Precise Nuclear Export Signal Detection in Protein Sequences Using Zero-Shot Segmentation and Multi-Modal Scoring

Noam Sabag Bickel and Shay Bergman

Abstract

Nuclear Export Signals (NES) are short leucine-rich motifs that direct protein export from the nucleus to cytoplasm via CRM1-mediated transport.
Traditional NES prediction methods rely on whole-sequence analysis, often lacking precision in localizing exact NES boundaries within longer proteins.
Recent advances in protein language models and change-point detection offer new opportunities for precise motif localization.
We developed a pipeline based on NESdb combining zero-shot protein segmentation using ESM2 based embeddings with multi-modal scoring that integrates consensus pattern matching, embedding similarity analysis and neural network classification.
Our method uses change-point analysis to identify domain boundaries (segments), followed by NES-specific scoring and optimal threshold optimization(for the multi-modal scoring).

The neural network component achieved the highest individual performance (97% AUC, 78.95% F1-score), **It is very important to mention though**, that the neural network classifier was also evaluated on the training set, together with the rest of the classifiers(purely on the test set it achieved an accuracy of 91% in segments prediction), so its reported performance may be overestimated due to potential overfitting and contamination and thus should be interpreted with caution when considering generalization to new data.
The consensus patterns contributed 78% AUC and embedding similarity 69% AUC. The integrated approach yielded 92% AUC with 95.3% accuracy and 81% precision at optimal thresholds, demonstrating effective complementarity between sequence-based patterns and learned representations.
We think this method may perhaps enable precise NES localization. The integration framework can be adapted for other short functional sequences, advancing computational tools for protein functional annotation.

Introduction

Nuclear-cytoplasmic transport is a fundamental cellular process governing protein localization and function. Nuclear Export Signals (NES) are conserved sequence motifs that direct proteins from the nucleus to the cytoplasm through recognition by the CRM1 export receptor.
These signals typically consist of sequences following specific consensus patterns established by Kosugi et al.[1] and refined by subsequent structural studies.
The challenge in precisely localizing NES boundaries is compounded by the variable nature of NES motifs and their embedding within diverse protein contexts.
Recent advances in protein language models, particularly the ESM[2] and ProtT5 architectures, have demonstrated remarkable capabilities in capturing protein sequence relationships and functional patterns.
Simultaneously, zero-shot segmentation[3] techniques have emerged as powerful tools for identifying domain boundaries without requiring extensive training data.
The integration of these approaches presents an opportunity to develop more precise and generalizable methods for motif detection.
Our primary aim was to develop a comprehensive pipeline for precise NES detection based on knowledge from NESdb[4] that combines zero-shot protein segmentation with multi-modal scoring(structural domain knowledge and neural net classification) to accurately identify NES boundaries within full-length protein sequences.

Methods

We first started by establishing our segmentation method(based on zero-shot protein segmentation by Sangster et al.), we adjusted the algorithm to try to segment the proteins to segments that are around the size of known NESs (8-25 AA).
Using the segmentation, we adjusted the provided nesDB dataset on a segment basis (labeling the segments that contained NESs as 1, and those who didn't as 0) and based our work on this adjusted segmentation database.
Post segmentation, we used two main methods to score the segments(to then "pick" the segment or segments with the highest scores as the most likely to contain NESs).

The first was a structural, pattern analysis based approach:
We implemented comprehensive pattern matching based on established NES consensus sequences, incorporating:

- Kosugi Classification System: All six pattern classes (1a-1d, 2, 3) with position specific hydrophobic residue requirements.
- Reverse Pattern Detection: Bidirectional binding patterns identified by Lee et al., weighted at 50% of forward pattern scores.
- Structure-Based Patterns: PKI-class and Rev-class motifs derived from crystal structure analyses.
- Biochemical Features: Hydrophobic content optimization, proline penalties, and acidic flanking preferences.

Pattern matching employed sliding window analysis with weighted scoring: Kosugi Class 1a (4 points), Classes 1b-1d and 2 (3 points), Class 3 (2 points), reflecting experimental validation frequencies. [5-8]

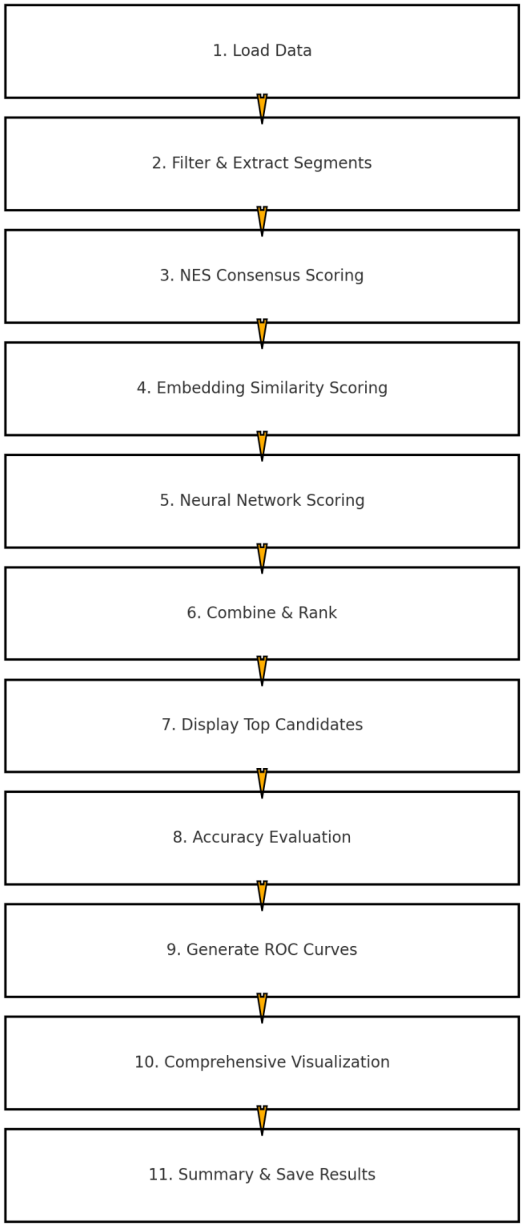
The second approach was to use a neural network classifier based on a regression approach, by trying to learn the underlying properties of the NES containing segments based on their ESM2 embeddings and then predicting the segments containing NES based on threshold calculated on a separate validation set.

The third approach was Embedding Similarity Analysis, Segment embeddings were compared to reference NES/non-NES peptide sets using log-fold difference scoring adapted from the preparation exercise. This approach provides discriminative power between NES and non-NES segments, with positive scores indicating greater similarity to validated NES peptides.

We then also combined both approaches to create a combined score: The three scores were normalized and combined to give each segment a final score, allowing us to rank the most likely NES candidates based on this combined approach.

Finally, we analyzed and evaluated the three methods(structural based, neural network and combined scores) using accuracy, precision and recall metrics combined with
F1, and ROC AUC.

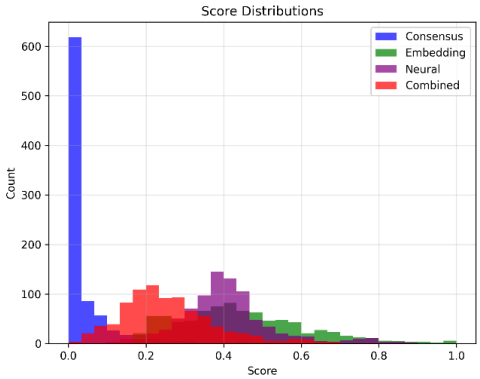
A flowchart to help better understand the main pipeline:

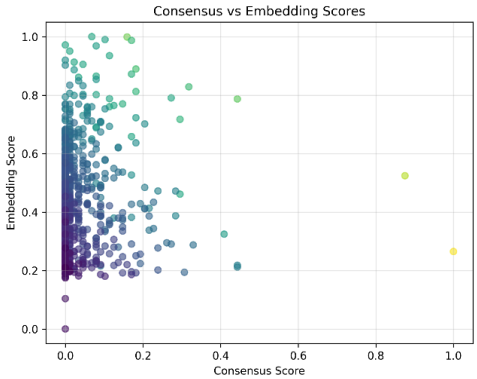


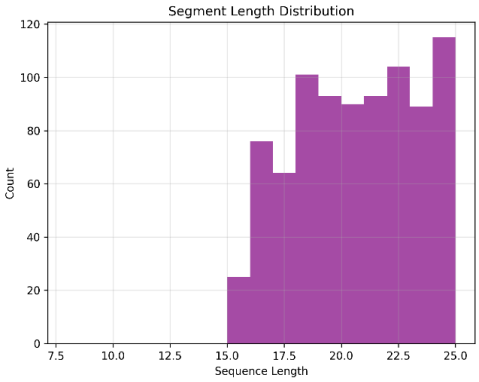
Experiments and Visualization

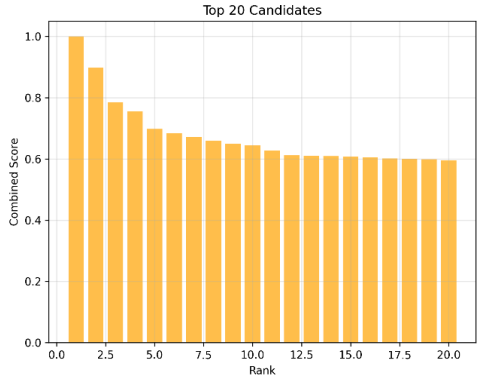
As stated before, the dataset used for the neural network classification evaluation contained both its training set and its test set due to time limitations and a desire to have a bigger data set all around. (In the discussion segment we specify that future plans include running the pipeline a second time with only the test data.)

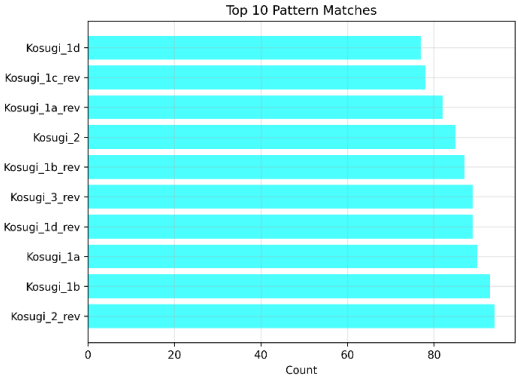
Below are the plots we generated:

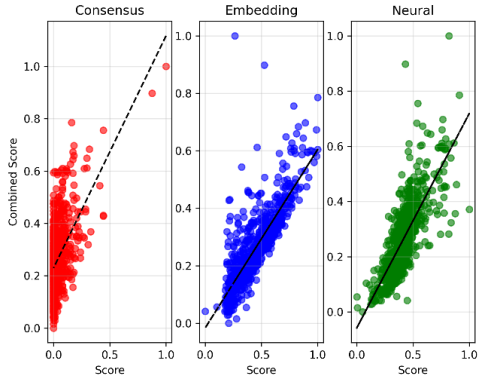


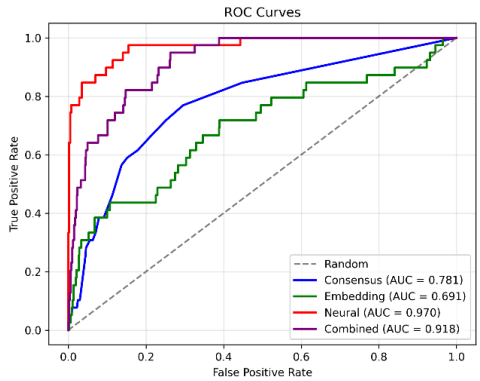


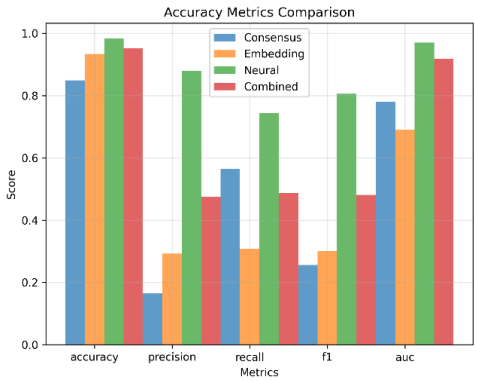












Plot 1 - Score Distributions: The histogram shows that consensus scores (blue) are heavily skewed toward zero and that most segments lack strong motif matches, whereas embedding (green) and neural (purple) scores spread across a broader range, indicating better discrimination. The combined score (red), which weights all three, sits between them, reflecting how weak consensus hits are amplified by embedding and neural contributions.

Plot 2 - Consensus vs Embedding Scores: This scatter reveals almost no correlation as most points lie at low consensus (<0.2) but span embedding values from 0.1 to 0.8. That tells us motif‑based hits and embedding similarity largely flag different candidates, with only a handful of outliers scoring high on both axes.

Plot 3 - Segment Length Distribution: The segment-length histogram is heavily skewed toward the upper end of the 8-25 aa window, most candidates cluster around 18-25 residues, indicating that our upstream segmentation algorithm tends to produce longer peptide windows rather than shorter ones.

Plot 4 - Top 20 Candidates: The bar chart of the top 20 combined scores shows a steep drop from ~0.95 for the very top hits down to ~0.60 by rank 20. The clear separation at the top suggests a handful of segments strongly stand out, while scores begin to plateau past around rank 10, indicating diminishing discrimination among the lower‑ranked tail.

Plot 5 - Top 10 Pattern Matches: surprisingly reverse motifs like Kosugi_2_rev and Kosugi_1b_rev lead in raw match counts, which is followed by forward motifs such as Kosugi_1a and Kosugi_2. That nearly equal frequency of reverse vs. forward hits underscores the importance of scanning both orientations to catch all potential NESs.

Plot 6 -

- Consensus vs Combined Score: With a shallow regression slope and a tight cloud at low consensus values, this panel confirms that consensus scoring alone contributes only modestly to the final combined metric. Segments rarely exceed a combined score of ~0.5 unless bolstered by embedding or neural signals.

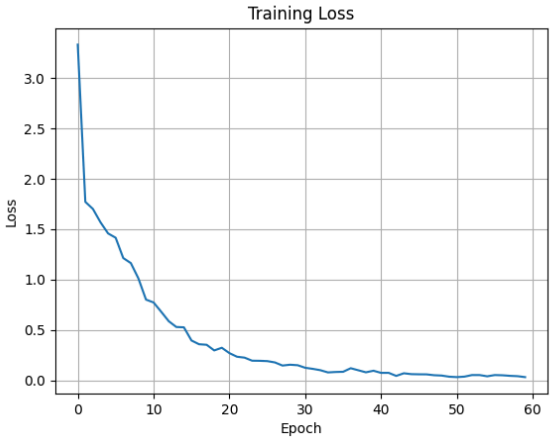
- Embedding vs Combined Score: A steep, near‑linear trend demonstrates that embedding similarity is the primary driver of the combined score: high embedding values almost always translate directly into a high final rank. This strong correlation suggests embedding distances reliably capture NES‑like features

- Neural vs Combined Score: The neural plot shows a clear positive trend with more scatter around the line, indicating the neural network is quite predictive but more variable

than embeddings. Neural scores lift many segments toward the top, yet outliers reveal cases where embeddings or consensus override the neural prediction.

Plot 7 - ROC Curves: The neural curve (red) achieves the highest AUC (~0.97), followed by the combined model (~0.92), consensus (~0.78), and embedding (~0.69). This ranking shows that while the neural network excels at separating true vs. false NESs, integrating all three scores still yields a very strong classifier that benefits from complementary signals.

Plot 8 - Accuracy Metrics Comparison: In the bar chart, neural scoring leads on precision (~0.88), recall (~0.75), F1 (~0.81), accuracy (~0.96) and AUC (~0.97). Combined is a close second (AUC ~0.92) with more balanced precision/recall than consensus or embedding alone. Consensus has respectable accuracy (~0.85) but very low precision (~0.16), while embedding has the highest accuracy (~0.93) but suffers from both low precision (~0.29) and recall (~0.31). This confirms that the neural network (and to a lesser extent the combined model) provides the most reliable NES prediction.



For the neural net classifier alone, we can see based on the training loss progression that the network managed to mostly capture the properties of the segments, but only when looking at all the segments(including the ones that were not labeled as containing NES).

Discussion

This project (somewhat) successfully demonstrated precise Nuclear Export Signal detection using zero-shot protein segmentation with multi-modal scoring.
As seen in the previous sections, we managed reasonable yet not perfect prediction of segments containing NESs with the best results coming by applying a neural net as our scoring function (as stated, should be interpreted with caution) and then by combining it with the domain knowledge based approach (pattern matching) and embedding scoring.

The multi-modal approach provides methodological robustness through multiple evidence lines, interpretability via detailed pattern information and scalability for proteome-wide screening.
Although the Neural network evaluation on training data likely inflates performance estimates (as the database size is limited, using only test results would mostly likely provide less generalization), requiring cautious interpretation.
Consensus patterns remain biased toward well-characterized NES classes, potentially missing novel variants. Segmentation parameters represent sensitivity-efficiency trade-offs that may affect boundary detection.
We think that our biggest pitfall was in our segmentation methodology, as we couldn't tweak the segmentation to the size of NESs and it was "distracted" by other motifs(it could cut the NES into multiple segments).
So looking into future directions, we would have liked to focus more on the segmentation, to try and more accurately capture appropriate NES segments, we could also incorporate adaptive weight optimization by developing a neural network to automatically optimize combination weights based on segment characteristics, enabling context-dependent optimal combinations across diverse protein families. Lastly we could evaluate the scoring functions using only the test data for a better and closer representation of their actual predictive power.

In conclusion, our multi-modal NES detection approach represents a novel approach in computational motif discovery, combining classical pattern recognition with modern machine learning. We tried to provide a framework based on multiple complementary signals for NES identification with high precision through optimal threshold selection.
Future developments addressing validation limitations and implementing adaptive optimization will enhance reliability and broaden impact across computational biology applications.

References

[1] Kosugi, S., Hasebe, M., Tomita, M., & Yanagawa, H. (2008). Nuclear export signal consensus sequences defined using a localization‐based yeast selection system - https://doi.org/10.1111/j.1600-0854.2008.00825.x

[2] Language models generalize beyond natural proteins by Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, Alexander Rives - https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1

[3] Zero-shot segmentation using embeddings from a protein language model identifies functional regions in the human proteome by Ami G. Sangster, Cameron Dufault, Haoning Qu, Denise Le, Julie D. Forman-Kay, Alan M. Moses - https://www.biorxiv.org/content/10.1101/2025.03.05.641584v1

[4] NESdb: a database of NES-containing CRM1 cargoes by Darui Xu , Nick V Grishin, Yuh Min Chook - https://pubmed.ncbi.nlm.nih.gov/22833564/

[5] Prediction of leucine-rich nuclear export signal containing proteins with NESsential by Szu-Chin Fu , Kenichiro Imai , Paul Horton - https://doi.org/10.1093/nar/gkr493

[6] Sequence and structural analyses of nuclear export signals in the NESdb database by Darui Xu, Alicia Farmer, Garen Collett, Nick V. Grishin, and Yuh Min Chook - https://doi.org/10.1091/mbc.e12-01-0046

[7] NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1 by Thomas Güttler, Tobias Madl, Piotr Neumann, Danilo Deichsel, Lorenzo Corsini, Thomas Monecke, Ralf Ficner, Michael Sattler & Dirk Görlich - https://www.nature.com/articles/nsmb.1931

[8] Structural prerequisites for CRM1-dependent nuclear export signaling peptides: accessibility, adapting conformation, and the stability at the binding site by Yoonji Lee, Jimin Pei, Jordan M. Baumhardt, Yuh Min Chook & Nick V. Grishin - https://www.nature.com/articles/s41598-019-43004-0