International Journal of Life Sciences and Biotechnology
Omar, J. M., Attention-Enhanced Bidirectional Long Short-Term Memory (Bi-LSTM) Ensembles with Frozen Evolutionary Scale Modeling-2 (ESM-2) Embeddings Achieve Competitive Performance in Protein Subcellular Localization. International Journal of Life Sciences and Biotechnology, 2026 9(2): p. 90-105.
https://doi.org/10.38001/ijlsb.1882985
Araştırma Makale/ Research article

Attention-Enhanced Bidirectional Long Short-Term Memory (Bi-LSTM) Ensembles with Frozen Evolutionary Scale Modeling-2 (ESM-2) Embeddings Achieve Competitive Performance in Protein Subcellular Localization

Johaimen M. Omar[1]*

ABSTRACT
Protein subcellular localization prediction is a fundamental challenge in computational biology, and while recent deep learning approaches have achieved strong performance on this task, independent validation on genuinely novel proteins remains uncommon. Here, an attention-enhanced Bidirectional Long Short-Term Memory (Bi-LSTM) ensemble, trained exclusively on frozen ESM-2 protein language model embeddings, attained 86.81% accuracy (MCC = 0.825) on a homology-reduced four-class test set (CD-HIT, 40% threshold), surpassing an SVM baseline by 22.27 percentage points (MCC = 0.530). On 86 prospectively collected eukaryotic proteins released in 2024 and unseen during training, the ensemble reached 88.37% accuracy (MCC = 0.827), outperforming DeepLoc 2.1 (80.23%, MCC = 0.714, p = 0.007) and MULocDeep (77.91%, MCC = 0.682, p = 0.019). The architecture achieves this without complex multi-stage designs. Quantitative alignment of attention peaks against UniProt-annotated biological features confirmed class-dependent enrichment patterns consistent with known targeting mechanisms, providing database-grounded evidence that the model captures biologically relevant sequence features. These findings suggest that embedding quality, rather than architectural complexity, may be a primary determinant of localization prediction performance, though the limited benchmark size and four-compartment scope warrant cautious generalization.

ARTICLE HISTORY
Received
05 February 2026
Accepted
20 April 2026

KEY WORDS
attention mechanisms,
Bi-LSTM,
deep learning,
ESM-2 embeddings,
Protein subcellular localization

Introduction

Subcellular localization determines the functional context of every protein in the cell. The compartment in which a protein resides governs its access to substrates, interaction partners, and regulatory machinery, making correct spatial organization a prerequisite for metabolic pathway integrity, signal transduction, and gene regulation [1, 2]. Disruption of normal localization is a recognized mechanism in a broad range of diseases, including cancer, neurodegeneration, and metabolic disorders, where mislocalized proteins either acquire aberrant interactions or fail to fulfill their intended roles [3, 4]. From a translational perspective, a protein's spatial distribution and movement between subcellular compartments, as well as its presence in the plasma, are directly relevant to identifying potential therapeutic candidates. Accessible targets, including membrane-bound receptors and circulating plasma proteins, are central among validated pharmacological treatments, and characterizing these locations is therefore a vital step in drug target research [5, 6]. Despite this biological and clinical significance, large-scale proteomes remain incompletely annotated, and experimental localization assays are low-throughput and resource-intensive [1, 2]. Computational prediction thus serves as an indispensable annotation tool. However, while current studies routinely use predicted localization in target selection, rigorous independent validation of predictors on genuinely novel proteins is still uncommon in large-scale applications [7, 8].

Prediction of protein subcellular localization is a fundamental task in computational biology, essential for annotating large-scale proteomic data. However, the reliability of current predictors is often compromised by inadequate homology partitioning, where insufficient separation between training and testing datasets can lead to inflated performance estimates [9–11]. While traditional methods relied on manual feature engineering,

[1] Institute of Graduate Education, Kastamonu University, Kastamonu 37200, Türkiye
Mindanao State University-Main Campus, Marawi City, Lanao del Sur 9700, Philippines
e-mail: johaimen.omar@msumain.edu.ph / 241057007@ogr.kastamonu.edu.tr

recent advances have shifted toward protein language models (PLMs) like evolutionary scale modeling-2 (ESM-2), which capture evolutionary context directly from raw sequences [9, 11].

State-of-the-art predictors, such as DeepLoc 2.1 [12], have successfully integrated embeddings within sophisticated multi-stage architectures that include specialized components like attention pooling with DCT regularization to maximize performance. In this study, we aim to investigate the sufficiency of a streamlined deep learning approach. We hypothesize that frozen ESM-2 embeddings, when paired with a simple Bidirectional Long Short-Term Memory (Bi-LSTM) classifier, can match the performance of complex multi-stage architectures without requiring manual feature engineering or elaborate architectural designs.

To validate this hypothesis, we implemented a standard Bi-LSTM network with a simple attention mechanism. Rather than aiming to outperform established tools, our objective is to determine if this generalized architecture can achieve robustness and generalizability comparable to the state-of-the-art DeepLoc 2.1 and MULocDeep [13]. We evaluate this approach using a temporal benchmark of novel proteins released in 2024, providing an engineering perspective on the trade-off between architectural complexity and embedding quality.

Materials and Methods

Data Acquisition and Preprocessing

Dataset Construction

The primary dataset for this study was constructed by retrieving protein sequences from the UniProt Knowledgebase (UniProtKB), encompassing diverse eukaryotic organisms (Metazoa, Viridiplantae, Fungi) [14–16]. This taxonomic restriction is inherent to the target classes (Nucleus, Cytoplasm, Mitochondrion, and Cell Membrane), as prokaryotic organisms lack the nuclear and mitochondrial compartments defined in our study schema.

To generate a balanced initial dataset, we utilized the UniProt REST API to fetch approximately 2,500 sequences per class. To ensure high-fidelity annotations, we utilized the UniProt Subcellular Location (SL) controlled vocabulary rather than free-text searches. We enforced strict inclusion criteria using specific SL identifiers: SL-0191 (Nucleus), SL-0086 (Cytoplasm), SL-0173 (Mitochondrion), and SL-0039 (Cell membrane). While datasets were managed in tabular formats for this study, the pipeline is format-agnostic and natively accepts standard sequence representations, including FASTA.

To exclude prokaryotic sequences, which lack defined membrane-bound organelles, we applied a strict taxonomic filter for Eukaryota (Taxonomy ID: 2759). Furthermore, to prevent label ambiguity, we employed explicit negative filtering (e.g., *Nucleus* queries explicitly excluded *Cytoplasm*, *Membrane,* and *Mitochondrion* annotations) to guarantee single-label multiclass classification.

Filtering and Quality Control

We filtered raw sequences to ensure compatibility with the deep learning pipeline. We restricted sequence lengths to a range of 50 to 600 amino acids to accommodate the fixed input window of the model while filtering out fragments. Furthermore, sequences containing non-standard or ambiguous amino acid codes (B, J, Z, X, U, O) were removed to ensure a standard vocabulary of the 20 canonical amino acids.

Homology Reduction and Data Splitting

To evaluate the model's generalization capability and prevent performance inflation due to sequence homology (data leakage), we applied a strict redundancy reduction strategy [9]. We employed a sequence identity threshold of 40% via Cluster Database at High Identity with Tolerance (CD-HIT) [17]. While some specialized signal peptide predictors utilize a stricter 30% cutoff [9, 10], we opted for a 40% threshold to preserve the structural diversity essential for training high-capacity deep learning models.

The dataset was subsequently split into training, validation, and testing sets using a cluster-aware partitioning strategy (GroupShuffleSplit). This ensured that all sequences within a specific CD-HIT cluster were assigned to the same subset, thereby guaranteeing that the model would not encounter near-identical sequences between the training and testing phases. The CD-HIT clustering process yielded 4,662 distinct protein clusters. The final dataset consisted of 9,959 unique protein sequences, distributed in Table 1.

Table 1 Dataset Distribution after Quality Filtering and Homology Reduction

Partition	Nucleus	Cytoplasm	Mitochondrion	Membrane	Total
Training	1,812	1,804	1794	1801	7211
Validation	180	180	240	177	777
Test	506	505	454	506	1971
Overall	2498	2489	2488	2484	9959

Benchmark of Novel Proteins

To rigorously assess generalizability beyond standard cross-validation, we constructed a prospective

benchmark of novel proteins released in UniProtKB after January 1, 2024. Unlike standard static benchmarks, this dataset represents an "open-world" evaluation, capturing sequences deposited after the model's training cutoff.

The benchmark was retrieved via the UniProt REST API using a strict Boolean query strategy to ensure label purity and taxonomic consistency.

(i) Temporal Cutoff: date_created:[2024-01-01 TO *] (Capturing all entries from Jan 1, 2024, to the present).
(ii) Taxonomic Filter: taxonomy_id:2759 (Eukaryota) to exclude prokaryotic contaminants.
(iii) Label Exclusivity: To prevent ambiguity, we enforced mutual exclusivity using NOT operators. For example, the query for Nucleus explicitly excluded other compartments:

... AND (cc_scl_term:SL-0191) AND NOT (cc_scl_term:SL-0086) AND NOT (cc_scl_term:SL-0039).

To guarantee that performance metrics reflected genuine generalization rather than memorization, we likewise applied a strict homology reduction threshold of 40% against the training set. This rigorous filtering resulted in a high-quality but limited test set of 86 distinct proteins (40 Nucleus, 21 Cytoplasm, 20 Membrane, 5 Mitochondrion). The resulting test set was used to compare our proposed model with the established predictors, the DeepLoc 2.1 and MULocDeep.

Reproducibility and Deterministic Training

To ensure reproducibility, strict deterministic protocols were enforced across all computational pipelines. We fixed the global random seed to 42 for Python, NumPy, TensorFlow, and PyTorch operations, and disabled hash randomization (PYTHONHASHSEED=42). Furthermore, to ensure that performance metrics reflected architecture stability rather than lucky initialization, the statistical analysis was derived from five independent training runs where the model weights were re-initialized while keeping data splits constant.

Evolutionary Scale Modeling

We utilized ESM-2 (Evolutionary Scale Modeling) for feature extraction, a transformer-based protein language model developed by Lin et al. [11]. ESM-2 was trained via masked language modeling, approximately 250 million protein sequences drawn from UniRef [11], enabling it to learn high-dimensional representations of protein structure and function directly from amino acid sequences without requiring pre-computed features or multiple sequence alignments.

Model selection

For this study, we employed the esm2_t33_650M_UR50D checkpoint. This specific architecture consists of 33 transformer layers and approximately 650 million parameters. We selected this model because it provides a favorable trade-off between computational efficiency and representational power. It demonstrated performance comparable to larger, multi-billion-parameter models on downstream structure prediction and residue-level classification tasks [11].

Embedding generation

Feature extraction was performed in inference mode with the pre-trained ESM-2 model parameters frozen (torch.no_grad()). The extraction process followed a strict protocol:

a) Tokenization: Each protein sequence was tokenized using the standard ESM tokenizer.
b) Forward Pass: The sequences were passed through the model to generate hidden states for all layers.
c) Layer Selection: We extracted the last hidden state (Layer 33) to serve as the sequence representation, as the final layers of protein language models are known to capture the most contextually rich information relevant for prediction tasks.
d) Token Pruning: The start-of-sequence (<cls>) and end-of-sequence (<eos>) tokens were explicitly removed to align the embedding matrix strictly with the biological sequence.

This converted each protein sequence of length L into a dense matrix of size $L \times 1280$, where 1280 represents the embedding dimension of the 33-layer model. These embeddings were stored as memory-mapped arrays (numpy.memmap) to efficiently handle the large-scale high-dimensional data during the training of the downstream bidirectional LSTM network.

To maintain numerical stability while optimizing memory, we utilized FP16 solely for disk storage (via memory mapping) while strictly enforcing FP32 for all gradient computations, as full mixed-precision training resulted in numerical underflow.

a) Storage (Float16): Output embeddings were immediately cast to half-precision floating-point format (FP16) and serialized to disk using memory-mapped arrays (numpy.memmap). This reduced storage requirements by 50% (from ~15 GB to ~7.5 GB) and minimized I/O bottlenecks during batch loading

[18].

b) Computation (Float32): All model training operations, including gradient computation and weight updates, were performed in full precision (FP32). During the training phase, embeddings were cast from FP16 to FP32 in real-time using a "Just-In-Time" conversion within the data generator. This approach ensured numerical stability for the Bi-LSTM layers and prevented gradient underflow issues often associated with mixed-precision training [18].

While mixed-precision training (FP16 computation) can accelerate training on modern GPUs, we observed numerical instabilities (NaN loss values) during preliminary experiments with the mixed_float16 TensorFlow policy. These instabilities are well-documented for recurrent networks due to the accumulation of small gradients across long sequences [19]. Our hybrid approach, FP16 storage with FP32 computation, achieves memory efficiency without sacrificing training stability.

Model Architecture

The core classification pipeline was constructed as a hybrid deep-learning model to process variable-length embedding sequences generated by ESM-2. The architecture consists of four distinct stages: input masking, sequential feature extraction, attention-weighted aggregation, and final classification.

Input processing and masking

The model accepts an input tensor of shape $(B, L, 1280)$, where B is the batch size and $L = 600$ is the maximum sequence length. To handle variable-length protein sequences within fixed-size tensors, a Masking Layer was applied to ignore zero-padded time steps (mask_value=0.0), ensuring that padding tokens do not contribute to the loss gradients or attention weights.

Bidirectional sequence modeling

To capture long-range dependencies and the sequential context of amino acid residues, the masked embeddings were fed into a Bi-LSTM network [20, 21]. The layer consisted of 512 hidden units (256 per direction) with a dropout rate of 0.3 applied to the LSTM layer to prevent overfitting. This configuration allows the model to learn dependencies from both N-terminal and C-terminal directions simultaneously.

Context-aware attention mechanism

A custom Attention Mechanism was implemented to allow the model to focus on specific sorting signals (e.g., signal peptides or nuclear localization signals) rather than treating all residues equally [22].

a) Score Calculation: The LSTM hidden states h_t were passed through a dense layer with 64 units and a *tanh* activation function to compute a non-linear alignment score.
b) Weight Generation: These scores were projected to a scalar and normalized via a Softmax function to produce attention weights α_t.
c) Context Aggregation: The final context vector c was computed as the weighted sum of the hidden states:

$$c = \sum_{t=1}^{L} \alpha_t h_t \quad (1)$$

$$\alpha_t = softmax(W_a \cdot \tanh(W_s \cdot h_t + b_s)) \quad (2)$$

where c is the final context vector (aggregated representation), L represents the sequence length, and h_t is the hidden state vector of the Bi-LSTM at time step t. The variable α denotes the normalized attention weight for the t-th residue, computed using learnable weight matrices W_a and W_s, and the bias vector b_s. This mechanism not only improves classification performance but also provides interpretability by highlighting the residues most critical for localization.

Classification Head

The 512-dimensional context vector was processed by a multi-layer perceptron (MLP) classifier:

(i) Dense Layer 1: 512 units, ReLU activation, L2 Regularization $\lambda = 0.01$, followed by Dropout (0.5).
(ii) Dense Layer 2: 256 units, ReLU activation, L2 Regularization $\lambda = 0.01$, followed by Dropout (0.3).
(iii) Output Layer: A final dense layer with Softmax activation to output probabilities for the four target classes (Nucleus, Cytoplasm, Mitochondrion, Membrane).

Training strategy and focal loss

To address class imbalances and concentrate learning on hard-to-classify examples, specifically, nuclear proteins carrying short, degenerate Nuclear Localization Signals (NLS) that are difficult to distinguish from the cytoplasmic background, and cytoplasmic proteins with spurious hydrophobic patches that may superficially mimic transmembrane domains, the model was trained using Focal Loss (FL) [23], a technique successfully adapted for subcellular localization prediction of biological sequences to mitigate dataset skew

[24]. It should be noted that dual-localized proteins were explicitly excluded from the dataset via Boolean NOT filtering during data acquisition (see Data Acquisition). The focal loss was applied to address within-class ambiguity among single-label sequences, not multi-localization ambiguity.
The loss function was parameterized with a focusing parameter $\gamma = 2.0$ and a balancing factor $\alpha = 0.25$:

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma}\log(p_t) \quad (3)$$

where p_t is the model's predicted probability for the true class, γ is the focusing parameter that modulates the loss for well-classified examples, and α is the balancing factor introduced to address class imbalance.
Optimization was performed using the Adam optimizer with a learning rate of 1×10^{-4} and gradient clipping (clipnorm=1.0) to ensure stable convergence. Fig 1 shows the schematic representation of the proposed pipeline.

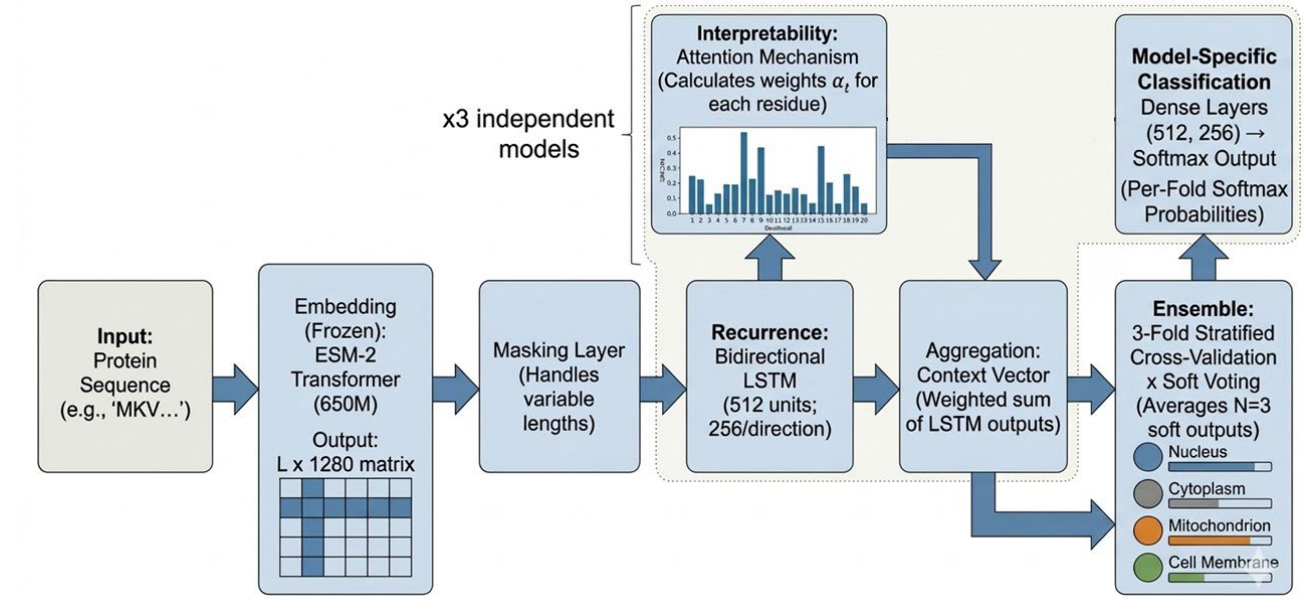

Fig 1 Schematic representation of the proposed pipeline. ESM-2 embeddings are extracted directly from primary amino acid sequences, enabling the ingestion of standard bioinformatics file formats such as FASTA. Bi-LSTM layers capture sequential dependencies. The attention mechanism assigns weights to key residues, and fully connected layers predict subcellular localisation.

Ensemble Strategy

To enhance model robustness and mitigate the variance often associated with deep learning models trained on biological sequences, we implemented a Stratified Cross-Validation Ensemble strategy [25]. Rather than relying on a single model, which may bias towards specific training subsets, we trained three independent instances of the Bi-LSTM architecture.

Training protocol

We utilized 3-Fold Stratified Cross-Validation, which partitions the training dataset into three distinct, non-overlapping folds while preserving the percentage of samples for each class. In each iteration, two folds were allocated for training (66.7%) and one-fold for validation (33.3%). This stratification ensures that every model observes a representative distribution of the minority and majority classes, preventing class imbalance bias during training [26].

To ensure optimal convergence and prevent overfitting, an adaptive training strategy was employed. The learning rate was dynamically adjusted using a ReduceLROnPlateau scheduler (factor=0.5, patience=4) whenever validation loss stagnated. Furthermore, EarlyStopping with a patience of 8 epochs to terminate training automatically was utilized when generalization performance plateaued. The final model for each fold was not taken from the last epoch but restored from the checkpoint with the lowest validation loss.

Prediction Aggregation (Soft Voting)

For the final inference on the independent test set, a probability averaging (soft voting) mechanism was adopted. Instead of using majority voting on discrete class labels, which discards confidence information, we

averaged the softmax probability distributions output by all N=3 models:

$$P_{final}(y|x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y|x) \quad (4)$$

where $P_{final}(y|x)$ is the aggregated probability for class y given input sequence x, N is the total number of independent models in the ensemble, and $P_i(y|x)$ represents the softmax probability vector output by the i-th model instance.

The final class label was determined by selecting the index with the maximum average probability. An averaging-based ensemble approach has been demonstrated to significantly improve accuracy and reduce overfitting in protein secondary structure prediction tasks compared to single-model approaches [27, 28].

Results

Quantifying Evolutionary Embedding Values

To isolate the performance advantage provided by ESM-2 embeddings, a classical machine learning baseline was evaluated through a support vector machine (SVM) with an RBF kernel, utilizing 3-mer (tripeptide) composition as the sole input feature. This serves as a technical control, restricting the input to "local chemical composition" (e.g., hydrophobicity counts) while strictly excluding sequential order and evolutionary context. The SVM baseline achieved a global test accuracy of 64.54% and an MCC of 0.5295. While superior to random guessing, the performance ceiling highlights the insufficiency of compositional features for complex localization tasks.

As detailed in Table 2, our proposed architecture (ESM-2 + Bi-LSTM) achieved a test accuracy of 86.81% (+22.27%) and an MCC of 0.825 (+0.296). The class-specific breakdown reveals exactly where embeddings provide value.

The Mitochondrial Leap (+0.35 Recall): The baseline struggled most with mitochondria (Recall: 0.56), likely because mitochondrial targeting peptides (mTPs) rely on amphipathic structural patterns rather than specific amino acid counts [29]. The Deep Learning model raised this to 0.91, proving that the Bi-LSTM successfully captures these distributed structural motifs.

The Membrane Improvement (+0.19 Recall): While membrane proteins are chemically distinct (hydrophobic) [30], the SVM only identified 71% of them. Our model improved this to 0.90, suggesting that the attention mechanism detects subtle or non-canonical transmembrane domains that simple composition statistics miss [31].

The improvement in cytoplasmic recall from 0.63 to 0.86 confirms that identifying the default cytoplasmic localization of many proteins requires a deep contextual understanding of the entire sequence, rather than relying solely on the detection of signal peptides [9].

Nuclear protein recall improved from 0.67 (SVM baseline) to 0.81 (ESM-2 + Bi-LSTM), representing a moderate gain of +0.14. This improvement was the smallest among all four classes, a pattern consistent with the biological nature of NLS. Unlike mitochondrial targeting peptides (mTPs) or transmembrane domains (TMDs), which are structurally well-defined and physicochemically distinct, NLS sequences are short (typically 7–20 residues), degenerate, and often embedded within longer disordered regions. The SVM's compositional features (3-mer counts) fail to capture the positional context required for NLS recognition, partially explaining the lower baseline. When the attention mechanism fails to sufficiently weight these short, degenerate NLS motifs against the broader sequence context, the model defaults to cytoplasm, the signal-absent class, consistent with the 16% nuclear-to-cytoplasm misclassification rate observed in the confusion matrix (Fig. 2).

Table 2 Quantification of the "Embedding Advantage" (Baseline vs. Proposed)

Metric / Class	SVM Baseline (3-mer)	Our Model (ESM-2)	Technical Gain
Global Accuracy	64.54%	86.81%	+22.27%
Global MCC	0.5295	0.8254	+0.2959
Membrane Recall	0.71	0.90	+0.19 (Strong Improvement)
Mitochondrion Recall	0.56	0.91	+0.35 (Massive Gain)
Cytoplasm Recall	0.63	0.86	+0.23 (Strong Improvement)
Nucleus Recall	0.67	0.81	+0.14 (Moderate Gain)

Model Performance

The proposed ensemble Deep Learning model was evaluated on the independent test set of 1,971 sequences, strictly isolated from the training process via cluster-aware partitioning using the internal test set. The model

achieved a final accuracy of 86.81% and a Matthews Correlation Coefficient (MCC) of 0.8254. We prioritized MCC as the critical metric [32] because, unlike accuracy, it remains robust against class imbalance. The observed MCC of 0.8254 confirms that the model has learned robust, non-linear decision boundaries and is not merely biasing its predictions toward the majority class. Fig. 2 presents the normalized confusion matrix, revealing distinct technical characteristics in how the model processes sorting signals.

Contrary to the baseline SVM, the deep learning model excelled at identifying Mitochondrial proteins (Recall: 0.91) and Membrane proteins (Recall: 0.90). The high performance on Mitochondria (up from 56% in the baseline to 91%) validates the Bi-LSTM's efficacy. mTPs are non-local and amphipathic, relying on distributed physicochemical patterns rather than strict sequence motifs [33, 34]. The Bi-LSTM successfully aggregates these distributed features into a coherent global context, enabling the attention mechanism to effectively "attend" to the N-terminal presequence.

The Nucleus class showed the lowest sensitivity (0.81), with a significant 16% of nuclear proteins misclassified as Cytoplasm. This represents a False Negative error likely driven by signal vanishing. Cytoplasmic localization is widely considered the default outcome for eukaryotic proteins that lack specific signals for other organelles. The model's failure to detect the NLS in these 16% of cases suggests that the attention mechanism occasionally fails to weight short, degenerate NLS motifs heavily enough against the global background of the protein sequence. When the NLS is missed, the model defaults to the signal-less class (Cytoplasm).

Cytoplasmic proteins were correctly identified 86% of the time. The confusion was distributed across Nucleus (7%) and Mitochondrion (6%). This represents a False Positive error or signal hallucination. In these cases, the model over-interpreted a random hydrophobic or basic patch in a cytoplasmic protein as a functional sorting signal (mTP or NLS). However, the precision remains high compared to the baseline, indicating that the ESM-2 embeddings provide a sufficiently high signal-to-noise ratio to minimize these hallucinations in most cases.

Global, structural signals (Membrane TMDs, Mitochondrial mTPs) are learned with greater than 90% reliability, while short, local motifs (Nuclear NLS) are slightly more prone to signal dropout (81%), defaulting to the cytoplasmic state. This confirms that the architecture is primarily driven by explicit signal detection rather than memorization.

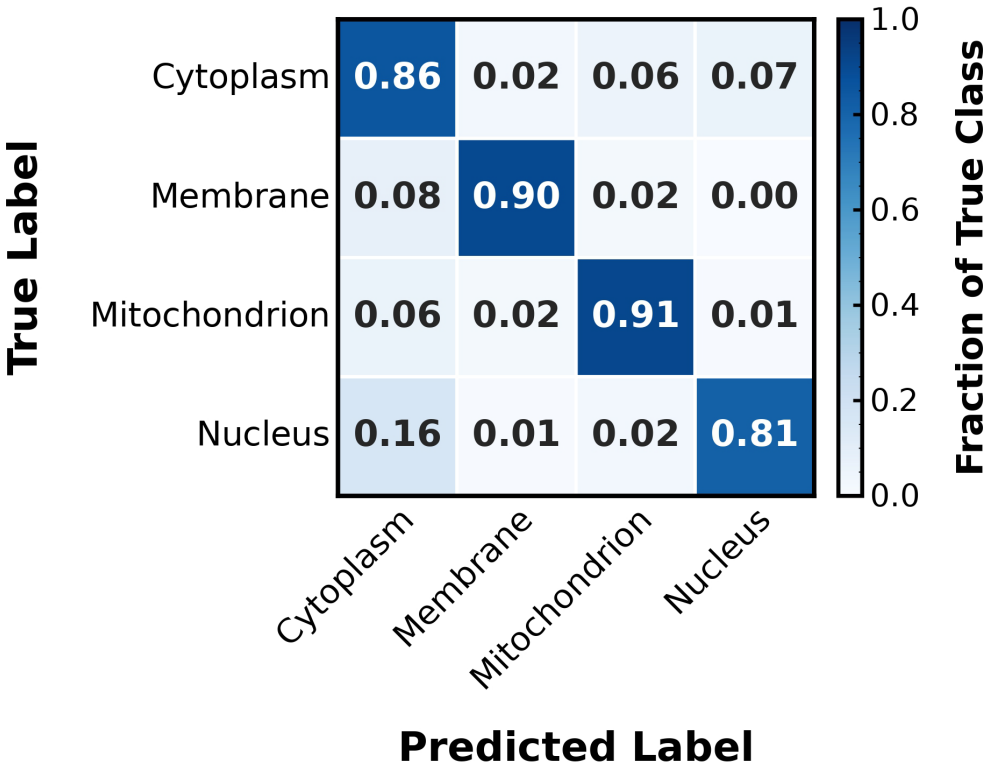

Fig 2 Normalized Confusion Matrix of the ensemble model. The matrix illustrates the model's per-class recall (sensitivity) on the independent test set. The model achieves peak performance on Mitochondrion (0.91) and Membrane (0.90) classes, demonstrating the architecture's ability to capture complex distributed signals. The primary error source is the misclassification of Nuclear proteins as Cytoplasm (0.16), indicating a tendency to default to the "null" localization when short nuclear signals are not strongly attended to.

Statistical Significance and Stability Analysis

To validate that the reported performance metrics were robust to stochastic weight initialization and not the result of a "lucky" seed, we conducted a rigorous stability analysis. While the data partition (CD-HIT clustering and GroupShuffleSplit) was held fixed to ensure a consistent benchmark, the entire model training and evaluation pipeline was executed five times independently (N=5 runs).

In each iteration, the Deep Learning model was re-initialized with a different random state and trained from scratch on the training fold. Through these five experiments, the architecture demonstrated exceptional stability with minimal variance. The mean accuracy was 85.09% (95% CI: [84.50%, 85.68%]), and the mean MCC was 0.8031 (95% CI: [0.7954, 0.8108]).

The narrow standard deviation ($\sigma \approx 0.0048$ for accuracy) indicates that the model converges to a consistent global minimum regardless of the initial weight distribution. The low variance across independent runs confirms the stability of the optimization convergence.

Table 3 Statistical Summary of 5 Independent Training Runs

Metric	Mean (μ)	Standard Deviation (σ)	95% Confidence Interval
Accuracy	85.09%	± 0.48%	(84.50%, 85.68%)
MCC	0.8031	± 0.0062	(0.7954, 0.8108)

Note: Confidence intervals were calculated using the t-distribution method ($t_{crit} = 2.776$), which is suitable for small sample sizes (N = 5).

Interpretability via Attention Mechanisms

To validate that the model's classification performance stems from learning genuine sequence features rather than statistical artifacts, we analyzed the internal attention weights (α_t) generated during inference. The attention mechanism provides a transparent view of the model's decision-making process by assigning a scalar importance score to each amino acid residue. Fig. 3 illustrates the attention profiles for two representative correctly classified proteins.

Case 1: C-Terminal Signal Localization (Membrane)

For the membrane protein CEAM1_RAT (sp|P16573, 519 amino acids), the attention mechanism exhibited a highly sparse and localized activation pattern. As shown in the attention map (Fig. 3A), the model assigned near-zero weights (≈ 0.00) to the first ~420 residues (~96% of the sequence), indicating these regions contributed negligibly to the final classification decision. A massive, sharp accumulation of attention mass (peak ≈ 0.074) was observed exclusively in the C-terminal tail (positions ~420–450). This pattern confirms that the Bi-LSTM learned to function as a precise anchor detector. Despite the length of the protein, the attention distribution indicates that the model assigns high weight exclusively to the C-terminus, consistent with findings that the C-terminal tail region harbors crucial signals for membrane protein insertion [35, 36].

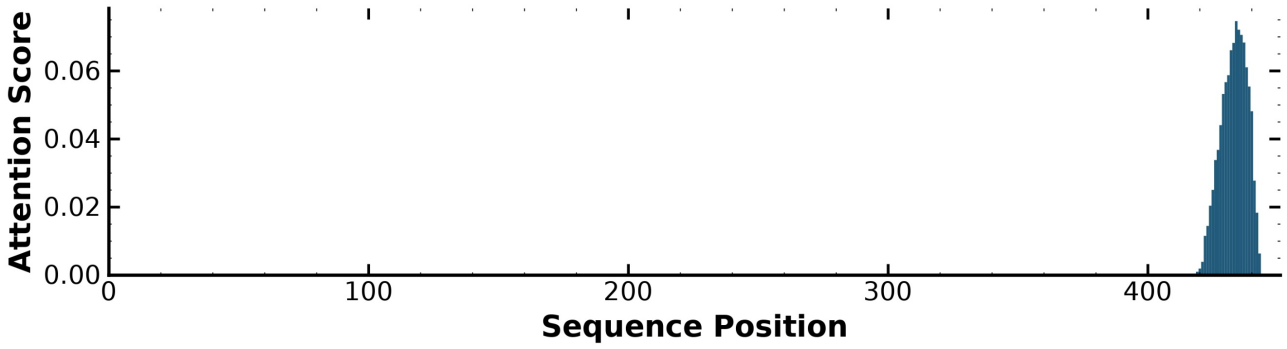

Fig 3A Membrane Protein (CEAM1_RAT): The model exhibits a sharp, localized attention peak (max ≈ 0.074) exclusively at the C-terminus (positions 420–470), effectively identifying the hydrophobic transmembrane anchor while suppressing the remainder of the sequence.

Case 2: N-Terminal Distributed Attention (Mitochondrion)

In contrast, the mitochondrial protein NDUAA_MOUSE (sp|Q99LC3, 355 amino acids) elicited a fundamentally different attention distribution (Fig. 3B), highlighting the model's architectural flexibility. Instead of a single sharp spike, the attention weights formed in the N-terminal region (positions ~5–28), which represents approximately 57% of this short protein sequence. The attention intensity rose from position 2–5,

peaked in the 10–20 range (≈ 0.058), and declined sharply after position 25, with the remainder of the sequence falling to near-zero.

This broad distribution indicates that the model did not search for a single key residue but rather integrated information across a contiguous window of ~25 amino acids. This pattern aligns closely with the properties of mitochondrial targeting sequences (MTS), which are typically 15–60 residues long and defined by distributed physicochemical properties, particularly amphipathicity, positive charge distribution, and the capacity to form amphipathic helices, rather than a strict consensus sequence [29, 37]. The higher model confidence (90.97%) may indicate that this distributed N-terminal signal provided more robust or less ambiguous features, though confidence scores can also reflect factors such as training data distribution and embedding quality.

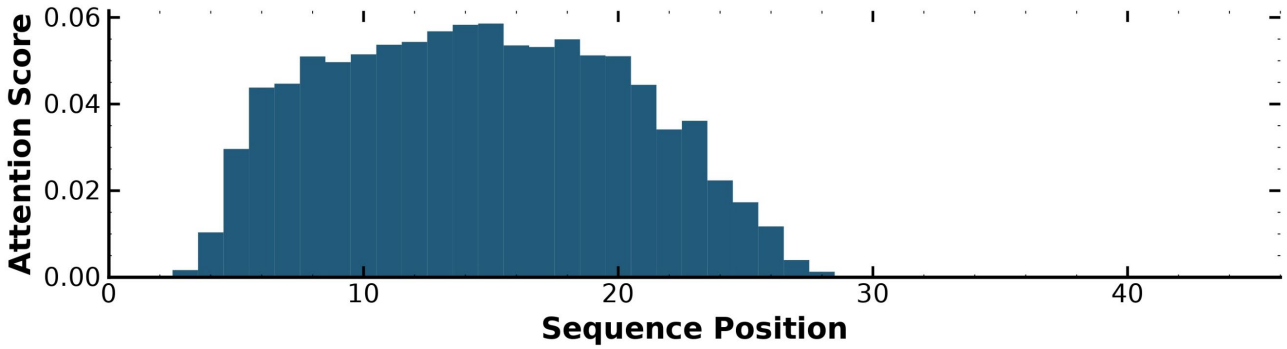

Fig 3B Mitochondrial Protein (NDUAA_MOUSE): In contrast, the model displays a broad, distributed attention profile across the N-terminal region (positions 3–28), consistent with the recognition of a non-motif-specific amphipathic mitochondrial targeting sequence (mTP).

The contrast between these two profiles, a sharp C-terminal spike for membrane proteins versus a broad N-terminal window for mitochondrial proteins, demonstrates that the attention mechanism is not applying a static heuristic. Instead, it dynamically adapts its receptive field based on the learned context of the ESM-2 embeddings.

This adaptive behavior validates that the model has learned biologically meaningful sequence features that correspond to known protein localization mechanisms, rather than exploiting spurious correlations or dataset biases. The attention weights provide interpretable evidence that the Bi-LSTM classifier operates as a context-aware feature integrator on top of the ESM-2 foundation, extracting location-specific signals that align with the established understanding of subcellular targeting pathways.

We acknowledge that these two examples represent qualitative observations rather than systematic validation. To address this limitation quantitatively, the top-1 attention peak of each test protein was cross-referenced against annotated biological motifs retrieved from UniProtKB, including signal peptides, transit peptides, topological domains, and regions of interest. The alignment rate was calculated as the fraction of proteins whose single highest attention weight position fell within a UniProt-annotated targeting feature. As shown in Fig. 4a, Nucleus proteins exhibited the highest top-1 peak alignment rate (~57%), well above the random chance baseline (~33%), followed by Cytoplasm proteins (~50% vs. ~32% random). Membrane and Mitochondrion proteins showed alignment rates at or below random expectation. Attention enrichment analysis (Fig. 4b) is consistent with these findings: Nucleus proteins showed the highest enrichment (approximately 2.35-fold), followed by Cytoplasm (approximately 2.0-fold), while Mitochondrion proteins showed sub-random enrichment (~0.35-fold). The high enrichment observed for Cytoplasm likely reflects the model attending to broadly annotated functional regions in UniProtKB, such as active sites, binding sites, and regions of interest, rather than canonical targeting sequences, which are absent in this class. The sub-random enrichment for Mitochondrion, despite high class recall (0.91), is consistent with the distributed physicochemical nature of mitochondrial targeting sequences, which are defined by amphipathicity and charge distribution across the N-terminal region rather than a discrete annotated motif [34]. A representative example, Membrane protein O35186 (Fig. 4c), illustrates a 21.2x individual enrichment; however, the class average enrichment for Membrane proteins (~1.1-fold) indicates substantial variability across the test set.

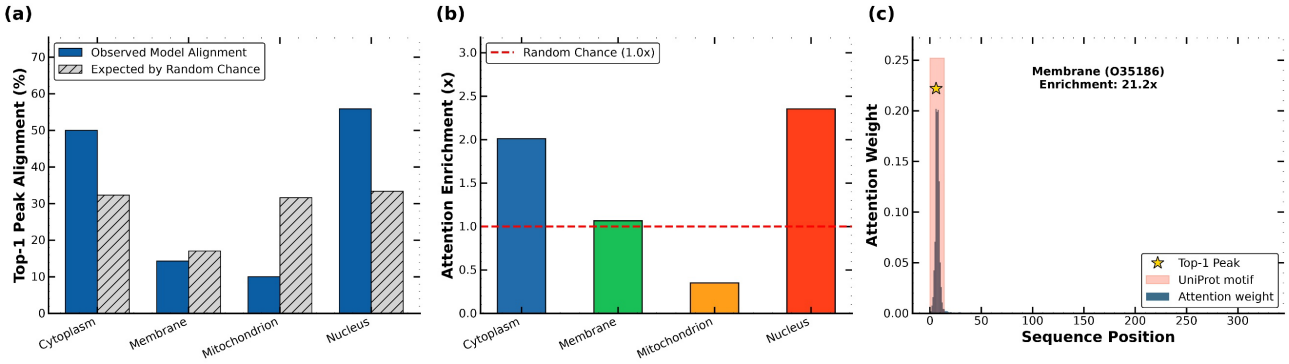

Fig 4 Quantitative alignment of model attention peaks with UniProt-annotated sequence features across the test set. **a** Top-1 peak alignment rate per subcellular class, showing the percentage of proteins whose highest attention weight position falls within a UniProt-annotated biological feature, compared against the expected rate under random attention assignment. **b** Attention mass enrichment ratio per class, defined as the fraction of total attention concentrated within UniProt-annotated regions relative to random expectation (dashed line at 1.0x). **c** Representative attention profile for Membrane protein O35186 (enrichment = 21.2x), illustrating precise colocalization of the dominant attention peak with the UniProt-annotated biological motif.

Discrimination Capability and Receiver Operating Characteristic Analysis

To evaluate the model's ability to separate each subcellular class from the background, we computed the Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) using a One-vs-Rest (OvR) strategy. As illustrated in Fig. 5, the model demonstrates high discriminative power across all four compartments, with all class-specific AUC values exceeding 0.94.

Class-Specific Discrimination Hierarchy:

Membrane (AUC = 0.989): This class exhibited the sharpest decision boundary. The curve shows a rapid initial ascent, reaching a True Positive Rate (TPR) of ~0.93 at a False Positive Rate (FPR) of only 0.05. This indicates that the embeddings for membrane proteins, likely driven by distinct transmembrane domains, form a tightly clustered manifold that is easily separable from other classes.

Mitochondrion (AUC = 0.981): The model maintained robust separation for mitochondrial proteins, achieving greater than 0.90 sensitivity at 95% specificity. The minimal drop in performance relative to the membrane class suggests that the distributed N-terminal signals (mTPs) are captured effectively by the Bi-LSTM despite their variable length.

Nucleus (AUC = 0.968) & Cytoplasm (AUC = 0.948): While still highly effective, these classes displayed a more gradual ascent in the ROC space. The Cytoplasm class, being the "default" localization with the least distinct sorting signals, required a higher FPR threshold (~0.10) to achieve the same sensitivity levels (greater than 0.90) observed for membrane proteins at stricter thresholds.

At a clinically relevant operating point of 5% False Positive Rate (95% Specificity), the model retains high sensitivity across the board: 95.85% for Membrane, 92.07% for Mitochondrion, 87.15% for Nucleus, and 75.45% for Cytoplasm. This confirms that the Deep Learning architecture effectively minimizes false alarms, a critical requirement for high-throughput proteomic screening where false positives can lead to costly experimental validation.

Feature Space Visualization

To validate that the model learned discriminative biological representations rather than memorizing training data, we visualized the 256-dimensional latent feature space using t-Distributed Stochastic Neighbor Embedding (t-SNE). As shown in Fig. 6, the topological structure of the embeddings directly correlates with the classification performance observed in the confusion matrix and ROC analysis.

The Membrane class forms the most distinct and compact cluster (isolated vertical column), confirming that the model effectively maps strong hydrophobic transmembrane signals into a unique, nearly uncontaminated region of the manifold. This geometric isolation explains the near-perfect AUC (0.989).

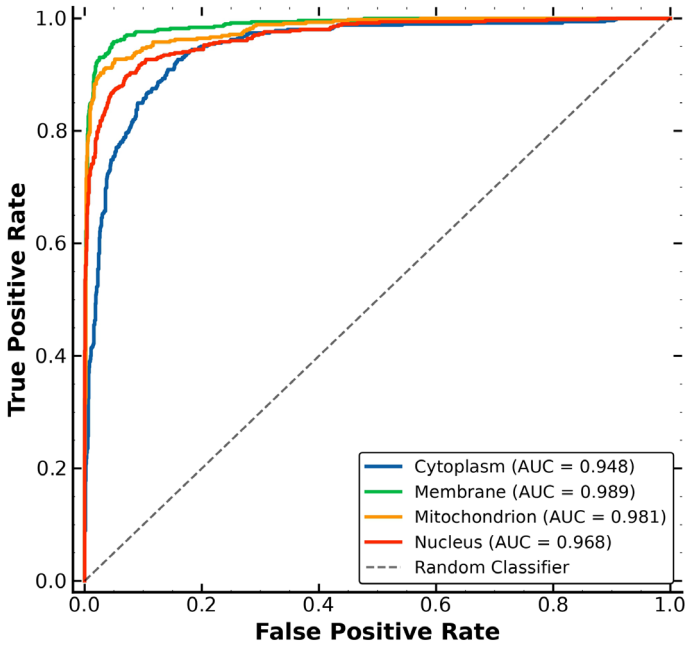

Fig 5 One-vs-Rest Receiver Operating Characteristic (ROC) analysis. Curves illustrate the trade-off between sensitivity and false positive rate for each subcellular class on the independent test set. The model demonstrates robust separation across all compartments (AUC > 0.94), with Membrane proteins (AUC=0.989) showing the highest discriminative separability and Cytoplasm (AUC=0.948) representing the most challenging boundary due to signal heterogeneity.

In contrast, the Cytoplasm class appears as a dispersed distribution with noticeable overlap against the Nucleus cluster. This lack of compactness technically validates the default nature of cytoplasmic localization. Because these proteins lack a unifying sorting signal, their latent representations are more heterogeneous [38, 39]. The observed overlap provides a structural explanation for the 16% misclassification rate of nuclear proteins, as distinct nuclear signals occasionally fade into this diffuse cytoplasmic background.

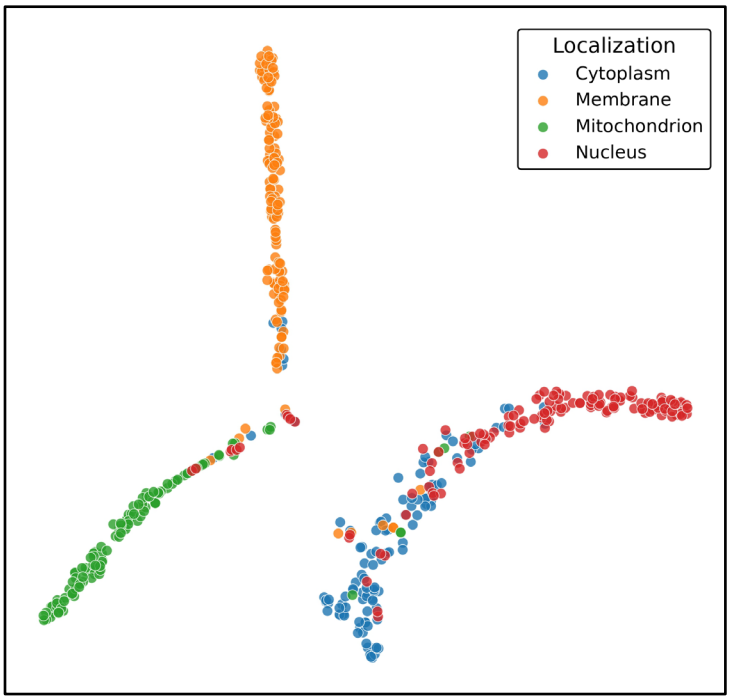

Fig 6 Visualization Projection of Learned Embeddings. The visualization reveals a hierarchical feature space where cluster compactness correlates with signal strength. Membrane proteins (Orange) and Mitochondria (Green) form tight, distinct clusters driven by strong sorting signals, while Cytoplasm (Blue) forms a dispersed cloud, reflecting its biological heterogeneity and overlap with the Nucleus (Red).

Prospective Evaluation on Novel Non-Homologous Proteins

In a preliminary evaluation against established predictors, DeepLoc 2.1[12] (the current state-of-the-art) and MULocDeep [13] (a comprehensive generalist tool), we benchmarked the model on the 86 novel sequences. For DeepLoc 2.1, we normalized the prediction space by considering only the probabilities of the four relevant compartments and assigning the label with the highest relative confidence. For MULocDeep, the standard predicted class labels was utilized as provided by the tool. On this specific dataset, the proposed ESM-2 + Bi-LSTM architecture observed a higher classification accuracy (88.37%, MCC = 0.827) compared to DeepLoc 2.1 (80.23%, MCC = 0.714) and MULocDeep (77.91%, MCC = 0.682).

While a paired t-test indicated statistical significance ($p = 0.007$ vs. DeepLoc 2.1; $p = 0.019$ vs. MULocDeep), we interpret these results with caution due to the limited sample size. The performance gap suggests that our specialized 4-class architecture offers a statistically significant performance advantage over state-of-the-art generalist tools for this specific subset of eukaryotic proteins, though larger-scale benchmarks would be required to establish definitive superiority.

Table 4 Performance Comparison on Prospective Benchmark (N=86)

Model	Backbone	Accuracy	MCC	Significance
Our Model	ESM-2 (650M)	88.4%	0.827	—
DeepLoc 2.1	ProtT5	80.2%	0.714	p = 0.0074
MULocDeep	BiLSTM+Attention	77.9%	0.682	p = 0.0192

To address the interpretive limitations imposed by the benchmark sample size (N = 86), a post-hoc power analysis was conducted for each head-to-head comparison using a paired t-test framework with a two-sided significance level of 0.05 and Cohen's *d* as the effect size measure (Fig. 7).

For the comparison against DeepLoc 2.1 (Fig. 7a), the analysis revealed an observed statistical power of 1.00 (100%), with a minimum sample size of only N = 19 required to achieve 80% power at the observed effect size. This result confirms that the performance advantage over DeepLoc 2.1 ($p = 0.0074$) supported by the current benchmark, and that the effect size is large enough that the conclusion would hold even with a substantially smaller dataset.

The comparison against MULocDeep (Fig. 7b) produced a different picture. At N = 86, the observed power was 0.66 (66%), falling below the conventional 80% threshold. A sample size of approximately N = 120 would be required to achieve adequate power at the observed effect size. While the paired t-test remained statistically significant ($p = 0.0192$), the underpowered comparison means there is an elevated risk of a Type II error, and the reported advantage over MULocDeep should therefore be interpreted with greater caution. Future work with an expanded prospective benchmark is warranted to confirm this result.

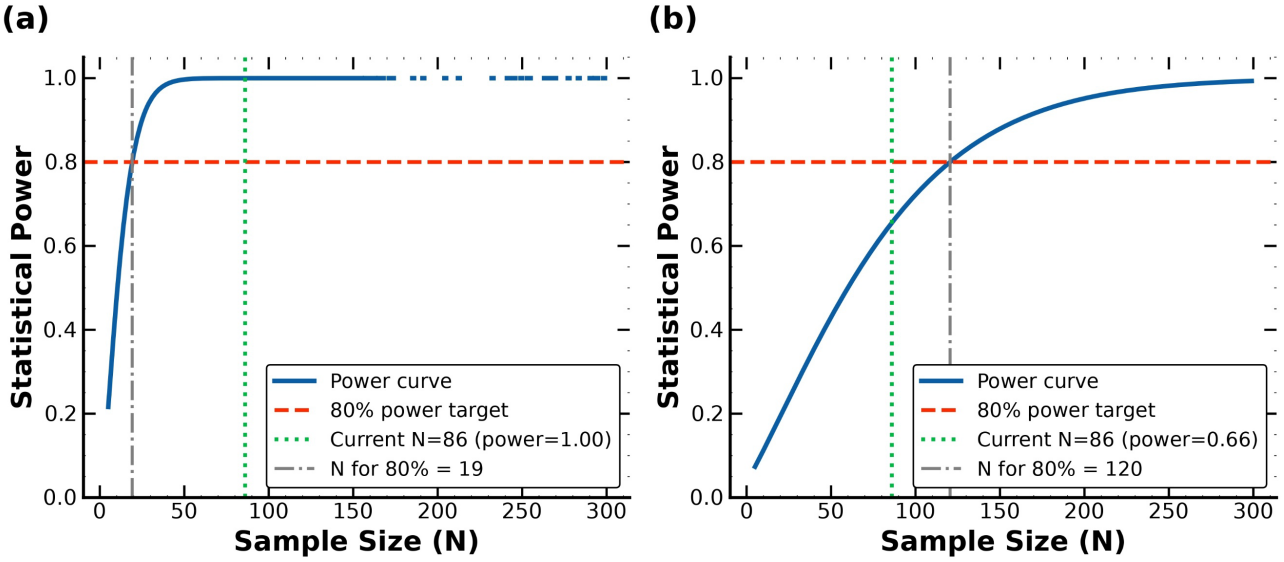

Fig 7 Post-hoc statistical power curves for head-to-head benchmark comparisons (paired t-test, two-sided, $\alpha = 0.05$). **a.** Comparison against DeepLoc 2.1 (observed power = 1.00; minimum adequate N = 19). **b.** Comparison against MULocDeep (observed power = 0.66; minimum adequate N = 120).

Discussions

The Feasibility of Streamlined Architectures

Present protocol suggests that, for targeted classification tasks, a streamlined Bi-LSTM network coupled with ESM-2 embeddings can achieve performance comparable to that of more complex architectures. In our internal testing, the model effectively decoded evolutionary signals without requiring the extensive distinct layers found in generalist predictors. This suggests that the richness of the ESM-2 embedding space may reduce the need for highly complex downstream classifiers, potentially lowering the barrier to developing custom proteomic tools.

Performance on Novel Sequences

On the prospective benchmark of 86 novel proteins released in the year 2024, the present model demonstrated encouraging generalization capabilities, achieving an accuracy of 88.37%. It is again emphasized that while the paired t-test indicates statistical significance ($p=0.007$), the limited sample size ($N=86$) and severe class imbalance (Mitochondrion: $N=5$) constrain the generalizability of these findings. A single misclassification in the mitochondrial class would result in a 20% change in class-specific recall. Therefore, the present model is preliminary evidence of competitive performance rather than definitive superiority. The true value of this approach lies in its architectural simplicity and interpretability, which is demonstrated on the larger internal test set ($N=1,971$), rather than claiming state-of-the-art performance on this small prospective benchmark.

Interpretability Findings

A key advantage of the attention-based architecture is its transparency. As visualized in the attention maps (Fig. 3), the model appeared to focus on biologically relevant regions, specifically, the C-terminal tails for membrane proteins and distributed N-terminal regions for mitochondrial precursors. These patterns align with established biological sorting rules, providing qualitative confidence that the model is leveraging genuine sequence motifs rather than learning statistical artifacts, although these interpretations are qualitative.

Limitations

The present study was developed under the assumption that each protein resides at one, and only one, subcellular compartment, and thus cannot handle multi-location proteins that may simultaneously exist at, or move between, two or more sites. This restriction was enforced during data curation by applying explicit Boolean NOT filtering across all SL identifiers, ensuring label purity and preventing annotation ambiguity. However, this approach excludes a biologically significant fraction of the proteome, as more than one-third of eukaryotic proteins are known to exhibit multiple subcellular localizations [40]. Such multiplex proteins are biologically important and should not be ignored; yet, consistent with several existing tools and benchmarks [9, 41, 42], the present framework was constructed exclusively for single-location proteins. Consequently, the reported accuracy metrics reflect performance on a single-localization subpopulation and may not generalize to proteins with multiple localizations [9, 43]. A series of methods have addressed this by framing subcellular localization as a multi-label problem across humans, plants, viruses, bacteria, and eukaryotes, often reporting improved performance when multi-location proteins are explicitly modeled [43–47]. Recent deep and multimodal models, such as DeepLoc 2.0, ProLoc-IHS, and several IHC-image-based predictors, employ multi-label architectures with independent outputs per location. These models were trained with binary cross-entropy or focal loss-based objectives for each label, establishing multi-label learning as a central direction for future work in subcellular localization prediction [9, 46]. Extending the present framework to this paradigm, replacing the softmax output layer with independent sigmoid units per compartment, remains a primary goal for future work.

While the model demonstrates that a streamlined Bi-LSTM architecture performs effectively with ESM-2 embeddings, it did not systematically evaluate alternative design choices due to computational constraints. Specifically, was not tested: (1) alternative protein language models (e.g., ProtT5, ProtBERT, Ankh), (2) different ESM-2 layer selections (we used Layer 33 but did not compare to intermediate layers), (3) architectural variants such as removing the attention mechanism or varying LSTM depth, or (4) fine-tuning ESM-2 versus using frozen embeddings. Therefore, the proposed model of 'ESM-2 embeddings are sufficient' to be interpreted as 'ESM-2 Layer 33 embeddings with this specific architecture perform well,' rather than a systematic comparison across the design space. These ablations remain important directions for future work to fully characterize each component's contribution.

Conclusion

In this study, the utility of combining ESM-2 language model embeddings with a Bi-LSTM classifier for protein subcellular localization was explored. Through a rigorous evaluation on strictly non-homologous proteins released in 2024, it was observed that:

1. The proposed architecture achieved an accuracy of 88.4% on the prospective dataset, demonstrating that it can generalize to novel sequences within the four targeted classes.
2. The study suggests that pre-trained evolutionary embeddings (ESM-2) can enable simpler, resource-efficient models to perform effectively, potentially offering a viable alternative to heavier architectures for specific tasks.
3. The model's attention mechanism highlighted sequence regions consistent with known sorting signals (transmembrane domains and signal peptides), suggesting the network successfully captured relevant biological features.

The present study suggests that established deep learning architectures combining pre-trained protein language models with attention-based bidirectional recurrent layers may offer a computationally efficient alternative to complex multi-stage predictors for targeted classification tasks. However, this study represents a single point in the architectural design space. We focused on a single well-specified configuration (ESM-2 Layer 33 + Bi-LSTM + attention) to assess whether this streamlined approach can achieve competitive performance. We demonstrate the feasibility of this approach and provide a rigorous temporal validation methodology that can be adopted in future studies, not architectural optimality.

Acknowledgment
The author acknowledges the UniProt Consortium for maintaining the publicly accessible UniProt Knowledgebase and Meta AI for releasing the ESM-2 protein language model under an open-source license. All computational analyses were conducted using Google Colab.

Abbreviations
AUC: Area Under the Curve/Eğri Altında Kalan Alan, Bi-LSTM: Bidirectional Long Short-Term Memory/Çift Yönlü Uzun Kısa Süreli Bellek, CD-HIT: Cluster Database at High Identity with Tolerance/Toleranslı Yüksek Benzerlikte Küme Veritabanı, ESM: Evolutionary Scale Modeling/Evrimsel Ölçekli Modelleme, FL: Focal Loss/Odak Kaybı, FP16: Half-Precision Floating-Point/Yarı Hassasiyetli Kayan Nokta, FP32: Full Precision Floating-Point/Tam Hassasiyetli Kayan Nokta, FPR: False Positive Rate/Yanlış Pozitif Oranı, MCC: Matthews Correlation Coefficient/Matthews Korelasyon Katsayısı, MLP: Multi-Layer Perceptron/Çok Katmanlı Algılayıcı, mTP: Mitochondrial Targeting Peptide/Mitokondriyal Hedefleme Peptidi, MTS: Mitochondrial Targeting Sequence/Mitokondriyal Hedefleme Dizisi, NLS: Nuclear Localization Signal/Nükleer Lokalizasyon Sinyali, OvR: One-vs-Rest/Biri-Diğerlerine-Karşı, PLM: Protein Language Model/Protein Dil Modeli, ReLU: Rectified Linear Unit/Doğrultulmuş Doğrusal Birim, ROC: Receiver Operating Characteristic/Alıcı İşletim Karakteristiği, SL: Subcellular Location/Hücre Altı Konumu, SVM: Support Vector Machine/Destek Vektör Makinesi, TMD: Transmembrane Domain/Transmembran Alanı, TPR: True Positive Rate/Doğru Pozitif Oranı, t-SNE: t-Distributed Stochastic Neighbor Embedding/t-Dağılımlı Stokastik Komşuluk Gömme, UniProtKB: UniProt Knowledgebase/UniProt Bilgi Tabanı.

Funding
The author did not receive support from any organization for the submitted work.

Data Availability statement
All protein sequences used in this study were retrieved from the UniProt Knowledgebase (UniProtKB; https://www.uniprot.org) using the Boolean query strategy and SL controlled vocabulary identifiers described in the Materials and Methods section (SL-0191, SL-0086, SL-0173, SL-0039). The curated sequence accession list, class labels, and train/validation/test split assignments, and the 86-protein prospective benchmark sequences (N=86, released in UniProtKB after January 1, 2024) are publicly available in the project repository at https://github.com/johmar-22/protein-subcellular-localization. ESM-2 embeddings can be regenerated from the provided accession lists using the code in the repository and the publicly available ESM-2 model (facebook/esm2_t33_650M_UR50D; https://github.com/facebookresearch/esm). Pre-trained fold checkpoints and the SVM baseline model are available as release assets in the same repository. All analysis code is available at the repository link above.

Compliance with ethical standards
Conflict of interest
The author declares no conflict of interest.

Ethical standards
The study is proper with ethical standards.

References

1. Villanueva, E., et al., System-wide analysis of RNA and protein subcellular localization dynamics. Nat Methods, 2024. 21: p. 60–71. Doi: 10.1038/s41592-023-02101-9
2. Martinez-Val, A., et al., Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution. Nat Commun, 2021. 12: p. 7113. Doi: 10.1038/s41467-021-27398-y
3. Pasha, T., et al., Karyopherin abnormalities in neurodegenerative proteinopathies. Brain, 2021. 144: p. 2915–2932. Doi: 10.1093/brain/awab201
4. Calabrese, G., C. Molzahn, and T. Mayor, Protein interaction networks in neurodegenerative diseases: From physiological function to aggregation. Journal of Biological Chemistry, 2022. 298: p. 102062. Doi: 10.1016/j.jbc.2022.102062
5. Lavoie, H., J. Gagnon, and M. Therrien, ERK signalling: a master regulator of cell behaviour, life and fate. Nat Rev Mol Cell Biol, 2020. 21: p. 607–632. Doi: 10.1038/s41580-020-0255-7
6. Jiang, Y., et al., Exploring potential therapeutic targets for asthma: a proteome-wide Mendelian randomization analysis. J Transl Med, 2024. 22: p. 978. Doi: 10.1186/s12967-024-05782-8
7. Shen, Y., et al., Critical evaluation of web-based prediction tools for human protein subcellular localization. Brief Bioinform, 2020. 21. Doi: 10.1093/bib/bbz106
8. Xiao, H., et al., A Review for Artificial Intelligence Based Protein Subcellular Localization. Biomolecules, 2024. 14: p. 409. Doi: 10.3390/biom14040409
9. Thumuluri, V., et al., DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res, 2022. 50. Doi: 10.1093/nar/gkac278
10. Teufel, F., et al., SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol, 2022. 40. Doi: 10.1038/s41587-021-01156-3
11. Lin, Z., et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (1979), 2023. 379. Doi: 10.1126/science.ade2574
12. Ødum, M.T., et al., DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. Nucleic Acids Res, 2024. 52: p. W215–W220. Doi: 10.1093/nar/gkae237
13. Jiang, Y., et al., MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. Comput Struct Biotechnol J, 2021. 19. Doi: 10.1016/j.csbj.2021.08.027
14. Bateman, A., et al., UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res, 2025. 53. Doi: 10.1093/nar/gkae1010
15. Bateman, A., et al., UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res, 2023. 51. Doi: 10.1093/nar/gkac1052
16. Bateman, A., et al., UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res, 2021. 49. Doi: 10.1093/nar/gkaa1100
17. Li, W. and A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006. 22. Doi: 10.1093/bioinformatics/btl158
18. Narang, S., et al., Mixed precision training, in 6th International Conference on Learning Representations, ICLR 2018 – Conference Track Proceedings. 2018.
19. Hochreiter, S., The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowldege-Based Systems, 1998. 6. Doi: 10.1142/S0218488598000094
20. Hochreiter, S. and J. Schmidhuber, Long Short-Term Memory. Neural Comput, 1997. 9. Doi: 10.1162/neco.1997.9.8.1735
21. Schuster, M. and K.K. Paliwal, Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997. 45: p. 2673–2681. Doi: 10.1109/78.650093
22. Bahdanau, D., K.H. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, in 3rd International Conference on Learning Representations, ICLR 2015 – Conference Track Proceedings. 2015.
23. Lin, T.Y., et al., Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell, 2020. 42. Doi: 10.1109/TPAMI.2018.2858826
24. Deng, Y., J. Jia, and M. Yi, EDCLoc: a prediction model for mRNA subcellular localization using improved focal loss to address multi-label class imbalance. BMC Genomics, 2024. 25: p. 1252. Doi: 10.1186/s12864-024-11173-6
25. Dietterich, T.G., Ensemble methods in machine learning, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2000.
26. Kohavi, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in IJCAI International Joint Conference on Artificial Intelligence. 1995.
27. Vignesh, U., R. Parvathi, and K. Gokul Ram, Ensemble deep learning model for protein secondary structure prediction using NLP metrics and explainable AI. Results in Engineering, 2024. 24. Doi: 10.1016/j.rineng.2024.103435
28. Yuan, L., Y. Ma, and Y. Liu, Ensemble deep learning models for protein secondary structure prediction using bidirectional temporal convolution and bidirectional long short-term memory. Front Bioeng Biotechnol, 2023. 11. Doi: 10.3389/fbioe.2023.1051268
29. Boob, A.G., et al., Design of diverse, functional mitochondrial targeting sequences across eukaryotic organisms using variational autoencoder. Nature Communications, 2025. 16. Doi: 10.1038/s41467-025-59499-3
30. Zhou, Z., et al., Advances in solubilization and stabilization techniques for structural and functional studies of membrane proteins. PeerJ, 2025. 13.
31. He, L. and X. Liu, The Development and Progress in Machine Learning for Protein Subcellular Localization Prediction. Open Bioinforma J, 2022. 15. Doi: 10.2174/18750362-v15-e2208110
32. Chicco, D. and G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 2020. 21: p. 6. Doi: 10.1186/s12864-019-6413-7
33. Gavel, Y., L. Nilsson, and G. von Heijne, Mitochondrial targeting sequences why “non-amphiphilic” peptides may still be amphiphilic. FEBS Lett, 1988. 235. Doi: 10.1016/0014-5793(88)81257-2

34. Calvo, S.E., et al., Comparative analysis of mitochondrial N-termini from mouse, human, and yeast. Molecular and Cellular Proteomics, 2017. 16. Doi: 10.1074/mcp.M116.063818
35. Kalinin, I.A., et al., Features of membrane protein sequence direct post-translational insertion. Nature Communications, 2024. 15. Doi: 10.1038/s41467-024-54575-6
36. Sun, S. and M. Mariappan, C-terminal tail length guides insertion and assembly of membrane proteins. Journal of Biological Chemistry, 2020. 295. Doi: 10.1074/jbc.RA120.012992
37. Reed, A.L., et al., Interactions of amyloidogenic proteins with mitochondrial protein import machinery in aging-related neurodegenerative diseases. Front Physiol, 2023. 14. Doi: 10.3389/fphys.2023.1263420
38. Hein, M.Y., et al., Global organelle profiling reveals subcellular localization and remodeling at proteome scale. Cell, 2025. 188: p. 1137–1155.e20. Doi: 10.1016/j.cell.2024.11.028
39. Zhang, X., et al., Prediction of protein subcellular localization in single cells. Nat Methods, 2025. 22: p. 1265–1275. Doi: 10.1038/s41592-025-02696-1
40. Briesemeister, S., J. Rahnenführer, and O. Kohlbacher, Going from where to why-interpretable prediction of protein subcellular localization. Bioinformatics, 2010. 26. Doi: 10.1093/bioinformatics/btq115
41. Wan, B., et al., EGA-Ploc: An Efficient Global-Local Attention Model for Multi-Label Protein Subcellular Localization Prediction on the Immunohistochemistry Images. IEEE J Biomed Health Inform, 2025. 29: p. 9299–9311. Doi: 10.1109/JBHI.2025.3613205
42. Zhang, Q., et al., MpsLDA-ProSVM: Predicting multi-label protein subcellular localization by wMLDAe dimensionality reduction and ProSVM classifier. Chemometrics and Intelligent Laboratory Systems, 2021. 208: p. 104216. Doi: 10.1016/j.chemolab.2020.104216
43. Wang, F. and L. Wei, Multi-scale deep learning for the imbalanced multi-label protein subcellular localization prediction based on immunohistochemistry images. Bioinformatics, 2022. 38: p. 2602–2611. Doi: 10.1093/bioinformatics/btac123
44. Guo, X., et al., Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier. Sci Rep, 2016. 6: p. 28087. Doi: 10.1038/srep28087
45. Sahu, S.S., C.D. Loaiza, and R. Kaundal, Plant-mSubP: a computational framework for the prediction of single- and multi-target protein subcellular localization using integrated machine-learning approaches. AoB Plants, 2020. 12. Doi: 10.1093/aobpla/plz068
46. Liu, F., S. Xin, and Y. Liu, ProLoc-IHS: Multi-label protein subcellular localization based on immunohistochemical images and sequence information. Int J Biol Macromol, 2025. 313: p. 144096. Doi: 10.1016/j.ijbiomac.2025.144096
47. Zhang, Q., et al., Accurate prediction of multi-label protein subcellular localization through multi-view feature learning with RBRL classifier. Brief Bioinform, 2021. Doi: 10.1093/bib/bbab012