

# Scientific inquiry, inference and critical reasoning in the macromolecular crystallography curriculum

Bernhard Rupp

q.e.d. Life Science Discoveries Inc., USA. Correspondence e-mail: br@ruppweb.org

The astounding progress in methods and technology that led to the undisputed success and impact of biomolecular crystallography on areas ranging from essential structural biology to therapeutic drug discovery or the study of molecular complexes of ever increasing size and beauty has brought with it new requirements for the education of students in the field. With the great power of modern crystallography comes great responsibility for its appropriate use, and a modern curriculum must extend beyond the delivery of required technical expertise. The complexity of macromolecular models and the sometimes low determinacy, combined with local variety in structure quality, requires that the student is provided with means of critical analysis and hypothesis testing that extend beyond classical validation, developing a mindset that remains robust against mental bias towards finding what one seeks. The increasing neglect of critical analysis and hypothesis testing in many undergraduate curricula requires that the modern crystallography curriculum itself addresses such fundamental analytical tools of the scientific method to avoid high-profile structure retractions that might tarnish the otherwise unrivalled contributions of macromolecular crystallography to our understanding of the molecular basis of life.

© 2010 International Union of Crystallography  
Printed in Singapore – all rights reserved

## 1. Motivation

'The scientist must be the judge of his own hypotheses, not the statistician' (Edwards, 1992, p. 34).

A number of recent protein structures published in high-impact journals, some of them analyzed in editorials (Matthews, 2007; Petsko, 2007), indicate that not all is well in the way macromolecular crystallography and biological interpretation of structure models is taught to students. The negative impact of severely flawed crystal structures extends beyond mere nuisance. Crystallographic structure models carry great persuasive power – a wrong or misinterpreted structure that contradicts correct experimental findings can make it impossible for others to obtain funding for work related to such a structure. With the great power of modern crystallography comes great responsibility for its appropriate use. The mistaken notion that crystallography is just a basic analytical technique only amplifies the risk of uncritical use – and its propagation to the next generation – of increasingly powerful crystallographic methods. Interestingly, a closer inspection of incorrect structures almost always shows that an even deeper and more concerning general misconception and ignorance of the process of scientific inquiry led to the faulty structures and misinterpretations. These findings should be taken as an encouragement to teach crystallography in a conceptual framework that consistently emphasizes the role of experimental evidence and prior probability – in essence, treating structure model interpretation and validation as hypothesis testing in a Bayesian framework of inductive inference (Bayes, 1763). In such a curriculum, even once the

details of crystallographic theory are long forgotten, the fundamentals of proper inquiry and inference will remain invariably useful for students and future scientists.

The need for more education in the general aspects of scientific inquiry has also come up in the lively discussions during the crystallography education summit organized by the American Crystallographic Association and the US National Committee for Crystallography (USNC/Cr) which took place in 2005. The outcome of the discussions is the consensus policy statement on crystallography education and training available from the USNC/Cr (<http://sites.nationalacademies.org/pga/biso/IUCr/>). In an attempt to address some of these common concerns I will in the following provide a discussion of related ideas, viewpoints and examples that I have, after much discussion with critical colleagues and peers, also introduced in my textbook (Rupp, 2009; all figures from the book and other supplementary material are freely available from <http://www.ruppweb.org/>).

Some of my views are necessarily opinionated, all are open to discussion, and none are intended to dictate in any rigid form how critical thinking should be taught in a crystallographic curriculum – each educator must distill what serves one's own students best. As Albert Einstein phrased it, 'I never teach my pupils; I only attempt to provide the conditions in which they can learn.'

## 2. Some historic remarks and observations

One fascinating observation is that almost all of the significant controversies of recent times – where clearly questionable

structures have been published, whether intentionally (rare) or as a result of wishful thinking – could have been detected using the simple tools of inductive Bayesian inference (Bayes, 1763) or just common sense. Basic Bayesian reasoning almost always convincingly shows a low posterior probability for model correctness in such cases (which also sheds some doubt on the editorial and review process in these examples). It therefore seems appropriate to introduce a top-down view by treating macromolecular structure models, or the propositions deduced from those models, as a hypothesis that needs testing through critical validation.

To emphasize in any macromolecular crystallography curriculum the point of critical hypothesis testing early on – that the results of crystallographic work are models, not images or ‘real’ molecules of gospel-like truth – seems to be important to me for a number of reasons. Given the associated effort and expense, protein structures are usually determined to test some general or specific biological hypothesis. The element of unplanned discovery – which has contributed at least as much to our knowledge of biomolecular structures (Perutz, 2002) as strictly hypothesis-driven studies – has unfortunately taken backstage and seems at present underappreciated and hardly fundable. This has led to a development that is cause for concern, namely that structure determination is not always objectively conducted as a test for a hypothesis, but instead is sought as the ultimate confirmation of a biological hypothesis. This seemingly small difference between objectively testing *versus* seeking to confirm a hypothesis has widespread consequences. Almost all spectacular blunders in recent structure determinations originated from the desire to confirm or establish a pre-existing proposal and resulted from neglect to critically investigate the underlying molecular structure models in the face of available evidence. This can be described as mental model bias, which is quite insidious as no objective measure exists to determine its degree. As an example, the often unidentifiable electron density originating from the propensity of binding sites to accumulate whatever detritus is floating in the crystallization cocktail beckons to be filled with the desired ligand. The scientific method, fortunately, is quite robust, and carelessness or ignorance will not withstand public scrutiny, as recent history has shown. In fact, promoting training in critical thinking as a form of insurance might well increase its acceptance with students.

A second point that may need early emphasis is that we cannot derive the experimental evidence term without data, and as a consequence our ability to judge the merits of the model is severely impacted. Full evaluation of deposited structure models is only possible on the basis of deposited structure factor amplitudes. Yet, despite being the recommended practice of the International Union of Crystallography for a decade (Guss, 2000), it took until February 2008 to reach agreement on mandatory deposition of experimental structure factor amplitudes. Moreover, disclosure of primary data constitutes an essential foundation of science. Without data a structure model cannot be falsified (Popper, 2002) and, in Karl Popper’s sense, the work is not in compliance with

fundamental scientific methodology – that is, it is on a par with pseudo-science and quackery. One can therefore argue that out of scientific principle no protein structure model without supporting experimental data should ever have been accepted for publication, emphasizing to the disciple the importance of plausibly interpreted evidence in the form of minimally biased electron density.

The origin of the historic reluctance for depositing original data remains somewhat obscure, but it is a fact, quantified by Kleywegt & Jones (2002), that crystallographic trade journals such as *Acta Crystallographica Section D* already had a very high voluntary deposition rate of structure factors (>90%), while high-impact-factor journals (often affectionately called the magazines or vanity journals) such as *Nature*, *Cell* and *Science* enjoyed a much lower experimental data deposition rate. Revealed also was a (perhaps not surprising) correlation between poor *R* values and the tendency not to submit structure factor amplitudes. In this context, the additional preservation of the diffraction images should questions of data provenance arise (in addition to good scientific reasons) has been repeatedly discussed (Janssen *et al.*, 2007; Androulakis *et al.*, 2008). Such historical recount illustrates to students the fundamental and invariable importance of supporting experimental evidence in any scientific endeavor, irrespective of how complex (or maybe ambiguous) the interpretation ultimately may be.

### 3. Model interpretation as hypothesis testing

#### 3.1. Some origins of the problem

The crystallographic procedures leading to refined structure models are highly nonlinear and hard to parameterize, and the multivariate parameter landscape is infested with local minima. In addition, macromolecular refinement is sometimes on the brink of being underdetermined. As a consequence, known restraints such as universally valid stereochemistry are applied to keep the model in check with reality, and the resulting model is to a large degree dependent on – and therefore reflective of – our prior knowledge. It seems quite intuitive then that both the evidence supporting the model and the entire body of prior knowledge will determine the quality of the model and both must be subject to critical validation. Moreover, protein structure quality is a local property, and any hypothesis or model feature that is local in nature must be validated against local evidence. Global quality indicators such as *R* values are notoriously and necessarily insensitive to local errors.

#### 3.2. Bayes simplified

The common way to treat such a problem of inductive inference is to express the model likelihood, that is, the likelihood that our model (*M*) is described by the observed data (*D*), as a Bayesian posterior probability. The derivation and applicability of Bayes’ theorem is given in many textbooks (Edwards, 1992; Sivia, 1996; Jaynes, 2003; Rupp, 2009), but such depth of detail is not necessary for the following quali-

tative discussion. In simple terms, the sought-after model likelihood  $L_M(M|D)$  measuring the credibility of our model or hypothesis  $M$  will be a posterior probability in the form of the joint probability of the data likelihood function  $L_D(D|M)$  (which can be more or less readily calculated given the model) and all of the prior knowledge terms  $P_i(M)$  based on the plausibility of the model. The emphasis here is on ‘all’ – every piece of established knowledge (as well as any additional independent experimental evidence) is fair game in the context of validation.

$$L_M(M|D) \propto L_D(D|M) \times P_1(M) \times P_2(M) \times P_3(M) \dots \quad (1)$$

Equation (1) represents a joint probability and thus is subject to corresponding qualifiers. First, the multiplied individual probabilities need to be independent. This is not always correct for the prior probability terms  $P_i(M)$  of our model: a model that violates one fundamental law of physics will likely be incompatible with various other instances of established prior knowledge and thus the probability terms will not be independent. In a strict quantitative sense this may be imperfect, but for the purpose of estimating the overall plausibility of the model, the joint prior probabilities still provide a powerful tool. On a similar note, for our qualitative assessment, we do not worry too much about the absolute scaling or weighting of the data likelihood *versus* the prior probabilities (the problem does, however, exist for example in establishing proper restraint weights during refinement). We can just normalize the prior probabilities on a scale of one (highly probable, fully compatible with knowledge) to zero (highly improbable, severely contradicts prior expectations). Any hypothesis must provide correspondingly strong supporting evidence in data likelihood terms. A model that contradicts prior knowledge will have to provide very convincing and solid experimental evidence to persuade us to accept it as probable and eventually change our generally well established prior perceptions. Convincing and reproducible evidence is the means by which science eventually corrects its own misperceptions (Kuhn, 1970). An excursion into Thomas S. Kuhn’s *The Structure of Scientific Revolutions* may be suggested further reading for interested students at this point (compare also *Author’s notes* in Appendix A).

Of course one is permitted to present additional biological or biochemical data in support of the model or hypothesis: they just go into an additional likelihood term. Also those data need to be critically checked against all available knowledge. For example, one might have built a ligand of choice with acceptable geometry into some more or less featureless blob of density (possible), but the ligand has no reasonable nonbonded contacts to the protein (improbable). The posterior probability will not be convincing, not even in view of binding data, a low  $R$  value or great geometry for the rest of an otherwise superb structure: if the hypothesis is local, validation must be local. At this point it might also be sensible to review that, when faced with two or more competing models or hypotheses based on exactly the same data (the same data is important here), a Bayes factor  $L_M(M_1|D)/L_M(M_2|D)$  or a likelihood ratio (Edwards, 1992) can provide a qualitative

measure for how competing hypotheses compare (*cf.* Rupp, 2009, Sidebars 12–10 and 13–1).

### 3.3. Model validation as a responsibility

One can also approach the subject of model validation and analysis from the viewpoint of how to present the structure to a competent user or reader of a journal, and how to make sure that it withstands the scrutiny of a critical reviewer in the first round and the merciless scrutiny of time in the foreseeable future. More importantly, it is a matter of individual responsibility for the crystallographer to deposit a structure that is as close to a polished model as reasonably achievable, given the tools available at the time of deposition. Imagine the harm that a ‘hot’ but flawed structure published in a high-impact journal can cause to fellow researchers trying to obtain funding for a study based on preliminary results that contradict an incorrect structure (Petsko, 2007) or disagree with a flawed hypothesis developed from it. Crystallographic protein structure models of important drug targets or key molecules carry great weight. Their always pretty images carry strong persuasive power, and with great power comes great responsibility. The most prevalent misconception by users and also new depositors is perhaps that once a structure is accepted by the Protein Data Bank (PDB; Berman *et al.*, 2000) it is of sufficient quality and nothing can be wrong. This is a grave mistake. While the PDB and particularly its Validation Task Force work on improving and streamlining the validation process (Baker *et al.*, 2010), acceptance by the PDB means only that the model conforms to formal standards, making it acceptable and processable by the PDB. The depositor personally remains the final authority and responsible party regarding the correctness of the structure model and hypotheses.

Irrespective of the usefulness of prior-knowledge-based stereochemical, geometric or other validation criteria, a complete evaluation of a protein structure must include specific evidence and cannot be solely based on generic prior knowledge. It is the irrefutable evidence of a clear, minimally biased electron-density map that determines whether a reported outlier is the result of the inability to model a region of poor density correctly or a true feature of the structure. It is clear that the more a feature deviates from prior knowledge, the stronger the support in its favor must be. It is also true that such points of higher local energy often do carry significance, either for specific interactions at binding sites or perhaps as trigger points for conformational rearrangement in response to environmental change. The most powerful and complete analysis and validation of a structure will be inspection of the real-space electron density with special attention to the nature and location of stereochemical outliers or other low-prior-probability features.

## 4. Bayes in action

In the preceding section the case was made for using all available information, including experimental support as well

as all other independent and established physicochemical information, to determine the probability of a structure-based hypothesis. The purpose of the following examples is that analysis of a highly unlikely proposition will prompt students to critically question their own models and hypotheses until either convinced beyond reasonable doubt that they are correct or, if not, the students can clearly point out where they feel that more experiments are required or where follow-up studies may clarify the propositions. There is nothing wrong with seeking or proposing further support for a hypothesis in a publication – but there is reason for great caution should the model be incompatible with the known laws of physics. The former is a normal process of scientific investigation, the latter almost always a demonstration of poor judgment. One may be able to mask deficiencies and slip them by the reviewers in the first round, but the inevitable scrutiny by other researchers in the field will eventually expose any shortcomings.

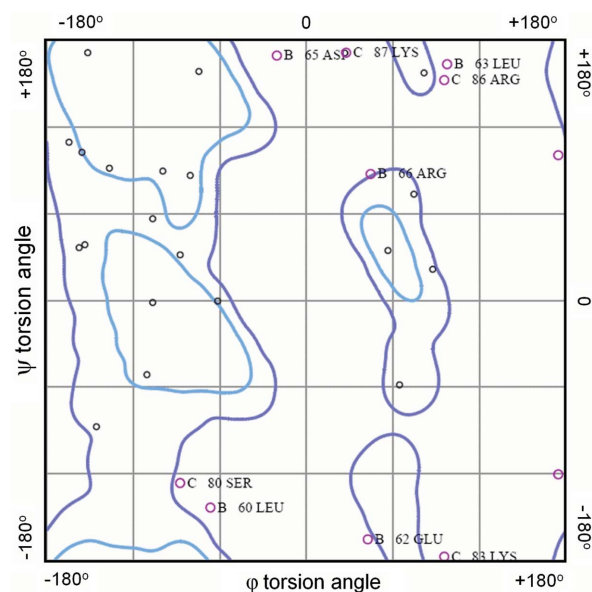
#### 4.1. Case study I: a highly improbable proposition

A controversial (Hanson & Stevens, 2000; Rupp & Segelke, 2001) and recently retracted (Hanson & Stevens, 2009) publication provides excellent illustration of the power of critical reasoning. The study describes the cocrystal structure of a 41-residue target peptide bound to a protease, the light chain of *Clostridium botulinum* neurotoxin B (BotLC/B). Because of their catalytic function botulinum toxins are the most toxic substances known and thus of great relevance and always good for a prime-time story in a high-impact magazine. After ingestion, the toxins enter the blood stream and are transported to the neuromuscular junction where they enter the presynaptic neuron and disrupt vesicle–membrane fusion by blocking the acetylcholine exocytosis pathway. Specifically, BotLC/B catalytically cleaves the synaptobrevin II (Sb2) peptide of the SNARE (soluble *N*-ethylmaleimide-sensitive-factor attachment receptor) complex mediating vesicle exocytosis in motor neurons thus leading to the flaccid paralysis effects in botulinum poisoning.

**4.1.1. Examination of prior probability.** The experimental section of the publication reveals the following ligand-soaking procedure. A 41-residue peptide (residues 38–88) was soaked into the crystals as follows: ‘Crystals of the Sb2–BoNT/B-LC complex were obtained by soaking pre-formed BoNT/B-LC crystals in a solution containing 15% (v/v) ratio of 2,4-methylpentanediol to mother liquor and a 2.5 molar excess of the Sb2 peptide fragment. The apo-crystals were soaked for approximately 5 s, followed by immediate freezing in liquid nitrogen.’ This is of course a most interesting procedure because diffusion is a slow process, and students may be well aware from soaking experiments that even small-molecule ligands take a long time – often hours – to diffuse through the solvent channels into pre-formed crystals (one can actually calculate the diffusion speed of such a peptide in random conformation from its radius of gyration and the second of Fick’s laws). The diffusion distance of such a 41-residue, random conformer peptide in pure water within 5 s is less than the reported dimensions of the crystal, not to mention threading of this

peptide through solvent channels. We give the presence of this peptide throughout the crystal therefore only a small probability based on simple physical chemistry considerations (noting that a small portion of protease molecules located at the outside of the crystal might still partially bind some peptide). The prior probability  $P_{PC}(Sb2)$  is thus a small number, say about 5% or 0.05.

The next peculiar observation results from the analysis of the backbone torsion-angle plot (Ramachandran plot) of the Sb2 peptide from the deposited coordinates (Fig. 1). It seems that this random peptide also has near random backbone torsion angles, again a highly improbable proposition. As the Ramachandran plot is in essence a projection of a potential energy surface, this peptide must be sitting in the binding site in an exceptionally high internal conformational energy (like a wound-up spring). Although some receptor-induced strain in the ligand is possible, this extreme scenario would require exceptionally strong nonbonded contacts to the protease and represents a highly improbable and so far not observed situation. In addition, most angle violations are located in regions where the peptide makes no contact with the protease, excluding any fit-induced high-energy conformation. The prior probability of this high-energy conformation  $P_{CF}(Sb2)$  is therefore very low as well, maybe 0.01. Remember: Extraordinary claims require extraordinary proof (and perhaps one might remind students that random peptide conformation does not imply nor allow random backbone torsion angles).



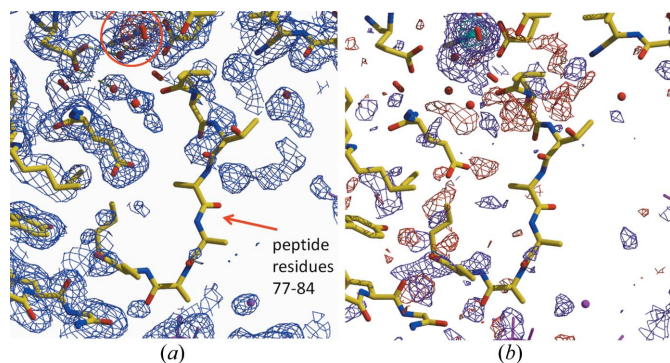
**Figure 1**

Backbone torsion-angle distribution for an improbable peptide. The panel shows a *MolProbity* plot (Davis *et al.*, 2007) of a purported random peptide with only 28% of residues in the favored region, 38% in additional allowed regions and 34% total outliers. The conformational energy of this peptide is extraordinarily high, and it is highly improbable that it can exist in this conformation. In addition (and as a consequence of its improbable geometry), the *MolProbity* all-atom clash score, a measure for serious steric overlaps, places the peptide into the zeroth percentile (worst) of available reference structure model entries. Reproduced from Rupp (2009) by permission of Garland Science/Taylor & Francis LLC.

Consistent with the Ramachandran plot appearance is the clash score for the peptide. According to *MolProbity* (Davis *et al.*, 2007), the clash score as a measure for improbable conformations puts the Sb2 model into the zeroth percentile of comparable structures. We are generous and give  $P_{CS}(Sb2)$  a probability of 0.01, simply because we notice that the clash score will not be independent of the backbone torsion violations. Strictly speaking, using clash score and backbone torsion outliers jointly in estimation of the prior probability is in violation of the independence requirements of probability algebra. However, even when treated as a single warning sign, either alone should be cause for great concern.

The biochemists among the students may also find that the direction of the peptide in the binding site is opposite to the well established canonical binding direction of a peptide in Zn proteases. While it has been shown that at low pH and high concentration at least the BotLC/A protease can be forced to digest itself in noncanonical binding mode (Segelke *et al.*, 2004), noncanonical binding again is a strong claim that would require strong support. We cautiously assign a prior probability for noncanonical binding  $P_{NC}(Sb2)$  of 0.2.

In search of the extraordinary proof now required to overcome the highly improbable prior probabilities, we first resort to the coordinate entry and observe that the mean  $B$  factor for the peptide is  $\sim 130 \text{ \AA}^2$ , approximately four times the average  $B$  factor for neighboring protease atoms. Based on



**Figure 2**

Bias-minimized electron density for an improbable peptide. The left map is a *REFMAC* (Murshudov *et al.*, 1997)  $2mF_o - DF_c$  omit map (*i.e.* the Sb2 peptide omitted in the model) reconstructed from  $\sigma_A$ -based maximum-likelihood coefficients (Read, 1986). The map contoured at the  $0.8\sigma$  density level (blue) shows no trace of the peptide (stick model, residues 77–84 shown in the figure). Density reconstruction indicates a possible water network in the vicinity of the binding site, but no trace of the peptide. The strong density (red,  $5\sigma$ ) for the also omitted Zn atom is emphasized by the red circle. The electron density reconstructed by other methods looks the same (Rupp & Segelke, 2001; Breidenbach & Brunger, 2004). Compare the corresponding electron-density real-space correlation coefficient (Rupp, 2009, Fig. 13–7). The right panel shows a *CNS* (*Brunger*, 2007) difference map reconstructed from cross-validated  $\sigma_A$   $F_o - F_c$  maximum-likelihood coefficients. Both the Sb2 peptide and the Zn ion were omitted in the map calculations. No positive difference density for the peptide at a  $2\sigma$  density level can be found in the  $F_o - F_c$  difference map. Note in contrast the strong positive difference density (blue, cyan) for the omitted Zn ion in the difference map. Negative difference density in red, positive difference density in blue (which should reveal Sb2 peptide density, if it were present). Structure factor amplitudes and model from obsolete PDB entry 3g94. Image created using *Xtalview* (McRae, 1999) and rendered with *Raster3D* (Merritt & Bacon, 1997).

these  $B$  factors, the relative scattering contribution will be low and we suspect – but have no proof as of yet – that not much electron density for the peptide might be visible in the electron-density map. This again is not conducive to proving the strong claim of the high conformational energy of the Sb2 peptide. We give the prior probability of a strong scattering contribution  $P_{SC}(Sb2)$  a correspondingly low value, 0.05.

**4.1.2. Examination of primary evidence.** So far our analysis has been limited to estimating the joint prior probability of the model, and has been solely based on comparing procedure and model features against independently acquired prior knowledge. Despite the fact that the result does not bode well for the model, strong proof might still require us to change a few fundamental laws of physics. We remain open minded and concede that we might be faced with a scientific revolution of Kuhnian dimensions (*cf. Author's notes*), and fairness requires that we must now consider the experimental support, that is, the data likelihood term supporting the model. We fully recognize that nearly all scientific revolutions result from initially unexplainable and often contradictory experiments that finally – when sufficiently supported by evidence – force us to change our theoretical beliefs and refine the underlying beliefs or theory.

This is the point where the ‘rubber meets the road’, or crystallographically speaking, the model meets the map, and availability of experimental structure factor amplitudes becomes crucial. If we did not have the structure factor file available, we could never support or disprove the claimed proposition based on primary evidence. It is a fundamental principle of science that others should be able to come to more or less the same conclusions given the data and proper procedures.

The examination of bias-minimized electron density as the cornerstone of crystallographic evidence leads to a result unfortunately in agreement with first impressions: the electron density – irrespective of the programs used to generate the maps or difference maps – shows practically no evidence for a peptide in the purported location. Fig. 2 illustrates this with a maximum-likelihood electron-density map as well as a difference map for the alleged BotLC/B–Sb2 complex. Based on the absence of electron density,  $L_M(D|Sb2)$  shall be about 0.01. The procedure for how (not) to generate at first sight convincing but (self)deceiving electron-density figures for non-existent ligands has been published and cautioned against (Rupp & Segelke, 2001).

**4.1.3. The verdict: joint posterior probability.** We finally summarize our findings qualitatively (subject to the limitations we discussed in §3.2) in the form of a simple joint probability for the posterior probability of the Sb2 peptide actually being present – that is, its model likelihood  $L_M(Sb2|D)$ :

$$\begin{aligned}
 L_M(Sb2|D) &\simeq L_D(D|Sb2) \times P_{PC}(Sb2) \times P_{CF}(Sb2) \times P_{CS}(Sb2) \\
 &\quad \times P_{NC}(Sb2) \times P_{SC}(Sb2) \\
 &\simeq 0.01 \times 0.05 \times 0.01 \times 0.01 \times 0.2 \times 0.05 \\
 &= 0.01 \times 5 \times 10^{-8} = 5 \times 10^{-10}. \quad (2)
 \end{aligned}$$



The clear qualitative result here is that – irrespective of a precise or absolute scale of the numbers – in order to overcome the extremely low prior probability of the Sb2 peptide being actually bound to the protease given the body of prior knowledge, we would have to provide at least an inversely huge supportive term to put this hypothesis up for discussion. This is certainly not the case.

Following this example, students should be encouraged to judge their model and hypotheses derived from the model with utmost scrutiny against all available independent prior information. It does not really matter what value one assigns to the respective evidence and probabilities; any honest guess will suffice if the propositions are truly questionable. This procedure will inevitably prevent major embarrassment but will also provide students with the confidence to boldly propose unconventional new insights as soon as their evidence supports it (again, all supporting evidence is fair game too, not just the electron density). After all, this is the stuff scientific revolutions are made of.

The original model (PDB entry 1f83) discussed above was superseded prior to the final retraction by a different model with partially occupied ( $n = 0.34$ ) Sb2 peptide (3g94) based on the same experimental data (and with phases of unspecified origin). As an exercise, students may compute their own biased-minimized electron-density maps by whatever accepted method they see fit and perform the same analysis on both obsolete entries.

The structure of the BotLC/A serotype protease bound to the SNAP 25 peptide has been determined by Breidenbach & Brunger (2004), who also confirmed the absence of the Sb2 peptide in the BotLC/B complex based on the deposited data. They avoided the problem of soaking a huge peptide into a crystal by co-crystallizing the target peptide with an inactive mutant of the protease. Another possibility is to link a nonhydrolyzable target peptide to the protease with a flexible linker and crystallize the ‘self-complex.’

#### 4.2. Case study II: perplexing features abound

An example emphasizing to students the importance of checking all published tables of crystallographic data – often and unfortunately condemned to the supplemental material section – for consistency is provided in the following. One of three concurrently published high-impact studies of a complex and flexible multi-domain molecule that is proposed to undergo large conformational changes stands out (Ajees *et al.*, 2006). Selected supplemental data collection and refinement statistics are shown in Table 1.

What strikes one immediately in Table 1 is the exceptionally low merging  $R$  value,  $R_{\text{merge}}$ , given the low signal-to-noise ratio,  $\langle I/\sigma(I) \rangle$ , in the last resolution shell. For an  $\langle I/\sigma(I) \rangle$  of 1.32, we would expect from statistics (Rupp, 2009) a corresponding  $R_{\text{merge}}$  in this shell of  $0.8/1.32 = 0.61$ , but instead we find an exceptionally low  $R_{\text{merge}}$  for the last resolution shell of 0.11. This is highly improbable and deserves some explanation. Similarly, for a 2.3 Å structure of purported high flexibility, the  $R$  values and an  $R_{\text{free}} - R$  gap of 1.4% are quite remarkable:

**Table 1**

Selected data collection and refinement statistics with interesting anomalies.

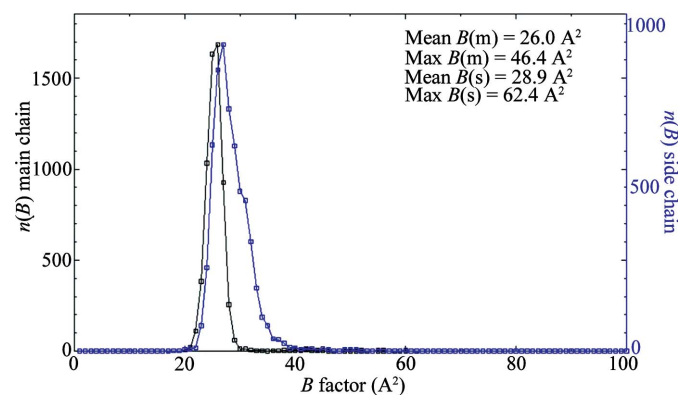
Evaluate the data-collection statistics for the last refinement shell (parameters in parentheses), and compare resolution,  $B$ -factor distribution (Fig. 3) and the  $R$  values for the refinement of a reportedly flexible molecule with disordered domains.

Data collection	
Space group	C2
Cell dimensions	
$a, b, c$ (Å)	151.20, 142.70, 203.70
$\beta$ (°)	98.9
Resolution (Å)	45.3–2.3
$R_{\text{merge}}$	0.07 (0.11)
Mean $I/\sigma(I)$	5.36 (1.32)
Completeness (%)	97.3
Refinement	
No. of reflections	194 135
$R_{\text{work}}/R_{\text{free}}$	18.0/19.4

again not impossible, but quite unusual and deserving attention (see Rupp, 2009, Fig. 12–24). Finally, if we plot a  $B$ -factor distribution for both the main-chain and side-chain atoms, we find an exceptionally narrow distribution (Fig. 3). This is highly improbable, particularly so in view of the claimed molecular rearrangement, the missing residues, and the flexibility and partly disordered domains in the structure.

Further examination has revealed (Janssen *et al.*, 2007) that the structure factors reflect no bulk solvent attenuation and that packing contacts exist in only two dimensions. As a result of all these unexplained and improbable irregularities, the origin of the structure factors has been called into question and the deposited structure factors ultimately proved to be (poorly) fabricated (Baker *et al.*, 2010). We are again faced with a highly improbable situation that would require strong experimental support, ultimately in the form of the original raw data, which could not be provided.

The forensics presented above should serve as an encouragement for students (and reviewers alike) to carefully check



**Figure 3**

Unusually narrow  $B$ -factor distribution. The  $B$ -factor distributions for the main chain (black graph) and side chain (blue graph) are extremely narrow and quite improbable for a 2.3 Å structure of a flexible molecule with missing side chains and entire missing domains. PDB entry 2hr0 (Ajees *et al.*, 2006). Compare this super-sharp distribution with a ‘normal’  $B$ -factor distribution in panel 13–21–A (Rupp, 2009). Reproduced by permission of Garland Science/Taylor & Francis LLC.

available data tables to ensure that they make sense and are consistent with proposed model features (and derived hypotheses) and with prior knowledge. In the vast majority of cases one will be able to prevent by simple consistency checks nuisance errors such as using refinement tables from intermediate stages, or depositing data collection statistics and structure factor files that do not belong to the model. The *B*-factor distribution plots are also a quick means to assure that neither scaling nor *B*-factor restraints have produced any irregularities that defy basic physical chemistry (see Rupp, 2009, §13.8).

## 5. Concluding remarks

Engaging students in an objective and analytical way of dissecting acknowledged mistakes in the recent literature can leave a lasting and habit-forming impression extending beyond the biomolecular crystallography core curriculum. A refresher of the methods of scientific inference and reasoning, particularly the power of a simplified Bayesian approach to the probabilistic judgment of claims, will benefit the disciple's scientific pursuits in general.

## APPENDIX A Author's notes

(1) In the abstract, I mention the failure of the undergraduate curriculum – particularly in the liberal arts – to provide adequate training in the scientific method. This does not come out of nowhere, but reflects the opinion of the practicing science educators who were kind enough to explain the US system of higher education to me. In addition, general education experts have voiced the same concerns; the essays in *Declining by Degrees* (Hersh & Merrow, 2005) draw a similar picture and call for corrective action.

(2) I argue that a strong correspondence exists between Thomas S. Kuhn's concept of scientific revolutions (Kuhn, 1970) and Bayesian inference. The difference seems to be that, in normal scientific progression, the prior knowledge terms – in the face of new evidence – experience a gradual refinement, whereby the new knowledge gained by the initially contradictory experiments becomes now part of the established prior knowledge. The scientific revolution is characterized by a fundamental and far reaching disagreement with prior knowledge terms, resulting in and in fact requiring what Kuhn calls a paradigm change. Even if such a discontinuity is the trademark of a scientific revolution, the basic Bayesian principle of balancing strong evidence with or against prior knowledge provides reliable guidance in the process.

(3) On occasion, in defence of hypotheses based on weak evidence, arguments such as the following may be voiced: 'In scientific pursuit we develop and follow a hypothesis until it has been proven to be flawed and we have not reached that point with our study.' This argument – essentially an abuse of Popper's (2002) falsification requirement and however 'open-minded' it may sound – is deeply flawed. The point of Baye-

sian inference – which actually and effectively guides us in everyday life as it does in scientific review – is exactly that not every outlandish hypothesis (e.g. being able to walk through walls) must be pursued until falsification. In a similar fashion, fringe scientists, UFO acolytes and other transcendentalists provide the argument that 'the absence of evidence is not the evidence of absence'. Bayes of course defeats this equally superficially open-minded argument, which ignores any prior probability terms: The crucial difference between the absence of a fossil find in an evolutionary sequence and the absence of UFO remnants is the strikingly different level of prior probability. In the same sense, prior knowledge dictates that discrete features of a crystallographic model must be associated with evidence of electron density. Absence of electron density is then evidence of absence.

(4) Another point related to the evasion of falsification is the introduction of unwarranted complexity. The equivalence of complexity (or over-parameterization) and non-falsifiability relates to the principle of parsimony, also known as Occam's razor, and is briefly discussed in the context of refinement and cross-validation by Rupp (2009, Sidebar 12.3).

(5) In Appendix XVII of his seventh German edition (Popper, 1982) of *Logik der Forschung* Popper argues against the usefulness of Bayesian inductive inference as a tool in the context of falsification. I find the arguments somewhat unsatisfactory, but unfortunately in the English Routledge edition (Popper, 2002) the appendices beyond XII are omitted. Compare also Edwards (1992, ch. 4) for a far more thorough discussion of the applicability of Bayes theorem.

(6) To properly respond to members of the open-minded crowd inclined towards anarchy, who often like to misquote from *Against Method* (Feyerabend, 1975), I suggest a helpful amended quote: 'Anything goes – as long as you have no idea what you are talking about'.

I wish to express my deep gratitude to the active educators and researchers in the crystallographic community who have kindly shared their concerns and opinions with me.

## References

- Ajees, A. A., Gunasekaran, K., Volanakis, J. E., Narayana, V. L. S., Kotwal, G. J. & Murthy, H. M. K. (2006). *Nature (London)*, **444**, 221–225.
- Androulakis, S. *et al.* (2008). *Acta Cryst.* **D64**, 810–814.
- Baker, E. N., Dauter, Z., Einspahr, H. & Weiss, M. S. (2010). *Acta Cryst.* **D66**, 115.
- Bayes, T. (1763). *Philos. Trans. R. Soc. London*, **53**, 370–418.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
- Breidenbach, M. A. & Brunger, A. T. (2004). *Nature (London)*, **432**, 925–929.
- Brunger, A. T. (2007). *Nat. Prot.* **2**, 2728–2733.
- Davis, I. W. *et al.* (2007). *Nucleic Acids Res.* **35**, W375–W383.
- Edwards, A. W. F. (1992). *Likelihood – An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*, expanded ed. Baltimore: Johns Hopkins University Press.
- Feyerabend, P. K. (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge*. New York: New Left Books.

- Guss, M. (2000). *Acta Cryst.* **D56**, 2.
- Hanson, M. A. & Stevens, R. C. (2000). *Nat. Struct. Biol.* **7**, 687–692.
- Hanson, M. A. & Stevens, R. C. (2009). *Nat. Struct. Mol. Biol.* **16**, 795.
- Hersh, R. H. & Mellow, J. (2005). *Declining by Degrees*. New York: Palmgrave Macmillan.
- Janssen, J. C., Read, R. J., Brunger, A. T. & Gros, P. (2007). *Nature (London)*, **448**, E1–E3.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.
- Kleywegt, G. J. & Jones, T. A. (2002). *Structure*, **10**, 465–472.
- Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*, 2nd ed. University of Chicago Press.
- Matthews, B. W. (2007). *Protein Sci.* **16**, 1013–1016.
- McRae, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
- Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.
- Perutz, M. F. (2002). *I Wish I'd Made You Angry Earlier: Essays on Science, Scientists, and Humanity*. Oxford University Press.
- Petsko, G. A. (2007). *Genome Biol.* **8**, 103–105.
- Popper, K. (1982). *Logic der Forschung*. Tuebingen: J. C. B. Mohr.
- Popper, K. (2002). *The Logic of Scientific Discovery*, 14th printing ed. New York: Routledge.
- Read, R. J. (1986). *Acta Cryst.* **A42**, 140–149.
- Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st ed. New York: Garland Science.
- Rupp, B. & Segelke, B. W. (2001). *Nat. Struct. Biol.* **8**, 643–664.
- Segelke, B., Knapp, M., Kadhkodayan, S., Balhorn, R. & Rupp, B. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 6888–6893.
- Sivia, D. S. (1996). *Data Analysis – A Bayesian Tutorial*. Oxford University Press.