Natural Language Processing Applied to Bioinformatics

Y. Sarahi García González[*]

Centro de Investigación en Matemáticas CIMAT[**]

July, 2024

In this work, we explore applying large language models for predicting protein solubility and fluorescence using only protein sequences datasets from the TorchDrug library. For solubility prediction, we employed a modified version of the PeptideBERT architecture. This model, configured with 12 hidden layers and 12 attention heads, was adjusted to handle sequences of up to 512 amino acids, achieving 67.98 %accuracy. In contrast, the ProtBERT model used in PEER was trained with longer sequences and a more robust configuration, achieving slightly higher accuracy. In the fluorescence prediction task, we used a similar model based on PeptideBERT but modified for regression. The performance evaluation showed a test Spearman's rank correlation coefficient of 0.23, indicating a low performance. This result can be attributed to the dataset's nature, the model's inadequacy for this specific task, and the limited number of training epochs.

I. INTRODUCTION

Proteins are fundamental components of all living cells and play a critical role in many biological processes. They act as enzymes and structural elements essential for the structure, function, and regulation of the body's tissues and organs [1] . They are also key players in the immune system. Antibodies are proteins that recognize and neutralize pathogens such as bacteria and viruses.[2]

Every protein is made up of one or more chains of amino acids (called polypeptides) that fold into complex ensembles of three-dimensional structures that determine its functions [1]. The DNA sequence of the protein-encoding gene determines these chains or sequences: mRNA is read and translated into the string of amino acid chains that make up the synthesized protein (this process is called translation). [3]

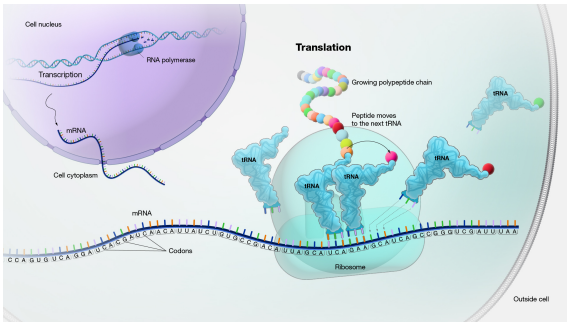

Figura 1. Translation is the process through which information encoded in messenger RNA directs the addition of amino acids during protein synthesis. Translation occurs on ribosomes in the cell cytoplasm. [3]

Protein modeling is not just an academic pursuit. It is a crucial tool for understanding biological processes at the molecular level, with significant implications for disease management and the development of new drugs. This task is an essential field in Bioinformatics that focuses on understanding protein structure, function, and interactions.

If we know a protein's structure, we can infer how it interacts with other molecules and how it can perform specific tasks. The problem of predicting protein structure from its amino acid sequence is the polypeptide's flexibility, which can fold into many different shapes. Also, small perturbations can change the final shape. Another ways to determine a protein's structure requires advanced experimental approaches, such as X-ray crystallography, Nuclear Magnetic Resonance Spectroscopy, and cryo-electron microscopy. However, these methods are not only expensive but also limited in their applicability to certain proteins.

On the other hand, new sequencing technologies have led to an exponential growth of protein sequence databases, but not on annotated subsets, so there is a growing gap between those two. [4] However, because protein sequences can be represented as text, with each of the twenty amino acids represented by one letter, large language models have enabled a new way of protein modeling. Advances in NLP techniques have shown that self-supervised learning is beneficial for extracting information from unlabeled protein sequences. [5][4]

Numerous NLP models have been adapted to extract biological information from extensive protein sequence datasets. In this work, we utilize an architecture proposed in PeptideBert [6], which employs the ProtBERT [5] pretrained transformer model with 12 attention heads and 12 hidden layers. Then, they fine-tune the pre-trained model for the three downstream tasks: prediction of hemolysis, non-fouling, and solubility. Here, we focus on one of these three: the prediction of solubility (a classification task)

[*] yesenia.garcia@cimat.mx

[**] PLN a cargo de Dr. J. Fernando Sánchez Vega

and an additional one, the prediction of fluorescence (a regression task)

Predicting protein solubility and fluorescence are critical tasks in bioinformatics and molecular biology. Protein solubility is a fundamental property affecting its functionality and stability, influencing various applications from drug development to industrial biotechnology. Insoluble proteins can lead to aggregation, reducing their effectiveness and potentially causing diseases. Accurate prediction of protein solubility can aid in designing more effective therapeutic proteins and in optimizing bioprocesses for industrial enzyme production. [7][6]

Fluorescence prediction, on the other hand, is pivotal for understanding protein interactions and dynamics. Fluorescent proteins are extensively used as molecular and cellular biology markers, allowing scientists to visualize and track biological processes in real-time. Predicting fluorescence properties helps in engineering proteins with enhanced or novel fluorescence characteristics, which can improve the sensitivity and specificity of bioimaging techniques and biosensors. These advancements have significant implications for medical diagnostics and environmental monitoring.[8][9]

II. BACKGROUND

Notions and terms

- **Aminoacid:** Building block of proteins is an alpha amino acid which contains a basic amino group, an acidic carboxyl group, and a hydrogen or organic side chain attached to the central carbon atom. There are 20 different alpha amino acids commonly found in nature that can covalently link with each other to form short peptides or longer proteins.
- **Sequence/primary structure:** The linear sequence of amino acids in a peptide or protein (Sanger, 1952). Any sequence of polypeptides is reported starting from the single amine (N-terminus) end to carboxylic acid (C-terminus).
- **RNA:** A polymeric molecule essential in various biological roles, most often single-stranded.
- **Secondary structure (SS)**: The 3D form of local segments of proteins. The two most common secondary structural elements are α-helix (H) and β-strand (E);
- **Tertiary structure**: The 3D shape of a protein.
- **Quaternary structure**: The 3D arrangement of the subunits in a multisubunit protein.
- **Protein function prediction:** A Task that uses techniques to assign biological or biochemical roles to proteins. Gene Ontology (GO) annotations classify functions into three main categories of molecular function, biological process, and cellular component
- **Fluorescence**: refers to the emission of light by fluorescent proteins when they absorb light at one wavelength and emit light at a different, typically longer, wavelength. This property is utilized extensively in molecular and cellular biology to visualize and track biological processes. Fluorescent proteins, such as GFP (Green Fluorescent Protein), are genetically encoded and can be fused to other proteins of interest, allowing researchers to study protein localization, interactions, and dynamics in living cells.
- **Solubility**: refers to the ability of a protein to remain dissolved in a solution, maintaining a stable and functional form without precipitating.

Architectures

- **ProtTrans** is a suite of protein language models that leverages self-supervised learning techniques from Natural Language Processing (NLP) to process protein sequences. Utilizing architectures such as Transformer-XL, BERT, and XLNet, these models are trained on massive datasets comprising billions of amino acid sequences.
- **ProtBERT-BFD** is a pre-trained model on protein sequences using a masked language modeling (MLM) objective. This model was introduced in the ProtTrans project and is trained on uppercase amino acids ProtBERT-BFD is based on the BERT architecture and was trained on the Big Fantastic Database (BFD), which contains over 2.1 billion protein sequences. The model was pre-trained in a self-supervised fashion on raw protein sequences. Each protein sequence is treated as a complete document, and during training, 15 % of the amino acids in the input are randomly masked, following the original BERT training methodology. The training process was conducted on NVIDIA V100 GPUs and required several weeks to complete due to the massive size of the dataset and the complexity of the model architecture. ProtBERT-BFD captures important biophysical properties governing protein shape, making it a powerful tool for various bioinformatics tasks.
- **PeptideBERT** is a language model designed to predict hemolysis, solubility, and non-fouling characteristics in peptides. PeptideBERT utilizes the ProtBERT pretrained transformer model with 12 attention heads and 12 hidden layers.

III. RELATED WORK AND STATE OF ART

The PEER (Protein sEquence undERstanding) benchmark is a comprehensive and multi-task benchmark designed to evaluate the performance of various deep learning methods on protein sequence understanding tasks. This includes a diverse set of tasks such as protein function prediction, protein localization prediction, protein structure prediction, protein-protein interaction prediction, and protein-ligand interaction prediction. The PEER benchmark uses designed training, validation, and test splits to ensure robust evaluation of model generalization capabilities.

The benchmark results for the fluorescence and solubility tasks were extracted from the PEER paper. The table below summarizes the three best results for each task, the performance of ProtBERT, and the state of the art (SOTA) as reported in the literature.

Fluorescence	
Model	**Performance (Spearman)**
Transformer	0.643 (0.005)
ProtBERT*	0.679 (0.001)
ESM-1b	0.679 (0.002)
CNN	0.682 (0.002)
SOTA	0.69
Solubility	
Model	**Performance (ACC)**
ProtBERT	68.15 (0.75)
Transformer	70.12 (0.31)
LSTM	70.18 (0.63)
ESBM-1b	70.23 (0.75)
SOTA	77.0

Cuadro I. Benchmark results on single-task learning for Fluorescence and Solubility tasks. We report mean (std) for each experiment. The top three performances and the state of the art (SOTA) from the literature.

IV. MODELS AND DATASET

We chose to use PEER for our experiments because it is one of the most cited benchmarks in the field and offers ease of implementation through the TorchDrug library. This accessibility and the comprehensive nature of PEER make it an ideal choice for evaluating and comparing different models on protein sequence analysis tasks. However, due to a lack of computational resources, we could not implement the PEER models, which use additional information extracted from the sequences and requires a lot of memory to process. The configuration for ProtBERT used in PEER is as follows:

- Attention Probs Dropout Prob: 0.0
- Hidden Act: GELU
- Hidden Dropout Prob: 0.0
- Hidden Size: 1024
- Initializer Range: 0.02
- Intermediate Size: 4096
- Max Position Embeddings: 40000
- Num Attention Heads: 16
- Num Hidden Layers: 30
- Type Vocab Size: 2
- Vocab Size: 30

And for protein function (as Fluorescence and Solubility), localization, and structure prediction tasks, applies a 2-layer MLP with a ReLU nonlinearity in between to perform the prediction.

The model was trained during 50 epochs for solubility and 100 for fluorescence.

Solubility

For this task we employed the architecture proposed by PeptideBERT. The PeptideBERT model utilized ProtBERT for this experiment, with 12 attention heads and 12 hidden layers followed by a single connected layer with 480 nodes. The output of the regression head is passed through a Sigmoid function to obtain the final binary prediction. The configuration for ProtBERT used in this work for this task is as follows:

- Vocab Size: 25
- Hidden Size: 480
- Num Hidden Layers: 12
- Num Attention Heads: 12
- Hidden Dropout Prob: 0.15
- Max Position Embeddings: 512
- Prediction Layer: Fully connected layer with 480 nodes followed by a sigmoid activation function is used for classification.

We trained the model during 8 epochs.

We used solubility dataset from the PEER benchmark from TorchDrug library. Originally, this dataset comes from the DeepSol [10] project. The dataset is structured to remove training sequences with more than 30 % sequence identity to any sequence in the test set, ensuring

the evaluation of the model's generalization ability across dissimilar protein sequences.

This dataset consists of 71419 protein sequences labeled and eexhibits a wide range of sequence lengths. However, as illustrated in the histograms above, the training, validation, and test sets show that significant concentration of sequences are below 500 amino acids. The bar plots on the right side of each row depict the distribution of binary labels (0 for insoluble and 1 for soluble), showing a balanced distribution across the different subsets of the dataset.

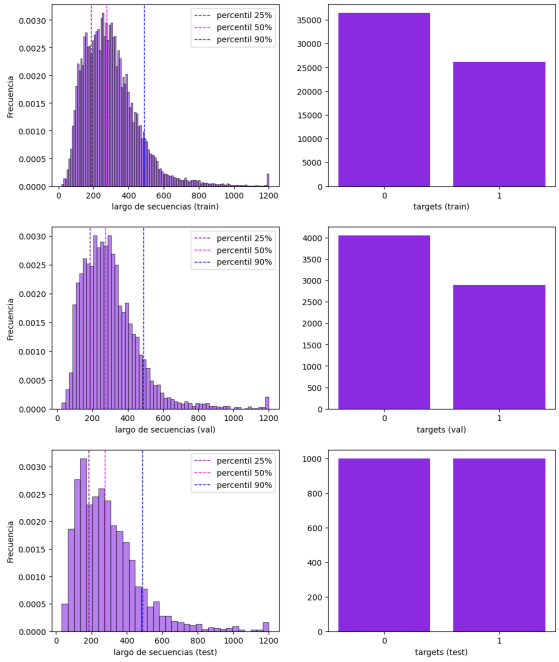

Figura 2. Sequence length distributions of the training, validation and test sets. The 25th, 50th, and 90th percentiles of sequence lengths for the training, validation, and test sets are marked by dashed lines in the histograms.

Figure 3 shows the distribution of the three subsets after taking all the protein sequences above 500 amino acids; as we can see, the binary labels still show balanced distributions.

Given that 90 % of the sequences in the dataset are below 500 amino acids in length, we decided to limit the training to this subset. This decision was driven by practical considerations related to computational resources. Training on the original dataset with the full range of sequence lengths would require significant memory, potentially leading to inefficiencies and difficulties in model training.

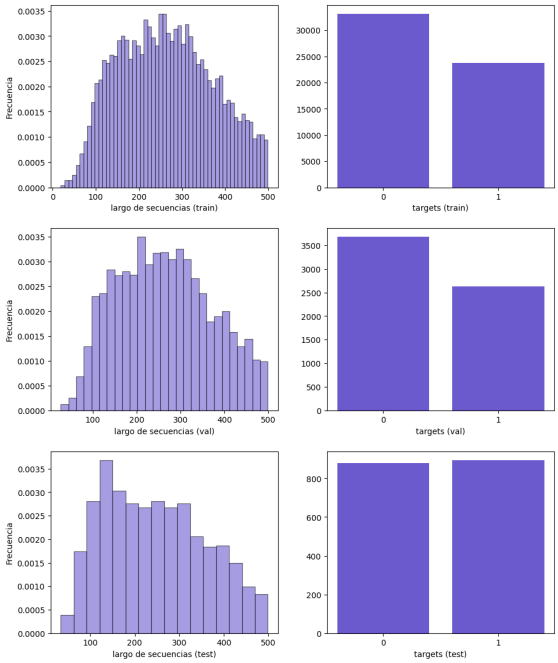

Figura 3. Sequence length distributions of the training, validation and test sets after taking all the sequences above 500 amino acids.

Fluorescence

For the fluorescence prediction task, we utilized a modified version of the PeptideBERT model. The following describes the architecture and configuration of the model used:

- Vocab Size: 25
- Hidden Size: 512
- Num Hidden Layers: 16
- Num Attention Heads: 16
- Hidden Dropout Prob: 0.15
- Max Position Embeddings: 256
- Regression: a fully connected layer that maintains the hidden size of 512, followed by a ReLU activation function to introduce non-linearity. To prevent overfitting, a dropout layer with a dropout probability of 0.15 is included, and the final linear layer outputs a single continuous value for the regression task.

For the fluorescence prediction task, we used the fluorescence dataset from the PEER benchmark, which is also accessible through the TorchDrug library. This dataset

is used to predict the fitness of green fluorescent protein mutants. The target is the logarithm of fluorescence intensity annotated by Sarkisyan et al. The dataset splits are from TAPE [4]

The histogram above shows the distribution of sequence lengths in the entire dataset. All sequences are of approximately the same length, around 236 to 237 amino acids. The second histogram illustrates the distribution of fluorescence values (labels) in the dataset. The values range from approximately 1.5 to 4.0 on the logarithmic scale.

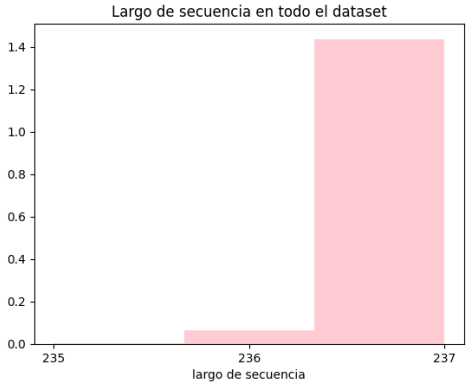

Figura 4. Sequence length distribution of the entire dataset.

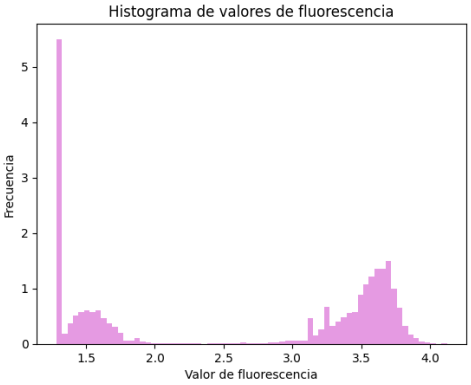

Figura 5. Label distribution of the entire dataset.

V. RESULTS

Solubility

In the solubility prediction task, we evaluated the performance of our model using the modified PeptideBERT architecture. The classification report for the test set shows the following metrics:

```
Test Accuracy: 67.98%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.63      0.86      0.73       879
         1.0       0.79      0.50      0.61       895

    accuracy                           0.68      1774
   macro avg       0.71      0.68      0.67      1774
weighted avg       0.71      0.68      0.67      1774
```

Figura 6. Classification report for the test set in solubility task

The ProtBERT model in the PEER benchmark achieves a slightly higher accuracy of 68.15 %, but it uses the complete dataset, including sequences up to 1200 amino acids, and is trained for 50 epochs.

Despite the differences in dataset size, training epochs, and model configuration, this model demonstrates competitive performance. The choice to limit sequence length to 512 amino acids was made to optimize memory usage. This decision, while reducing the overall dataset size, still allowed us to achieve a high level of accuracy close to that of the state-of-the-art models reported in the PEER benchmark.

Fluorescence

In the fluorescence prediction task, we evaluated the performance of our model using the modified PeptideBERT architecture. The test Spearman's rank correlation coefficient was 0.23, which is relatively low. This outcome can be attributed to several factors, including the dataset's inherent nature, the model's inadequacy for this specific task, and the limited number of training epochs (30).

```
100%|██████████| 1702/1702 [02:59<00:00,  9.46it/s]
Test Accuracy: 0.23889437267209304%
```

Figura 7. Test sSpearman's for the test set in fluorescence task

The best results reported in the PEER benchmark were achieved using a shallow CNN model, which obtained a Spearman's ρ of 0.69. The PEER benchmark also includes models trained for a more extended period, allowing them to better capture the complex relationships within the data.

VI. CONCLUSION

The modified PeptideBERT model demonstrated competitive performance in solubility prediction but is limited by computational resources. For fluorescence prediction, the low performance suggests exploring alternative models and extending the training epochs to capture the complex relationships in the data better.

Future work could consider architectures like the shallow CNN used in PEER, increasing training epochs and applying data augmentation strategies to improve model performance in predicting protein properties.

---

[1] National Human Genome Research Institute, Genetics home reference: Protein, https://www.genome.gov/genetics-glossary/Protein (2024), accessed: 2024-07-30.
[2] B. HU, C. TAN, L. WU, J. ZHENG, J. XIA, Z. GAO, Z. LIU, F. WU, G. ZHANG, and S. Z. LI, SCIENCE CHINA Information Sciences (2024).
[3] National Human Genome Research Institute, Genetics home reference: Translation, https://www.genome.gov/genetics-glossary/Translation (2024), accessed: 2024-07-30.
[4] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, and Y. S. Song, arXiv preprint arXiv:1906.08230 (2019), arXiv:1906.08230 [cs.LG].
[5] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, D. Bhowmik, and B. Rost, bioRxiv , 2020.07.12.199554 (2020).
[6] C. Guntuboina, A. Das, P. Mollaei, S. Kim, and A. B. Farimani, arXiv preprint arXiv:2309.03099 (2023), arXiv:2309.03099 [q-bio.BM].
[7] M. Xu, Z. Zhang, J. Lu, *et al.*, NeurIPS 2022 Datasets and Benchmarks (2022), arXiv:2206.02096.
[8] K. S. Sarkisyan *et al.*, Nature **533**, 397 (2016).
[9] Q. Zhang, B. Liu, G. Cai, J. Qian, and Z. Jin, Journal of Theory and Practice of Engineering Science **4** (2024).
[10] S. Khurana *et al.*, Bioinformatics **34**, 2605 (2018).