PROTEIN STRUCTURE VALIDATION AND QUALITY ASSESSMENT

Luigi Vitagliano

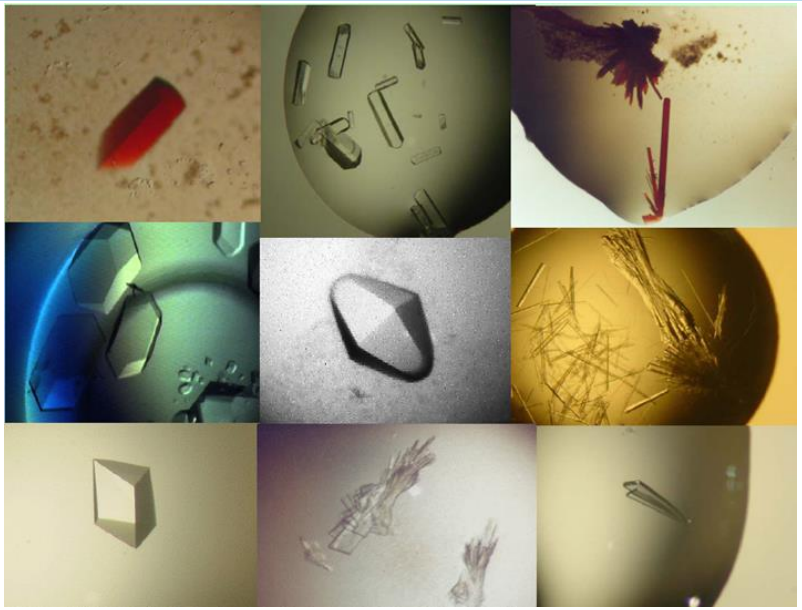
Istituto di Biostrutture e Bioimmagini
Consiglio Nazionale delle Ricerche
Via Mezzocannone 16, Napoli
E-mail: luigi.vitagliano@unina.it

Naples, Italy
29 Aug – 3 Sep 2019

Outline

- Protein crystals: strengths and limitations
- The overfitting problem in protein crystallography
- Major failures in protein crystallography
- Refinement and crystallographic indicators
- Protein structure validation: why and how
- The Protein Data Bank validation report
- From protein structure validation to quality assessment

Protein crystals



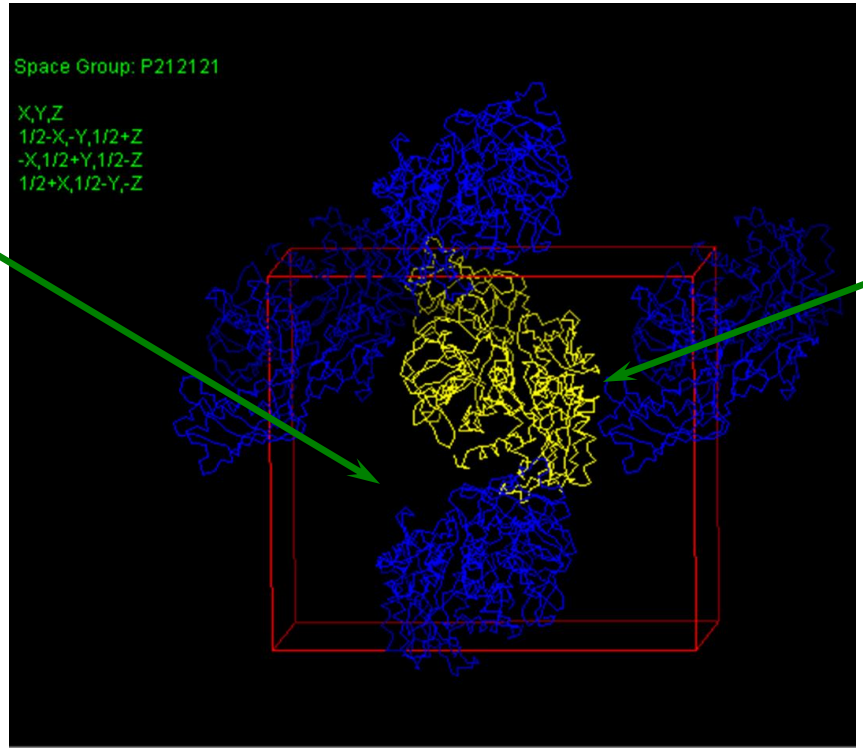
- High solvent (tipically water plus buffer) content (20-80%)
- Large solvent channels
- Limited level of order and size
- High fragility and sensitivity to external conditions

Protein crystals

Positive outcome of protein crystal properties

The crystal packing of the Elongation Factor G

Solvent channels

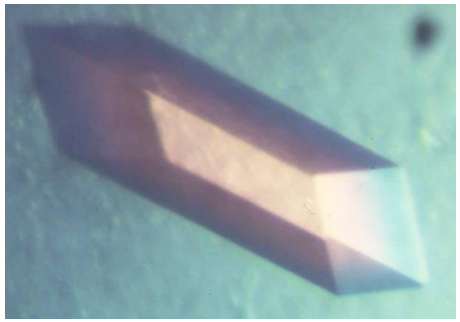


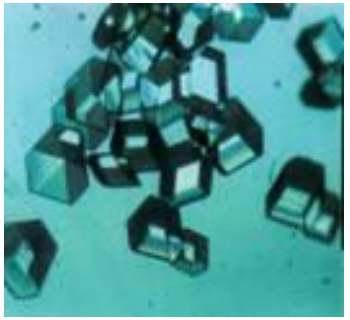
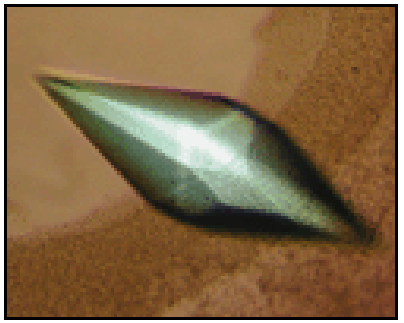
Limited number of intermolecular contacts *per* atom

The protein may be active in the crystal state

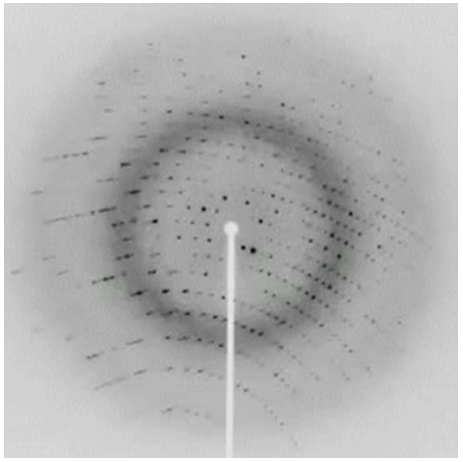
It is possible to diffuse small molecules in the crystals as ligands and inhibitors (soaking)

Protein crystals



The appealing morphology of protein crystals does not warrant high resolution diffraction



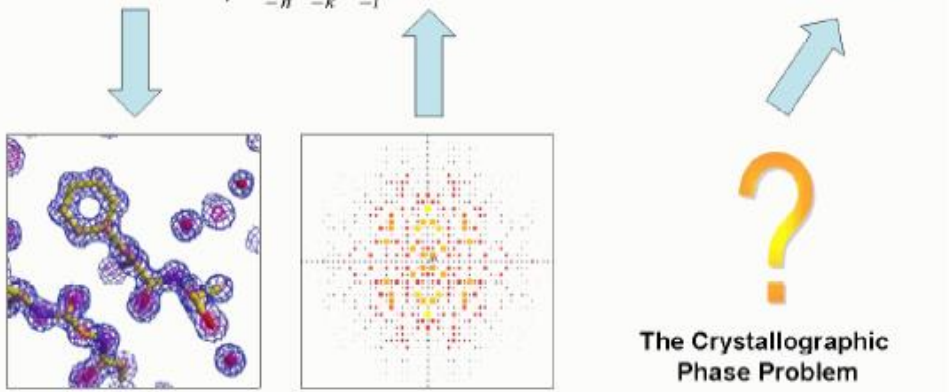
The power of protein crystals is typically rather limited

The data that can be used for structure determination and refinement is also rather limited

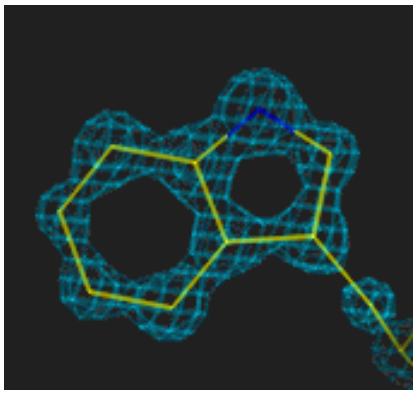
Amount of data and resolution

The number of measured reflections has a major impact on the ensuing electron density and on the accuracy of the final three-dimensional structure

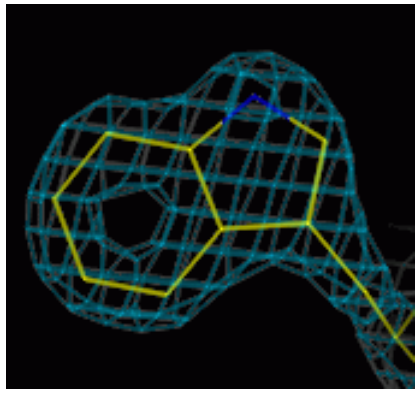
$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l}|F_{hkl}|\cdot\exp[-2\pi i(hx+ky+lz-\alpha_{hkl})]$$



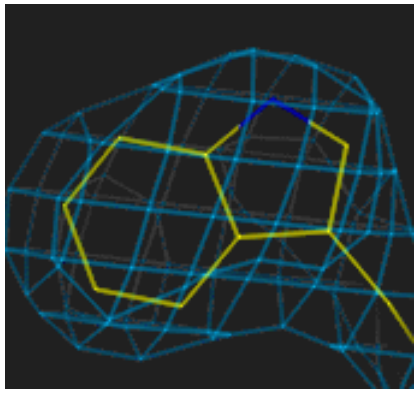
Electron density as function of the resolution



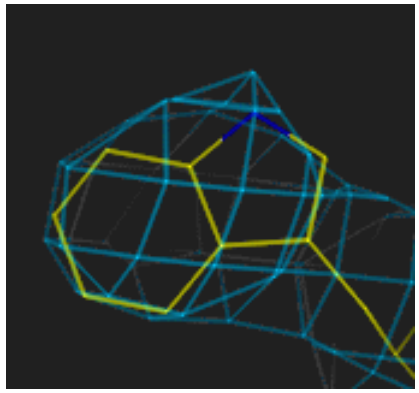
1.0 Å



1.8 Å



2.3 Å



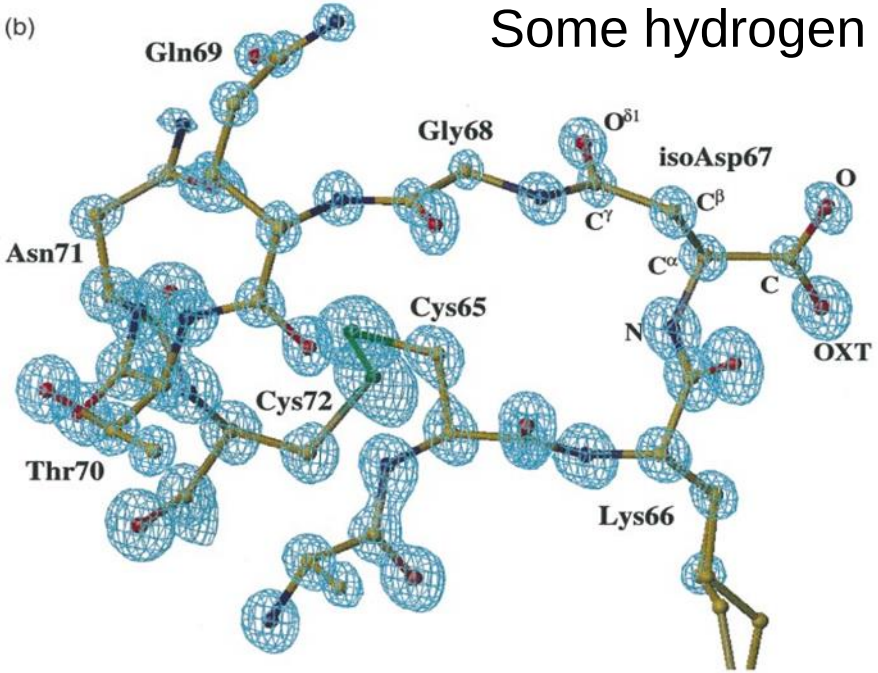
3.0 Å

High resolution structures

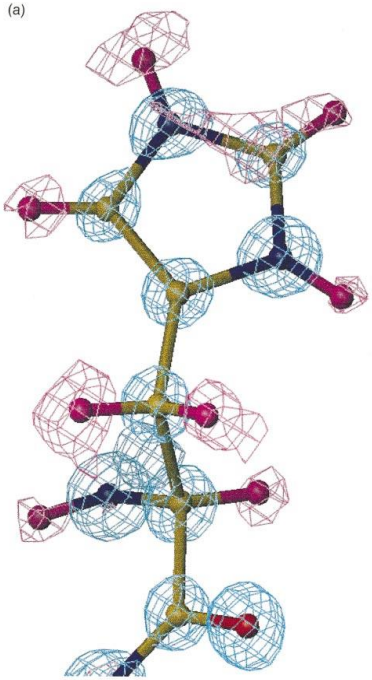
Only very high resolution structures can provide atomic level views

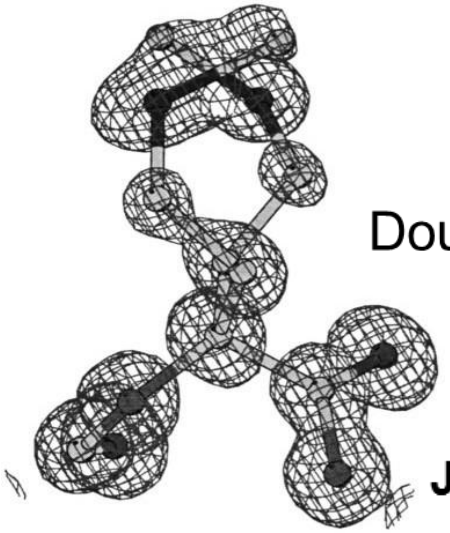
Atomic resolution: better than 1.2 Å

"ultra-high resolution": the example of RNase A 0.87 Å



Some hydrogen atoms can be identified →





Double conformations

Individual atoms can be seen in the density

JMB, 297, 713-732, 2000

Data/parameter ratio

The ratio of reflections used in the refinement to the parameters (the r/p ratio) is important for ensuring precise atomic and geometric parameters. Higher r/p ratios generally improve the precision.

For non-centrosymmetric space groups when only elements lighter than argon are present, it is expected that the r/p ratio should be at least 8:1.

In data collection on small molecules this ratio is generally > 10

Typical case for a protein structure refined at medium/resolution

Crystal structure of papain

Number of spots: 25,000 at 2.0 Å resolution

Number of non-H atoms 2000

x 4 parameters (x, y, z, B) = 8000 parameters

data/parameters = 25,000/8000 ≈ 3

For proteins refined using data at 3 Å the ratio <1 !

Data/parameter ratio

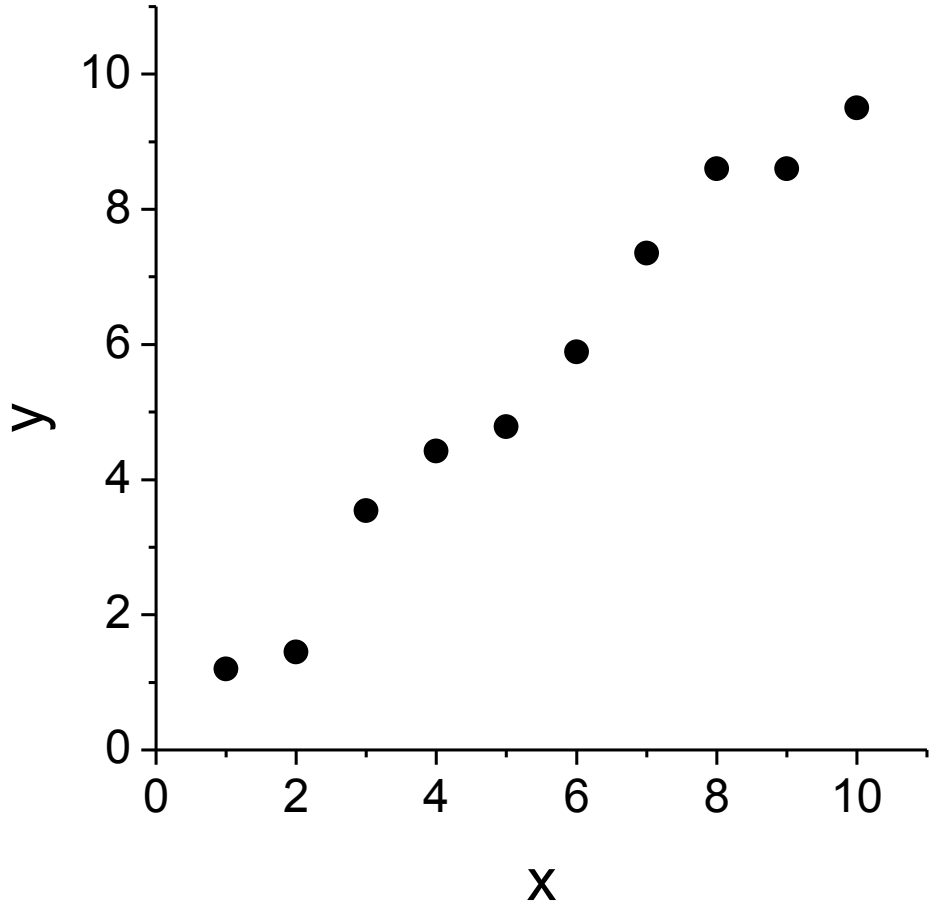
Crystallographic data are largely insufficient for determining the position of the remarkable number of atoms that constitute a protein.

Refinement of protein structures relies on additional information essentially stereochemical data derived from small molecule crystallography and quantum chemistry calculations

The risk of data overfitting

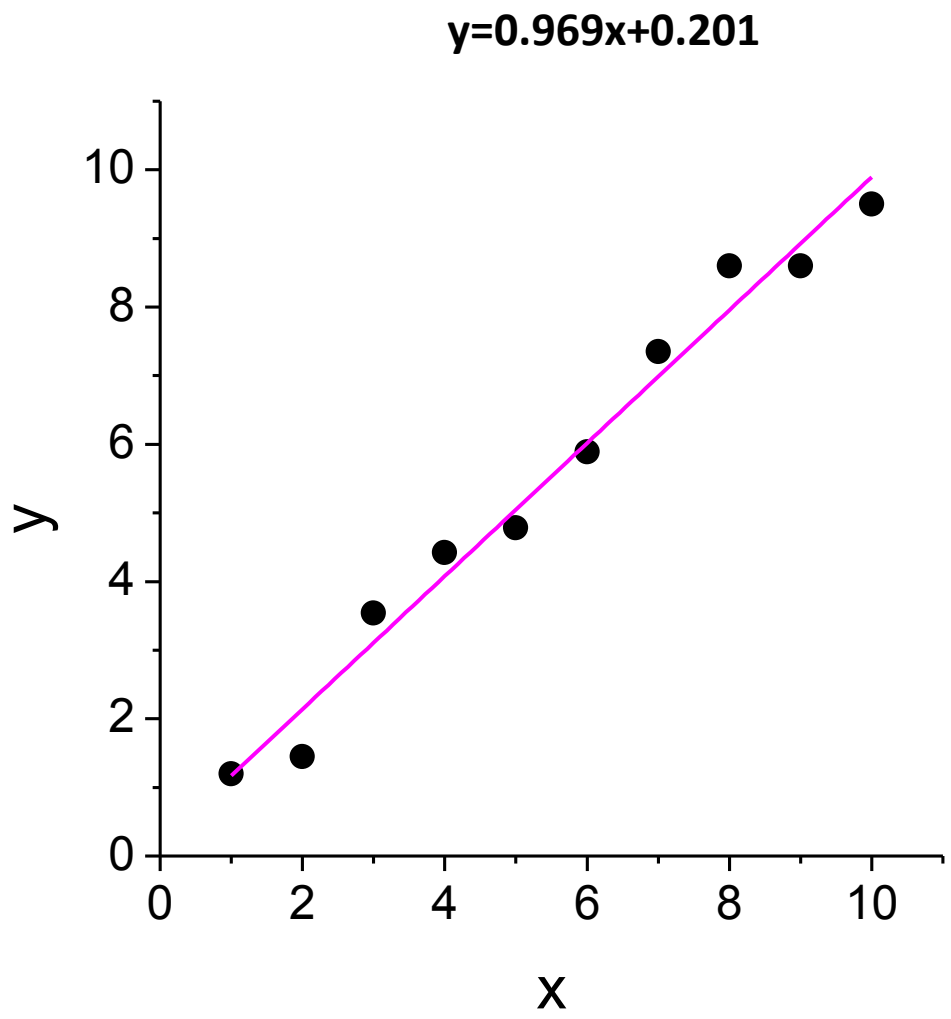
Data fitting and overfitting

An ensemble of data plotted as function of two parameters



Data fitting and overfitting

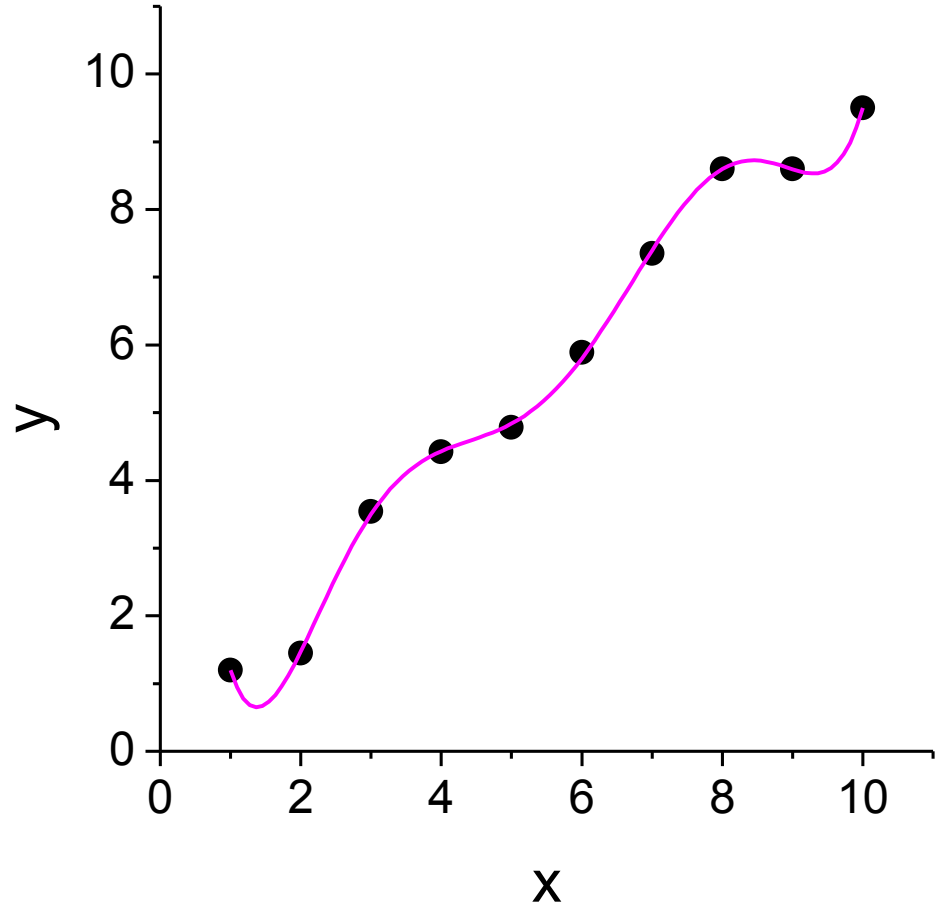
Fitting with a linear function – A low parameter choice



Data fitting and overfitting

Fitting with a polynomial function that contains a larger number of parameters

$$y=-24.829x^6+17.373x^5-5.409x^4+0.850x^3-0.065x^2+0.002x+13.274$$



Data fitting and overfitting

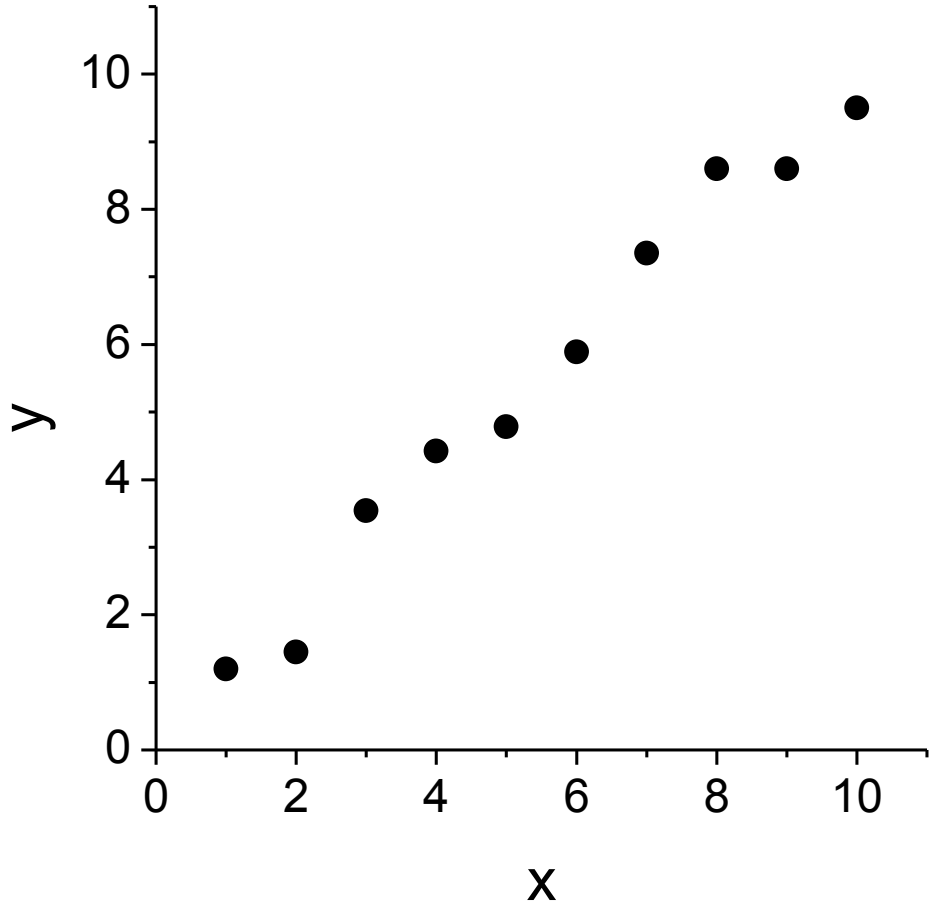
Which is the best fitting?

Best mathematical fitting *versus* the best description of the physical phenomenon

The cross validation approach

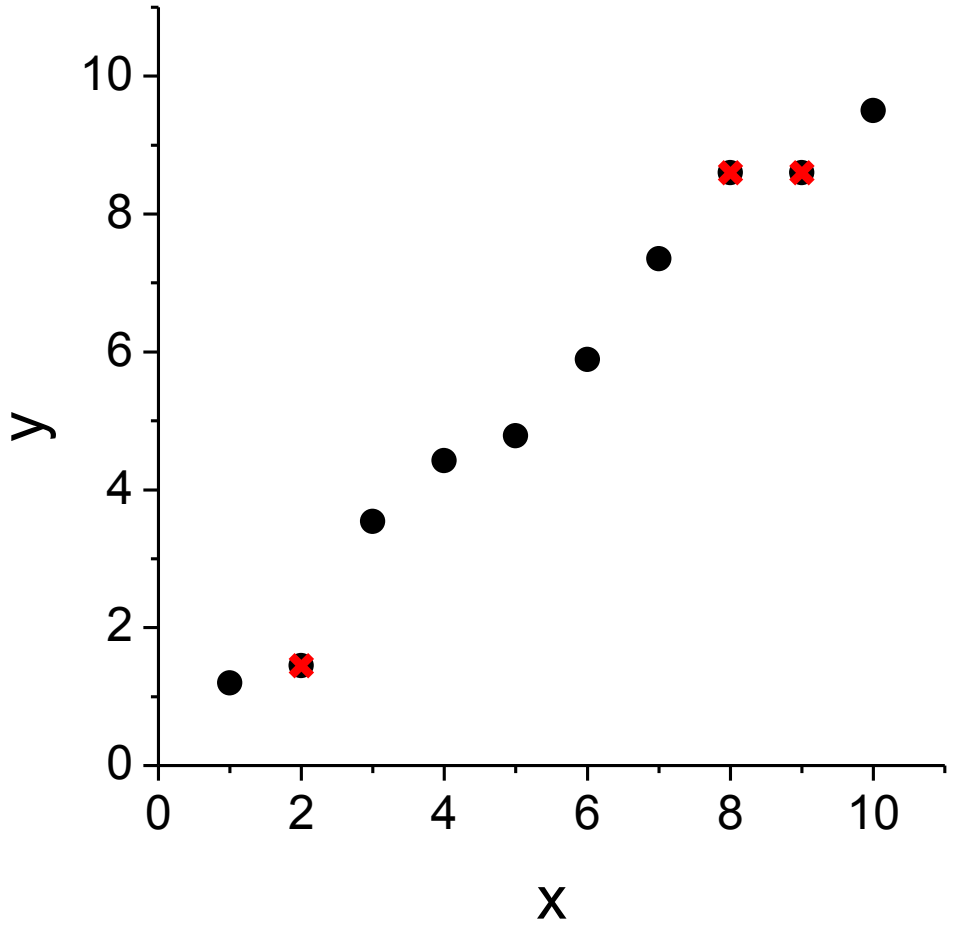
Data fitting and overfitting

Back to the data



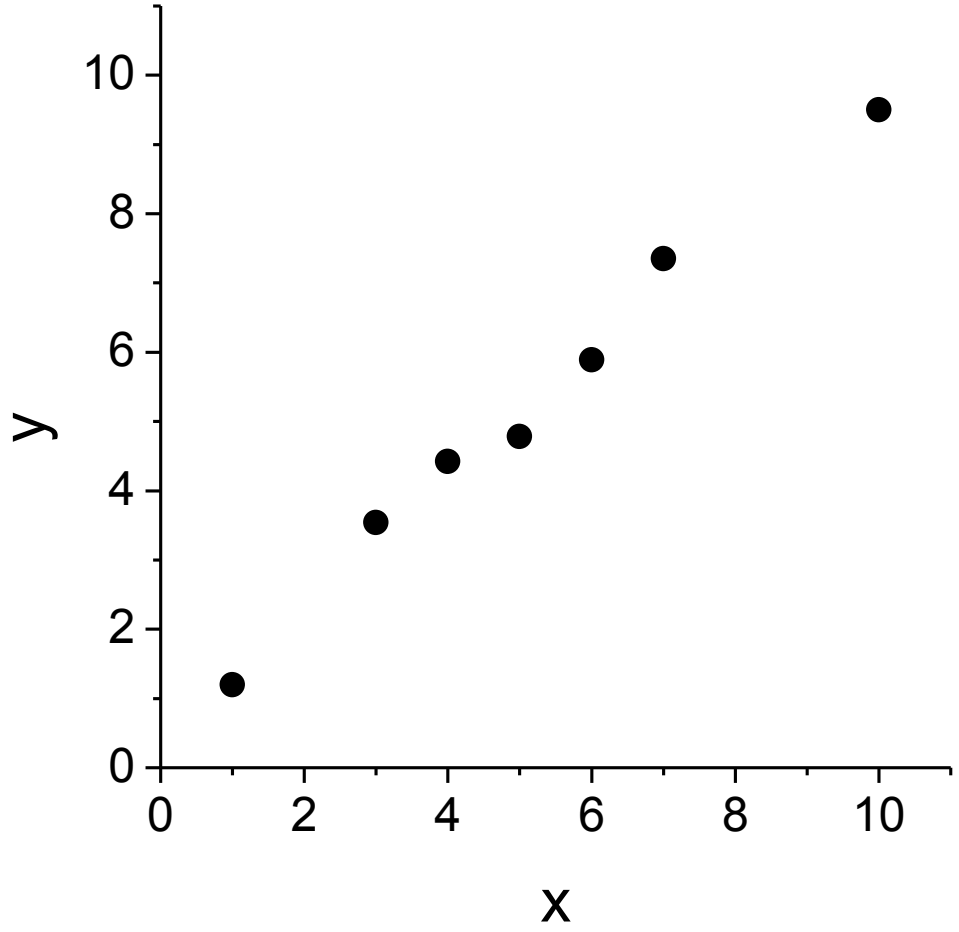
Data fitting and overfitting

Randomly selection of some data

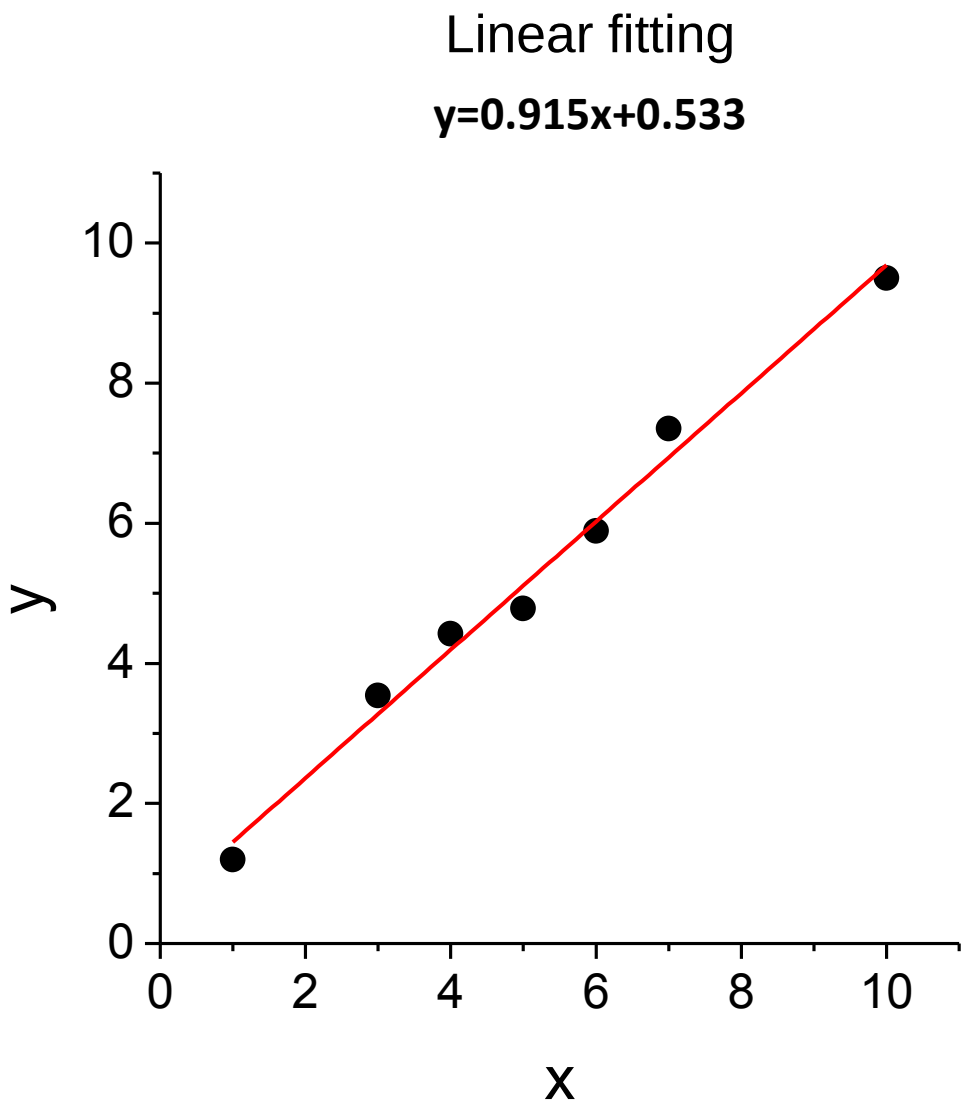


Data fitting and overfitting

Set them apart



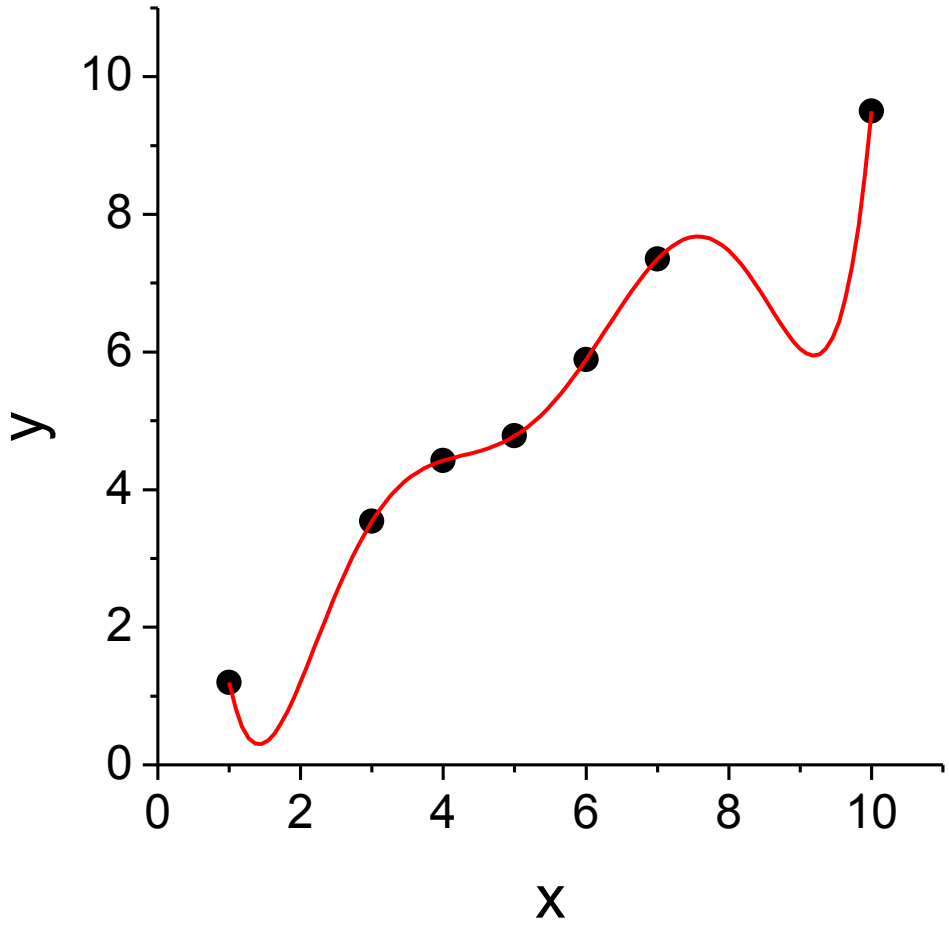
Data fitting and overfitting



Data fitting and overfitting

Polynomial fitting

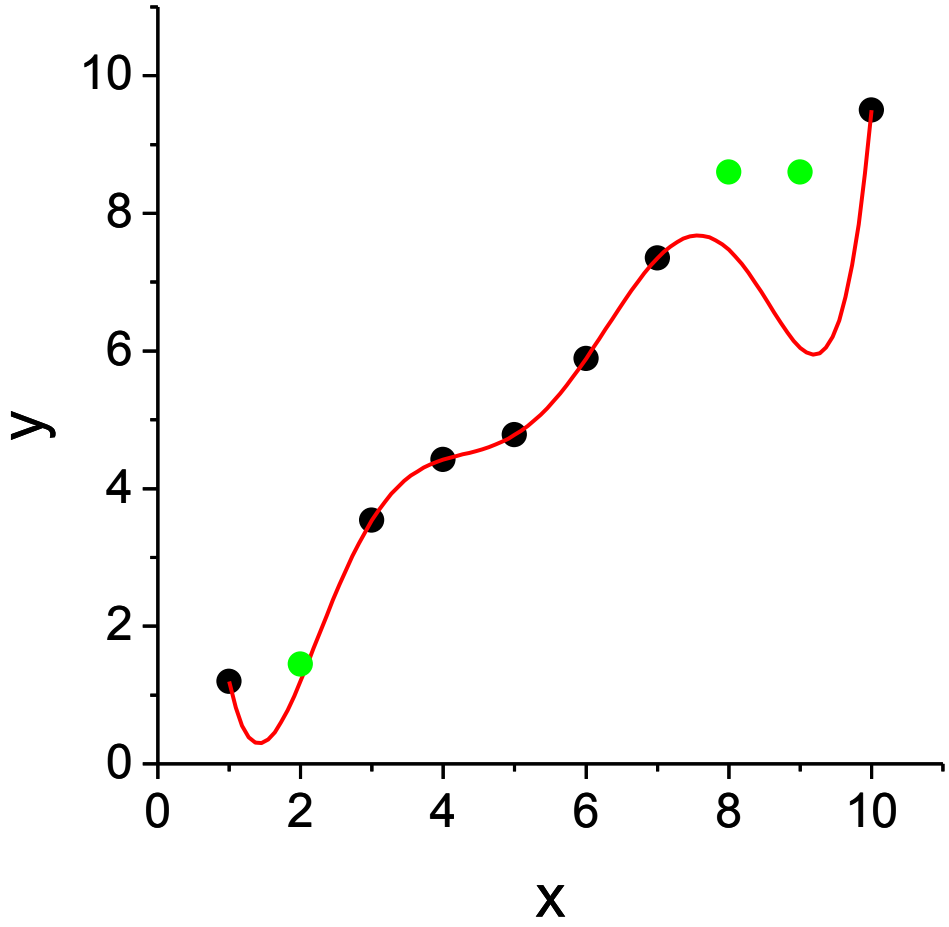
$$y=-36.149x^6+25.455x^5-8.145x^4+1.325x^3-0.106x^2+0.003x+18.816$$



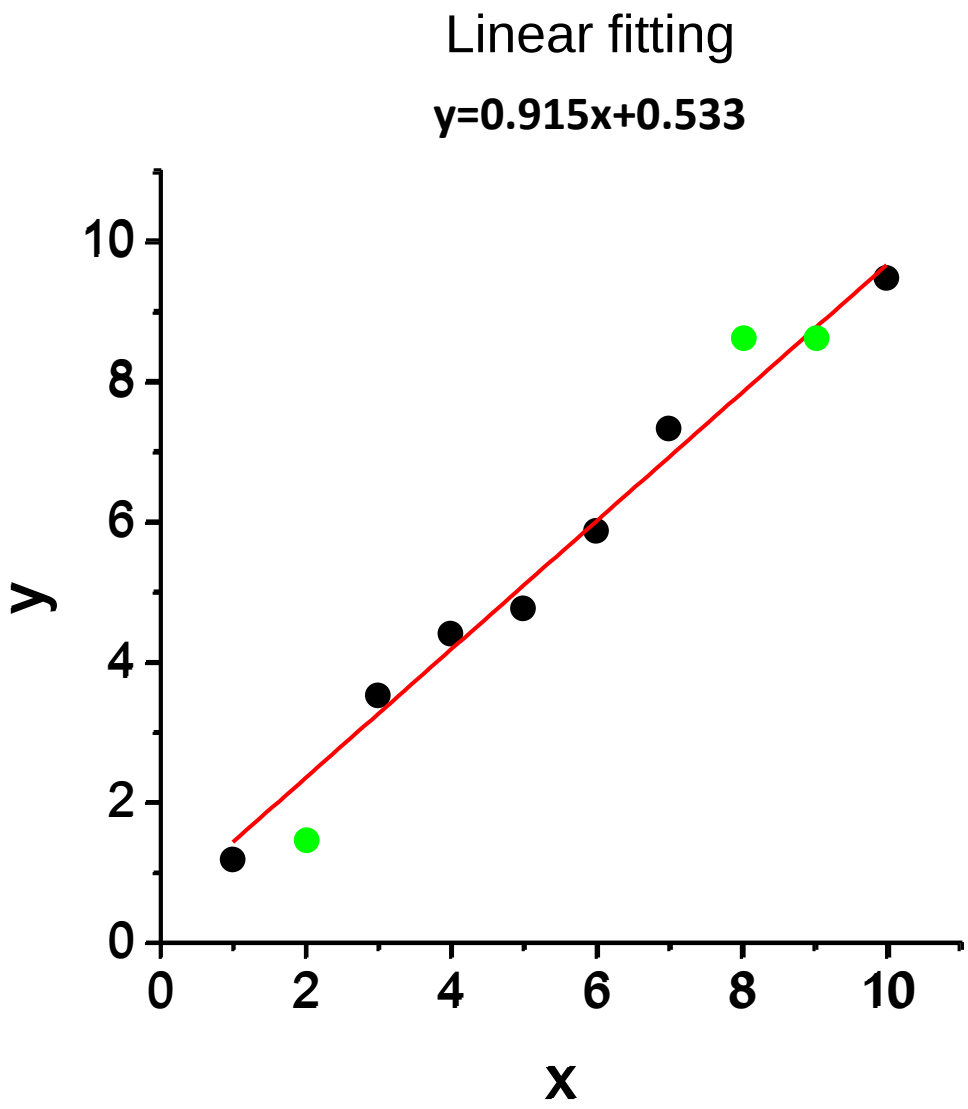
Data fitting and overfitting

Polynomial fitting

$y=-36.149x^6+25.455x^5-8.145x^4+1.325x^3-0.106x^2+0.003x+18.816$

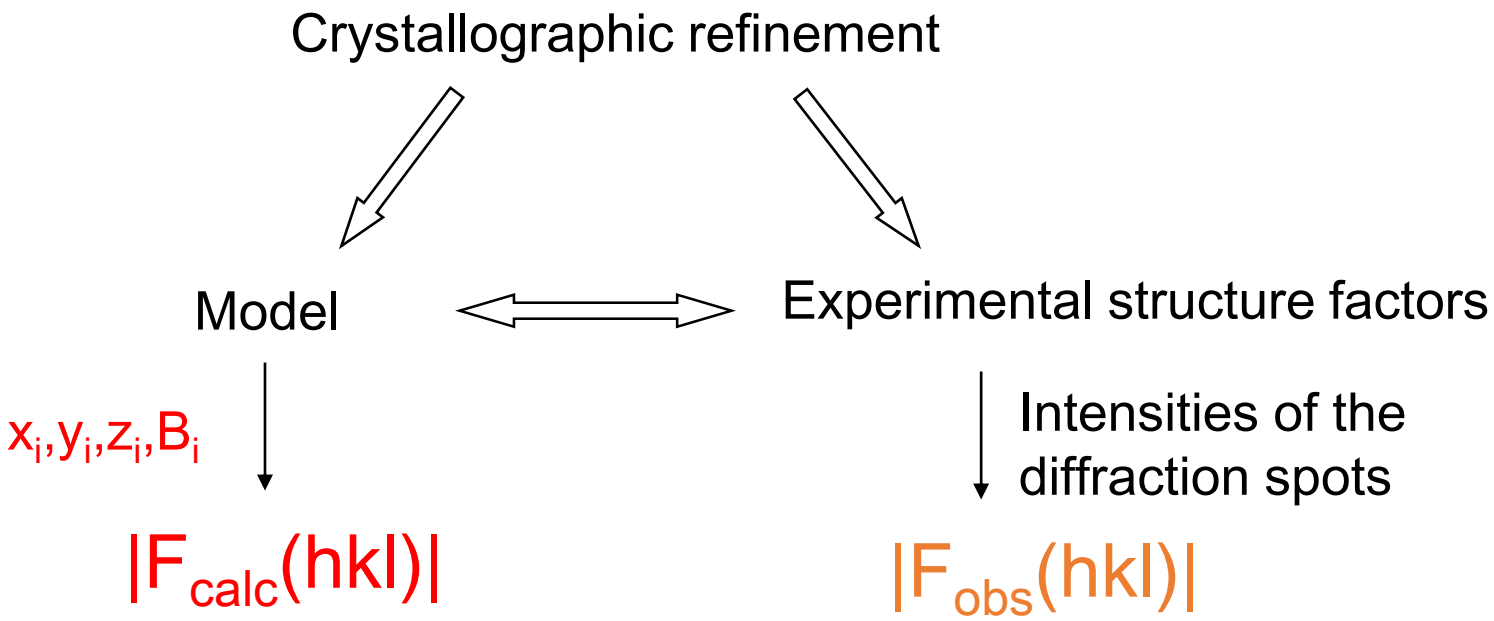


Data fitting and overfitting



Protein structure refinement

The 3D model requires for each atom of the molecule three parameters for the position (x_i,y_i,z_i) + 1 parameter for the thermal factor B_i



Minimization of the function

$$Q = \sum_{hkl} w(h,k,l)\,(|F_{obs}(h,k,l)| - |F_{calc}(h,k,l)|)^2$$

Target function

Protein structure refinement

This approach will not work with the amount of data that can be typically derived from protein crystals

A limited number of data with an impressive number of parameters to be determined

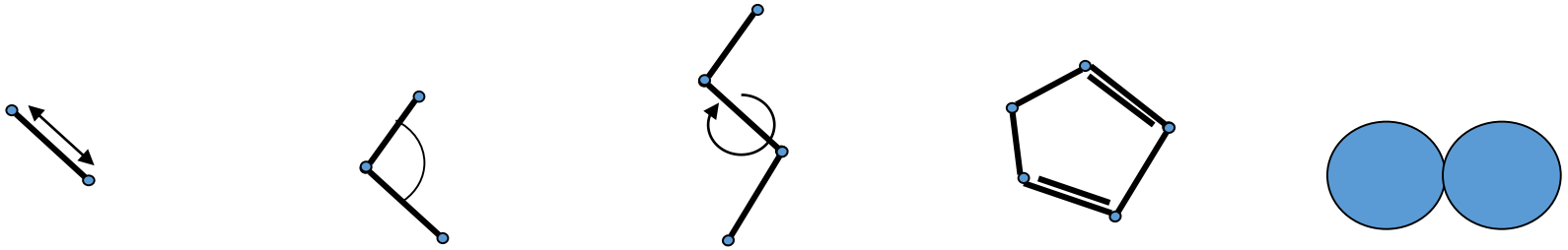
The crystallographic data must be integrated with some additional information

Solution

Application of stereochemical restraints

Protein structure refinement

Application of stereochemical restraints



Bond distances Valence bond Torsion angles Planarity van der Waals

$$Q = \sum_{hkl} w(h,k,l)\ (|F_{obs}(h,k,l)| - |F_{calc}(h,k,l)|)^2 + \Sigma_q w(q)\ (g_{eq} - g_q)^2$$

Where g_q is the q-restraint

g_{eq} is the ideal target for the restraints (es: C-O 1.231±0.02 Å)

g_{eq}= *a priori* knowledge of protein structure.
Ideal values are derived from small molecules and calculations

w(h,k,l) and w(q) are the weight that are difficult to estimate *a priori*

Crystallographic indicators

How can the refinement protocol be evaluated ?

The R-factor index , R-factor

$$R\text{-factor} = \sum_{hkl} \left| |F_{obs}(hkl)| - k|F_{calc}(hkl)| \right| / \sum_{hkl} |F_{obs}(hkl)|$$

R-factor values usually decrease when the resolution increases

It is commonly stated that the R-factor should be < 0.20 for well-refined structures. This is not enough as several wrong models presented good values of the R-factor.

Other fundamental checks

The number of electron density peaks that are not interpreted is low

The model must present a correct stereochemistry

Application of the cross validation analysis – The R-free

Crystallographic indicators

The real space R-factor

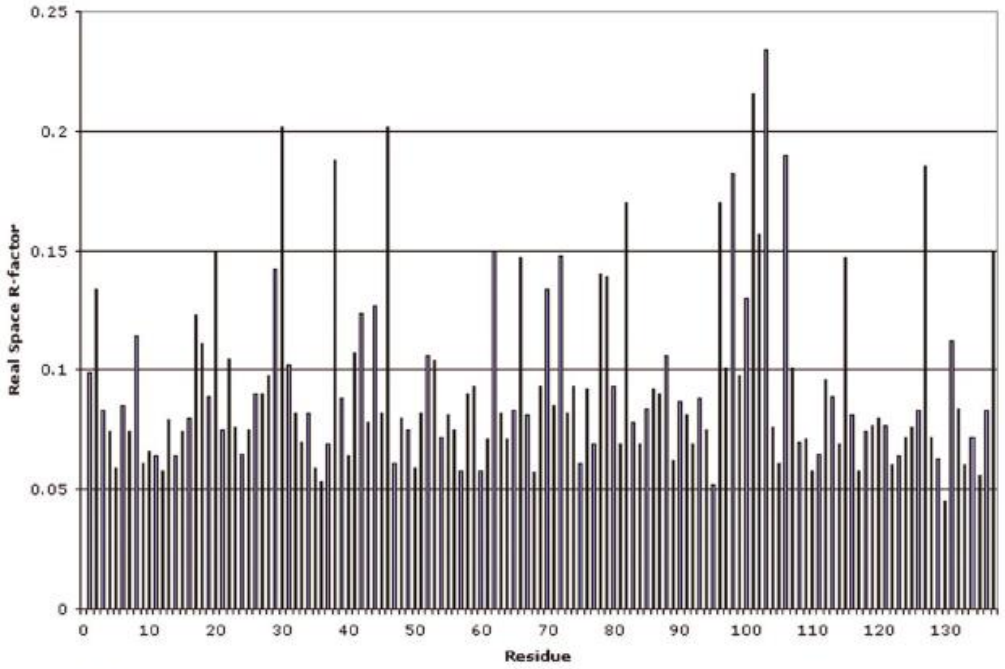


Figure 3
Example of a residue-based plot of real-space *R* values (for entry 1cbs). The bar of every residue is clickable in the browser and will launch the density viewer, load and display the appropriate model and map and centre on the selected residue.

Jones 1991, Kleywegt et al. 2004.

Famous “bad” classical structures

- Azobacter ferredoxin (wrong space group)
- Zn-metallothionein (mistraced chain)
- Alpha bungarotoxin (poor stereochemistry)
- Yeast enolase (mistraced chain)
- Ras P21 oncogene (mistraced chain)
- Gene V protein (poor stereochemistry)

Need for Protein structure validation

In the early nineties the community became aware of the problem

COMMENTARY

Between objectivity and subjectivity

Carl-Ivar Brändén and T. Alwyn Jones

Protein crystallography is an exacting trade, and the results may contain errors that are difficult to identify. It is the crystallographer's responsibility to make sure that incorrect protein structures do not reach the literature.

Nature **343**, 687 - 689 (22 February 1990)

Protein crystallography is an exacting trade, and the results may contain errors that are difficult to identify.

It is the crystallographer's responsibility to make sure that incorrect protein structures do not reach the literature.

Need for protein structure validation

Cross correlation in Structural Biology

Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures

Axel T. Brünger

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, 06511, USA

THE determination of macromolecular structure by crystallography involves fitting atomic models to the observed diffraction data[1]. The traditional measure of the quality of this fit, and presumably the accuracy of the model, is the *R* value. Despite stereochemical restraints[2], it is possible to overfit or 'misfit' the diffraction data: an incorrect model can be refined to fairly good *R* values as several recent examples have shown[3]. Here I propose a reliable and unbiased indicator of the accuracy of such models. By analogy with the cross-validation method[4,5] of testing statistical models I define a statistical quantity (R_T^{free}) that measures the agreement between observed and computed structure factor amplitudes for a 'test' set of reflections that is omitted in the modelling and refinement process. As examples show, there is a high correlation between R_T^{free} and the accuracy of the atomic model phases. This is useful because experimental phase information is usually inaccurate, incomplete or unavailable. I expect that R_T^{free} will provide a measure of the information content of recently proposed models of thermal motion and disorder[6–8], time-averaging[9] and bulk solvent[10].

$$R\text{-factor} = \sum_{hkl} \left| |F_{obs}(hkl)| - k\,|F_{calc}(hkl)| \right| / \sum_{hkl} |F_{obs}(hkl)|$$

$$R\text{-}Free = \sum_{hkl} \left| |F_{obs}(hkl)| - k\,|F_{calc}(hkl)| \right| / \sum_{hkl} |F_{obs}(hkl)|$$

The R-free is calculated on reflections that are not included the refinement

Need for Protein structure validation

The value of the R-free is always larger than that of the R-factor

The difference between R-free and R-factor should decrease upon resolution increases. It should be low (4-5 percentage units) for well refined structures

R-factor *versus* R-free

Table 13.1. Statistics of data collection and refinement for hevamin at pH 2.0. The structure was refined with TNT against data from 15.0 to 1.9 Å resolution.*

Data processing	
Number of observations	109372
Number of unique reflections	19169
R_{merge} (1.93–1.90 Å)	0.065 (0.213)**
Completeness (1.93–1.90 Å)	0.941 (0.414)
Refinement	
Resolution range (Å)	15.0–1.90
Completeness of working set	0.83
Completeness of test set	0.09
R-factor	0.157
R_{free}	0.199
Number of protein atoms	2087
Number of solvent atoms	140
RMS deviations from ideality	
Bond lengths (Å)	0.010
Bond angles (°)	1.48
Dihedrals (°)	23.0
B-value correlations for bonded atoms (Å^2)	2.1
Average *B*-values (Å^2)	
All protein atoms	16.5
Main chain atoms	13.2
Side chain atoms	20.1
Solvent atoms	33.2

* From Terwisscha van Scheltinga, thesis 1997, University of Groningen, with permission

** Values in parentheses are for the high resolution shell

Resolution= 1.9 Å

Data/parameters
19169/8908=2.15

R-factor=0.157

R-free=0.199

Need for protein structure validation

Errors in protein structures

This short note in NATURE describes how more than 1,000,000 problems were detected in the PDB using the program WHAT_CHECK. This article was submitted with the title *1000000 outliers in protein structures* but NATURE changed the title without asking us.

We called all these problems deliberately outliers because in most cases we can only be 90%, 99%, 99.9%, etc., sure that a detected problem is really an error. A small but significant fraction of the problems could represent actual features of the protein structure. One should keep in mind that a values that deviate three sigma from the mean should show up in about 1 per 1000 of all cases.

Errors *versus* outliers

A FEW OF THE ERRORS IN THE LITERATURE...	
Inconsistent symmetry information	19 files
Transformation matrix has determinant not equal to 1.0	5 cases
D amino acid	183 cases
Atom too close to symmetry axis leading to a clash	332 cases
Structure probably solved in wrong space group	24 files
Much too high Matthews' coefficient (V_m>7.0)	69 files
B-factors over-refined	533 files
Cell dimension off by more than 0.5%	1,914 files
Atomic occupancies negative or larger than 1.0	43,934 cases
Bond length deviates more than 4σ	61,051 cases
Bond angle deviates more than 4σ	309,186 cases
Atoms more than 0.4 Å too close to each other	265,290 cases
Side chain of His, Asn of Gln needs 180° flip	19,906 cases

The 1,159,804 outliers in Protein Data Bank data sets reflect discrepancies with conventions, statistical outliers and probable errors. Of the 76 classes of problems only 13 are listed in this table. The complete tables, full reports about every entry that we tested and detailed descriptions of all tests are available from http://www.sander.embl-heidelberg.de/pdbreport/

R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, **Nature** (1996) 381, 272-272

Protein structure validation

The ABC Transporter Debacle

A letter by Chang et al (2006) Science 314:1875 retracted the structures of the membrane proteins MsbA and EmrE multidrug transporters as reported in five papers:

Chang and Roth (2001) Science 293:1793

Rees and Chang (2005) Science 308:1028

Pornillos, Chen, Chen and Chang (2005) Science 310:1950

Chang (2003) J.Mol.Biol. 330:419

Ma and Chang (2004) Proc.Natl.Acad.Sci USA 101:2852

Protein structure validation

The ABC Transporter Debacle

Retraction

WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP•vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

The error in the topology of the original MsbA structure was a consequence of the low resolution of the data as well as breaks in the electron density for the connecting loop regions. Unfortunately, the use of the multicopy refinement procedure still allowed us to obtain reasonable refinement values for the wrong structures.

The Protein Data Bank (PDB) files 1JSQ, 1PF4, and 1Z2R for MsbA and 1S7B and 2F2M for EmrE have been moved to the archive of obsolete PDB entries. The MsbA and EmrE structures will be recalculated from the original data using the proper sign for the anomalous differences, and the new Cα coordinates and structure factors will be deposited.

We very sincerely regret the confusion that these papers have caused and, in particular, subsequent research efforts that were unproductive as a result of our original findings.

GEOFFREY CHANG, CHRISTOPHER B. ROTH, CHRISTOPHER L. REYES, OWEN PORNILLOS, YEN-JU CHEN, ANDY P. CHEN

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

References

1. G. Chang, C. B. Roth, *Science* **293**, 1793 (2001).
2. C. L. Reyes, G. Chang, *Science* **308**, 1028 (2005).
3. O. Pornillos, Y.-J. Chen, A. P. Chen, G. Chang, *Science* **310**, 1950 (2005).
4. R. J. Dawson, K. P. Locher, *Nature* **443**, 180 (2006).
5. G. Chang, *J. Mol. Biol.* **330**, 419 (2003).
6. C. Ma, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).

Questioning the retraction

COMMENTARY

Five retracted structure reports: Inverted or incorrect?

BRIAN W. MATTHEWS

Institute of Molecular Biology, Howard Hughes Medical Institute, and Department of Physics, University of Oregon, Eugene, Oregon 97403, USA

B.W. Matthews Protein Sci. 2007 Jun; 16(6): 1013–1016.

Protein structure validation

Protein structure properties typically used for validation

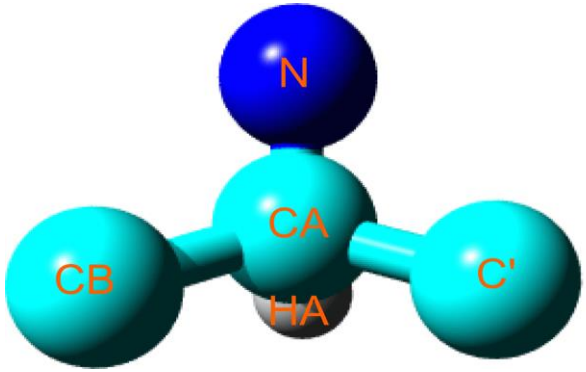
Bond lengths, bond angles, chirality, omega angles,

Ramachandran plot, rotameric states, packing quality, backbone conformation, side chain planarity

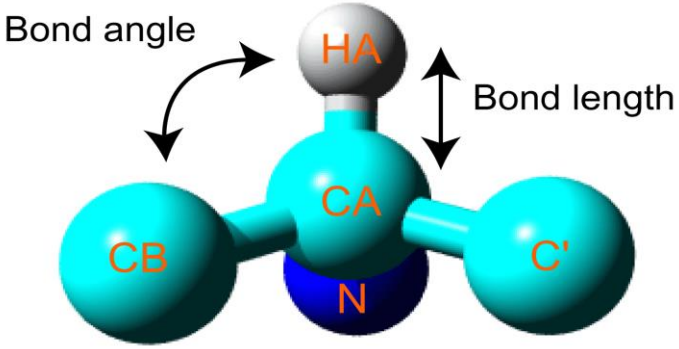
Inter-atomic bumps, buried hydrogen-bonds, electrostatics

Protein structure validation

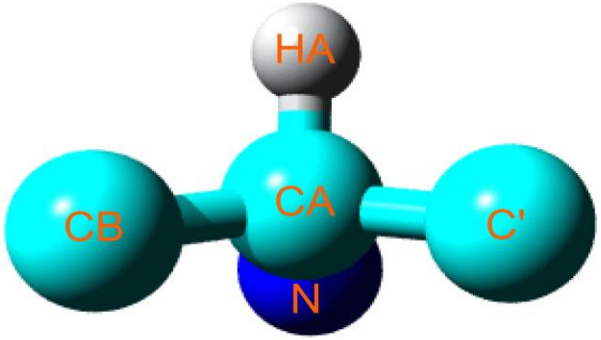
Bonded geometry



D-amino acid



L-amino acid

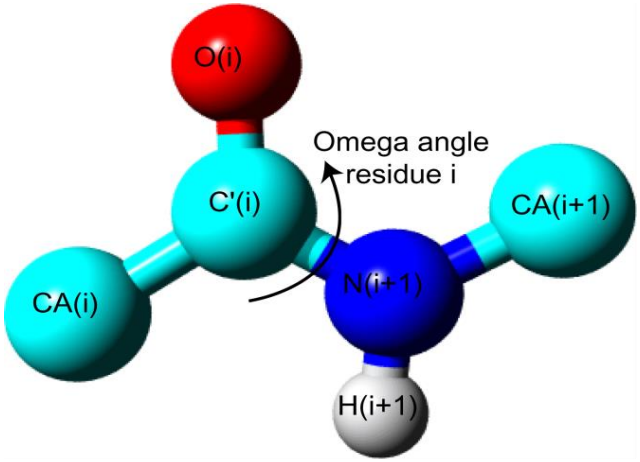


Distorted Cα-chirality

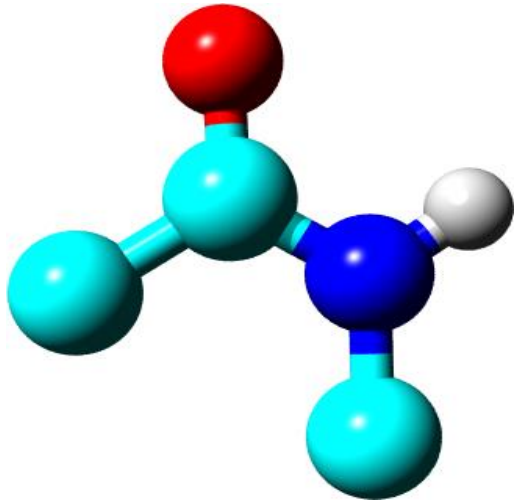
Protein structure validation

Planarity of the peptide bond

The omega angle



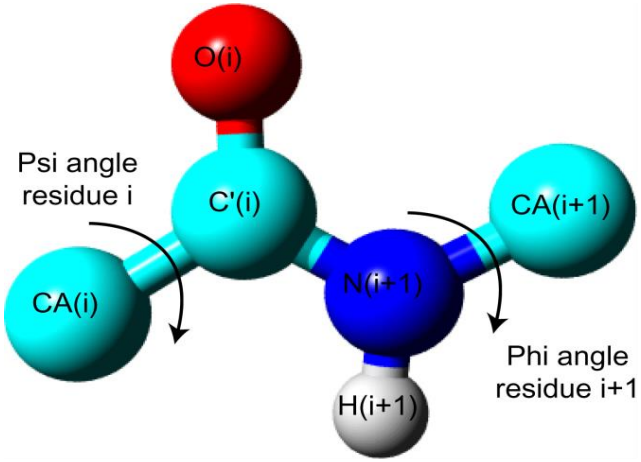
Trans-conformation
(omega=180°)



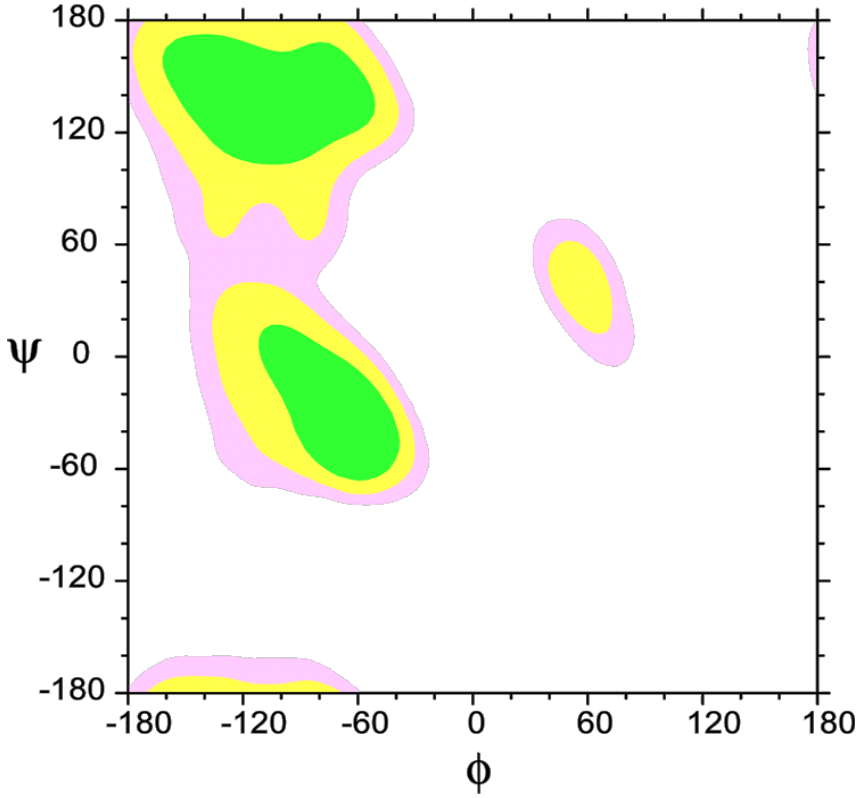
Cis-conformation
(omega=0°)

Protein structure validation

The Ramachandran Plot



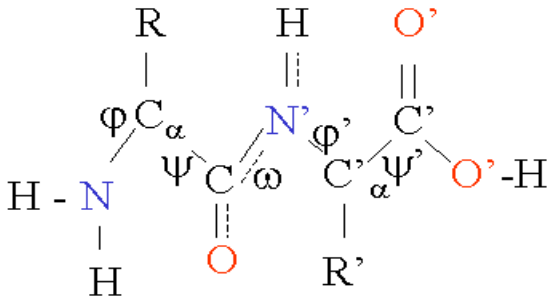
Phi and Psi angles



Ramachandran plot

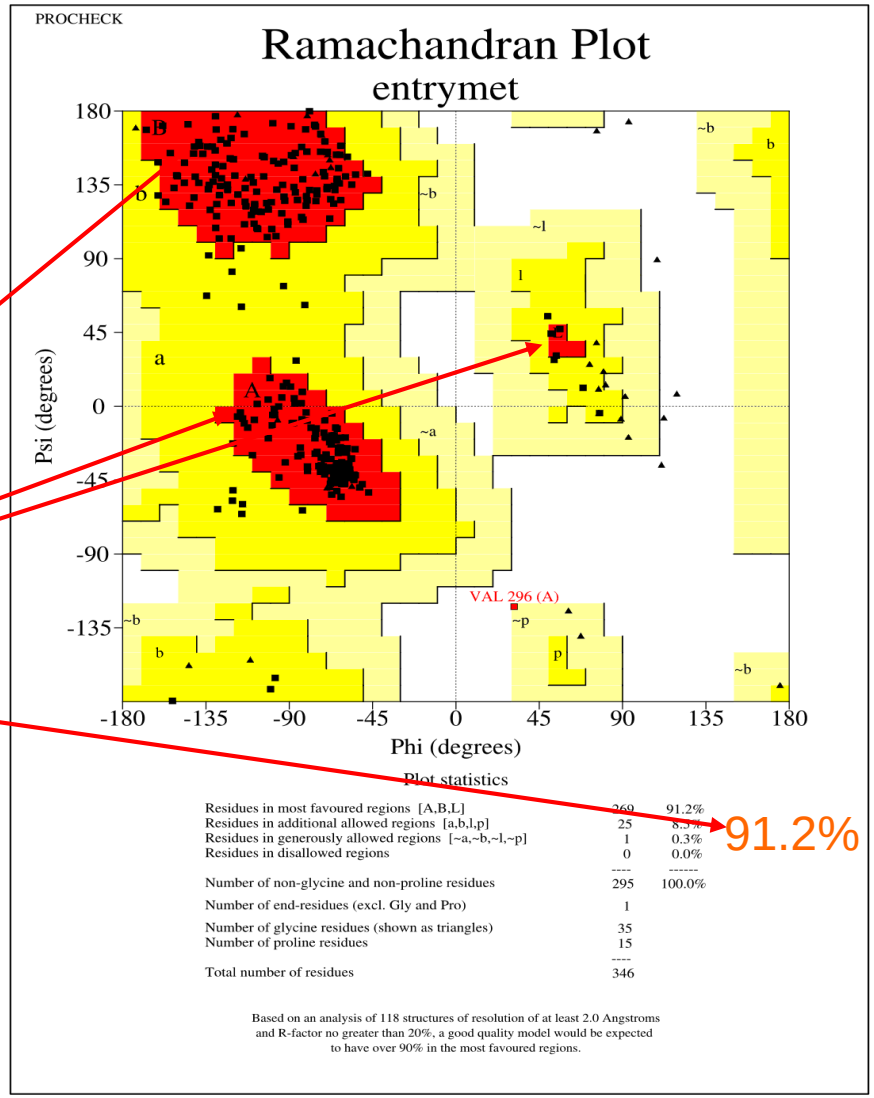
Protein structure validation

The Ramachandran Plot Analysis in PROCHECK



Most favoured regions

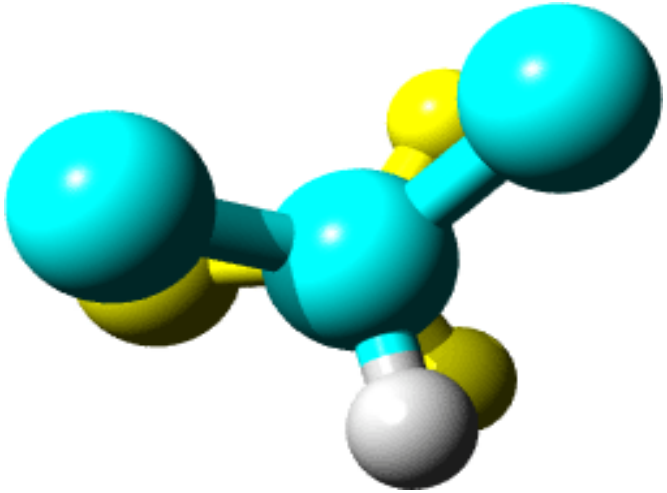
> 90%



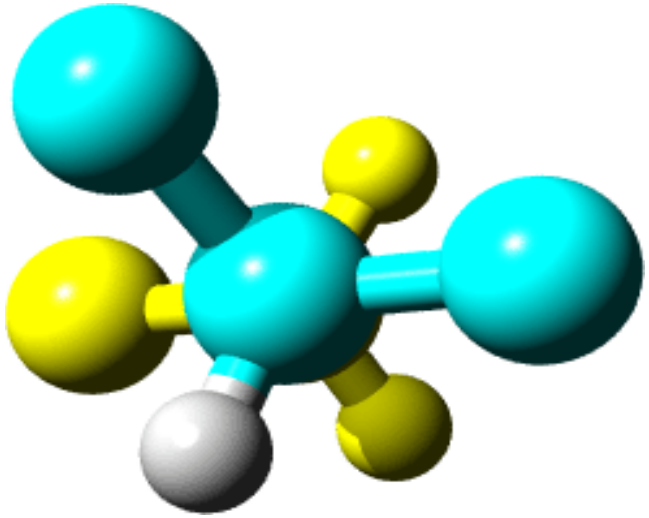
91.2%

Protein structure validation

Rotameric states



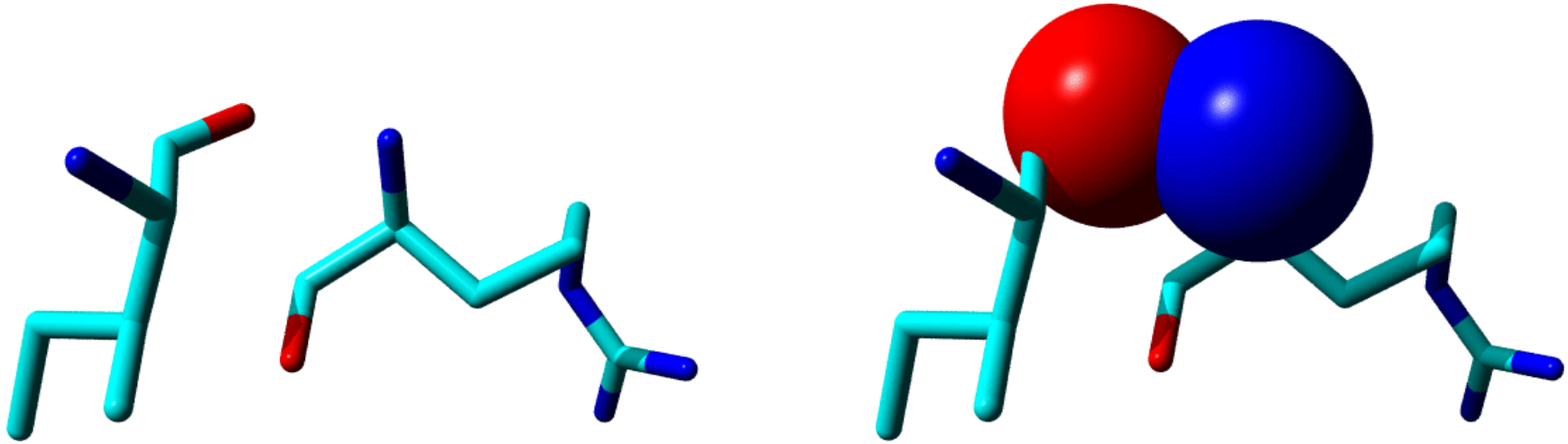
Eclipsed



Staggered

Protein structure validation

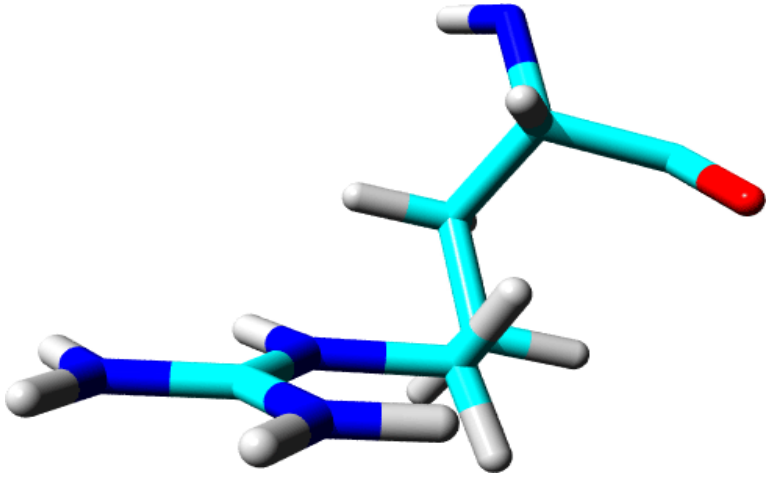
Inter-atomic bumps



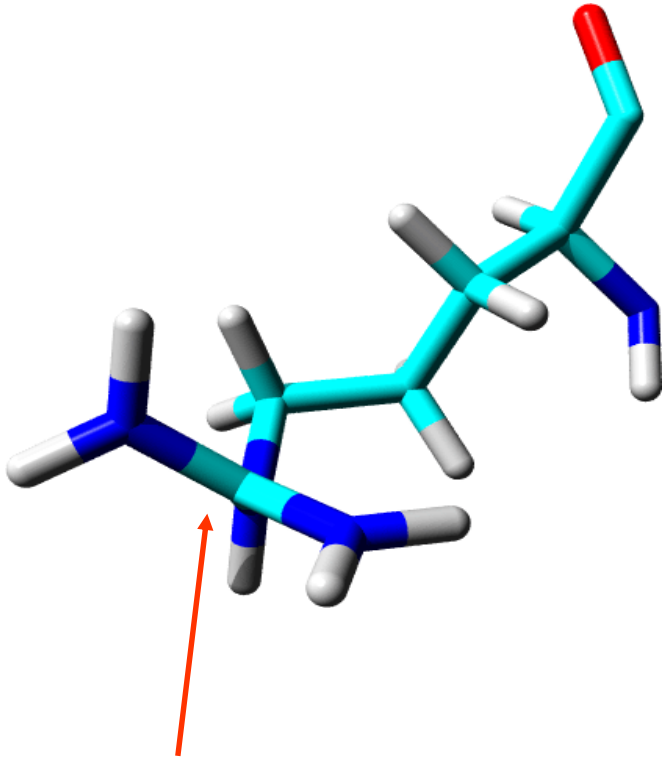
Overlap of protein atoms

Protein structure validation

Side chain planarity



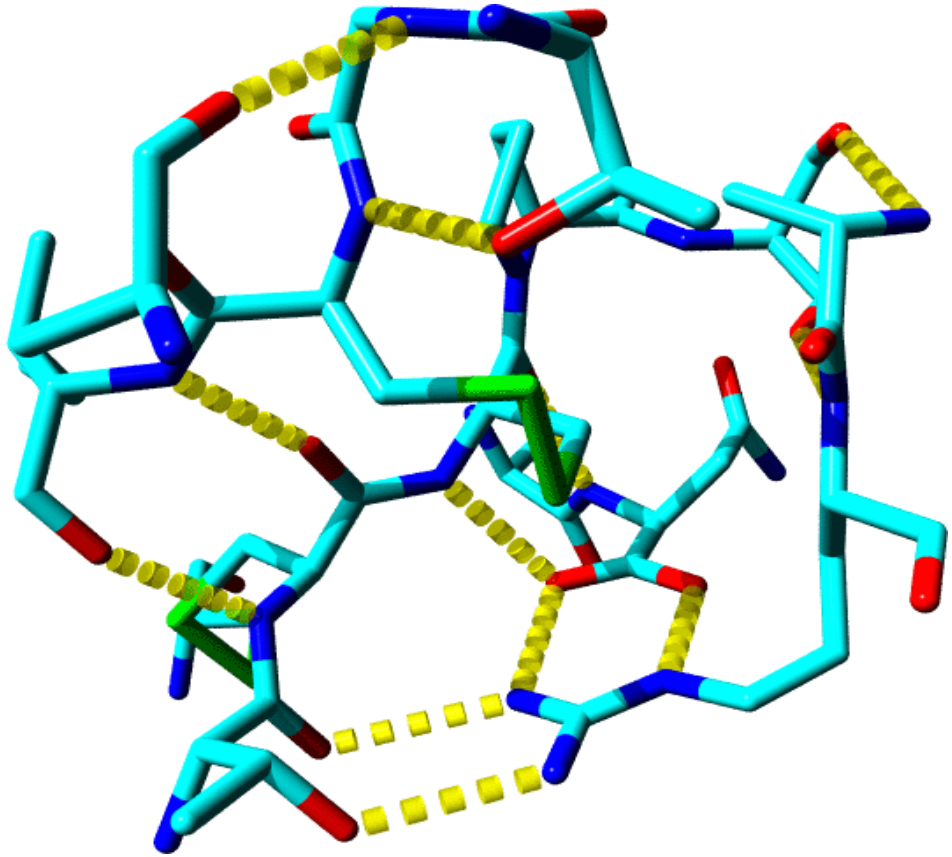
Planar ARG side-chain



Non-planar ARG side-chain

Protein structure validation

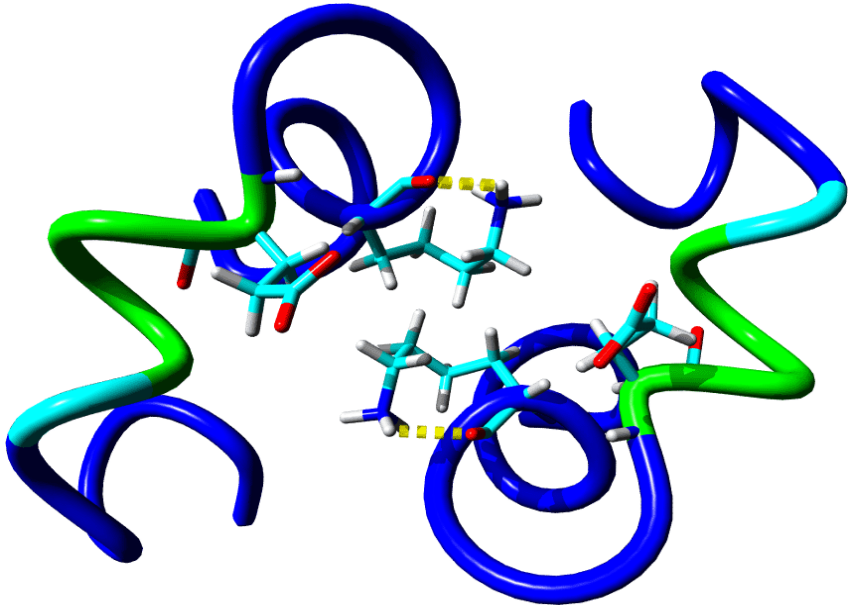
Satisfied/unsatisfied internal hydrogen bonding donors and acceptors



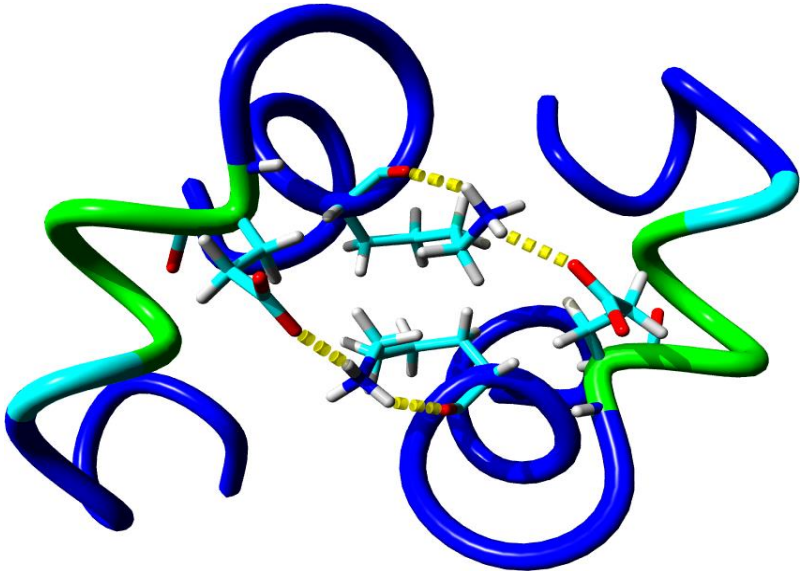
Internal hydrogen bonds in Crambin

Protein structure validation

Optimization of the electrostatic/polar interactions



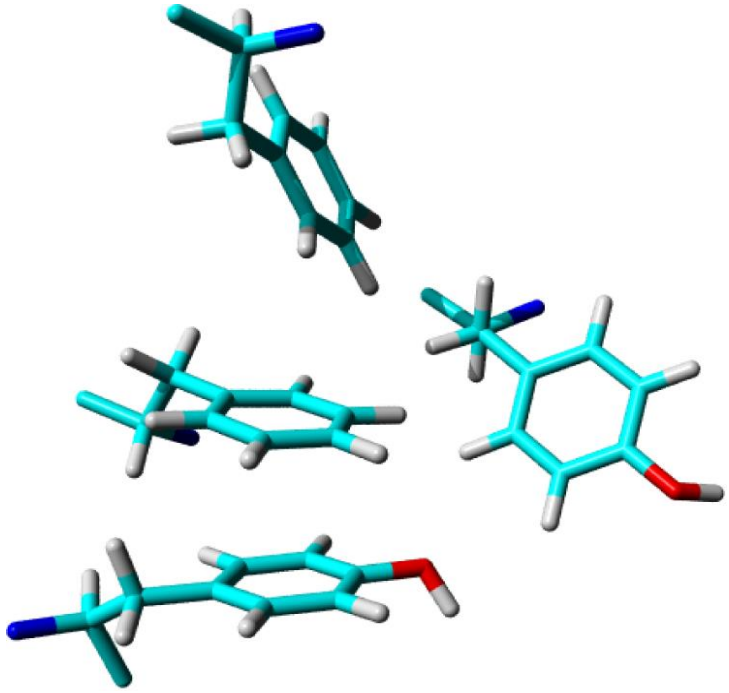
Non-optimal electrostatics



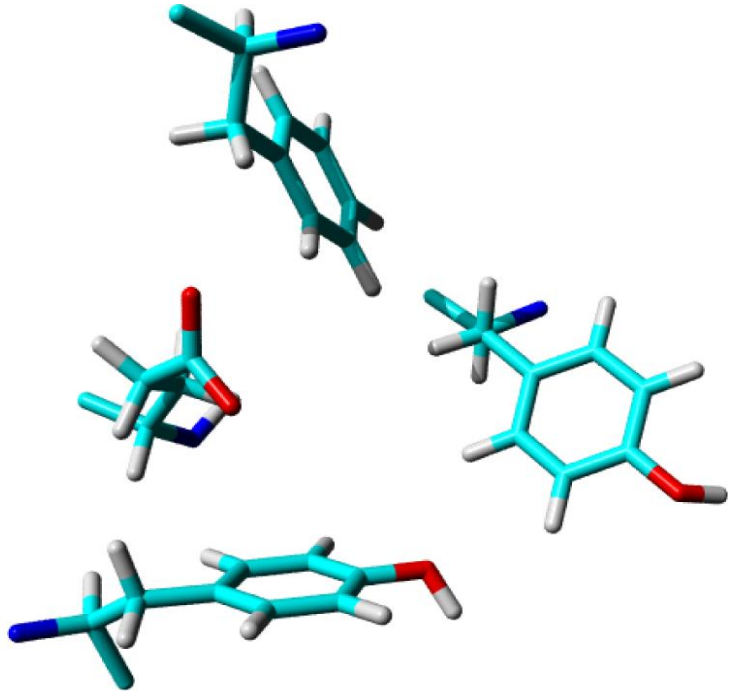
After energy minimization
including electrostatics

Protein structure validation

Packing quality



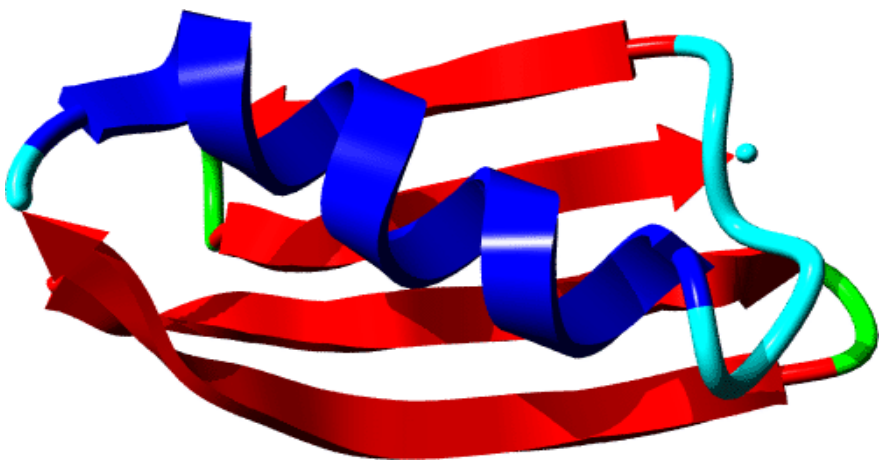
Optimal packing



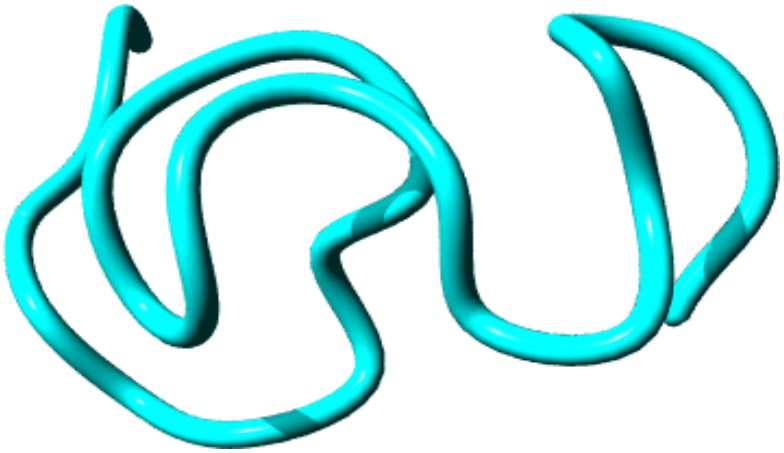
Loose packing

Protein structure validation

Secondary structure content



Typical case

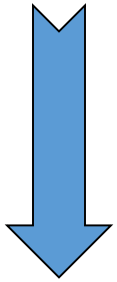


Very unusual

Errors and outliers

Warning!

Identification of *outliers* (unusual properties)
Not necessarily errors



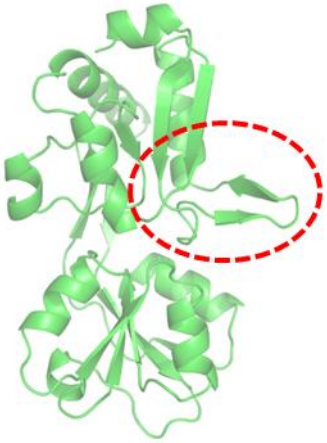
outliers → Look back to the electron density

Electron density maps can be viewed and downloaded for each deposited Pdb entry at the PDB-entry pages (PDBe; http://pdbe.org/). Infos at http://www.ebi.ac.uk/pdbe/eds

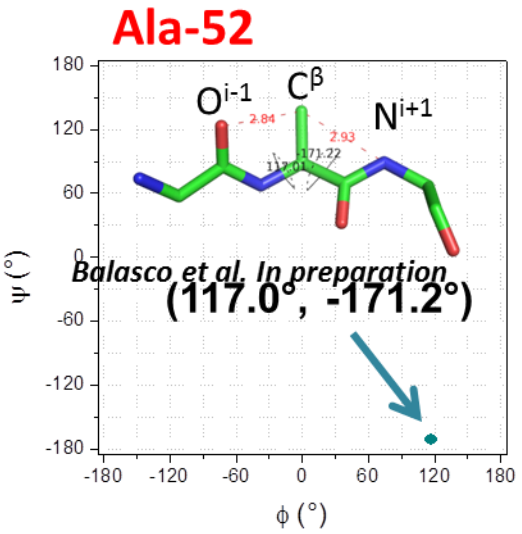
Errors and outliers

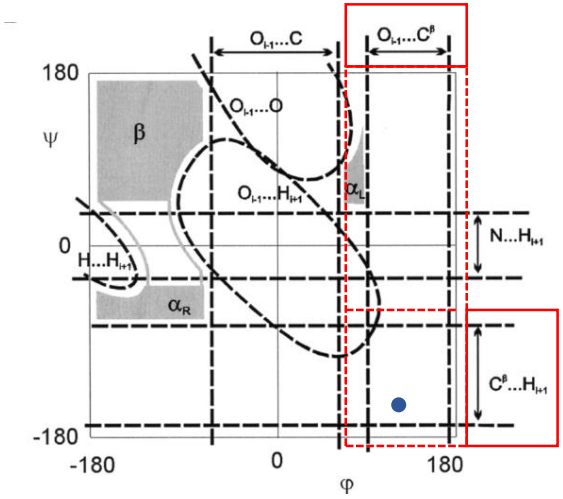
An example of a real outlier

Cristallographic Structure of G52A ArgBP mutant



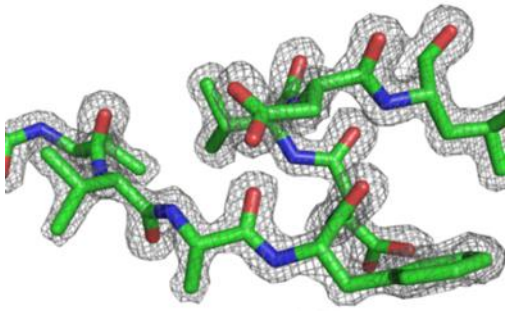
PDB code 6Q3U





Steric clashes of the C^{β} atom

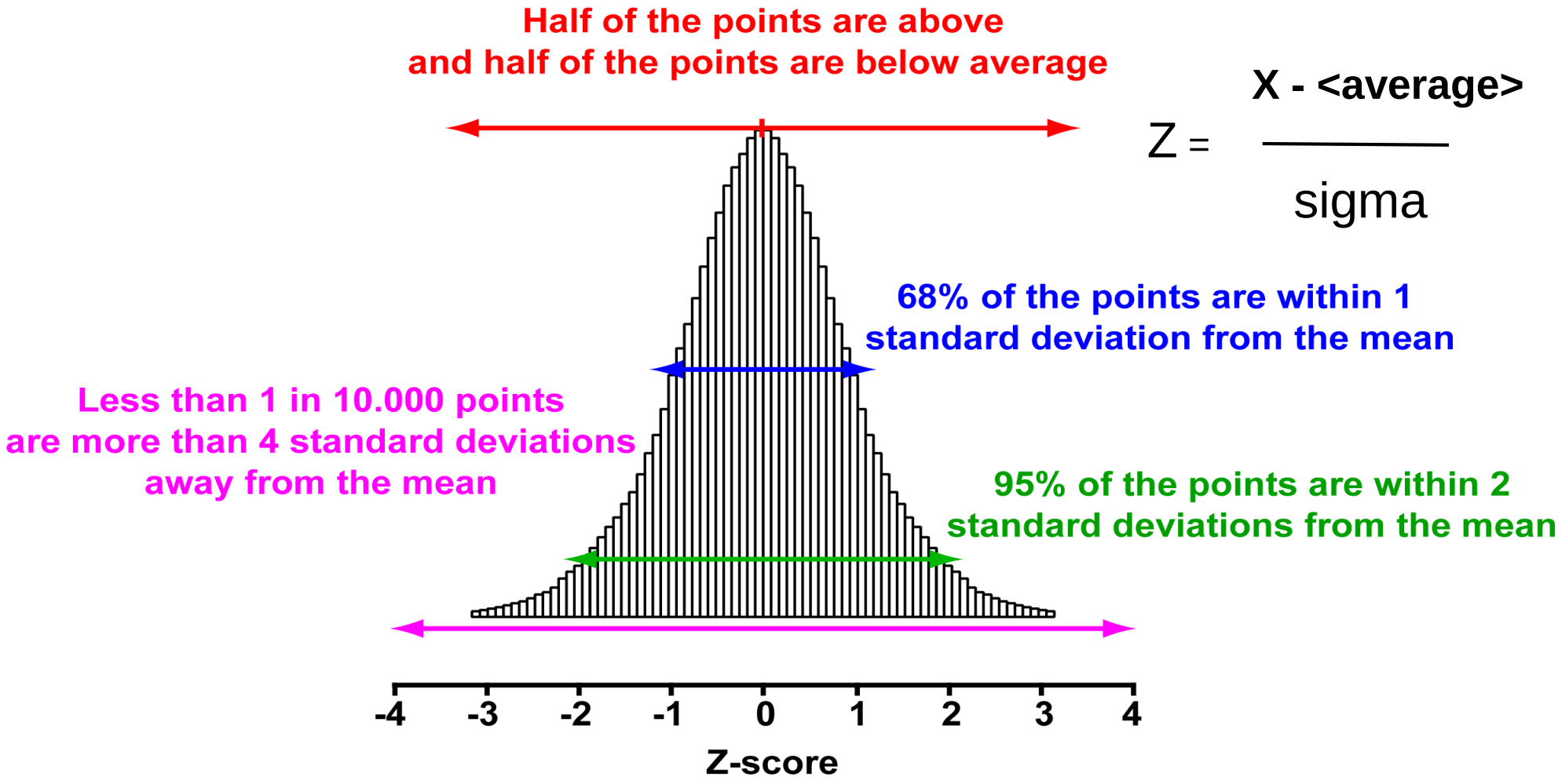
Electron density



Balasco et al. Sci. Rep. 2019, 9, 6617

Errors and outliers

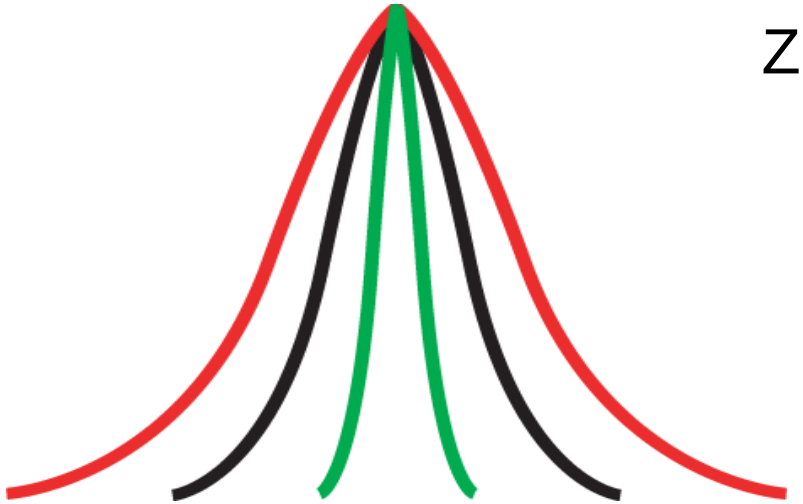
Normal distributions and Z-scores



Errors and outliers

Normal distributions and RMS Z-scores

$$Z = \frac{\text{X - <average>}}{\text{sigma}}$$



RMS Z-score~0.5

RMS Z-score=1.0 (reference)

RMS Z-score~2

$$\text{RMS Z-score} = \sqrt{\frac{sum(Z^2)}{n}}$$

Errors and outliers

Z-scores and RMS Z-scores

Local geometry RMS Z-scores

- Too tight restraining of geometry ➔ 0 < RMS Z-score < 1
- Proper Gaussian distribution ➔ RMS Z-score 1
- Too loose restraining of geometry ➔ 1 < RMS Z-score

Errors and outliers

A WHAT IF summary report

RMS Z-scores

Bond lengths	0.654 (tight)
Bond angles	1.006
Omega angle restraints	3.829 (loose)
Side chain planarity	1.556
Improper dihedral distribution	0.620
Inside/Outside distribution	0.921

Validation servers

General-purpose structure validation

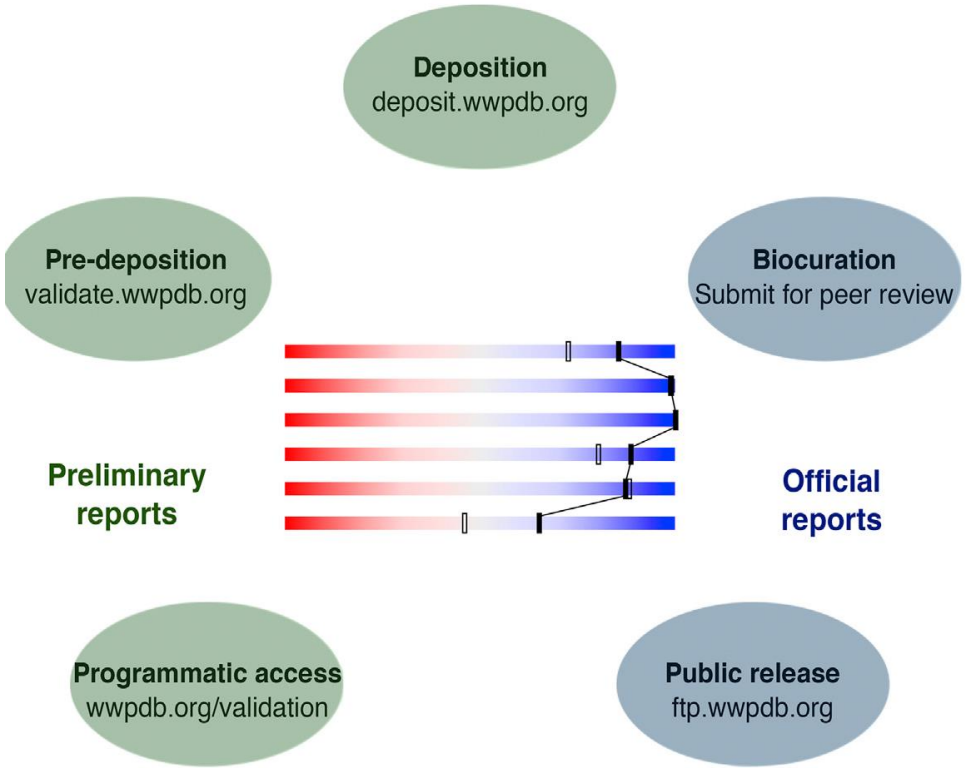
PDB validation/deposition site
MolProbity web service
PDBREPORT - Protein structure validation database
What_Check software
ProCheck software
pdb-care (carbohydrate validation)
Privateer (carbohydrate validation)
ProSA web service
Verify-3D profile analysis

X-ray specific

Coot - modeling software (built-in validation)
PDB_REDO - X-ray model optimization: rebuilding and refining PROSESS - Protein Structure Evaluation Suite & Server
VADAR - Volume, Area, Dihedral Angle Reporter

PDB Validation report

Validation of Structures in the Protein Data Bank



Gore et al., 2017, Structure 25, 1916–1927

PDB Validation report

Table 1. Key Validation Metrics Reported in the wwPDB Structure Validation Reports and Used for Percentile Rank Calculation

Metric	Details	Software Package and References
R_{free}	cross-validation of goodness of fit between the model and the experimental diffraction data not used for refinement. Applicable to crystallographic structures	DCC (Yang et al., 2016)
Clashscore	number of too-close contacts in an entry normalized per 1,000 atoms	MolProbity (Chen et al., 2010)
Ramachandran outliers	fraction of polypeptide residues deemed to have very unusual backbone conformation (<0.5% of those observed in a high-quality reference set)	MolProbity (Chen et al., 2010), Maxit (Z.F., https://sw-tools.rcsb.org/apps/MAXIT)
Side-chain outliers	fraction of polypeptide residues in non-rotameric side-chain conformations (<0.5% of those observed in a high-quality reference set)	MolProbity (Chen et al., 2010)
RSRZ outliers	fraction of polypeptide and/or polynucleotide residues that do not fit the electron density well when compared with other instances of the same residues in structures at similar resolution. Applicable to crystallographic structures	EDS (Kleywegt et al., 2004)
RNA backbone	average score over all RNA nucleotides in the entry indicating the quality of the observed RNA backbone conformation	MolProbity (Chen et al., 2010)

Gore et al., 2017, Structure 25, 1916–1927

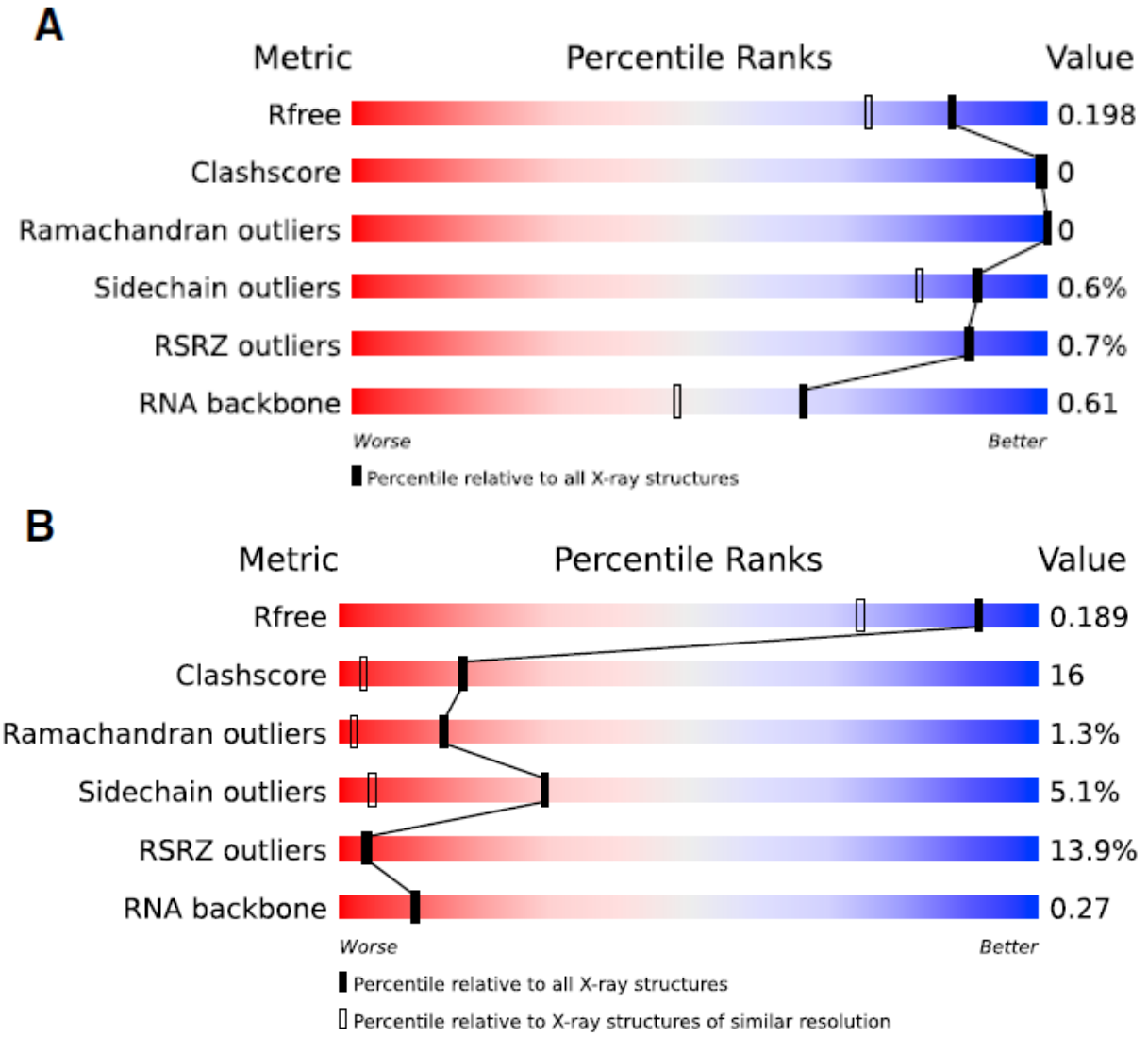
PDB Validation report

Table 3. Component Software Packages Included in the 2017 Version of the Validation Pipeline

Software Package	Which Section and Metric of the Report the Package Is Used for	Reference
MolProbity	model geometry: bond lengths and bond angles of standard protein residues and nucleotides, too-close contacts, Ramachandran outliers, rotamer outliers, RNA suiteness	Chen et al., 2010
MAXIT	model geometry: symmetry-related too-close contacts, stereochemistry issues, identification of *cis*-peptides	Maxit (Z.F., https://sw-tools.rcsb.org/apps/MAXIT/index.html)
Mogul	model geometry: bond-length and bond-angle outliers in small molecules	Bruno et al., 2004
Xtriage (Phenix)	crystallographic data and refinement statistics: signal-to-noise, twinning	Adams et al., 2010
DCC	crystallographic data and refinement statistics: R, R_{free} fit to crystallographic data: R_{free}	Yang et al., 2016
EDS	fit to crystallographic data: real-space R outliers	Kleywegt et al., 2004
Cyrange	NMR ensemble composition: identification of well-defined protein cores	Kirchner and Güntert, 2011
RCI	NMR chemical shifts: prediction of protein backbone order parameter from chemical shifts	Berjanskii and Wishart, 2005
PANAV	NMR chemical shifts: suggested referencing corrections in chemical shift assignments	Wang et al., 2010

Gore et al., 2017, Structure 25, 1916–1927

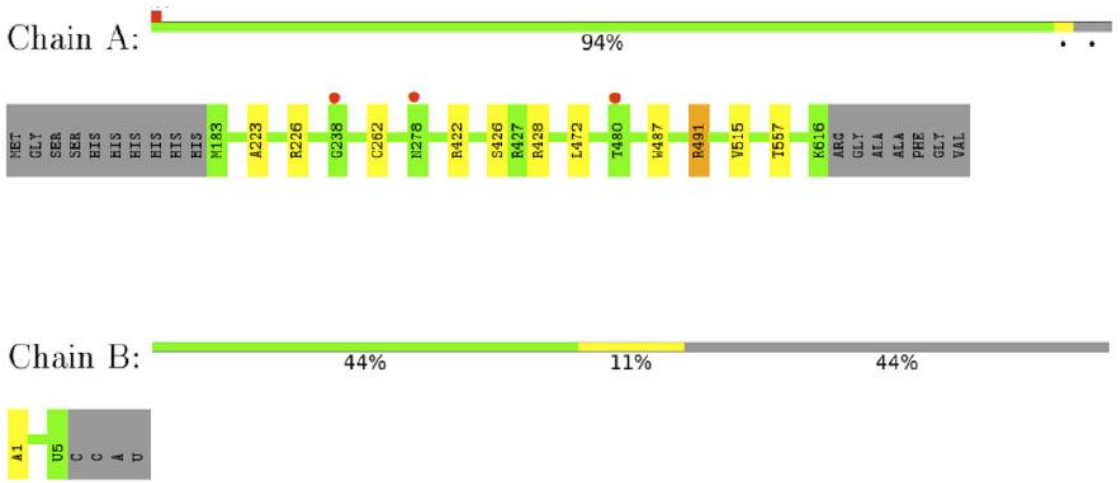
PDB Validation report



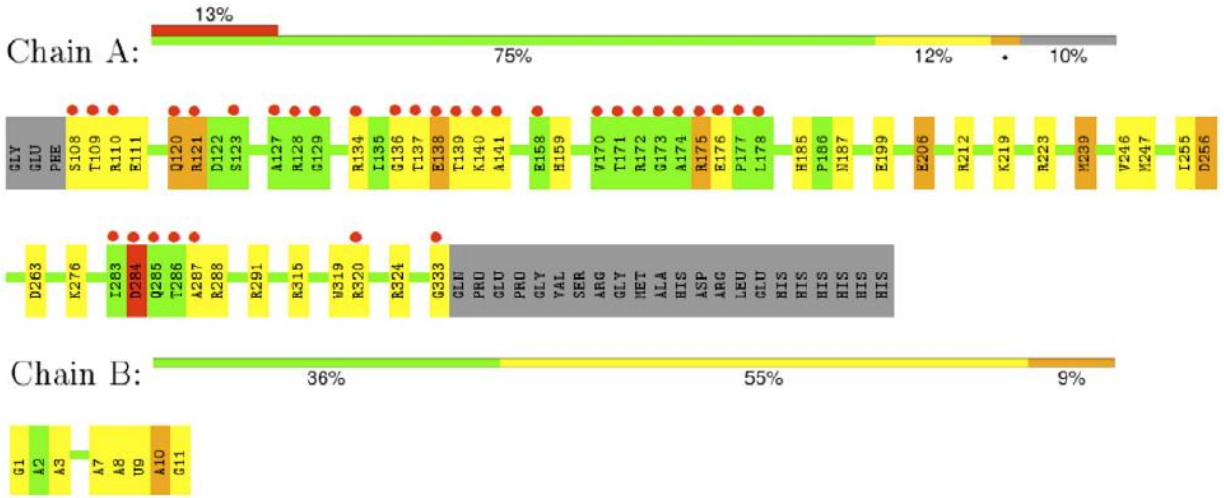
Gore et al., 2017, Structure 25, 1916–1927

PDB Validation report

A



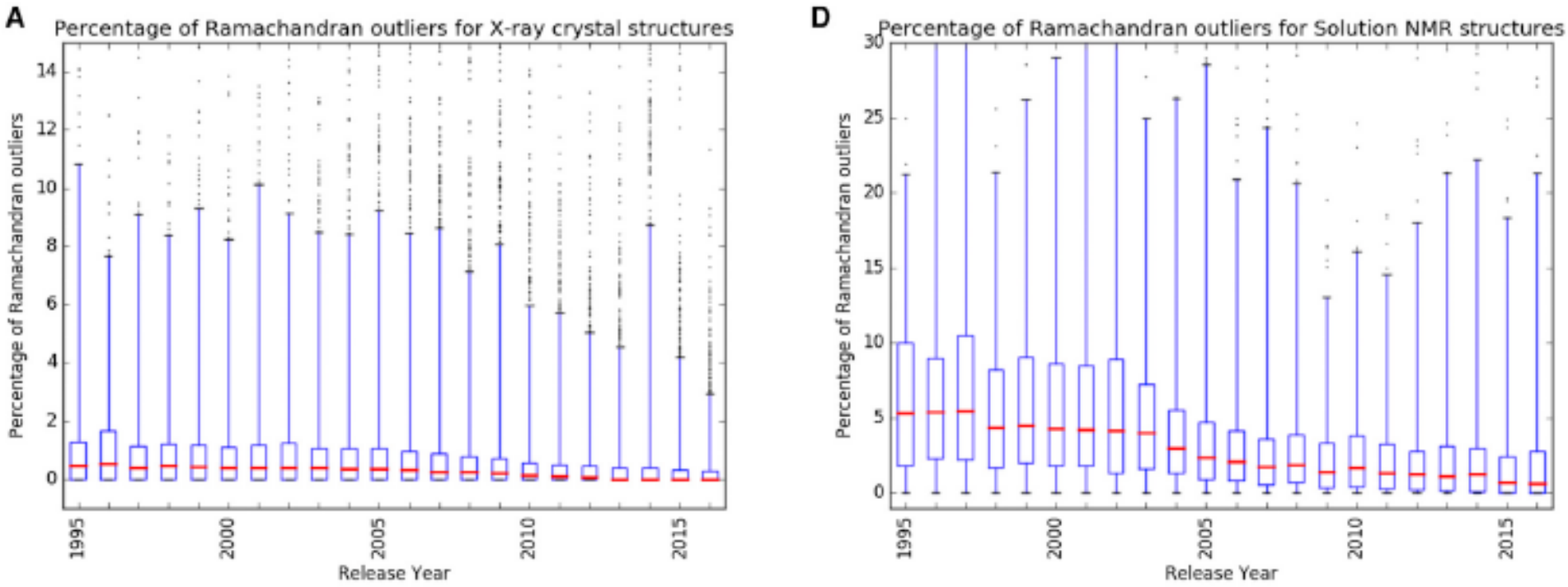
B



Gore et al., 2017, Structure 25, 1916–1927

PDB Validation report

Ramachandran plot outliers as function of time



Gore et al., 2017, Structure 25, 1916–1927

PDB Validation report

Summary table for bond lengths and angles

Mol	Chain	Bond lengths		Bond angles	
		RMSZ	#|Z| >5	RMSZ	#|Z| >5
1	A	1.67	39/5385 (0.7%)	3.87	1158/7301 (15.9%)
1	B	1.67	40/5385 (0.7%)	3.87	1154/7301 (15.8%)
1	C	1.67	38/5385 (0.7%)	3.87	1154/7301 (15.8%)
1	D	1.67	38/5385 (0.7%)	3.87	1151/7301 (15.8%)
1	E	1.67	39/5385 (0.7%)	3.87	1156/7301 (15.8%)
1	F	1.67	39/5385 (0.7%)	3.87	1154/7301 (15.8%)
All	All	1.67	233/32310 (0.7%)	3.87	6927/43806 (15.8%)

Protein structure quality assessment

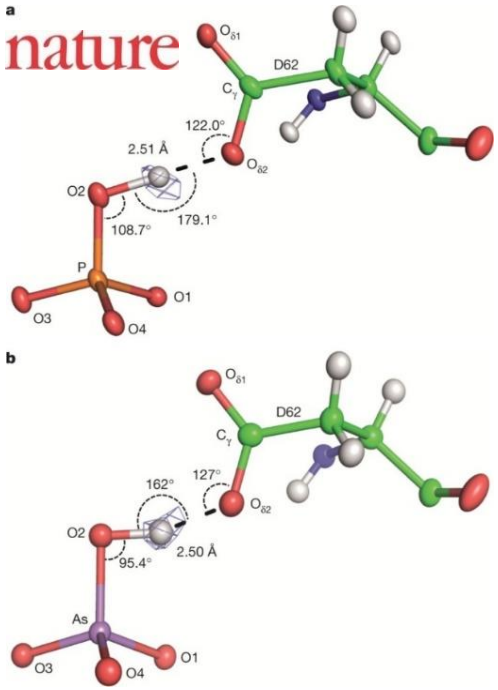
Proteins combine molecular complexity with a very fine structural regulation

Fundamental protein activities often rely on extremely subtle structural details that may fall in the low (or even sub) picometer scale.

Questions involving large scale properties such as the overall fold of a protein, or its topological similarity to other proteins → model essentially correct even though of fairly low precision

Questions involving reaction mechanisms, detailed active-site geometry, ligand binding, protonation states → model with the greatest accuracy and precision as possible

Protein structure quality assessment



The molecular basis of phosphate discrimination in arsenate-rich environments

Periplasmic **phosphate-binding protein** PBP: arsenate-bound and phosphate-bound structures determined at 0.96 Å and 0.88 Å resolution.

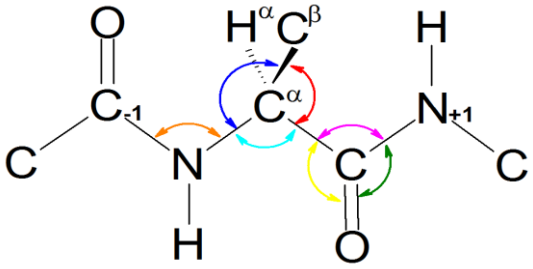
The low-barrier Hydrogen Bond (negative-charge-assisted HB) angles are optimal in the phosphate-bound structure but distorted with arsenate. **This is the consequence of the longer As-O bond than the P-O bond.**

→Anion selectivity (at least 10^3 excess)

Elias *et al. Nature* 491, 34-137 (**2012**)

Protein structure quality assessment

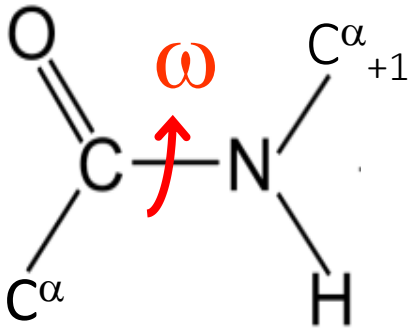
The fine structure of proteins



Polypeptide chain: Backbone Bond Angles

$NC^{\alpha}C$, $C^{\alpha}CO$, $C^{\alpha}CN_{+1}$, $C_{-1}NC^{\alpha}$, $C^{\beta}C^{\alpha}C$, $NC^{\alpha}C^{\beta}$, OCN_{+1}

Peptide bond planarity: Difference of Dihedral Angles



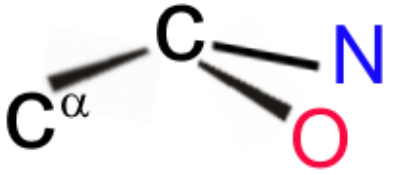
$\Delta\omega = \omega - 180°$ (mod360°)

$\omega \rightarrow C^{\alpha}CN_{+1}C^{\alpha}_{+1}$

$\Delta\omega° > 0$; $\omega > 180°$

$\Delta\omega° < 0$; $\omega < 180°$

θ_C C-pyramidalization

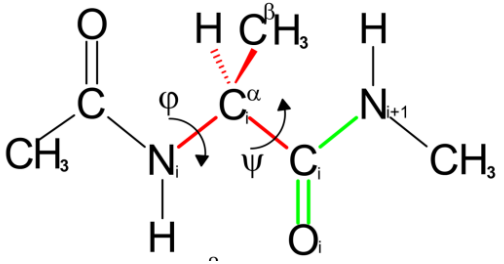


θ_C is the angle between the planes $C^{\alpha}CN_{+1}$ and $N_{+1}CO$.

Polypeptide chain: Backbone Bond Distances

C-N, C-O

N-C^{α}, C^{α}-C^{β} , C^{α}-C

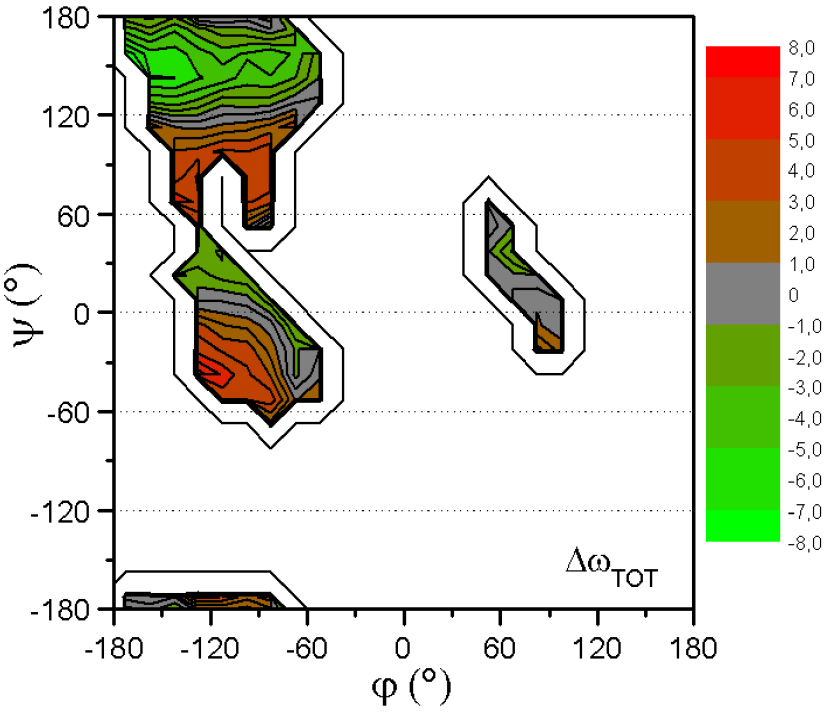


Protein structure quality assessment

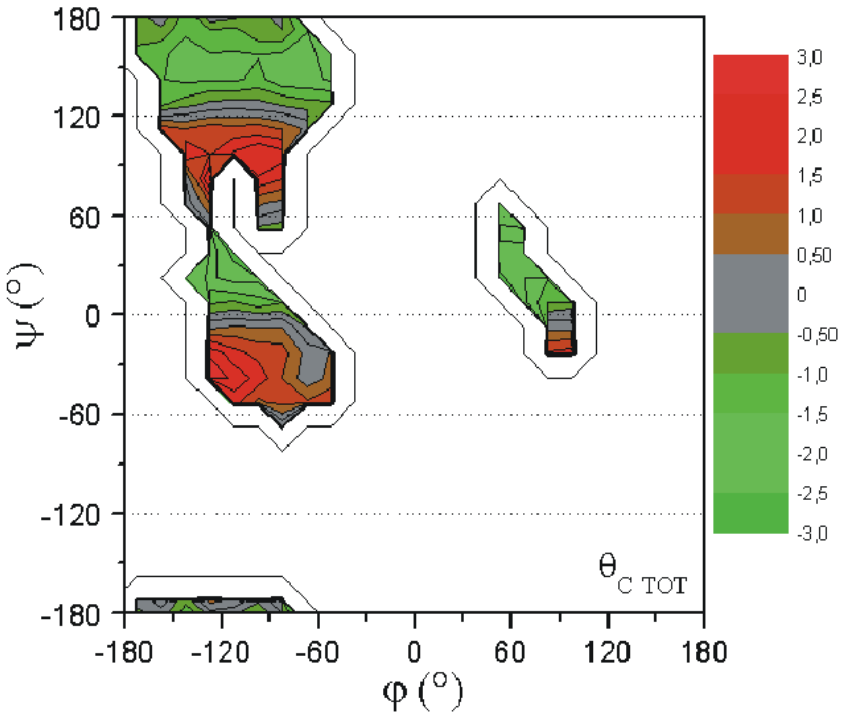
Peptide bond planarity distortion

There is a clear trend of alternating signs of the Δω (θ_C) deviations with the periodic 60° variation in the ψ angle.

Δω



θ_C

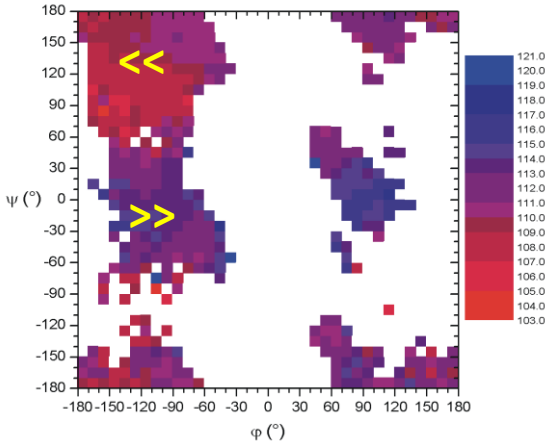


Protein structure quality assessment

Bond angles and Bond distances

$N\hat{C}^{\alpha}C$ angle

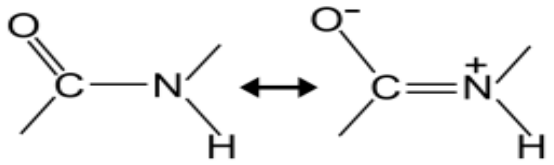
- *Higher values of the angle are found around ψ=0° and in α-helix region, whereas lower values are observed in the β-strand region.*

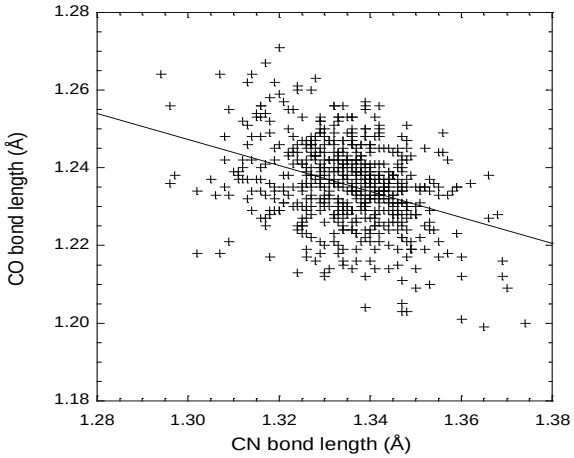


C-O/C-N bond length correlation

- *The C-O shortens when the C-N lengthens according to Pauling's resonance model. There is a statistically significant negative correlation between CN and CO bond distances in protein peptide groups.*

Pauling's resonance model





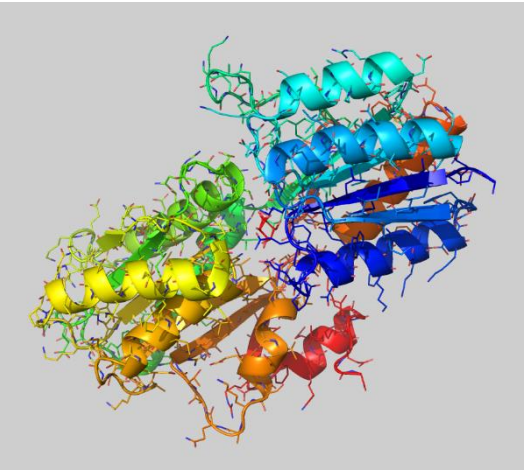
Protein structure quality assessment

Is it possible to detected these trends in individual structures of the PDB?

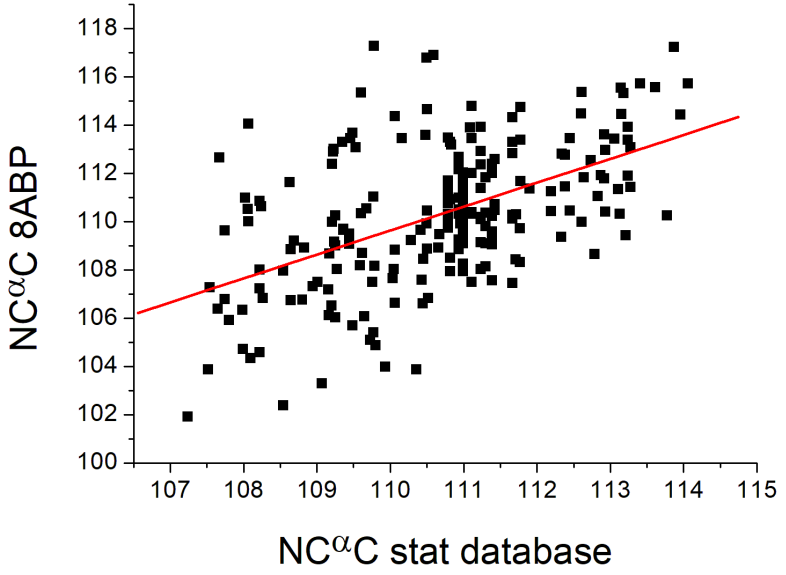
What can we learn from the analysis of PDB structures?

Protein structure quality assessment

Comparison of the expected values derived from databases of well refined structures with those derived from individual PDB structure



8ABP arabinose-binding
Alpha Beta protein
Resolution: 1.49 Å
305 residues



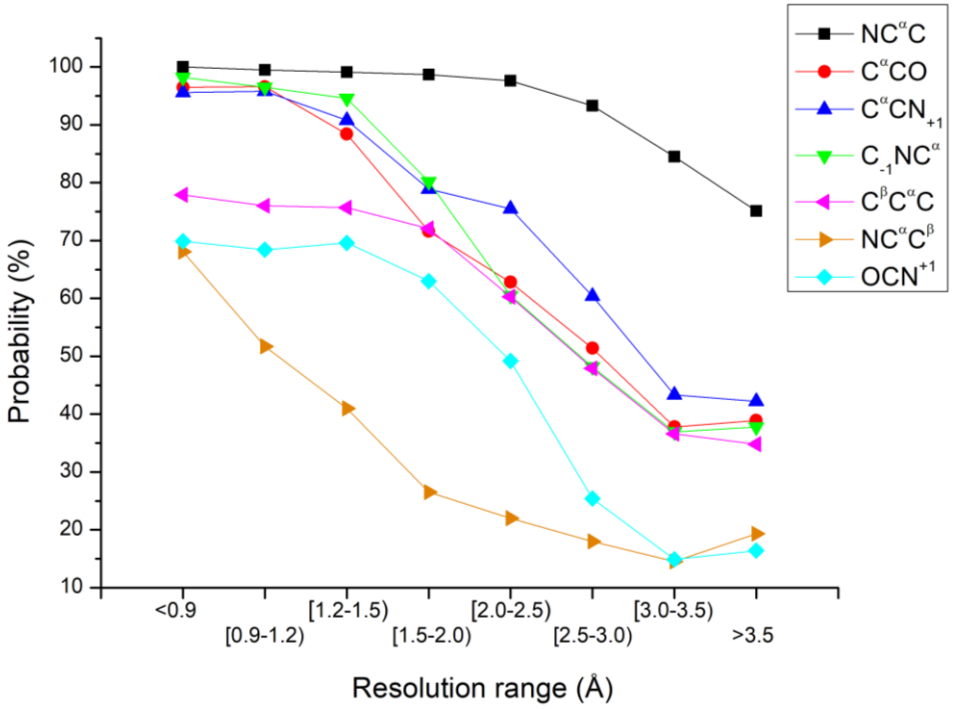
A threshold of P-value<0.001 to consider that the parameter follows the expected trend

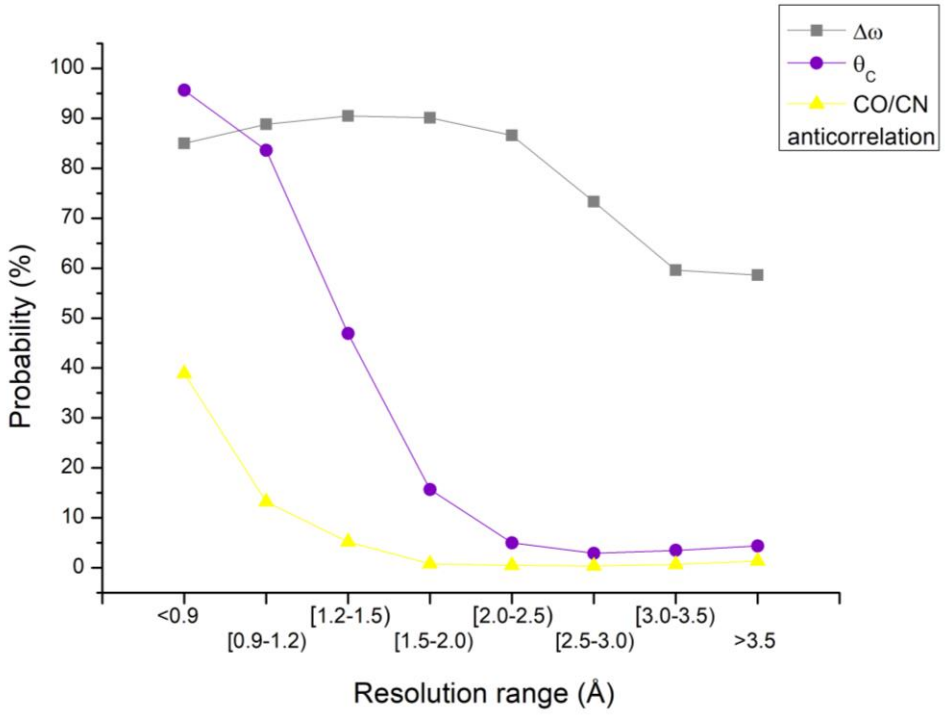
Linear Regression:
Y = 0.59 + 0.99 * X

R-corr.	N	**P-value**
0,521	224	**$<10^{-5}$**

Protein structure quality assessment

Analysis of the entire PDB content





Protein structure quality assessment

Implications

The variability of the **NC$^{\alpha}$C** angle may be considered a sort of universal property of protein structure that is detected in the vast majority of the protein structures.

In very high resolution structures (<1.5 Å) the variability of other bond angles (**C$^{\alpha}$CO**, **C$^{\alpha}$CN$_{+1}$**, and **C$_{-1}$NC$^{\alpha}$**) is well-reproduced.

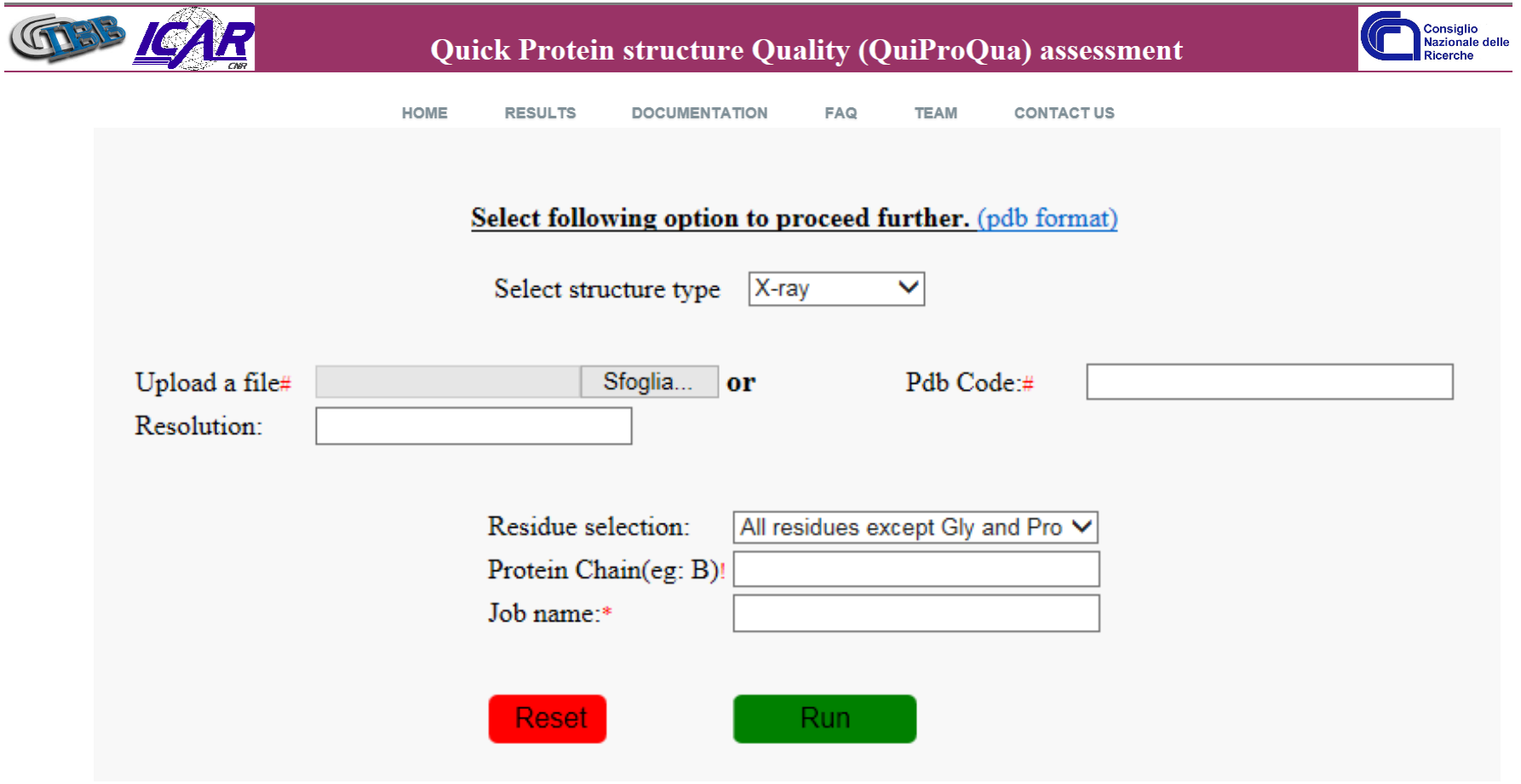
The $\Delta\omega$ trends are reproduced in most of the structures even at low resolution.

The statistical trends for **C pyramidalization** is highly resolution dependent.

Significant trends for distances are evident only for some structures at ultra-high resolution.

Protein structure quality assessment

Protein structure quality assessment using the conformation-dependent geometrical variability



http://study.ibb.cnr.it/quiproqua/index.php