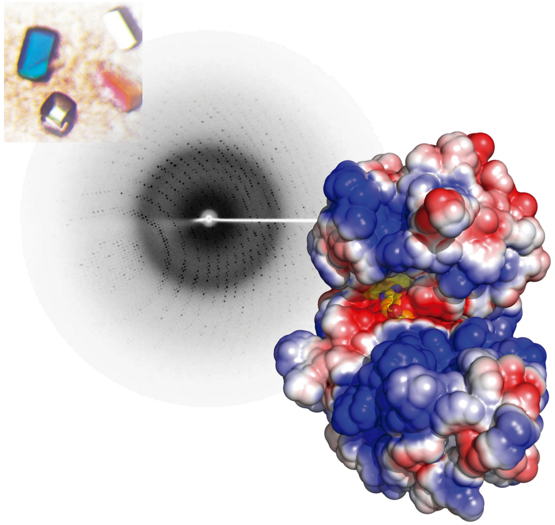
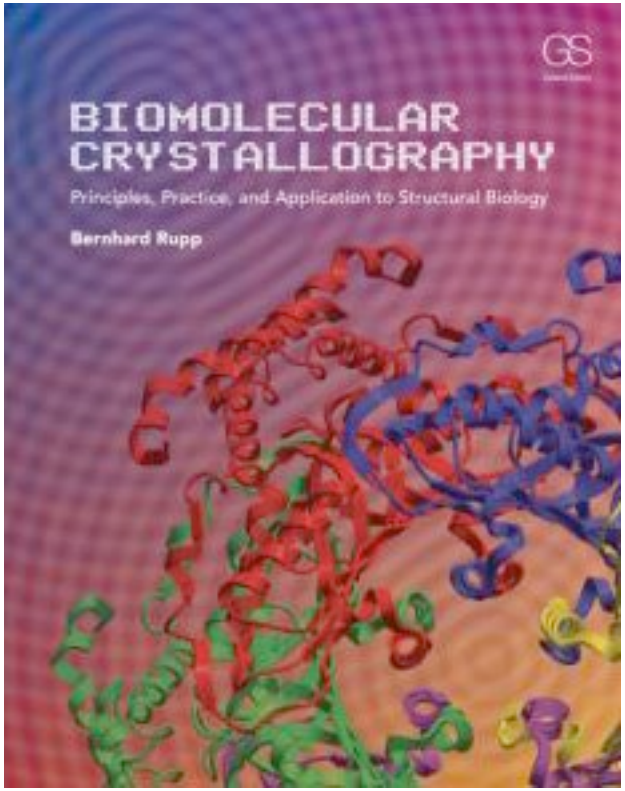
Unit of Protein Crystallography

Institut Pasteur
de Montevideo

Model quality:
concepts & statistics
(validation)



Macromolecular Crystallography School 2018
November 2018 - São Carlos, Brazil


GS
BIOMOLECULAR CRYSTALLOGRAPHY
Principles, Practice, and Application to Structural Biology
Bernhard Rupp

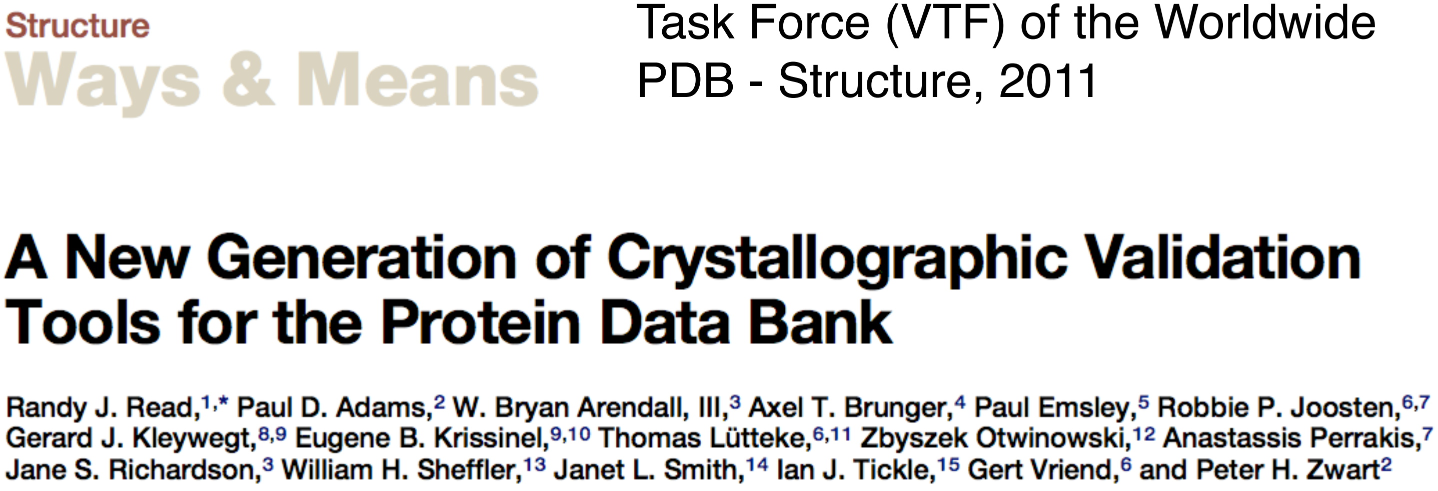
Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology

Bernhard Rupp

Tutorial by Gerard Kleywegt

http://xray.bmc.uu.se/embo2001/modval/

Conclusions of the X-ray Validation Task Force (VTF) of the Worldwide PDB - Structure, 2011



Structure

Ways & Means

A New Generation of Crystallographic Validation Tools for the Protein Data Bank

Randy J. Read,[1,*] Paul D. Adams,[2] W. Bryan Arendall, III,[3] Axel T. Brunger,[4] Paul Emsley,[5] Robbie P. Joosten,[6,7] Gerard J. Kleywegt,[8,9] Eugene B. Krissinel,[9,10] Thomas Lütteke,[6,11] Zbyszek Otwinowski,[12] Anastassis Perrakis,[7] Jane S. Richardson,[3] William H. Sheffler,[13] Janet L. Smith,[14] Ian J. Tickle,[15] Gert Vriend,[6] and Peter H. Zwart[2]

SUMMARY

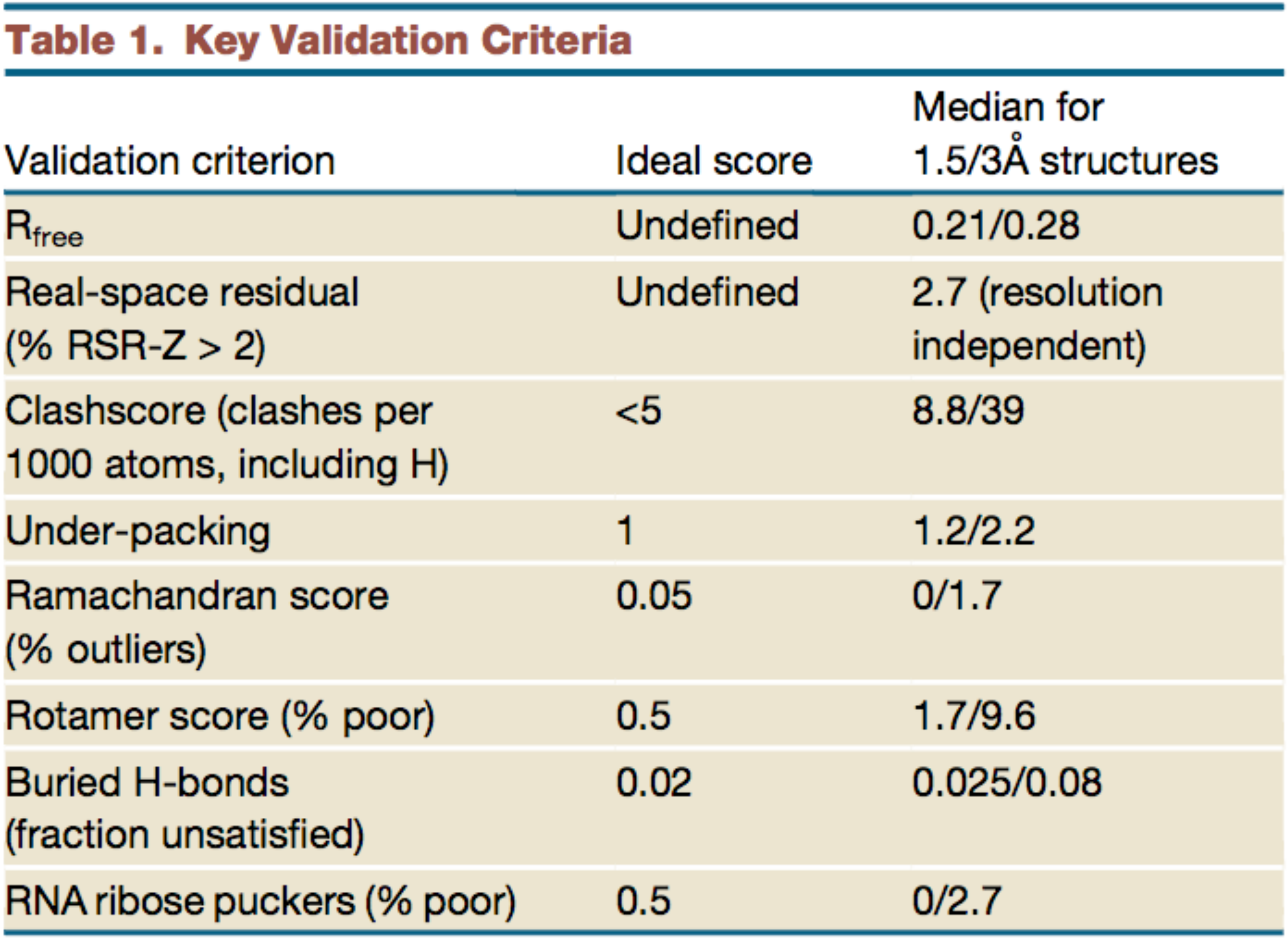


Table 1. Key Validation Criteria

Validation criterion	Ideal score	Median for 1.5/3Å structures
R_{free}	Undefined	0.21/0.28
Real-space residual (% RSR-Z > 2)	Undefined	2.7 (resolution independent)
Clashscore (clashes per 1000 atoms, including H)	<5	8.8/39
Under-packing	1	1.2/2.2
Ramachandran score (% outliers)	0.05	0/1.7
Rotamer score (% poor)	0.5	1.7/9.6
Buried H-bonds (fraction unsatisfied)	0.02	0.025/0.08
RNA ribose puckers (% poor)	0.5	0/2.7

Topics

1. What is validation, and what's validation in crystallography?

2. Overview of quality checks in PX : global vs local; the data, the model, the model AND data

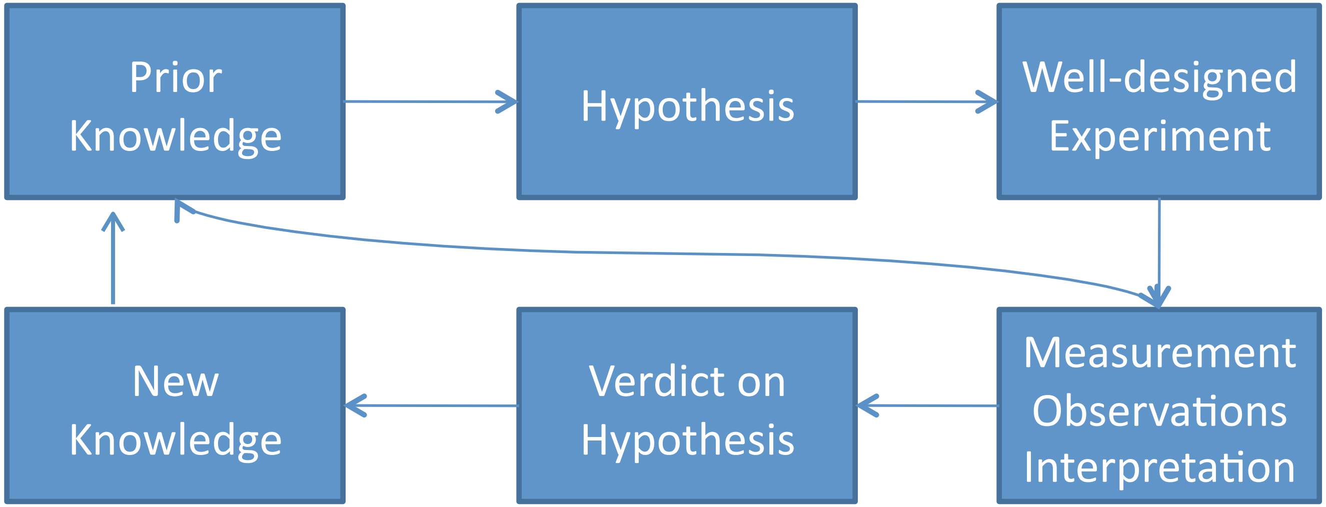
3. Data only (briefly; already introduced in data processing lectures/tutorials)

4. Model only : stereochemistry, dihedrals, packing

5. Model vs data : amount of data, R factors, map quality, model:map fit, crystal packing, B factors

Validation in crystallography : quality control

…within the general scientific scenario: hypothesis testing



Prior Knowledge

Hypothesis

Well-designed Experiment

Measurement Observations Interpretation

Verdict on Hypothesis

New Knowledge

- Prior knowledge aids (or somehow affects) interpretation.

- Measurements should conform to prior knowledge, or be strong and repeatable enough to refute it.

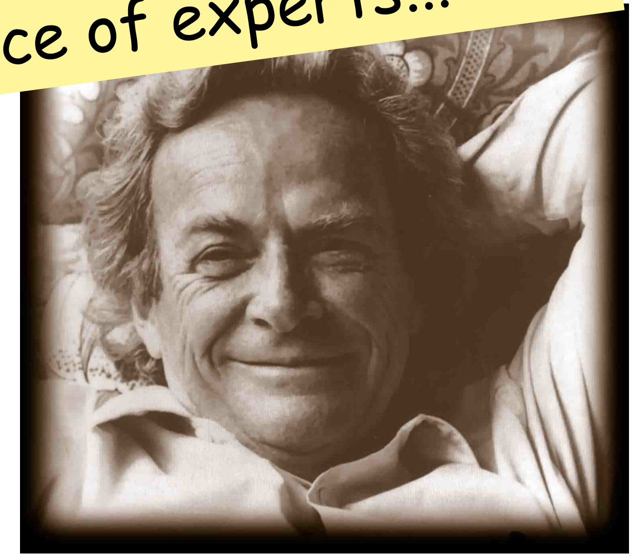
Model quality control

= Validation = establishing the truth or accuracy of

* Theory
* Hypothesis
* Model
* Claim ... etc

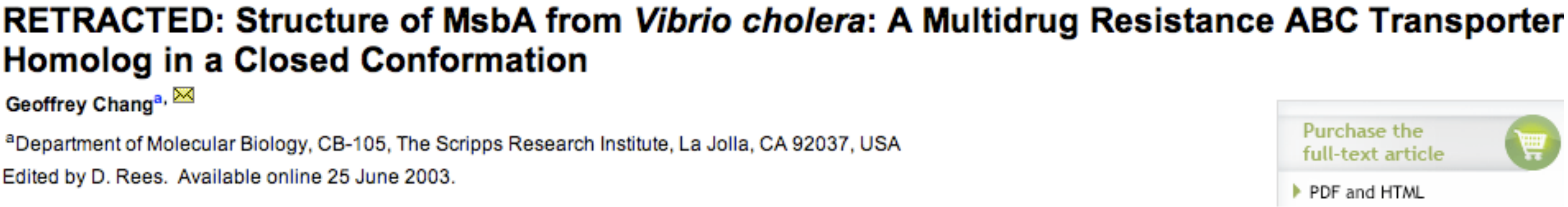
Science is the belief in the ignorance of experts...

"Science is a way of trying not to fool yourself. The first principle is that you must not fool yourself, and you are the easiest person to fool."
(Richard Feynman)



Model quality control

is also a means of ensuring responsibility : withstanding the scrutiny of a critical reader (including reviewers, PDB annotators, fellow scientists, and the whole community!)



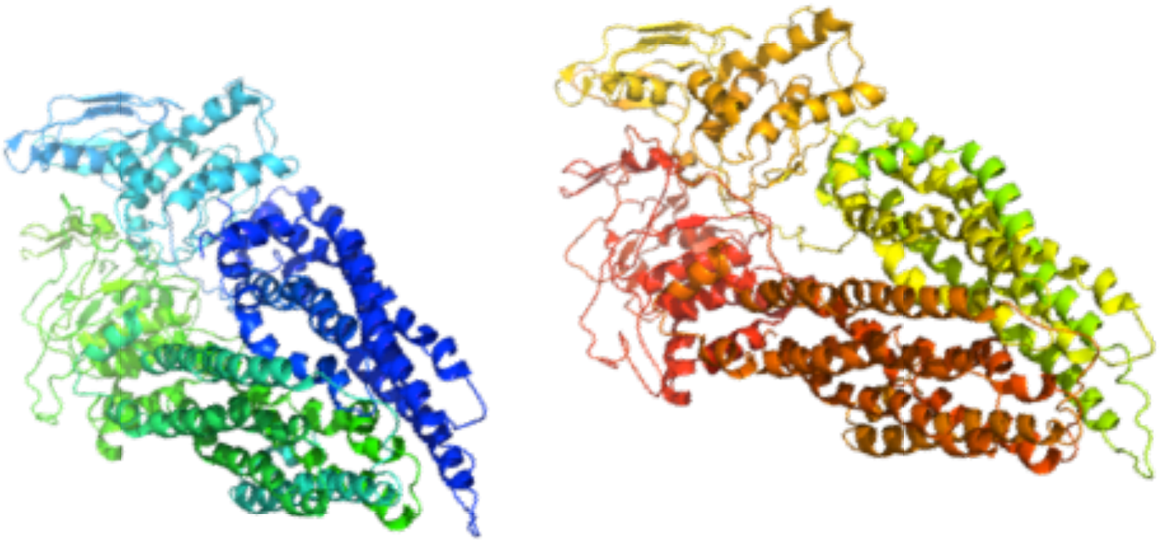
RETRACTED: Structure of MsbA from *Vibrio cholera*: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation

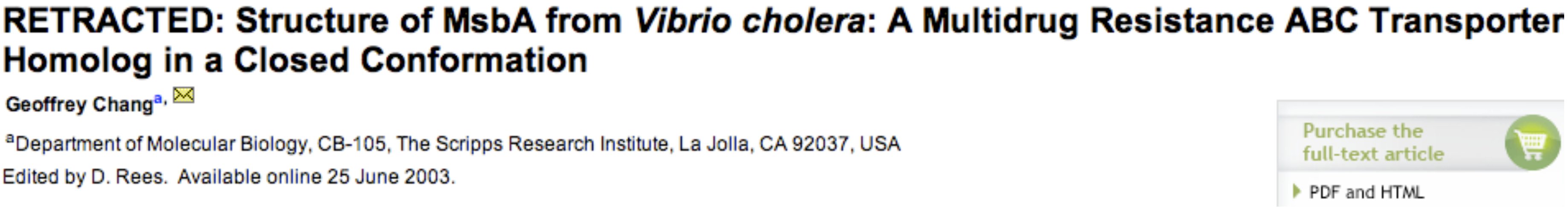
Geoffrey Chang[a]

[a]Department of Molecular Biology, CB-105, The Scripps Research Institute, La Jolla, CA 92037, USA

Edited by D. Rees. Available online 25 June 2003.

"were incorrect in both the hand of the structure and the topology. Thus, the biological interpretations based on the inverted models for MsbA are invalid."





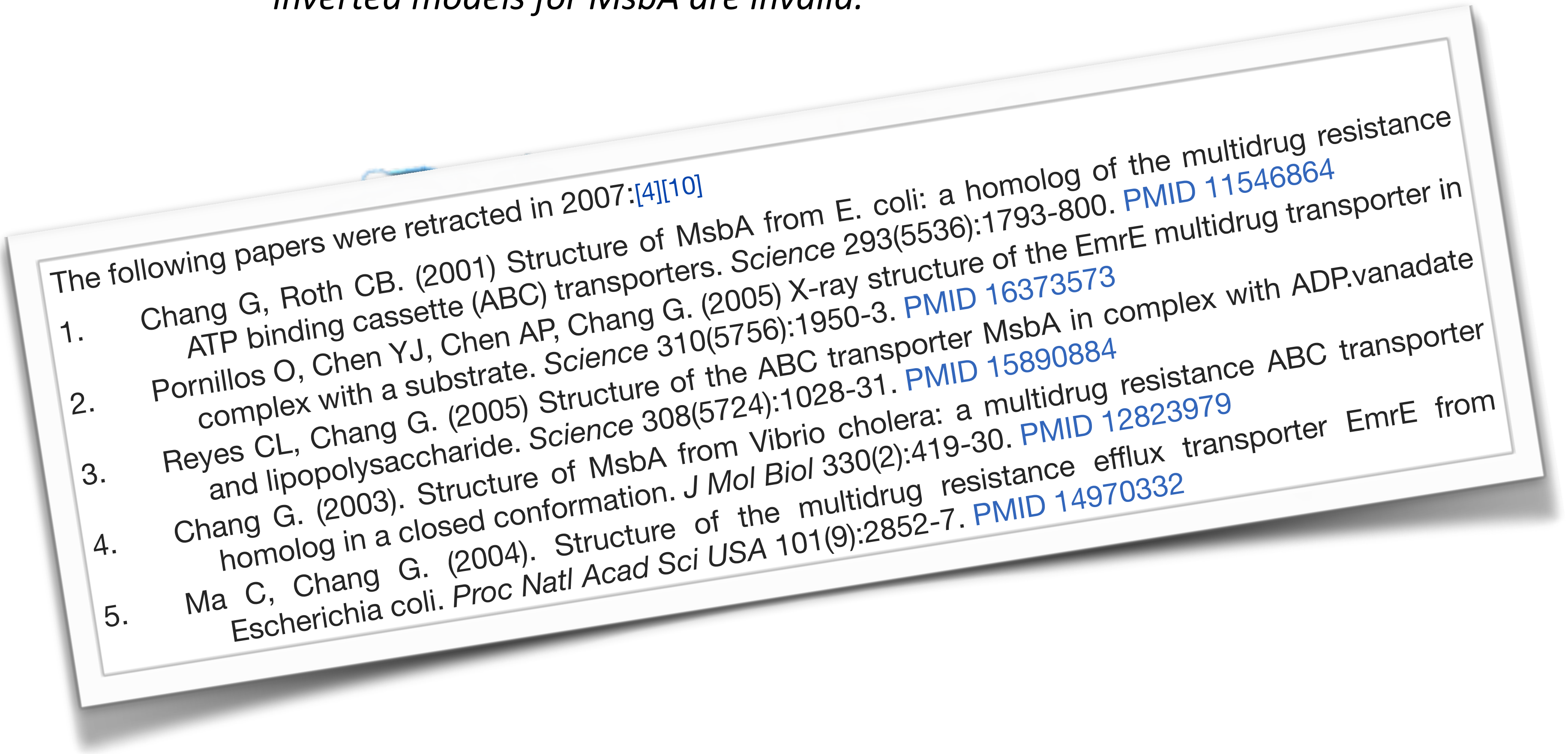
RETRACTED: Structure of MsbA from *Vibrio cholera*: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation

Geoffrey Chang[a]

[a]Department of Molecular Biology, CB-105, The Scripps Research Institute, La Jolla, CA 92037, USA

Edited by D. Rees. Available online 25 June 2003.

"were incorrect in both the hand of the structure and the topology. Thus, the biological interpretations based on the inverted models for MsbA are invalid."



The following papers were retracted in 2007:[4][10]

1. Chang G, Roth CB. (2001) Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* 293(5536):1793-800. PMID 11546864
2. Pornillos O, Chen YJ, Chen AP, Chang G. (2005) X-ray structure of the EmrE multidrug transporter in complex with a substrate. *Science* 310(5756):1950-3. PMID 16373573
3. Reyes CL, Chang G. (2005) Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide. *Science* 308(5724):1028-31. PMID 15890884
4. Chang G. (2003). Structure of MsbA from Vibrio cholera: a multidrug resistance ABC transporter homolog in a closed conformation. *J Mol Biol* 330(2):419-30. PMID 12823979
5. Ma C, Chang G. (2004). Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli. *Proc Natl Acad Sci USA* 101(9):2852-7. PMID 14970332

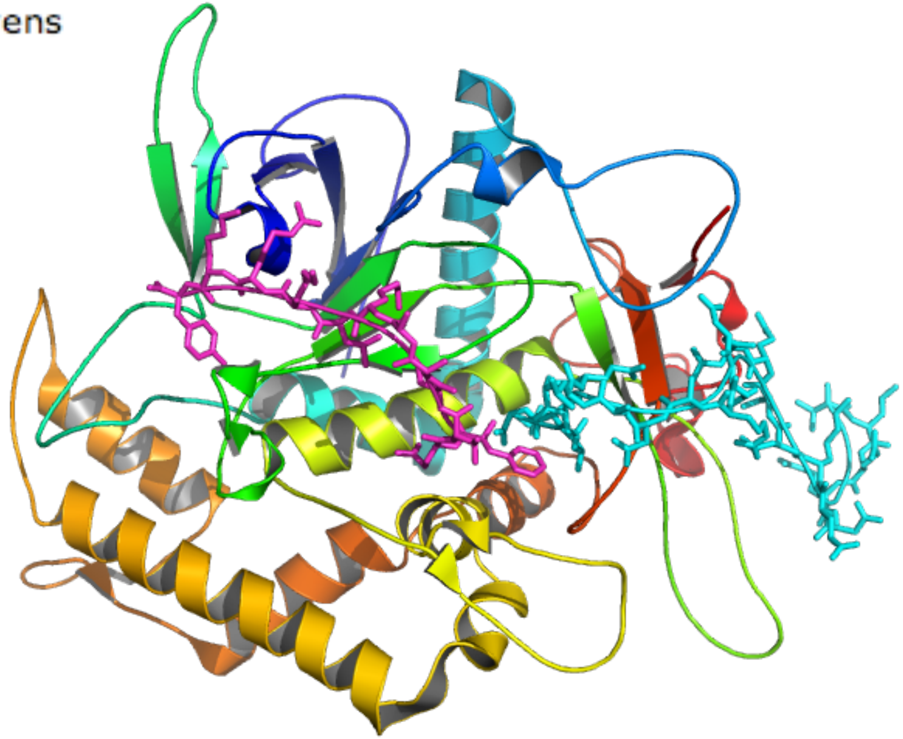
Nature Structural & Molecular Biology **16**, 795 (2009)
doi:10.1038/nsmb0709-795

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

Michael A Hanson & Raymond C Stevens

“However, because of the lack of clear and continuous electron density for the peptide in the complex structure, the paper is being retracted.”

1F83



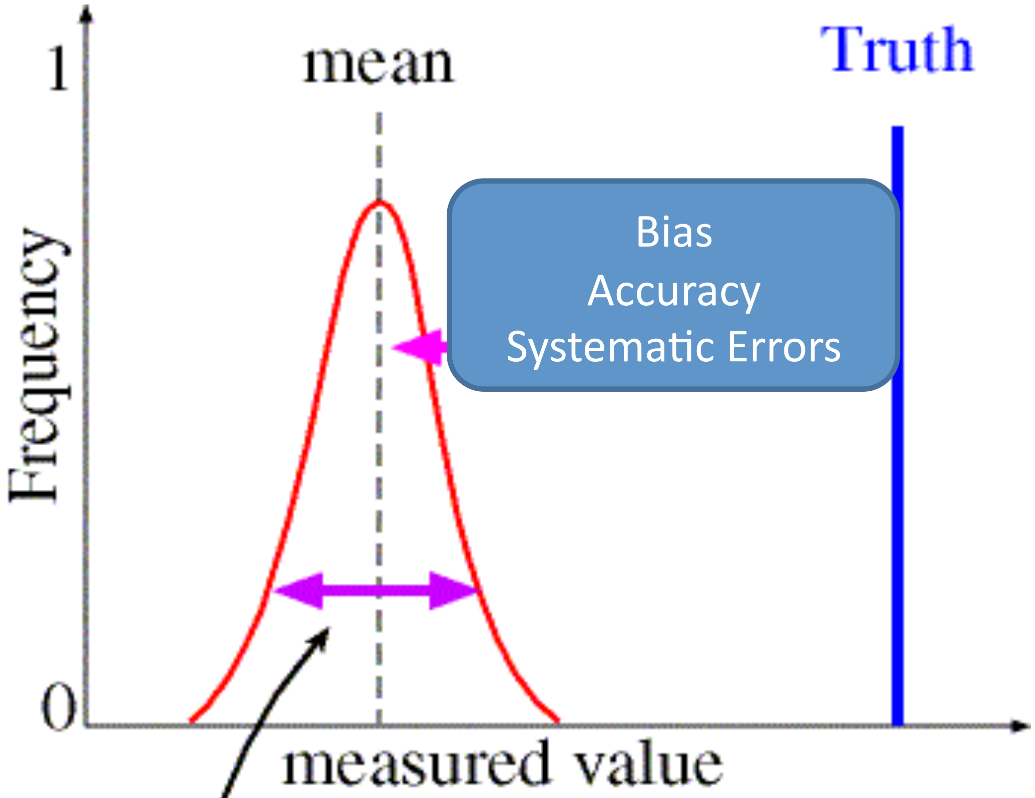
Model quality control

is also a means of ensuring responsibility : withstanding the scrutiny of a critical reader (including reviewers, PDB annotators, fellow scientists, and the whole community!)

but, it's important to note

- the complexity of defining "error" (mistake), when it comes to evolving interpretation of results!
- the need for judicious analysis of the outputs of validation programs and statistics (outliers are less probable, but not necessarily impossible!) : checking against expectation values

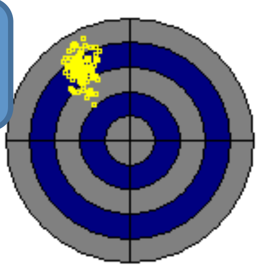
Errors affect measurement (and interpretation)



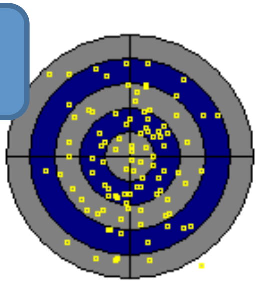
Bias
Accuracy
Systematic Errors

Consistency
Precision
Random Errors

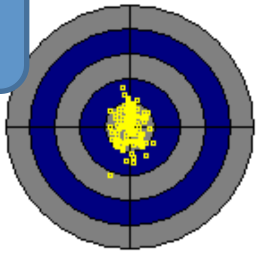
Precise, but inaccurate



Accurate, but imprecise



Accurate and Precise



several of the most important parameters that define a crystallographic model

1) Biochemical entities :

- Biopolymers
  (polypeptides, polynucleotides, carbohydrates)
- Small-molecule ligands (ions, organic)
- Crystallographic additives, e.g. GOL, PEG
- Solvent

2) Coordinates, Displacement

- Unique x,y,z
- Partial, multiple, absent (occupancy)
- Isotropic or anisotropic B factors
- TLS approximation

3) Bulk solvent model (Ksol, Bsol)

4) Crystallographic parameters

- Cell, symmetry, NCS

A high-quality MX model makes sense in all respects

- Chemical

Bond lengths, angles, planarity, chirality

- Physical

Good packing, sensible interactions, reasonable atomic displacement distribution

- Crystallographic

Low crystallographic residual, residues fit density, flat difference map

- Protein Structure

Ramachandran, peptide bonds torsion angles, rotamers, disulphides, salt bridges, pi-interactions, hydrophobic core

- Statistical

Best possible hypothesis to fit data, no over-fitting, no under-modelling

- Biological

Explains observations (activity, mutants, inhibitors, cell phenotype, protein:protein interactions data)
Is predictive

Model quality control

important misconception to highlight : “a structure that has been deposited in the PDB is of sufficient quality and cannot be wrong”... actually, the author is ultimately responsible (not the annotators!)

Beyond mere geometry checking...

- Global vs local

global descriptors (e.g. refinement R factors, overall stereochemical deviations from target values, bulk solvent model, avg and Wilson B factors, etc) are first quality indicators, and not proof of absence of (even important) mistakes

certainty (coordinates, B factors, etc) varies along a single model, so reliability of models is mostly a local property! (most relevant for biological aims)

Beyond mere geometry checking...

- Global vs local

local descriptors : rotamers, model:map correlation, values of 2mFo-DFc and mFo-DFc at and around atomic positions, sequence register, ligand identity, individual B-factors and distribution, occupancies, etc

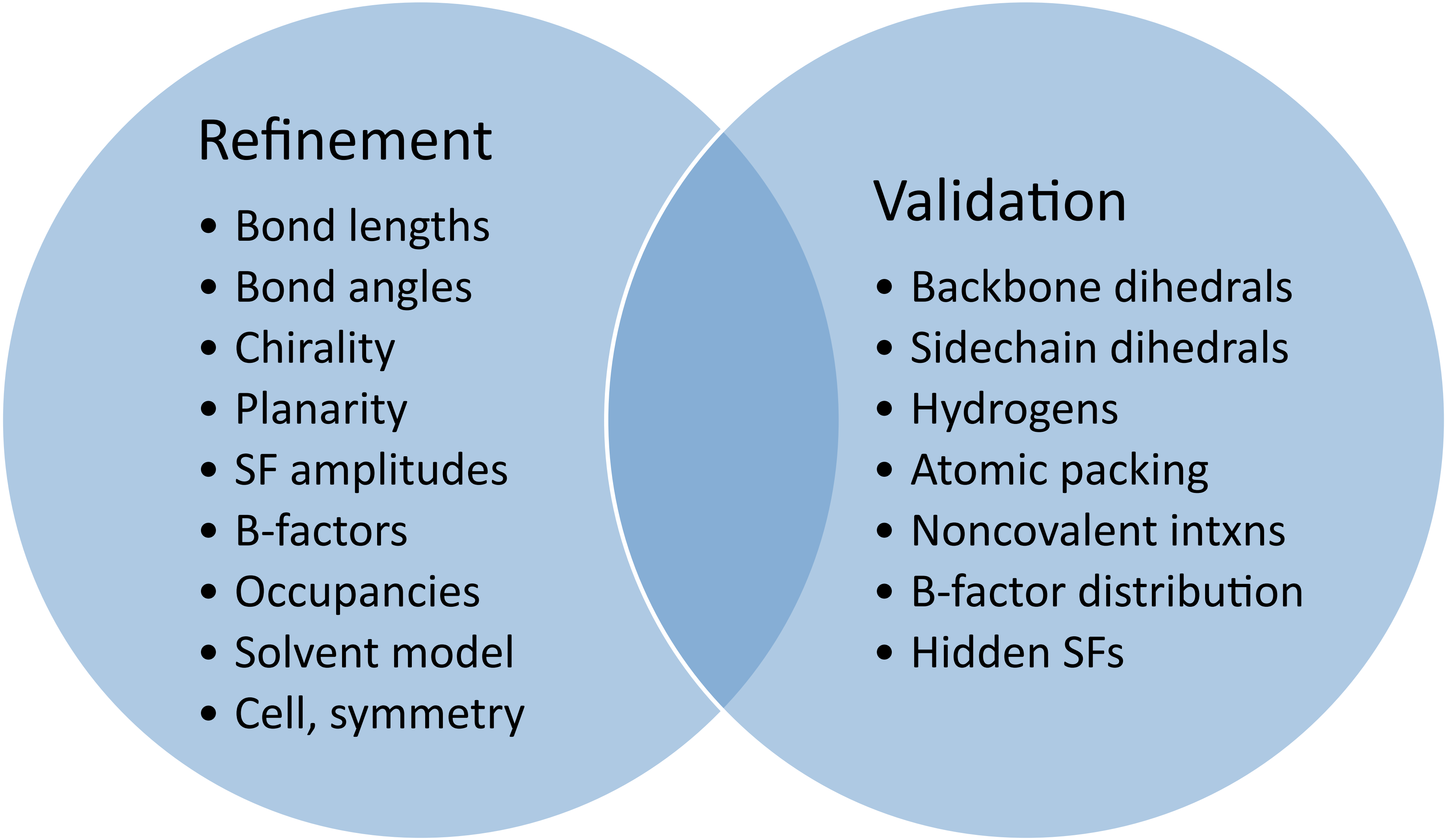
Beyond mere geometry checking...

- REMEMBER : validation criteria that examine properties that have been restrained during refinement (bond distances, angles, planarity, etc) or purposefully sought to be modified (refinement programs seek for Rcryst minimization!), are somehow tautologic, reflecting what we searched for!!!

- they are still useful to examine outliers, and most importantly to judge on the progress (and eventual convergence) of the refinement procedure itself...

- but they need to be combined with evidence-based confirmation : electron density map!!

Validation done against unrefined entities is powerful



Refinement

- Bond lengths
- Bond angles
- Chirality
- Planarity
- SF amplitudes
- B-factors
- Occupancies
- Solvent model
- Cell, symmetry

Validation

- Backbone dihedrals
- Sidechain dihedrals
- Hydrogens
- Atomic packing
- Noncovalent intxns
- B-factor distribution
- Hidden SFs

Types of quality criteria for macromolecular crystallography

- Global vs local

- Model-only

How good is model irrespective of experiment?

Only coordinates are used

Simple, intuitive

- Model and data

How well does the model fit the data?

Crucial! Sets your model apart from theoretical model!

- Data-only

Data-Quality + Crystallographer = Model Quality

Good data necessary for reliable model

Can be understood readily only by expert crystallographer

Data only

Data only

Principles of Protein X-Ray
Crystallography, Jan Drenth, Springer 3rd ed

R-Factor for Comparing the Intensity of Symmetry-Related Reflections

$$R_{\mathrm{sym}}(I) = \frac{\sum\limits_{hkl}\sum\limits_{i} |I_i(h\,k\,l) - \overline{I(h\,k\,l)}|}{\sum\limits_{hkl}\sum\limits_{i} I_i(h\,k\,l)}$$

Precision Indicating Merging *R*-Factor for Determining the Precision of an Average Measurement

$$R_{p.i.m.} = \frac{\sum\limits_{hkl} \frac{1}{(N-1)^{1/2}} \sum\limits_{i} |\,I_i(h\,k\,l) - \overline{I\ (h\,k\,l)}\,|}{\sum\limits_{hkl}\sum\limits_{i} I\ (h\,k\,l)},$$

where N is the redundancy of the data and $\overline{I\ (h\,k\,l)}$ the average intensity. This *R*-factor has the advantage over $\mathrm{R}_{\mathrm{sym}}$, which it is redundancy independent

Data only

Principles of Protein X-Ray
Crystallography, Jan Drenth, Springer 3rd ed

R-Factor for Comparing the Intensity of Symmetry-Related Reflections

$$\sum \sum |I_i(h\,k\,l) - \overline{I(h\,k\,l)}|$$

R_{sym} is obsolete now, but useful to understand the meaning of Rs

Precision Indicating Merging *R*-Factor for Determining the Precision of an Average Measurement

$$R_{p.i.m.} = \frac{\sum\limits_{hkl} \frac{1}{(N-1)^{1/2}} \sum\limits_{i} |\, I_i(h\,k\,l) - \overline{I\;(h\,k\,l)}\,|}{\sum\limits_{hkl} \sum\limits_{i} I\;(h\,k\,l)},$$

where N is the redundancy of the data and $\overline{I\;(h\,k\,l)}$ the average intensity. This *R*-factor has the advantage over R_{sym}, which it is redundancy independent

Data only checks

Quality of the X ray diffraction data is essential for eventually achieving a good quality model !

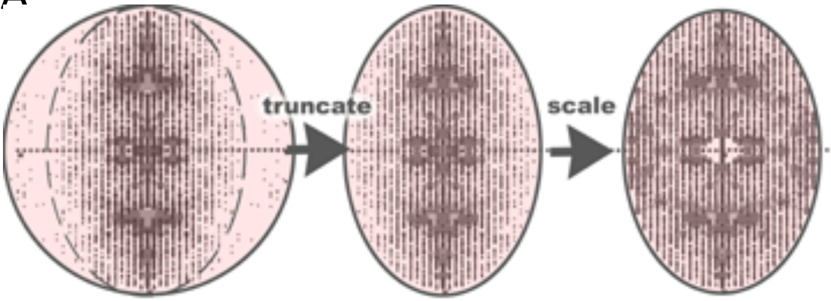
- Wilson plot (phenix.xtriage, truncate, etc to analyze)
  - Average intensity in resolution bins
  - Has a characteristic shape
  - too high a mean intensity at low resolution, or increasing mean intensity at high resolution, can indicate problems with data processing
  - twinning, translational NCS, extreme solvent content : can modify the plot

- Twinning: Padilla-Yeates plot and others

Data-only quality checks

- Anisotropy
  - Break-up of Wilson plot for diff h, k, l directions
  - Model can probably be better refined using data with resolution anisotropically truncated (UCLA — Diffraction Anisotropy Server http://services.mbi.ucla.edu/anisoscale)

http://staraniso.globalphasing.org



- Data quality
  - Completeness
    - Completeness reduces towards higher resolution shells
    - I / σ(I), signal to noise, drops at higher resolution
  - Rmerge: how well do reflections agree across frames.
  - Rmeas/Rpim/CC(1/2): how well do the symmetry-related reflections agree.
  - Has the the right resolution cutoff been chosen?

Model only criteria

Model only criteria

- Stereochemistry

Covalent bonds, angles, dihedrals, chirality, planarity, ring geometry

- Dihedral angle distributions

Ramachandran, sidechains, RNA backbone

Derived distributions from small-molecule datasets

…don't forget ligands!!! They are molecules too… :)

- Packing

Bad vdw clashes

Underpacking

Hydrogen bonds and environment

Examining model stereochemistry

Geometric model validation compares model properties such as stereochemistry, local chemical environment and packing propensity, against their empirical expectation values based on prior knowledge.

* Z-values (multiples of their esu, when the target values have well-established uncertainties)

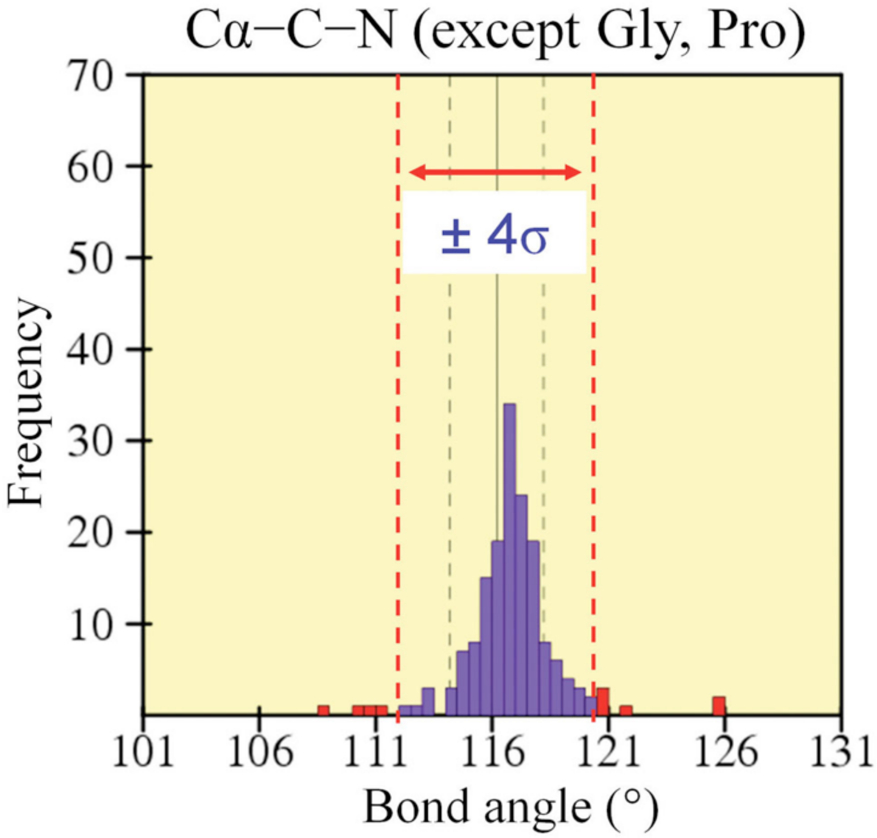
* the clashscore in MolProbity (total number of close contacts per 1000 atoms)

* the error function in Errat (statistics of non-bonded atom-atom interactions, compared to a database of reliable high-resolution structures)

Examining model stereochemistry

Many programs : Procheck, Whatcheck, MolProbity, Errat, Verify3d

http://molprobity.biochem.duke.edu
http://nihserver.mbi.ucla.edu/ERRAT/
http://nihserver.mbi.ucla.edu/Verify_3D/
… and others



very useful tools from within Coot!!

Stereochemistry outliers (e.g. using Procheck)

Covalent geometry

- **Reference sources for bonds and angles**
  - for Proteins and Nucleotides
    - ‣Small-molecule crystallography
      - ✳does not suffer from the phase problem!
      - ✳Numerous expt-structures (CSD > 875000)
    - ‣Ultra-high resolution MX structures
    - ‣Mean, variability = refinement target, force constants
    - ‣Engh & Huber (1991,2001), Parkinson et al (1996)
  - Small-molecules
    - ‣Comparable fragments from small-molecule database
    - ‣Mogul, JLigand, AceDRG among others to create topology, define geometry parameters

Covalent geometry

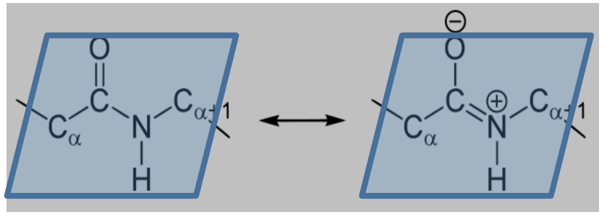
- Small variation -> highly restrained in refinement

  - Bond length variation ~ 0.02 Å, angle variation ~ 2°, etc etc

  - But still useful to check for large deviations

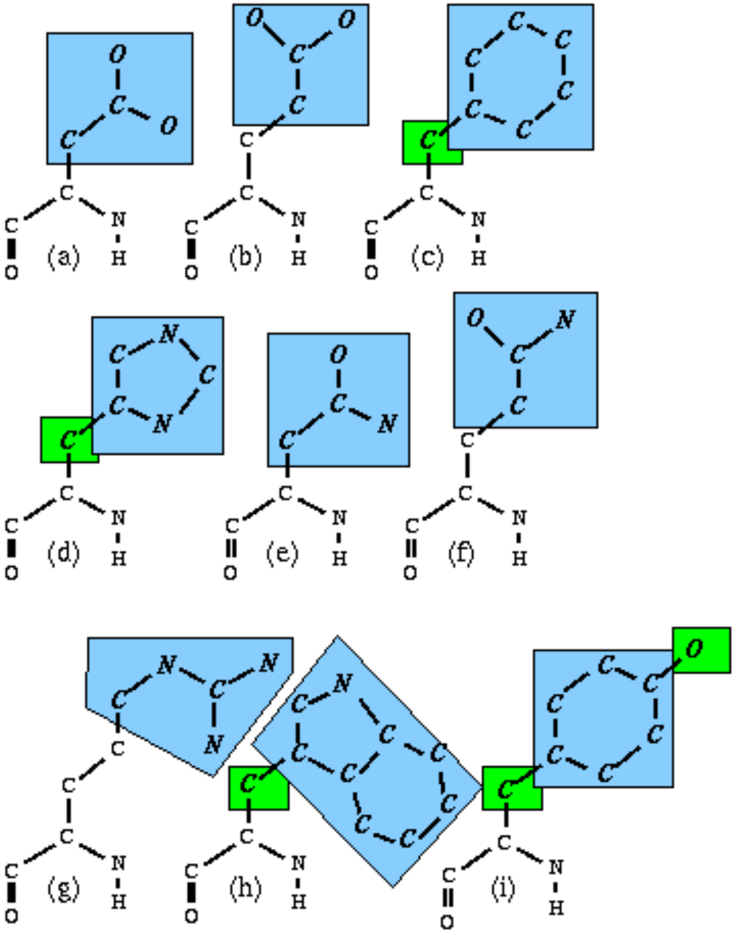
    - ▸refinement problems, incorrect parameters
    - ▸Systematic directional error in lengths due to wrong cell

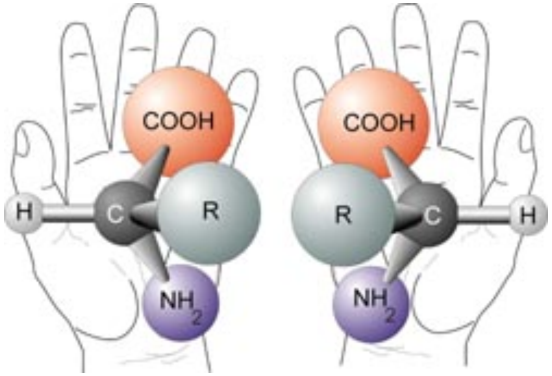
Covalent geometry of proteins



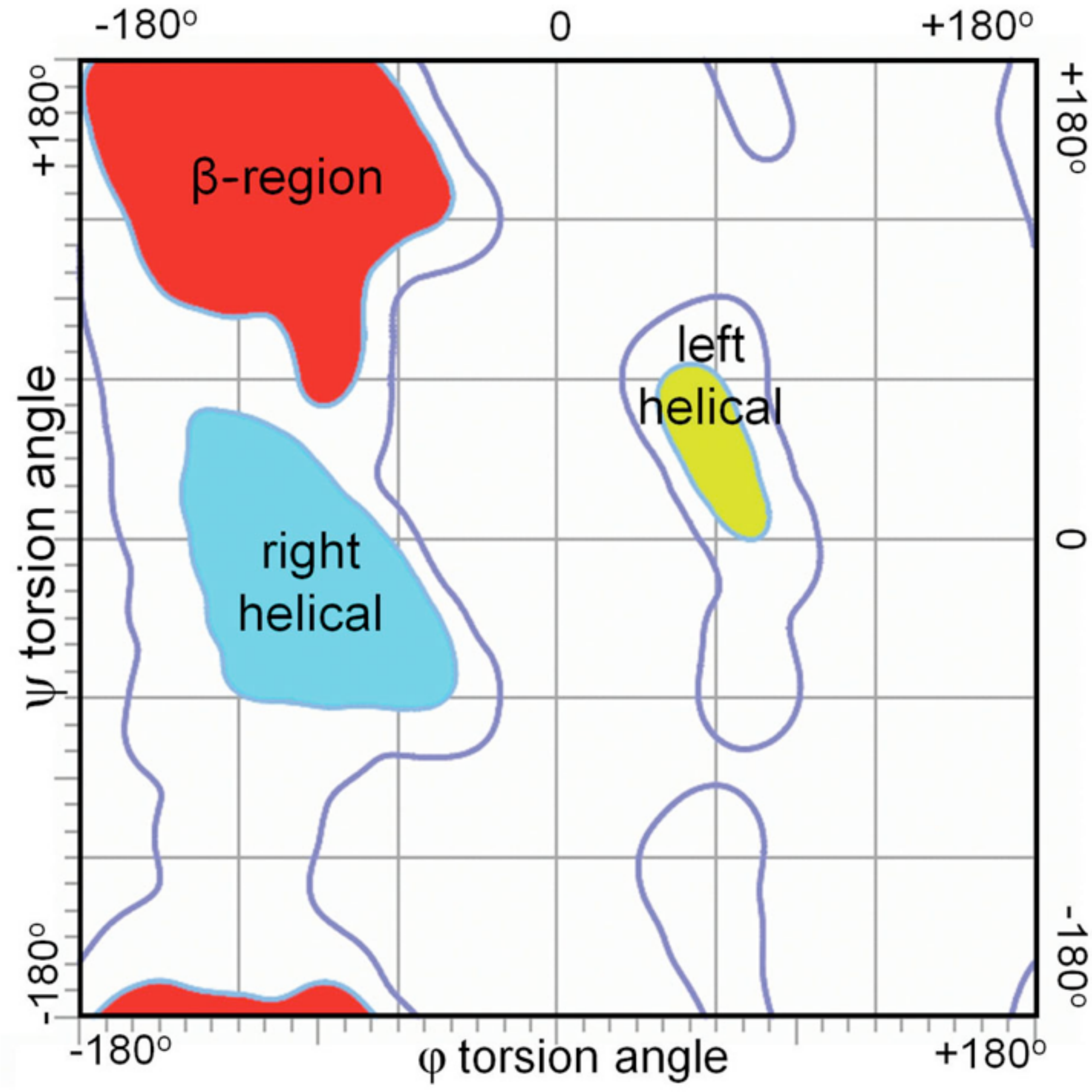
- Planarity
  - Peptide bond
  - Phe, Tyr, Trp, His, nucleotide bases
  - Arg, Gln, Asn, Glu, Asp

- Chirality
  - Should be always L at CA
  - Gly is not chiral!
  - CB in Ile is (2S,3S) and in Thr (2S,3R)
  - CA-N-C-CB ~ 34°, chiral volume ~ 2.5 Å^3





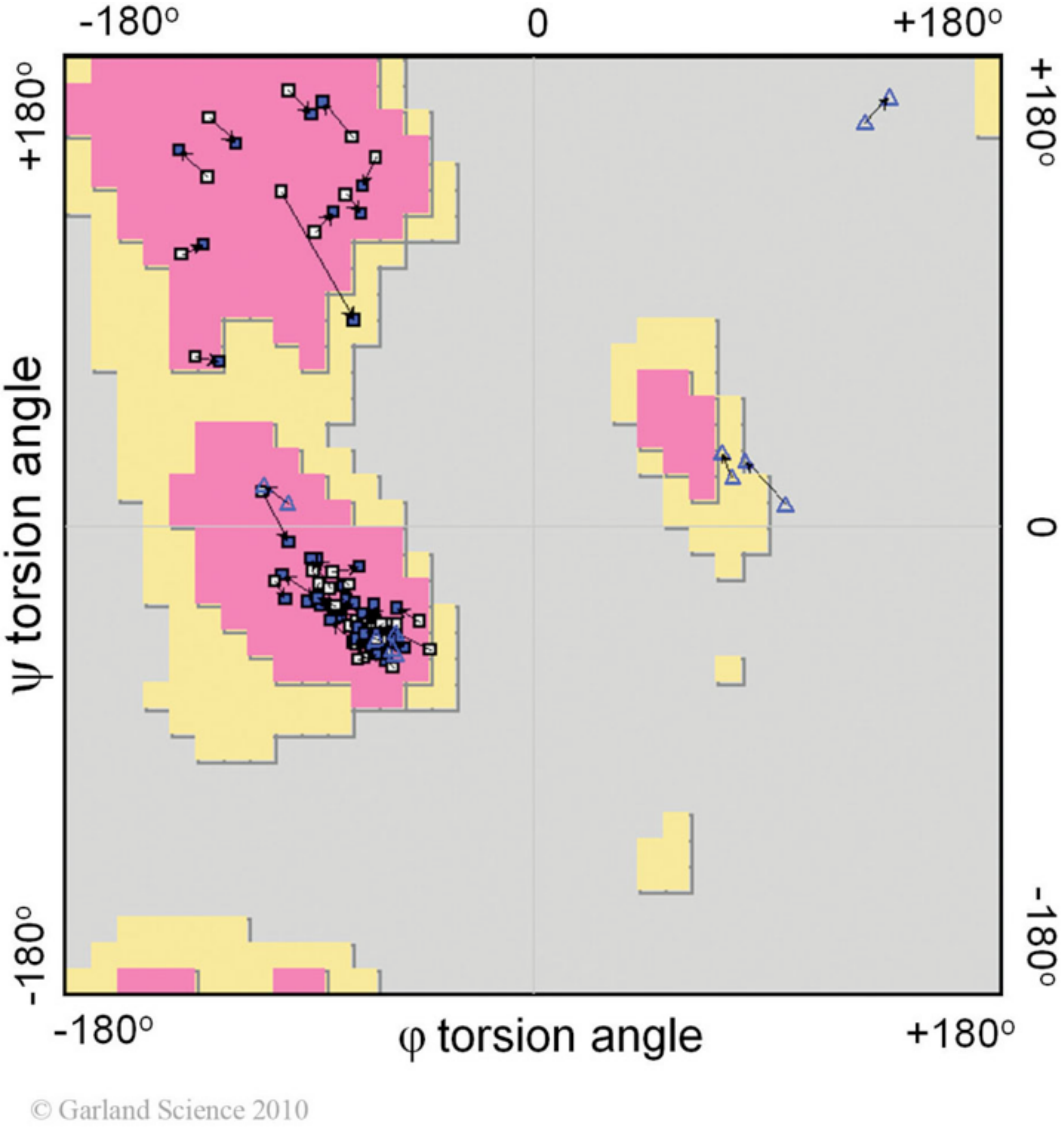
Dihedral angles : distributions



Ramachandran plot

on average, 98% of the residues are expected to lie within the core regions, and 0.2% outside the second boundary

different distributions for Gly, pre-Pro & Pro

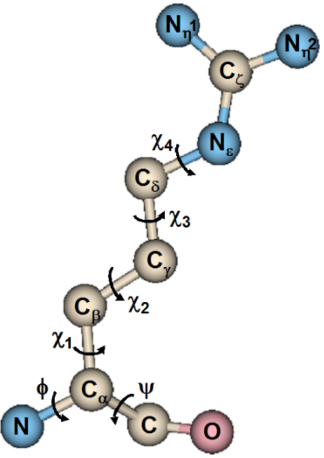


© Garland Science 2010

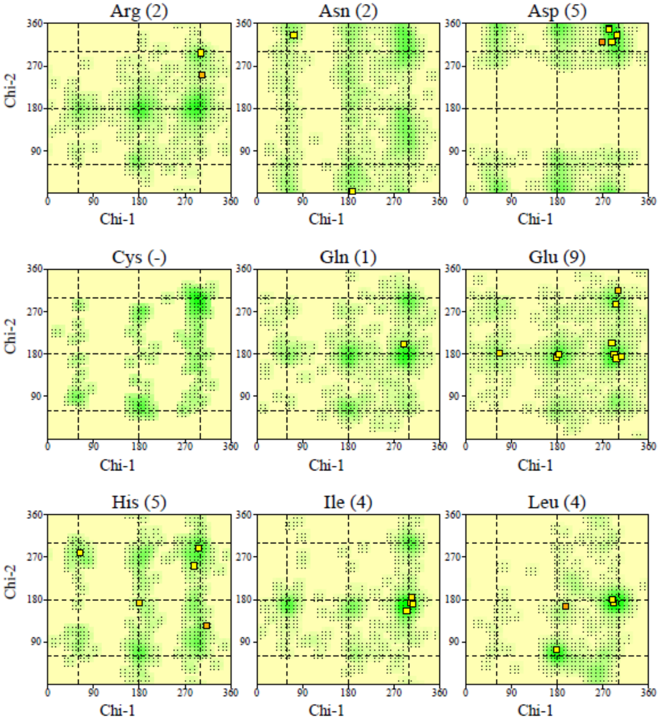
Backbone torsion angle distribution for NCS-related molecules

("Kleywegt plots")

...even random coil peptides do not have random φ/ψ torsions!



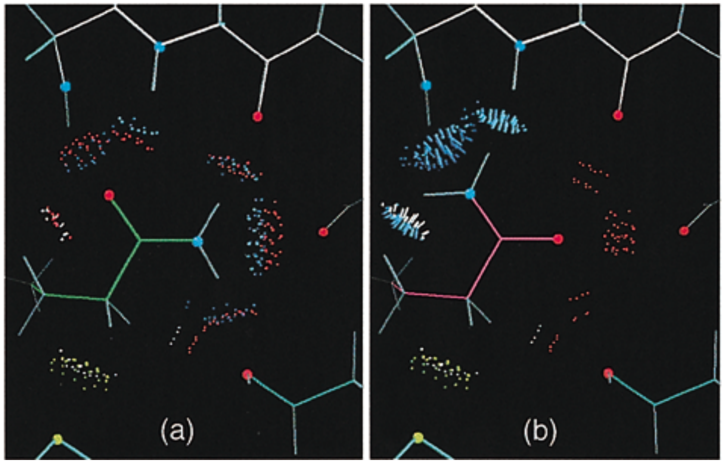
Side chain quality



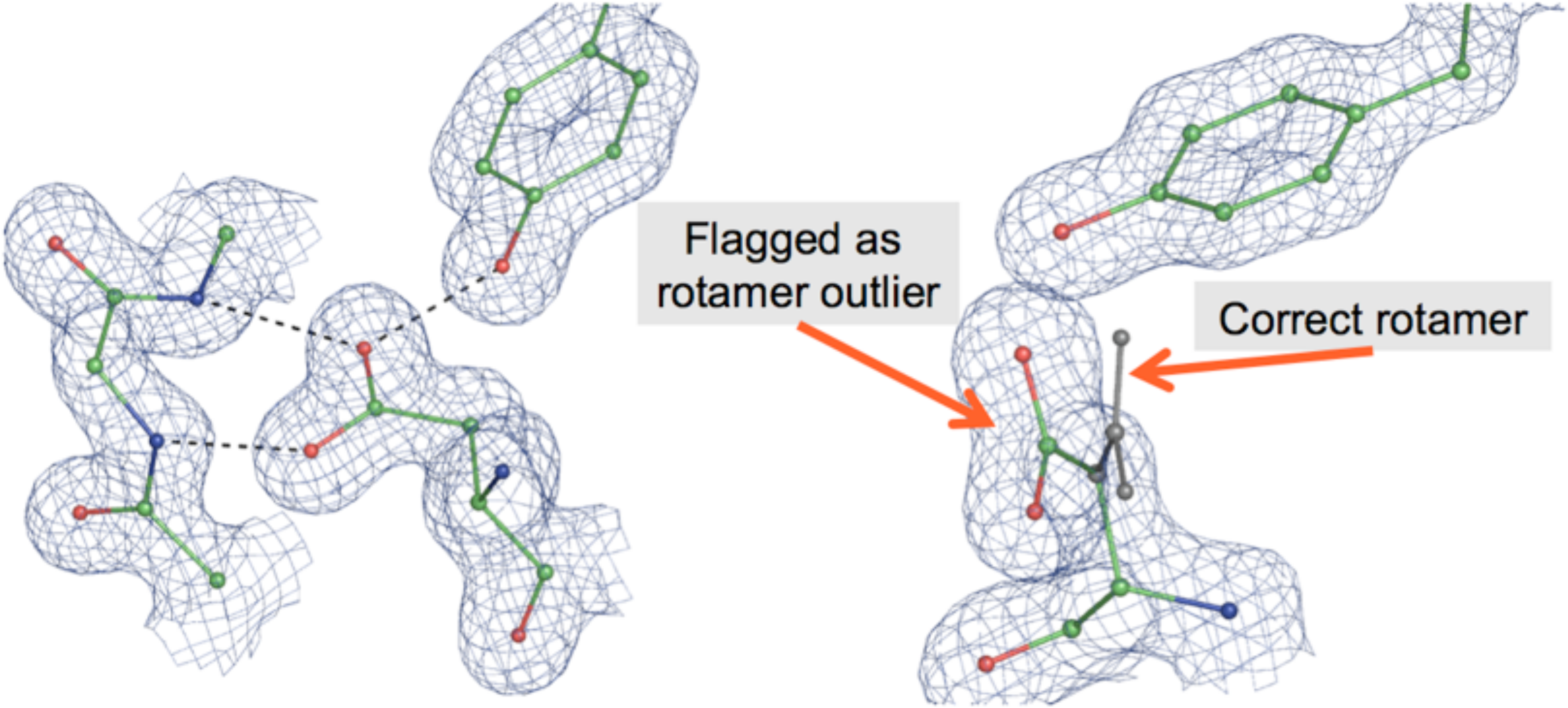
- Fraction of rotameric sidechains
  - Rotamericity calculations vary slightly between MolProbity, ProCheck, WhatCheck

- Non-rotameric
  - Does not mean incorrect
  - But is there clear density to justify the modelled conformation?
  - Does the conformation make sense in the environment?

- Can the sidechain be flipped?
  - Asn (ND1, OD2), Gln (NE1,OE2), His (ND2, NE2) are not unambiguously defined by electron density
  - Does flipping make the model better?
    - E.g. Gln90 in 1REI : Better H-bonds and reduced bad contacts after flip



Look at the maps!! not all outliers are wrong: evidence, when strong, can refute expected prior knowledge



Flagged as rotamer outlier

Correct rotamer

- 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

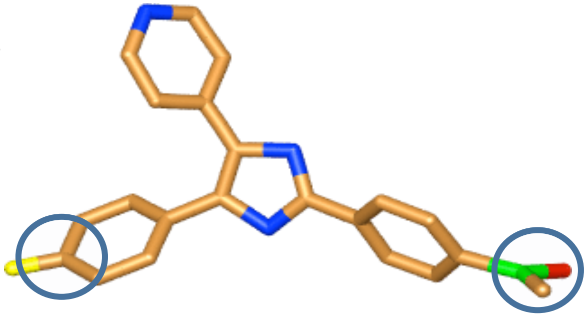
adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010

Covalent geometry of ligands

- Small molecule ligands have huge variety
  - They can get modified on soaking.

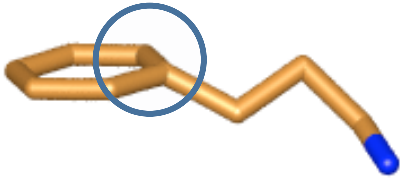
- Few geometric rules other than the basic rules
  - Chirality (when known)
  - planarity of aromatics and conjugated systems
  - almost invariant bond lengths and angles
  - CCDC preferences for fragments of molecules

- Wrong ligand geometry does not result in overall bad crystallographic statistics for the complex
  - Very often ligands end up having a poor geometry.



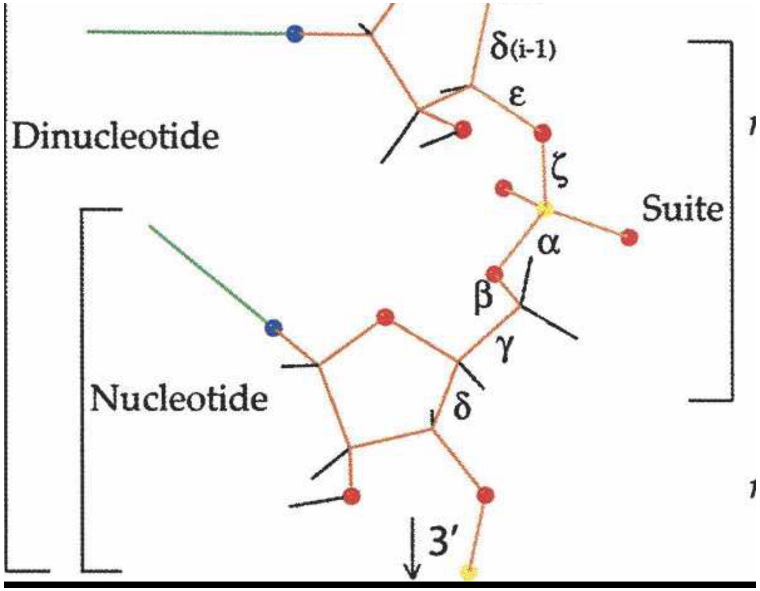
- SB-203580 in 1PME, 1998, 2.0Å, *Prot. Sci.*

- 3-Phenylpropylamine, in 1TNK, 1994, 1.8Å, *Nature Struct. Biol.*



Nucleic acid validation

- Essential to check quality of nucleic acids as much as proteins!
- Prominent tetrahedral phosphates and planar bases
- Sugar-phosphate backbone defined by 6 dihedrals
  - ~ 50 frequent ‘suites’
- Dominant puckers are C3’-endo, C2’-endo
- Implemented in MolProbity
- Quality metrics
  - Percentage of unfavorable backbone suites
  - Percentage of unlikely ribose puckers



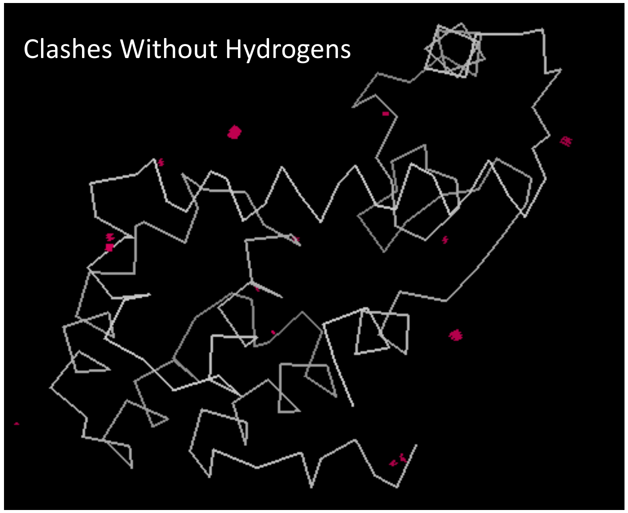
Packing as a powerful validation criterium : clashes

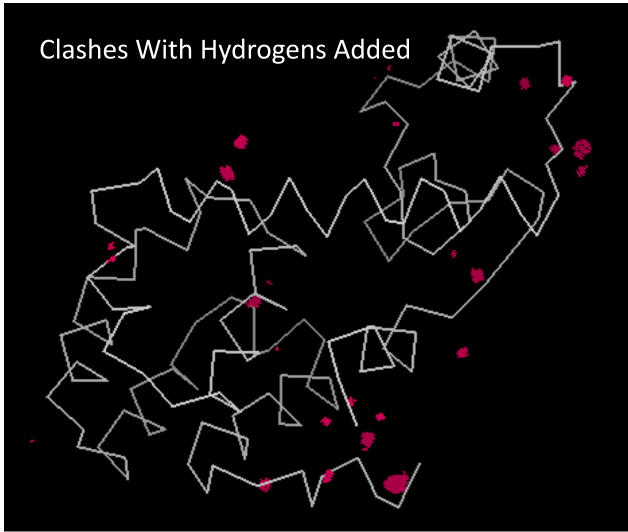
- D(A,B) < vdwR(A) + vdwR(B)
  - Covalent bonding? Noncovalent interaction?
  - Steric clash! Unrelated atoms cannot get arbitrarily close

- Heavy atom clashes are rare and avoided in refinement

- Hydrogens
  - generally absent in refinement.
  - Clashes on rebuilt hydrogens is a powerful validation check!

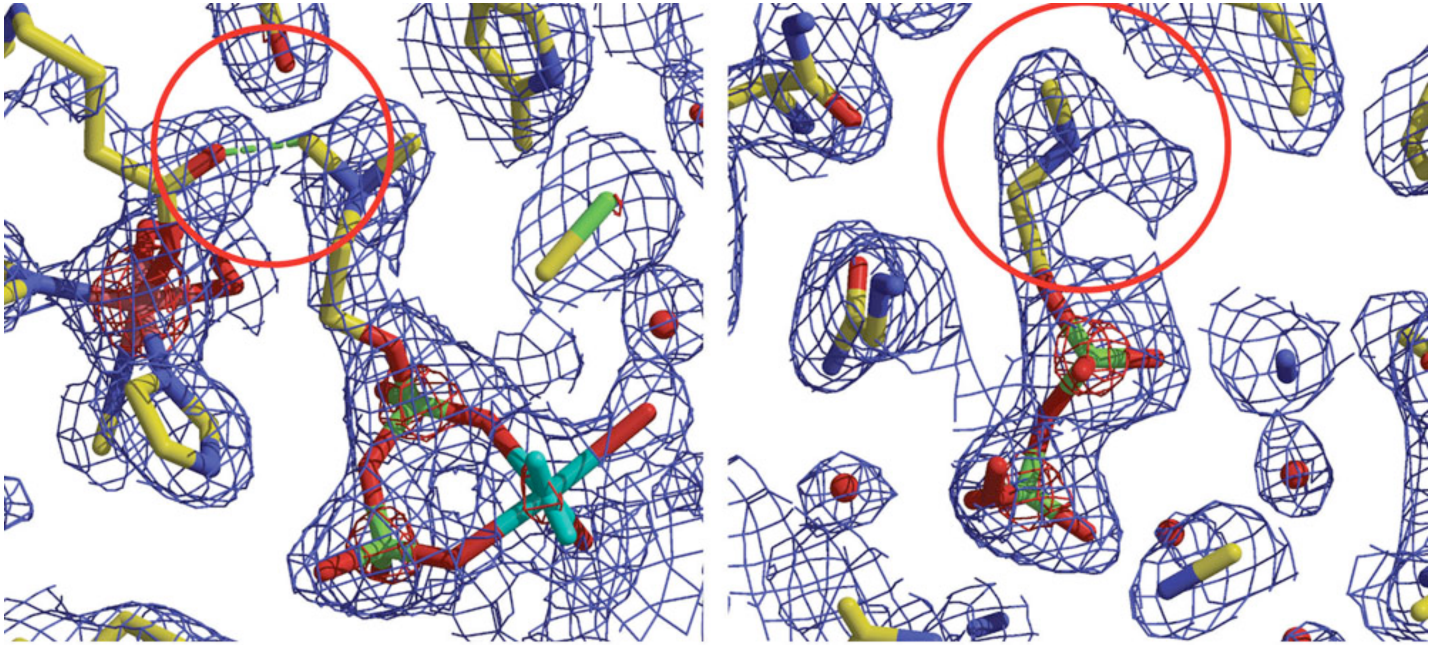
- Quality metric
  - Number of bumps per 1000 atoms after adding hydrogen atoms
  - Local: per residue clashes
  - Completeness of model: Fraction of non-solvent atoms present in the model with decent occupancy and B-factors





MolProbity all-atom contact analysis

- it adds hydrogen atoms for all residues in riding positions, and then evaluates all-atom contacts
- enables better judgement of clashes

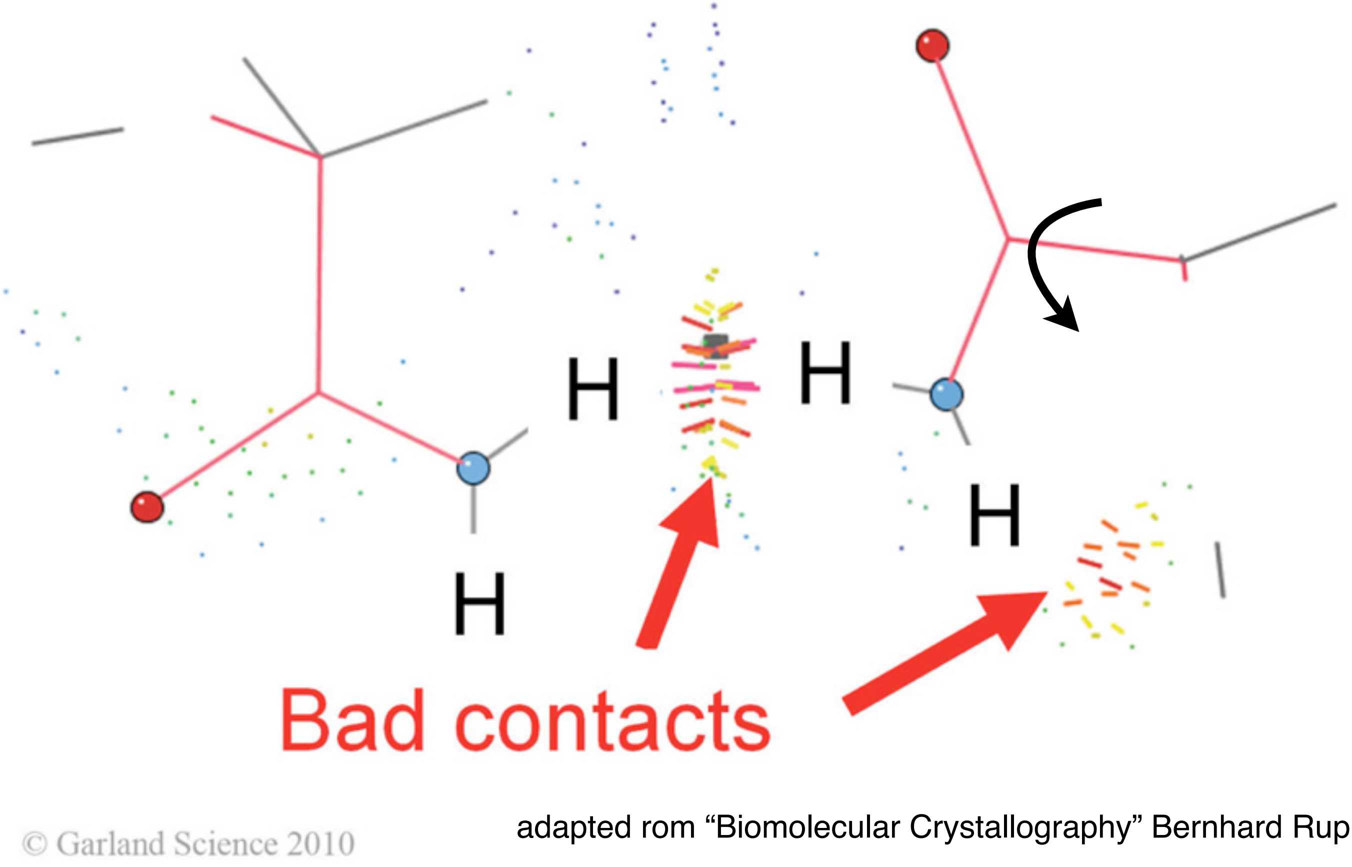


© Garland Science 2010

adapted rom "Biomolecular Crystallography" Bernhard Rupp, Garland Science 2010

MolProbity all-atom contact analysis

- ...and H-bond networking analysis (particularly useful to guide NQH side-chain flipping)



adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010

Judging on packing quality

- Protein interiors
  - well-packed with complementary surfaces
  - satisfied H-bond donors, acceptors
  - don't have voids

- Interior voids can be due to inflated unit cell dimensions, e.g. T4 lysozyme identified by RosettaHoles (Sheffler & Baker, 2008)

- Interaction quality for residues
  - Count fraction of unsatisfied buried H-bond donors/acceptors
  - Report atypical neighborhood not observed previously in the database
  - e.g. What_check, DACA, verify3D

Model vs data criteria

Model vs data criteria

1- Data sufficiency for model parameterization

Resolution and its effect on the data-to-parameters ratio

2- R factors

Match between observed and calculated structure factor amplitudes

3- Map quality (clarity, all features explained) and quality of mutual fit between model and map

4- Validation of protein-ligand complexes

5- B factors (distribution, variation)

1- Is the model plausible with respect to the amount of data available in the experiment?

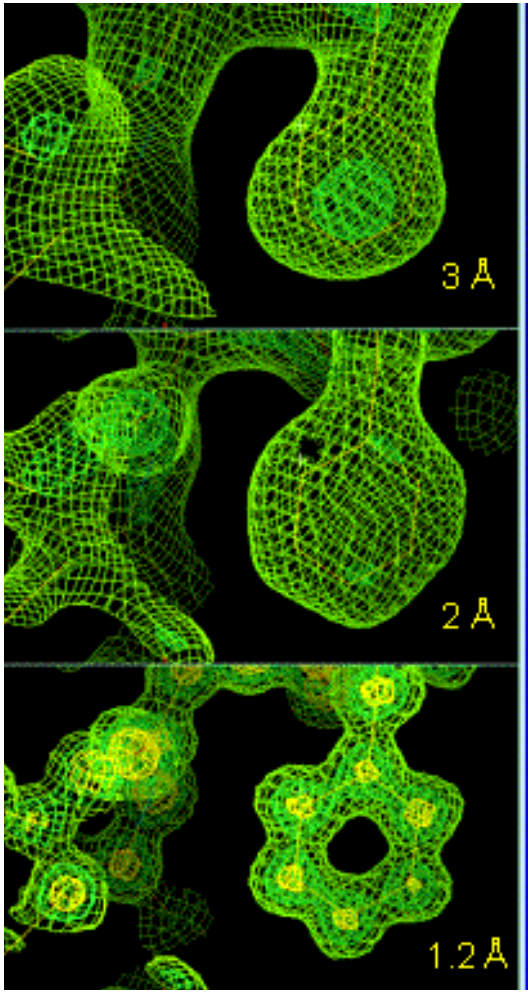
The model can be constructed at various levels of detail

CA-only all the way to explicit hydrogens

Macromolecule only or waters & small molecules also

Overall or TLS or atomic (isotropic or anisotropic) B factors

Single or multiple conformers with partial occupancies

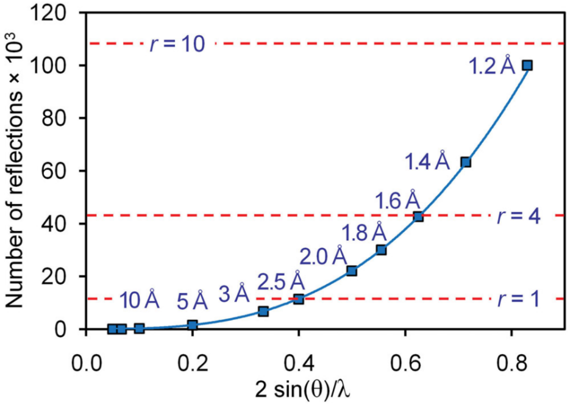


The same amount of detail cannot be modelled across all resolutions

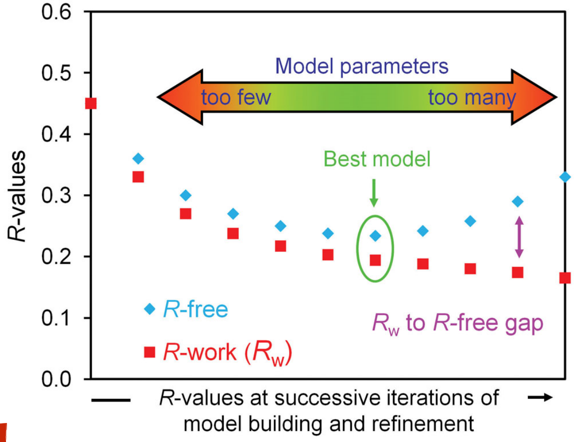
- Higher resolution = more information
- A good model has just enough detail to explain the observed data without overfitting it
- A model with high data to params ratio is more reliable
- Low data:parameters ratio can lead to overfitting which manifests as model errors

Beware of a model...

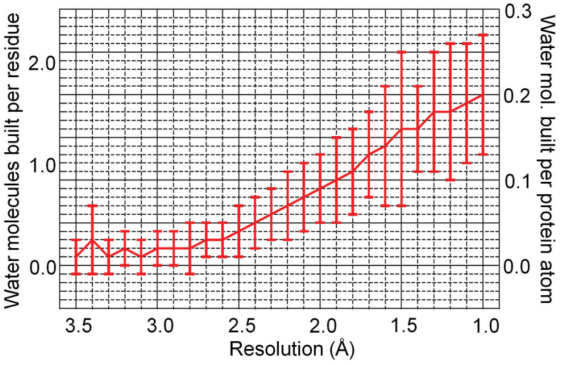
- With anisotropic B factors at 3Å res
- With multi-model refinement at 4.5Å (e.g. Chang, Roth 2001)
- With hydrogens or many waters modelled at 2.7Å



© Garland Science 2010

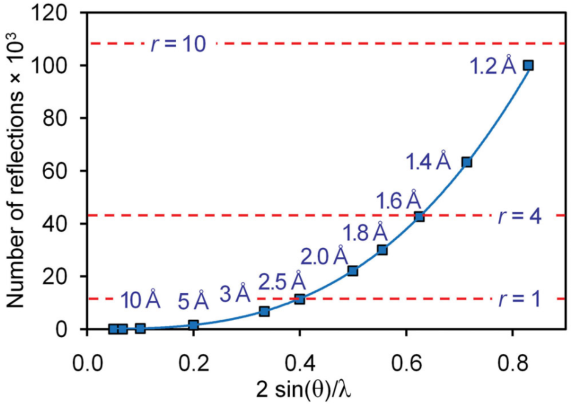


© Garland Science 2010



© Garland Science 2010

The same amount of detail cannot be modelled across all resolutions



© Garland Science 2010

- Higher resolution = more information
- A good model has just enough detail to explain the observed data without overfitting it
- A model with high data to params ratio is more reliable
- Low data:parameters ratio can lead to overfitting which manifests as model errors

Beware of a m

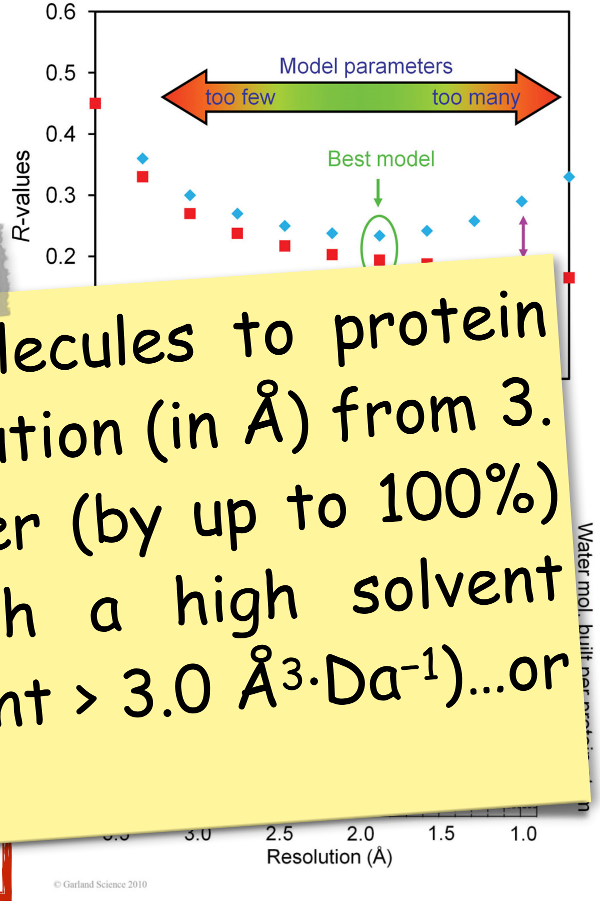
- With anis

- With mult

Chang, Roth

- With hydr

2.7Å

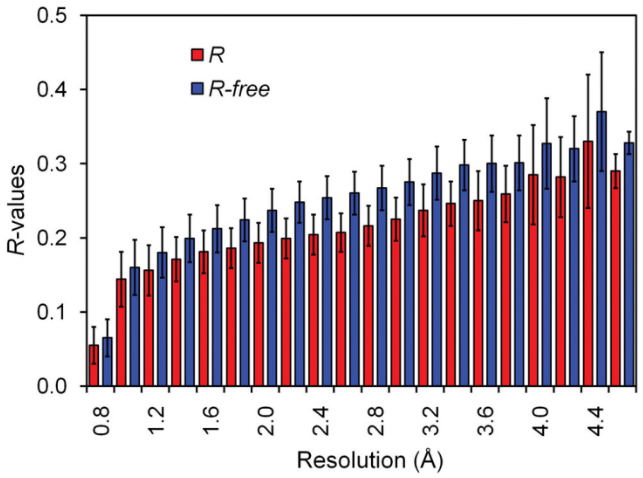


© Garland Science 2010

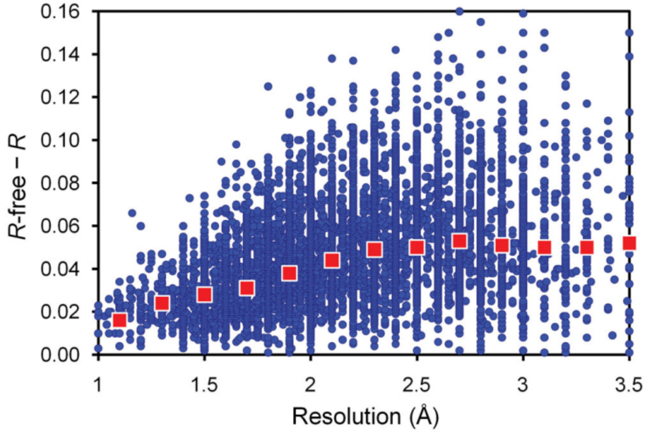
expected ratio of water molecules to protein residues : subtract the resolution (in Å) from 3. This indicator could be higher (by up to 100%) for crystal structures with a high solvent content (Matthews coefficient > 3.0 $Å^3 \cdot Da^{-1}$)...or lower as B_{ave} gets higher...

2- Crystallographic R factors

$$R = \frac{\sum\limits_{reflections} \left| F_{\mathrm{OBS}} - \left| F_{\mathrm{MODEL}} \right| \right|}{\sum\limits_{reflections} F_{\mathrm{OBS}}}$$



© Garland Science 2010



© Garland Science 2010

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%

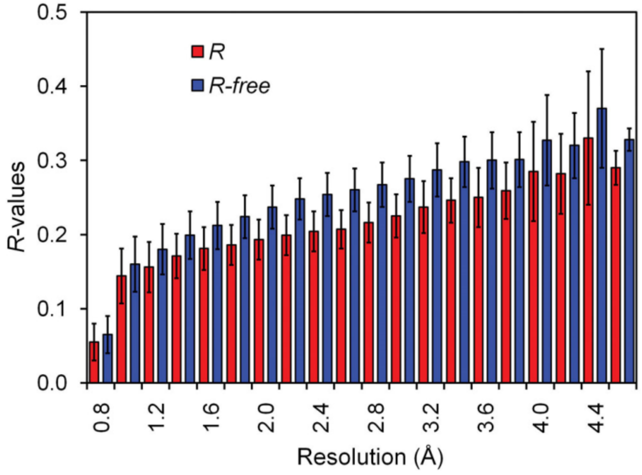
Sometimes the R-factor looks very good (you would expect a good model) but the model-to-map fit is terrible... Overfitting!!

Crystallographic R factors

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}$$

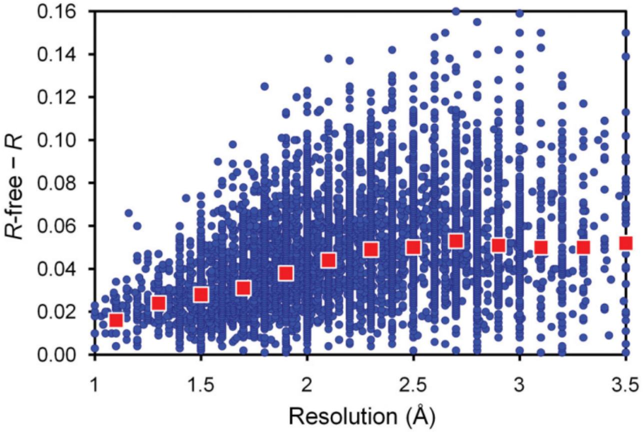
Before refinement, Fobs are divided into a working and a 'free' set.

- The free set should not relate with the working set via symmetry-related reflections.
- R_{work}: R calculated on Fo's exposed to refinement.
- R_{free}: R calculated on Fo's free of refinement.
- $R_{free} > R_{work}$: is problematic if difference is large.



© Garland Science 2010

Resolution-dependence of R_{free} , R_{work} and difference



© Garland Science 2010

R-factors increase in higher resolution shells

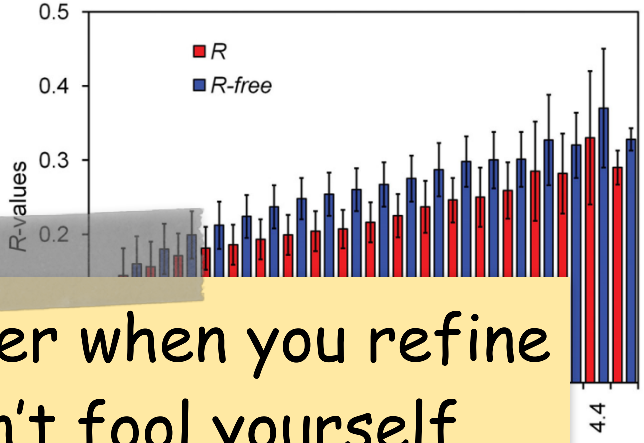
- Greater detail to fit and higher chance of not getting it right
- High R-factor at low resolution: is bulk solvent model correct?

Crystallographic R factors

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}$$

Before refinement, Fobs are divided into a working and a 'free' set.

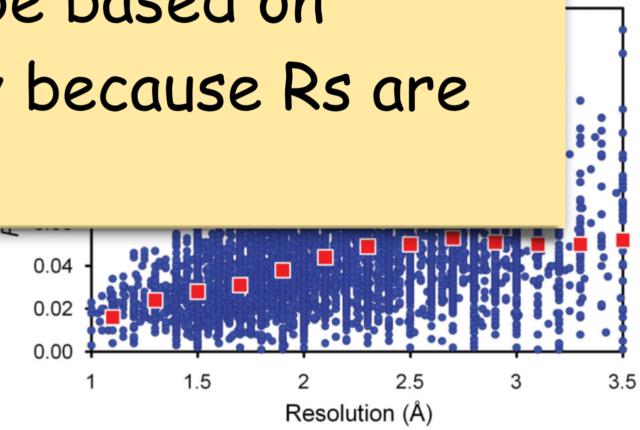
- The free set should not relate with the working set via symmetry-related reflections.
- R_{work}: R calculated on Fo's exposed to refinement.
- R_{free}: R calculated on F
- $R_{free} > R_{work}$: is problem large.



Resolution-dependence of R_{free}

R-factors increase in higher re

- Greater detail to fit and higher chance of not getting it right
- High R-factor at low resolution: is bulk solvent model correct?



© Garland Science 2010

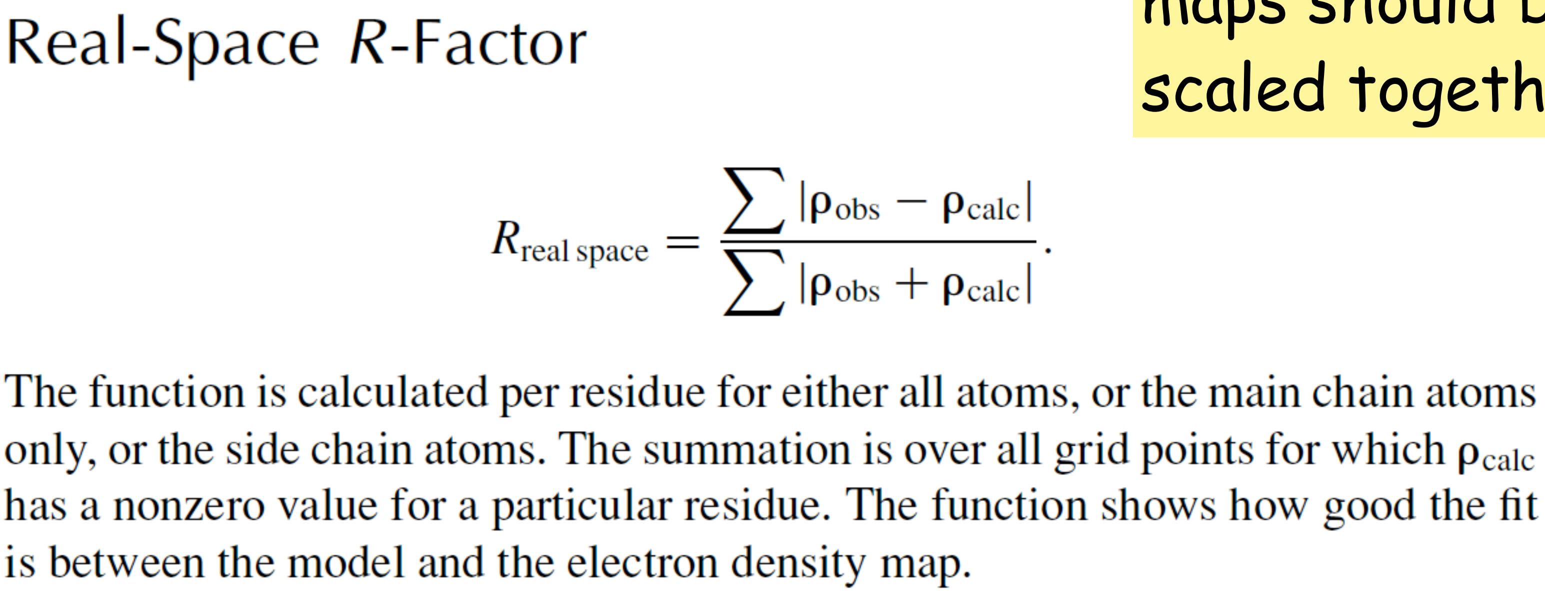
Rs are always lower when you refine with twinning: don't fool yourself… (twinning should be based on evidence, not only because Rs are lower)

3- Electron density-based model validation

Importance of depositing structure factors!! (…and raw diffraction images : SBGrid Data Bank https://data.sbgrid.org)

Real-space R values (RSR) and real-space correlation coefficients (RSCC)

maps should be scaled together!



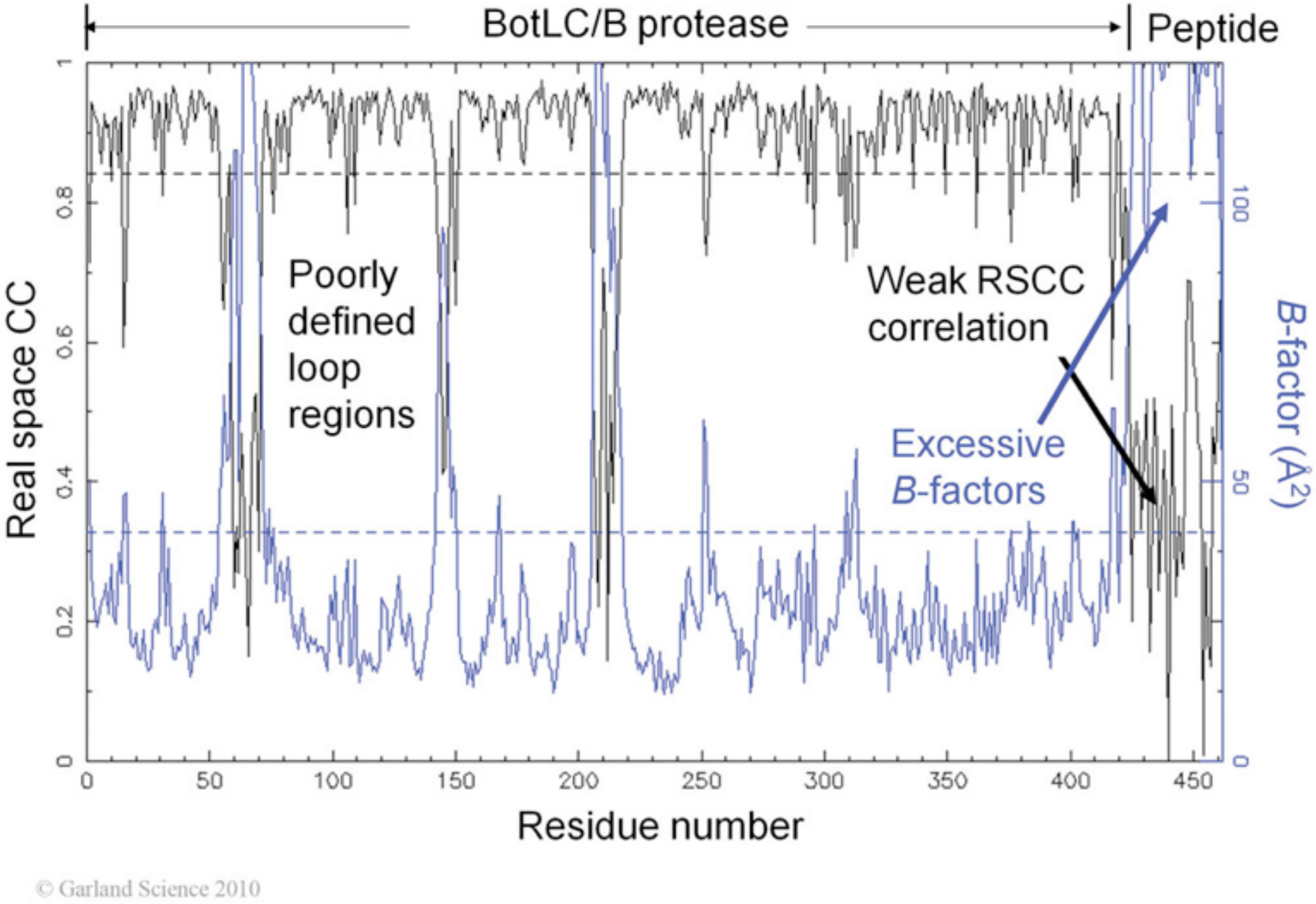
Real-Space *R*-Factor

$$R_{\text{real space}} = \frac{\sum |\rho_{\text{obs}} - \rho_{\text{calc}}|}{\sum |\rho_{\text{obs}} + \rho_{\text{calc}}|}.$$

The function is calculated per residue for either all atoms, or the main chain atoms only, or the side chain atoms. The summation is over all grid points for which ρ_{calc} has a nonzero value for a particular residue. The function shows how good the fit is between the model and the electron density map.

Standard Linear Correlation Coefficient Between Two Electron Density Maps, $\rho_1(xyz)$ and $\rho_2(xyz)$

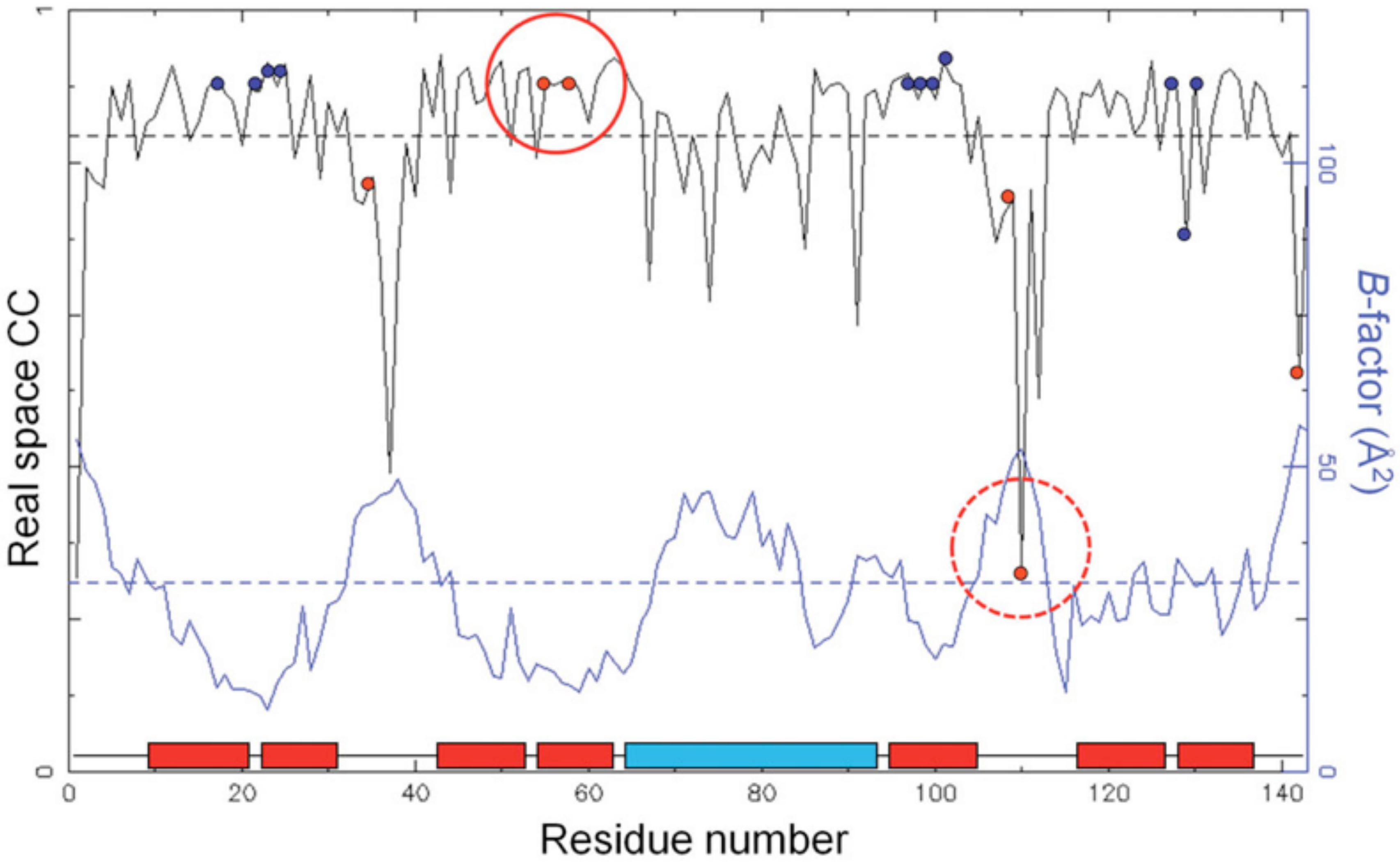
$$C = \frac{\Sigma(\rho_1(xyz) - \overline{\rho_1(xyz)}) \times (\rho_2(xyz) - \overline{\rho_2(xyz)})}{\left[\Sigma(\rho_1(xyz) - \overline{\rho_1(xyz)})^2 \times \Sigma(\rho_2(xyz) - \overline{\rho_2(xyz)})^2\right]^{1/2}}.$$



adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010

CCP4 Overlapmap, SFCheck, edstats; validation tools in Phenix

EDS web server
http://eds.bmc.uu.se/eds/
…switching to PDBe !!



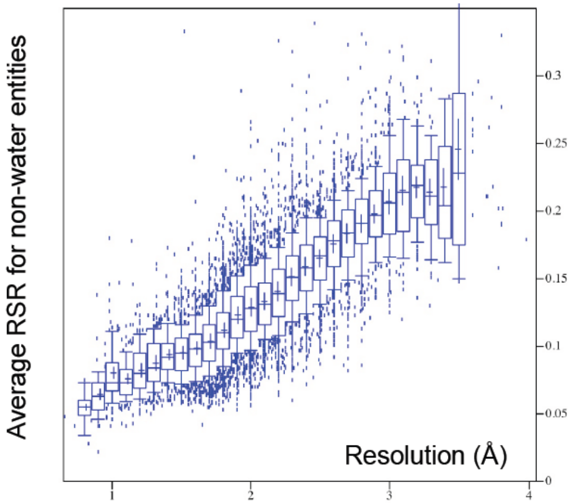
© Garland Science 2010

Red dots = Ramachandran outliers

Blue dots = xtal contacts

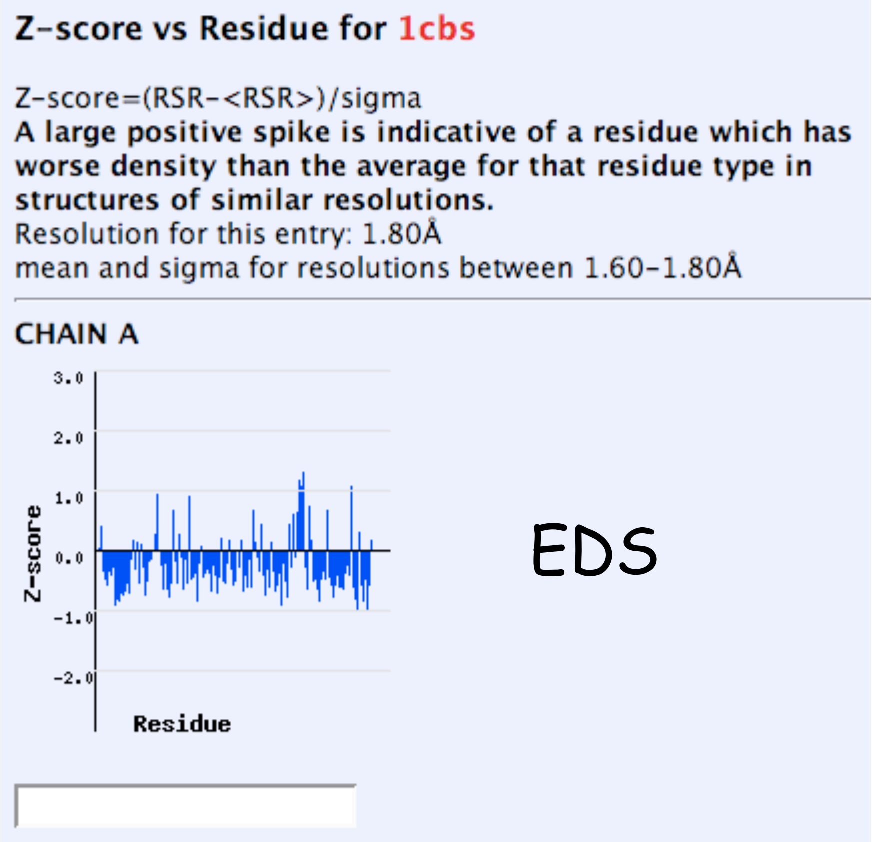
adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010

Maps, RSR, RSCC



(Data from ~14,000 EDS entries, December, 2005)

- RSR is dependent on residue type
  - Different flexibility and levels of solvent exposure
- RSR depends on resolution
  - Calculated electron density will be poorer at lower resolution
- RSR-Z
  - Brings RSRs of residues on same scale, by removing the effects of resolution and residue type
  - Z(RSR, residue-type, resolution) =
    (RSR - <RSR(aa,d)>) / σ(RSR(aa,d))



Z-score vs Residue for 1cbs

Z-score=(RSR-<RSR>)/sigma

A large positive spike is indicative of a residue which has worse density than the average for that residue type in structures of similar resolutions.

Resolution for this entry: 1.80Å

mean and sigma for resolutions between 1.60-1.80Å

CHAIN A

EDS

4- Validation of protein-ligand complexes

Extremely important (and exquisitely linked to local indicators!!)

Use of automated (more objective) algorithms, such as ARP/wARP and others

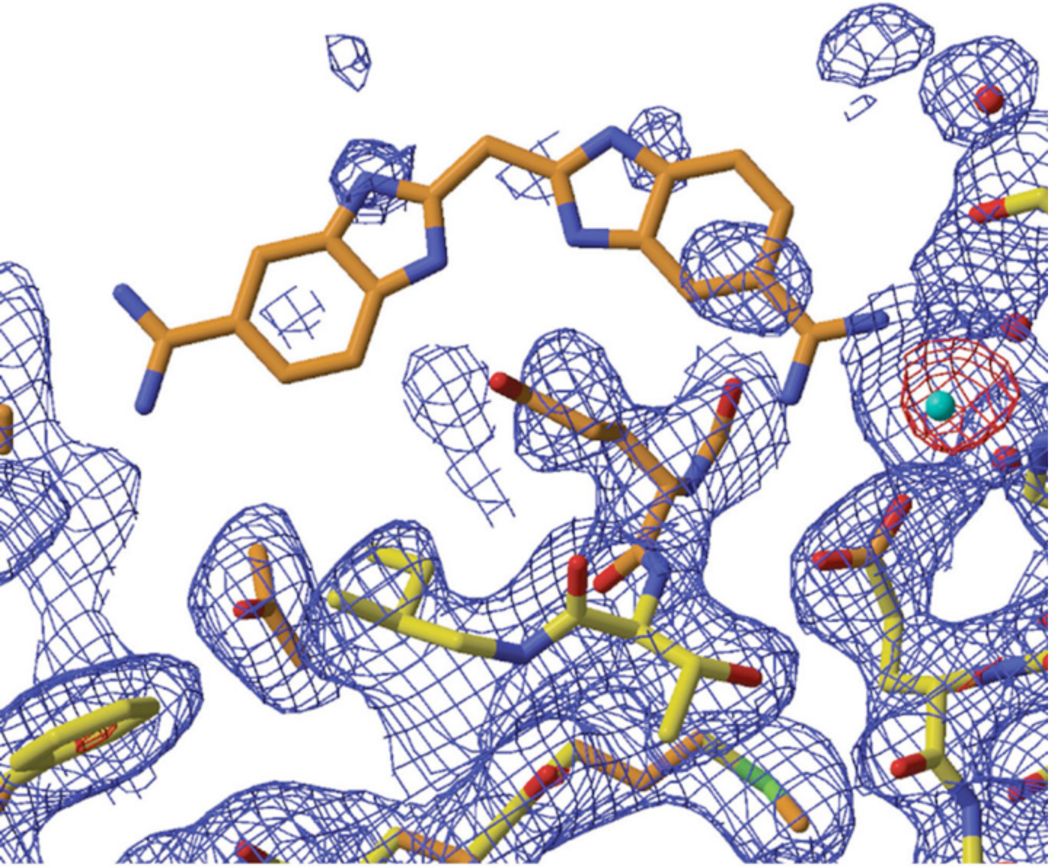
Look at the electron density!!!

Occupancy and B-factor adjustment

Generating (or revising) proper ligand stereochemical restraints (Jligand/Acedrg, Grade/Mogul, etc)

Chemical plausibility and binding pocket analysis (Ligplot, electrostatic potential mapping on surface APBS, etc)

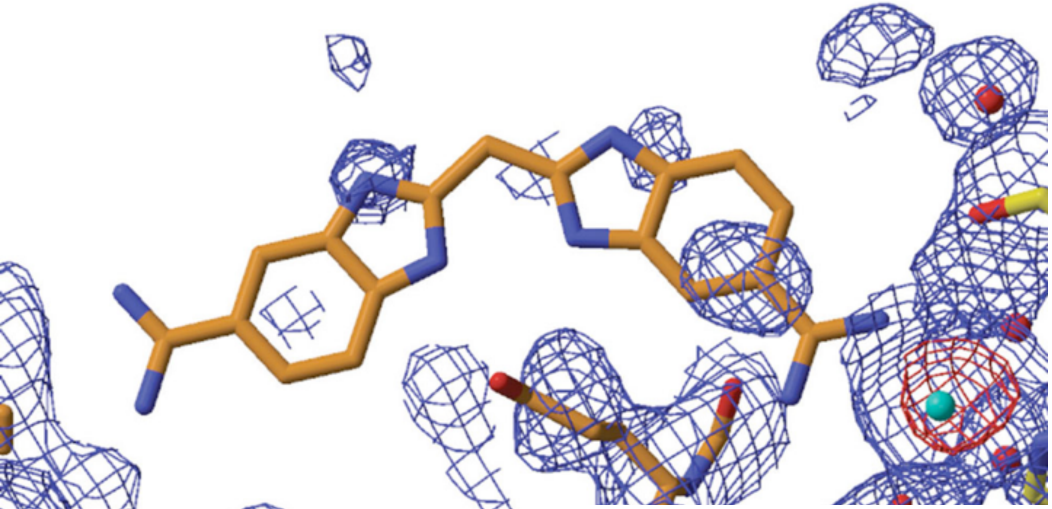
adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010



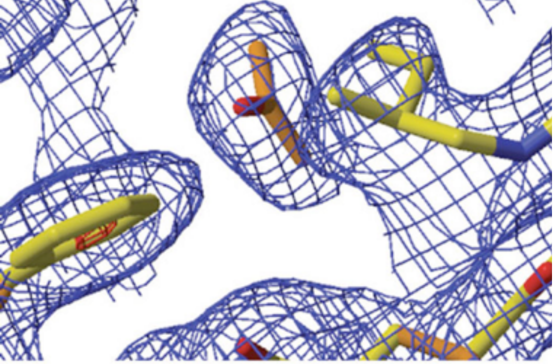
© Garland Science 2010

Calculate omit maps, and compare B factors of ligand and surrounding atoms

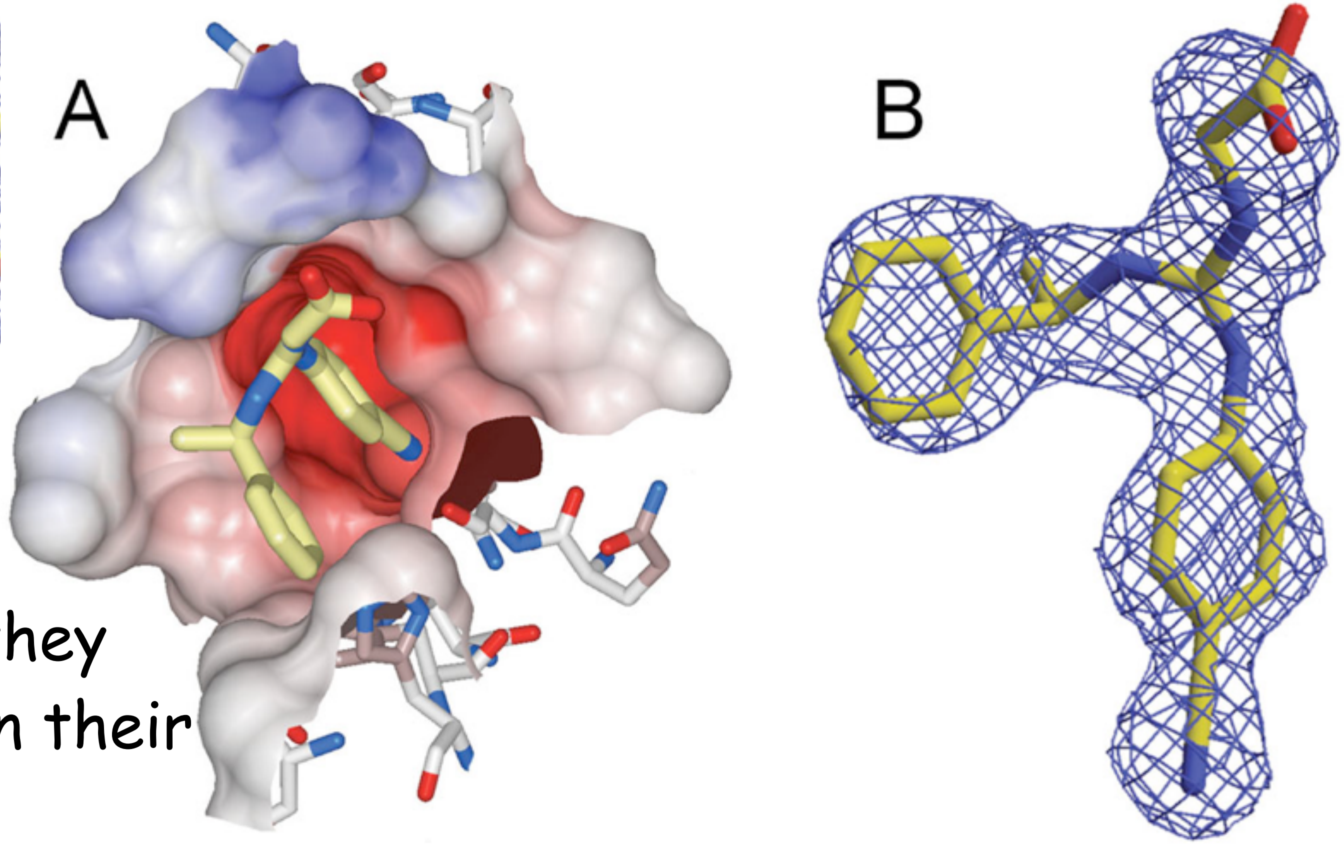
adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010



Calculate omit maps, and compare B factors of ligand and surrounding atoms

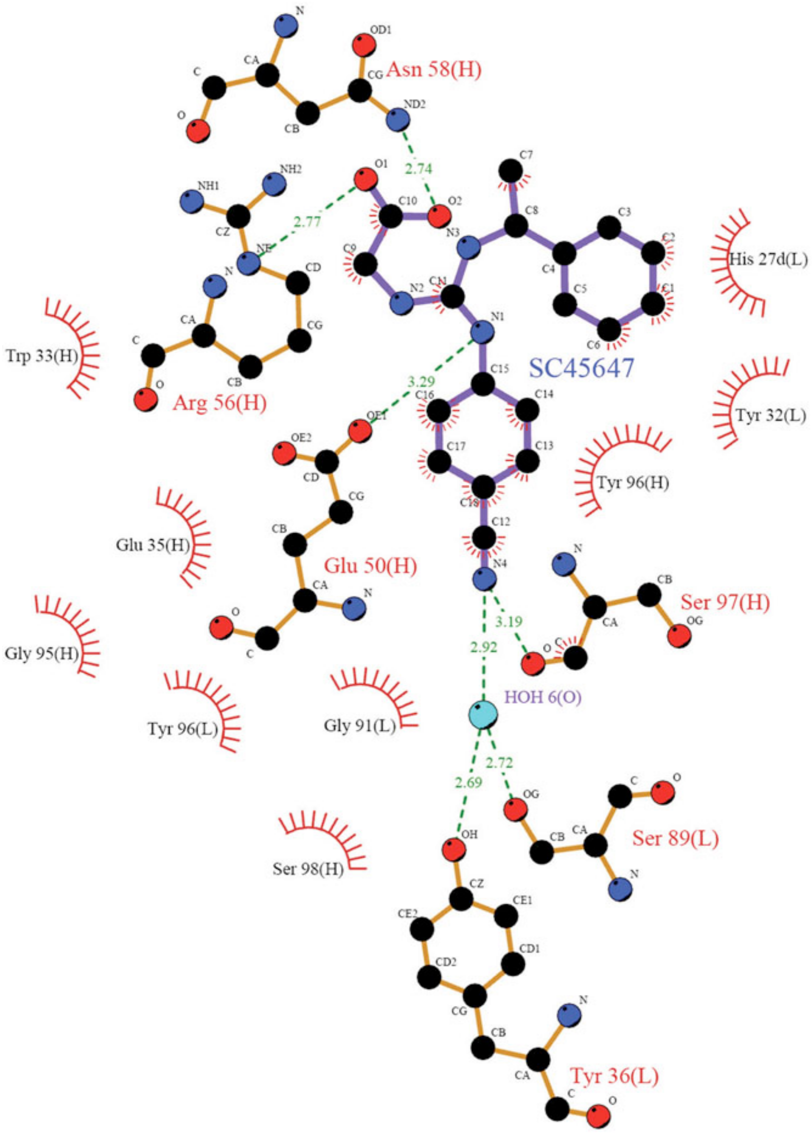


© Garland Science 2010

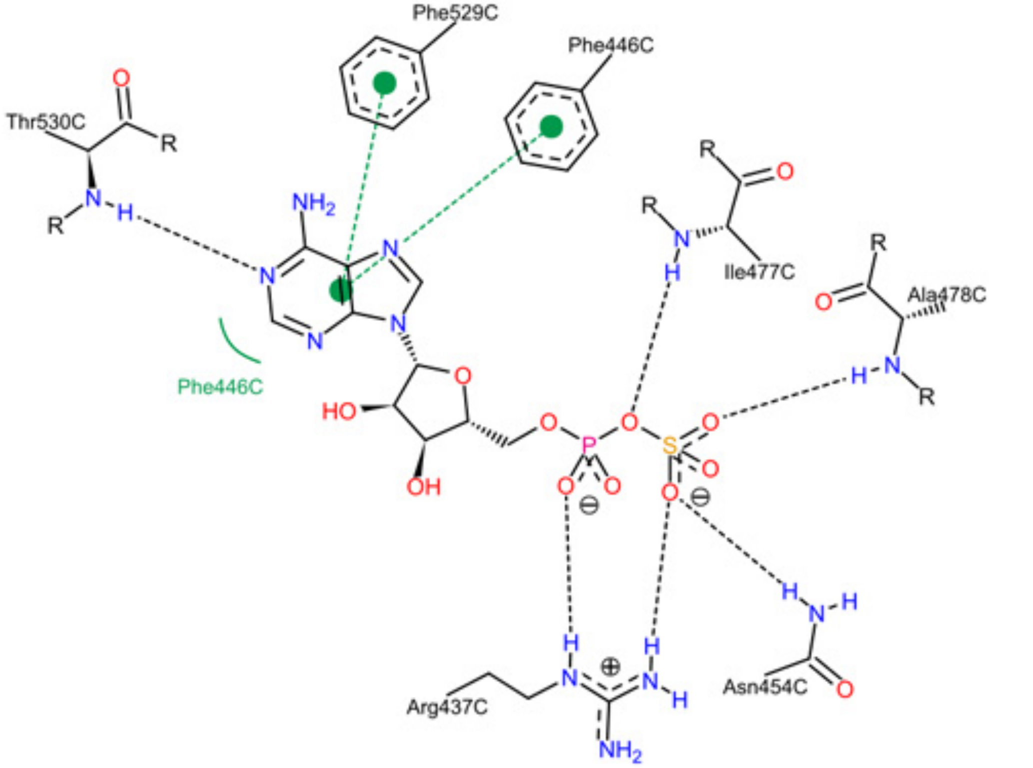


analyze how well they fit/interact within their binding pocket...

2D-sketches of interactions are extremely useful (Ligplot, PoseView, etc)



© Garland Science 2010



5- B factor or atomic displacement parameter

$$F_{(h)} = \Sigma\ f_i\ \exp^{(2\pi i h.x_i)}\ \exp^{(-B\ \sin^2\theta\ /\ \lambda 2)}$$

Higher B factors imply faster decay in scattering intensity with resolution (i.e. atoms with higher B factors contribute less to higher resolution reflections)

$$B_i = 8\ \pi^2\ \langle U_i \rangle^2$$

B = 20 => <U> = 0.5Å,
B = 50 => <U> = 0.8Å,
B = 100 => <U> = 1.13Å,
B = 200 => <U> = 1.6Å

<U> = average RMS displacement of the atom, uncertainity in coordinates

Can be modelled as an anisotropic ellipsoid (using 6 parameters instead of 1 isotropic)

B factor or atomic displacement parameter

Although one has to be cautious with overinterpretation (B factors can become “error sinks”), they do provide valuable information on atom displacement (electron density spread)

B factor or atomic displacement parameter

Although one has to be cautious with overinterpretation (B factors can become “error sinks”), they do provide valuable information on atom displacement (electron density spread)

Reasons behind the “error sink” role:

Refinement increases B factor to explain the absence of strong density…maybe occupancy is low!

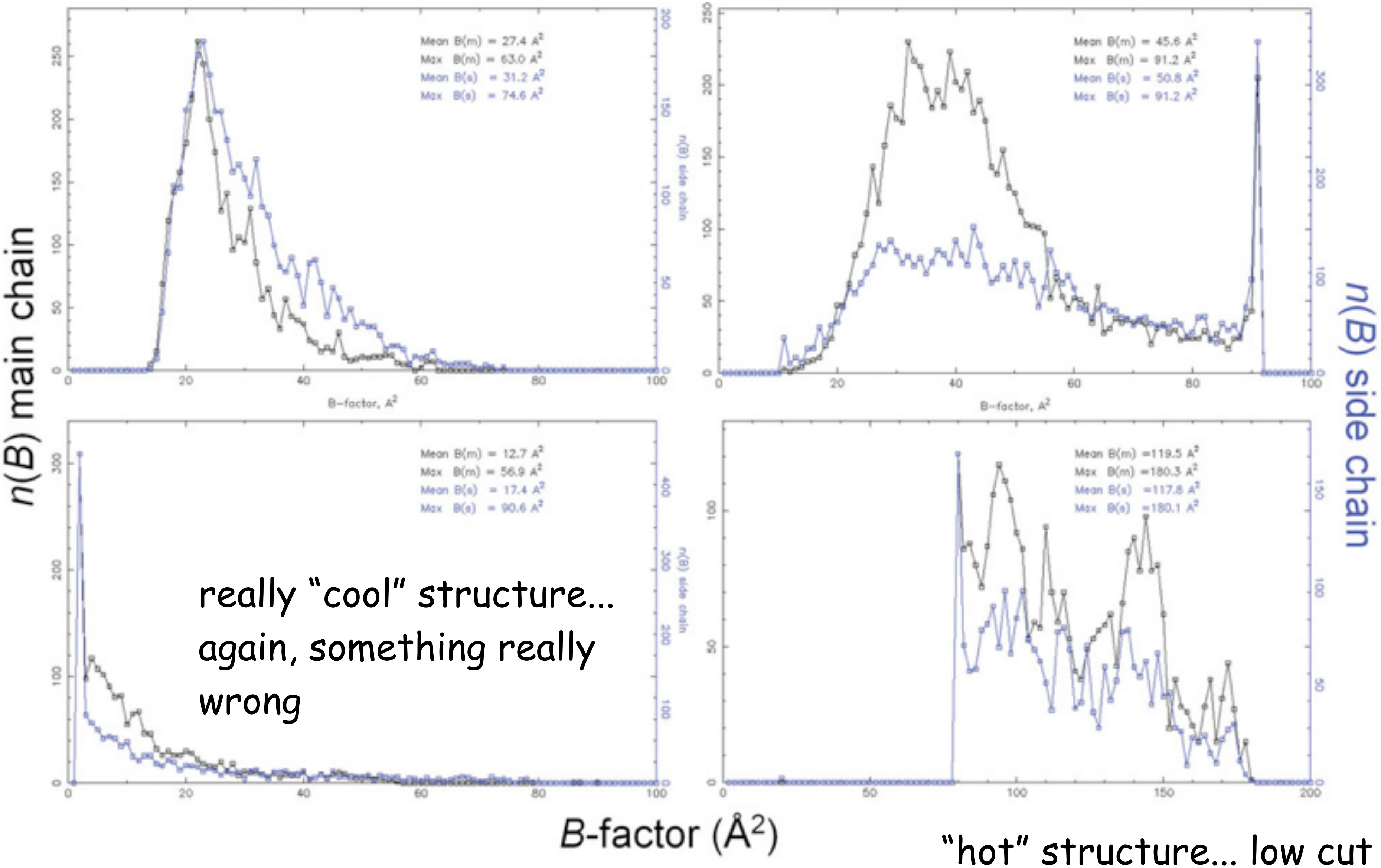
…or wrong conformation, non-existent molecules, wrong atomtype

Could be static disorder with not well defined alternate conformations

When corresponding atoms don’t obey strict NCS, this can lead to high B

…it is thus essential to look at B factor distributions typical distribution

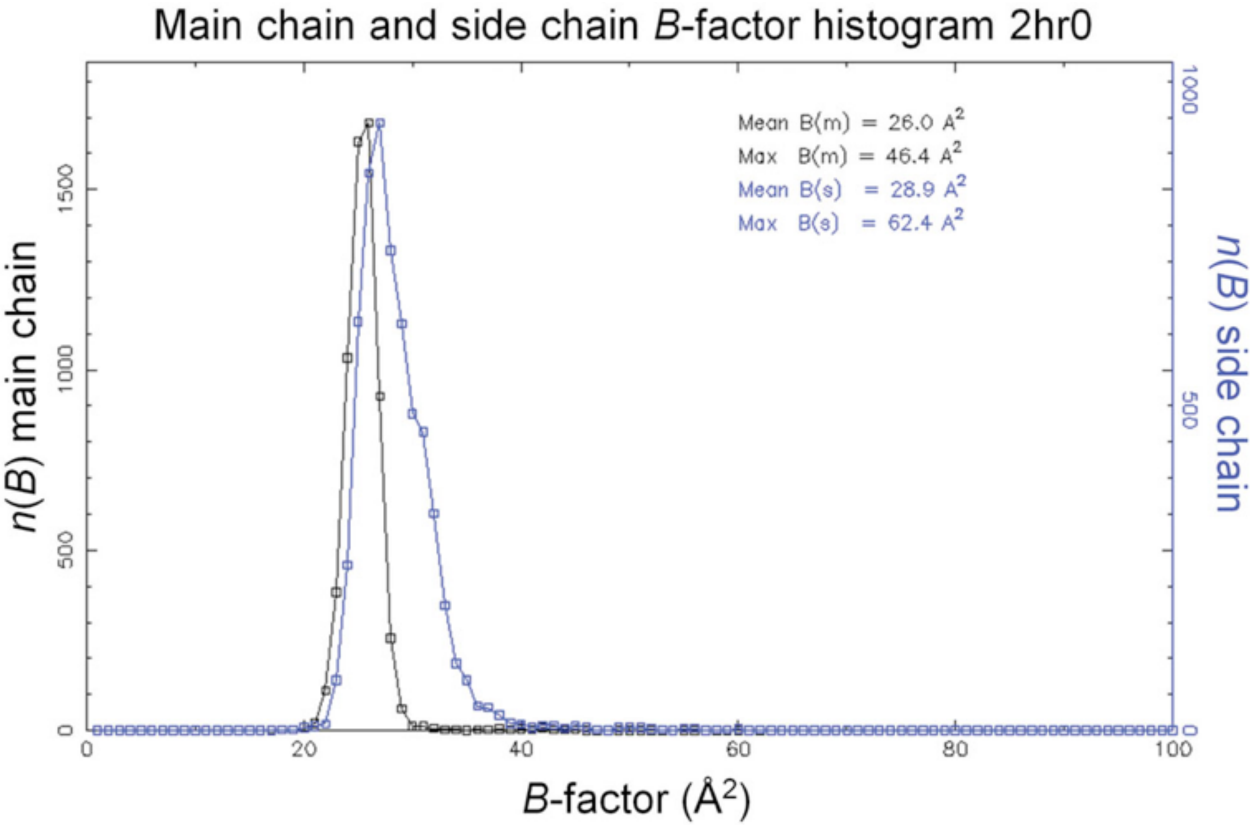
wrong strategy: high B cut-off at 92Å2
weird behavior mc/sc



really “cool” structure...
again, something really
wrong

“hot” structure... low cut
off? (it’s a 3.9Å res)

adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010



© Garland Science 2010

...or yet too tight restraints may lead to unusually sharp distributions

adapted rom “Biomolecular Crystallography” Bernhard Rupp, Garland Science 2010

SUMMARY

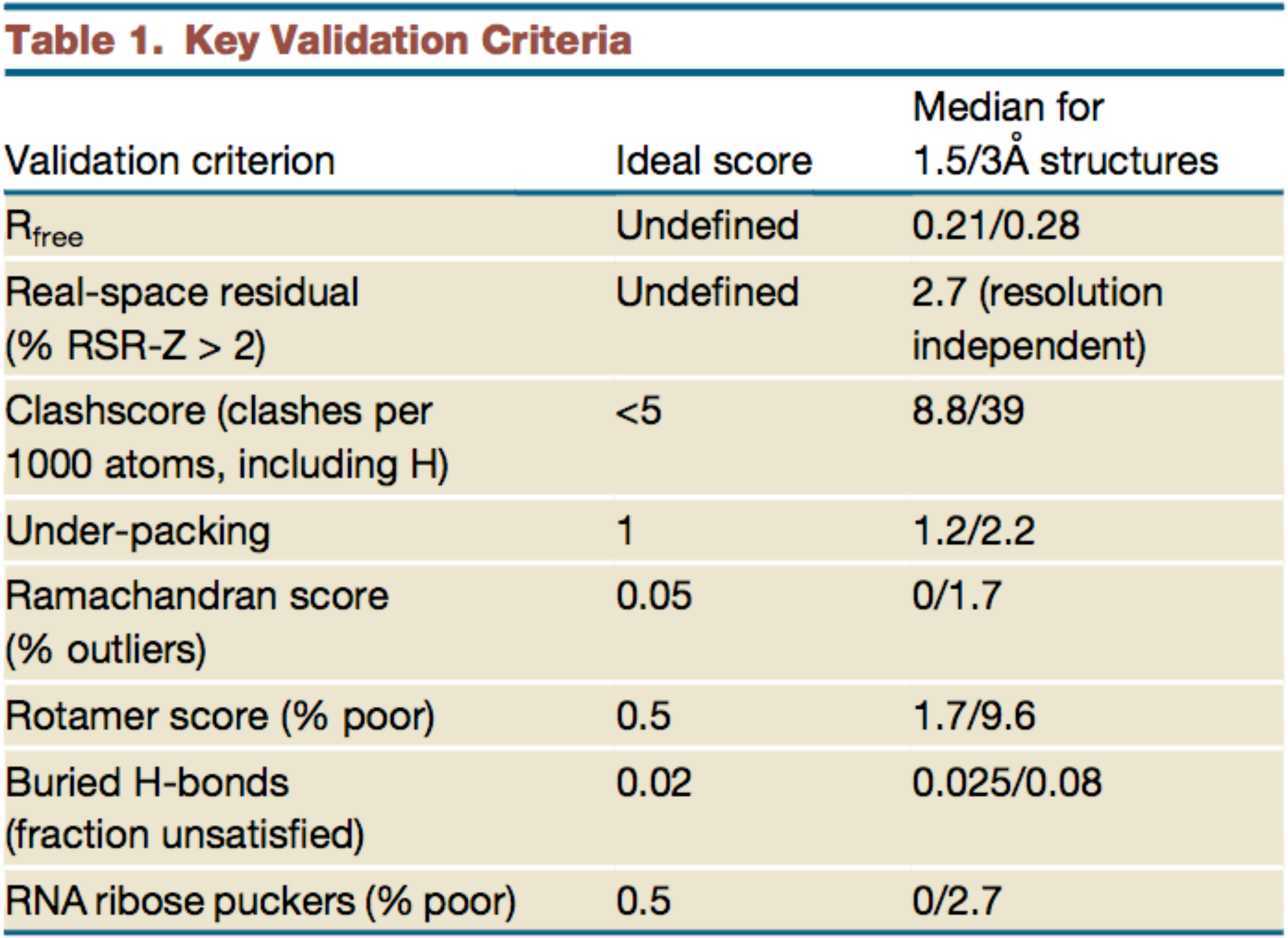


Table 1. Key Validation Criteria

Validation criterion	Ideal score	Median for 1.5/3Å structures
R_{free}	Undefined	0.21/0.28
Real-space residual (% RSR-Z > 2)	Undefined	2.7 (resolution independent)
Clashscore (clashes per 1000 atoms, including H)	<5	8.8/39
Under-packing	1	1.2/2.2
Ramachandran score (% outliers)	0.05	0/1.7
Rotamer score (% poor)	0.5	1.7/9.6
Buried H-bonds (fraction unsatisfied)	0.02	0.025/0.08
RNA ribose puckers (% poor)	0.5	0/2.7

Some important messages...

✓A good model makes sense from all perspectives

chemical, physical, structural, crystallographic, statistical, biological

✓Mistakes can always happen! but, this emphasizes the need to perform careful validation of model quality

✓Comparison against other structures of similar resolution and size is useful (red-blue sliders in the PDB and Coot; polygon within phenix GUI : Graphical comparison of statistics versus the PDB)

Some important messages...

✓Special attention should be given to non-standard entities like small molecules, carbohydrates etc.

✓Current criteria and tools catch majority of errors and help building high quality models ; filters: you (maybe rushing), your (often too busy) supervisor and colleagues, up-to-date (& bug-free) software tools

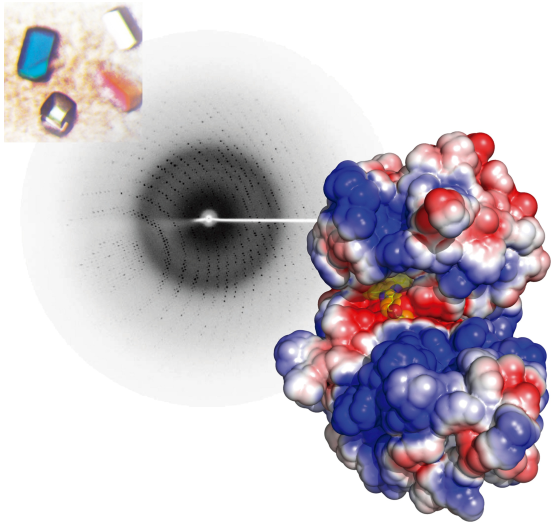
✓Depositing in the PDB (also deposit raw diffraction images!), please deposit unmerged intensities (plus amplitudes): and follow the validation requirements, answer to the PDB annotator!

✓**use PDB-redo** to look at pdb's (often improved models!!)

Institut Pasteur
de Montevideo

Unit of Protein Crystallography

Muchas gracias!!



Macromolecular Crystallography School 2018
November 2018 - São Carlos, Brazil