AlphaFold2 and RoseTTAFold Predict Posttranslational modifications. Chromophore Formation in GFP-like Proteins.

Sophia M. Hartley[†b], Kelly A. Tiernan[†b], Gjina Ametaj[‡ b], Adriana Cretu[‡ b], Yan Zhuang[a] and Marc Zimmer[b]

[a] Department of Mathematics and Statistics, Connecticut College, New London, CT 06320

[b] Chemistry Department, Connecticut College, New London, CT 06320

Abstract

AlphaFold2 and RoseTTAfold are able to predict, based solely on their sequence whether GFP-like proteins will post-translationally form a chromophore or not. Their training has not only taught them protein structure and folding, but also chemistry. The structures of 21 sequences of GFP-like fluorescent proteins that will post-translationally form a chromophore and of 23 GFP-like non-fluorescent proteins that do not have the residues required to form a chromophore were determined by AlphaFold2 and RoseTTAfold. The resultant structures were mined for a series of geometric measurements that are crucial to chromophore formation. Statistical analysis of these measurements showed that both programs conclusively distinguished between chromophore forming and non-chromophore forming proteins. A clear distinction between sequences capable of forming a chromophore and those that do not have the residues required for chromophore formation can be obtained by examining a single measurement - the RMSD of the overlap of the central alpha helices of the crystal structure of S65T GFP and the AlphaFold2 determined structure.

[†] Both authors contributed equally, first authors, names in alphabetic order
[‡] Both authors contributed equally, names in alphabetic order

Introduction

AlphaFold2[1, 2] and RoseTTAfold[3] are two freely available programs that can predict three-dimensional protein structures from their amino acid sequence with atomic accuracy. Both programs were created by machine learning and the ~180,000 structures in the protein data bank (pdb) [4, 5] were used as an important training set. The three-dimensional structures of many of the protein structures in the pdb have been influenced by the proteins binding ligands and post-translation modifications. Agirre and co-workers have shown that AlphaFold2 predicts the folding of glycosylated proteins solely from the amino acid sequence of the proteins because the program was trained on the pdb in which the glycosylated proteins are found in their full or partially glycanated forms.[6] In the absence of its heme cofactor AlphaFold2 will also fold a hemoglobin subunit so that it has a cavity complementary with a heme subunit.[6] [7] Finally, AlphaFold2 has been show to generate protein-peptide complex structures without multiple alignments of the peptide fragment.[8]

Fluorescent Proteins (FPs) are commonly used molecular tracer molecules. It is their post-translational autocatalytic chromophore formation that makes these genetic tracer molecules so useful. They are all folded into a β-barrel composed of eleven β sheets surrounding the chromophore that is located in a central alpha helix. Some Green Fluorescent Proteins-like (GFP-like) proteins fold into the characteristic β-barrel shape but do not form a chromophore. The training set is heavily weighted towards structures with a chromophore, less than 2% of the GFP-like structures in the pdb have no chromophore. There are 578 GFP-like proteins in the pdb, 568 of these structures have a fully formed chromophore, they are the GFP-like fluorescent proteins (GFP-FPs) and 10 have no chromophore.

The commonly accepted mechanism for chromophore formation in GFP-like proteins is shown in Figure 1. In order for the autocatalytic chromphore formation to occur the immature GFP-like protein has to be in the tight-turn conformation Figure 1-**I**.[9] This tight turn breaks the canonical i to i+4 hydrogen bonding arrangement found in alpha helices,[10] forming a kink that removes the intra main-chain hydrogen bonds that are commonly associated with an α-helix (see Figure S1). It is presumed that this aids in chromophore formation because the hydrogen bonds would otherwise have to be broken during maturation. Tyr66, Gly65, Arg96 and Glu222 (GFP numbering) are involved in chromophore formation and are highly conserved in all naturally

occurring GFP-like FPs.[11, 12] The central tyrosine of the chromophore, Tyr66, is conserved in all naturally-occurring GFP-like FPs, although any aromatic residue in that position will auto-catalytically form a chromophore. [11] Thus, we have chemical knowledge that can be used to predict whether a GFP-like protein will form a chromophore or not. A GFP-like protein has to have a glycine65, an aromatic residue in position 66 and an arginine96 to form a chromophore. This is chemical knowledge that was not directly provided to AlphaFold2 or RoseTTAFold.
We have used AlphaFold2 and RoseTTAFold to generate the structures of two series of GFP-like sequences; one that can form a chromophore and the other that can't. And have shown that both programs fold the chromophore forming GFP-like sequences into a distinctly different conformation compared to the sequences that cannot form a chromophore. The structures in the training set taught the AI programs chemistry to distinguish which proteins will be post-translationally modified from those that will not.

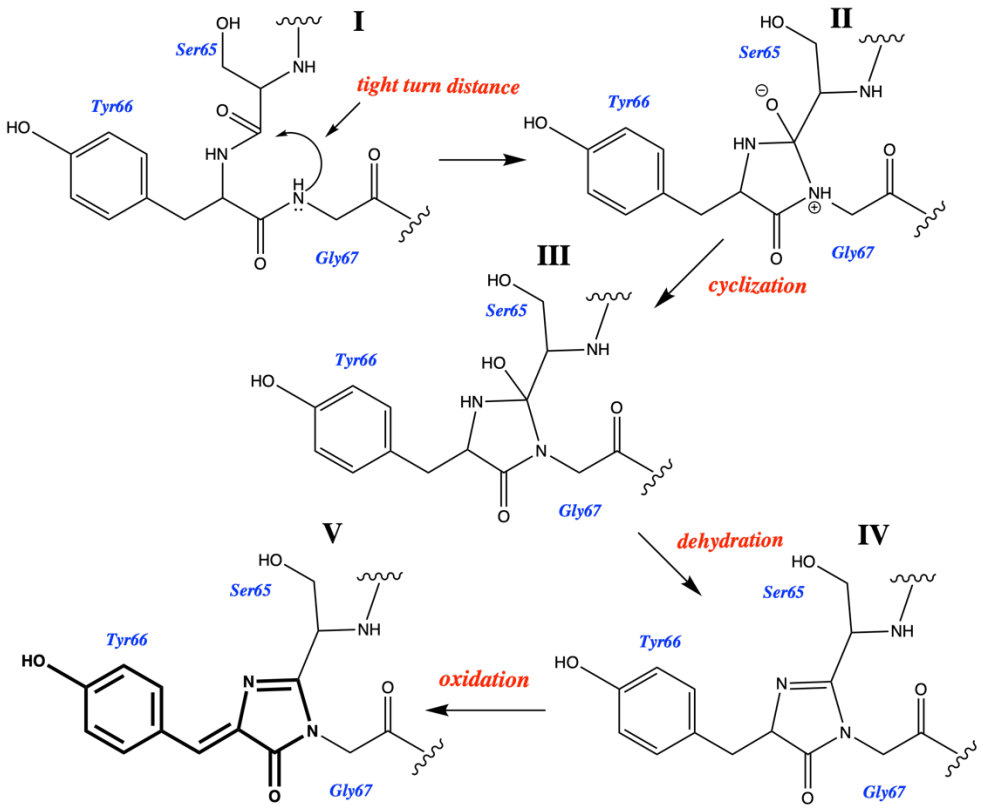

Figure 1. The cyclization-dehydration-oxidation pathway for chromophore formation. [13] Structure **I** is known as the immature precyclized form of the protein. The tight-turn distance depicted in structure **I** was measured for all sequences and is presented in Table S1. Structure **V** shows the fully formed mature chromophore (bolded).

Methods

1. Sequence Selection

Standard Structures

Three GFP-like proteins with known solid-state structures were selected: 1EMA, 2AWJ, 1H4U (pdb codes).

1EMA is the crystal structure of the S65T mutant of GFP. Together with 1GFL[14] it was the first crystal structure of GFP to be solved.[15] It is the structure of a fluorescent protein that has a mature chromophore formed from the 65TYG67 triad of residues, see Figure 1-**V**.

2AWJ is an immature precyclized GFP-like protein structure. It is the R96M mutant that takes about 3 months to form the chromophore.[16] Arginine and methionine have similar sizes, allowing us to safely assume that the conformation of immature wild-type GFP and its R96M mutant would be the same. The crystal structure of the R96M mutant is the closest solid-state structure to the immature GFP precyclized structure, that is Figure 1-**I**.

1H4U is the structure of the G2 domain of mouse nidogen. It is a GFP-like structure with a central alpha helix in an 11 stranded beta barrel. It has no chromophore and cannot form a chromophore.[17]

Sequences of Chromophore Forming GFP-like Proteins

Amino acid sequences of GFP-like fluorescent proteins that do not appear in the pdb were obtained from FPbase[18]. The chromophore forming GFP-like FP sequences were chosen from the FPBase in a way that they spanned as much phylogenetic space possible and had not been crystalized i.e. they did not appear in the pdb. Protein sequences were selected from each FP lineage, starting from the parent protein. If the parent protein's crystal structure was already solved, then the next protein sequence in the lineage closest to the parent was selected. No sequence was used from lineages in which all structures have been solved (DrCBD) and far-red FPs with a tetrapyrroloid chromphores as these FPs are not GFP-like β-barrels.

Sequences of GFP-like Proteins That Cannot Form a Chromophore

Sequences of GFP-like non-fluorescent proteins were taken from Haddock's "Non-excitable fluorescent protein orthologs found in ctenophores" paper[19] and a series of nidogen-G2 domains were obtained from a standard protein BLAST search (BLASTP) using the 1H4U

FASTA sequence as a protein query.* Only the G2 fragment region of the nidogen sequences were used in the AI alignments and the subsequent analyses. See Supplementary Table S2 for sequences studied and their sequence similarity to S65T GFP (1EMA).

2. **AI predictions**

RoseTTAFold was accessed via the RosettaCommons web-interface server Robetta [20] and simplified AlphaFold(v2.1.0) was accessed through AlphaFold's Colab notebook.[21]

3. **Alignments**

MAFFT DASH[22] was used for the multiple alignment and Clustal Omega[23] to find the percent identity matrix of all of the proteins relative to 1EMA, see Table S2. All residues aligned with residues 60 to 74 of 1EMA were considered part of the central alpha-helix, see Figure S2.

4. **Measurements**

Maestro Version 12.9.123[24] was used to perform hydrogen bond measurements of the structures obtained from pdb[5], Alphafold2 and RoseTTAFold. RMSD values for the alpha helical overlaps were obtained by determining the pairwise distances between structures, using the rms displacement after optimal rigid-body superposition between pairs of non-hydrogen backbone atoms of residues highlighted in the alignment shown in Figure S2 of the supplementary material.[25, 26] all the measurements are presented in Table S1.

Results and Discussion

Structures Examined

AlphaFold2 and RoseTTAFold were used to predict the 3-dimensional structure of a series of GFP-like proteins - a set of 3 standard structures with known solid state structures; a group of 21 GFP-like fluorescent proteins that will post-translationally form a chromophore and whose solid state structures have not been determined; as well as a group of 23 GFP-like non-fluorescent proteins that do not have the residues required to form a chromophore and whose solid state structures have not been determined. Table S2 lists all the sequences and how they were obtained.

How does AI deal with the chromophore?

* The KAB1270852.1 Nidogen-1 *Camelus dromedarius* sequence was chosen to honor the fact that the Connecticut College mascot is the camel.

Since the chromophore formation involves post-translational modification of the 65TYG67 triad we can't expect the AI programs to model the structure of GFP-like FPs with a fully formed chromophore (Structure **V** in Figure1). The central alpha-helical strand (residues 60 to 74) of the structure obtained from the AI predicted folding of the S65T GFP sequence (as found in 1EMA) was overlapped with the pdb structure of 1EMA. Both AlphaFold2 and RoseTTAFold predict a structure closest to the immature precyclized form (structure **I** in Figure 1, a structure similar to the one obtained if one was to graphically mutate the chromophore back to 65TYG67 and then minimize), a conformation most closely represented by crystal structure of the R96A mutant 2AWJ (RMSD for AF2 = 0.32Å and RF = 0.40Å). The alpha helical overlaps with the regular alpha helix observed in the nidogen G2 domains (RMSD vs 1H4U for AF2 = 1.65Å and RF = 1.68Å) or even vs. its own crystal structure (RMSD vs 1EMA for AF2 = 0.78Å and RF = 0.80Å) are much higher, see Table S1.

Can AI Folding Programs Predict Whether GFP-like Proteins Will Form a Chromophore Or Not?

AlphaFold2 and RoseTTAFold were used to predict the 3-dimensional structures of 21 GFP-like fluorescent proteins that will post-translationally form a chromophore as well as a group of 23 GFP-like non-fluorescent proteins that do not have the residues required to form a chromophore. All sequences folded into GFP-like 11 stranded β-barrels with a central alpha helix. The resultant structures were analyzed to establish whether according to the chromophore forming mechanism discussed in the introduction and shown in Figure 1, they were in conformations geometrically primed for chromophore formation or whether they could not form a chromophore. We overlapped their central alpha helical strands with the strands obtained from the crystal structures of 1EMA (has chromophore), 2AWJ (has no chromophore but is in the right conformation to form one) and 1H4U (has no chromophore and is in the wrong conformation to form chromophore); their tight-turns were measured, as were the i to i+4 hydrogen bonding distances in their alpha helices.

Statistical Analysis of the AlphaFold2 Predictions

From the boxplot, Figure 2, as well as the data presented in Table S1 and Table S3, it is apparent that the RMSD for α-helix overlap of the AlphaFold2 predicted structures with 1EMA-crystal is much higher for those that will not form a chromophore than those that do. Moreover, the Welch two sample t-test shows there is a significant difference in the mean values of α-helix overlap

with 1EMA-crystal between GFP-like proteins that will form a chromophore and those that do not (t-value=10.646 and p-value=7.77 x 10^{-12}). The average RMSD for GFP-like proteins that will not form a chromophore is 1.372Å while the average value for GFP-like proteins that will form a chromophore is 0.799Å.

For the structures predicted by AlphaFold2 one only needs to compare the predicted alpha helical structure with that of the S65T GFP pdb structure to know whether the GFP-like protein will form a chromophore or not.

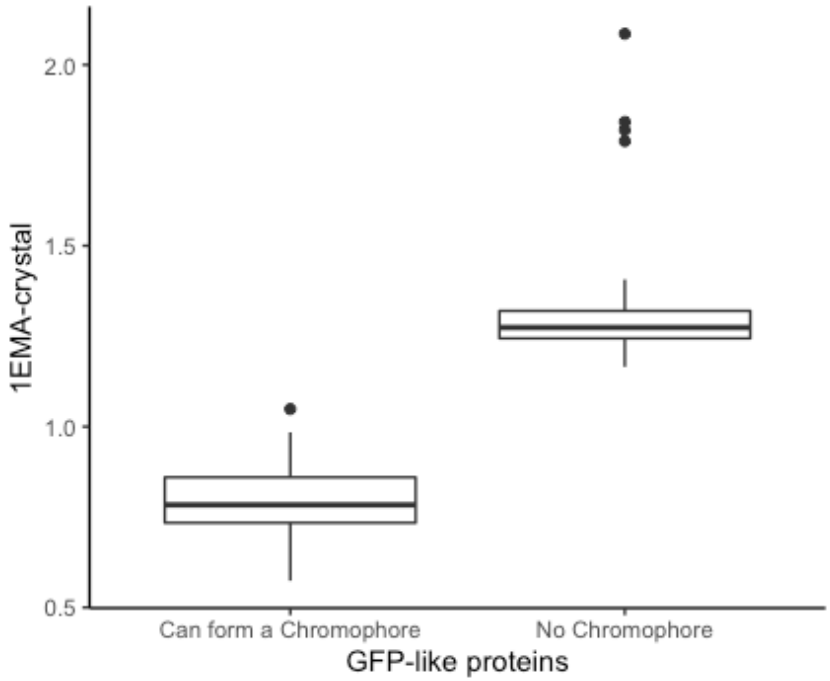

Figure 2: RMSD overlap of the α-helix of the 1EMA-crystal structure with the α-helix of 1EMA as determined by AlphaFold2 for GFP-like proteins that will form a chromophore (SYG) and those that do not (no SYG).

A LASSO regression model was built using "can form chromophore" (=1) and "cannot form chromophore" (=0) as the response variable with all the measurements including Alpha helix overlap, tight turn distance, as well as H-Bond distances in angstrom. The final model includes the following predictors - the alpha helix overlap with 1EMA-crystal as well as H-bond distances in Angstrom between residues 61-65 (HD2), 62-66 (HD3), 70-74 (HD11). The model shows there is a clear distinction between the sequences capable of chromophore forming and those that

don't have the residues required for chromophore formation, see Figure S3 and Table S4. A model built on just the measurements of the hydrogen bonding distances is also able to distinguish between chromophore and non-chromophore forming structures.

Statistical Analysis of the RoseTTAFold Predictions

The RMSD for the α-helix overlap of the RoseTTAFold predicted structures with 1EMA-crystal structures is generally higher for those structures that will not form a chromophore than those that do. However the appropriate boxplot, Figure S4, as well as the data presented in Table S1 and Table S5, show that while RoseTTAFold can distinguish between posttranslationally modified and non-posttranslationally structures it is not as distinctive as AlphaFold2.

The Welch two sample t-test shows that there is a significant difference in the mean values for alpha helix overlap with 1EMA-crystal between GFP-like proteins that will form a chromophore and those that do not (t-value=4.699 and p-value=3.526×10^{-5}). The average value for GFP-like proteins that will not form a chromophore is 1.317 while the average value for GFP-like proteins that will form a chromophore is 1.003.

For data collected using RoseTTaFold, we built a LASSO regression model using "can form chromophore" (=1) and "cannot form chromophore" (=0) as the response variable. The selected predictors of final model are: the RMSD of the alpha helix overlap with 1EMA-crystal and 2AWJ-crystal, the tight turn distances, as well as H-bond distances in Angstrom between residues 61 and 65 (HD2), 62-66 (HD3), 63-67 (HD4), 64-68 (HD5), see Table S6 and Figure S5. From the prediction results, we can see there is a clear distinction between the sequences capable of forming a chromophore and those that can't. A model built on just the measurements of hydrogen bonding distances is also able to distinguish between chromophore and non-chromophore forming structures.

Conclusion

Only 10 of the 578 GFP-like proteins in the pdb have no chromophore, yet when AlphaFold2 and RoseTTAFold are presented with the sequences of 44 GFP-like proteins that are not in the pdb they fold the proteins in such a way that one can unequivocally distinguish between those that can and cannot form a chromophore. They predict the conformation of the immature protein with a kink in the α-helix (as expected from machine learning vs. memorization) and have used

their training set to learn some chemistry, they can distinguish between GFP-like proteins that will form a chromophore and those that do not. We suspect that the pdb training set and multiple sequence alignments enable AlphaFold2 and RoseTTAFold to “think” like chemists and look for the presence of residues equivalent to Arg96, Gly65 and an aromatic residue at position 66 in GFP – the residues required for chromophore formation – and use those to fold the sequence in a chromophore forming or non-chromophore forming conformation.

Supplementary Materials

Figure S1 Main chain i to i+4 "hydrogen bonding" alpha helical interactions measured and presented in Table S1.

Figure S2 Alignment of all sequences used.

Figure S3 Predicted possibility that structure forms a chromophore or cannot form a chromophore using LASSO model from Table S3 on AlphaFold2 data.

Figure S4 RMSD overlap of the α-helix of the 1EMA-crystal with the α-helix of 1EMA as determined by RoseTTAFold for GFP-like proteins that will form a chromophore and those that do not.

Figure S5 Predicted possibility that structure forms a chromophore or cannot form a chromophore using LASSO model from Table S5, which is based on RoseTTAFold data.

Table S1 RMSD overlap of the α-helix of the 1EMA, 2AWJ, 1H4U crystal structures with the α-helix of the RoseTTAFold and AlphaFold determined structures, tight turn distances of RoseTTAFold and AlphaFold determined structures and the main chain i to i+4 "hydrogen bonding" alpha helical interactions. These are the geometric measurements that are crucial to chromophore formation. The structures are divided into two groups, GFP-like proteins that will form a chromophore and those that do not.

Table S2 Description of all sequences used in this study.

Table S3 Summary statistics of the RMSD overlap (in Angstrom) of the α-helix of the 1EMA-crystal with the α-helix of 1EMA as determined by AlphaFold2 for GFP-like proteins that will form a chromophore and those that do not.

Table S4 LASSO model results for the alpha helix overlap with 1EMA-crystal as well as H-bond distances in angstrom between residues 61-65 (HD2), 62-66 (HD3), 70-74 (HD11) collected using AlphaFold2.

Table S5 Summary statistics of the RMSD overlap (in Angstrom) of the α-helix of the 1EMA-crystal with the α-helix of 1EMA as determined by RoseTTAFold for GFP-like proteins that will form a chromophore and those that do not.

Table S6 LASSO model results for the RMSD of the alpha helix overlap with 1EMA and 2AWJ-crystal structures, the tight turn distances, as well as H-bond distances in Angstrom between residues 61 and 65 (HD2), 62-66 (HD3), 63-67 (HD4), 64-68 (HD5) collected using RoseTTAFold simulations.

References:

1. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
2. Tunyasuvunakool, K., et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021. **596**(7873): p. 590-596.
3. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
4. Burley, S.K., et al., *RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy.* Nucleic Acids Research, 2018. **47**(D1): p. D464-D474.
5. Burley, S.K., et al., *RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences.* Nucleic Acids Research, 2021. **49**(D1): p. D437-D451.
6. Bagdonas, H., et al., *The case for post-predictional modifications in the AlphaFold Protein Structure Database.* Nature Structural & Molecular Biology, 2021. **28**(11): p. 869-870.
7. Hekkelman, M.L., et al., *AlphaFill: enriching the AlphaFold models with ligands and co-factors.* bioRxiv, 2021: p. 2021.11.26.470110.
8. Tsaban, T., et al., *Harnessing protein folding neural networks for peptide–protein docking.* Nature Communications, 2022. **13**(1): p. 176.
9. Branchini, B.R., A.R. Nemser, and M. Zimmer, *A Computational Analysis of the Unique Protein-Induced Tight Turn That Results in Posttranslational Chromophore Formation in Green Fluorescent Protein.* J. Am. Chem. Soc., 1998. **120**: p. 1-6.
10. Barondeau, D.P., et al., *The case of the missing ring: Radical cleavage of a carbon-carbon bond and implications for GFP chromophore biosynthesis.* Journal of the American Chemical Society, 2007. **129**(11): p. 3118-3126.
11. Ong, W.J.H., et al., *Function and structure of GFP-like proteins in the protein data bank.* Molecular BioSystems, 2011. **7**(4): p. 984-992.
12. Lemay, N.P., et al., *The role of the tight-turn, broken hydrogen bonding, Glu222 and Arg96 in the post-translational green fluorescent protein chromophore formation.* Chemical Physics, 2008. **348**(1-3): p. 152-160.
13. Grigorenko, B.L., A.I. Krylov, and A.V. Nemukhin, *Molecular Modeling Clarifies the Mechanism of Chromophore Maturation in the Green Fluorescent Protein.* Journal of the American Chemical Society, 2017. **139**(30): p. 10239-10249.
14. Yang, F., L.G. Moss, and G.N. Phillips, *The molecular structure of green fluorescent protein.* Nature Biotechnol., 1996. **14**: p. 1246-1251.
15. Ormoe, M., et al., *Crystal structure of the Aequorea victoria green fluorescent Protein.* Science, 1996. **273**: p. 1392-1395.
16. Wood, T.I., et al., *Defining the role of arginine 96 in green fluorescent protein fluorophore biosynthesis.* Biochemistry, 2005. **44**(49): p. 16211-16220.

17. Hopf, M., et al., *Mapping of binding sites for nidogens, fibulin-2, fibronectin and heparin to different IG modules of perlecan.* Journal of Molecular Biology, 2001. **311**(3): p. 529-541.
18. Lambert, T.J., *FPbase: a community-editable fluorescent protein database.* Nature Methods, 2019. **16**(4): p. 277-278.
19. Francis, W.R., et al., *Non-excitable fluorescent protein orthologs found in ctenophores.* BMC Evolutionary Biology, 2016. **16**(1): p. 167.
20. Robetta. January, 2022]; Available from: https://robetta.bakerlab.org/.
21. AlphaFoldColab. January 2022]; Available from: https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb.
22. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Briefings in Bioinformatics, 2019. **20**(4): p. 1160-1166.
23. Sievers, F. and D.G. Higgins, *Clustal Omega for making accurate alignments of many protein sequences.* Protein Science, 2018. **27**(1): p. 135-145.
24. Maestro, *Schroedinger LLC, New York, NY.* 2022.
25. Kabsch, W., *Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1978. **34**(SEP): p. 827-828.
26. Kabsch, W., *Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1976. **32**(SEP1): p. 922-923.