# **NES-Finder: A Transformer-based Pipeline for Identifying Novel Nuclear Export Signals**

**Authors:** Daniel Levin, Imri Shuval, Shira Gelbstein, Ron Levin

### **Abstract**

The transport of proteins between the cell nucleus and cytoplasm is mediated by short amino acid sequences, including the Nuclear Export Signal (NES) which binds to the export protein CRM1. Identifying functional NES motifs computationally is challenging due to their degenerate sequence patterns. To address this, we developed a computational pipeline using embeddings from the ESM-2 protein language model to train a Transformer-based classifier. The model was trained to distinguish between proteins containing known NES motifs and a curated set of human proteins experimentally shown not to bind CRM1.

Initial baseline tests using bacterial proteins as a negative control confirmed the pipeline's functionality, with the model achieving near-perfect accuracy on this simplified task. When trained on the more challenging human non-binder dataset, the model achieved an Area Under the Curve (AUC) of **[Enter AUC value]**. A positive correlation was also observed between the model's prediction confidence and the experimentally measured binding strength of the NES to CRM1.

This work establishes a functional framework for NES prediction and underscores the critical importance of a well-curated negative dataset for this task. The resulting model serves as a tool for generating hypotheses about NES function and prioritizing candidates for experimental validation.

### **1. Introduction**

#### **General Biological Background**

Living organisms are made of cells, the basic units of life. In complex organisms like humans, these are called eukaryotic cells. A key feature of these cells is a compartment called the **nucleus**, which encloses the cell's genetic blueprint, DNA.

The site of most protein synthesis and cellular activities is the **cytoplasm**, the region outside the nucleus. For the cell to function, there must be a constant, regulated flow of molecules between the nucleus and the cytoplasm. This traffic moves through gateways in the nuclear membrane called **Nuclear Pore Complexes (NPCs)**.

#### **Specific Background: The NES Signal**

For a protein to exit the nucleus through an NPC, it often needs to carry a specific signal. This signal is a short amino acid sequence within the protein called a **Nuclear Export Signal (NES)**. The NES motif serves as a binding site for a transport protein called **CRM1** (also known as XPO1). When CRM1 binds to a protein's NES, it transports that protein out of the nucleus into the cytoplasm.

This process is important for normal cell function and is also implicated in diseases. For instance, many viruses use the CRM1 pathway to export their own proteins, and in some cancers, CRM1 is overactive. Because of this, CRM1 is a therapeutic target in several diseases.

A key challenge is that the NES is not a single, fixed sequence. It is a degenerate pattern, generally rich in hydrophobic (water-repelling) amino acids like Leucine (L) at specific spacings. This ambiguity makes it difficult to reliably search for NES motifs in a protein sequence.

#### **Project Goal**

The main objective of this project is to develop and apply a deep learning-based pipeline to address the challenge of ambiguous NES patterns. By learning from known examples, we aim to create an accurate classifier that can distinguish between proteins that contain a functional NES and those that do not bind to CRM1.

### **2. Methods**

#### **Dataset and Preprocessing**

**Positive Dataset:** Our positive samples were sourced from the supplementary materials of *Lee et al., "NESsential: a database of nuclear export signals." Nature Scientific Reports, 2019*. As this source provided only the short NES sequences, we programmatically retrieved their corresponding full-length protein sequences from the UniProt database to construct our dataset.

**Negative Dataset:** We took an iterative approach to constructing our negative dataset.

1. **Initial Baseline:** For an initial baseline validation, a negative set was constructed from randomly sampled subsequences of bacterial proteins.
2. **Final Negative Set:** For our final model, we used a more challenging dataset sourced from a collaborator (file name: DB\_Tanya). This dataset consists of human proteins that have been experimentally shown *not* to bind to the CRM1 exportin, providing a set of validated true negatives.

The final dataset was split into training and test sets for model development and evaluation.

#### **Computational Approach**

Our pipeline consists of two main stages:

**1. Sequence Representation:** Full-length protein sequences were first converted into numerical embeddings using the ESM-2 protein language model. Subsequently, subsequences corresponding to the known NES motifs (for positives) or randomly selected segments (for negatives) were extracted from these full-sequence embeddings based on annotated start positions.

**2. Model Training and Testing:** A preliminary model was trained using the bacterial negative set to validate the pipeline's functionality. Based on these initial results, we proceeded to train our final **Transformer-based classifier** (transformer\_NES\_classifier.py) on the human non-binder dataset. The model was trained using a binary cross-entropy loss function. For evaluating a single novel sequence, our method uses a sliding window of 20 amino acids to generate predictions across its entire length.

All code for this project is available on our GitHub repository: github.com/shiragelb/Bio-3D-Hackathon-2025

### **3. Results (Experiments)**

1. Initial Sanity Check: Human vs. Bacterial Proteins

Our first experiment served as a baseline validation for the pipeline. The model was trained to distinguish human positive NES sequences from randomly selected bacterial protein sequences, and it achieved near-perfect accuracy (accuracy ≈ 1.0). A t-SNE visualization of the pre-training embeddings (Figure 1) shows that the two groups were already highly separable, indicating that this was a simple classification task and motivating our move to a more realistic negative dataset.

***[Placeholder for Figure 1: graphs/human-positive NES vs Bacteria (negative)/2d\_k\_means.png. The caption should read: "Figure 1: t-SNE visualization of protein embeddings before training. The vectors for human NES proteins (positives) and bacterial proteins (negatives) are clearly differentiable, suggesting this is not a challenging classification task." ]***

**2. Correlation with Binding Affinity**

### **TODO: align the test results with the power of binding to see if it is correlated**

***[Placeholder for Figure 2. The caption should read: "Figure 2: An analysis of the test set predictions suggests a correlation between the model's prediction confidence and the experimentally measured binding strength of the NES to CRM1. Peptides with stronger binding affinities generally received higher prediction scores." ]***

**3. Final Model Performance**

### **TODO: AUC graph. Explain results**

***[Placeholder for Figure 3: Your final ROC Curve plot. The caption should read: "Figure 3: ROC curve for the final Transformer classifier on the held-out test set of human non-binders. The model achieved an AUC of [Your AUC Value], demonstrating its ability to distinguish between NES-containing proteins and true non-binding proteins." ]***

### **4. Discussion**

Our project developed a deep learning pipeline for the classification of Nuclear Export Signals. Our final model, trained on a curated dataset of human non-binders, showed a [TODO complete here based on results] ability to distinguish between positive and negative examples.

**Limitations:** The main bottleneck of our project was the acquisition and definition of a high-quality negative dataset. We struggled with defining the negative control data set – would it be a human protein that binds to other protein? If so, what would be the sub-sequence we take from it? Due to the fact that NES motifs are so varied and ill-defined, it is genuinely hard to understand what sequence features differentiate them from other parts of the proteome. This lack of clear biological definition for a "non-NES" makes creating a perfectly representative negative dataset a significant challenge in this field.

Future Work:

The primary avenue for future work is the continued improvement of the negative dataset. This could involve finding other databases of proteins with known subcellular localizations or interaction partners to further refine the set of true non-binders.

### **TODO: if result 2 is good, adding this binding strength parameter as a feature under consideration while training the model could improve performance.**

Broader Implications:

Despite the data challenges, our work establishes a functional framework for tackling this problem. A well-trained NES classifier can be a tool for cell biologists to generate hypotheses and prioritize proteins for experimental study, ultimately accelerating our understanding of nucleocytoplasmic transport.