# Towards a greener AlphaFold2 protocol for protein complex modeling: Insights from CAPRI Round 55

Büşra Savaş[1,2], İrem Yılmazbilek[1], Atakan Özsan[1], Ezgi Karaca[1,2]

[1]İzmir Biomedicine and Genome Center, İzmir, Türkiye; [2]Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir, Türkiye

*Corresponding Author: Ezgi Karaca (ezgi.karaca@ibg.edu.tr)*

## ABSTRACT

In the 55th round of CAPRI, we primarily used enhanced AlphaFold2 (AF2) sampling and data-driven docking. Our AF2 protocol relies on Wallner's aggressive sampling approach, which combines different AF2 versions and sampling parameters to produce thousands of models per target. For T231 (an antibody peptide complex) and T232 (PP2A:TIPRL complex), we employed a 50-fold reduced mini aggressive sampling and a custom ranking approach, leading to a top-ranking medium prediction in both cases. In the cases of T233 and T234 (two antibody bound MHC I complexes), we followed data-driven docking, which did not lead to an acceptable solution. Our post-CAPRI55 analysis showed that if we would have used our mini aggressive sampling on T233 and T234, we could have submitted a medium-quality model for T233 as well. For the scoring challenge, we utilized the scoring function of FoldX, which was effective in selecting acceptable models for T231 and medium quality models for T232. Our success, especially in predicting and ranking a medium quality model in the case of antibody-peptide T231 underscores the feasibility of greener and accurate enhanced AF2 sampling. It also highlights the ongoing need for alternative ranking approaches for antibody complexes.

## INTRODUCTION

In the CASP14 competition, AlphaFold2 (AF2) achieved an unprecedented success in producing highly accurate protein structural models [1]. The release of AF2's source code motivated the community to adapt AF2's use in multimer modeling. In response to this interest, DeepMind extended the training regime of AF2 to support protein multimer modeling (AFv2.1) [2]. The interfacial clash issues produced by this version was resolved in AF v2.2. The next and final AF2 version was announced after the CASP15-CAPRI round, where the community achieved an extraordinary 90% success rate in multimer modeling by using AF2 [3]. In CASP15-CAPRI, one of the top-performers, the Wallner team, devised an aggressive sampling strategy to generate thousands of models. For this, they employed AF v2.1 and v2.2 and played with dropout and template use, while increasing the number of recycles and seeds [4,5]. The latter two was adopted in AF v2.3 [6]. Expanding on this, in CAPRI55, we followed a 50-fold reduced, i.e., mini version of Wallner's aggressive sampling to predict the binding of two targets (T231-2) and data-driven docking to model the other two (T233-4). This effort led to top-ranking medium quality predictions for T231 and T232 and incorrect models for T233 and T234. In the scoring round, we ranked the model pool with the interface scoring of FoldX, which led to the selection of acceptable and medium quality models for T231 and T232 targets, respectively [7]. In the following sections, we provide a detailed analysis of our performance in CAPRI55, highlighting our strategies and their outcomes.

## METHODS

### Prediction Category

The summary of our modeling strategies is illustrated in Figure S1.

- Throughout the 55th round, the ColabFold web resource was accessed through Colab Notebook: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb [8]. In the case of T232, half of our AF2 runs were generated on a lab-scale workstation equipped with an A4000 GPU.
- AF3 was run on https://alphafoldserver.com/ [9]
- HADDOCK was accessed via https://rascar.science.uu.nl/haddock2.4/submit/1 and its refinement protocol via https://rascar.science.uu.nl/haddock2.4/refinement/1 [10].
- AF2Rank is accessed via a Colab Notebook, introduced in https://github.com/jproney/AF2Rank [11].
- The molecular dynamics (MD) simulations were run using GROMACS (2020.4) with CHARMM forcefield [12,13].
- The contacts and interface energies are calculated with our in-house interface analysis tool, DynaBench [14].
- The Cα root mean square deviations (RMSD) are calculated with the `MULTI` command of ProFitV3.1, which is an implementation of the McLachlan algorithm [15].

### Scoring Category

We first extracted models from the ensemble file as provided by CAPRI with the *pdb_splitmodel* function of pdb-tools python library [16]. Then, we mapped the scrambled chain IDs onto their original ones, where necessary (e.g., ABCD to ABHL for T233). We then scored these interfaces with the *AnalyseComplex* command of FoldX [7]. We identified 100 low-energy models per target and visually inspected them with PyMOL 2.5.2 to make the final selection [17].

### Post-CAPRI analysis

All structural illustrations on this paper were created Pymol 2.5.2 [17]. For further evaluation of our models, we used the DockQ tool (https://github.com/bjornwallner/DockQ) [18]. We followed the standard CAPRI protocol for DockQ calculation.

## RESULTS

### T231: Our mini aggressive sampling combined with peptide pLDDT ranking was our key highlight in CAPRI55

Target T231 was an antibody bound to an eight amino acids long peptide (PDB ID: 8RMO [19]). Since this target lacked the typical co-evolution signal, we decided to use an enhanced AF2 sampling. For this, we constructed a mini version of Wallner's aggressive sampling protocol. In the regular aggressive sampling, Wallner generates 1000 models per probed sampling scenario [5]. In our mini version, we reduced this number by 50-fold and kept the dropout layers active to increase the model variability. We also employed three AF2 versions by using four seeds, while playing with the template usage and utilizing 3, 9, 21 recycles, while adjusting MSA pairing (Table 1). As an outcome, we run 11 different scenarios on ColabFold [8] (Methods).

**Table 1. Our mini aggressive sampling scheme.**

| Scenario | AF2 version | Dropout | Templates | No of Recycles | MSA Pairing | No of Seeds |
|---|---|---|---|---|---|---|
| **1** | v2.1 | Yes | Yes | 3 | No | 4 |
| **2** | v2.2 | Yes | Yes | 3 | No | 4 |
| **3** | v2.3 | Yes | Yes | 3 | No | 4 |
| **4** | v2.1 | Yes | No | 3 | No | 4 |
| **5** | v2.2 | Yes | No | 3 | No | 4 |
| **6** | v2.3 | Yes | No | 3 | No | 4 |
| **7** | v2.1 | Yes | No | 21 | No | 4 |
| **8** | v2.2 | Yes | No | 9 | No | 4 |
| **9** | v2.3 | Yes | No | 9 | No | 4 |
| **10** | v2.3 | Yes | No | 21 | No | 4 |
| **11** | v2.1 | Yes | No | 21 | Yes | 4 |
| **Total number of models generated** | | | | | | **220** |

To find the top performing scenario, we focused on the average peptide pLDDT (pep-pLDDT) instead of the regular AF2 confidence score. This was done since we observed overinflated ipTM and pTM values due to the domination of off-interface antibody regions. The top pep-pLDDT was generated by the 7$^{th}$ scenario, which used AF v2.1 with 21 recycles and no template information. In this scenario, the top three pep-pLDDT scores were recorded as 55, 62, and 68, with almost identical ipTM values (Fig. 2A-C). The top pep-pLDDT model was then subjected to HADDOCK 2.4 refinement [10] (since the relaxation module of ColabFold was down at the time of competition [20]). The final pep-pLDDT of our refined model was acquired through AF2rank, which calculates AF2 scoring metrics of a given structure [11]. While running AF2rank, we enabled amino acid and sidechain unmasking and used four recycles to

ensure accurate reproduction of the input. As an outcome, we acquired a pep-pLDDT of 80 and submitted this model as our top-ranking prediction (Fig. 2D). This model was classified as medium, placing us among the top seven predictors (Kozakov/Vajda, Pierce, Karaca, Huang, Venclovas, Zou, and Vakser) for this target. Our post-CAPRI analysis showed that the first two ranking models of the 7$^{th}$ scenario were indeed incorrect, while the third ranking one was acceptable. Our submission nicelybecame a medium model after refinement and rescoring (Table S1).

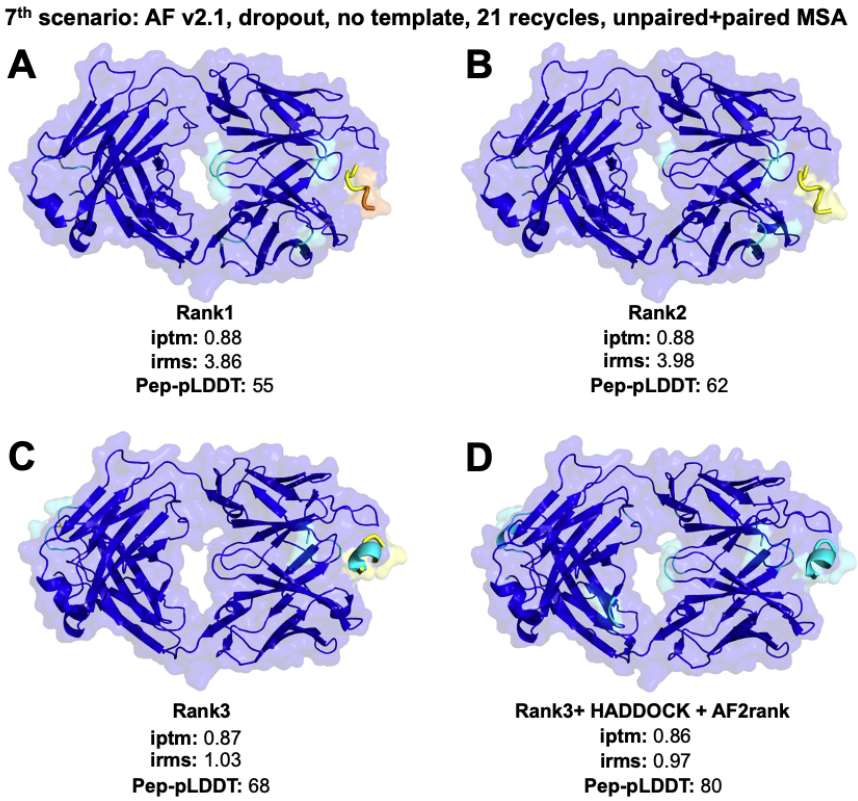


**Figure 1**. **Evaluation of the top models from the 7$^{th}$ mini aggressive sampling scenario. A.** The first, **B.** second, and **C.** third ranking models of the 7$^{th}$ scenario. **D.** Our top submission obtained after refinement and rescoring of the panel C model. All the models are colored according to their pLDDT scores and depicted in cartoon and transparent surface representation (pLDDT coloring: blue (highest), cyan, yellow, orange (lowest) confidence). ipTM, irms, and pep-pLDDT values of each model are provided below their associated structures.

Our other nine submissions came from a 450 ns long MD trajectory, which we started from a structure obtained through a default ColabFold-AF v2.2 run (Methods). We consulted MD, since we hypothesized that MD could aid us in finding a stable peptide pose. Accordingly, we saved a snapshot at each nanosecond and so analyzed 450 conformers. Our submissions were selected according to their peptide-antibody

contact profiles (the higher interface contacts were ranked higher, Figure S1). All these models were classified as "acceptable," except for the last incorrect one.

**Mini aggressive sampling could explore a conformational space range up to 12 Å for T232**

T232 was made of regulatory and catalytic subunits of a serine/threonine-protein phosphatase 2A (PP2A) in complex with a TIP41-like (TIPRL) protein. In an available complex structure, TIPRL was found to interact with both PP2A subunits (Fig. 2A, PDB ID: 5W0W [21]). The main challenge in this target was to identify an alternative heterotrimeric organization of this complex. To find this conformation, we employed our mini-aggressive sampling while skipping 3$^{rd}$, 9$^{th}$, and 10$^{th}$ scenarios to save time (ColabFold was down at the time of competition [20]) (Table 1). Among the generated conformations, we wanted to submit models with a wide RMSD range to increase our chances in capturing the alternative binding mode. Our modeling scenarios resulted in Cα–RMSD values ranging up to 11.7 Å (Fig. 2B), primarily due to the clockwise movement of TIPRL (Fig. 2C). The models with the highest conformational freedom came from the 8$^{th}$ scenario (AF2.2 with 9 recycles) (Fig. 2C). Based on the observed RMSD variation, we manually selected our first eight submissions, all of which are classified as medium (Figure S1).

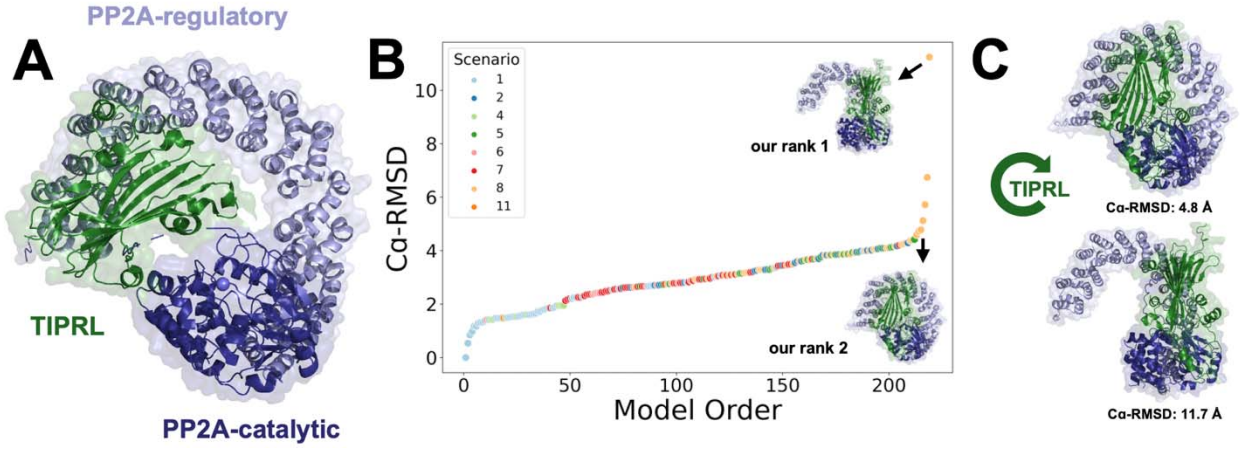


**Figure 2. Structural comparison of T232 models to the reference structure (PDB ID: 5W0W [20]). A.** TIPRL in complex with the regulatory and catalytic subunits of PP2A. **B.** The distribution of Cα-RMSD values across all generated models colored according to their scenarios. The reference state was chosen as the top-ranking structure of the 1$^{st}$ scenario. Our top two submission are highlighted on the plot. C. Structural depiction and the RMSD values of our top two submissions RMSD of 11.7 Å and 4.8 Å.

As a secondary approach, we aimed to generate a conformationally diverse set of models using HADDOCK in the *ab initio* mode. In this case, the catalytic subunit of PP2A and TIPRL were merged into a single unit to follow a two-body docking with the

following parameters: center of mass restraints=on, it0 = 5000, it1 = 400, and water = 400. All models generated in it1 and water stages converged to the arrangement presented in 5W0W. We, therefore, selected our last two submissions from it0 manually from the high Cα-RMSD regime. Unfortunately, these models were incorrect.

**Mini aggressive sampling could have prevented our failure in T233, but not in T234**

T233 and T234 were about predicting two different antibody binding sites on a major histocompatibility class (MHC) I complex. MHC I complexes are essential for the immune system, where its HLA-A component recognizes an 8-10 amino acids long antigenic peptide (Fig. 3A) [22]. Before the presentation of the antigenic peptide, the peptide-binding site of MHC is stabilized with a tapasin (TAP) molecule [23]. Since, for both targets, the presence of an antigenic peptide was not reported, we assumed that the antibodies could bind on HLA-A either on its peptide binding site (Ab(1) in Fig. 3A), its TAP binding site (Ab(2) in Fig. 3A) or a free spot opposite of its TAP binding site (Ab(3) in Fig. 3A). Expanding on this assumption, we devised three data-driven docking scenarios, where we inferred the HLA-A binding sites from previously published complexes (PDB ID: 1QVO [24] for Ab(1) and PDB ID: 5WER [25] for Ab(2) and Ab(3)). Data-driven sampling was carried out with HADDOCK with the following parameters it0 = 5000, it1 = 400, and water = 400. As the interface sizes of each site differed, we used FoldX to rank the HADDOCK structures. We made sure that our submitted models contained at least one solution from each targeted binding site. Unfortunately, none of our submissions could lead to an acceptable solution.

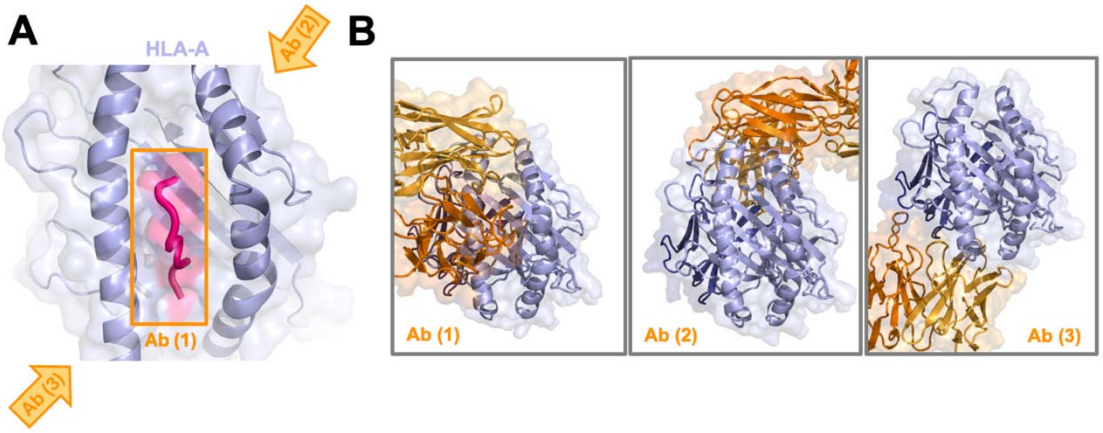


**Figure 3. The antibody (Ab) binding sites used in the modeling of T233 and T234. A.** The HLA-A and antigenic peptide binding shown in top view. The targeted Ab binding sites are shown by orange arrows and a box on the reference structure. **B.** Examples from the generated models for each targeted Ab binding site. MHC-I and antibodies are represented in lilac and orange cartoon and shades, respectively.

As post-CAPRI analysis, we run our mini aggressive sampling on T233 and T234. Due to the large size of these targets, we reduced the number of seeds to two for each modeling, leading to 110 models per target. For T233, even with this reduced sampling, we could obtain one medium and one acceptable solution. The medium-quality model came from the 2$^{nd}$ scenario, while the acceptable model came from the 1$^{st}$ (Table 1, Table S1). Interestingly, the medium model had the highest confidence score among all models, while the acceptable one was ranked at 95. So, by using AF v2.2 with 3 recycles and template information enabled, we could have submitted a top-ranking medium prediction. So, contrary to T231, for T233, the standard AF2 scoring worked wonderfully. For T234, on the other hand, our protocol could not generate a meaningful solution.

**Scoring challenge**

For the scoring challenge, we used the interface scoring function of FoldX (Methods). This helped us to select two acceptable models for T231, though they were our 4$^{th}$ and 9$^{th}$ ranking submissions. For T232, we scored eight medium-classified models out of our ten scoring submissions, while having our 1$^{st}$ and 5$^{th}$ submissions as incorrect.

## DISCUSSION

In this work, we demonstrated that a 50-fold reduced version of Wallner's aggressive sampling could generate medium quality models for T231, T232, and T233. For ranking T231, though, replacing the standard AF2 scoring was needed since ipTM and pTM values of this target was dominated by the high confidence off-interface antibody regions. This highlights the importance of adapting the AF2 scoring strategy when dealing with small antigens. For us, such an adaptation was achieved by using pep-pLDDT as the scoring metric. We further showed that employing mini aggressive sampling could generate a wide range of conformations in the case of T232, demonstrating another application area of our protocol. Importantly, incomplete information on T233 and T234 made us design targeted docking strategies, which did not result in a successful model generation. Interestingly, a simple AF2.2 run could have produced a medium solution for T233.

After CAPRI55, DeepMind released AlphaFold3 (AF3), making it accessible to researchers via a web server [9]. We assessed AF3's performance on T231 and T233-4 antibody targets using a single seed. Notably, AF3 failed to generate an acceptable model for T231 (Table S1). This may be attributed to the fact that the original AF3 study utilized 1000 seeds in antibody-antigen modeling [9]. In contrast, AF3 successfully produced high-quality models for T233, surpassing the performance of AF2 (Table S1). However, T234 still presented a challenge, as AF3 attempted to fill the peptide-binding site with random protein segments. This suggests that T234 is particularly difficult to model in the absence of the peptide. Based on these results, we recommend continued use of enhanced AF2 sampling for difficult antibody-antigen targets, at least until AF3's code becomes publicly available.

To conclude, CAPRI55 provided an excellent platform to test our mini aggressive sampling approach. The accuracy achieved was better than we had anticipated, demonstrating that this method could serve as a green, user-friendly option for protein complex modeling where the typical co-evolution signal is missing. With its reduced computational cost and accessibility via ColabFold's web service, it holds great promise for widespread use.

## REFERENCES


1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2
2. Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. Published online October 5, 2021. doi:10.1101/2021.10.04.463034.
3. Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. *Proteins*. 2023;91(12):1636-1657. doi:10.1002/prot.26598
4. Wallner B. Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins*. 2023;91(12):1734-1746. doi:10.1002/prot.26562
5. Wallner B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*.2023;39(9):btad573. doi:10.1093/bioinformatics/btad573
6. https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md
7. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33(Web Server issue):W382-W388. doi:10.1093/nar/gki387
8. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi:10.1038/s41592-022-01488-1
9. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500. doi:10.1038/s41586-024-07487-w
10. Honorato RV, Trellet ME, Jiménez-García B, et al. The HADDOCK2.4 web server for integrative modeling of biomolecular complexes. *Nat Protoc*. Published online June 17, 2024. doi:10.1038/s41596-024-01011-0
11. Roney JP, Ovchinnikov S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys Rev Lett*. 2022;129(23):238101. doi:10.1103/PhysRevLett.129.238101
12. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. *J Comput Chem*. 2005;26(16):1701-1718. doi:10.1002/jcc.20291
13. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, et al. CHARMM: the biomolecular simulation program. *J Comput Chem*. 2009;30(10):1545-1614. doi:10.1002/jcc.21287
14. https://github.com/CSB-KaracaLab/DynaBench
15. McLachlan A.D, Rapid Comparison of Protein Structres. Acta Cryst A38, 871-873. 1982. as implemented in the program ProFit (Martin, A.C.R. and Porter, C.T., http://www.bioinf.org.uk/software/profit/)
16. Rodrigues JPGLM, Teixeira JMC, Trellet M, Bonvin AMJJ. pdb-tools: a swiss army knife for molecular structures. *F1000Res*. 2018;7:1961. Published 2018 Dec 20. doi:10.12688/f1000research.17456.1

17. The PyMOL Molecular Graphics System, Version 2.5.2, Schrödinger, LLC.
18. Mirabello C, Wallner B, DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *bioRxiv*. Published online June 2, 2024. https://doi.org/10.1101/2024.05.28.596225
19. Beugelink JW, Sweep E, Janssen BJC, Snijder J, Pronker MF. Structural Basis for Recognition of the FLAG-tag by Anti-FLAG M2. *J Mol Biol*. 2024;436(16):168649. doi:10.1016/j.jmb.2024.168649
20. https://x.com/sokrypton/status/1735574651466650011
21. Wu CG, Zheng A, Jiang L, et al. Methylation-regulated decommissioning of multimeric PP2A complexes. *Nat Commun*. 2017;8(1):2272. Published 2017 Dec 22. doi:10.1038/s41467-017-02405-3
22. Nguyen AT, Szeto C, Gras S. The pockets guide to HLA class I molecules. *Biochem Soc Trans*. 2021;49(5):2319-2331. doi:10.1042/BST20210410
23. Wieczorek M, Abualrous ET, Sticht J, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front Immunol*. 2017;8:292. Published 2017 Mar 17. doi:10.3389/fimmu.2017.00292
24. Li L, Bouvier M. Structures of HLA-A*1101 complexed with immunodominant nonamer and decamer HIV-1 epitopes clearly reveal the presence of a middle, secondary anchor residue. *J Immunol*. 2004;172(10):6175-6184. doi:10.4049/jimmunol.172.10.6175
25. Jiang J, Natarajan K, Boyd LF, Morozov GI, Mage MG, Margulies DH. Crystal structure of a TAPBPR-MHC I complex reveals the mechanism of peptide editing in antigen presentation. *Science*. 2017;358(6366):1064-1068. doi:10.1126/science.aao5154