ESMfluc: Predicting flexible regions in a protein using language models

Ngoc K. Tran[1], Mandy Korpusik[1], Zara Alavi[2]

[1]Department of Computer Science, Loyola Marymount University, Los Angeles, USA
[2]Department of Physics, Loyola Marymount University, Los Angeles, USA

Introduction

Proteins are dynamic entities whose functions are determined by their structural flexibility, encoded directly in their sequences. Although machine learning has successfully mapped sequences to structures, capturing this flexibility remains challenging. In this study, we address this gap by training a biLSTM model on molecular dynamics simulation data, utilizing embeddings from the ESM model, to predict protein dynamics. Our model enables the user to obtain flexibility data without the need for computationally expensive simulations.

Methods

Training data is obtained from ATLAS library[1], which contains molecular dynamics simulation results for 1390 protein chains. Local fluctuations data are reported in the library, including an entropy-based measure of flexibility N_{eq} which is defined per residue as:

$$N_{eq} = 2^H \quad \text{where} \quad H = -\sum_{i=1}^{16} p_i log_2 p_i$$

and p_i is the probability of the residue visiting protein-block i during the simulation time. Protein-blocks (PBs) are structural prototypes.[2] In principle, any conformation of any amino acid could be represented by one of the 16 available PBs. A N_{eq} of 1.0 means the amino acid is very rigid and a N_{eq} of 16 means it visits all possible PBs equally frequently and thus is highly flexible. We chose N_{eq} as the target for our classification task. Fig. 1 shows the distribution of N_{eq} values in the training set.

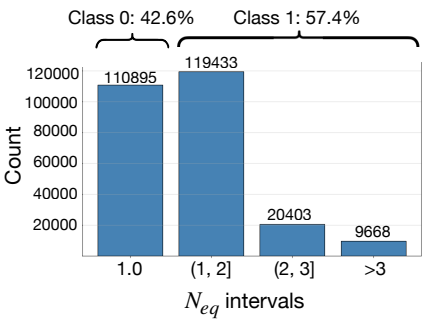

Fig. 1. Histogram of N_{eq} values in training data.

Input sequences were tokenized using ESM tokenizer[3], embeddings were extracted and fed into a biLSTM[4] model with one attention head followed by a FC layer for classification. Fig. 2 shows the architecture of the best model.

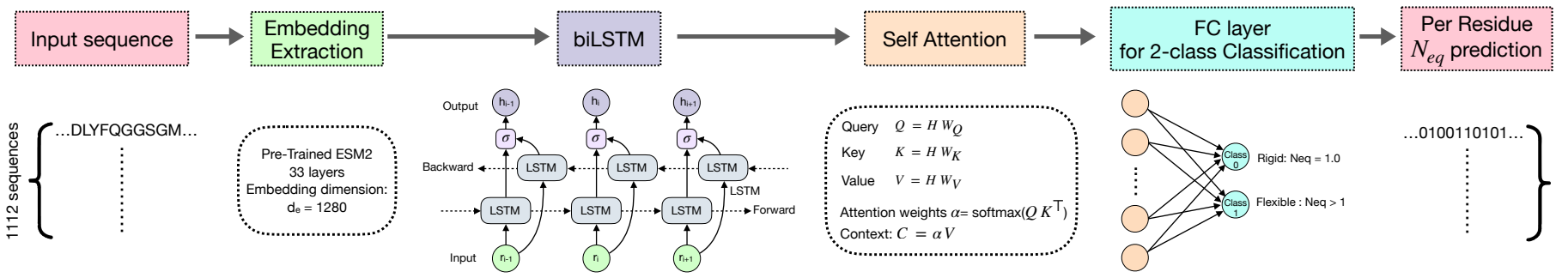

Fig. 2. Best model architecture

To address class imbalance we used Focal loss: $\mathscr{L}_{\text{Focal}} = -(1-p_t)^\gamma \log(p_t)$ where $log(p_t)$ is the log-probability from cross-entropy and the focusing parameter γ was set to 2. Other hyperparameters are shown in Table. 1.

Optimizer	LR Scheduler	Epochs	Dropout	Layers Frozen	Precision
AdamW	Reduce on Plateau	80 (early stopping patience = 3)	0.3	0-4	FP16 (AMP)

Table 1. Model hyperparameters

Results & Discussion

Model Performance: Table 2 shows the effect of ESM checkpoint sizes and model architecture on training performance. The optimal performance is achieved using ESM checkpoint t33 with 650 M parameters as features to train a BiLSTM + Attention model. Fig. 3 shows the distribution of N_{eq} values in the test set along with misclassified instances for each N_{eq} interval using the best model.

Architecture	ESM Checkpoint	Precision (%)		Recall (%)		Macro Avg F1(%)
		Class 0	Class 1	Class 0	Class 1	
BiLSTM + Attention	t33 *650 M*	**81.33**	**81.73**	**73.77**	**87.40**	**81.44**
BiLSTM + Attention	t30 *150 M*	77.73	80.59	72.66	84.51	78.81
BiLSTM + Attention	t12 *35 M*	78.02	77.06	65.50	86.26	76.31
BiLSTM + Attention	t6 *8 M*	77.49	75.99	63.37	86.30	75.27
Transformer[5]	t33 *650 M*	80.61	81.38	73.31	86.87	80.41
Logistic Regression	t33 *650 M*	78.30	79.80	71.00	85.40	78.50

Table 2. Model performance for various architectures and ESM checkpoints.

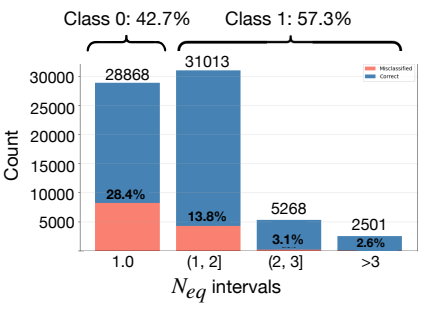

Fig. 3. Histogram of N_{eq} values in test set with misclassifications in pink.

Attention Mechanism: The attention weights matrix α reveals residues strongly influencing each other, including those far apart in the sequence. Fig. 4 displays the attention heat-map for protein PDB: 1LRI, highlighting key residue interactions. Fig. 5 shows the corresponding crystal structure.

To quantify attention homophily, we look at "what proportion of the query's attention is allocated to the same type of Secondary Structure (SS) or N_{eq} class as itself?" Table 3 shows (a): on average, how a coil/helix/β-strand distributes its attention across the three secondary structure classes and (b): on average, how a N_{eq} class of 1/0 distributes its attention across the two N_{eq} classes. The model has internalized a modular view of the protein: residues preferentially consult partners that share both structural state and dynamic behavior, even when those partners are distant along the chain.

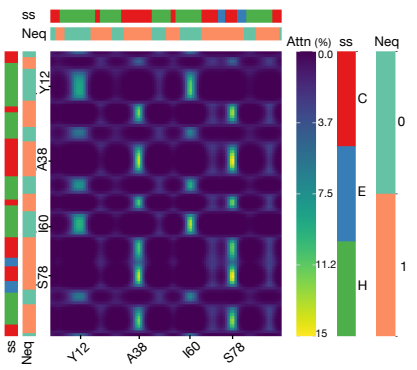

Fig. 4. Residue level attention mechanism in a test set sample (PDB 1LRI). Annotation bars: top / left are secondary-structure prediction (red = C, green = H, blue = E) obtained from NetSurfP[6], bottom / right are predicted Neq class (orange = 1, teal = 0)

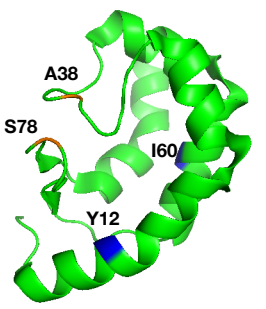

Fig. 6.Crystal Structure of PDB 1LRI.

a)

	A_C	A_H	A_E
$\bar{a}^C$	**0.78**	0.13	0.09
$\bar{a}^H$	0.25	**0.68**	0.07
$\bar{a}^E$	**0.52**	0.25	0.23

b)

	B_0	B_1
$\bar{b}_0$	**0.79**	0.21
$\bar{b}_1$	0.08	**0.92**

Table 3. Attention distribution, averaged over the entire test set (a) in ss-space and (b) in neq-space.

Future Work:

- Multi-class Classification: Identification of residues with high N_{eq} values (e.g., N_{eq}>3). Such residues may play critical roles in conformational dynamics and function.
- Variant Effect: Studying how mutations alter attention patterns to predict functional changes.
- Expansion of Training Data: Performing molecular dynamics simulations, focusing particularly on building a library of strain measurements induced by ligand binding. Training models on such data would identify remote allosteric modulators and residues critical for conformational transitions involved in ligand binding.

References:
1. "ATLAS: protein flexibility description from atomistic molecular dynamics simulations" by Y. V. Meersche et al., Nucleic Acids Research 2024
2. "Evolutionary-scale prediction of atomic-level protein structure with a language model" by Z Lin et al., Science 2023
3. "Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks" by de Brevern et al., Proteins 2000
4. "Long Short-Term Memory" by S. Hochreiter & J. Schmidhube, Neural Computation, 1997
5. "Attention Is All You Need" by A. Vaswani et al., NeurIPS 2017
6. "NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning" by M. H. Høie et al., Nucleic Acids Research 2022