

Bioinformatics Final Assignment

Fabian Flores Gobet

Universitat della Svizzera Italiana - USI

MSc in Artificial Intelligence

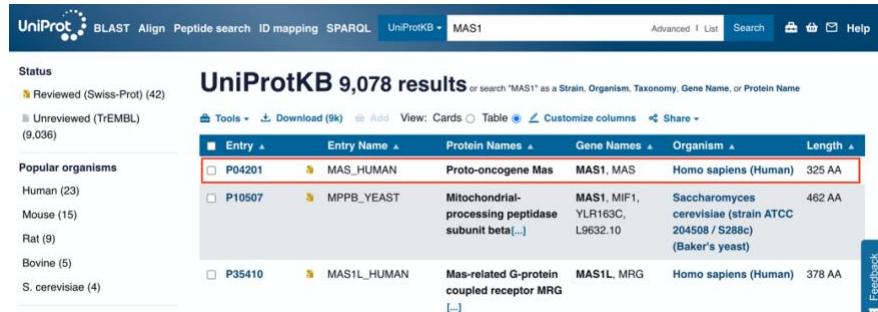
Table of Contents

TEMPLATE SELECTION	3
UNIPROT	3
PSI-BLAST	4
Decidability on sequences of the same subtype	7
Decidability for all sequences	9
HHPRED	10
MODELLER	18
VALIDATION	19
VISUAL CHECK	19
Terminals	19
Intracellular Loops	20
Extracellular Loops	20
Transmembrane helices	20
A-PRIORI ANALYSIS	23
UNIPROT	23
PSIPRED	23
DEEPTMHMM	25
Analysis	25
A-POSTERIORI ANALYSIS	26
Ramachandran plot	27
LOOP MODELLING	36
ALPHAFOLD COMPARISON	40
FINAL SUMMARY	42

TEMPLATE SELECTION

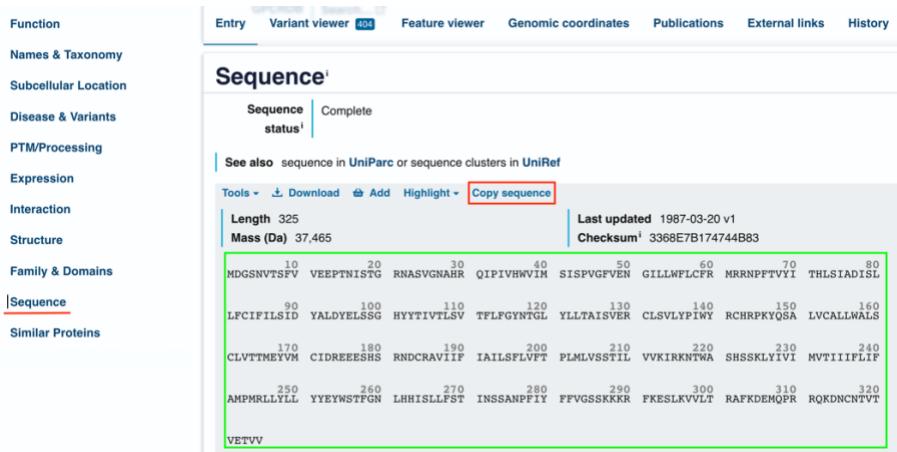
UNIPROT

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**.



The screenshot shows the UniProtKB search results for the query "MAS1". The results table has columns for Entry, Entry Name, Protein Names, Gene Names, Organism, and Length. The first result is P04201, which corresponds to MAS_HUMAN (Proto-oncogene Mas). Other results include P10507 (Mitochondrial-processing peptidase subunit beta) and P35410 (Mas-related G-protein coupled receptor MRG).

Figure 1: UniProtKB search results



The screenshot shows the UniProtKB entry page for P04201 (MAS1). The left sidebar includes links for Function, Names & Taxonomy, Subcellular Location, Disease & Variants, PTM/Processing, Expression, Interaction, Structure, Family & Domains, Sequence (which is highlighted), and Similar Proteins. The main content area shows the protein sequence with numbered positions from 10 to 325. The sequence is:

```

  10  MDGSNVTSTFV VEEPTNSTG 20  RNASVGNNAH 30  QIPIVHWIIM 40  SISPVGFVEN 50  GILLWFLCFR 60  MRRNPFTVYI 70  THLSIADISL
  20  VEEPTNSTG 30  RNASVGNNAH 40  QIPIVHWIIM 50  SISPVGFVEN 60  GILLWFLCFR 70  MRRNPFTVYI 80  LFCIFILSID
  30  RNASVGNNAH 40  QIPIVHWIIM 50  SISPVGFVEN 60  GILLWFLCFR 70  MRRNPFTVYI 80  YALDYELSSG
  40  QIPIVHWIIM 50  SISPVGFVEN 60  GILLWFLCFR 70  MRRNPFTVYI 80  HYYTIVTSLV 100  TFLFGYNTGL
  50  SISPVGFVEN 60  GILLWFLCFR 70  MRRNPFTVYI 80  YLLTAISVER 110  CLSVLVPIWY
  60  GILLWFLCFR 70  MRRNPFTVYI 80  RCHRPKYOSA 120  PLMLVSSTIL
  70  MRRNPFTVYI 80  LVCALLWALS 130  VVKIRKNTWA 140  SHSSKLYIVI 150  MVTIIIFLIF
  80  THLSIADISL 160  CLVTTMEYVM 170  CIDREEEHS 180  RNDCRAVIIF 190  IAILSLFVFTPLM 200  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  90  LFCIFILSID 100  YALDYELSSG 110  HYYTIVTSLV 120  PLMLVSSTIL 210  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  100  YALDYELSSG 110  HYYTIVTSLV 120  PLMLVSSTIL 210  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  110  HYYTIVTSLV 120  PLMLVSSTIL 210  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  120  PLMLVSSTIL 210  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  130  VVFKIRKNTWA 220  SHSSKLYIVI 230  MVTIIIFLIF
  140  SHSSKLYIVI 230  MVTIIIFLIF
  150  MVTIIIFLIF
  160  LVCALLWALS
  170  CLVTTMEYVM
  180  CIDREEEHS
  190  RNDCRAVIIF
  200  IAILSLFVFTPLM
  210  VVFKIRKNTWA
  220  SHSSKLYIVI
  230  MVTIIIFLIF
  240  VVFKIRKNTWA
  250  AMPMRLLLYLL
  260  YYEYWSTFGN
  270  LHHISLLFST
  280  INSSANPFIY
  290  FFVGSSKKR
  300  FKESELKVLT
  310  RAFKDEMOPQR
  320  RQKDNCNTVT
  330  VETVV
  
```

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

The found sequence for MAS1 is:

```

MDGSNVTSTFV VEEPTNSTG RNASVGNNAH QIPIVHWIIM SISPVGFVEN GILLWFLCFRM RRNPF TVYIT
HLSIADISLLFCIFILSIDYALDYELSSGHYYTIVTSLVTFLFGYNTGLYLLTAISVERCLSVLYPIWYRCHRPKY
QSALVCALLWALSCLVTTMEYVMCIDREEEHSRNDCRAVIIFIAILSFLVFTPLMVLVSSTILVVVKIRKNTWAS
HSSKLYIVIMVTIIIFLIFAMPMRLLYLLYEWSTFGNLNHHISLLFSTINSSANPFIYFFVGSSKKRFKESLK
VLTRAFKDEMOPRRQKDNCNTVTETVV
  
```

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).



P04201 · MAS_HUMAN

Protein¹ Proto-oncogene Mas
Gene¹ MAS1
Status¹ UniProtKB reviewed (Swiss-Prot)
Organism¹ Homo sapiens (Human)

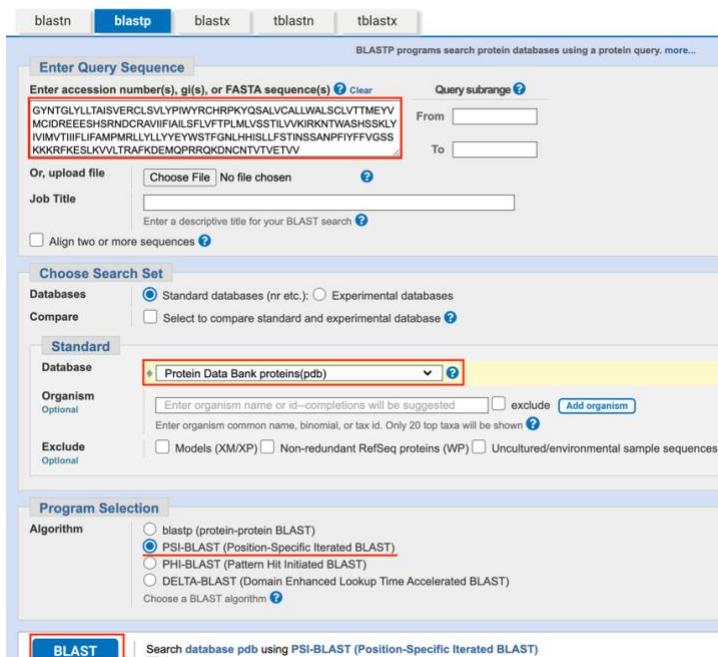
Amino acids 325 (go to sequence)
Protein existence¹ Evidence at protein level
Annotation score¹ 66

Organism names
Taxonomic identifier¹ 9606 NCBI ↗
Organism¹ Homo sapiens (Human)

Function¹
Receptor for angiotensin 1-7 (By similarity).
Acts specifically as a functional antagonist of AGTR1 (angiotensin-2 type 1 receptor), although it up-regulates AGTR1 receptor levels. Positive regulation of AGTR1 levels occurs through activation of the G-proteins GNA11 and GNAQ, and stimulation of the protein kinase C signaling cascade. The antagonist effect on AGTR1 function is probably due to AGTR1 being physically altered by MAS1 By Similarity 2 Publications

Figure 3: Additional information about P04201

I copy the sequence using *UniProtKB* ‘Copy sequence’ functionality and head over *PSI-BLAST* webserver² 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.



blastn **blastp** blastx tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Or, upload file No file chosen
Job Title
 Align two or more sequences ?

Choose Search Set
Databases Standard databases (nr etc.) Experimental databases
Compare Select to compare standard and experimental database ?

Standard
Database Protein Data Bank proteins(pdb)
Organism Enter organism name or id—completions will be suggested exclude
Exclude Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?
 Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection
Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
 Choose a BLAST algorithm ?

BLAST | Search database pdb using PSI-BLAST (Position-Specific Iterated BLAST)

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%.

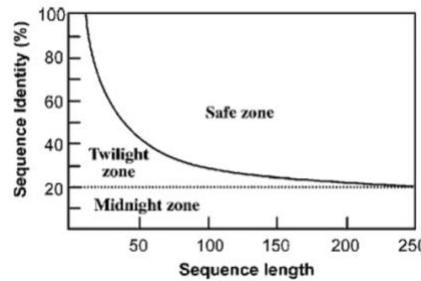


Figure 5: Doolittle's Rule for homology modelling

- **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 is chimeric or not.

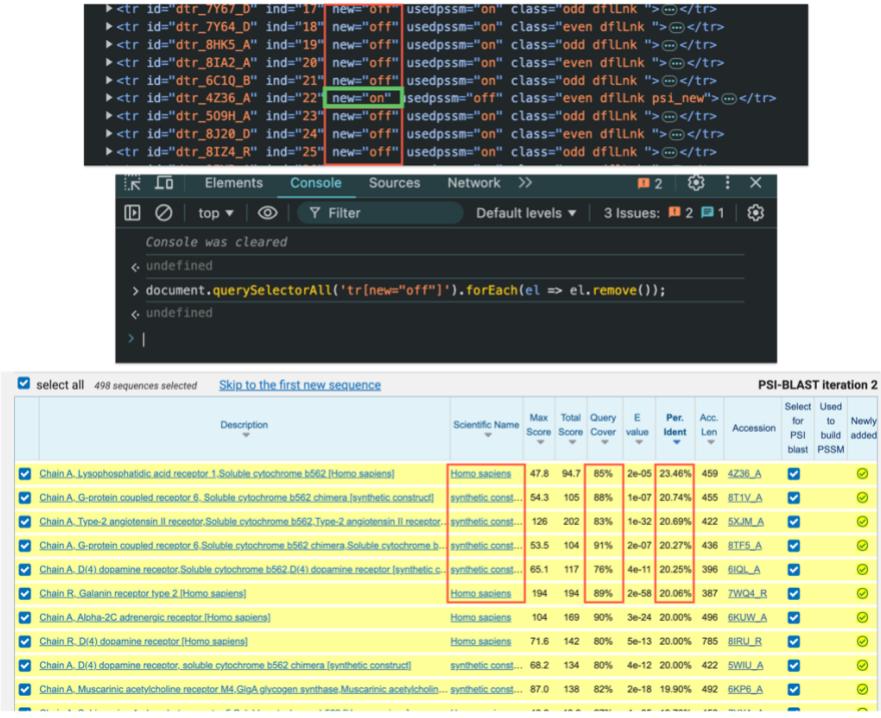
The most restrictive criteria here is the identity percentage. Thus, on each iteration I will sort the results in descending order of identity percentage and see if there is at least one that fits the criteria.

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.

PSI-BLAST iteration 1												
<input checked="" type="checkbox"/> select all 112 sequences selected	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build	Newly added PSSM
<input checked="" type="checkbox"/>	Chain E_Soluble cytochrome b562 Mas-related G-protein coupled receptor member X4 Gree... synthetic cons...		149	149	82%	2e-39	39.11%	710	8K4S_E	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain F_Soluble cytochrome b562 Mas-related G-protein coupled receptor member X1 LgBII... synthetic cons...		172	172	81%	3e-48	38.43%	637	8HJ5_F	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Mas-related G-protein coupled receptor member X1 [Homo sapiens]	Homo sapiens	171	171	82%	2e-50	38.01%	319	8JGB_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Mas-related G-protein coupled receptor member X1 [Homo sapiens]	Homo sapiens	171	171	82%	2e-50	38.01%	322	8JGG_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Mas-related G-protein coupled receptor member X1 [Homo sapiens]	Homo sapiens	171	171	82%	2e-50	38.01%	323	8QWC_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Soluble cytochrome b562 Mas-related G-protein coupled receptor member X4 chim... synthetic cons...		150	150	86%	4e-41	36.14%	472	7SBP_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain C_G-protein coupled receptor 35 [Homo sapiens]	Homo sapiens	49.3	49.3	32%	4e-06	36.04%	308	8H8J_C	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	330	7VDH_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	331	7S8L_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R_Soluble cytochrome b562 Mas-related G-protein coupled receptor member X4 [Homo sapiens]	Homo sapiens	149	149	86%	2e-40	35.09%	472	8YRG_R	<input checked="" type="checkbox"/>		

Figure 6: *PSI-BLAST* 1st iteration

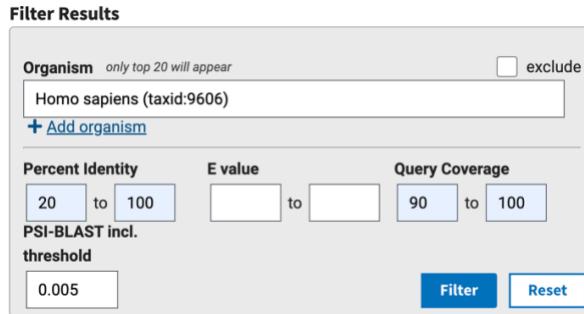
The second iteration of *PSI-BLAST* results in 498 sequences (not all newly added). I need to assess if any of the new added sequences fit my criteria. To retrieve an honourable picture for this report of just the newly added sequences (since this filtering functionality is not implemented in the webserver),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



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 A_Lysophosphatidic acid receptor_1 Soluble cytochrome b562 [Homo sapiens]	Homo sapiens	47.8	94.7	85%	2e-05	23.46%	459	4Z36_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_G-protein coupled receptor_8_Soluble cytochrome b562 chimera [synthetic construct]	synthetic construct	54.3	105	88%	1e-07	20.74%	455	8TIV_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A-Type-2 adenylyl II receptor_Soluble cytochrome b562-Type-2 adenylyl II receptor	synthetic construct	126	202	83%	1e-32	20.69%	422	5XJM_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_G-protein coupled receptor_8_Soluble cytochrome b562 chimera_Soluble cytochrome b-	synthetic construct	53.5	104	91%	2e-07	20.27%	436	8TE5_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_D(4) dopamine receptor_Soluble cytochrome b562 D(4) dopamine receptor [synthetic construct]	synthetic construct	65.1	117	76%	4e-11	20.25%	399	6IQ1_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain R_Galanin receptor type 2 [Homo sapiens]	Homo sapiens	194	194	89%	2e-58	20.06%	387	7WQ4_R	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_Alpha-2C adrenergic receptor [Homo sapiens]	Homo sapiens	104	169	90%	3e-24	20.00%	496	6KUW_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain R_D(4) dopamine receptor [Homo sapiens]	Homo sapiens	71.6	142	80%	5e-13	20.00%	785	8IRU_R	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_D(4) dopamine receptor_solid cytochrome b562 chimera [synthetic construct]	synthetic construct	68.2	134	80%	4e-12	20.00%	422	5WUJ_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A_Muscarinic acetylcholine receptor M4 GtpA glycogen synthase Muscarinic acetylcholin-	synthetic construct	87.0	138	82%	2e-18	19.90%	492	6KP6_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Figure 7: First picture shows the pattern and flags I must look for. Second picture denotes the console and the command³ 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.



Organism only top 20 will appear exclude

[+ Add organism](#)

Percent Identity	E value	Query Coverage
20 to 100	[] to []	90 to 100

PSI-BLAST incl.
threshold
 Filter Reset

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 our 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*.

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 PSSM
<input checked="" type="checkbox"/>	Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	330	7VDH_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R, Mas-related G-protein coupled receptor member X2 [Homo sapiens]	Homo sapiens	137	137	90%	3e-37	35.13%	331	7S8L_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain A, Mas-related G-protein coupled receptor member D [Homo sapiens]	Homo sapiens	152	152	98%	6e-43	34.04%	322	8DQH_A	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	85.5	85.5	92%	3e-18	26.81%	348	8XXV_A	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain A, Prostaglandin D2 receptor 2 [Homo sapiens]	Homo sapiens	85.5	85.5	92%	4e-18	26.81%	346	8XXU_A	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	73.6	73.6	95%	5e-14	24.15%	378	7EJU_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R, fMet-Leu-Phe receptor [Homo sapiens]	Homo sapiens	72.4	72.4	90%	1e-13	23.88%	354	7T6T_R	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	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	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain F, Galanin receptor type 2 [Homo sapiens]	Homo sapiens	41.2	41.2	90%	0.002	22.84%	332	7XJK_F	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/>	Chain R, Melanin-concentrating hormone receptor 1 [Homo sapiens]	Homo sapiens	40.8	40.8	90%	0.003	22.02%	402	8WSS_R	<input checked="" type="checkbox"/>		

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; Hominoidea; 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](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.

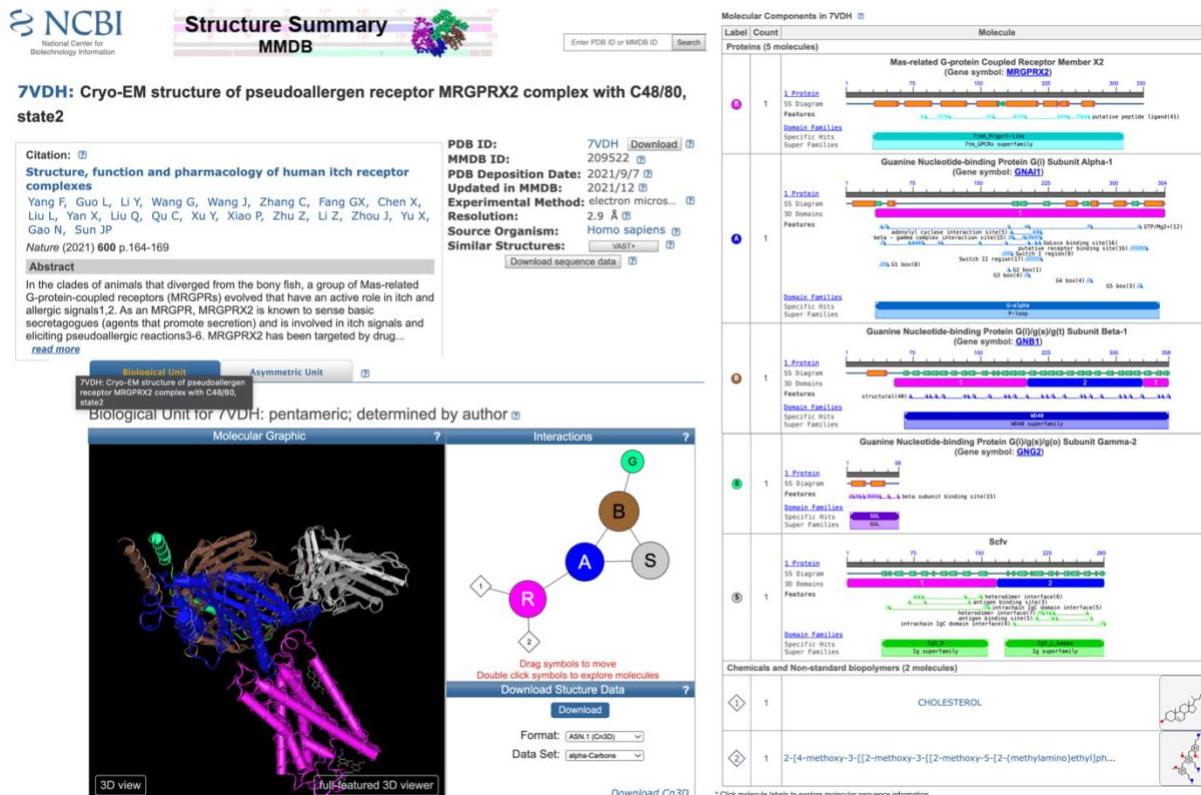


Figure 11: <https://www.ncbi.nlm.nih.gov/Structure/pdb/7VDH>

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.

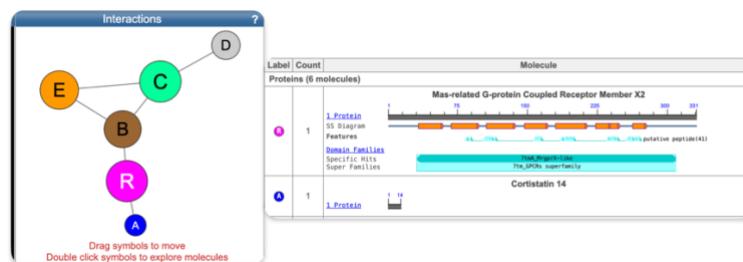


Figure 12: *7S8L_R* information on MMDB.

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.

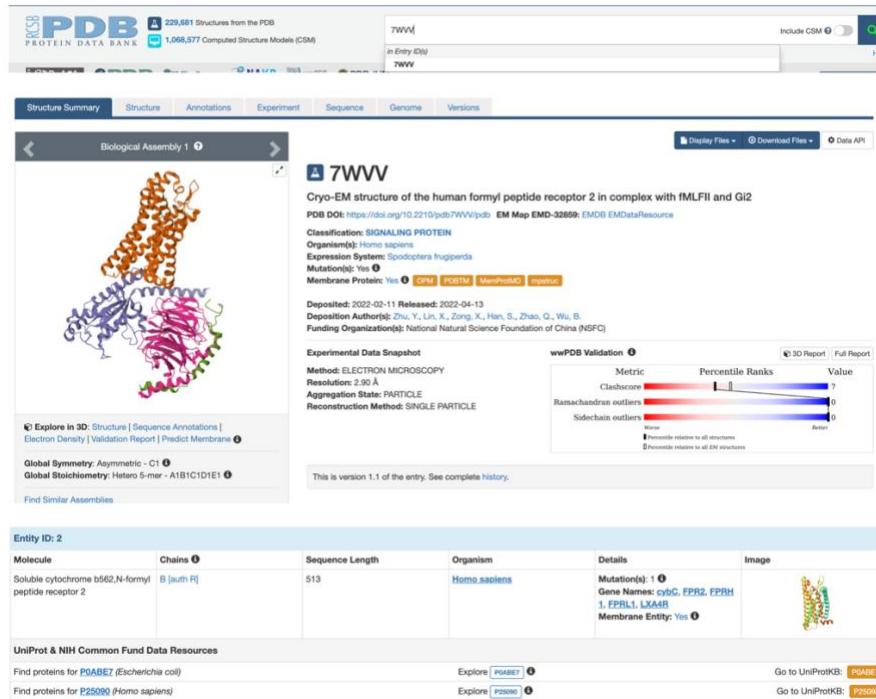


Figure 13: Case example for *7WVV_R*, which is chimeric.

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 webserver⁴ 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>

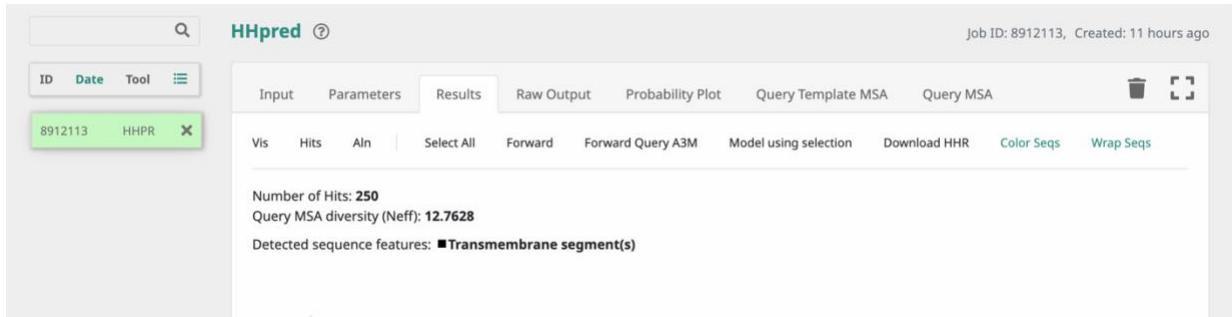


Figure 16: HHPRED job termination.

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*.

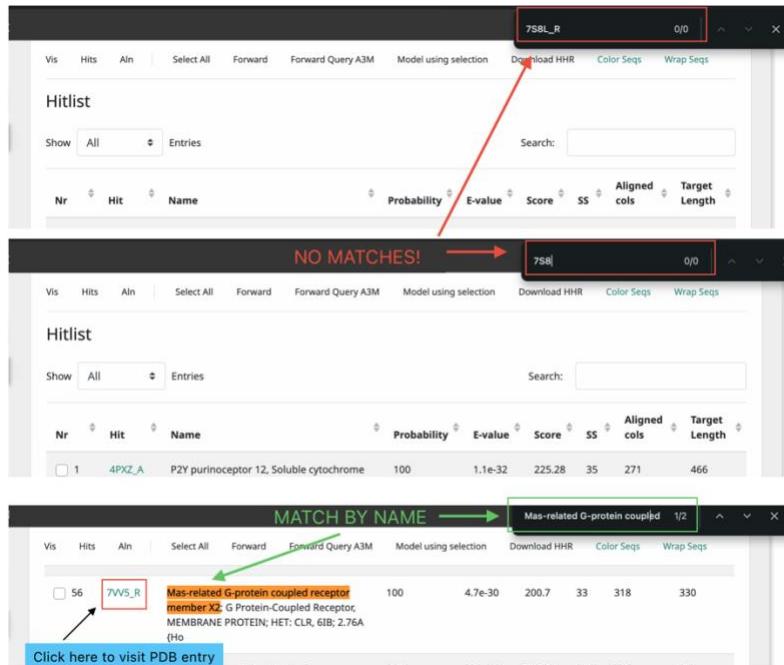


Figure 17: HHPRED crossmatch example with *7S8L_R*.

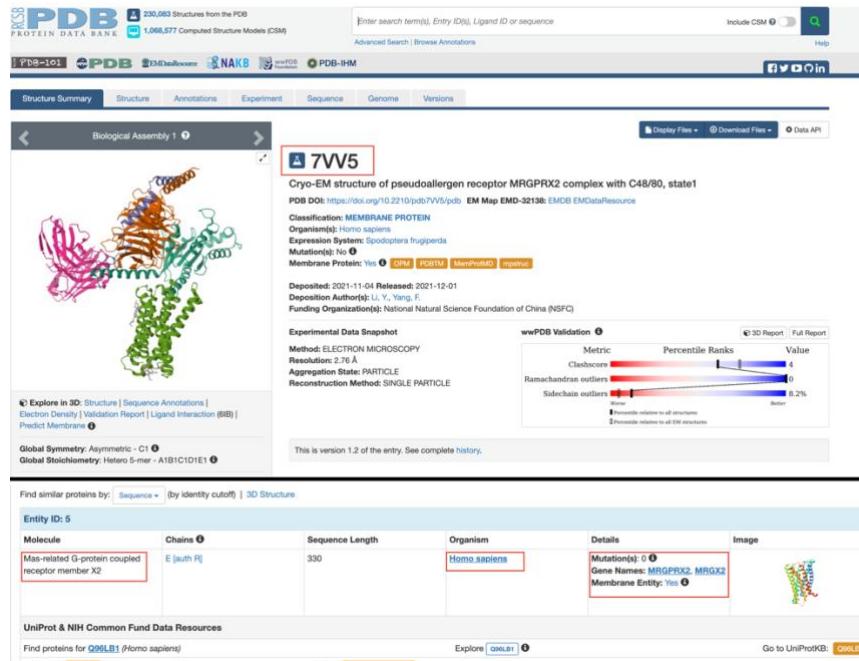


Figure 18: HHPRED crossmatch example with 7S8L_R, visiting PDB entry for information.

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

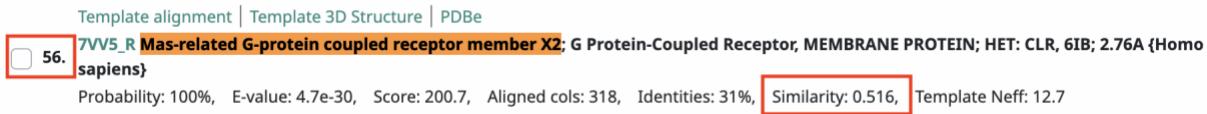


Figure 19: HHPRED crossmatch example with 7S8L_R, entry index and similarity score.

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 (\AA) 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.

<input type="checkbox"/>	56	7VVS_R	Mas-related G-protein coupled receptor member X2; G Protein-Coupled Receptor; MEMBRANE PROTEIN; HET: CLR, 6IB; 2.76A {Ho	100	4.7e-30	200.7	33	318	330
Click here to go to PDB entry page									

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

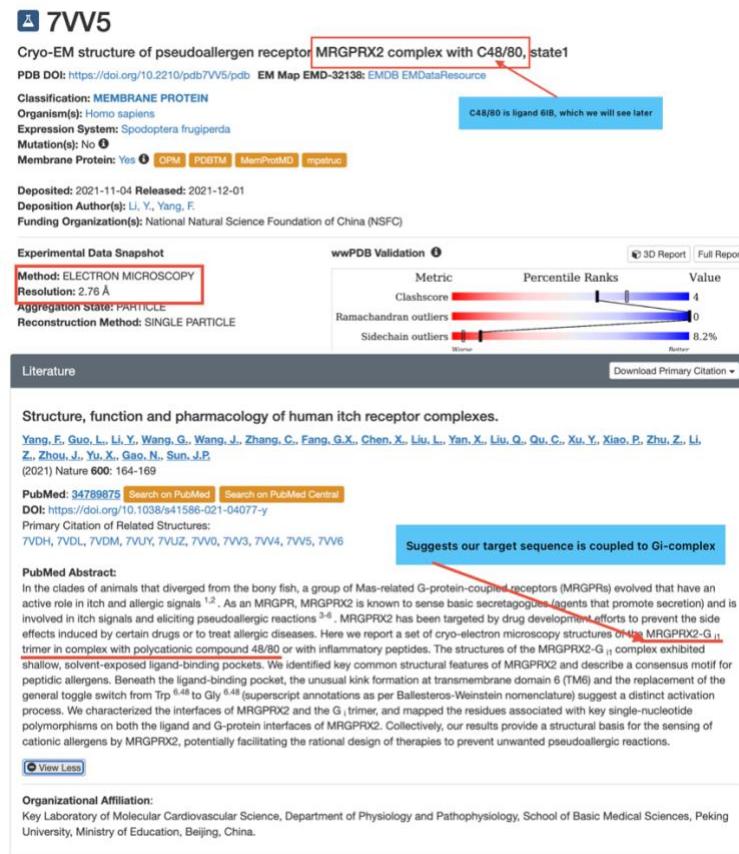


Figure 21: Information gathering example, preliminary information of PDB.

On the bottom of the page is a section for the ligands. For this protein we have two ligands, *6B* and *CLR*. I can click on *6B* 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 *6B* 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)

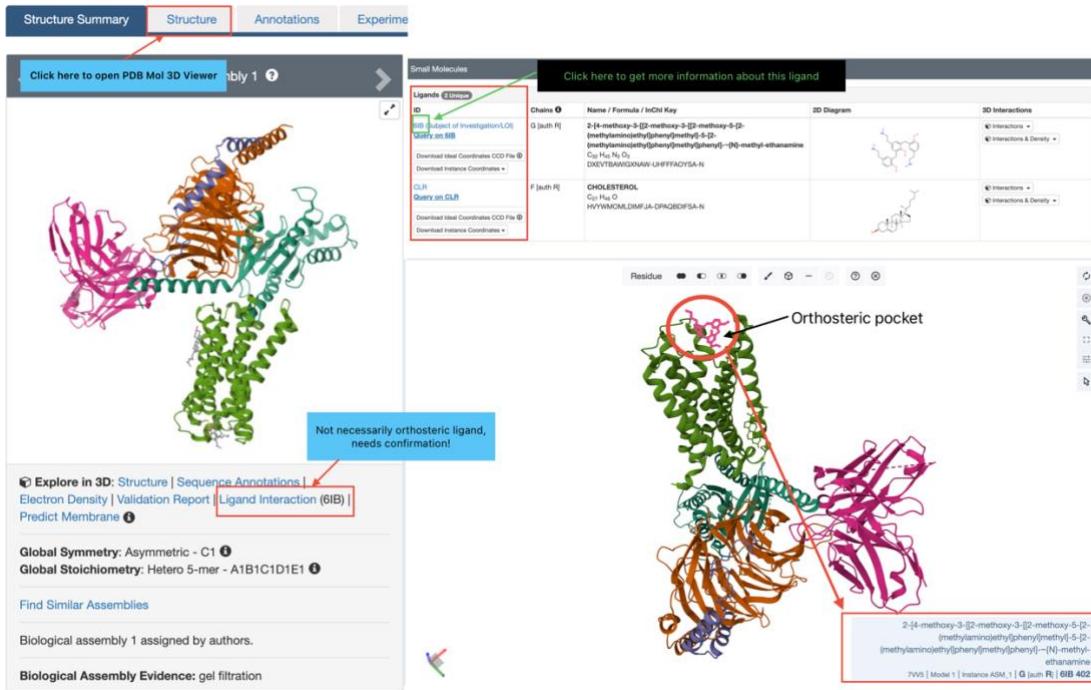


Figure 22: Information gathering example, ligand and 3D analysis.

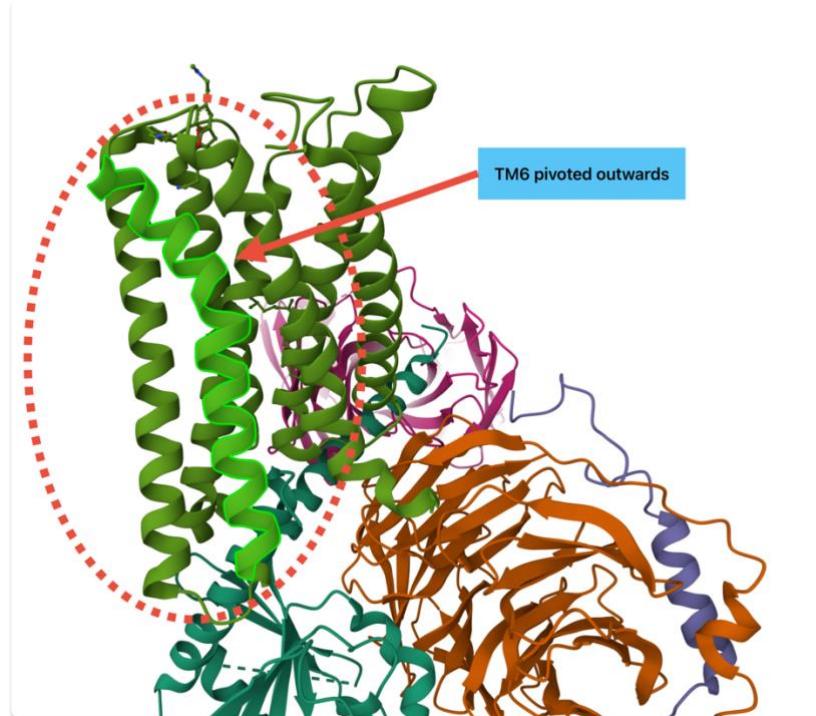
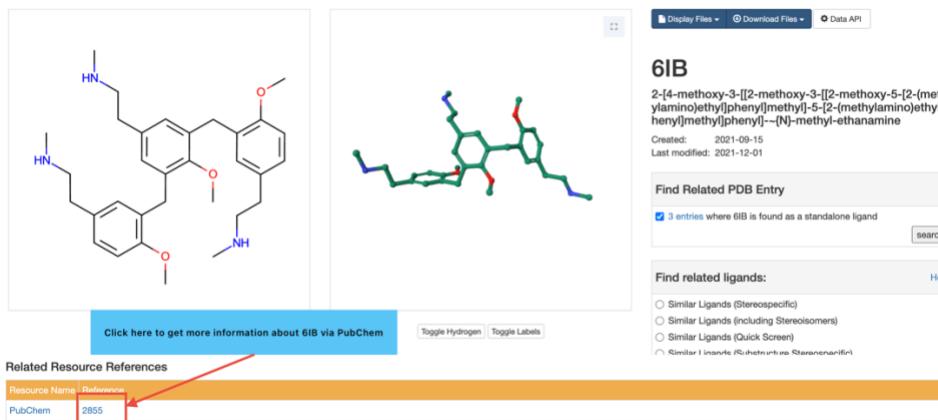


Figure 23: Information gathering example, TM6 positioning.

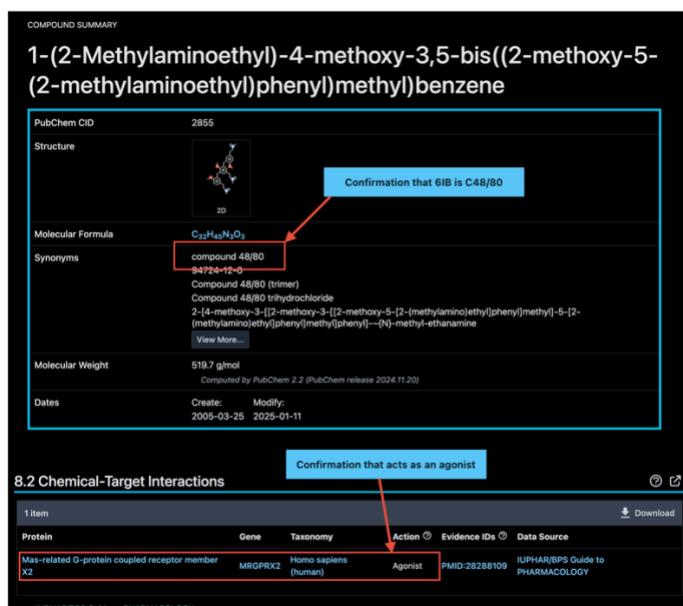
Now I want to confirm if 6IB is the same as C48/80. Relating back to [Figure 22](#), I can click on 6IB hyperlink, and this takes me to the PDB entry regarding this ligand. From here I can infer some information about it, but nothing that immediately tells me it is the same as C48/80. At the bottom of the page there is a hyperlink to the respective PubChem entry⁵.



The screenshot shows the PDB entry for 6IB. It features a 2D chemical structure on the left and a 3D ball-and-stick model on the right. Below the structures are several buttons: "Display File", "Download File", and "Data API". The chemical structure is labeled "Click here to get more information about 6IB via PubChem". There are also "Toggle Hydrogen" and "Toggle Labels" buttons. At the bottom, there's a "Related Resource References" section with a table. The table has columns for "Resource Name" and "Reference". A red arrow points from the "Reference" column to the number "2855", which is highlighted with a red box.

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.



The screenshot shows the PubChem compound summary for 6IB (PubChem CID: 2855). It includes the following details:

- Structure:** Shows the chemical structure of 6IB.
- Molecular Formula:** C21H24N2O3
- Synonyms:** Includes "compound 48/80", "Compound 48/80 (trimer)", "Compound 48/80 trihydrochloride", and "2-[4-methoxy-3-[(2-methoxy-5-[(2-(methylamino)ethyl]phenyl)methyl]-5-[2-(methylamino)ethyl]phenyl]-(N)-methyl-ethanamine".
- Molecular Weight:** 519.7 g/mol
- Dates:** Create: 2005-03-25, Modify: 2025-01-11

A blue box highlights the synonym "compound 48/80" with the text "Confirmation that 6IB is C48/80". Another blue box highlights the "Action" column in the "8.2 Chemical-Target Interactions" table with the text "Confirmation that acts as an agonist".

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.

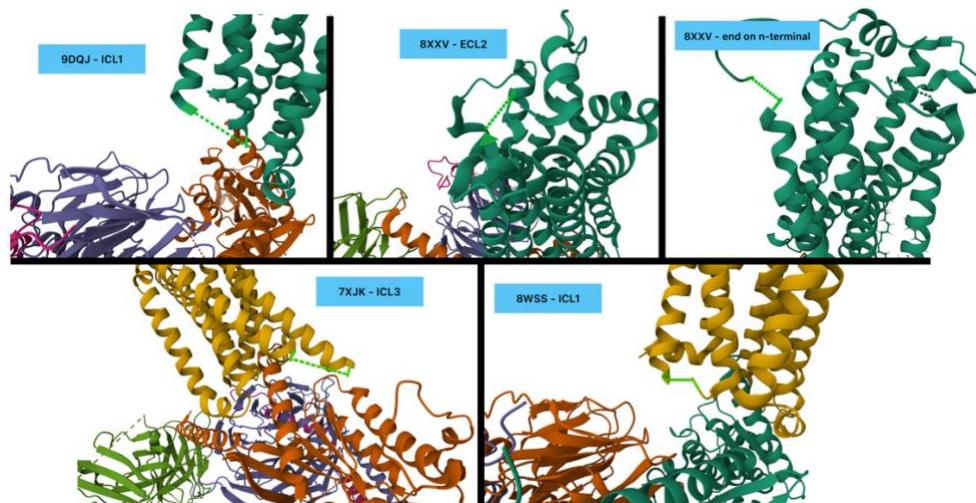


Figure 26: Preliminary selection missing structures.

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.

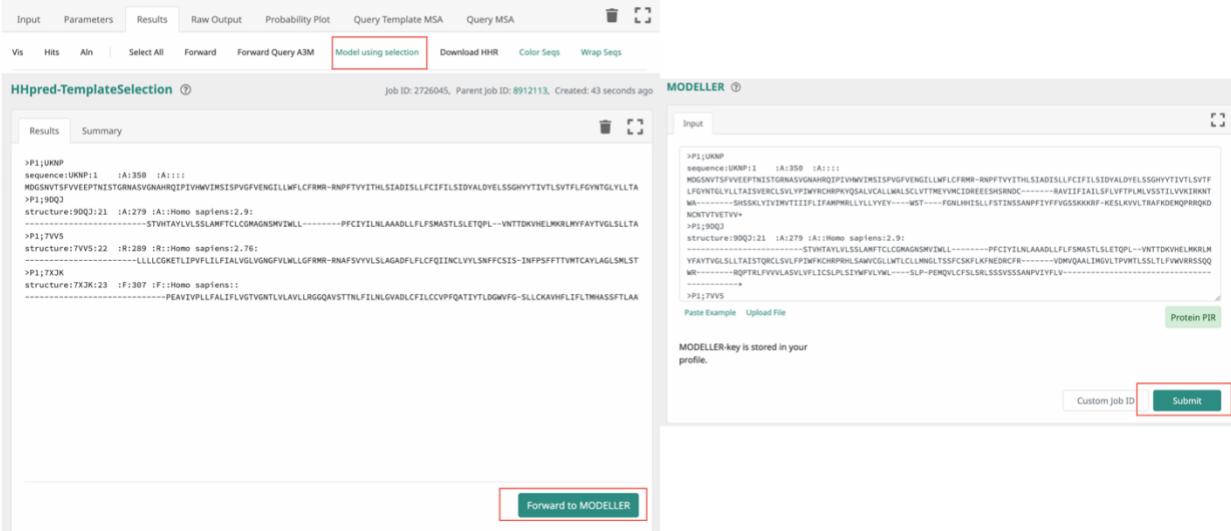
index	New PDB	% identity	% similarity	% query cover	Missing	Resolution Å
1	7VV5	35.13	51.60	90.00	-	2.76
2	9DQJ	34.04	46.70	98.00	ICL1	2.90
4	7T6T	24.15	26.90	95.00	-	3.20
5	7XJK	22.84	21.15	90.00	ICL3	3.30
6	8WSS	22.02	20.01	90.00	ICL1	3.01

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.



The screenshot shows two adjacent web pages. On the left, the 'HHpred-TemplateSelection' page displays a sequence alignment between four template proteins (P1;1;UKNP, P1;1;90Q3, P1;1;7V5, P1;1;7XK) and the target protein (P1;1;90Q2). The alignment highlights identical and similar amino acids. On the right, the 'MODELLER' page shows the input sequence from the alignment. A red box highlights the 'Submit' button at the bottom right of the MODELLER interface.

Figure 27: Submission to MODELLER

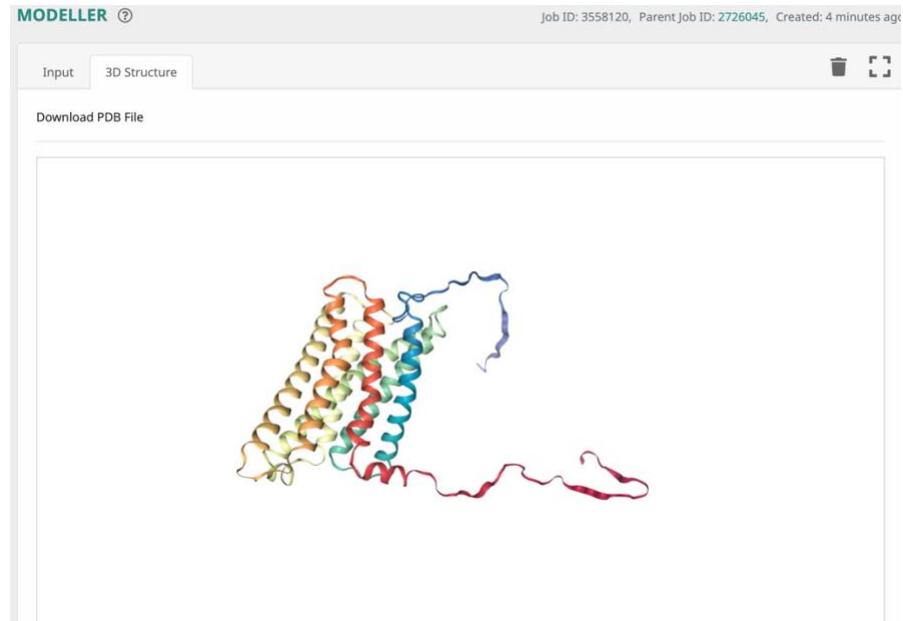


Figure 28: Result from MODELLER

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.

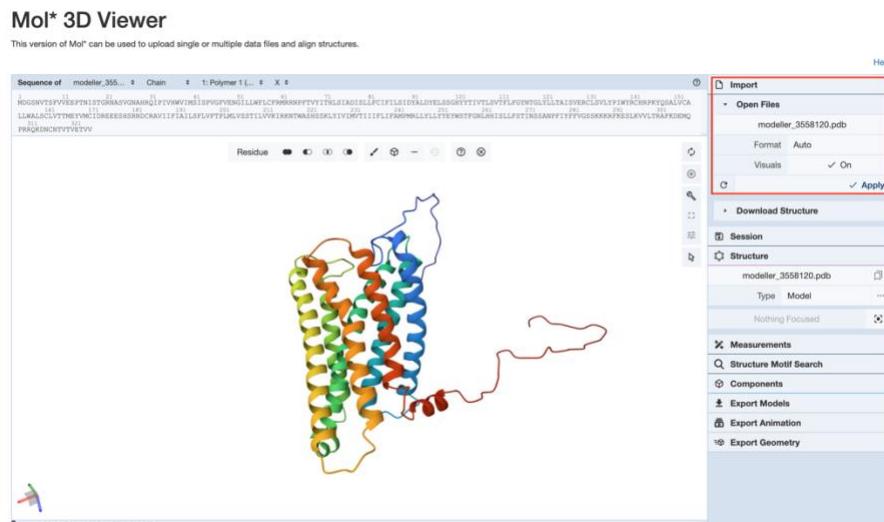


Figure 29: <https://www.rcsb.org/3d-view>

Terminals

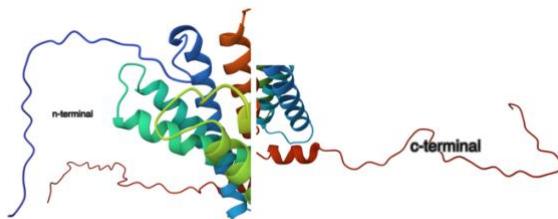


Figure 30: c-terminal and n-terminal

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

Intracellular Loops

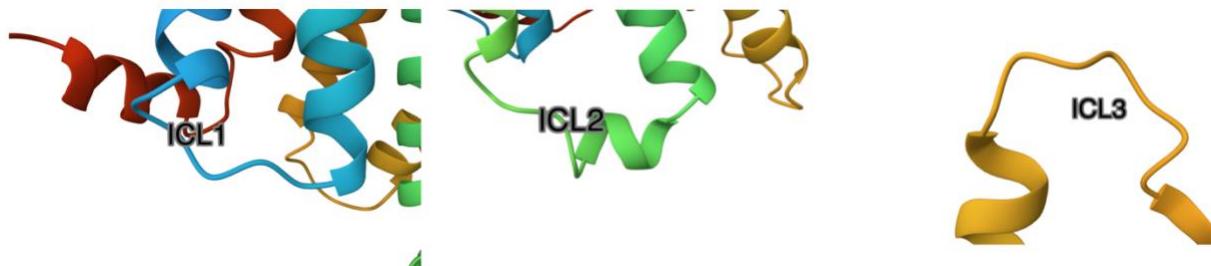


Figure 31: Model intracellular loops.

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

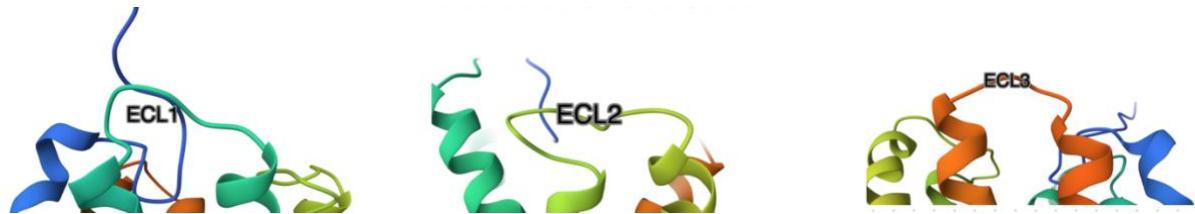


Figure 32: Model extracellular loops.

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

Transmembrane helices

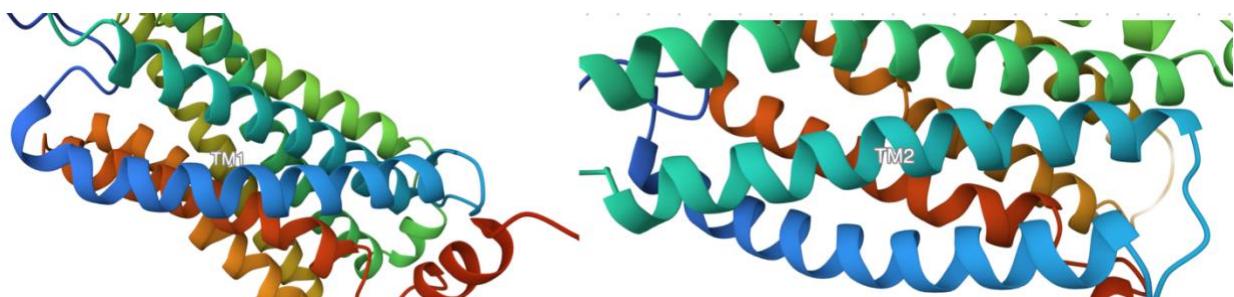


Figure 33: Model TM1 and TM2

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.

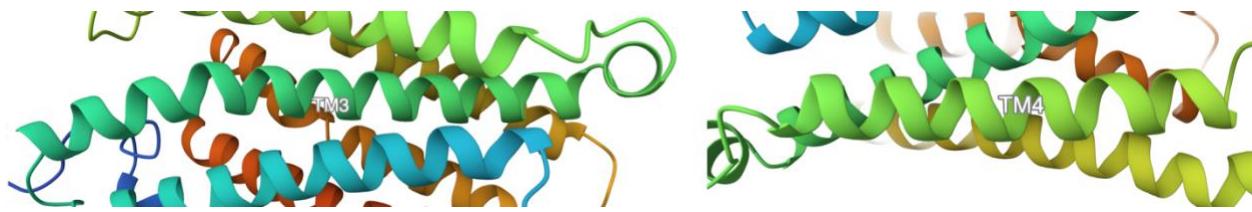


Figure 34: Model TM3 and TM4.

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.

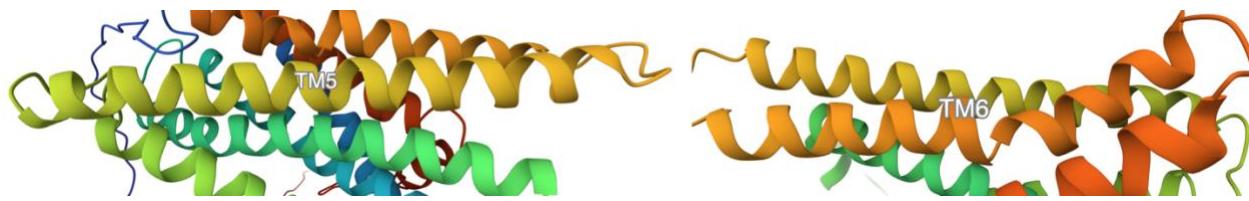


Figure 35: Model TM5 and TM6.

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.

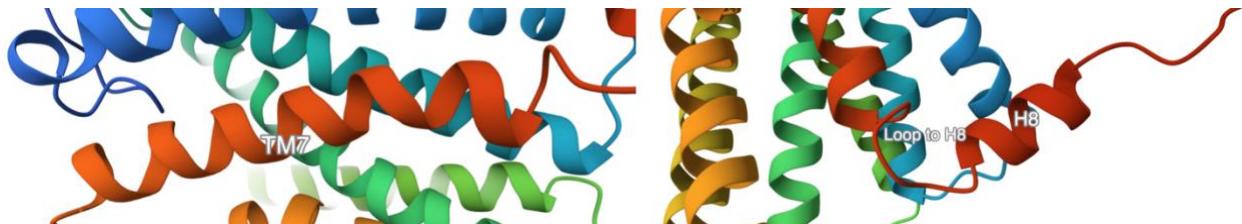


Figure 36: Model TM7, loop to H8 and H8.

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.

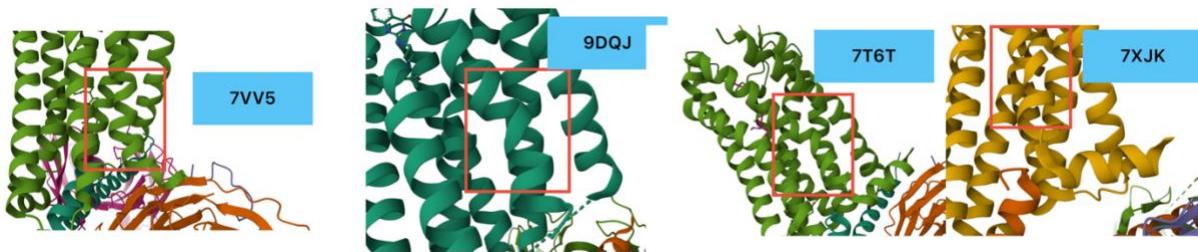


Figure 37: Templates bent on TM

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

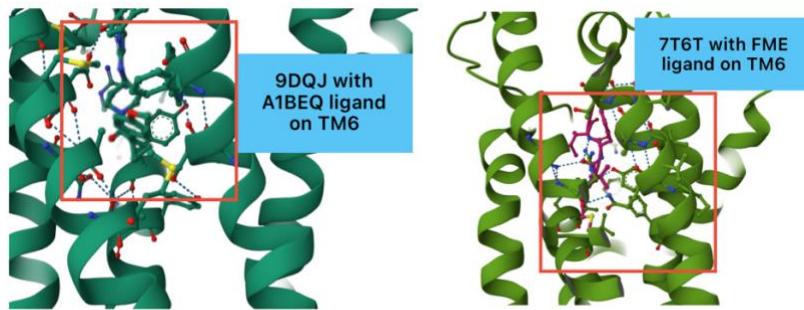


Figure 38: 9DQJ and 7T6T ligands on TM6 bent.

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.

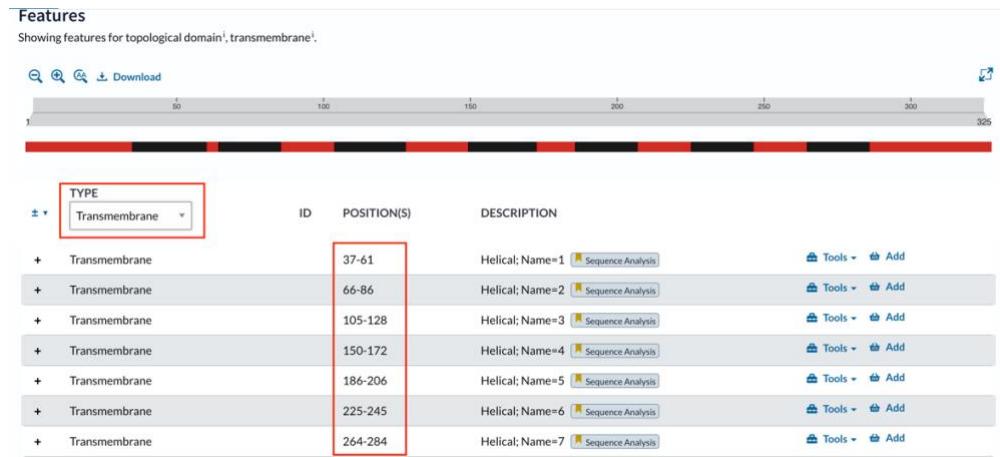


Figure 39: P04201 TM start-end residues.

PSIPRED

Next, I go to PSIPRED⁶, I select both PSIPRED and MEMSAT-SVM from the analysis menu, paste the sequence of **P04201** and click submit.

The figure shows a screenshot of the PSIPRED analysis submission interface. The interface is divided into several sections:

- Popular Analyses:** Includes checkboxes for PSIPRED 4.0 (Predict Secondary Structure) and MEMSAT-SVM (Membrane Helix Prediction), which are both checked.
- Structure Modelling:** Includes a checkbox for DMPFold 2.0 (Protein Structure Prediction).
- Single Sequence Prediction:** Includes a checkbox for SiAPred (Single Sequence SS Prediction).
- Contact Analysis:** Includes a checkbox for DeepMetaPSICOV 1.0 (Structural Contact Prediction).
- Fold Recognition:** Includes checkboxes for genTHREADER (Rapid Fold Recognition) and pDomTHREADER (Protein Domain Fold Recognition).
- Domain Prediction:** Includes a checkbox for DomPred (Domain Boundary Prediction).
- Function Prediction:** Includes checkboxes for FFPred 4 (Eukaryotic Function Prediction) and DMPmetal (Metal Binding Site Prediction).
- Submission details:** A button labeled "Submit" is located here.
- Protein Sequence:** A text input field containing the protein sequence: MGSNVTSPVVEEPTNISTGRNASVGNAHRQIPIVHWVIMSISPVGFVENGILLWFLCFCRMRRNPFTVYIHLISIADISLLFCIFILSIDYALDYELSSGHYYITVTLSTVFLFG YNTGLYLLTAISVERCLSVLYPIWRCHRPKYQSALVCALLVALSCLVTTMEYMCIDREEEHSRNDRCRAVIIIFIALSFLVFTPLMLVSSTILVVKIRKNTWASHSSKLYIVI MVTIIIFLIFAMPMLLRLYYEYWSTFGNLHISLFLSTINSSANPFIYFFVGSSKKRKFESLKVVLTRAFTKDEMOPRRQKDNCNTVTVETV\|.

Figure 40: PSIPRED analysis submission.

⁶ <http://bioinf.cs.ucl.ac.uk/psipred/>

Sequence Plot



Figure 41: PSIPRED results.

Sequence Plot

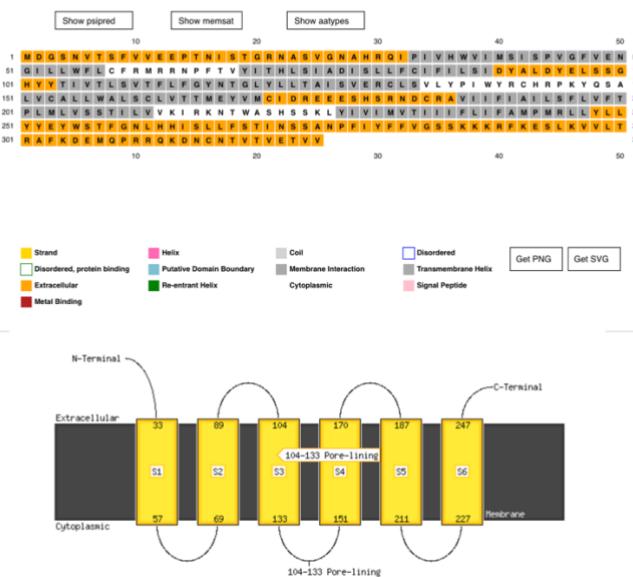
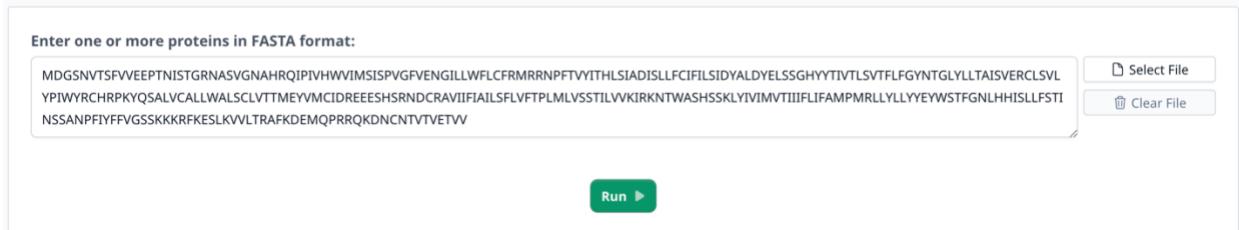


Figure 42: MEMSAT-SVM results.

DEEPTMHMM

Finally, I go to DEEPTMHMM⁷, paste the sequence of P04201 and click run.



Enter one or more proteins in FASTA format:

```
MDGSNVTSFVVEEPTNISTGRNASVGNNAHRQIPIVHWVIMISPVGFVENGILLWFLCFMRMRNPFTVYTHLSIADISLLFCIFILSIDYALDYELSSGHYTIVTLSVTFLFGYNTGLYLLTAISVERCLSVL
YPIWYRCHRPKYQSALVCALLWALSCLVTTMEYVMCIDREEEHSRNDCRAVIIJIALSFLVFTPLMLVSSTILVVKIRKNTWASHSSKLYIVIMVTIIIFLIFAMPMRLLYLYEYWSTFGNLHHISLLFSTI
NSSANPFIYFFVGSSKKRKFESLKVVLTRAFKDEMQPQQRKDNCTNTVETVV
```

Run ➔

Select File

Clear File

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				avg.
	start	end	start	end	start	end	start	end	start	end	
H1	8	3	0	0	4	-1	8	-1			3.125
H2	1	-7	0	0	4	-4	4	-4			3.000
H3	5	-7	1	-2	4	-2	8	-7			4.500
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			

legend:
 abs(diff) ≤ 2 (green)
 3 ≤ abs(diff) ≤ 4 (yellow)
 abs(diff) ≥ 5 (red)

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 LAB⁸. Once I access this webserver, I start by submitting the .pdb file I got from MODELLER and press 'Run programs'.

UCLA-DOE LAB – SAVES v6.1

To run any or all programs:
upload your structure, in PDB format only

The server is slower, please be patient. Send any questions or complaints to

modeler_3558120.pdb
Customize job name:
modeler_3558120.pdb

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.

UCLA-DOE LAB – SAVES v6.1

Job 140427 has been created

[New Job](#)

job #140427: modeller_3558120.pdb [job link] [3D Viewer]

ERRAT	Verify3D	PROVE
Analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures. Start	Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. Start	Temporarily down at the moment
WHATCHECK	PROCHECK	OPEN
Derived from a subset of protein verification tools from the WHATIF program (Vriend, 1990), this does extensive checking of many stereochemical parameters of the residues in the model. Start	Checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. Start	We are open to suggestions for a 6th program to operate in this window. If you know of a program that we could run locally on our server that would be most useful, please let us know: email holton at mbi dot ucla dot edu with your suggestion

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

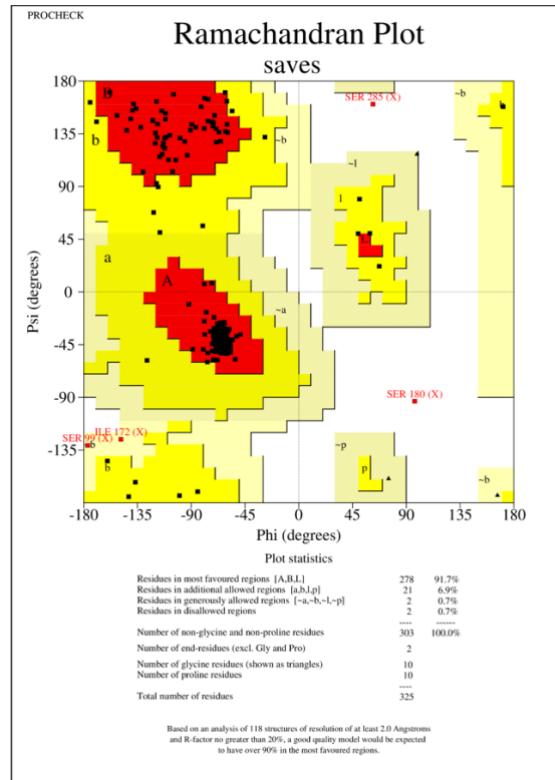


Figure 49: Ramachandran plot of my model.

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 $\geq 90\%$ 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.

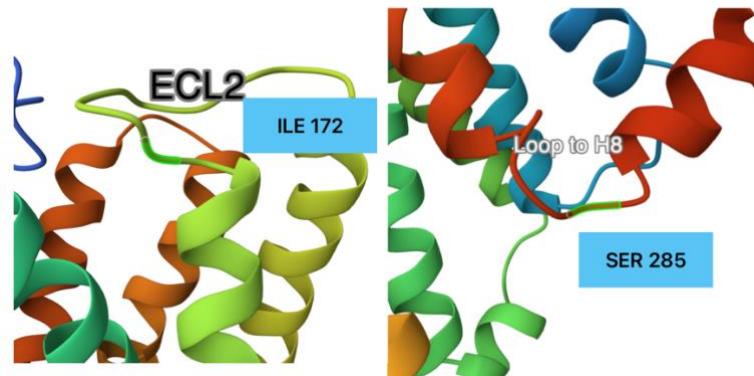


Figure 50: ILE 172 and SER 285 locations

ILE 172 and SER 285 are determined to be in disallowed regions in the Ramachandran plot, but they are contained within 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

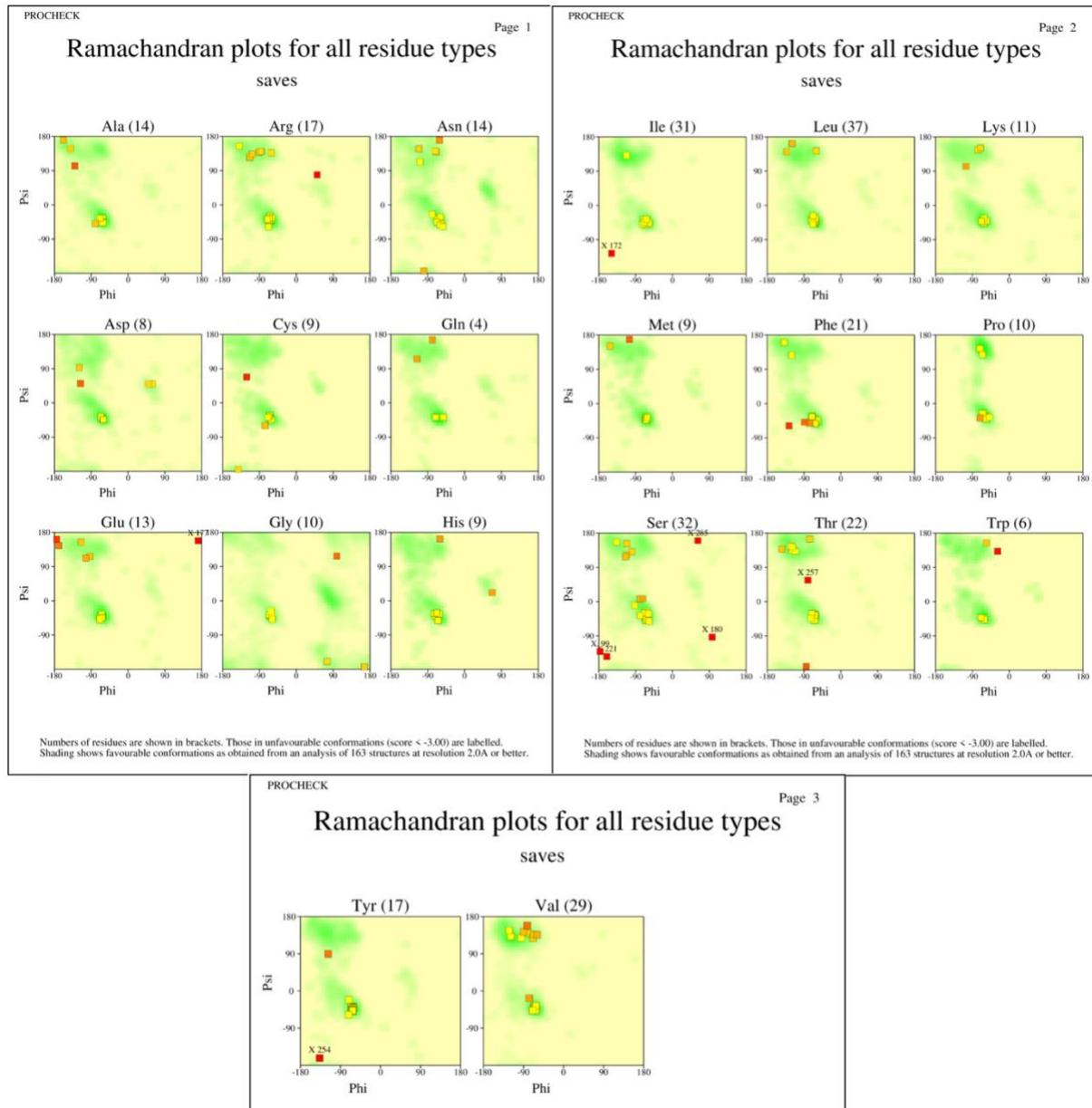


Figure 51: Ramachandran plots for all residues of my model.

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 $\leq 2.0 \text{ \AA}$).
- 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.

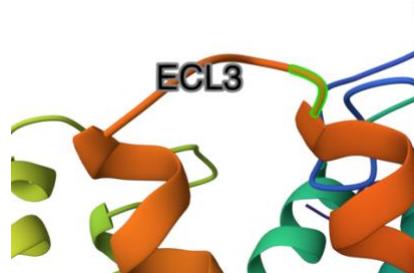


Figure 53: THR 257

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 (χ_1 and χ_2) 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):

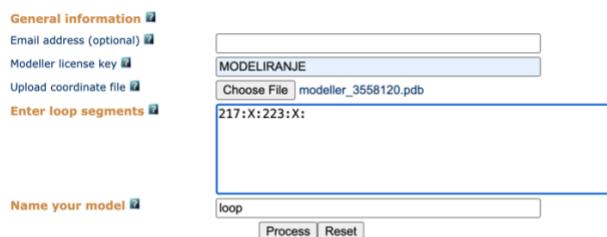
- A few residues are flagged as outliers, requiring further investigation.

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 webserver⁹, 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.

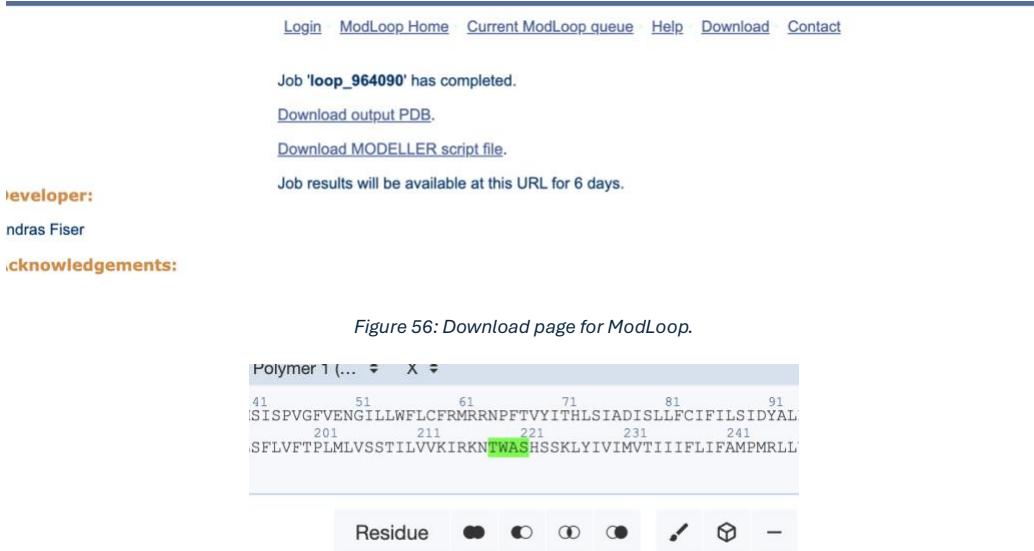


The screenshot shows a web-based form for job submission. On the left, there's a sidebar with "General information" and links for "Email address (optional)", "Modeler license key", "Upload coordinate file", and "Enter loop segments". The "Enter loop segments" link is highlighted in orange. The main area has several input fields: a text input for "MODELIRANJE", a file selection input for "Choose File | modeller_3558120.pdb", and a large text input for "217:X:223:X:" which is highlighted with a blue border. At the bottom, there's a "Name your model" input with "loop" typed in, and two buttons: "Process" and "Reset".

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.



The screenshot shows the 'ModLoop Home' page with the following content:

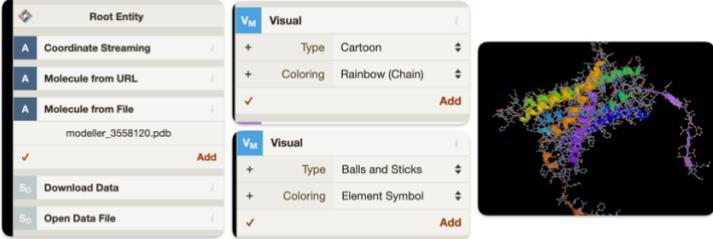
- [Login](#) [ModLoop Home](#) [Current ModLoop queue](#) [Help](#) [Download](#) [Contact](#)
- Job 'loop_964090' has completed.
- [Download output PDB.](#)
- [Download MODELLER script file.](#)
- Job results will be available at this URL for 6 days.
- Developer:** ndras Fiser
- Acknowledgements:**

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 LiteMol¹⁰, added cartoons for balls and sticks and started analysing them.



Root Entity

- Coordinate Streaming
- Molecule from URL
- Molecule from File
- modeler_3558120.pdb
- Add**
- Download Data
- Open Data File

V_M Visual

- Type: Cartoon
- Coloring: Rainbow (Chain)
- Add**

V_M Visual

- Type: Balls and Sticks
- Coloring: Element Symbol
- Add**

Figure 58: LiteMol initializer

¹⁰ <https://glyco3d.cermav.cnrs.fr/LitemolViewer/advancedView.php>

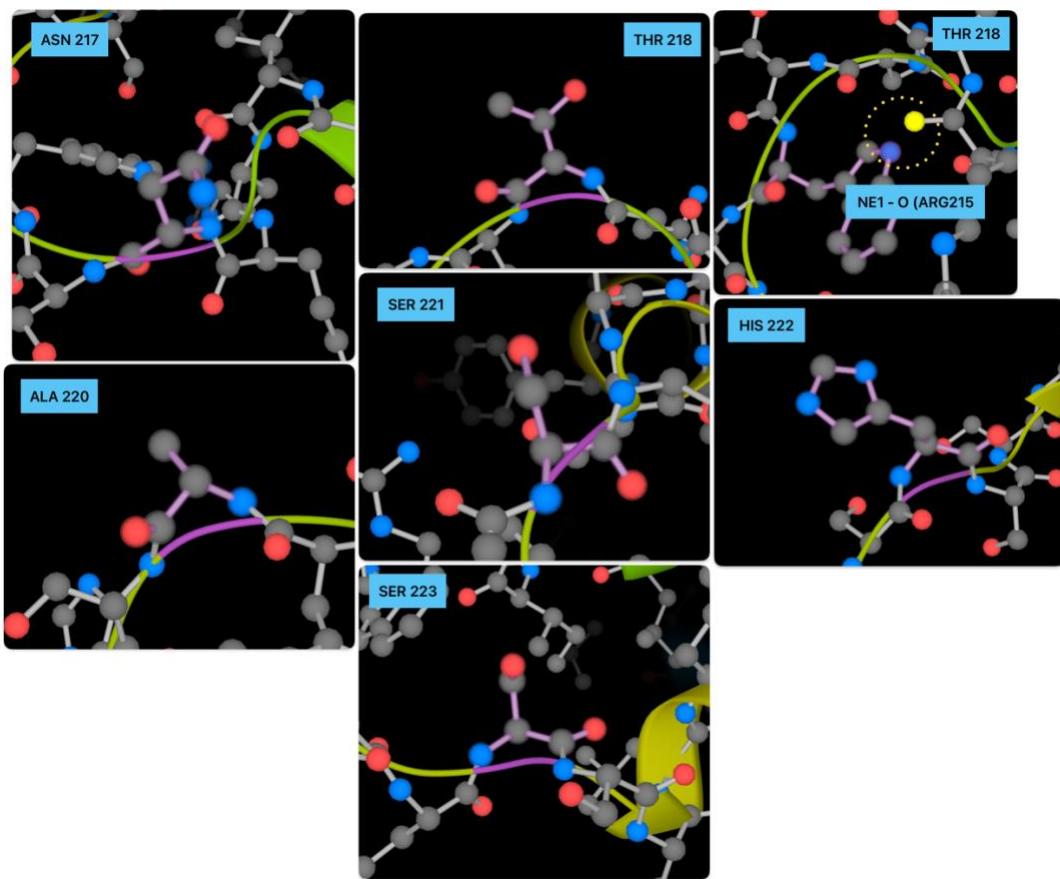


Figure 59: My model and side chain interactions in ICL3.

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

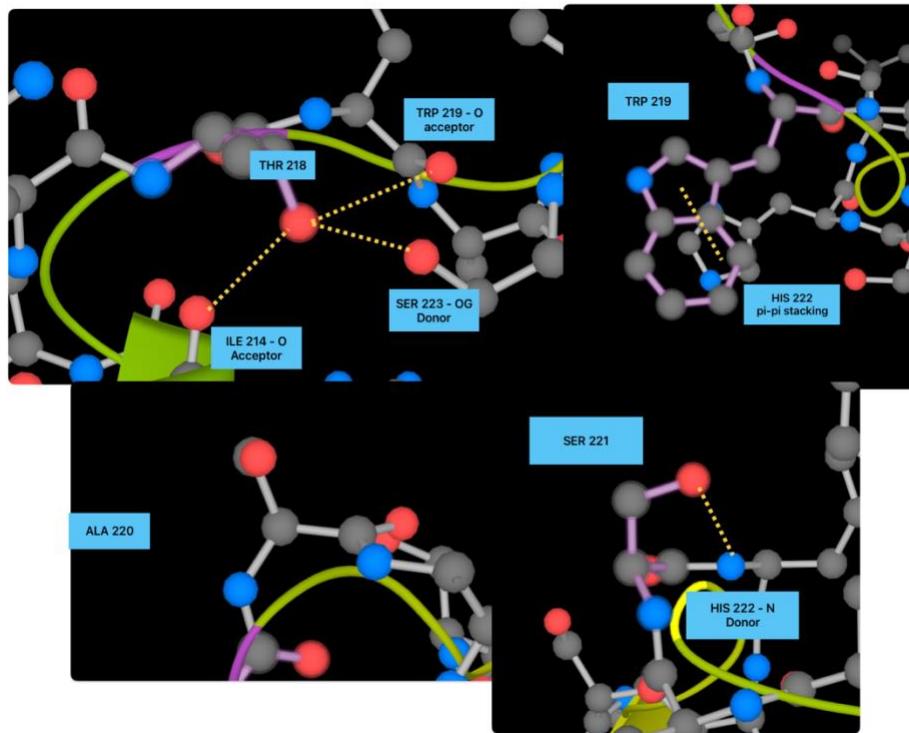


Figure 60: ModLoop model ICL3 interactions

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 webserver¹¹, paste the sequence of **P04201** and submit the job.

The screenshot shows the AlphaFold webserver interface. At the top, there are dropdown menus for 'Entity type' set to 'Protein' and 'Copies' set to '1'. Below this is a text input area containing the amino acid sequence of P04201, with residue numbers 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 325, and 335 aligned in a grid. At the bottom left is a button '+ Add entity', and at the bottom right is a button 'Save job'.

Figure 61: Alphafold sequence submission.

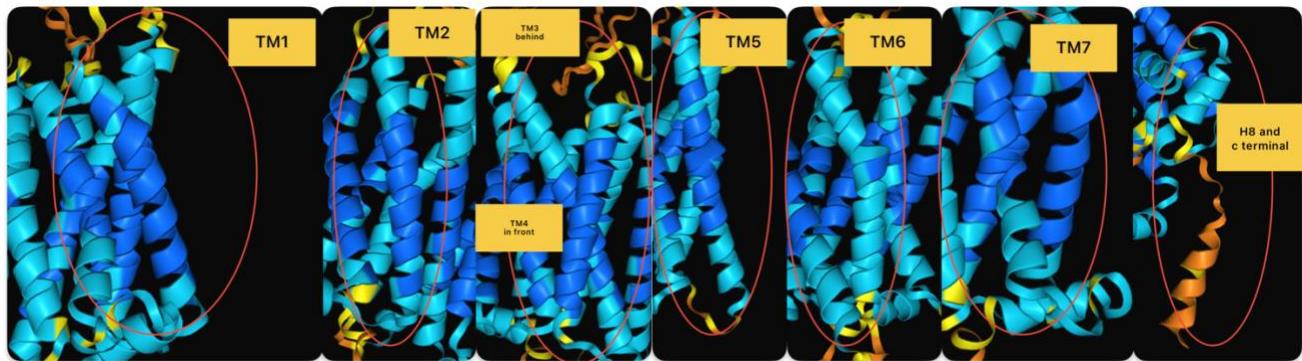


Figure 62: AlphaFold visual check.

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 bends. 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 bends. Overall, the barrel like structure of the GPCR is achieved.

¹¹ <https://alphafoldserver.com/>

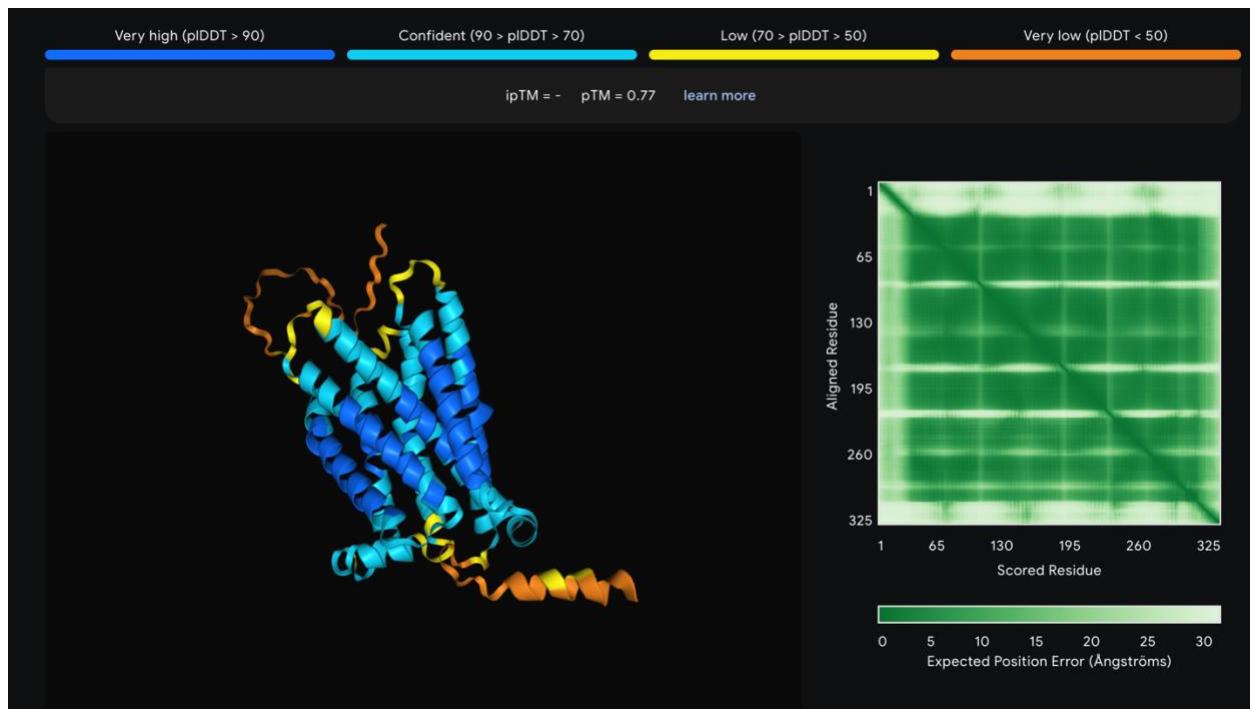


Figure 63: Results from AlphaFold.

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 ($p\text{LDdT} > 90$) representing very high confidence, light blue regions ($70 < p\text{LDdT} \leq 90$) indicating confident predictions, yellow regions ($50 < p\text{LDdT} \leq 70$) showing low confidence, and orange/red regions ($p\text{LDdT} \leq 50$) 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.