# NLSeer: Nuclear Localization Signal Prediction Project

Erdős Data Science Boot Camp Spring 2024

Scott Auerbach, Ukamaka Nnyaba, Ming Zhang, Yingyi Guo,
Hemaa Selvakumar, Çisil Karagüzel

April 26, 2024

# Why nuclear localization signals?

- Nuclear localization signals are sequences within a protein that directs it towards the nucleus
- Cellular processes require certain proteins to move to nucleus at the right time
- Losing/gaining this ability dysregulates these processes and can lead to cancer (example: TRPS1 in triple negative breast cancer)
- Goal: For any given protein, identify potential NLS

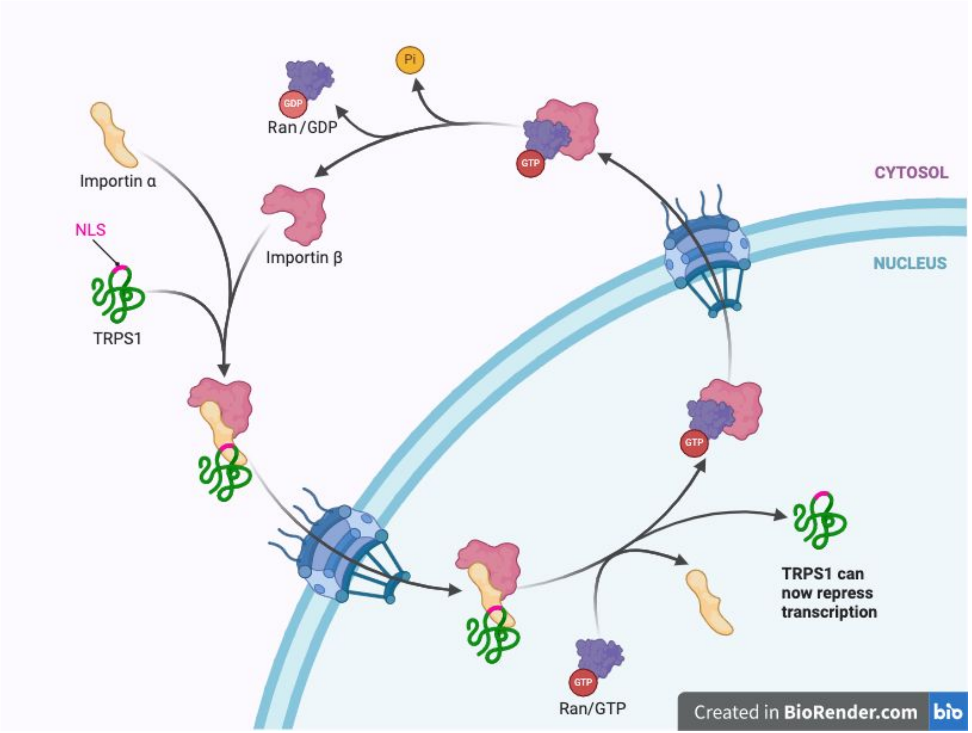
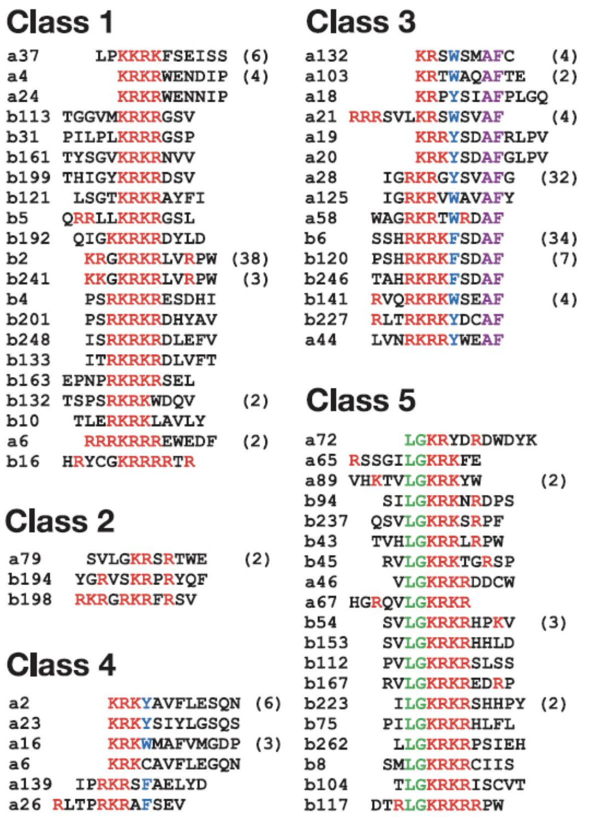


Image created with BioRender.com

# What does a NLS look like?

- Set of amino acids represented by letters, can vary in length from several to couple dozen
- They can vary considerably, but tend to have fairly consistent patterns, thus making them a potent feature for prediction tools



### Class 1

| | | |
|---|---|---|
| a37 | LPKKRKFSEISS | (6) |
| a4 | KRKRWENDIP | (4) |
| a24 | KRKRWENNIP | |
| b113 | TGGVMKRKRGSV | |
| b31 | PILPLKRRRGSP | |
| b161 | TYSGVKRKRNVV | |
| b199 | THIGYKRKRDSV | |
| b121 | LSGTKRKRAYFI | |
| b5 | QRRLLKRKRGSL | |
| b192 | QIGKKRKRDYLD | |
| b2 | KRGKRKRLVRPW | (38) |
| b241 | KKGKRKRLVRPW | (3) |
| b4 | PSRKRKRESDHI | |
| b201 | PSRKRKRDHYAV | |
| b248 | ISRKRKRDLEFV | |
| b133 | ITRKRKRDLVFT | |
| b163 | EPNPRKRKRSEL | |
| b132 | TSPSRKRKWDQV | (2) |
| b10 | TLERKRKLAVLY | |
| a6 | RRRKRRREWEDF | (2) |
| b16 | HRYCGKRRRRTR | |

### Class 2

| | | |
|---|---|---|
| a79 | SVLGKRSRTWE | (2) |
| b194 | YGRVSKRPRYQF | |
| b198 | RKRGRKRFRSV | |

### Class 4

| | | |
|---|---|---|
| a2 | KRKYAVFLESQN | (6) |
| a23 | KRKYSIYLGSQS | |
| a16 | KRKWMAFVMGDP | (3) |
| a6 | KRKCAVFLEGQN | |
| a139 | IPRKRSFAELYD | |
| a26 | RLTPRKRAFSEV | |

### Class 3

| | | |
|---|---|---|
| a132 | KRSWSMAFC | (4) |
| a103 | KRTWAQAFTE | (2) |
| a18 | KRPYSIAFPLGQ | |
| a21 | RRRSVLKRSWSVAF | (4) |
| a19 | KRRYSDAFRLPV | |
| a20 | KRKYSDAFGLPV | |
| a28 | IGRKRGYSVAFG | (32) |
| a125 | IGRKRVWAVAFY | |
| a58 | WAGRKRTWRDAF | |
| b6 | SSHRKRKFSDAF | (34) |
| b120 | PSHRKRKFSDAF | (7) |
| b246 | TAHRKRKFSDAF | |
| b141 | RVQRKRKWSEAF | (4) |
| b227 | RLTRKRKYDCAF | |
| a44 | LVNRKRRYWEAF | |

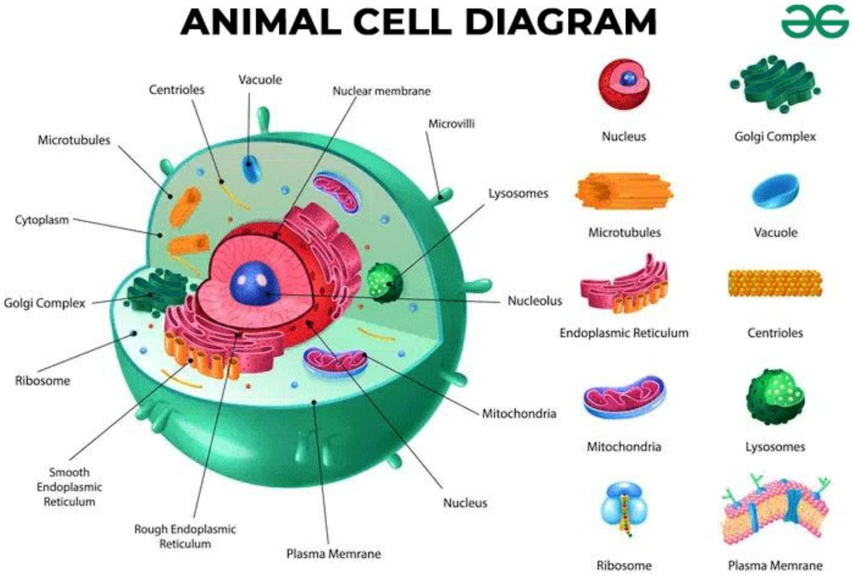
### Class 5

| | | |
|---|---|---|
| a72 | LGKRYDRDWDYK | |
| a65 | RSSGILGKRKFE | |
| a89 | VHKTVLGKRKYW | (2) |
| b94 | SILGKRKNRDPS | |
| b237 | QSVLGKRKSRPF | |
| b43 | TVHLGKRRLRPW | |
| b45 | RVLGKRKTGRSP | |
| a46 | VLGKRKRDDCW | |
| a67 | HGRQVLGKRKR | |
| b54 | SVLGKRKRHPKV | (3) |
| b153 | SVLGKRKRHHLD | |
| b112 | PVLGKRKRSLSS | |
| b167 | RVLGKRKREDRP | |
| b223 | ILGKRKRSHHPY | (2) |
| b75 | PILGKRKRHLFL | |
| b262 | LLGKRKRPSIEH | |
| b8 | SMLGKRKRCIIS | |
| b104 | TLGKRKRISCVT | |
| b117 | DTRLGKRKRRPW | |

Source for image:
https://www.novoprolabs.com/support/articles/will-nuclear-localization-signal-nls-be-removed-after-protein-maturation-201902221563.html

# Datasets

- Main dataset is curated collection of proteins with NLS compiled by Yamagishi et al.
  - ~1,400 proteins, some of which had multiple NLS inside the same sequence
  - This dataset was later padded with an additional 1,400 randomly selected non-NLS proteins from UniProt
- For later classification, additional dataset from DeepLoc  (another prediction tool for general localization) was used
  - This DeepLoc dataset contained several types of sorting signals, including NLS

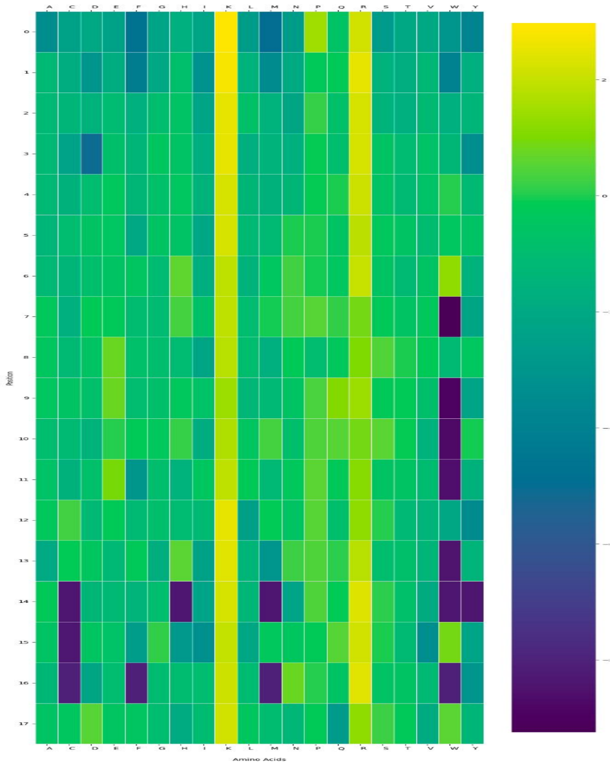


Source for image: https://www.geeksforgeeks.org/animal-cell/

# Methodology: Position specific scoring matrix (PSSM)

- PSSM with 18 rows and 20 columns for amino acid scoring in NLSs.
- Utilizing PSSM to evaluate the likelihood of a protein region starting a NLS.
- Applying traditional machine learning models like Random Forest to predict NLS presence based on PSSM scores.

NLS PSSM heatmap



# Methodology: Convolutional Neural Network (CNN) + XGBoost

- One-hot encoding amino acids: Representing amino acids as binary values to indicate the presence or absence of desired signals.
- Enhancing feature sets with PSSM scores for a richer analysis.
- Processing the encoded data through a Convolutional Neural Network and then combining it with XGBoost for refined predictions.

# Methodology: Long Short-term memory (LSTM) model

- LSTM is a variant of recurrent neural network (RNN) that is capable of learning dependencies.
- LSTM has edge over the conventional feed-forward NN and RNN because it selectively remembers patterns for long duration of time.

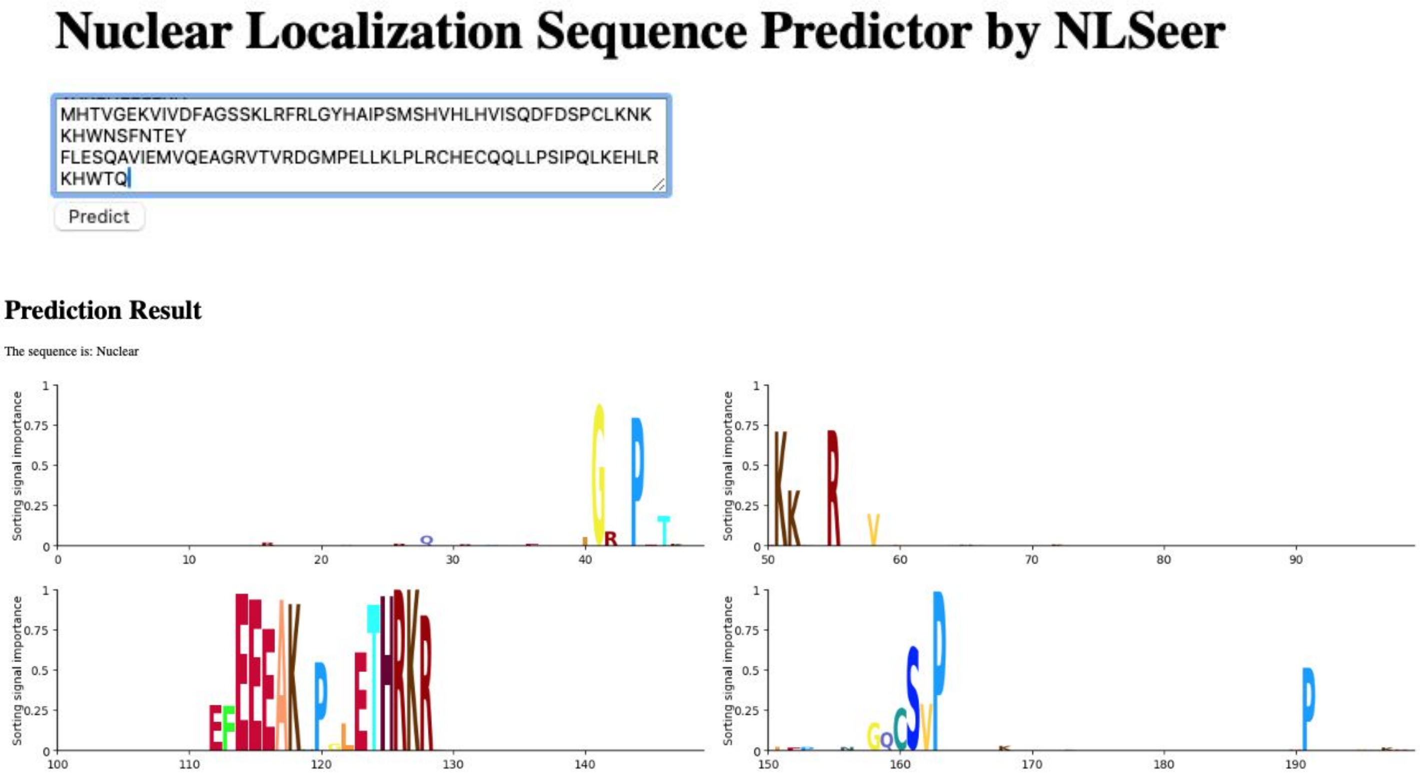
## Architecture of the Model

```
LSTMModel(
  (lstm): LSTM(20000, 200)
  (fc): Linear(in_features=200,
out_features=1, bias=True)
  (sigmoid): Sigmoid()
)
```

# Results: Accuracy Table

| **Method** | **Training_Test Dataset**<br>(main dataset) | **Holdout Dataset**<br>(DeepLoc dataset) |
|---|---|---|
| PSSM + Random Forest | 76% | 89% |
| PSSM + CNN + XGBoost | 91% | 73.5% |
| LSTM | 91.2% | 85.5% |

# Flask Web Application: Interface and Results



# Nuclear Localization Sequence Predictor by NLSeer

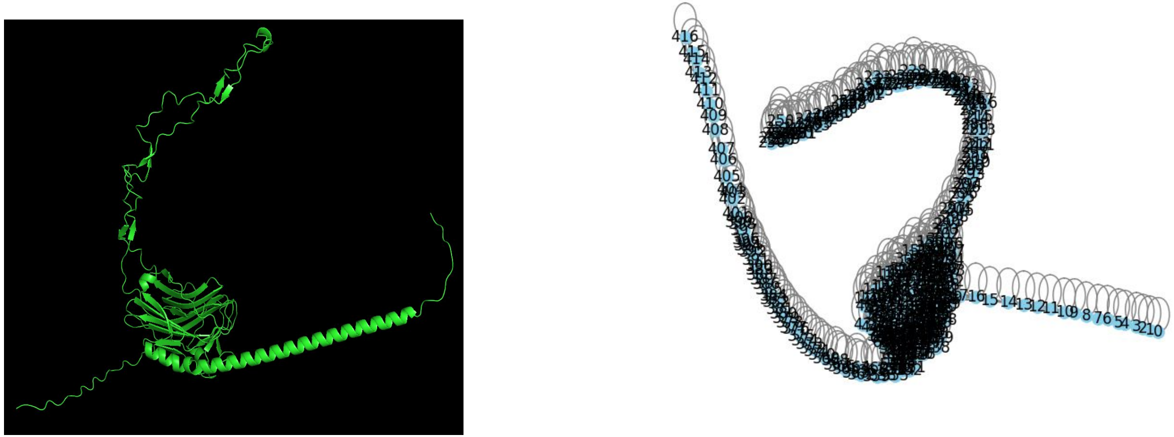
```
MHTVGEKVIVDFAGSSKLRFRLGYHAIPSMSHVHLHVISQDFDSPCLKNK
KHWNSFNTEY
FLESQAVIEMVQEAGRVTVRDGMPELLKLPLRCHECQQLLPSIPQLKEHLR
KHWTQ
```

## Prediction Result

The sequence is: Nuclear

# Future Directions

- Graph convolutional network to incorporate vectorized nodes of  graph embedded 3-D structures as feature alongside PSSM and sequence
- ProtBERT - pretrained NLP model adapted to parse through amino acid sequences without prior labelling



AlphaFold 3-D structure of calreticulin (left) and its graph embedding generated by Networkx (right)

# Acknowledgements:

- Roman Holowinsky
- Steven Gubkin
- Alexis Johnson
- Rost Lab at the Technical University of Munich
- Ryosuke Yamagishi and Hiroki Kaneko at Nihon University
- Danish Department of Health Technology (DeepLoc)