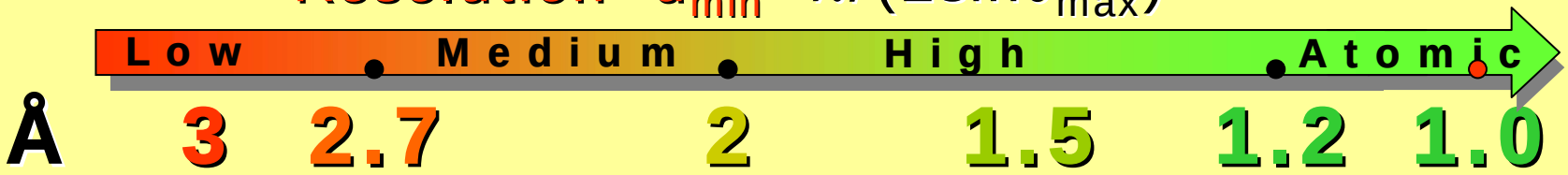
Ultra-high resolution structures in validation (and not only...)

Mariusz Jaskolski

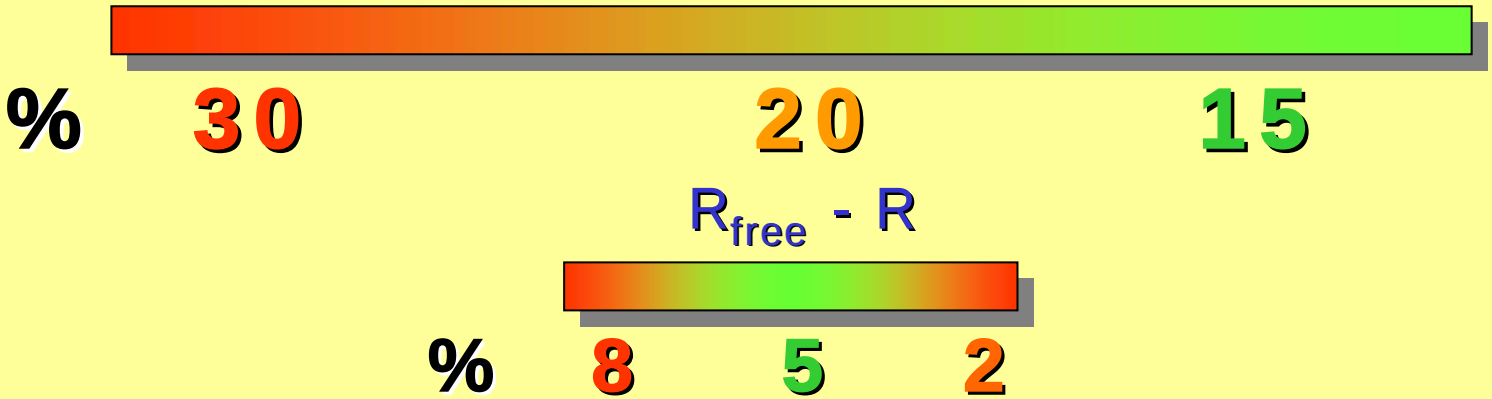
Department of Crystallography, A. Mickiewicz University
Center for Biocrystallographic Research,
Polish Academy of Sciences, Poznan, Poland

How to assess crystallographic structures?

Resolution $d_{min} = \lambda/(2\sin\theta_{max})$



R (residual), R_{free} , R_{sleep}: $R = \Sigma||F_o|-|F_c||/\Sigma|F_o|$



R_{free} - R

Deviations from „ideal" geometry

rms(bond-length – standard)

In **small-molecule crystallography**, **atomic resolution** is nearly always achieved and **atomic coordinates** are determined directly as peak coordinates in Fourier or difference-Fourier maps

In **macromolecular crystallography**, this is rarely the case; the **primary product** is a (lower resolution) **electron density map**, and an **atomic model** of a macromolecule is (at least to some degree) a result of its **subjective interpretation** by a crystallographer

small-molecule crystallography

(full Cu sphere, λ=1.54 Å)
d_{min} = 0.77 Å

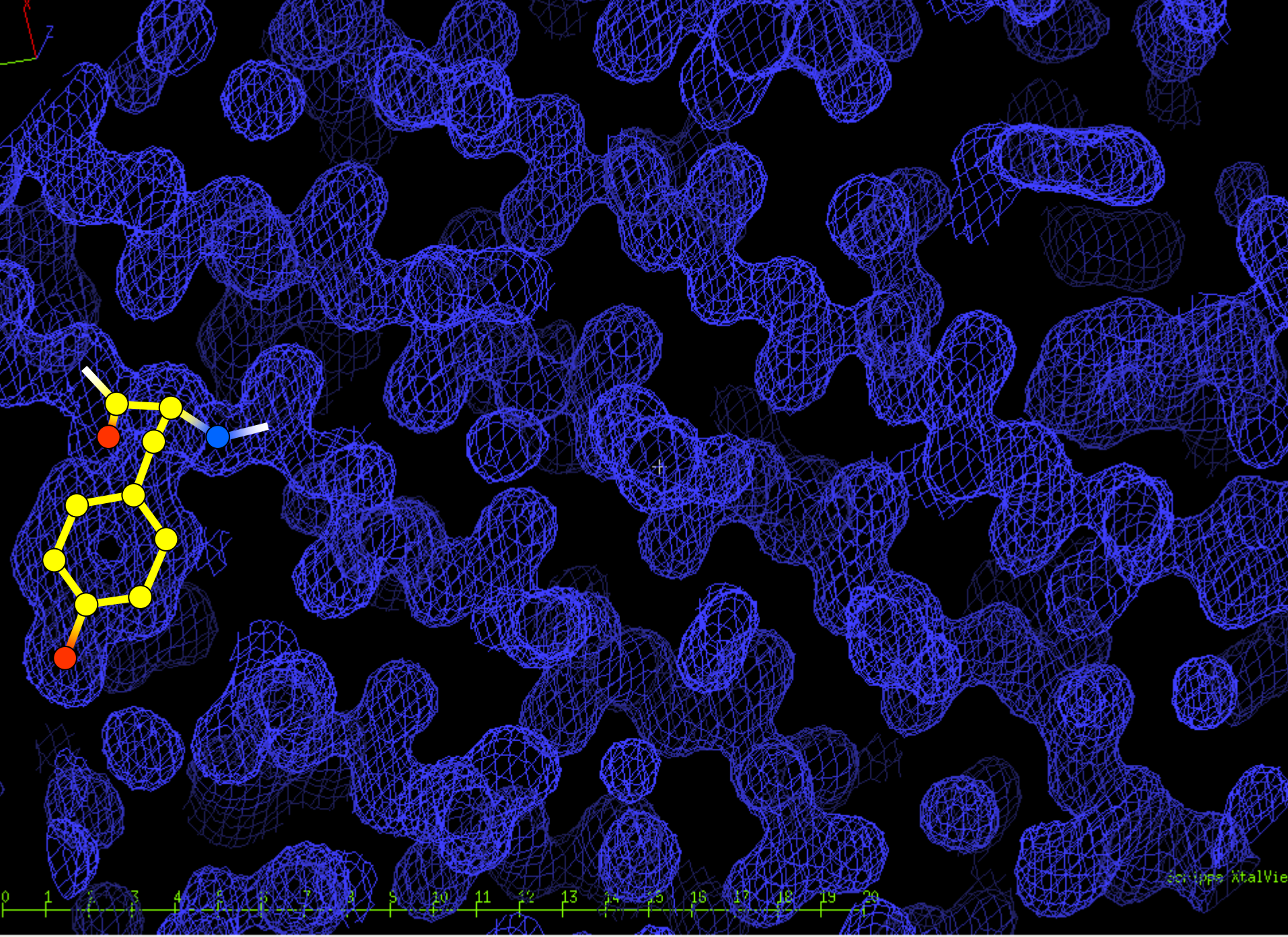
Mo radiation, λ=0.71 Å
$2\theta_{max}$ = 60º
d_{min} = 0.71 Å

macromolecular crystallography

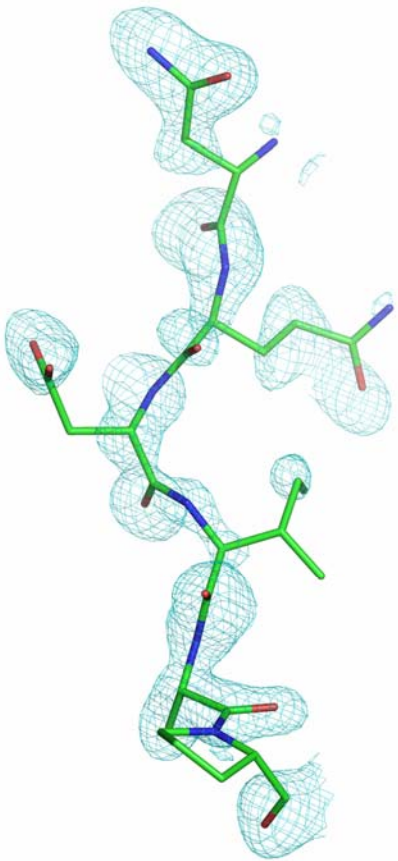
d_{min} = 1.2 Å (Sheldrick criterion)
(atomic resolution)

already a big success

but even if atomic resolution is not achieved – atomic models can be built


Shell - Konsole

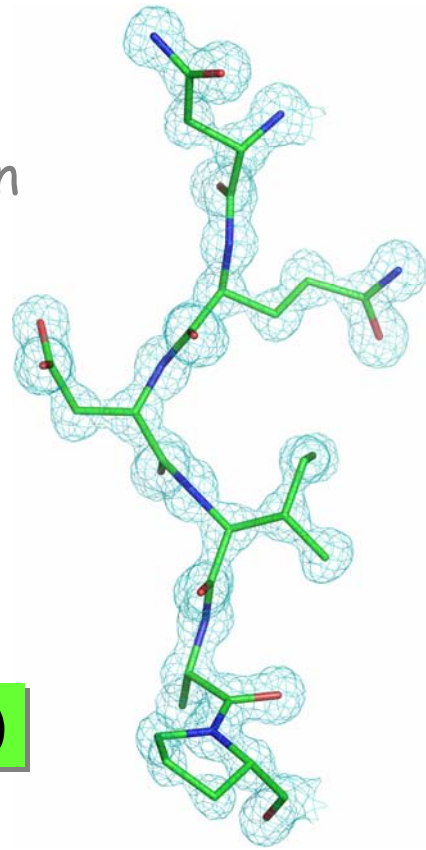
Not all map interpretations are non-problematic, errors are inevitable.
They can be insignificant, trivial, serious, severe or
even devastating or (extremely rarely) intentional (fraud)



a degree of subjectivity
is involved in map interpretation
especially with poor data
therefore

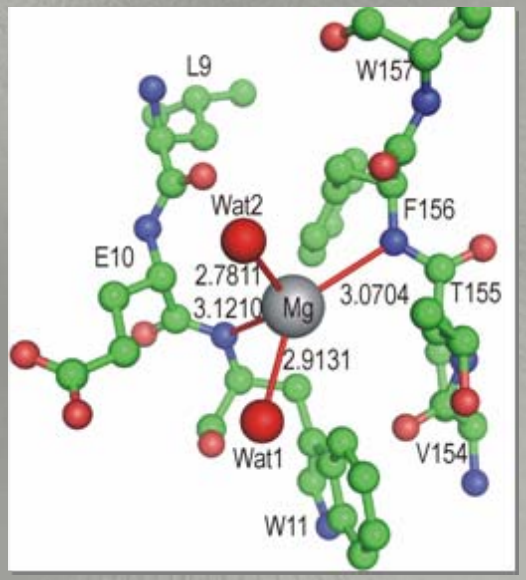
**it is so essential
that crystallographers
should be well trained**

(for instance, at Como Schools)



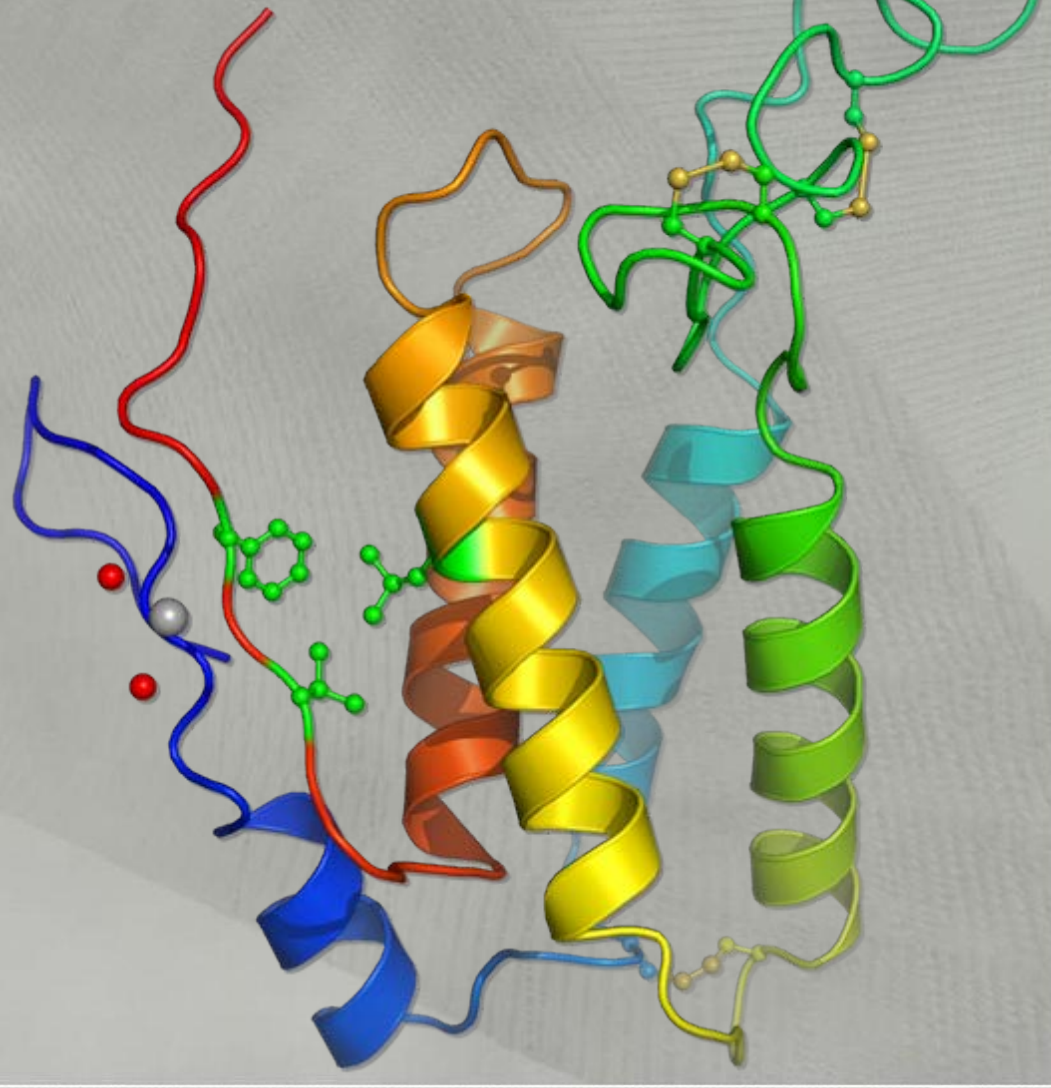
figures courtesy of Z. Dauter

Appearances may be deceptive.....



- unrealistic apolar "active site"
- "metal site" without proper coordination by protein
- funny placement of "S-S" bridges

Frankensteinase



Error types

- scientific fraud, data fabrication (extremely rare)
- totally wrong model (honest error), e.g. ABC transporters, RuBisCO subunit
- wrong connections between secondary structure elements
- register error – sequence shift
- wrong residue assignment
- wrong side chain conformation
- wrong **metal**/water assignment
- unjustified solvent modeling
- unjustified/unreasonable **B-factors**
- unjustified modeling of H atoms
- fictitious modeling in bad **maps at low contour**

Error source

- **paucity of data** (reflections) - model "overinterprets" available data
- bad data quality
- negligence of experimenter

Observation/parameter ratio in protein crystallography

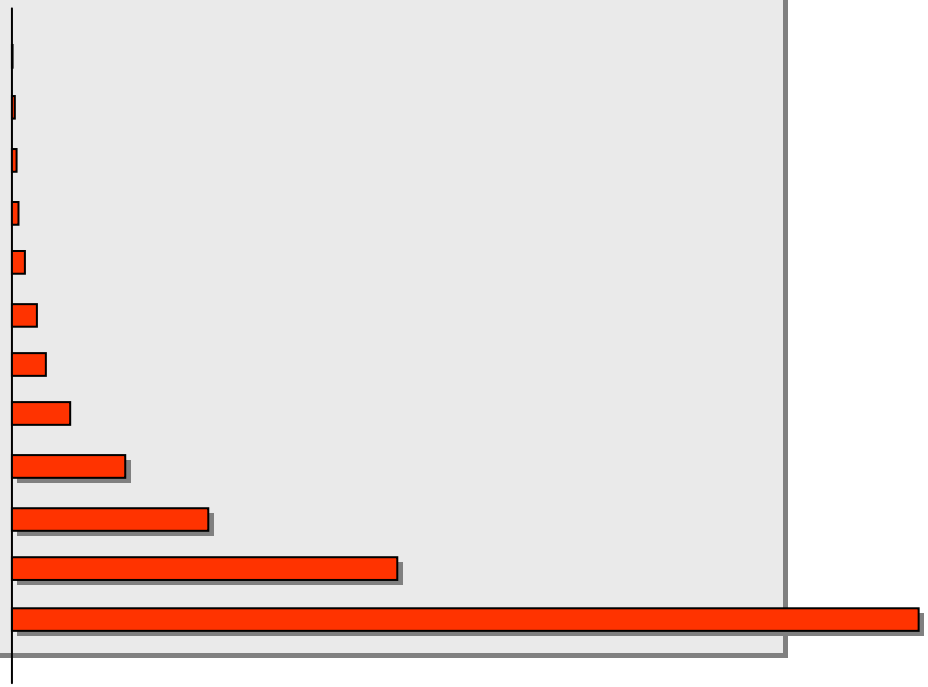
Typical protein crystal:

Matthews volume:	Vm=2.1 Å^3/Da → 40% solvent
protein in ASU:	500 aa (60 kDa) → 4000 non-H (5000 H) atoms
H_2O molecules in ASU:	1500

model parameters:	12 000 non-H coordinates, Biso fixed (no water)
(assuming no disorder)	**16 000** non-H coordinates + individual Biso (no H_2O)
	20 000 non-H coordinates + individual Biso (1000 H_2O)
	45 000 non-H coordinates + Baniso (1000 H_2O)
	64 500 aniso non-H + H coordinates/Biso fixed (1500 H_2O)
	70 000 aniso non-H + iso H (1500 H_2O)
	(*152 000 multipole model*)

unique reflections:

6.0 Å	1200
4.0 Å	4000
3.5 Å	6100
3.0 Å	9700
2.5 Å	17000
2.0 Å	33000
1.8 Å	45000
1.5 Å	77500
1.2 Å	150000
1.0 Å	260000
0.8 Å	510000
0.6 Å	1200000



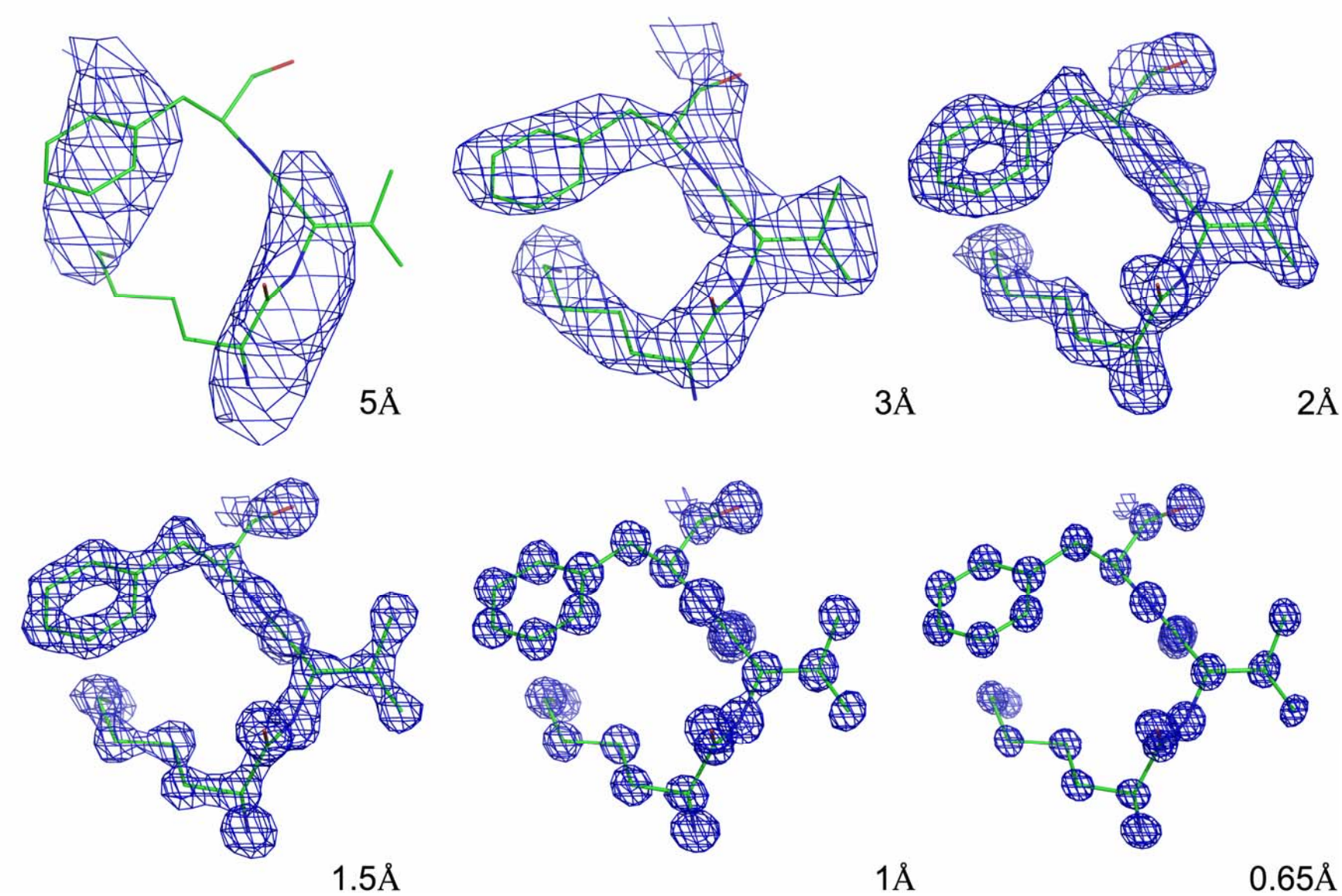
hkl **not enough for high (or any, for that matter) resolution**

criteria for high-resolution limit of experimental data
(parameters for high-resolution data shell)

% completeness at least	80*
(data with $I > 2\sigma(I)$	50%)
$\langle I/\sigma(I) \rangle$	2.0

*but don't reject reflections in sparsely populated data shells!
each and every one reflection is very precious;
the true "optical resolution" will be simply lower

Model building at low and high resolution



figures courtesy of Z. Dauter

Resolution of electron-density maps and the corresponding model detail

Resolution	Features Observed
5.0 Å	Overall shape of the molecule
3.5 Å	Cα trace
3.0 Å	Side chains
2.7 Å	Carbonyl O atoms (bulges) First chance to see water molecules
2.5 Å	Side chains well resolved, Peptide bond plane resolved
1.4 Å	Holes in aromatic rings Anisotropic B-factors possible
1.2 Å	True atomic resolution
1.0 Å	First chance to see H atoms
0.6 Å	Current limit for best protein crystals

Structure Refinement

may also involve elements of subjectivity in choosing a refinement path and especially in manual model rebuilding, solvent interpretation, etc.

Minimize (e.g. in LSQ) the following function:

$$S1=\Sigma w(|F_o|-|F_c|)^2$$

Notorious **problem** in macromolecular refinement, especially at low resolution: insufficient amount of data

Solution: reduce number of parameters (constraints) or (better) increase number of "observations" (**restraints**)

Stereochemical restraints as additional equations:

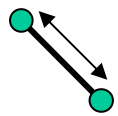
$$S2=\Sigma(\sigma)^{-2}[p(ideal)-p(calc)]^2$$

weight

stereochemical target

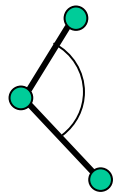
(analogy to "energy" minimization)

Types of stereochemical restraints

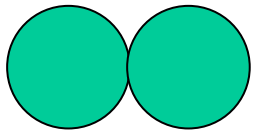


bond lengths

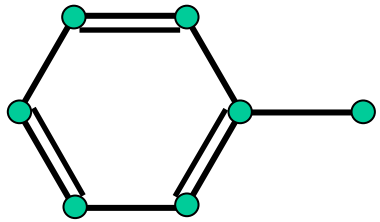
Bond lengths, angles, and other parameters that must be reproduced by the model



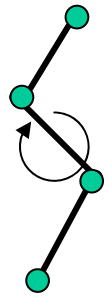
bond angles



van der Waals contacts



planar groups



torsion angles

The source of restraints used in almost all refinement programs

Table 2. *Bond parameters*

Bond type	σ	Bond length (Å)
C5W–CW	0·018	1·433
CW–C[illegible]	0·017	1·409
C[illegible]–CH[illegible]	0·021	1·525
C[illegible]–CH[illegible]E	0·014	1·497
CW–CH2E	0·031	1·498
CF–CH2E	0·023	1·502
CY–CH2E	0·022	1·512
C–CH[illegible]	0·025	1·516
CN–C[illegible]2E	0·019	1·5[illegible]
C[illegible]	[illegible]	1·5[illegible]
C5W–CR1E	0·025	1·36[illegible]
CY–CR1E	0·026	1·39[illegible]
CW–CR1W	0·021	1·394
CF–CR1E	0·021	1·384
CY–CR1E	0·021	1·389
CY2–CR1E	0·024	1·378
C5–CR1H	0·011	1·354
C5–CR1E	0·011	1·356
C–N	0·016	1·341
C–NC2	0·018	1·326
C5–NH1	0·011	1·378
CW–NH1	0·011	1·370
C–NH1	0·014	1·329
C–NH2	0·021	1·328
C5–NR	[illegible]	1·371
[illegible]	[illegible]	1·231
[illegible]	[illegible]	1·208
[illegible]	[illegible]	1·249
[illegible]	[illegible]	1·376
[illegible]	[illegible]	1·304
[illegible]	[illegible]	1·540
[illegible]	[illegible]	1·530
[illegible]	[illegible]	1·521
[illegible]	[illegible]	1·466
[illegible]	[illegible]	1·458
[illegible]	[illegible]	1·491
[illegible]	[illegible]	1·433
[illegible]	[illegible]	1·520
[illegible]	[illegible]	1·492
[illegible]	[illegible]	1·503
[illegible]	[illegible]	1·513
[illegible]	[illegible]	1·473
[illegible]	[illegible]	1·451
CH2E–NH1	0·018	1·460
CH3E–NH1	0·018	1·460
CH2E–NH3	0·030	1·489
CH2E–OH1	0·020	1·417
CH2E–S	0·020	1·822
CH2E–SM	0·034	1·803
CH2E–SH1E	0·033	1·808
CH3E–SM	0·059	1·791
CR1E–CR1E	0·030	1·382
CR1E–CR1W	0·025	1·400
CR1W–CR1W	0·019	1·368
CR1E–NH1	0·021	1·374
CRH–NH1	0·020	1·345
CRHH–NH1	0·010	1·321
CR1H–NH1	0·011	1·374
CRH–NR	0·013	1·319

Acta Cryst. (1991). A**47**, 392–400

Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement

By Richard A. Engh and Robert Huber

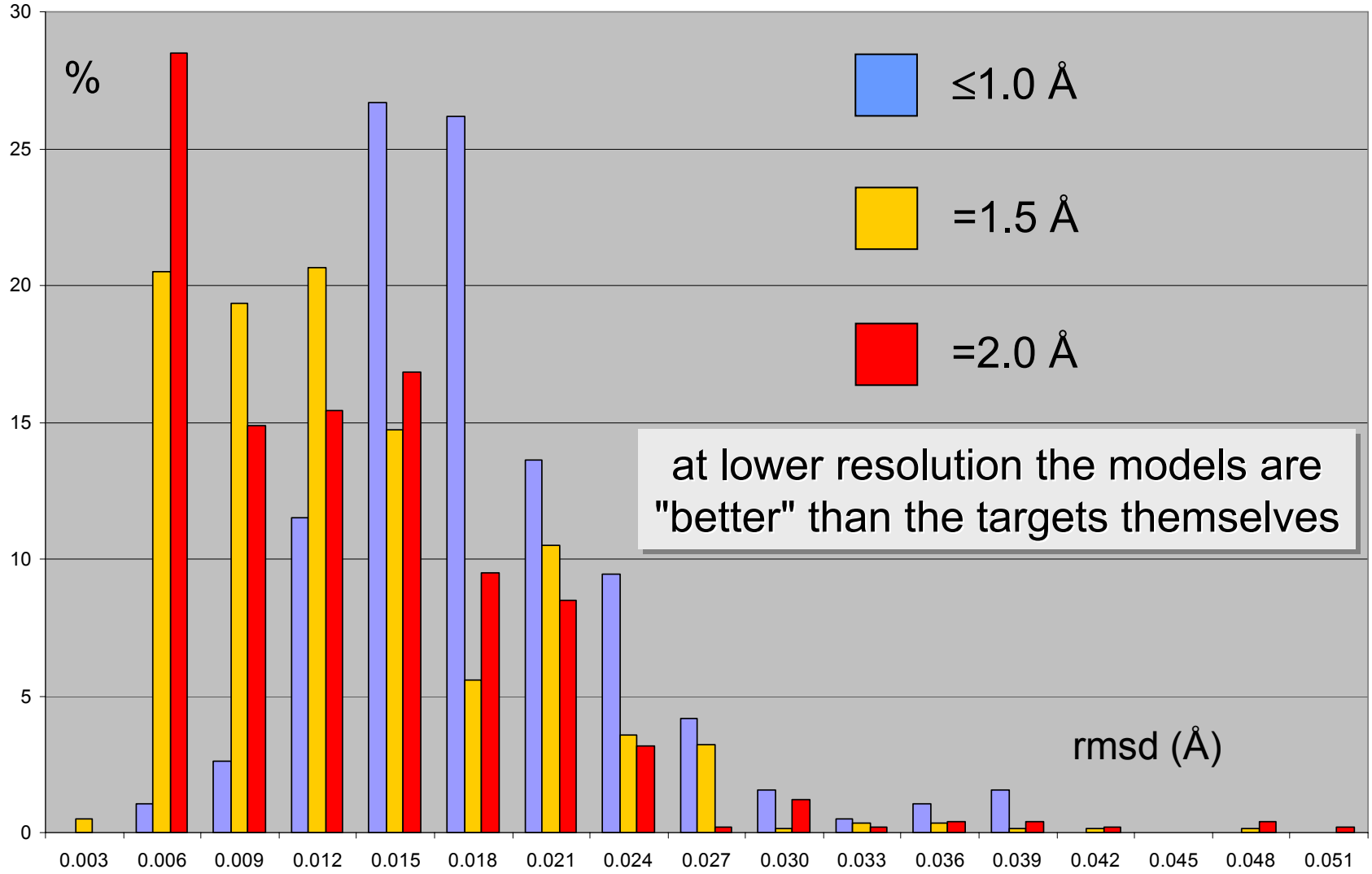
Max Planck Institut für Biochemie, D8033 Martinsried bei München, Germany

(*Received* 14 *November* 1990; *accepted* 17 *January* 1991)

derived from
small-molecule structures
(CSD)

Average σ
0.022 Å

Rmsd of bond lengths in PDB-deposited structures



The influence of restraints on the final model

- guarantee stereochemical correctness of the model
- restraints absolutely necessary at low resolution
- with increasing resolution compete for control of refinement with diffraction terms
- should not be enforced more strictly than warranted by *their* errors
- at atomic resolution diffraction terms dominate anyway, rmsd(b)=0.03 Å normal
- at atomic resolution can be restricted to flexible parts of the model only
- their weight should be coupled with B/occ of affected atoms
- are the targets reliable? New info from CSD *and* PDB
- are their standard uncertainties (weights) reliable?

Expansion of structural databases

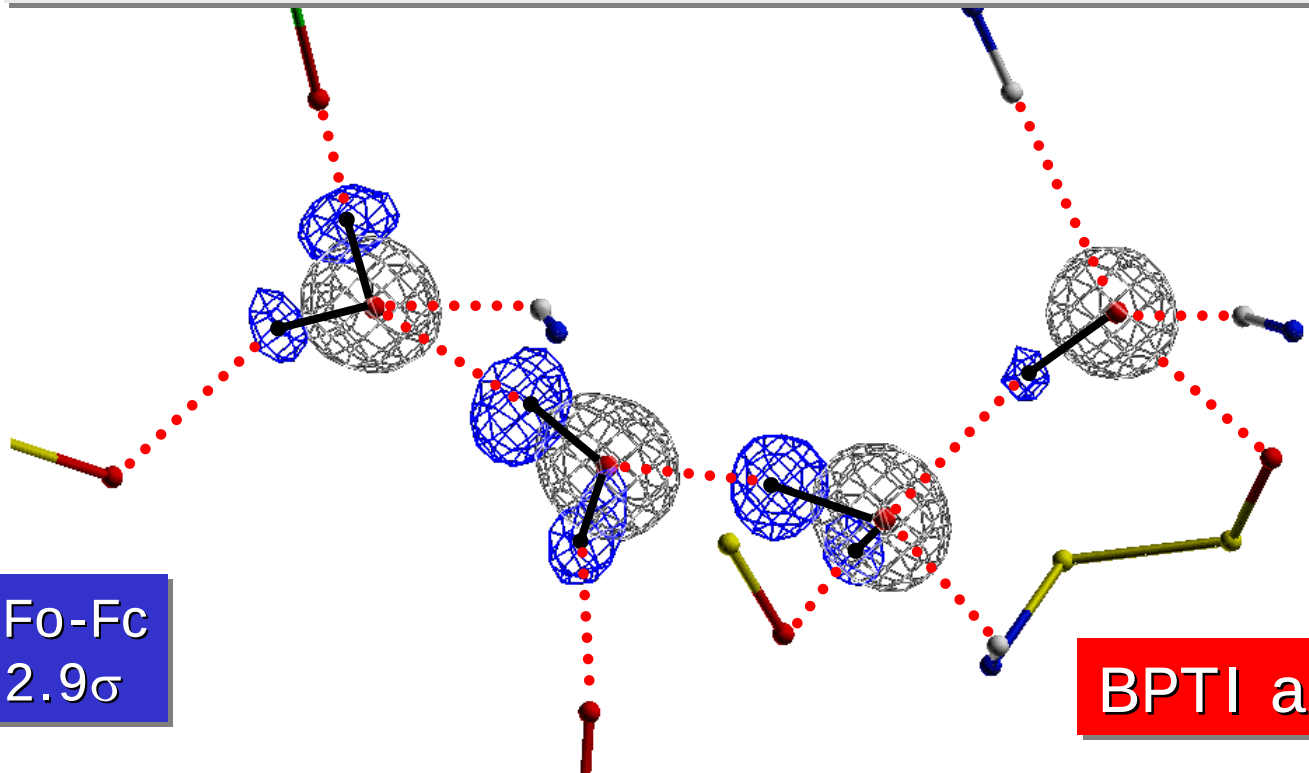
	Jan 1991	Sep 2008	Increase
Cambridge Structural Database (CSD)	~80000	~480000	6x
PDB	709	>53000	~75x
atomic	15	~870	~60x
0.8 Å	1997→ 0	21	∞

Availability of ultra-high resolution protein structures

- at 1.0 Å number of data sufficient for unrestrained refinement
- but there is increased level of resolved multiple conformations (more params.)
- unrestrained refinement of best regions (main-chain) possible
- restraints still necessary in fragments poorly defined by diffraction (disorder)
- diffraction terms dominate refinement
- full-matrix and estimated standard uncertainties possible
- rmsd(target) values reflect combined error of model and targets
- 0.02-0.03 Å deviations from target bonds acceptable

Importance of high resolution

- confidence in unusual model features
- accurate modeling of multiple conformations
- possibility to "see" H atoms (essential for enzymes etc.)
- removal of model bias (better observation/parameter ratio)
- "protein stereochemistry from proteins"



2Fo-Fc 2.0σ

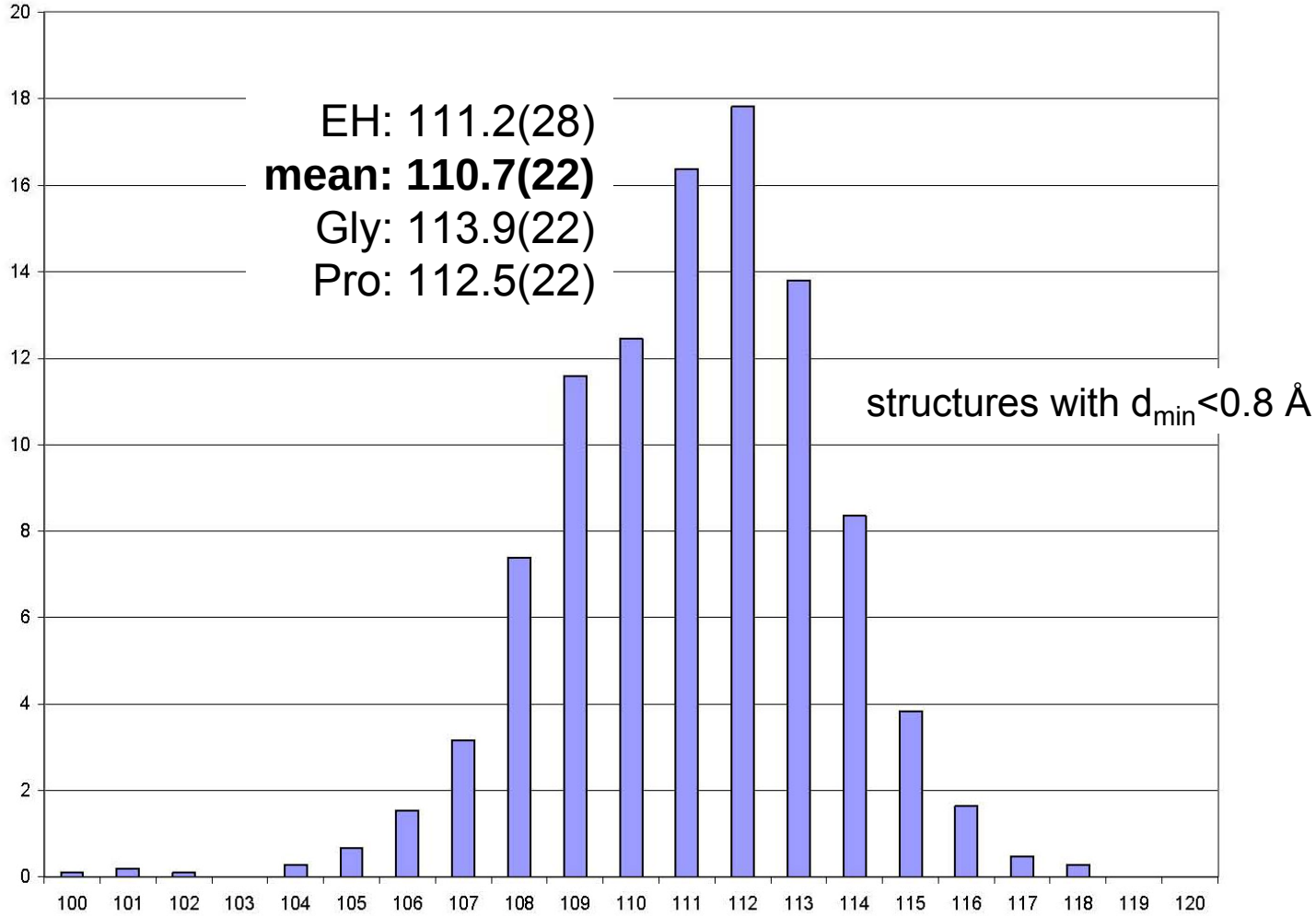
Fo-Fc 2.9σ

BPTI at 0.86 Å

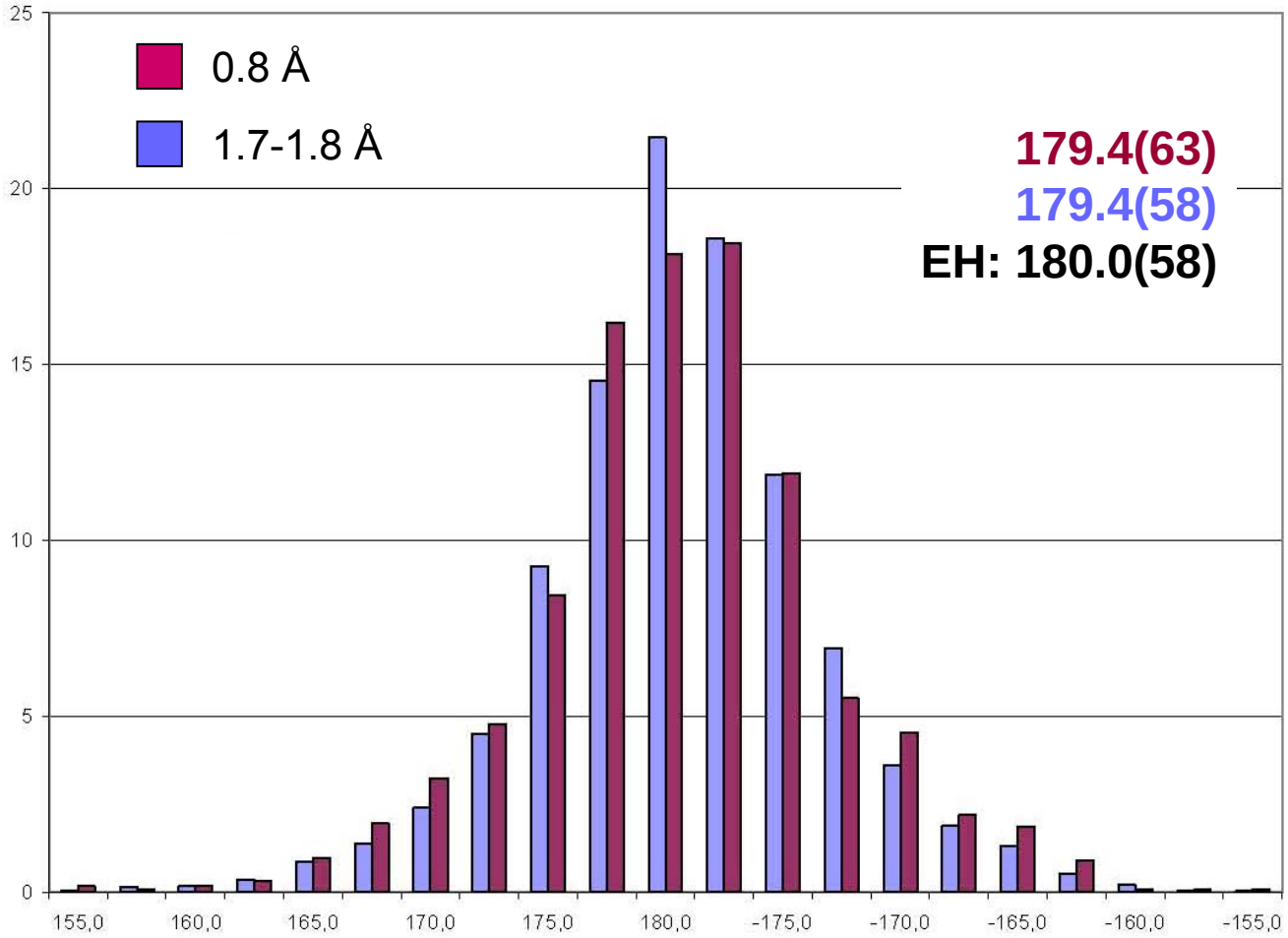
Ten highest-resolution structures in PDB

	1EJG	1UCS	1US0	1YK4	1R6J	1HJE	3AL1	2B97	1GCI	1X6Z	<d>(sd)	EH
Resolution (Å)	0.54	0.62	0.66	0.69	0.73	0.75	0.75	0.75	0.78	0.78		
R/R_{free}	0.090/ 0.094	0.137/ 0.155	0.094/ 0.103	0.100/ 0.108	0.075/ 0.087	0.127/ xxx	0.130/ 0.145	0.130/ 0.148	0.099/ 0.103	0.143/ 0.157		
Rmsd(d) (Å)	0.022 0.011	0.015 0.014	0.067 0.014	0.022 0.018	0.016 0.012	0.023 0.011	0.036 0.016	0.028 0.027	0.016 0.014	0.019 0.017	all atoms no disorder, B<40Å^2	
N-Cα (Å)	1.460(28) 1.456(11)	1.459(10) 1.455(10)	1.455(8) 1.453(9)	1.467(25) 1.462(9)	1.455(9) 1.454(8)	1.463(26) 1.453(6)	1.454(8) 1.454(8)	1.455(23) 1.455(23)	1.457(10) 1.457(9)	1.449(13) 1.449(13)	1.456(15) 1.454(12)	1.458(19)
Cα-C (Å)	1.529(13) 1.530(9)	1.528(11) 1.528(10	1.524(11) 1.524(10)	1.531(22) 1.534(11)	1.525(9) 1.525(10)	1.517(22) 1.524(13)	1.519(8) 1.519(8)	1.532(22) 1.532(21)	1.527(13) 1.528(10)	1.523(11) 1.523(11)	1.526(14) 1.527(13)	1.525(21)
C-N (Å)	1.336(10) 1.337(9)	1.336(12) 1.337(11)	1.333(15) 1.334(10)	1.336(12) 1.337(12)	1.333(13) 1.334(9)	1.334(11) 1.330(9)	1.333(8) 1.333(8)	1.333(38) 1.334(35)	1.336(11) 1.336(10)	1.331(12) 1.331(12)	1.334(18) 1.334(13)	1.329(14)
C=O (Å)	1.237(15) 1.234(7)	1.235(12) 1.235(12)	1.231(11) 1.231(9)	1.240(17) 1.240(10)	1.234(10) 1.234(10)	1.233(15) 1.229(9)	1.229(6) 1.229(6)	1.229(18) 1.229(18)	1.237(9) 1.237(9)	1.236(13) 1.236(13)	1.234(13) 1.234(12)	1.231(20)
long/short(+/-)				+			-			-		
Remark	Molly					unrestr.	L/D			Refmac		

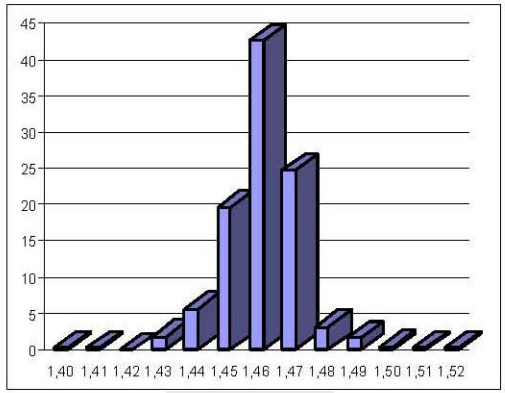
N-Cα-C angles (°) in PDB



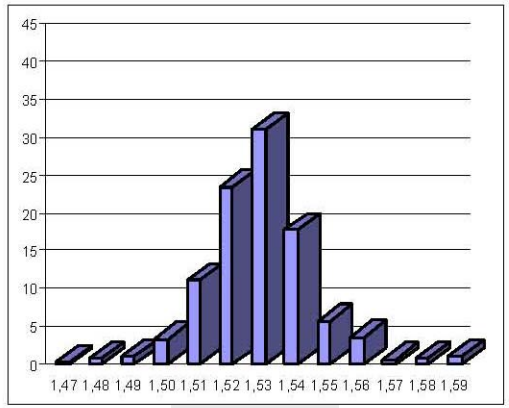
Torsion angles ω (°) in PDB



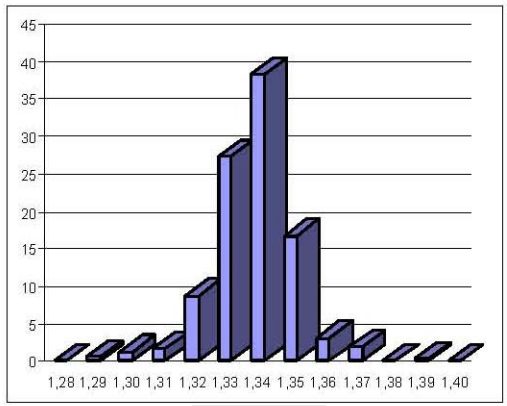
Peptide bonds (Å) in current CSD (R<0.075)



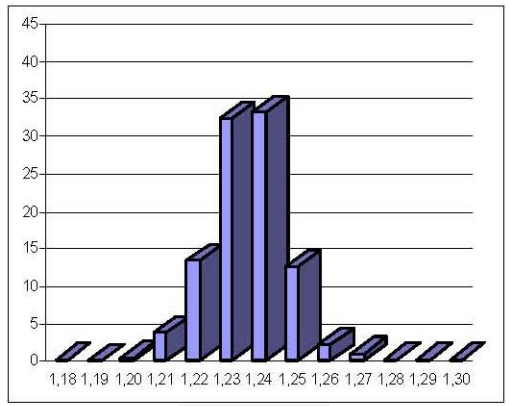
N-Cα



Cα-C



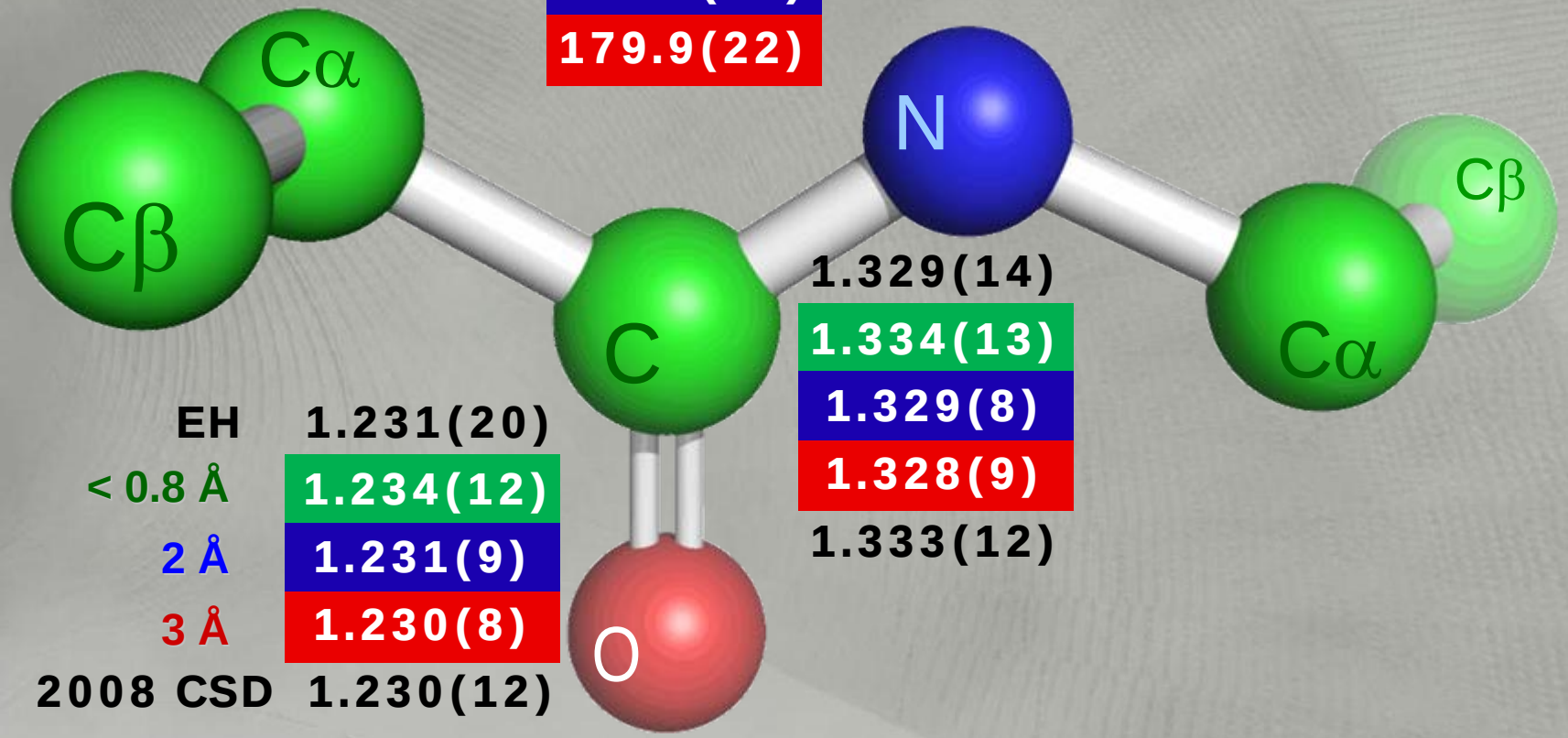
C-N



C=O

Summary of results (Å, °)

ω	180.0(58)
	179.4(63)
	179.8(58)
	179.9(22)



	1.329(14)
	1.334(13)
	1.329(8)
	1.328(9)
	1.333(12)

EH	1.231(20)
< 0.8 Å	1.234(12)
2 Å	1.231(9)
3 Å	1.230(8)
2008 CSD	1.230(12)

Take-home messages

- Engh & Huber parameters are generally correct but some small adjustments are indicated
- making the restraints too tight gives the *impression* that the structure is more accurate; it simply fits the parameters more precisely
- a correctly set restraint weight will give rmsd for bonds about 0.015-0.020 Å
- at ultra-high resolution the diffraction terms should be allowed to dominate; the rmsd value can be even 0.02-0.03 Å
- a single grossly erroneous parameter can severely bias rmsd

Collaboration

Alex Wlodawer
Zbyszek Dauter

NCI Frederick
NCI/APS Argonne

Mirek Gilski

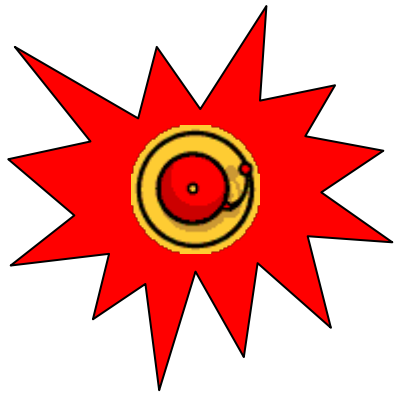
A.Mickiewicz Univ. Poznan

want to learn more? Click here...

Sample experimental table for well and poorly determined structure

Data collection and processing statistics		
Space group	$P4_12_12$	P41212
Unit cell paramters (Å,°)	a=133.1, c=223.1	133.0978, 133.0978, 223.1445 90.0000, 90.0000, 90.0000
Temperature (K)	100	
Resolution (Å)	50-1.40 (1.45-1.40)	6.00-3.234 (3.5-3.234)
Completeness (%)	100(97.6)	76.2(13.1)
<I/σ(I)>	45.2 (2.1)	3.21 (1.3)
R_{merge} (%)	4.6 (36.6)	13.9 (177.9)
Refinement statistics		
No. working set/test reflections	433398/1331	24501/2731
R/R_{free} (%)	13.8/18.6	27.69/38.21
No. water molecules / waters	611	998
Rmsd from ideal bonds (Å)	0.018	0.037
<B> protein/water ($Å^2$)	21/28	89.856/85.32
Ramachandran most favored/allowed φ/ψ (%)	93.2/6.8	77/14.72
PDB code	2GUD	-

Warning signs



- high R (>25%)
- large R_{free}-R (>7%)
- large rmsd(bonds) (>0.03 Å)
- violation of chemical common sense
- unusual B-factors (outside 4<B<60 $Å^2$)
- high fluctuations of B-factors
- too many water molecules (>(3-resol.) per residue)
- poor Ramachandran statistics (<90% favored; disallowed angles)
- zero-occupancy atoms
- funny metal/water sites
- H atoms pretending to be part of the model
- too low contour of electron density (2Fo-Fc: <1σ, Fo-Fc: <2.5σ)
- reporting/discussing numbers with unjustified precision (1.2111 Å)