**Dear Reviewer**:

Thank you for your comments concerning our project proposal. Those comments are all valuable and very helpful for revising and improving our manuscript, as well as the important guiding significance to our researches. Here are our response to each comments.

**Clarity and Structure**: The proposal is well-organized, with a clear problem statement and solution. However, some technical terms (e.g., ESM-2) could benefit from brief explanations to improve clarity.

The ESM-2 model is an advanced protein language model developed by Meta AI, designed to predict protein structure and function from amino acid sequences with high accuracy and efficiency. Using this model, we would like to extract the global sequence pattern and macro-level features. The manuscript has already modified on line 19-20.

**Originality and Significance**: The project addresses an important challenge in protein activity prediction, but the novelty could be emphasized further. CNNs are a solid choice, but exploring more innovative methods like transformers could add originality.

Protein activity prediction is an important task in the feild of protein design, which allow us to guide varieties of applications, such as enzyme engineering, biological orthogonal system design, etc. The manuscript has already modified on line 35-37.

Thanks for your advice on transformer model. The Transformer model can learn complex patterns and dependencies in protein sequences, which may effectively capture structural and functional information within the sequences. Actually, ESM-2 model is based on transformer, which will be used to embed our data into the model.

**Methodology**: The methodology is appropriate, using CNNs and embedding algorithms for protein sequence analysis. The use of different loss functions is well thought out. However, more details on the adaptive weighting mechanism, especially its role in feature integration, would improve clarity.

The adaptive weighting mechanism is a method used to dynamically adjust the importance, or “weight,” of different features, layers, or inputs based on their relevance to a specific task or context. By assigning variable weights, this mechanism allows models to emphasize more relevant information and reduce the impact of less useful data, enhancing model flexibility and

performance. Actually, this approach has already widely been used in multi-task learning and ensemble field, where it helps optimize predictions by balancing various sources of information according to changing conditions or objectives. The manuscript has already modified on line 25-27.

**Feasibility**: The project seems feasible, with a manageable dataset and a sound plan for dataset processing. However, strategies such as cross-validation could be included to ensure the model's generalizability and reduce overfitting.

Although we haven't emphasize validation process, validation is a necessary step in our project, especially when our dataset is not big enough. We plan to split dataset with 8:1:1 for trainning, validation and test. The manuscript has already modified on line 17 and line 31-32.

**Evaluation Plan**: The evaluation plan is solid, with appropriate loss functions and optimization methods. However, including external benchmarks for comparison (e.g., AlphaFold) would strengthen the assessment.

We would have a try to compare our model's performance with any other models if we have enough time to do so. Even more, we would like to test our results by experiment if the results are good enough.

**Interpretability and Impact**: The focus on model interpretability through feature regression is commendable and will likely enhance the project's impact, especially in optimizing fluorescent proteins for microscopy applications.

As we mentioned in proposal (line 32-34), we would like to have a try in order to achieve protein directed evolution.

**Writing Style and Presentation**: The writing is clear, though some grammatical issues and a lack of detailed explanations (e.g., adaptive weighting) should be addressed.

The manuscript has been double checked to avoid language error. Adaptive weighting has been expained in line 25-27.

Jinwang LU

Nov. 11th, 2024

# DSAA6000M Project Proposal

Ying Wang Tingyu ZHU Jinwang LU

13. November 2024

Any chemical substances, including proteins, follows the principle that structure determines properties. In recent years, protein structure predicting models, like AlphaFold, have made significant progress in predicting protein higher order structures from protein primary sequencing. However, predicting protein biological activity remains a challenge. Fluorescent proteins (FPs) are a great starting point for deep learning models to connect these two kinds of information due to the availability of datasets and various properties related to biological function. In this project, we would like to leverage deep learning network to link FPs sequence information with theoretical physical properties related to biological activity, specifically focusing on optimizing FPs' properties for specific microscopy experiments.

We plan to use amino acid sequences and their associated physical parameters, such as mutations, fluorescence properties and structural data as training dataset for deep learning models, which could be collected from FPbase containing detailed information of 1,073 FPs. Our goal is to predict specific physical parameters of FPs, such as fluorescence colors, intensities and lifetimes. From our model, we would like to screen for brighter or acid-resistant FPs, which have significant potential in single-molecule imaging or correlative light and electron microscopy.

We will normalize the FPbase dataset, drop missing values, encode categorical variables and split dataset with 8:1:1 for trainning, validation and test. The model will start with embedding algorithm to convert protein sequences into numerical representations, followed by ESM-2 for extracting global sequence patterns and macro-level features, since ESM-2 model has shown a great ability on predicting protein structure and function from amino acid sequences with high accuracy and efficiency. We propose using convolutional neural networks (CNNs) to process one-dimensional protein sequences, which has shown strong performance with amino acid sequences. CNNs will refine the extracted global and non-local features through local feature extraction, capturing more detailed sequence patterns. An adaptive weighting mechanism is then applied to balance and integrate the features obtained from previous steps, optimizing the model's focus on different characteristics emphasizing more relevant information and reduce the impact of less useful data, enhancing model flexibility and performance Lastly, the model concludes with a multi-task prediction Layer, which uses the processed features to predict multiple protein parameters.

We plan to use different loss functions to predict different FPs' physical parameters, such as MSE to fluorescence intensity and KL divergence to fluorescence spectrum. Adam optimizer will be used during the training process. Then model validation and fine-tuning hyperparameters will be conducted, in order to improve accuracy and avoid overfitting. Meanwhile, we will implement feature regression to identify key protein characteristics associated with specific physical properties to improve the model's interpretability. This will not only help us understand the model's predictions but may also suggest new approaches in FPs design. Further, our model connecting the protein sequencing and biological activity, will provide a new guidance on varieties of applications, such as enzyme engineering, biological orthogonal system design, etc