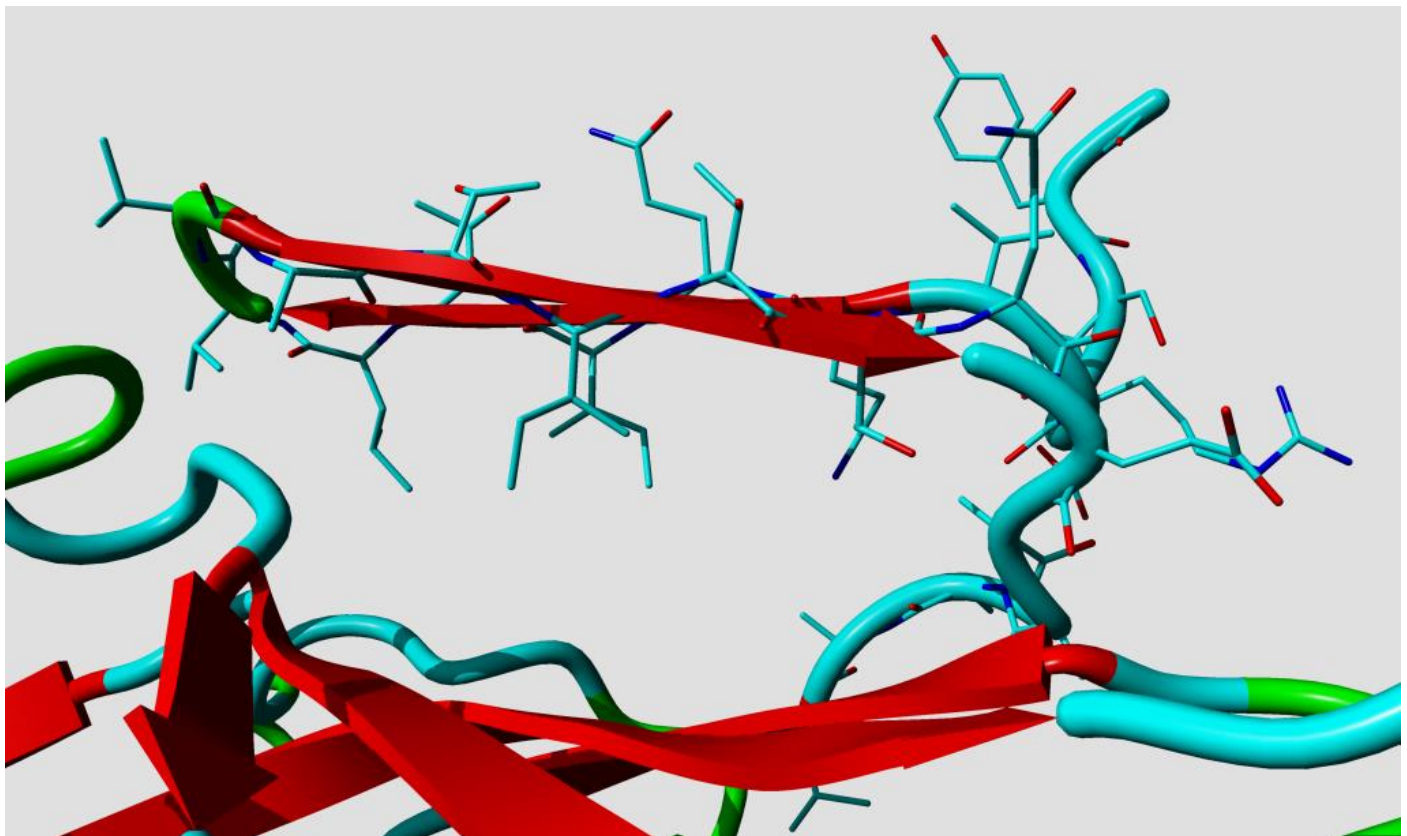
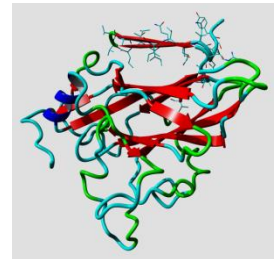


Take your time; three hours really is enough to finish everything. Read the questions well. If questions seem very simple, they probably are. Use common sense. Write legible; it is easy to oversee an error when the answer is written clearly and legibly. Keep the answers short (!) so think before you write; but always explain your answer. You can answer in Dutch, English, or Hochdeutsch. The amount of white space is all you get to answer. Try to not use more space; shorter answers are always better (and take less time to grade). All questions are worth equally many points. **Answers are in red.**

1) To the right you see an overview picture of molecule X, and below you see a blow-up of the N-terminal region. The whole molecule is shown as a ribbon with the full atomic model of the 25 N-terminal amino acids superposed as a stick-model. The sequence of the 25 N-terminal residues is: GYSDRVQQITLVVATITTQEANAV (the whole protein is about 250 amino acids long).



a) Indicate the secondary structure of the N-terminal 25 amino on the sequence:

GYSDRVQQITLVVATITTQEANAV
 -----SSSSSTSSSS-----???

b) These N-terminal 25 amino acids can be mutated easily by protein engineering. If you had to make the whole protein more thermostable by modifying a couple (say maximally 5) amino acids in this N-terminal stretch, what would you do?

Red Q -> I VV in turn -> GP oid. T (that looks like S) -> V.

2) On the next two pages you see 8 pictures of metal ions in eukaryotic proteins with their local surrounding indicated. There are many atoms that you cannot see because they are behind other atoms, but all atoms that you cannot see are what you expect them to be, and all atoms that carry a charge are

visible. The 8 metal ions are: two times Ca, two times K, two times Zn, once Pb, and once Mg. Protein is coloured by atom type, nucleic acids are shown in purple. In the space to the right of each plot:

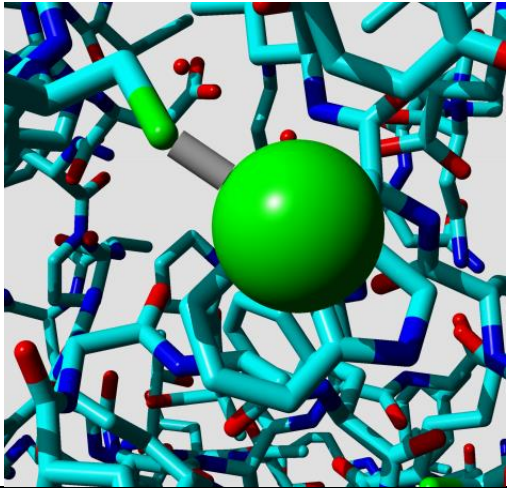
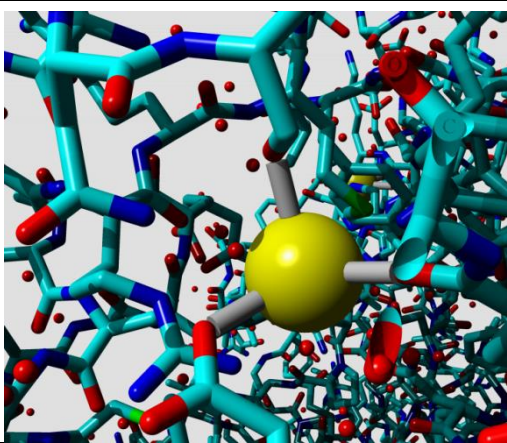
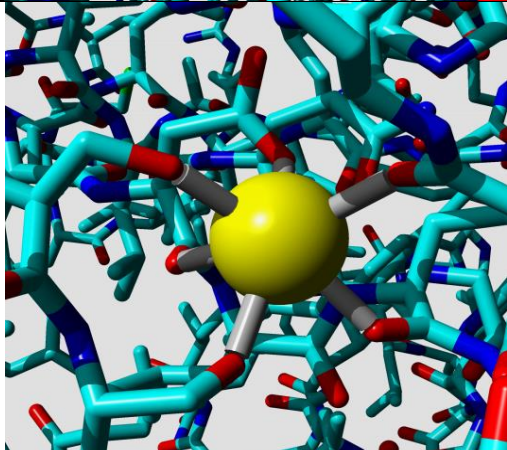
a) Write the ion type;

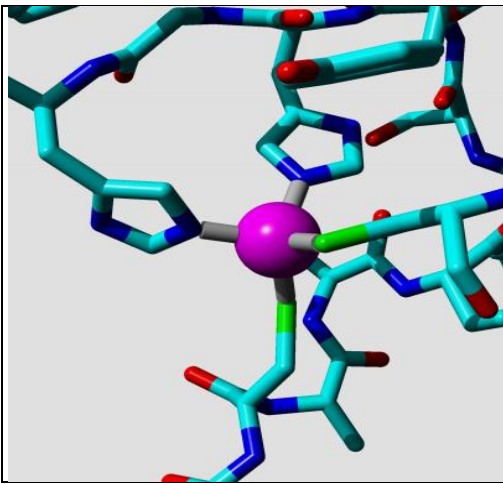
b) Indicate what you think the ion is doing there (crystallisation artefact; part of active site; role in stability, etc.);

c) Indicate if the crystallographer made some (very obvious) error.

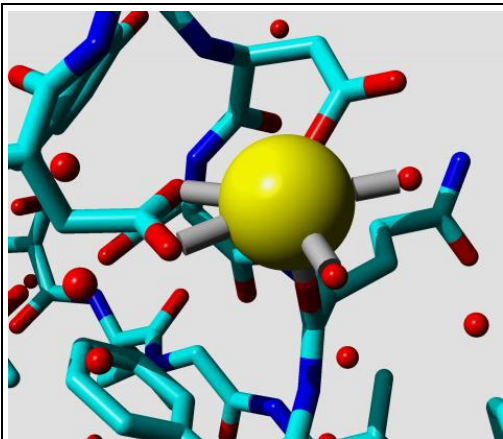
d) Where is this protein located? Cytosol? Extra-cellular? Somewhere else?

Be aware that you cannot give each answer in all cases; just see how far you get. Explain your answer(s)!

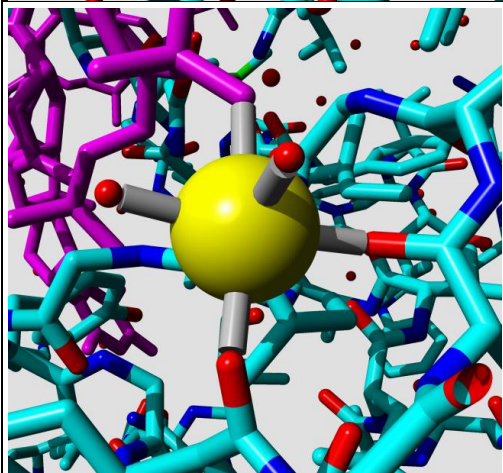
	<p>Pb, artefact. Needed for solving structure. No hints about function or location. There seems nothing else around to bind the Pb, and that is funny, albeit that there could be things in a crystal-neighbouring protein binding.</p>
	<p>K. Surrounding seems incomplete (that is not an error). Might be bound just because it was present in the crystallisation conditions; function hard to determine. (must be 1+ ion as there is only one Asp or Glu present). Not active site as K is never in active site. K indicates that this is a cytosolic protein. The Asn to the left seems to have its O and N the wrong way around.</p>
	<p>K. (must be 1+ ion as there is only one Asp or Glu present). All six liganding atoms seem present, so it might be biologically relevant. Not active site as K is never in active site. K indicates that this is a cytosolic protein.</p>



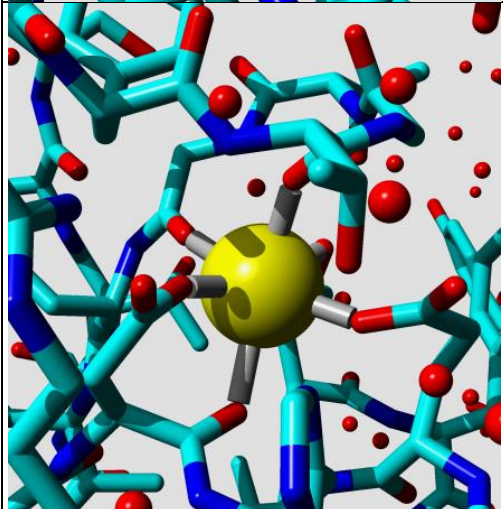
Zn. Surrounded by two His and two Cys, so not likely active site. Stability? Not function in charge relay or so as Zn is always 2+...



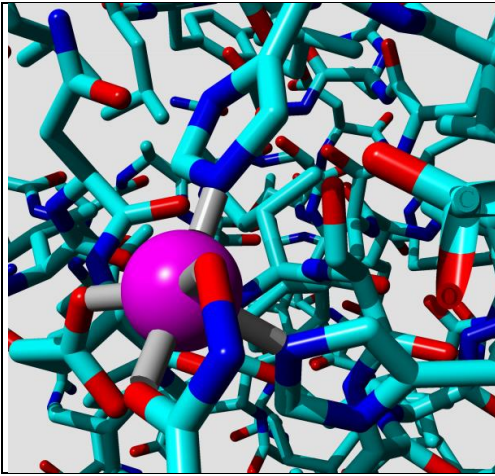
Ca. Surrounded by two Asp, backbone O, and two waters. Calcium means that this is extra-cellular, and its function is most likely to stabilize the protein once it gets outside the cell.



Mg. Because it sits between protein and DNA. This is likely to happen in the nucleus. Its 'functional role' is to stabilize the protein-DNA interaction. It seems to have all six ligands, but I am missing one charge!



Ca. Surrounded by one Asp one glu, 3 backbone O, and one undetermined O in the back. Calcium means that this is extra-cellular, and its function is most likely to stabilize the protein once it gets outside the cell.



Zn. Surrounded by one his, one asp or glu, and one 'funny thing' that binds with an O and an O that seems bound to an N. This is probably a modified N-terminus...
Not sure of course, but this seems likely an active site.

3) Lipinski invented, a long time ago, his famous 'rule of five'. The original rule (it got modified a bit over the years) implies that good ligands to try as a drug in a drug design project will have a series of properties:

- A. a molecular weight less than 500
- B. fewer than 5 hydrogen-bond donors and fewer than 5 hydrogen-bond acceptors
- C. fewer than 5 internal degrees of freedom
- D. logP below 5 (which means that the compound must be rather hydrophobic).
- E. There must be fewer than 5 hydrogen-bond donors and fewer than 5 acceptors, but why would it be bad to have no hydrogen-bond donors and acceptors at all?

Now that you went through the whole SFB course, do you think this comes as a surprise? Let me help you, I think it does not come as a surprise. But why not? Can you explain for these four properties (A-D) why none of them is a surprise to you? (Some of the answers might overlap a bit...). Sub-question E might be a bit difficult...

A) The simple answer is that most medicines need to replace (sit in the pocket of) an endogenous ligand. And those aren't very big either. If the medicine gets too big, it will either stick out in the solvent or bump into something.

B) When the ligand binds, it loses a lot of entropy (its own 6-dimensional motion, and probably some internal degrees of freedom). This must be compensated, and that goes best by the entropy of water, both the waters in the pocket, and the waters that are unhappy around the hydrophobic parts of the ligand.

C) Every degree of freedom is lost upon binding. Too many of that might be so much energy loss that the entropy of water (see B) can no longer compensate for it.

D) See B.

E) You need some H-bonding to get specificity in the binding. Not so much for the medicine, but for the natural ligand that is 'replaced' by the medicine.

4) a) What is the positive-in rule?

When (membrane) proteins are produced by ribosome they first move to ER membrane. Translocon helps put it there is bad at crossing positive charge through ER membrane, so positive charge stays outside, in cytosol. Golgi buds off from ER and moves to outer membrane. Golgi merging with outer membrane leaves cytosolic side cytosolic side, so outside of ER becomes inside of outer membrane. Bacteria are much simpler, but seem to have a similar 'problem' when putting helices through (outer) membrane.

b) What can we do with this rule?

Predicting TM helices in a sequence is easy, but predicting in/out side is difficult. This rule helps with that.

c) The common household bacterium *bacterius dirtyus* can kill a person by causing an immune overreaction. The small bacterial peptide:

AGNYLHSPGPAGYAALAAYMFLILVLPVSFLTLYVKLQ**HKKPRT**PLNYILLLLAVAILFMVLAGFLALMYTSM

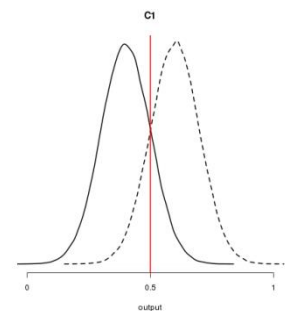
is known to elicit this immune response, and it is known that a histidine in this peptide is crucially important for this response. The company Vaccinus Inc wants to make a vaccine against *bacterius dirtyus* by injecting a goat with a dodecameric peptide. Which peptide is the most likely candidate?

I coloured (roughly) the two TM helices red. The little loop in-between (blue) is rather positive, and thus likely in the cytosol of the bacterium and will not be seen by any antibody. So it should be either the N-terminal, or the C-terminal loop, and the latter is way too short. So I suggest the underlined peptide (or take 1 more or fewer).

I am always surprised when people do not think about predicting TM helices in question c after answering a and b (more or less) correctly.

5) Suppose you have used some machine learning technique and created a two-class classifier for a bioinformatics classification problem, in which we have sequences that elicit a response (solid line) and sequences that don't (dashed line).

a) Given the decision threshold, indicate in the figure the true negative, true positive, false negative, and false positive fractions by writing in the figure TP, FP, TN, and FN at logical places. From left to right you see TP, FP, FN, and TN.



b) Macromolecular structure validation software 'criticizes' structures solved by crystallographers and NMR spectroscopists. Suppose the output of a structure checking algorithm is shown in the figure above. Do you think the decision threshold needs to be set differently? Why (not) and how (not)? Ps, I can imagine that you can equally well defend that the red line must move to the right as that it should move to the left. Do you have equally much imagination as I have?

Obviously, the line should stay where it is to maximise the chance of a correct answer. But, ... When validating somebody else's software or data, you might want to minimize the number of false negatives (= incorrect error messages). But when we are re-refining structures, or looking at our own work, then we want the software to point out anything and everything that might be wrong.

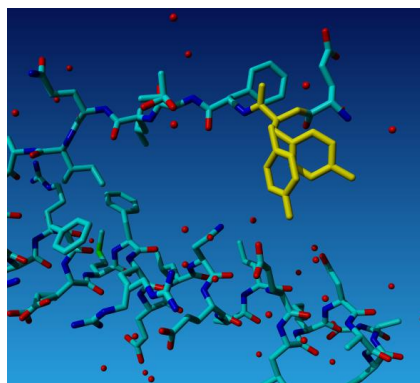
Obviously, there are many other lines of reasoning that might give you points. Actually, this was not a very good question as it was hard to call many answers wrong (partly because I am too nice...).

6) Question 5 actually described a deep-learning neural-network experiment to predict transmembrane helices in bacterial proteins. That works very well, but once we have the results it is very difficult to learn something from those results. I would therefore like to repeat this work the classical way, i.e., using a force field.

a) in less than 20 words, tell me what is a force field (in bioinformatics)?

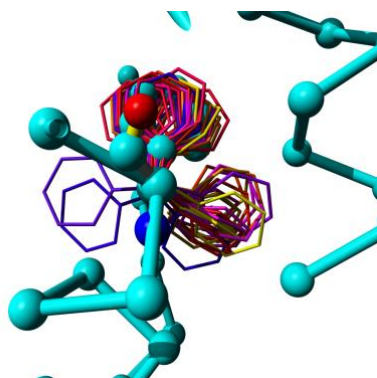
b) Use the scratch paper labelled "Question 6" to tell me in great detail how you would set-up such a force field based computational method. (There are two pages labelled "Question 6", use just one, and scratch out the other).

7) Some questions about side chain conformations



a) In this picture the tyrosine in the top right seems to have two side chains. Can you explain why that is?

Both conformations are observed in a fraction of the molecules in the crystal. The final structure is the 'average' of all conformations in the crystal.



b) The rotamer distribution is shown for Phe at position 38 in PDB-file 6TBP (Phe 38 itself is also shown as a ball-and-stick model). What can we learn from this picture?

We see that the three primary rotamers (i.e. the three that have different chi-1 angles) are observed to different extents. The Phe-38 itself sits in the most populated rotamer cloud. So, probably, we can conclude that there is nothing to worry about.



c) Whose tombstone is this? Boltzmann

d) Which formula is printed above his head?

The Boltzmann equation: Entropy = $k \cdot \log W$

e) What is the rule of 10?

1 kCal/Mol gives about a factor 10 shift in equilibrium

e) How do we get from this formula to that rule of 10?

fill everything in (with $K=10$) in $\Delta G = -RT \ln(K)$

f) How is the rule of ten related to the questions 8a and 8b?

8a) Both positions are (perhaps only roughly, occupancies are not given) energetically equally likely in crystal, so both show up. YASARA then shows both possibilities at the same time. 8b) A rotamer density (=number of observed rotamers in same cloud) difference of a factor 10 indicates an energy advantage of 1 kCal/Mol for the more densely populated rotamer.

Funny how many people answered d) with $S = -k \cdot \log(W)$, which is the correct answer, but didn't bring them many points...

8) When we want to make a protein more stable, we can use the concepts *entropic* stabilisation and *enthalpic* stabilisation.

a) Explain what is meant with these two terms. b) What are the differences, and what do they have in common?

Entropic when you aim at reduction of mobility of U so Gly->X or X-Pro, or introduction Cys-Cys bond;
Enthalpic when you introduce interactions (H-bond, salt-bridge, capping, etc). In common that it is best done at surface, new residue must fit, if Gly or Pro involved backbone angles should be checked.
Differences are many...

b) I want to make my protein more stable, so I decide to mutate a very exposed isoleucine into an aspartic acid, to make the surface of my protein more hydrophilic. Explain why this is a stupid plan.

Same differences in folded and unfolded form...

c) What is helix capping and how can it be used to make a protein more stable by mutagenesis? And why has this method proven so successful in practice?

Interaction between charge and helix dipole. So good because helix doesn't exist in unfolded form so it is guaranteed gained in F.

d) When working on increasing the stability of my protein by mutagenesis, I often consult a table that holds all the (backbone) torsion angles. Why?

New residue must fit. Especially important for the often applied Gly->X or X-Pro.

9) Describe in at most ten words per term what the term is/means:

Feel free to look up the answers in the canon

Z-score

B-factor

R-factor

Force Field

MD

BLAST

Homology Modelling

PDB

CSD

Salt Bridge

Bond angle

Torsion angle

NMR

SCOP, CATH, DALI

HSSP

DSSP

Occupancy

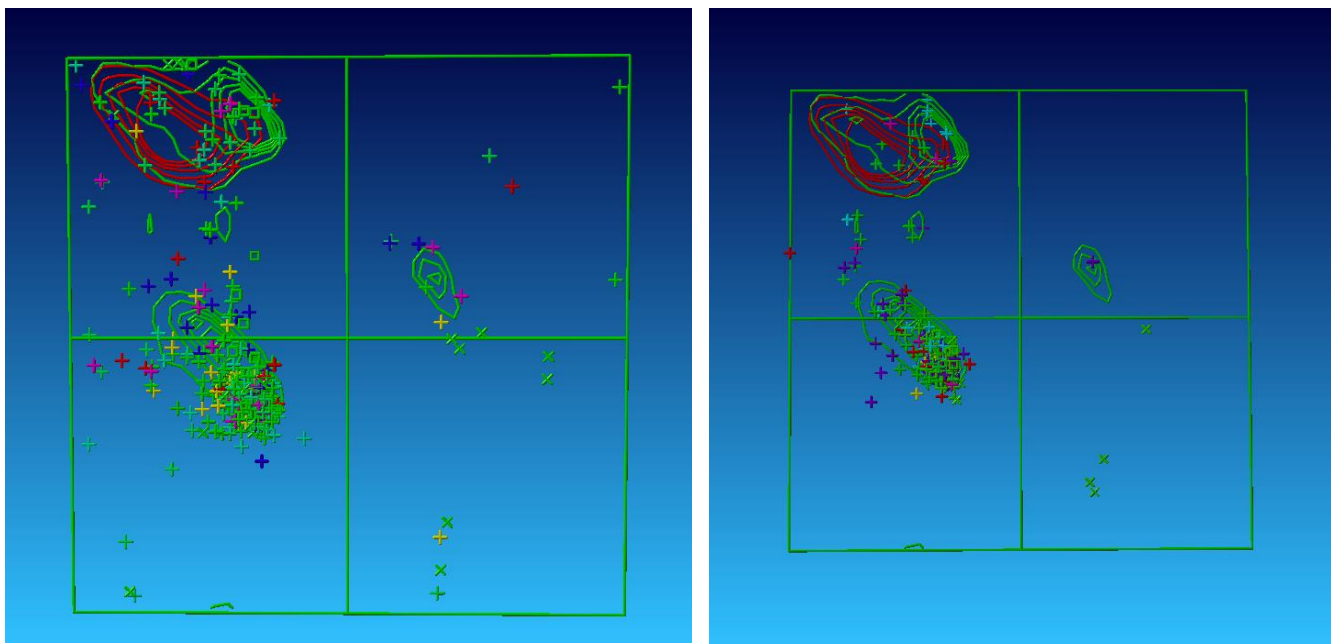
Resolution

NOE

Crystal packing artefact

MD Time step

10) Below you see two times two Ramachandran plots.



a) The top two are for bovine rhodopsin, and haemoglobin. The residues are represented by crosses or squares and coloured by characteristic (positive=blue; polar=purple; negative=red; hydrophobic=yellow/green/light-blue). Unfortunately I forgot which one is which. Can you figure that out?

Haemoglobin and rhodopsin both have 7 helices, so that doesn't help/ But, rhodopsin is a membrane protein that, thus, has lower resolution and thus a more chaotic picture. Left rhodopsin, thus.

