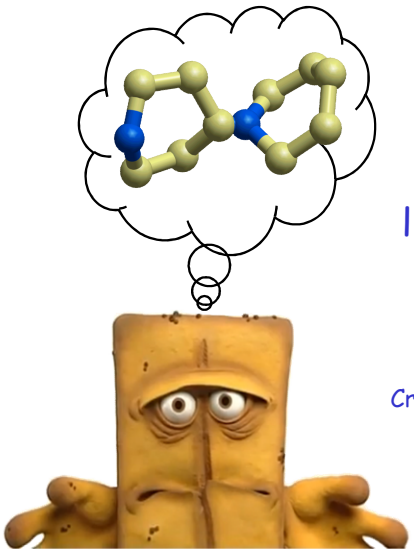
Protein-Ligand Complexes: Only Seeing Is Believing



About technical reasons and human factors for routinely encountered difficulties in protein-ligand structure modelling

Bernhard Rupp
Med. Univ. Innsbruck, Austria
Crystallographiae Vindicis Militum Ordo
San Diego, CA
http://www.hofkristallamt.org/

What is so special about crystallography?

- For example: compare to cell biology experiments

a) we have an enormous amount of data per experiment – several 10,000 to 100,000s of data points

b) we have very clear expectations about the principles of stereochemistry for each ultimate experimental outcome (structure model), again from 100,000s of crystal structures

c) as a consequence, we can reasonably apply and use statistics. Let's start.

A simple statistical student test (t)

a) Statistics show that 93% of all talks reporting statistics are perceived as more important than those without….

b) Smart Audience: How may of you believe they are above average – and how many below ?

c) Epic fail. CLT says 50% of a population will be below and 50% above the sample mean (avg)

d) What have we learned: Non-empirical, cognitive factors or biases can enter scientific judgement:

The human mind is its own greatest enemy
(Francis Bacon, 1620)

Ok, then:

The average human being has :

a) One breast and one testicle

b) Epic fail. Averages deceive (even when the law of large numbers remains valid) when the population is pathologically non-normal…

c) What have we learned:

Never trust a statistic(ian).

Scoring ligand quality is non-trivial

- Born out of curiosity and anecdotal evidence [1,2]
- Simple real space correlation coefficient based score *S*
- In *Twilight*, bona fide small molecule ligands (no buffers, peptides (*Twilight II*), but includes glycans)

$$S = 2 / (\mathrm{RSCC} / 0.6 + RESOL / 1.3)$$

Various similar programs/metrics exist, e.g. VHELIBS [3] evaluates also binding site electron density

A summary of various measures and statistics: [4]

[1] Pozharski, E., Weichenberger, C.X. & Rupp, B. (2013). *Acta Crystallogr.* **D69**, 150-167.

[2] Weichenberger, C.X., Pozharski, E. & Rupp, B. (2013). *Acta Crystallogr.* **F69**, 195-200.

[3] Cereto-Massague, A., Ojeda, M.J., Joosten, R.P., Valls, C., Mulero, M., Salvado, M.J., Arola-Arnal, A., Arola, L., Garcia-Vallve, S. & Pujadas, G. (2013). J. Cheminform. 5, 36.

[4] Deller, M.C. & Rupp, B. (2015). *J. Comp. Aided Mol. Des.* **29**, 817-836.

The results are consistent:

(spoiler alert)

1. Most ligand models are of good quality/fit
2. Some interpretations qualify as mildly optimistic
3. And some can be charitably described as delusional

Table 1 Electron density-based validation of protein–ligand models [4]

Scores			Classification	
RSCC	% of structures	Predicted number of PDB	Twilight (abstains from judgement)	VHELIBS
1.0–0.9	67	~46,900	Ligand fits density	'Good'
<0.9–0.8	21	~14,700	Ligand fits density partially	'Dubious'
<0.8–0.7	7	~4,900	Significant parts of ligand not in density	'Bad'
<0.7–0.6	3	~2,100	Very poor fit of ligand to density	
<0.6–0.5	1	~700	Improbably poor fit of ligand to density, density almost absent	
<0.5	1	~700	Catastrophic fit of ligand to density, density completely absent	

What do this numbers mean

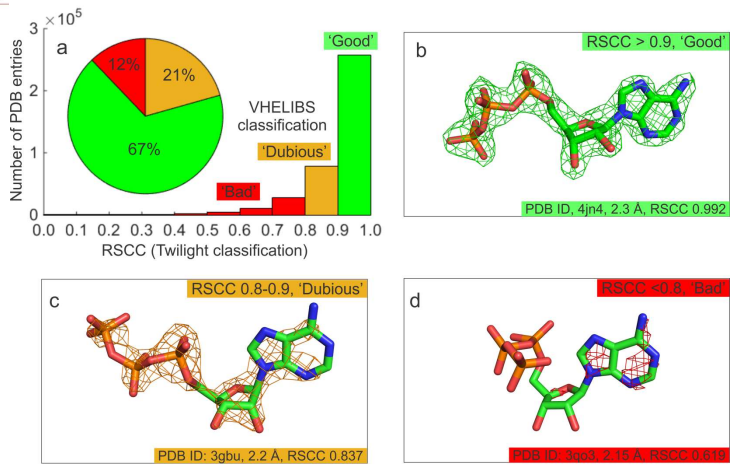


Figure 2. (a): Classification and distribution of RSCC values of ligands.
(b,c,d): Example for the same ligand at comparable resolutions with decreasing RSCC values, contoured at 2σ $2mF_o$-DF_c [4].

As always: resolution matters

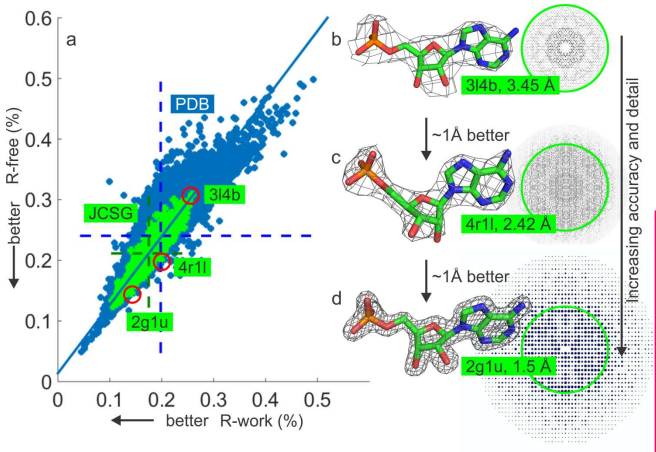


Figure 1. Global R-work and R-free for 85000 ligand structures: As always, the amount of experimental data determines model detail – also true for good (well-modelled) ligands [4].

It is to some degree subjective what a 'good' ligand structure model is: it very much depends on your intended use – e.g. just to demonstrate a ligand is there versus detailed lead compound improvement

Resolution, B-factor and RSCC correlate

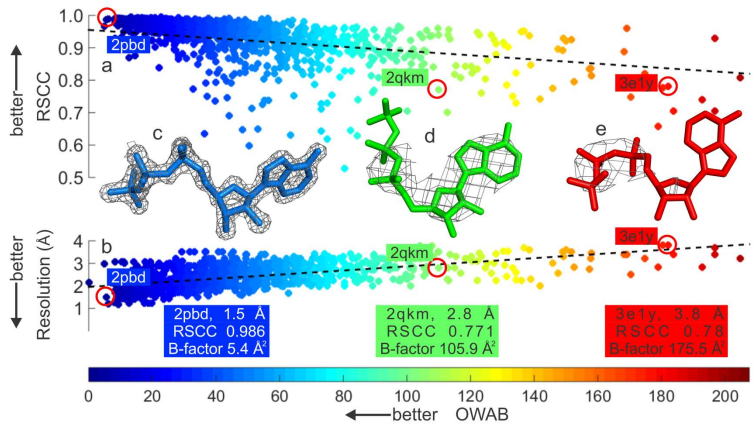


Figure 3. Same ligand, modelled at different resolutions (2σ $2mF_o$-DF_c). Not surprising, at better resolution the ligand generally will fit the density better – think what causes low resolution in the first place [4].

Many measures, metrics and statistics exist

Table 2 *Local* and *global* structure validation measures [4]

Validation measure		*Local*	*Global*	Reference
Quality and fit of ligand	Real Space Correlation Coefficient (RSCC)	●	●	N/A
	Real Space R-value (RSR)	●	●	[43, 44]
	Real Space Observed Density Z-Score (RSZO)	●	●	[40]
	Occupancy-Weighted Average B-factor (OWAB)	●		[15, 20]
	S score	●		[15]
	Q score	●		[13]
	Local Ligand Density Fit (LLDF)	●		http://bit.ly/1si1ZeL
	deviation of bond lengths	●		[35]
	deviation of bond angles	●		
Quality of model and data	R-value		●	N/A
	R-free		●	[38]
	CC*		●	[39]
	Real Space R-value (RSR)	●	●	[44]
	Diffraction-data Precision Indicator (DPI)		●	[90]
Quality of model and stereochemistry	RMSD/Z bond lengths	●	●	[35]
	RMSD/Z bond angles	●	●	
	G-factor	●		[91]
	Clashscore	●	●	[53, 54]
	Ramachandran outliers	●		[92]
Other measures of model and data quality	Volumetric packing scores	●	●	[93, 94]
	B-factor	●	●	[95, 96]
	Resolution		●	N/A

Local validation measures are metrics used to validate the quality of the ligand model, and *global* validation measures are used to assess the overall quality of the protein–ligand model. Many *local* measures can also be averaged to provide a (less informative) *global* measure

Local is a relative term – merging a local metric (say RSR for each atom of a ligand) into a global one means loss of information – most common reason for unjustified or at least not meaningful score (LLDF – is such a measure really local?).

Many validation programs can be used…

…but a bad model does not become any better by picking another validation suite/metric

Table 3 Protein-ligand validation software

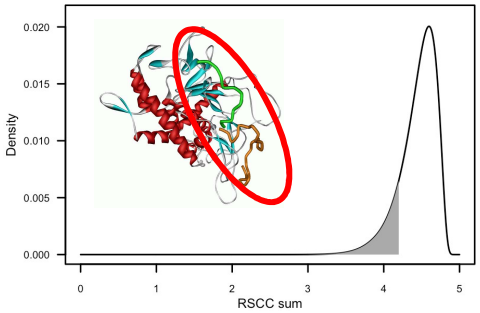
Validation	Package	Description	URL	Reference
Quality and fit of ligand	Twilight	Visual analysis of ligand density	http://bit.ly/1shcwu4	[12, 15]
	VHELIBS	Fit of ligands and binding sites	http://bit.ly/1t5szxl	[14]
	ValLigURL server	Compare conformations of ligands in the PDB	http://bit.ly/1v80yoS	[13]
	MotiveValidator and ValidatorDB	Interactive web-based validation of ligands and residues	http://bit.ly/1tNd7Vs	[70]
	PDB_REDO	Updated and optimised X-ray structure models and maps	http://bit.ly/1pVYQQA	[21, 22]
	wwPDB	wwPDB Validation Server	http://bit.ly/1xdfoZN and http://bit.ly/1si1ZeL	[45]
	PDB-CARE and CARP	Checks glycan nomenclature and stereochemistry	http://bit.ly/1pW9gQ0	[97, 98]
	LIGPLOT	Ligand–protein interaction diagrams	http://bit.ly/1qwZ7cS	[85]
	EDS	Electron density server	http://bit.ly/1Ddnlkg	[20]
	EDSTATS	Statistical quality indicators of electron density maps	N/A	[40]
	OVERLAPMAP	Average of two maps	http://bit.ly/ZZUSk3	[99]
	BUSTER	Refinement of proteins and ligands	http://bit.ly/1w8NX1Q	[100]
	WHAT_IF/WHAT_CHECK	Protein and ligand verification tools	http://bit.ly/1F0j19Y	[55, 101]

Software that is particularly useful for ligand validation is highlighted. For general protein structure validation software see Table 4

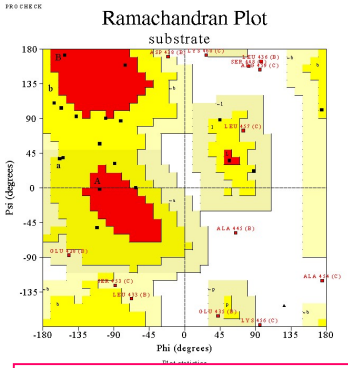
It is almost impossible to cheat this way, and if you have to, there is probably already a problem….

Peptide ligands are similar…

…with the advantage of simple stereochemistry



Probability density functions of RSCC **values and sums.** Graph of f_5, which is the p.d.f. of RSCC sums S_5, calculated by iterative convolutions of f_1. The gray area represents $P(S_n \leq 4.2)$, the probability to observe a sum of five RSCC values less than or equal to 4.2, which is given by integrating f_5 from $-\infty$ to 4.2, which is equivalent to evaluating the respective cumulative distribution function $F(4.2) = 0.1426$. The distribution of the random variable S_5 peaks at 4.6 and $F(4.6) = P(S_n \leq 4.6) = 0.71$.



NB: A random coil peptide does not have a random backbone torsion angle distribution – this is an energy surface!

Why are about 1/3 of protein-ligand models less than excellent?

(A) Technical & chemical subtleties

Crystallography is difficult. Still. And entropy works against you. Always.

(B) Human factors

The human understanding is not composed of dry light, but is subject to influence from the will and the emotions, a fact that creates fanciful knowledge; (wo)man prefers to believe what (s)he wants to be true.

Francis Bacon, Novum Organum Scientiarum, Aphorism 49 (1620)

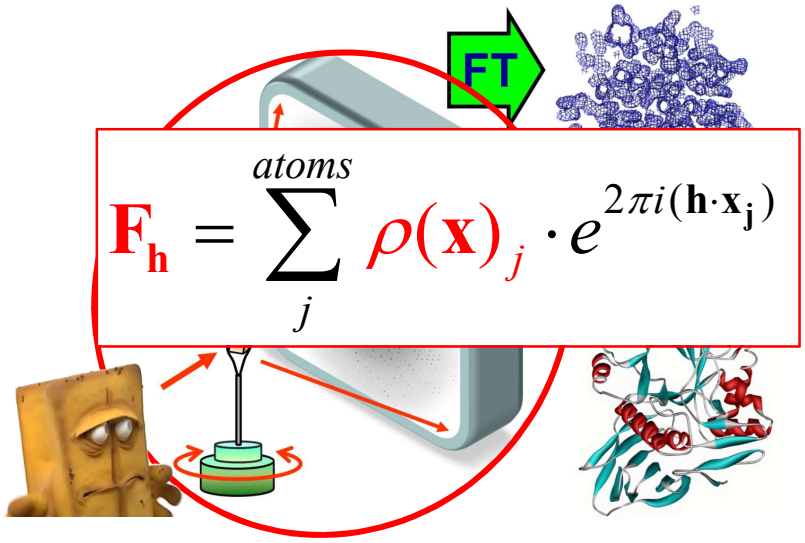
Science is a way of trying not to fool yourself: The first principle is that you must not fool yourself - and you are the easiest person to fool.

Richard Feynman (1974)

(A) Back to basics (i): What is crystallography?

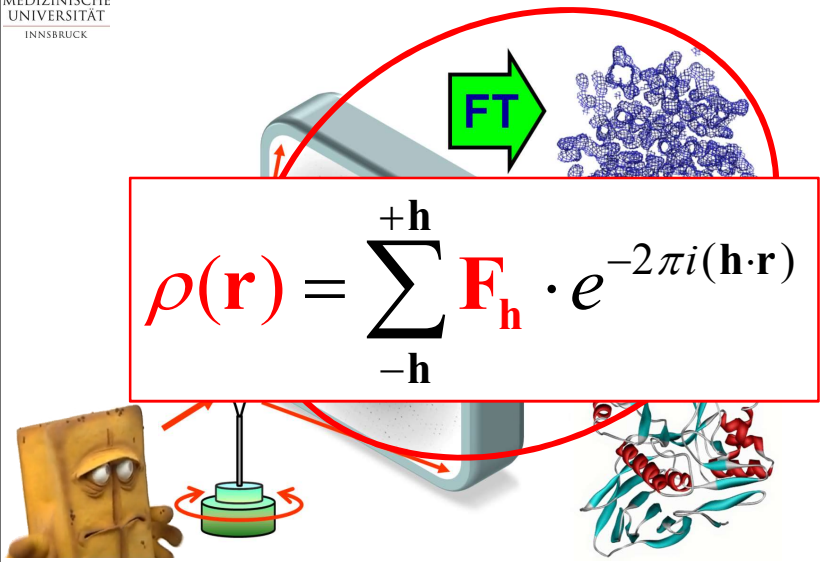
Crystallography: getting real space electron density from reciprocal space diffraction data

First we need to get good data



$$\mathbf{F_h} = \sum_{j}^{atoms} \rho(\mathbf{x})_j \cdot e^{2\pi i(\mathbf{h}\cdot\mathbf{x_j})}$$

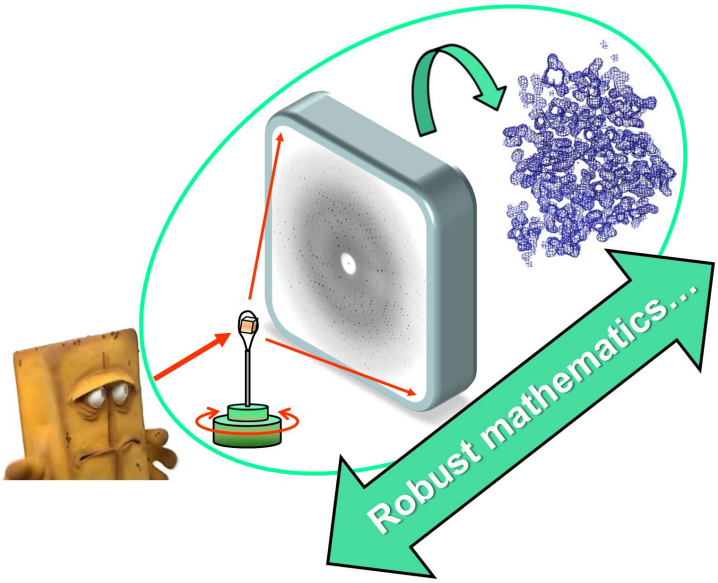
Then we reconstruct the electron density



$$\rho(\mathbf{r}) = \sum_{-\mathbf{h}}^{+\mathbf{h}} \mathbf{F_h} \cdot e^{-2\pi i(\mathbf{h}\cdot\mathbf{r})}$$

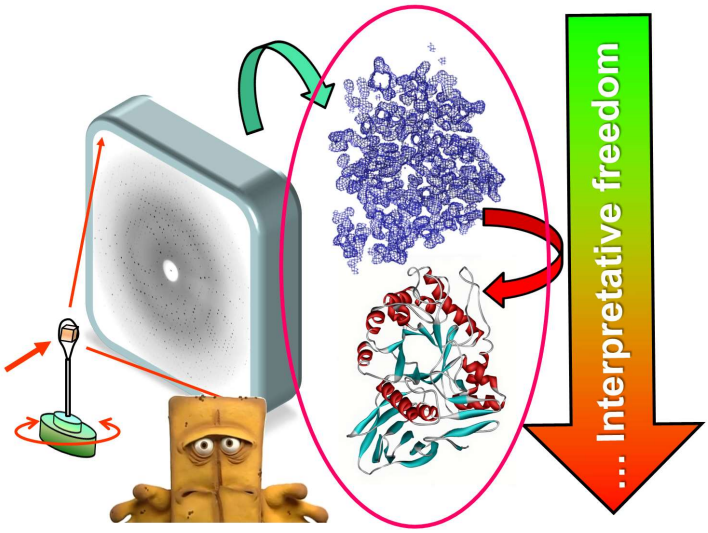
MEDIZINISCHE UNIVERSITÄT INNSBRUCK

So far, a fairly robust process



MEDIZINISCHE UNIVERSITÄT INNSBRUCK

Then we build a model – and the trouble starts…c.f. (B)



We need phases to complete the FT

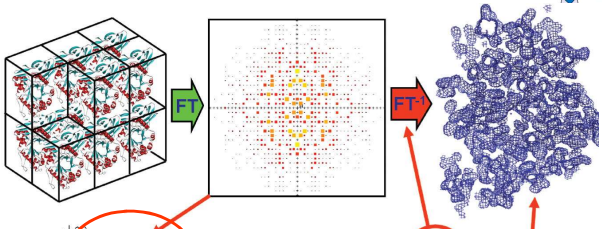


Figure 9-15 The crystallographic phase problem. The measurable component of the Fourier transform of the crystal is only the scalar structure factor amplitude $F(\mathbf{h})$ proportional to the square root of $I(\mathbf{h})$. The missing phases $\varphi(\mathbf{h})$ must be supplied by additional phasing experiments or in the form of model phases via molecular replacement. The two necessary Fourier coefficients in the back-transform formula are emphasized in blue.

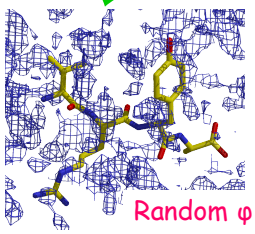
$$\sum_{\mathbf{h}=-\infty}^{+\infty} F(\mathbf{h}) \cdot \exp[-2\pi i(\mathbf{h}\cdot\mathbf{r}) + i\varphi(\mathbf{h})] = \rho(\mathbf{r})$$

(C) Bernhard Rupp 2010

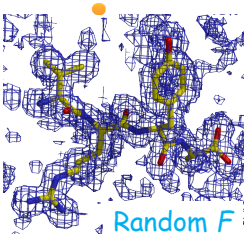
Which more important? Amplitudes or phases?

Phases dominate the Fourier reconstruction of the electron density map

But intensity differences determine quality of the phases!



Random φ



Random *F*

Electron density reconstruction

We can split the complex electron density summation into an amplitude part *F*

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{\mathbf{h}=+\infty} \mathbf{F}_{\mathbf{h}} \cdot e^{-2\pi i(\mathbf{hr})} = \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{\mathbf{h}=+\infty} F_{\mathbf{h}}(obs) \cdot \exp i\varphi_{\mathbf{h}}(calc) \cdot \exp\left[-2\pi i(\mathbf{hx})\right]$$

and a separate phase term φ, which are the so called Fourier coefficients for the map. Depending on the type of Fourier coefficients used there are different types of maps:

Fc, φc: calculated structure factor amplitudes, calculated phases
= model map

Fo, φc: observed structure factor amplitudes, calculated phases
= model biased map (more or less a model map)

mFo-DFc, φc: shows differences between 'observed' and calculated map

2Fo-Fc, φc: shows differences (Fo-Fc) over Fo map

2mFo-DFc, φwgt: as above, but with maximum likelihood (sigmaA) coeff.

Fo*fom, φ(ave): figure of merit weighted, phase averaged map

F^+-F^-, φ(exp): anomalous difference map with experimental phases

Truth

The choice of (diff) map type is important!

	F_o (data) with ligand contribution	F_o (data) without ligand contribution
F_c (model) with ligand contribution	No significant difference density (good)	Negative ligand difference density (bad)
F_c (model) without ligand contribution: omit ligand (or low occupancy and/or, high B factor)	Positive ligand difference density (good)	No or poor difference density (meaningless noise subtraction)
F_o without ligand contribution – isomorphous apo-structure	Positive ligand difference density (good)	No or noise difference density for ligand

So…what determines the **Fs** which we need to reconstruct electron density?

…on their position (determining phase)

$$\mathbf{F_h} = \sum_{j}^{atoms} n_j \cdot f_j(s) \cdot e^{i \cdot 2\pi(\mathbf{h} \cdot \mathbf{x_j})} \cdot e^{-B_j s^2}$$

One (!) single data point (diffraction spot)

…has information about ALL *j* atoms

…depends on their occupancy *n*

…..depends on their scattering power *f*

…and a probability measure *B* indicating whether the atom is really at its believed mean position

Consequences of the F summation

1) All atoms contribute to each **F**, therefore any global measure based on **F** can not indicate whether a few atoms are modelled correctly or not.

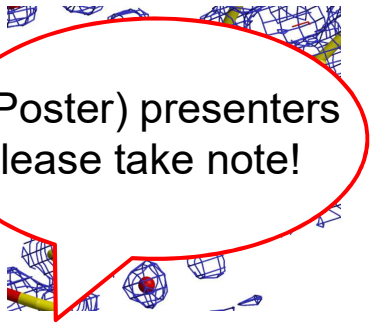
2) Because the occupancy of a ligand, out of principle, can never be 100% and often is significantly lower, the already small contribution of a few ligand atoms to F is further reduced.

3) Ligands are generally not covalently bound and have a large degree of conformational freedom, leading to increased positional uncertainty (larger B-factors) and additional reduction of contributions to F.

Global reciprocal space indicators of data fit like R-values are completely uninformative for ligand validation. Ligand scattering mass is often only 1/1000 of the protein. Combine this with high B-factors and partial occupancies and it becomes even worse (btw, good protein geometry alone means nil).

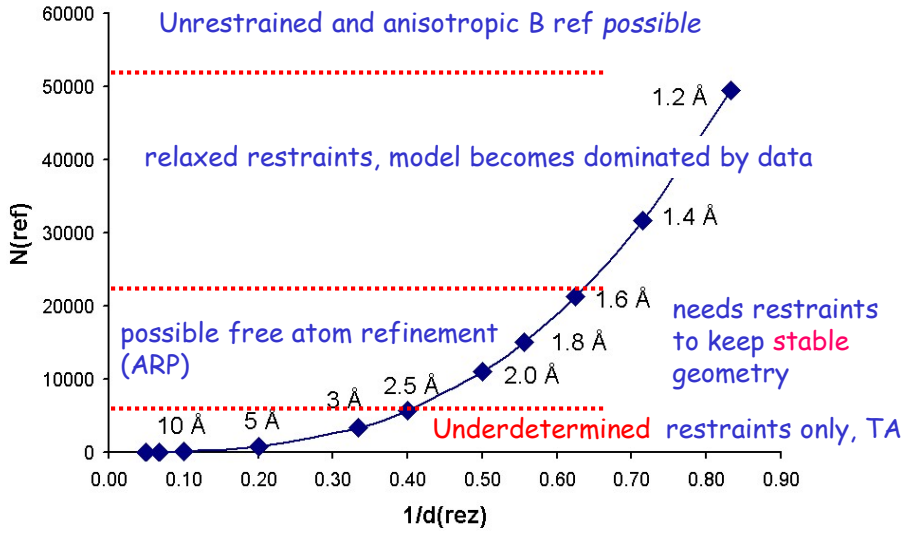
The need for local real space validation

As reciprocal space global indicators are essentially worthless, we need local (real space) indicators that show t fit between model and electro density. The electron density – preferably minimally biased positive omit difference density - is the primary evidence !



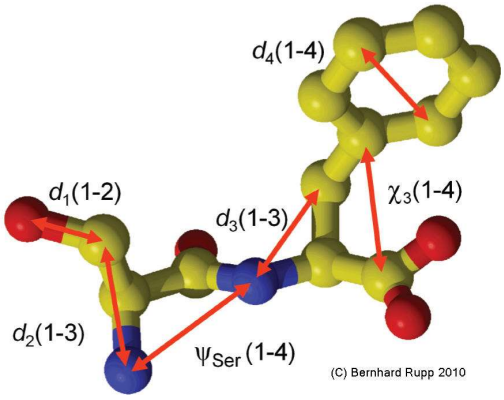
How does that work out in reality (i.e. the published structure models) ?

Technical subtleties (iii): you do not have enough data to freely refine: restraints



Stereochemical restraints add generally valid observations and keep the model plausible

Figure 12-12 Distance and torsion restraints. Bond distances, bond angles, and fixed dihedrals in planar groups are implemented in a common restraint function as specific 1-2, 1-3, and 1-4 distances (d_n). In contrast to dihedral angles d_n(1-4) restrained around one single value, torsions χ_n(1-4) can assume multiple preferred angles or minima. The special backbone torsions such as the ψ-torsion of the serine ψ_{ser} depend on the secondary structure environment and are therefore not, or only weakly, restrained through energy or vdW terms, depending on the type of refinement parameterization. The absence of any strong restraints makes the backbone φ/ψ torsions quite useful for "geometric cross-validation" in the Ramachandran plots.



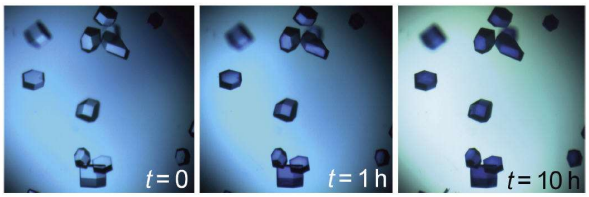
$$Q_{total} = w_{xray} \sum_{\mathbf{h}} \frac{\left(F_{\mathbf{h}}^{obs} - F_{\mathbf{h}}^{calc}(\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_{atoms})\right)^2}{\left(\sigma_{\mathbf{h}}^{obs}\right)^2 + \left(\sigma_{\mathbf{h}}^{calc}\right)^2} + \sum_{i=1}^{restraints} \frac{\left(r_i^{obs} - r_i^{calc}\right)^2}{\left(\sigma_{r(i)}^{obs}\right)^2}$$

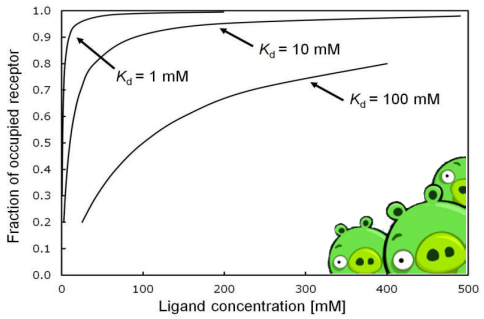
Total residual | X-ray weight | ML X-ray residual | Geometry residual

Biochemical subtlety: Binding sites suck (up stuff)

Figure 3-38 Soaking of blue dye into lysozyme crystals. The crystals were dyed by adding 0.5 µl of methylene blue solution to a 2 µl drop, and imaged immediately after dye addition, after 1 h, and after 10 h. The solution becomes successively lighter while the crystals absorb nearly all of the dye. It is also important to realize that it takes quite a while even for small molecules to diffuse into a crystal. It is not possible to soak large ligands such as peptides into crystals in seconds and expect to observe any electron density.[27]

Note: diffusion is a slow process ! (movie)





Fraction of occupied receptor sites plotted against ligand equilibrium concentration for three different binding constants. While at mM and lower K_d range small concentrations of ligand suffice to achieve reasonable binding site occupancy (between 70-90%), quite impractical concentrations of ligand in the crystallization drop are required for poor binders. On the other hand, given sufficiently high concentration, even weakly binding and non-native ligands can be forced into a binding site.

What to do about (B) Human factors

First the bad news.

1) Most of us are (at least in a genetic sense) human

2) Our cognitive defects are Expectation Bias and Confirmation Bias

3) I did not make that up [5-7]

Now the good news.

The two pillars of scientific epistemology – the relations between newly gained evidence and prior knowledge have been incorporated in an inductive framework of formal logic by Rev. T. Bayes in 1763 [8] and in its current modern form by Laplace [9]

[5] Koehler JJ (1993) The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality. *Org. Behavior Human Decision Proc.* **56**(1): 28-55.

[6] Frey BS (2003) Publishing as Prostitution? Choosing Between One's Own Ideas and Academic Failure. *Public Choice* ***116****, 205-223 (ETHZ)*

[7] Simmons JP, Nelson LD and Simonsohn U (2011) False-Positive Psychology: **Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant**. *Psychological Science* **22**, 1359-1366.

[8] Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53**: 370-418.

[9] Laplace, P S (1814). Essai philosophique sur les probabilités. Paris Bachelier, Paris.

Bayesian inference and our models

$$prob(A \mid B,n) \propto prob(B \mid A,n) \times prob(A \mid n)$$

Reformulate above tool (Bayes' Theorem, derived from the product rule for independent conditional probabilities) in terms of Model (M) and Data (D):

$$prob(model \mid data) \propto prob(data \mid model) \times prob(model)$$

Model Likelihood: The final posterior probability of our structure model given the experimental data: this is what we ultimately need to know

Data Likelihood: (sampling probability): how well are our experimental data reproduced by a given model – that is, the strength of experimental evidence for the given model

Prior Probability of that model given all prior knowledge of chemistry, physics, biology, but without consideration of the data

Consequence of Bayes (common sense)

Model Likelihood ≈ Quality of Evidence x Prior probability

It is best if both are large – good fit to data and no violation of stereochemistry or other laws of physics.

Poor fit to data and violation of stereochemistry or other laws of physics is really bad (but correlates…)

There has to be a balance between the terms – strong claim with little prior basis needs strong evidence !

This means we need to obtain clear evidence (+DD electron density) and examine how consistent our protein ligand complex models is with prior expectations

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

The hunger (for density) games

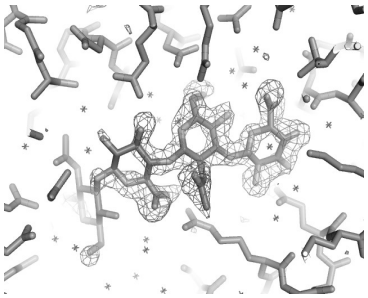


Figure 1: (**Author:**) Clear electron density unambiguously confirms the presence of the terminal poly-saccharide units. Figure made with PyMol.

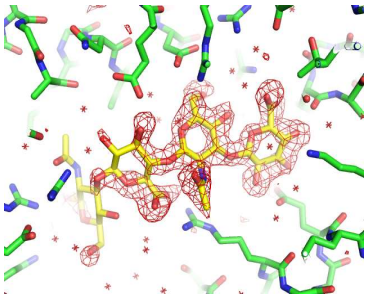


Figure 1: Clear mF_o-DF_c negative difference electron density contoured at -3 sigma unambiguously confirms the absence of the terminal poly-saccharide units. Figure made with PyMol.

Any (reviewer) comments?

Be clear about what you are looking at

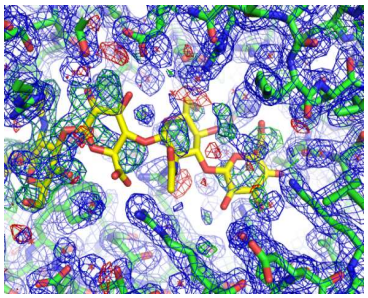
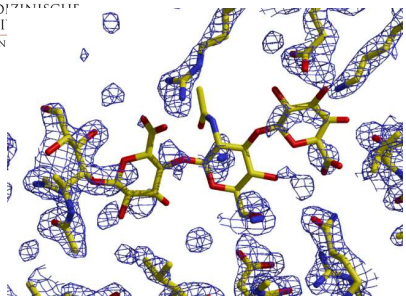


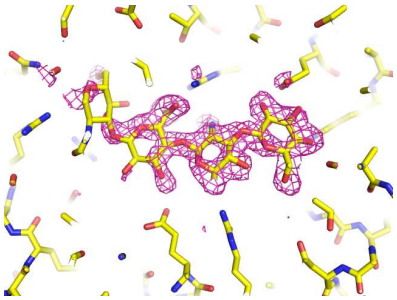
Figure 1: (**Reality:**) Neither ligand omit electron density maps ($2mF_o$-DF_c, blue,1σ) nor difference density maps (mF_o-DF_c, green, +3σ, red -3σ) calculated by *REFMAC* from deposited coordinates (less ligand) and structure factors show any signs of positive density for the terminal poly-saccharide units.

For your own safety, state:

- Type of map (omit, difference?)
- Map Coefficients (ML)
- Contour levels
- Somewhere, program and source of data used for map calculation
- Beware of the b&w trap
- Do not use the blob&noise trick!
- Look at the real space measures

Abuse of fixed B-factors for cosmetics





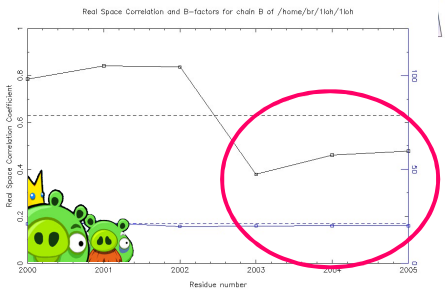
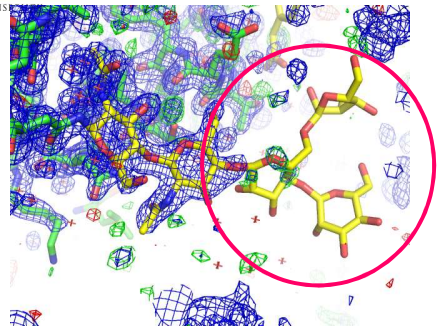
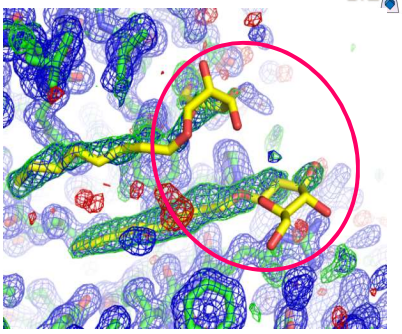


Figure 1. Real space correlation, *B*-factors, and electron density for hexa-saccharide in 1loh. Top right: The real space correlation (black) for saccharide units 4, 5 and 6 is distinctly lower than that for the units 1-3. The reported *B*-factors remain inconsistently low. Top left Panel: the Shake&wARP electron density reconstruction contoured at 1 sigma, showing no density for saccharides 4-6. There is some density (clipped) for unit 4, which is not correctly placed. The electron density Fig.2 in [1] could not be reproduced. Bottom left panel: the difference density map contoured at -3 sigma, showing negative density (indicating absence of the model) for saccharides 4-6. Calculated by REFMAC using original model deposited in the PDB without modification.

Partially visible ligands – the most common problem

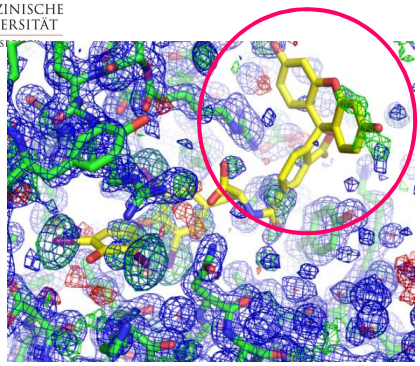


Missing density: extended glycosylations, omit density maps (only the last three sugar monomers were excluded from the omit map calculation). The specific conformation of the extended branched glycosylation (A5-A7) in PDB entry 3ib0 (Mir *et al.*, 2009) is unsupported by electron density in the structure of the bovine lactotransferrin.



Missing density: Detergents, ligand omit maps. Two detergent molecules placed into the models of membrane proteins. The plant SLAC1 anion channel structure, PDB entry 3m73, (Chen *et al.*, 2010) shows two molecules (BOG A317/A318) that have clear density for the hydrophobic acyl chain but not for the head groups.

Partially disordered ligands – a common problem

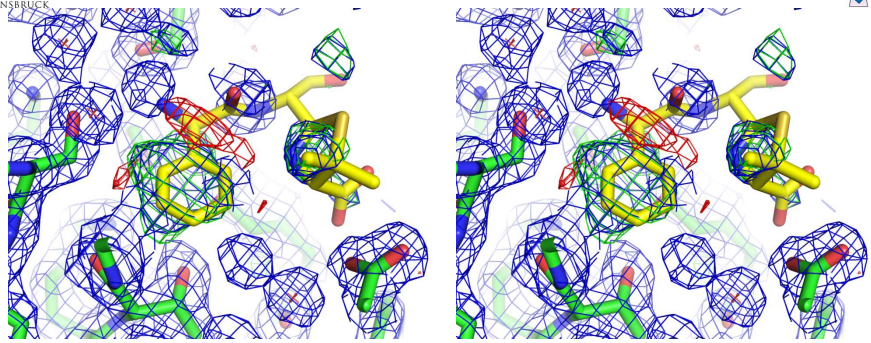


Partially disordered ligand, ligand omit density. The fluorescein moiety of the ligand molecule (F6Z A1356) is missing in the electron density of the thyroxine-binding globulin, PDB entry 2xn7, (Qi *et al.*, 2011), even at 0.4 σ noise level $2mF_o$-DF_c density.

For your own safety

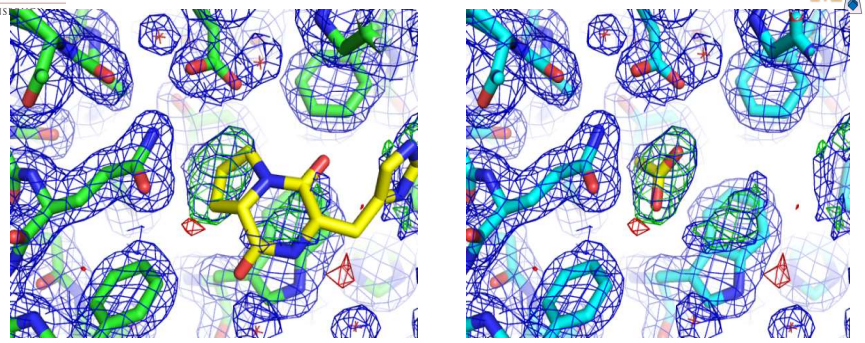
- Use sound judgment
- Am I misleading myself?
- Am I misleading the reader?
- Can I really say what is there?
- Would you take a drug based on that structure?
- Would you bet your money ($3B for a drug) on that specific ligand pose?

Ligands that are cocktail components



Ligands placed into mother liquor density, ligand omit maps. In the structure of the penicillin binding protein 4 from *S. aureus*, PDB entry 3hun, (Navratna *et al.*, 2010) the phenyl moiety of the ampicillin (ZZ7 B501) is placed in the region of the electron density that based on difference density analysis could be better interpreted as a sulphate ion . The re-refined model that includes sulphate ion is shown in the left panel.

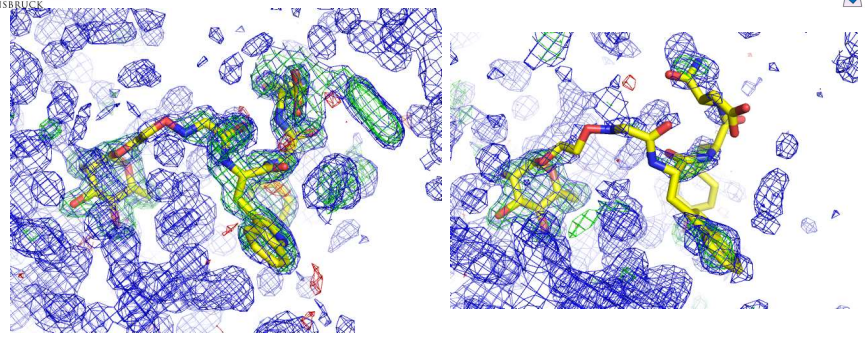
Ligands that are cocktail components



Ligands placed into mother liquor density, ligand omit maps. A: In the structure of the *B. cereus* chitinase, PDB entry 3n1a, (Hsieh *et al.*, 2010), the cyclo-(L-His-L_Pro) molecule (CHQ A1514) is placed into low level electron density that is difficult to interpret, and which may be plausibly interpreted as an acetate molecule present in crystallization cocktail at 200 mM.

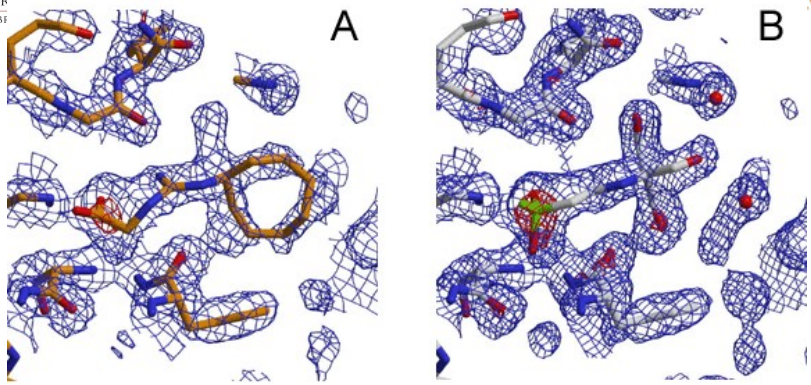
Very tempting and very common – check your imagination!

NCS means chemically not equivalent!



A peptide molecule that appears to be projected from the binding site of a non-crystallographic symmetry (NCS) related copy. Shown are electron density maps from the 1.9 Å structure of concanavalin A in complex with a glycomimetic peptide, PDB entry 4czs. Two peptide molecules (chain E, F, rank 3105, 2528) are well traceable in electron density (chain E shown in panel A), while rank chains G and H (rank 254 and 99, respectively) are only partly visible in electron density (chain H shown in panel B). (Ng et al., 2015). The original publication shows only a surface rendering with a ball and stick model of the glycopeptide (unspecified chain ID, without electron density).

Binding sites want to bind – anything they can

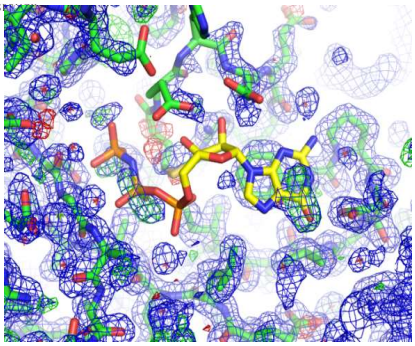


TES buffer in ligand binding site. 2.1 Å maps contoured at 1σ (blue) and 5σ (red). (A) presumed **supersweetener** built into *CNS* ML $2mF_o$-DF_c **taste receptor** map; (B) *Shake&wARP* map, with TES buffer built into density. Map has less noise and cleaner connectivity and reveals the true nature of the ligand. A questionable VdW contact is also obvious between 'ligand' and protein in the left panel (A).

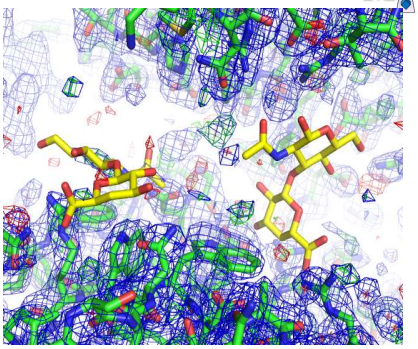
Lack of supervision and training may often be responsible!

PS: TES buffer does NOT taste sweet :-/

…and ligands that just are not there



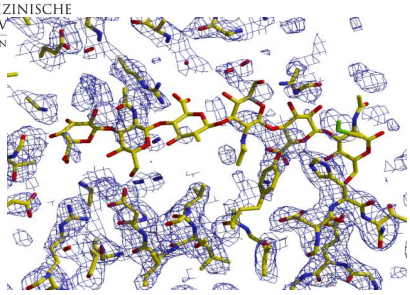
Absent ligand density in the omit map. In the structure of the Nudix hydrolase, PDB entry 1sz3, (Ranatunga *et al.*, 2004), the non-hydrolyzable GDP analogue (GNP 3030A) is placed in a conformation and position entirely unsubstantiated by $2mF_o$-DF_c electron density.

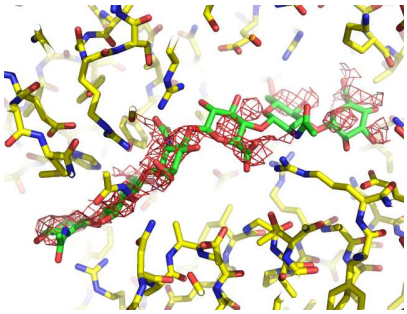


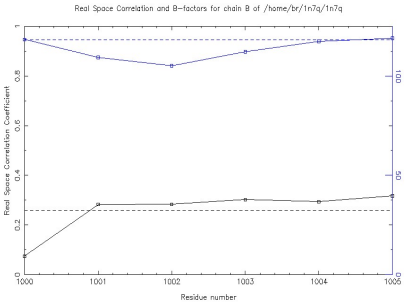
Missing ligands. Two di-saccharide molecules in the structure of the hyaluronate lyase from *S. agalactiae*, PDB entry 1i8q, (Li & Jedrzejas, 2001) are not supported by the omit electron density maps.

Did you deposit the right files? Check your records!

Almost absence of difference density in noise case



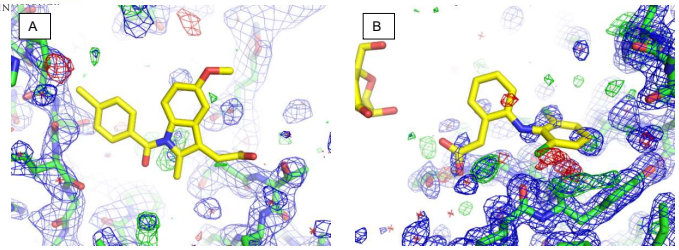




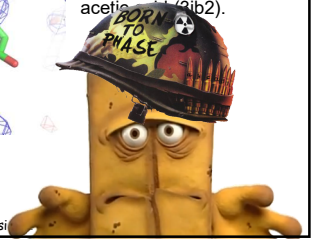
EDS identical assessment

Figure 2. Real space correlation, *B*-factors, and electron density for hexa-saccharide in 1n7q. Top Panel: The real space correlation (black) for all saccharide units 1- 6 of the hexa-saccharide is abysmally low and the *B*-factors correspondingly high. Top left Panel: the Shake&wARP electron density reconstruction contoured at 1 sigma, showing only noise and solvent density for saccharides 1-6. Bottom left panel: the difference density map contoured at -2 sigma, showing negative density (red) that coincides with the ligand.

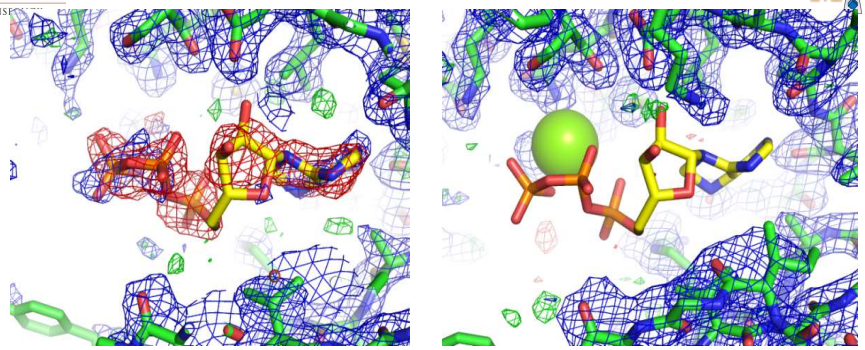
Some are serial (drug) offenders



Absent ligands. Four protein-ligand complex structures presented in (Mir *et al.*, 2009) include ligands that are not supported by electron density. All panels show the omit maps for complex structures with the following ligands: A. indomethacin (PDB entry 3ib1); diclofenac (3ib0); C. aspirin (3iaz); D. α-methyl-4-(2-methylpropyl) benzene acetic [illegible] (3ib2).



Worst: distorted, no density, but important!



Negative difference density for a ligand. The electron density in the structure of the mutant of the human kinase ERK2 ,PDB entry 1gol, (Robinson *et al.*, 1996) contradicts the modeled position and provides no evidence for a severely distorted conformation of the ATP molecule. The difference density map from *EDS* (A) shows the negative density that coincides with the ligand position. The omit difference map (B) shows no difference density above 3σ level that would suggest ATP presence. The green sphere represents a magnesium ion in the original model.

...and more peptides that just are not there

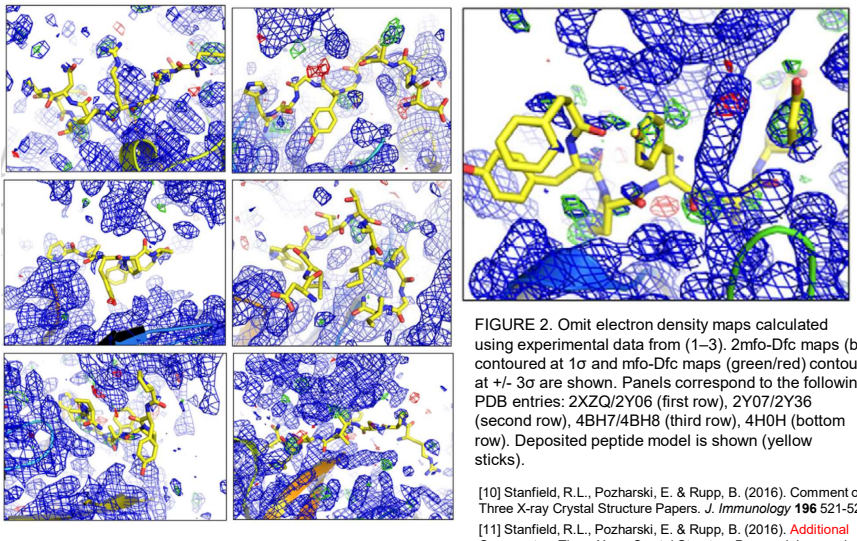


FIGURE 2. Omit electron density maps calculated using experimental data from (1–3). 2mfo-Dfc maps (blue) contoured at 1σ and mfo-Dfc maps (green/red) contoured at +/- 3σ are shown. Panels correspond to the following PDB entries: 2XZQ/2Y06 (first row), 2Y07/2Y36 (second row), 4BH7/4BH8 (third row), 4H0H (bottom row). Deposited peptide model is shown (yellow sticks).

[10] Stanfield, R.L., Pozharski, E. & Rupp, B. (2016). Comment on Three X-ray Crystal Structure Papers. *J. Immunology* **196** 521-524.

[11] Stanfield, R.L., Pozharski, E. & Rupp, B. (2016). Additional Comment on Three X-ray Crystal Structure Papers.*J. Immunology* **196**, 528-530.

...with the internal energy of a molecular nuclear device...

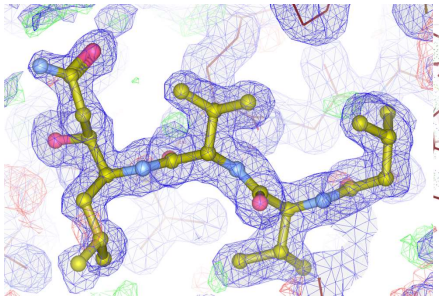
Table I. Summary of PDB validation reports for structure models and cited reference structures

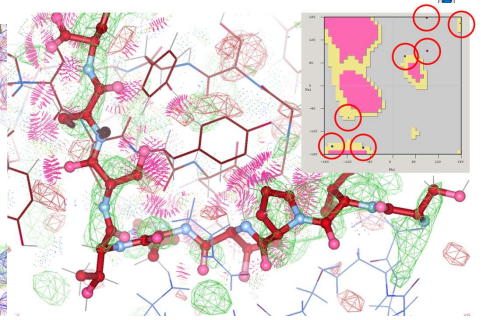
PDB ID	Chain	Ramachandran favoured (%)	Ramachandran allowed (%)	Ramachandran outliers (%)	Ramachandran percentile	Rotamer percentile	Real space R value Z-score RSRZ	RSRZ > 2 outliers %	Author
1pww	C	44%	11%	44%	0(0)	0(2)	3.1	72%	Liddington
	D	44%	11%	44%	0(0)	0(0)	3.3	72%	Liddington
1uvi	D	RNA	-	-	-	-	2.9	75%	Grimes
	E	RNA	-	-	-	-	3.9	100%	Grimes
	F	RNA	-	-	-	-	3.2	75%	Grimes
2xzq	P	29%	0%	71%	0(0)	0(0)	5.3	100%	Salunke
2y06	P	0%	38%	62%	0(0)	5(8)	4.3	100%	Salunke
2y07	P	25%	0%	75%	0(0)	0(0)	3.3	100%	Salunke
2y36	P	11%	22%	67%	0(0)	0(0)	5.9	100%	Salunke
4bh7	P	29%	29%	43%	0(0)	100(100)	3.2	55%	Salunke
4bh8	P	14%	14%	71%	0(0)	1(1)	6.3	100%	Salunke
4h0h	D	20%	20%	40%	0(0)	0(0)	9.7	100%	Salunke
3fn0	P	86%	14%	0%	100(100)	100(100)	-0.1	11%	Wilson
3ggw	E	100%	0%	0%	100(100)	100(100)	0.3	9%	Bentley
	F	100%	0%	0%	100(100)	16(3)	0.6	8%	Bentley

Falsehood flies, and truth comes limping after it, so that when men come to be undeceived, it is too late; the jest is over, and the tale hath had its effect.

— Jonathan Swift (1667-1745)

Good ligand….bad ligand





A peptide inhibitor with optimal fit to electron density. The figure shows the peptide inhibitor pepstatin A (ball and stick model) bound to a retroviral protease in clear electron density (blue grid), contoured at a 2mFo-DFc electron density level of 1 sigma (PDB entry 3sm1, chain J, by Alexander Wlodawer and coworkers [32]). There is (i) no doubt about the presence and pose of the inhibitor, (ii) the peptide inhibitor assumes a stereochemically plausible conformation, and (iii) no steric clashes or implausible interactions with the retroviral protease are indicated. **Positive difference omit density practically overlaps with the displayed density**. The image was prepared using the graphical model building program Coot [33] displaying the 2mFo-DFc electron density map reconstructed from maximum likelihood coefficients computed by the refinement program Refmac [26].

An improbable peptide ligand. The main figure shows the model of a peptide antigen (ball and stick model) bound to an Fab antibody fragment (thin sticks), together with its positive mFo-DFc omit difference density (meaning that the ligand was omitted during maximum likelihood difference density map coefficient calculation). The peptide model should be surrounded by clear difference density (green grid, contoured at 2.5 sigma above mean noise density) resembling its distinct shape. There are **only discontinuous fragments** visible which in part can be explained by **ordered solvent** (the round green spheres are typical for water molecules). In addition to (i) not being placed in any recognizable electron density, the antigen model (ii) has a multitude of unreasonably close contacts (**steric clashes, visualized as the nasty red spikes**) with the antibody fragment and (iii) has an utterly implausible high energy backbone conformation as evidenced by all the backbone torsion angle pairs being located in unfavourable regions in the Ramachandran plot (top right insert).

Personal defense against the abstruse ligand is based on a simple epistemological idea:

View your structure model as a proposition or hypothesis that should withstand scrutiny against a body of evidence **AND** prior knowledge

In other words: you determine structures to TEST a structural hypothesis but NOT to PROVE it

Such prevents you from the tendency to find what one seeks…(peer pressure, nagging stressed supervisors, grant deadlines, promising the cure for cancer…)

A final old notice to the young scientist:

Do not believe in anything simply because you have heard it. Do not believe in anything simply because it is spoken and rumored by many. Do not believe in anything simply because it is found written in your books. Do not believe in anything merely on the authority of your teachers and elders. Do not believe in traditions because they have been handed down for many generations. But after observation and analysis, when you find that anything agrees with reason and is conducive to the good and benefit of one and all, then accept it and live up to it.

This is a quote from….?

A final old notice to the young scientist:

Do not believe in anything simply because you have heard it. Do not believe in anything simply because it is spoken and rumored by many. Do not believe in anything

Gautama Buddah, ca. 500 BC.

generations. But after observation and analysis, when you find that anything agrees with reason and is conducive to the good and benefit of one and all, then accept it and live up to it.

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

A final old notice to the young scientist:

Do not believe in anything simply because you have heard it. Do not believe in anything simply because it is spoken and rumored by many. Do not believe in anything

And this, my friends, is the difference between knowledge and wisdom.

generations. But after observation and analysis, when you find that anything agrees with reason and is conducive to the good and benefit of one and all, then accept it and live up to it.

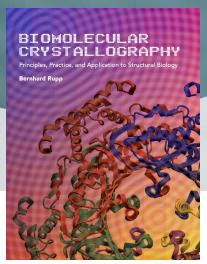
Model wisely and prosper.

FWF Der Wissenschaftsfonds.

Edwin Pozharski
Chris Weichenberger
Marc Deller
Robyn Stanfield

Structure validation, analysis, and presentation

The scientist must be the judge of his own hypotheses, not the statistician.
A. F. W. Edwards (1992) in *Likelihood – An account of the statistical concept of likelihood and its application to scientific inference*, p 34



In discrete form: Structure factor summation to electron density summation

General Fourier transformation (FT) integral – from real space domain (x, Å) into reciprocal domain (1/x or x*, $Å^{-1}$). Another example is time, in sec, to frequency, sec^{-1}.

$$\mathbf{F_h} = \sum_{j}^{atoms} \rho(\mathbf{x})_j \cdot e^{2\pi i(\mathbf{h}\cdot\mathbf{x_j})}$$

In the discrete case, the complex Fourier integral becomes a complex Fourier summation

$$\rho(\mathbf{x}) = \sum_{-\mathbf{h}}^{+\mathbf{h}} \mathbf{F_h} \cdot e^{-2\pi i(\mathbf{h}\cdot\mathbf{x})}$$