HALO: Hybrid Attention Model for Subcellular Localization

Shafayat Ahmed, Nazifa Ahmed Moumi, and Liqing Zhang
Virginia Tech,
Blacksburg, VA, USA
E-mail: shafayatpiyal@vt.edu, moumi@vt.edu, lqzhang@vt.edu

Subcellular localization prediction is critical for understanding protein functions and interactions, providing insights into cellular mechanisms and potential therapeutic targets. We propose HALO (Hybrid Attention model for subcellular LOcalization), a framework that integrates semantic embeddings from fine-tuned protein language models (e.g., ESM) with structural information derived from AlphaFold. HALO uses a graph attention network (GAT) to incorporate biochemical, structural, and sequence-derived features into a unified representation, while dynamically balancing their contributions. Crucially, the design allows HALO to operate in two modes: (i) a sequence-only mode, where predictions are made from the fine-tuned protein language model (PLM) when structural data are unavailable, and (ii) a hybrid mode, where structural adjacency and biochemical features complement PLM predictions, especially in low-confidence regions. We evaluate HALO on multiple datasets with minimal homology between training and test sets, where it achieves competitive performance across key metrics. By flexibly combining sequence-based and structure-informed predictions, HALO addresses the limitations of relying on a single modality and offers an adaptable framework for accurate and generalizable subcellular localization.

Keywords: Subcellular Localization, LLM, ESM, AlphaFold, Protein Structure, Graph Attention Network

1. Introduction

The localization of proteins within a cell is an essential aspect of their biological roles and interactions within the cell environment.[1] The ability to predict subcellular localization could provide benefits for functional annotation, drug discovery, and understanding disease mechanisms.[2,3] Traditionally, subcellular localization prediction has relied on alignment-based methods, such as BLAST,[4] that use sequence similarity to homologous proteins to infer localization. Although effective for well-annotated databases, these methods are not generalizable for novel or poorly characterized proteins. Increasingly, machine learning models have been developed to predict the subcellular location of a given protein. For example, CELLO[5] and WoLF PSORT[6] use features such as amino acid composition and physicochemical properties to predict subcellular localization. However, the features used in these models do not necessarily capture the intricate relationships between the amino acids within the protein sequences. To address this problem, DeepLoc[7] was developed, integrating convolutional and recurrent neural networks to capture the spatial relationships of amino acids for subcellular localization prediction. While DeepLoc is generally effective due to its ability to model sequence-based spatial patterns, it relies solely on sequence information and does not incorporate protein structural data, which is critical for improving subcellular localization predictions.[8,9] The advent of protein language models, such as ESM (Evolutionary Scale Modeling)[10] and ProtT5,[11] has further revolutionized the field. Luo et al.[12] showed how to maximize the efficacy of ESM-2 embeddings for subcellular localization tasks. These models, trained on large-scale protein databases, extract rich, context-aware embeddings that capture evolutionary and functional information. Furthermore, fine-tuning these language models has proven to improve the accuracy even further.[13] Similarly, AlphaFold[9] and ESM-fold[14] have transformed structural biology by providing accurate predictions of protein 3D structures, offering opportunities to integrate structural information into predictive models. The utilization of protein structures for downstream function annotation has proven to be effective for subcellular localization prediction tasks.[15] GPSFun[16] employs large language models to predict 3D conformations of protein sequences and extract informative sequence embeddings. These embeddings, combined with structural features processed by a geometric featurizer, are used in a graph neural network for subcellular localization prediction tasks. The AlphaFold structure database, known for its extensive repository of 3D protein structures, is widely regarded as more reliable than ESMFold due to its superior accuracy and validation against experimentally resolved structures, as demonstrated in studies like Lin et al. (2021).[17] AlphaFold's reliance on evolutionary covariation and sequence alignments provides robust structural predictions, particularly for highly conserved proteins, making it a gold standard in structural bioinformatics. In contrast, ESMFold, while faster and alignment-free, may exhibit lower accuracy for proteins with less sequence conservation or sparse evolutionary information. Fine-tuning large language models like ESM or ProtT5 on task-specific datasets further enhances their ability to capture nuanced functional and localization signals, as shown by Schmirler et al. (2024).[13] Integrating these fine-tuned models into an ensemble framework, which combines predictions from sequence embeddings, structural features, and graph neural networks, offers an unprecedented opportunity to improve predictive accuracy. Such an approach not only leverages the complementary strengths of language models and structural information but also mitigates individual model biases, providing a unified and robust solution for subcellular localization prediction.

In this study, we introduce HALO, a novel framework that combines sequence embeddings from a fine-tuned ESM model with structural features derived from AlphaFold-predicted 3D structures. Unlike existing methods, HALO leverages a Graph Attention Network (GAT)[18] to integrate these diverse feature modalities. By constructing a graph representation of proteins, where residues are nodes enriched with sequence, structural, and biochemical features, HALO captures both local and global dependencies within the protein. Additionally, using learnable weights for different feature types ensures adaptive prioritization of relevant information. Our framework bridges the gap between sequence- and structure-based predictions, addressing the limitations of previous methods that rely on either modality alone. By incorporating multimodal data and leveraging graph-based learning, HALO offers a significant advance in the ability to predict subcellular localization, paving the way for improved functional annotation of proteins, particularly in datasets with limited prior annotations.

2. Methods

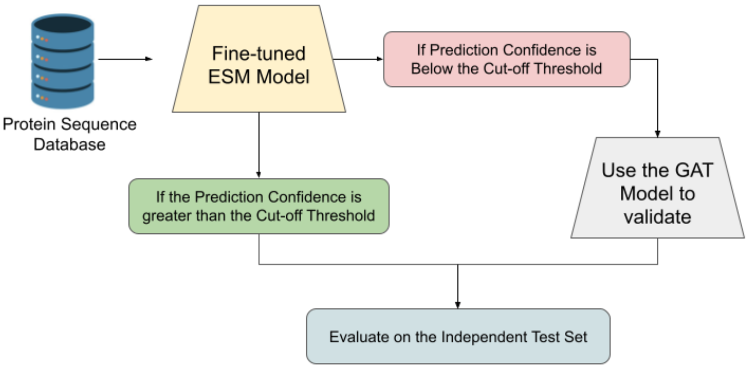
Protein Sequence Database
Fine-tuned ESM Model
If Prediction Confidence is Below the Cut-off Threshold
If the Prediction Confidence is greater than the Cut-off Threshold
Use the GAT Model to validate
Evaluate on the Independent Test Set

Fig. 1. A hybrid model leveraging Fine-tuned ESM and AlphaFold structure-based GAT attention model for subcellular localization.

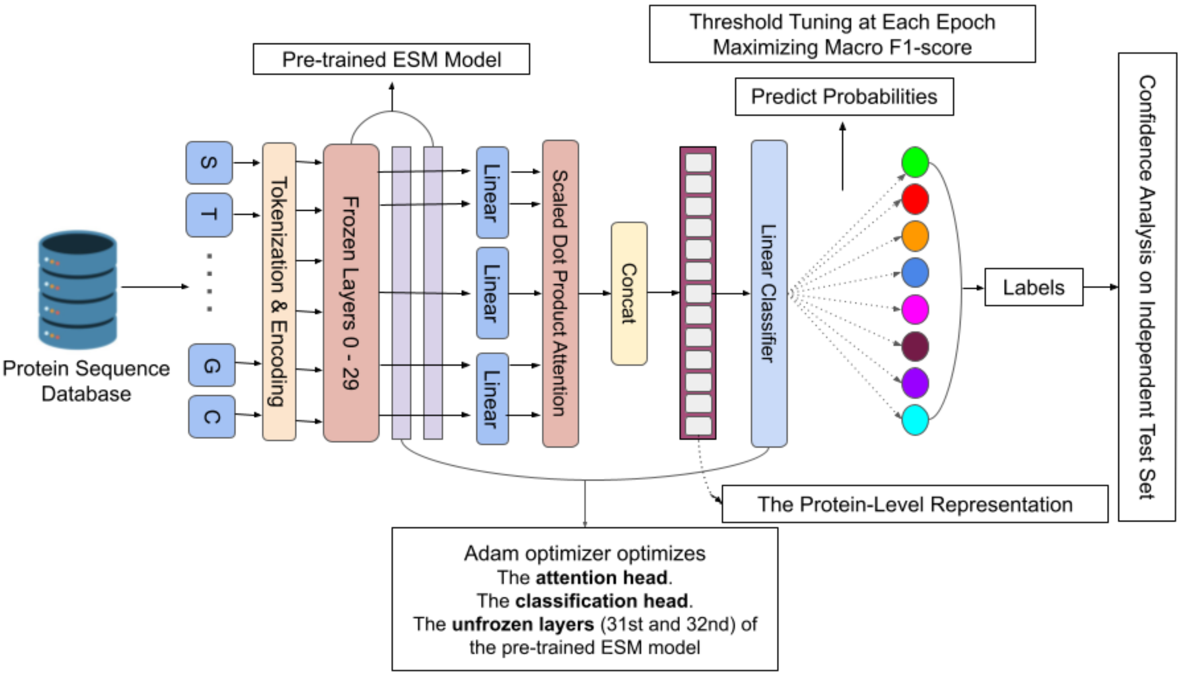
Pre-trained ESM Model
Threshold Tuning at Each Epoch Maximizing Macro F1-score
Predict Probabilities
Confidence Analysis on Independent Test Set
Protein Sequence Database
S
T
G
C
Tokenization & Encoding
Frozen Layers 0 - 29
Linear
Linear
Linear
Scaled Dot Product Attention
Concat
Linear Classifier
Labels
The Protein-Level Representation
Adam optimizer optimizes
The attention head.
The classification head.
The unfrozen layers (31st and 32nd) of the pre-trained ESM model

Fig. 2. Fine-tuned pre-trained ESM model for subcellular localization prediction.

Overview

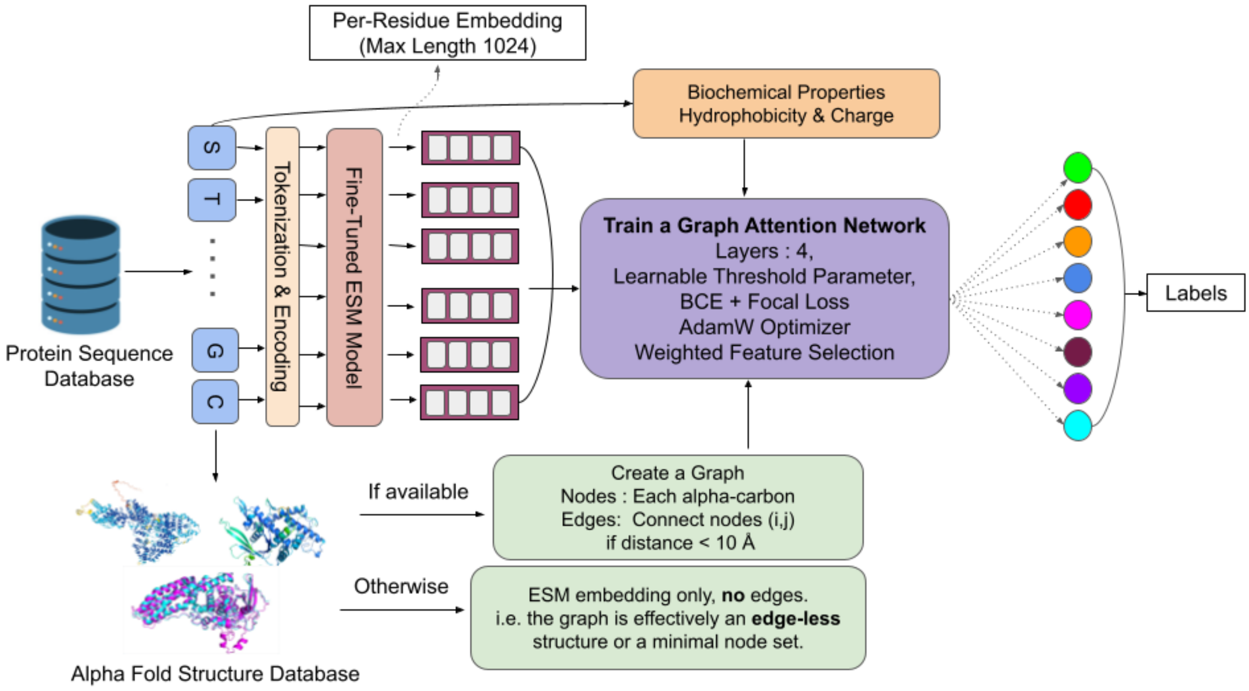
Per-Residue Embedding
(Max Length 1024)
Biochemical Properties
Hydrophobicity & Charge
S
T
G
C
Tokenization & Encoding
Fine-Tuned ESM Model
Protein Sequence Database
Train a Graph Attention Network
Layers : 4,
Learnable Threshold Parameter,
BCE + Focal Loss
AdamW Optimizer
Weighted Feature Selection
Labels
If available
Create a Graph
Nodes : Each alpha-carbon
Edges: Connect nodes (i,j)
if distance < 10 Å
Otherwise
ESM embedding only, **no** edges.
i.e. the graph is effectively an **edge-less** structure or a minimal node set.
Alpha Fold Structure Database

Fig. 3. Leveraging the Fine-Tuned ESM with AlphaFold Structures and Biochemical Properties for training a weighted graph attention model.

We developed HALO, a hybrid framework that integrates features derived from fine-tuned ESM embeddings with those extracted through a Graph Attention Network (GAT) model utilizing AlphaFold-predicted structural data. Below, we detail the individual components of these models and the novel pipeline used to ensemble them.

2.1. *Fine-Tuning ESM Model*

This model (Figure 2) is designed for multi-label protein subcellular localization. It uses the Hugging Face ESM2 backbone to generate embeddings, which are then processed by an attention-based aggregator to highlight important features. Finally, a linear classifier predicts the localization labels. Each protein's sequence is at first tokenized via the ESMTokenizer (up to length 1024 to align with the ESM2 capacity). Batches of sequences are forwarded into the ESM2 base model. Rather than directly using the final classifier layer, we introduce an attention head. Within this head, it learns a soft weight (softmax over positions) to highlight important tokens, producing a single aggregated embedding by summing each token's hidden state scaled by its attention score. On top of that aggregated embedding, a fully connected layer (the classifier) outputs multi-label logits (one dimension per subcellular class). Because the ESM2 base is large, most of its layers are frozen (except the last two encoder blocks and the classifier) to reduce computational cost and risk of overfitting. This design keeps the core ESM2 features intact while still providing some fine-tuning capability on the final layers.

Training proceeds with batches fetched from PyTorch DataLoaders, guided by a focal loss (a variant of binary cross-entropy that upweights hard-to-classify instances). During each training epoch, only the unfrozen ESM2 layers, the attention-based aggregation, and the classification heads are updated, with a checkpoint saved at the end of each iteration. After training, the code performs threshold tuning: it systematically tests thresholds from 0.1 to 0.95 in small increments, computing macro F1 at each threshold and logging the results. The best threshold is chosen by maximum macro F1, which is then used for the final evaluation. In parallel, the script groups each prediction's confidence into bins (e.g., 95–100%, 90–95%, etc.), tracking how often the model is correct vs. incorrect in each bin to analyze calibration. Early stopping is driven by a plateau in the learning rate scheduler (ReduceLROnPlateau). Based on the threshold analysis, 0.45 yielded the maximum F1 score on the validation set.

2.2. *The GAT Model Training*

The GAT model (Figure 3) is trained by generating fine-tuned ESM embeddings for each protein sequence. Sequences are truncated to 1,024 tokens to fit ESM2's maximum input length extracting the final hidden layer on a per-residue basis. For three-dimensional structural information, we utilize AlphaFold-predicted Protein Data Bank (PDB) files obtained from the AlphaFold structure database. We choose AlphaFold structures over, for example, ESMFold predictions because AlphaFold's large-scale, standardized dataset offers broader coverage and consistency. If the file is absent, that is, no structural information is available, we will use only fine-tuned ESM embeddings (no structural coordinates or edges). As a result, proteins lacking structural information are still included in training and inference, represented solely by node features. This approach avoids discarding valuable sequence-level data simply because an AlphaFold structure is unavailable. For each protein, we build a graph where each residue's alpha-carbon is a node and these nodes are connected with edges if the distance is less than 10 Å. We rely on learned attention coefficients within the GAT to model residue-residue interactions. Each node's feature vector comprises scaled three-dimensional coordinates (coord_weight), fine-tuned ESM embedding (esm_weight), and biochemical properties (bio_weight). A multi-layer GAT with global mean pooling transforms these features into graph-level embeddings, which undergo a focal loss that uses a learnable threshold for multi-label predictions. Early stopping is triggered if macro F1 on the validation set ceases to improve. We varied weights on coordinates, ESM embeddings, and biochemical properties to balance structural and sequence information, tested combinations, and selected the one that maximized validation accuracy within our computational limits.

2.3. *HALO*

We introduce HALO, a hybrid architecture that harnesses the complementary strengths of two separate models: a fine-tuned ESM (sequence-based) model and a GAT (structure-based) model. Our empirical comparisons revealed that the fine-tuned ESM model delivers higher overall accuracy and often outperforms the GAT model on a per-class basis. However, a detailed confidence analysis (see Figure 4) indicates that the ESM model's predictive reliability decreases substantially (Accuracy drops below 60%) for the DeepLoc dataset below a 55% confidence threshold. In other words, for predictions labeled with less than 55% confidence, the fine-tuned ESM model is more prone to misclassification. This shortcoming motivated us to incorporate the GAT model's alternative perspective, particularly its ability to leverage 3D structural adjacency through residue-residue edges, to boost performance on those uncertain predictions.

Hence, HALO operates by initially relying on the fine-tuned ESM model's logits to classify each protein sample. When the ESM model's confidence for a label is above 55%, the label is taken as the final prediction. When confidence dips below 55%, the GAT model's prediction is considered as well and if both models predict the same label, we incorporate it in the final multi-label output, otherwise, we discard that prediction. This cross-checking mechanism effectively rescues uncertain predictions, mitigating the ESM model's confidence drop. Figure 1, provides an overview of HALO's pipeline: sequences flow first into ESM for classification; afterward, low-confidence candidates are passed to the GAT for structural verification, leading to robust synergy in multi-label subcellular localization tasks. By combining sequence-derived semantics (ESM) with structural adjacency insights (GAT), HALO achieves higher overall precision across different confidence levels than either model alone.

2.4. *Implementation Details*

The framework was implemented in Python, leveraging PyTorch and PyTorch Geometric libraries. All computations were performed on an NVIDIA RTX 4090 GPU to handle the computational demands of graph-based learning and large protein embeddings. Detailed logging of training and evaluation metrics was maintained for reproducibility, and the model dynamically adjusted weights for each feature type during training to optimize the contribution of sequence, structural, and biochemical information. The complete code, data, and, results are included in the repository - https://github.com/Shafayat115/HALO.

3. Experiments & Results

3.1. *Evaluation Metrics*

We split the data into training, validation, and test sets with minimal inter-set homology to increase predictive difficulty. We used the validation set for hyperparameter tuning and the test set for final evaluation. Our evaluation employed a broad suite of metrics, including accuracy, macro and micro-averaged F1-scores, Jaccard Index, precision, recall, and Matthews Correlation Coefficient (MCC). To further assess the contribution of each model component, we conducted an extensive comparative analysis of the sub-models. Additionally, we benchmarked our hybrid approach against state-of-the-art methods across two distinct datasets, emphasizing its performance under challenging conditions.

3.2. *Performance Analysis for Component Models*

Table 1, compares the performance of three GAT models on the DeepLoc test set: one using the pre-trained ESM embeddings (ESM), one with the fine-tuned ESM embeddings (F-ESM), and the last version is with a weighted combination of the fine-tuned ESM embeddings, biochemical features, and structural coordinates (WF-ESM). The WF-ESM model demonstrates the best performance across most metrics, showcasing the benefits of incorporating additional

Table 1. Performance comparison of three GAT models on the DeepLoc test set.

Metric	**ESM**	**F-ESM**	**WF -ESM**
macro F1	0.5420	0.55	0.5942
micro F1	0.6023	0.61	0.6803
Accuracy	0.2671	0.24	0.3819
Jaccard Index	0.5375	0.5375	0.6164
Precision	0.4706	0.48	0.5771
Recall	0.8361	0.85	0.8283
ROC AUC	0.9293	0.93	0.9399
Label-wise MCC			
Cytoplasm	0.4608	0.38	0.4798
Nucleus	0.6806	0.67	0.5731
Extracellular	0.8209	0.84	0.8389
Cell Membrane	0.4516	0.47	0.5401
Mitochondrion	0.6973	0.74	0.7371
Plastid	0.7935	0.83	0.8456
Endoplasmic Reticulum	0.4088	0.4	0.5314
Lysosome/Vacuole	0.2798	0.25	0.2674
Golgi Apparatus	0.2043	0.22	0.3214
Peroxisome	0.2259	0.25	0.3112

structural and biochemical information. It achieves the highest macro F1-Score (0.5942), micro F1-Score (0.6803), and Jaccard Index (0.6164), reflecting a better balance and accuracy in multi-label classification compared to the other models. The ROC AUC of WF-ESM (0.9399) is also slightly higher than the other two models, highlighting its robust ability to distinguish between classes. Furthermore, the WF-ESM model shows notable improvements in Precision (0.5771), indicating a reduction in false positives, while maintaining a strong Recall (0.8283).

The F-ESM model improves over the baseline ESM model in terms of Recall (0.85 vs. 0.8361) and maintains competitive performance for most metrics. However, its macro F1-Score (0.55) and micro F1-Score (0.61) are only slightly better than the ESM model, indicating that fine-tuning alone provides moderate improvements without incorporating additional features.

The ESM model shows relatively weaker performance, with lower macro F1 (0.5420), micro F1 (0.6023), and precision (0.4706). Despite these shortcomings, it demonstrates strong recall (0.8361) and robust performance in high-confidence predictions, as indicated by its ROC AUC (0.9293).

Regarding label-wise MCC, the WF-ESM model consistently outperforms the other two models for most subcellular localization labels. Notably, it achieves the highest MCC for Cell Membrane (0.5401), Endoplasmic Reticulum (0.5314), and Golgi Apparatus (0.3214), demonstrating its ability to accurately predict labels that are typically more challenging due to their sparse representation. All models perform competitively for well-represented labels such as Plastid and Extracellular, with MCC values above 0.8 for both labels across all three models. However, the WF-ESM model still leads among them with marginal improvements in these categories.

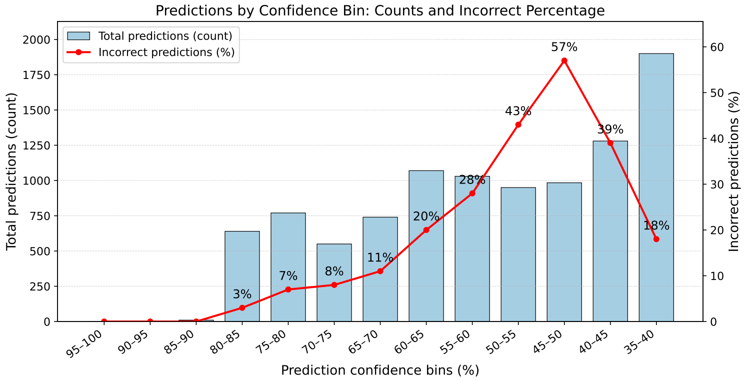

Fig. 4. Confidence distribution of the fine-tuned ESM model on the evaluation set, with total predictions and misclassifications per confidence range. Errors concentrate at lower confidence ranges, indicating that confidence is informative about reliability.

Figure 4, illustrates the distribution of total and incorrect predictions across varying confidence ranges for the fine-tuned ESM model, highlighting the relationship between prediction confidence and model accuracy. Predictions with high confidence (70-100%) demonstrate remarkable reliability, with negligible incorrect predictions, underscoring the robustness of the model in these intervals. However, as confidence decreases, the proportion of incorrect predictions increases significantly, particularly in the 55-65% range, marking a critical threshold for model reliability. In the lowest confidence ranges (55-45% and 45-40%), nearly half of the predictions are incorrect, indicating a substantial decline in model performance. These results emphasize the importance of using confidence thresholds to assess prediction reliability, suggesting that predictions below 55% confidence should be subjected to additional validation. This analysis provides critical insights into the strengths and limitations of the model across confidence levels, guiding its effective application in practical scenarios.

Table 2, highlights the comparative performance of HALO, fine-tuned ESM, and GAT models on the DeepLoc test set, showcasing HALO's ability to bridge the strengths of both sequence and structure-focused approaches. HALO achieves a macro F1 of 0.6782 and a micro F1 of 0.7364, indicating robust performance that effectively balances the complementary modalities of sequence and structural information. Notably, HALO surpasses the fine-tuned ESM model in key metrics such as Jaccard Index (0.7187 vs. 0.7075) and ROC AUC (0.9412 vs. 0.9254), while also demonstrating strong precision (0.7098) and overall MCC (0.7131), reflecting its capacity for multi-label classification across diverse localization labels.

The fine-tuned ESM model, while slightly ahead in metrics like macro F1 (0.6688) and micro F1 (0.7471), emphasizes nuanced sequence-level features, benefiting from attention-based aggregation. On the other hand, the GAT model leverages structural adjacency and 3D relationships to deliver strong MCC values for Extracellular (0.8389) and Plastid (0.8456), although it falls behind in overall metrics like accuracy (0.3819) and macro F1 (0.5942).

HALO's hybrid approach integrates high-confidence predictions from the fine-tuned ESM

Table 2. Comparison of overall performance metrics and MCC per subcellular localization label across HALO, fine-tuned ESM, and GAT models on the DeepLoc test set.

Metric	**HALO**	**Fine-tuned ESM**	**GAT**
macro F1	0.6782	0.6688	0.5942
micro F1	0.7364	0.7471	0.6803
Accuracy	0.5756	0.5621	0.3819
Jaccard Index	0.7187	0.7075	0.6164
Precision	0.7098	0.6996	0.5771
Recall	0.6511	0.6537	0.8283
ROC AUC	0.9412	0.9254	0.9399
Matthews Correlation Coefficient (MCC) per Label			
Cytoplasm	0.6125	0.6162	0.4798
Nucleus	0.6819	0.6821	0.5731
Extracellular	0.8559	0.8554	0.8389
Cell Membrane	0.6752	0.6660	0.5401
Mitochondrion	0.7921	0.7999	0.7371
Plastid	0.8918	0.8758	0.8456
Endoplasmic Reticulum	0.5735	0.5864	0.5314
Lysosome/Vacuole	0.2621	0.2982	0.2674
Golgi Apparatus	0.3349	0.3283	0.3214
Peroxisome	0.5835	0.5991	0.3112

model with structural insights from the GAT model, particularly excelling in cases where sequence- or structure-only models face limitations. This synergy allows HALO to enhance robustness and maintain precision across challenging labels, such as Extracellular (MCC: 0.8559) and Plastid (MCC: 0.8918). However, the model shows room for improvement in complex localizations such as Lysosome/Vacuole (MCC: 0.2621) and Golgi Apparatus (MCC: 0.3349), suggesting that further optimization in the ensemble mechanism could enhance performance for these challenging categories. HALO's ability to integrate sequence and structure-derived features establishes it as a versatile and adaptable framework for subcellular localization prediction.

3.3. *Comparison with other tools*

We compared our model's performance with two of the state-of-the-art tools for subcellular localization prediction using two different datasets curated by them. Here is a detailed comparison of our tool in these datasets -

3.3.1. *DeepLoc*

Table 3, showcases the comparative performance of HALO, DeepLoc 1.0, DeepLoc 2.0 (ESM1b and ProtT5), and other models on the DeepLoc test set, underscoring HALO's strong performance relative to state-of-the-art methods. HALO achieves a macro F1 score of 0.6782 and a micro F1 score of 0.7364, significantly outperforming DeepLoc 1.0 (0.47 and 0.58, respectively)

Table 3. Comparison of Performance Metrics and Matthews Correlation Coefficient (MCC) per Subcellular Localization Label across All Models

Metric	**DeepLoc 1.0**	**DeepLoc 2.0 (ESM1b)**	**DeepLoc 2.0 (ProtT5)**	**HALO**	**Fine-tuned ESM**	**GAT**
Accuracy	0.48	0.53	0.55	0.5756	0.5621	0.3819
Jaccard Index	0.56	0.68	0.69	0.7187	0.7075	0.6164
micro F1	0.58	0.72	0.73	0.7364	0.7471	0.6803
macro F1	0.47	0.64	0.66	0.6782	0.6688	0.5942
Precision	–	–	–	0.7098	0.6996	0.5771
Recall	–	–	–	0.6511	0.6537	0.8283
ROC AUC	–	–	–	0.9412	0.9254	0.9399
Matthews Correlation Coefficient (MCC) per Label						
Cytoplasm	0.45	0.61	0.62	0.6125	0.6162	0.4798
Nucleus	0.46	0.66	0.69	0.6819	0.6821	0.5731
Extracellular	0.78	0.85	0.85	0.8559	0.8554	0.8389
Cell Membrane	0.53	0.64	0.66	0.6752	0.6660	0.5401
Mitochondrion	0.58	0.73	0.76	0.7921	0.7999	0.7371
Plastid	0.69	0.88	0.90	0.8918	0.8758	0.8456
Endoplasmic Reticulum	0.32	0.52	0.56	0.5735	0.5864	0.5314
Lysosome/Vacuole	0.06	0.24	0.28	0.2621	0.2982	0.2674
Golgi Apparatus	0.20	0.36	0.34	0.3349	0.3283	0.3214
Peroxisome	0.15	0.48	0.56	0.5835	0.5991	0.3112

and closely approaching DeepLoc 2.0 (ProtT5), which achieves a macro F1 of 0.66 and a micro F1 of 0.73. HALO's Jaccard Index (0.7187) and accuracy (0.5756) also exceed DeepLoc 1.0 (0.56 and 0.48, respectively) and closely align with ProtT5's performance (0.69 and 0.55).

HALO demonstrates consistent improvement over DeepLoc 1.0 in several subcellular localization labels, such as Cytoplasm (0.6125 MCC), Nucleus (0.6819 MCC), and Cell Membrane (0.6752 MCC). For complex localizations like Extracellular (0.8559 MCC) and Plastid (0.8918 MCC), HALO performs on par with DeepLoc 2.0 models, maintaining robust predictions. However, HALO shows a slight performance gap in more challenging labels like Lysosome/Vacuole (0.2621 MCC) and Golgi Apparatus (0.3349 MCC), where it lags behind ProtT5 (0.28 and 0.34, respectively).

Overall, HALO bridges the gap between DeepLoc 1.0 and the more advanced DeepLoc 2.0 models by effectively integrating sequence and structural insights through its hybrid approach. While it does not universally surpass DeepLoc 2.0 (ProtT5) in every metric, HALO offers a balanced and versatile framework for subcellular localization prediction, with the potential for further optimization to improve predictions in complex localization categories.

3.3.2. *Single-Label Swissprot Dataset*

We also compared our model with the work of Luo et. al.[12] They proposed different ESM2 representation feature extraction strategies, considering both the character type and position within the ESM2 input sequence. They benchmarked their models curated from the SwissProt section of the UniProt database,[19] focusing on entries with a "Reviewed" status and confirmed protein existence at the protein level (updated as of April 15, 2023). Proteins selected for their study had amino acid sequence lengths ranging from 15 to 4000 and were associated with a single classification label. The dataset was randomly shuffled and split into two subsets: a training set (60% of the total) and a test set (40%). Additionally, protein localization data were obtained from the TrEMBL section of the UniProt database, serving as independent test datasets. De-homology datasets were subsequently generated by excluding proteins that showed homology with those in SwissProt. We performed using the exact train and test set for HALO and compared it with the results they put up in their paper. We added some additional metrics to better represent our model's performance in that data split, which are represented as a gap for their methodologies.

Table 4. Performance comparison across the three datasets S (Swiss-Prot), T (TrEMBL Independent), and N (non-homology TrEMBL Independent) for our HALO model and the interpretable LLM models.[12]

Metric	HALO_S	RF_S	MLP_S	HALO_T	RF_T	MLP_T	HALO_N	RF_N	MLP_N
Macro Precision	0.9178	0.722	0.644	0.8712	–	–	0.7860	–	–
Macro Recall	0.5319	0.563	0.606	0.4426	–	–	0.3364	–	–
macro F1	0.5722	0.596	0.621	0.4900	–	–	0.3827	–	–
micro F1	0.8358	–	–	0.7182	–	–	0.5672	–	–
Jaccard Index	0.7940	–	–	0.6503	–	–	0.4912	–	–
Accuracy	0.7924	0.9555	0.963	0.6468	–	–	0.4880	–	–
ROC AUC	0.9284	0.766	0.792	0.9072	–	–	0.8242	–	–
Overall MCC	0.8198	–	–	0.6943	–	–	0.5338	–	–
MCC Per Label									
Cell Junction	0	0.377	0.187	0	0.06	0.07	0	0	0
Cell Membrane	0.83	0.780	0.824	0.72	0.62	0.67	0.50	0.21	0.26
Cell Projection	0	0.069	0.151	0	0.21	0.12	0	0	0
Cytoplasm	0.66	0.558	0.620	0.44	0.40	0.43	0.26	0.19	0.19
ER	0.71	0.618	0.699	0.74	0.65	0.71	0.58	0.19	0.18
Golgi Apparatus	0.62	0.605	0.612	0.43	0.51	0.48	0.34	0.41	0.25
Lysosome	0.24	0.380	0.314	0.14	0.68	0.63	0	0	0
Mitochondrion	0.89	0.878	0.865	0.77	0.73	0.64	0.67	0.47	0.23
Nucleus	0.87	0.730	0.836	0.76	0.61	0.68	0.55	0.26	0.30
Secreted	0.91	0.866	0.896	0.81	0.69	0.72	0.60	0.53	0.52

Table 4, represents the performance comparison across the three datasets SwissProt (S), TrEMBL Independent (T), and Non-Homology TrEMBL Independent (N). It highlights the effectiveness of different models, including HALO, and the interpretable LLM's,[12] two final models comprising all their features with Random Forest (RF), and Multi-Layer Perceptron (MLP) model. HALO consistently demonstrates superior performance in most metrics across the datasets, particularly excelling in Macro Precision, Macro Recall, and ROC AUC, with scores of 0.9178, 0.5319, and 0.9284, respectively, on the SwissProt dataset. This indicates its ability to make highly confident and accurate predictions while capturing class-specific variability. In contrast, RF and MLP models show comparable but slightly lower performance, particularly in Precision and F1 Score for the SwissProt dataset, with MLP slightly outperforming RF in macro Recall and macro F1. On the TrEMBL Independent dataset, HALO maintains its lead with a Macro Precision of 0.8712 and an ROC AUC of 0.9072, indicating its robustness in generalizing to independent datasets. However, RF and MLP performance metrics for this dataset were not fully available, suggesting potential limitations in model generalizability or dataset-specific challenges.

For the Non-Homology TrEMBL Independent dataset, which presents the most challenging scenario due to the elimination of homologous proteins, HALO still performs reasonably well, achieving a macro F1 of 0.3827 and an ROC AUC of 0.8242. These results underscore HALO's ability to make accurate predictions in scenarios with reduced homology information. In comparison, RF and MLP show worse performance across most metrics, particularly for complex labels like Cell Membrane and Mitochondrion, highlighting their limitations in handling non-homologous data.

Overall, HALO's superior performance, particularly in ROC AUC and Macro Precision, indicates its strength in capturing complex patterns within protein localization data. While RF and MLP provide competitive results, especially for simpler datasets, their performance declines in independent and non-homologous datasets, emphasizing the importance of advanced modeling techniques like HALO.

4. Conclusion

HALO represents a novel hybrid approach that integrates sequence and structural information to enhance subcellular localization prediction. By combining fine-tuned protein language model (PLM) embeddings, biochemical properties such as residue charge and hydrophobicity, and AlphaFold-predicted structural information, HALO delivers competitive performance across multiple datasets with minimal homology between train and test splits. Importantly, HALO demonstrates that structural information, even in a relatively simple representation, can play a critical role in rescuing low-confidence sequence-based predictions, highlighting the value of complementary modalities for functional annotation. These findings underscore the importance of flexible frameworks that can operate effectively in both sequence-only and hybrid sequence–structure modes.

Future work will focus on several directions. First, more sophisticated structural encodings, such as contact maps, torsion angles, and geometric message-passing layers, could capture richer spatial dependencies and further improve predictive accuracy, especially for challenging compartments like Lysosome/Vacuole and Golgi Apparatus. Second, systematic benchmarking of different fine-tuning strategies for PLMs and alternative graph neural network architectures will be pursued as resources permit, extending beyond our current constraint of fine-tuning only the final two ESM layers. Third, sensitivity analyses of design parameters such as the 55% gating threshold and the 10 Å distance cutoff will provide a more comprehensive justification of these choices. Finally, future work will explore feature attribution and modality importance analyses to improve the biological interpretability of HALO, clarifying which inputs contribute most to accurate predictions. Collectively, these enhancements will improve robustness, scalability, and interpretability, expanding HALO's utility to broader protein functional annotation tasks and advancing the integration of multi-modal data in bioinformatics.

Acknowledgments and Appendices

This work was partly funded by the National Science Foundation (NSF grants #2319522, #2125798, and #2004751).

References

1. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson *et al.*, *Molecular biology of the cell* (Garland New York, 1994).
2. H. Yin and A. D. Flynn, Drugging membrane protein interactions, *Annual review of biomedical engineering* **18**, 51 (2016).
3. J. P. Overington, B. Al-Lazikani and A. L. Hopkins, How many drug targets are there?, *Nature reviews Drug discovery* **5**, 993 (2006).
4. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic acids research* **25**, 3389 (1997).
5. C.-S. Yu, Y.-C. Chen, C.-H. Lu and J.-K. Hwang, Prediction of protein subcellular localization, *Proteins: Structure, Function, and Bioinformatics* **64**, 643 (2006).
6. P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-Collier and K. Nakai, Wolf psort: protein localization predictor, *Nucleic acids research* **35**, W585 (2007).
7. J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen and O. Winther, Deeploc: prediction of protein subcellular localization using deep learning, *Bioinformatics* **33**, 3387 (2017).
8. G. Wang, Y.-J. Zhai, Z.-Z. Xue and Y.-Y. Xu, Improving protein subcellular location classification by incorporating three-dimensional structure information, *Biomolecules* **11**, p. 1607 (2021).
9. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with alphafold, *nature* **596**, 583 (2021).
10. D. J. Beal, Esm 2.0: State of the art and future potential of experience sampling methods in organizational research, *Annu. Rev. Organ. Psychol. Organ. Behav.* **2**, 383 (2015).
11. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger *et al.*, Prottrans: Toward understanding the language of life through self-supervised learning, *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112 (2021).
12. Z. Luo, R. Wang, Y. Sun, J. Liu, Z. Chen and Y.-J. Zhang, Interpretable feature extraction and dimensionality reduction in esm2 for protein localization prediction, *Briefings in Bioinformatics* **25**, p. bbad534 (2024).
13. R. Schmirler, M. Heinzinger and B. Rost, Fine-tuning protein language models boosts predictions across diverse tasks, *Nature Communications* **15**, p. 7407 (2024).
14. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido *et al.*, Language models of protein sequences at the scale of evolution enable accurate structure prediction, *bioRxiv* (2022).
15. G. Dubourg-Felonneau, A. Abbasi, E. Akiva and L. Lee, Improving protein subcellular localization prediction with structural prediction & graph neural networks, *bioRxiv* , 2022 (2022).
16. Q. Yuan, C. Tian, Y. Song, P. Ou, M. Zhu, H. Zhao and Y. Yang, Gpsfun: geometry-aware protein sequence function predictions with language models, *Nucleic Acids Research* , p. gkae381 (2024).
17. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (2023).
18. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, Graph attention networks, *arXiv preprint arXiv:1710.10903* (2017).
19. U. Consortium, Uniprot: a hub for protein information, *Nucleic acids research* **43**, D204 (2015).