Published in final edited form as:

Curr Opin Struct Biol. 2013 October ; 23(5): 707–714. doi:10.1016/j.sbi.2013.06.004.

Crystallographic Model Validation: from Diagnosis to Healing

Jane S. Richardson, Michael G. Prisant, and David C. Richardson

Department of Biochemistry, Duke University, Durham, North Carolina 27710 USA

Abstract

Model validation has evolved from a passive final gatekeeping step to an ongoing diagnosis and healing process that enables significant improvement of accuracy. A recent phase of active development was spurred by the worldwide Protein Data Bank requiring data deposition and establishing Validation Task Force committees, by strong growth in high-quality reference data, by new speed and ease of computations, and by an upswing of interest in large molecular machines and structural ensembles. Progress includes automated correction methods, concise and user-friendly validation reports for referees and on the PDB websites, extension of error correction to RNA and error diagnosis to ligands, carbohydrates, and membrane proteins, and a good start on better methods for low resolution and for multiple conformations.

Introduction

Historical Background and Definitions

The general presumption behind crystallographic model validation is that in addition to explaining the experimental data, macromolecular coordinates must also be consistent with basic physics and chemistry and with prior experience of what holds true for molecular structures. If ailments are diagnosed, healing is called for.

Model validation as practiced since the early 1990s (reviewed in [1]) has three primary components: geometry, conformation, and sterics. Traditionally, geometry includes covalent bond lengths and angles ([2] for proteins; [3] for nucleic acids), plus planarity and chirality where appropriate. Deviations from geometric ideality primarily reflect the weighting terms in refinement, so usually they affect atomic coordinates only slightly. However, WhatCheck utilizes overall directional bond-length deviations to diagnose errors in unit cell dimensions [4], and some combinations of bond angles are sensitive to local fitting errors, such as the C deviation to backward-fit sidechains [5]. Conformation covers validity of dihedral angle values, assessed in multi-dimensional sets such as 2-D Ramachandran plots [6], up to 4-D protein sidechain rotamers [7], 2-D glycoMaps for carbohydrate linkages [8], and 7-D backbone conformers for RNA [9]. Sterics encompass non-covalent interactions: unsatisfied H-bonds [4], steric clashes as best measured with all hydrogens explicit [10], and preferred local environments around sidechains [11].

The above criteria are local in the model: each outlier is within a residue or between two residues. Global scores for an entire structure are obtained from the number of outliers,

© 2013 Elsevier Ltd. All rights reserved.

Corresponding author: Jane S. Richardson, 1-919-684-6010, jsr@kinemage.biochem.duke.edu, 132 Nanaline Duke Bldg, DUMC 3711, Durham, NC 27710 USA.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

suitably normalized by total residues, parameters, or atoms. For criteria with well-behaved reference data distributions, such as covalent geometry, outliers are usually declared at 4 σ , or Z-score=4. Conformational outliers are traditionally defined by a boundary enclosing nearly all filtered reference data, such as outside the 99.95% or 1-in-2000 contour for Ramachandran outliers (equivalent to Z=3.5) [1]. Clashes are declared at overlap>0.4Å, and overall “clashscore” = clashes per 1000 atoms [12]. Most global validation scores depend strongly on resolution, the single number most indicative of crystal structure quality. This relationship is sometimes made explicit by reporting percentile scores relative to PDB x-ray structures within a resolution cohort [12].

Theory & Practice of Traditional Model Validation

Most validation outliers are errors, but a few are functionally significant anomalies where evolution has found it worthwhile to spend some net folding energy on stabilizing an unfavorable local conformation. Since active stabilization is required to hold a group in an unfavorable position, valid outliers should have clear electron density and be constrained by H-bonds or packing interactions [13] (Figure 1a).

Diagnosing problems is a much more productive endeavor if anomalies can be distinguished from errors and some large fraction of the latter can be corrected, or “healed”. This reversed meaning of “molecular medicine” is one motivation for many recent developments in model validation. The errors worth correcting in macromolecular models are qualitative misfittings into the wrong local minimum (Figure 1b). Rather than random, those are nearly always systematic errors arising from a particular misinterpretation to which either people or programs are prone [13].

The earliest form of reliable, automated correction of model-fitting errors is for 180° “flips” of sidechain amide or histidine ring orientations [14, 15, 16]). Flips are common because of indistinguishable electron density for N vs O or C – which also means they can be corrected without reference to diffraction data. Flip state is especially important for histidines, because it dominates the determination of protonation state [17]. Since 2002, the incidence of incorrect Asn/Gln/His flips in new PDB depositions worldwide has decreased by 45% [18], a conspicuous win for the community.

A theoretical issue recognized by early contributors [19, 20] is that validation criteria are stronger if independent of the target functions used in refinement. This produces an inherent conflict between using all relevant information to produce the best attainable model and still being able to reliably evaluate the accuracy of the result. There are several partial answers to this dilemma. First, the argument applies to global measures, not to individual problems, and to complacency over good scores, not to distress over bad ones. Ideal covalent geometry does not reflect general model quality, but it is inexcusable to retain a 10 σ bond-length outlier – those occur disturbingly often even at very high resolution if geometry terms are downweighted [18]. The second point is that we should not (and probably cannot) prevent crystallographers from using all available information in order to succeed for difficult cases. Rather, validators should continue developing new criteria and methods, perhaps including full cross-validation enabled by the computational power to routinely perform parallel structure determinations with different information left out each time. The third, most important, point is that any one validation criterion can be “gamed”, but the entire set of model plus data criteria (at least for protein interiors, currently, at reasonable resolutions) provides cooperative constraints that cannot be satisfied simultaneously without having the right answer.

For the practitioner of model validation and healing, we suggest two guiding principles. The first, from the classic T-shirt of the 2010 Cold Spring Harbor crystallography course (Figure

1b), is “At least one person should look at the map!” Even the best automated methods will occasionally make a silly mistake for some new case the programmer did not expect. The second principle is the “Zen of Model Anomalies” [18]:

- Consider each outlier and correct most.
- Treasure the meaningful, valid few.
- Live serenely with the small inscrutable remainder.

Current Developments (2009-2013)

New Access to Concise and Detailed Quality Metrics

The current era in crystallographic model validation and healing was initiated in 2008, when the worldwide Protein Data Bank (wwPDB) required deposition of structure factors and constituted the wwPDB X-ray Validation Task Force committee, the first of the wwPDB VTFs [21]. Reliable availability of diffraction data not only allows validation of the data itself, such as detection of twinning [22, 23], and of model-to-data match [24], but also enables better quality-filtering of reference data, more nuanced model validation, and potential future improvement of already-deposited structures as methods become more powerful. Analogously to the high-profile errors [25] that prompted the start of structure validation, both the PDB and journals (e.g., [26]) were motivated to strengthen validation by several incorrect [27] or even fabricated [28] structures, as well as by the more productive new possibilities for making all crystal structures more accurate.

The X-ray VTF was tasked with recommending what should be included in a report for journal referees, as a summary on each structure's PDB web page, and as detailed global and per-residue scores for depositors and users. The VTF report [1] reviews pre-existing validation tools as applied to the entire x-ray PDB, considers extensions and updates, and advises on how to present them clearly and intuitively without requiring deep knowledge of either crystallographic or validation methodology. A central feature is the idea of expressing global scores as percentiles, now under implementation by the wwPDB [29]. Scores relative to structures at similar resolution allow referees and depositors to judge how capably the available data was handled, and scores relative to all PDB x-ray structures give end-users a more absolute comparison between structures at different resolutions. Both types can be shown together on slider bars, as in Figure 2. One complication is that a standardized definition of resolution would be desirable to prevent the spikiness in validation-score vs resolution plots ([1] supplement) that results from some users subjectively rounding down to even tenth-Å values.

Per-residue outlier flags are important for evaluating local biological questions and for enabling crystallographers to correct problems. They work best when showing clusters of multiple criteria along the sequence or in 3D. Initial implementation by the wwPDB [29], soon available, will probably use lists for referee reports but will provide sequence plots of outliers on the web sites. An especially valuable feature will be the diagnosis of “wishful thinking” ligands without adequate density [20] such as [30], by comparing their average real-space RSR-Z score [31] with that for nearby protein.

Updated Parameters: Reinforcing the Foundations

One aspect of current progress involves using newly detailed knowledge to reinforce the basic foundations of validation – the reference data and distributions against which models are judged – to hold up under newly demanding expectations of universal usage and model correction. A simple example is expansion of the new wwPDB Ramachandran criteria to six types, adding plots for Ile/Val and cis-Pro [1].

Previously, mean values of the $\angle$ (N-C-C) were shown to be a function of ϕ , [32], and RNA bond angles and dihedrals to be a function of ribose pucker [9]. Such considerations have now been expanded into a full conformation-dependent library for backbone covalent geometry as a function of ϕ , [33] and into pucker-dependent refinement targets for RNA [34].

Now that explicit H atoms are central in correcting as well as validating sterics [12] and are often used in refinement [34], the traditional crystallographic parameters for parent-atom-to-hydrogen effective bond lengths have been re-examined, along with re-optimization of hydrogen van der Waals radii [35]. The biggest issue is the difference of $>0.1\text{\AA}$ between the distance to the H nucleus (appropriate for neutron or NMR structures) and the shorter distance to the center of the H electron cloud (appropriate for x-ray structures). Most existing software uses the wrong one of these alternatives at some point. Within each alternative, some smaller changes were recommended [35], based on quantum electron density calculations, new surveys of small-molecule neutron coordinates in the Cambridge Structural Database (CSD) [36], and surveys of H difference peaks for small-molecule x-ray structures in the Crystallography Open Database (COD) [37] and for protein structures at $<1\text{\AA}$ in the PDB. Tests of the new parameters showed improved convergence of all-atom clashscores to zero for the best protein structures. The resulting H positions and radii are implemented for x-ray and neutron refinement and validation in Phenix v1.8.2 and MolProbity4 [35].

New Criteria and Coverage

The X-ray VTF identified a promising new criterion for underpacking, or cavities, in proteins. Steric clashes are a straightforward and sensitive measure of overpacking, but underpacking is more difficult. Although methods exist for identifying cavities, for instance using Voronoi or related volume concepts [38], none have provided practical diagnostics adopted into validation. RosettaHoles [39] is a less mathematically elegant but more practical system of filtered overlapping spheres. Extensive tuning to distinguish predicted from experimental structures has given it good traction on suboptimal experimental structures, identifying known problems with a low false-positive rate [1, 40].

Another challenge for validation is membrane proteins, with standard geometry, conformation, and clashes but unusual packing and sidechain environments [41]. It was found that glycines are strongly disfavored at surfaces exposed to the membrane, giving a statistical check for incorrectly modeled lipid interfaces. The most significant issue is that many media for crystallizing membrane proteins, such as detergents or micelles, risk seriously distorting the structure from its compact functional form [41]. The resulting underpacking spreads helices apart, which can be identified visually but will require new measures for automated recognition.

Model validation is also being extended to carbohydrates, biologically important and increasingly prevalent constituents with a variety of monomers and complexity of linkages that renders them highly error-prone. This area was pioneered with checks for nomenclature and linkage types [8, 42] and was recently reviewed [43]. Newer force fields for carbohydrates [44] should prove helpful in both refinement and error detection, and a new system facilitates data-mining of glycan conformations in the PDB [45]. Carbohydrate validation is under active development in Phenix, in Rosetta, and at the PDB.

The PDB will now also be able to improve their validation of small-molecule ligand geometry and conformation, because the CSD has agreed to make their Mogul system [46] available for that purpose [29]. A recent survey [47] showed that ions, especially important in nucleic acid structures, are frequently misidentified. Crystallographic waters are also

notoriously unreliable, whether manually or automatically placed, because of partial occupancy and because those isolated density peaks may actually be ions, or noise, or unfit alternate conformations [48]. Suitable peak shape and position for real waters can be diagnosed by DDQ [49], and tools are under development for re-classification by the nature and pattern of contacts [48].

At very high resolution, a difficulty is the wealth of suggestive information on multiple conformations. Two new systems identify potential alternates, one by analyzing electron density just above noise level but in the right position [50], and the other by analyzing shifts after unrestrained refinement trials [51]. Building those starting points into consistent alternative models is still a considerable challenge. Also, the end-points of alternates should be less arbitrarily defined to avoid unjustified geometrical distortion [18].

Several developments exploit expanded reference data and computational capabilities to move validation into higher dimensions – adding either continuous dimensions or discrete ones such as residue identity. A hierarchical Dirichlet process was used to define 20×20 neighbor-dependent Ramachandran plots [52], spreading the effect of related information to counteract small sample size. The combination of backbone and sidechain dimensions is also being explored [53, 54], uncovering relationships not seen in the simpler versions.

New Correction Methods and More Automation

A main goal of current model validation is either avoiding or correcting problems. Modern model-building software takes into account validation criteria and quality-filtered rotamer or fragment libraries. Thus the user starts off with favorable, regular conformations and should only allow distortion if the data requires it later. Coot [55] provides both its signature user-tugged real-space optimization and also automated rotamer search with backrub shifts [56], displaying validation traffic-light summaries afterward. Phenix [34] has full MolProbity validation [12] built in, clickable from any outlier to center in Coot at the right spot for fix-up. It also has an automated rotamer refitting option in phenix.refine, using real-space match for choice and minimization, but checking clashes and rotamericities for acceptance. Even more ambitious rebuilding is done by the second version of the PDB-redo project [57], which re-refines all PDB crystal structures with modern default methods from the CCP4 suite and now also does automated real-space sidechain corrections. Each of these sidechain rebuilding systems can make significant improvements to most structures, but for future use at lower resolutions where electron density is insufficiently informative, steric and conformational terms will be needed in the inner loop. Automated methods should be pursued whenever feasible, but we must realize that total automation is unrealistic for such complex and evolving problems: automation can only converge toward correctness.

An especially exciting push toward automated and effective correction of model problems is the application of energy functions and sampling strategies separately developed for structure prediction and design. Systems such as Rosetta have achieved significant success at the extremely difficult prediction problem, but recently it became clear that they can deliver consistently impressive results when provided with modest amounts of experimental data. Structure validation and correction is now benefiting from two such efforts, one for proteins and one for RNA. Rosetta and Phenix together have been used to improve low-resolution protein crystallographic models [58], with success at least on par with other low-resolution refinement approaches. Another effective hybrid system called Erraser tackles the recalcitrant problem of correcting RNA backbone conformations [59]. Fitting RNA backbone is difficult and error-prone because there are 6 torsional parameters per nucleotide, with only the phosphate and the base clearly defined at typical resolutions. Erraser uses MolProbity diagnosis of clashes, ribose puckers, and backbone conformations to decide which nucleotides need work, and applies cycles of Phenix refinement, Rosetta relax

refinement, and a new “step-wise assembly” Rosetta-based exhaustive local search protocol. It nearly always generates valid backbone conformers [9] without using them as explicit targets, and it corrects validation outliers of all types with a thoroughness unprecedented for RNA, as can be appreciated in Figure 3.

Future Directions

We will soon see the appearance, and then the effects, of more complete, understandable, and obvious validation at the PDB. Similarly, we will see whether more dimensions are the right answer for validation, or only for design. High-productivity tools will deliver robust diagnosis and easy correction for previously neglected components – ligands, ions, waters, carbohydrates, nucleic acids, and alternate conformations. Model healing will be increasingly integrated into our automated protocols, to become model wellness.

Accurate Low-Resolution Structures

Even more visionary possibilities are on the horizon. The concerted use of tuned diagnostics, exhaustive conformational sampling, and more outside information may make possible quite accurate models even at resolutions poorer than 3 Å. That would be revolutionary, since typical low-resolution structures not only have steric clashes everywhere, implausibly irregular secondary structure and missing loops, but are also at risk for serious problems such as sequence misregister. For example, 6 of 27 ribosomal proteins originally fit at 3.5 Å needed 7- to 46-residue sequence-shifts rebuilt ([60] supplement), perhaps amenable to automated diagnosis. At low resolution some atoms are genuinely outside density, which leads both people and programs to scrunch sidechains, misorient peptides and break H-bonds. A possible corrective is the new CaBLAM parameter space, using C α dihedrals to diagnose what secondary structure or motif is present, and CO dihedrals to diagnose impossible local conformation [18]. Once large-scale problems are corrected, other techniques can take over: better low-resolution refinement [61-63] and better sampling for repacking [58, 59].

Realistic Ensembles

Dynamics and structural ensembles are of growing interest, crystallographically relevant all the way from alternate conformations at atomic resolution to the role of molecular motions in causing low resolution. Model building would be transformed by achieving the aggressive sampling and aggressive validation needed to reliably fit the best few sidechain conformations into confusing density, or to calculate the few most physically realistic conformations of a short disordered loop or terminus. Promising starts have been made for local alternates [64] and for loop modeling, reviewed in [65]. There are already ensemble methods that make explicit the local uncertainty of current procedures [66, 67], but they refine each model separately against the data and seldom capture large changes. Refinement of the entire ensemble together, along with wide sampling, is clearly needed but multiplies the number of parameters greatly. Multicopy ensemble refinement has been tested with initial variability from simulated annealing [68], and early work [69] is being revisited for MD refinement with an exponentially decaying weight on earlier ensemble snapshots (available in Phenix 1.8.2). Such techniques rely heavily on accuracy of the force field. The right combination of sufficient conformational variability with individual models that can pass validation has not yet been achieved but is coming closer. In another contribution toward ensembles, a recent manifold-mapping algorithm has been generalized to extract multiple as well as single conformations from many low-signal, randomly oriented snapshots such as free-electron-laser data [70].

Paragon Structures

At a final extreme, validation and correction methods might aspire to create what we will call “paragon” structures, free of features current methods can show are wrong: no serious all-atom clashes, near-ideal H-bonding and packing, no significant unexplained difference density, no Ramachandran or rotamer outliers unsupported by clear density and stabilizing interactions, and R-factors commensurate with measured experimental accuracy. A paragon would have perfect scores in well-ordered core regions and an internally consistent model (as in Figure 3b), at least for the “A” alternative conformation and preferably for several conformations. Each such model could then be considered a valid, high-probability member of that molecule's dynamic ensemble. Paragons would educate us about concerted motions and would make optimal test cases for methods development in structural and computational biology.

Acknowledgments

Funding for this analysis is from NIH research grants R01-GM073930 and R01-GM073919.

References

1. Read RJ, Adams PD, Arendall WB III, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lütke T, Otwinowski Z, et al. A new generation of crystallographic validation tools for the protein data bank. *Structure*. 2011; 19:1395–1412. [PubMed: 2200512] **Recommendations of the wwPDB X-ray Validation Task Force, extensive explanations behind those, and a review of the field prior to VTF work
2. Engh RA, Huber R. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallogr*. 1991; A47:392–400.
3. Parkinson G, Vojtechovsky J, Clowney L, Brünger AT, Berman HM. New parameters for the refinement of nucleic acid containing structures. *Acta Crystallogr*. 1996; D52:57–64.
4. Hoof RWW, Vriend G, Sander C, Abola EE. Errors in crystal structures. *Nature*. 1996; 381:272. [PubMed: 8692262]
5. Lovell SC, Davis IW, Arendall WB III, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C geometry - , and C deviation. *Proteins: Struct Funct Genet*. 2003; 50:437–450. [PubMed: 12557186]
6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. ProCheck - a program to check the stereochemical quality of protein structures. *J Appl Cryst*. 1993; 26:283–291.
7. Dunbrack RL Jr. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*. 2002; 12:431–440. [PubMed: 12163064]
8. Frank M, Lütke T, von der Lieth CW. GlycoMapsDB - a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res*. 2007; 35:287–290. [PubMed: 17202175]
9. Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LW, et al. RNA backbone - consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium Contribution). *RNA*. 2008; 14:465–481. [PubMed: 18192612]
10. Word JM, Lovell SC, LaBean TH, Zalis M, Presley BK, Richardson JS, Richardson DC. Visualizing and quantitating molecular goodness-of-fit - small-probe contact dots with explicit hydrogens. *J Mol Biol*. 1999; 285:1711–1733. [PubMed: 9917407]
11. Vriend G, Sander C. Quality control of protein models - directional atomic contact analysis. *J Appl Cryst*. 1993; 26:47–60.
12. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity - all-atom structure validation for macromolecular crystallography. *Acta Crystallogr*. 2010; D66:12–21. **Most recent review of MolProbity validation tools for protein and RNA

13. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Proteins: Struct Funct Genet.* 2000; 40:389–408. [PubMed: 10861930]
14. McDonald IK, Thornton N. Satisfying hydrogen bonding potential in proteins. *J Mol Biol.* 1994; 238:777–793. [PubMed: 8182748]
15. Hoof RWW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct Funct Genet.* 1996b; 26:363–376. [PubMed: 8990493]
16. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine -using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol.* 1999; 285:1735–1747. [PubMed: 9917408]
17. Richardson JS, Prisant MG. Fitting Tip #3: Histidine flips and protonation. *Comput Crystallogr Newsletter.* 2012; 3(1):3–4.
18. Richardson JS, Richardson DC. Doing molecular biophysics: Finding, naming, and picturing signal within complexity. *Annu Rev Biophys.* 2013; 42:14.1–14.28. on-line. *Introduces alternate conformation treatment, CaBLAM low-resolution validation, and improvement in PDB clashscores
19. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Proteins: Struct Funct Genet.* 1992; 12:345–364. [PubMed: 1579569]
20. Kleywegt GJ. Validation of protein crystal structures. *Acta Crystallogr.* 2000; D56:249–265.
21. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The future of the Protein Data Bank. *Biopolymers.* 2012; 99:218–222. [PubMed: 23023942]
22. Zwart PH, Grosse-Kunstleve RW, Adams PD. Xtriage and Fest – automatic assessment of x-ray data and substructure structure factor estimation. *CCP4 newsletter.* 2005; 43:99–107.
23. Lebedev AA, Vagin AA, Murshudov GN. Intensity statistics in twinned crystals with examples from the PDB. *Acta Crystallogr.* 2006; D62:83–95.
24. Afonine P, Grosse-Kunstleve RW, Chen VB, Headd JJ, Moriarty NW, Richardson JS, Richardson DC, Urzhumtsev A, Zwart PH, Adams P. phenix.model_vs_data: a high-level tool for the calculation of crystallographic model and data statistics. *J Appl Cryst.* 2010; 43:677–685.
25. Branden CI, Jones TA. Between objectivity and subjectivity. *Nature.* 1990; 343:687–689.
26. Baker EN, Dauter Z, Einspahr H, Weiss MS. In defence of our science – validation now! *Acta Crystallogr.* 2010; D66:115.
27. Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ, Chen AP. Retraction. *Science.* 2006; 314:1875. [PubMed: 17185584]
28. Janssen BJC, Read RJ, Brunger AT, Gros P. Crystallography: crystallographic evidence for deviating C3b structure. *Nature.* 2007; 448:E1–E3. [PubMed: 17687277]
29. Gore S, Velankar S, Kleywegt GJ. Implementing an x-ray validation pipeline for the Protein Data Bank. *Acta Cryst.* 2012; D68:478–483. **In-progress work on the new wwPDB validation system, including a statement of goals, software, strategy and timing, and specific representation for users at PDBe
30. Hanson MA, Stevens RC. Retraction: Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0Å resolution. *Nat Struct Mol Biol.* 2009; 16:795. [PubMed: 19578378]
31. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. The Uppsala electron-density server. *Acta Cryst.* 2004; D60:2240–2249.
32. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* 1996; 5:1406–1420. [PubMed: 8819173]
33. Berkholtz DS, Shapalov MV, Dunbrack RL Jr, Karplus PA. Conformation dependence of backbone geometry in proteins. *Structure.* 2009; 17:1316–1325. [PubMed: 19836332] *The “CDL”: a library of bond lengths and angles as a function of ,
34. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, et al. PHENIX - a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* 2010; D66:213–221. *Includes description of comprehensive validation built into the Phenix GUI with one-click Coot links; also use of CDL geometry library and pucker-specific RNA parameters

35. Deis LN, Verma V, Videau LL, Prisant MG, Moriarty NW, Headd JJ, Chen VB, Adams PD, Snoeyink J, Richardson JS, Richardson DC. Phenix/MolProbity hydrogen parameter update. *Comput Crystallogr Newsletter*. 2013; 4(1):9–10. *Adjusted parameters of effective x-H distances for refining and validating atomic interactions in x-ray vs neutron structures, and of re-tuned H van der Waals radii
36. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr*. 2002; B58:380–388.
37. Grazulis S, Chateigner D, Downs RT, Yokochi AT, Quiros M, Lutterotti L, Manakova E, Butkus J, Moeck P, Le Bail A. Crystallography Open Database – an open-access collection of crystal structures. *J Appl Cryst*. 2009; 42:726–729. [PubMed: 22477773]
38. Esque J, Oguey C, de Brevern AG. A novel evaluation of residue and protein volumes by means of Laguerre tessellation. *J Chem Inf Model*. 2010; 50:947–960. [PubMed: 20392096]
39. Sheffler W, Baker D. RosettaHoles - rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci*. 2009; 18:229–239. [PubMed: 19177366] *Validation for cavities or other abnormal underpacking
40. Tronrud DE, Matthews BW. Sorting the chaff from the wheat at the PDB. *Protein Sci*. 2009; 18:2–5. [PubMed: 19177345]
41. Zhou HF, Cross TA. Influences of membrane mimetic environments on membrane protein structures. *Annu Rev Biophys*. 2013; 42 now available on-line. **Proposed criteria for deciding when a membrane protein structure has been distorted by its crystallization environment
42. Nakahara T, Hashimoto R, Nakagawa H, Monde K, Miura N, Nishimura S. Glycoconjugate Data Bank: Structures – an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res*. 2008; 36:D368–D371. [PubMed: 17933765]
43. Lütke T. Analysis and validation of carbohydrate three-dimensional structures. *Acta Cryst*. 2009; D65:156–168. **Review of current problems and tools for carbohydrate validation
44. Foley BL, Tessier MB, Woods RJ. Carbohydrate force fields. *WIREs Comput Mol Sci*. 2012; 2(4)10.1002/wcms89 *Review of carbohydrate force fields, and new one, for refinement and validation
45. Jo S, Im W. Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Res*. 2013; 41:D470–D474. [PubMed: 23104379]
46. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE, Orpen AG. Retrieval of crystallographically-derived molecular geometry information. *J Chem Inf Comput Sci*. 2004; 44:2133–2144. [PubMed: 15554684]
47. Zheng H, Chruszcz M, Lasota P, Lebioda L, Minor W. Data mining of metal ion environments present in protein structures. *J Inorg Biochem*. 2008; 102:1765–1776. [PubMed: 18614239]
48. Headd JJ, Richardson JS. Fitting Tip#5: What's with water? *Comput Crystallogr Newsletter*. 2013; 4(1):2–5.
49. van den Akker F, Hol WGJ. Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structure. *Acta Cryst*. 1999; D55:206–218.
50. Lang PT, Ng HL, Fraser JS, Corn JE, Echols N, Sales M, Holton JM, Alber T. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Sci*. 2010; 19:120–1431. *Ringer program for finding new alternates around an atom by analysis of electron density just above noise level
51. Sobolev OV, Lunin VY. Detection of alternative conformations by unrestrained refinement. *Acta Crystallogr*. 2012; D68:1118–1127. *Automated method for locating needed alternates at high resolution, from shifts after unrestrained refinement trials
52. Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL Jr. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput Biol*. 2010; 6(4):e1000763. [PubMed: 20442867]
53. Lundgren M, Niemi AJ. Correlation between protein secondary structure, backbone angles, and sidechain orientations. *Physical Review*. 2012; E86:021904. [PubMed: 23005782]
54. Richardson JS., Keedy DA., Richardson DC. The Plot thickens – more data, more dimensions, more uses. In: Bansal M., Srinivasan N., editors. *Biomolecular Forms and Functions: A*

Celebration of 50 Years of the Ramachandran Map. World Scientific Publishing; Singapore: p. 13-29.

55. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr.* 2010; D66:486–501.
56. Davis IW, Arendall WB III, Richardson JS, Richardson DC. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure.* 2006; 14:265–274. [PubMed: 16472746]
57. Joosten RP, Joosten K, Murshudov GN, Perrakis A. PDB_REDO: constructive validation, more than just looking for errors. *Acta Cryst.* 2012; D68:484–496. **A second version of the all-PDB crystal structure re-refinement project, now with specific conformational corrections
58. Adams PD, Baker D, Brunger AT, Das R, DiMaio F, Read RJ, Richardson DC, Richardson JS, Terwilliger TC. Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. *Annu Rev Biophys.* 2013; 42:11.1–11.23. on-line 15 Feb 2013. *Describes experience with the new Rosetta protein correction and Erraser RNA correction software at low resolution
59. Chou FC, Sripakdeevong P, Dibrov SM, Hermann T, Das R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods.* 2013; 10:74–76. [PubMed: 23202432] **Ground-breakingly effective Erraser correction method for difficult RNA backbone fitting. Combines local exhaustive sampling of all backbone dihedrals (plus syn-vs-anti bases) with MolProbity diagnosis and Phenix and Rosetta refinements
60. Dunkle JA, Wang L, Feldman MB, Pulk A, Chen VB, Kapral GJ, Noeske J, Richardson JS, Blanchard SC, Cate JHD. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science.* 2011; 332:981–984. in supplemental on-line material. [PubMed: 21596992]
61. Schroeder GF, Levitt M, Brunger AT. Super-resolution biomolecular crystallography with low-resolution data. *Nature.* 2010; 464:1218–1222. [PubMed: 20376006]
62. Terwilliger TC, Read RJ, Adams PD, Brunger AT, Afonine PV, Grosse-Kunstleve RW, Hung LW. Improved crystallographic models through iterated local density-guided model deformation and reciprocal space refinement. *Acta Crystallogr.* 2012; D68:861–870.
63. Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, Sharff A, Vornrhein C, Bricogne G. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in Buster. *Acta Crystallogr.* 2012; D68:368–80.
64. van den Bedem H, Dhanik A, Latombe JC, Deacon AM. Modeling discrete heterogeneity in x-ray diffraction data by fitting multi-conformers. *Acta Crystallogr.* 2009; D65:1107–1117. *qFit software for modeling short regions of multiple conformation
65. Gipson B, Hsu D, Kavraki LE, Latombe JC. Computational models of protein kinematics and dynamics: beyond simulation. *Annu Rev Anal Chem.* 2012; 5:273–291.
66. Depristo MA, de Bakker PIW, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by protein crystallography. *Structure.* 2004; 12:831–838. [PubMed: 15130475]
67. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Adams PD, Moriarty NW, Zwart P, Read RJ, Turk D, Hung LW. Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. *Acta Cryst.* 2007; D63:597–610.
68. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips GN Jr. Ensemble refinement of protein crystal structures: Validation and application. *Structure.* 2007; 15:1040–1052. [PubMed: 17850744]
69. Gros P, van Gunsteren WF, Hol WG. Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. *Science.* 1990; 249:1149–1152. [PubMed: 2396108]
70. Schwander P, Fung R, Phillips GN Jr, Ourmazd A. Mapping the Conformations of Biological Assemblies. *New J Phys.* 2010; 12:035007–15. *New algorithm for producing single-conformation or ensemble structures from data at many single-shot orientations

Highlights

- Robust evaluation: distinguishing errors from valid anomalies
- New validation at the PDB
- Updated infrastructure: reference data, parameters, and methods
- Diagnostics for more structure components
- Automated correction for protein sidechains and RNA backbone

Figure 1. Outliers: valid anomalies versus errors. a) A valid rotamer outlier with well-fit electron density and 3 H-bonds (PDB: 3KYV). b) An erroneous rotamer outlier with spectacularly ill-fit density and serious steric clashes. (PDB: 1QW9)

Figure 2.

“Slider-bar” representation of percentile scores at the wwPDB. Mockup of percentiles vs all-PDB xray entries (bars) and vs similar-resolution cohort (ellipses), for a very good structure at 2.7 Å. RSR-Z has only one symbol, since that score is inherently normalized by resolution. (PDB: 1XK8)

Figure 3.

Automated correction of RNA backbone with Erraser. a) Tertiary contact of A9 and G63 at a Mg site, with pucker outliers, clashes, bond angle outliers, and unrecognized backbone conformers (!!). b) After an especially successful Erraser run, with all outliers corrected and better ion geometry. (PDB: 2GIS)