Navigating Trustworthiness of Deep Learning in ΔΔG prediction : Addressing Data Bias, Model Evaluation, and Interpretation

Ruochi Zhang * [1 2] Ningning Chen * [1 3] Fengfeng Zhou[4] Xin Gao[1 5]

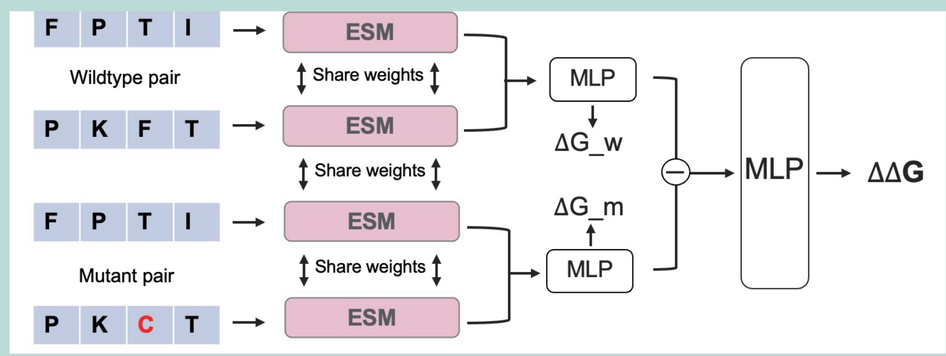
*Equal contribution [1]Syneron Technology, Guangzhou, China [2]School of Artificial Intelligence, Jilin University, Changchun, China [3]ETH Zurich, Department of Biosystems Science and Engineering, Basel 4056, Switzerland [4]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China [5]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center

1 Introduction

- Understanding the change of binding affinity (ΔΔG) caused by mutations in protein-protein interaction(PPI) is important in protein engineering & drug design.
- The trustworthiness of Deep Learning is a concern due to the limited and biased experimental data in reality.
- A comprehensive guideline is needed to assess reliability of models which consists of data analysis, model evaluation and interpretation.

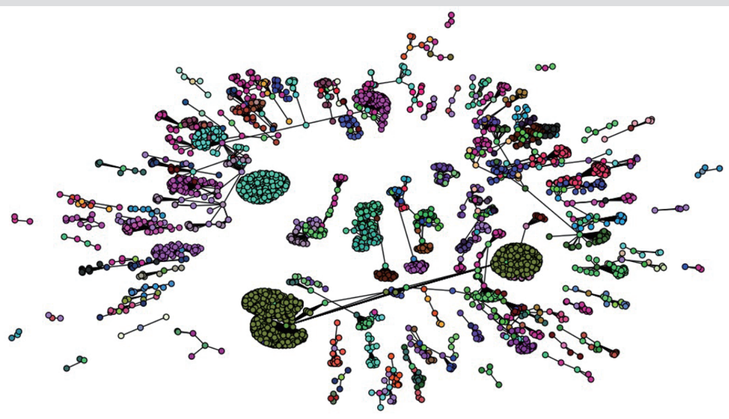
Set up & Model architecture

Models:
- MuPIPR[1]
- ProtMut_p (right figure)
- ProtMut_c (cross attantion)

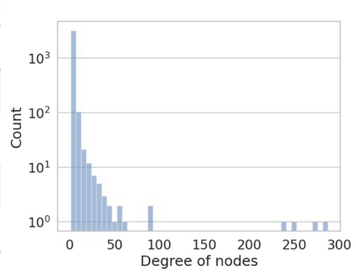


2 Biased dataset

SKEMPI v2 dataset visualization



SKEMPI v2 3562 Statistics	
Proteins (nodes)	3302
Interactions (edges)	3708
Protein family	138
Community	138



– The dataset is highly imbalanced that some nodes are aggregated to form large clusters while others are segregated.

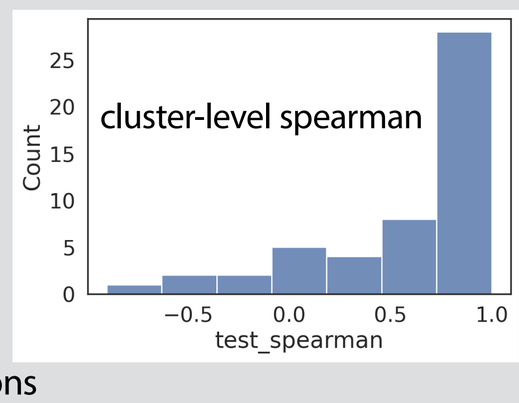
3 Model evaluation methods

Dataset Split Methods

– Random split vs Cluster-level split

Metrics determination

– Align with the real world needs this task: Spearman correlation
– Aggregate metric -> granular evaluations this task: Metrics in each protein family



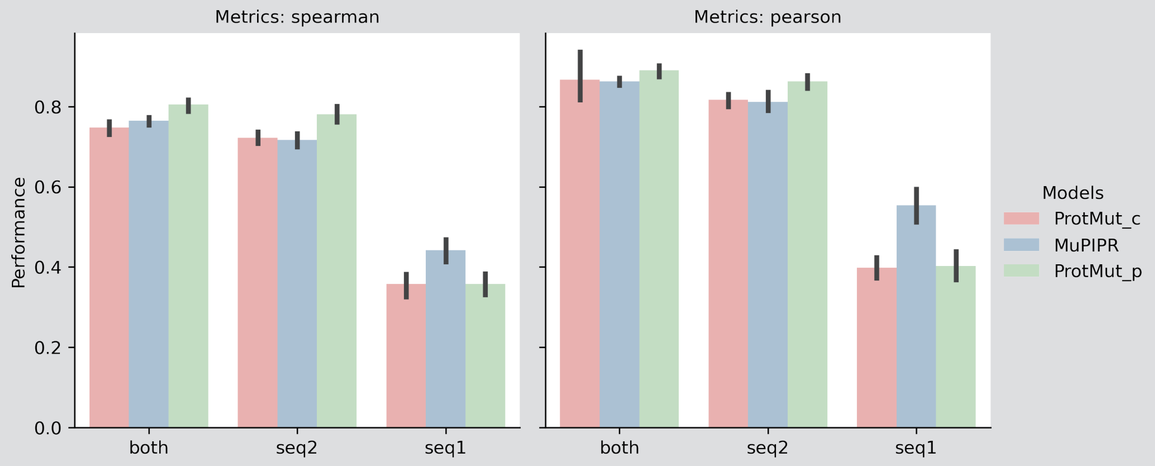
Held-out sanity check

– Dataset features regression test
– Held out one sequence in the protein sequence pair test

Model performance on different dataset splitting strategy.

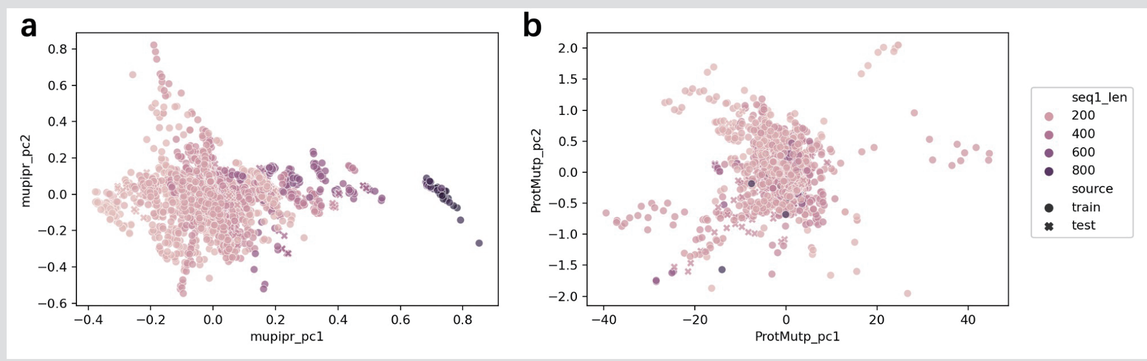
4 Model interpretation

Model performance on held-out inputs sanity check



Intriguingly, when only seq2 (mutated sequences) information is supplied, the performance remains largely unaltered.

Embeddings of models in the cluster-level split



PC 1 of MuPIPR(a) shares a correlation with sequence length, implying it learns dataset feature like sequence length instead the underlying protein-protein interaction mechanism

5 Conclusion & Discussion

– Deep learning models might be learning unintentional biases present in the dataset rather than the actual biological relationships
– Biased data in biology is a significant issue that requires careful attention and resolution
– The effectiveness of complex model architecture is limited when the dataset is insufficient. (cross-attention doesn't improve the performance)
– Data bias, model evaluation and model interpretation should be paid more attention in bioscience field

Reference

1. Zhou, G., Chen, M., Ju, C. J. T., Wang, Z., Jiang, J.Y., and Wang, W. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. NAR Genomics and Bioinformatics, 2(2):lqaa015, June 2020.

Contact:
zrc720@gmail.com
ningning.chen@bsse.ethz.ch