Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information

Rapid Prediction of Protein Natural Frequencies using Graph Neural Networks

Kai Guo[1,2] and Markus J. Buehler[1,3,4*]

[1] Laboratory for Atomistic and Molecular Mechanics (LAMM), Massachusetts Institute of Technology, 77 Massachusetts Ave. 1-165, Cambridge, Massachusetts 02139, United States of America

[2] Institute of High Performance Computing, A*STAR, Singapore 138632, Singapore

[3] Center for Computational Science and Engineering, Schwarzman College of Computing, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, United States of America

[4] Center for Materials Science and Engineering 77 Massachusetts Ave, Cambridge, Massachusetts 02139, United States of America

[*]Address correspondence to: mbuehler@MIT.EDU, +1.617.452.2750

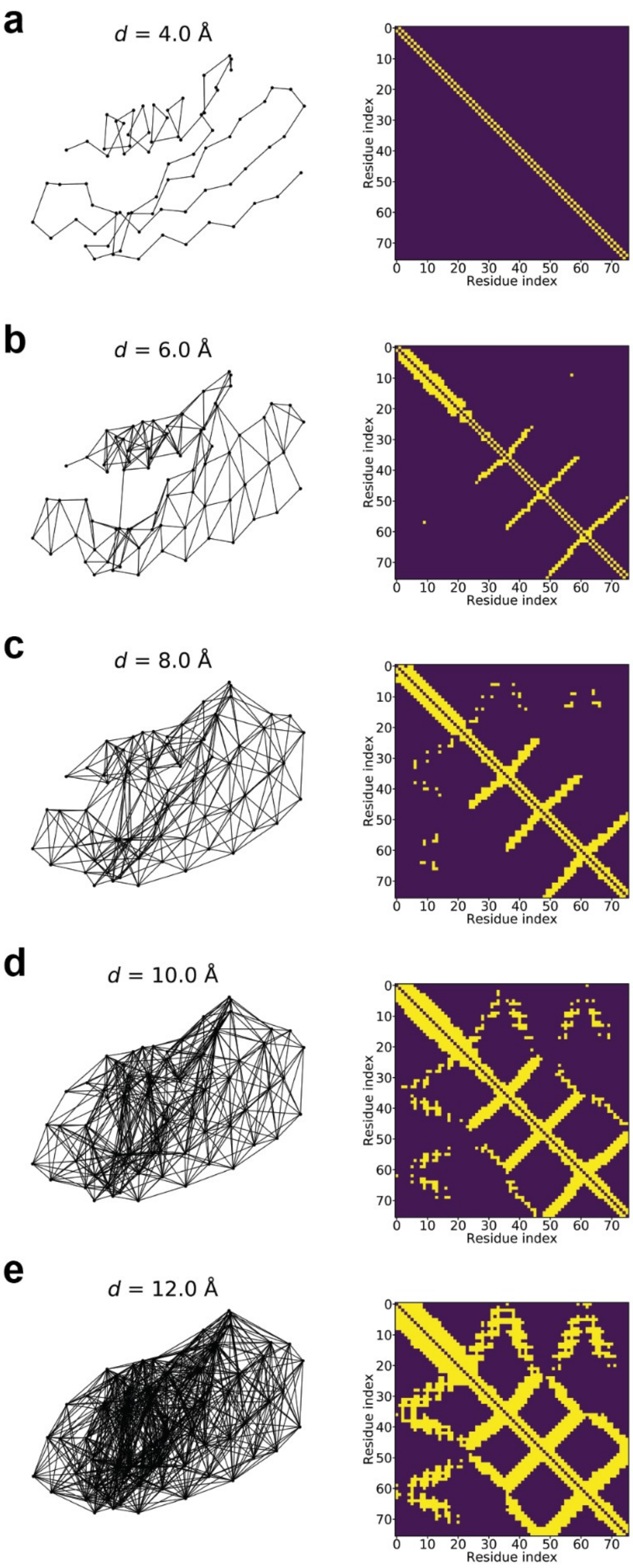

Figure S1: Graph and adjacency matrix of an example protein (PDB ID: 4R80) with a threshold distance of (a) 4 Å; (b) 6 Å; (c) 8 Å; (d) 10 Å; (e) 12 Å.

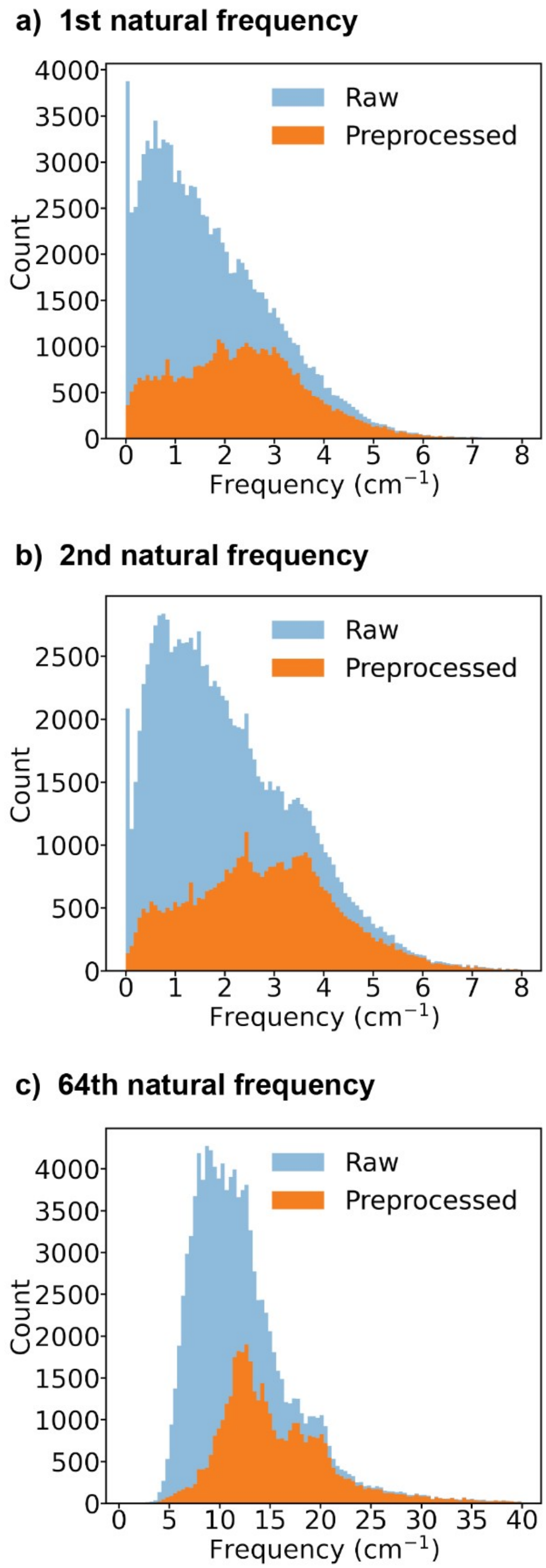

Figure S2: Comparison between the frequency distributions in the raw database and in the preprocessed protein graphs for the (a) 1st, (b) 2nd, (c) 64th natural frequency.

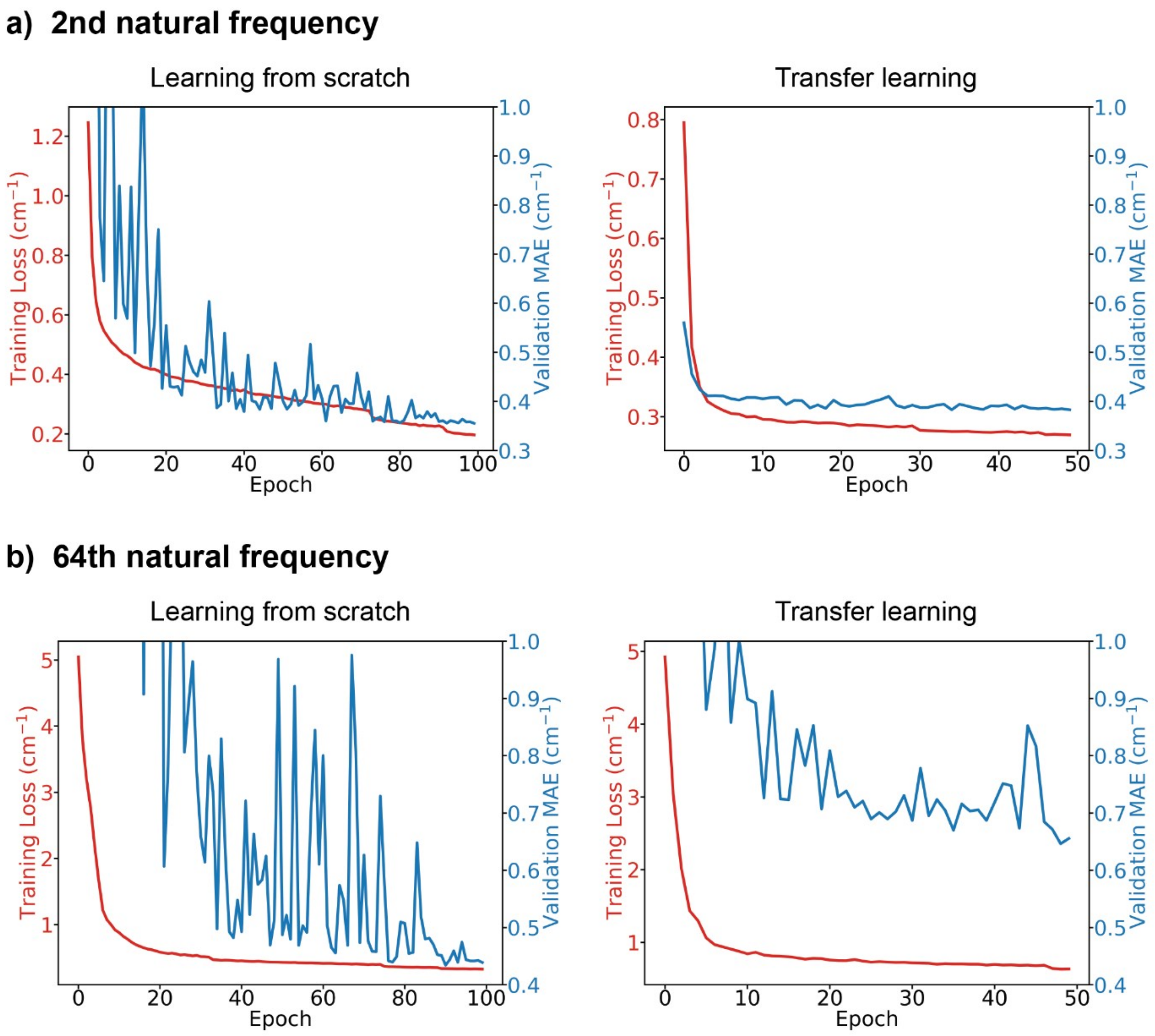

Figure S3: Learning curves of the models trained from scratch or trained via transfer learning for the (a) 2nd, and (b) 64th natural frequency.

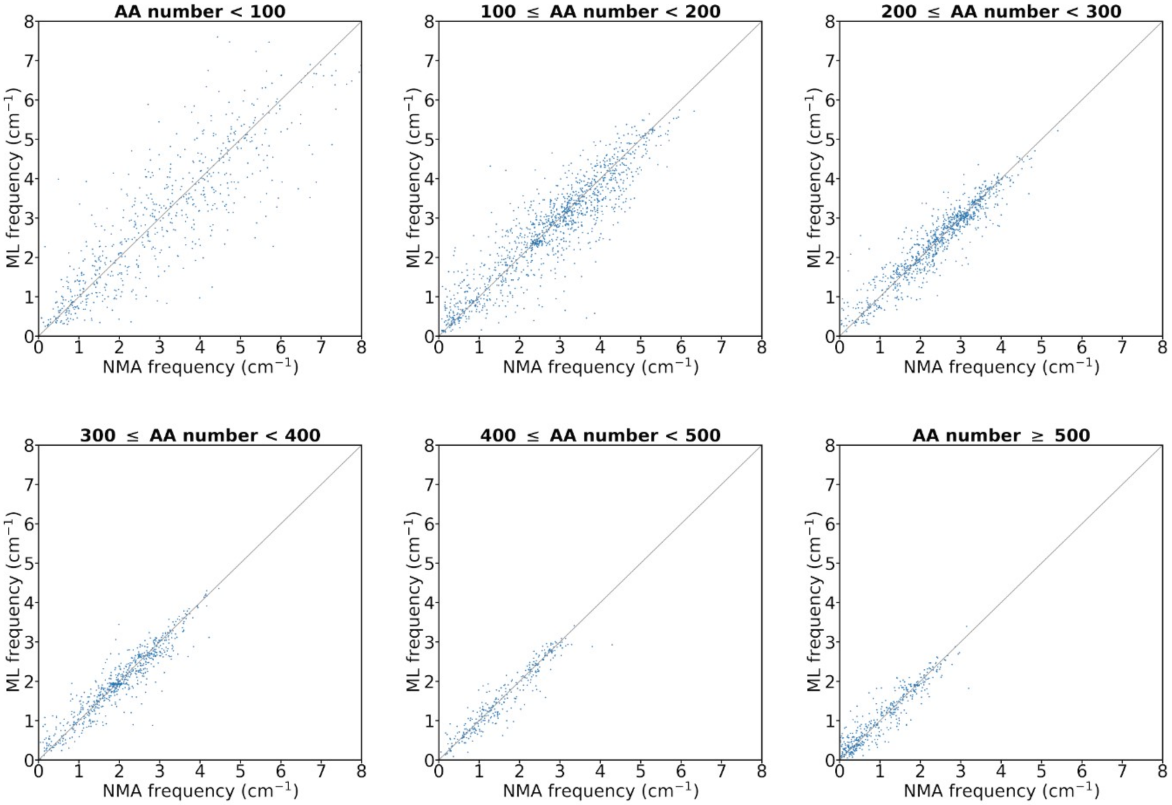

Figure S4: Comparison between the ML-predicted and NMA-calculated 1st natural frequency of proteins with different numbers of amino acids (AA) in the test set.

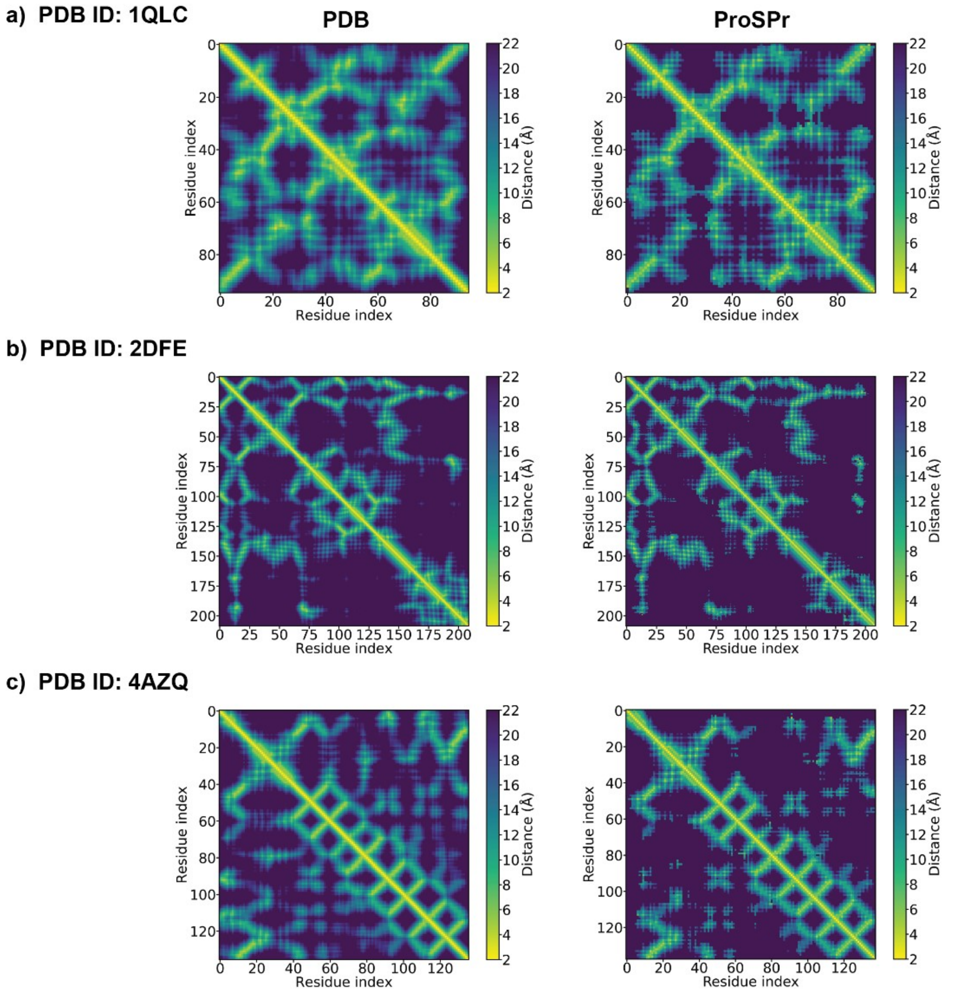

Figure S5: Comparison of the distance maps generated from PDB structures (left) and predicted by ProSPr (right) of a test protein with a PDB ID of (a) 1QLC, (b) 2DFE, (c) 4AZQ.

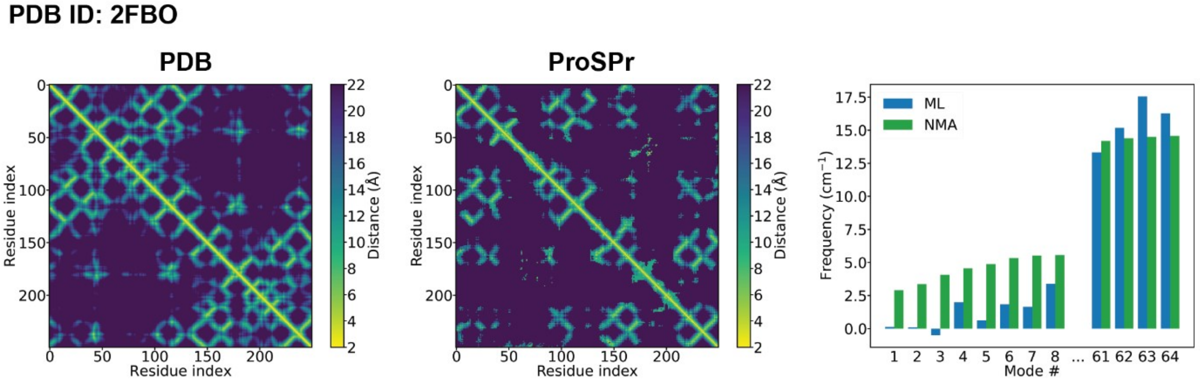

Figure S6: The distance maps generated from PDB structures and predicted by ProSPr, and the 1st-8th and 61-64th frequencies of a test protein with a PDB ID of 2FBO.

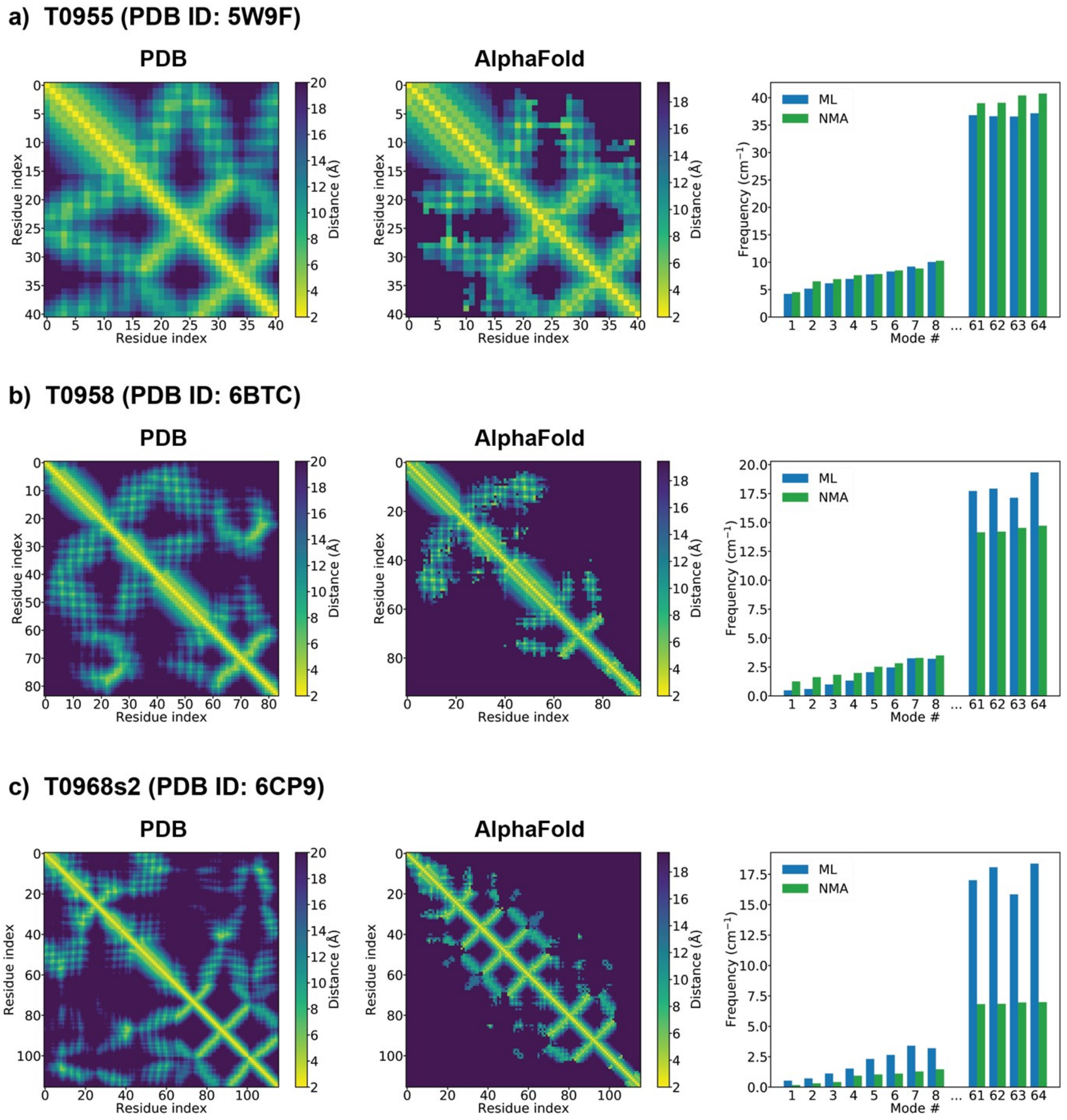

Figure S7: Comparison of the distance maps generated from PDB structures (left) and predicted by AlphaFold 1 (middle), and the 1st-8th and 61-64th frequencies (right) of a CASP13 target (a) T0955 (PDB ID: 5W9F), (b) T0958 (PDB ID: 6BTC), (c) T0968s2 (PDB ID: 6CP9).

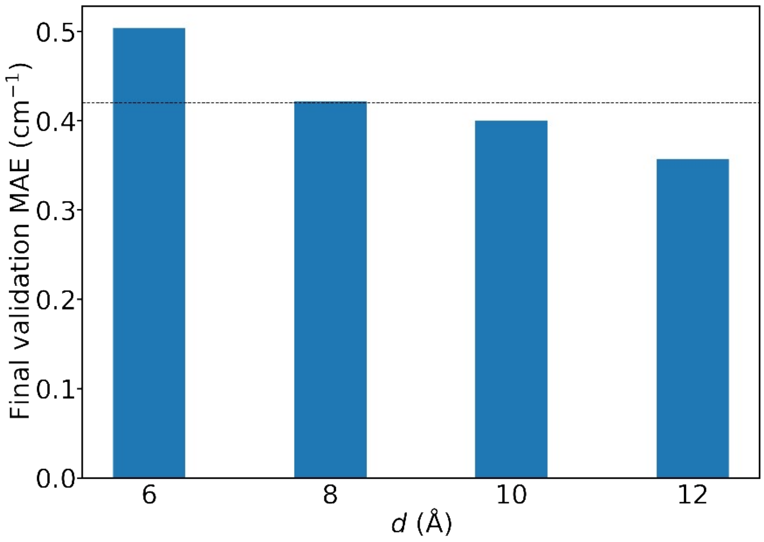

Figure S8: Comparison of the final validation mean absolute error (MAE) of the models trained with different threshold distances to define edges in protein graphs. The horizontal baseline denotes the mean value.

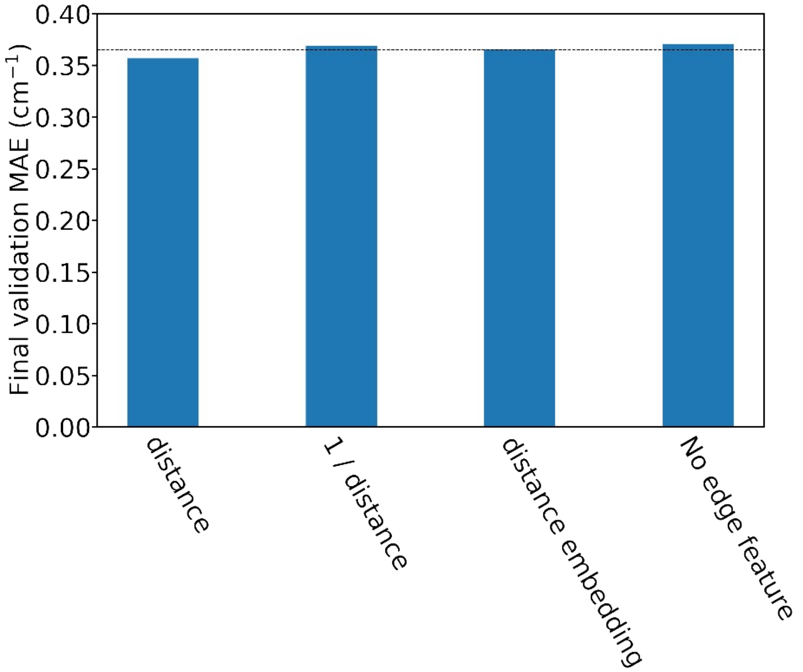

Figure S9: Comparison of the final validation MAE of the models trained with different representations of edge feature. The horizontal baseline denotes the mean value.

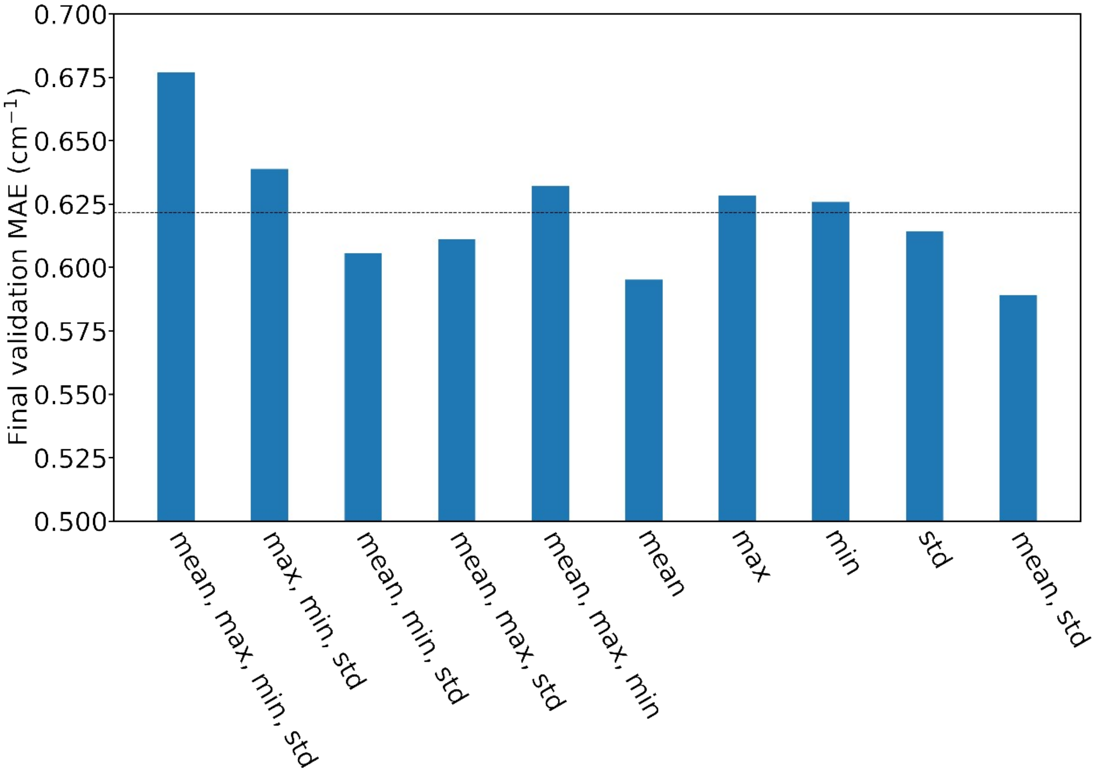

Figure S10: Comparison of the final validation MAE of the models trained with various combinations of aggregators in the PNA operator. In this computational experiment, the models were trained for 200 epochs with ~5000 protein graphs obtained using a threshold distance of 6 Å to define edges. The horizontal baseline denotes the mean value.