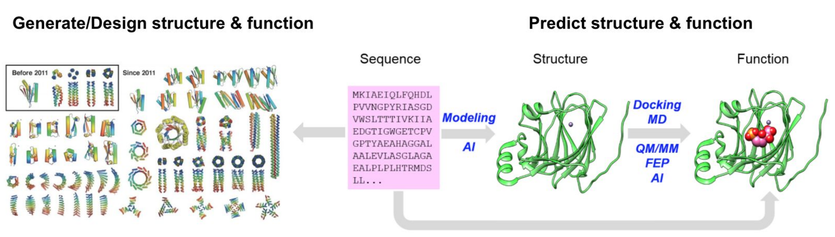
Protein Embedding Analysis

Students: Arjun Dharma, Rahil Dedhia, Thomas Waldschmidt
Advisor: Dr. Peter Rose - Director of the Structural Bioinformatics Laboratory at SDSC

Problem Statement and Context

Deep Learning transformer models such as Bidirectional Encoder Representations from Transformers (BERT) have been widely successful in a variety of natural language based tasks. Recently, BERT has been applied to protein sequences and has shown some success in protein prediction tasks relevant to biologists, such as secondary structure, fluorescence, and stability. To continue the investigation into BERT, we examined a new prediction task known as subcellular location, first described in DeepLoc (2017). Our goal was to we see if we can apply the transformer architecture to additional tasks like subcellular location prediction and achieve similar accuracy to past experiments.



Proteins are a fundamental building block of life. They perform complex functions ranging from transporting oxygen in the body, detecting stimuli, providing structure to cells, and even DNA replication. The rate at which protein structures are being identified is exponentially trailing behind the rate at which protein sequences are being discovered. This is important, especially in designing medicines, because protein structure strongly correlates with protein function, but current methods for identifying the structure of a single protein cost >$100k.

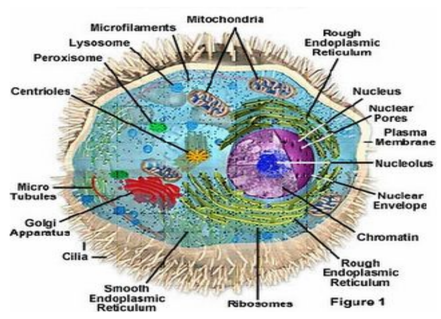
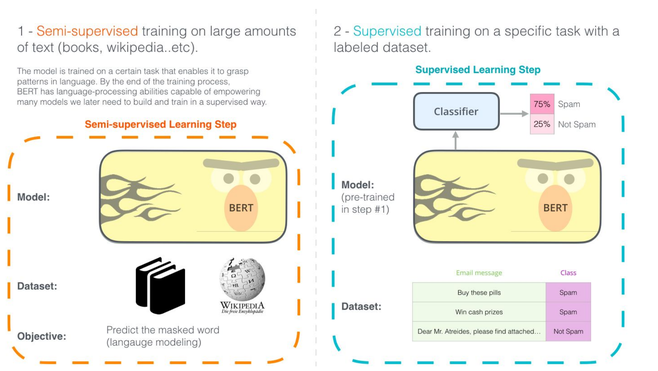
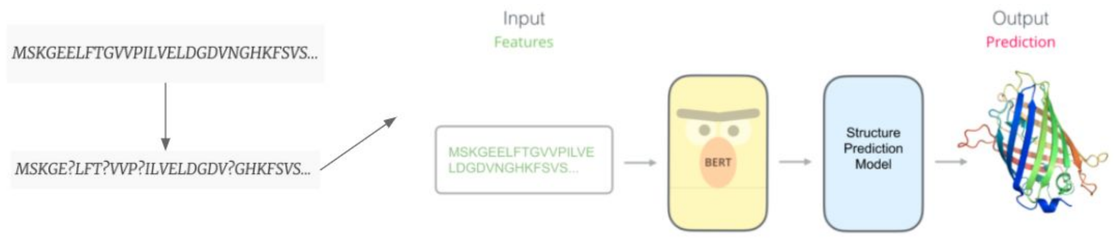


Figure 1

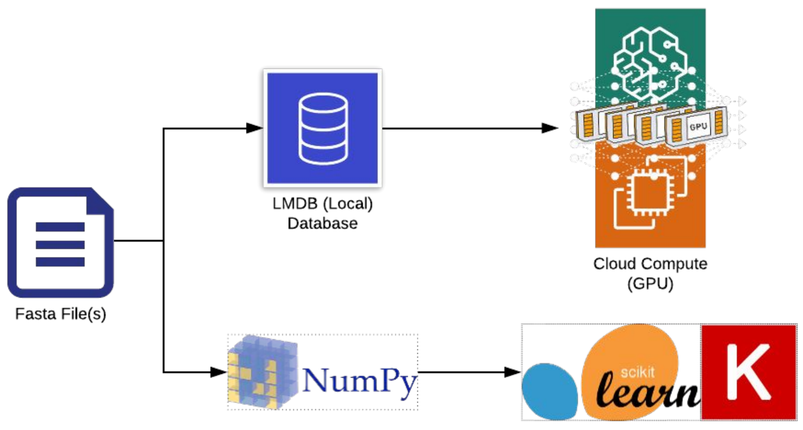


Through the use of natural language processing (NLP) and machine learning (ML) we can begin to close this gap and give researchers tools to learn information on new proteins using purely the sequences. In the NLP space, transformer models and specifically Bidirectional Encoder Representations from Transformers (BERT) have shown state of the art results in general language modeling, as well as for protein prediction tasks.



Data Science Pipeline

We took two approaches in benchmarking our DeepLoc models against the state of the art. First, we leveraged the training capabilities of TAPE in order to add the DeepLoc dataset as a task in order to run against their BERT model, as well as to benchmark it against other models they have available. This system leverages the deep learning framework pytorch and involves training over GPUs. Next, we used the embedding capabilities of TAPE to embed the DeepLoc dataset to use to train less computationally intensive machine learning models like logistic regression and xgboost.

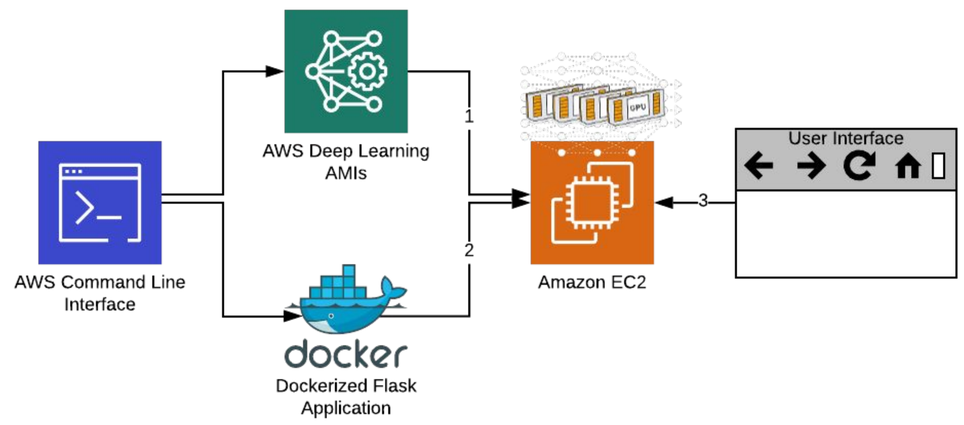


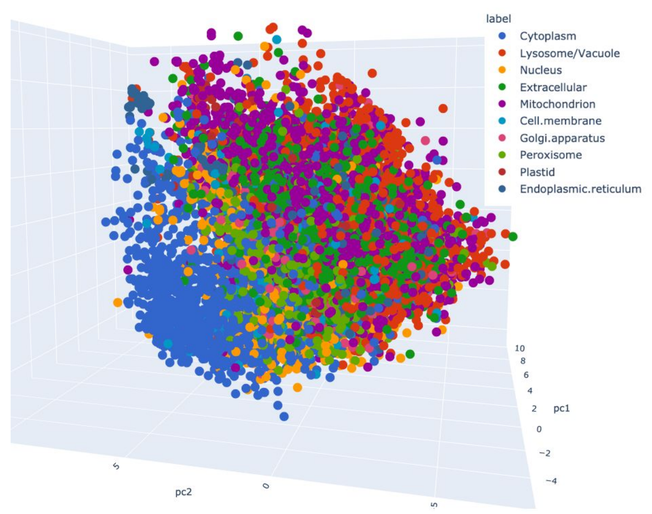
We evaluated the TAPE transformer embeddings for the DeepLoc dataset on four downstream models: xgboost, logistic regression, support vector classification (SVC), and a Keras DNN with two hidden layers of 32 nodes. The DeepLoc dataset also has a binary classification problem for membrane bound vs water soluble proteins, which we benchmarked the logistic regression and Keras DNN models against. Furthermore, we attempted to use multitask learning to train the two tasks together. In terms of using the models available in the TAPE repository for downstream tasks, we had the best success with the transformer model. We think that we could improve this result with more intuition about hyperparameter tuning, and possibly by reducing the number of trainable embeddings.

Model	Subcellular Location (Q10) Test Set Accuracy	Membrane Bound vs Water Soluble (Q2) Test Set Accuracy
Transformer	61%	
Logistic Regression	66.2%	87.2%
SVC	67.5%	
XGBoost	63.8%	
Keras DNN	64.5%	89.6%
Multitask DNN	66.2%	72.6%

Final Solution Architecture

For our final deliverable, we wanted to make our learnings on the power of pretrained embeddings for protein prediction tasks, especially transformer models, more accessible. To do this, we built an application that can be deployed with Flask + Docker w/ GPU compatibility to provide two major capabilities to researchers or the general public. The first is to easily embed protein sequences using the BERT transformer from the TAPE repository. The second is to perform PCA and visualize the embeddings in 3D space. This provides intuition on whether the TAPE embeddings will provide value for a given prediction task. We also built a simple user interface in front of this application to make it more user friendly.





Conclusion

Using BERT embeddings from a Berkeley research project titled Tasks Assessing Protein Embeddings (TAPE) as features for downstream modeling, we achieved a 67% test set accuracy using a support vector classifier for the 10 class classification task, and 89% using a Keras deep neural network for the binary classification task (membrane bound vs water soluble protein).

References

- Protein Structure: https://en.wikipedia.org/wiki/Protein_structure
- Natural Language Processing: https://en.wikipedia.org/wiki/Natural_language_processing
- BERT github: https://github.com/google-research/bert
- http://jalammar.github.io/illustrated-bert/
- https://www.biorxiv.org/content/10.1101/622803v2
- https://www.biorxiv.org/content/10.1101/589333v1
- TAPE Repository: https://github.com/songlab-cal/tape
- TAPE: https://arxiv.org/abs/1906.08230
- https://www. genecopoeia.com/product/subcellular_localization
- DeepLoc: https://academic.oup.com/bioinformatics/article/33/21/3387/3931857