text{grid points}} |\rho_{\text{CALC}} - \langle \rho_{\text{CALC}} \rangle|}{\left(\sum_{\text{grid points}} |\rho_{\text{OBS}} - \langle \rho_{\text{OBS}} \rangle|^2 \sum_{\text{grid points}} |\rho_{\text{CALC}} - \langle \rho_{\text{CALC}} \rangle|^2 \right)^{1/2}}$$

- Scale independent
- Can be computed for the whole structure (not really interesting – you already have R-factor) or locally (most interesting; typically computed per residue)
- Values greater than ~0.8 indicate good correlation
- May give high correlation for weak densities
- Map CC is correlated with B-factor: poorly defined regions typically have low map CC and high B-factors

Real-space

- In practice it is helpful to look at {B, map CC, 2mFo-DFc, mFo-DFc}

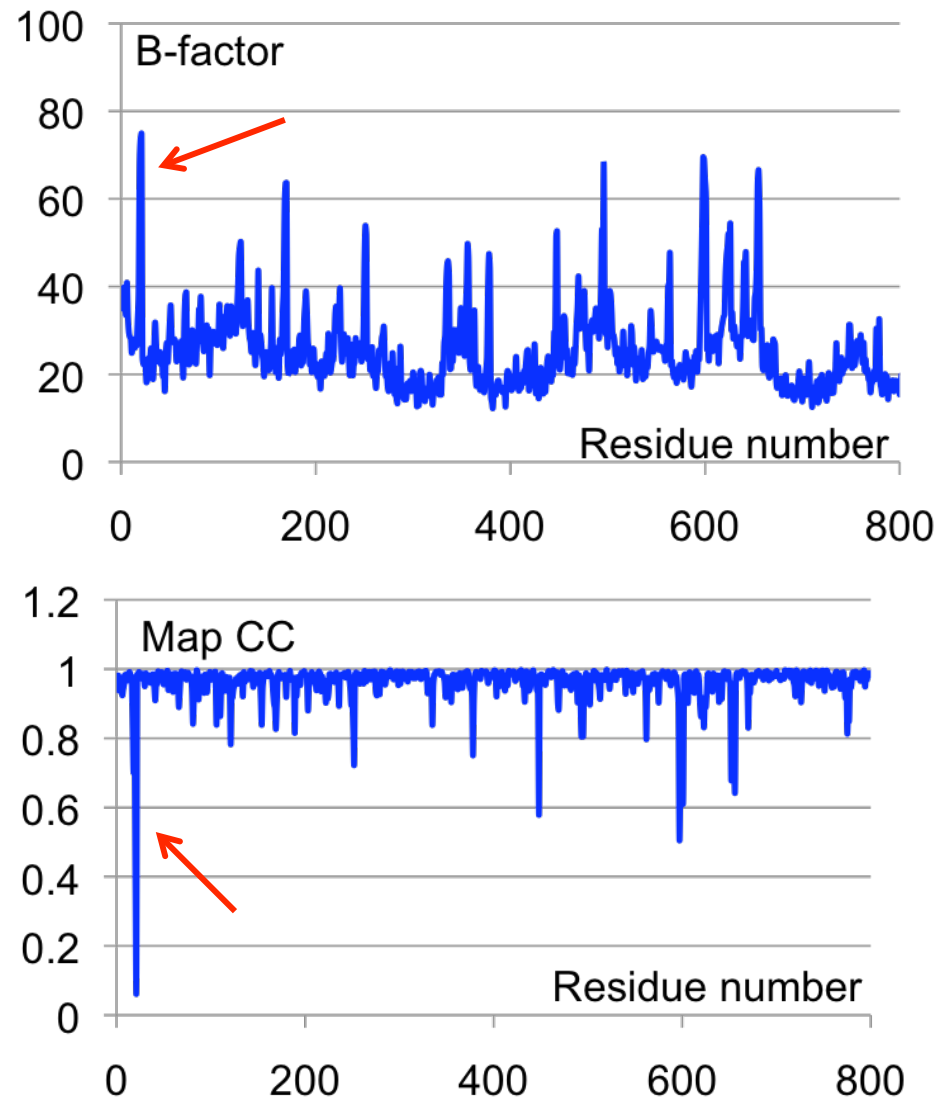


← Indicates problem places

Real-space

- In practice it is helpful to look at {B, map CC, 2mFo-DFc, mFo-DFc}

No	B	CC	2mFo-DFc	mFo-DFc
12	27.95	0.9677	1.64	0.74
13	26.74	0.9894	3.57	0.03
14	25.98	0.9909	3.41	0.14
15	26.55	0.9795	3.36	0.39
16	26.56	0.9793	3.21	2.06
17	32.46	0.8418	2.22	2.20
18	39.13	0.7003	1.36	-0.23
19	68.25	0.8350	-0.10	-5.57
20	73.73	0.3791	-0.23	-3.65
21	74.83	-0.0825	-0.41	-3.01
22	23.87	0.9831	4.00	0.35
23	22.26	0.9874	4.07	0.16
24	23.35	0.9910	2.87	0.62



Model, data and model-to-data fit quality indicators

▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- **Sequence register (incorrect residue identity)**
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

Sequence register

- Check actual sequence with one derived from PDB file with final model:
 - Extract sequence from PDB file:

```
phenix.print_sequence model.pdb > model.seq
```
 - Align actual sequence with `model.seq`
- Example of a problem:

```
MASTER  GFVDLTLHDQVSMEHPVKLLFGKCEGMVEIVYTFLSSTLKSLE
Chain A  GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTF-----KSLE
Chain B  GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTFVVSSTLKSLE
Chain C  GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFVVSSTLKSLE
Chain D  GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFLSSTLKSLE
***** ***** ***** ********** *****
```

Model, data and model-to-data fit quality indicators

▪ Global:

- R-factor (R_{WORK} and R_{FREE})
- Geometry (stereochemistry):
 - Deviation from ideal (rmsds): bond, angles, planarities,...
 - Non-bonded clashes, Molprobit clashscores
 - Ramachandran plot statistics
- Average B -factor and Wilson B
- Comparison statistics (to similar structures in the database)
- Bulk-solvent parameters (k_{SOL} and B_{SOL})

▪ Local:

- Geometry and environment (rotamers, etc, main- side-chain conformations)
- Real-space: map correlation, values of $2m\text{Fo-DFc}$ and $m\text{Fo-DFc}$ at and around atomic positions
- Sequence register (incorrect residue identity)
- Naming for ligands
- Other parameters (B -factors and their variations, occupancies).

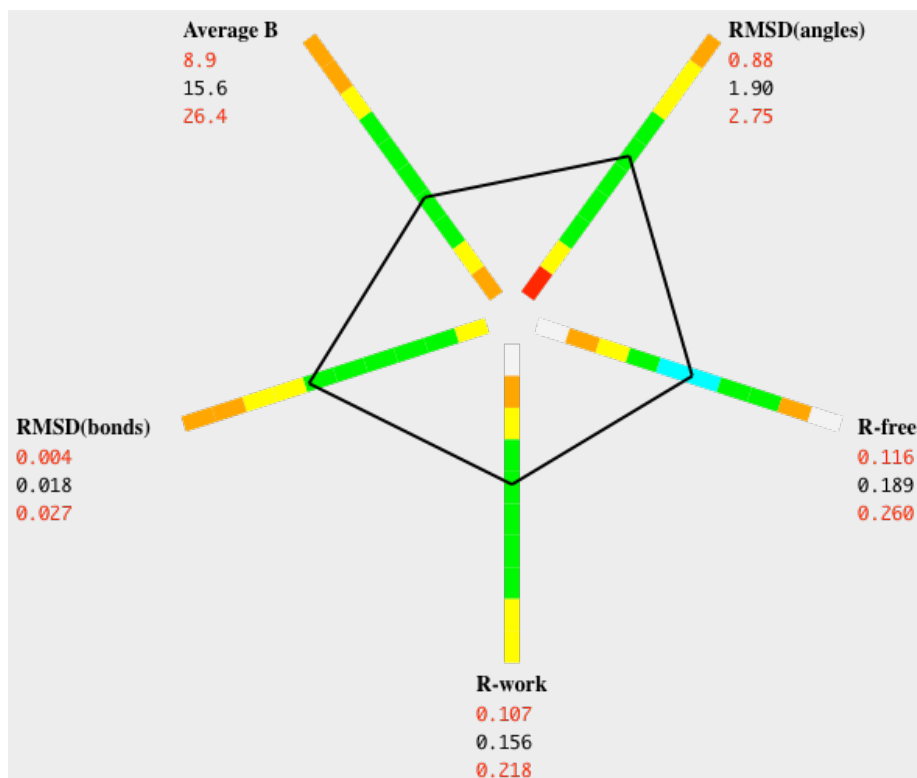
Model, data and model-to-data fit quality indicators

- **Comparative statistics:** typically global model quality figures are in agreement with corresponding values found in similar structures:
 - If it is not the case it does not necessarily mean the structure is wrong, but it is a good reason to stop and think
- **How it is done?**
 - Select similar (refined at similar resolution, for example) structures in the database (PDB)
 - Obtain the distribution of a parameter in question
 - See where the corresponding parameter of your structure is w.r.t. the distribution

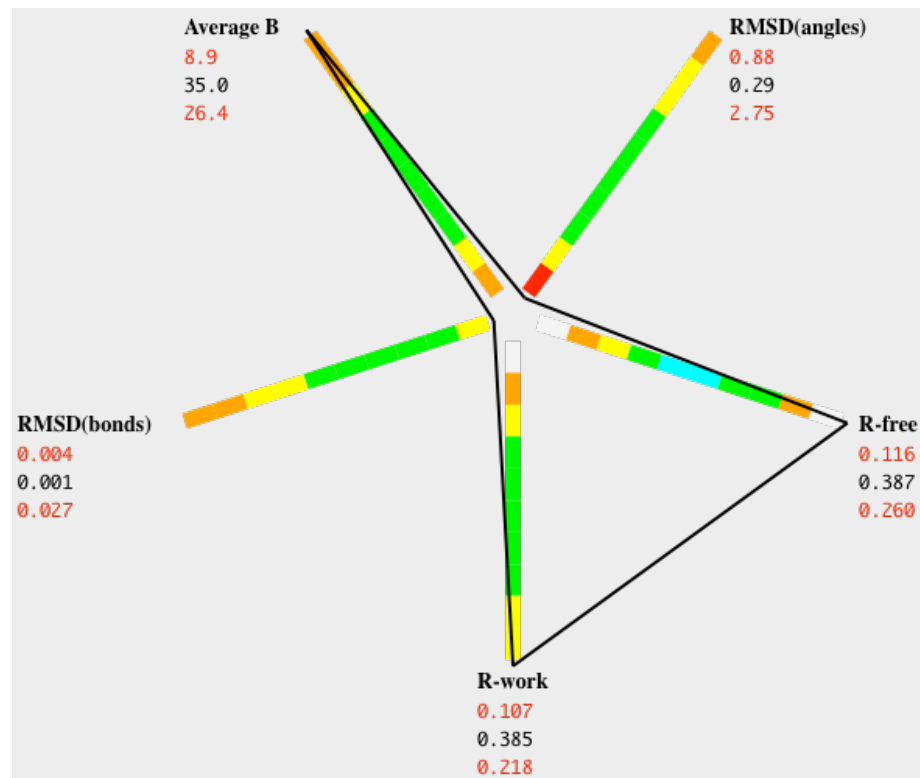
R_{WORK}	Number of structures
0.10 - 0.12:	68
0.12 - 0.14:	94
0.14 - 0.16:	73
0.16 - 0.18:	17 <<< your structure
0.18 - 0.20:	12
0.20 - 0.21:	3
0.21 - 0.23:	5
0.23 - 0.25:	0
0.25 - 0.27:	0
0.27 - 0.29:	2

New tool in PHENIX: POLYGON

Likely good model



This model needs some attention



Crystallographic model quality at a glance.

L. Urzhumtseva, P.V. Afonine, P.D. Adams, A. Urzhumtsev. *Acta Cryst.* **D65**, 297-300 (2009)

Example of under-refined model

■ PDB: **1eic**

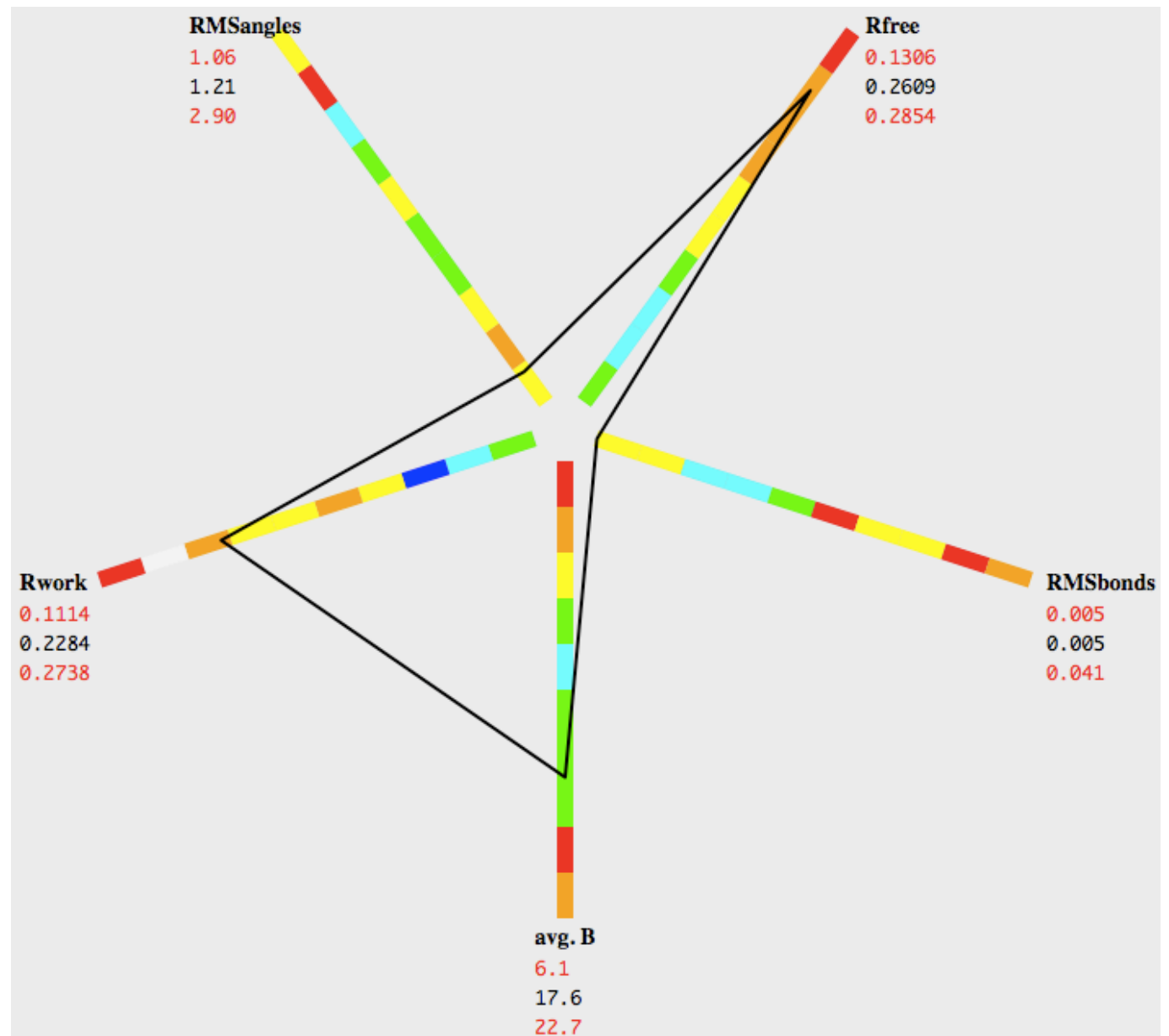
Resolution: 1.4Å

Deposition year:
2000

PUBLISHED:

Rwork = 20%

Rfree = 25%



Under-refined models or why automation is important

- Structure from PDB: **1eic** (resolution = 1.4Å; deposition year: 2000)

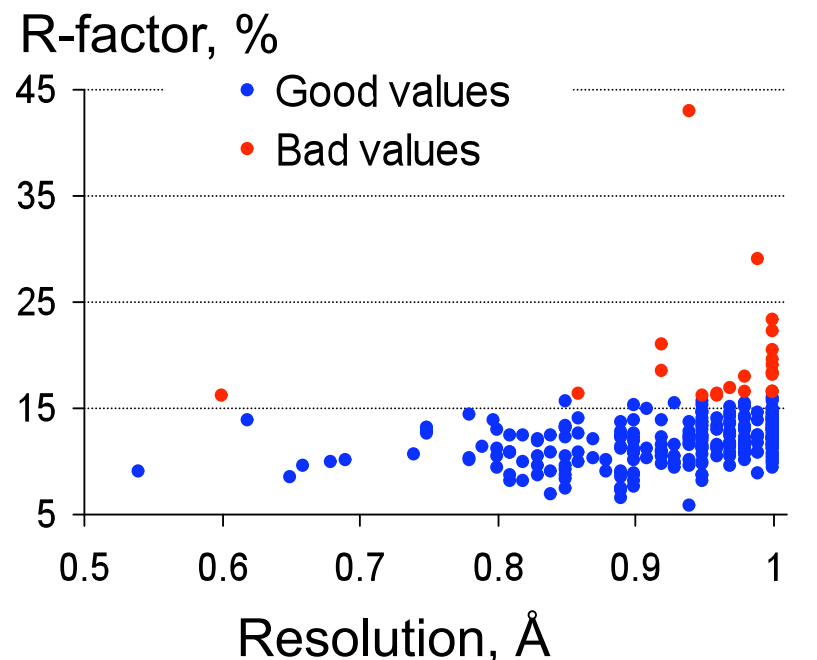
PUBLISHED: Rwork = 20% Rfree = 25%

- Clear problems:
 - No 'riding' H atoms;
 - All atoms are isotropic;
- Potential problems
 - Inoptimal weights, refinement is not converged, incomplete solvent model
- Fixing the model with PHENIX:
 - Add and refine H as riding model
 - Update ordered solvent
 - Refine atoms as anisotropic (except H and water)
 - Optimize X-ray/Restraints weights

FINAL MODEL: Rwork = 14% Rfree = 17%

- ✓ All this could be done by the software automatically, preventing deposition of under-refined models into PDB

R-factors (all models in PDB at resolution < 1 Å)



PDB code (year)	R-work, %	
	published	phenix.refine
2ppn (2007)	20.9	11.7
1g2y (2000)	19.5	12.3
1zlb (2005)	16.8	12.0
2g6f (2006)	18.4	12.9
2elg (2007)	23.2	13.0
1aho (1997)	16.3	9.6
1zf5 (2005)	29.0	16.9

- There are ~25 models out of 324 that have suspiciously high or very high R-factors.
 - For most of them the R-factors can be decreased to typical for this resolution values (~10-15%) in one phenix.refine run.
- Automated software with integrated validation would immediately flag these models as suspicious.

phenix.model_vs_data

- Database used by POLYGON is created using **phenix.model_vs_data** tool.
- **phenix.model_vs_data** is a tool that reports a page long summary about data, model and model-to-data fit:
 - Easy to run: **phenix.model_vs_data model.pdb data.hkl**
 - Any data type: X-ray or neutron
 - Most of reflection data file formats (CNS, SHELX, MTZ, ...)
 - Automatic twinning detection
 - Unknown ligands are handled automatically
 - Input model can be spread across multiple file (case of huge structures)
 - Regularly exercised by running through the whole PDB
 - Refmac style files with separated TLS (in REMARK 3) and residual B-factors are ok

Running `phenix.model_vs_data` for whole PDB

- Histogram of differences between reported (in PDB file header) and re-computed with `phenix.model_vs_data` *R*-work :

	-23.34	-	-17.95	3
	-17.95	-	-12.56	5
Worse than published:	-12.56	-	-7.17	47
	-7.17	-	-1.79	1106
	-1.79	-	3.59	30990
	3.59	-	8.98	1215
	8.98	-	14.36	242
Better than published:	14.36	-	19.75	96
	19.75	-	25.14	58
	25.14	-	30.53	18

Why reported R-factors may not match the re-computed ones?

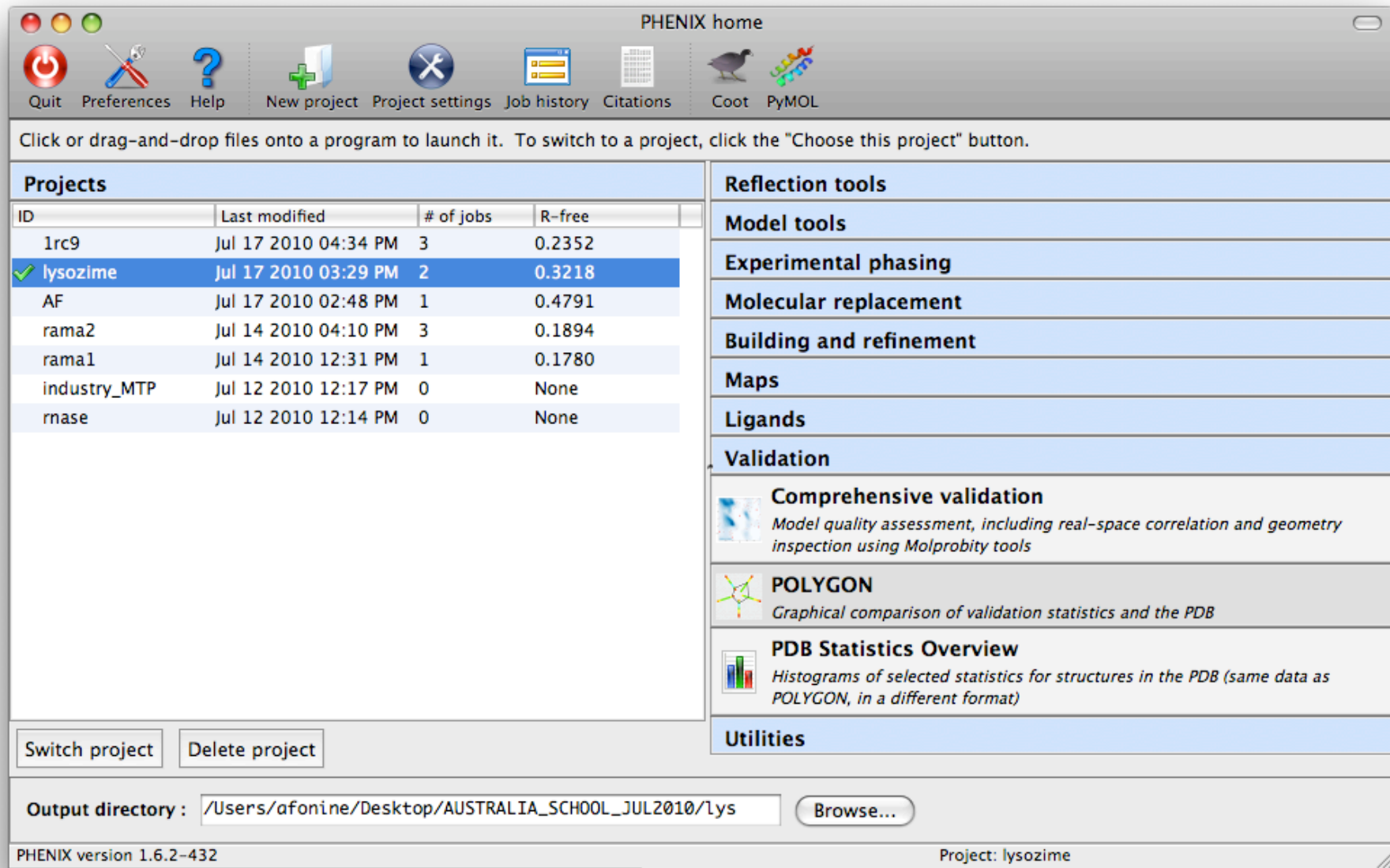
- Removing hydrogen atoms (up to 2.0%)
- Missing anisotropic ADPs (~5%)
- Nonsensical ADPs (~2...10%)
- Overlooked twinning (~5...20%)
- Missing water (~2...5%)
- Results of IAS or multipolar refinement are not preserved (~1%)
- Variations in bulk solvent model and anisotropic scaling parameters (~1...5%)
- Occupancies of atoms at special positions
- Test flags are missing
- Use f' and f'' in structure factor calculations for anomalous scatterers
- Removing Fobs outliers
- Incorrect structure factor data deposited (or correct data but incorrectly labeled).
- Corrupted TLS records (up to 10%, ~700 entries as of spring 2009).
- Different scattering tables? No!
- FFT vs direct summation? No!

PHENIX tools for model validation

- **Comprehensive validation** option available from PHENIX GUI:
 - MolProbity scores;
 - Real-space correlation (map CC), 2mFo-DFc and mFo-DFc listed for each atom or residue;
 - Basic geometry statistics (rmsd and max deviation for bonds, angles, ...)
 - phenix.model_vs_data report;
 - POLYGON.
- phenix.refine .log file contains lots of information.
- Tools to create various maps (iterative build omit maps, SA omit maps, Average kick maps, i^*mFo-j^*DFc maps)...
- Getting uncertainties by building multiple models.

PHENIX tools for model validation

- **Comprehensive validation** option available from PHENIX GUI:



The screenshot shows the PHENIX GUI interface. At the top, there is a menu bar with icons for Quit, Preferences, Help, New project, Project settings, Job history, Citations, Coot, and PyMOL. Below the menu bar is a text instruction: "Click or drag-and-drop files onto a program to launch it. To switch to a project, click the 'Choose this project' button." The main area is divided into two panes. The left pane, titled "Projects", contains a table with the following data:

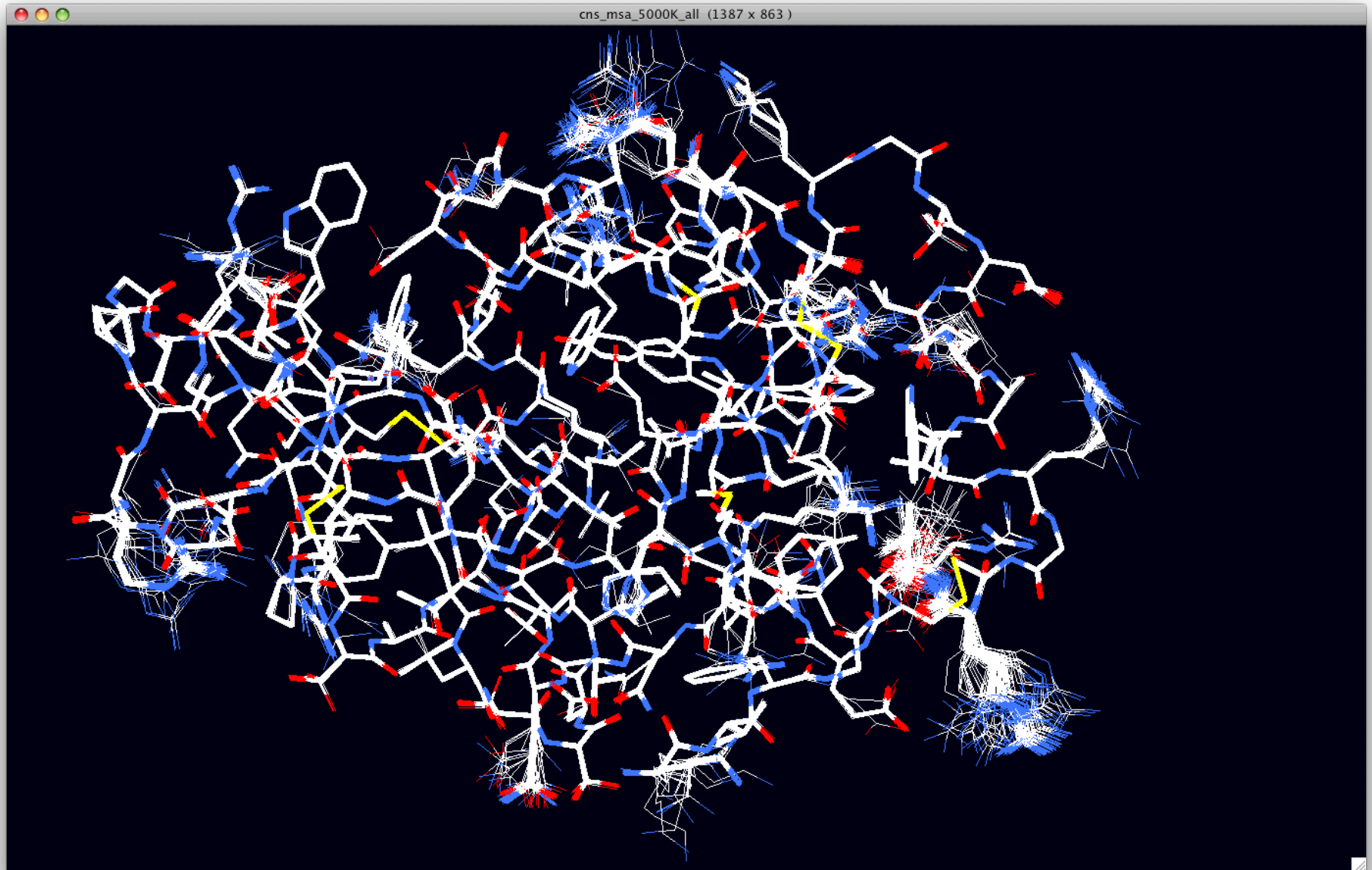
ID	Last modified	# of jobs	R-free
1rc9	Jul 17 2010 04:34 PM	3	0.2352
✓ lysozyme	Jul 17 2010 03:29 PM	2	0.3218
AF	Jul 17 2010 02:48 PM	1	0.4791
rama2	Jul 14 2010 04:10 PM	3	0.1894
rama1	Jul 14 2010 12:31 PM	1	0.1780
industry_MTP	Jul 12 2010 12:17 PM	0	None
rnase	Jul 12 2010 12:14 PM	0	None

Below the table are "Switch project" and "Delete project" buttons. The right pane, titled "Validation", contains several tool options:

- Reflection tools**
- Model tools**
- Experimental phasing**
- Molecular replacement**
- Building and refinement**
- Maps**
- Ligands**
- Validation**
 - Comprehensive validation**: Model quality assessment, including real-space correlation and geometry inspection using Molprobity tools
 - POLYGON**: Graphical comparison of validation statistics and the PDB
 - PDB Statistics Overview**: Histograms of selected statistics for structures in the PDB (same data as POLYGON, in a different format)
- Utilities**

At the bottom, there is an "Output directory" field with the path "/Users/afonine/Desktop/AUSTRALIA_SCHOOL_JUL2010/lys" and a "Browse..." button. The status bar at the very bottom shows "PHENIX version 1.6.2-432" on the left and "Project: lysozyme" on the right.

Uncertainties



PDB deposition dos and don'ts (I)

- Do not change anything in PDB file with refined model.
- If you did change something, then re-run the refinement to update statistics.
- Deposit the data: the one used in refinement. If this data was modified (resolution truncated, corrected for anisotropy, etc), then deposit the original data as well.
- Some people vote for depositing F_{calc} (F_{model}). Personally, I think this is not necessarily: if the data and PDB file are complete and accurate the statistics (and therefore F_{model}) should be reproducible.
- Once you have sent the data and model to PDB, they will come back to you with modified (reformatted) files for your approval. Check it carefully. Make sure you can reproduce the statistics (R-factors) using these files and make sure the PDB file header still contain the data and model stats that you originally submitted.
- Deposit free-R flags and phase information (HL coefficients, if available).
- If depositing multiple datasets indicate which one was used for obtaining the final structure.
- Once your files are publicly available at PDB site, download it and check.

PDB deposition dos and don'ts (II)

- Some programs and people tend to interpret unknown density using “dummy atoms”. In PDB files it typically looks like this:

ATOM	10	O	UNK	2	6.348	-11.323	10.667	1.00	8.06	X
ATOM	11	O	UNK	2	6.994	-12.600	10.740	1.00	7.16	X
ATOM	12	O	UNK	2	6.028	-13.737	10.607	1.00	6.58	X
ATOM	13	DUM	UNK	2	6.796	-15.043	10.583	1.00	8.28	
ATOM	14	DUM	UNK	2	5.099	-13.727	11.792	1.00	7.15	

- Do not deposit this in PDB, especially if chemical element type is undefined (rightmost column)

Conclusions

- The software should be as much automated as possible to minimize user errors.
- People should be skilled enough to solve structures: continuous education through workshops is important.
- Deposition tools should be smart to deal with broad variety of situations and not only with “standard” ones.
- No-one knows your structure better than you. Make sure this knowledge is correct and makes sense (use validation tools) and it is properly documented.