Bioinformatics Final Assignment

Fabian Flores Gobet

Universitat della Svizzera Italiana - USI

MSc in Artificial Intelligence

16 January 2024

Table of Contents

TEMPLATE SELECTION

UNIPROT

For this project, the protein I am trying to model is the *GPCR* proto-oncogene MAS receptor, also known as **MAS1** receptor. Information regarding said protein, namely the amino acid sequence can be found in the *UniProtKB* database with entry *ID* **P04201**.

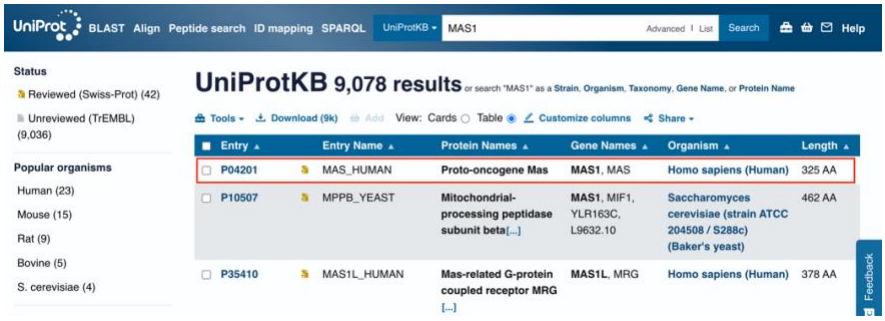


Figure 1: UniProtKB search results

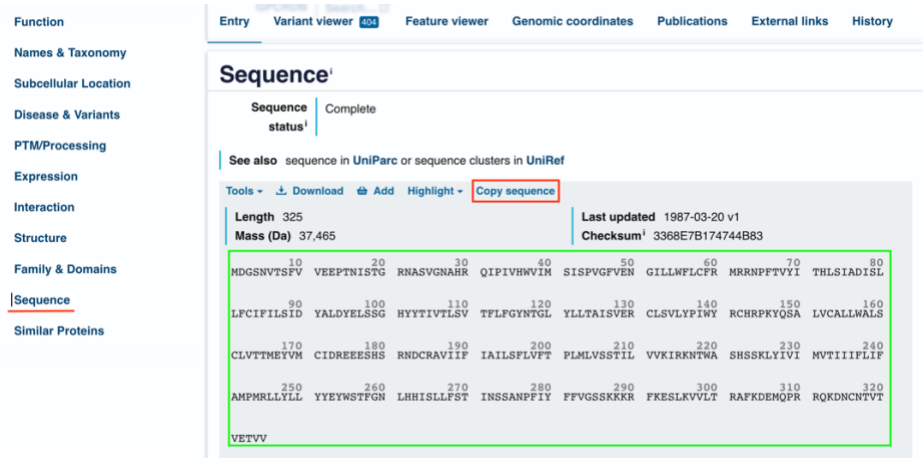


Figure 2: Amino acid sequence of P04201 (MAS1)[1]

The found sequence for MAS1 is:

M D G S N V T S F V V E E P T N I S T G R N A S V G N A H R Q I P I V H W V I M S I S P V G F V E N G I L L W F L C F R M R R N P F T V Y I T
H L S I A D I S L L F C I F I L S I D Y A L D Y E L S S G H Y Y T I V T L S V T F L F G Y N T G L Y L L T A I S V E R C L S V L Y P I W Y R C H R P K Y
Q S A L V C A L L W A L S C L V T T M E Y V M C I D R E E E S H S R N D C R A V I I F I A I L S F L V F T P L M L V S S T I L V V K I R K N T W A S
H S S K L Y I V I M V T I I I F L I F A M P M R L L Y L L Y Y E Y W S T F G N L H H I S L L F S T I N S S A N P F I Y F F V G S S K K K R F K E S L K V
V L T R A F K D E M Q P R R Q K D N C N T V T V E T V V

[1] https://www.uniprot.org/uniprotkb/P04201/entry

PSI-BLAST

Important information that should be considered can also be seen on the former webpage. We can see that the sequence is composed by **325 amino acids**, belonging to ***Homo Sapiens* species** (taxonomic identifier 9606 *NCBI*). Furthermore, it is a *GPCR* and acts as a receptor for angiotensin-(1-7).

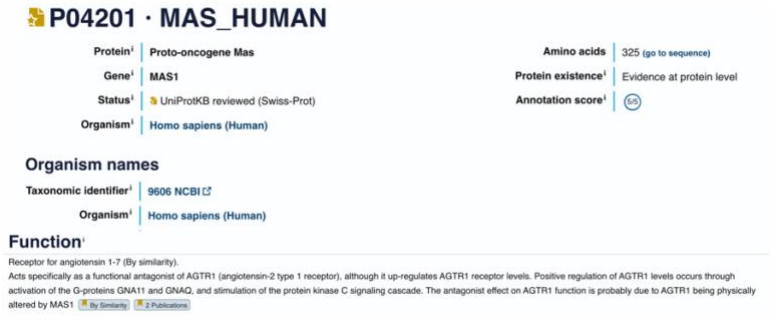


Figure 3: Additional information about P04201

I copy the sequence using *UniProtKB* '*Copy sequence'* functionality and head over *PSI-BLAST* webserver[2] from the National Library of Medicine (*NCBI*). Then, I proceed to paste the sequence into *'Query sequence'* box, select the Protein Data Bank (*PDB*) as a standard search set and click *BLAST*. By default, *PSI-BLAST* is set in program selection.

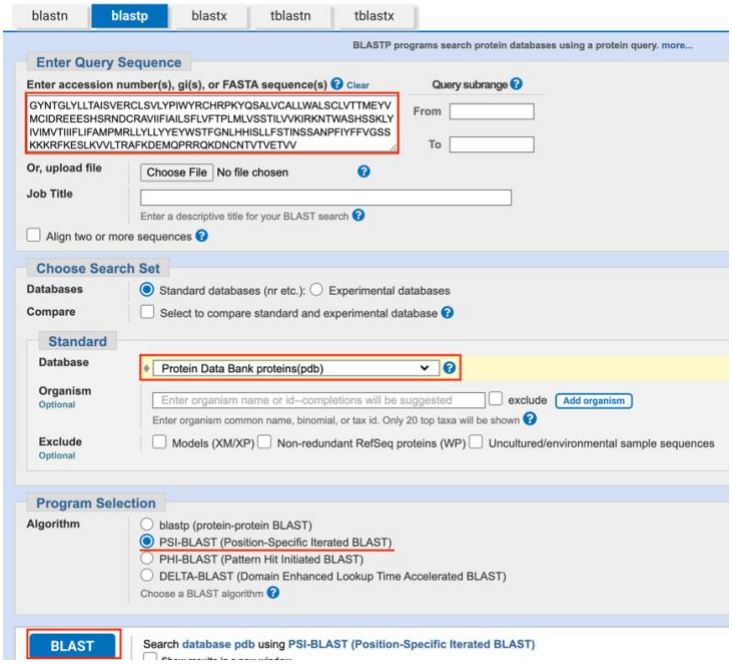


Figure 4: PSI-BLAST inputs

[2] https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROGRAM=blastp&BLAST_PROGRAMS=psiBlast

Before proceeding into the results and consideration of further *PSI-BLAST* iterations, some filtering criteria must be established, namely the percentages of query coverage and identity, the species and the type of protein (chimera/wild-type).

- **Query coverage percentage**: Ideally, the query coverage percentage should be at least 90%.
- **Identity percentage**: since the sequence is composed of 325 amino acids, according to Doolittle's Rule of Thumb it is safe to consider templates in which the identity percentage is strictly above 20%.

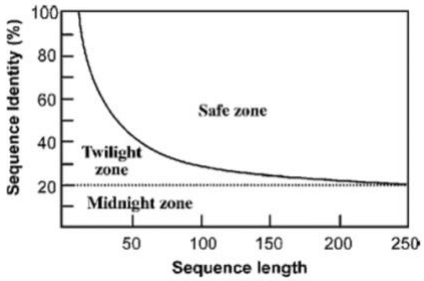


Figure 5: Doolittle's Rule for homology modelling

- **Species**: the target species is *Homo Sapiens*, as previously observed in *UniProtKB*.
- **Type of protein**: some proteins come labelled as *Homo Sapiens*, but they are chimeric in nature. To assess this, once the initial selection of our results in *PSI-BLAST* is done, I must go over each, consulting the available data in *NCBI* and *PDB* to determine if the target chain in chimeric or not.

The most restrictive criteria here is the identity percentage. Thus, on each iteration I will sort the results in descending order of identity percentage and see if there is at least one that fits the criteria.

The first iteration of *PSI-BLAST* results in a total of 112 sequences. After sorting in descending order of percentage of identity I can see that there are already some sequences that fit the criteria, making the process eligible for a second iteration exploration.

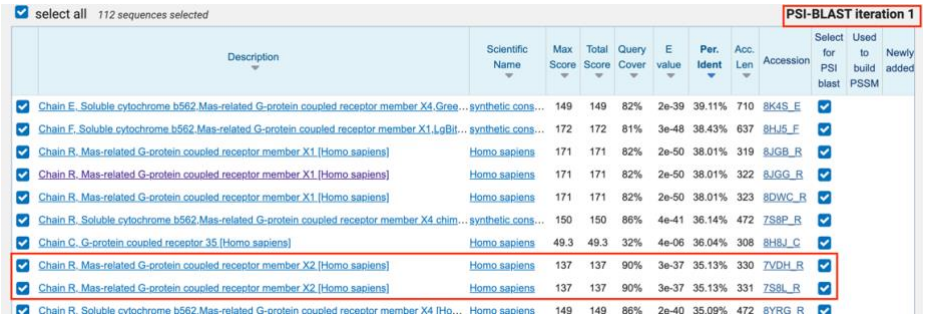


Figure 6: PSI-BLAST 1st iteration

The second iteration of *PSI-BLAST* results in 498 sequences (not all newly added). I need to assess if any of the new added sequences fit my criteria. To retrieve an honourable picture for this report of just the newly added sequences (since this filtering functionality is not implemented in the webserver), I inspect the page source code and run a *JS* command to remove such sequences.

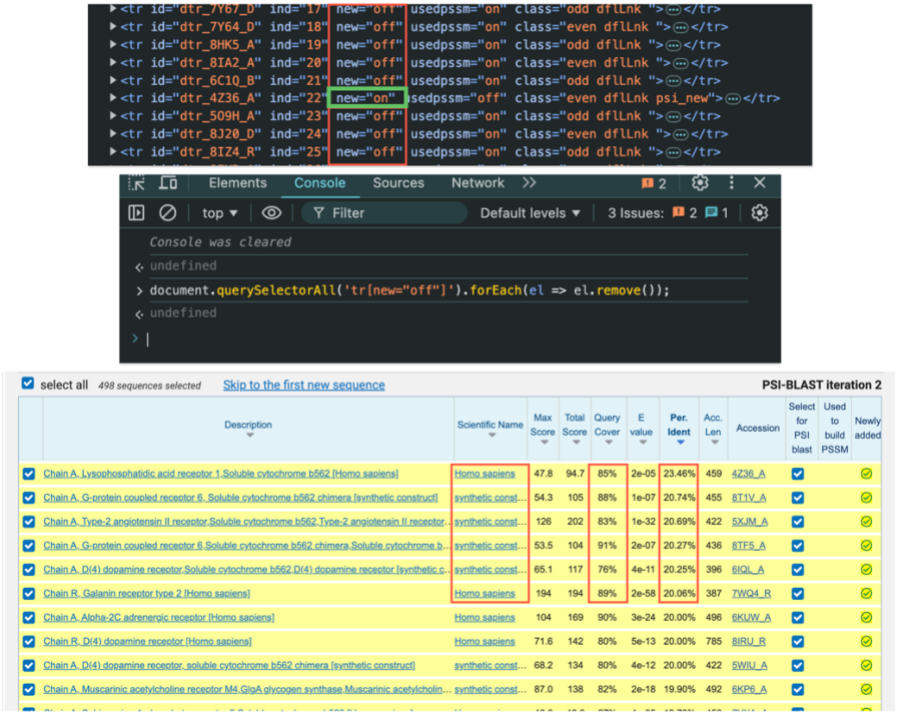


Figure 7: First picture shows the pattern and flags I must look for. Second picture denotes the console and the command[3] to remove non new sequences. Third picture shows the result of running the command.

From the third picture of *Figure 7* we can see that newly added sequences within the identity percentage threshold don't fit the criteria due to either being a synthetic construct or not having enough query coverage. This analysis then concludes that our sequences are to be considered from the first iteration of *PSI-BLAST*. Therefore, we revert to the first iteration and apply the filtering criteria using the webserver's tool.

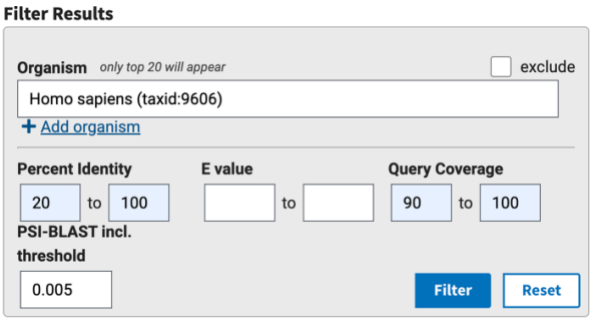
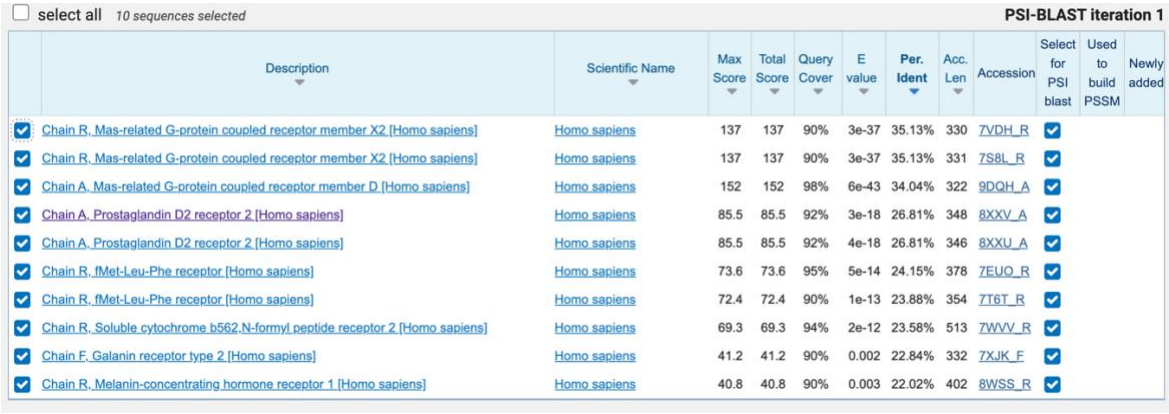


Figure 8: Filters applied to 1st iteration of PSI-BLAST

[3] *document.querySelectorAll('tr[new="off"]').forEach(el => el.remove());*

The later procedure results in a total of 10 found sequences, from which now I must select those with unique features, e.g. one protein per template per subtype, whilst also searching the respective information in *NCBI* and *PDB* to make sure tour target chains are not chimeric in nature. It's worth noting that now I have not decided the state of the target homology model (apo/halo, active/inactive) as this decision will depend upon the results over cross validation with *HHPRED*.



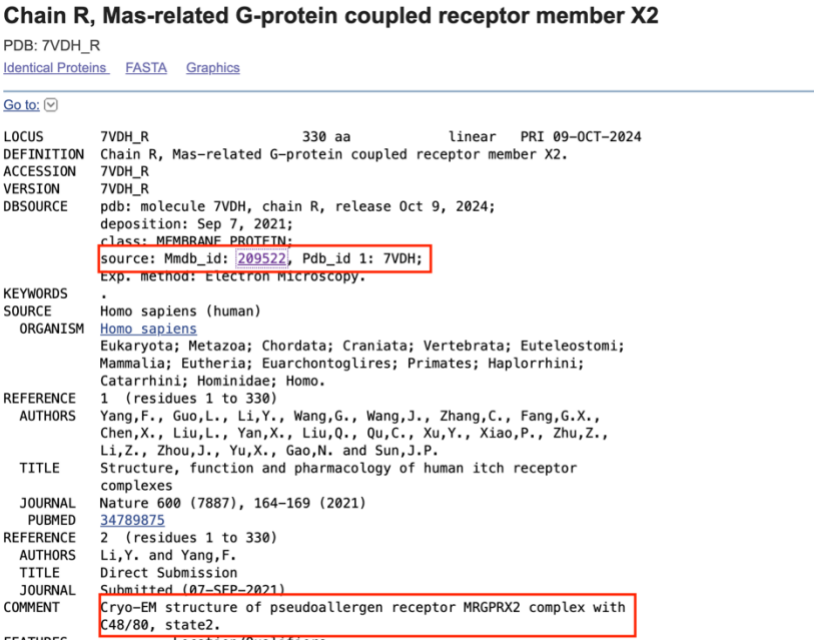
select all *10 sequences selected* **PSI-BLAST iteration 1**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	330	7VDH_R	☑		
Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	331	7S8L_R	☑		
Chain A, Mas-related G-protein coupled receptor member D [Homo sapiens]	Homo sapiens	152	152	98%	6e-43	34.04%	322	9DQH_A	☑		
Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	85.5	85.5	92%	3e-18	26.81%	348	8XXV_A	☑		
Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	85.5	85.5	92%	4e-18	26.81%	346	8XXU_A	☑		
Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	73.6	73.6	95%	5e-14	24.15%	378	7EUO_R	☑		
Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	72.4	72.4	90%	1e-13	23.88%	354	7T6T_R	☑		
Chain R, Soluble cytochrome b562,N-formyl peptide receptor 2 [Homo sapiens]	Homo sapiens	69.3	69.3	94%	2e-12	23.58%	513	7WVV_R	☑		
Chain F, Galanin receptor type 2 [Homo sapiens]	Homo sapiens	41.2	41.2	90%	0.002	22.84%	332	7XJK_F	☑		
Chain R, Melanin-concentrating hormone receptor 1 [Homo sapiens]	Homo sapiens	40.8	40.8	90%	0.003	22.02%	402	8WSS_R	☑		

Figure 9: PSI-BLAST pre-analysis result.

Decidability on sequences of the same subtype

7VDH_R and *7S8L_R* have the same query coverage and identity metrics. By clicking on the accession hyperlink, I can get access to more information regarding each conformation. Regarding *7VDH_R*, this template represents the Cryo-EM structure of MRGPRX2 complex with C48/80 (agonist).



Chain R, Mas-related G-protein coupled receptor member X2

PDB: 7VDH_R

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       7VDH_R                330 aa            linear   PRI 09-OCT-2024
DEFINITION  Chain R, Mas-related G-protein coupled receptor member X2.
ACCESSION   7VDH_R
VERSION     7VDH_R
DBSOURCE    pdb: molecule 7VDH, chain R, release Oct 9, 2024;
            deposition: Sep 7, 2021;
            class: MEMBRANE PROTEIN;
            source: Mmdb_id: 209522, Pdb_id 1: 7VDH;
            Exp. method: Electron Microscopy.
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 330)
  AUTHORS   Yang,F., Guo,L., Li,Y., Wang,G., Wang,J., Zhang,C., Fang,G.X.,
            Chen,X., Liu,L., Yan,X., Liu,Q., Qu,C., Xu,Y., Xiao,P., Zhu,Z.,
            Li,Z., Zhou,J., Yu,X., Gao,N. and Sun,J.P.
  TITLE     Structure, function and pharmacology of human itch receptor
            complexes
  JOURNAL   Nature 600 (7887), 164-169 (2021)
   PUBMED   34789875
REFERENCE   2  (residues 1 to 330)
  AUTHORS   Li,Y. and Yang,F.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-SEP-2021)
COMMENT     Cryo-EM structure of pseudoallergen receptor MRGPRX2 complex with
            C48/80, state2.
```

Figure 10: https://www.ncbi.nlm.nih.gov/protein/7VDH_R?report=genbank&log$=prottop&blast_rank=1&RID=SCSKFR93016

To get more information, I can access the Molecular Model Data Base (*MMDB*) with the *ID* 209522 (depicted in *Figure 10*), retrieving the resolution of the experimental method, the citation to the paper relative to this structure, the molecular components and the link to the respective *PDB* page.

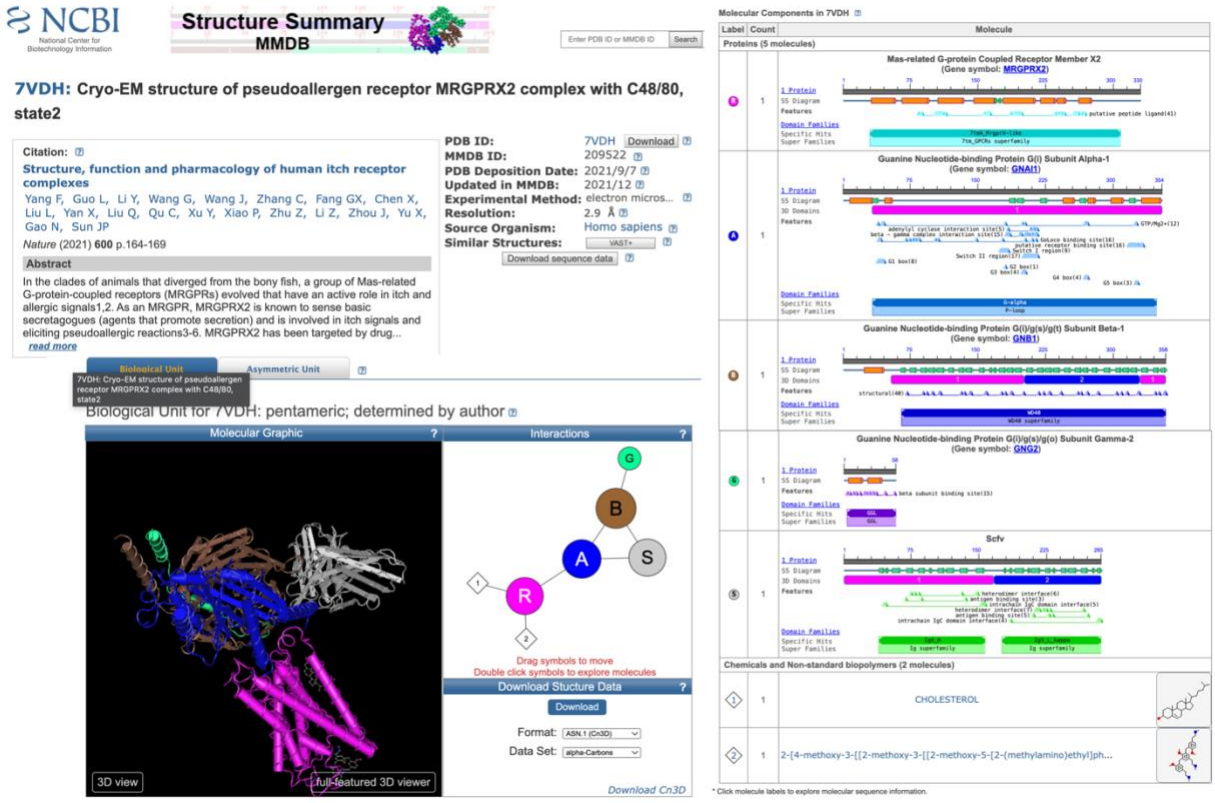


Figure 11: https://www.ncbi.nlm.nih.gov/Structure/pdb/7VDH

The same procedure can be applied to *7S8L_R*, revealing it is the *CryoEM* structure of *Gq-coupled MRGPRX2* with peptide agonist *Cortistatin-14*.

7VDH_R has a resolution of 2.9 Å, whereas *7S8L_R* has a resolution of 2.45 Å. Both structures represent *holo* states in active conformations. When looking closer at the molecule and interaction panels of 7VDH_R, it is not clear where the C48/80 agonist ligand is, whereas for *7S8L_R* this is very clear. All this information can be further validated in the respective *PDB* entry.

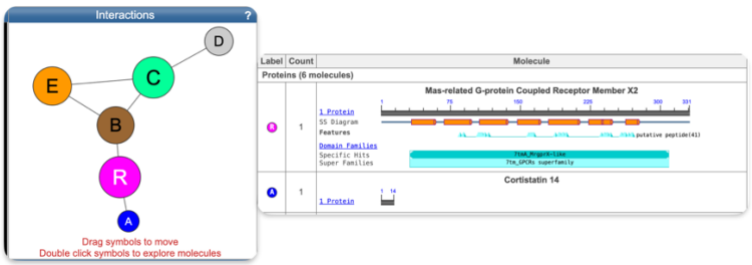


Figure 12: 7S8L_R information on MMDB.

Since the resolution of *7S8L_R* is more favorable, and the availability of structural information is clearer, I decided to rule out *7VDH_R* and keep *7S8L_R*.

Regarding *8XXV_A* and *8XXU_A,* by similar analysis to the prior case, I found that the only key differences were an increase in 2 amino acids for *8XXV_A* and a lower resolution for *8XXV_A*. All other analytical components appeared to be equal. Given this, the choice I made was to keep *8XXV_A* due to better resolution.

Regarding *7EUO_R* and *7T6T_R*, we can see in *PSI-BLAST* that the former has better query coverage and identity percentage than the latter. Hence the choice here is simpler than the other cases, and I keep *7EUO_R*.

Decidability for all sequences

For all non-discarded sequences, I must ensure that the target chain is non chimeric even though the species are labeled as *Homo Sapiens*.

To do this, for each of the sequences I go to PDB and see if the target chain is composed by a fusion of structures.

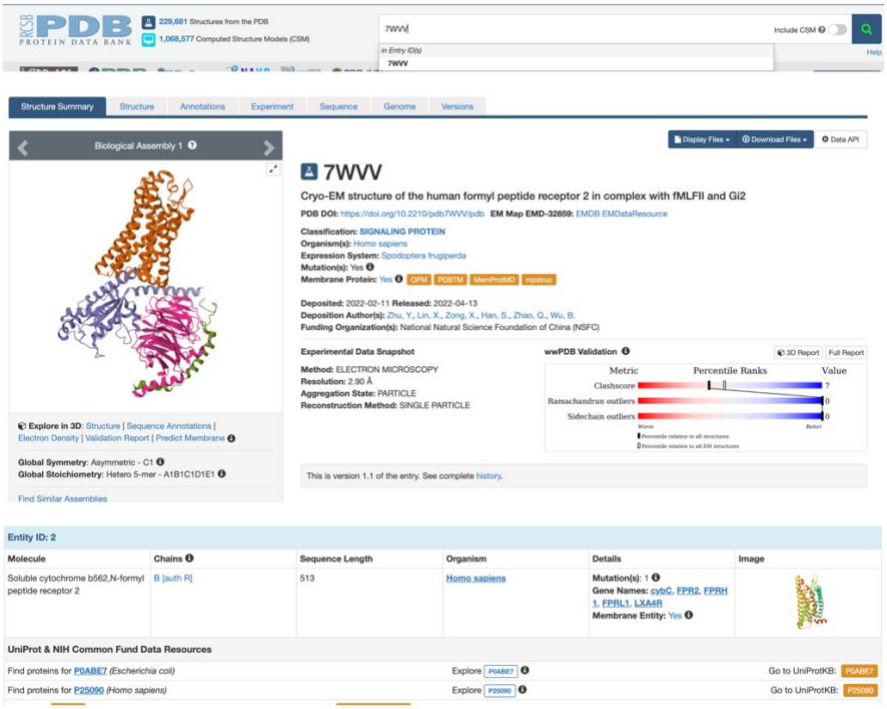


Figure 13: Case example for 7WVV_R, which is chimeric.

In all cases except for *7WVV_R*, the target chain is non chimeric and belongs to the *Homo Sapiens* species. In *Figure 13*, on the bottom image, we can see that the target chain is a fusion of soluble cytochrome b562 and N-formyl peptide receptor 2.

At this time of the procedure, I have finally chosen 6 templates which I will use to crossmatch with the results from HHPRED, namely: *7S8L_R, 9DQH_A, 8XXV_A, 7EUO_R, 7XJK_F and 8WSS_R*. I proceed to download the descriptions table CSV file.

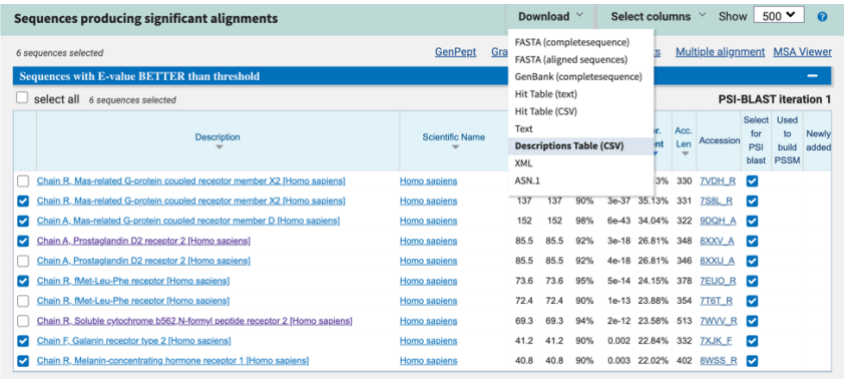


Figure 14: PSI-BLAST candidate templates.

index	Description	Species	Name	% identity	% query cover
1	Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	7S8L_R	35.13	90
2	Chain A, Mas-related G-protein coupled receptor member D [Homo sapiens]	Homo sapiens	9DQH_A	34.04	98
3	Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	8XXV_A	26.81	92
4	Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	7EUO_R	24.15	95
5	Chain F, Galanin receptor type 2 [Homo sapiens]	Homo sapiens	7XJK_F	22.84	90
6	Chain R, Melanin-concentrating hormone receptor 1 [Homo sapiens]	Homo sapiens	8WSS_R	22.02	90

Table 1: PSI-BLAST initial candidates

HHPRED

For this part, I head over to HHPRED webserver[4] and paste the same sequence found in the beginning in *UniProtKB* (*Figure 2*) and submit the job.

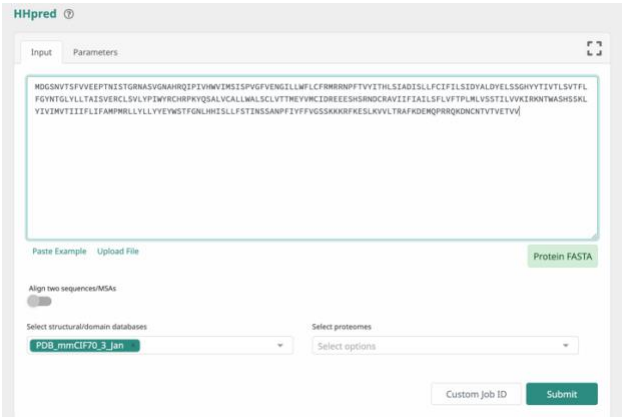


Figure 15: HHPRED Protein FASTA.

HHPRED returns 250 hits given the prior query sequence.

[4] https://toolkit.tuebingen.mpg.de/tools/hhpred

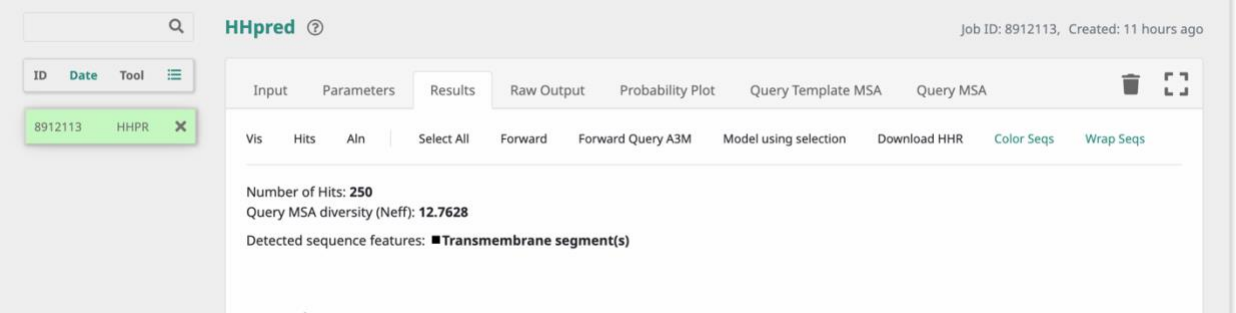


Figure 16: HHPRED job termination.

Next, I will search for a direct match between the name of the sequence and the results or a match for the first four or three digits of the sequence name, e.g. if *7S8L_R* is not found I search for *7S8L and 7S8 if need be*, which is usually an indicative that these proteins come from the same research study. The former information can be confirmed via visiting the respective webpage within *PDB*. In case I don't find a match using the protein name, I still try to find it via the description, e.g. if the description is '*Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]'* I search for '*Mas-related G-protein coupled receptor member X2'*.

For each match found, I always visit the *PDB* webpage for that sequence entry and confirm if the species remains Homo Sapiens. I will illustrate an example in images of the process which is done for all sequences in *Table 1*.

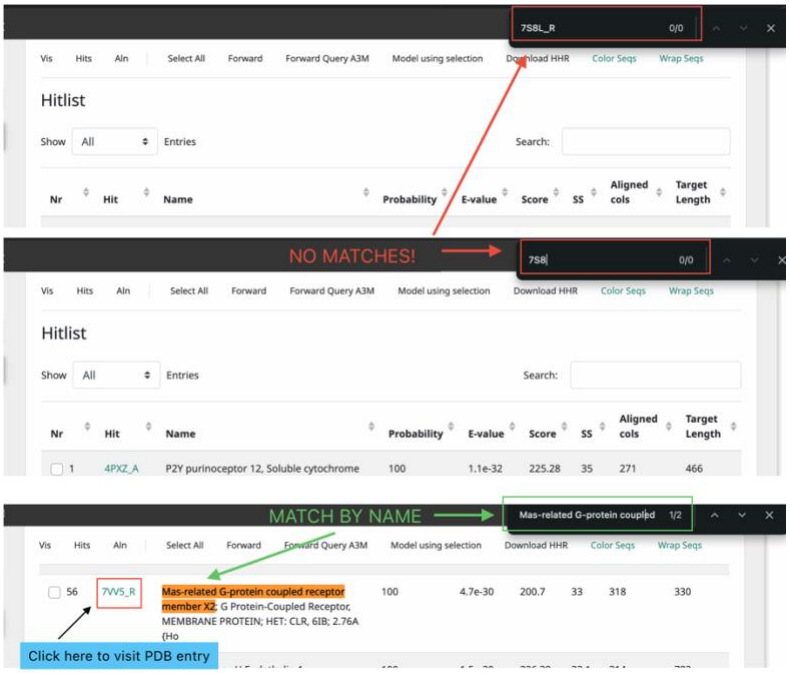


Figure 17: HHPRED crossmatch example with 7S8L_R.

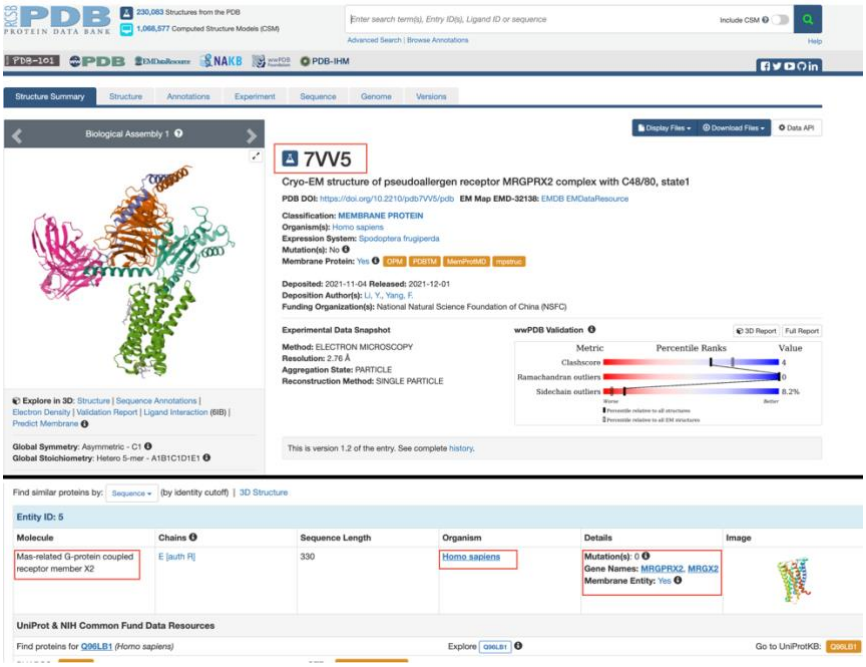


Figure 18: HHPRED crossmatch example with 7S8L_R, visiting PDB entry for information.

Furthermore, on the HHPRED results page I also take notice of the result index and the similarity score associated with it.

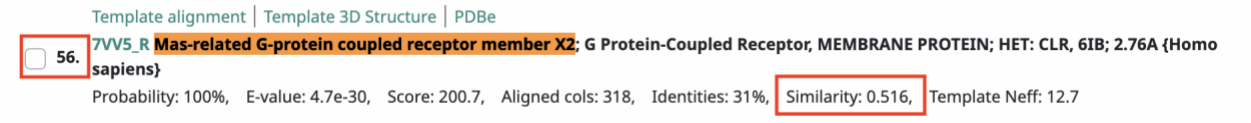


Figure 19: HHPRED crossmatch example with 7S8L_R, entry index and similarity score.

index	Description	Species	Name	% identity	% query cover
1	Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	7S8L_R	35.13	90
2	Chain A, Mas-related G-protein coupled receptor member D [Homo sapiens]	Homo sapiens	9DQH_A	34.04	98
3	Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	8XXV_A	26.81	92
4	Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	7EUO_R	24.15	95
5	Chain F, Galanin receptor type 2 [Homo sapiens]	Homo sapiens	7XJK_F	22.84	90
6	Chain R, Melanin-concentrating hormone receptor 1 [Homo sapiens]	Homo sapiens	8WSS_R	22.02	90

index	% similarity	HHPRED PDB ID	Protein name	New PDB	HHPRED position
1	51.60	No	MRGPRX2	7VV5	56
2	46.70	9DQJ	MRGPRD	-	18
3	27.70	Yes	PTGDR2	-	16
4	26.90	No	FPR1	7T6T	354
5	21.15	Yes	GALR2	-	332
6	20.01	Yes	MCHR1	-	90

Table 2: Results from crossmatch between PSI-BLAST and HHPRED

Table 2 highlights the results from crossmatching the sequences from *PSI-BLAST* with the sequences found in HHPRED. Sequences with indices 1 and 4 were not found via the *PDB ID*, but rather through name search, as suggested in column *'HHPRED PDB ID'*. The sequence with index 2 was found through a variation of the *PDB ID*. Finally, sequences with indices 3, 5 and 6 were found via a direct match. For all sequences, the protein name remained the same, hence the description column remains unchanged.

Notice that I didn't include a column for the type of protein (chimera or wild type) because they are currently all wild type. Furthermore, because tables can get very wide to uphold all the information, an indexing system is in place to make a proper reference, as observed in the first columns of *Table 2*.

The next step in this process is to assess if the proteins in consideration are in holo/apo state, with an active/inactive conformation. To do this, I need to check the respective *PDB* entry and look at the ligands and effectors in place, if they exist. I can also read the abstract from the study associated with the protein to get a clue and retrieve the resolution (Å) of the methodology used. Additionally, *GPCRs* in holo state have a ligand bound in the orthosteric pocket, located within the transmembrane bundle, and a typically an active conformation results in the outward pivoting of TM6. Hence, visualizing the 3D structure may help to consolidate previously gathered information, as well as provide information on missing segments of the target chain in relation to the sequence we are trying to model, improving the decision upon which templates to use. To illustrate this, I will provide a full example of the procedure in images, which I will then apply to all sequences from *Table 2*.

First, I find my sequence in HHPRED and click on the protein name to access the respective *PDB* entry.

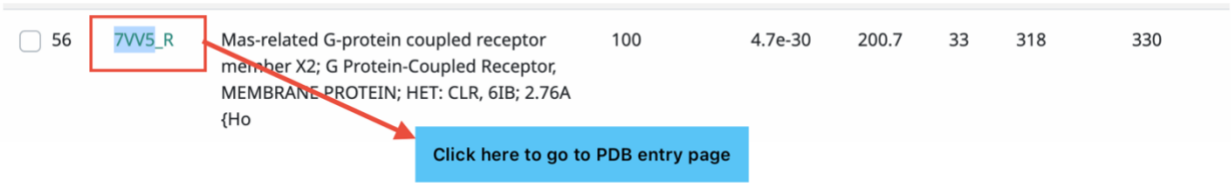


Figure 20: Information gathering example, accessing PDB entry.

From the *PDB* entry page I can immediately gather suggestive information about a ligand in the title name of the protein. Scrolling down, I can get the type of methodology used to infer the structure and the resolution. On the literature section I can further confirm that this protein is liganded to *C48/80* and bound to a *Gi*-complex

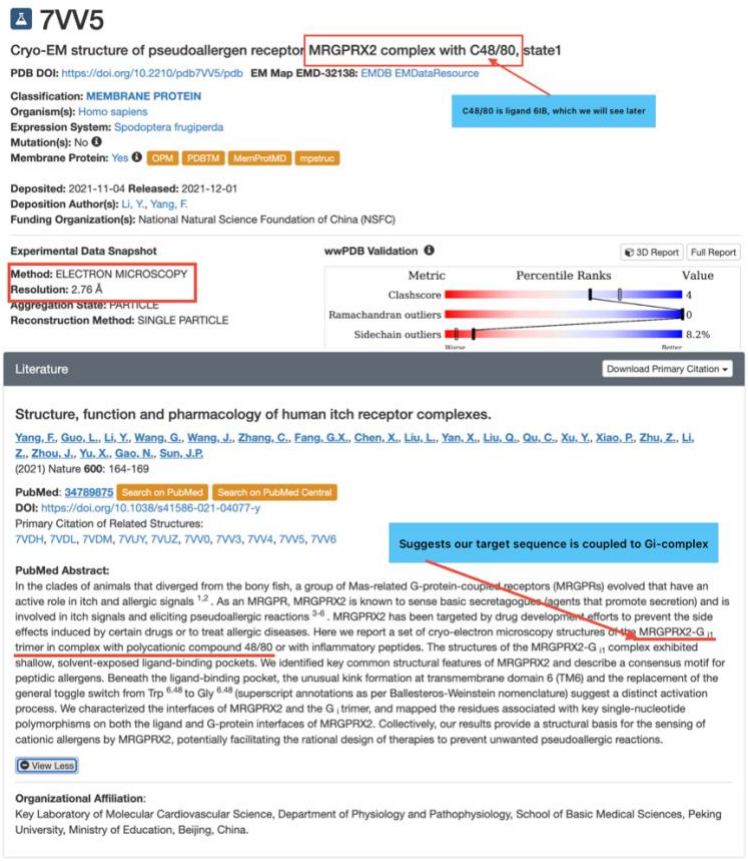


Figure 21: Information gathering example, preliminary information of PDB.

On the bottom of the page is a section for the ligands. For this protein we have two ligands, *6IB* and *CLR*. I can click on *6IB* to get more info (which I will show later), and I can scroll back up to explore the 3D structure in *PDB Mol Viewer*. At the top of the webpage there is a suggestion on the 3D preview regarding ligand interactions. The ligand that appears in this section might not always be the one in the orthosteric pocket, which ultimately will determine the holo/apo state, so I must be careful. To investigate further, I click on structure to open the 3D viewer, and I can effectively see that *6IB* is in the orthosteric pocket. With the same reasoning it is also possible to infer whether we have a *Gi*, *Go* or *Gq* complex bound to the protein. Additionally, I can inspect how TM6 is pivoted and if there are any missing sections on the *GPCR*.

(related images on next page)

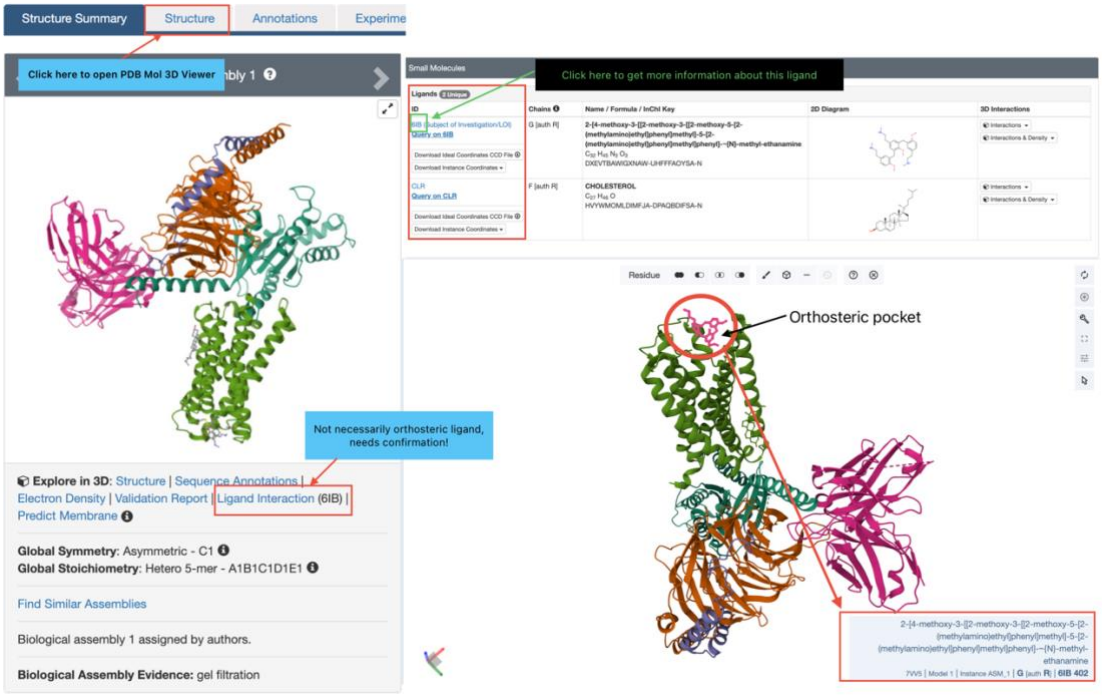


Figure 22: Information gathering example, ligand and 3D analysis.

Figure 23: Information gathering example, TM6 positioning.

Now I want to confirm if *6IB* is the same as *C48/80*. Relating back to *Figure 22*, I can click on *6IB* hyperlink, and this takes me to the *PDB* entry regarding this ligand. From here I can infer some information about it, but nothing that immediately tells me it is the same as *C48/80*. At the bottom of the page there is a hyperlink to the respective *PubChem* entry[5].

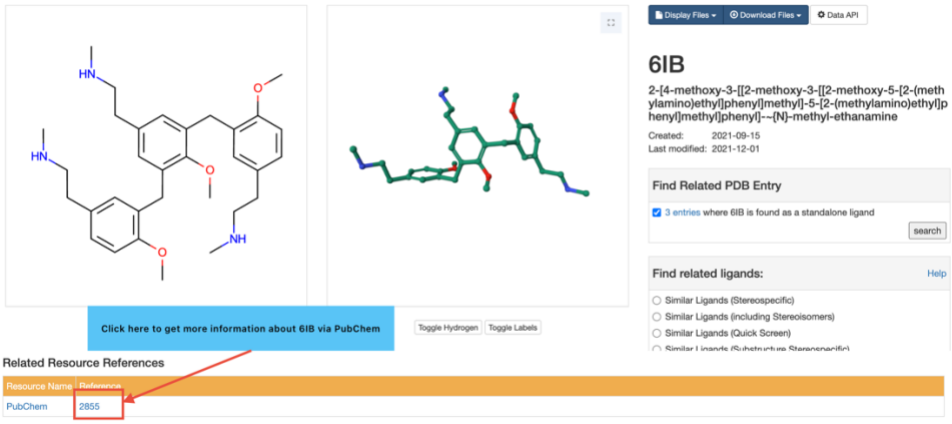


Figure 24: Information gathering example, 6IB PDB info.

In *PubChem* I can effectively see that *C48/80* is a synonym for *6IB*. Furthermore, at the bottom of this webpage there is information on how this ligand acts as an agonist in our current protein under investigation.

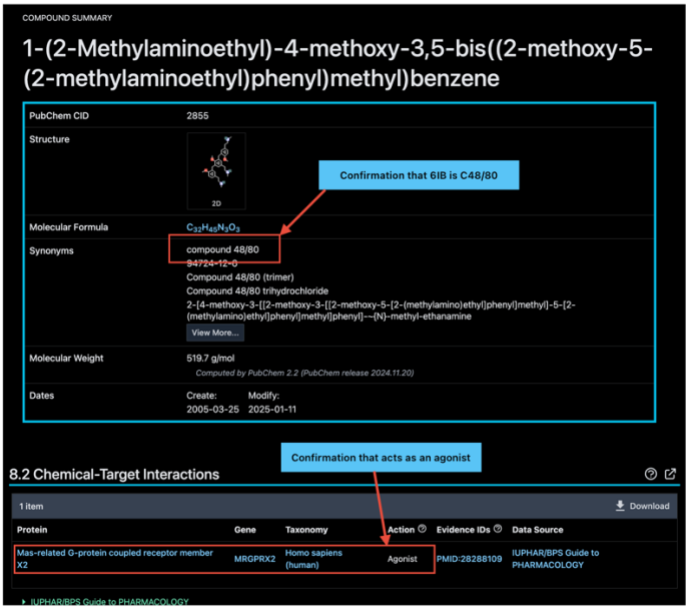


Figure 25: Information gathering example, 6IB PubChem info.

[5] https://pubchem.ncbi.nlm.nih.gov/compound/2855#section=Ligands-from-Protein-Bound-3D-Structures

This process is repeated for all 6 of our templates in current consideration. The gathered information goes on the following table.

index	New PDB	Resolution Å	Orthosteric ligand ID/Type	Effector	TM6	State & conformation	WT/Chimera
1	7VV5	2.76	6IB - agonist	Gi	outwards	holo, active	WT
2	9DQJ	2.90	A1BEQ - agonist	Gq	outwards	holo, active	WT
3	8XXV	2.33	-	Gi	inwards	apo, non-active	WT
4	7T6T	3.20	FME - agonist	Gi	outwards	holo, active	WT
5	7XJK	3.30	Galanin - agonist	Gq	outwards	holo, active	WT
6	8WSS	3.01	MCH - agonist	Gi	outwards	holo, active	WT

Table 3: Information gathered on preliminary selection of templates.

From this table I can see that I have five templates in holo state with an active conformation, and one template in apo state with an inactive conformation. Furthermore, some proteins were missing structural parts, depicted in the following image. All of them were missing the n-terminal and the c-terminal.

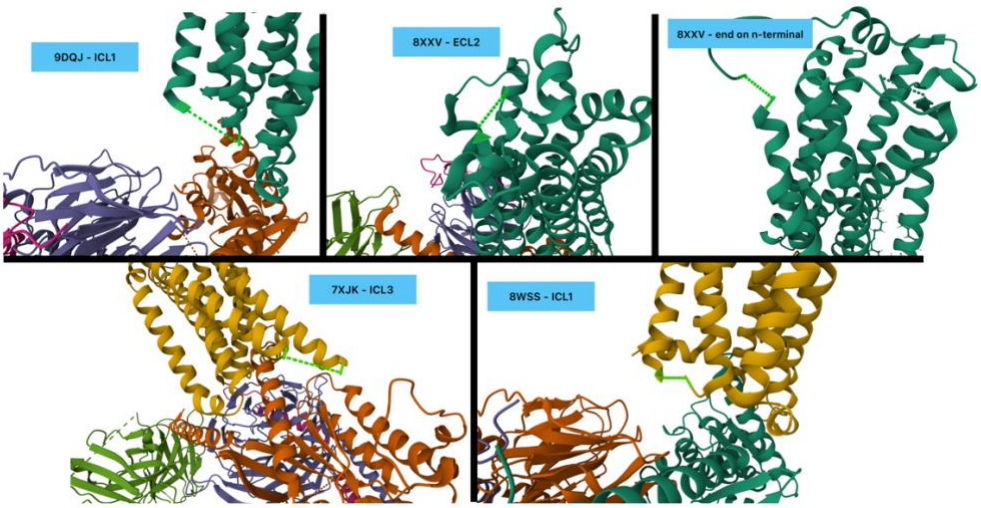


Figure 26: Preliminary selection missing structures.

From the gathered information, my choice for homology modelling will be the holo state in active conformation of MAS1 sequence. Furthermore, I will consider the parameters of choice from most to least important to be percentage of identity, percentage of similarity, percentage of query coverage, missing parts and resolution.

index	New PDB	% identity	% similarity	% query cover	Missing	Resolution Å
1	7VV5	35.13	51.60	90.00	-	2.76
2	9DQJ	34.04	46.70	98.00	ICL1	2.90
4	7T6T	24.15	26.90	95.00	-	3.20
5	7XJK	22.84	21.15	90.00	ICL3	3.30
6	8WSS	22.02	20.01	90.00	ICL1	3.01

Table 4: Final templates ordered in descending by priority of parameters.

Conveniently, the first four elements only break descending order on query coverage. Furthermore, by choosing the first four I can make sure I incorporate all elements with non-missing structural parts and no repeated missing structural parts. Hence, my choice will be *7VV5*, *9DQJ*, *7T6T* and *7XJK*.

MODELLER

Given the choice of templates, I proceed to the HHPRED webpage, select the four previously mentioned elements and click on *'Model using selection'*, then 'Forward to MODELLER'. For the MODELLER key I use 'MODELIRANJE' and click submit.

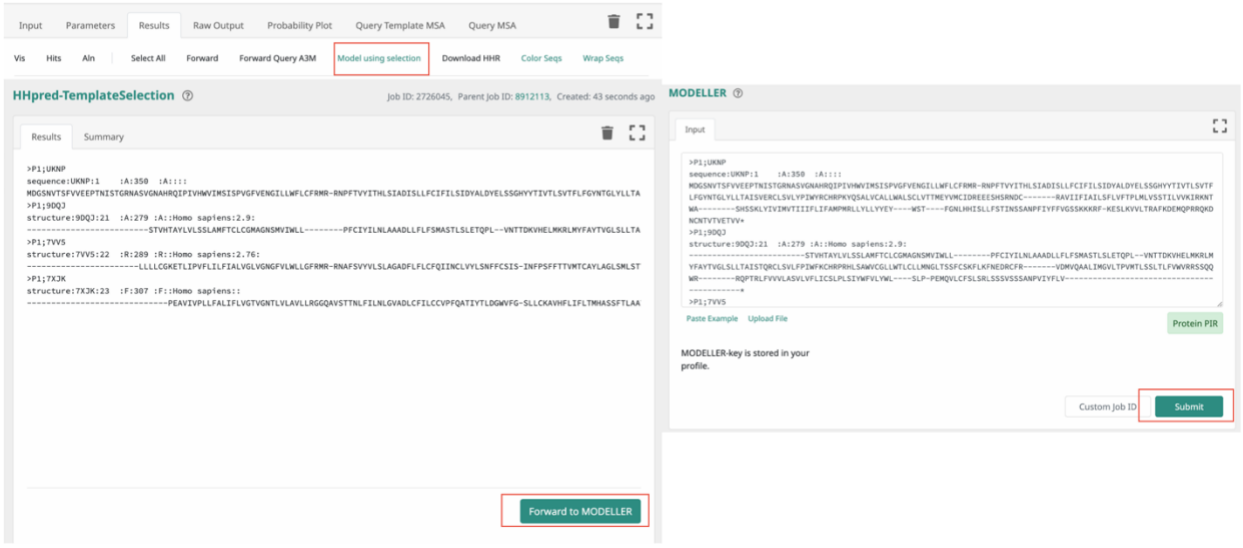


Figure 27: Submission to MODELLER

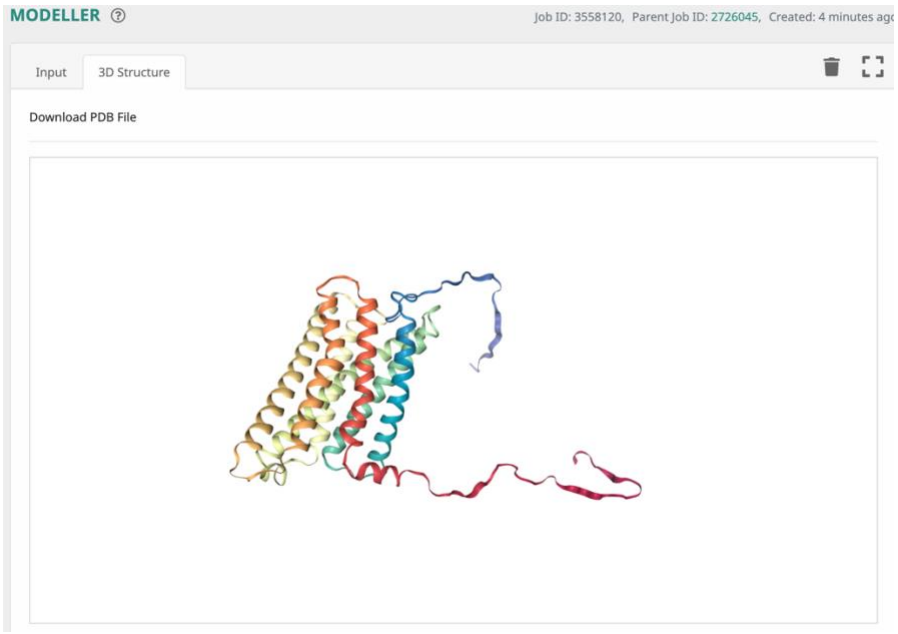


Figure 28: Result from MODELLER

From the resulting webpage I also download the PDB file for the next steps.

VALIDATION

VISUAL CHECK

To perform visual checking of the modelled protein, I loaded the .pdb file into PDB Mol Viewer.

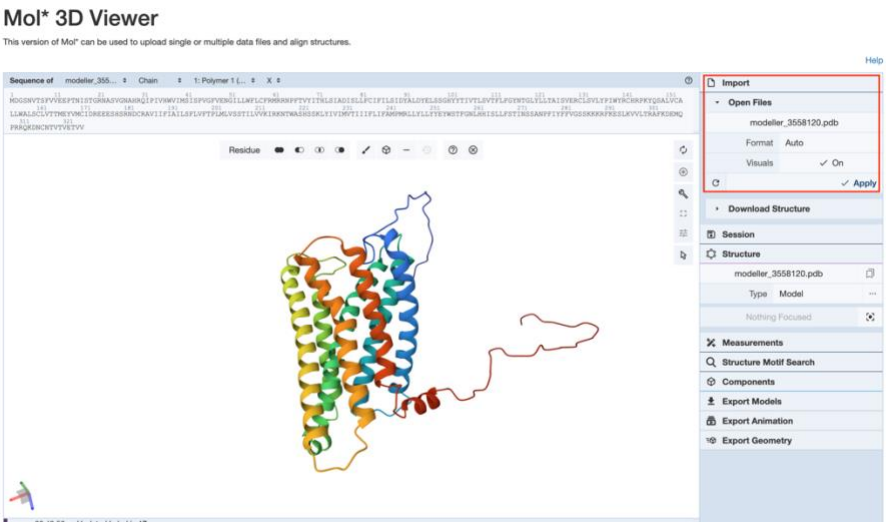


Figure 29: https://www.rcsb.org/3d-view

Terminals

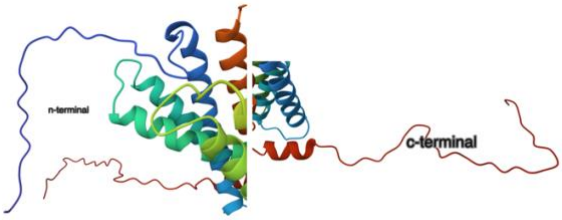


Figure 30: c-terminal and n-terminal

Both terminals are unstructured as expected. As previously mentioned, these were absent in all our considered templates.

Intracellular Loops

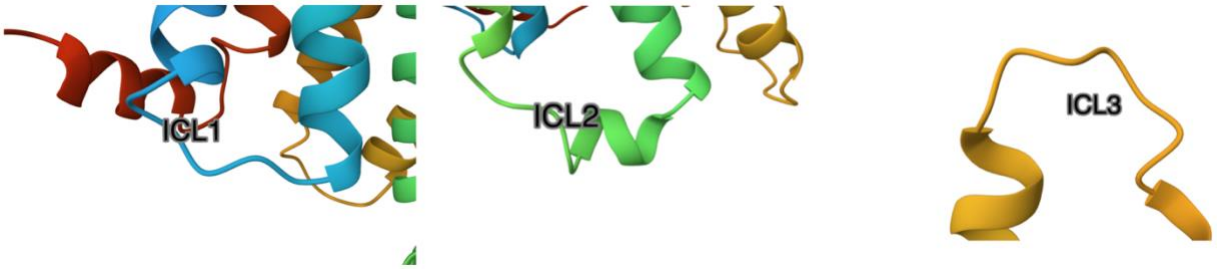


Figure 31: Model intracellular loops.

ICL1 and ICL3 present no irregularities or accentuated loops, whereas ICL2 presents a helix structure typical of GPCRs in holo state with an active conformation, as ICL2 is a binding site for the Gi-complex.

Extracellular Loops

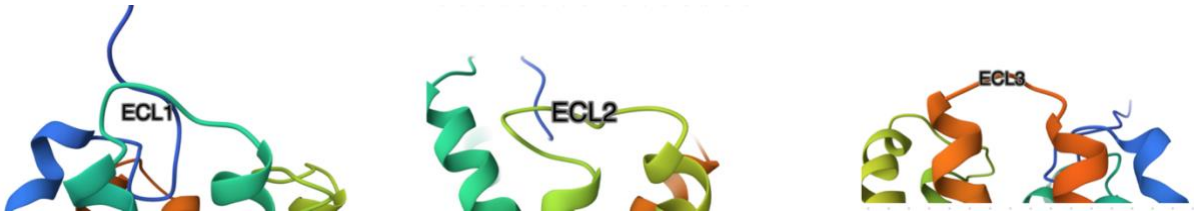


Figure 32: Model extracellular loops.

All extracellular loops don't present any irregularities or accentuated loops.

Transmembrane helices

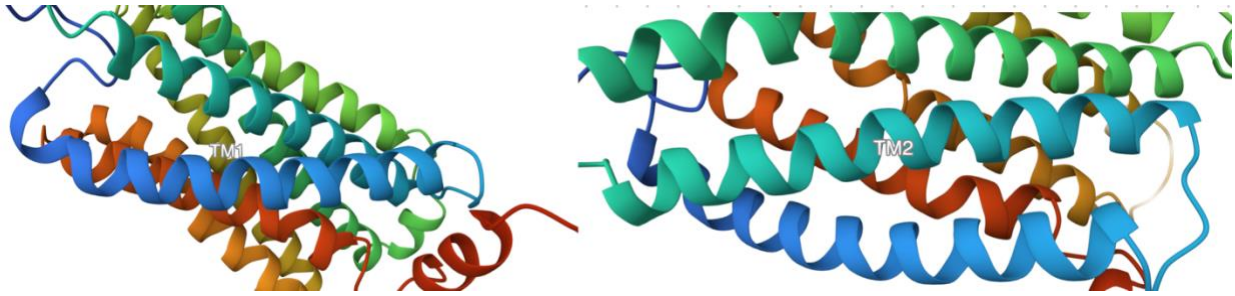


Figure 33: Model TM1 and TM2

TM1 shows a slight elongation towards the extracellular region, but overall helices are well structured with no bents. TM2 is overall well-structured with no bents or elongations.

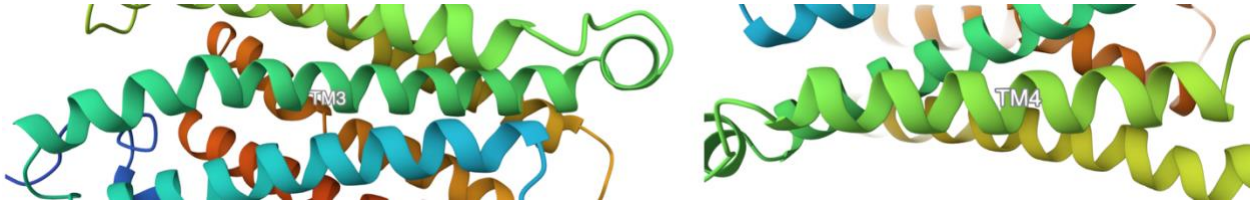


Figure 34: Model TM3 and TM4.

TM3 has a slight bent in the middle part nearest to the extracellular region, but overall helices are well defined. TM4 shows no bents or elongations, with well-formed helices.

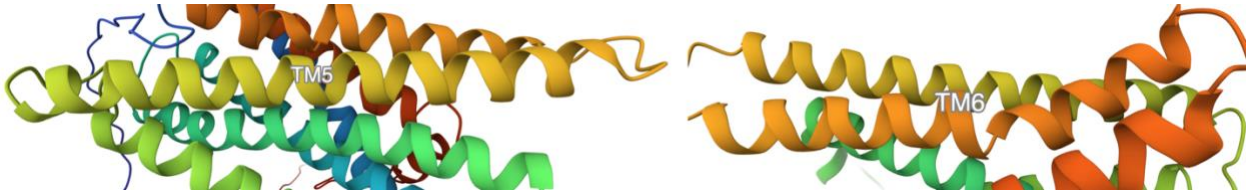


Figure 35: Model TM5 and TM6.

TM5 shows no bents or elongations, with well formed helices. TM6 shows a kink, bending outwards, typical of GPCRs in holo-active conformation. There is a secondary structure anomaly where the bent happens, this might be due to bound ligands in that region, deep in the orthosteric pocket as we will see later.

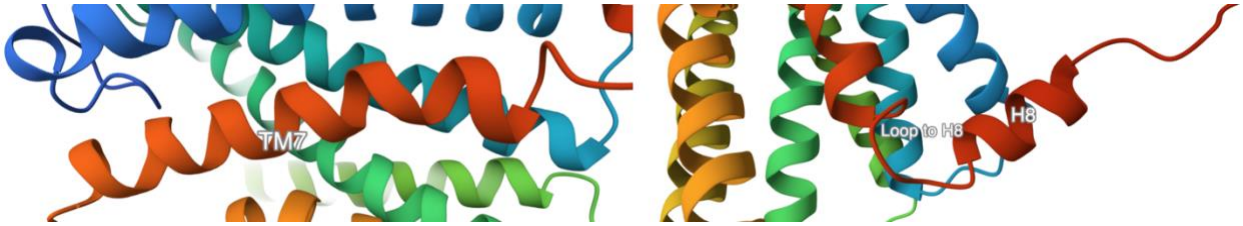


Figure 36: Model TM7, loop to H8 and H8.

TM7 shows a slight bent towards the end region of the intracellular regions, but overall helices are well defined. As expected from GPCRs in holo-active conformation, we have a small loop after TM7, followed by H8, typical of modelling GPCRs coupled to a G-complex.

Overall, we have the expected barrel like conformation of GPCRs and the structural holo-active conformation.

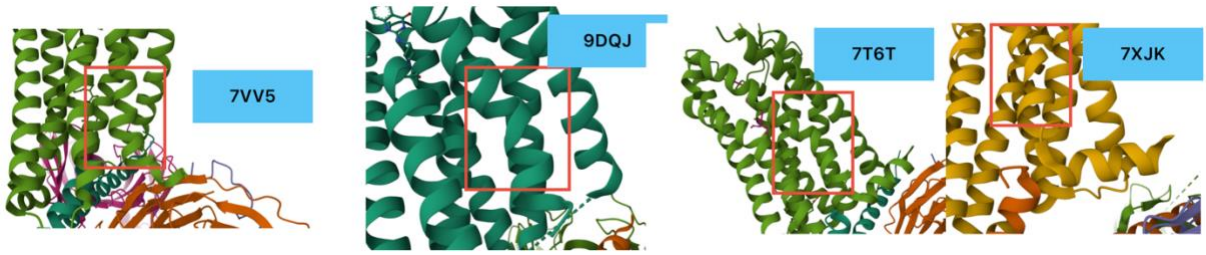


Figure 37: Templates bent on TM7

Like in my model, we can see a bent appearing on all my templates, which justifies why it appears on my model too.

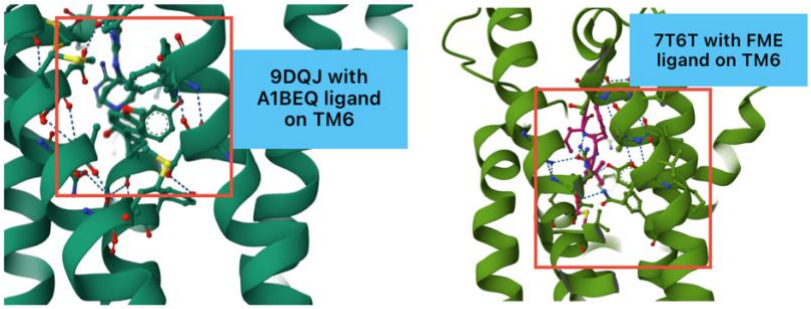


Figure 38: 9DQJ and 7T6T ligands on TM6 bent.

In the above figure we can see how the ligands to the kink region might be affecting the secondary structure of my model.

A-PRIORI ANALYSIS

UNIPROT

To perform an a-priori analysis, I head over to **P04201** initial webpage and extract the residues indexes for the transmembrane helices.

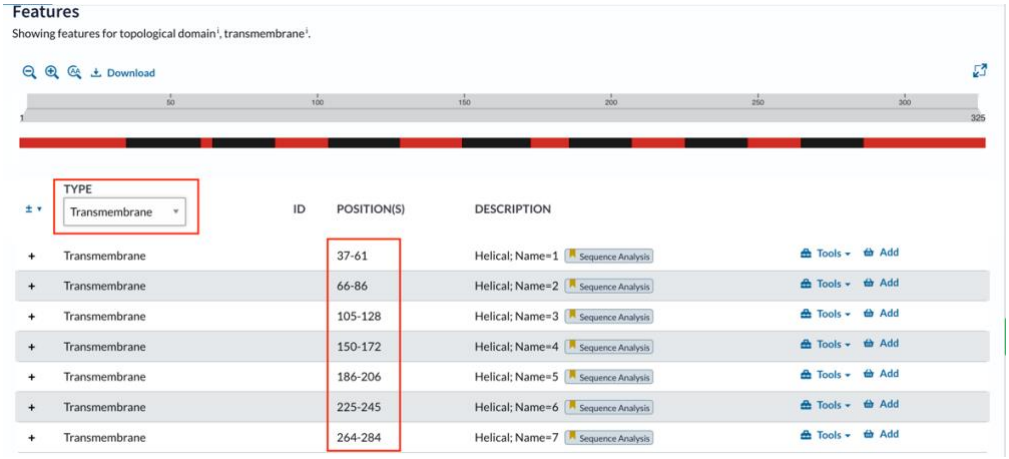


Figure 39: P04201 TM start-end residues.

PSIPRED

Next, I go to PSIPRED[6], I select both PSIPRED and MEMSAT-SVM from the analysis menu, paste the sequence of **P04201** and click submit.

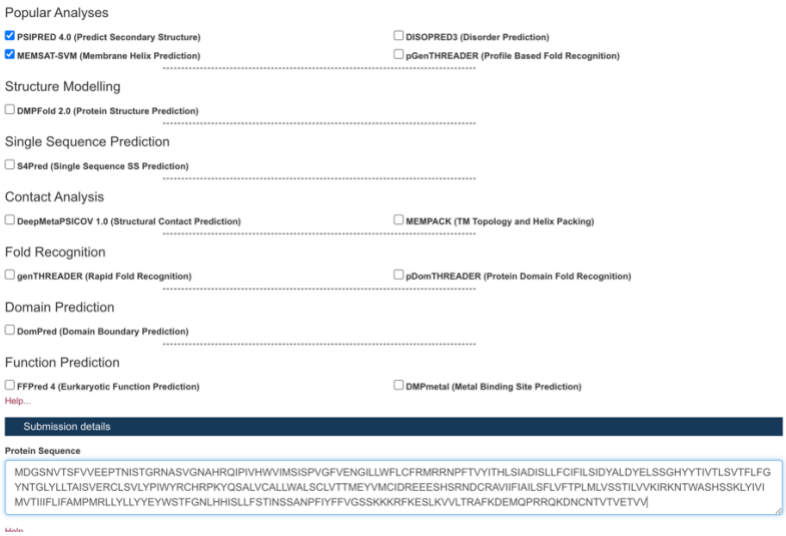


Figure 40: PSIPRED analysis submission.

[6] http://bioinf.cs.ucl.ac.uk/psipred/

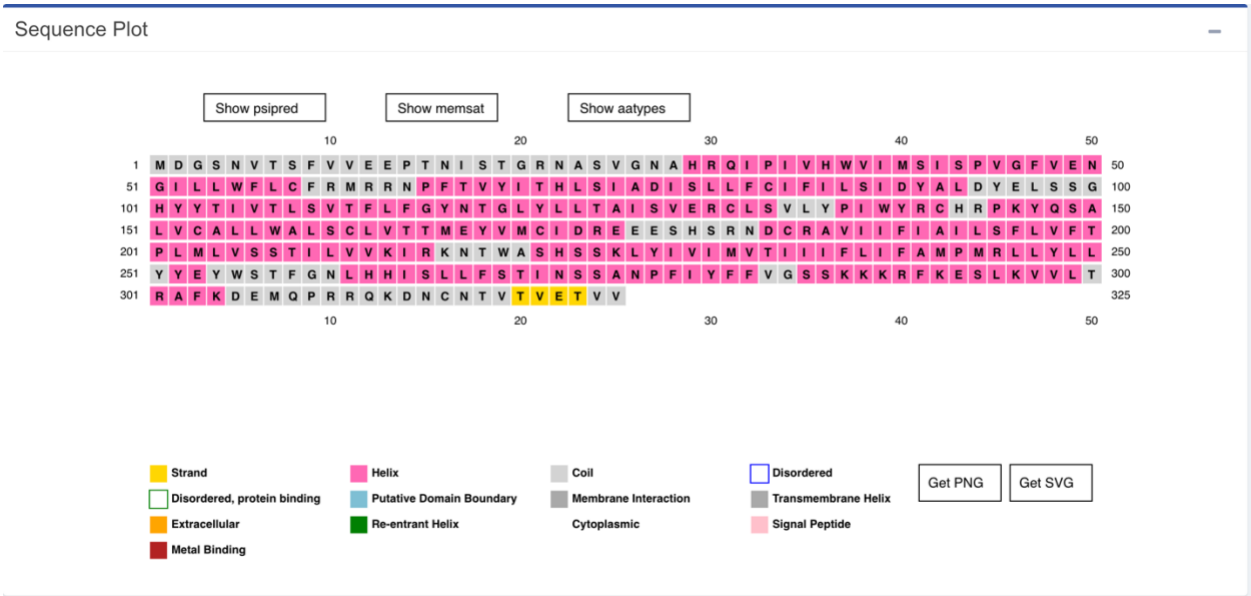


Figure 41: PSIPRED results.

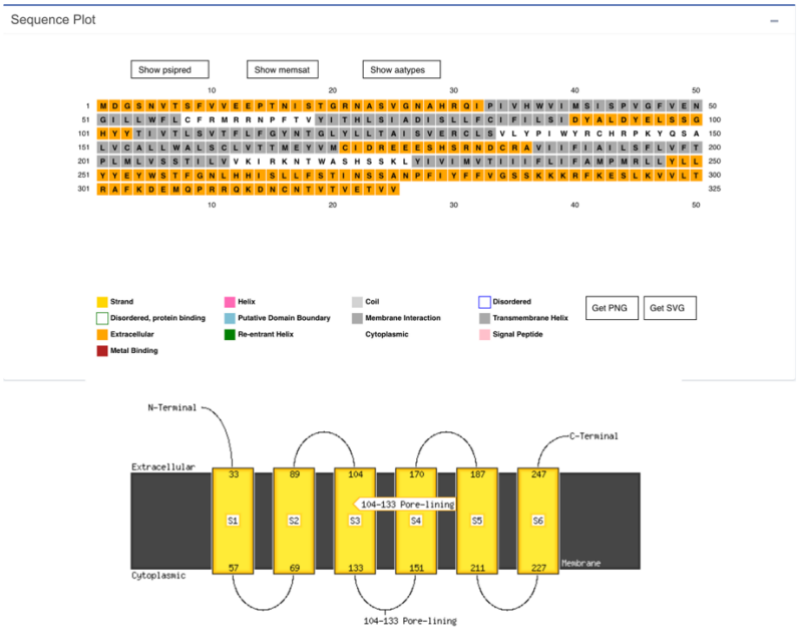


Figure 42: MEMSAT-SVM results.

DEEPTMHMM

Finally, I go to DEEPTMHMM[7], paste the sequence of **P04201** and click run.

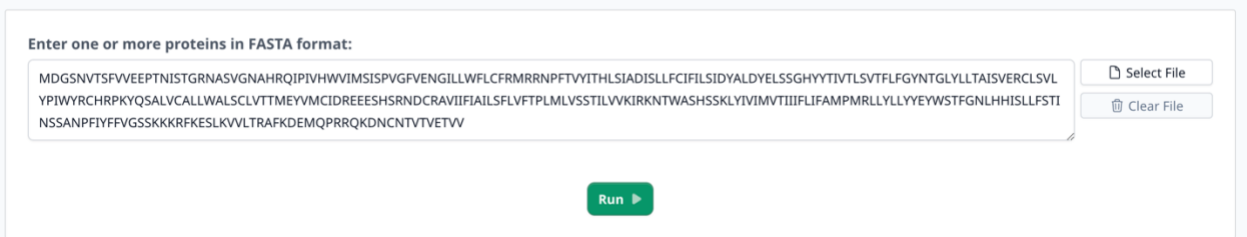


Figure 43: DEEPTMHMM analysis submission.

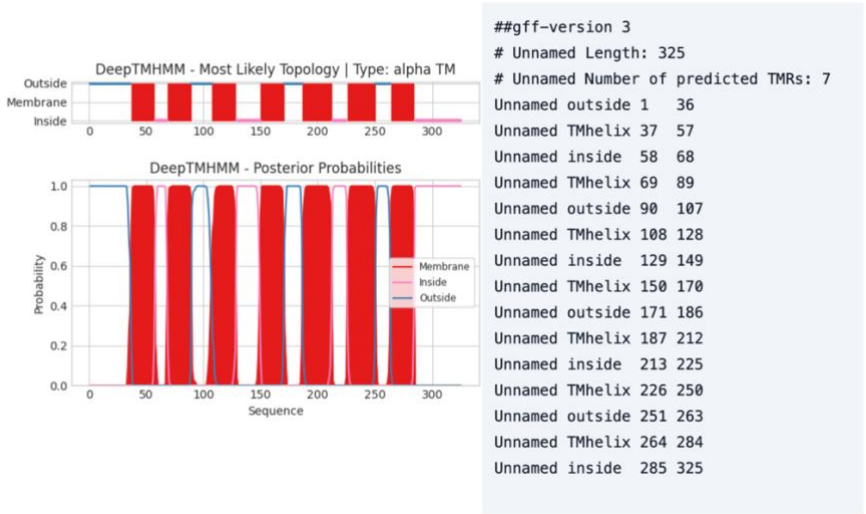


Figure 44: Results from DEEPTMHMM.

Analysis

Having gathered all the data, I proceed to PDB MOL 3D Viewer to also extract the residues indices for the transmembrane helices and compile everything in an excel file, like we did in class, to assess which tool provides the best and worst a-priori analysis, as well as which helix is worse/best predicted.

(tables on the next page)

[7] https://dtu.biolib.com/DeepTMHMM

	My model		*UniProt*		*PSIPRED*		*MEMSAT-VSM*		*DeepTMHMM*	
	start	end	start	end	start	end	start	end	start	end
H1	29	58	37	61	29	58	33	57	37	57
H2	65	93	66	86	65	93	69	89	69	89
H3	100	135	105	128	101	133	104	133	108	128
H4	148	170	150	172	145	175	151	170	150	170
H5	181	216	186	206	183	215	187	211	187	212
H6	224	253	225	245	221	250	227	247	226	250
H7	258	280	264	284	261	282			264	284

Figure 45: Summary of the helix's indices gathered.

		worst		*best*								
		UniProt		***PSIPRED***		***MEMSAT-VSM***		***DeepTMHMM***				
		start	end	start	end	start	end	start	end	**avg.**		**legend:**
	H1	8	3	0	0	4	-1	8	-1	3.125		abs(diff) ≤ 2
	H2	1	-7	0	0	4	-4	4	-4	3.000		3 ≤ abs(diff) ≤ 4
	H3	5	-7	1	-2	4	-2	8	-7	4.500		abs(diff) ≥ 5
best	**H4**	2	2	-3	5	3	0	2	0	2.125		
worst	**H5**	5	-10	2	-1	6	-5	6	-4	4.875		
	H6	1	-8	-3	-3	3	-6	2	-3	3.625		
	H7	6	4	3	2			6	4	4.167		
	avg.	4.929		1.786		3.500		4.214				

Figure 46: Results of a-priori analysis.

From the a-priori analysis we can see that PSIPRED was the best tool in predicting the helices was PSIPRED, whereas the worst was UniProt. Furthermore, across all models the best predicted helix was TM4, and the worst was TM5.

Regarding PSIPRED, even though in general the best predicted helix was TM4, this was it's worst one. On the other hand, it perfectly predicted TM1 and TM2.

A-POSTERIORI ANALYSIS

To perform an a-posteriori analysis using Ramachandran plots, I'll be using UCLA-DOE LAB[8]. Once I access this webserver, I start by submitting the .pdb file I got from *MODELLER* and press '*Run programs*'.

Figure 47: UCLA-DOE LAB analysis submission.

[8] https://saves.mbi.ucla.edu/

Once the webserver processes the request, a set of tools appear of which I choose procheck.

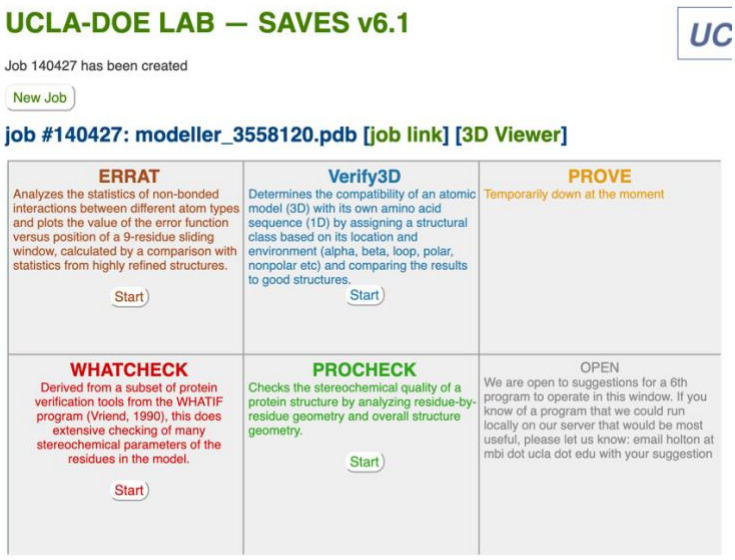


Figure 48: Procheck on UCLA-DOE LAB.

Ramachandran plot

The Ramachandran Plot is a graphical representation of the ϕ (phi) and ψ (psi) torsion angles of the polypeptide backbone in a protein structure. These angles determine the spatial arrangement of residues in a protein and are critical for understanding its conformation.

Key Elements of the Plot:

1. Favoured Regions:

- Represented by red areas. These are the most favoured ϕ and ψ angle combinations, typically observed in common secondary structures like α-helices, β-sheets, and left-handed helices.
- Residues in these regions are **structurally stable** and energetically favourable.

2. Allowed Regions:

- Represented by yellow areas. These are less common but still permissible ϕ and ψ angle combinations.

3. Disallowed Regions:

- Represented by white areas. These regions correspond to ϕ and ψ angle combinations that lead to steric clashes or unfavourable interactions, making them energetically unfavourable.

4. Outliers:

- Residues that fall into disallowed regions are flagged as structural outliers and could indicate errors in the model or unusual structural features.

5. Special Cases:

1. Glycine: Represented as triangles, it has greater flexibility due to the absence of a side chain and can occupy broader regions of the plot.
2. Proline: Represented as squares, it has restricted ϕ angles due to its rigid cyclic structure.

Interpretation of the Ramachandran Plot

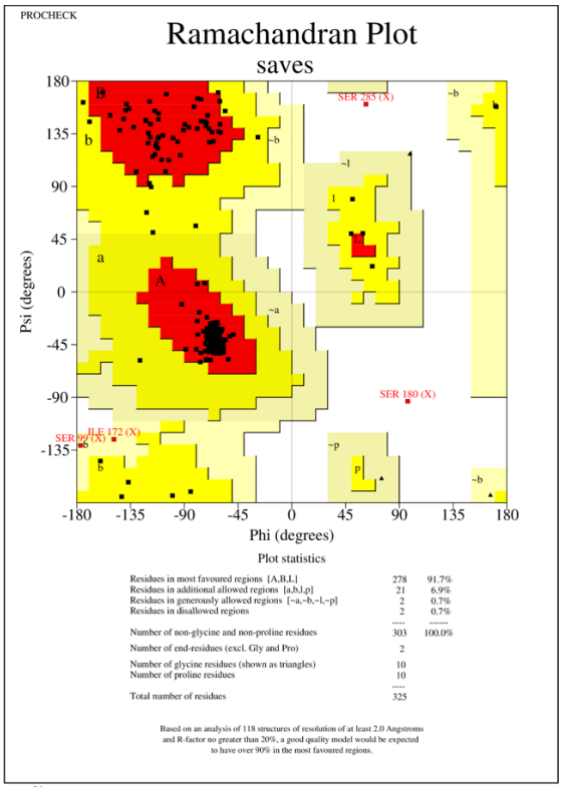


Figure 49:Ramachandran plot of my model.

Key Statistics:

- Residues in Most Favoured Regions (A, B, L): 91.7% (278 residues)
- Residues in Additional Allowed Regions (a, b, l, p): 6.9% (21 residues)
- Residues in Generously Allowed Regions (~a, ~b, ~l, ~p): 0.7% (2 residues)
- Residues in Disallowed Regions: 0.7% (2 residues)
- Total Residues: 325 (303 non-glycine/non-proline residues)

Interpretation:

1. Residues in Most Favoured Regions (91.7%):

- This is excellent, as high-quality models typically have ≥90% of residues in the most favoured regions. This value indicates that the model is well-structured and reliable.

2. Residues in Additional Allowed Regions (6.9%):

- This percentage is within a reasonable range and accounts for residues in less-constrained regions, such as loops and turns.

3. Residues in Generously Allowed Regions (0.7%):

- While small, these residues may require further investigation. They might represent flexible or unique regions.

4. Residues in Disallowed Regions (0.7%):

- Two residues (SER 285 and ILE 172) fall into disallowed regions. These could indicate:
- Modelling artifacts.
- Unusual conformations (e.g., in loops or turns).
- Errors in backbone geometry.
- These residues should be checked and refined, especially if they are in critical functional regions.

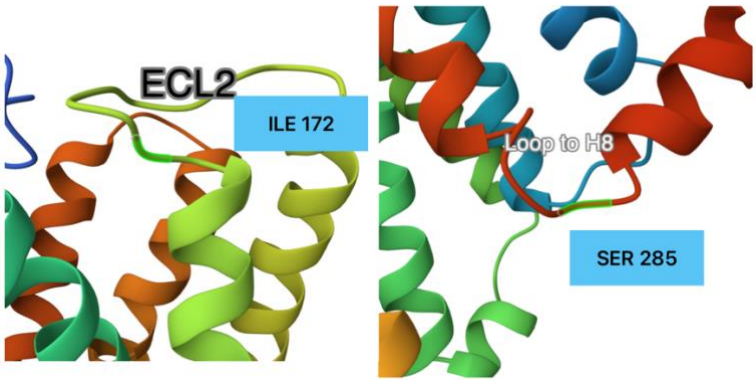


Figure 50: ILE 172 and SER 285 locations

ILE 172 and SER 285 are determined to be in disallowed regions in the Ramachandran plot, but they are contained withing flexible regions of the protein, namely ECL2 and the loop to H8.

5. Special Residues:

- Glycine (10 residues): Glycine is expected to have broad distribution due to its flexibility. Ensure none of these glycine residues fall into disallowed regions.
- Proline (10 residues): Proline's restricted ϕ angles are expected and consistent with its structural role.

The overall quality of the model is good, with 91.7% of residues in the most favoured regions, which meets the standard for high-quality structural models.

Ramachandran plots for all residues

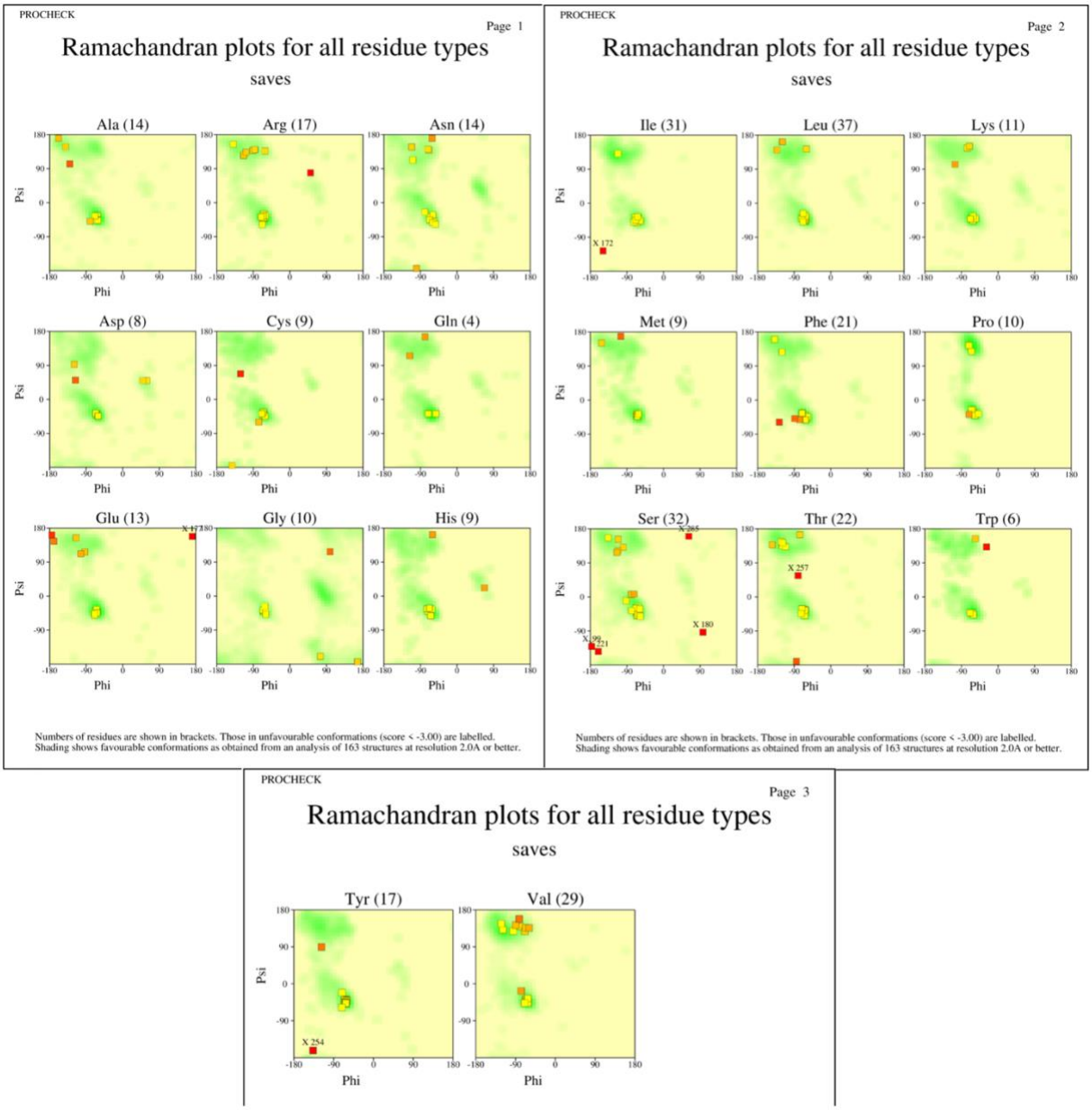


Figure 51: Ramachandran plots for all residues of my model.

Meaning of the Ramachandran Plot for All Residues

These Ramachandran plots provide residue-specific ϕ (phi) and ψ (psi) angle distributions for individual amino acids in your structure. Each amino acid type is analysed separately to assess how well their conformations align with favourable regions of the Ramachandran map.

Key Features of These Plots:

1. Residue-Specific Analysis:

- Each plot corresponds to a specific residue type (e.g., ALA, ARG, etc.), showing the ϕ and ψ angles of all instances of that residue in your protein.
- The number in parentheses indicates the total count of that residue type in the structure.

2. Shaded Regions:

- Green regions: Favourable conformations as determined from high-quality protein structures (resolution ≤2.0 Å).
- Yellow regions: Allowed conformations that are less common but permissible.
- White areas: Disallowed regions where ϕ and ψ angles are sterically or energetically unfavourable.

3. Highlighted Points:

- Squares or circles: Represent individual residues and their ϕ/ψ angles.
- Red points (X): Residues in disallowed regions (outliers). These should be investigated for structural accuracy.
-

Interpretation of the Plots

1. Residues in Favourable Regions:

- Most residues are concentrated in the green (favourable) or yellow (allowed) regions, which is expected for a well-constructed model.

2. Outliers:

- Residues marked as red squares (X) fall into disallowed regions, indicating potential structural anomalies or unusual conformations. These should be closely examined and may require refinement or justification if biologically relevant.

Analysis:

Page 1:

1. Alanine (ALA):

- All residues are in favourable or allowed regions, indicating no anomalies for this residue type.

2. Arginine (ARG):

- One residue (red square) is in a disallowed region. This needs further investigation.

3. Asparagine (ASN):

- One residue is flagged in a disallowed region, suggesting a potential outlier.

4. Aspartic Acid (ASP) and Cysteine (CYS):

- All residues are in favourable regions, showing no issues.

5. Glutamine (GLN):

- No outliers observed; all residues are in favourable or allowed regions.

6. Glutamic Acid (GLU):

- One residue is in a disallowed region (red square). This should be investigated.

7. Glycine (GLY):

- Glycine's flexibility allows a broad range of conformations, but no outliers are observed.

8. Histidine (HIS):

- One outlier in a disallowed region.

Page 2:

1. Isoleucine (ILE):

- ILE 172 is flagged as an outlier in a disallowed region but was previously assessed to belong to a flexible region ECL2.

2. Leucine (LEU):

- All residues are in favourable or allowed regions.

3. Lysine (LYS):

- All residues are in favourable or allowed regions.

4. Methionine (MET):

- One outlier in the disallowed region.

5. Phenylalanine (PHE) and Proline (PRO):

- All residues are within favourable or allowed regions.

6. Serine (SER):

- SER 285 and SER 180 are in disallowed regions. SER 285 was previously assessed and belongs to a flexible region, the loop just before H8. It is not the same case with SER 180, as it belongs to the beginning of TM5.

Figure 52: SER 180

7. Threonine (THR):

- THR 257 is flagged as an outlier but is in a flexible region ECL3.

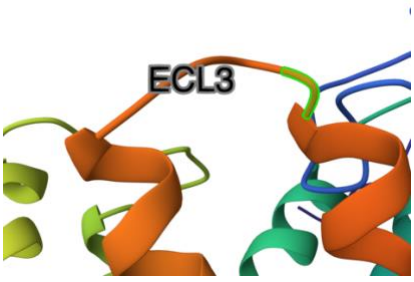


Figure 53: THR 257

8. Tryptophan (TRP):

- All residues are within favourable or allowed regions.

Page 3:

1. Tyrosine (TYR):

- TYR 254 is in a disallowed region. Like THR 257, it also belongs to ECL3.

Figure 54: TYR 254

2. Valine (VAL):

- All residues are in favourable or allowed regions, with no outliers.

Conclusion:

- The structure overall has good agreement with favourable regions for most residues, indicating a well-refined model.
- Outliers: SER 180 is the only one belonging to a helix, whereas all the others are in flexible regions

Chi1-Chi2 plots

The Chi1-Chi2 plots display the side-chain dihedral angles (χ1 and χ2) for residues with flexible side chains in the protein structure. These plots help evaluate the conformational preferences of amino acid side chains and detect any unfavourable side-chain conformations.

Key Features of the Plots:

1. Chi1 and Chi2 Dihedral Angles:

- Chi1 and Chi2 are the first and second dihedral angles of the amino acid side chains, respectively.
- These angles describe the rotation of the bonds within the side chains.

2. Shaded Regions:

- Green regions: Represent the most favourable side-chain conformations based on high-quality crystal structures.
- Yellow regions: Allowed conformations that are less favourable but permissible.
- White regions: Disallowed regions where steric clashes or energetically unfavourable interactions occur.

3. Data Points:

- Squares represent the residues and their Chi1 and Chi2 angles.
- Red squares indicate residues in disallowed regions (outliers) that may require further investigation.

Interpretation of the Plots:

Page 1:

1. Arginine (ARG):

- Most residues are in favourable regions, but a few are outliers in disallowed regions.

2. Asparagine (ASN):

- Several residues fall in disallowed regions, indicating potential structural issues with their side-chain orientations.

3. Aspartic Acid (ASP):

- All residues are in favourable or allowed regions.

4. Cysteine (CYS):

- No outliers detected. Side-chain orientations are within expected regions.

5. Glutamine (GLN):

- One residue falls in a disallowed region.

6. Glutamic Acid (GLU):

- One or two residues are in disallowed regions, suggesting potential refinement is needed.

7. Histidine (HIS):

- A few residues fall in disallowed regions. These need to be verified or refined.

8. Isoleucine (ILE):

- Residue ILE 235 is flagged as an outlier in a disallowed region. This residue is in TM6, near the ligand site of FME and A1BEQ from the considered templates.

9. Leucine (LEU):

- Most residues are in favourable regions, with a few falling into disallowed areas.

Page 2:

1. Lysine (LYS):

- All residues are in favourable or allowed regions, with no significant outliers.

2. Methionine (MET):

- A few residues fall in disallowed regions.

3. Phenylalanine (PHE):

- Residues PHE 240 and PHE 281 are flagged as outliers. PHE 240, like ILE235 is in TM6 near the ligand site of FME and A1BEQ. PHE 281 is the last residue of TM7.

4. Tryptophan (TRP):

- All residues are in favourable or allowed regions, indicating no issues.

5. Tyrosine (Tyr):

- A few residues are flagged as outliers, requiring further investigation.

Conclusion:

Most residues are in favourable or allowed regions for Chi1 and Chi2 angles, indicating a well-refined model overall. The Chi1-Chi2 analysis supports a structurally sound model, with room for refinement in specific outlier residues to improve accuracy and reliability.

LOOP MODELLING

My objective here is to model the loop ICL3 and assess whether the resultant model is better or not by performing analysis of sidechain interactions. To do this, I access ModLooP webserver[9], I upload my .pdb file from MODELLER, I input the key *MODELIRANJE* and specify the range of the ICL3 of my model, which goes from residue 217 to 223.

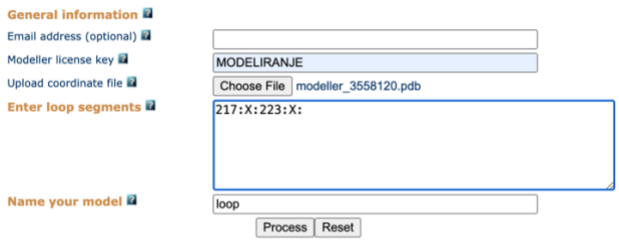


Figure 55: ModLooP job submission.

[9] https://modbase.compbio.ucsf.edu/modloop/

After the process was complete, I downloaded the resulting PDB and uploaded it to PDB Mol Viewer to see the new loop arrangement before going to LiteMol, where I noticed the loop was now shortened, ranging from residues 218 to 221.

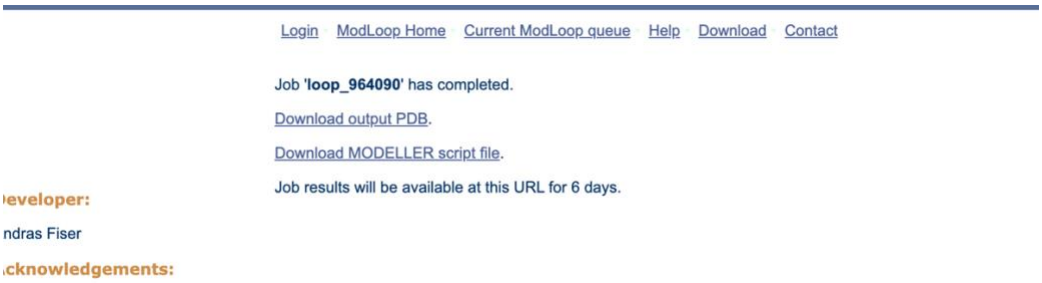


Figure 56: Download page for ModLoop.

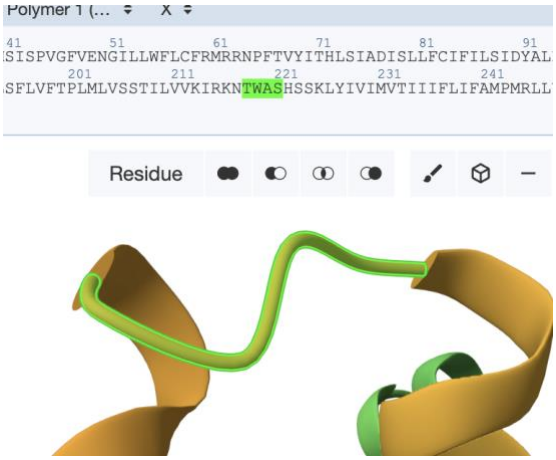


Figure 57: New ICL3 from ModLooP

To start investigating on possible interactions between in each of the loops, from my model and the ModLooP one, I loaded both .pdbs into LiteMol[10], added cartoons for balls and sticks and started analysing them.

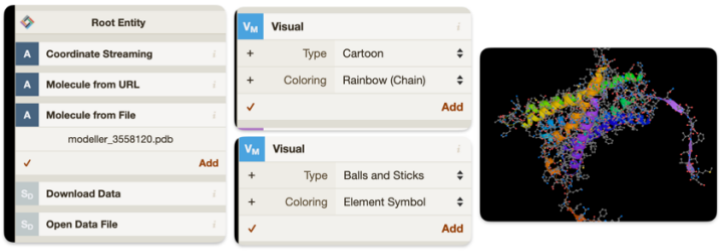


Figure 58: LiteMoL initializer

[10] https://glyco3d.cermav.cnrs.fr/LitemolViewer/advancedView.php

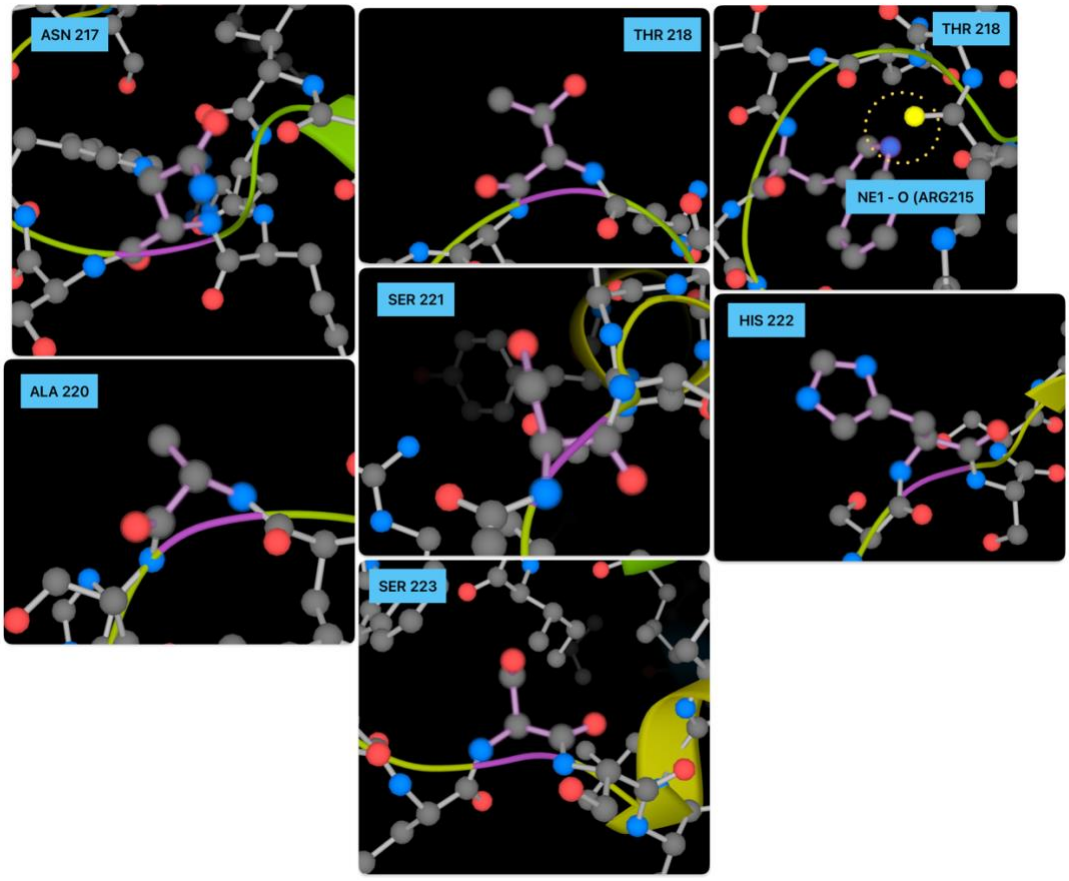


Figure 59: My model and side chain interactions in ICL3.

The only feasible interaction that I found in my model was a hydrogen bond between THR 218 NE1 (donor) and ARG 215 O (receiver).

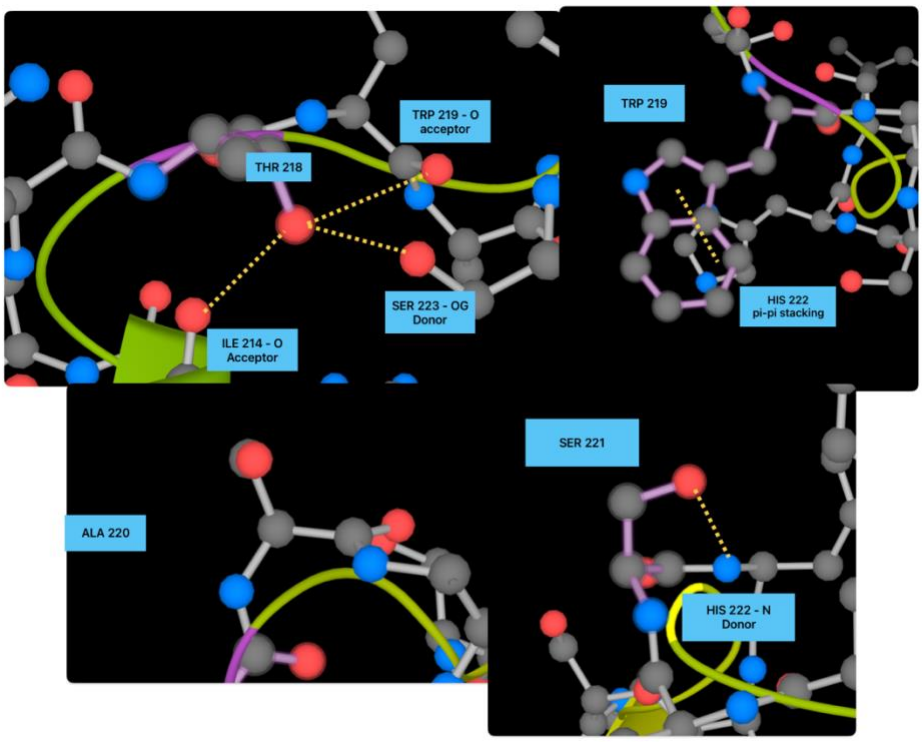


Figure 60: ModLoop model ICL3 interactions

On the other hand, in the ModLoop model I found there to be many residual interactions. THR 218 – OG1 acts as both as a donor and acceptor for surrounding atoms of oxygen or doubles bonded-oxygen. TRP 219 has a pi-pi interaction with HIS 222 and SER 221 – OG acts as a hydrogen acceptor with HIS 222 – N. Thus, counting a total of 5 interactions.

The ModLoop model contains a total of 5 interactions vs. 1 from my model. Furthermore, the loop was shortened in the ModLoop model. This is an indicator that the total free energy of the conformation is lessened, contributing to a more stable version of the protein. Hence, the ModLoop model is better.

Important note: I saw that LiteMol considers different intervals for the loops. I crossed matched the results from PDB Mol Viewer with HHPREDs output and they were equal. Hence, I considered those to be the real intervals and not the ones proposed by LiteMol.

ALPHAFOLD COMPARISON

To make an assessment regarding the AlphaFold model, I access its webserver[11], paste the sequence of **P04201** and submit the job.

Figure 61: Alphafold sequence submission.

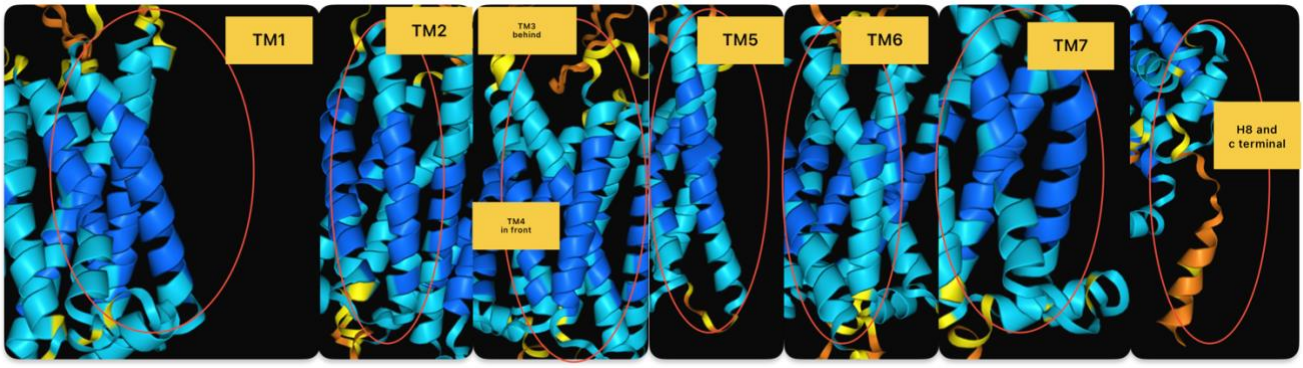


Figure 62: AlphaFold visual check.

From the results of AlphaFold we can see slight disturbances along TM1 and more noticeably in TM7, where the deformity is very pronounced. TM2, TM3 and TM4 present well-formed helices with no bents. TM6 has the expected outward kink deformation expected from GPCRs in holo-active conformation. Finally, the loop before H8 and the c-terminal present unlikely secondary structures and accentuated bents. Overall, the barrel like structure of the GPCR is achieved.

[11] https://alphafoldserver.com/

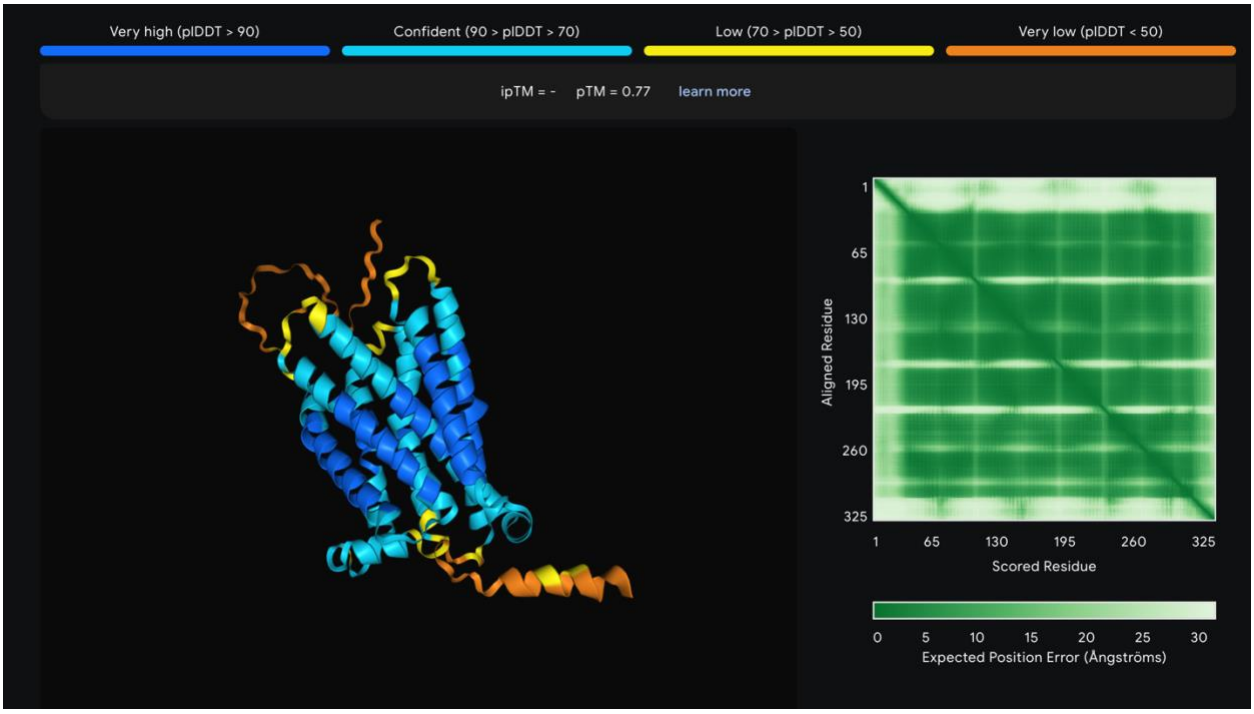


Figure 63: Results from AlphaFold.

The model presents the structural prediction results from AlphaFold, including the 3D protein model and confidence metrics such as pLDDT and pTM. The pLDDT (per-residue Local Distance Difference Test) measures the confidence in the predicted structure for each residue. The scale is color-coded, with dark blue regions (pLDDT > 90) representing very high confidence, light blue regions (70 < pLDDT ≤ 90) indicating confident predictions, yellow regions (50 < pLDDT ≤ 70) showing low confidence, and orange/red regions (pLDDT ≤ 50) reflecting very low confidence. In this model, most of the structure is in the blue range, indicating high to very high confidence in the predicted backbone and side-chain conformations. However, some regions, such as loops or termini, are coloured yellow or orange, suggesting low confidence in these flexible or poorly modelled regions. The pTM (Predicted Template Modelling Score) reflects the overall confidence in the global topology of the structure. A score of 0.77 suggests a relatively high confidence that the global structure is accurate.

The colouring indicates confidence per residue, with the core helices having the highest confidence and loops or termini having lower confidence. The orange and yellow regions likely correspond to disordered regions or flexible loops, which are inherently difficult to model accurately. These regions may also be involved in binding or undergo conformational changes, explaining the variability in confidence.

The Predicted Aligned Error (PAE) plot shows the expected positional error between residue pairs. The green diagonal line indicates low error for residues that are spatially close, such as those within the same secondary structure. Off-diagonal regions show error predictions between distant residues. Low error across most of the plot indicates strong confidence in the relative positions of residues. Some areas show higher error, which corresponds to the low-confidence regions in the 3D model, such as loops or flexible regions.

FINAL SUMMARY

The homology model for the MAS1 receptor (GPCR proto-oncogene) was carefully developed and validated through a series of bioinformatics analyses, ensuring that the resulting structure is robust and reliable for further research.

The initial template selection process began with the UniProt entry P04201, confirming the sequence details and species information. Using PSI-BLAST, templates were identified and filtered based on criteria such as query coverage, identity percentage, and structural resolution. The iterative analysis led to a refined selection of non-chimeric templates, prioritizing sequences with the best resolution and coverage.

Complementing PSI-BLAST, HHPRED provided cross-validation to ensure the selected templates were consistent and reliable. Six high-quality templates were identified, including both apo and holo states, with a final decision favouring holo-active conformations. MODELLER was then utilized to construct the homology model based on these templates, focusing on structural completeness and alignment accuracy.

The validation process highlighted the structural integrity of the model. Visual inspections using tools like PDB Mol Viewer confirmed the expected GPCR features, such as the barrel-like conformation, well-formed transmembrane helices, and characteristic outward pivoting of TM6 for the holo-active state. Despite minor bends in TM7 and TM3, these features were consistent with the selected templates and likely reflect the inherent dynamics of the receptor.

A-priori and a-posteriori analyses, including secondary structure predictions (PSIPRED, MEMSAT-SVM, DEEPTMHMM) and Ramachandran plots, provided further confidence in the model. The Ramachandran analysis revealed 91.7% of residues in the most favoured regions, meeting high-quality modelling standards. Outliers were limited to flexible loop regions, minimizing concerns about functional impact. Chi1-Chi2 plots corroborated these findings, indicating favourable side-chain conformations with minimal steric clashes.

Loop modelling was another critical aspect. The refinement of ICL3 through ModLoop demonstrated a significant improvement in stability, with a shortened loop and increased hydrogen bonding interactions. This refinement enhances the overall energy profile and structural robustness of the model.

Finally, a comparative analysis with the AlphaFold structure validated the quality of the homology model. While AlphaFold offered high confidence for core helices, regions such as TM7 exhibited pronounced deformations not observed in the homology model. The homology model demonstrated better alignment with experimental structural expectations, particularly in the transmembrane domain and loop regions.

In summary, the homology model of the MAS1 receptor is a robust representation, validated through rigorous computational and structural analyses. It aligns well with known GPCR features and provides a reliable foundation for subsequent studies, such as ligand binding or functional analysis. Minor areas, including specific loop regions and side-chain conformations, may benefit from additional refinement, but the model's overall quality is excellent and suitable for advanced applications.