Supporting information: How to select the best model from AlphaFold2 structures?

Yuma Takei and Takashi Ishida

1 Supplementary methods

1.1 Exclusion of targets with low domain interaction

For several targets that have multiple domains, we found the cases that the AlphaFold2 structures which were accurate on individual domain-level but failed to be predicted because of the relative positions between the domains. Proteins with few interactions between domains may not have a stable whole structure even in vivo, and their relative positions may change. Thus, it is inappropriate to evaluate the prediction accuracy of the whole structure using a structural similarity metric such as GDT_TS. Therefore, we excluded such proteins from the data set.

For excluding inappropriate targets, we used the following procedure.

1. Calculate the percentage of contacts from all combinations of the residues up to the i-th residue and the residues after i-th residues.

2. If the percentage of contacts is less than a threshold, consider the interaction to be low.

We defined contact between residues where the distance between CA atoms is less than 12Å. 1% was used as the threshold for the percentage of contacts.

As a result, not only multi-domain proteins but also other proteins such as those with a long loop region at the terminal were excluded. Using this procedure, some proteins with sufficient domain interactions were also excluded. In this study, false positives were not a critical problem so that such targets were also excluded.

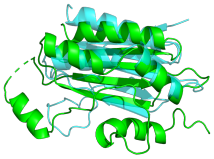
2 Supplementary results

2.1 Selected targets

2.2 Tertiary structure of the targets with low accuracy

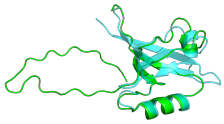
Fig S1 shows the superposition of the native structure of a target with GDT_TS less than 0.75 and the predicted structure with the maximum GDT_TS within that target.

(a) Target: 6Z4U_A[1], maximum GDT_TS: 0.259

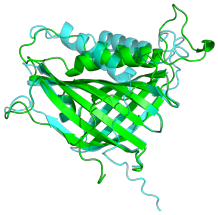


(b) Target: 6BJG_A[2], maximum GDT_TS: 0.448

(c) Target: 7EL1_E[3], maximum GDT_TS: 0.475

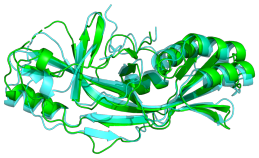


(d) Target: 6NEK_A[4], maximum GDT_TS: 0.680



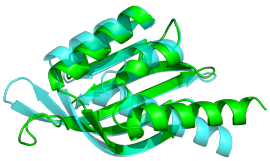
(e) Target: 6NNW_A[5], maximum GDT_TS: 0.695

(f) Target: 6RO0_B[6], maximum GDT_TS: 0.712



(g) Target: 7N50_A[7], maximum GDT_TS: 0.712

(h) Target: 6BS3_B[8], maximum GDT_TS: 0.724



(i) Target: 6W40_A[9], maximum GDT_TS: 0.742

Figure S1: **Superposition of the best structure in the target and the native structure with low structure prediction accuracy.** The green color indicates the native structure of the target, and the cyan translucent color indicates the predicted structure with the maximum GDT_TS in the target. The superimposition was created using TM-score[10].

2.3 Difference in the accuracy of predicted structures for the single target

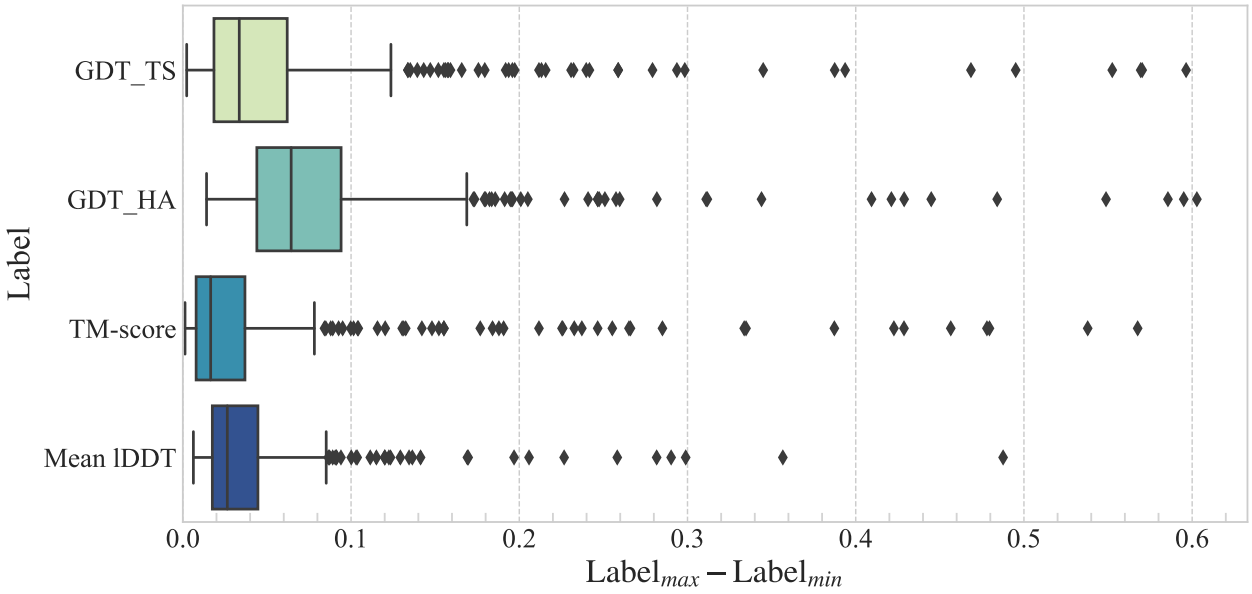


Figure S2: **Box plot of the difference between the maximum and minimum value of labels in each target.**

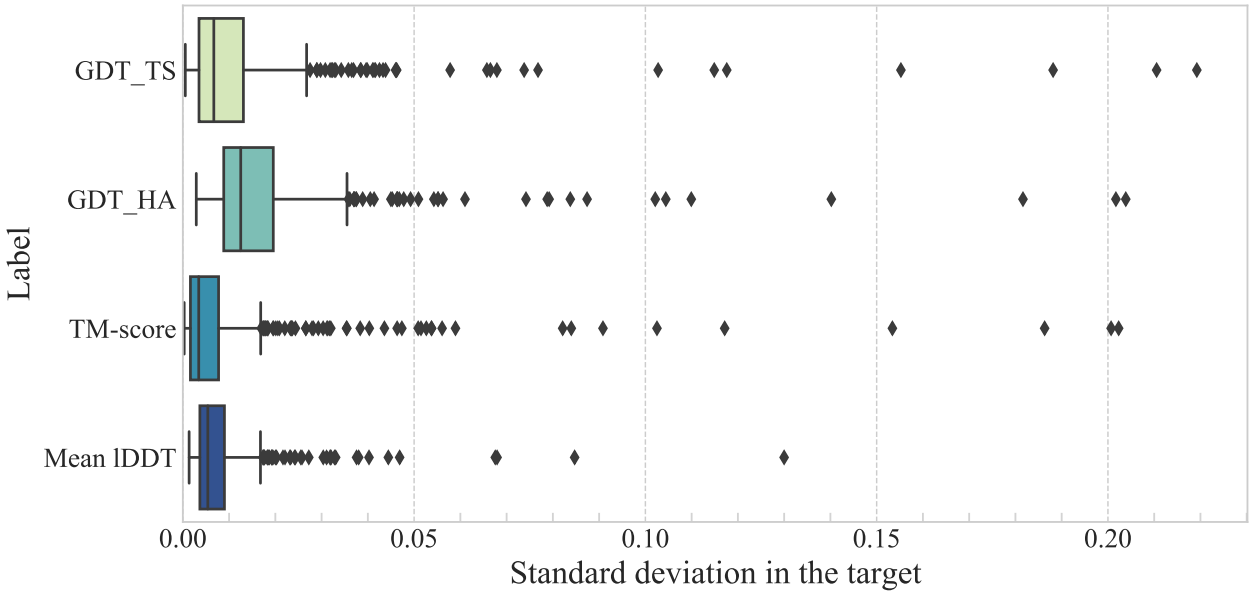


Figure S3: **Box plot of the standard deviation of labels in each target.**

2.4 Difference in the accuracy by parameters

2.4.1 Prediction model

We examined the difference in accuracy among the ten different prediction models of AlphaFold2. The distribution of the accuracy for each prediction model is shown in Fig S4. Although there was a slight difference in the accuracy of each prediction model, no particular prediction model was outstanding. However, there were cases where the accuracy of a particular model was superior for some targets. Fig S5 shows the difference in accuracy between the prediction models for the targets where there was a difference. There were targets for which a particular model was superior and targets for which there was a split between models with low accuracy and models with high accuracy. In addition, there were 62 targets for which the maximum GDT_TS difference between the models was 0.05 or more. Thus, the accuracy of the prediction models differed greatly depending on the target. Therefore, it is important to use multiple prediction models.

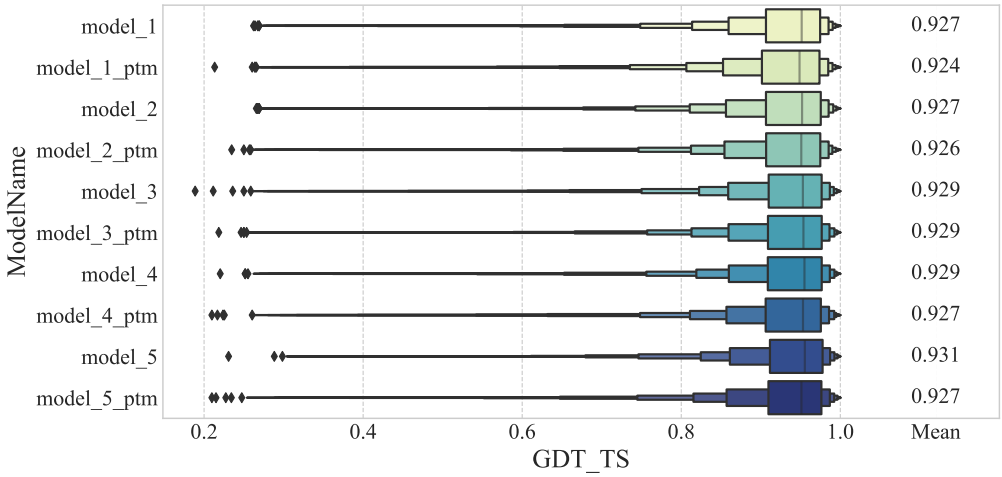


Figure S4: **Boxen plot of GDT_TS for each prediction model.** The X-axis shows GDT_TS and the Y-axis shows the name of the prediction model. All structures predicted by each model were used. The mean value is shown to the right of the boxen plot.

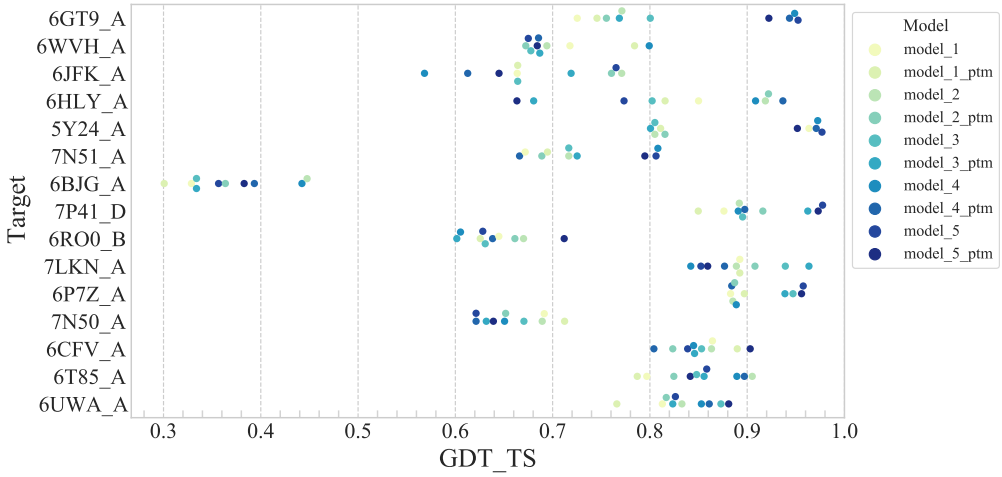


Figure S5: **Plot of maximum GDT_TS for each prediction model.** The X-axis shows GDT_TS and the Y-axis shows the name of the targets. Only targets where the maximum GDT_TS and the median of the maximum GDT_TS per prediction model within the target are greater than 0.05 are shown. A single point represents the maximum GDT_TS for a particular prediction model.

Next, we examined the accuracy difference between the normal model used in CASP14 and the ptm model fine-tuned to predict the pTM score. Fig S6A shows the scatter plot of GDT_TS for the structure

predicted by the normal model and the ptm model. In this figure, we compare the GDT_TS between the structures where all parameters are the same except for the prediction model. For most of the structures, there is no significant difference in GDT_TS between the normal model and the ptm model, but for some of them, there is a significant difference. Thus, as shown in Fig S6B, when the maximum GDT_TS values of the normal model and ptm model were compared for each target, there was no significant difference. Therefore, regardless of whether the normal model or ptm model is used to generate structures, it is possible to generate a structure with equal accuracy by combining other parameters. For this reason, it is almost sufficient to use either the normal model or the ptm model for the prediction model.

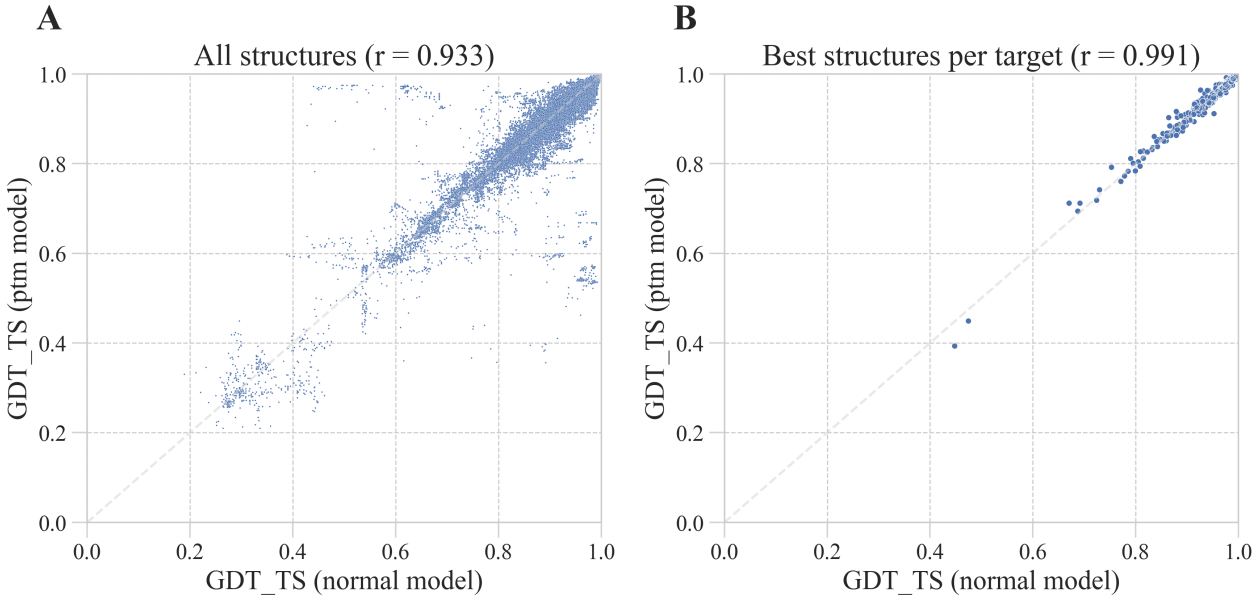


Figure S6: **Scatter plot of GDT_TS between normal model and ptm model.** (A) Difference when all prediction structures are used. (B) Difference when the best GDT_TS structure per target are used.

2.4.2 Random seed

In this study, two random seeds were used to generate the structures. The scatter plot of GDT_TS between the structures with the same parameters except for the random seed is shown in Fig S7. There is a large difference between some of the structures depending on the random seed, but the overall difference is small. When comparing the maximum GDT_TS for each target, there is almost no difference between them. Therefore, the influence of random seed is small.

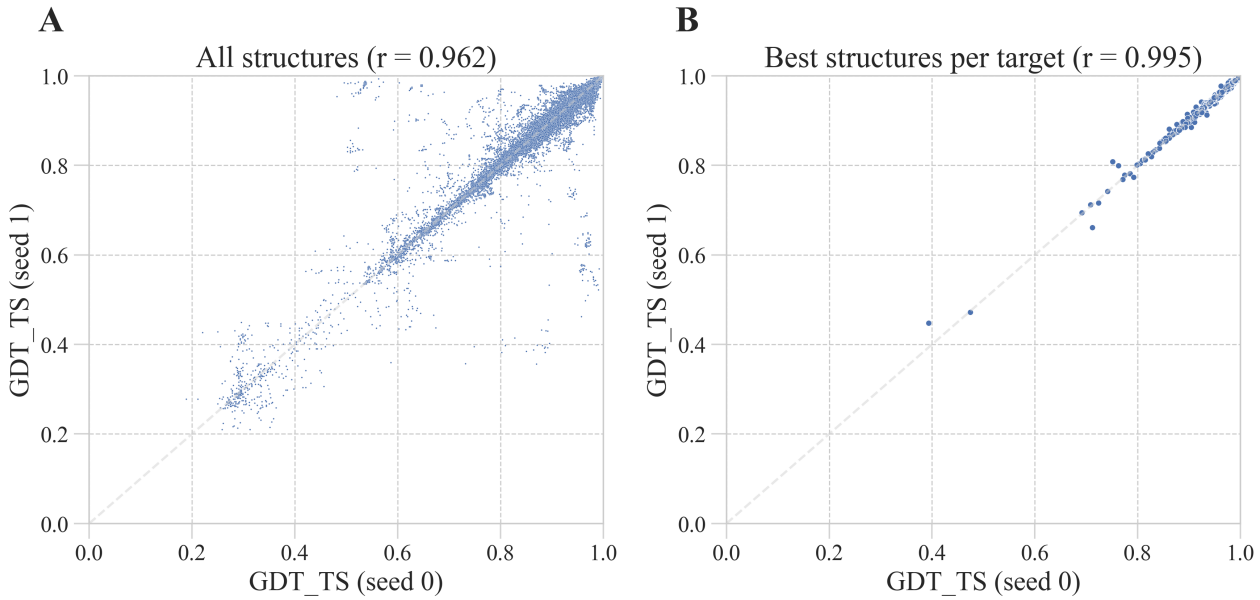


Figure S7: **Scatter plot of GDT_TS between random seeds.** (A) Difference when all prediction structures are used. (B) Difference when the best GDT_TS structure per target are used.

2.4.3 Ensemble

In this study, we generated prediction structures for two cases: when the number of ensembles was 1 and 8, i.e., with and without ensembles. The scatter plot of GDT_TS of the structures with and without ensemble is shown in Fig S8. In several cases, the accuracy was slightly higher with the ensemble than without the ensemble. The average GDT_TS was 0.928 and 0.927, respectively. However, there was almost no difference in the maximum GDT_TS for each target.

2.4.4 Recycle

The distribution of GDT_TS for recycling numbers 1 through 10 is shown in Fig S9. The mean value of GDT_TS shows a slight increase up to recycle number 3 but converges after that. The default recycling number of 3 for AlphaFold2 was found to be appropriate.

Overall, the accuracy did not change significantly by recycling, but for some targets, the prediction accuracy changed significantly by recycling. Fig S10 shows the transition of GDT_TS for the 32 targets whose GDT_TS changed more than 0.1 by recycling. The accuracy of some targets increased with repeated recycling. In some cases, only the structure of a particular prediction model improves the accuracy. In other cases, there were targets whose accuracy decreased by recycling.

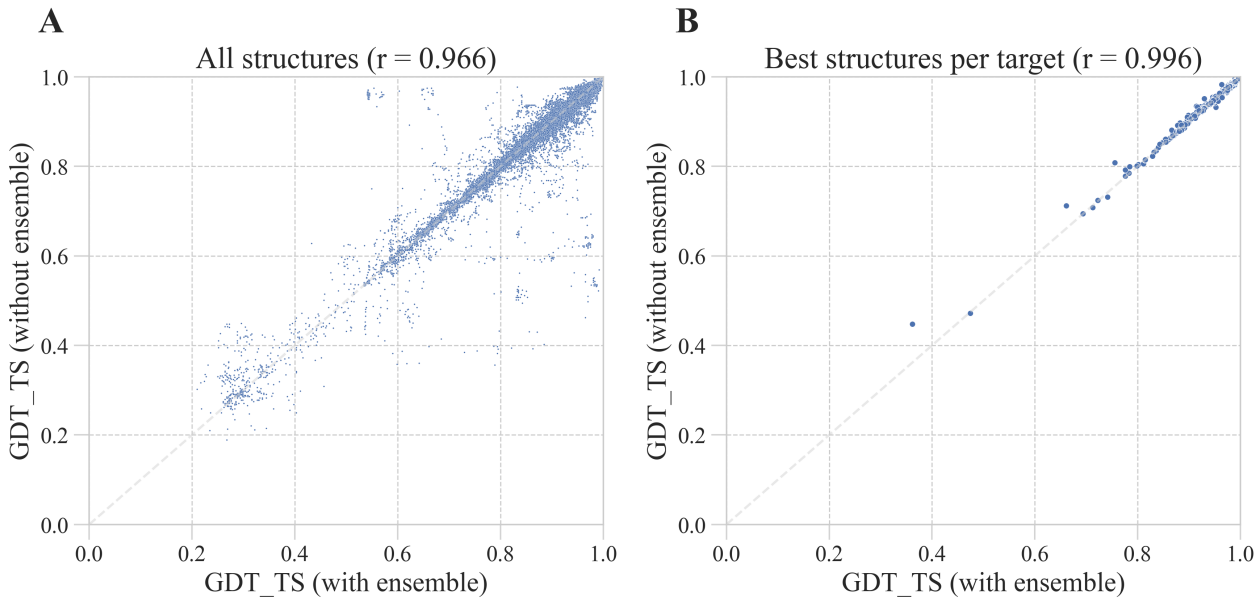


Figure S8: **Scatter plot of GDT_TS between with and without ensemble.** (A) Difference when all prediction structures are used. (B) Difference when the best GDT_TS structure per target are used.

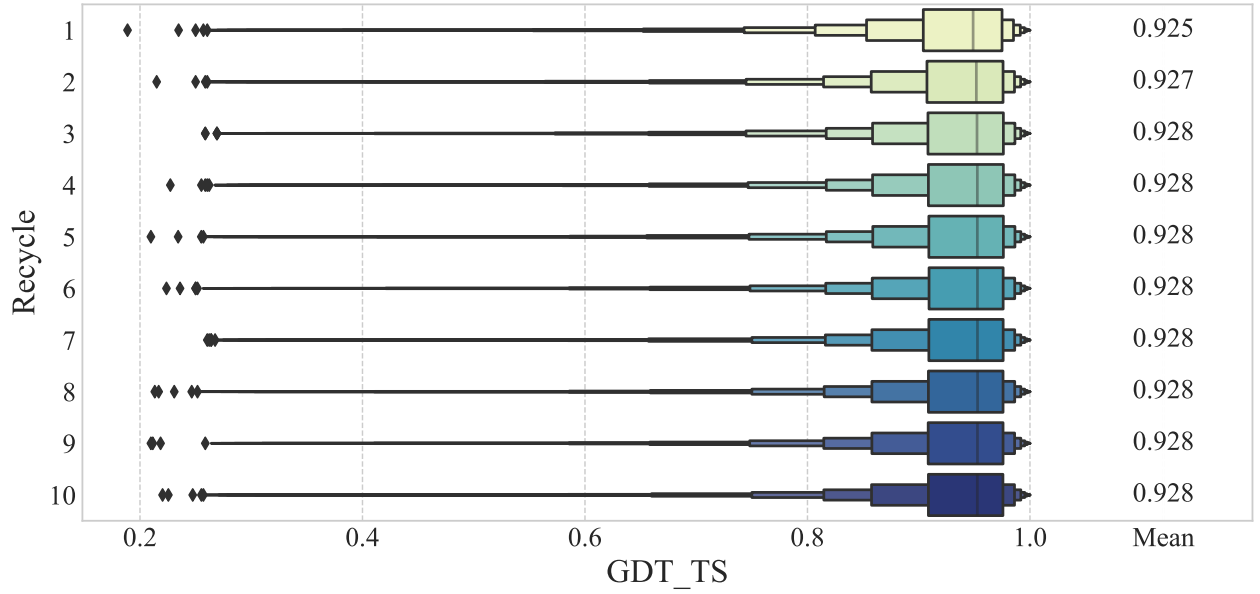


Figure S9: **Boxen plot of GDT_TS for each number of recycles.** The X-axis shows GDT_TS and the Y-axis shows the number of recycles. All structures by each recycling were used.

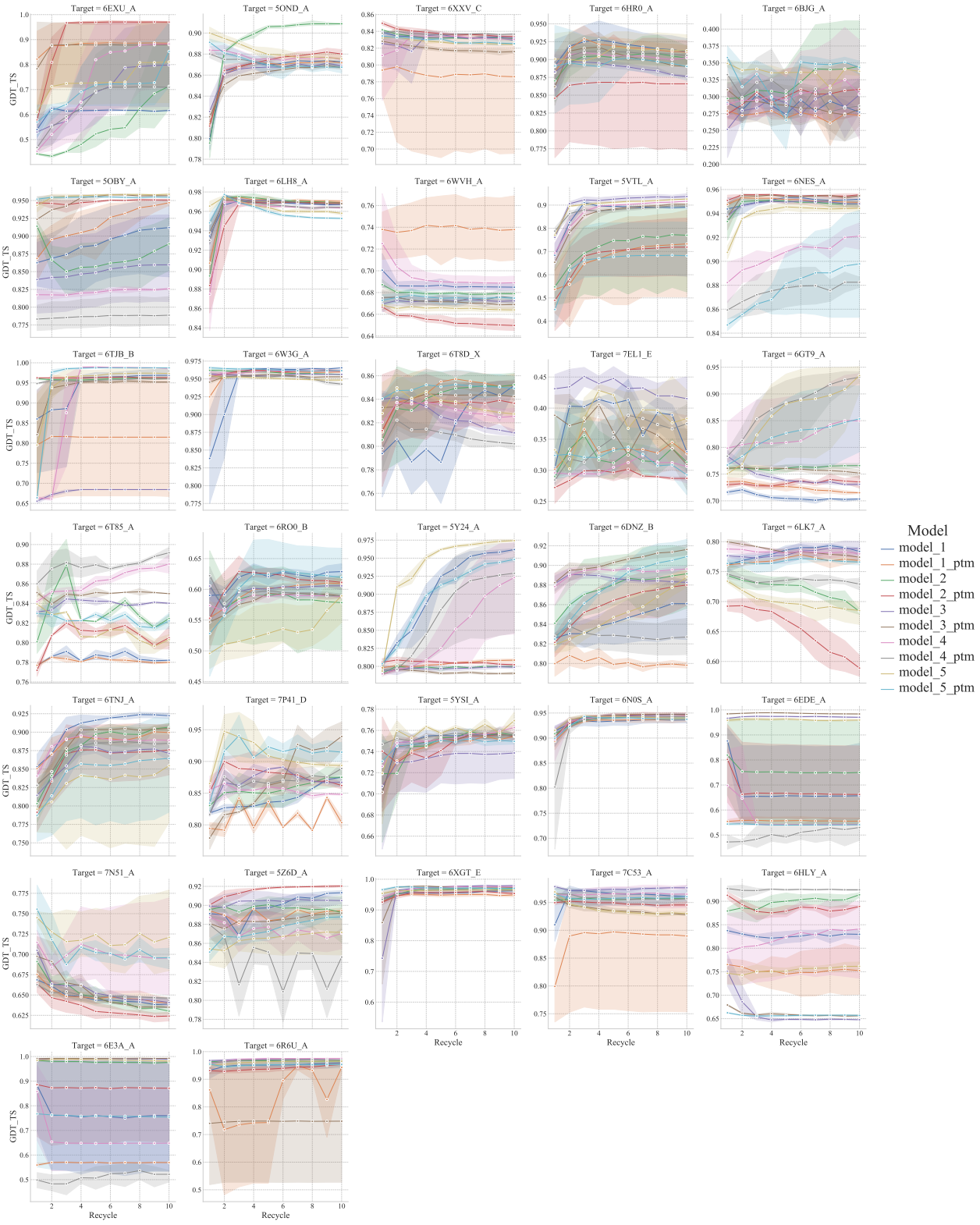


Figure S10: **Line plot of GDT_TS for each recycling.** The X-axis shows the number of recycles and the Y-axis shows GDT_TS. The color of the line represents the prediction model. The mean value of GDT_TS for the structures with the same number of recycles and the prediction model is represented by a point, and the range represents the 95% confidence interval.

2.5 Estimation performance of relative accuracy

The estimation performance of the relative accuracy when GDT_HA, TM-score, and lDDT are used as labels is shown in Table S1. For all labels, plDDT or pTM was the best.

Table S1: **Estimation performance of relative accuracy**

Label	GDT_HA			TM-score			Mean lDDT		
Target number	303			81			76		
Method	Loss	Pearson	Spearman	Loss	Pearson	Spearman	Loss	Pearson	Spearman
DOPE	∗4.010	∗0.229	∗0.192	4.086	0.404	0.303	2.050	0.595	0.491
SOAP	∗3.465	∗0.211	∗0.184	3.273	0.379	0.306	2.045	∗0.536	0.451
ProQ3D	∗4.381	∗0.119	∗0.087	5.216	∗0.263	0.185	∗3.199	∗0.315	∗0.231
SBROD	∗4.747	∗0.039	∗0.030	∗5.178	∗0.173	∗0.153	∗3.386	∗0.210	∗0.195
VoroCNN	∗4.192	∗0.108	∗0.085	3.953	0.283	0.233	∗2.861	∗0.343	∗0.295
P3CMQA	∗3.932	∗0.141	∗0.117	3.911	0.280	0.209	∗2.847	∗0.321	∗0.256
DeepAccNet	∗3.693	∗0.187	∗0.141	2.757	0.366	0.269	1.820	0.563	0.464
DeepAccNet-Bert	∗3.688	∗0.139	∗0.115	3.151	∗0.358	0.286	∗2.453	∗0.436	∗0.371
pLDDT	**3.075**	**0.341**	**0.283**	2.946	0.454	0.332	**1.350**	**0.663**	**0.528**
pTM	3.129	∗0.307	0.262	**2.723**	**0.477**	**0.349**	1.410	∗0.615	0.502
Random selection	∗4.363	-	-	∗5.111	-	-	∗3.413	-	-

The first row represents the label and the second row represents the number of targets. For each label, only targets with a difference between the maximum and minimum values within the target greater than 0.05 were used. The first column shows the method name. The second and subsequent columns represent the average value of the loss, Pearson correlation, and Spearman correlation for each label. The loss values are multiplied by 100 for clarity. An asterisk means that the p-value was less than 0.01 when conducting the Wilcoxon signed rank test against pLDDT.

The distribution of the estimation performance of the relative accuracy for each target when GDT_TS is used as a label is shown in Fig S11.

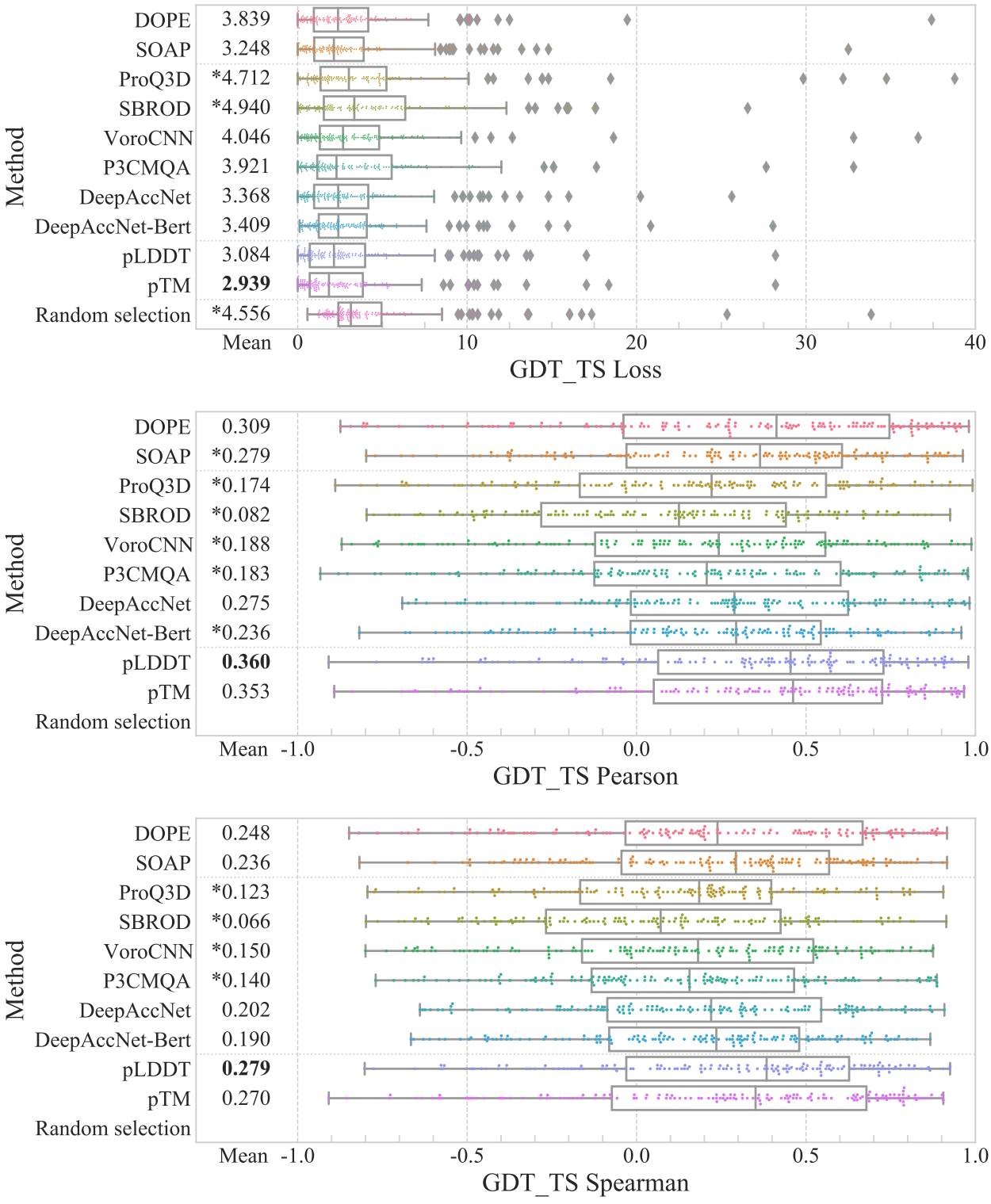


Figure S11: **Distribution of loss, Pearson correlation and Spearman correlation for GDT_TS.** The X-axis shows the value of each metric and the Y-axis shows the method name. A single point represents a single target. The mean values are shown on the left, with the best values in bold. An asterisk means that the p-value was less than 0.01 when conducting the Wilcoxon signed rank test against pLDDT.

2.6 Estimation performance of absolute accuracy

The results of estimation performance of absolute accuracy for the mean lDDT and TM-score are shown in Fig S12 and S13, respectively.

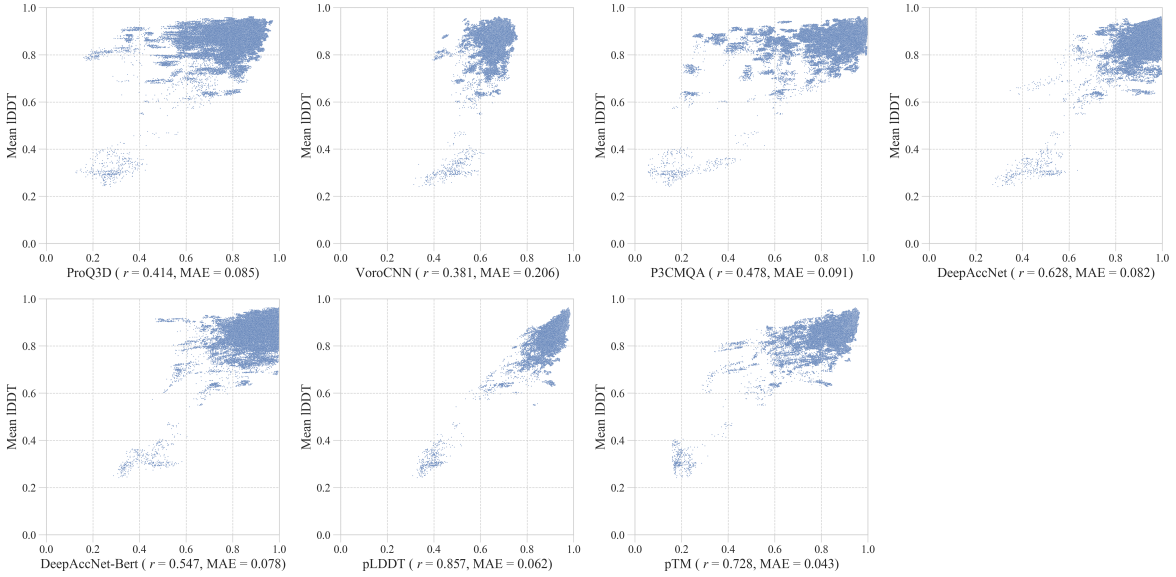


Figure S12: **Scatter plot between mean lDDT and MQA methods.**

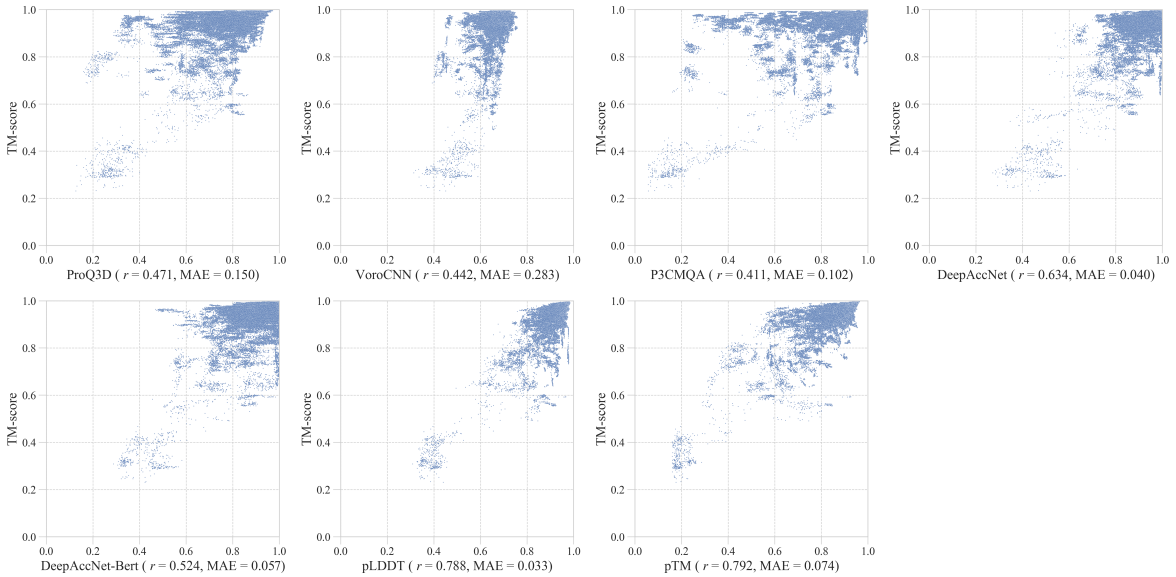


Figure S13: **Scatter plot between TM-score and MQA methods.**

2.7 Factors causing differences in the accuracy

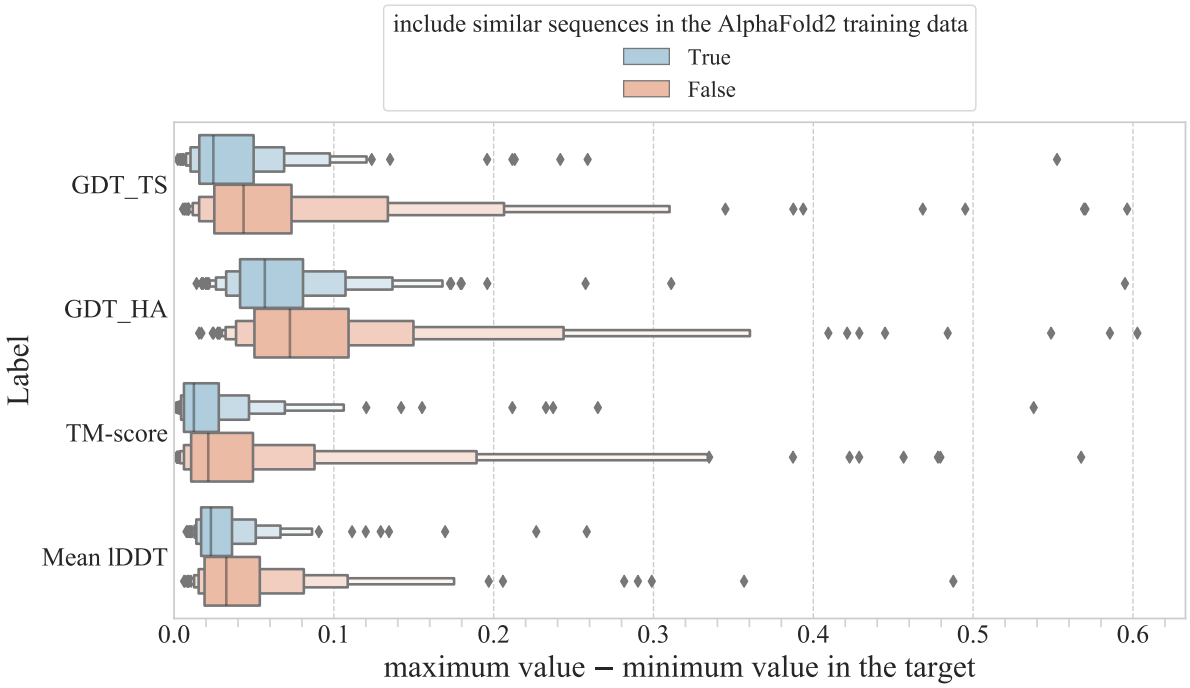


Figure S14: **Distribution of the difference in accuracy with and without similar sequences in the AlphaFold2 training data.** The X-axis shows the difference between the maximum and minimum values of the label. The Y-axis shows the label. The distribution for targets whose similar sequences are included in the AlphaFold2 training data are shown in blue, and those for not included targets are shown in red.

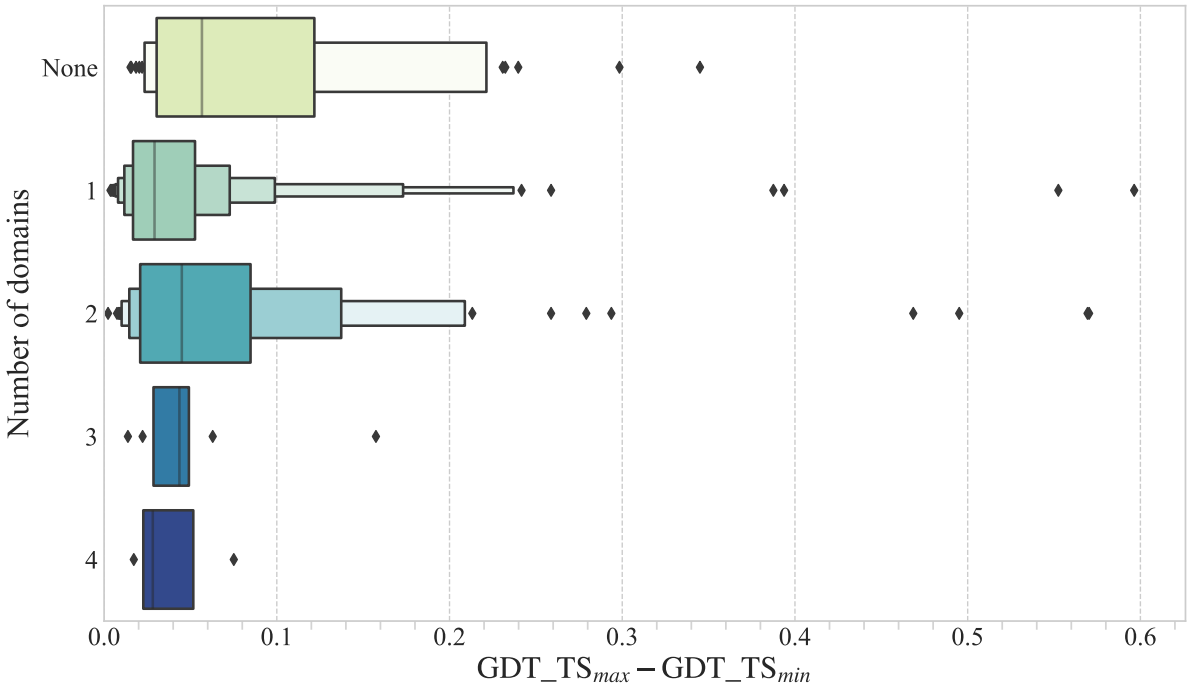


Figure S15: **Distribution of the difference in GDT_TS by domain number.** The X-axis shows the difference between the maximum and minimum GDT_TS. The Y-axis shows the domain number.

2.8 Analysis of the factors that cause AlphaFold2 to fail in accuracy estimation

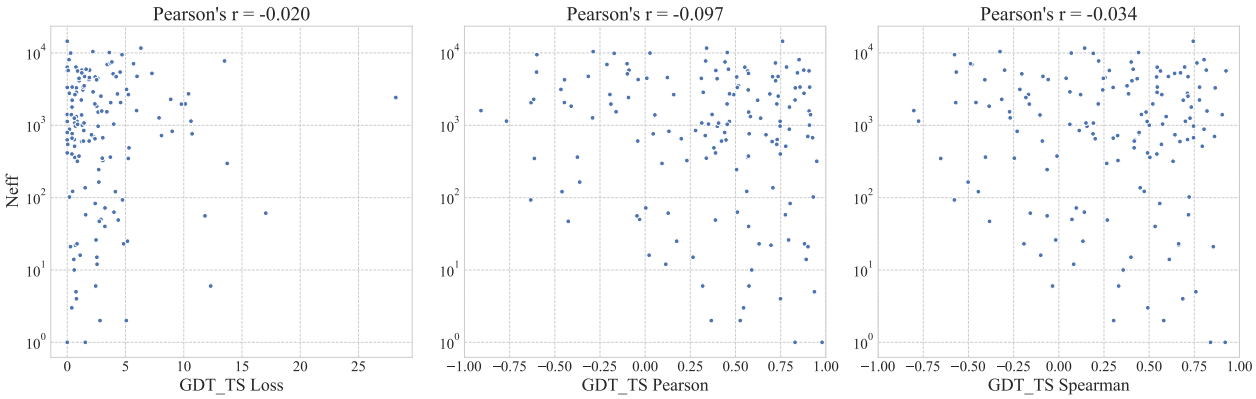


Figure S16: **Scatter plot between Neff and MQA performance of pLDDT.** The X-axis shows the value of each MQA evaluation metric. The Y-axis shows the Neff.

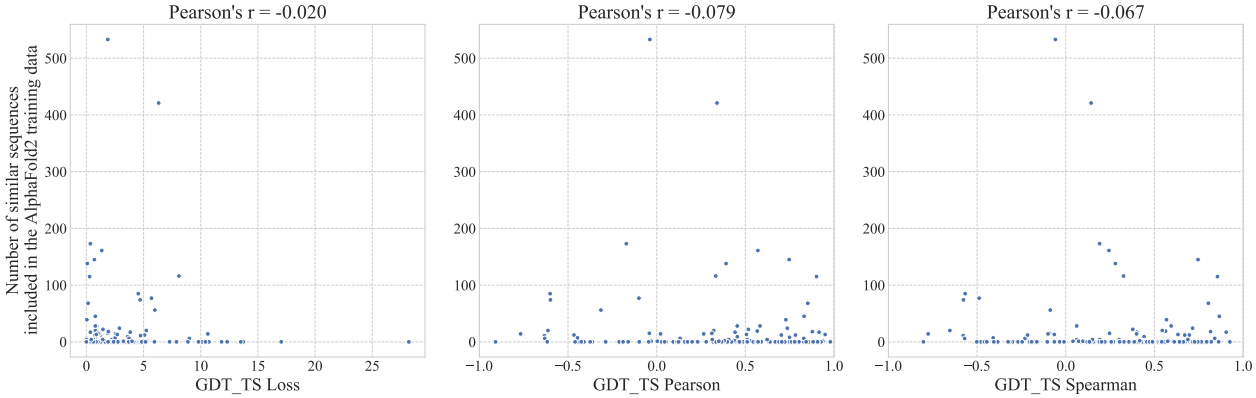


Figure S17: **Scatter plot between number of similar sequences in AlphaFold2 training data and MQA performance of pLDDT.** The X-axis shows the value of each MQA evaluation metric. The Y-axis shows the number of similar sequences included in the AlphaFold2 training data.

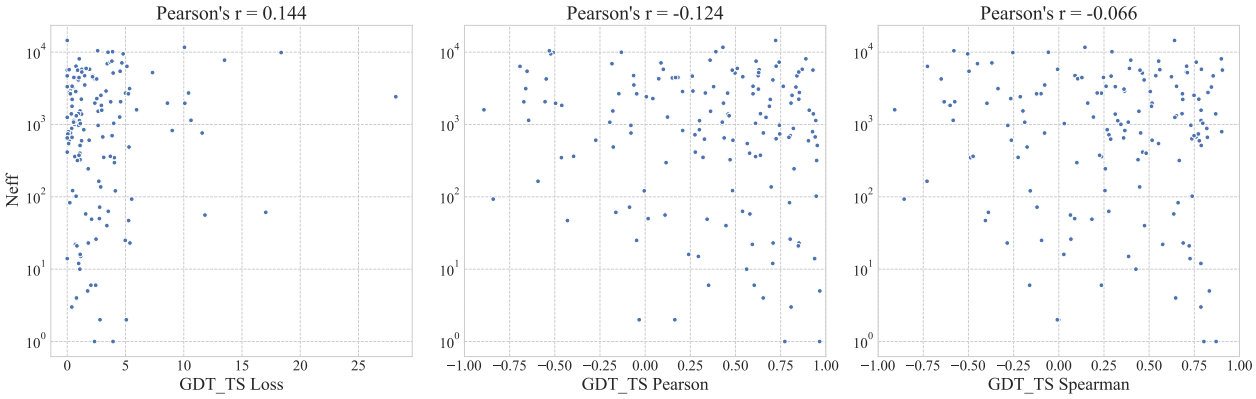


Figure S18: **Scatter plot between Neff and MQA performance of pTM.**

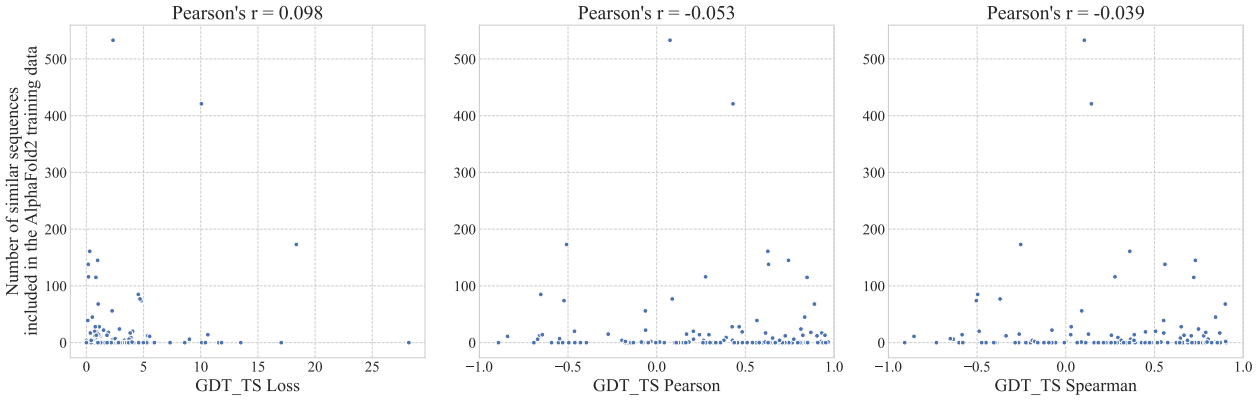


Figure S19: **Scatter plot between number of similar sequences in AlphaFold2 training data and MQA performance of pTM.**

2.9 Target for which pLDDT and pTM failed to select the best structure

Scatter plots between pLDDT and GDT_TS for targets with a GDT_TS loss of pLDDT greater than 10 are shown in Fig S20. For targets with large GDT_TS Loss, there was little correlation between pLDDT and GDT_TS, and some targets were inversely correlated. The best structures in these targets and the selected structures superimposed on the native structures are shown in Fig S21, S22, and S23. The selected structures of pLDDT for targets 5Z6D_A, 6UWA_A, 6RO0_B, 6WVH_A, 6CFV_A, 7LKN_A, and 6DNZ_B were successfully predicted per domain (GDT_TS > 0.9), but the positions between domains were displaced.

The scatter plots between pTM and GDT_TS in targets with GDT_TS loss greater than 10 are shown in Fig S24. Among the 10 targets with GDT_TS loss greater than 10 in pLDDT, 7 targets except 6RO0_B, 5Z6D_A, and 6DNZ_B overlapped, and 3 additional targets, 5OBY_A, 6NES_A, and 6SP9_A were included. The best structures in the 3 targets that do not overlap with pLDDT and the selected structure by pTM superimposed on the native structure are shown in Fig S25.

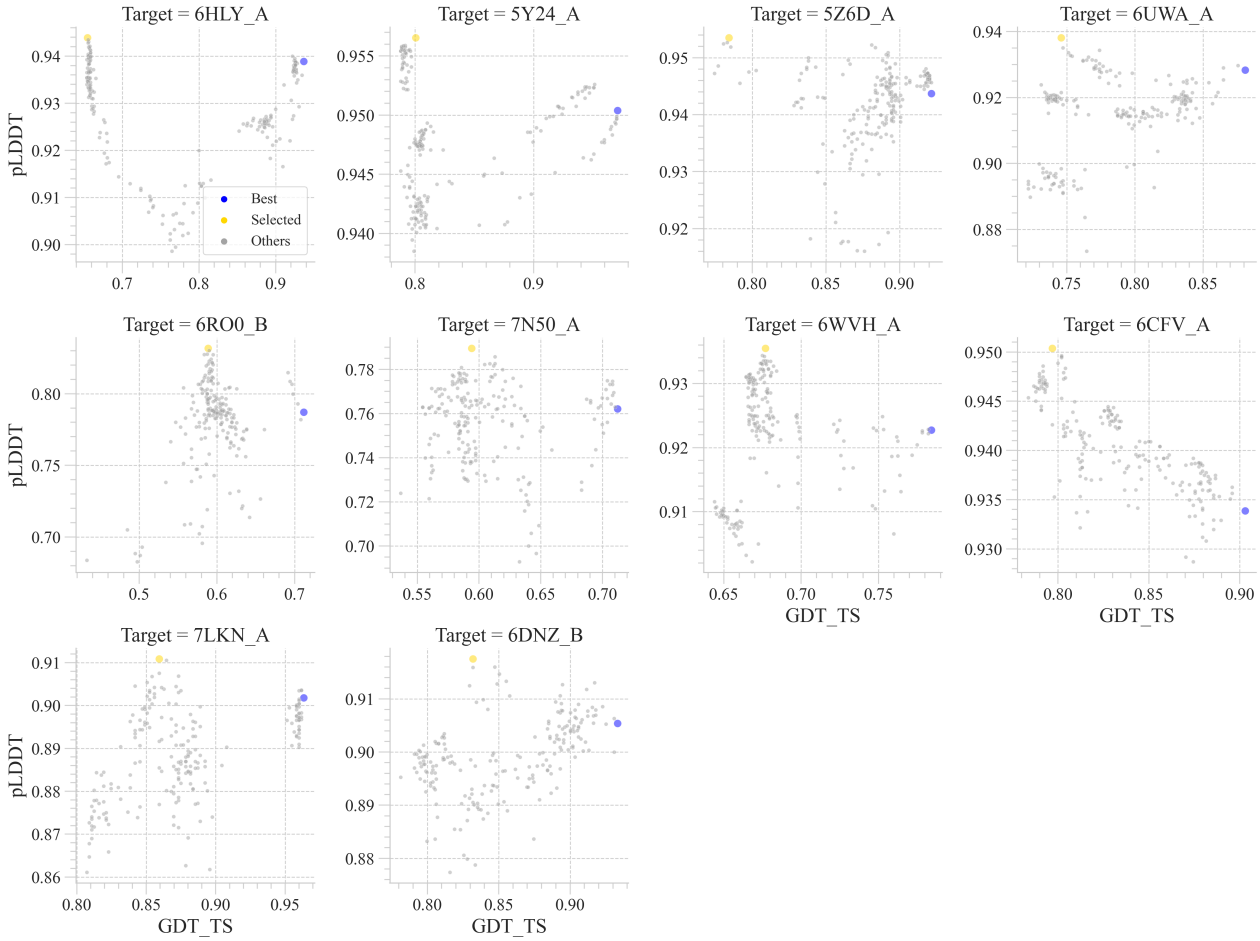


Figure S20: **Scatter plot between GDT_TS and pLDDT for targets where pLDDT failed to select the best structure.** The X-axis shows GDT_TS and the Y-axis shows pLDDT. The structure with the maximum GDT_TS is shown in blue, and the structure with the maximum pLDDT is shown in yellow.

Best model (GDT_TS: 0.937) Model selected by pLDDT (GDT_TS: 0.655)

(a) Target: 6HLY_A[11]

Best model (GDT_TS: 0.977) Model selected by pLDDT (GDT_TS: 0.801)

(b) Target: 5Y24_A[12]

Best model (GDT_TS: 0.922) Model selected by pLDDT (GDT_TS: 0.784)

(c) Target: 5Z6D_A[13]

Best model (GDT_TS: 0.881) Model selected by pLDDT (GDT_TS: 0.739)

(d) Target: 6UWA_A[14]

Figure S21: **Targets for which pLDDT failed to select the best structure[1].** The best structure is shown in blue, the structure selected by pLDDT in yellow, and the native structure in semi-transparent white. Native and predicted structures were superimposed using TM-score.

Best model (GDT_TS: 0.712) Model selected by pLDDT (GDT_TS: 0.589)

(a) Target: 6RO0_B[6]

Best model (GDT_TS: 0.712) Model selected by pLDDT (GDT_TS: 0.594)

(b) Target: 7N50_A[7]

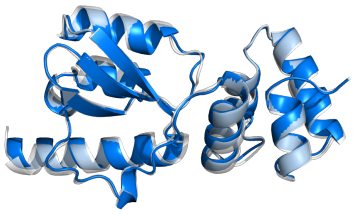
Best model (GDT_TS: 0.799) Model selected by pLDDT (GDT_TS: 0.673)

(c) Target: 6WVH_A[15]

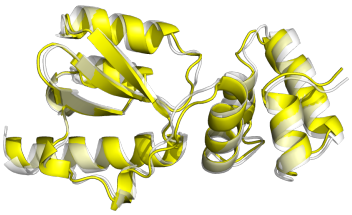
Best model (GDT_TS: 0.903) Model selected by pLDDT (GDT_TS: 0.802)

(d) Target: 6CFV_A[16]

Figure S22: **Targets for which pLDDT failed to select the best structure[2].**

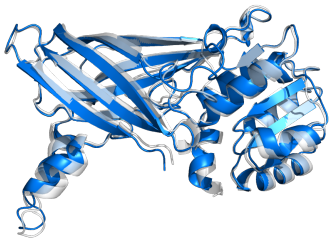


Best model (GDT_TS: 0.964)

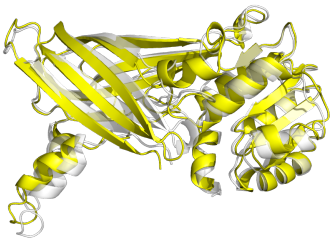


Model selected by pLDDT (GDT_TS: 0.859)

(a) Target: 7LKN_A[17]



Best model (GDT_TS: 0.712)



Model selected by pLDDT (GDT_TS: 0.594)

(b) Target: 6DNZ_B[18]

Figure S23: **Targets for which pLDDT failed to select the best structure[3].**

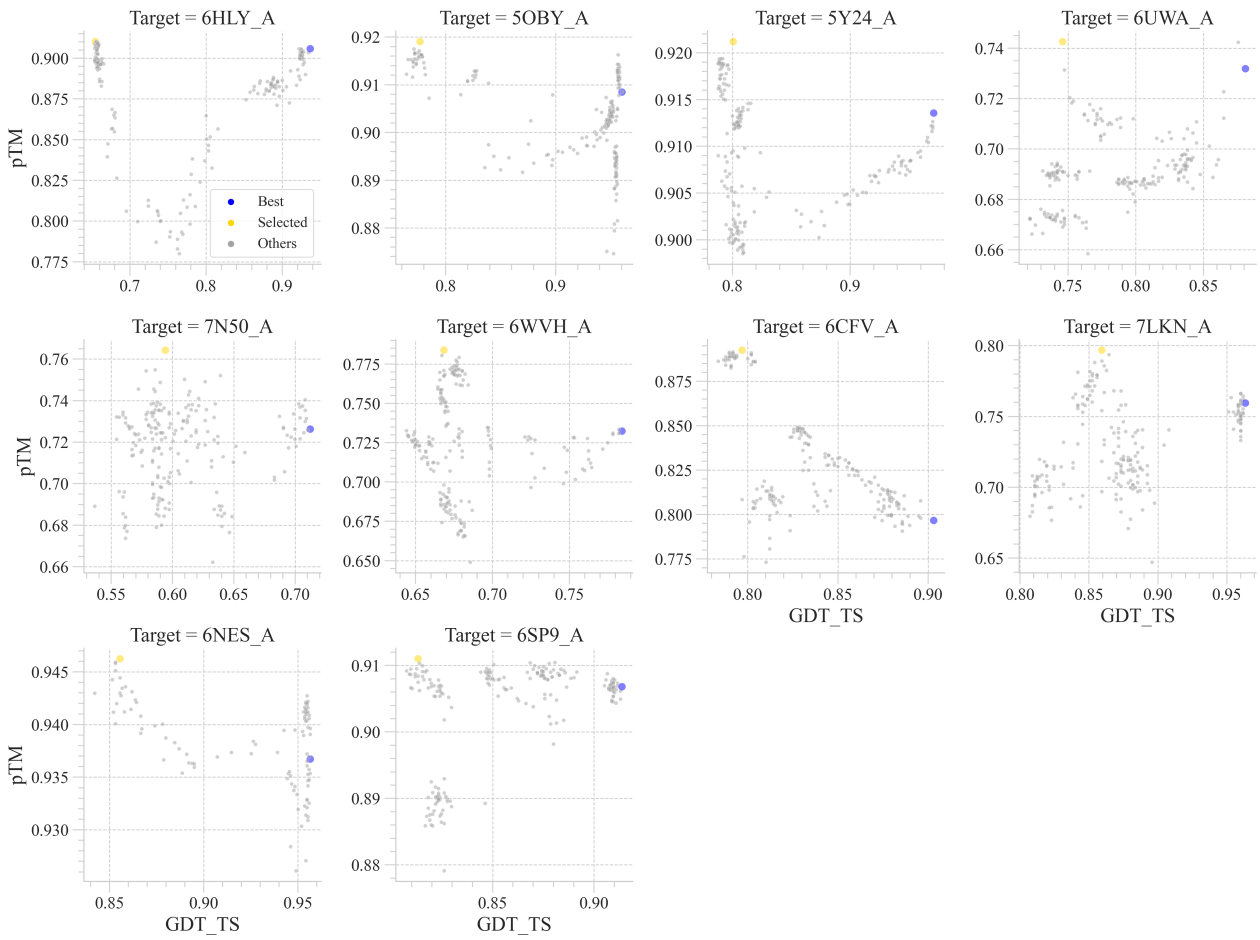
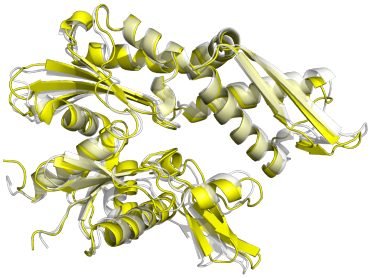


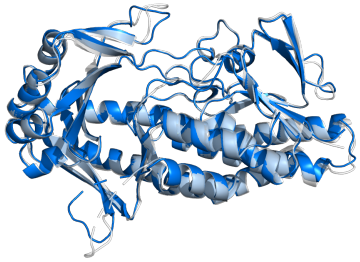
Figure S24: **Scatter plot between GDT_TS and pTM for targets where pTM failed to select the best structure.**

Best model (GDT_TS: 0.960)

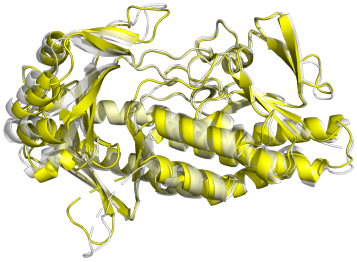


Model selected by pTM (GDT_TS: 0.777)

(a) Target: 5OBY_A[19]

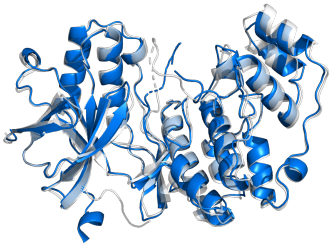


Best model (GDT_TS: 0.957)



Model selected by pTM (GDT_TS: 0.856)

(b) Target: 6NES_A[20]



Best model (GDT_TS: 0.914)

Model selected by pTM (GDT_TS: 0.813)

(c) Target: 6SP9_A[21]

Figure S25: **Targets for which pTM failed to select the best structure.**

2.10 Estimation performance of relative accuracy for single-domain targets

Table S2: Estimation performance of relative accuracy for single-domain targets

Label		GDT_TS			GDT_HA	
Target number		77			181	
Method	Loss	Pearson	Spearman	Loss	Pearson	Spearman
DOPE	2.901	0.373	0.295	∗3.516	∗0.235	0.192
SOAP	2.704	0.305	0.257	∗3.038	∗0.203	∗0.178
ProQ3D	∗4.220	∗0.177	∗0.118	∗4.012	∗0.101	∗0.067
SBROD	∗4.405	∗0.097	∗0.068	∗4.181	∗0.046	∗0.034
VoroCNN	3.317	∗0.201	0.144	∗3.606	∗0.096	∗0.067
Sato-3DCNN	3.012	0.291	0.228	∗3.542	∗0.148	∗0.122
P3CMQA	∗3.381	0.200	0.165	∗3.638	∗0.136	∗0.116
DeepAccNet	2.630	0.335	0.242	∗3.170	∗0.204	∗0.154
DeepAccNet-Bert	∗2.810	∗0.229	0.175	∗3.339	∗0.120	∗0.098
pLDDT	**2.018**	**0.397**	**0.303**	**2.495**	**0.341**	**0.276**
pTM	2.138	0.396	0.287	2.673	0.326	0.269
Random selection	∗3.799	-	-	∗3.887	-	-

Label		Mean lDDT			TM-score	
Target number		48			41	
Method	Loss	Pearson	Spearman	Loss	Pearson	Spearman
DOPE	2.025	0.575	0.484	3.084	0.492	0.367
SOAP	∗2.283	∗0.488	0.436	2.777	0.395	0.322
ProQ3D	∗3.123	∗0.278	∗0.199	4.797	0.281	0.173
SBROD	∗3.237	∗0.170	∗0.146	∗4.770	∗0.184	∗0.132
VoroCNN	∗3.001	∗0.286	∗0.250	3.626	0.311	0.238
Sato-3DCNN	2.515	∗0.468	∗0.368	3.277	0.402	0.300
P3CMQA	∗3.165	∗0.273	∗0.232	3.424	0.315	0.261
DeepAccNet	1.769	0.569	0.483	2.186	0.459	0.323
DeepAccNet-Bert	∗2.496	∗0.405	∗0.343	2.489	∗0.361	0.277
pLDDT	**1.128**	**0.675**	**0.543**	**1.526**	0.525	0.392
pTM	1.269	0.633	0.511	1.583	**0.537**	**0.393**
Random selection	∗3.381	-	-	∗4.087	-	-

The first row shows the label and the second row shows the number of targets. For each label, only targets with a difference between the maximum and minimum values within the target greater than 0.05 were used. The best values are shown in bold. An asterisk means that the p-value was less than 0.01 when conducting the Wilcoxon signed rank test against pLDDT.

References

[1] Weeks SD, De Graef S, Munawar A. X-ray Crystallographic Structure of Orf9b from SARS-CoV-2; 2020.

[2] Foss DV, Schirle NT, MacRae IJ, Pezacki JP. Structural insights into interactions between viral suppressor of RNA silencing protein p19 mutants and small RNAs. FEBS Open Bio. 2019;9(6):1042–1051. doi:https://doi.org/10.1002/2211-5463.12644.

[3] Liu H, Zhu Y, Lu Z, Huang Z. Structural basis of Staphylococcus aureus Cas9 inhibition by AcrIIA14. Nucleic Acids Research. 2021;49(11):6587–6595. doi:10.1093/nar/gkab487.

[4] Sun YJ, Gakhar L, Fuentes EJ. Crystal structure of a consensus PDZ domain; 2019.

[5] Little R, Paiva FCR, Jenkins R, Hong H, Sun Y, Demydchuk Y, et al. Unexpected enzyme-catalysed [4+2] cycloaddition and rearrangement in polyether antibiotic biosynthesis. Nature Catalysis. 2019;2(11):1045–1054. doi:10.1038/s41929-019-0351-2.

[6] Ausar SF, Zhu S, Duprez J, Cohen M, Bertrand T, Steier V, et al. Genetically detoxified pertussis toxin displays near identical structure to its wild-type and exhibits robust immunogenicity. Communications Biology. 2020;3(1):427. doi:10.1038/s42003-020-01153-3.

[7] Johnson AG, Wein T, Mayer ML, Duncan-Lowey B, Yirmiya E, Oppenheimer-Shaanan Y, et al. Bacterial gasdermins reveal an ancient mechanism of cell death. bioRxiv. 2021;doi:10.1101/2021.06.07.447441.

[8] Hu K, Jordan AT, Zhang S, Dhabaria A, Kovach A, Rangel MV, et al. Characterization of Guided Entry of Tail-Anchored Proteins 3 Homologues in Mycobacterium tuberculosis. Journal of Bacteriology. 2019;201(14):e00159–19. doi:10.1128/JB.00159-19.

[9] Basanta B, Bick MJ, Bera AK, Norn C, Chow CM, Carter LP, et al. An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proceedings of the National Academy of Sciences. 2020;117(36):22135–22145. doi:10.1073/pnas.2005412117.

[10] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics. 2004;57(4):702–710. doi:https://doi.org/10.1002/prot.20264.

[11] Marty L, Vigouroux A, Aumont-Nicaise M, Pelissier F, Meyer T, Lavire C, et al. Structural basis for two efficient modes of agropinic acid opine import into the bacterial pathogen Agrobacterium tumefaciens. Biochemical Journal. 2019;476(1):165–178. doi:10.1042/BCJ20180861.

[12] Wang Q, Guan Z, Pei K, Wang J, Liu Z, Yin P, et al. Structural basis of the arbitrium peptide–AimR communication system in the phage lysis–lysogeny decision. Nature microbiology. 2018;3(11):1266–1273.

[13] Swaroop Srivastava S, Raman R, Kiran U, Garg R, Chadalawada S, Pawar AD, et al. Interface interactions between $\beta\gamma$-crystallin domain and Ig-like domain render Ca2+-binding site inoperative in abundant perithecial protein of Neurospora crassa. Molecular Microbiology. 2018;110(6):955–972. doi:https://doi.org/10.1111/mmi.14130.

[14] Aggarwal A, Mire J, Sacchettini JC, Igumenova T. Mouse PKC C1B and C2 domains; 2020.

[15] Liu S, Li S, Shen G, Sukumar N, Krezel AM, Li W. Structural basis of antagonizing the vitamin K catalytic cycle for anticoagulation. Science. 2021;371(6524):eabc5667. doi:10.1126/science.abc5667.

[16] Ji T, Zhang C, Zheng L, Dunaway-Mariano D, Allen KN. Structural basis of the molecular switch between phosphatase and mutase functions of human phosphomannomutase 1 under ischemic conditions. Biochemistry. 2018;57(25):3480–3492.

[17] Llontop EE, Cenens W, Favaro DC, Sgro GG, Salinas RK, Guzzo CR, et al. The PilB-PilZ-FimX regulatory complex of the Type IV pilus from Xanthomonas citri. PLOS Pathogens. 2021;17(8):1–35. doi:10.1371/journal.ppat.1009808.

[18] Hashimoto H, Kafková L, Raczkowski A, Jordan KD, Read LK, Debler EW. Structural Basis of Protein Arginine Methyltransferase Activation by a Catalytically Dead Homolog (Prozyme). Journal of Molecular Biology. 2020;432(2):410–426. doi:https://doi.org/10.1016/j.jmb.2019.11.002.

[19] Adell M, Calisto BM, Fita I, Martinelli L. The nucleotide-bound/substrate-bound conformation of the Mycoplasma genitalium DnaK chaperone. Protein Science. 2018;27(5):1000–1007. doi:https://doi.org/10.1002/pro.3401.

[20] Rodríguez Benítez A, Tweedy SE, Baker Dockrey SA, Lukowski AL, Wymore T, Khare D, et al. Structural Basis for Selectivity in Flavin-Dependent Monooxygenase-Catalyzed Oxidative Dearomatization. ACS Catalysis. 2019;9(4):3633–3640. doi:10.1021/acscatal.8b04575.

[21] Nichols C, Ng J, Keshu A, Kelly G, Conte MR, Marber MS, et al. Mining the PDB for Tractable Cases Where X-ray Crystallography Combined with Fragment Screens Can Be Used to Systematically Design Protein–Protein Inhibitors: Two Test Cases Illustrated by IL1β-IL1R and p38α–TAB1 Complexes. Journal of Medicinal Chemistry. 2020;63(14):7559–7568.