



The three R's of a good structure,
Resolution, Refinement
and Reality

Eddie Snell

The structure from a crystal is the structure of the macromolecule?

Right?

Well yes, but

It's the average structure.

It's averaged over many individual macromolecules.

And it's averaged over time.



'Look, Watkins - I've invented a new prehistoric creature!'

Most importantly it's a **MODEL** that best explains the data

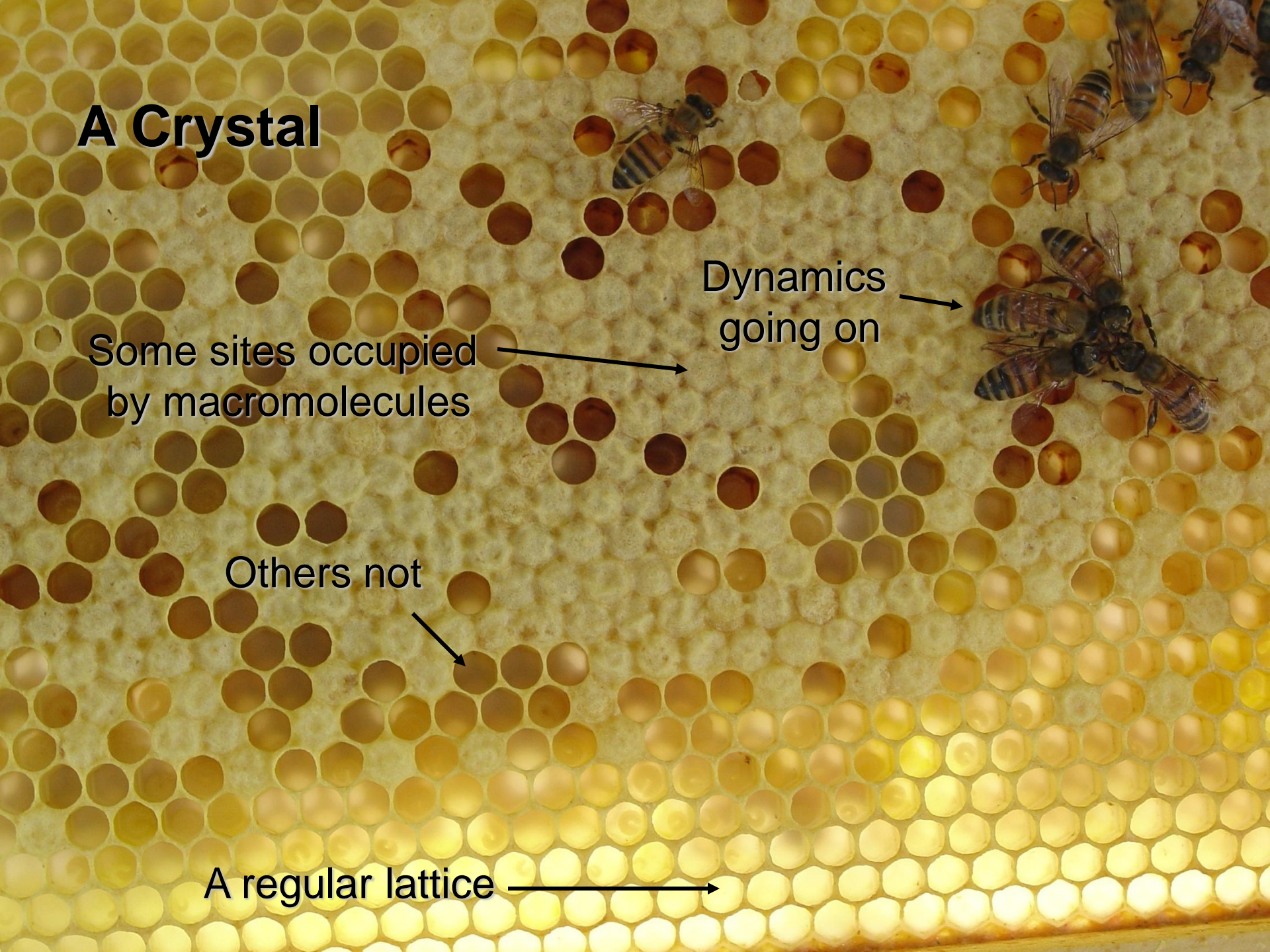
A Crystal

Some sites occupied
by macromolecules

Dynamics
going on

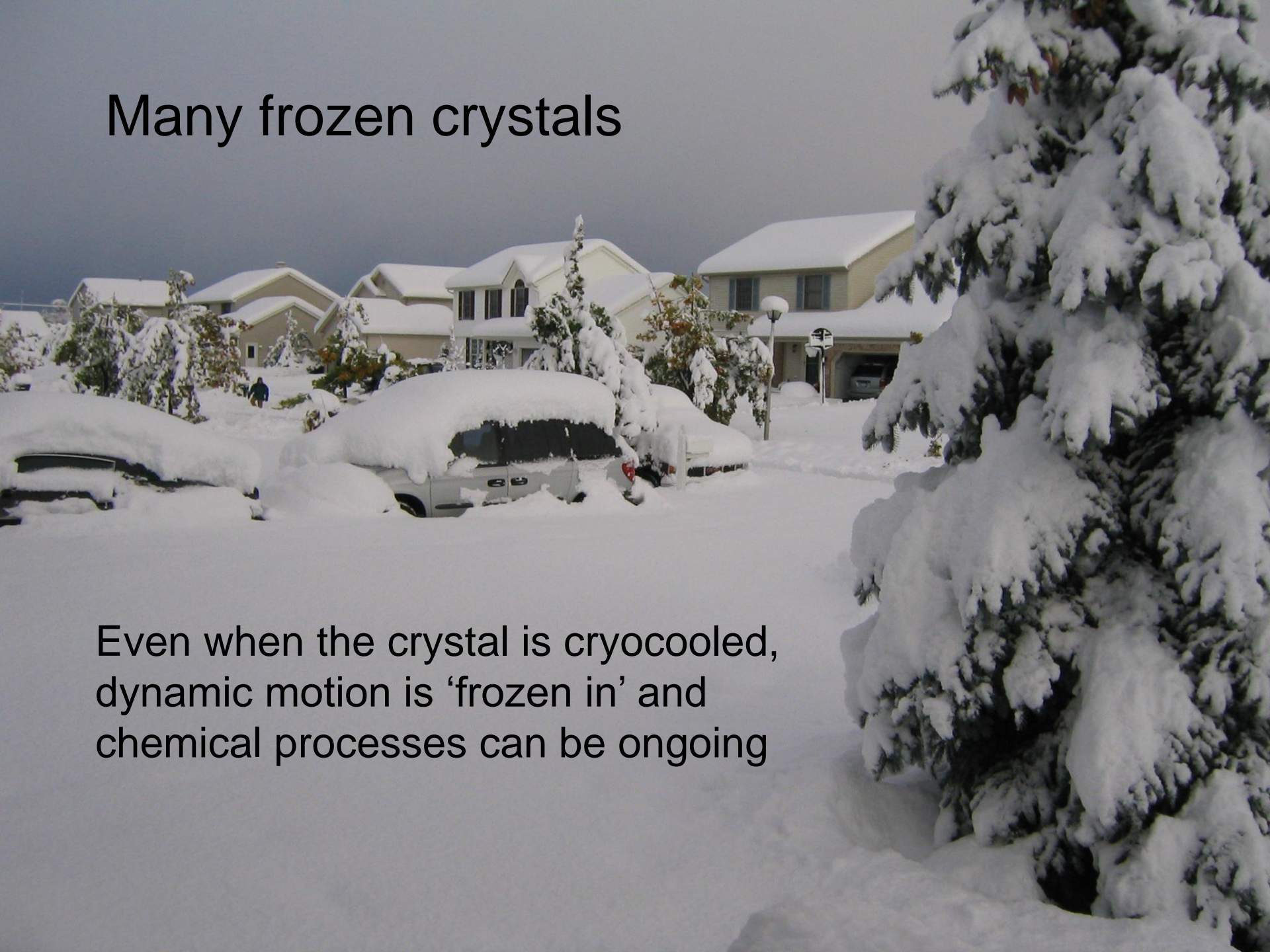
Others not

A regular lattice



Many frozen crystals

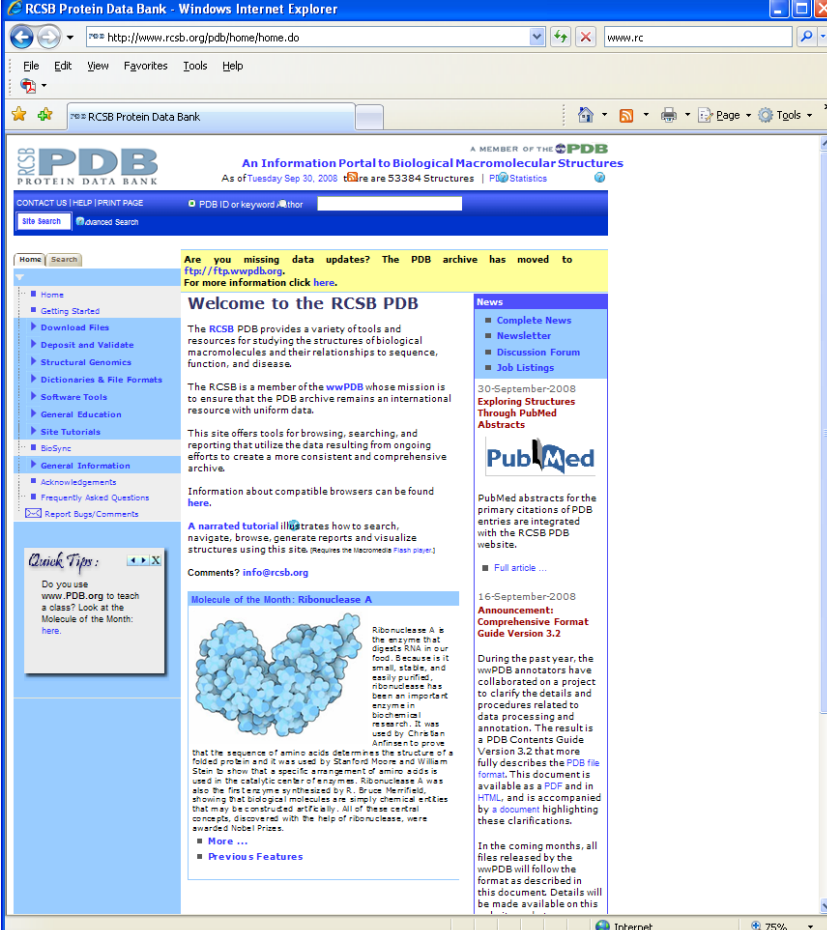
Even when the crystal is cryocooled, dynamic motion is 'frozen in' and chemical processes can be ongoing



How do we store structures?

Structures are deposited in the protein data bank or PDB (which also includes other biological macromolecules)

<http://www.rcsb.org>



The screenshot shows the RCSB Protein Data Bank homepage. The browser window title is "RCSB Protein Data Bank - Windows Internet Explorer". The address bar shows "http://www.rcsb.org/pdb/home/home.do". The page features a navigation menu on the left with options like Home, Getting Started, Download Files, Deposit and Validate, Structural Genomics, Dictionaries & File Formats, Software Tools, General Education, Site Tutorials, BioSync, General Information, Acknowledgements, Frequently Asked Questions, and Report Bugs/Comments. The main content area includes a search bar, a welcome message, a "Molecule of the Month" section for Ribonuclease A with a 3D molecular model, and a "News" section with announcements from September 2008. The footer shows the site is a member of the wwPDB and provides statistics as of Tuesday, Sep 30, 2008, with 53384 structures.

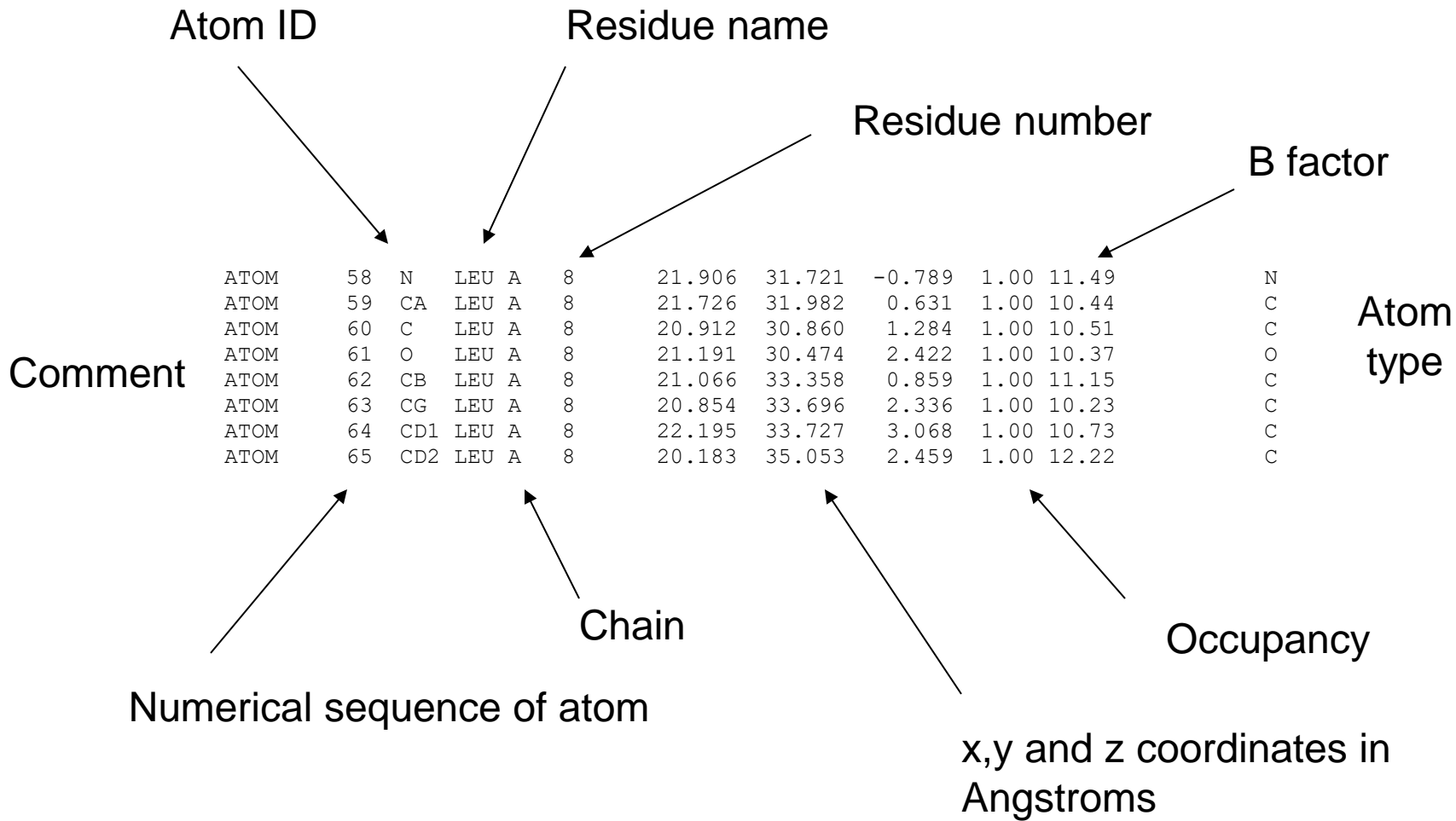
A tutorial is available at <http://www.rcsb.org/pdb/tutorials/tutorial.html>

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----

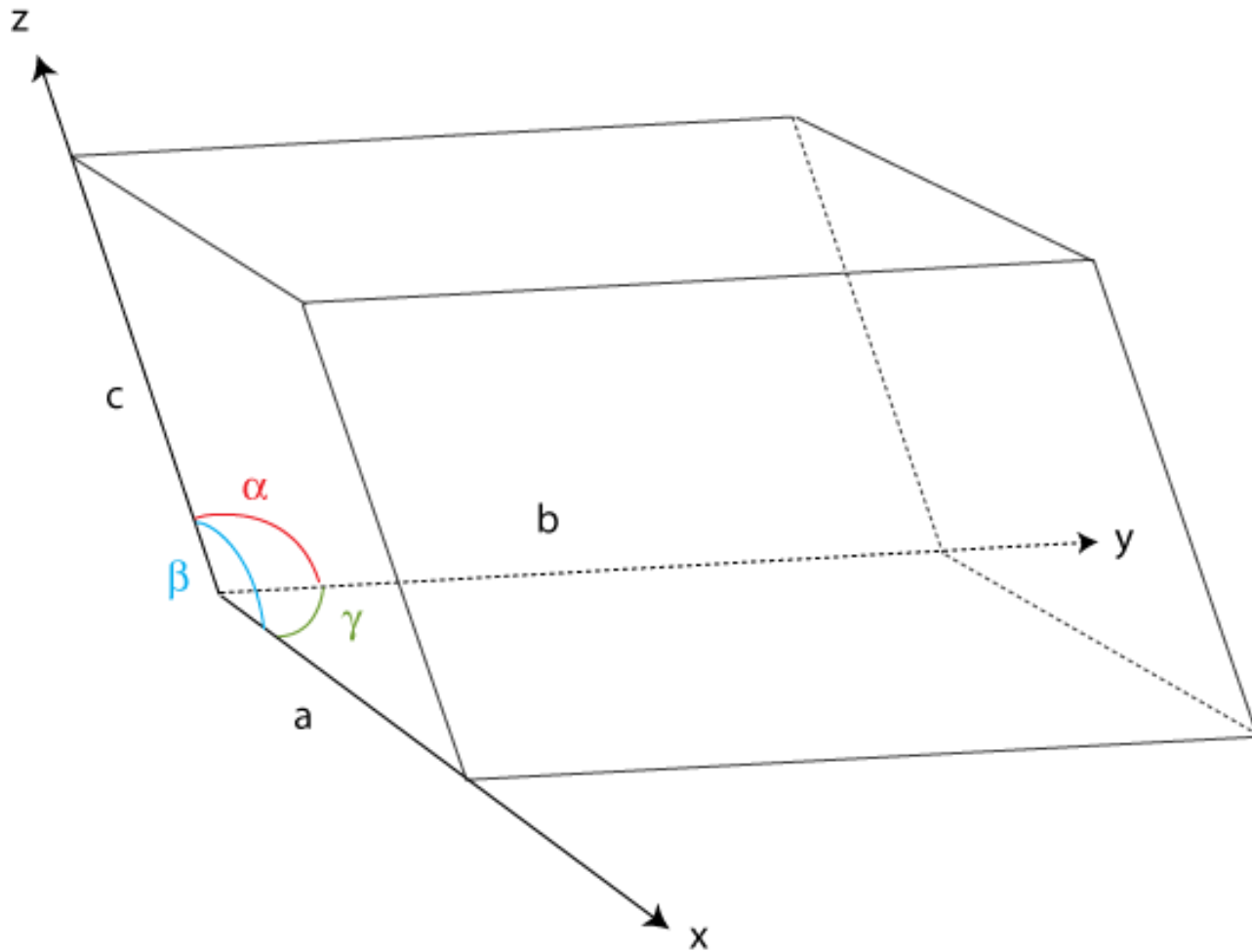
The PDB file

The PDB stores coordinates, experimental detail and comments and more recently the processed diffraction image data used to generate the coordinates.

```
CRYST1 77.453 77.453 37.183 90.00 90.00 90.00 P 43 21 2 8
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.012911 0.000000 0.000000 0.000000
SCALE2 0.000000 0.012911 0.000000 0.000000
SCALE3 0.000000 0.000000 0.026894 0.000000
ATOM 1 N LYS A 1 29.393 40.729 0.892 1.00 15.98 N
ATOM 2 CA LYS A 1 28.951 39.778 -0.166 1.00 15.98 C
ATOM 3 C LYS A 1 27.443 39.827 -0.378 1.00 15.87 C
ATOM 4 O LYS A 1 26.684 39.780 0.587 1.00 14.49 O
ATOM 5 CB LYS A 1 29.349 38.350 0.233 1.00 17.08 C
ATOM 6 CG LYS A 1 28.843 37.250 -0.704 1.00 20.49 C
ATOM 7 CD LYS A 1 29.418 35.906 -0.283 1.00 22.21 C
ATOM 8 CE LYS A 1 28.758 34.735 -0.990 1.00 22.43 C
ATOM 9 NZ LYS A 1 29.208 34.569 -2.386 1.00 25.22 N
ATOM 10 N VAL A 2 27.014 39.943 -1.635 1.00 15.27 N
ATOM 11 CA VAL A 2 25.584 39.924 -1.941 1.00 14.91 C
ATOM 12 C VAL A 2 25.281 38.513 -2.439 1.00 14.86 C
ATOM 13 O VAL A 2 25.698 38.133 -3.542 1.00 15.04 O
ATOM 14 CB VAL A 2 25.192 40.921 -3.057 1.00 15.23 C
ATOM 15 CG1 VAL A 2 23.695 40.820 -3.334 1.00 16.00 C
ATOM 16 CG2 VAL A 2 25.552 42.337 -2.640 1.00 15.40 C
ATOM 17 N PHE A 3 24.589 37.736 -1.603 1.00 14.21 N
ATOM 18 CA PHE A 3 24.203 36.353 -1.915 1.00 12.95 C
ATOM 19 C PHE A 3 23.063 36.267 -2.923 1.00 13.95 C
ATOM 20 O PHE A 3 22.212 37.154 -2.989 1.00 12.93 O
ATOM 21 CB PHE A 3 23.688 35.632 -0.659 1.00 13.15 C
ATOM 22 CG PHE A 3 24.750 34.999 0.195 1.00 11.18 C
ATOM 23 CD1 PHE A 3 25.551 35.764 1.046 1.00 12.22 C
ATOM 24 CD2 PHE A 3 24.905 33.620 0.193 1.00 10.60 C
ATOM 25 CE1 PHE A 3 26.480 35.161 1.885 1.00 12.38 C
ATOM 26 CE2 PHE A 3 25.831 33.000 1.026 1.00 11.09 C
ATOM 27 CZ PHE A 3 26.626 33.760 1.879 1.00 12.19 C
```



The Unit Cell



ATOM 58 N LEU A 8 21.906 31.721 -0.789 1.00 11.49 N

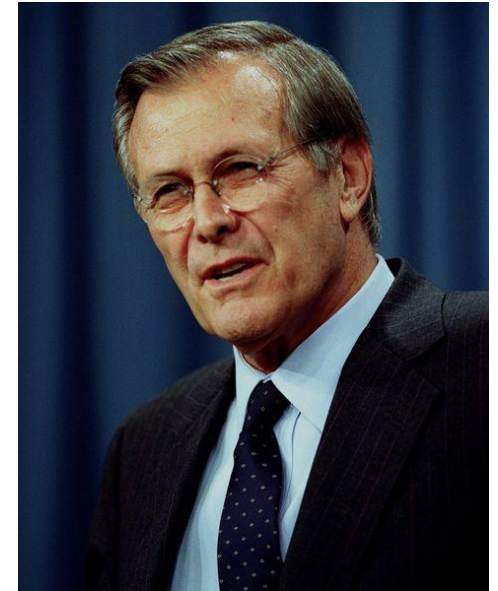
CRYST1	77.453	77.453	37.183	90.00	90.00	90.00	P	43	21	2	8
ORIGX1	1.000000	0.000000	0.000000			0.000000					
ORIGX2	0.000000	1.000000	0.000000			0.000000					
ORIGX3	0.000000	0.000000	1.000000			0.000000					
SCALE1	0.012911	0.000000	0.000000			0.000000					
SCALE2	0.000000	0.012911	0.000000			0.000000					
SCALE3	0.000000	0.000000	0.026894			0.000000					

Fractional coordinates in unit cell given by:

$$\begin{aligned}x_{\text{frac}} &= \text{Scale1}_1 \times X + \text{Scale1}_2 \times Y + \text{Scale1}_3 \times Z + U1 \\y_{\text{frac}} &= \text{Scale2}_1 \times X + \text{Scale2}_2 \times Y + \text{Scale2}_3 \times Z + U1 \\z_{\text{frac}} &= \text{Scale3}_1 \times X + \text{Scale3}_2 \times Y + \text{Scale3}_3 \times Z + U1\end{aligned}$$

The Essence of Structural Crystallography

- There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.
- Donald Rumsfeld, Feb 12th 2002.

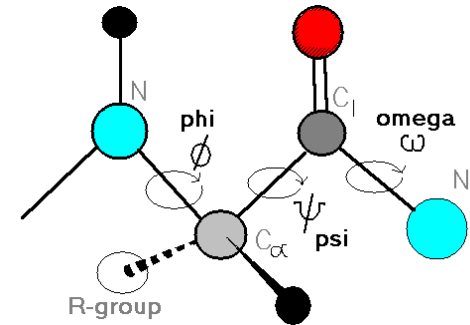
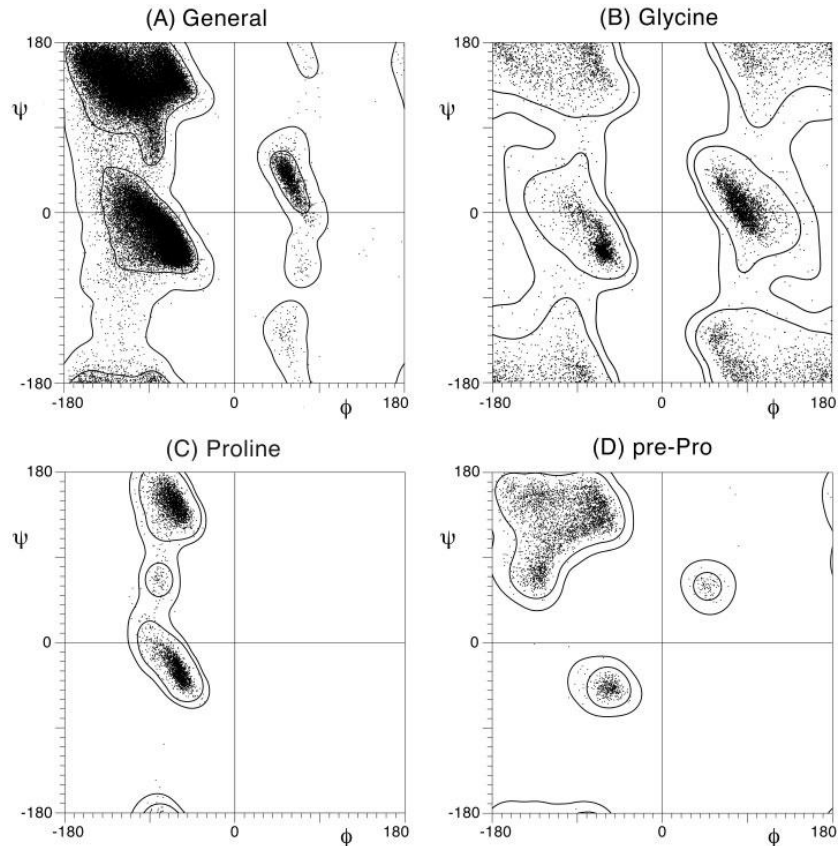


Shamelessly copied from a slide by Ted Baker

The three R's of a good structure

- A structure is a model that best represents the measured data.
- Think about what you are measuring:
 - The data is an average taken over many macromolecules. For example, a $100 \mu\text{m}^3$ crystal produced from a macromolecule that has a typical size of 200 \AA on edge will consist of $\sim 5,000$ molecules on edge or $125,000,000,000$ molecules in total.
 - The data is not static, it represents an average of those molecules over time.
 - The data is dynamic. X-rays cause chemical changes which can also be captured over time.
- Given these, how do we get a good structure.
 - **How do we know when we have a good structure?**

Known knowns – we know what to expect



Structure Validation by C_α Geometry: ϕ , ψ and C_β Deviation PROTEINS: Structure, Function, and Genetics 50:437–450 (2003)

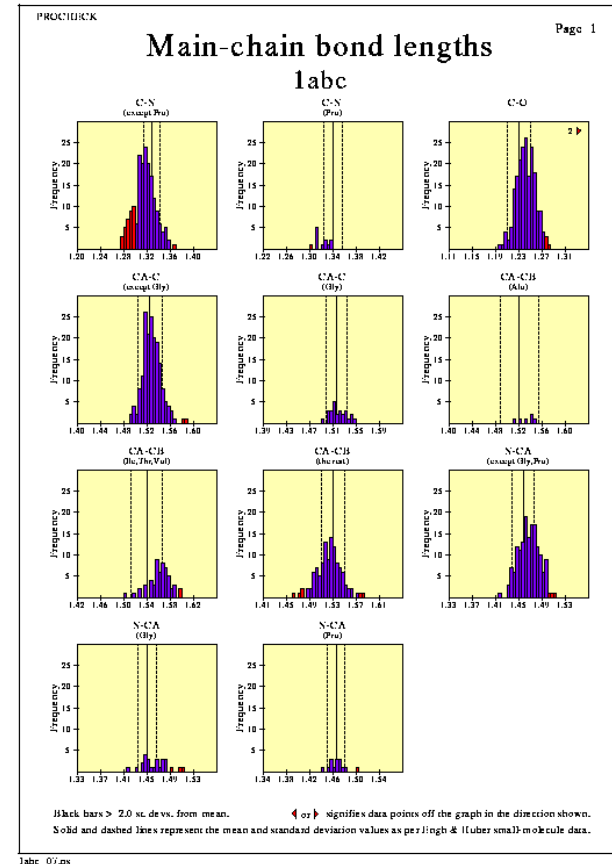
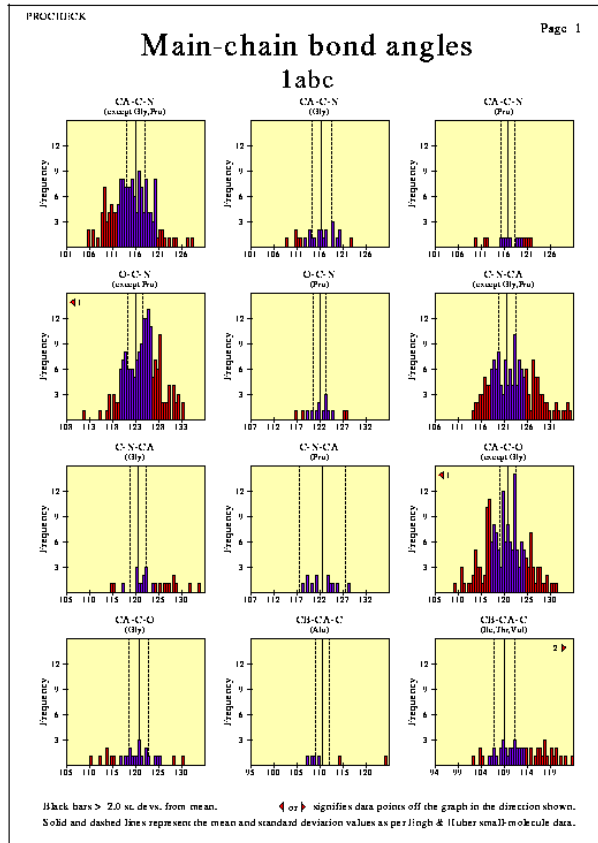
Simon C. Lovell, Ian W. Davis, W. Bryan Arendall III, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson

The dihedral angles in the main chain have allowed and disallowed regions that are well known – developed by Gopalasamudram Narayana Ramachandran and called the Ramachandren plot. Available as part of several software packages.

Known knowns – we know what to expect

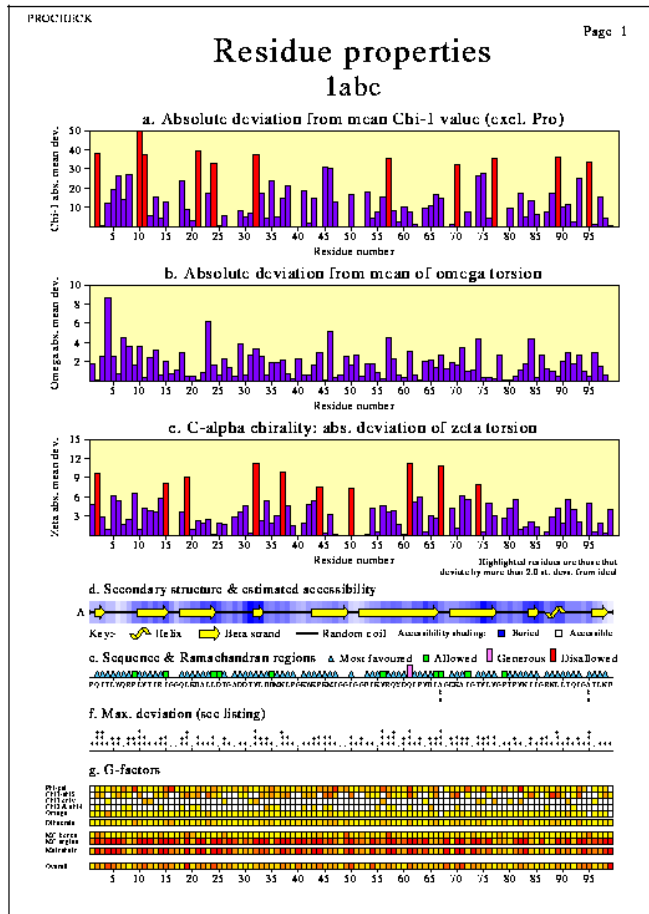


Known knowns – we know what to expect



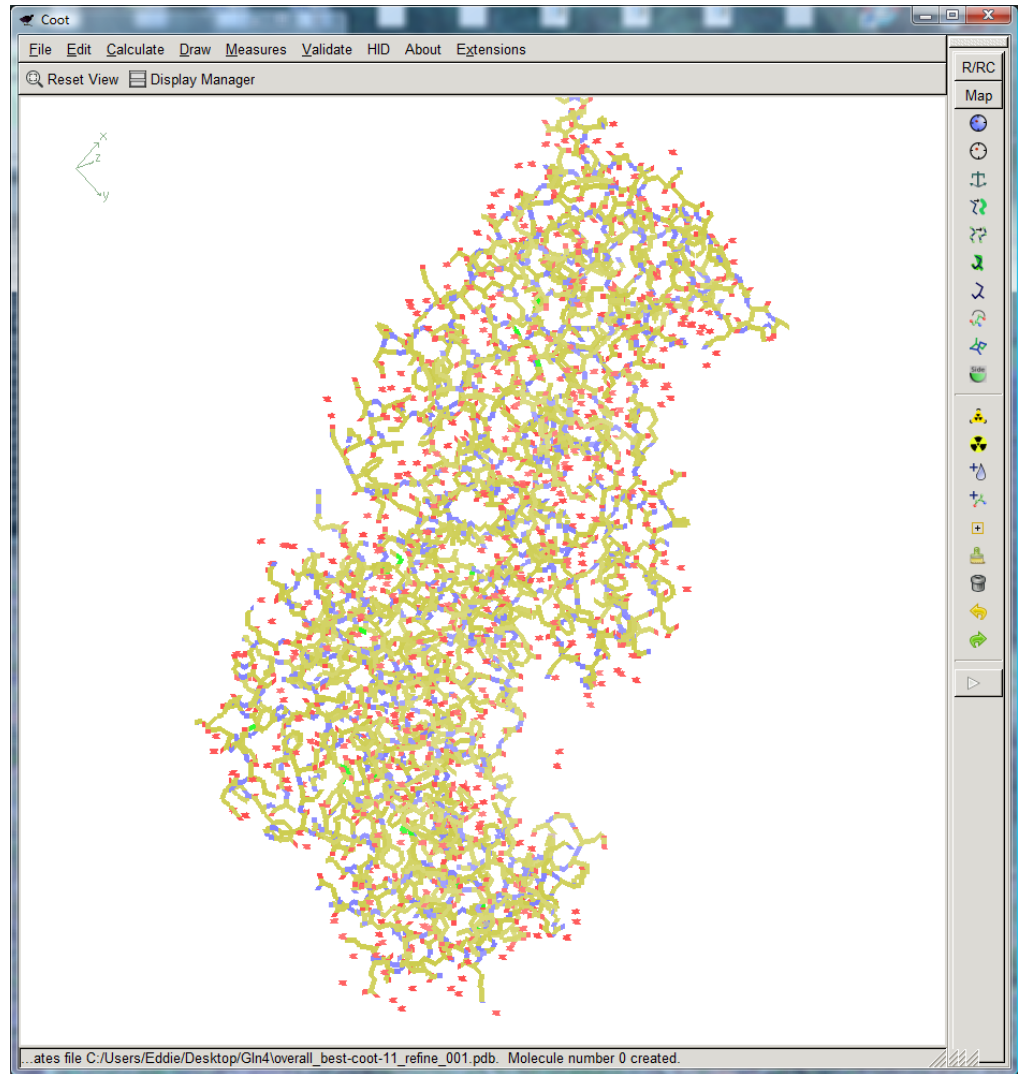
We can look for deviations in bond chain angles and lengths from previous data, e.g. Procheck <http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>

Known knowns – we know what to expect



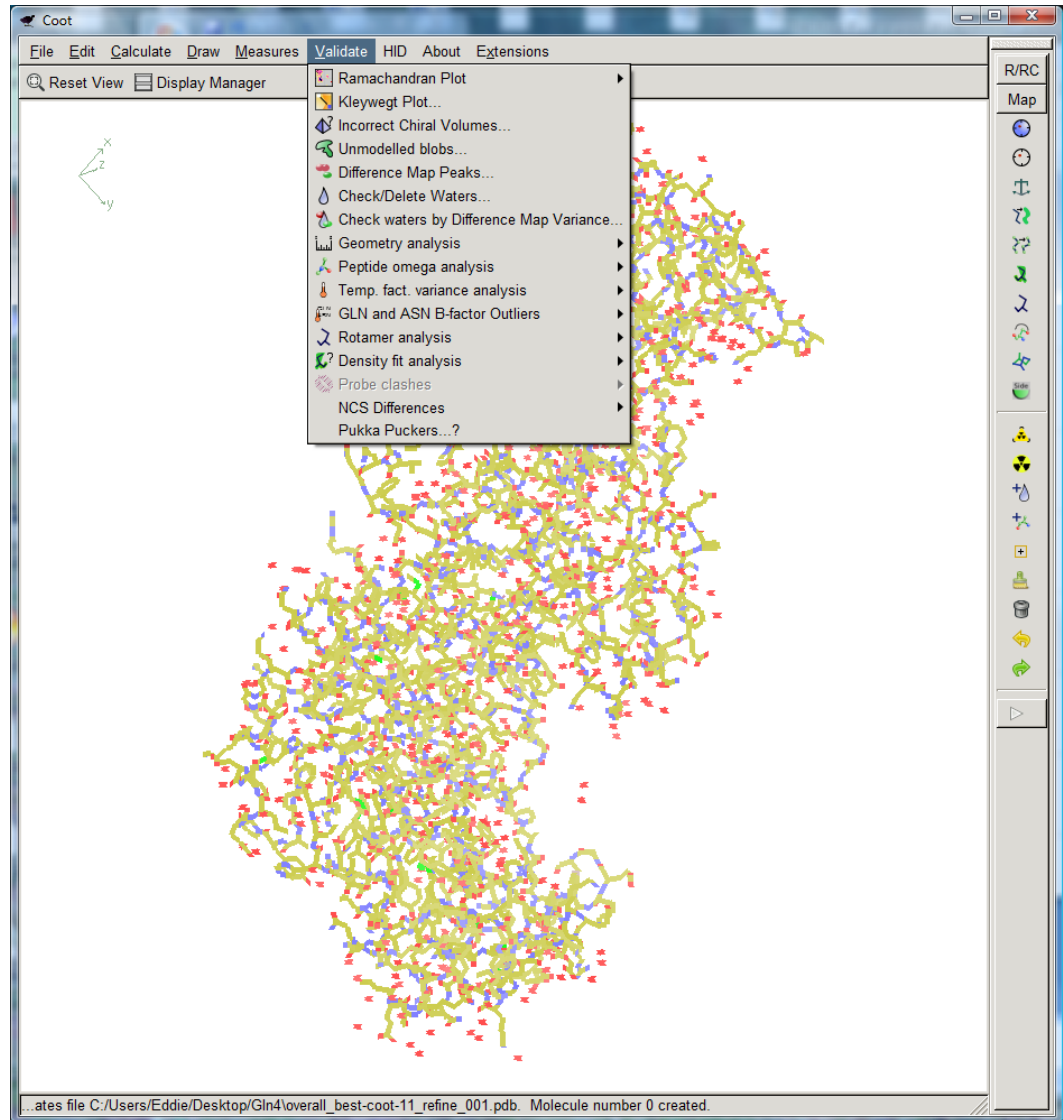
Similarly we can look for deviations in properties and geometry from previous data, e.g. Procheck <http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>

Some of these checks are incorporated within building and display programs, e.g. Coot



<http://www.ytbl.york.ac.uk/~emsley/coot/>

The validation option gives many options on how to check your structure against the data and against stereochemical restraints and previous structural information



Known knowns – we know what not to expect

Quick Guide to characteristics of metal sites in proteins - Windows Internet Explorer

http://tanna.bch.ed.ac.uk/qg3.htm

marjarie harding metal bond

File Edit View Favorites Tools Help

Quick Guide to characteristics of metal sites in proteins

QUICK GUIDE TO CHARACTERISTICS OF METAL SITES IN PROTEINS

metal	Na sodium	Mg ** magnesium	K potassium	Ca calcium	Mn manganese
atomic no. (= no. of electrons)	11	12	19	20	25
usual ion	Na ⁺	Mg ²⁺	K ⁺	Ca ²⁺	Mn ²⁺ (Mn ³⁺)
usual donor atoms ¹ (for more detail see.)	O m.chain O of asp.glu	O of asp.glu O m.chain	O m.chain O of asp.glu,	O of asp.glu, O m.chain	O of asp.glu N of his
other donors sometimes found (and occasionally found)	O of ser,thr O of asn,gln (N of his)	O of ser,thr O of asn,gln N of his	O of ser,thr O of asn,gln (N of his)	O of asn,gln O of ser,thr (N of his)	O m.chain O of asn,gln O of ser,thr
usual coordination number(s) ²	5,6	6	5,6	6	5,6
other coordination numbers ³	4,7	3,4,5	4,7,8	4,5,7,8	4,7
typical ⁴ distance (Å) more info	M-O 2.35-2.45 M-N M-S	2.05-2.15	2.75-2.85 ?+	2.35-2.45	2.15-2.20 2.21 2.35
relative abundance ⁵ in PDB	96	177	72	358	109
link to lists of examples ⁶	Na groups	Mg groups	K groups	Ca groups	Mn groups

** chlorophyll groups, containing Mg, are common, but not included here

Done Internet 100%

There are many sources of typical bond distances, e.g. Metal Coordination Sites in Proteins at:

<http://tanna.bch.ed.ac.uk/>

Known unknowns – we know when we have something we don't know



X-rays are diffracted due to interactions with electrons in the atoms.

The data we produce is a map of electron density.

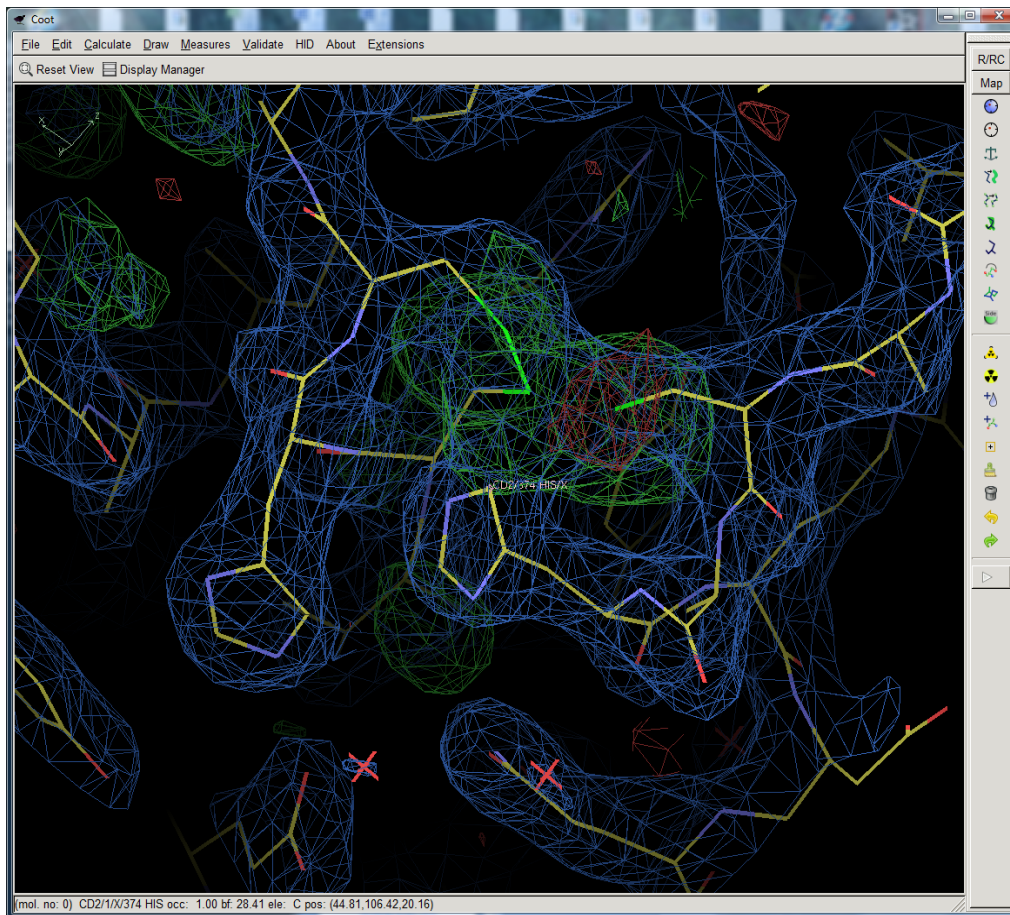
Known unknowns – we know when we have something we don't know

If we plot electron density maps as:

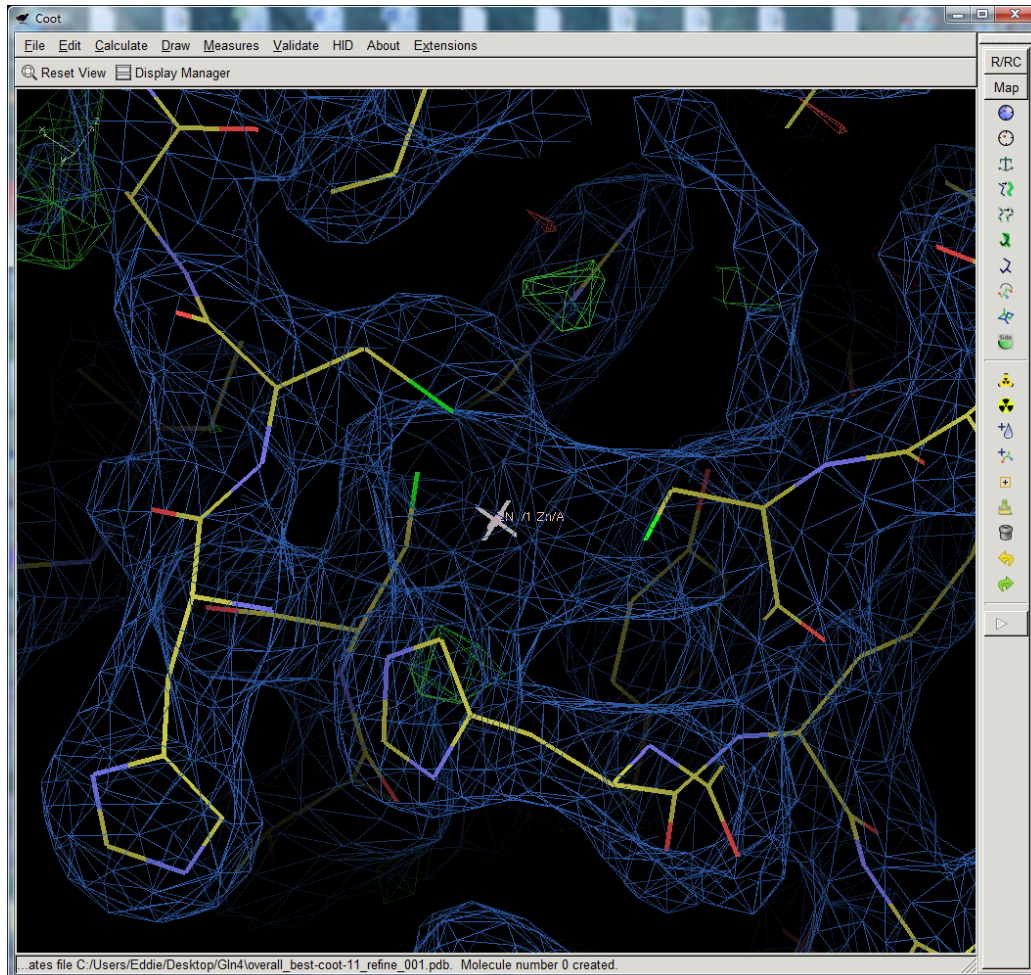
(Two times the observed data – the calculated data) in blue (2Fo-Fc)

And (the observed data – the calculated data) and color this according to negative (red) and positive (green) (Fo-Fc)

We know we have something we don't know.



Known unknowns – we know when we have something we don't know



If we model the correct atoms, i.e. a zinc finger motif then the maps tells us that we got it correct.

We can adjust our model to add the 'known something we didn't know'.

Unknown unknowns – The R-factor

$$I_{hkl} \propto |F(hkl)|^2$$

The X-ray intensity for a particular reflection (hkl), I_{hkl} , is measured, F_{obs} is related to the measured I.

The X-ray intensity for a particular reflection (hkl) can also be calculated from the model, I_{calc} , is calculated, F_{calc} is related to the calculated I and the model.

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}$$

The R factor – small is good.

A small R-factor indicates minimal differences between the electron density calculated for the model and that calculated from the observed data.

The R-factor – What does it mean?

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}$$

If the observed measurement is in complete agreement with the calculated measurement then R will equal zero.

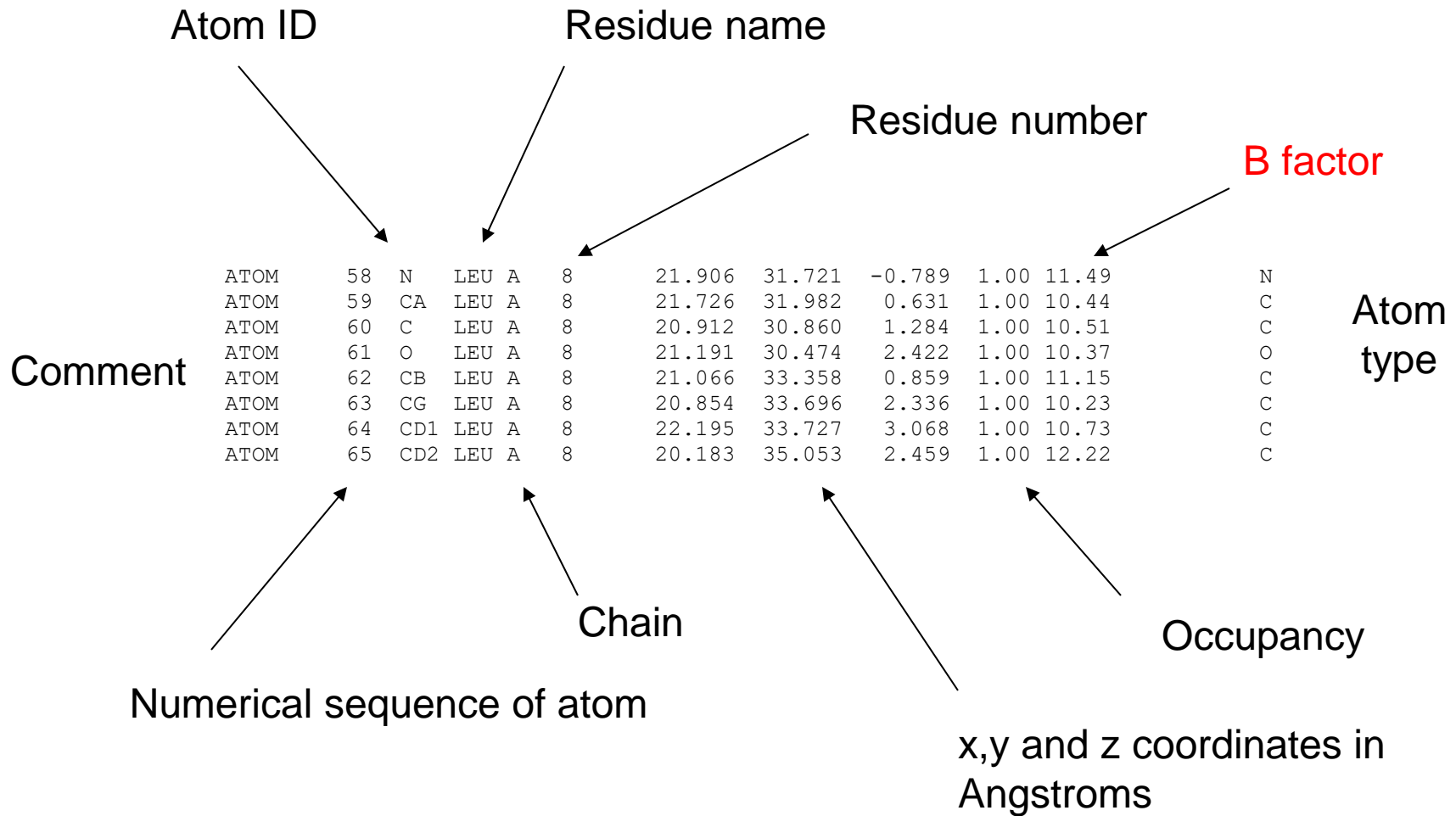
If there is disagreement then R will be finite. R is expressed as a percentage and is typically a little better than ten times the resolution for a good structure.

The R_{free} -factor – What does it mean?

$$R = \frac{\sum ||F_{\text{obs}}| - |F_{\text{calc}}||}{\sum |F_{\text{obs}}|}$$

Some 5-10% of the reflections are not used to calculate the model.
These are used to calculate an R_{free} .

When the model is improving, i.e. it is accurately explaining the data and both the R and R_{free} should reduce during refinement. Once the R_{free} stops reducing the model is being overfitted to the data – it is losing accuracy.



The B factor (or atomic displacement factor) describes how the electron is spread out in space. A high B factor would indicate a high degree of uncertainty in the atomic position and a potential warning sign.

A satellite-style map of a coastal region, likely the Gulf of Mexico and Florida peninsula. The land is shown in shades of green and brown, while the water is a deep blue. The map is characterized by very low resolution, with large, blocky pixels and a lack of fine detail. The text "Very low Resolution" is overlaid in yellow in the center-left area.

Very low Resolution

An aerial photograph of a river system. The river is dark blue and flows from the top right towards the bottom left. The surrounding land is a mix of green and brown, indicating vegetation and bare earth. The image is labeled 'Low Resolution' in yellow text on the left side. The river's path is somewhat irregular, with several bends and a small tributary joining from the bottom. The overall appearance is that of a low-resolution satellite or aerial image.

Low Resolution

An aerial photograph of a city, likely Los Angeles, showing a dense urban grid and a large river (the Los Angeles River) winding through the landscape. The image is labeled 'Medium Resolution' in yellow text. The river is dark blue and occupies the left side of the frame. The city grid is composed of numerous small, rectangular blocks. The overall color palette is dominated by greys, browns, and greens, with the river providing a stark contrast in blue. The text 'Medium Resolution' is centered horizontally and vertically, overlaid on the river and the city grid.

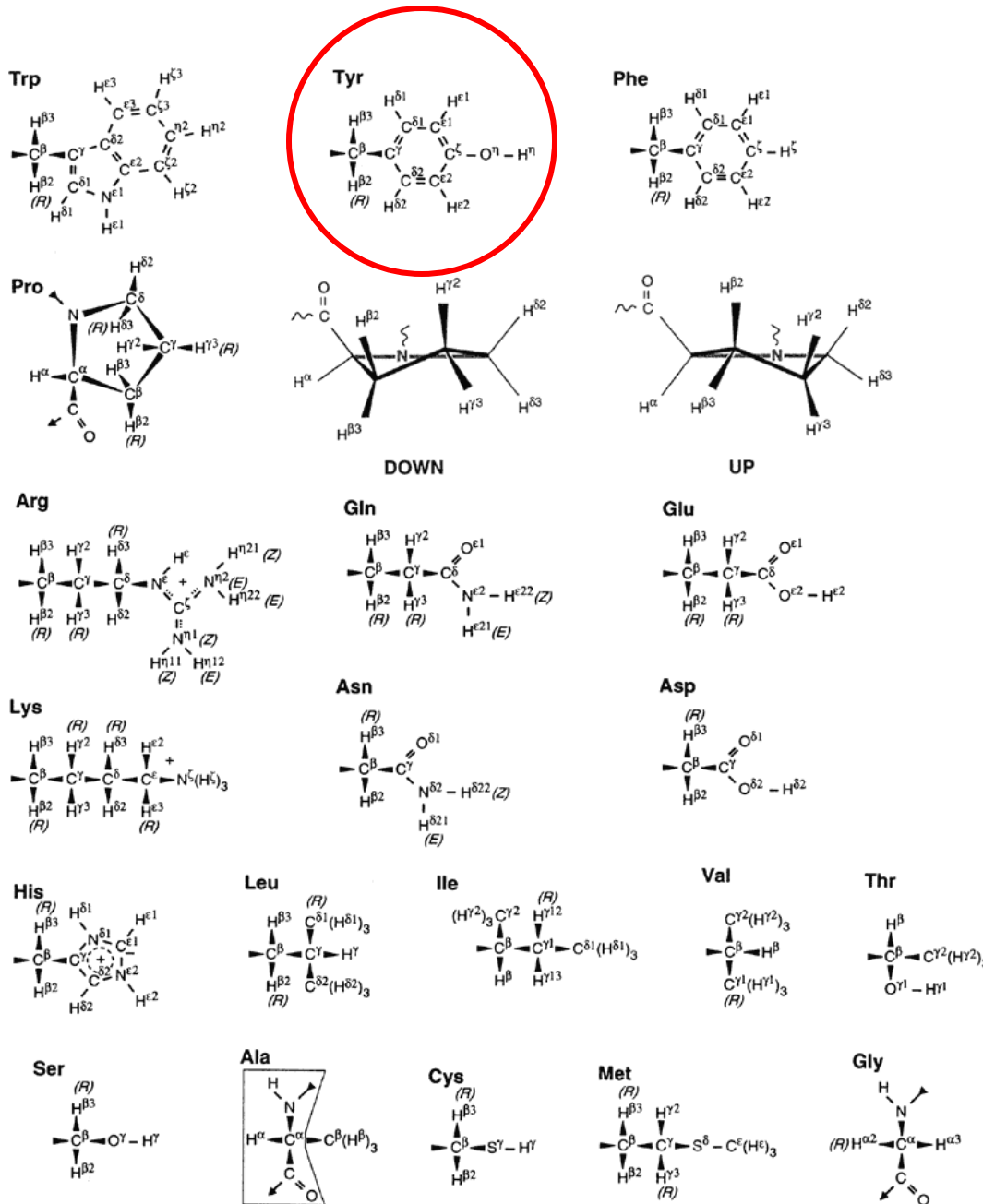
Medium Resolution

This is a high-resolution aerial satellite image of a city. The image shows a dense urban grid with numerous buildings, streets, and parking lots. A prominent feature is a large, irregularly shaped green area on the left side, which appears to be a park or a forested area. A river or stream flows through this green area. In the center-right, there is a large, multi-lane highway interchange. The overall scene is a detailed view of an urban environment.

High Resolution

Very-high Resolution



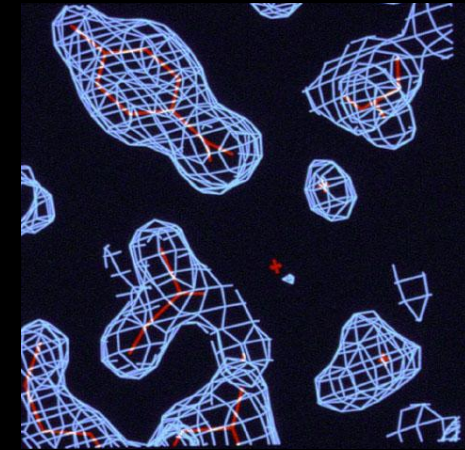


The amino acids that form the building blocks of biological molecules

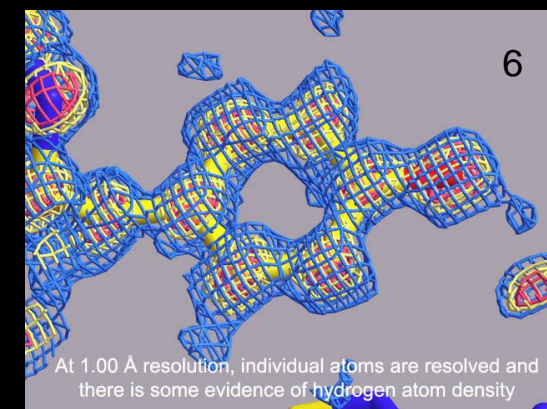
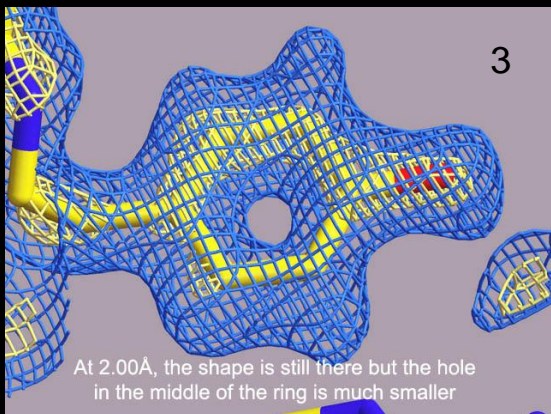
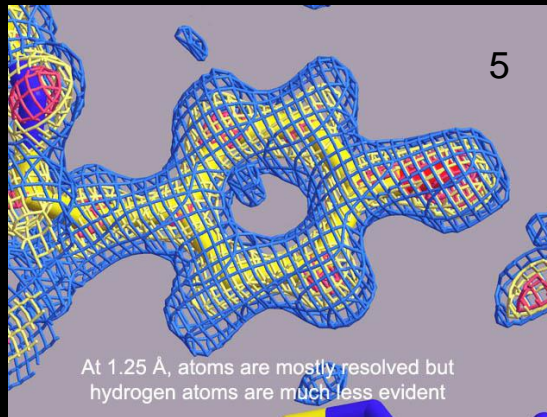
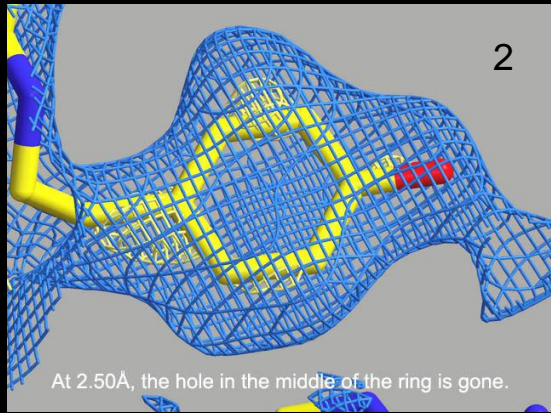
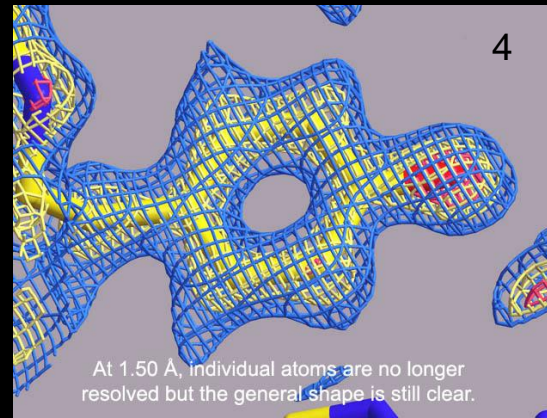
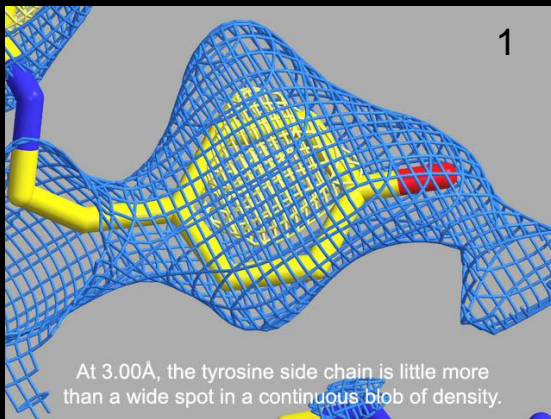
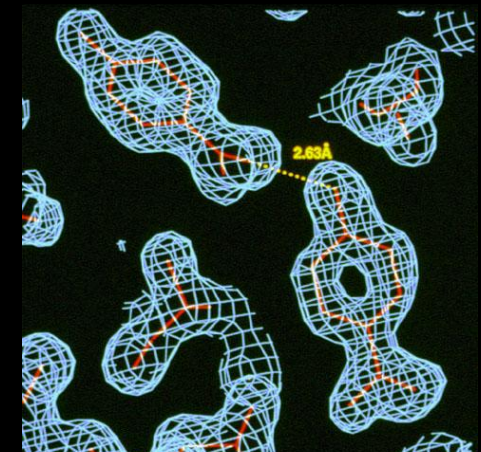
Let's pay particular attention to one of them and the concept of resolution

Quality (Resolution)

Low resolution



High resolution



Where does Reality come in?

- The model is accurate if it can be validated. The higher the resolution then the more precise the detail in the model.
- Remember, it is only a model that explains the data and not data that explains the model.

Examples of Publically available software

Refinement:

- Phenix, a software suite for doing just about everything with data but displaying the structure: <http://www.phenix-online.org/>
- CCP4, similar to Phenix but covering a broader area of application and including validation and display components: <http://www.ccp4.ac.uk/main.html>
- CNS, Crystallography and NMR system: <http://cns-online.org/v1.21/>

Validation:

- Excellent tutorial on validation: <http://xray.bmc.uu.se/gerard/embo2001/modval/index.html>
- Procheck is a good validation example: <http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html> and is incorporated with CCP4.
- MolProbity allows both validation and fixing the model: <http://molprobity.biochem.duke.edu/>
- Direct validation through the PDB: <http://sw-tools.pdb.org/apps/VAL/index.html>

Display:

- Coot (also includes validation routines): <http://www.yesbl.york.ac.uk/~emsley/cool/>

