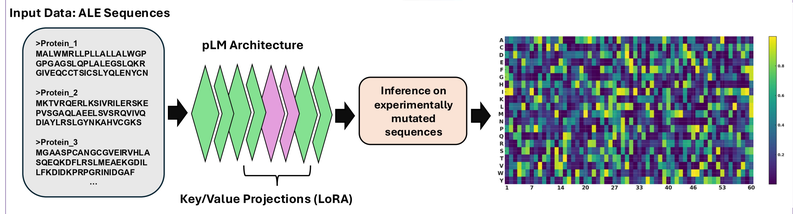
Learning Adaptive Lab Evolved Mutational Landscapes: Leveraging LoRA on a Protein Language Model

Authors: Silba Dowell & Shivanand Sheshappanavar
University of Wyoming

Introduction:

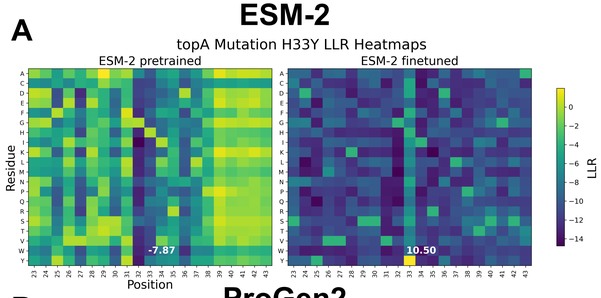
- ALE sequences are evolutionarily divergent from common protein sequences due to selection pressures.
- While beneficial in protein engineering, ALE is costly in resources and may not find beneficial mutations efficiently.
- Protein language models (pLMs) have been used for mutation effects.
- Using fine-tuning strategies, the goal is to implement a pipeline for a pLM to predict mutations of ALE data.

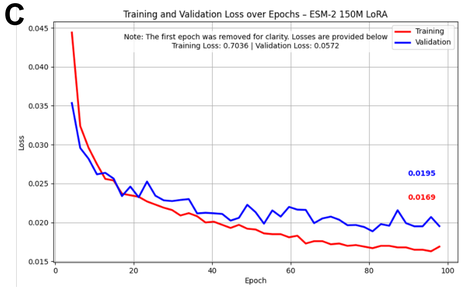
Methodology:



- Collect ALEdb sequences and preprocess for frequent mutations across trials. Synthetically mutate homologs to augment dataset.
- Inject LoRA modules into pLM architecture (ESM-2 and ProGen2) and fine-tune on sequence data.
- Infer mutation effects and assess through log-likelihood ratio between the mutant and wild-type sequence per residue per position.

Experimental Results:





E

Metric	ESM-2 Pre-trained	ESM-2 Fine-tuned	ProGen2 Pre-trained	ProGen2 Fine-tuned
Perplexity	6.050	1.021	3.717	1.025
Accuracy	0.4356	0.996	0.602	0.995

A) Predicted mutation effect heatmap from the ESM-2 LoRA model. **B)** Predicted mutation effect heatmap from the ProGen2 LoRA model. **C)** Training and validation losses for ESM-2. **D)** Training and validation losses for ProGen2. **E)** Perplexity and token accuracy for ESM-2 and ProGen2 models.

Conclusion:

- Fine-tuning ESM-2 and ProGen2 using LoRA was effective.
- ESM-2 performs marginally better than ProGen2 in accuracy but less in perplexity.

Contributions:

- First to implement pLMs on ALE dataset.
- Demonstrated effective mutation effect prediction using SOTA fine-tuning approaches such as LoRA.
- Expanded the pLM's protein landscape using evolutionary divergent sequences.

Future Works:

- Include a protein structure module for mutation effect enhancement.
- Implement a distillation pipeline for computational efficiency.
- Use more datasets of indels.

References:

[1] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, Mar. 2023.
[2] E. Nijkamp, J. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022.
[3] P. V. Phaneuf, D. Gosting, B. O. Palsson, and A. M. Feist. ALEdb 1.0: A database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Research, 47(Database issue):D1164–D1171, Jan. 2019.