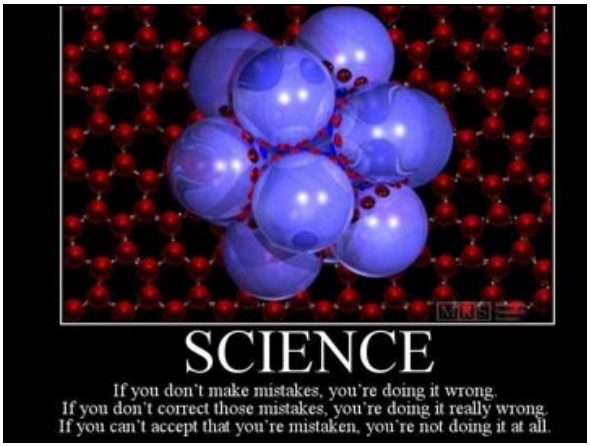



Winter School on Structural Biology, CEITEC, Brno, 13 February 2015

# Applied common sense

## The why, what and how of validation

Gerard J. Kleywegt
Protein Data Bank in Europe (pdbe.org)
EMBL-EBI, Cambridge, UK

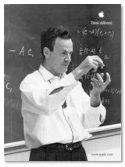
PDBe EMBL-EBI

## What *is* validation?

PDBe EMBL-EBI

## Validation according to the dictionary

- Validation = establishing or checking the truth or accuracy of (something)
  - Theory
  - Hypothesis
  - Model
  - Assertion, claim, statement
- Integral part of scientific activity!

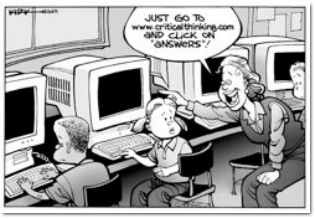


- "*Science is a way of trying not to fool yourself. The first principle is that you must not fool yourself, and you are the easiest person to fool.*" (Richard Feynman)

PDBe EMBL-EBI

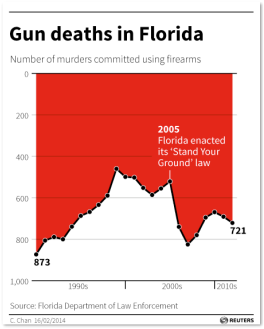
## Critical thinking

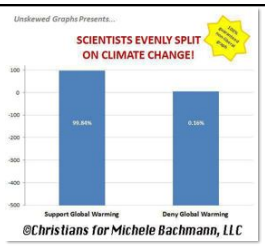
- Essential "24/7" skill for every scientist
  - And, in fact, for every non-scientist too
- Important aspect of validation

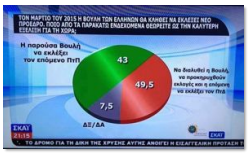



PDBe EMBL-EBI

## Critical thinking



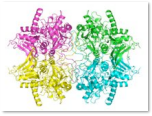




PDBe EMBL-EBI

## Critical thinking

- What is wrong here?
  - The tacR gene regulates the human nervous system
  - The tacQ gene is similar to tacR but is found in *E. coli*
  - ==> The tacQ gene regulates the nervous system in *E. coli*!



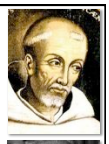
And here?

"*The tetramer has a total surface area of 81,616Å²*"

(Implies: +/- 0.5Å² …)



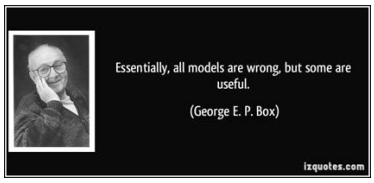
## Validation = critical assessment



- How good is my model, really?
- At the very least:
  - Does it explain all the data that I used?
  - Does it explain all the prior knowledge that I had?
- More importantly:
  - Does my model explain all the data that I didn't use?
  - Does my model explain all the prior knowledge that I didn't use?
  - Is my model the best possible, most parsimonious explanation for the data?
  - Are the testable predictions based on my model correct?
- If any of these questions is answered with "***no***", you have a problem!



## The *why* of validation






## Crystallography is great!!


- Crystallography can provide important biological insight and understanding!!



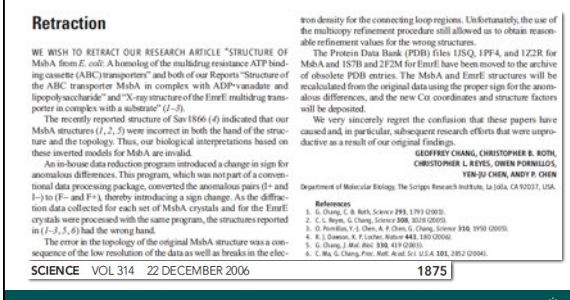
## Crystallography is great!!


- Crystallography can result in an all-expenses-paid trip to Stockholm (albeit in December)!!



## Nightmare before Christmas

… but sometimes we get it horribly wrong



**Retraction**

WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP·vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I−) to (F− and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

The error in the topology of the original MsbA structure was a consequence of the low resolution of the data as well as breaks in the electron density for the connecting loop regions. Unfortunately, the use of the multicopy refinement procedure still allowed us to obtain reasonable refinement values for the wrong structures.

The Protein Data Bank (PDB) files 1JSQ, 1PF4, and 1Z2R for MsbA and 1S7B and 2F2M for EmrE have been moved to the archive of obsolete PDB entries. The MsbA and EmrE structures will be recalculated from the original data using the proper sign for the anomalous differences, and the new Cα coordinates and structure factors will be deposited.

We very sincerely regret the confusion that these papers have caused and, in particular, subsequent research efforts that were unproductive as a result of our original findings.

GEOFFREY CHANG, CHRISTOPHER B. ROTH, CHRISTOPHER L. REYES, OWEN PORNILLOS, YEN-JU CHEN, ANDY P. CHEN

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

References

1. G. Chang, C. B. Roth, *Science* **293**, 1793 (2001).
2. C. L. Reyes, G. Chang, *Science* **308**, 1028 (2005).
3. O. Pornillos, Y.-J. Chen, A. P. Chen, G. Chang, *Science* **310**, 1950 (2005).
4. R. J. Dawson, K. P. Locher, *Nature* **443**, 180 (2006).
5. G. Chang, *J. Mol. Biol.* **330**, 419 (2003).
6. C. Wu, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).

SCIENCE VOL 314 22 DECEMBER 2006 1875

## Why do errors survive?

- "*Why do errors make it into the literature and the PDB?*"
- Suggestions from students
  - Cold Spring Harbor course, 2005
  - Copenhagen University course, 2006


PDBe Who/what do ***YOU*** think is to blame? EMBL-EBI

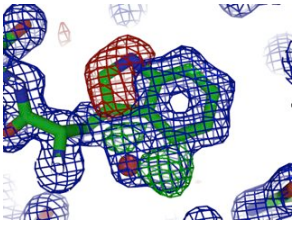
## Playing the Blame Game …

- Crystallographer
  - ignorance, lack of experience, incompetence, incorrect preconceptions/bias, cheating, laziness, "science by mouse-click", stress, can't be bothered to fix minor problems, no validation
- PI
  - pressure to publish/graduate fast, career interest, competition/scoops, grant writing, insufficient supervision
- Referees/editors
  - lazy, inadequate reviewing routines, no access to raw data, "validation by senior author name", lack of experience
- Software
  - misses or causes errors
- PDB
  - doesn't check
- "Nature"
  - limitations of the technique/resolution, errors hard to detect, poor data

PDBe EMBL-EBI

## Why do crystallographers make mistakes?

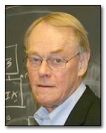
- Limitations to the data
  - Incomplete
  - Weak
  - Limited resolution
  - Space and time averaged
  - Phase errors
- The human factor
  - Subjectivity and bias involved in map interpretation and refinement (even at atomic resolution!)
  - Inexperienced people do the work, use of black boxes, …
  - Not everybody is a good chemist
  - Even experienced people make mistakes



PDBe Kleywegt, *Acta Cryst.* **D65**, 134 (2009) EMBL-EBI

## Crystallographer = Super(wo)man?

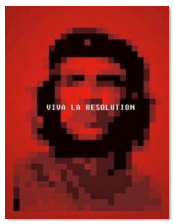
- The crystallographer ideally has
  - Knowledge of the history of the sample
  - Knowledge of the biology of the system
  - Knowledge of chemistry
  - Knowledge of physics
  - Understanding of data collection and processing
  - Understanding of the refinement process and software
  - Experience in map interpretation (preferably with a range of resolutions, space groups, *etc.*)
  - Read and remembered all the relevant literature
  - …

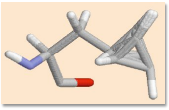


PDBe (Wayne Hendrickson) EMBL-EBI

## The odds are stacked against us

- Crystallographers produce models of structures that will contain errors
  - High resolution AND skilled crystallographer ➔ probably nothing major
  - High resolution XOR skilled crystallographer ➔ possibly nothing major
  - NOT (High resolution OR skilled crystallographer) ➔ pray for nothing major



PDBe *"I know the human being and fish can coexist peacefully"* EMBL-EBI

## A little experiment


- Hypothesis: "*If a card has a vowel on one side, then it has an even number on the other side*"
- Validate this hypothesis by turning as *few* cards as possible
- How many, and which, cards must you turn?

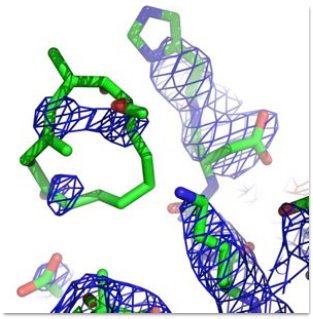
PDBe Wason selection task EMBL-EBI

## Confirmation bias

- A scientific model is a hypothesis to be shot down
- We should be looking for **disconfirming** evidence
- But we often don't! We tend to look for **supporting** evidence
  - Reasonable expectation to find a ligand + Any old density blob in a reasonable ligand-binding site => Model the ligand!
  - Even if it isn't really there…
  - Conversely: we don't expect a ligand, so we model waters

PDBe EMBL-EBI

## *"Believing is seeing…"*



PDBe Retracted "ligand complex" published in *Nature* EMBL-EBI

"A philosopher is a blind man in a dark room looking for a black cat that isn't there"

"A crystallographer is the man who finds it"

PDBe Paraphrasing HL Mencken EMBL-EBI

## Xtallography ≠ exact science

- Crystallographic models will contain errors
  - Crystallographers need to fix errors (if possible)
  - Users need to be aware of potentially problematic aspects of the model
  - Note: every crystallographer is also a user!
- Validation is important
  - Is the model as a whole reliable?
  - How about the bits that are of particular interest?
    - Active-site residues
    - Interface residues
    - Ligand, inhibitor, co-factor, …

PDBe EMBL-EBI

## Why don't people admit to their errors easily?

- To err is human
  - But so is denying that you erred
  - In some cases, "retraction battles" have raged for years
- Cognitive dissonance - discomfort caused by conflicting views of self
  - "*I am an intelligent, hard-working scientist who makes good decisions*"
  - "*There is an error in my structure*"


- How to resolve this discomfort?

PDBe EMBL-EBI

## Cognitive dissonance – ways of coping

- (1) Self-justification/denial/passing the buck
  - "There's nothing wrong with it"
  - "It doesn't change the conclusions"
  - "Everybody makes those kinds of errors"
  - "It's really a matter of interpretation"
  - "It's probably low occupancy/high mobility"
  - "There is strain in the active site"
  - "It fits other data/my chemical intuition"
  - "It was my student's first structure"
  - "Legacy software changed the signs of $\Delta F_{anom}$"
- (2) Depression – no need for that!
- (3) Acceptance/reconciliation – the grown-up thing to do
  - "I made an error, I'll fix it and learn from it"
  - Still an intelligent, hard-working scientist!
  - Doing yourself and science a favour

Proceedings of the CCP4 Study Weekend. Accuracy and Reliability of Macromolecular Crystal Structures (1990)

PDBe (David Eisenberg) EMBL-EBI

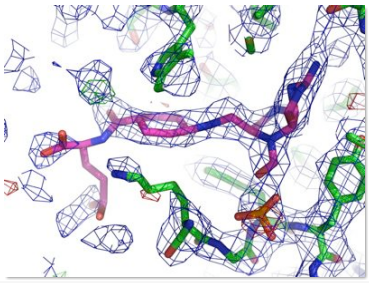
## Cognitive dissonance in action

THE LIGAND N5G IN THIS ENTRY IS N5-IMINIUM PHOSPHATE. HOWEVER, THERE IS SOME DISCREPANCY IN THE GEOMETRY. THE GEOMETRY FOR N5G IS SUGGESTED BY THE REFINEMENT. THE CO-ORDINATES FIT WELL IN THE ELECTRON DENSITY MAP. THE MAP WAS GENERATED USING A DATASET COLLECTED AT 2.8 ANGSTROM RESOLUTION. THE DENSITY FOR THE LIGAND IS UNAMBIGUOUS AND THEREFORE THE GEOMETRIES ARE CORRECT AND ARE AS THEY WOULD BE IN A BIOLOGICAL MOLECULE, WHERE THE MICRO ENVIRONMENT HAS A PROFOUND INFLUENCE ON THE GEOMETRIES OF THE LIGAND.

- Single N-C bonds of 1.1 and 1.6Å
- Non-bonded C…C contact of 2.0Å
- $PO_3$ moiety separated by 2.7Å from O



## The experimental "evidence"



*"Evidence that molecular-orbital theory breaks down in the presence of a protein crystallographer"* (K. Henrick)




| Mol | Type | Chain | Res | Link | Bond lengths | | | Bond angles | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Counts | RMSZ | #\|Z\| > 2 | Counts | RMSZ | #\|Z\| > 2 |
| 4 | N5G | A | 501 | - | 36,36,40 | 5.86 | 20 (55%) | 49,50,57 | 8.25 | 27 (55%) |
| 5 | PO4 | A | 502 | - | 1,3,4 | 4.54 | 1 (100%) | 0,3,6 | 0.00 | - |



## Errors and validation

- We need to take the drama out of the whole issue of errors and validation
- "*When a friend makes a mistake, the friend remains a friend and the mistake remains a mistake*" (S. Peres)
- Lao Tzu (more than 2500 years ago):

*A great nation is like a great man:*

*When he makes a mistake, he realises it*

*Having realised it, he admits it*

*Having admitted it, he corrects it*

*He considers those who point out his faults as his most benevolent teachers.*

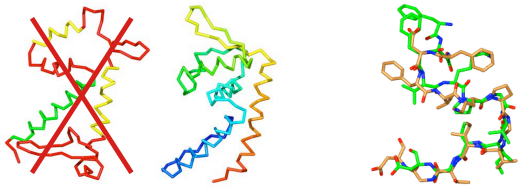


## What kinds of errors do crystallographers make?

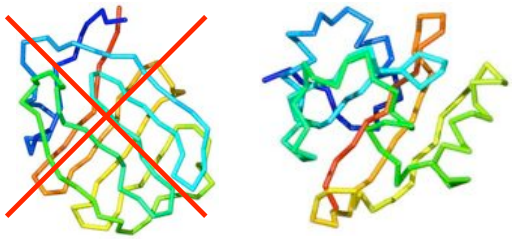


## Errors in protein structures

- Brändén & Jones (1990)
  - Mistracing an entire molecule or domain
  - Register errors
  - Local errors in the main chain



## Example of a tracing error

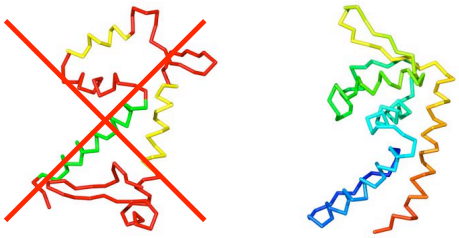


1PHY (1989, 2.4Å, *PNAS*)
Entire molecule traced incorrectly

2PHY (1995, 1.4Å)



## Example of a tracing error



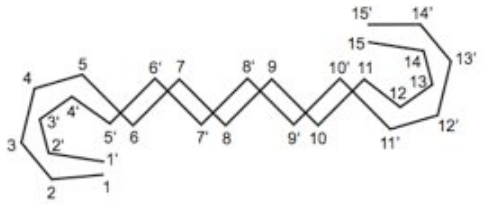
1FZN (2000, 2.55Å, *Nature*)
- One helix in register, two helices in place, rest wrong
- 1FZN obsolete, but complex with DNA still in PDB (1FZP)

2FRH (2006, 2.6Å)



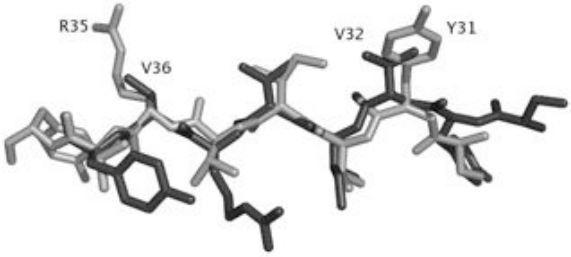
## What are register errors?

- For a segment of a model, the assigned sequence is out-of-register with the actual density






## Example of a register error




- 1CHR (light; 3.0Å, 1994, *Acta D*) *vs*. 2CHR (dark)

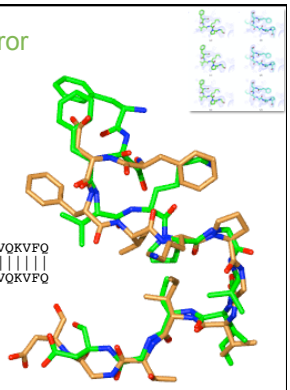


## Example of a register error

1ZEN (green carbons), 1996, 2.5Å, *Structure*

1B57 (gold carbons), 1999, 2.0Å

```
1B57 (A)      ---SKIFDFVKPGVITGDDVQKVFQ
.=ALIGN |=ID     .. .......... |||||||
1ZEN (_)      SKI-FD-FVKPGVITGD-DVQKVFQ
```



Confirmed by iterative build-omit maps
(Tom Terwilliger *et al.*, 2008)



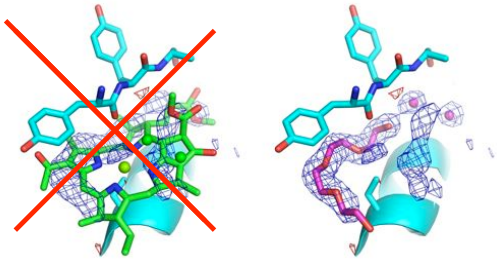
## Problems with ligands

## Reasonable assumptions?

- Typical assumptions
  - We know what the ligand is
  - The modelled ligand was really there
  - We didn't miss anything important
  - The observed conformation is reliable
  - At high resolution we get all the answers
  - The H-bonding network is known
  - We can trust the waters
  - We are good chemists
  - (The complex structure is relevant for drug design)

PDBe EMBL-EBI

## A case of mistaken identity…



3OEG – bacteriochlorophyll-*a* 3VDI – PEG fragment and waters

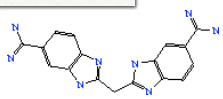
PDBe Tronrud & Allen, *Photosynth. Res.* **112**, 71 (2012) EMBL-EBI

## The ligand is really there?

**Structural Basis for BABIM Inhibition of Botulinum Neurotoxin Type B Protease** [*J. Am. Chem. Soc.* **2000**, *122*, 11268–11269]. Michael A. Hanson, Thorsten K. Oost, Chanokporn Sukonpan, Daniel H. Rich, Raymond C. Stevens*

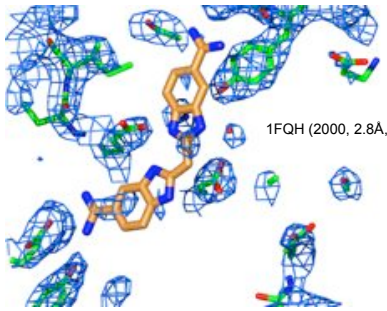
Page 11268. After a detailed analysis of the electron density maps for the structure of the inhibitor complex, we have concluded that the maps do not support the placement of the inhibitor as stated in the paper. Therefore, we are withdrawing the structural conclusions derived from PDB file 1FQH presented in the paper.

(J. Amer. Chem. Soc., August 2002)



PDBe EMBL-EBI

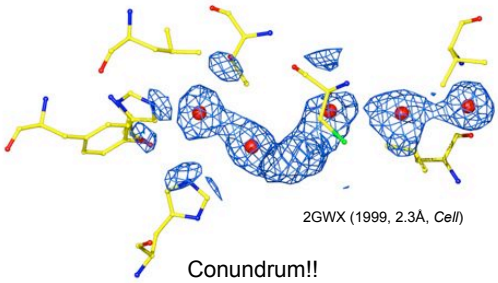
## Dude, where's my density?



1FQH (2000, 2.8Å, *JACS*)

PDBe EMBL-EBI

## We didn't miss anything?

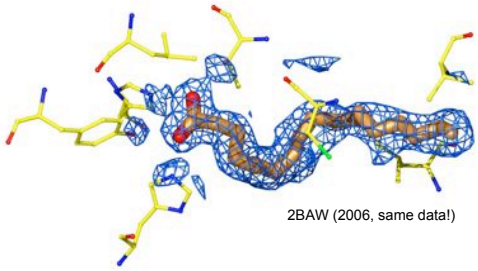


2GWX (1999, 2.3Å, *Cell*)

Conundrum!!

PDBe EMBL-EBI

## Oh, *that* ligand!



2BAW (2006, same data!)

PDBe EMBL-EBI

## Small-molecule anomalies

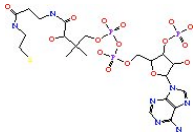
- 3-Phenylpropylamine in 1TNK, 1994, 1.8Å, *Nature Struct. Biol.*
- Aromatic carbon in between planar (0˚) and pyramidal (35˚) … 17˚

PDBe EMBL-EBI

## Oops-a-daisy!

- COA = coenzyme A
- 2.25Å, R 0.25/0.28, *Mol. Cell*
- Deposited 2003

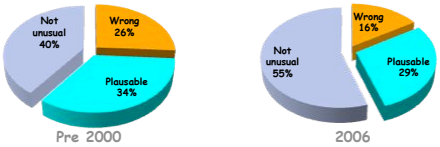
- Non-bonded contacts as close as 0.54Å
- Bond lengths up to 6.7Å
- Bond angles as low as 18˚
- Impropers of 160˚



PDBe EMBL-EBI

## Validation of PDB ligand structures by CCDC

- 16% of PDB entries deposited in 2006 had ligand geometries that were almost certainly in significant error *(in-house analysis using Relibase+/Mogul)*
- The good news - for structures before 2000 the figure was 26%




**(Jana Hennemann & John Liebeschuetz)**

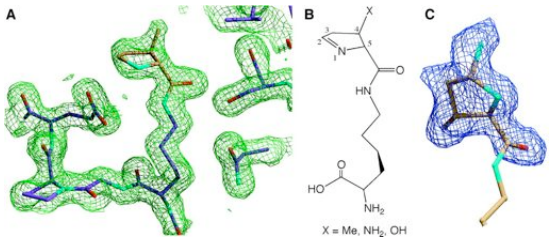
PDBe Liebeschuetz *et al.*, *J. Comput. Aid. Mol. Des.* **26**, 169 (2012) EMBL-EBI

## High resolution reveals all?

- Even at very high resolution there are sources of subjectivity and ambiguity
  - How to model temperature factors?
  - Is a blob of density a water or not?
  - How to model alternative conformations?
  - How to interpret density of unknown entities?
  - How to tell C/N/O apart?

PDBe EMBL-EBI

## The 22nd amino acid @ 1.55Å




Sodium chloride Ammonium sulfate

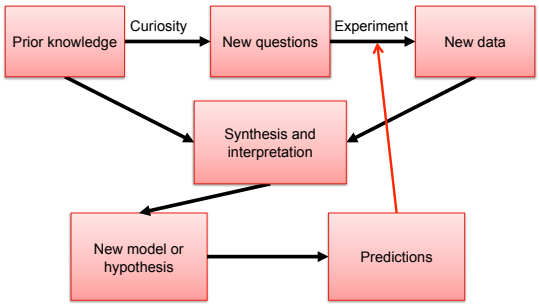
(Hao *et al.*, 2002; PDB entries 1L2Q and 1L2R)

PDBe EMBL-EBI

## The ***what*** of validation

PDBe EMBL-EBI

## How do we generate new knowledge?




PDBe EMBL-EBI

## Errors affect measurements

- Random errors (noise)
  - Affect precision
  - Usually normally distributed
  - Reduce by increasing nr of observations


- Systematic errors (bias)
  - Affect accuracy
  - Incomplete knowledge or inadequate design
  - Reproducible


- Gross errors (bloopers)
  - Incorrect assumptions, undetected mistakes or malfunctions
  - Sometimes detectable as outliers

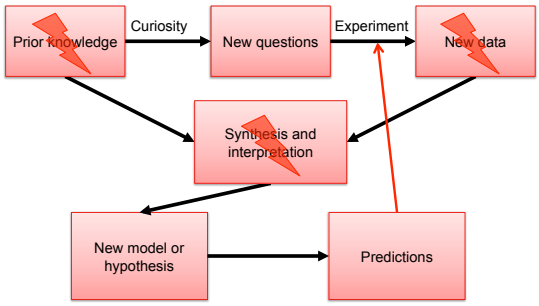
PDBe EMBL-EBI

## Errors affect measurements

- How tall is Gerard?
- 200 203 202 203 202 201 203 80
- Random error
- Systematic error
- Gross error

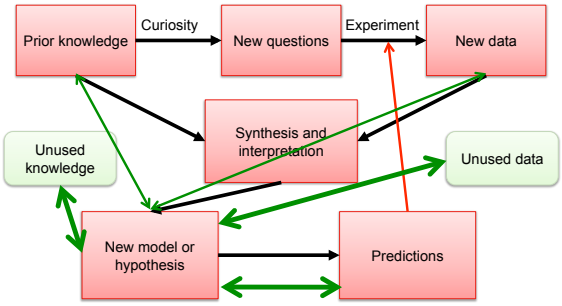
PDBe Anisotropic model of Gerard EMBL-EBI

## What can go wrong?




PDBe Sod's Law (a.k.a. Murphy's Law) EMBL-EBI

## Various kinds of validation




PDBe EMBL-EBI

## This model of hypothesis validation is entirely general for experimental sciences

How does it apply to protein crystallography?

PDBe EMBL-EBI

## The *how* of validation

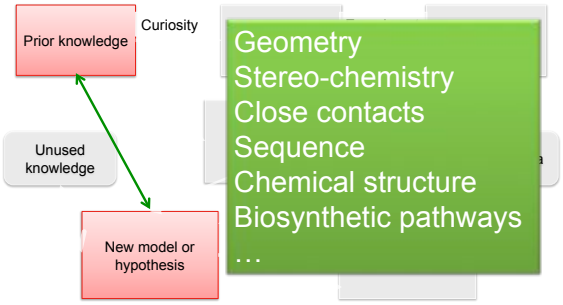
PDBe EMBL-EBI

## What is a good model?

- A good model makes SENSE in all respects!

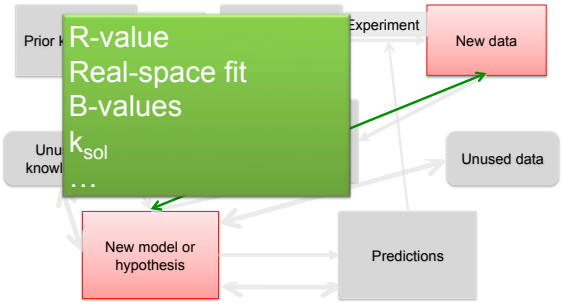
PDBe EMBL-EBI

## Various kinds of crystal structure validation



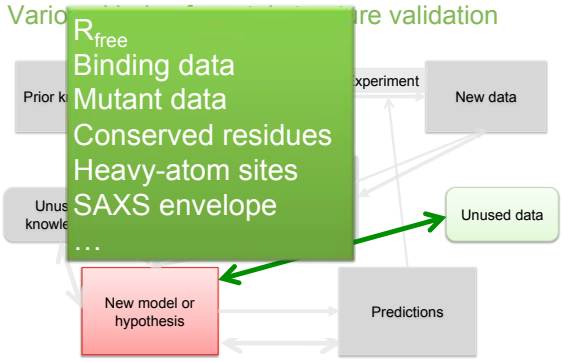

PDBe EMBL-EBI

## Various kinds of crystal structure validation



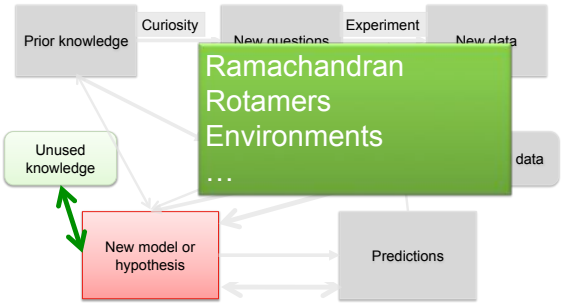

PDBe EMBL-EBI

## Various kinds of crystal structure validation



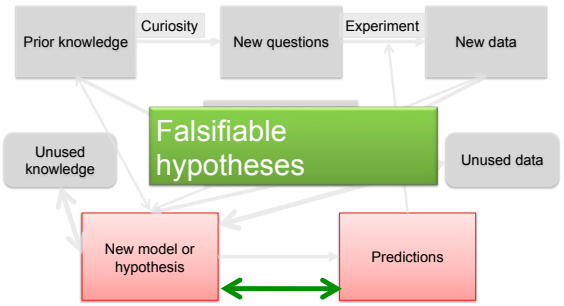

PDBe EMBL-EBI

## Various kinds of crystal structure validation




PDBe EMBL-EBI

## Various kinds of crystal structure validation




PDBe EMBL-EBI

## Validation in a nutshell

- Compare your model to the experimental data and to the prior knowledge. It should:
  - Reproduce knowledge/information/data used in the construction of the model
    - R, RMSD bond lengths, chirality, …
  - Predict knowledge/information/data not used in the construction of the model
    - $R_{free}$, Ramachandran plot, packing quality, …
  - Global and local
  - Model alone, data alone, fit of model and data
  - … and if your model fails to do this, there had better be a plausible explanation!

PDBe EMBL-EBI

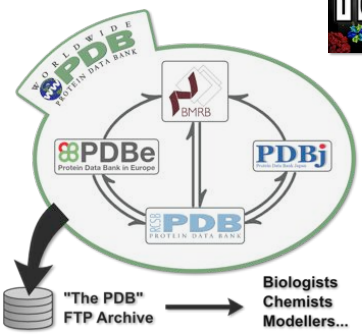
## What is "the PDB" doing about validation?


PDBe EMBL-EBI

## What is "the PDB"?


PDBe EMBL-EBI

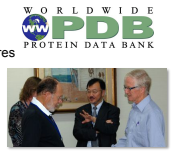
## wwPDB


PDBe wwpdb.org EMBL-EBI

## wwPDB partnership

- Collaborate on "*data in*"
  - Policy issues
  - Weekly releases
  - Validation standards
  - Format specifications
  - Chemical Component Dictionary
  - Deposition and annotation procedures
  - Archive quality and remediation
  - Journal interactions
  - Community interactions
- Friendly competition on "*data out*"
  - Serving PDB data with added-value
  - PDB-based services
  - Other services, resources and activities



PDBe wwpdb.org

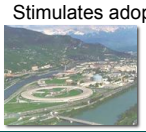
## Validation addresses important questions

- Entry-specific validation (quality control)
  - Is this model ready for archiving and publication?
  - Is this model a faithful, reliable and complete interpretation of the experimental data?
  - Are there any obvious errors/problems?
  - Are the conclusions drawn in the paper justified by the data?
  - Is this model suitable for my application?
- Archive-wide validation (comparative)
  - Is this model a better interpretation of the data?
  - What is the best model for this molecule/complex to answer my research question?
  - Which models should I select/omit when mining the PDB?

PDBe EMBL-EBI

## Validation by wwPDB - advantages

- Applies community-agreed methods uniformly
- Improves the quality and consistency of the PDB archive
- Supports editors and referees
- Helps users assess if an entry is suitable
- Helps users compare related entries
- Enables identification of outliers when mining the PDB
- Stimulates adoption of better protocols by the community

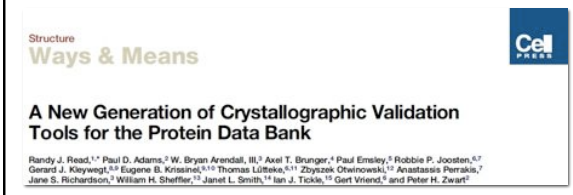


PDBe EMBL-EBI

## The future of validation

WORLDWIDE PDB PROTEIN DATA BANK

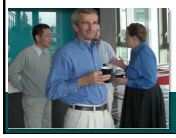
- wwPDB X-ray Validation Task Force


Structure
Ways & Means

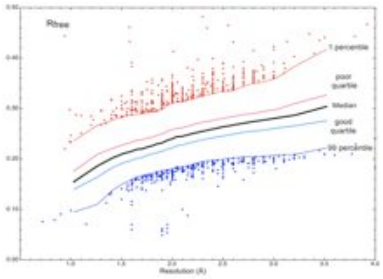
Cell PRESS

A New Generation of Crystallographic Validation Tools for the Protein Data Bank

Randy J. Read,[1,*] Paul D. Adams,[2] W. Bryan Arendall, III,[3] Axel T. Brunger,[4] Paul Emsley,[5] Robbie P. Joosten,[6,7] Gerard J. Kleywegt,[8,9] Eugene B. Krissinel,[9,10] Thomas Lütteke,[6,11] Zbyszek Otwinowski,[12] Anastassis Perrakis,[7] Jane S. Richardson,[3] William H. Sheffler,[13] Janet L. Smith,[14] Ian J. Tickle,[15] Gert Vriend,[9] and Peter H. Zwart[2]

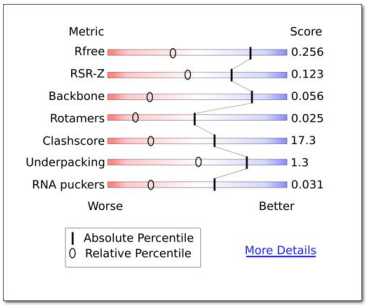


## Archive-wide analysis



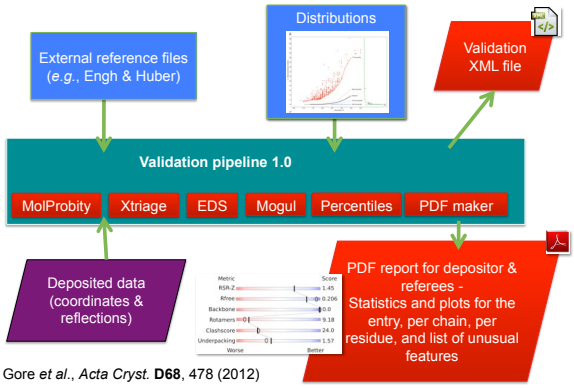

PDBe X-ray VTF: Read *et al.*, *Structure* **19**, 1395 (2011) EMBL-EBI

## Percentile scores




PDBe EMBL-EBI

## wwPDB X-ray validation pipeline




Gore *et al.*, *Acta Cryst.* **D68**, 478 (2012)

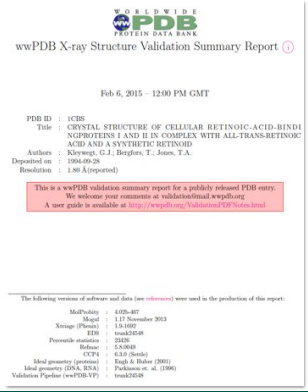
## What does it mean for a crystallographer?

- There are three uses of the validation pipeline
  - At deposition time
    - Not all checks can be run, e.g. some sequence and ligand checks
    - Report for depositor
  - At annotation time
    - Complete validation report, also suitable for editors/referees
  - Independently of deposition
    - Anonymous web-based server to use on models not (yet) in the PDB
    - Not all checks can be done
    - Will be developed once the production pipeline is up and running
    - Will not be available as a stand-alone software package

PDBe EMBL-EBI

## Validation reports

- Front cover
  - Deposition info
  - Software info

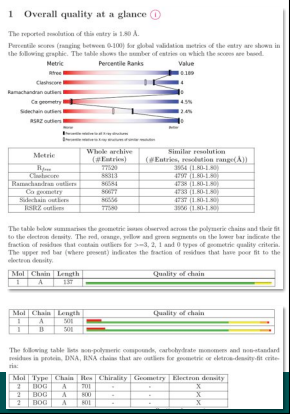



- wwpdb.org/validation-reports.html
- wwpdb.org/validation-servers.html

PDBe pdbe.org/valrep/1cbs EMBL-EBI

## Validation reports

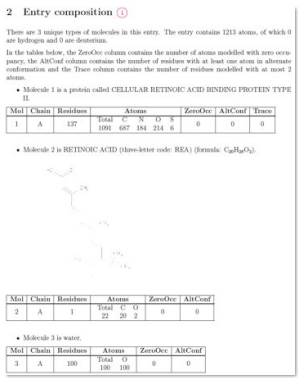
- Summary
  - Quality vs. all PDB X-ray
  - Quality vs. entries at similar resolution
  - Overview of residue-based quality for every polymer
  - Table of ligands that may need attention




PDBe

## Validation reports

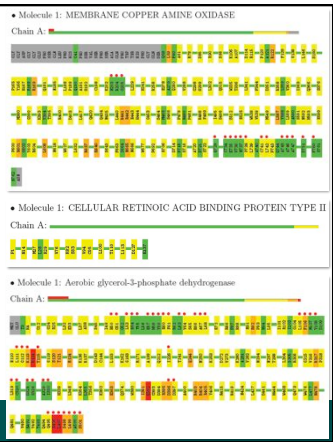
- Entry contents
  - Inventory




PDBe EMBL-EBI

## Validation reports

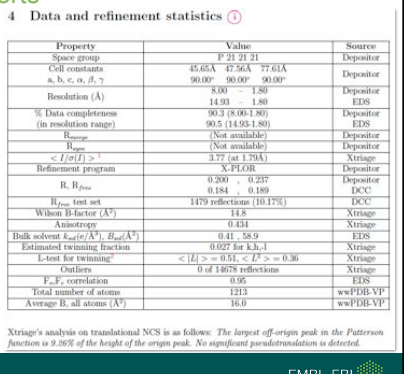
- Residue quality
  - One plot per polymer
  - Coloured by number of ***types*** of geometric outliers
  - Grey if not modelled
  - Red dots: poor density (RSR-Z > 2, as in EDS)




PDBe

## Validation reports

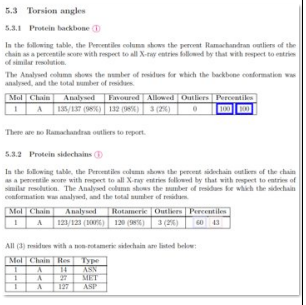
- "Table 1"
  - Xtriage




PDBe EMBL-EBI

## Validation reports

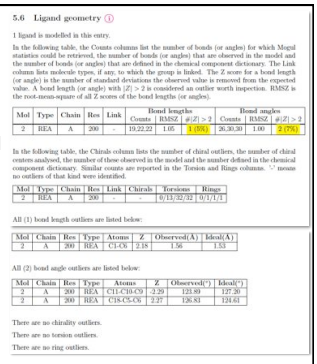
- Model quality
  - Bond lengths and angles
  - Torsion angles (Ramachandran, rotamers)
  - Clashes
  - Separately for standard residues, non-standard residues, ligands, carbohydrates
  - Generally: information about distribution, outlier stats, percentile scores, list of up to 5 (worst) outliers (full reports contain all outliers)



PDBe EMBL-EBI

## Validation reports

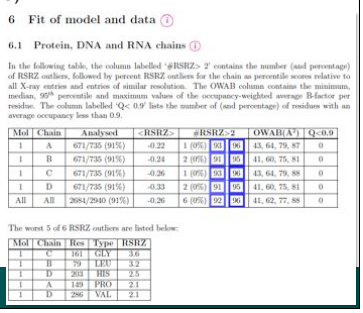
- Geometry validation of ligands and non-standard entities
  - Mogul (CCDC)
- wwPDB will get CSD coordinates for new and existing compounds (if they are available, of course)



PDBe EMBL-EBI

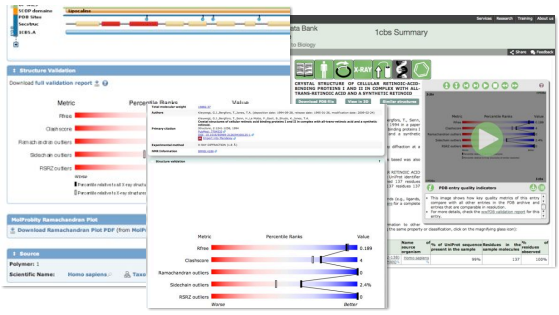
## Validation reports

- Model/data fit proteins, DNA, RNA
  - RSR and RSR-Z (EDS)
- Ligands *etc.*
  - RSR and LLDF



PDBe

## Public X-ray Validation Reports

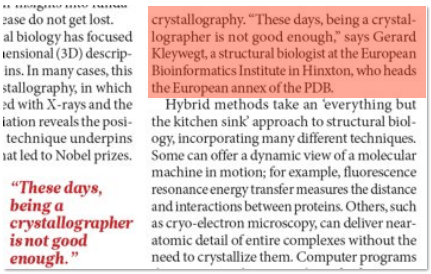


PDBe pdbe.org – rcsb.org – pdbj.org EMBL-EBI

## Beta site at PDBe

PDBe http://wwwdev.ebi.ac.uk/pdbe/entry/pdb/1cbs EMBL-EBI

## Other methods?



ease do not get lost.
al biology has focused
nensional (3D) descrip-
ins. In many cases, this
stallography, in which
ed with X-rays and the
iation reveals the posi-
technique underpins
hat led to Nobel prizes.

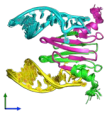
*"These days, being a crystallographer is not good enough."*

crystallography. "These days, being a crystallographer is not good enough," says Gerard Kleywegt, a structural biologist at the European Bioinformatics Institute in Hinxton, who heads the European annex of the PDB.

Hybrid methods take an 'everything but the kitchen sink' approach to structural biology, incorporating many different techniques. Some can offer a dynamic view of a molecular machine in motion; for example, fluorescence resonance energy transfer measures the distance and interactions between proteins. Others, such as cryo-electron microscopy, can deliver near-atomic detail of entire complexes without the need to crystallize them. Computer programs

PDBe *Nature* **514**, 416 (2014) EMBL-EBI

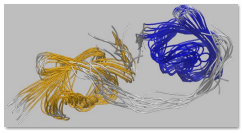
## Other Methods?



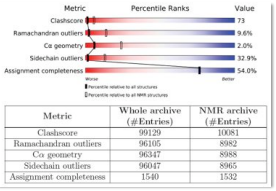
- Model validation using same criteria as X-ray
  - MolProbity, Mogul
- Some special model-related issues per technique
  - X-ray: alternative conformations
  - NMR: ensemble of models; well-defined regions
  - 3DEM: clashes of rigid-body fitted models; difference in species of model and sample sequence
- Data quality and model/data-fit assessment will be different for each technique

PDBe EMBL-EBI

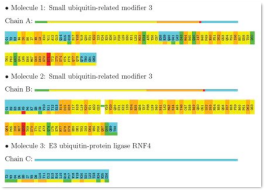
## NMR Validation



- NMR VTF recommendations published
- Global quality scores reported for "well-defined residues" only
  - As averages over the ensemble
  - Medioid model only

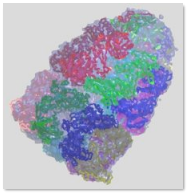



| Metric | Whole archive (#Entries) | NMR archive (#Entries) |
|---|---|---|
| Clashscore | 99129 | 10081 |
| Ramachandran outliers | 96105 | 8982 |
| Cα geometry | 96347 | 8988 |
| Sidechain outliers | 96047 | 8965 |
| Assignment completeness | 1540 | 1532 |




PDBe EMBL-EBI

## 3DEM Validation

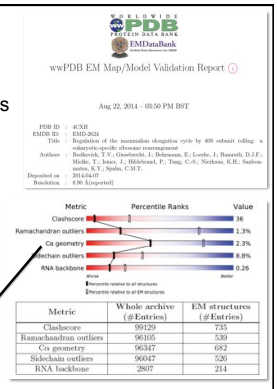


- Model validation
  - Clashes?
  - Taxonomy?
  - Homology models?
  - Non-atomistic models?
  - Cα-only models?
  - Rigid-body *vs.* flexible fitting *vs.* de novo modelling?
- Data and map validation
  - Per technique and resolution regime
  - Tilt-pair analysis; handedness; projections vs. raw data
- Map + model
  - Depending on resolution regime and model-building method?

PDBe EMBL-EBI

## EM Validation Reports

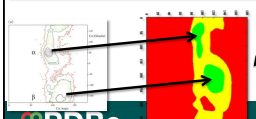
- Metrics relevant for EM models
- Define "Table 1" for EM




| Metric | Whole archive (#Entries) | EM structures (#Entries) |
|---|---|---|
| Clashscore | 99129 | 735 |
| Ramachandran outliers | 96105 | 539 |
| Cα geometry | 96347 | 682 |
| Sidechain outliers | 96047 | 526 |
| RNA backbone | 2807 | 214 |

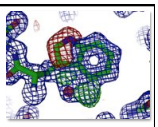
4 Experimental information

| Property | Value | Source |
|---|---|---|
| Reconstruction method | Not provided | Depositor |
| Imposed symmetry | 1 | Depositor |
| Number of images | 28993 | Depositor |
| Resolution determination method | FSC at 0.143 cut-off | Depositor |
| CTF correction method | Each particle | Depositor |
| Microscope | OTHER | Depositor |
| Voltage (kV) | 300 | Depositor |
| Electron dose ($e^-/Å^2$) | 25 | Depositor |
| Minimum defocus (nm) | 1000 | Depositor |
| Maximum defocus (nm) | 2500 | Depositor |
| Magnification | 59000 | Depositor |
| Image detector | Kodak SO 163 film | Depositor |

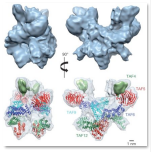


PDBe EMBL-EBI

## Validation by wwPDB



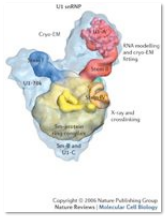
- By no means the end of the story!
  - Room for extension and improvement
    - Ligands, nucleic acids, carbohydrates, NCS, spacegroup errors, …
    - wwPDB ligand-validation workshop in 2015
  - X-ray
    - Re-convene X-ray VTF in 2015 to evaluate and update recommendations
  - NMR
    - Further development in progress
  - EM
    - Rudimentary at present, lots more work needed
  - All methods: annual re-compute of distributions
  - User feedback welcome at validation@mail.wwpdb.org



PDBe EMBL-EBI

## "Other other" methods

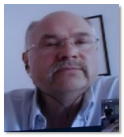

- SAS – wwPDB task force (2012, 2014)
- Hybrid methods – wwPDB task force (2014)
  - For example: solid-state NMR + EM + SAXS + solution NMR + homology modelling …
- Questions
  - What to archive and where?
  - What to accept?
  - What requirements for deposition?
  - How to validate?
  - What to do with non-atomistic models?
  - What to do with homology models?




PDBe EMBL-EBI

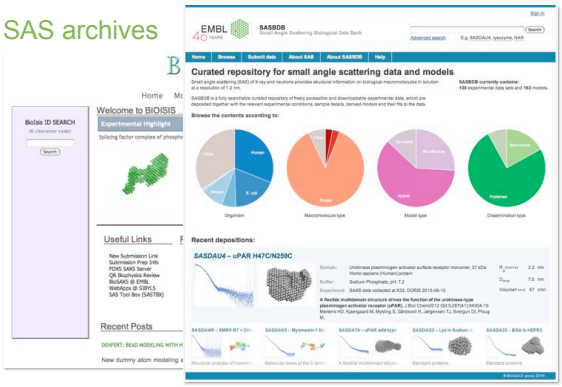
## SAS Task Force recommendations

- Need repository for SAXS and SANS data
- Need dictionary (data model) for SAXS and SANS
- Shape/bead and atomistic models should be archived (somewhere, somehow)
- Validation criteria need to be defined
- Archive of non-atomistic models from hybrid data
- What should (not) be in the PDB?




PDBe Trewhella *et al.*, *Structure* **21**, 875 (2013) EMBL-EBI

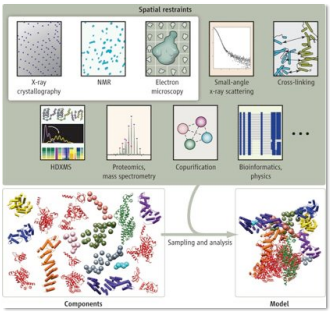
## SAS archives




PDBe bioisis.net – sasbdb.org EMBL-EBI

## Hybrid Methods

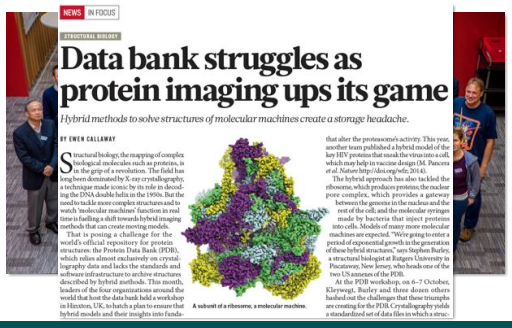
- Task Force met in October 2014
- Representatives of existing task forces, other methods, integrative modellers, and wwPDB
- Questions about what to archive where, what data and meta-data, how to validate




PDBe EMBL-EBI

## wwPDB Hybrid Methods Task Force

EMBL-EBI, Hinxton, 6-7 October, 2014



NEWS IN FOCUS

STRUCTURAL BIOLOGY

**Data bank struggles as protein imaging ups its game**

*Hybrid methods to solve structures of molecular machines create a storage headache.*

BY EWEN CALLAWAY

Structural biology, the mapping of complex biological molecules such as proteins, is in the grip of a revolution. The field has long been dominated by X-ray crystallography, a technique made iconic by its role in decoding the DNA double helix in the 1950s. But the need to tackle more complex structures and to watch 'molecular machines' function in real time is fuelling a shift towards hybrid imaging methods that can create moving models.

That is posing a challenge for the world's official repository for protein structures: the Protein Data Bank (PDB), which relies almost exclusively on crystallography data and lacks the standards and software infrastructure to archive structures described by hybrid methods. This month, leaders of the four organizations around the world that host the data bank held a workshop in Hinxton, UK, to hatch a plan to ensure that hybrid models and their insights into funda-

A subunit of a ribosome, a molecular machine.

that alter the proteasome's activity. This year, another team published a hybrid model of the key HIV proteins that sneak the virus into a cell, which may help in vaccine design (M. Pancera *et al. Nature* http://doi.org/nfz; 2014).

The hybrid approach has also tackled the ribosome, which produces proteins; the nuclear pore complex, which provides a gateway between the genome in the nucleus and the rest of the cell; and the molecular syringes made by bacteria that inject proteins into cells. Models of many more molecular machines are expected. "We're going to enter a period of exponential growth in the generation of these hybrid structures," says Stephen Burley, a structural biologist at Rutgers University in Piscataway, New Jersey, who heads one of the two US annexes of the PDB.

At the PDB workshop, on 6–7 October, Kleywegt, Burley and three dozen others hashed out the challenges that these triumphs are creating for the PDB. Crystallography yields a standardized set of data files in which a struc-

PDBe *Nature* **514**, 416 (2014) EMBL-EBI

## Key outcomes of discussion

- Be as inclusive as possible in collecting data from many different experimental methods
- Accommodate many types of structural representations
- Create a federated system to collect/curate data
- Use a common interface to collect data
- wwPDB should play a leadership role
- Whitepaper to describe vision

PDBe EMBL-EBI

## What have we learned?

PDBe EMBL-EBI

## Why do/did things sometimes go horribly wrong in X-ray?

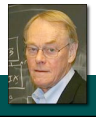
- Blind optimism/naïveté/ignorance
  - Belief in (wrong) numbers and in "magic" refinement programs
- Inappropriate (use of) modelling/refinement methods
  - Fitting too many parameters
- No/inappropriate quality control/validation
- "Believing is seeing"
- Large influx of non-experts

PDBe EMBL-EBI

## Of course, none of this should be news or surprising…

Hendrickson (CCP4 Proc., 1980) - "*That which is not restricted will take its liberties*"

Knight *et al.* (CCP4 Proc., 1990) - "*None of this evidence is dependent on a refined model and instead makes use of known facts about proteins in general and the S subunit of RuBisCO in particular*"




PDBe EMBL-EBI

## 1990

Protein crystallography is a highly competitive field where the same or similar structures are being worked on in a number of different laboratories. Although this may result in an urge to publish quickly and prematurely, it is the responsibility of crystallographers to check their preliminary models carefully before rushing into print. Journals must insist on the publication of enough data for crystallographers to convince the reader that they have a correct structure, and the readership should be sophisticated enough to judge the quality of the data. We strongly object to publication of structural work where authors supply a minimum amount of detail in the form of a cartoon. □

*Carl-Ivar Brändén and T. Alwyn Jones are in the Department of Molecular Biology, Uppsala Biomedical Center, PO Box 590, S-751 24 Uppsala, Sweden.*


PDBe Brändén & Jones, *Nature* **353**, 687 (1990) EMBL-EBI

## Lessons

- Have we learned anything from 25 years of errors?
  - Education is important
    - Avoid blind optimism, naïveté, belief in "magic" programs
    - Don't be afraid to ask a colleague's help or opinion
  - Use restraint and restraints when modelling
    - Consider the ratio of observations and parameters
    - Consider the information content of your data
  - ***Null-hypothesis: everything is normal!***
    - Trans-peptides, bond lengths/angles, rotamers, NCS, …
    - Unless your data shouts at you otherwise, or you have reliable prior knowledge

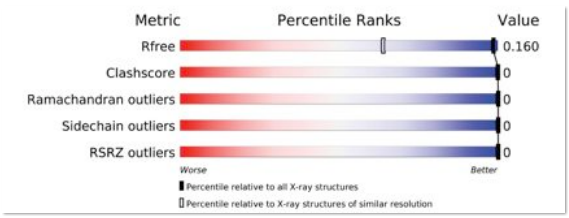
PDBe EMBL-EBI

## Lessons

- Have we learned anything from 25 years of errors?
  - Use (lots of) validation tools throughout, not just when you deposit
    - Or worse, rely on wwPDB annotators to tell you what's dodgy about your model…
  - Be your own fiercest critic!
    - Avoid confirmation bias - try to shoot down your own models and hypotheses
    - How will you deal with cognitive dissonance?

PDBe EMBL-EBI

## What you would like your plots to look like…




• Molecule 1: Potassium channel toxin ShK

Chain A:

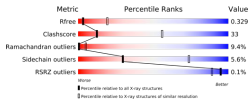
There are no outlier residues recorded for this chain.

PDBe pdbe.org/valrep/4lfq EMBL-EBI

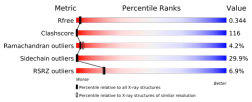
## Validation reports for today's structures

- New-style wwPDB X-ray validation reports are available for most of the structures shown or discussed in this lecture (even superseded ones) from http://www.ebi.ac.uk/~gerard/valrepcshl.html

- Examples:
- 1Z2R (part of the pentaretraction); 4.2Å

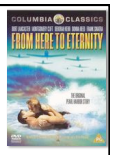



- 3LNA (imagined ligand); 2.7Å




PDBe EMBL-EBI

## Where to go from here?




- Download and read:
  - GJ Kleywegt. Validation of protein crystal structures. Acta Crystallographica D56, 249-265 (2000) (and many references therein)
  - GJ Kleywegt. On vital aid: the why, what and how of validation. Acta Crystallographica, D65, 134-139 (2009)

- Do this web-based tutorial:
  - http://xray.bmc.uu.se/embo2001/modval

PDBe EMBL-EBI

## Acknowledgements


- Alwyn Jones (Uppsala U)
- Randy Read (Cambridge U)
- Andy Davis (AstraZeneca)
- Members of the wwPDB and EMDataBank VTFs
- CCDC
- Colleagues
  - Uppsala, PDBe, wwPDB, EMDataBank, EBI, EMBL
- Everybody whom I have ever discussed validation and errors in protein structures with
- Many funding agencies in Sweden, UK, Europe and US as well as Uppsala University and EMBL


PDBe EMBL-EBI


Questions?

If you see this slide, I've gone too far

PDBe EMBL-EBI