# A Deep Learning Approach to Residue-Level NES–CRM1 Binding Prediction Using Pretrained Protein Language Models

**Group members:** Amitai Turkel, Yedaaya Zivan, Ori Kitsberg, Yifa Admoni

Github: https://github.com/amitaiturkel/Peptide-Sequence-Structure-Toolkit/


# Abstract

This project focuses on predicting CRM1-binding residues within Nuclear Export Signals (NESs) using deep learning. NESs are short, leucine-rich motifs that direct proteins out of the nucleus via the CRM1 export pathway, and their identification remains challenging due to their short, degenerate sequence patterns and dependence on structural context. We built a residue-level prediction model that combines ESM2 protein language model embeddings with a BiLSTM neural network. Training labels were derived from AlphaFold-predicted CRM1–peptide complexes, with each peptide annotated by its top 5 closest residues to CRM1. Despite modest performance (F1 ≈ 0.45), the model significantly outperformed a random baseline, indicating that pretrained sequence embeddings capture relevant biological signals. Limitations include reliance on predicted structural labels and the lack of explicit 3D input. This work demonstrates the feasibility of sequence-based approaches for modelling molecular interactions and provides a foundation for further refinement in identifying functional export signals.


# 1. Introduction

The export of proteins from the nucleus is a tightly regulated process crucial for cellular function. This export is mediated by short peptide motifs called **Nuclear Export Signals (NESs)**—typically 8–15 residues with spaced hydrophobic amino acids—that are recognized by the export receptor **CRM1 (also known as exportin-1 or XPO1)**. CRM1 binds NES-containing proteins in the nucleus, escorts them through the nuclear pore complex, and releases them into the cytoplasm. Identifying NESs from sequence alone is challenging: their short, degenerate consensus (e.g., LxxxLxxLxL) leads to many false positives, and their functionality depends on structural context such as solvent exposure and binding pocket compatibility.

Despite these difficulties, reliable identification of NESs—and specifically the residues that bind CRM1—is important. CRM1 controls the nuclear-cytoplasmic localization of key regulatory factors, including oncogenes and tumor suppressors and is a validated cancer drug target. Existing computational methods typically combine motif-searching with predicted biophysical features yet still suffer from limited data and false positives. Incorporating structural modelling has shown promise but is computationally expensive.

Here, we propose a **deep learning method** that predicts CRM1-binding residues directly from sequence. We leverage **ESM-2 protein language model embeddings**, which encode evolutionary and structural context, and a **BiLSTM classifier** to make residue-level predictions. Training labels are derived from AlphaFold-generated

CRM1–NES complexes. Our aim is a scalable, sequence-based alternative to motif and structure-centered methods.

# 2. Methods

## 2.1 Dataset Preparation

Our dataset consists of structural complexes containing NES peptides bound to CRM1, sourced from AlphaFold structural predictions (as generated in Exercise 4). The data processing pipeline extracts peptide sequences and generates residue-level binding labels through the following steps:

**Structure Processing**: We implemented a structure parser using BioPython that supports both PDB and mmCIF formats. In each complex, the CRM1 chain and the NES peptide chain were identified based on structural heuristics: according to the exercise, chain **B** was designated as the peptide, while chain **A** was selected as the CRM1 receptor, identified as the longest chain in the structure.

**Sequence Extraction**: Peptide sequences were extracted by converting three-letter amino acid codes to single-letter codes. Only standard amino acids were retained; any structures containing non-standard residues were filtered out to ensure consistency and compatibility with downstream embedding models.

**Binding Site Identification**: For each peptide residue, we computed the minimum heavy-atom distance (excluding hydrogen atoms) to any heavy atom in the CRM1 structure. We then applied a top-$k$ labelling strategy, where the $k$ residues with the smallest distances were labelled as **binding** (label = 1), and the rest as **non-binding** (label = 0). We used $k = 5$, consistent with the five hydrophobic pockets of CRM1 known to interact with NES motifs [Dong et al., 2009; Fung et al., 2017].

This top-$k$ approach was chosen for its simplicity and alignment with biological knowledge. We also considered a **distance thresholding** strategy (e.g., labelling residues within a fixed radius), but top-$k$ labelling ensured that exactly 5 residues per peptide were consistently labelled, as we assumed really happened.

**Data Validation**: We applied quality control checks, including ensuring exactly *5* residues were labelled per peptide, removing structures with parsing errors, and excluding samples with missing chains or formatting issues.

**Label Format**: For each peptide, the dataset contains:

- The amino acid sequence (length *L, different to each* sequence)
- A binary label vector of length $L$, where 1 indicates a binding residue and 0 otherwise

## 2.2 ESM Embedding Extraction

We leverage the ESM2 family of protein language models to generate high-dimensional residue embeddings that capture evolutionary and structural information relevant to protein binding.

**Model Selection**: Our pipeline supports all ESM2 model variants, with embedding dimensions ranging from 320 to 5120.

The model variant is configurable via a single parameter (`embedding_size`) in the pipeline.

**Layer Selection**: ESM2 models contain 6–48 transformer layers depending on size. Following previous literature, we extracted embeddings from layers 6, 8, and 20 for comparative experiments. Layer selection is handled via a configurable `layer` parameter.

**Batch Processing**: Sequences are tokenized using the `batch_converter` utility provided by ESM's API. This ensures consistent formatting and padding. Embeddings are extracted in a single batch and sliced to match sequence lengths.

**Embedding Caching**: To avoid redundant computation, embeddings are saved to disk after their first generation and reused across future runs. This caching mechanism significantly accelerates model development and hyperparameter tuning.

## 2.3 Model Architecture

Our binding prediction model integrates deep sequence modelling with residue-level classification to predict CRM1 binding sites from ESM-derived embeddings.

**Bidirectional LSTM**: The core of the model is a bidirectional Long Short-Term Memory (BiLSTM) network, which captures both forward and backward context in the peptide sequence. This is essential because the binding behaviour of a residue can depend on its surrounding amino acids.

**Architecture Parameters**:

- Input dimension: Matches the selected ESM embedding size (320–5120).
- Hidden dimension: Configurable (we tested values in the range 64–256).
- Output dimension: Twice the hidden dimension (due to bidirectionality).

**Classification Head**: A single linear layer maps each residue's hidden representation to a scalar logit, which is converted into a probability using a sigmoid activation function.

**Regularization**:

**Weight decay:** Applied via the Adam optimizer to reduce overfitting.

**Output Layer**: The model produces a probability for each residue using sigmoid activation. During training, predictions are compared to binary labels using binary cross-entropy loss. During evaluation and inference, residues are classified as binding or non-binding by selecting the top 5 scoring residues, ensuring consistency with the labelling strategy.

## 2.4 Training Procedure

We trained the model to perform residue-level binary classification using ESM-derived embeddings as input.

**Loss Function**: We use **Binary Cross-Entropy Loss (BCE)** with a sigmoid activation to compare predicted binding probabilities to ground truth labels. The loss is computed only on valid (unpadded) sequence positions using per-sample masks.

**Class Imbalance Handling**: NES binding sites represent a minority class (typically 20-30% of residues in a peptide), leading to class imbalance. We address this through:

1. **Positive Weighting**: Automatic calculation of positive class weights based on the training data distribution
2. **Masked Loss**: Proper handling of variable-length sequences through attention masking

**Optimization**: We optimized the model using the **Adam optimizer** with a learning rate of 1e-3.

**Training Loop**: The training procedure includes:

- Training was conducted using mini-batch gradient descent.
- For each batch, we computed the masked binary cross-entropy loss and updated model parameters.

**Evaluation:** After training, model performance was assessed on a held-out test set. Since labels always contain exactly five positives per sequence, we adjusted predictions at test time to also select the top 5 residues with the highest predicted probabilities.

**Hyperparameter Optimization**: We implement a comprehensive grid search framework that systematically explores:

We implemented a **grid search** to evaluate combinations of:

- ESM model variants (e.g., ESM2-t30-150M, ESM2-t33-650M)
- Embedding layers (6, 8, 20)
- LSTM hidden dimensions
- Batch sizes and number of training epochs

The grid search generates unique model identifiers and saves results in structured CSV format for analysis, and we selected the models with the best performance.

# 3. Results

## 3.1 Evaluation Metrics

To evaluate our model's performance in predicting CRM1-binding residues from NES peptides, we adopted common classification metrics at the **residue level**:

- **Precision**: The proportion of residues predicted as binding (label = 1) that are actually binding in the ground truth. High precision indicates low false positives.

- **Recall (Sensitivity)**: The proportion of actual binding residues that the model correctly identified. High recall ensures critical residues are not missed.

- **F1 Score**: The harmonic mean of precision and recall. This metric is particularly important in our imbalanced dataset, where binding residues are a minority.

- **Support**: The number of positive (binding) residues in the test set. This contextualizes performance and metric reliability.

All metrics were calculated on the **residue level**, using a **top-5 prediction scheme**: for each peptide, the five residues with the highest predicted probabilities were labeled as "binding" (1), and the rest as non-binding (0). This strategy reflects our dataset design, where each peptide was annotated with exactly 5 CRM1-binding residues.

Although we also logged **binary cross-entropy loss** during training, we found it to be a weak indicator of final model utility. Since BCE averages error across all residues, it is strongly influenced by the dominant class (non-binding residues) and does not guarantee correct identification of the top-5 residues in each peptide. Therefore, F1 and precision/recall were more informative for downstream performance.

## 3.2 Quantitative Results

We first trained our models without any class weighting. After running a systematic grid search over embedding sizes, ESM layers, epochs, and batch sizes, we found that the best model **without class weights** used:

  - ESM embedding size: **1280**
  - ESM layer: **6**
  - Epochs: **20**
  - Batch size: **16**

- This configuration achieved:
  - **Precision**: 0.45
  - **Recall**: 0.45
  - **F1 Score**: 0.45
  - (Residue-level support: 2159 test residues)

Despite reasonable performance, we observed that the model tended to favor the majority class (non-binding), especially when predicting on longer sequences. This limitation motivated the use of **class weighting** to emphasize the minority class (binding residues) during training.

After incorporating **class weighting** into the loss function (using PyTorch's pos_weight in BCEWithLogitsLoss), we observed:

- Improved recall across most settings
- Slight decrease in precision
- Increased average F1 score in several configurations
- Generally **better identification of true binding residues**, especially in peptides with weaker signal

However, not all configurations improved. For some hyperparameter settings, the added emphasis on positive examples led to more false positives. Thus, tuning remained important even with class weighting.

## 3.3 Per-Peptide Performance Comparison

To complement the residue-level metrics, we performed a per-peptide analysis focusing on how many binding residues (out of 5) were correctly predicted for each peptide.

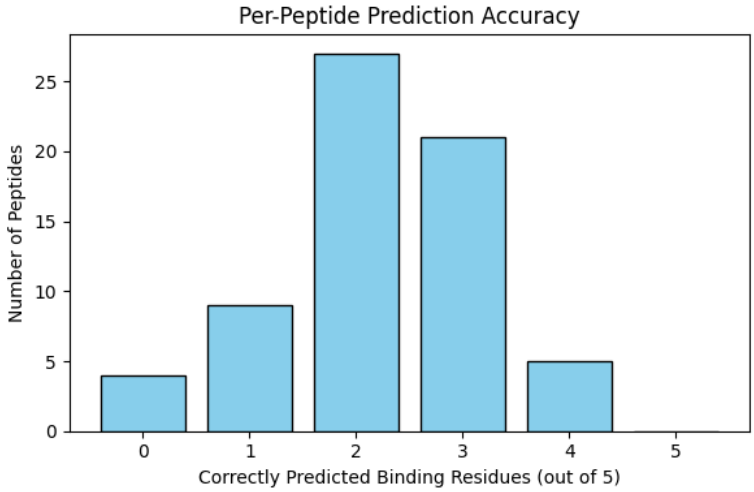
**Mean number of correctly predicted residues per peptide:**

- Without class weighting: 2.27
- With class weighting: 2.21

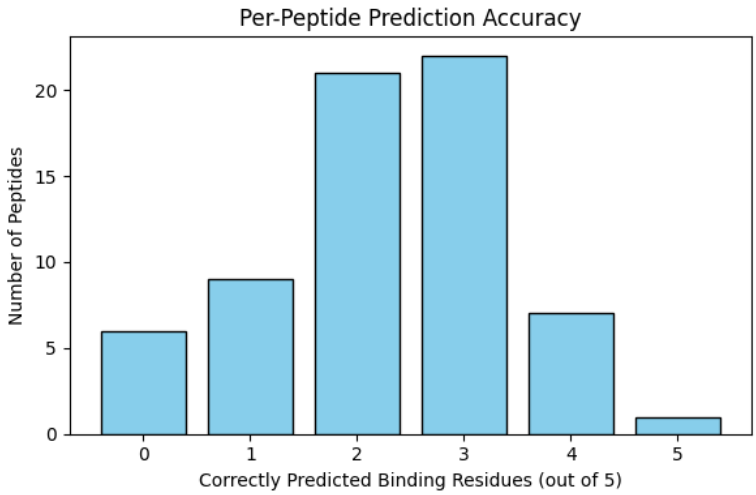
This small difference indicates that both models perform similarly at the peptide level, despite having different residue-level metrics.

The table below shows how many peptides had a given number of correct predictions (out of 5):

**Weight:**



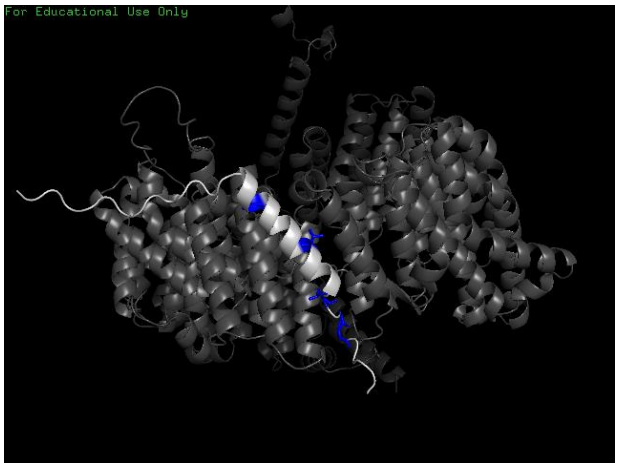

**without weight:**

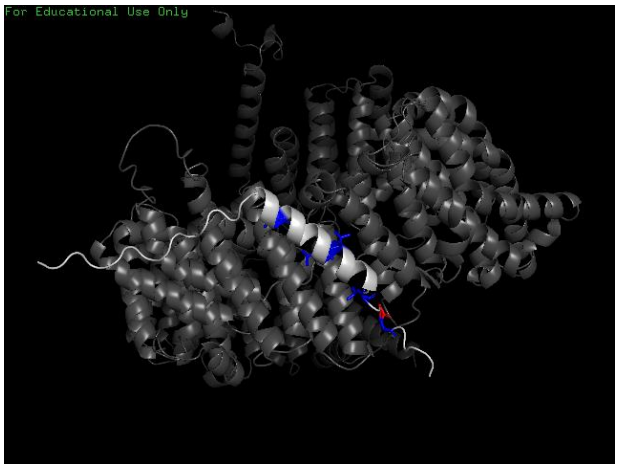



These results show:

- Most peptides had 2–3 correct predictions regardless of the model.
- The model without class weights had a slightly higher count of peptides with 4–5 correct predictions, suggesting better performance in high-confidence cases.
- The model with class weights tended to shift predictions slightly toward recall, but with a small trade-off in per-peptide precision.

**Example:** CRM1–NES Complex Visualization (PDB: pos_FA68_E357_al)




True binding residues are shown in red, and predicted residues are shown in blue. In this example, the unweighted model correctly predicted all 5 binding residues (left image), while the weighted model predicted 4 out of 5 correctly (right image).

# 4. Discussion

## 4.1 Performance Interpretation

While our model achieved a residue-level F1 score of approximately **0.45**, this performance is modest in absolute terms. However, it's important to contextualize this result relative to a **random baseline**. In our setup, each peptide contains exactly 5 true binding residues, and the model is constrained to predict 5 residues per peptide. For a typical peptide of 25 residues, randomly selecting 5 residues would yield an expected precision and recall of only **0.20**, resulting in an F1 score around **0.20**. Therefore, an F1 score of **0.45** indicates that the model is capturing meaningful signal from the sequence — more than **double** the performance of chance.

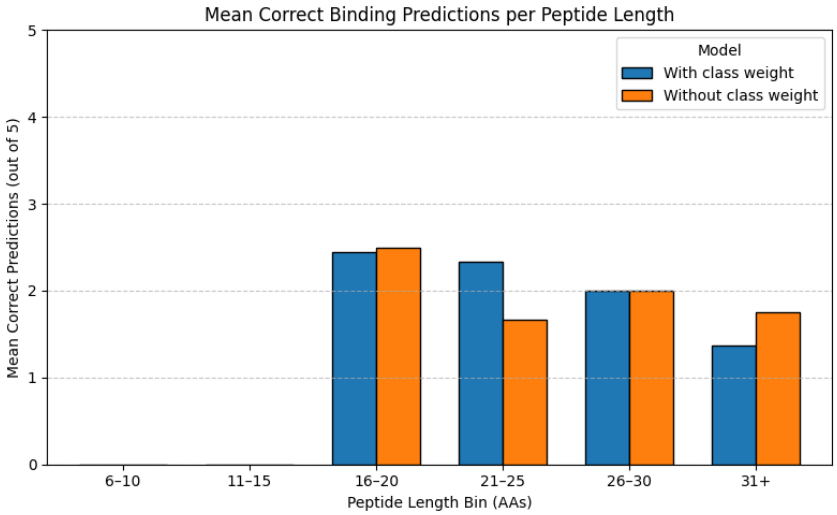
Nonetheless, the model's performance leaves substantial room for improvement. While it learns discriminative patterns from the amino acid sequence alone, the prediction accuracy remains far from perfect, especially in longer or more complex peptides. These limitations suggest that **sequence-based features alone may not fully capture the structural or contextual information required to accurately identify CRM1 binding residues**. Future iterations could benefit from incorporating structural features (e.g., AlphaFold coordinates), motif-aware priors, or context from known NES consensus patterns.

## 4.2 Length-Dependent Performance Drop

An analysis of prediction accuracy as a function of peptide length revealed a moderate negative correlation, indicating that model performance declines as peptide length increases. As shown in the figure below, the mean number of correctly predicted binding residues decreases for longer peptides in both the unweighted and class-weighted models. Specifically, we observed Pearson correlation coefficients of –0.32 (no class weighting) and –0.42 (with class weighting), confirming this downward trend.

This behavior likely reflects the increased challenge of selecting the correct five NES residues from a longer sequence, given that CRM1 accommodates only five hydrophobic pockets. Because our model enforces a fixed top-5 prediction strategy—aligned with biological constraints—it must choose five candidates regardless of sequence length. This increases the chance of false positives in longer peptides, where non-binding residues vastly outnumber true binders.

These results highlight a limitation of fixed-top-k strategies under structural constraints. Future work could explore incorporating position-aware mechanisms or motif-informed priors to improve residue selection, especially in longer and more structurally diverse peptides.




## 4.3 What Worked and What Didn't

While the overall performance of the model was modest (F1 ≈ 0.45), several aspects of the architecture and input features contributed to its ability to learn meaningful patterns. The use of ESM embeddings provided a rich representation of the sequence, enabling the model to outperform random guessing by a substantial margin. The BiLSTM architecture handled variable-length peptides effectively, capturing both local and broader sequence dependencies. Additionally, enforcing a top-k prediction constraint during inference helped align predictions with the known biological mechanism — where exactly five residues engage CRM1 binding pockets — and maintained consistency with the labelling strategy.

However, we found that the model still struggled with ambiguous sequence contexts, particularly in longer peptides. While class weighting partially improved recall, the inherent class imbalance — with only five binding residues per sequence — reduced the signal-to-noise ratio and made learning discriminative features difficult. The lack of explicit structural context may also limit the model's capacity to resolve spatially distant interactions that appear similar at the sequence level.

## 4.4 Limitations

Several limitations should be considered when interpreting these findings:

1. **Label Accuracy:** Our labels are derived from AlphaFold predictions, which may not accurately capture transient or weak CRM1 interactions. Manual curation or experimental structures would improve training signal.
2. **No Structural Input:** The model does not see 3D geometry directly during the training. Integrating structural representations (e.g., from AlphaFold-Multimer or geometric deep learning) could improve spatial reasoning.
3. **Binary Binding Assumption:** We assume exactly five residues bind in each peptide, aligned with CRM1's known pocket structure. In reality, some interactions may be weaker or more diffuse, which our labels do not reflect.
4. **Peptide Diversity:** Our dataset may not reflect the full diversity of NES sequences across species or contexts.

### 4.5 Opportunities for Improvement

With more time or data, several directions could enhance performance:

- **Structure-Aware Models:** Incorporate 3D features or distance maps alongside sequence embeddings to provide direct geometric context.
- **Multi-Class or Continuous Labels:** Instead of binary labels, predict per-residue binding confidence or pocket-specific contacts (Φ0–Φ4), allowing richer supervision.
- **Motif-Based Priors:** Integrate known NES motif patterns (e.g., Φ–X2–3–Φ–X2–3–Φ) into the model or loss function to guide learning.
- **Calibration of Probabilities:** Calibrate predicted probabilities and experiment with soft thresholds or expected number of binders, rather than always selecting the top

### 4.6 Implications

Despite modest overall performance (F1 ≈ 0.45), our results demonstrate that pretrained protein language models like ESM2 capture biologically relevant sequence patterns sufficient for predicting CRM1-binding residues. The findings suggest that language models contain useful evolutionary and motif information, but explicit structural context may be necessary to resolve spatial ambiguity. Our pipeline provides a generalizable framework for residue-level prediction tasks and highlights both the potential and limitations of sequence-only approaches in molecular interaction modelling.

## 5. References

**Dong, X. et al. (2009)**. "Structural basis for leucine-rich nuclear export signal recognition by CRM1." *Nature*, 458(7242), 1136–1141. https://www.nature.com/articles/nature07975

**Lee, B. J. et al. (2019)**. "CRM1 recognizes diverse nuclear export signals." *Cell*, 177(5), 1228–1242. https://elifesciences.org/articles/23961

**Fung, H. Y. J. et al. (2017)**. "Structural basis of CRM1 inhibition by selective inhibitors of nuclear export." *Nature*, 550(7676), 491–495. https://elifesciences.org/articles/23961

**Parikh, K. et al. (2014)**. "Selective inhibition of nuclear export with selinexor enhances anti-tumor activity and overcomes therapy resistance in multiple myeloma." *Blood*, 124(5), 1112–1121. https://ashpublications.org/blood/article/13/2/192/33344/Editorial-he-Di-Guglielmo-Syndrome

**Jumper, J. et al. (2021)**. "Highly accurate protein structure prediction with AlphaFold." *Nature*, 596(7873), 583–589. https://www.nature.com/articles/s41586-021-03819-2

**Xu, D. et al. (2012)**. "LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins." *Bioinformatics*, 28(10), 1357–1364. https://academic.oup.com/bioinformatics/article/31/9/1357/200064

**Rives, A. et al. (2021)**. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *PNAS*, 118(15), e2016239118. https://www.pnas.org/doi/10.1073/pnas.2016239118

**Lin, Z. et al. (2023)**. "Language models of protein sequences at the scale of evolution enable accurate structure prediction." *Nature Biotechnology*, 41(4), 546–553. https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1

**Hanson, J. et al. (2018)**. "Accurate prediction of protein contact maps using deep residual neural networks." *Bioinformatics*, 34(23), 4039–4045. https://academic.oup.com/bioinformatics/article/34/23/4039/5040307

**Gainza, P. et al. (2020)**. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning." *Nature Methods*, 17(2), 184–192. https://www.nature.com/articles/s41592-019-0666-6