A Physics-informed Inverse Folding Model for Protein Free Energy Prediction

Zhaoyang Li
Department of Bioengineering
Stanford University
Stanford, CA 94305
`zhaoyangli@stanford.edu`

Pengwei Sun
Department of Radiology
Stanford University
Stanford, CA 94305
`pengwei@stanford.edu`

1 Training

We construct the pre-training dataset from the Protein Data Bank (PDB). We discard PDB entries with a resolution worse than 3.5 Å and repair the remaining structures. Multimeric complexes are split into individual chains. Following the original ProteinMPNN data preprocessing pipeline, we further cluster sequences by sequence identity.

We adapt the ProteinMPNN implementation in PyG and modify the dataloader so that, in each epoch, PDB files are sampled randomly at the level of clusters, without sampling from the same cluster more than once. The model is trained on a single H100 GPU. The training and validation curves are shown in Figure 1.

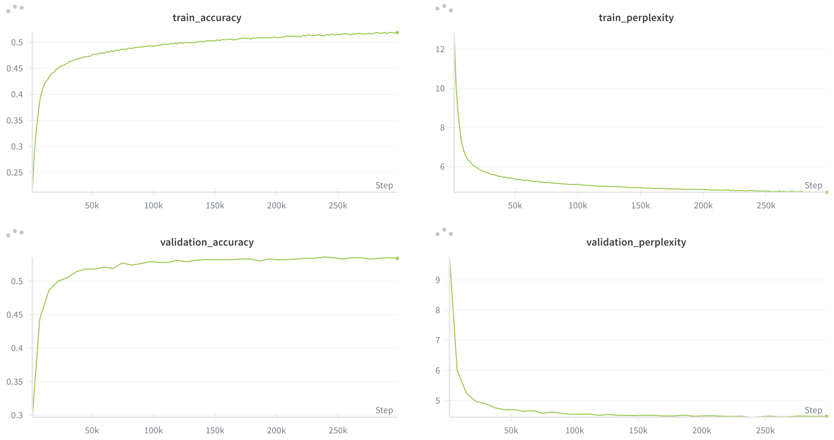

Figure 1: Training and validation curves for the pre-training phase.

2 Finetuning

We fine-tune the model on the MegaScale stability dataset. For each mutation, we use the difference in $\Delta\Delta G$ between the wild-type and mutant sequence as the target value. Fine-tuning is performed on a single H100 GPU. The training and validation curves are shown in Figure 2.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

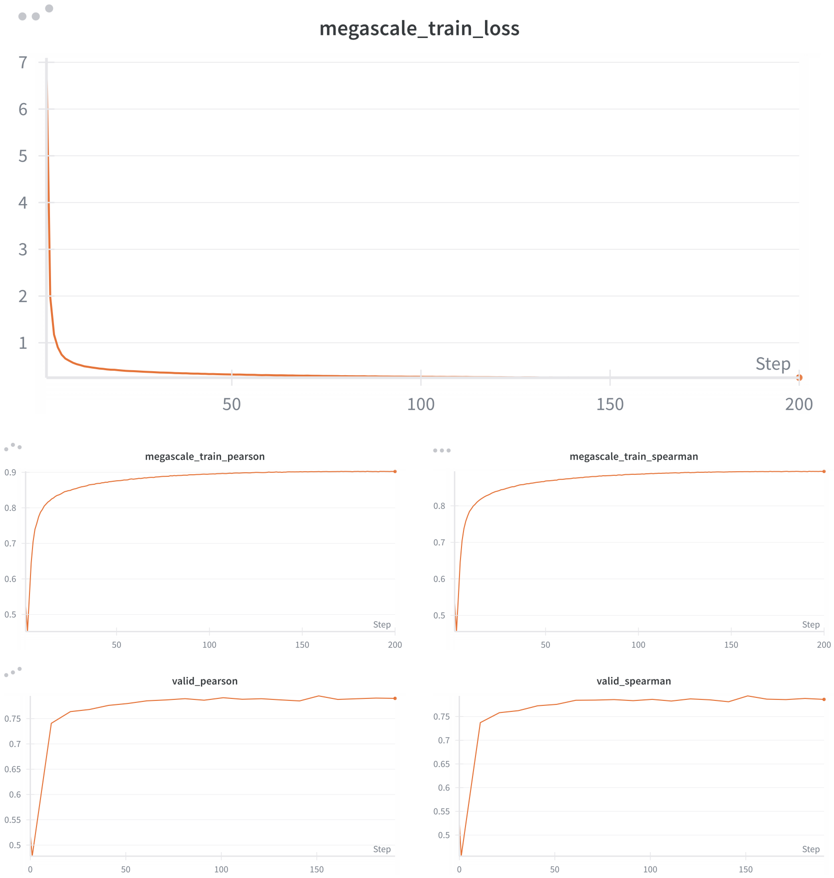

Figure 2: Training and validation curves for the finetuning phase.