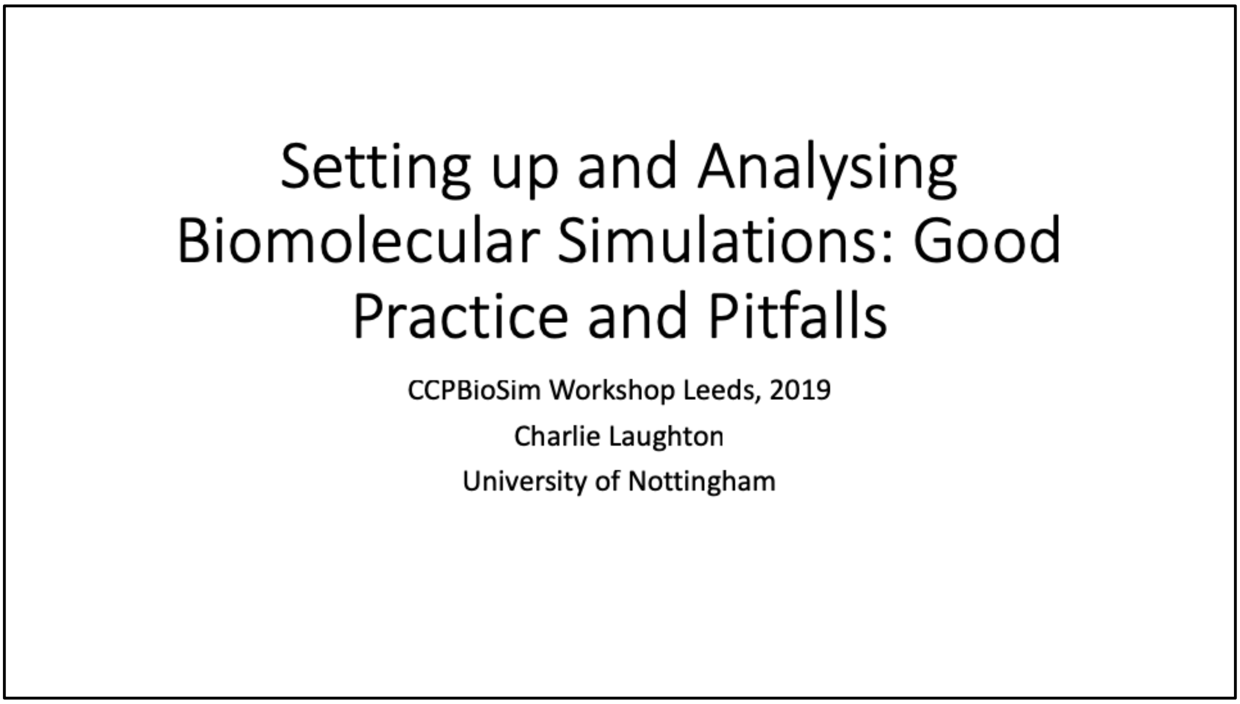
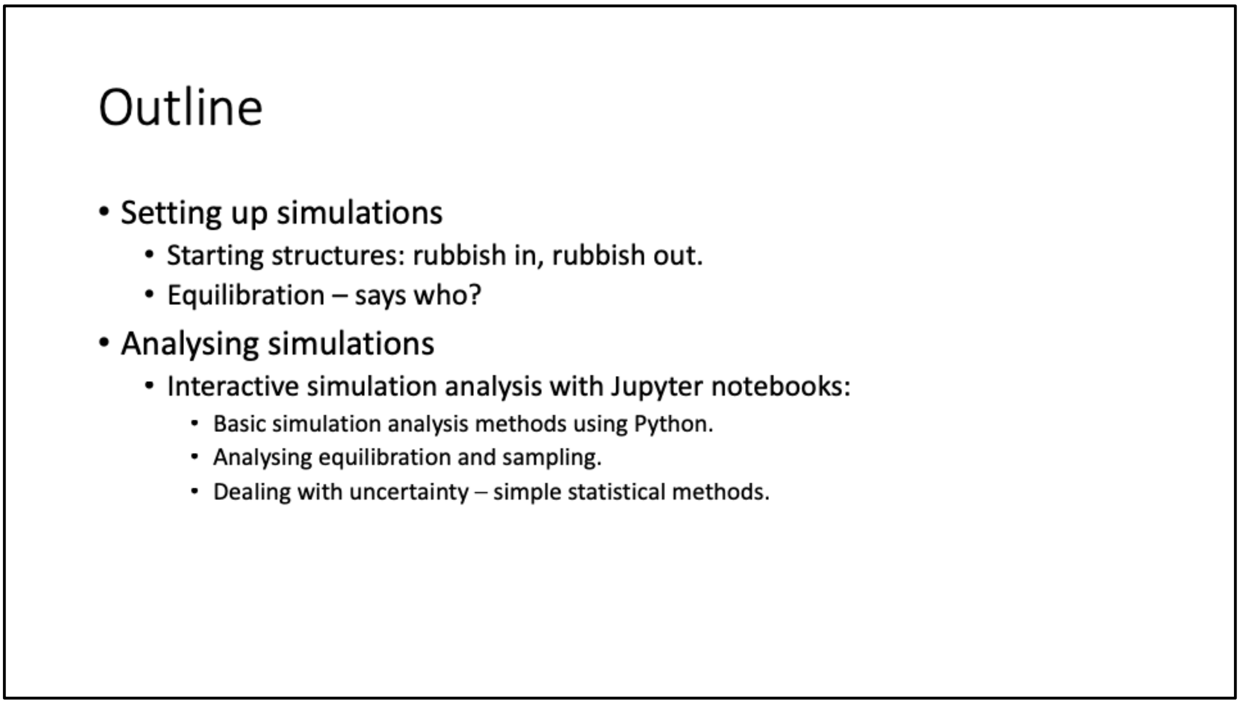


Setting up and Analysing Biomolecular Simulations: Good Practice and Pitfalls

CCPBioSim Workshop Leeds, 2019

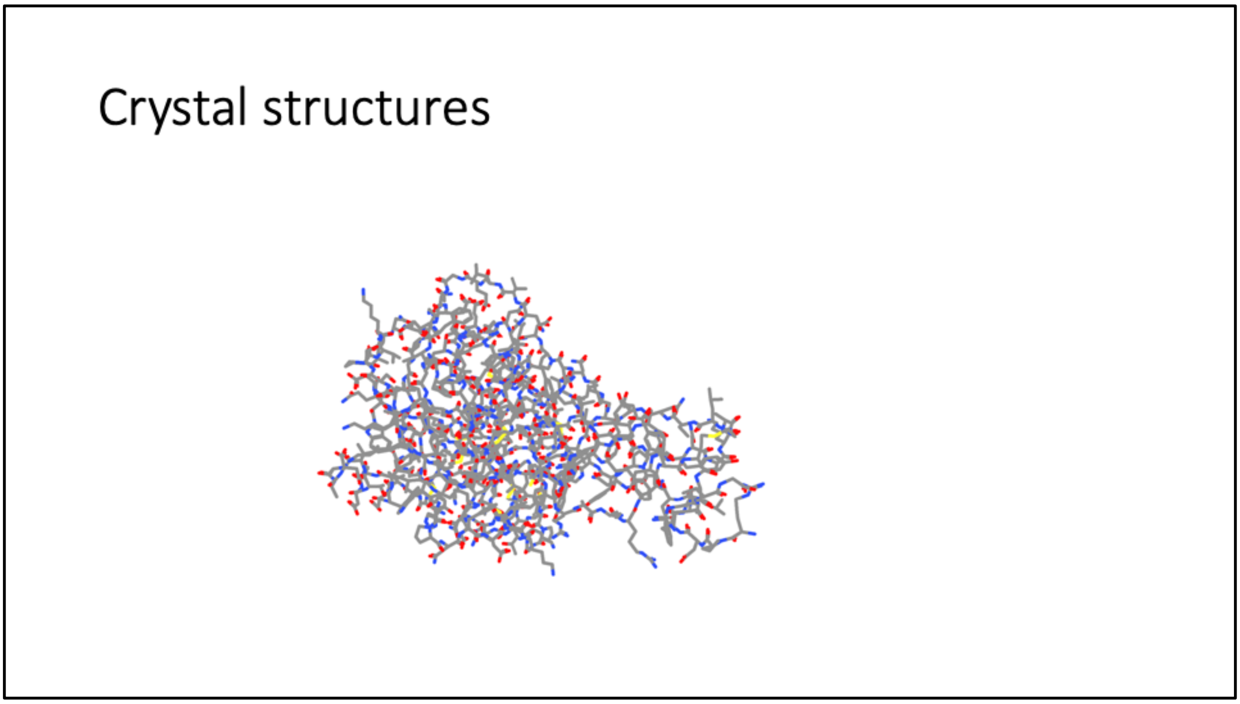
Charlie Laughton
University of Nottingham



Outline

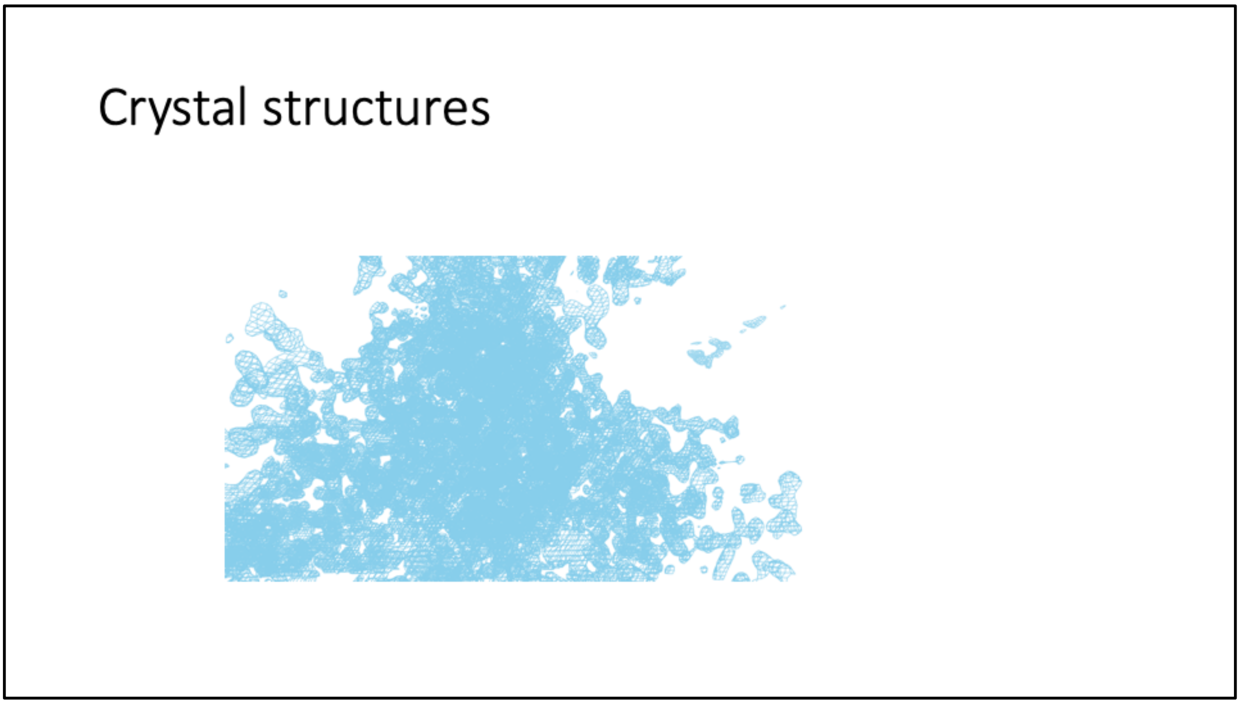
- Setting up simulations
  - Starting structures: rubbish in, rubbish out.
  - Equilibration – says who?
- Analysing simulations
  - Interactive simulation analysis with Jupyter notebooks:
    - Basic simulation analysis methods using Python.
    - Analysing equilibration and sampling.
    - Dealing with uncertainty – simple statistical methods.

What we will aim to cover. Part 1 is this informal lecture with some interspersed activities, part 2 is a set of self-paced Jupyter notebooks



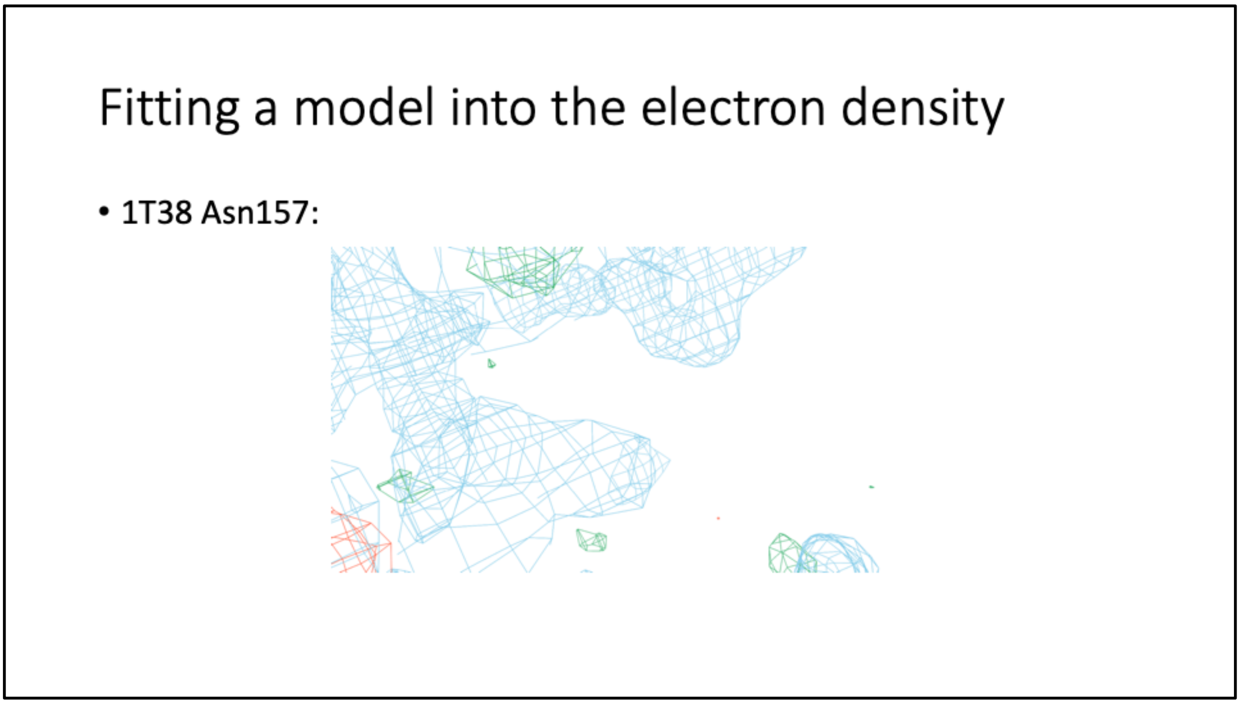
Crystal structures

For most modeling we need to start with a structure; very often this comes from Xray crystallography – e.g. this.



Crystal structures

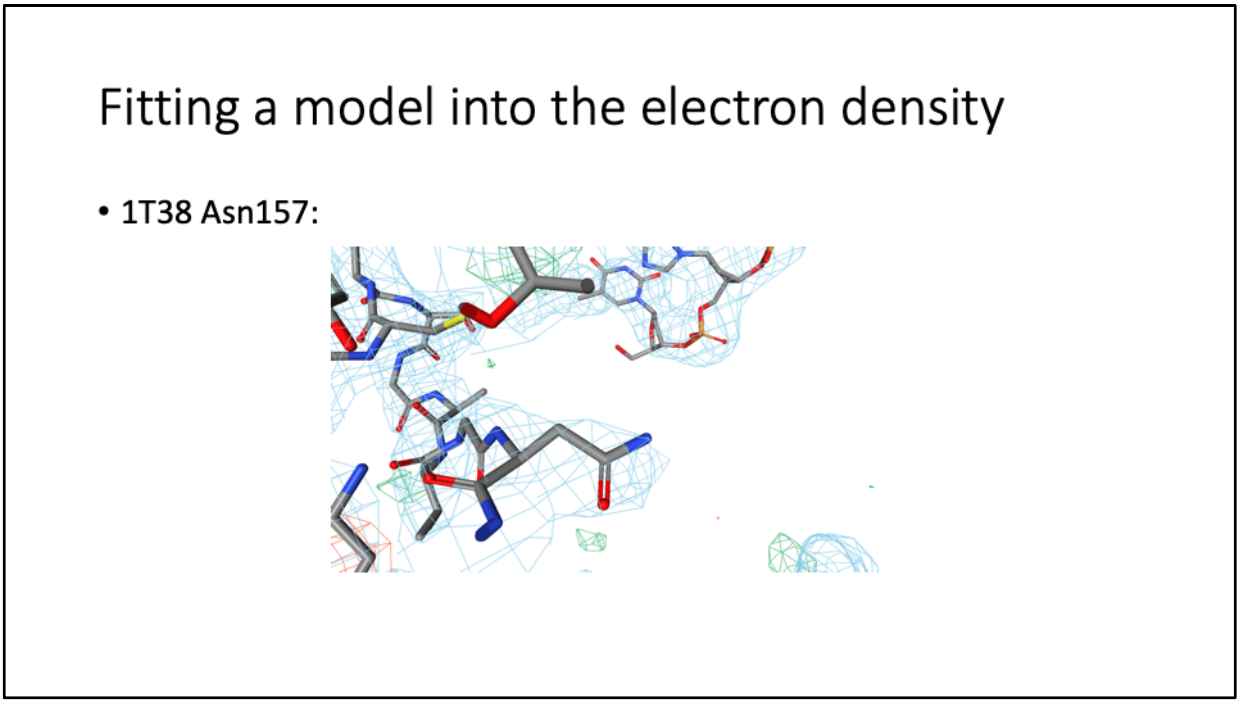
But that is NOT what the crystallographer saw – they saw an electron density map.
The “structure” you download is a MODEL that fits the data – it may not be accurate.



Fitting a model into the electron density

- 1T38 Asn157:

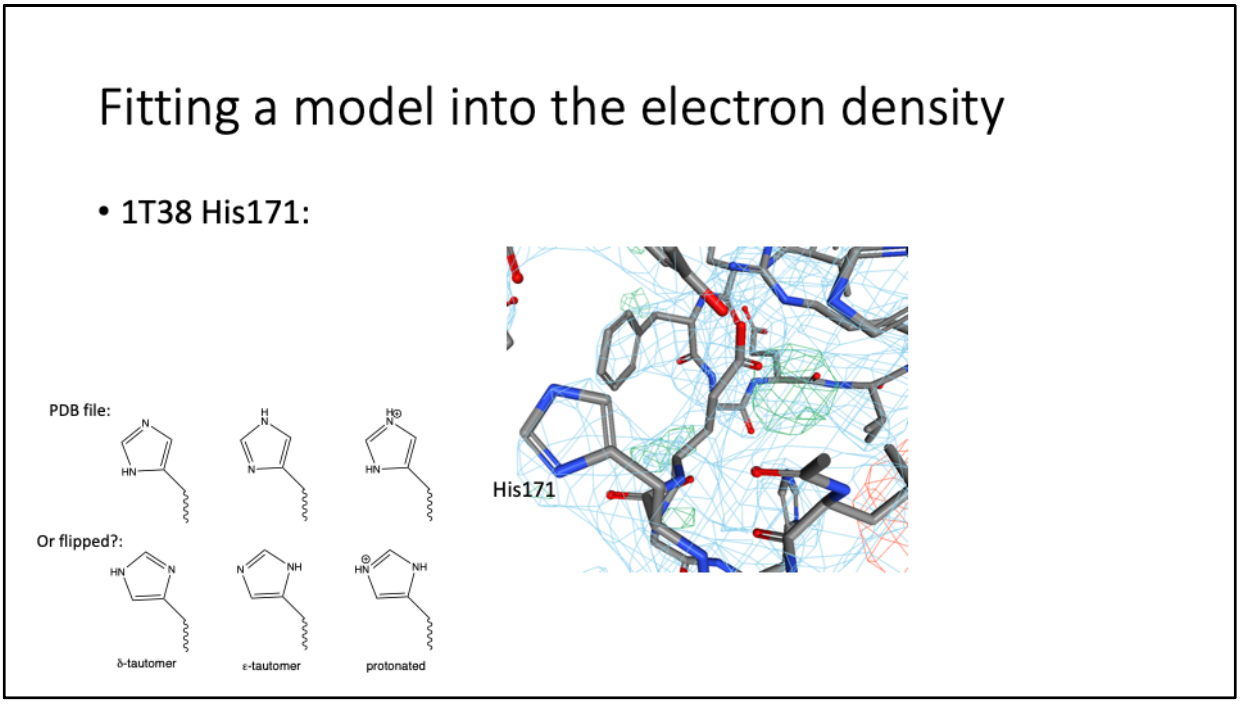
There is an aspargine side chain in this blob…



Fitting a model into the electron density

- 1T38 Asn157:

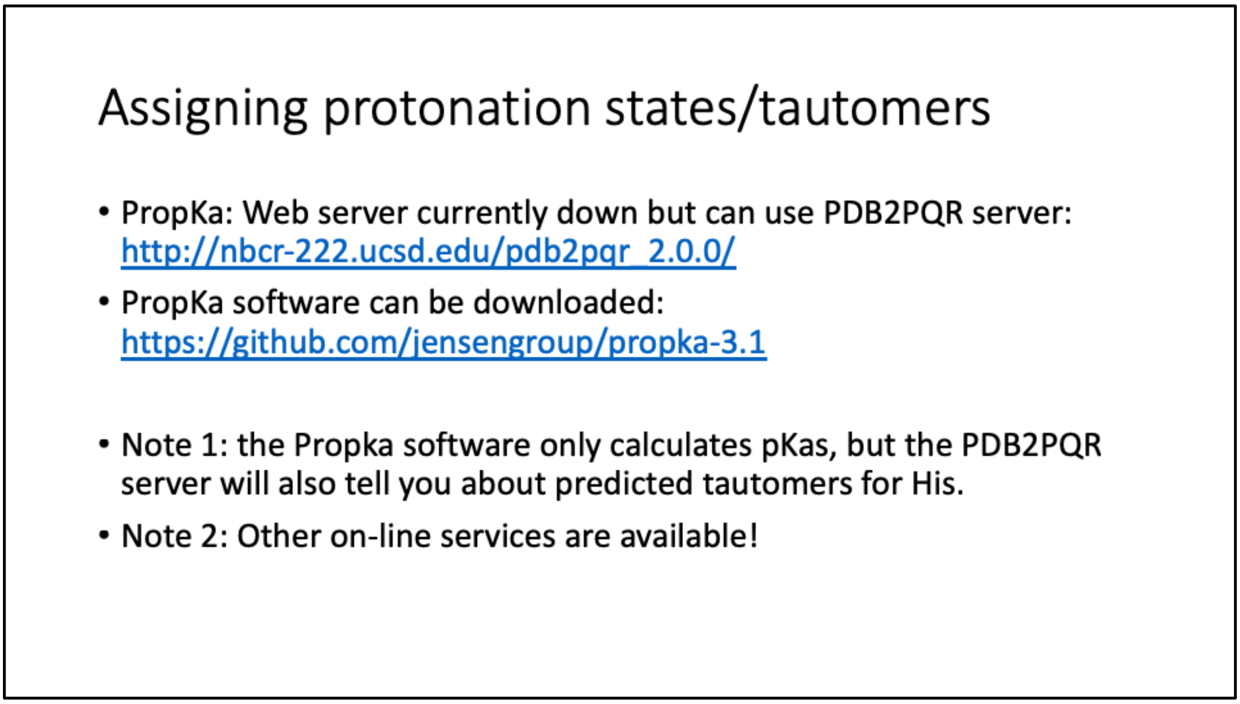
… but how do we know which way round the amide group is truly oriented?



Fitting a model into the electron density

- 1T38 His171:

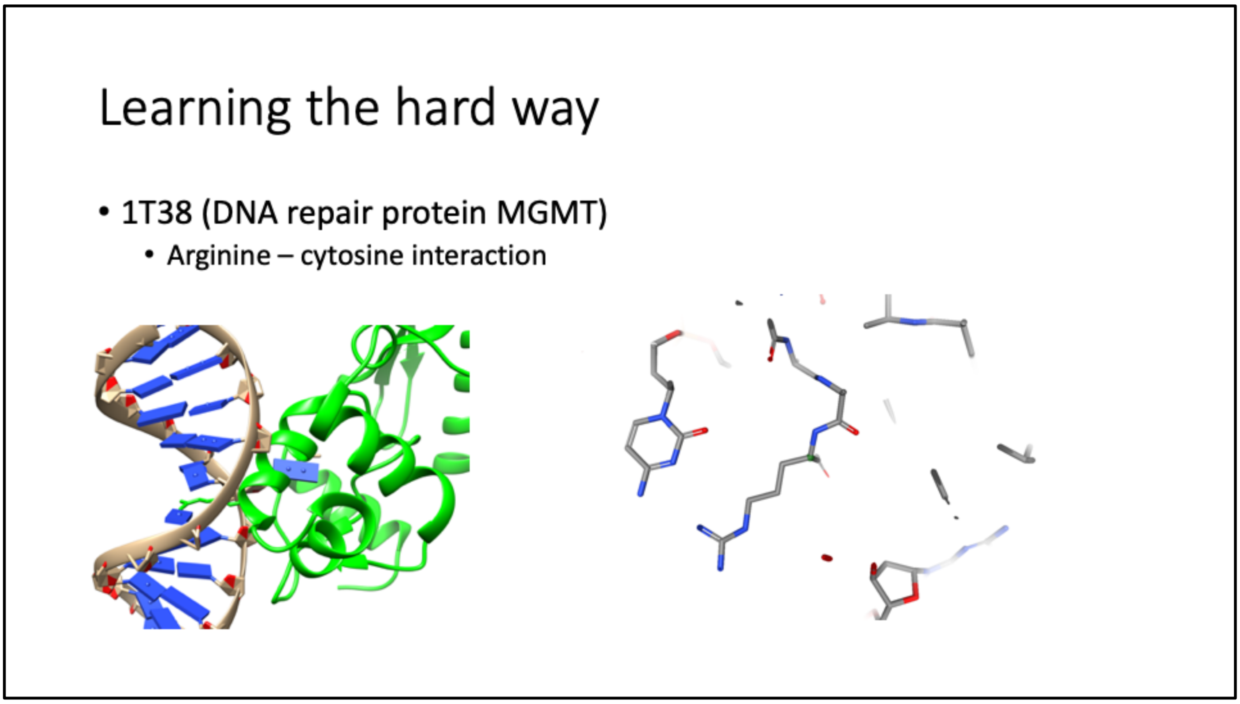
If a blob is supposed to be a histidine side chain, how do we know its orientation, or where the hydrogens are – there are six different solutions that might be compatible with the electron density data.



Assigning protonation states/tautomers

- PropKa: Web server currently down but can use PDB2PQR server: http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/
- PropKa software can be downloaded: https://github.com/jensengroup/propka-3.1
- Note 1: the Propka software only calculates pKas, but the PDB2PQR server will also tell you about predicted tautomers for His.
- Note 2: Other on-line services are available!

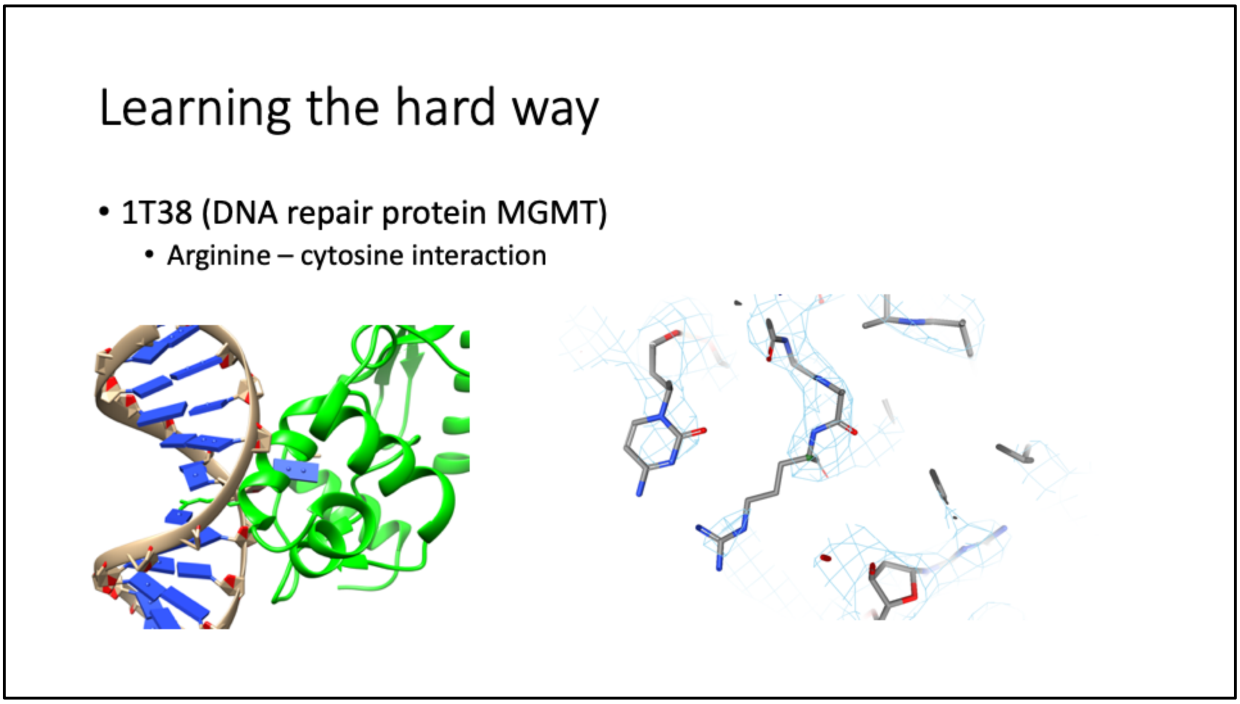
There are tools out there to help you assign protonation states, they may be better than the default methods built in to the modelling code you use (GROMACS, AMBER, etc). At least get a consensus opinion by using a couple of alternative methods.



Learning the hard way

- 1T38 (DNA repair protein MGMT)
  - Arginine – cytosine interaction

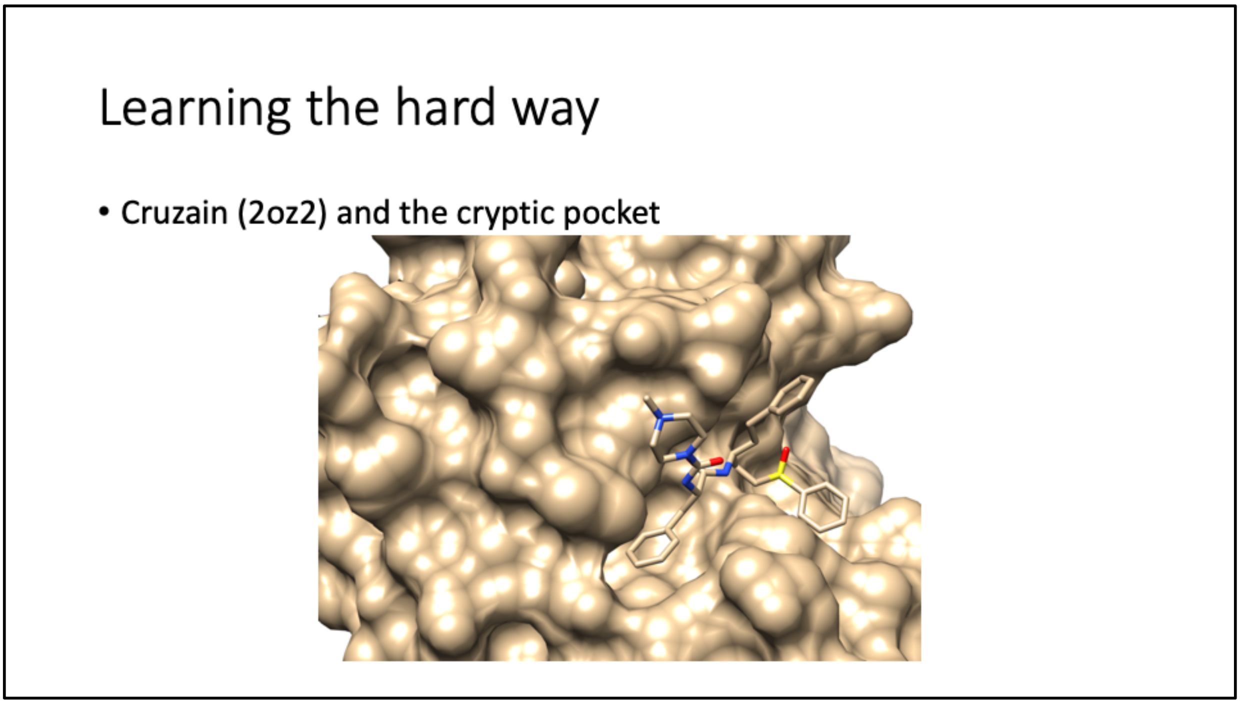
An example from our own work. We were interested in understanding how DNA repair proteins locate the site of damage. In the crystal structure of an MGMT complex we see an “arginine finger” in the protein that seems to be “sensing” the position of the damaged base (an alkylated guanine), by forming an interaction with its Watson-Crick partner (a cytosine). But the geometry looks a bit odd. We spent ages doing MD simulatons and umbrella sampling to try and understand how the “arginine finger” might operate, but got no sensible results.



Learning the hard way

- 1T38 (DNA repair protein MGMT)
  - Arginine – cytosine interaction

Very late in the day, we went back to the crystal structure, and the Electron Density Server (EDS). This shows that actually in the experiment the position of the arginine is very poorly resolved – it might be just about anywhere really, doing anything. A bad structure to try and hang a modelling investigation off!

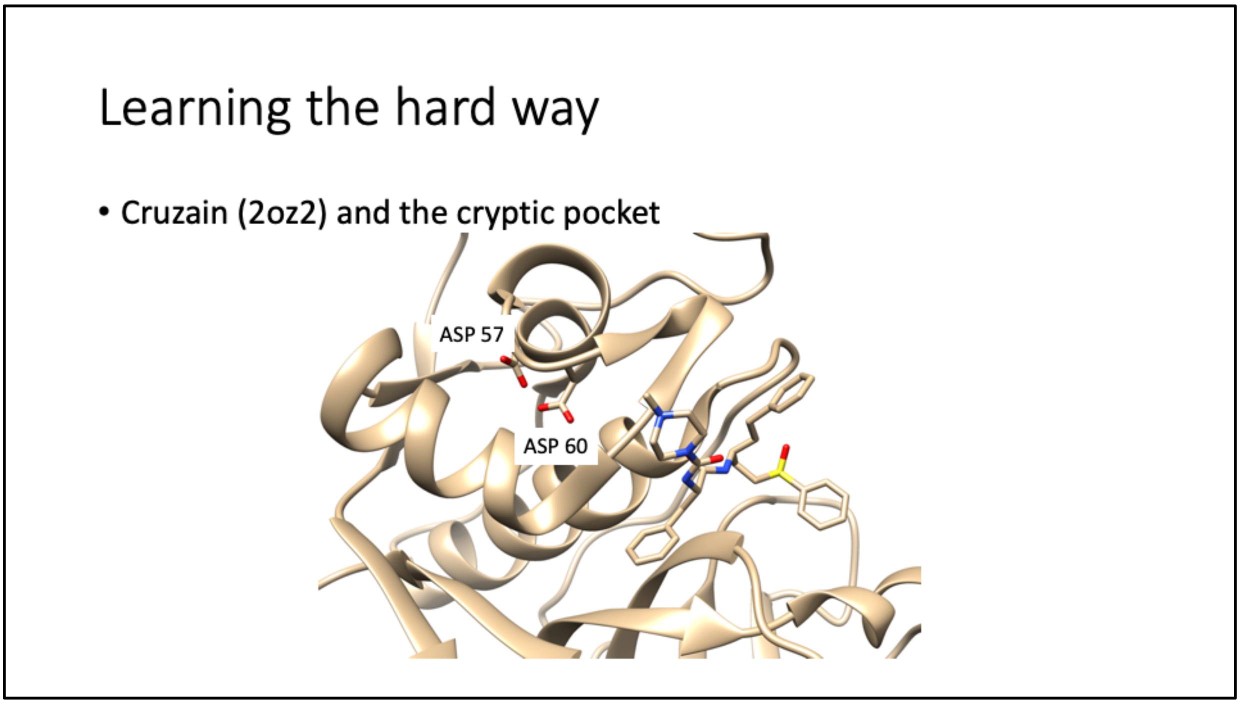


Learning the hard way

- Cruzain (2oz2) and the cryptic pocket

Another example. We have a project to design new inhibitors of cruzain, a cysteine protease found in the parasite that causes Chagas disease. Crystal structures show us it is very hard to design an inhibitor specific for this enzyme, as the host (us) also has many important cysteine proteases with very similar 3D structures that we need to avoid inhibiting at the same time.

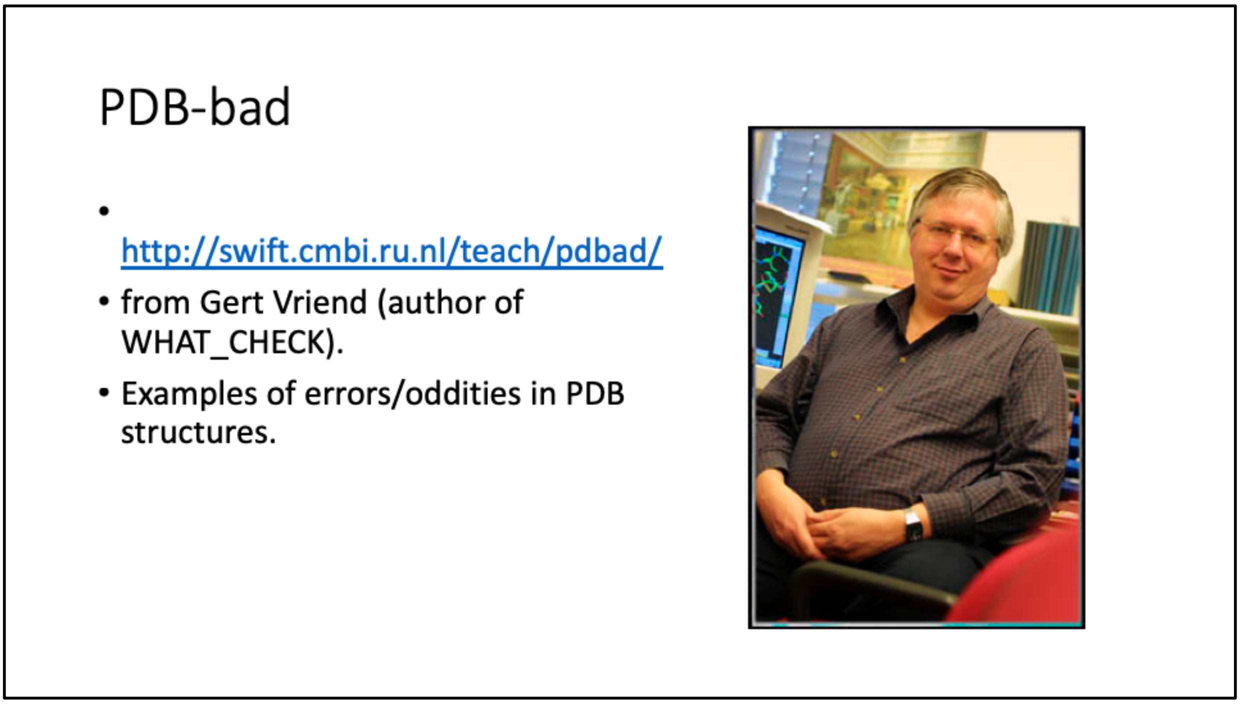
So we got very excited when MD simulations showed that cruzain, unlike other cysteine proteases, seemed to have a cryptic pocket that drug structures could be designed to interact with.



Learning the hard way

- Cruzain (2oz2) and the cryptic pocket

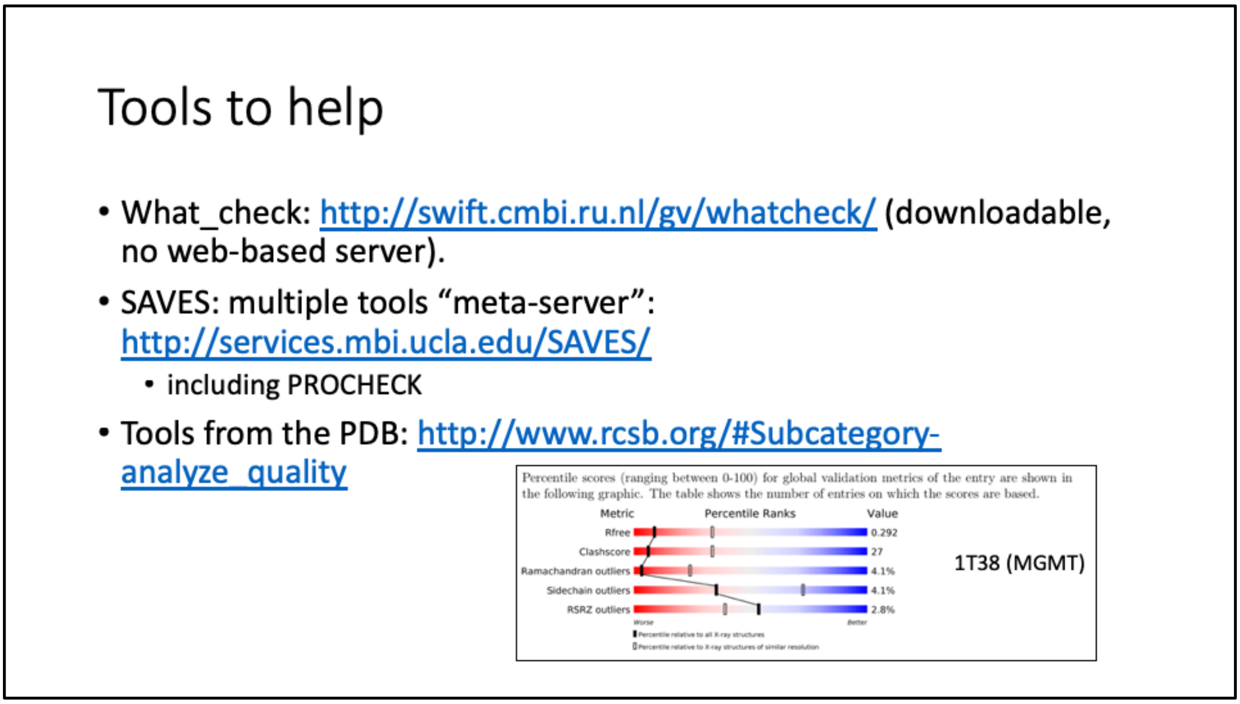
Then we looked a bit more carefuly, and saw that the flexible loop that moves to create the cryptic pocket contains two aspartic acid residues in very close proximity. It is well-known that this sort of geometry is found in aspartic proteases – it's a key feature of their catalytic mechanism – and that because of the unusual microenvironment one of them exists in the neutral, protonated, form while the other is in the normal, deprotonated form. Our default modelling package had assumed that the aspartates in cruzain would all be normal, deprotonated ones. The strong electrostatic repulsion between the two charged centres was driving the opening of the packet. When we submitted the structure to the PropKa server, it predicts that - just about – one of the aspartates here is also in the neutral form. When we repeated the simulations with a protonated aspartate, the cryptic pocket did not reappear!



PDB-bad

- http://swift.cmbi.ru.nl/teach/pdbad/
- from Gert Vriend (author of WHAT_CHECK).
- Examples of errors/oddities in PDB structures.

Gert Vriend has a good (quite funny) web site in which he describes some of the dodgy things he has found in deposited crystal structures – well worth a browse.



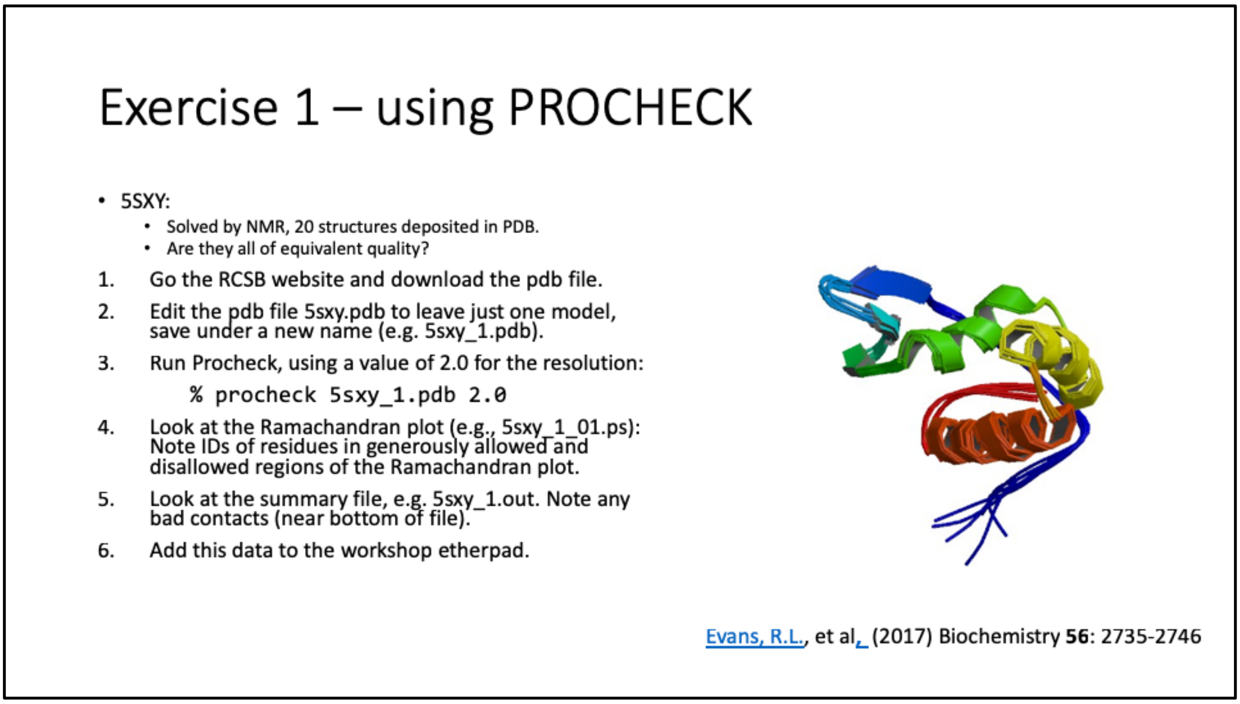
Tools to help

- What_check: http://swift.cmbi.ru.nl/gv/whatcheck/ (downloadable, no web-based server).
- SAVES: multiple tools "meta-server": http://services.mbi.ucla.edu/SAVES/
  - including PROCHECK
- Tools from the PDB: http://www.rcsb.org/#Subcategory-analyze_quality

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

1T38 (MGMT)

There are a number of tools available to check protein models for "quality", but to be honest many of them are pretty old. This doesn't make them wrong, just rather difficult to use. But a couple of web sites offer a decent service. Also, it is now much easier to asses the quality of a structure by looking at data in the PDB itself.



Exercise 1 – using PROCHECK

- 5SXY:
  - Solved by NMR, 20 structures deposited in PDB.
  - Are they all of equivalent quality?

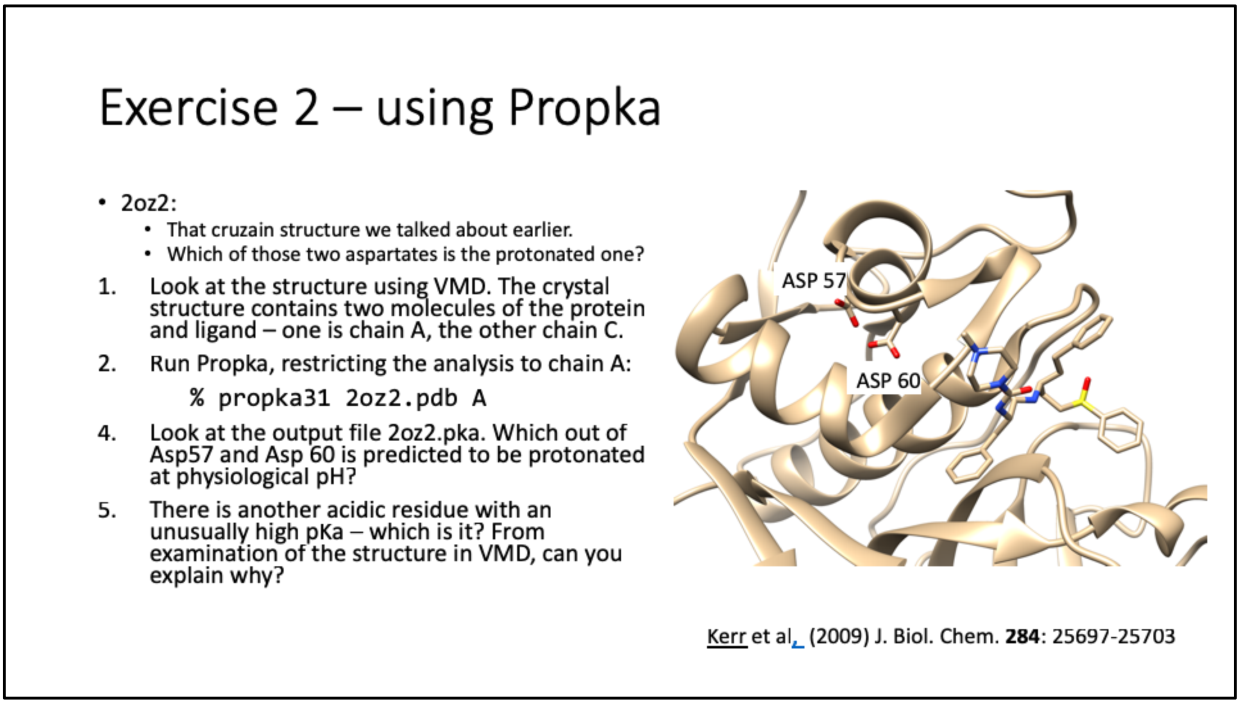
1. Go the RCSB website and download the pdb file.
2. Edit the pdb file 5sxy.pdb to leave just one model, save under a new name (e.g. 5sxy_1.pdb).
3. Run Procheck, using a value of 2.0 for the resolution:

```
% procheck 5sxy_1.pdb 2.0
```

4. Look at the Ramachandran plot (e.g., 5sxy_1_01.ps): Note IDs of residues in generously allowed and disallowed regions of the Ramachandran plot.
5. Look at the summary file, e.g. 5sxy_1.out. Note any bad contacts (near bottom of file).
6. Add this data to the workshop etherpad.

Evans, R.L., et al, (2017) Biochemistry **56**: 2735-2746

An exercise for the attendees. Compare the quality of each of the 20 different models in this NMR-derived PDB entry.



Exercise 2 – using Propka

- 2oz2:
  - That cruzain structure we talked about earlier.
  - Which of those two aspartates is the protonated one?

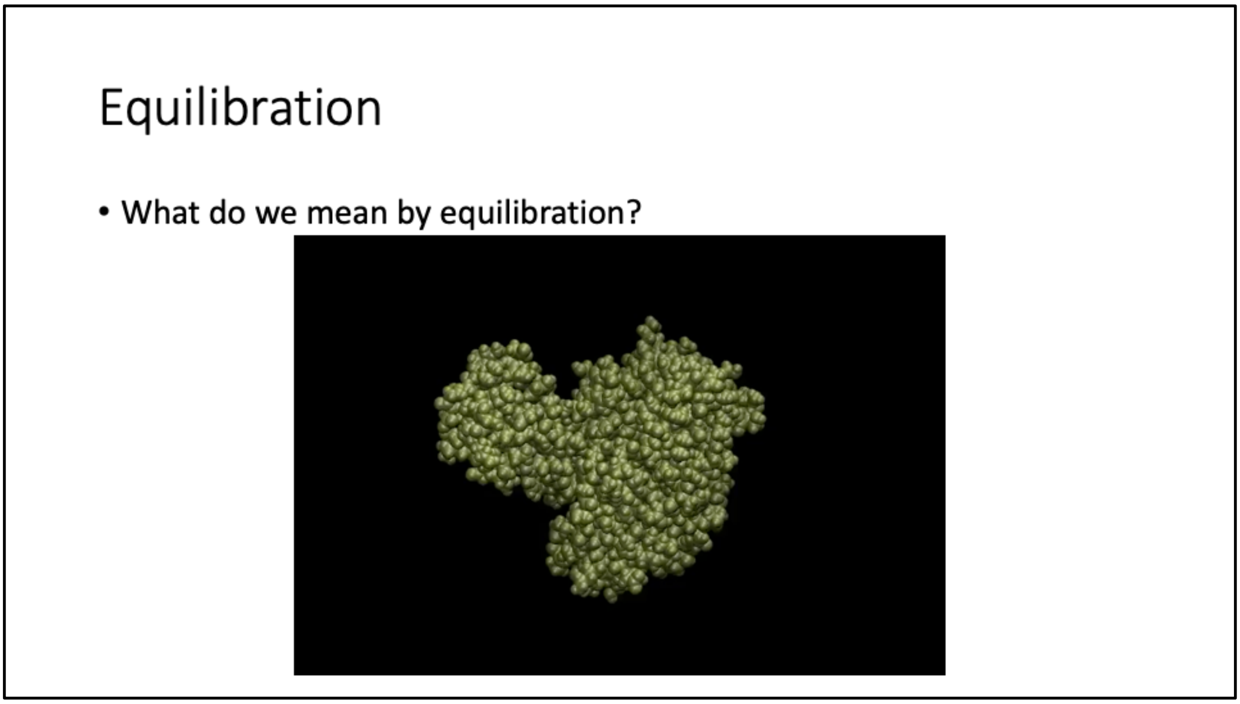
1. Look at the structure using VMD. The crystal structure contains two molecules of the protein and ligand – one is chain A, the other chain C.
2. Run Propka, restricting the analysis to chain A:

```
% propka31 2oz2.pdb A
```

4. Look at the output file 2oz2.pka. Which out of Asp57 and Asp 60 is predicted to be protonated at physiological pH?
5. There is another acidic residue with an unusually high pKa – which is it? From examination of the structure in VMD, can you explain why?

Kerr et al, (2009) J. Biol. Chem. **284**: 25697-25703

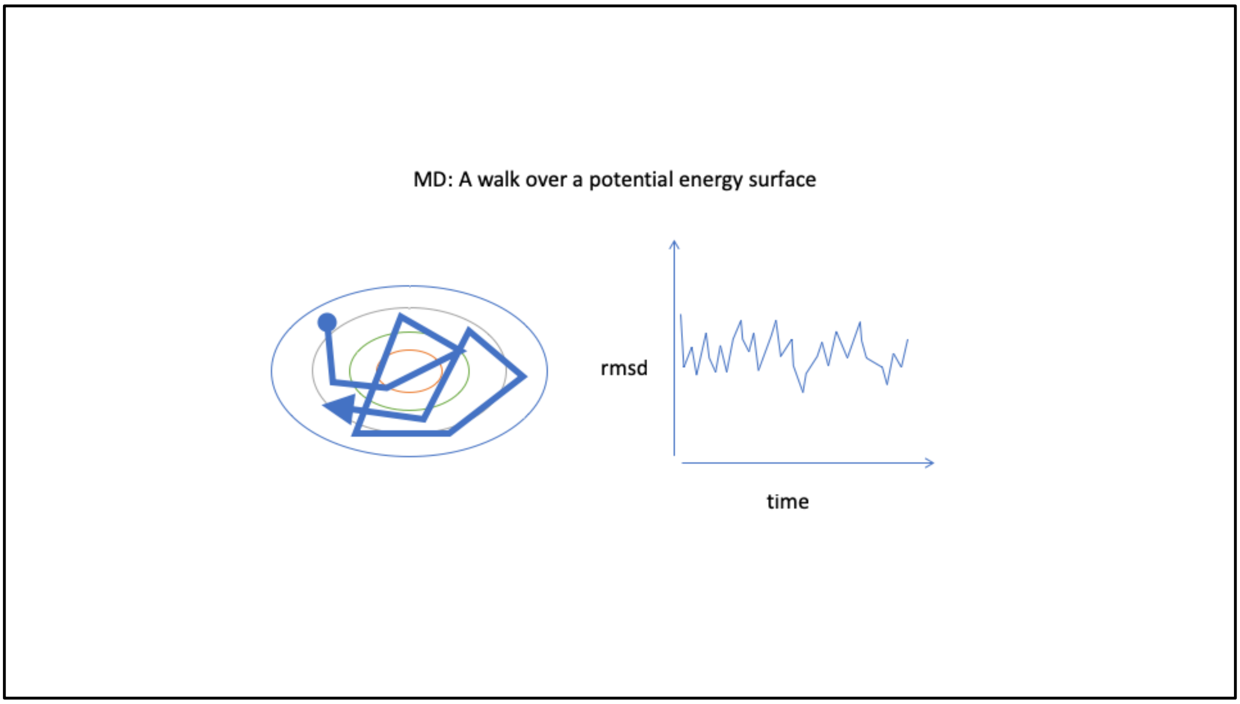
An exercise for the attendees. Use the Propka software and analyse the output files it produces.



Equilibration

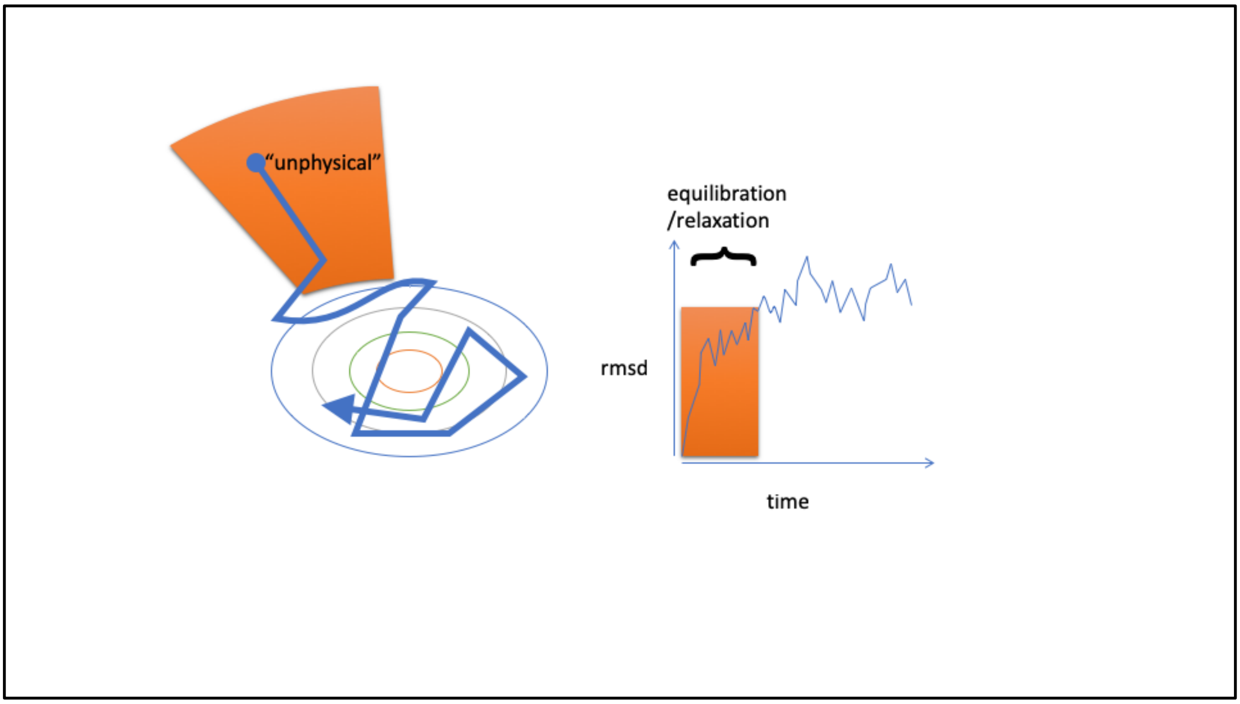
- What do we mean by equilibration?

We all have the idea that an MD simulation study can be divided into an “equilibration” phase and a “production” phase – but what exactly do we think is going on - how do we tell when we have moved from one phase to the other? This is actually a massive can of worms, and unfortunately even in many published papers the way “equilibration” is demonstrated is rather poor (to put it mildly).

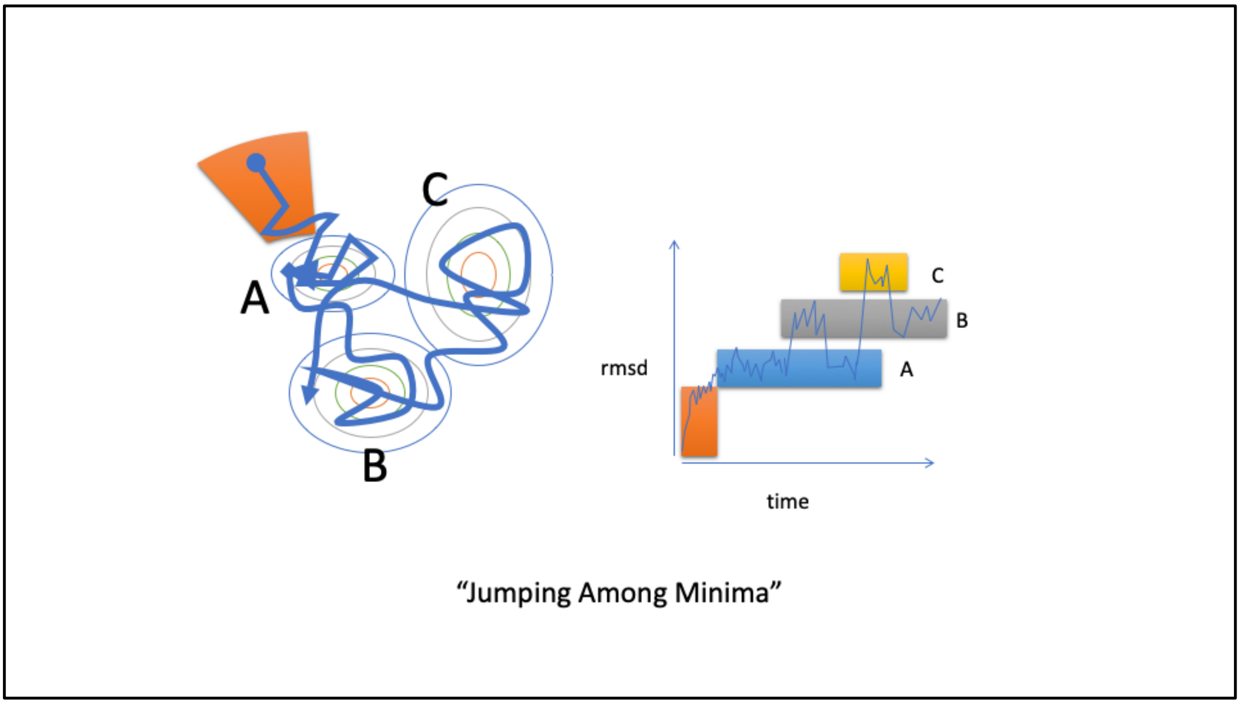


MD: A walk over a potential energy surface

A simple, ideal, simulation. The trajectory is contained within a single potential well. RMSD measures the ‘distance’ of each snapshot from some reference position (e.g. the centre of the well), and is more-or-less constant.

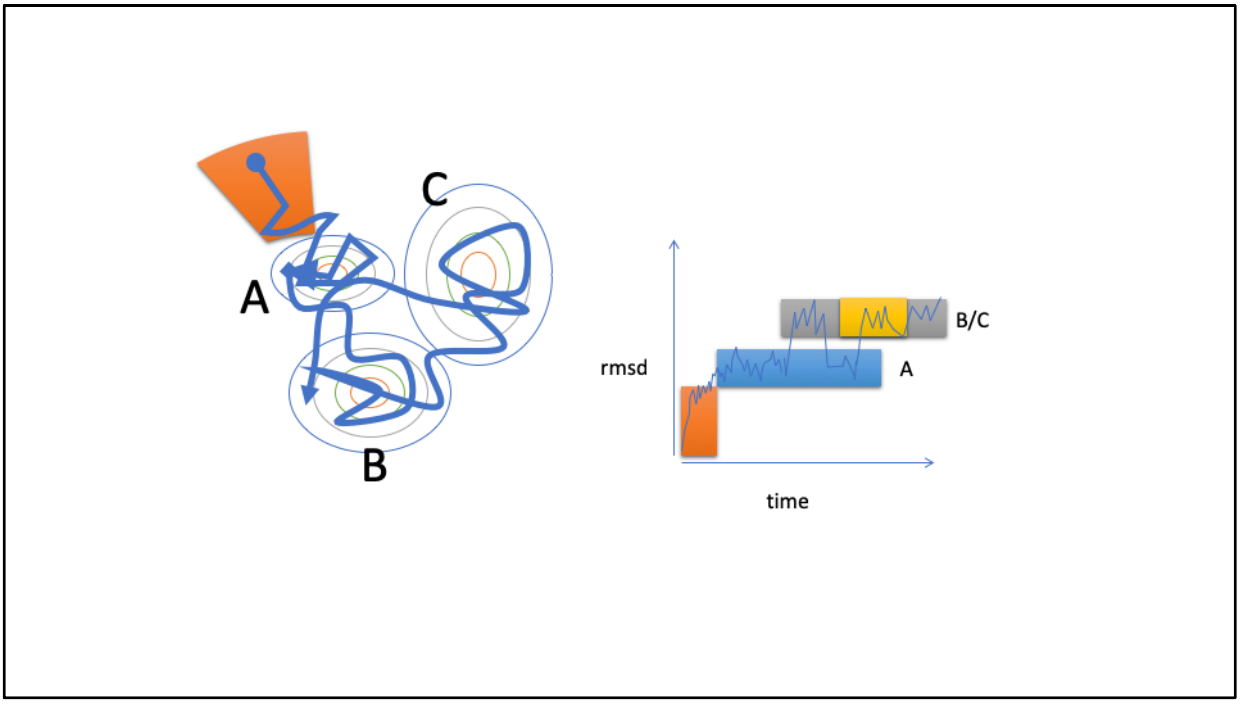


A simple, idealised, equiilibration process. The initial structure starts by ‘falling’, more or less irreversibly, into the potential well, which it then samples. The relationship to an RMSD trace is shown.

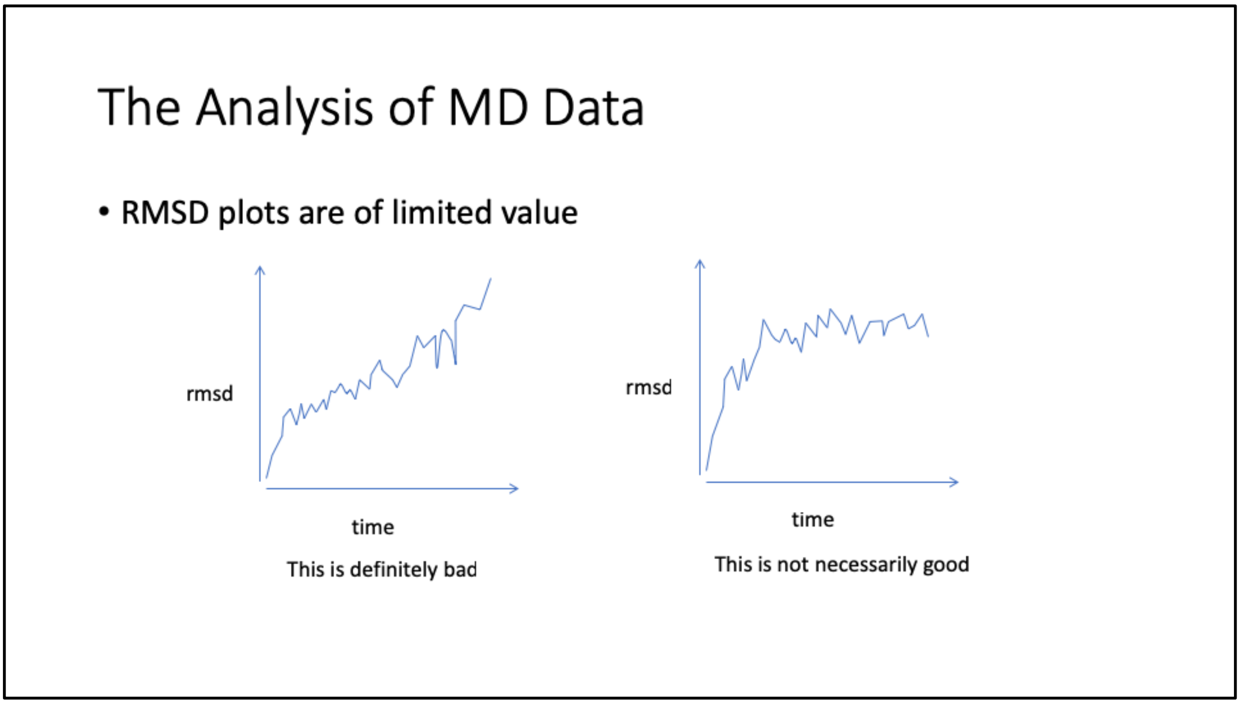


"Jumping Among Minima"

A more realistic model of a simulation, involving "jumping between minima". The plateaus and jumps in the RMSD plot reflect this.



The problem: two different local minima (here B and C) may be an equal ‘distance’ from the reference structure used for the RMSD analysis. The fact that the simulation jumps between basins B and C is not visible.



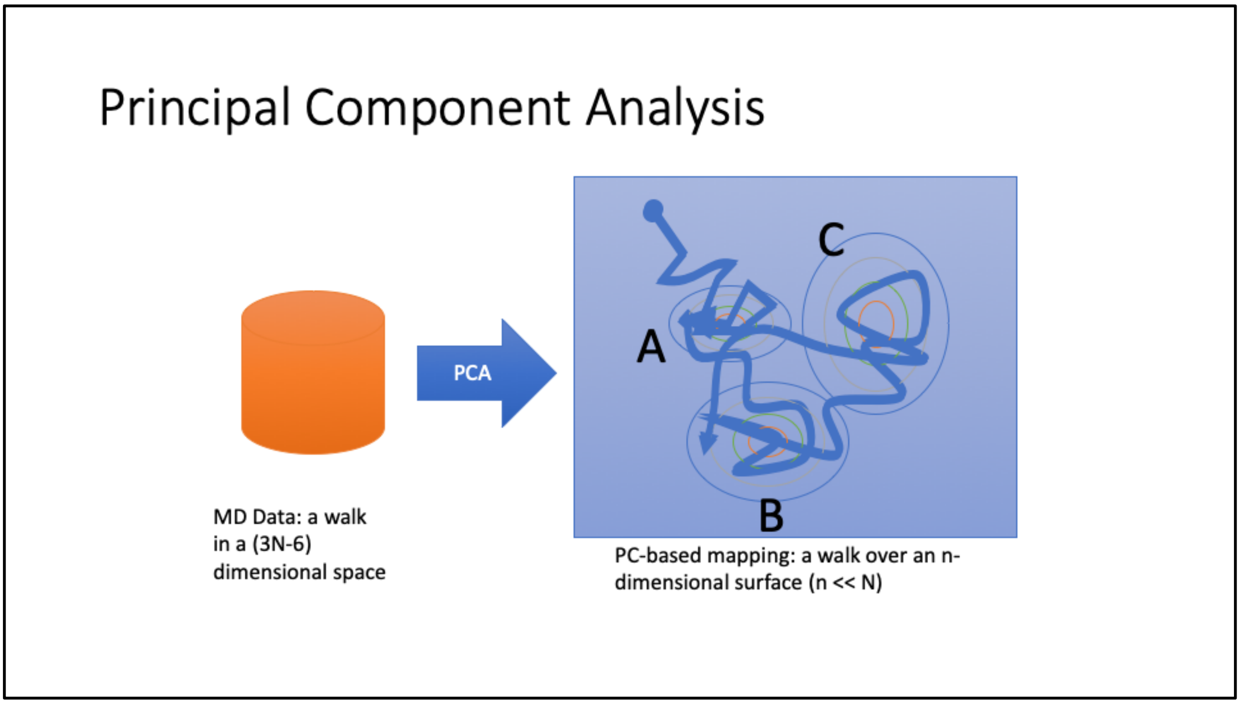
The Analysis of MD Data

- RMSD plots are of limited value

This is definitely bad

This is not necessarily good

The limitation of RMSD analysis to test for relaxation and good sampling of conformational space.

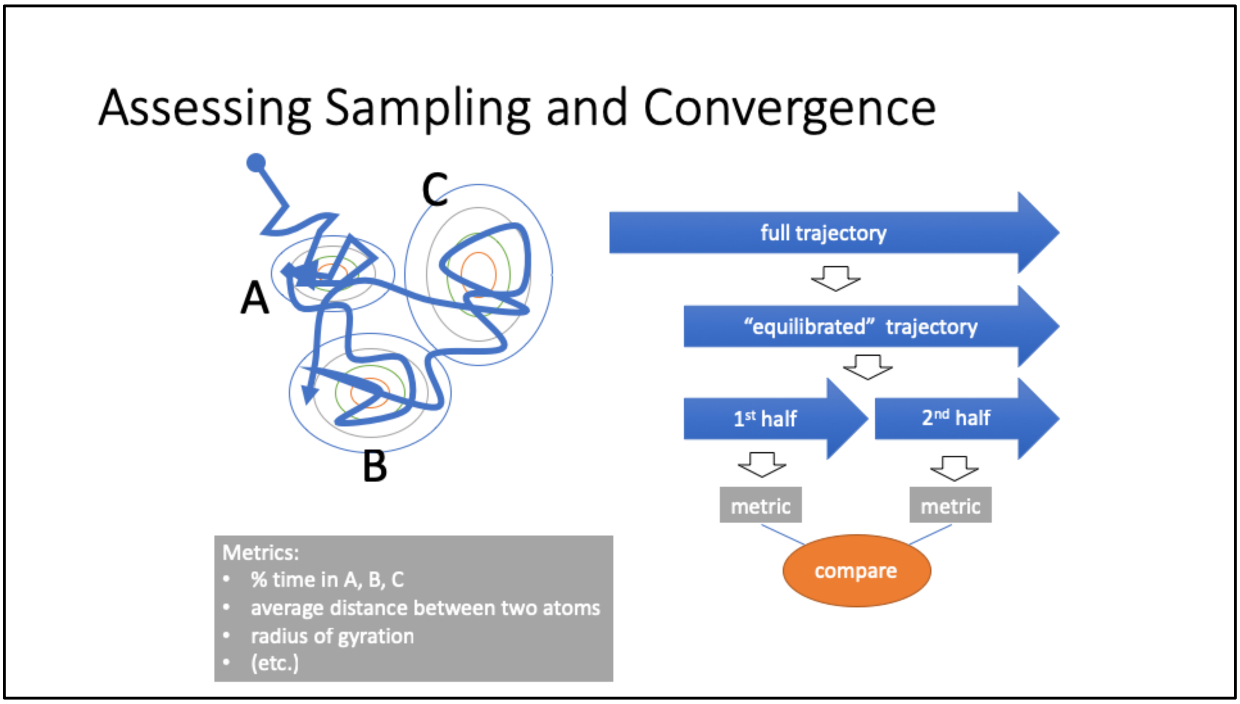


Principal Component Analysis

MD Data: a walk in a (3N-6) dimensional space

PC-based mapping: a walk over an n-dimensional surface (n << N)

The basis of PCA analysis: directly creates a low, but multi-, dimensional view of a trajectory. Much easier to see separate local minimum states.



Assessing Sampling and Convergence

C

A

B

full trajectory

"equilibrated" trajectory

1st half

2nd half

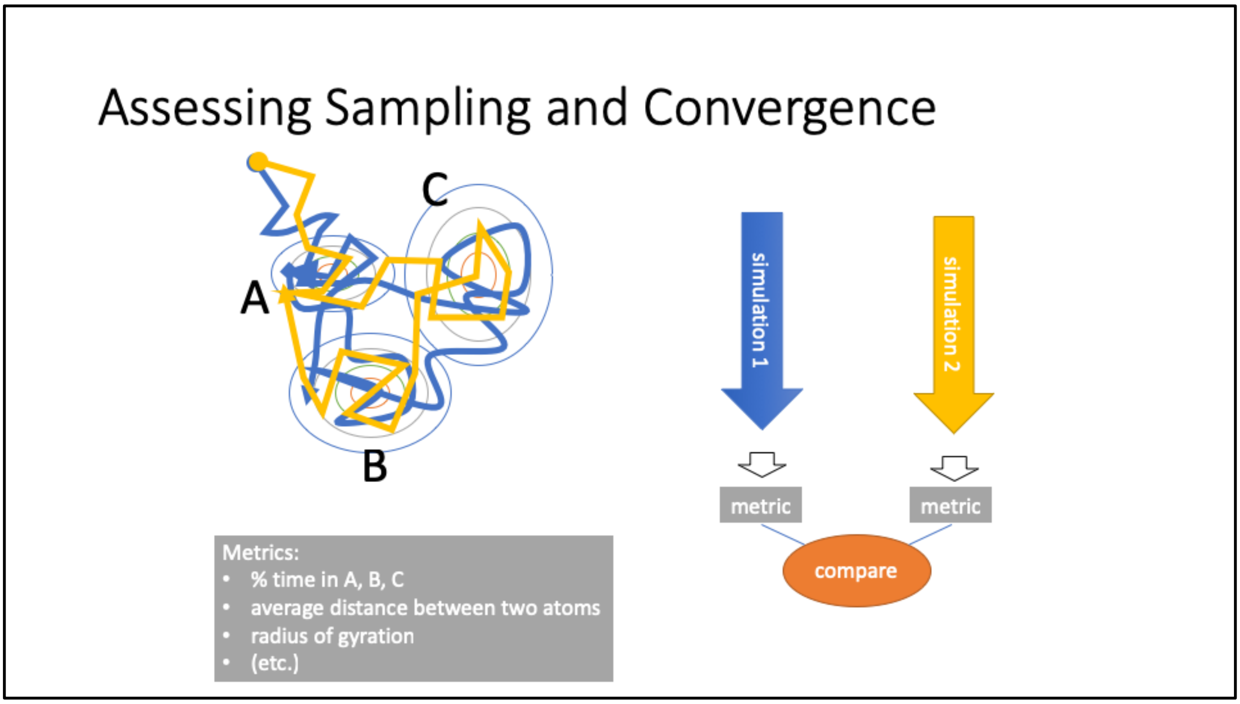
metric

metric

compare

Metrics:
- % time in A, B, C
- average distance between two atoms
- radius of gyration
- (etc.)

One approach (there are many others) to assess convergence. Remove the preliminary, unequilibrated, section of the MD and then split the remaining data in two. If the simulation is well sampled, metrics calculated from the fist half of the data should be in close agreement with those calculated using the second half of the data.



Assessing Sampling and Convergence

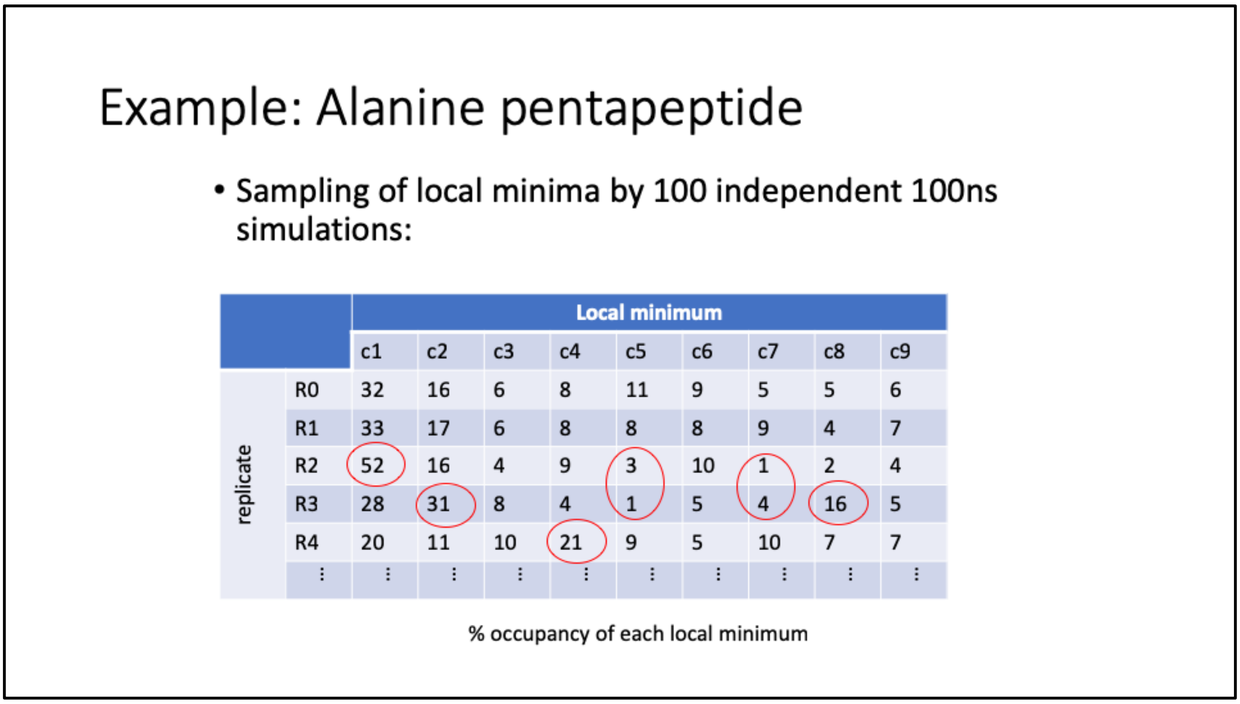
metric

metric

compare

Metrics:
- % time in A, B, C
- average distance between two atoms
- radius of gyration
- (etc.)

Another approach: compare metrics from two (or more) replicate simulations.



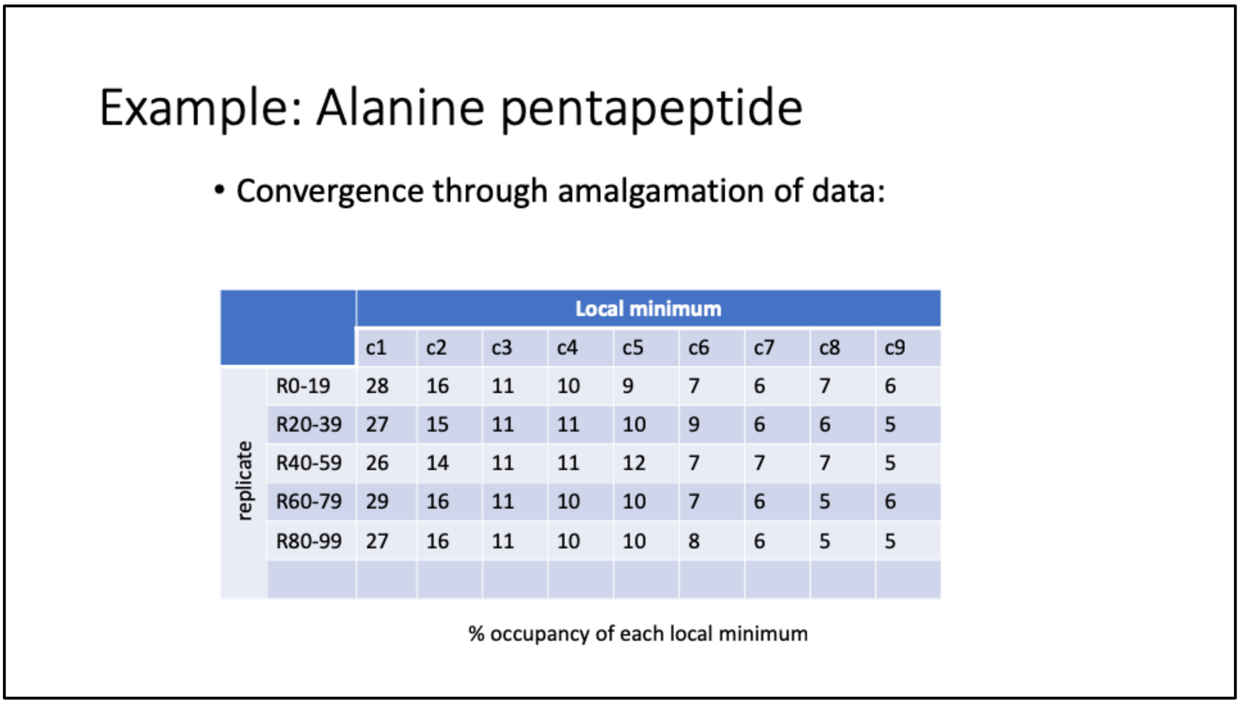
Example: Alanine pentapeptide

- Sampling of local minima by 100 independent 100ns simulations:

		Local minimum								
		c1	c2	c3	c4	c5	c6	c7	c8	c9
replicate	R0	32	16	6	8	11	9	5	5	6
	R1	33	17	6	8	8	8	9	4	7
	R2	52	16	4	9	3	10	1	2	4
	R3	28	31	8	4	1	5	4	16	5
	R4	20	11	10	21	9	5	10	7	7
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

% occupancy of each local minimum

Example data. From 100 independent simulations (R0-R99) of the alanine pentapeptide, nine local minima in the free energy surface are identified (c1-c9). Different replicas sample the different minima to vey different extents. e.g. replica 3 spends twice as much time in c2 than the other simulations shown here, and local minimum c5 is poorly sampled by replicates R2 and R3.



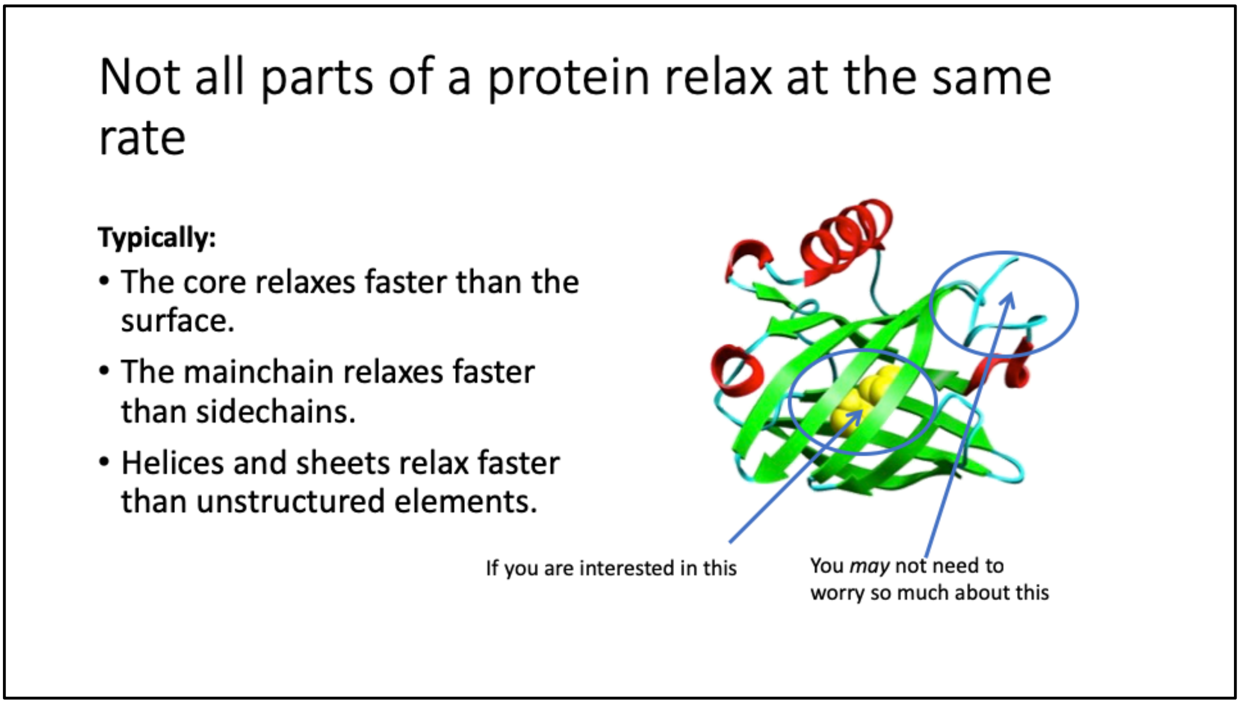
Example: Alanine pentapeptide

- Convergence through amalgamation of data:

		Local minimum								
		c1	c2	c3	c4	c5	c6	c7	c8	c9
replicate	R0-19	28	16	11	10	9	7	6	7	6
	R20-39	27	15	11	11	10	9	6	6	5
	R40-59	26	14	11	11	12	7	7	7	5
	R60-79	29	16	11	10	10	7	6	5	6
	R80-99	27	16	11	10	10	8	6	5	5

% occupancy of each local minimum

By lumping together replicate simulations into bigger ensembles (here 20 at a time), the sampling metrics become much more consistent.



Not all parts of a protein relax at the same rate

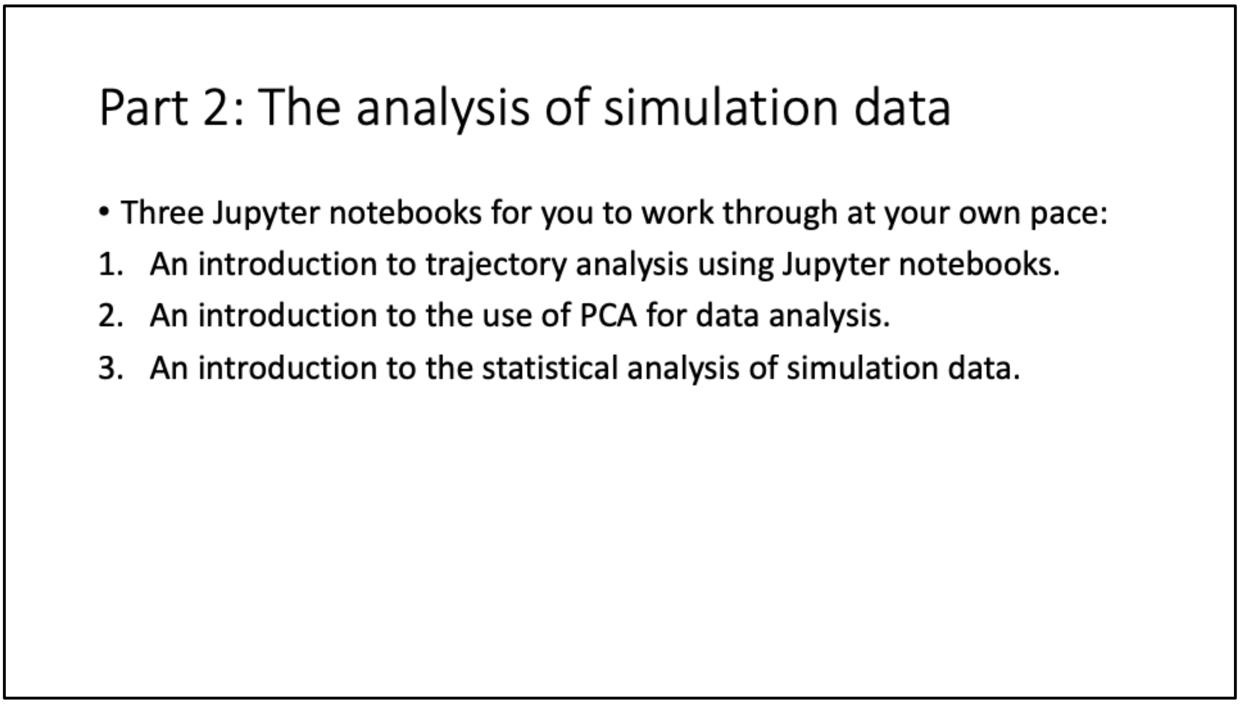
Typically:

- The core relaxes faster than the surface.
- The mainchain relaxes faster than sidechains.
- Helices and sheets relax faster than unstructured elements.

If you are interested in this

You *may* not need to worry so much about this

It is frequently useful to 'zoom in' in your analysis to focus on the key part of the biomolecule – sampling here may be better than over the molecule(s) as a whole.



Part 2: The analysis of simulation data

- Three Jupyter notebooks for you to work through at your own pace:

1. An introduction to trajectory analysis using Jupyter notebooks.
2. An introduction to the use of PCA for data analysis.
3. An introduction to the statistical analysis of simulation data.