The road to scientific serfdom*

Bernhard Rupp
Department of Forensic Crystallography
k.-k. Hofkristallamt
Vista, CA 92084, USA

Marie Curie Incoming International Fellow
Dept. Genetic Epidemiology
Innsbruck Medical University,
A-6020 Innsbruck, Austria

Sociology, epistemology and systemic human factors in the emergence of over-interpretation and errors in crystallographic structure models: Validation of your mind.

*with apologies to F. A. Hayek (1899–1992)

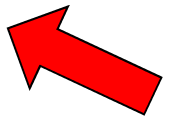
You are successful, young, and objective scientists, right?

Where did you actually learn about the epistemology of empirical science – the rules of acquiring knowledge in the process of inductive inquiry ?

Let's make this as simple as possible:

There are essentially 2 'things':
(a) things we know and (b) things we do not know*

*When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it: that is knowledge.
Confucius, Analects (ca. 500 BC)



What does the scientist do?

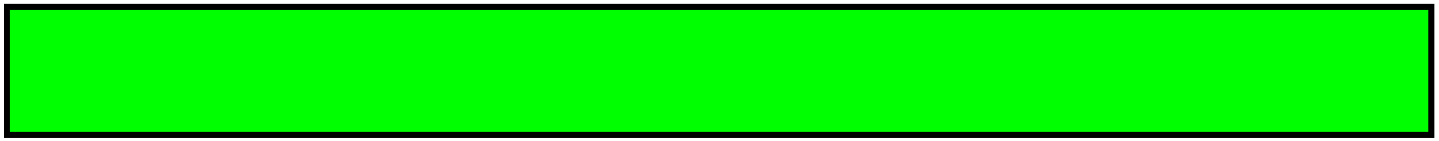
(a) things that are known: you should learn them – they become your prior knowledge

(b) things you do not know: you conduct an experiment and gather evidence (perhaps to support a model, a hypothesis, or just so)

Let us now examine a few scenarios of prior knowledge vs. new evidence:

The early mode of empirical Science….

Experiment



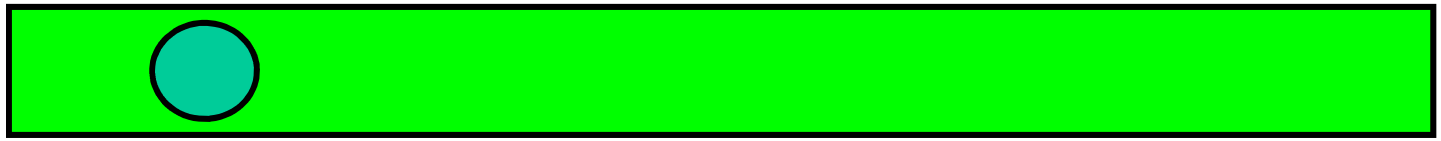
Adam & Eve

Prior knowledge

Unified Theory Of Everything

The early mode of empirical Science….

Experiment



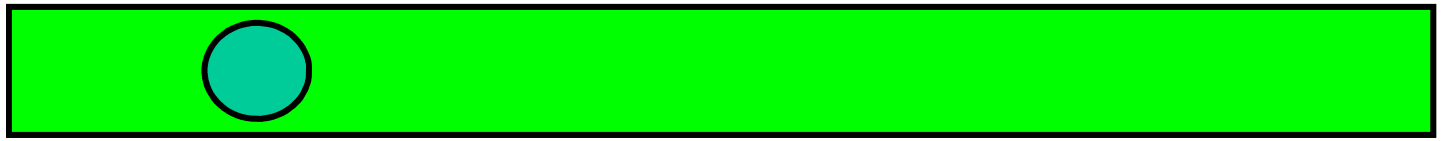
Adam & Eve

Prior knowledge

Unified Theory Of Everything

The early mode of empirical Science….

Experiment



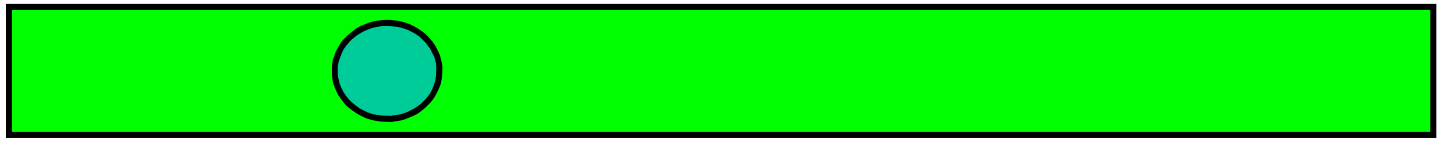
Adam & Eve

Prior knowledge

Unified Theory Of Everything

The early mode of empirical Science….

Experiment



Adam & Eve

Prior knowledge

Unified Theory Of Everything

This early stage (Baconian) 17th century method works to a certain degree…but:

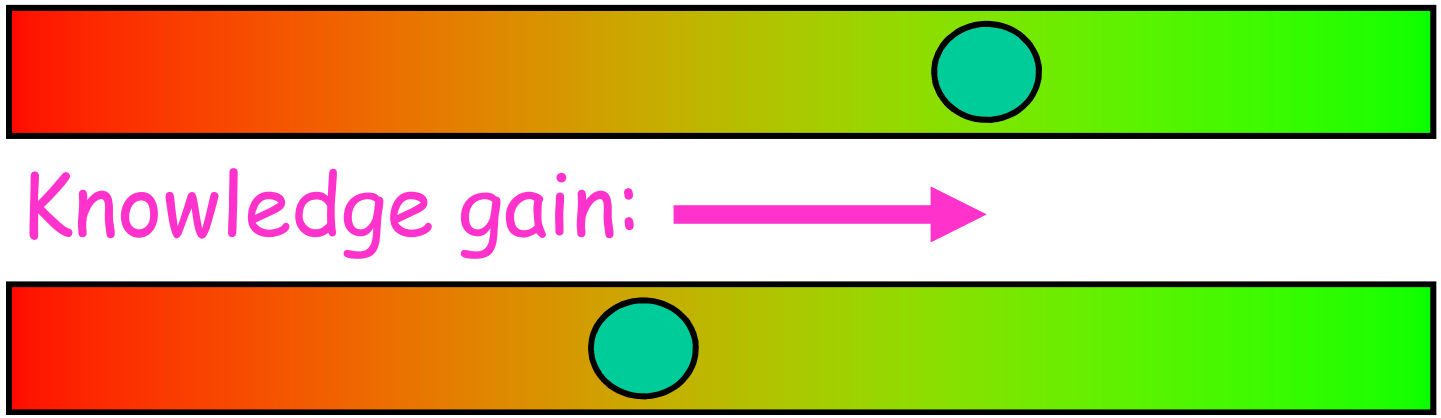
(a) It is essentially discovery based: almost any experiment adds to (disorganized) prior knowledge – i.e. let's sort it out later

(b) The 'System of the World' as a mere collection of all known observations becomes contradictory and unmanageable (example 'HEAT')

(c) But it is inherently safe – we have not much prior expectations when gathering evidence

Transition to model (hypothesis) based method (18th ct, R.Soc.):

New evidence



Prior knowledge – in form of laws and epistemological paradigms

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

The modern model or hypothesis based approach works great:

(a) BUT: A new necessity arises to deal with negative results (with dignity)!

Extreme case: The emergence of scientific revolutions*

Deny New evidence Confirm

Prior knowledge - paradigms

Such strong contradictions are generally (after denial) found by many and alter the underlying paradigms, i.e. lead to a scientific revolution

*The Structure of Scientific Revolutions, Thomas S. Kuhn (1962)

Problems tend to arise in less dramatic cases:

(a) BUT: A new necessity arises to deal with negative results (with dignity)!

(b) Purpose is potentially unsafe – we may have too much prior expectation when faced with negative or poor evidence

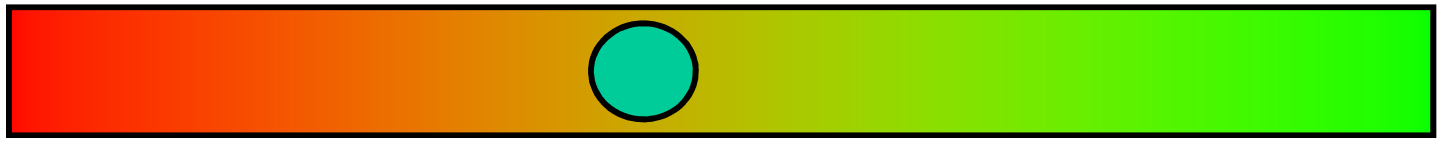
Koehler JJ (1993) The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality. *Organizational Behavior and Human Decision Processes* **56**(1): 28-55.

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

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*: DOI: 10.1177/0956797611417632.

Temptations of negatives (I):

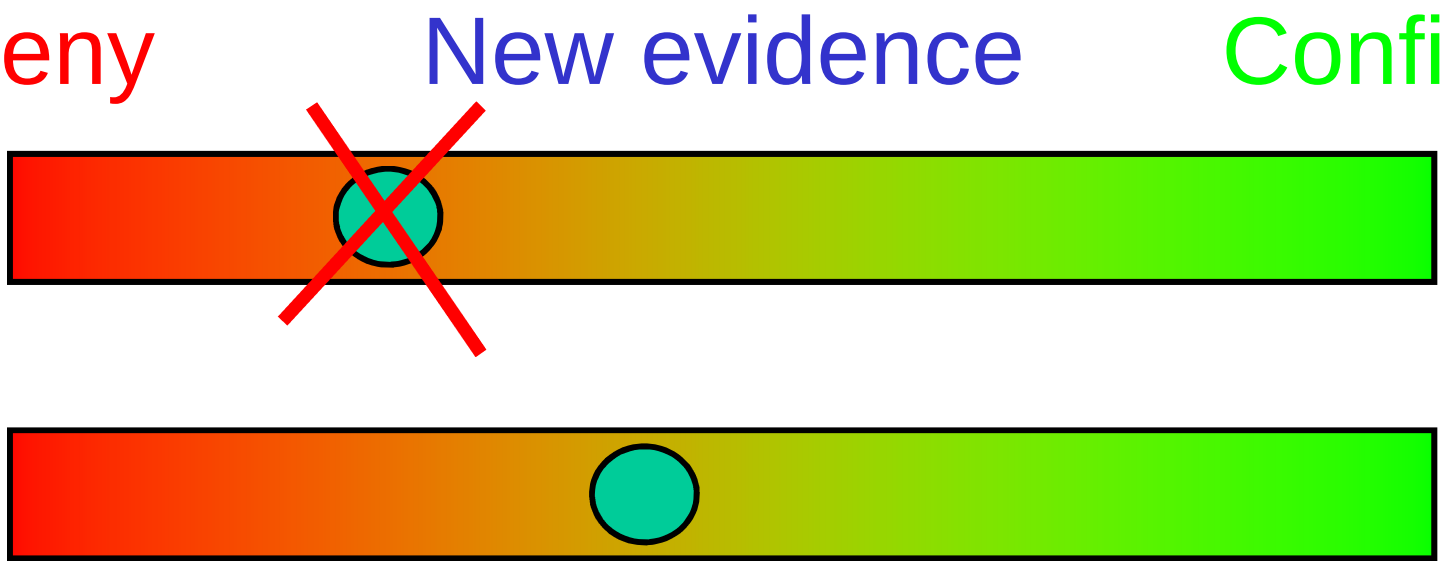
Deny New evidence Confirm



Your beloved idea/model/hypothesis

This is simple Fabrication. Almost all of us can resist this and therefore fabrication is rarely a problem. It is also very difficult to do it right…

Temptations of negatives (II):

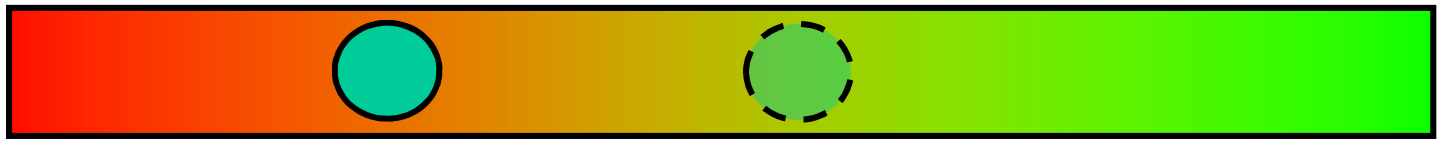


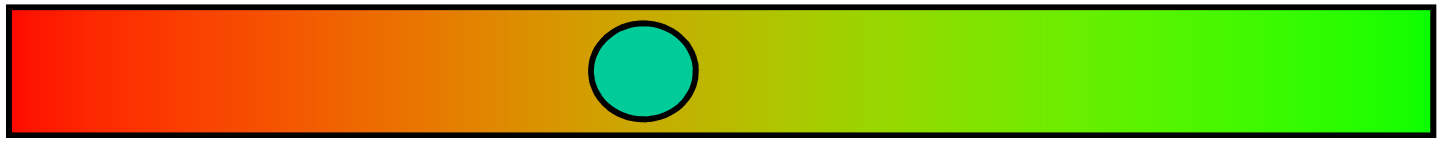
Your beloved idea/model/hypothesis

This is omission of negatives. The phenomenon is known as confirmation bias in the psychological literature. It is not very difficult to do…

Temptations of negatives (III):

Deny New evidence Confirm





Your beloved idea/model/hypothesis

This is overinterpretation, often supported by ad hoc assumptions. The phenomenon is known as expectation bias in the psychological literature. It is also not very difficult to do …

'Why' it happens is actually very old business (17th century):

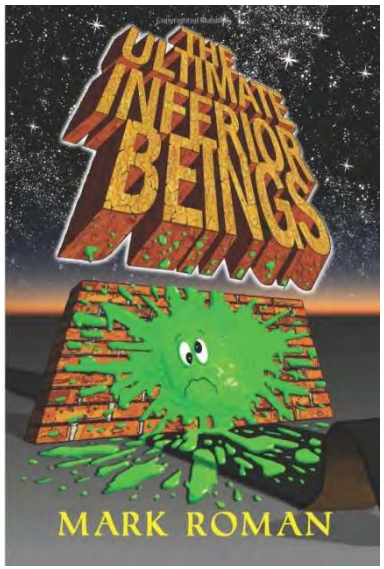
The human understanding is not composed of dry light, but is subject to influence from the will and the emotions, a fact that creates fanciful knowledge; man prefers to believe what he wants to be true……for what man had rather were true he more readily believes.

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

Men fall in love with particular pieces of knowledge and thoughts either because they believe themselves to be authors and inventors, or because they have put a great deal of labor into them, and they have got very used to them. Francis Bacon, ibid.

Why is it hard to be objective ?

- We are under permanent pressure
- We have financial or career interests
- We become susceptible to expectation bias (e.g. overinterpret spurious evidence)
- We become susceptible to confirmation bias (e.g. ignore negative results/evidence)



- In other words, (most of us) are human beings…

Modern (18th century!) scientific epistemology has also provided us with a most valuable survival guide:

(c) The incorporation of inductive inference into a framework of formal logic by Rev. Bayes provides a clear relation between prior knowledge and new evidence – and defense against 'human fancy'

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

Rev. T. Bayes, Phil Trans RS. 49 (1763)

How can we incorporate our inductive inference into a framework of formal logic ?

$$prob(A,B\,|\,n) = prob(A\,|\,n) \times prob(B\,|\,A,n) = prob(B,A\,|\,n)$$

Product rule for independent conditional probabilities

$$prob(A,B\,|\,n) = prob(A\,|\,B,n) \times prob(B\,|\,n) = prob(B\,|\,A,n) \times prob(A\,|\,n)$$

$$prob(A\,|\,B,n) = \frac{prob(B\,|\,A,n) \times prob(A\,|\,n)}{prob(B\,|\,n)}$$

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

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

Aha...and what exactly does that do for us and our validation?

Reformulate that tool (Bayes' Theorem) in terms of Model (M) and Data (D):

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

Final posterior probability of the model given the data – the Model Likelihood

The Data Likelihood (sampling prob.): how well are experimental data reproduced by a given model – the strength of experimental evidence for the given model

The Prior Probability of that model based on ALL prior knowledge without considering the data (geometry, chemistry, physics, biological evidence)

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 common…)

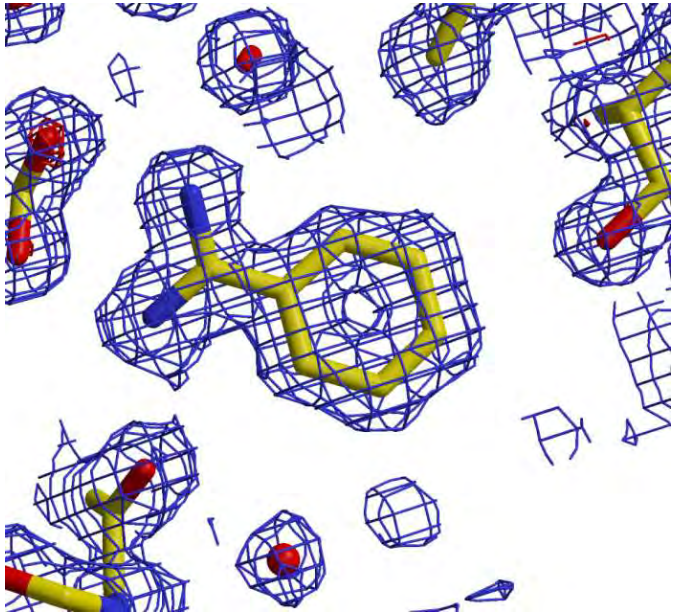
There has to be a balance between the terms – strong claim with little prior basis needs strong evidence !

Now, let's look at the evidence term for a class of very interesting structure models – protein ligand complexes

Why ligand-models may be dangerous to your career

1. Global indicators of (reciprocal space) data fit like R-values are completely insensitive. 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. Ditto, protein geometry means nil.

2. Therefore we need local (real space) indicators that show the fit between model and electron density. The electron density – preferably minimally biased positive omit difference density - is the primary evidence !

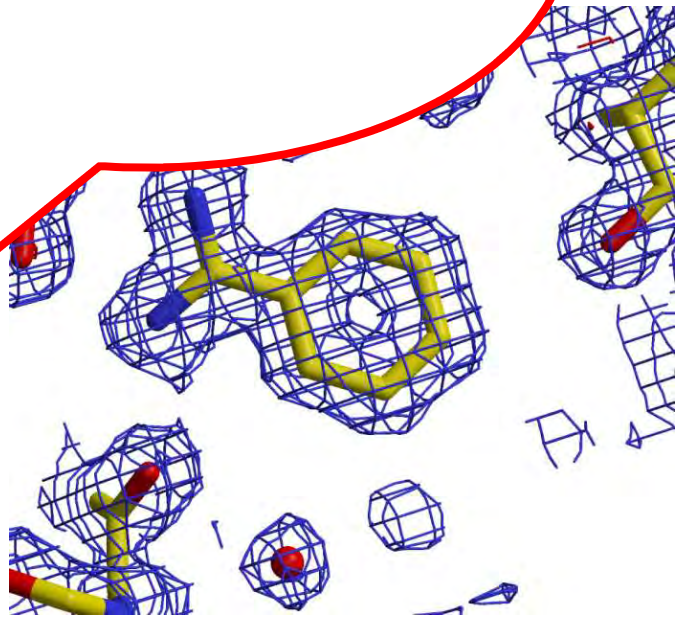


Why ligand-models may be dangerous to your career

1. Global indicators of (reciprocal space) data fit like R-values are completely insensitive. Ligand scattering mass is often only 1/1000 of th… …e this with high B-factors and po… becomes even worse. Dit… …il.

2. Therefore we need loca… space) indicators that show the fit between model and electron density. The electron density – preferably minimally biased positive omit difference density - is the primary evidence !

(Poster) presenters please take note!



The hunger (for density) games

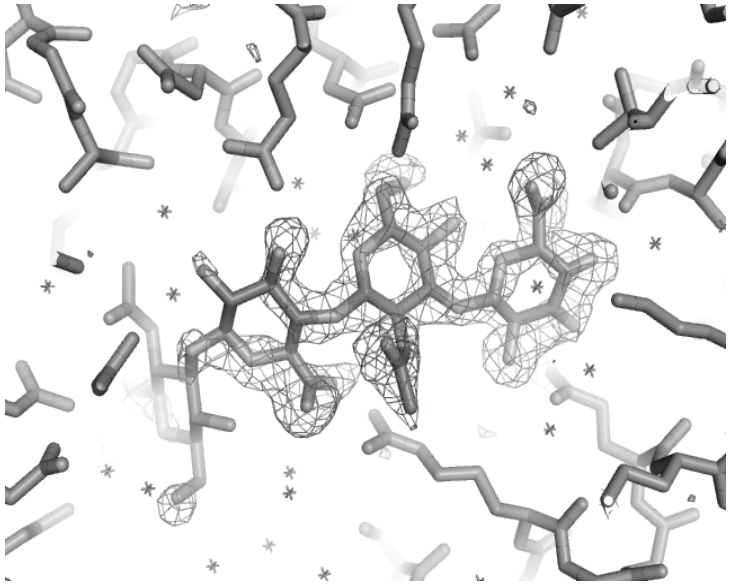


Figure 1: Clear electron density unambiguously confirms the presence of the terminal poly-saccharide units. Figure made with PyMol.

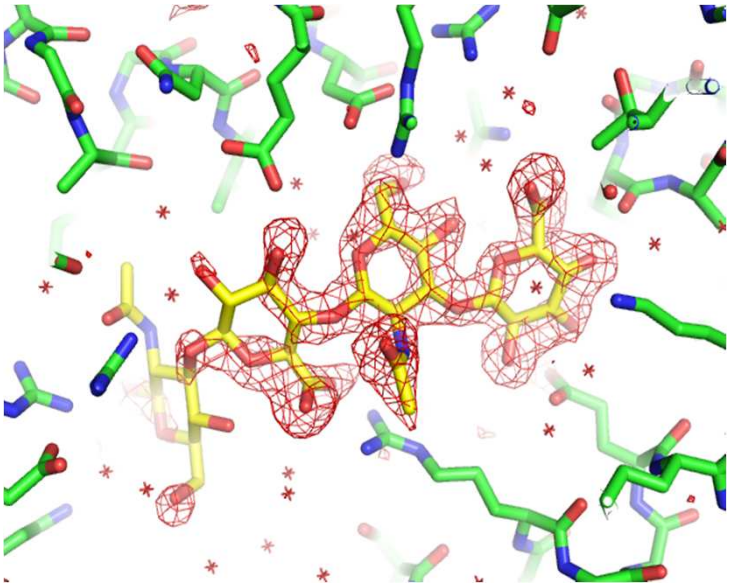


Figure 1: Clear mF_o-DF_c negative difference electron density contoured at -3 sigma unambiguously confirms the absence of the terminal poly-saccharide units. Figure made with PyMol.

Any (review) comments?

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 difference density (bad)
F_c (model) **without** ligand contribution: **omit** ligand (or low occupancy and/or, high B factor)	Positive difference density (good)	No or poor difference density (meaningless noise subtraction)

Be clear about what you are looking at

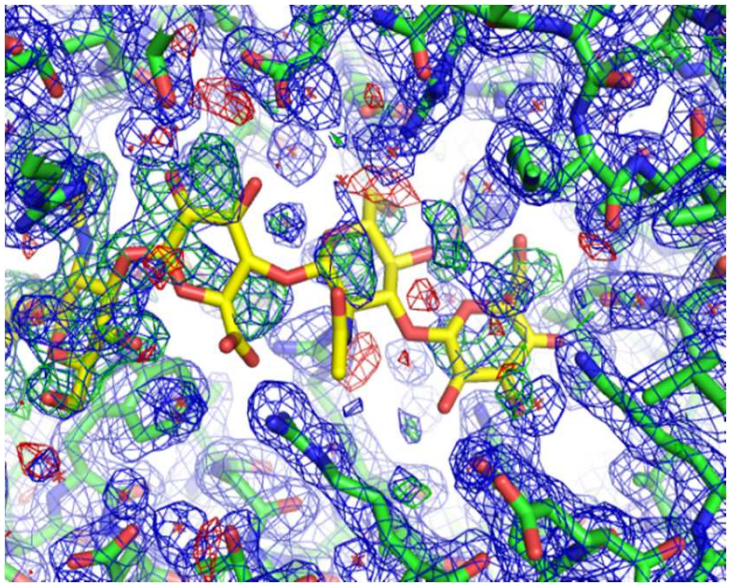
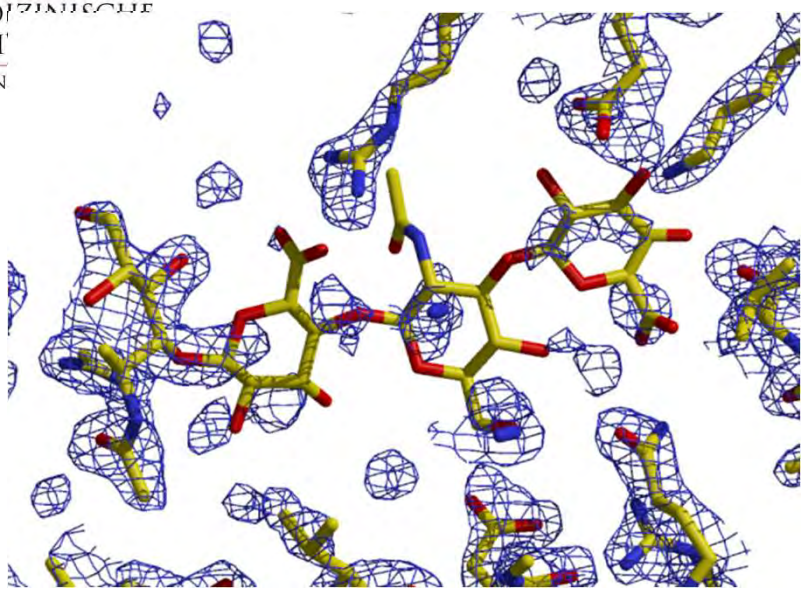


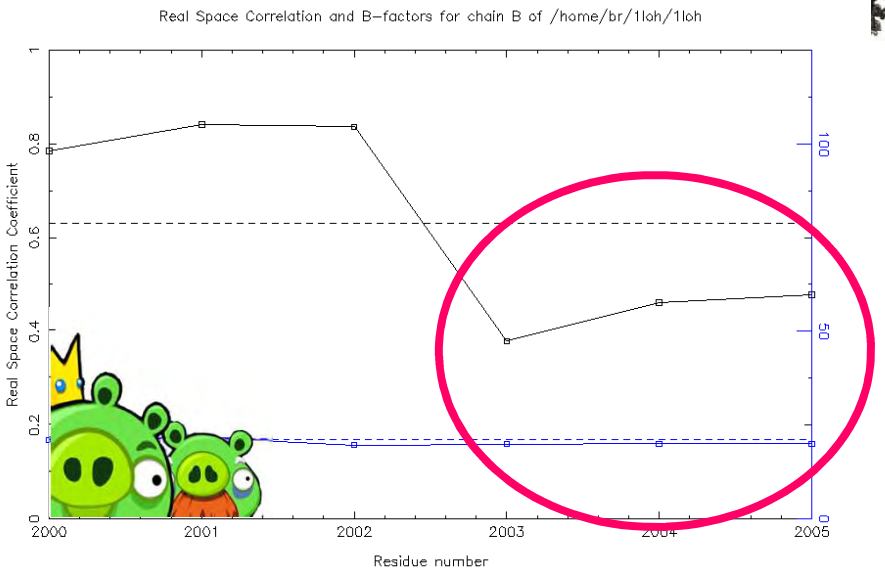
Figure 1: Neither ligand omit electron density maps ($2mF_o$-DF_c, blue,1σ) nor difference density maps (mF_o-DF_c, green, +3σ, red -3σ) calculated by *REFMAC* from deposited coordinates (less ligand) and structure factors show any signs of positive density for the terminal poly-saccharide units.

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
- Look at the RSCC
- Do not use the blob&noise tick!

Abuse of fixed B-factors for cosmetics





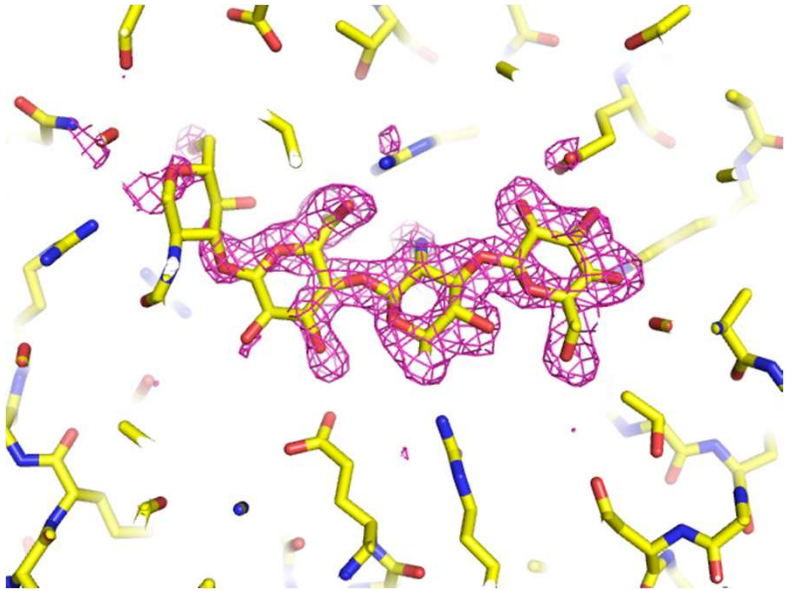
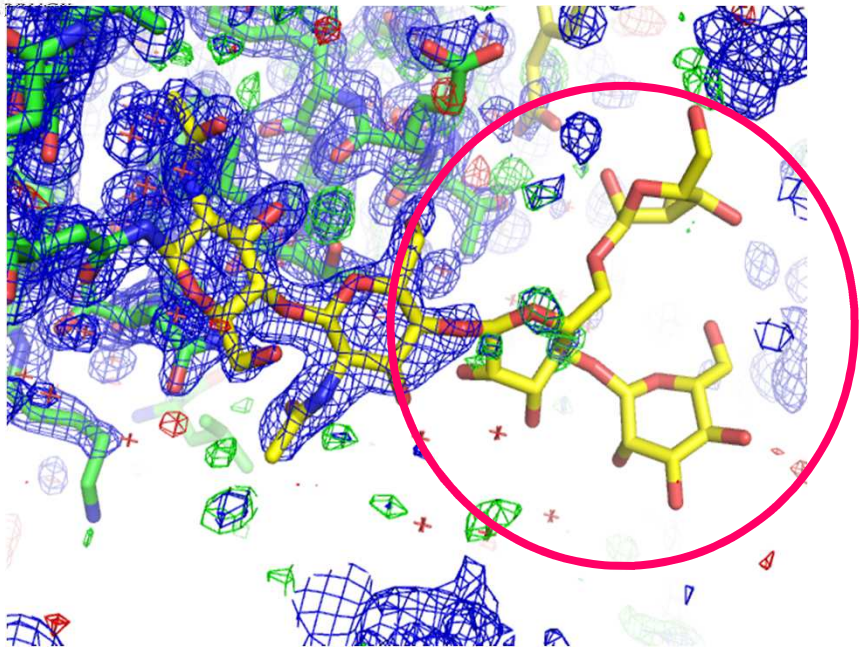
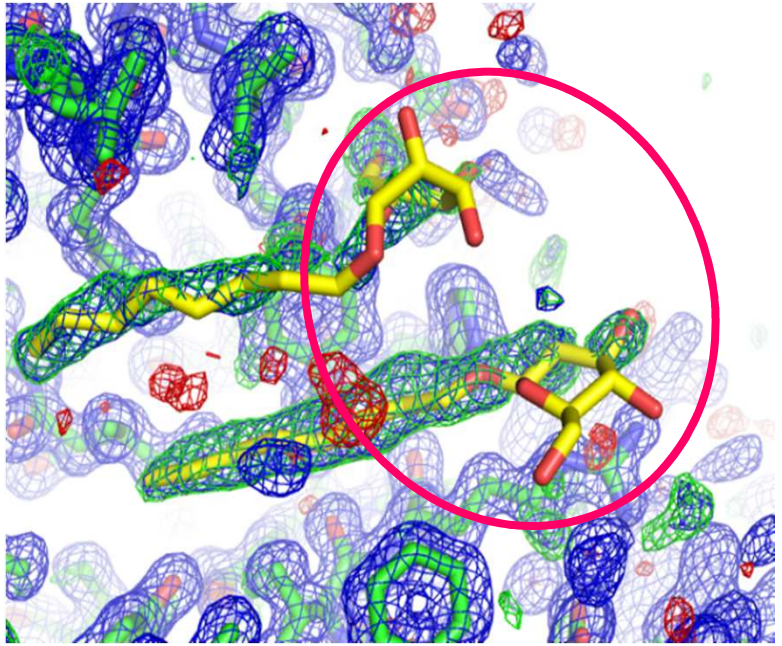


Figure 1. Real space correlation, *B*-factors, and electron density for hexa-saccharide in 1loh. Top right: The real space correlation (black) for saccharide units 4, 5 and 6 is distinctly lower than that for the units 1-3. The reported *B*-factors remain inconsistently low. Top left Panel: the Shake&wARP 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 – a common problem

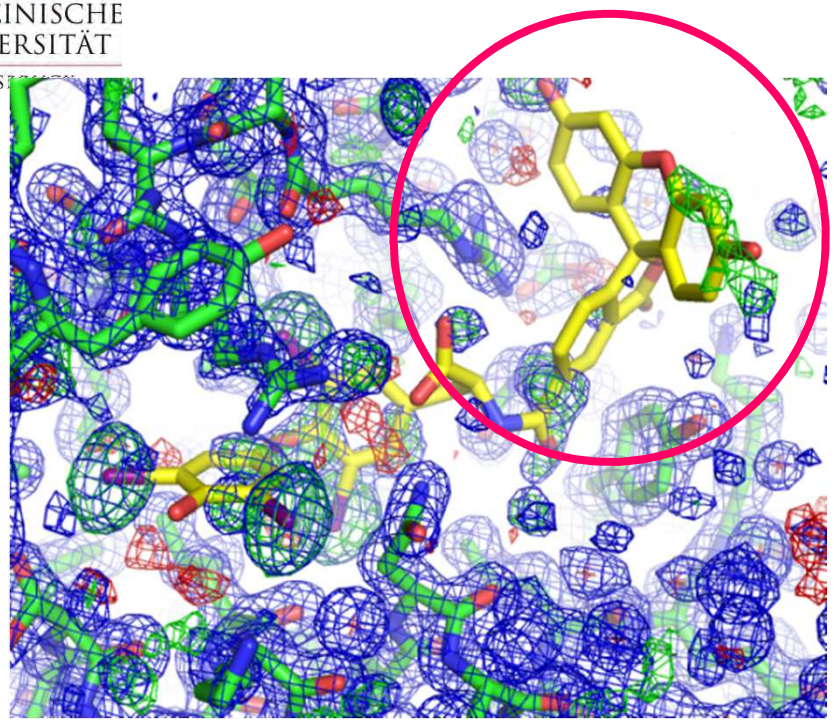


Missing density: extended glycosylations, omit density maps (only the last three sugar monomers were excluded from the omit map calculation). The specific conformation of the extended branched glycosylation (A5-A7) in PDB entry 3ib0 (Mir *et al.*, 2009) is unsupported by electron density in the structure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



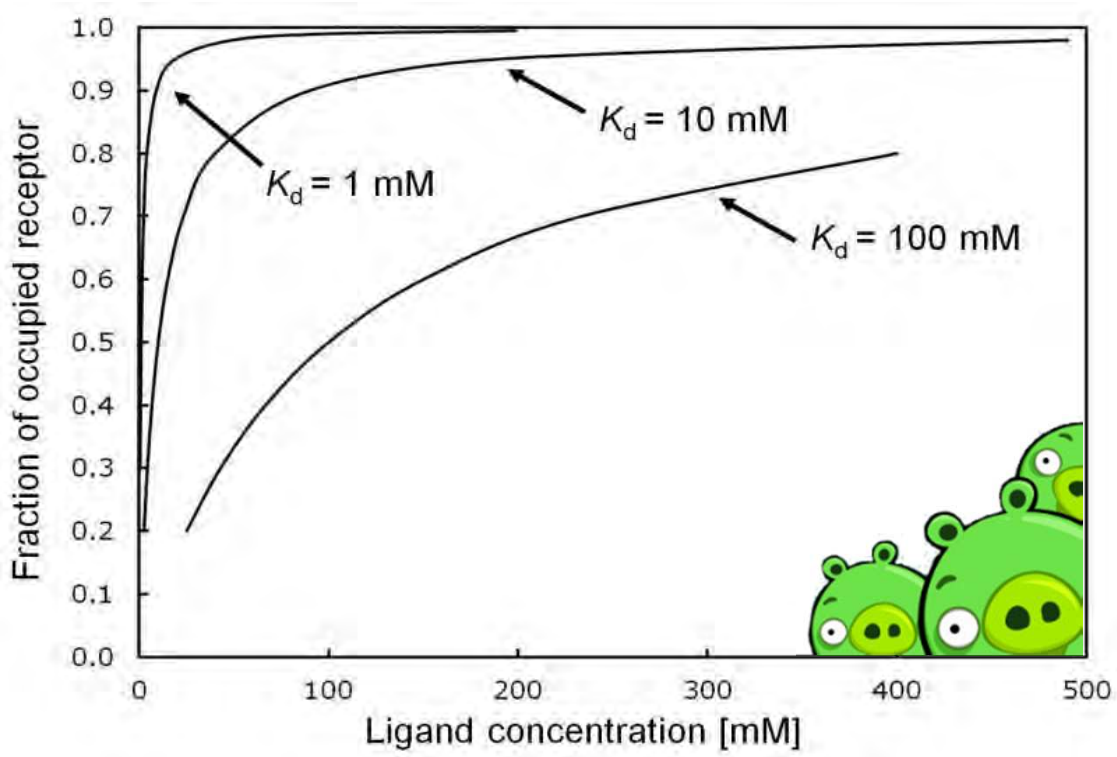
Partially disordered ligand, ligand omit density. The fluorescein moiety of the ligand molecule (F6Z A1356) is missing in the electron density of the thyroxine-binding globulin, PDB entry 2xn7, (Qi *et al.*, 2011), even at 0.4 σ noise level $2mF_o$-DF_c density.

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?

Bindings sites suck (up stuff)

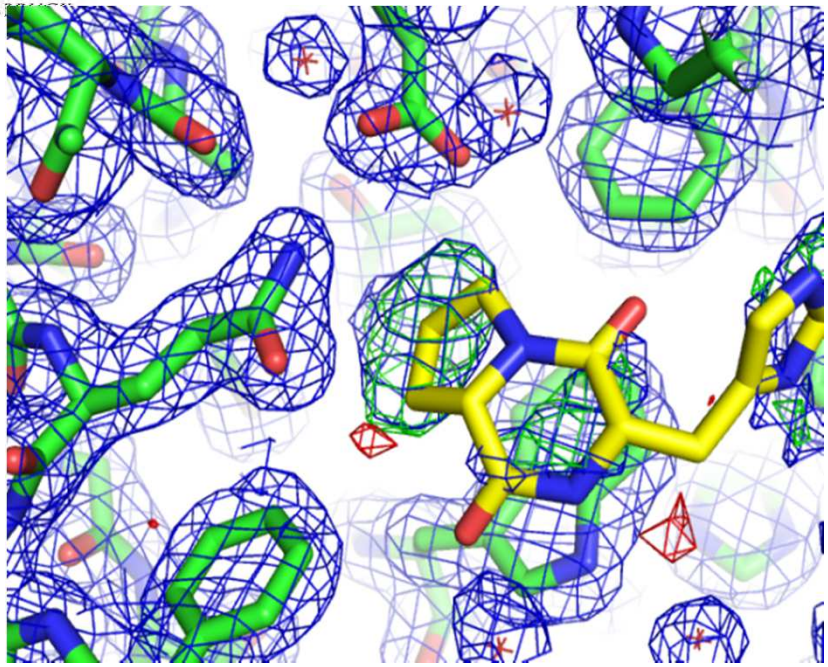
Out of principle, binding sites are never fully occupied:

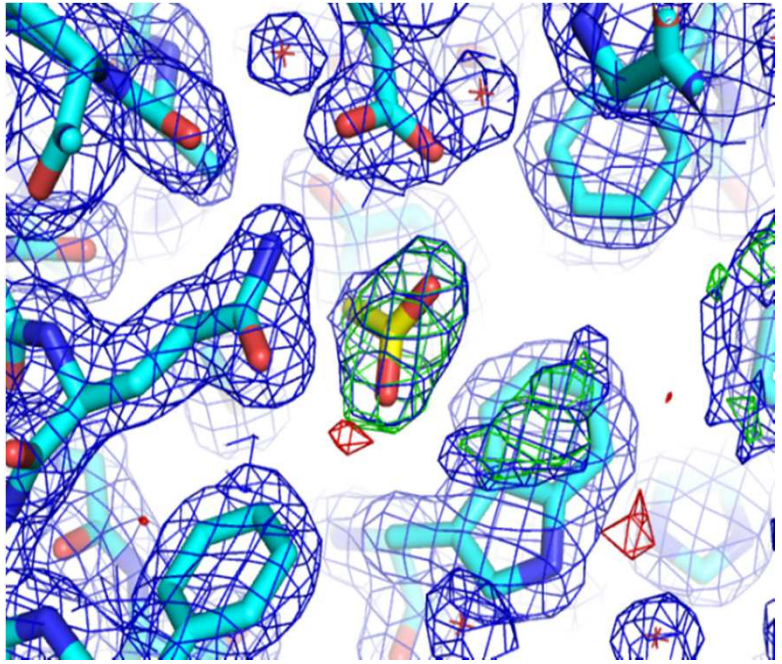


Fraction of occupied receptor sites plotted against ligand equilibrium concentration for three different binding constants. While at mM and lower K_d range small concentrations of ligand suffice to achieve reasonable binding site occupancy (between 70-90%), quite impractical concentrations of ligand 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.

-> There is almost always some obscure density in sites

Ligands that are cocktail components

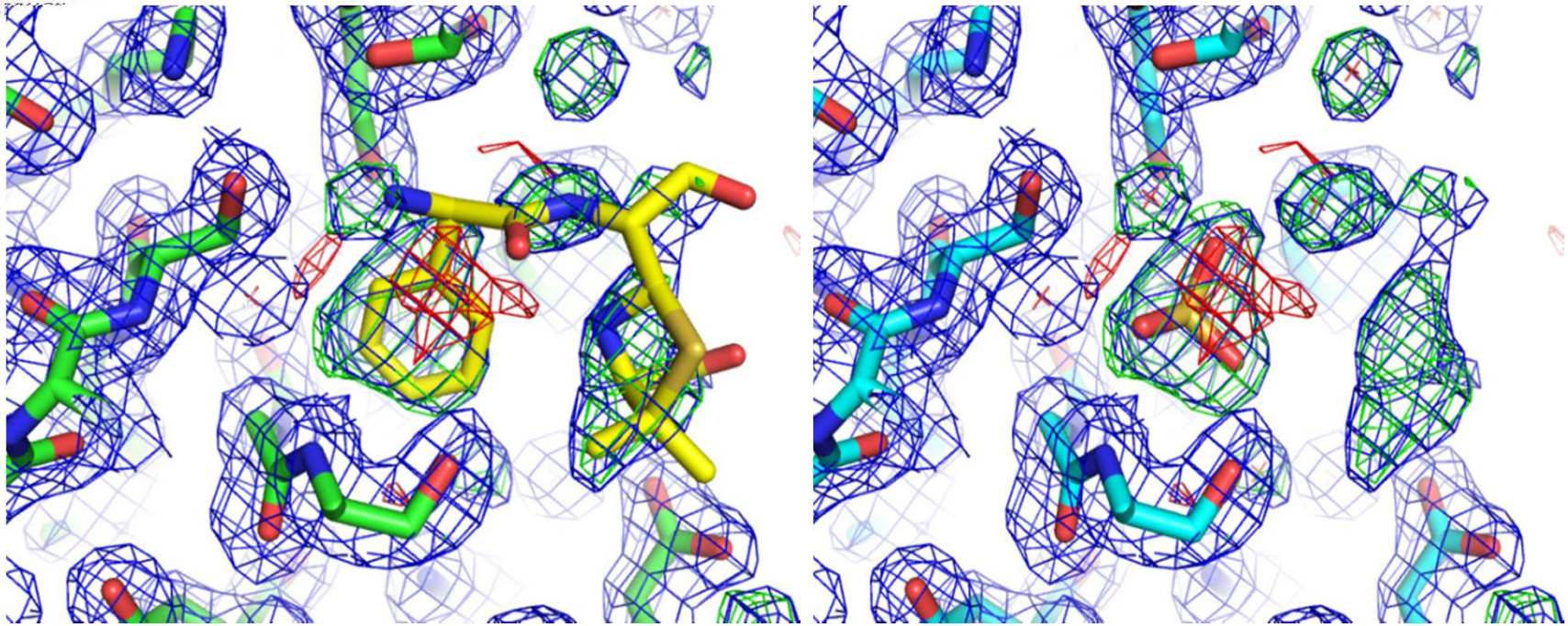




Ligands placed into mother liquor density, ligand omit maps. A: In the structure of the *B. cereus* chitinase, PDB entry 3n1a, (Hsieh *et al.*, 2010), the cyclo-(L-His-L_Pro) molecule (CHQ A1514) is placed into low level electron density that is difficult to interpret, and which may be plausibly interpreted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

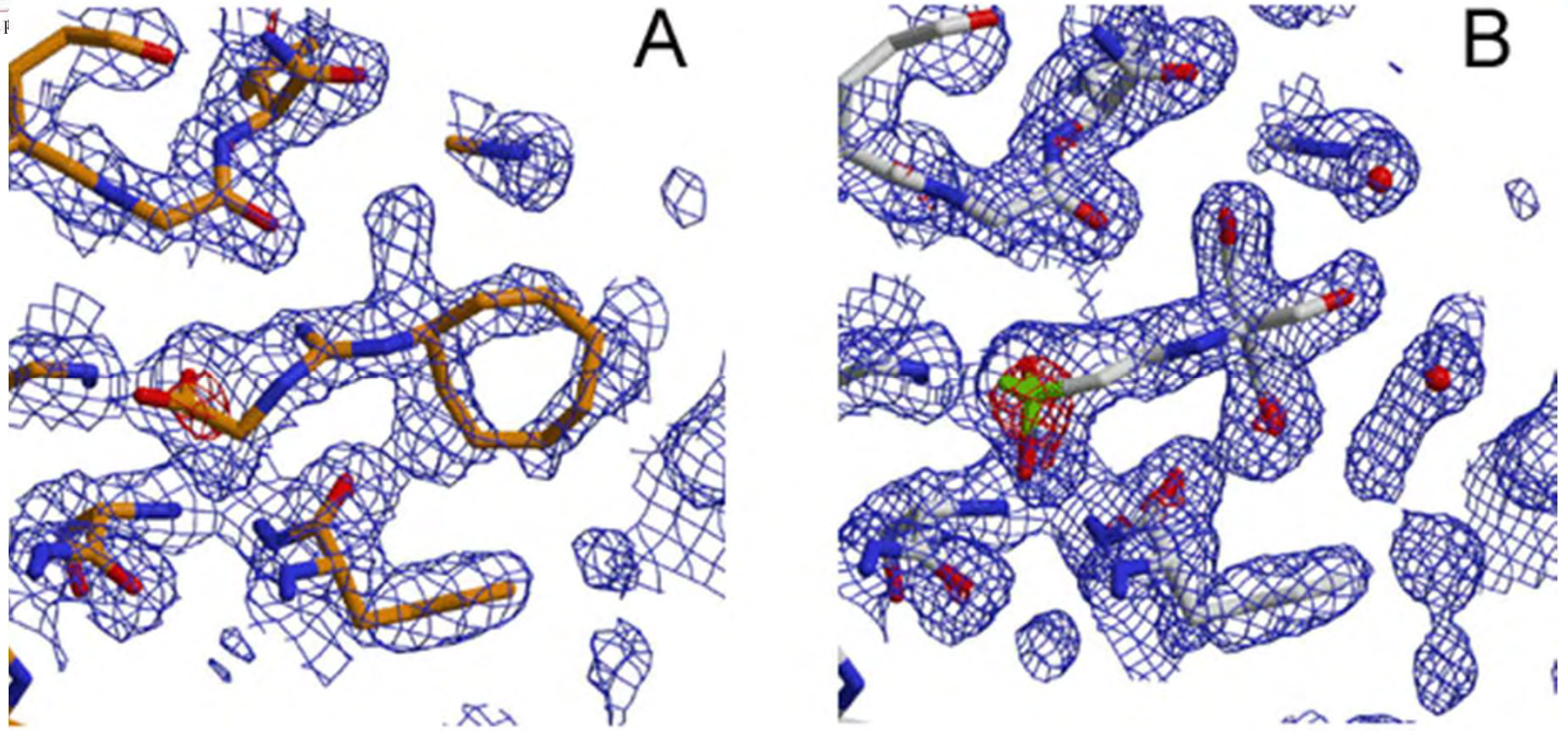
Very tempting and very common – check your imagination!

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.

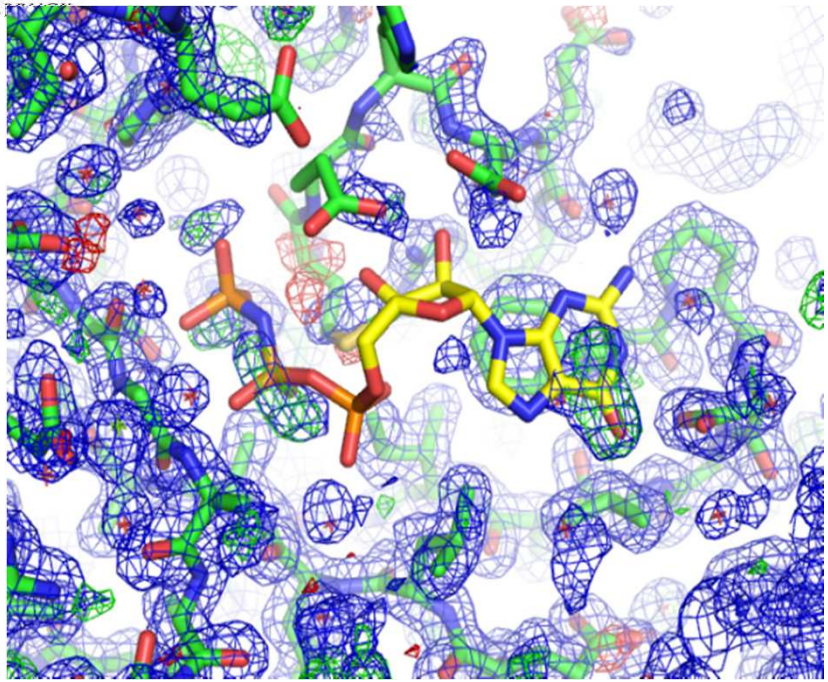
Binding sites want to bind – anything they can



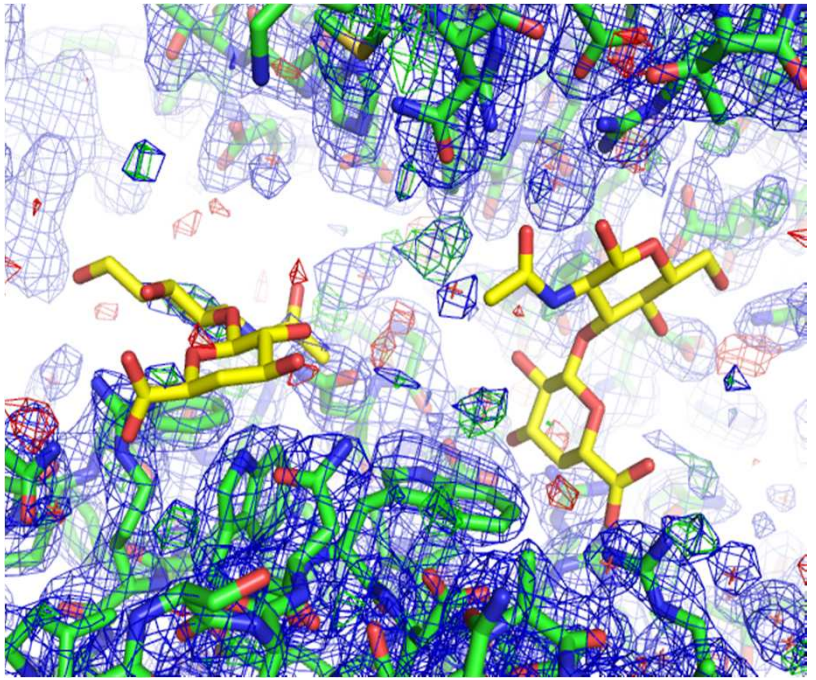
TES buffer in ligand binding site. 2.1 Å maps contoured at 1σ (blue) and 5σ (red). (A) presumed ligand built into *CNS* ML $2mF_o$-DF_c map; (B) *Shake&wARP* map, with TES buffer built into density. Map has less noise and cleaner connectivity and reveals the true nature of the ligand. A questionable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

Lack of supervision and training may often be responsible!

…and ligands that just are not there



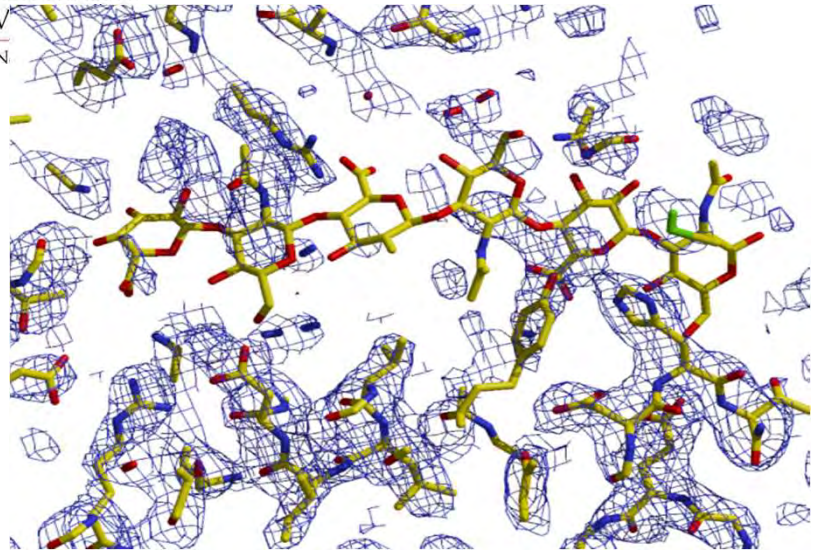
Absent ligand density in the omit map. In the structure of the Nudix hydrolase, PDB entry 1sz3, (Ranatunga *et al.*, 2004), the non-hydrolyzable GDP analogue (GNP 3030A) is placed in a conformation and position entirely unsubstantiated by $2mF_o$-DF_c electron density.

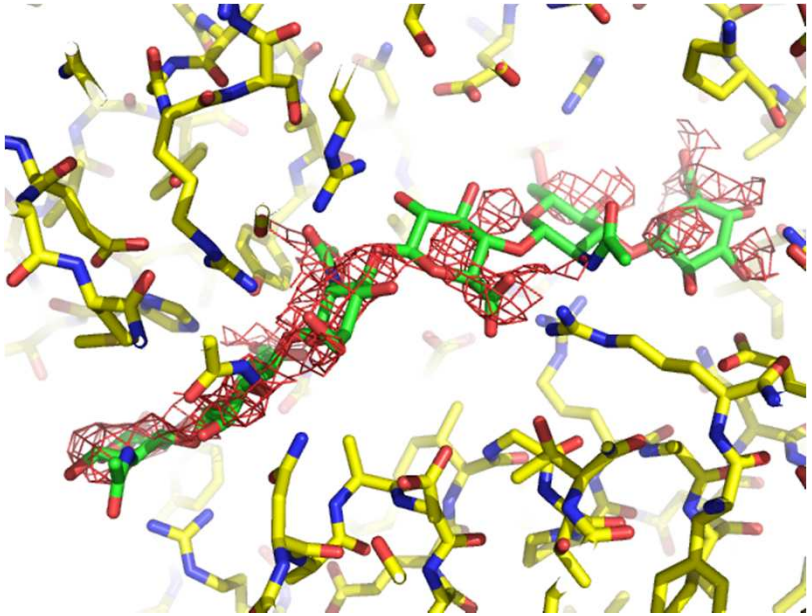


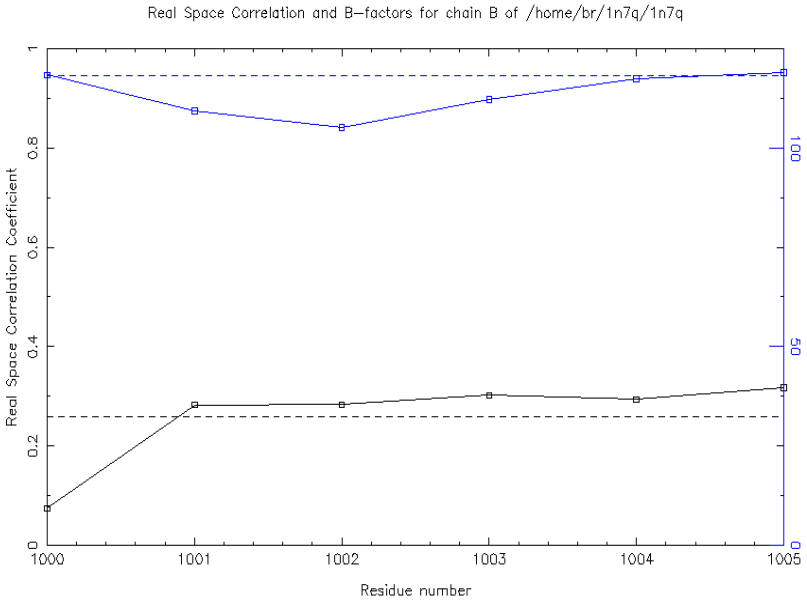
Missing ligands. Two di-saccharide molecules in the structure of the hyaluronate lyase from *S. agalactiae*, PDB entry 1i8q, (Li & Jedrzejas, 2001) are not supported by the omit electron density maps.

Did you deposit the right files? Check your records!

Almost absence of difference density in noise case



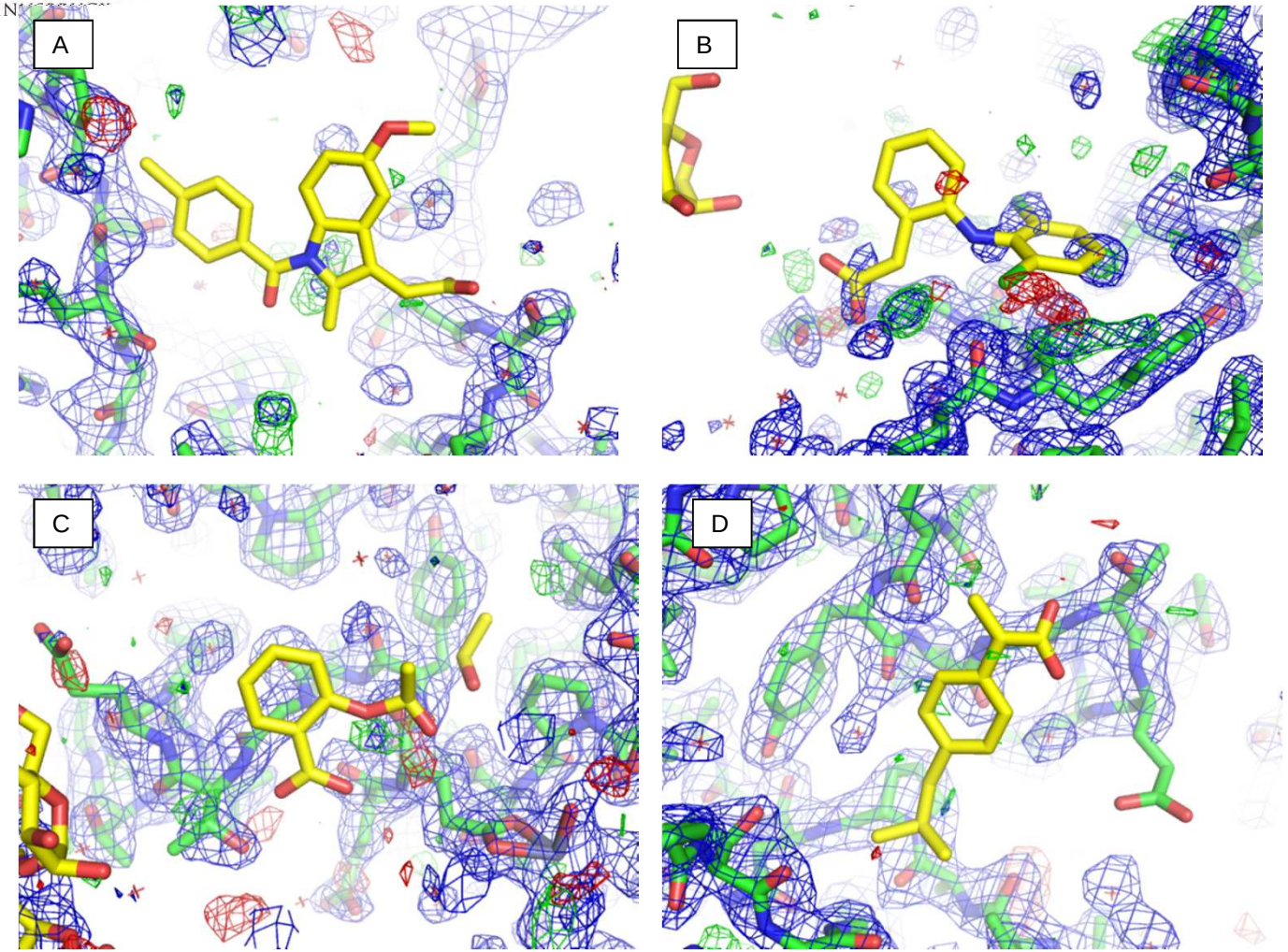




EDS identical assessment

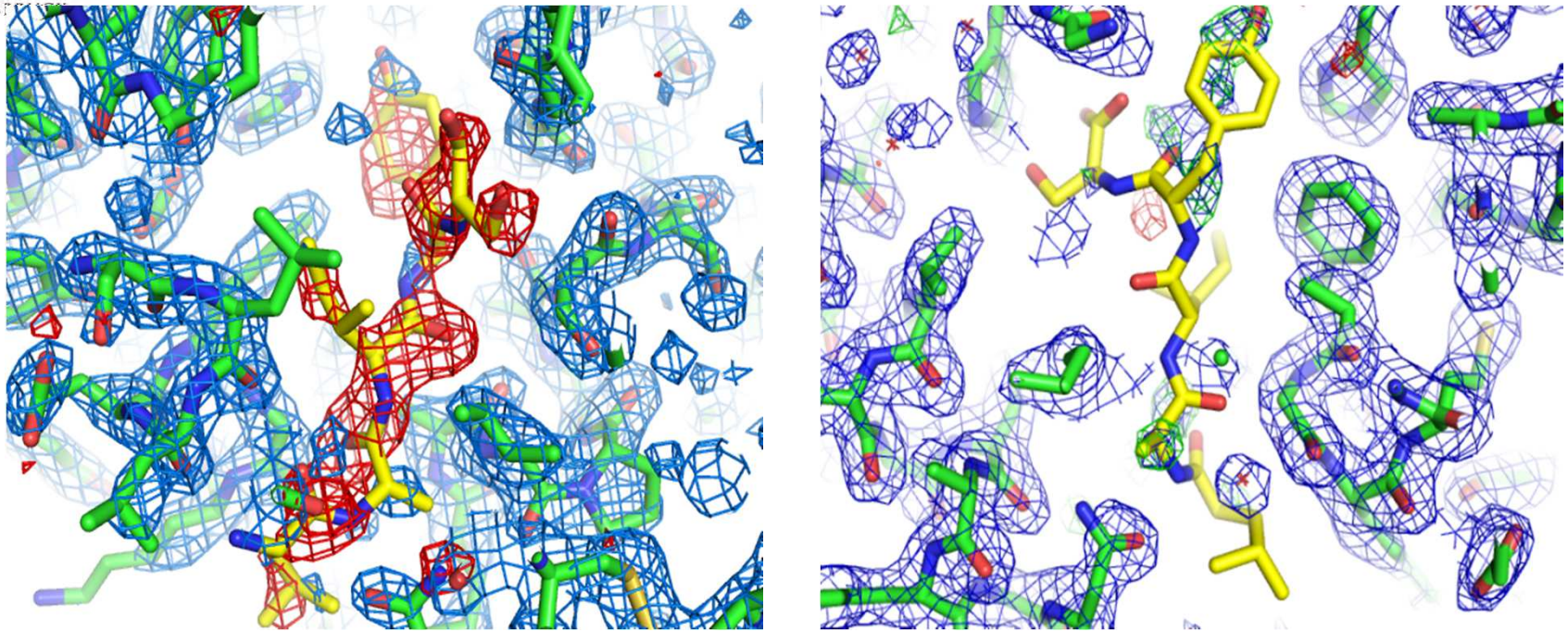
Figure 2. Real space correlation, *B*-factors, and electron density for hexa-saccharide in 1n7q. Top Panel: The real space correlation (black) for all saccharide units 1- 6 of the hexa-saccharide is abysmally low and the *B*-factors correspondingly high. Top left Panel: the Shake&wARP electron density 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 acid (3ib2).

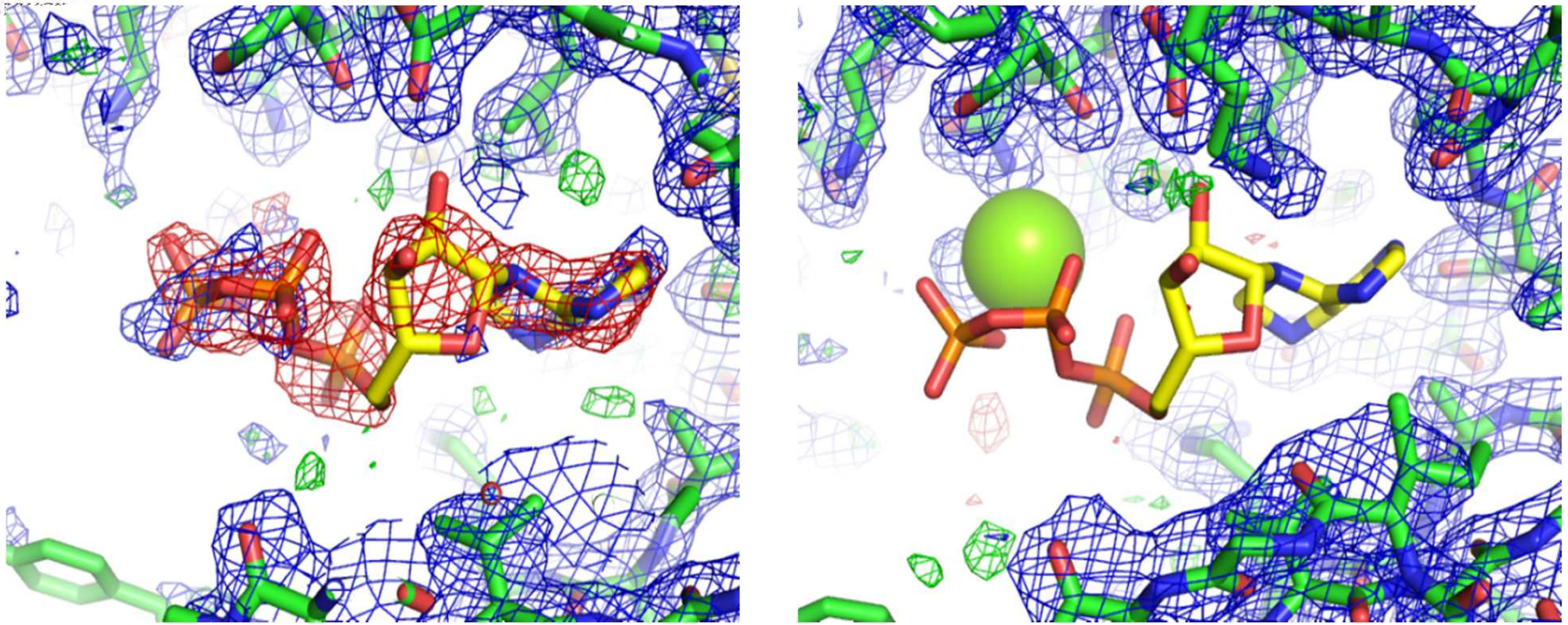
…and more ligands that just are not there



Absent inhibitor: the peptide inhibitor in the structure of phospholipase 2, PDB entry 1jq8, (Chandra *et al.*, 2002). The electron density maps downloaded from *EDS* show that the placed ligand overlaps with negative difference density below -3σ level (A) while the omit maps do not support ligand presence in the active site of the enzyme via positive difference density (B).

If one map type fails, others will likely too!

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.

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

The structure model must be viewed as a 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, grants…)

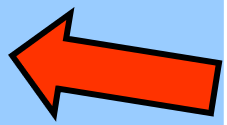
A final notice to the young scientist:

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.

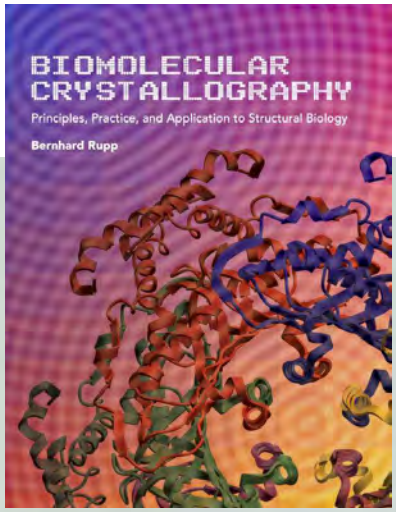
A final notice to the young scientist:

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 500BC.



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.



Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures.
E. Pozharski, C. X. Weichenberger and B. Rupp, Acta Crystallogr D69, 150-167 (2013)

Visualizing Ligand Molecules in Twilight Electron Density.
C. X. Weichenberger, E. Pozharski and B. Rupp, Acta Crystallogr. F69(2), 195-200(2013)

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