Supplementary Materials

An Interpretable Deep Geometric Learning Model to Predict the Effects of Mutations on Protein-Protein Interactions Using Large-scale Protein Language Model

Caiya Zhang[1], Yan Sun[1,2,3], Pingzhao Hu[1,2,3,4,5,6,*]

[1]Department of Computer Science, Western University, London, Ontario, Canada
[2]Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada
[3]Department of Biochemistry, Western University, London, Ontario, Canada
[4]Department of Oncology, Western University, London, Ontario, Canada
[5]Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada
[6]The Children's Health Research Institute, Lawson Health Research Institute, London, Ontario, Canada

1. Evaluation of the Similarity between Training and Test Sets

In constructing the graph input, we encoded the elemental information of protein atoms, aromatic bonds, the number of neighboring atoms (including hydrogen atoms), and implicit

valence to create feature vectors for each atom. These atomic feature vectors were then combined to generate a comprehensive feature matrix representing the protein. To align with our experimental setup, we pruned this matrix to focus exclusively on the mutated residue and its two adjacent residues for training and testing.

To analyze data similarity, we traced back the pruned graph matrix to specific positions in the protein sequence, corresponding to the mutated residue and its two adjacent residues. The sequence derived from the matrix was then entered to Protein BLAST (NCBI BLAST) for alignment, allowing us to evaluate sequence similarity between training and test datasets. We applied the same approach to the unpruned sequences. The pairwise similarity distributions for the training and test sets, for both the prunned sequences and the full sequences, are shown in **Figure S1** and **Figure S2**, resepctively.

Building on previous work, we adopted three similarity thresholds (1%, 5%, 10%), which measure the maximum sequence similarity of each test sequence to any sequence in the training set [1]. We then iteratively removed a test sample if its similarith to any training sample exceeded the given threshold. The remaining dataset sizes are summarized in **Table S1**, while **Table S2** presents the impact of similarity-based filtering on the prediction performance. As shown in **Table S2**, the removal of sequences with varying levels of similarity to the training set affected prediction performance. Although eliminating more similar sequences from the test set resulted in a slight decrease in predictive accuracy, the Pearson correlation coefficient (Rp) showed only a modest decline compared to its performance on the full dataset. This suggests that, even with the gradual removal of data, the model's performance remains comparable to, or even better than, the performance of the baseline methods observed with the complete dataset (**Table 2** in the main manuscript).

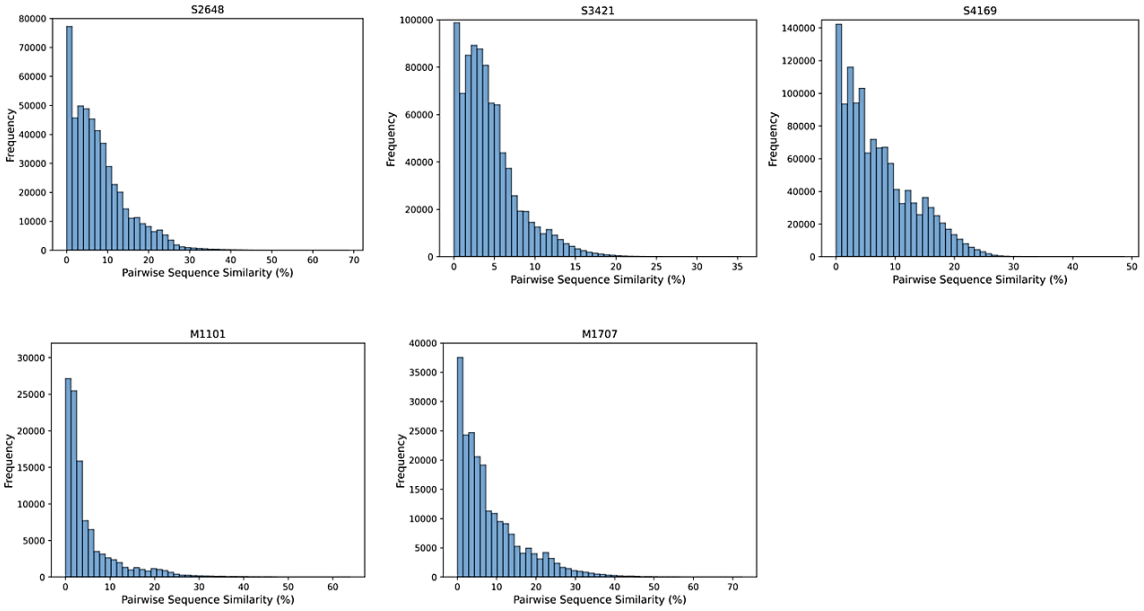

Figure S1. Pairwise Sequence Similarity Distribution with Pruned Sequence.

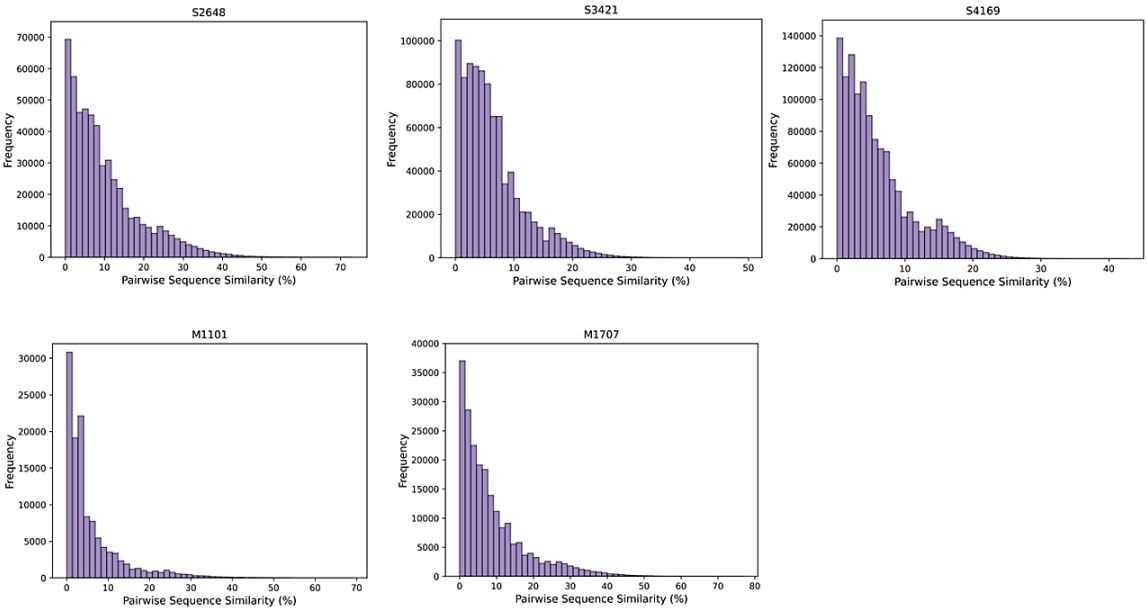

Figure S2. Pairwise Sequence Similarity Distribution with Whole Sequence.

Table S1. Number of Samples in Each Test Set after the Removal of Partial Test Set based on Its Similarity to the Training Set

	S2648	S3421	S4169	M1101	M1707
All Sequences[1]	529	684	833	220	341
Sequence Similarity ≤ 10%[2]	512	660	797	217	332
Sequence Similarity ≤ 5% [3]	456	632	774	197	311
Sequence Similarity ≤ 1% [4]	398	627	771	186	307

Table S2. Assessment of GES_PPI's Average R_P Score Following the Removal of Partial Test Set based on Its Similarity to the Training Set

	S2648	S3421	S4169	M1101	M1707
All Sequences[1]	0.6491	0.7166	0.6892	0.5679	0.7538
Sequence Similarity ≤ 10%[2]	0.6485	0.7007	0.6775	0.5647	0.7519
Sequence Similarity ≤ 5% [3]	0.6338	0.6986	0.6682	0.5591	0.7468
Sequence Similarity ≤ 1% [4]	0.6297	0.6954	0.6653	0.5606	0.7398

[1]All Sequences: test with all sequences in the test set in each dataset.
[2]Sequence Similarity ≤ 10%: test with sequences in the test set having at most 10% similarity to the training set.
[3]Sequence Similarity ≤ 5%: test with sequences in the test set having at most 5% similarity to the training set.
[4]Sequence Similarity ≤ 1%: test with sequences in the test set having at most 1% similarity to the training set.

References

[1] T. Sanavia *et al.*, "Limitations and challenges in protein stability prediction upon genome variations: Towards future applications in Precision Medicine," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1968–1979, 2020. doi:10.1016/j.csbj.2020.07.011