Supplementary File for

Discovering nuclear localization signal universe through a novel deep language learning network with attention as the neurons

Yi-Fan Li, Xiaoyong Pan, and Hong-Bin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China

Contents

List of Supplementary Figures

List of Supplementary Tables

Existing Nuclear Localization Signal (NLS) predictors

Supplemental Table 1. Overview of existing Nuclear Localization Signal (NLS) predictors.

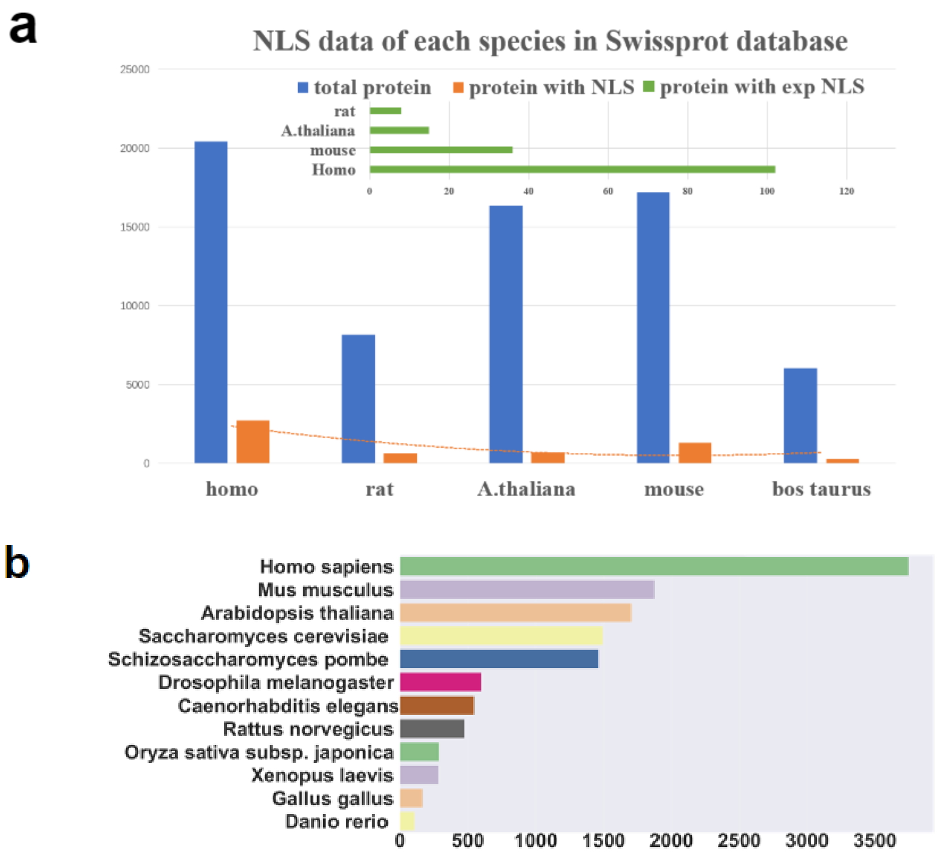
Method	Website	Base Model	published time
cNLS Mapper[1]	http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi	Template matching	2009
PSORT II[2]	https://psort.hgc.jp/form2.html	Machine learning	1999
PredictNLS[3]	https://www.predictprotein.org/	Electronic mutation	2000
NLStradamus[4]	http://www.moseslab.csb.utoronto.ca/NLStradamus	Machine learning	2009
NucImport[5]	https://omictools.com/nucimport-tool	Machine learning	2011
SeqNLS[6]	http://mleg.cse.sc.edu/seqNLS	Frequent patterns	2013
INSP[7]	http://www.csbio.sjtu.edu.cn/bioinf/INSP/	Machine learning	2020

Template matching-based method, i.e. cNLS Mapper, assesses the cumulative score of amino acid segments based on their individual activities as outlined in the activity profile. This profile, derived from diverse standard NLS templates, calculates the total score following the additivity principle, where each amino acid's score is treated independently. Consequently, the NLS score is obtained through the summation of amino acid activity scores across different positions.

Machine learning-based method represents another effective approach for NLS prediction. Current methods employ classic machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbors, and Bayesian networks, to establish NLS scoring models. These models typically predict NLS based on protein sequence characteristics such as NLS patterns, alkaline amino acid content, and amino acid frequency distribution.

Electronic mutation-based method involves the initial deletion or substitution of segments within experimentally validated NLS sequences. Subsequently, these mutated protein sequences are assessed to determine whether they retain the ability to enter the nucleus, thereby confirming the functionality of the mutated NLS in mediating sequence entry into the nucleus.

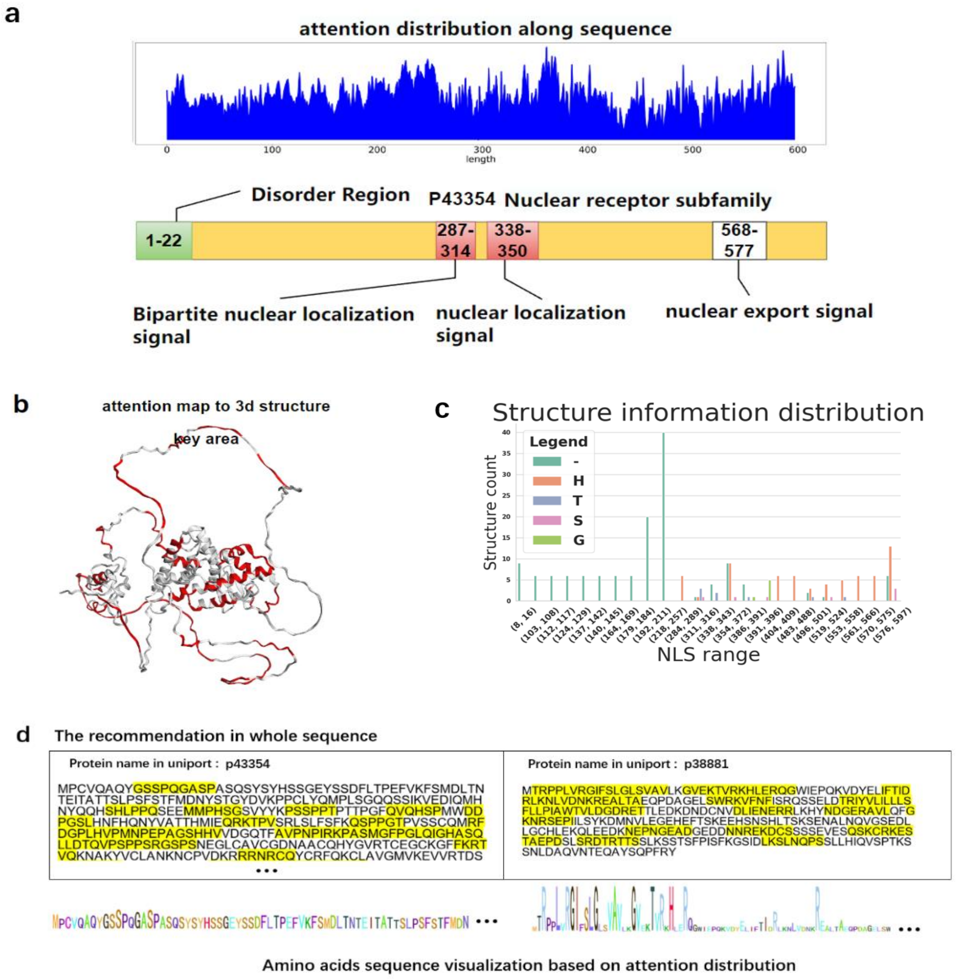
Dataset of different species



Supplemental Figure 1. Nuclear Localization Signal (NLS) landscape in SwissProt and nuclear-localized proteins across the top 12 species. (**a**) the scarce landscape of experimentally validated NLS data in SwissProt, displaying only the top 4 instances (**b**) Dominant nucleus-localized proteins in the top 12 species according to SwissProt.

The Function of NLSExplorer

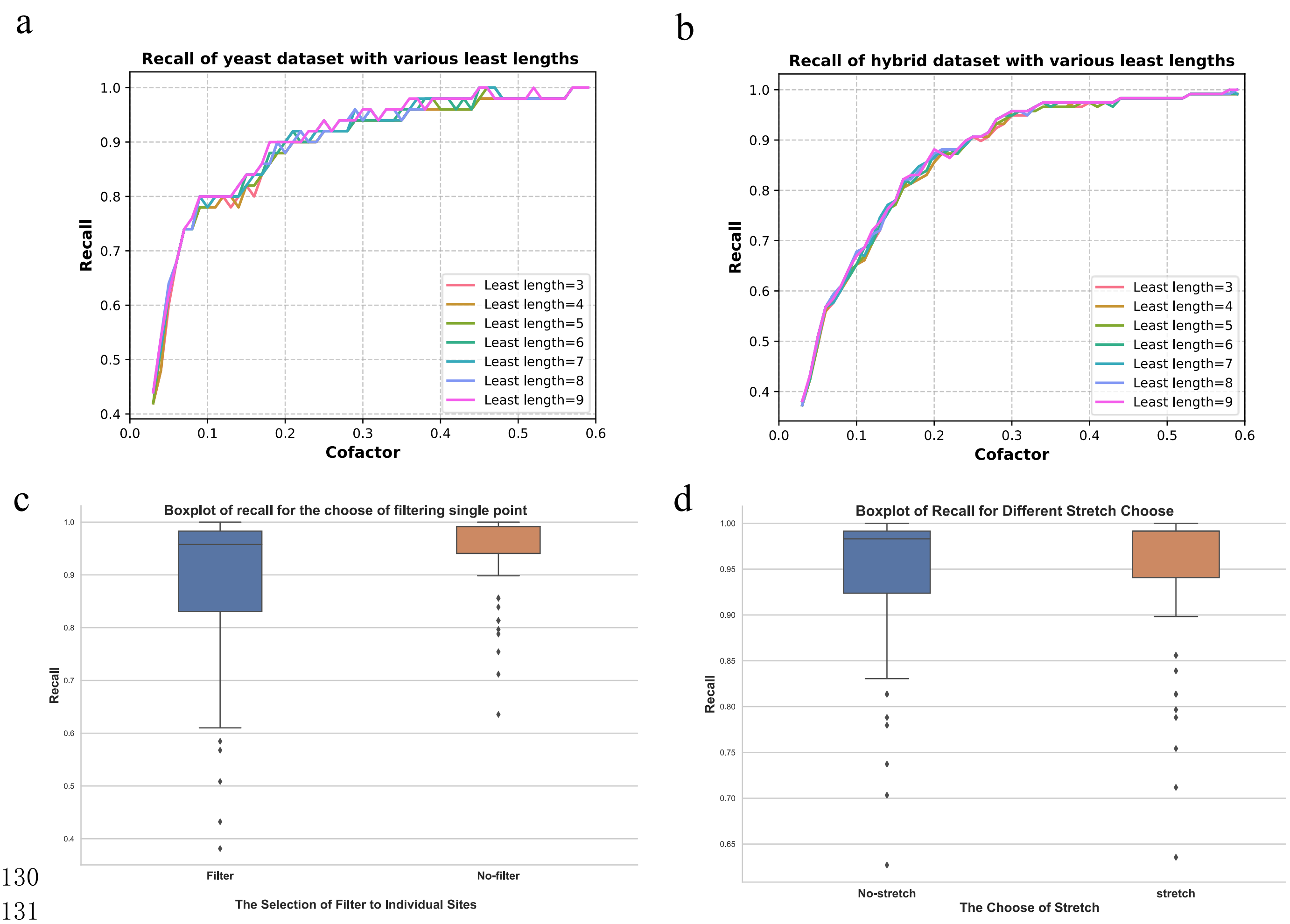
Starting from the attention pattern distribution, NLSExplorer derives an attention distribution map that exhibits a significant correlation with specific patterns crucial for nuclear localization (Supplemental Figure 2-a). Additionally, if the prediction process involves 3D structural information, the model provides users with the three-dimensional structural representation of the recommended segments (Supplemental Figure 2-b) and showcases statistical insights in the structural domain (Supplemental Figure 2-c) . Ultimately, our model will generate a visual representation of the sequence image by incorporating attention weights to the height of amino acids and the recommendations from the entire segment (Supplemental Figure 2-d). In summary, our attention module effectively recapitulates segments with a substantial impact on nuclear localization prediction through the distribution of attention.



Supplemental Figure 2. The functions of NLSExplorer. (**a**) NLSExplorer displays attention across the entire protein sequence. (**b**) NLSExplorer can illustrate recommended segments in the protein 3D structure. (**c**) NLSExplorer provides structural information statistics for different segments. (**d**) Visualization of protein sequence based on the attention distribution generated by NLSExplorer.

The parameter optimization of NLSExplorer

As our model NLSExplorer initially recommends segments and subsequently scores them, NLSExplorer exhibits dual capabilities. The cofactor is a pivotal parameter in the model, we first investigate the relationship between the model's recall and the cofactor. Supplemental Figure 3-a, b illustrates the relationship between the cofactor and recall across various minimum recommended segment lengths. On the hybrid datasets of INSP, as the cofactor varies from 0.05 to 0.6, the recall changes from the minimum of 0.63 to the maximum of 1. This signifies that as the recommendation magnitude increases, NLSExplorer suggests more segments, enhancing the likelihood of capturing segments of NLS.

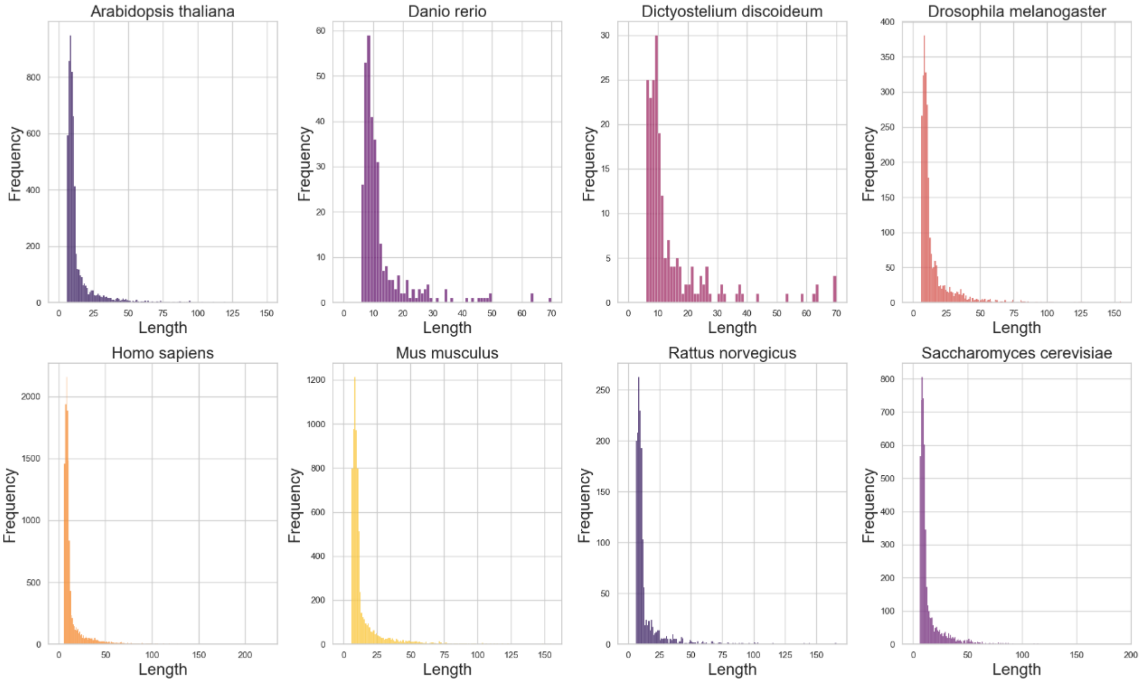


Supplemental Figure 3. The impact of various parameters on NLSExplorer's performance. (**a**) Variation in recall on the yeast datasets with different random work threshold. (**b**) Variation in recall on the hybrid datasets with different random work threshold. (**c**) The recall is influenced by the decision to filter single amnio acid and stretch recommended segment by random work. The least stretch length is 5.

It is noteworthy that the NLSExplorer model sometimes recommends a single residue. Considering no NLS is derived from single residue, we propose random walks[8, 9] based method to address it. It extends the single residue until a length threshold. Our rationale for this choice is that, although a single residue may not connect with surrounding residues to form a segment, it may signify the presence of an area around the single residue that is crucial for predicting NLS.

Supplemental Figure 3-a,b shows the impact of different random walk thresholds on the recall of NLSExplorer. Larger thresholds lead to an improvement in the recall but a decline in the accuracy. In Supplemental Figure 3-c we test the influence of filtering single amino acid and stretch on the model performance. Improvements in recall are observed when opting not to filter single amino acids or utilizing random selection within recommended segments, which indicates that there is a substantial likelihood of the presence of NLS around individually recommended residues. Compared to the strategy that filter out the single residue, we choose random walk-based strategy, which can not only prevent the model from neglecting individual points with high attention scores, but also aids in directing focus towards the vicinity surrounding these points.

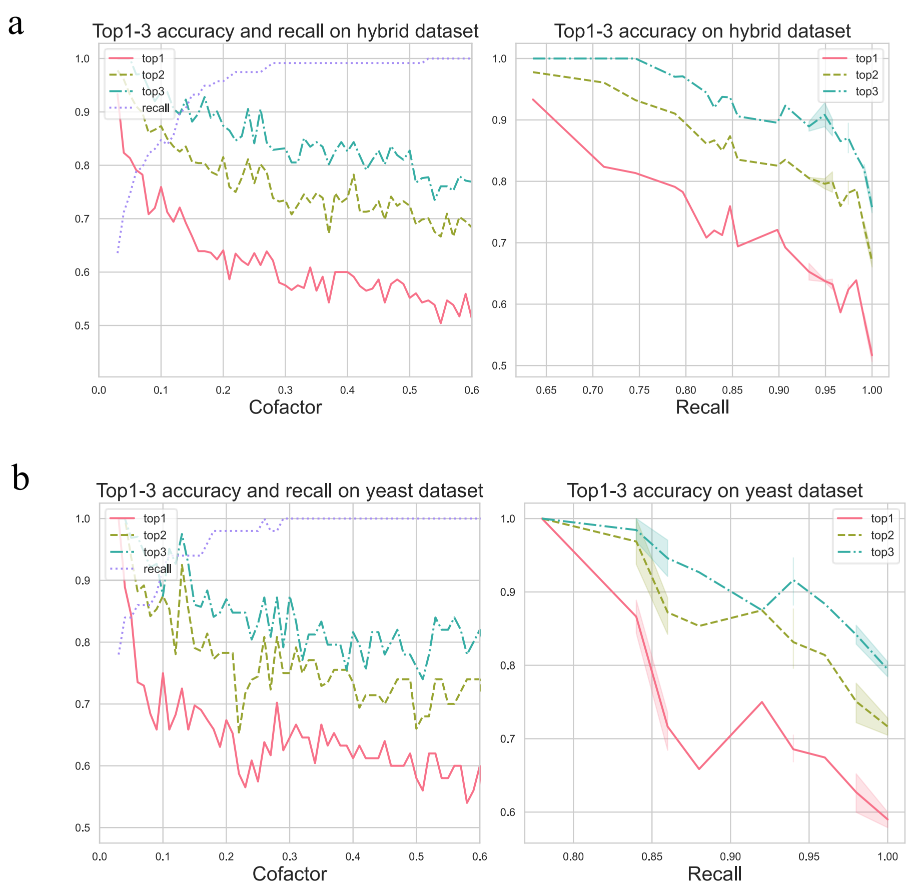
Length distribution of recommended peptides



Supplemental Figure 4. The distribution of peptide length recommended by NLSExplorer is in agreement with known NLS.

Supplemental Figure 4 illustrates the length distribution of all recommended segments by NLSExplorer. The recommended segment lengths predominantly fall below 100, with the majority falling within the range of 3 to 40, this aligns with the known distribution and patterns of NLS lengths[10], indicating that the top-3 recommended segments align well with the experimentally verified NLS segments.

Effect of cofactor variation on recall and accuracy curves

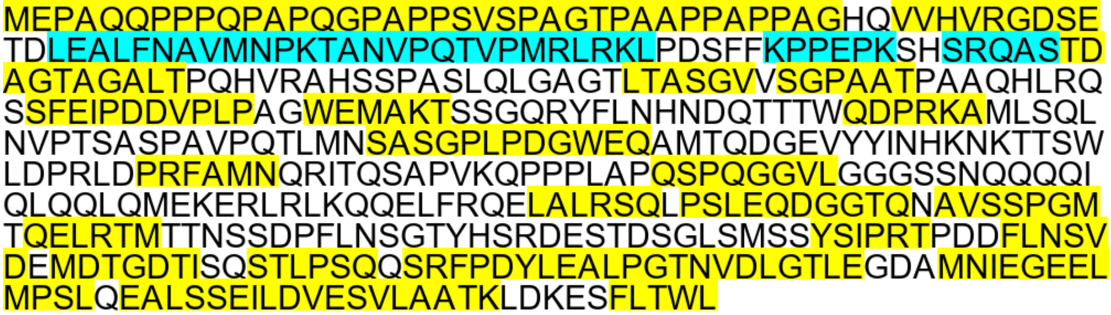


Supplemental Figure 5. In hybrid and yeast dataset Top1-3 accuracy and recall curve with the cofactor, top1-3 accuracy curve with the recall.

Supplemental Figure 5 illustrates the performance of the factor in relation to the accuracy of the recommended segments. For datasets of two distinct species, the overall accuracy exhibits a decline as the scale of exploration (cofactor) increases. Take hybrid dataset as an example, as the cofactor varies from 0.05 to 0.6, the accuracy changes with its maximum value of 0.93 to the minimum of 0.43. This indicates that as the recommendation magnitude increases, the model suggests more segments with some noise segments

We mainly take Top-1 performance on hybrid dataset to determine the parameter, because the hybrid dataset contains more species, which can provide comprehensive perspectives, and the performance tendency with different cofactors between these two datasets are highly similar. As the recommendation magnitude increases, the change in recall gradually slows down when approaching its zenith. Within the range of 0.05 to 0.25, the recall increases by 0.27. However, in the interval of 0.25 to 0.6, the increase in the recall is only 0.07. This result suggests that within the range between 0.25 to 0.6, the recommendation magnitude introduces mostly irrelevant noise into our model. By making tradeoff between the accuracy and recall, we determined the optimal recommendation magnitude for the NLSExplorer model on the INSP dataset to be 0.3 with the high F1 score and recall close to 1 that can help avoid missing potential NLS.

NLS prediction result for YAP1

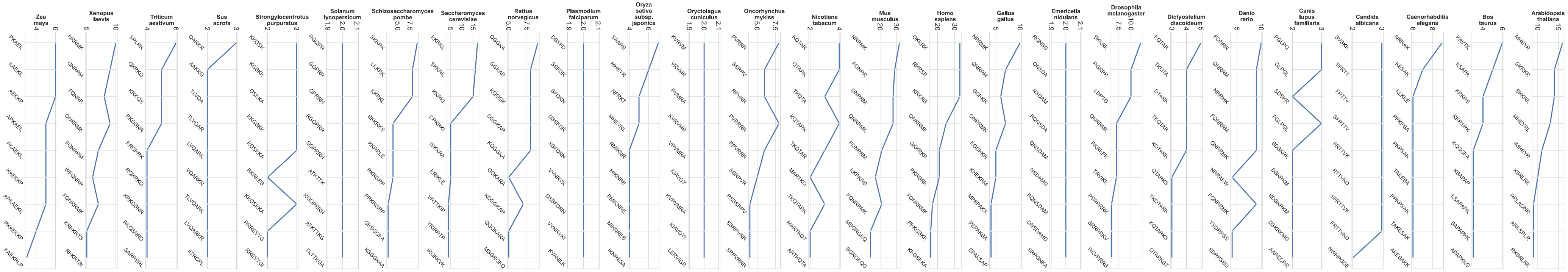


```
MEPAQQPPPQPAPQGPAPPSVSPAGTPAAPPAPPAGHQVVHVRGDSE
TDLEALFNAVMNPKTANVPQTVPMRLRKLPDSFFKPPEPKSHSRQASTD
AGTAGALTPQHVRAHSSPASLQLGAGTLTASGVVSGPAATPAAQHLRQ
SSFEIPDDVPLPAGWEMAKTSSGQRYFLNHNDQTTTWQDPRKAMLSQL
NVPTSASPAVPQTLMNSASGPLPDGWEQAMTQDGEVYYINHKNKTTSW
LDPRLDPRFAMNQRITQSAPVKQPPPLAPQSPQGGVLGGGSSNQQQQI
QLQQLQMEKERLRLKQQELFRQELALRSQLPSLEQDGGTQNAVSSPGM
TQELRTMTTNSSDPFLNSGTYHSRDESTDSGLSMSSYSIPRTPDDFLNSV
DEMDTGDTISQSTLPSQQSRFPDYLEALPGTNVDLGTLEGDAMNIEGEEL
MPSLQEALSSEILDVESVLAATKLDKESFLTWL
```

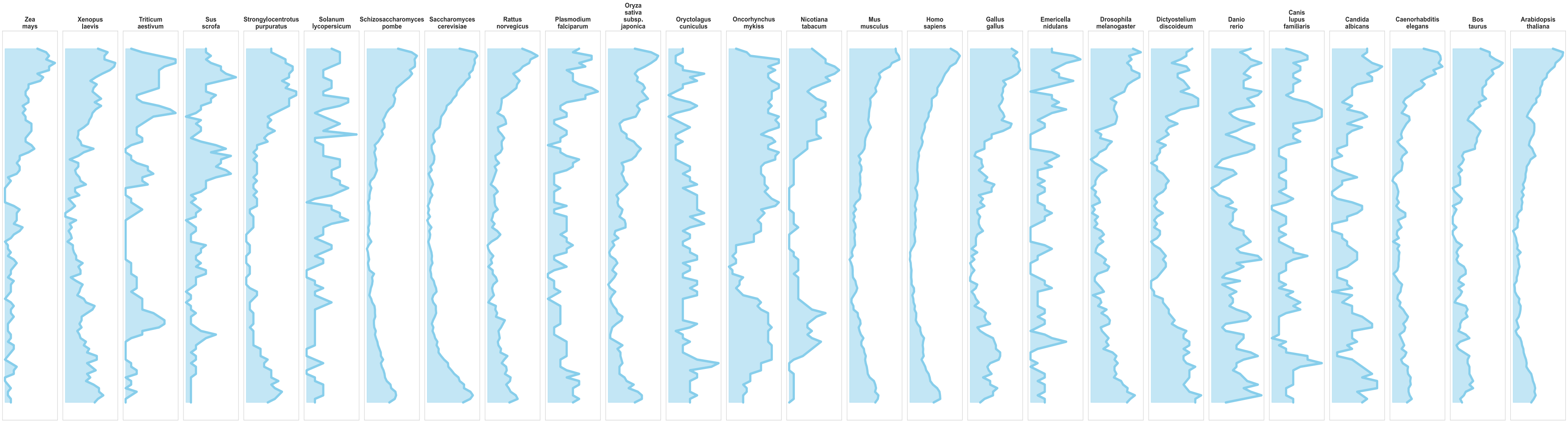
Supplemental Figure 6. The predicted result on YAP1 using NLSExplorer(cofactor 0.4, stretch length 5).

In YAP1, two distinct WW-NLS motifs are present: "WEMAKT...W" and "WEQ...TTSW". Supplemental Figure 5 illustrates our model's precise identification of these domains, depicted with a yellow background, at positions "WEQ" and "WEMAKT".

Pattern and positional characteristic



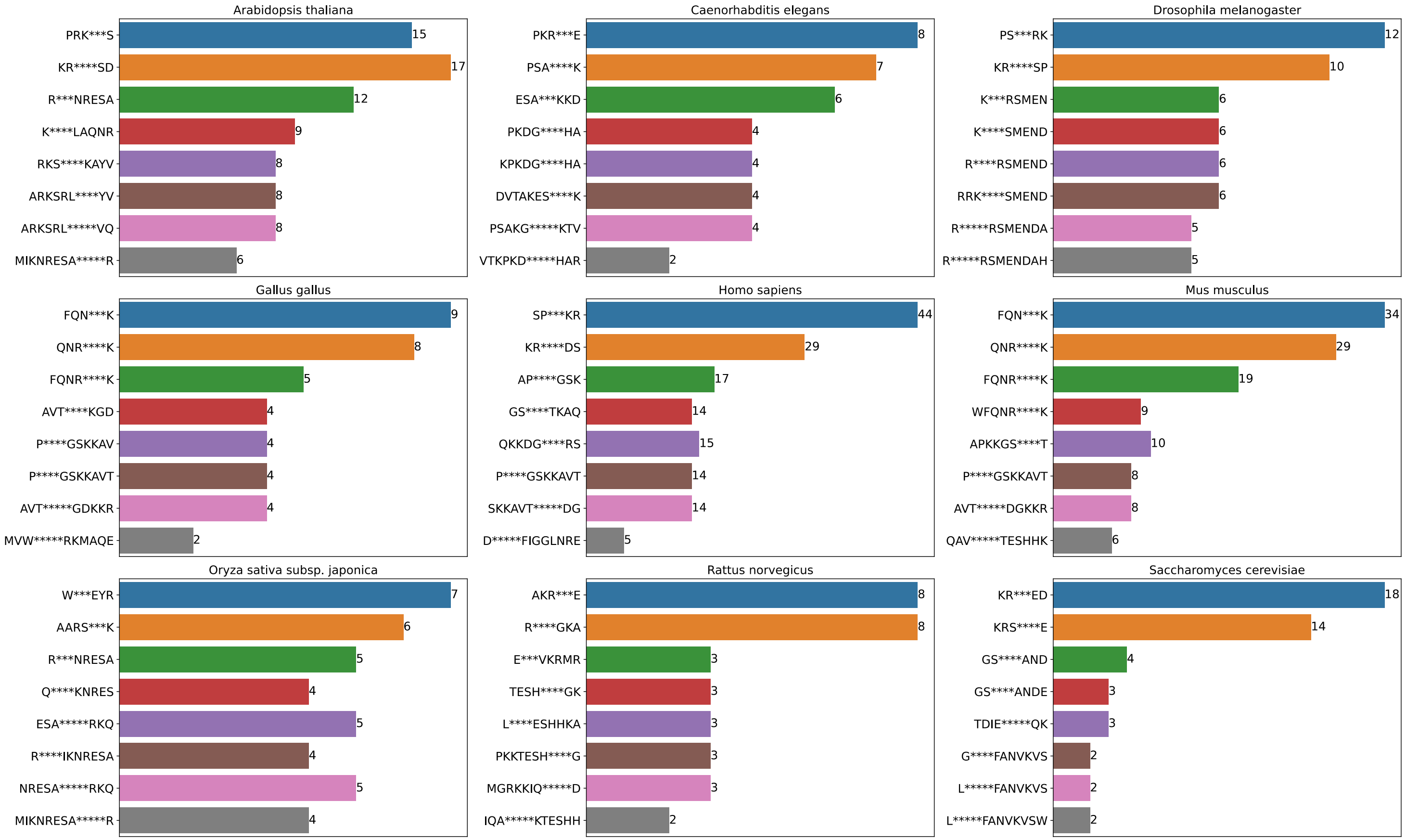
Supplemental Figure 7. The frequent pattern length ranges from 5-8 (diversity > 50%) in 5 species.



209

210 **Supplemental Figure 8**. The position pattern varies across different species.

211



Supplemental Figure 9. The non-contiguou patterns in 9 species, each with the largest number of proteins located in the nucleus among all species (pattern diversity > 50%).

Evaluation of the model and ablation experiment

In both the hybrid and yeast datasets, we collected protein structure information from the Alphafold online database, which has successfully predicted the structures of the majority of proteins in SwissProt. Since NLSExplorer has the capability to make predictions based on both sequence and structure information, in the ablation experiment, we intentionally exclude the structure information from our recommendation system while keeping other aspects of NLSExplorer unchanged. We then compared its performance against the NLSExplorer model to show the influence of structure information.

In order to fairly compare with other predictors, we select the same metrics as INSP dataset to ensure accurate assessment. NLSExplorer consists of two parts, namely the segment recommendation part based on the A2KA network and the segment ranking recommendation based on the EGNN network. Firstly, we assume that the total number of experimentally validated NLS fragments is *A0.* After the first part of the recommendation system processing, our recommendation system obtains *B0* segments, of which the number of non-repetitive hits of NLS sequences is *A1.* According to the definition of the recall score, the recall score of NLSExplorer is:

$$Recall = \frac{A1}{A0} \tag{S1}$$

For the segment ranking recommendation of NLSExplorer, we sort *B0* segments. Assuming a top-n sorting calculation method, a total of *A2* NLS segments are hit. According to the definition of accuracy, the accuracy equals the ratio of the predicted positives to the actual positives in the test set. The accuracy $accuracy_s$ of the segment ranking recommendation is:

$$accuracy_s = \frac{A2}{A1} \tag{S2}$$

Although this calculation method fully conforms to the definitions of recall and accuracy, it fails to reflect the actual performance of the entire prediction system. Consequently, the F1 score derived from this approach cannot be fairly compared with other predictors. The reason for this phenomenon lies in the fact that we calculate metrics for the two parts of the entire system independently. We need to replace the actual number of positives in the predicted set from *A1* to *A0* to obtain the prediction results based on the entire prediction system:

$$accuracy_t = \frac{A2}{A0} = \frac{A2}{A1} * \frac{A1}{A0} = accuracy_s * Recall \tag{S3}$$

where $accuracy_t$ refers to the accuracy of the entire prediction system.

Based on the above derivation, we obtain the final metric：

$$\begin{aligned} F1 = 2 * \frac{accuracy_t * Recall}{accuracy_t + Recall} &= 2 * \frac{accuracy_s * Recall^2}{accuracy_s * Recall + Recall} \\ &= 2 * \frac{accuracy_s * Recall}{accuracy_s + 1} \end{aligned} \tag{S4}$$

where the *F1* refers to the F1 score of the entire system.

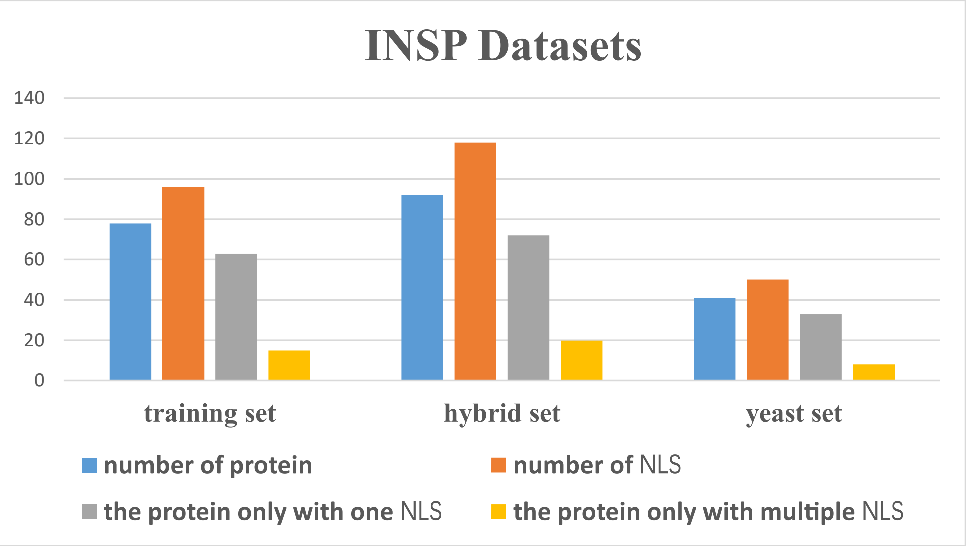
1. When optimizing the network architecture of the ranking recommendation system, we do so independently of the first part. At this stage, $accuracy_s$ enables us to conveniently optimize each part separately.
2. The accuracy for the entire system only requires multiplying $accuracy_s$ by the previously calculated recall factor. This process is convenient and easy to implement.

Based on the calculation of *F1* score, we can draw the following conclusion: the model's recall acts as a coefficient factor controlling the overall *F1* score. Compared to $accuracy_s$, improving the recall score not only ensures that the model explores and predicts more NLS sequences but also maximizes the performance of the model. This is why we adjust model parameters to improve the accuracy when maintaining a high recall score.

Supplemental Table 2. The ablation experiment on INSP dataset, "-ab" suffix denotes the prediction is made without structural information. the black font represents the best performance for the respective metric.

	Hybrid			Yeast		
	$accuracy_s$	recall	F1	$accuracy_s$	recall	F1
Top1	**0.64**	0.99	**0.77**	**0.65**	1	**0.79**
Top1-ab	0.59		0.73	0.56		0.72
Top2	**0.78**		**0.87**	**0.77**		**0.87**
Top2-ab	0.73		0.84	0.75		0.86
Top3	**0.84**		**0.90**	0.85		0.92
Top3-ab	0.80		0.88	**0.85**		**0.92**

We separately tested the performance of NLSExplorer on the INSP dataset with and without structural information (Supplemental Figure 10 and Table 2). It can be observed that, except for the top3 predictions on the yeast dataset, there is a significant enhancement in the predictions for top1, top2, and top3 on the hybrid and yeast datasets when structural information is added. Especially in the hybrid datasets with more species, the model with structural information performs better in all metrics. The ablation result indicates that the integration of structural information helps our model more accurately identify NLS.



Supplemental Figure 10. The statistic of INSP datasets.

A python-based package A2KA network

We developed a Python-based package that offers an extensible and comprehensive PyTorch support framework for the foundational architecture of A2KA. It supports the assembly of various key components of the A2KA network, which can be retrained as a unified module or independently embedded into existing workflows. A2KA is a customizable augmentation framework, which serves as a standalone framework while also accommodating diverse language model inputs and varied structural representations.

Reference

1. Kosugi, S., et al., *Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs.* Proceedings of the National Academy of Sciences, 2009. **106**(25): p. 10171-10176.
2. Nakai, K. and P. Horton, *PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.* Trends in biochemical sciences, 1999. **24**(1): p. 34-35.
3. Cokol, M., R. Nair, and B. Rost, *Finding nuclear localization signals.* EMBO reports, 2000. **1**(5): p. 411-415.
4. Nguyen Ba, A.N., et al., *NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction.* BMC bioinformatics, 2009. **10**: p. 1-11.
5. Mehdi, A.M., et al., *A probabilistic model of nuclear import of proteins.* Bioinformatics, 2011. **27**(9): p. 1239-1246.
6. Lin, J.-r. and J. Hu, *SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring.* PloS one, 2013. **8**(10): p. e76864.

7. Guo, Y., et al., *Discovering nuclear targeting signal sequence through protein language learning and multivariate analysis.* Analytical biochemistry, 2020. **591**: p. 113565.

8. Karlin, S. and J. McGregor, *Random walks.* Illinois Journal of Mathematics, 1959. **3**(1): p. 66-81.

9. Xia, F., et al., *Random walks: A review of algorithms and applications.* IEEE Transactions on Emerging Topics in Computational Intelligence, 2019. **4**(2): p. 95-107.

10. Nair, R., P. Carter, and B. Rost, *NLSdb: database of nuclear localization signals.* Nucleic acids research, 2003. **31**(1): p. 397-399.