Supporting Information

Enhancing Protein Aggregation Prediction: A Unified Analysis Leveraging Graph Convolutional Networks and Active Learning

Jiwon Sun[1,+], JunHo Song[1,+], Juo Kim[1,+], Seungpyo Kang[1], Eunyoung Park[2], Seungwoo Seo[2,*], and Kyoungmin Min[1,*]

[1]School of Mechanical Engineering, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Republic of Korea

[2]AinB, 160 Yeoksam-ro, Gangnam-gu, Seoul 06249, Republic of Korea

[+]These authors contributed equally to this work.

[*]Corresponding author: seungwoo.seo@ainbsci.com (S. Seo), kmin.min@ssu.ac.kr (K. Min)

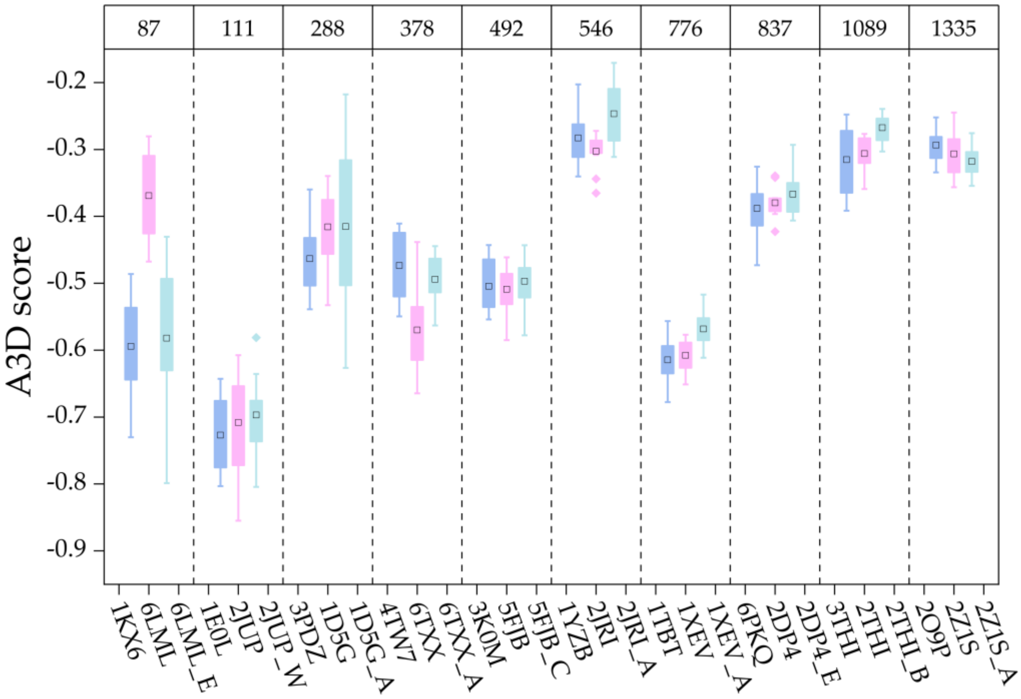


Figure S1. Represents the box plot of the A3D score of protein structures with the same AA sequence. Protein structures that have the same AA sequence were categorized by the AA sequence lengths represented at the top. The light blue box means the original single-polypeptide chains; the pink box means the single polypeptide chains within multi-polypeptide chains; and the cyan box means the single-polypeptide chains that were divided from the multi-polypeptide chains. The square box represents the average A3D score of 12 different protein structures in dynamic mode. The box in a box plot represents data covering around 75%, the portions marked with a rectangle signify outliers.

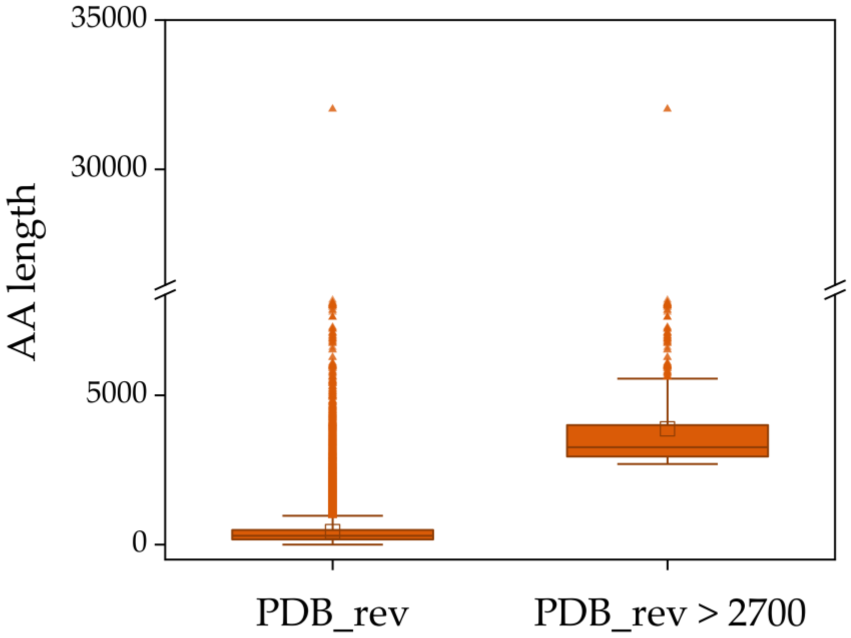


Figure S2. Represents box plot of AA length for PDB_rev data and PDB_rev data having over 2700 AA length. The box in a box plot represents data covering around 75%, the portions marked with triangles signify outliers, and the parts marked with squares indicate the average value.

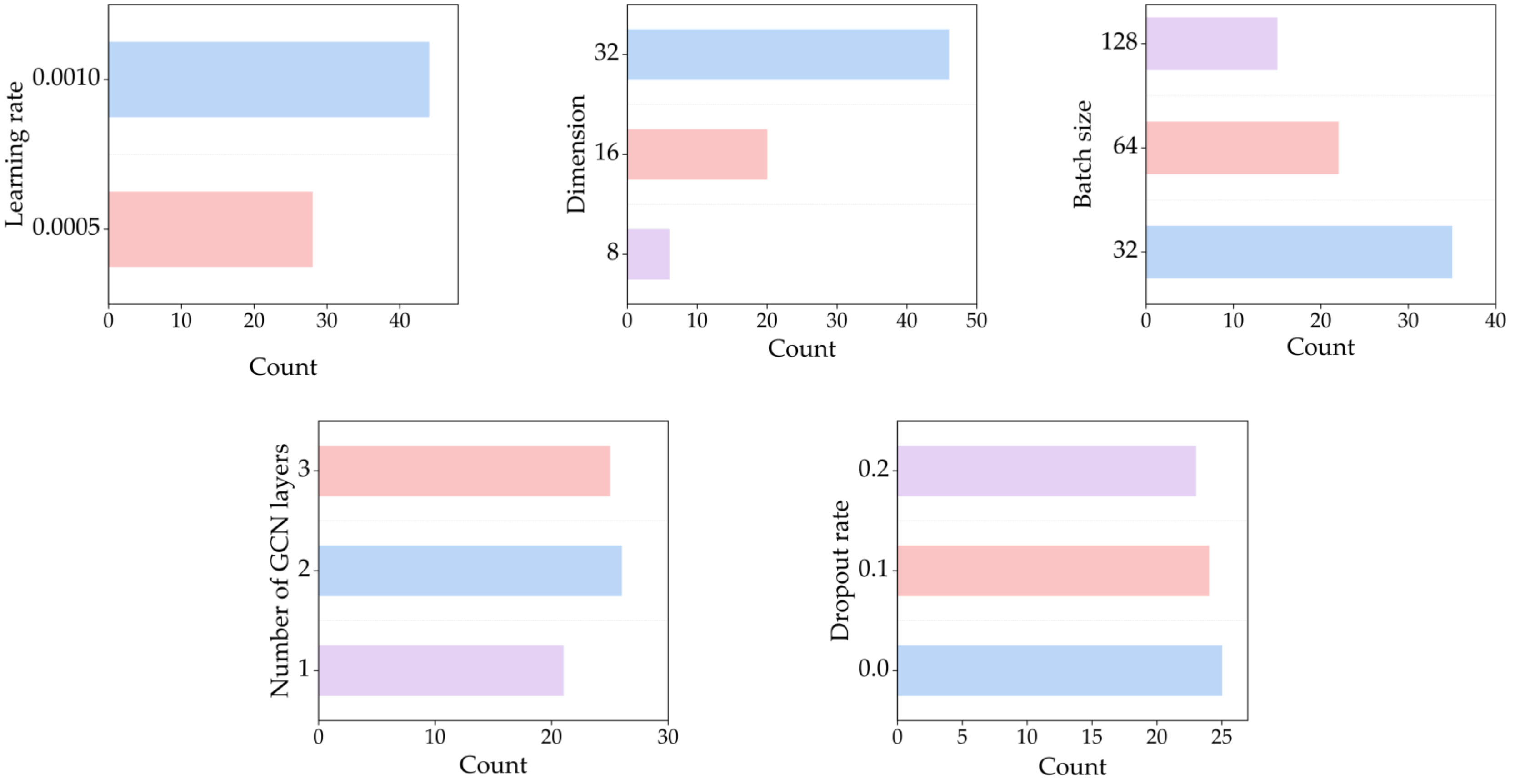
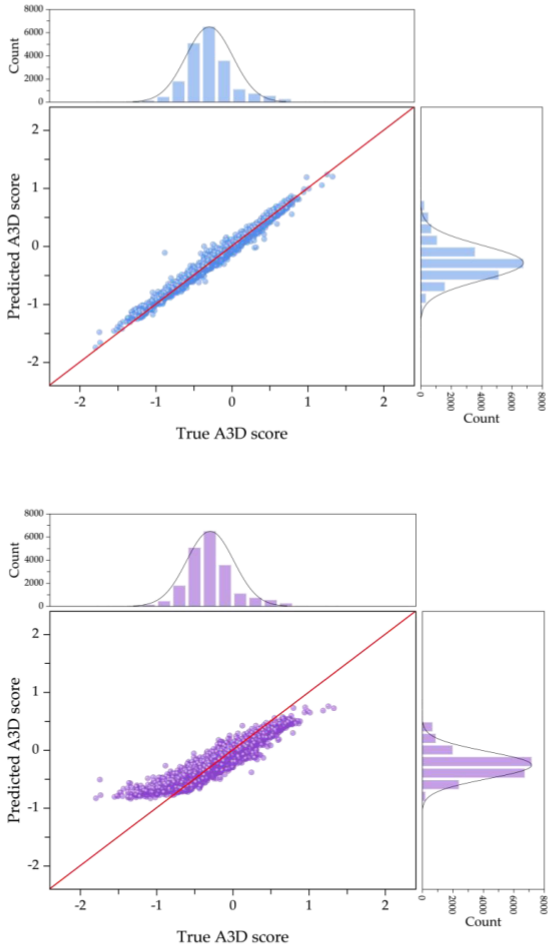


Figure S3. Distribution of hyperparameters in the top ten percent performing models: (a) learning rate, (b) dimension, (c) batch size, (d) number of GCN layers, and (e) dropout rate.



(a)

	R^2	MAE
Average	0.9820	0.0350
Standard deviation	0.0031	0.0029
Max	0.9849	0.0381
Min	0.9742	0.0278

(b)

	R^2	MAE
Average	0.8466	0.0786
Standard deviation	0.2190	0.0431
Max	0.9645	0.2305
Min	0.0110	0.0417

Figure S4. (Left) Comparison of predicted versus calculated A3D score value with GCN model, and (right) corresponding R^2 and MAE values and their statistical output (from 20 different prediction models from a randomly chosen training set) for (a) train size 80%, and (b) train size 0.1%, model.

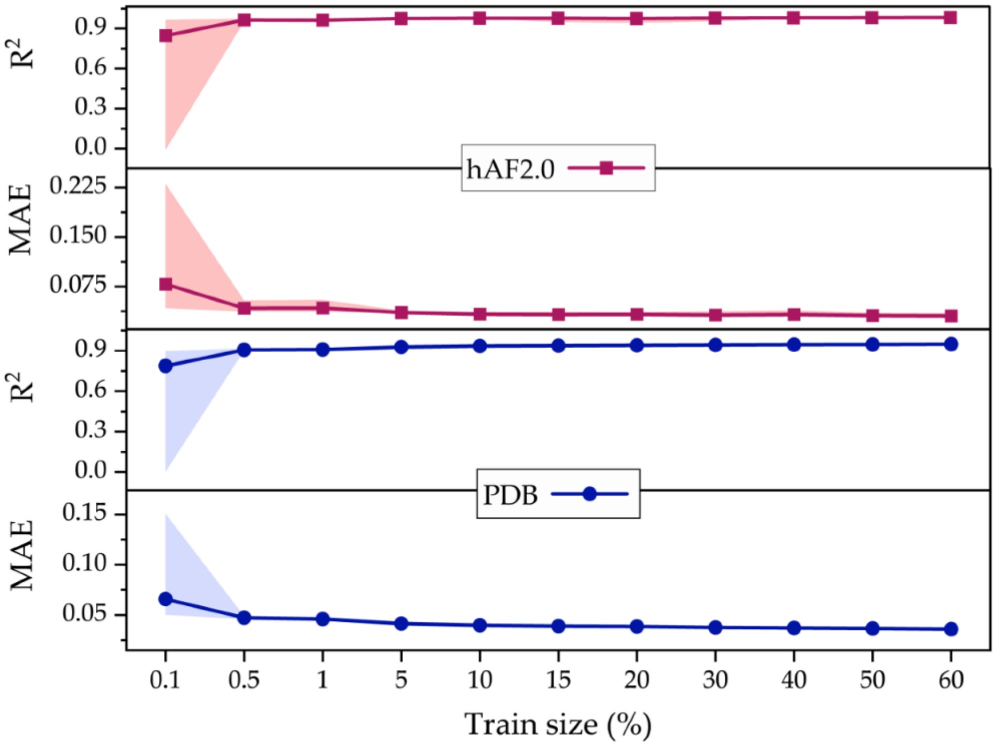


Figure S5. Performance visualization of MAE and R^2 with error range for each train size (0.1% to 60%) of PDB and hAF2.0.

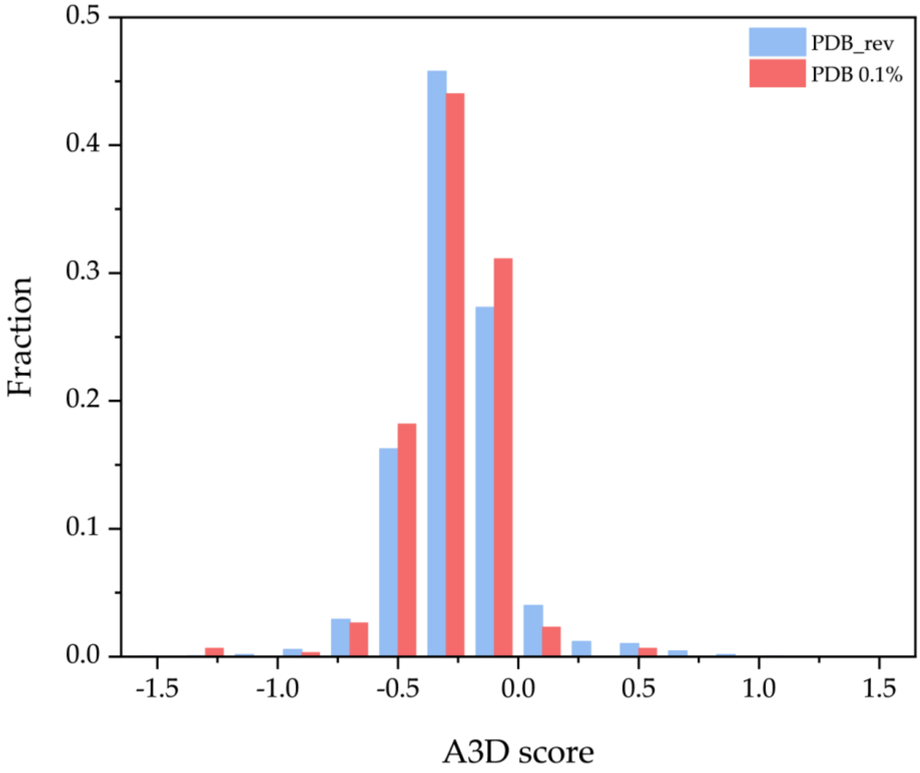


Figure S6. A3D score distribution from PDB_rev and selected PDB for 0.1% datasets

AA lengths	PDB	Standard deviation	average A3D score
87	1KX6	0.0715	-0.5945
	6LML	0.0643	-0.3692
	6LML_E	0.1130	-0.5820
111	1E0L	0.0548	-0.7269
	2JUP	0.0756	-0.7083
	2JUP_W	0.0551	-0.6970
288	3PDZ	0.0531	-0.4634
	1D5G	0.0546	-0.4159
	1D5G_A	0.1185	-0.4153
378	4TW7	0.0475	-0.4735
	6TXX	0.0594	-0.5698
	6TXX_A	0.0355	-0.4942
492	3K0M	0.0383	-0.5046
	5FJB	0.0350	-0.5089
	5FJB_C	0.0359	-0.4970
546	1YZB	0.0384	-0.2830
	2JRI	0.0252	-0.3027
	2JRI_A	0.0456	-0.2467
776	1TBT	0.0337	-0.6144
	1XEV	0.0229	-0.6078
	1XEV_A	0.0251	-0.5683
837	6PKQ	0.0393	-0.3881
	2DP4	0.0217	-0.3797
	2DP4_E	0.0305	-0.3668
1089	3THI	0.0469	-0.3151
	2THI	0.0263	-0.3061
	2THI_B	0.0205	-0.2675
1335	2O9P	0.0244	-0.2936
	2Z1S	0.0305	-0.3067
	2Z1S_A	0.0215	-0.3180

Table S1. Represents the average and standard deviation (std) of the A3D score that have the same Amino Acid (AA) sequences. The proteins sharing the same AA were categorized by the AA lengths as same as in **Figure S1**

Model	Inference time (sec)
GCN	0.5
Aggrescan 3.0	21,240

Table S2. The inference time for randomly selected 30 different proteins. Both benchmarks were performed using a CPU core of Intel(R) Xeon(R) Gold 6240 @ 2.60 GHz. The GCN model utilized two RTX 3090 24GB GPUs, while the Aggrescan 3.0 used only the CPU.