ESMnrg: Transformer-Based Decoding of Changes in Protein Stability due to Mutations

Angad Sandhu[1], Benjamin Chang[1], Soumya Prakash Behera[1], Claire Jung[1]

[1]*Johns Hopkins University*

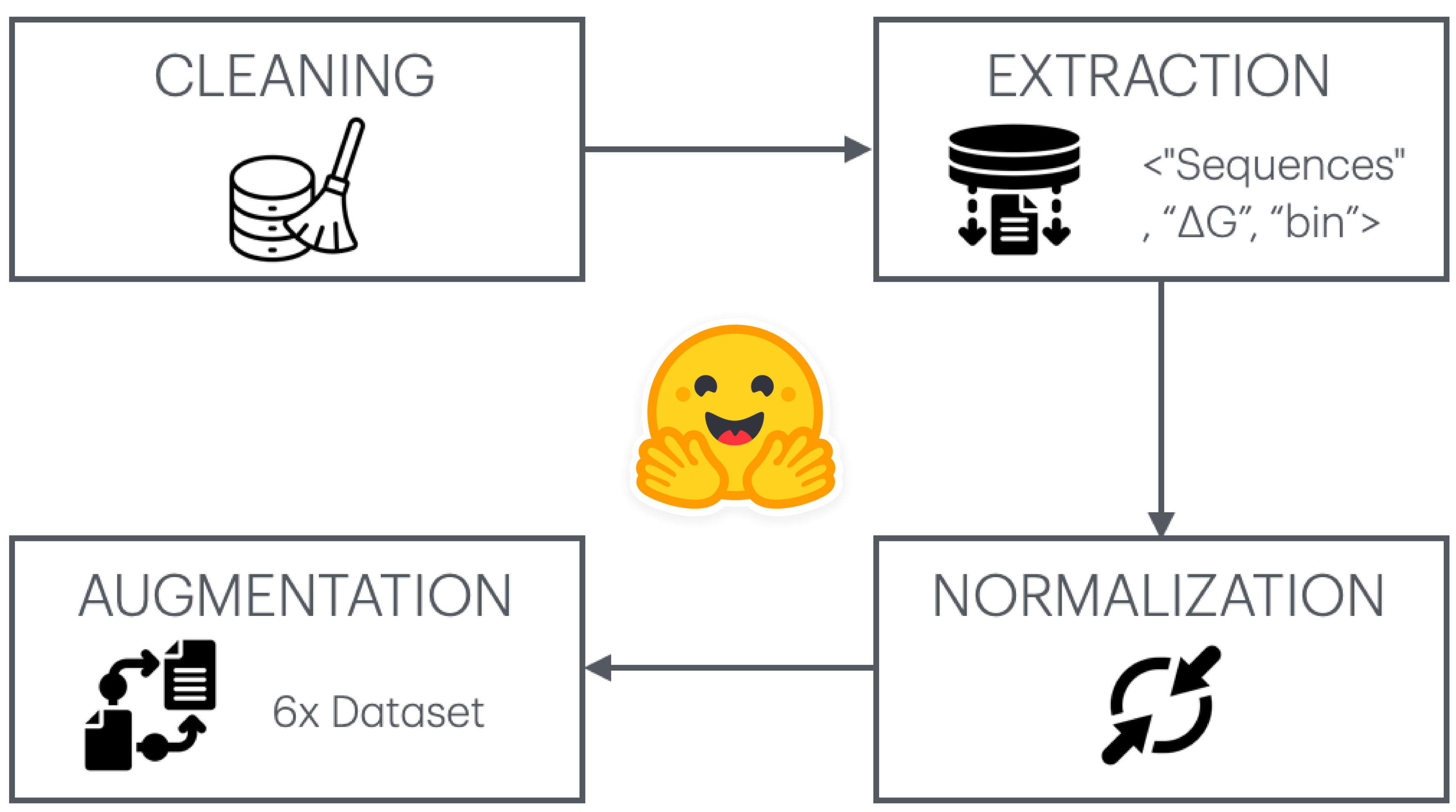
JOHNS HOPKINS UNIVERSITY

GOAL

Develop and validate ESMnrg, a transformer-based model that integrates ESM2 structural embeddings with protein stability data to predict mutation effects

DATA PREPARATION

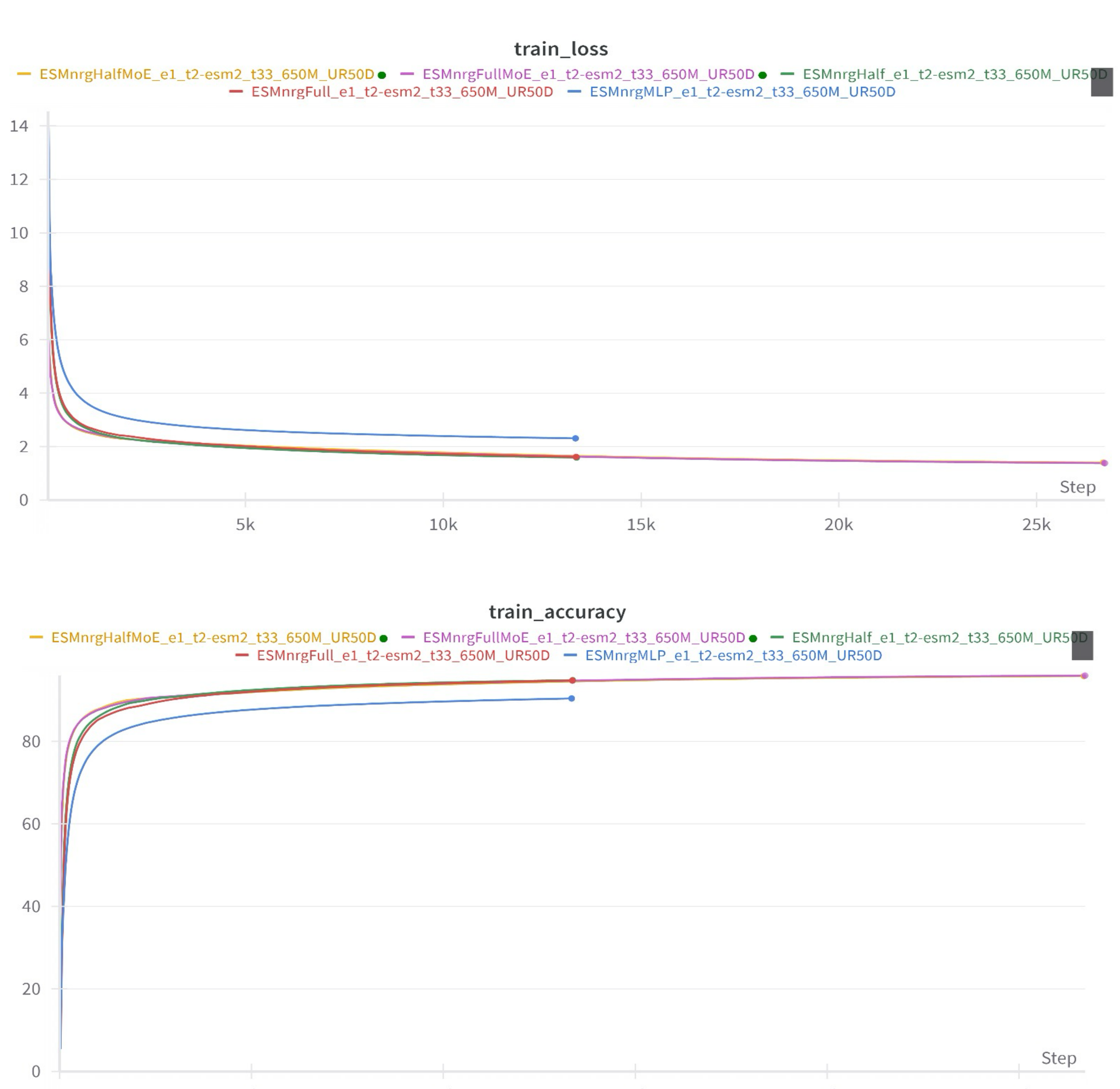
ESMnrg: Protein Stability Prediction Dataset



EXPERIMENTAL RESULTS

Models	Train Loss	Train Accuracy (%)	Validation Loss	Validation Accuracy (%)	Test Loss	Test Accuracy (%)
1	1.610	94.72	1.326	96.34	1.330	96.32
2	1.587	94.81	1.292	96.57	1.291	96.57
3	2.305	90.39	2.535	89.52	2.535	89.51
4	1.377	**95.94**	1.203	**97.08**	1.202	**97.08**
5	1.398	95.82	1.179	97.04	1.177	97.06

EXPERIMENTAL RESULTS



Train loss and train accuracy trends for 5 different ESMnrg models across training steps [top and down, respectively]. Shows each model version, from full configurations to specialized MoE and MLP versions, performs over time

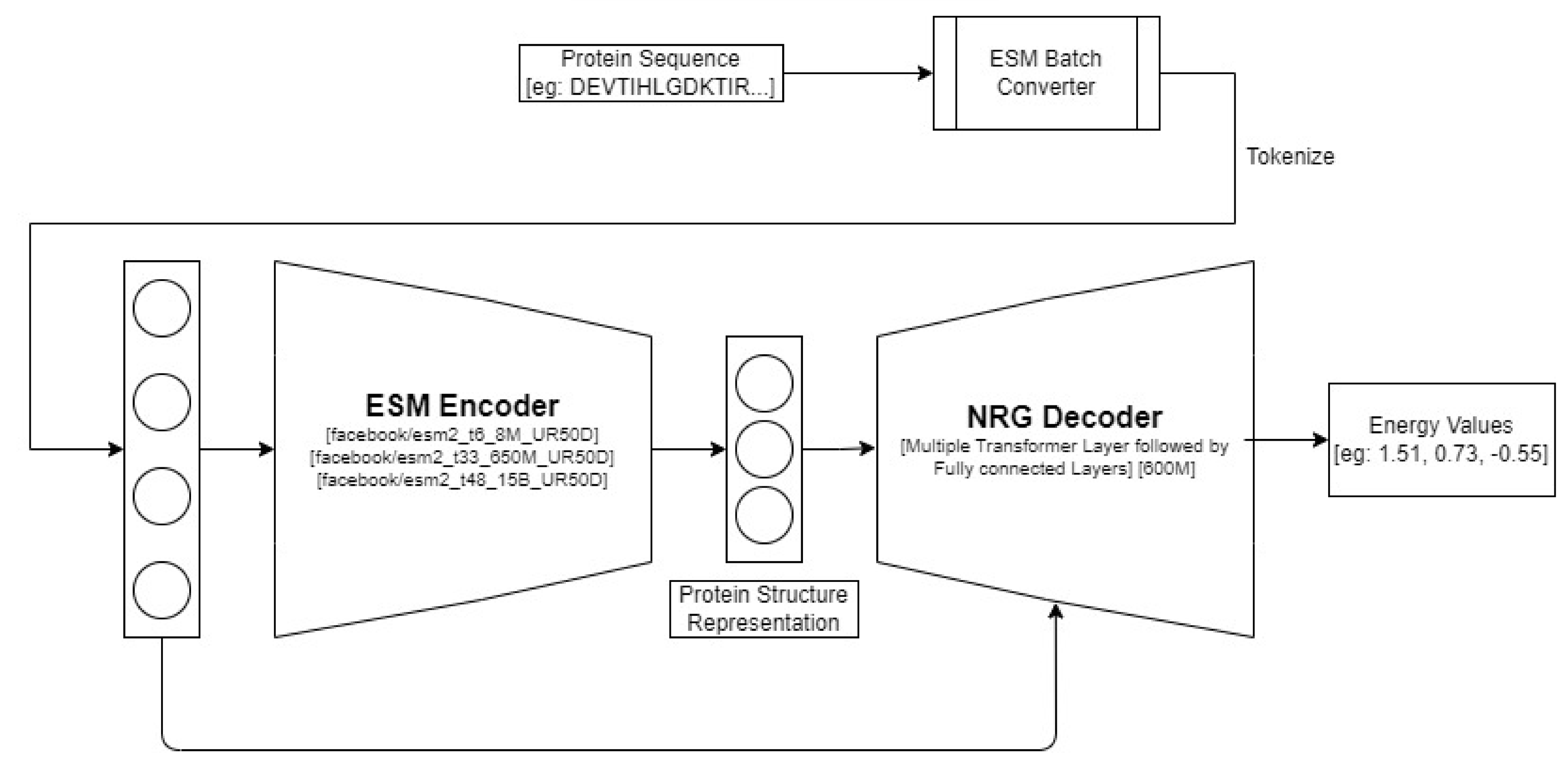
Reference

[1] Rives, A., et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." Proceedings of the National Academy of Sciences 118.15 (2021): e2016239118.

MOTIVATION

- ❑ Predicting protein stability changes due to mutations is crucial for understanding diseases, guiding therapies, and advancing protein engineering.
- ❑ Traditional methods rely on static structural data or evolutionary signals but fail to account for the dynamic nature of protein sequences.
- ❑ Rives et al. demonstrated how transformer-based models pre-trained on protein sequences can capture contextual embeddings of amino acids [1].

ARCHITECTURE



MODELS

1. ESMnrgFull_e1_t2-esm2_t33_650M_UR50D
2. ESMnrgHalf_e1_t2-esm2_t33_650M_UR50D
3. ESMnrgMLP_e1_t2-esm2_t33_650M_UR50D
4. ESMnrgFullMoE_e1_t2-esm2_t33_650M_UR50D
5. ESMnrgHalfMoE_e1_t2-esm2_t33_650M_UR50

CONCLUSION

- ❑ Empirical evidence from model training demonstrates varied performance across configurations, aided by regularization techniques such as batch normalization, dropout, and residual connections
- ❑ Performance of MoE-enhanced models outperform standard MLP configurations, underscoring their capability to integrate intricate protein sequence dynamics more effectively

FUTURE DIRECTIONS

- ❑ Incorporate additional data modalities (e.g., protein-protein interactions, evolutionary conservation scores)
- ❑ Expand ESMnrg to predict other protein properties (e.g., binding affinity, solubility, enzymatic activity)

LIMITATIONS

- ❑ Dependence on the availability and quality of protein stability data
- ❑ High computational resource requirements for training and deployment