# Hackathon - Finding NES Properties using PLM

## 

## Abstract

Nuclear Export Signals (NESs) are hydrophobic peptide motifs that facilitate CRM1-mediated protein transport from the nucleus to the cytoplasm, playing critical roles in cell cycle regulation and signaling pathways. Despite their biological importance, accurate NES detection remains challenging due to structural and sequence diversity, with computational methods suffering from high false positive rates. This study developed an integrated approach combining protein language model (PLM) embeddings with structural information to improve NES prediction accuracy. We trained a dense neural network using ESM2 embeddings on a curated dataset of 168 proteins containing 317 experimentally validated NES motifs. The model achieved high performance (AUC 0.9) on test data but exhibited significant positive bias when applied to full protein sequences, generating 12,131 false positives via moving window analysis. To address this limitation, we incorporated structural filtering based on DSSP-derived secondary structure and solvent accessibility data. Our analysis confirmed that NES regions are significantly enriched in helical (56.65% vs 30.12% in controls) and exposed (75.22% vs 72.20%) residues, with combined helical and exposed characteristics showing strong discrimination (40.69% vs 18.20%). Additionally, PLM-based protein segmentation revealed that NES motifs preferentially localize within single structural segments (≥85% overlap). Integrating these structural constraints reduced false positives by approximately 50% (from 12,131 to 5,843) while maintaining reasonable sensitivity. This multi-modal approach demonstrates the value of combining sequence-based predictions with structural validation for improved NES detection in proteome-wide analyses.

## Introduction

A cell’s nucleus is surrounded by Nuclear Pore Complexes (NPCs), which regulate the transport of macromolecules such as proteins into and out of the nucleus (Lin & Hoelz, 2019; Dickmanns, Kehlenbach, & Fahrenkrog, 2015). One well-characterized export pathway involves the export receptor CRM1, which recognizes short peptide motifs known as Nuclear Export Signals (NESs) on cargo proteins (Nguyen, Holloway, & Altura, 2012). Upon binding an NES-containing protein, CRM1 facilitates its transport through the NPC into the cytoplasm.

NES motifs are typically hydrophobic, often form part of an alpha-helix, and loosely follow consensus patterns (Chook et al., 2012). However, detecting NESs in protein sequences is challenging due to their structural and sequence diversity (Fung et al., 2017; Kosugi et al., 2014). Despite the significance of NES-CRM1 interactions in various cellular processes, including cell cycle regulation, signaling pathways, and RNA processing, precise identification of NES motifs remains difficult.

This hackathon aimed to develop and refine methods to accurately detect NES regions and their characteristics in protein sequences. To achieve this goal, we leveraged segmentation techniques using pretrained Language Model (pLM) embeddings and integrated structural information to reduce false positives. This approach builds upon prior research by Moses Lab (Ami G. Sangster et al., 2025) and draws insights from human proteome screenings for NES motifs (Kirli, Koray et al, 2015), highlighting the broader relevance of precise NES detection methods in cellular and structural biology.

## **Methods**

### Data

Our dataset from Lee et al, includes 317 experimentally validated NES annotations across 168 human proteins, each record includes protein information (sequence, name, uniprotID) and the NES information (position and sequence).   
To train our model we used an earlier version of this dataset, based on NESdb, which includes similar information, but also negative sequences from the protein that don’t include a NES peptide.

### Modeling with PLM embeddings

In order to train our NES prediction model first extracted the sequence embedding of each peptide from ESM2. This is done by extracting the embedding of each amino acid from the last layer of the model, then averaging them together to a sequence embedding.

The model was a basic dense neural network, with one hidden layer of dimension 256, run for 50 epochs with a learning rate of 0.01.

### Finding potential NES using a moving window

In order to find a potential NES peptide in a protein, we scanned through the sequence with a moving window of size 22, running inference on the model and retrieving the probability the model gave.

We then defined the following metrics:

* True positive: if the model predicted a window with a probability over 0.5 to be a NES and that window overlapped enough with a known NES (at least 10 amino acids)
* False positive: if the model predicted a window as a NES and that window did not overlap with a known NES.
* False negative (missed NES): if a NES was not detected at any point by the model

### Extracting structural information using DSSP

Secondary structure and solvent accessibility were extracted using the DSSP (Dictionary of Secondary Structure of Proteins) program via the Bio.PDB interface. Protein structures in PDB format were parsed, and residue-level annotations were generated using DSSP.

Residues were classified as:

* **Helical**: DSSP-assigned codes "H" (α-helix), "G" (3₁₀-helix), or "I" (π-helix).
* **Exposed**: Relative Solvent Accessibility (RSA) ≥ 0.25.

For each protein, boolean vectors were created to indicate whether each residue is helical (is\_helix), exposed (is\_exposed), or both.

### Cross-referencing annotations

After extraction of characteristics (exposed, helical, in segment) we filtered out the predicted NESs of the model by:

* Exposed: if the predicted window had more than 5 exposed amino acids
* Helical: if the predicted window had more than 5 helical amino acids
* Segmentation: if the predicted window fell at least 85% inside a single segment

### PLM-Based Segmentation and Overlap Analysis Pipeline

All modifications were implemented as an extension of the zero-shot-protein-segmentation framework (Sangster et al), without altering its core algorithms or data structures.

### Embedding Generation

Two protein language models were used via separate modules. The existing ESM-1b wrapper was extended to load ESM-2 weights, adjust the embedding dimensionality to 1,280, and batch sequences up to 4,000 residues.

### Sequence Segmentation

Change-point detection on each embedding matrix was performed using the Window algorithm from the “ruptures” library (width = 30, model = “rbf”, jump = 1). The maximum number of breakpoints per sequence was set to . Identified change points partitioned each protein into contiguous segments, which formed the basis for subsequent overlap analysis.

### NES Overlap Computation

For each experimentally validated NES window, overlap with each segment was computed as the length of their intersection divided by the NES window length, expressed as a percentage. The segment yielding the highest overlap was recorded for each motif, capturing its index, length, and overlap percentage for downstream evaluation.

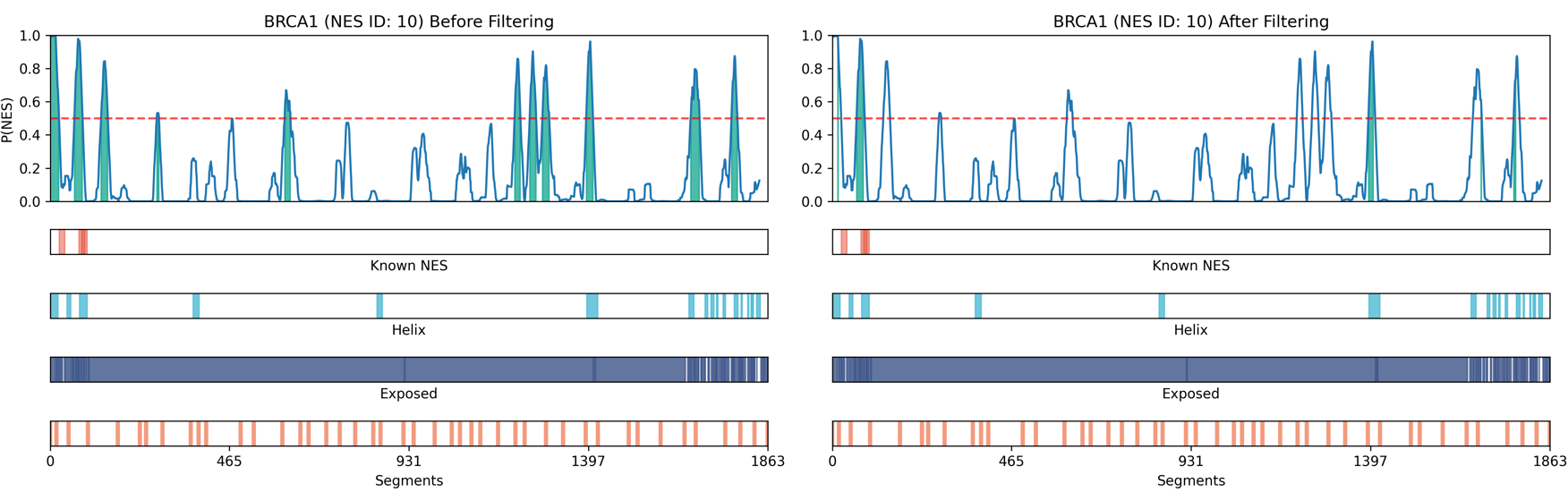
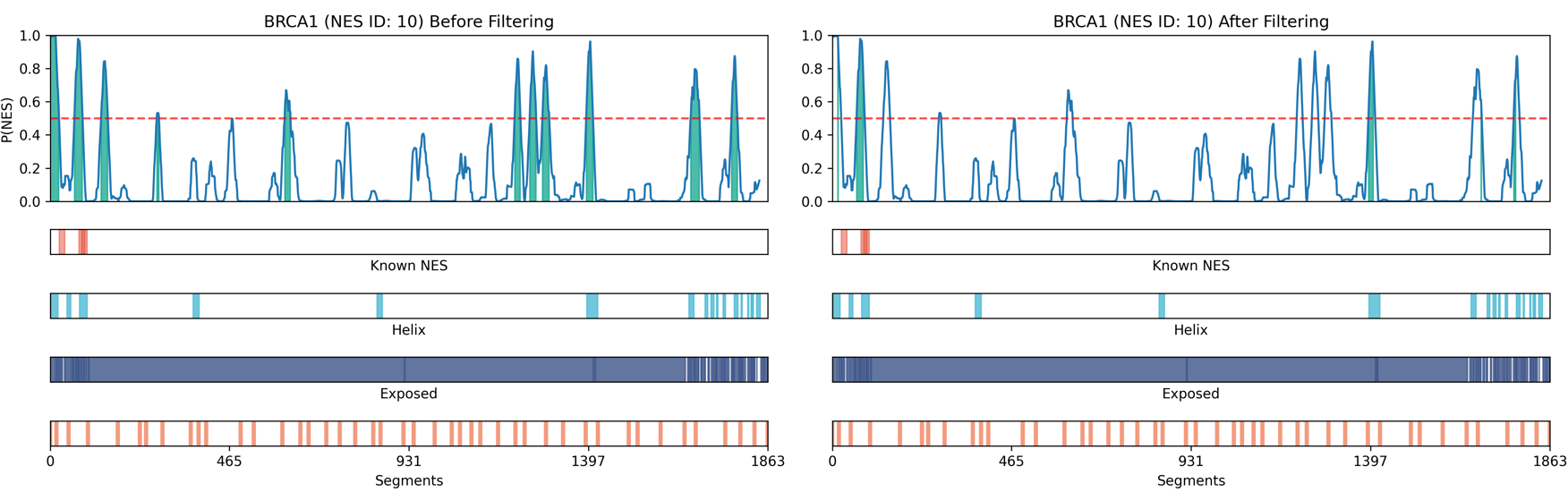
is the length of the intersection between the annotated NES window and the predicted segment. is the length of the annotated NES window.

## **Results**

### Large positive bias when predicting along entire protein sequences:

Even though the model’s results were very good (AUC 0.9) on the test dataset, when predicting using the model on entire protein sequences using a moving window, large amounts of false positives are detected. When running on our entire dataset a total of 12131 FP were detected. The number of false negatives was much lower (the mode discovered most of the known NESs) which implies a strong positive bias.

Figure 1- Top: Model NES probability over the protein BRCA1, Bottom: Known NES annotations across the protein



### NES regions show structural characteristics

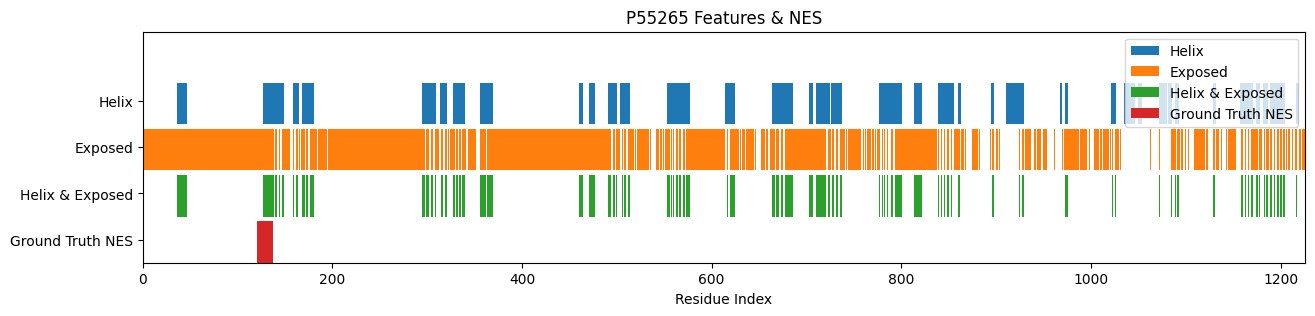
In order to filter out false positives, we tried to verify some known structural characteristics of a NES peptide, specifically being in exposed regions and having a partially helical structure.

Figure 2 **-** Example of feature block plot for helix, exposed, and NES regions of the ADAR gene (uniprot id: P55265)

After extracting the results from the database and comparing to randomly chosen peptides from the protein (the “Control”) we found that NES are significantly more in helical and exposed regions (when compared using a pair t-test). Just exposed regions were not enough to differentiate, however when combining the two we saw a strong differentiation between the classes (Figure 4).

Table 1 - NES helical and exposed region occurrences. The control was a randomly sampled 15 AA section in each protein

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | NES (%) | Control (%) | NES (Avg. Count) | Control (Avg. Count) |
| Exposed | 75.22% | 72.20% | 15.48 | 18.95 |
| Helix | 56.65% | 30.12% | 12.25 | 8.68 |
| Helix + Exposed | 40.69% | 18.20% | 8.32 | 5.07 |

### 

Figure 3: Box-and-strip plot showing distribution of helical residues in NES vs. control regions across all proteins in the DB

### PLM-Based Segmentation and Overlap Analysis

After the segmentation and selection of the segments best matching to the experimentally validated NES (Figure 4) we ended up with overall average NES matching score of 94.3% for ESM-2 based segmentation and 93.5% for ProtT5 based segmentation, suggesting that our method was probably able to reliably identify segments that correspond to functional NES motifs. We also explored the option of introducing a cutoff based on segment length - however, the distribution of the lengths of the identified segments was broad and did not support a clear threshold (Figure 5).

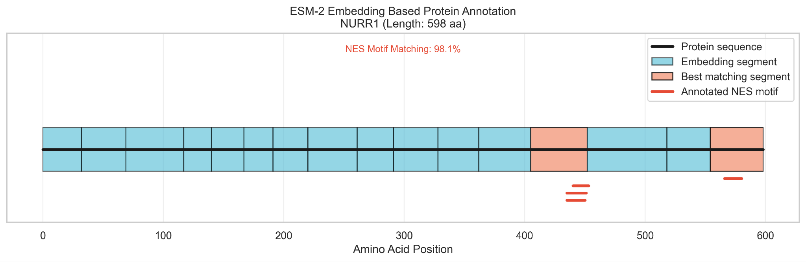
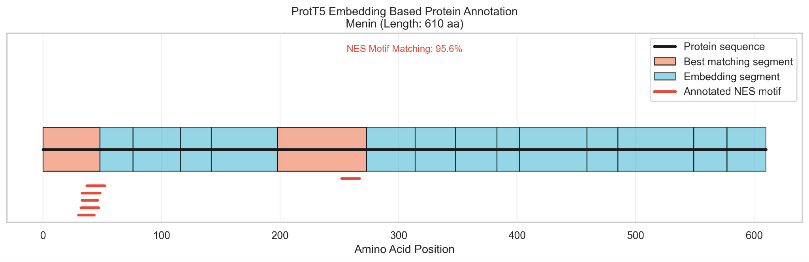
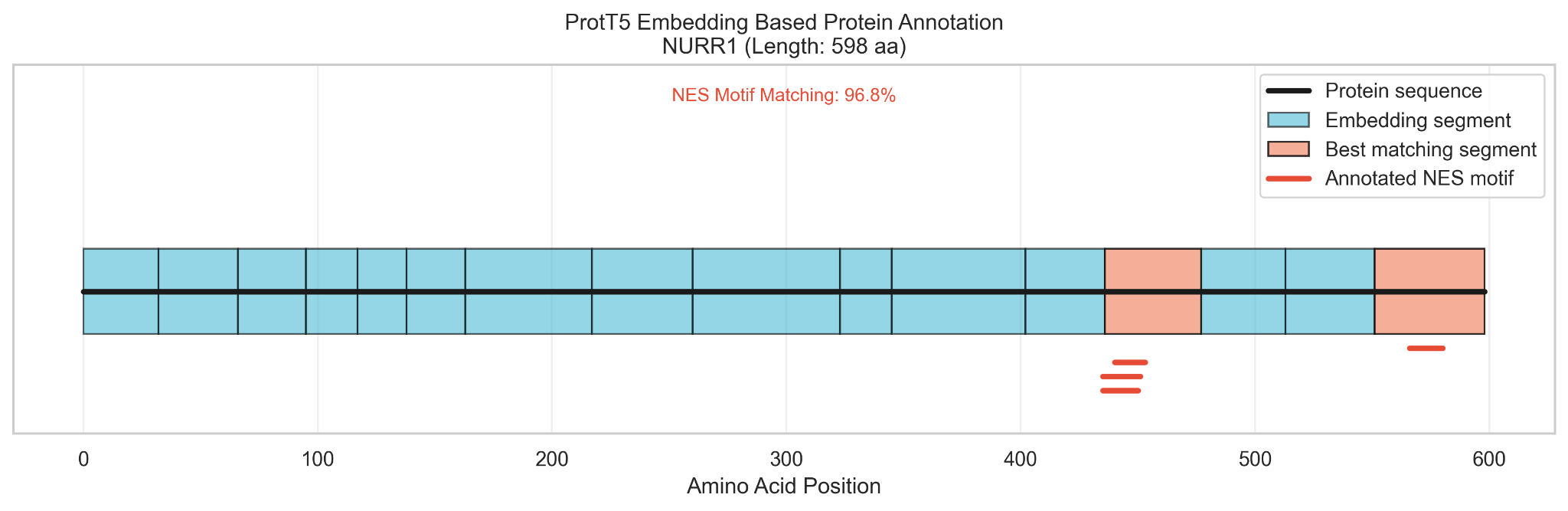
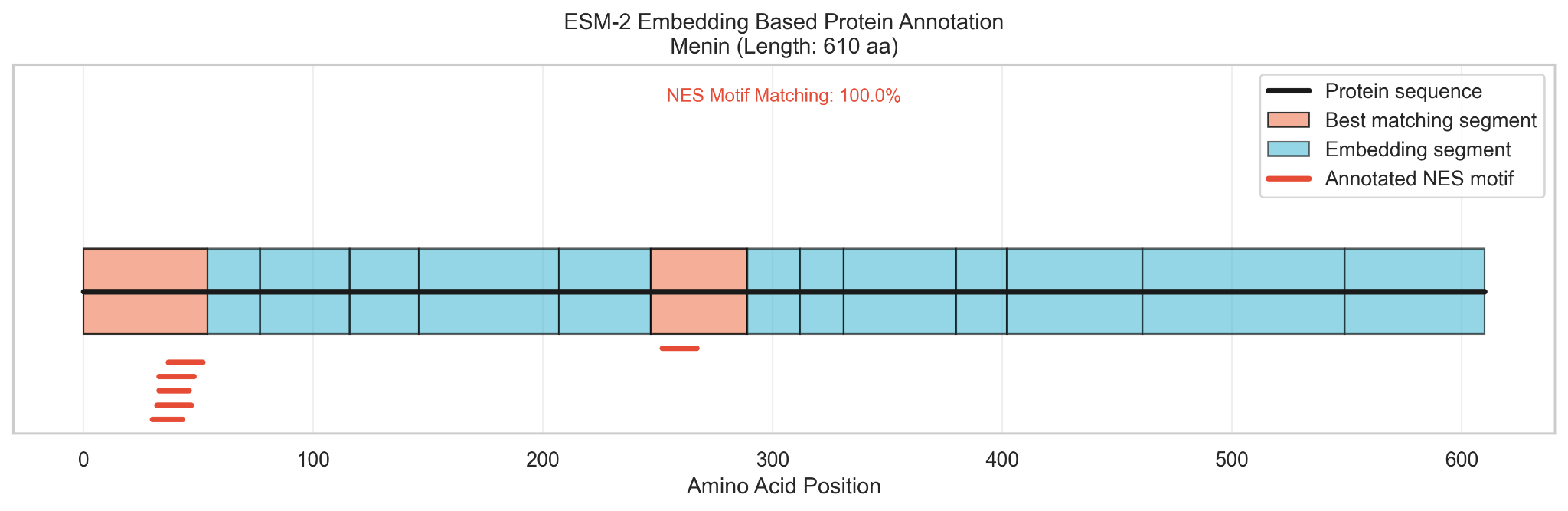


Figure 4- protein annotation plots - segmentation by the embedding, with highlighting of the segments selected by the highest match

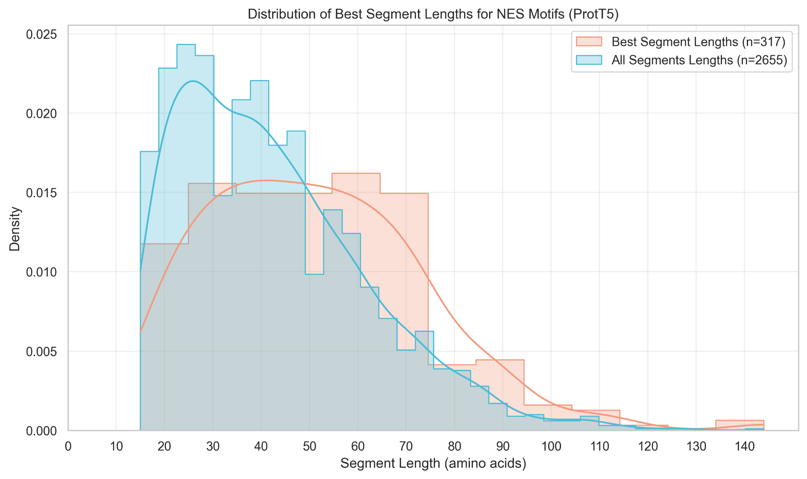
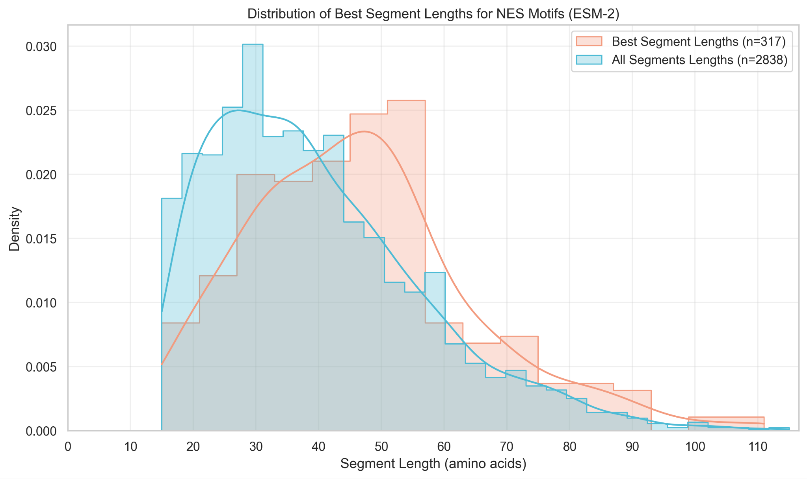


Figure 5 - Comparison of the length distributions of all identified segments versus those best matching NES motifs. Left: ProtT5, Right: ESM-2.

### 

### Cross referencing with NES characteristics reduces positive bias

Since we extracted information that characterizes NES motifs, we were then able to layer the information over the model prediction in order to reduce the false positive rate.

Figure 6 - Examples from genes RanBP1 (top) and BRCA1 (bottom) of NES cross before (left) and after (right) filtering. Each line is an annotation across the protein’s amino acids. From to bottom: Model probability prediction, Known NES annotation, helical annotation, exposed annotation, embedding segments.

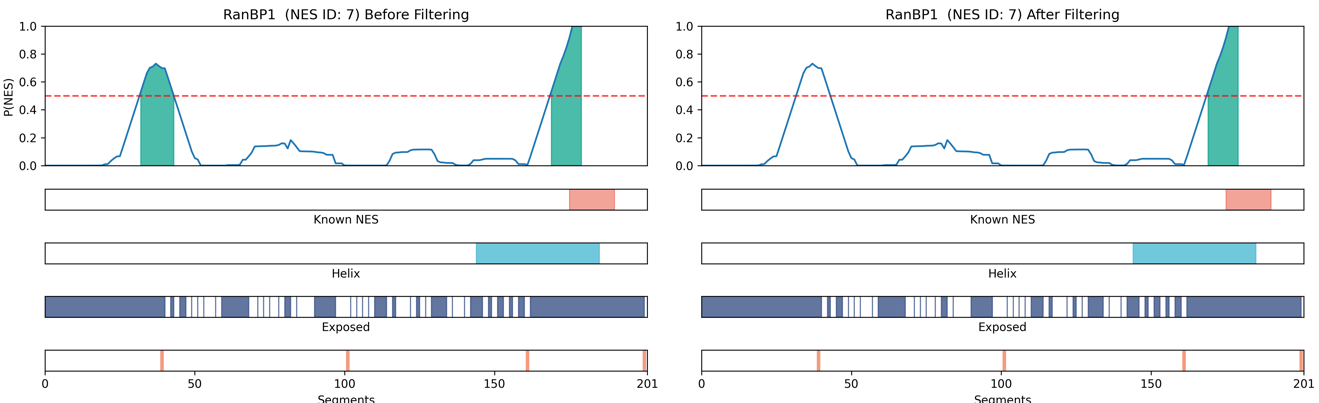
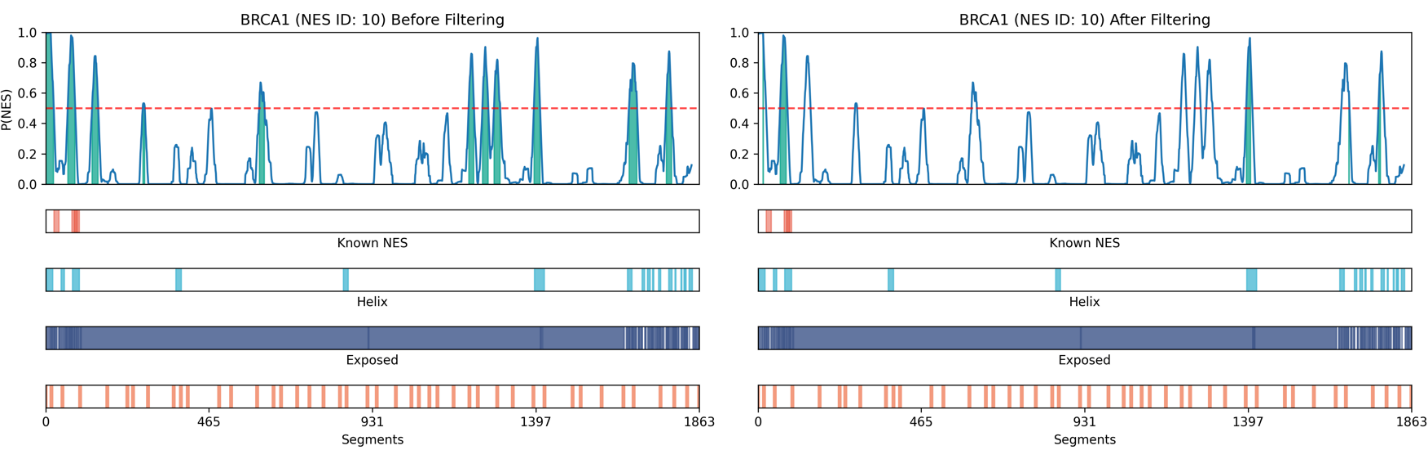


Table 2 - Model results on entire dataset before and after filtration

|  |  |  |
| --- | --- | --- |
| Metric | Before Filtration | After Filtration |
| **TP** | 2094 | 1116 |
| **FP** | 12131 | 5843 |
| **Missed NES** | 48 | 88 |

We managed to reduce the FP count two fold. The TP count also decreased (number of missed NES increased), which could be because of harsh filtering or problematic tagging in the dataset we used (as pointed out HERE). Overall, being able to produce a large number of “high quality” false positives could prove useful in future training and prediction.

## Discussion

In this hackathon we tried negate the large positive bias in our model we tried validating and finding new characteristics of NES proteins from the structure and the sequence. We managed to validate and find that NESs tend to be in exposed areas with helical structures. In addition, when segmenting the protein amino acids using a PLM’s embedding, we found that NESs tend to fall inside boundaries of a single segment.   
Taking all this information into account, we were able to filter out many false positives.

Using this filtering we could also create a new dataset with higher quality false labels. As our base dataset uses arbitrary protein segments, using the false positives, especially the “harder” ones that have all the characteristics we found, could be more informative for a model, and reduce the positive bias.

## References

1. Dickmanns, A., Kehlenbach, R. H., & Fahrenkrog, B. (2015). Nuclear Pore Complexes and Nucleocytoplasmic Transport: From Structure to Function to Disease. *International Review of Cell and Molecular Biology*, 320, 1-71.
2. Fung, H. Y. J., Fu, S. C., Chook, Y. M., & Süel, K. E. (2017). Nuclear export receptor CRM1 recognizes diverse conformations in nuclear export signals. *eLife*, 6, e23961.
3. Kosugi, S., et al. (2014). NESmapper: Accurate Prediction of Leucine-Rich Nuclear Export Signals Using Activity-Based Profiles. *PLoS Computational Biology*, 10(6), e1003664.
4. Lin, D. H., & Hoelz, A. (2019). The Structure of the Nuclear Pore Complex (An Update). *Annual Review of Biochemistry*, 88, 725-783.
5. Nguyen, K. T., Holloway, M. P., & Altura, R. A. (2012). The CRM1 nuclear export protein in normal development and disease. *International Journal of Biochemistry and Molecular Biology*, 3(2), 137–151.
6. Xu, Darui et al. “NESdb: a database of NES-containing CRM1 cargoes.” Molecular biology of the cell vol. 23,18 (2012): 3673-6. doi:10.1091/mbc.E12-01-0045.
7. Sangster, Ami G., et al. "Zero-shot segmentation using embeddings from a protein language model identifies functional regions in the human proteome." bioRxiv (2025): 2025-03.‏
8. Kırlı, Koray, et al. "A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning." elife 4 (2015): e11466.‏
9. Sangster, A. (2025). zero-shot-protein-segmentation [Computer software]. GitHub. <https://github.com/moses-lab/zero-shot-protein-segmentation>.