3D Structural Biology Hackathon – Report

Sivan mor Yosef – 212877468

Yair Mahfud - 207807082

Lior Zatz – 209226448

Yuval Cohen - 208305052

Introduction

In our previous work, we developed a binary classifier for Nuclear Export Signal (NES) prediction using ESM protein language model embeddings trained on a dataset of approximately 500 peptide sequences. The model was designed to distinguish between peptides containing NES motifs (positive class) and those lacking such signals (negative class), utilizing the rich contextual representations provided by ESM-2 embeddings to capture the complex sequence patterns characteristic of nuclear export signals. This peptide-level classification approach provided an initial framework for examining NES recognition at the local sequence level, demonstrating the potential of transformer-based protein language models for this specific biological prediction task.

The motivation for developing such a classifier extends beyond individual peptide classification toward the more practical and biologically relevant goal of protein-level NES prediction. In biological systems, there is typically greater interest in determining whether an entire protein contains nuclear export capability rather than focusing on isolated peptide fragments. Protein-level classification enables more meaningful biological interpretations and practical applications, such as predicting protein subcellular localization, understanding nuclear-cytoplasmic transport mechanisms, and identifying proteins that may be involved in disease processes. This shift from peptide-level to protein-level prediction represents a crucial step toward developing clinically and research-relevant tools for studying nuclear transport biology.

The biological significance of accurate protein-level NES prediction lies in its direct relevance to CRM1 nuclear export, a critical cellular transport mechanism. Proteins containing NES motifs are recognized and exported from the nucleus by CRM1, and disruption of this pathway is implicated in cancer where altered nuclear-cytoplasmic transport contributes to oncogenesis and therapeutic resistance. Beyond binary classification, our approach enables the identification of specific NES positions within protein sequences, providing detailed spatial information about where these critical motifs occur. This positional information is invaluable for understanding structure-function relationships, designing targeted interventions, predicting the effects of mutations on nuclear export efficiency, and developing therapeutic strategies that modulate specific transport pathways while preserving normal cellular function.

However, our current peptide-level classifier faces several important limitations that must be addressed for reliable protein-level applications. The training dataset, while carefully curated, remains relatively small with approximately 500 examples, which may limit the model's ability to capture the full diversity of NES motifs and their contextual variations across different protein families. More critically, the negative examples in our dataset were generated through random sampling rather than being derived from experimentally validated non-NES sequences, potentially creating a weaker and less representative negative class that may not reflect the true complexity of non-NES peptides found in natural protein sequences. Additionally, the transition from peptide-level to protein-level classification introduces a significant challenge in terms of prediction reliability, as each protein prediction requires aggregating hundreds of individual peptide classifications. This aggregation process amplifies any errors or uncertainties in the underlying peptide classifier, making the demand for highly accurate peptide-level predictions even more critical when the ultimate goal is confident protein-level classification.

Methods

In this project, we aimed to develop a robust and systematic workflow to achieve accurate classification of proteins containing nuclear export signals (NES). Throughout the process, we encountered various challenges due to the scale and complexity of the data, and we continuously refined our approach to address them.

We began by working with a dataset of 4,000 proteins — a substantial volume, particularly considering the need to segment each protein into its corresponding peptides. The first phase involved thorough pre-processing of the dataset: extracting peptides associated with each protein, and accurately labeling these peptides according to their true classification (presence or absence of NES elements).

Following this data preparation, we proceeded to the evaluation stage. Our strategy was to advance the model by iteratively analyzing its performance and predictive accuracy. The processing pipeline involved several key steps:

segmenting proteins into peptides,

generating embeddings for each peptide using the ESM-2 protein language model,

applying a scoring function with a dynamic threshold to produce predictions,

and computing overall classification accuracy at the protein level.

During this phase, we also conducted an additional experiment: we introduced a modification to control the number of peptides required for a protein to be classified as positive (i.e., containing NES). This adjustment was motivated by the observation that some proteins contained a disproportionately large number of peptides, which could bias the classification outcome. By imposing a limit on the minimum number of peptides needed to assign a positive label, we improved classification stability. Although this step was successful, it is considered more of an experimental adjustment rather than a core component of our methods.

A key part of our overall approach was the iterative refinement of the training data. After each evaluation cycle, we enhanced the dataset — in particular, by curating improved sets of negative examples — in order to create a more balanced and representative training set. This process helped the model generalize better and achieve higher accuracy over time.

The final objective of our workflow is to deploy the optimized model on a larger dataset of 20,000 proteins, enabling us to determine, with improved reliability, whether each protein contains NES elements or not. The design of this workflow — centered around evaluation, iterative learning, and correction — provides a well-balanced framework for progressing toward this goal.

Experiments

**NES Prediction Pipeline Overview**

The following plots summarize the key stages and evaluation metrics of our Nuclear Export Signal (NES) prediction pipeline:

* **Dataset Preparation:** We began with a dataset of approximately 4,000 proteins experimentally annotated for NES content. Each protein sequence was segmented into overlapping peptides using a sliding window of 22 amino acids (stride = 1). Ambiguous peptides were excluded to ensure high-confidence supervision, resulting in a labeled dataset of NES-positive and NES-negative peptides.
* **Feature Extraction:** Each peptide was embedded using the ESM-2 protein language model (layer 9), producing a 2,560-dimensional vector representation per peptide.
* **Peptide Classification:** A neural network classifier (SimpleDenseNet) was trained to predict NES presence at the peptide level based on these embeddings.
* **Protein-Level Prediction:** Protein classification was initially determined using a simple rule: if **any** peptide within a protein was classified as NES-positive, the entire protein was labeled as NES-containing.
* **NES Localization:** Predicted NES regions were extracted by mapping the coordinates of positively predicted peptides back to their original positions within the protein sequence.

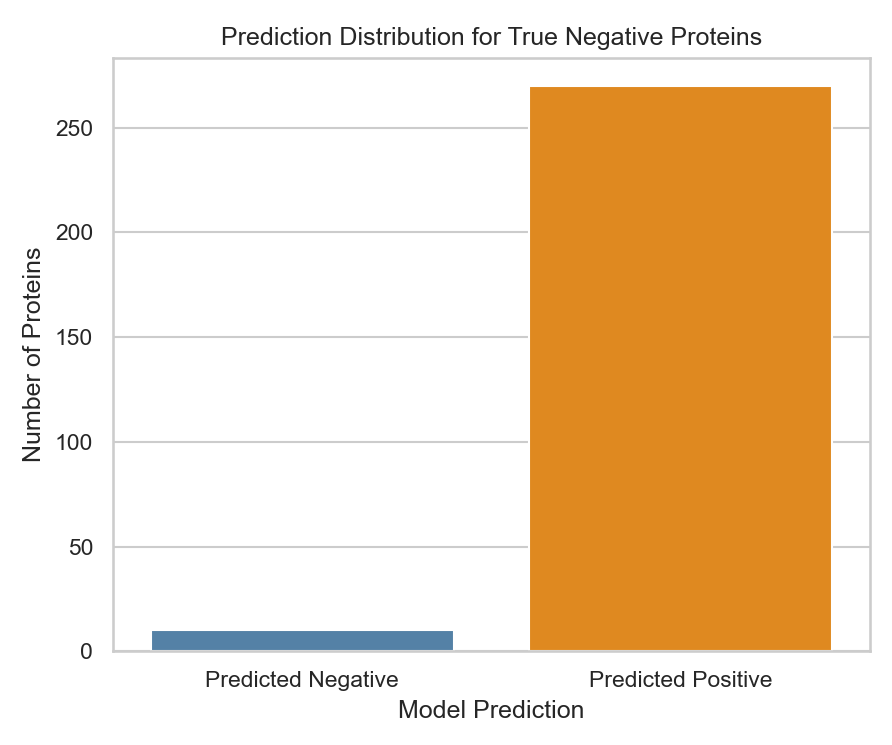
These visualizations demonstrate the model’s ability to detect NES-containing proteins and accurately localize NES regions with high sensitivity.

🟦 Figure 1: **Prediction Distribution for True Positive Proteins**

A graph with a bar and a number of numbers

AI-generated content may be incorrect.

This plot displays the model’s performance on proteins that are truly NES-positive. Here, the model performs **very well**, correctly identifying the vast majority of them. Only a small number of false negatives are observed, indicating **high recall (sensitivity)**.

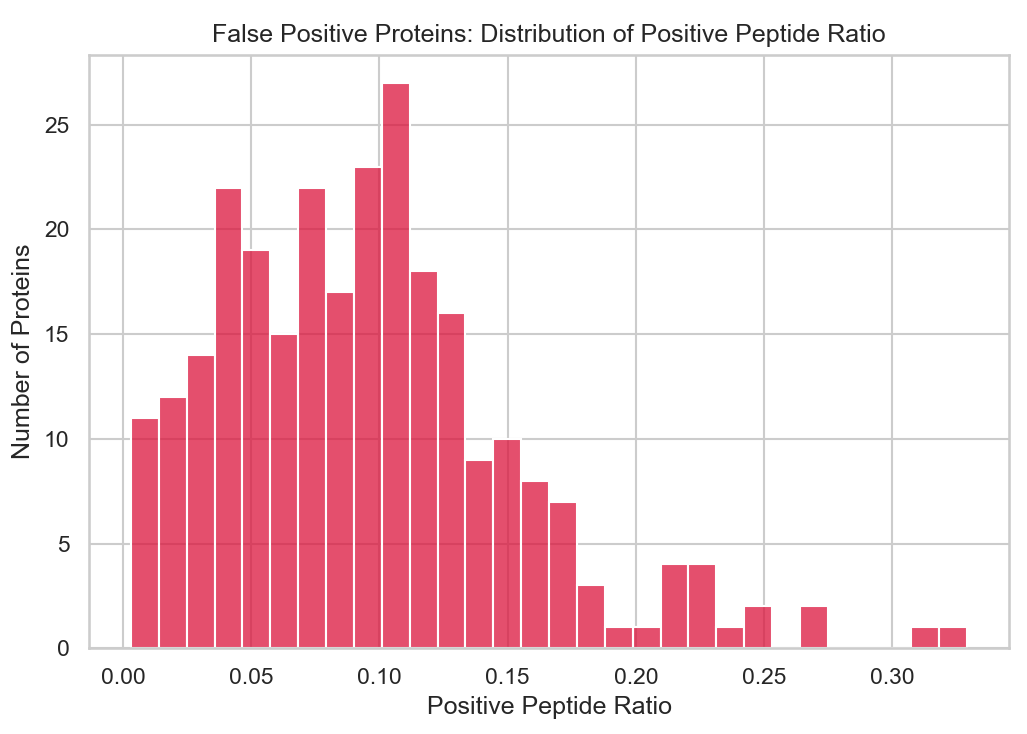
🟧 Figure2: **Prediction Distribution for True Negative Proteins**

This plot illustrates the model’s predictions for proteins that are truly negative—that is, proteins that do not contain a Nuclear Export Signal (NES). Despite their negative labels, most of these proteins were misclassified as positive by the model, indicating a high false positive rate.

This outcome, while concerning, is a direct result of the model’s “one-positive-peptide” decision rule: a protein is classified as NES-positive if even a single peptide within it receives a high prediction score. Given that proteins typically consist of hundreds of overlapping peptides, a single noisy or erroneously high-scoring peptide is often sufficient to cause an incorrect positive classification.

This observation naturally motivates the next plot, where we explore whether setting a threshold on the **number or proportion** of positive-scoring peptides could improve prediction reliability—by requiring more than just one peptide to classify a protein as positive.

🟥 Figure 3**: False Positive Proteins – Distribution of Positive Peptide Ratio**



This plot shows how many negative proteins were wrongly predicted as positive, based on the proportion of their peptides classified as positive.

We observe an inverse relationship: the fewer positive peptides a protein has, the more likely it is a false positive. Most of these errors occurred when only a small fraction (under 10%) of peptides were predicted as positive—often due to a single noisy peptide.

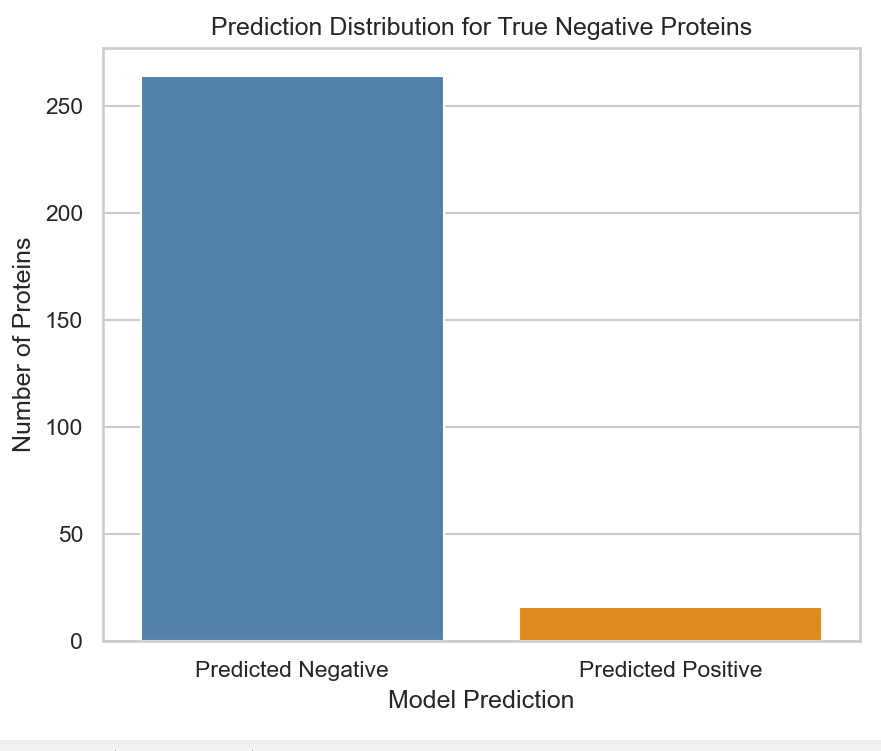
This reflects a weakness in the model’s original “one-positive-peptide” rule: if any peptide is positive, the whole protein is marked positive. Since true NES signals span several overlapping peptides, isolated positives are often errors.

To address this, we revised the rule: a protein is now predicted as positive only if at least **20%** of its peptides are classified as positive—reducing false positives without severely harming recall.

**Updated Results After Applying the 20% Peptide Threshold Rule-**

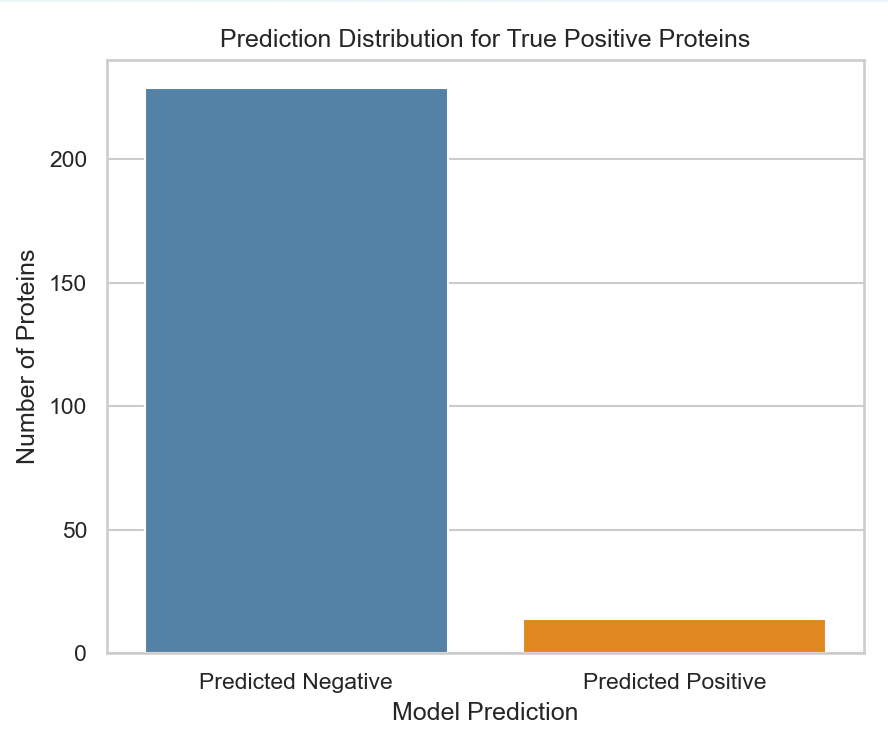
To mitigate the high false positive rate observed under the original “one-positive-peptide” policy, we revised the protein-level decision rule. Instead of marking a protein as NES-positive if any peptide is classified as positive, we now require **at least 20%** of the protein's peptides to be predicted as positive.

Figure 4: Prediction Distribution for True Negative Proteins



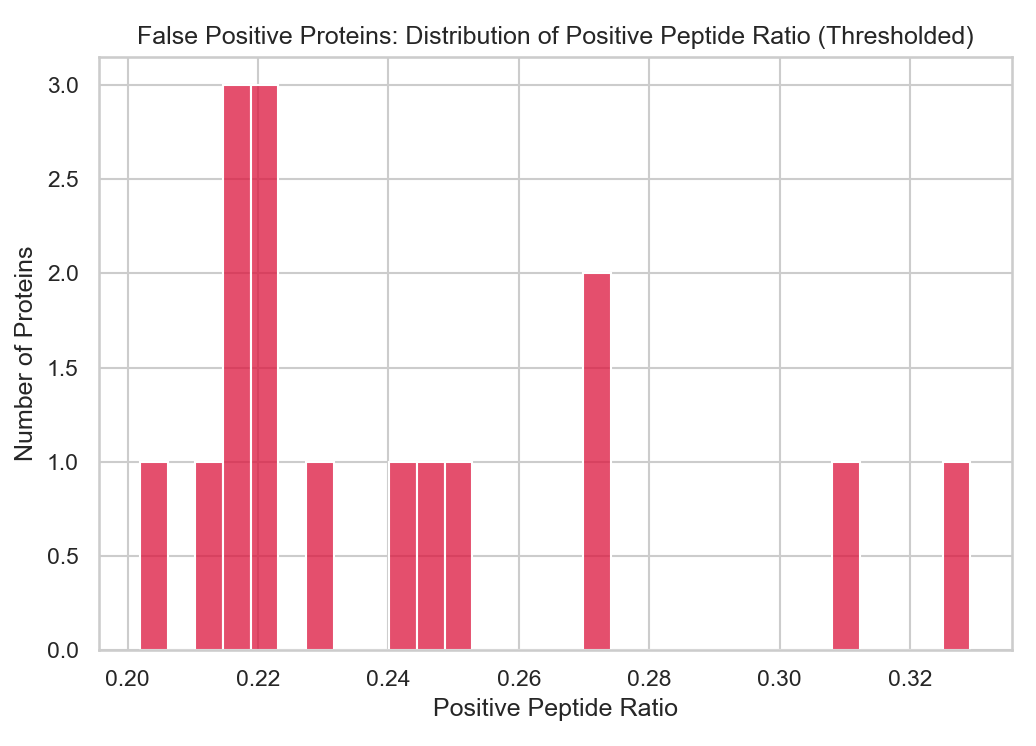
This plot shows a significant improvement in specificity: the vast majority of truly negative proteins are now correctly classified as negative. The false positive count dropped drastically, indicating that requiring multiple positive peptides effectively filters out spurious signals from isolated noisy peptides

**Figure 5: Prediction Distribution for True Positive Proteins**

****

While false positives decreased, this improvement came at a cost. The number of false negatives (true positives incorrectly predicted as negative) increased considerably. This suggests a **loss in recall**, as some true NES proteins no longer meet the 20% threshold—even though they may still contain valid NES regions.

**Figure 6: False Positive Proteins – Positive Peptide Ratio**



This histogram now shows that only a handful of false positives remain, and they barely pass the 20% threshold. This supports the effectiveness of the revised rule in eliminating weak, noisy signals. However, the sharp drop also implies that the threshold might be **too strict**, and that **a dynamic or learned threshold** could be a promising direction for future improvement.

**Conclusion**

While applying the 20% peptide threshold rule substantially improved specificity and reduced false positives, this improvement came at the expense of sensitivity. The model now fails to detect a significant portion of truly NES-positive proteins—suggesting that the revised rule, although effective in filtering noise, is overly conservative for many borderline cases.

This trade-off highlights an inherent limitation of the current model architecture and peptide-to-protein aggregation strategy. We tested both extremes—liberal (≥1 positive peptide) and conservative (≥20% positive peptides)—and observed that neither yields balanced performance. This suggests that the model, as currently trained, lacks sufficient discriminative capacity to support reliable protein-level NES classification.

**To achieve better performance, future work should explore retraining or rearchitecting the model, potentially incorporating more sophisticated aggregation logic or peptide context-aware mechanisms.**