Supplementary Information for

Distance-AF Improves Predicted Protein Structure Models by AlphaFold2 with User-Specified Distance Constraints

Yuanyuan Zhang[1], Zicong Zhang[1], Yuki Kagaya[2], Genki Terashi[2], Bowen Zhao[3], Yi Xiong[3] & Daisuke Kihara[1,2*]

[1] Department of Computer Science, Purdue University, West Lafayette, Indiana, 47907, USA

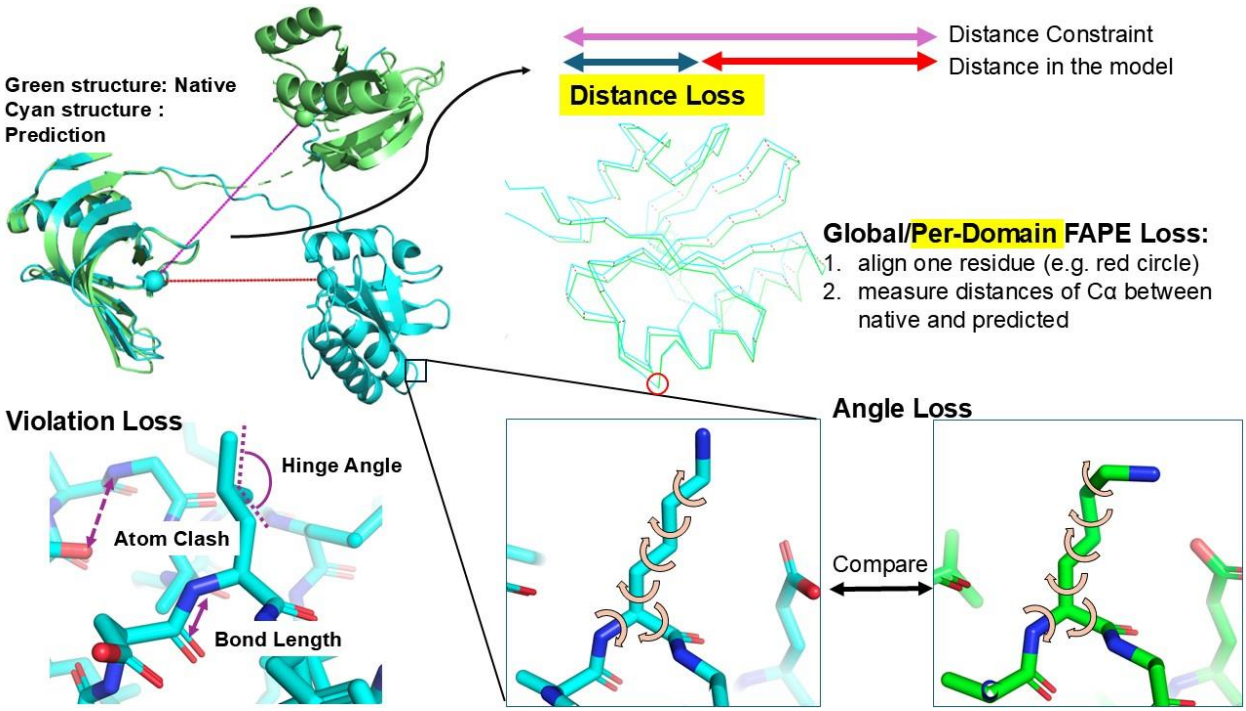
[2] Department of Biological Sciences, Purdue University, West Lafayette, Indiana, 47907, USA

[3] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

Algorithm 1 Intra-Domain FAPE Loss($domain_1$, $domain_2$, $\vec{x}_{pred}$, $\vec{x}_{initial}$, T_{pred}, $T_{initial}$,$\epsilon = 10^{-4}\mathring{A}^2$)

$\vec{x}_{pred\_local} = T_{pred}^{-1} \circ \vec{x}_{pred}$
$\vec{x}_{initial\_local} = T_{initial}^{-1} \circ \vec{x}_{initial}$
$\vec{x}_{pred\_local}, \vec{x}_{initial\_local} \in R^{L \times L \times 3}$
$d = \sqrt{\|\vec{x}_{pred\_local} - \vec{x}_{initial\_local}\|^2 + \epsilon}$
$d_{i,j} = 0, i \in domain_1, j \in domain_2$ & $i \in domain_2, j \in domain_1$
$L_{FAPE} = \frac{1}{10} mean_{i,j}(min(10, d_{i,j}))$
return L_{FAPE}

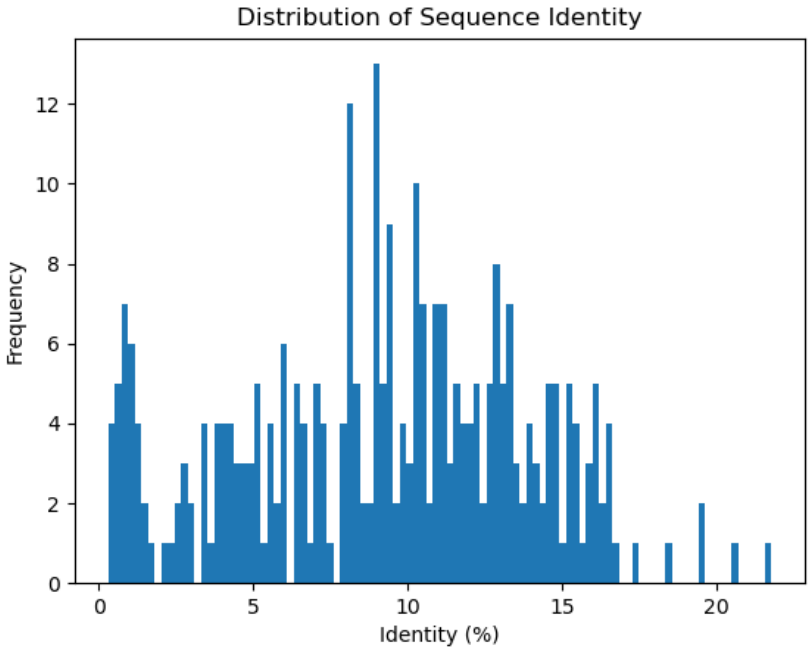
Supplementary Figure 1. The pseudo code of the modified FAPE loss that only considers the residues belonging to intra-domain.



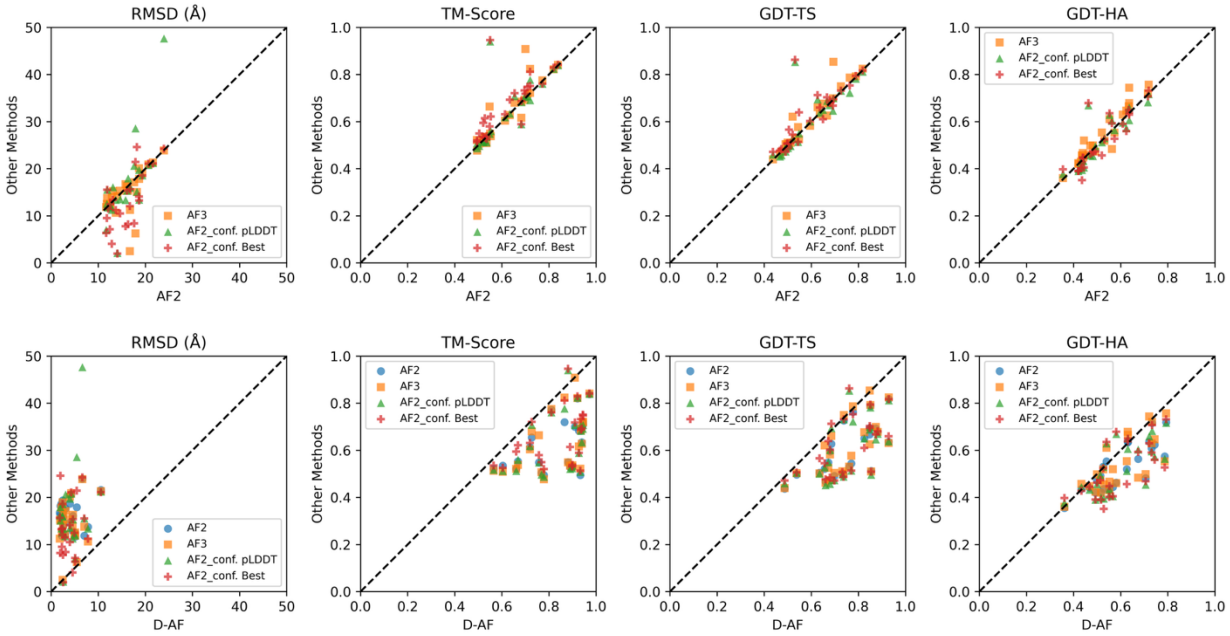
Supplementary Figure 2. Loss functions used in AF2 and in the Distance-AF.

Loss function terms used in the structural module of AF2 and Distance-AF are visualized. They are listed in Eq. 1 and Eq. 2in the main text. New loss terms introduced in Distance-AF, Distance Loss, and Per-Domain FAPE loss, are highlighted in yellow. The other loss terms are common between AF2 and Distance-AF.

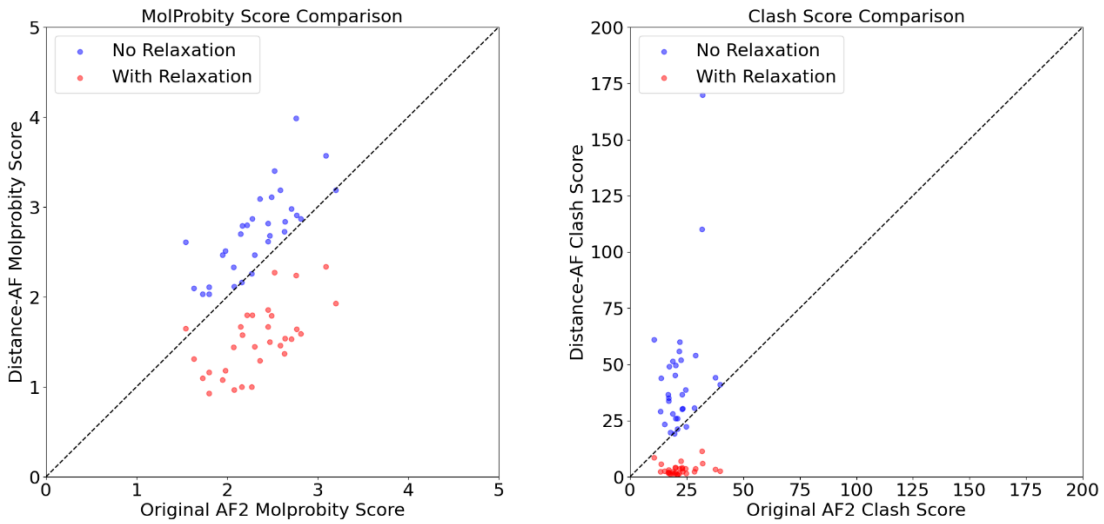
The Distance loss (Eq. 1) represents the difference between the (sum of) the user-provided distance information (magenta arrow) and the current distance (red arrow) in the model (blue arrow). In AF2, during training, the (global) FAPE loss considers distances between residues to try to construct the correct (i.e. native) structure. In Distance-AF, Domain level FAPE loss is for making individual domain structures close to the original domain structures in the starting structure, which may be distorted while iteratively modifying the conformation of the structure model. The angle loss in the Distance-AF tries to keep rotation angles in the model close to those in the starting structure. The violation loss is to keep the structure physically protein-like, i.e. avoiding deviation of hinge angles, bond lengths from the statistics of proteins, and avoid atom clashes.



Supplementary Figure 3. The sequence identity between pairs of the 25 targets. As shown, all pairs have less than 25% sequence identity.



Supplementary Figure 4. Structure prediction results with AlphaFold2 (AF2), AlphaFold3 (AF3), and AlphaFold2-Conformations. Top panel, comparison between AF2 and AF3, AF2-Conformations (pLDDT) and AF2-Conformations (Best). 130 models were generated by AF2-Conformations. Among them, the highest pLDDT model was selected for AF2-Conformations (pLDDT), while the best (i.e. highest) TM-Score model was selected for AF2-Conformations (Best). Bottom panel comparison with Distance-AF. Individual results are provided in Supplementary Table 2.



Supplementary Figure 5. Molprobity score of Distance-AF models. Left, Comparison of the Molprobity score of the starting AF2 models and Distance-AF models. Blue, before Amber relaxation; red, after relaxation. Molprobity score considers all-atom steric clashes, chain geometry, and Ramachandran, side-chain rotamer, and backbone outliers. The smaller, the better. Right, atom clash scores of the original AF2 models and corresponding Distance-AF models. The smaller, the better.

Distance-AF models have slightly larger Molprobity scores (almost all of them are within a margin of 0.2) (left panel, blue). After applying structure relaxation using molecular dynamics (MD), Molprobity scores become lower (better) than AF2 models for all but one case.

In terms of atom clash score, most of the Distance-AF models had a score 50 or lower, which is consistent with the score distribution in PDB (experimental structures) (blue). But there are a few models that have a larger clash score. After the MD relaxation, all the Distance-AF models lowered their clash scores, even lower than AF2 models.

Supplementary Table 1.

PDB ID	Chain ID	Length	AF2		Distance-AF		Rosetta
			Time	GPU memory	Time	GPU memory	Time
2BJ7	A	138	1:02	3.5 GB	0:9.5	1.8 GB	2:17
1MKM	A	249	1:10	4.1 GB	0:10.1	2.3 GB	1:54
2VLD	A	251	1:11	4.1 GB	0:9.6	2.2 GB	2:43
1NT2	B	258	1:19	4.1 GB	0:9.8	2.3 GB	2:35
1IXC	A	294	1:28	4.3 GB	0:14.2	2.5 GB	2:24
3BA0	A	365	2:15	4.8GB	1:35	4.2 GB	4:51
4TX8	A	439	2:09	5.3 GB	1:47	6.6 GB	4:19
3IPL	A	501	4:10	5.8 GB	2:58	9.8 GB	8:25

Computational time is shown in the format of hours: minutes.