NES-Finder: A Transformer-classifier for Identifying Novel Nuclear Export Signals

Daniel Levin[1*], Imri Shuval[1*], Shira Gelbstein[1*] , Ron Levin[1*]

[1]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel.

*Corresponding author(s). E-mail(s): Daniel.Levin2@mail.huji.ac.il; imri.shuval@mail.huji.ac.il; shira.gelbstein@mail.huji.ac.il; Ron.Levin1@mail.huji.ac.il;

Abstract

The transport of proteins between the cell nucleus and cytoplasm is mediated by nuclear pores complexes (NPCs) and transporter proteins. One family of such proteins is the Nuclear Export Signals (NES), which are proteins that bind the export protein CRM1. Identifying functional NES motifs is challenging due to their degenerate sequence patterns. To address this, we developed a method which uses 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.

Keywords: Machine Learning, Deep Learning, Protein Language Models, NES, CRM1

1 Introduction

1.1 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 genome. 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).

1.2 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 complex 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 at specific spacings. This ambiguity makes it difficult to reliably search for NES motifs in a protein sequence.

1.3 Project Goal

The main objective of this project is to develop and apply a deep learning-based method to address the challenge of identifying 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 which do not.

2 Methods

2.1 Dataset and Preprocessing

Positive Training Dataset

Our positive training samples were sourced from the supplementary materials of "NESsential: a database of nuclear export signals" [1]. Since they only provide the specific NES sequence, we retrieved their corresponding full-length protein sequences from the UniProt[2] database to construct our dataset.

Negative Dataset

At first, we constructed a negative set from randomly sampled sub-sequences of bacterial proteins. While they (almost certainly) guaranteed to not contain NES sequences, they proved to be a poor negative class,

as they are very different from proteins found in humans. Training a network to distinguish between our positive set and the bacterial proteins proved to be a trivial task. This was further confirmed when we looked at the ESM-embeddings of the two sets after using t-SNE [3] to visualize them in 2D

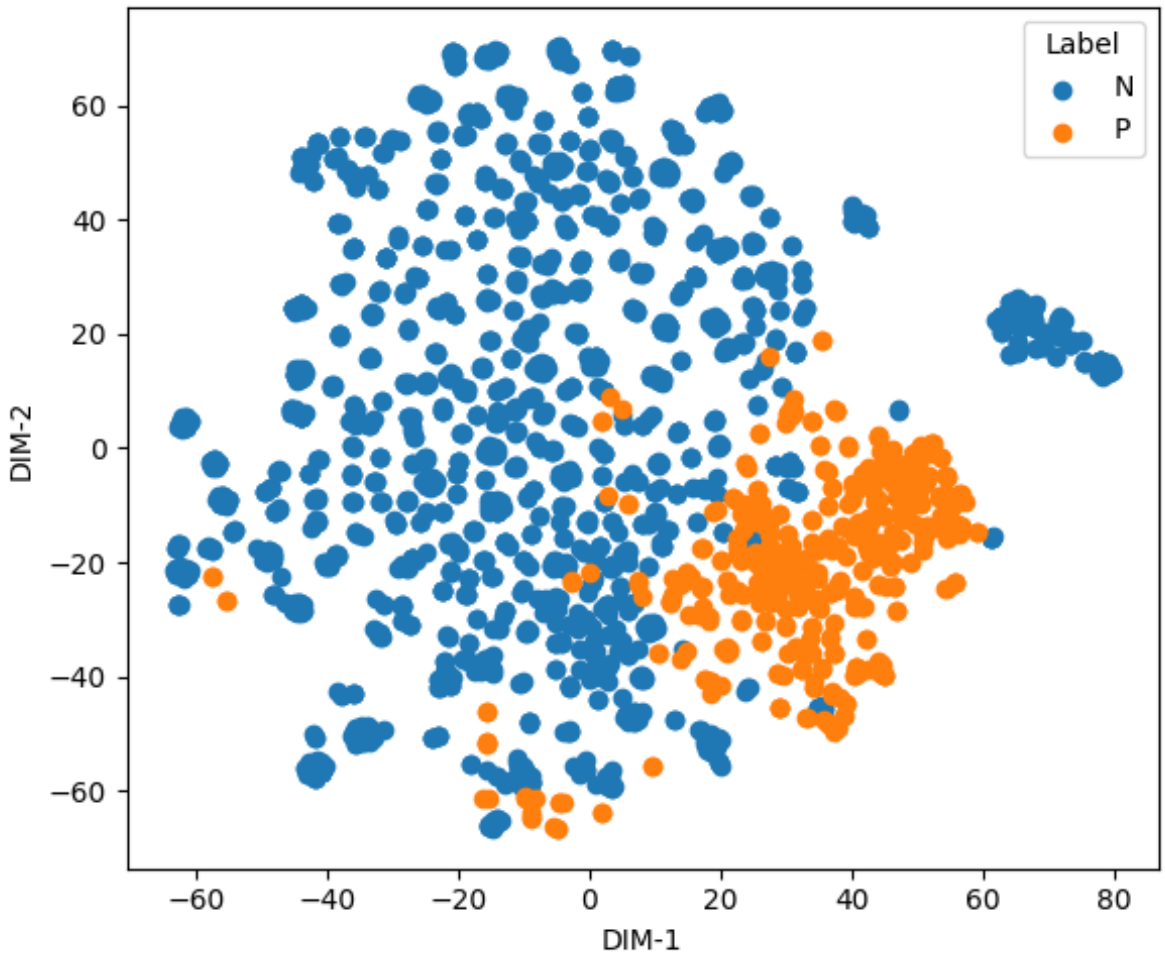

Fig. 1 t-SNE of embeddings of human-positive NES vs Bacteria (negative)

Essentially, it would be enough for the model to learn to distinguish between human and bacterial proteins, with no regards to whether they contain NES proteins. Consequently, we moved to find a more meaningful dataset, which we sourced from Koray Kırlı et al [4]). 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. **We partitioned this data to create our test set**.

2.2 Computational Approach

Our pipeline consists of two main stages:

1. Sequence Representation The full protein sequences were first converted into numerical embeddings using the ESM-2[5] protein language model. Subsequently, sub-sequences corresponding to the known NES motifs (for positives) or randomly selected segments (for negatives) were extracted from these full-sequence embeddings.

2. Model Training and Testing: We trained a Transformer-based classifier (transformer_NES_classifier.py) on the human non-binder dataset. Using the binary cross-entropy loss function. To evaluate whether a sequence contained a NES motif, our method uses a sliding window of 20 amino acids to generate predictions across its entire length. We then take the maximum of those predictions as our score for the entire sequence.

2.3 Code

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

3 Experiments and results

Initial attempt: Human vs. Bacterial Proteins

While our first experiment did not provide directly useful results, it served as a 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). As seen in (Figure 1), a t-SNE visualization of the embeddings shows that the two groups were already highly separable, indicating that this was a trivial classification task and motivating our move to a negative dataset which is in-distribution (human proteins).

Second attempt: Deep Proteomics

We moved to use the human proteome as described in 2.1.

As shown in Figure 2, the Receiver Operating Characteristic (ROC) curve demonstrates the prediction performance of our Transformer classifier on the held-out test set of human non-binders. The model achieved an area under the curve (AUC) of 0.79, indicating good overall ability to distinguish between NES-containing proteins and true non-binding proteins. Notably, the shape of the ROC curve suggests that positive samples (NES-containing proteins) are predicted with higher confidence than negative samples. This result confirms that the classifier successfully captures meaningful patterns for NES detection, while maintaining reasonable specificity.

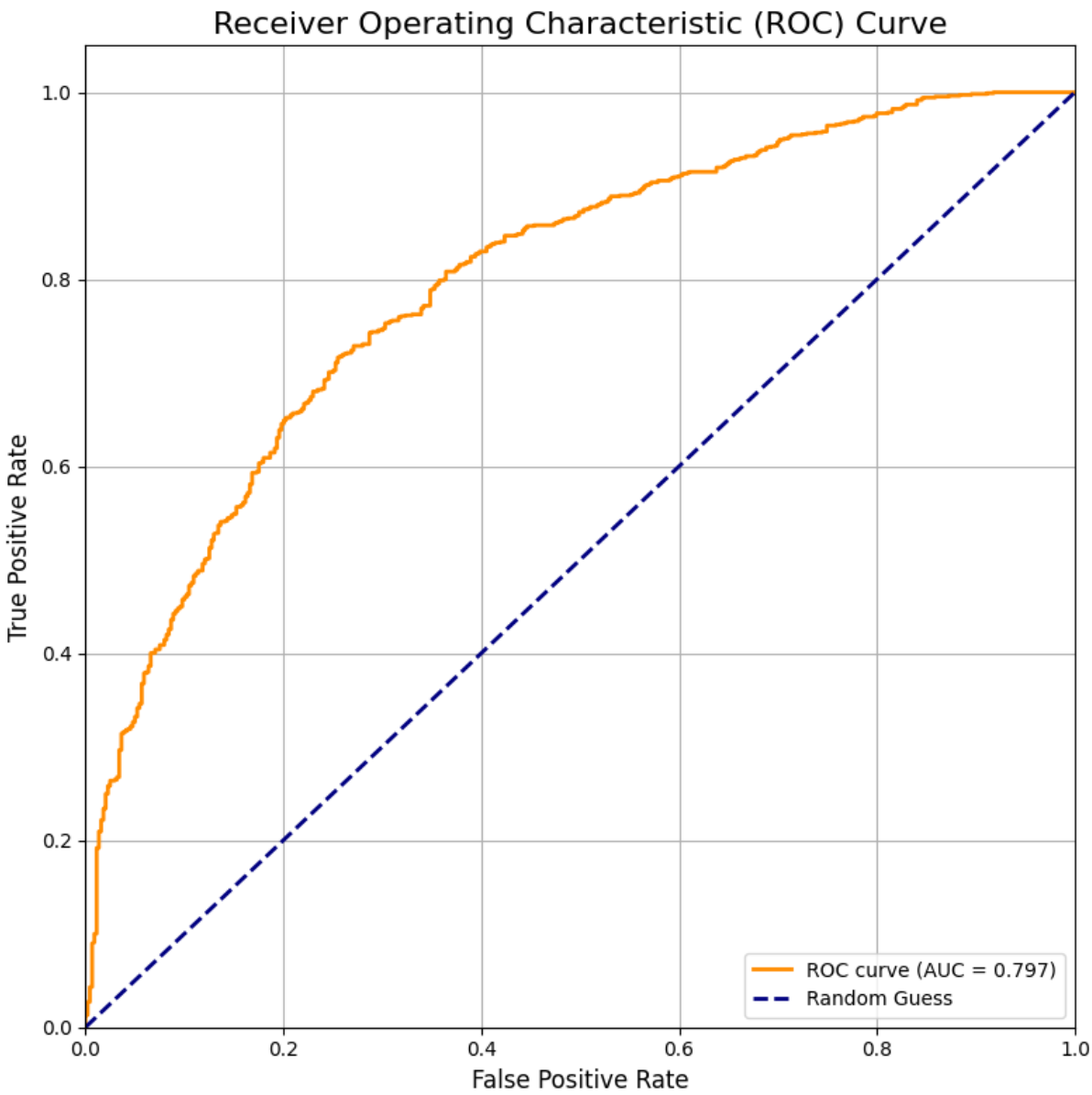

Fig. 2 Figure 2: ROC curve for the final Transformer classifier on the test set of human non-binders.

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 proteins, showed a strong ability to distinguish between positive and negative examples, achieving an AUC of 0.79 on a held-out test set.

4.1 Limitations

The main bottleneck of our project was the acquisition and definition of a high-quality dataset. To train our model, rather than running our model on entire protein sequences, we used sliding windows which cover 20 amino acids, as we expect any NES sequence to fit within that window. We do this as we theorized that feeding the (embedding) of the entire complex will introduce too much noise, in addition to being very demanding computationally. This introduces the issue that we cannot train on complexes which contains

a NES motif if we do not know where exactly in the sequence that motif is, limiting our choice of data. Moreover, while positive data is easy to identify, it is not clear how to produce a high-quality negative data. 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 protein complex" makes creating a perfectly representative negative dataset a significant challenge. Even when we have a protein complex in mind, the choice of which windows of its sequence to choose is ambiguous. While any window would produce a negative sample, as it is guaranteed not to have a NES motif within it, it seems unlikely that choosing at random will produce a high-quality negative sample which will help our model distinguish between hard examples at test time.

5 Future work

We believe the primary avenue for future work is the improvement of the training (and testing) data. This could involve finding other databases of proteins with known subcellular localizations or interaction partners to further refine the set of true non-binders.

We came up with a few ideas to overcome the challenge of finding high-quality negative samples, which we did not attempt due to the small time-scope of the project:

1. Use human proteins which bind to other transporter proteins. This would still require knowing the binding sites so that we can take those as our window, as well as requiring the sites to not be any larger than our sliding window of size 20, so that the separation would not be trivial.

2. Data Augmentation – "mutate" known NES-Sequences to a non-NES-sequence. While we can not know for sure that our "mutation" is guaranteed to produce a non-NES-sequence, it seems unlikely that switching multiple amino acids will still produce a NES protein. Moreover, since we know that the amino acids which bind to CRM1 are hydrophobic, we could change those to guarantee the newly produced sequence will not bind it.

5.1 Broader Implications

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

References

[1] Cour, T., Gupta, R., Rapacki, K., Skriver, K., Poulsen, F., Brunak, S.: Nesbase version 1.0: a database of nuclear export signals. Nucleic acids research **31**, 393–6 (2003). https://doi.org/10.1093/nar/gkg101

[2] Consortium, T.U.: Uniprot: the universal protein knowledgebase in 2025. Nucleic Acids Research **53**(D1), 609–617 (2024) https://academic.oup.com/nar/article-pdf/53/D1/D609/60719276/gkae1010.pdf. https://doi.org/10.1093/nar/gkae1010

[3] van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)

[4] Kırlı, K., Karaca, S., Dehne, H.J., Samwer, M., Pan, K.T., Lenz, C., Urlaub, H., Görlich, D.: A deep proteomics perspective on crm1-mediated nuclear export and nucleocytoplasmic partitioning. eLife **4**, 11466 (2015). https://doi.org/10.7554/eLife.11466

[5] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023) https://www.science.org/doi/pdf/10.1126/science.ade2574. https://doi.org/10.1126/science.ade2574